Explaining Walking Duration to Transit Stations by Socio-Demographic Characteristics

Qi Chen[1], Go Higuchi[2], Feifan Xu[1], and Shichen Zhao[3]

[1]Doctor Candidate, Graduate School of Human-Environment Studies, Kyushu University, Japan

[2]Graduate Student, Graduate School of Human-Environment Studies, Kyushu University, Japan

[3]Professor, Dr. Eng. Faculty of Human-Environment Studies, Kyushu University, Japan

ABSTRACT

This study works on how socio-demographic characteristics can influence the walking duration to transit stations. The original dataset is the survey of rail transit trip in Fukuoka including 4254 records. The walking duration is explained by examining the probability of accepting a longer walking duration than the given threshold. Three thresholds of walking duration (5, 8 and 13 minutes) are estimated in this study. The valid features that can influence the walking duration are selected by using analysis of variance (ANOVA) in terms of each given threshold of walking duration. The selected features are used for estimating walking duration by using random forests model. The model is trained with 2/3 of the original dataset and tested by the rest 1/3. As a result, trip purposes play the most important role in determining the willingness at any threshold of walking duration, the feature of peak hour also shows importance; the other features have different performance at different thresholds. The accuracy of the results from random forests model is not enough to predict individual behavior, nevertheless, it can provide an acceptable accuracy in predicting the overall trend of the probability for a specific group of people. The results are expected available for planning the catchment area of transit stations or estimating the catchment area of existing transit stations.

INTRODUCTION

Mass Rapid Transit (MRT) has been one of the most important parts in public transit system in modern cities. How to shift more people from private car user to public transit passenger is always a key topic for both urban planning and management. There are many factors that have influences on the use of public transit, including the factors of land use functions, accessibility and socio-demographic characteristics, most of which are based on the aggregate data and are considered from the existing environmental features of transit stations. Some factors like money costs, acceptability of walking duration, travel purpose and personal socio-demographic characteristics also contribute to people's individual motivation to use public transit. For this study, it focuses on the individual rather than transit stations to examine the walking duration to transit stations and try to explain how it is affected by socio-demographic characteristics. The 800m (half-mile) walking distance has been widely accepted as a principal reference of the catchment area for the planning of Transit-Oriented Development (TOD). Planners and researchers also use transit catchment areas to make prediction of the transit ridership. This 800m walking distance is loosely obtained from the sampling survey that how far people are willing to walk to transit stations, but the same reasoning has been used to justify other transit catchment areas and even in different cities and countries. People with different socio-demographic characteristics generally have different preferences of walking duration. Even people with different socio-demographic characteristics have the same preference of walking duration, their walking distance is not the same because of different walking speeds. It can be assumed that potential public transit users have their own preference and tendency for a given threshold of walking duration. The willingness of using public transit will raise if the expected walking duration is less than the acceptable walking duration, otherwise, the willingness will decrease. The diversity in preference of walking duration can also be reflected at the travel survey data. Based on which, this study attempts to give explanations on how socio-demographic characteristics influence the acceptability of walking duration by using the case of Fukuoka subway.

LITERATURE REVIEW

This section reviews the literature on the issue of walking distance to transit stations, some problems still not well addressed are collated and summarized as well. The review is arranged in two parts, research object (dependent variable) and influencing factors (independent variables). For the research object, there is always a difficult point in obtaining the accuracy walking distance by questionnaire because of the discrepancy between perceived values and objective values (Badland, Hannah M and Schofield, Grant M and Schluter 2007; McCormack et al. 2008). Also, the observed walking distance is not the reflection of how long people are willing to spend on walking to transit stations, but a description of the distance between the departure and the station. Some studies selected a different perspective to explain the walking distance or duration by introducing a threshold of walking duration (Besser and Dannenberg 2005; McCormack et al. 2008). Indeed, using a threshold can decrease the discrepancy between observation and reality in some extent, but this disposal also brought some new problems in. For instance, it can lead to a great loss of information in the raw data, on the other hand, how to decide the threshold is also a key problem. To date, there have been amount of studies working on the relationship between walking distance and social-demographic characteristics, however, some studies also argued that an individual has a limited amount of time spending on traveling during a day, people tend to accept further walking distance as the speed of travel increases (Marchetti 1994; Larsen and El-Geneidy 2010). That means, people with the same social-demographic characteristics may have the similar acceptability of walking duration, but they generally have different acceptability of walking distance due to different travel speed. In the influencing factors, socio-demographic characteristics are generally thought to be the key that can affect walking distance (Besser and Dannenberg 2005; Weinstein Agrawal et al. 2008; Krygsman et al. 2004; Yang and Diez-Roux 2012; Daniels and Mulley 2013; Guerra et al. 2012), however, there are few studies having clearly verified the relationship between socio-demographic characteristics and people's activities, even there is a study suggesting the walking distance should not be viewed as a function of socio-demographic characteristics (Krygsman et al. 2004). Several studies have confirmed the role of travel purposes in determining walking distance, the commute trip showed particularity from the other purposes, people with the purpose of commute tend to accept a longer walking distance to go to transit stations (Larsen and El-Geneidy 2010). However, the definite relationship between trip purposes and walking distance is still unclear. The same situation for other categories of factors, such as the factors of transportation environment, land use, and willingness (Guerra et al. 2012; Krygsman et al. 2004; Weinstein Agrawal et al. 2008). The only thing that has been confirmed to date is that the walking distance should be influenced by some factors, such as socio-demographic characteristics, trip chaining behavior, or built-environment, but the problem is how to be influenced. In summary, perhaps because of the problems in either the research object or influencing factors, most of the existing studies did not find the significant relevance between the dependent variable (research object, either walking distance or walking duration) and independent variables (influencing factors). The studies working on the qualitative description for the distribution of walking distance accounted for the majority, although some of the existing studies attempted regression model on this issue, the desired results were not obtained. To avoid such problems mentioned above, this study shifts the research object to the threshold of walking duration. Additionally, instead of finding the statistical relevance between dependent and independent variables, this study will introduce the approach of machine learning into this issue, and try to explain the walking duration by using probability.

DATA AND METHODS

Data for Fukuoka

For the study case of Fukuoka, it is the sixth largest city in Japan which has the largest population in Kyushu Island of Japan (more than 1.5 million). The data source in this study is the Northern Kyushu Area Person Trip, which is conducted about every 12 years, the latest available data is the 4th survey by the year of 2005, and the 5th survey is already in preparation from September 2017. Figure 1 is the research area and the distribution of rail transit stations. By the year of 2005, which is the year that the 4th Northern Kyushu Area Person Trip Survey was conducted, there are more than 70 rail transit stations located within the city area of Fukuoka, of which the amount of JR Kyushu stations is 27, Fukuoka Subway is 35, and West Japan Railway is 16. Till now, some new rail transit lines and stations are still under planning and construction. According to the “Survey on the current state of public transportation” published by the Ministry of Land, Infrastructure, Transport and Tourism of Japan, the rail transit system in Fukuoka carries a daily average of more than 0.4 million passengers by 2015 accounting for more than 20% in total motorized travel. Although the rail transit system of Fukuoka is still not a large-scale one at now, it has been playing a crucial role in people’s daily travel.

Data for trip

The main purpose of person trip survey is to master the travel trends, thus making a better living environment and providing support for traffic planning. The original data covered the range of all the main cities in Northern Kyushu Area, which has more than 483,000 records of trip chaining behavior. The available data in this study mainly included trip chaining behavior and socio-demographic characteristics, as shown in the table 1. To analyze the walking duration between departure to transit station in Fukuoka, the first step is to extract the valid records of rail transit trip within the city area of Fukuoka from the enormous data of more than 480,000 records. The procedure of extracting the valid data for this study is divided into 3 steps. Firstly, extracting all the person trip data that surveyed within the city area of Fukuoka; secondly, selecting the trip chaining behavior which contains the rail transit mode; thirdly, filtering the invalid data that without null value or abnormal value. The procedure of data cleaning is shown in figure 2. For analyzing the relationship between walking duration and socio-demographic characteristics, only the trip chaining behavior with unique individual identifier can be reserved at last. Finally, the valid dataset is reduced to a size of 4254 trips. Figure 3 shows the age distribution for walking trips to transit stations based on the finally valid dataset. The dataset is subject to normal distribution in general, and it has a good coverage for almost all the ages. Figure 4 shows the distribution of real walking duration to transit stations, it is generally subject to normal distribution with the mean 8.32, and the standard deviation 2, merely there are several peak values at the time of 5 multiples. It is speculated that the peak values may be caused by deviation occurred in the investigation. Because it seems that people are inclined to reply a looser answer when they are being asked some questions about details. This inclination will count some of the real walking duration that is near to 5 multiples as the 5 multiples and finally expressed in the result of investigation.

Methods

For the issue of this study, the dependent object is the walking duration to transit stations which is a continuous variable, while the independent objects are multi-categorical variables. An issue like this form is not judged to be suitable for multiple linear regression models. Even the multiple linear regression is the most accepted model for analyzing the relevance of various factors and is also used to estimate the walking duration, the coefficient obtained from the regression model cannot work well on the prediction of walking duration yet (Krygsman et al. 2004). Nevertheless, because what we are concerned with is the threshold of walking duration to transit stations, this issue can be converted into a binary choice problem, which can be also viewed as a classification problem. Giving a specific threshold of walking duration to go to a transit station, people can make the decision of whether choosing the transportation mode of rail transit, and this decision should also be affected by people's socio-demographic characteristics. That is, if the walking duration of someone is longer than a given threshold of walking duration, it can be considered that this given threshold of walking duration is an acceptable one, and he or she may have a relatively high probability of choosing to use rail transit. This willingness or tendency of using rail transit can be expressed as the probability of walking longer than the given threshold or not, which is obtained from the survey data. This study will examine the relevance of socio-demographic characteristics and this probability. This probability is used as the dependent variable in this study, given as Equation 1.

$$D = P^i(t^i > T) \tag{1}$$

Where:

D is the dependent variable,

P^i is the probability of walking longer than the given threshold for travel individual i,

t^i is the expected walking duration for travel individual i.

The essence of such kind of binary choice problems is also can be viewed as a classification problem, the models of decision tree, Bayesian, support vector machine (SVM), logistic regression, and neural network are widely adopted for analyzing this kind of problem. In this study, limited by the volume of samples and features, also the unknown for the inner relationship among the features, the decision tree model seems to be a good choice because of the good generalization for different forms of data. Furthermore, to avoid the structure of the tree being too complicated, and to improve the robustness of the model, this study will use an improved model of decision tree, which is the random forest model, to train and test the samples. Random forests model is an ensemble learning method mainly for classification and prediction. It is an extension and improvement for the decision tree model, that operate by constructing a multitude of decision trees and randomly selecting the features at training time (Ho 1995; Ho 1998). Another point, for improving the efficiency and accuracy of the model, also reducing the number of invalid branches in the model, it is necessary to filter the invalid features before estimating the model. Briefly, the general process of this study can be summarized as follows. Data is extracted on the Fukuoka city-wide, including 4254 samples, and 6 main categories of features. The features with importance in explaining the acceptable walking duration are selected by using analysis of variance (ANOVA). Then random forest model for each threshold of walking duration is estimated. Finally, the result obtained from the random forest model can be used for predicting the probability of people walking to transit station within a specific threshold of walking duration.

FEATURE SELECTION

The procedure of feature selection includes two parts, the first is the disposal of features with low variance, and the second is the univariate feature selection. The first one is used for filtering the features accounting for very little in the total samples, because a too small quantity cannot present significant influence on the result. In this study, the multi-categorical features in the original dataset are reclassified into some larger groups, thus dealing with the features with low variance. The second step is to select the valid features that indeed have relevance with the independent variable, this step is processed based on statistical tests. In this study, the Analysis of Variance (ANOVA) is used for estimating the importance of each feature.

Reclassifying features

The feature of trip purpose has 15 subcategories in the original dataset, it is reclassified into commuting to work, commuting to school, official business, private purpose (such as shopping, entertainment), and going home, 5 categories. The feature of occupation is reclassified from 14 subcategories into 5 categories as well, they are service, technology, administration, student, and the other. Table 2 reports some of the statistical description for each feature. Overall, the average walking duration to transit stations is 8.32 minutes, more than 75% are less than 10 minutes, most people walk to transit stations costing 5 10 minutes. In detail, there are some significant differences in the statistical description for each feature, the differences are summarized as follows.

1. The walking duration in peak hour is longer than off peak hour.
2. Young and old people are inclined to spend less time on walking to transit stations.
3. Trips with purposes of official business and going home tend to take shorter time in walking to transit stations.
4. Walking trips for commuting to work are significantly longer than walking trips for other purposes.

Univariate feature selection

According to the achievements from previous studies, the walking distance to transit stations ranged generally from 400 meters to 1000 meters in terms of different city types, travel habits, and even the needs of research purpose (Guerra et al. 2012; Murray et al. 1998; O'Sullivan and Morrall 1996; Keijer and Rietveld 2000; Zhao et al. 2003; Alshalalfah and Shalaby 2007). If converting this walking distance into walking duration by calculating the walking speed of 4.8 km/h, which has been argued in the existing study (Bohannon 1997), the walking duration would range from 5 minutes to 13 minutes. Back to this study, more than 40% of the walking duration are less than 5 minutes, and the walking duration within 13 minutes covers about 85% of the total. The

5 minutes' range can be viewed as the acceptable walking duration of the main users of transit stations, while a walking duration longer than 13 minutes cannot be accepted by most of the rail transit users. In addition, the mean walking duration in this study is about 8 minutes, therefore, the three more representative thresholds of 5, 8, 13 minutes are picked as the typical threshold for estimating how socio-demographic characteristics can influence the acceptable walking duration. The features with significance in explaining the walking duration (the p-value in ANOVA less than 0.05, which means this feature relevant with the dependent variable at the confidence level of 95%) are picked out and listed in Table 3. For the threshold of 5 minutes, the features of trip purposes and peak hour play the most important role in determining the willingness of walking duration. Situations are changed a little in the case of 8 minutes, the importance of age and sex raised in some extent, while the features of trip purposes change little from that of 5 minutes. For the threshold of 13 minutes, the feature importance shows a big difference from 5 and 8 minutes. This result of feature selection is also consistent with known characteristic of rail transit use. Most of the walking duration is distributed around the average walking duration 8 minutes, it can be considered that walking duration between 5 and 13 minutes is sensitive to socio-demographic characteristics. The walking duration more than 13 minutes is not accepted by most people even if they have different socio-demographic characteristics, and the threshold less than 5 minutes is also not sensitive to socio-demographic characteristics since the 5 minutes threshold is a generally acceptable one for most passengers.

ESTIMATION OF FEATURES

The valid features (Table 3) are used in the random forest model to estimate the probability of accepting the given threshold of walking duration in terms of people with different socio-demographic characteristics. For estimating the model, the dataset is divided into two parts, 67% of the samples are used for fitting the model and obtaining the coefficients, the rest 33% are used for testing the ability of prediction. The probabilities of walking less than the given threshold of walking duration can be estimated by the decision tree obtained by the random forest model, the prediction process based on the decision tree from the random forest model is presented in Figure

5. The prediction in random forest model is not obtained from the only one decision tree but the multitude of decision trees constructed by random selection of features and samples. Finally, the dependent variable at the given threshold of walking duration to transit stations is calculated by equation 1 based on the mean prediction. Figure 6 is the trend line of probabilities of walking less than the given threshold. The trend line of the surveyed values based on the test set is calculated by the mean probability of a group people who have close predicted values. The trend line is drawn by a descending order of predicted values. From the comparison of predicted values and surveyed values, it can be known that the prediction at the threshold of 5 minutes has the best fitness, and the prediction for the threshold of 13 minutes is also slightly good, while it's not so good in the case of 8 minutes. In fact, if checking the prediction for individuals, the accuracy of this model is still not enough to explain the individual behavior. But trend line shown in Figure 6 partly proved that this result can be used for reflecting the behavior of a group of people at a given threshold of walking duration.

DISCUSSION AND CONCLUSION

This study described and analyzed more than 4200 samples for the trips of rail transit. Three thresholds of walking duration are examined. From the result of Table 3, trip purpose is the most important factor in determining the willingness of walking duration for all the 3 thresholds. The feature of peak hour is also significant in explaining the walking duration. People can accept a longer time to go to stations at peak hours, while people with private purpose or on the way going home are not willing to choose the stations far away. For the details of each threshold, the unemployed people and the elderly people tend to walk less than 5 minutes for accessing transit stations; it is a little difficult for females and the people whose age is between 25 and 44 to accept a walking duration more than 8 minutes; the people whose age is between 45 to 64 can accept the walking duration of more than 13 minutes more easily than the others. But what summarized above is only the description for the distribution of walking duration in terms of each feature. Even knowing these, it is difficult to make use of them, just as the achievement obtained by previous studies. In fact, the data in this study is obtained from a factual investigation but not a willingness survey. Once people have made a decision of walking to transit stations, the real explanation of walking duration is the location of departure and the walking speed. It is hard to say whether the social-demographic characteristics affected the walking durations. Therefore, for this kind of study, there should have a basic assumption that passengers with different walking durations have a uniform travel rate. If a passenger chose to walk to a transit station, it means this passenger can accept the walking duration from his/her departure to that transit station. The distribution of walking durations for each group of people with a specific feature can be considered as the reflection of the acceptability for that group of people. Perhaps due to the reasons of discrepancy in investigation and the basic assumption, few studies can explain the relevant between walking duration and social-demographic characteristics quantitatively. According to the achievement of previous studies and this study, a conclusion can be drawn that if examining the walking duration as a continuous variable, there is no linear relation between walking duration and social-demographic characteristics. Therefore, this study tried to explain the walking duration by examining some specific threshold. As the results, the model of 5 minutes' threshold has a better explanatory power, the model of 13 minutes' threshold is a little weak, and the model of 8 minutes' threshold is not good. Here are some possible reasons for explaining the results. The selected thresholds of walking duration 5, 8, 13 minutes represent the lower boundary, mean value, and upper boundary of the major distribution of walking durations respectively. The acceptable walking durations range from 5 to 13 minutes. It can be inferred that the group of people who are not willing to accept the threshold of 5 minutes and who are willing to accept the threshold of 13 minutes may have significant features. However, as the mean value of walking duration, it can be thought that most of the walking durations are distributed around 8 minutes, for passengers there may be some randomness in making the decision of whether walking to transit stations. For the other threshold values near the mean value of 8 minutes, it also can be inferred that people may have some ambiguity in choosing whether to accept or not. This explanation can also be partly confirmed by the result of valid features selection. There are 5 identical features in both thresholds of 5 and 8 minutes, which means people with those features are not sensitive to the threshold from 5 to 8 minutes. Overall, the results of the random forest model perform not well on individual predictions. But for the average probability of a group of people, it shows a good predictive ability at the threshold of 5 minutes. In the next stage of this research, we plan to apply the results on predicting the willingness at a specific threshold of walking duration for a group of people. Also, the results can be used in planning the catchment area of transit stations or estimating the catchment area of existing transit stations.

ACKNOWLEDGMENTS

My deepest gratitude goes first to my supervisor, Professor Zhao, for his constant encouragement and guidance. Also, I want to thank Mr. Higuchi and Miss. Xu for their great help in data collection.

REFERENCES

Alshalalfah, B. W. and Shalaby, a. S. (2007). “Case Study: Relationship of Walk Access Distance to Transit with Service, Travel, and Personal Characteristics.” *Journal of Urban Planning and Development*, 133(2), 114–118.

Badland, Hannah M and Schofield, Grant M and Schluter, P. J. (2007). “Objectively measured commute distance: associations with actual travel modes and perceptions to place of work or study in Auckland, New Zealand.” *Journal of physical activity and health*, 4(1), 80–86.

Besser, L. M. and Dannenberg, A. L. (2005). “Walking to Public Transit.” *American Journal of Preventative Medicine*, 29(4), 273–280.

Bohannon, R. W. (1997). “Comfortable and maximum walking speed of adults aged 20-79 years: Reference values and determinants.” *Age and Ageing*, 26(1), 15–19.

Daniels, R. and Mulley, C. (2013). “Explaining walking distance to public transport: The dominance of public transport supply.” *Journal of Transport and Land Use*, 6(2), 5.

Guerra, E., Cervero, R., and Tischler, D. (2012). “Half-Mile Circle.” *Transportation Research Record: Journal of the Transportation Research Board*, 2276(2276), 101–109.

Ho, T. K. (1995). “Random decision forests.” *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1, 278–282, <http://ieeexplore.ieee.org/document/598994/>.

Ho, T. K. (1998). “The Random Subspace Method for Constructing Decision Forest.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844.

Keijer, M. J. N. and Rietveld, P. (2000). “How do people get to the railway station? The dutch experience.

Krygsman, S., Dijst, M., and Arentze, T. (2004). “Multimodal public transport: An analysis of travel time elements and the interconnectivity ratio.” *Transport Policy*, 11(3), 265–275.

Larsen, J. and El-Geneidy, A. (2010). “Beyond the quarter mile: Re-examining travel distances by active transportation.” *Canadian Journal of Urban Research*, 19(1 SUPPL.), 70–88.

Marchetti, C. (1994). “Anthropological invariants in travel behavior.” *Technological Forecasting* *and Social Change*, 47(1), 75–88.

McCormack, G. R., Cerin, E., Leslie, E., Du Toit, L., and Owen, N. (2008). “Objective versus perceived walking distances to destinations: correspondence and predictive validity.” *Environment and behavior*, 40(3), 401–425.

Murray, A. T., Davis, R., Stimson, R. J., and Ferreira, L. (1998). “Public Transportation Access.” *Transportation Research Part D: Transport and Environment*, 3(5).

O’Sullivan, S. and Morrall, J. (1996). “Walking Distances to and from Light-Rail Transit Stations.” *Transportation Research Record*, 1538(1), 19–26.

Weinstein Agrawal, A., Schlossberg, M., and Irvin, K. (2008). “How Far, by Which Route and Why? A Spatial Analysis of Pedestrian Preference.” *Journal of Urban Design*, 13(1), 81–98.

Yang, Y. and Diez-Roux, A. V. (2012). “Walking distance by trip purpose and population subgroups.” *American Journal of Preventive Medicine*, 43(1), 11–19.

Zhao, F., Chow, L.-F., Li, M.-T., Ubaka, I., and Gan, A. (2003). “Forecasting Transit Walk Accessibility: Regression Model Alternative to Buffer Method.” *Transportation Research Record*, 1835(1), 34–41.

List of Tables

TABLE 1. Available Data Contents

Category	Feature
Trip chaining behavior	Departure location
	Departure time
	Destination location
	Arrival time
	Transport modes
	Time spent for each mode
	Location of bus stop or rail transit station
Socio-demographic attributes	Age
	Sex
	Occupation
	Trip purpose
	Vehicle/License holding
	Address

TABLE 2. Statistical Description of Walking Duration for Each Feature

Features	Categories	Count	Percentage	Mean	Std	10th	25th	50th	75th	90th
	Total	4254	-	8.32	4.63	3	5	8	10	15
Sex	male	2257	53.1%	8.41	4.63	3	5	8	10	15
	female	1996	46.90%	8.22	4.63	3	5	7	10	15
Peak hour	peak	2976	70.00%	8.51	4.66	3	5	8	10	15
	Off peak	1277	30.00%	7.88	4.52	3	5	7	10	15
Age	5-24	543	12.80%	8.18	4.7	3	5	7	10	15
	25-44	1992	46.80%	8.36	4.42	3	5	8	10	15
	45-64	1407	33.10%	8.43	4.83	3	5	8	10	15
	65-	311	7.30%	7.78	4.81	3	5	7	10	15
Occupation	service	1256	29.50%	8.32	4.57	3	5	8	10	15
	tech	721	17.00%	8.58	4.89	3	5	8	10	15
	office	1076	25.30%	8.44	4.48	4	5	8	10	15
	student	325	7.60%	8.12	4.73	3	5	7	10	15
	null	875	20.60%	8.03	4.62	3	5	7	10	15
Purpose	commuting to work	2697	63.40%	8.69	4.51	4	5	9	10	15
	commuting to school	287	6.70%	8.19	4.6	3	5	8	10	15
	official business	153	3.60%	7.1	5.05	2	3	5	10	15
	private purpose	789	18.60%	7.82	4.78	3	5	7	10	15
	going home	327	7.70%	7.17	4.67	2	5	5	10	15

TABLE 3. Valid Features and the Effect at Each Threshold

5 min threshold		8 min threshold		13 min threshold	
Features	Effect*	Features	Effect*	Features	Effect*
Age over 65	L	Female	M	Age 45-64	M
Peak hour	M	Age 25-44	M	Peak hour	M
O_Null	L	Peak hour	M	P_commuting to work	M
P_commuting to work	M	P_commuting to work	M	P_ private purpose	L
P_official business	L	P_official business	L		
P_private purpose	L	P_ private purpose	L		
P_going home	L	P_ going home	L		

***Note:** L means the individual with this feature tend to spend less time than the given threshold of walking duration on accessing transit stations, while M is the opposite meaning.

List of Figures

Fig. 1. Distribution of Transit Stations

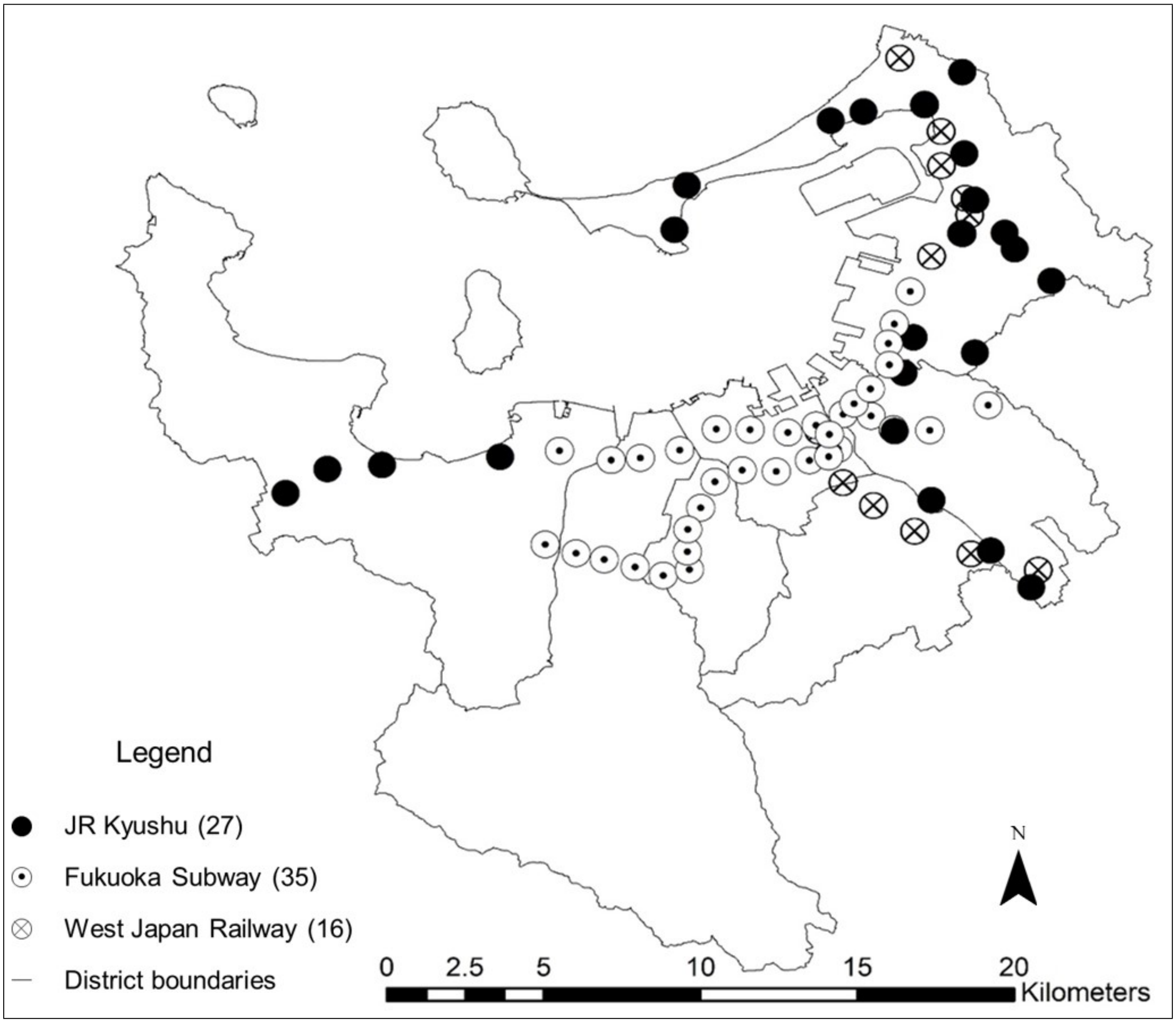


Fig. 2. Process of Data Cleaning

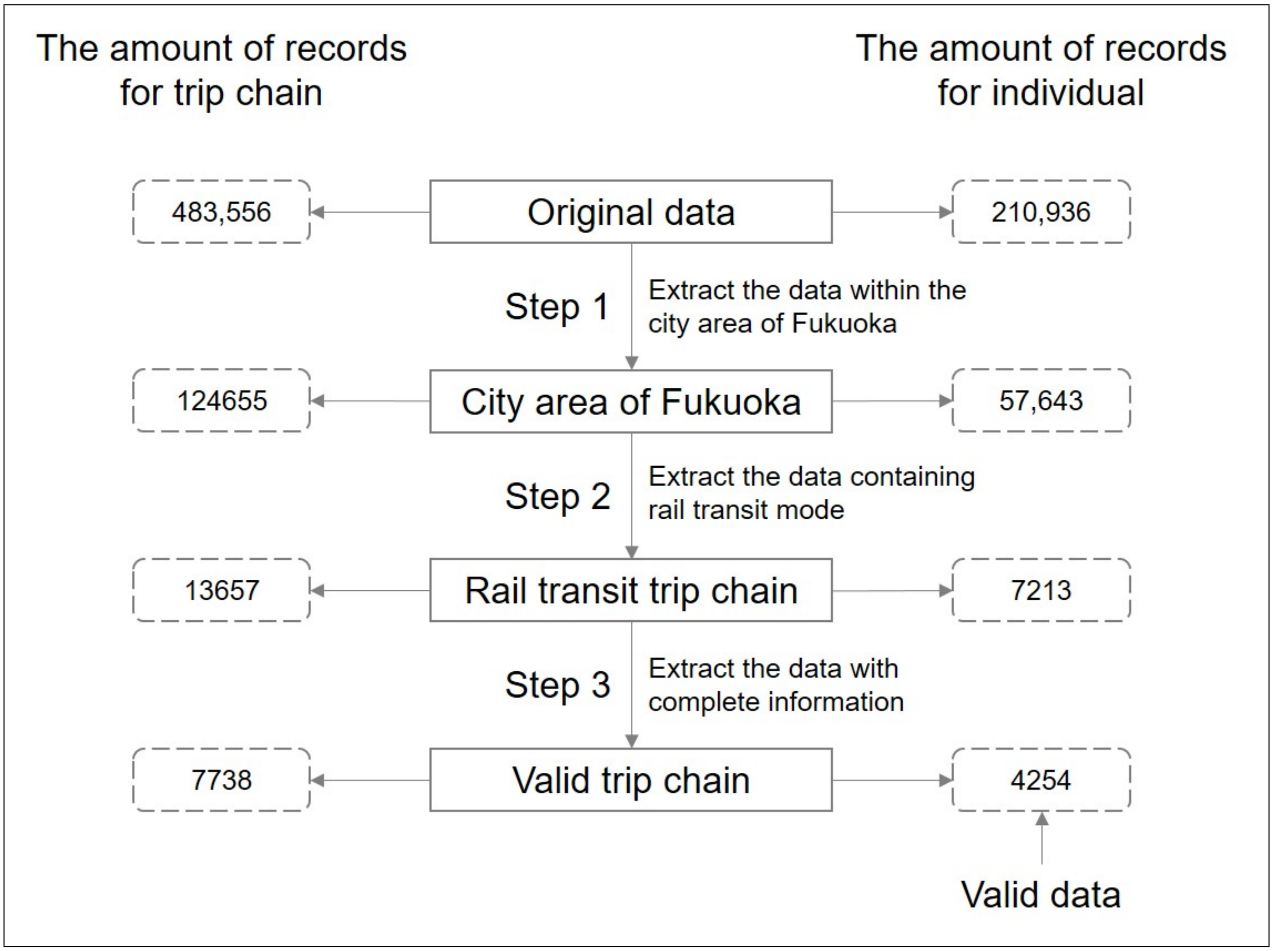

The amount of records for trip chain
The amount of records for individual
483,556
Original data
210,936
Step 1
Extract the data within the city area of Fukuoka
124655
City area of Fukuoka
57,643
Step 2
Extract the data containing rail transit mode
13657
Rail transit trip chain
7213
Step 3
Extract the data with complete information
7738
Valid trip chain
4254
Valid data

Fig. 3. The Age Distribution of Passengers

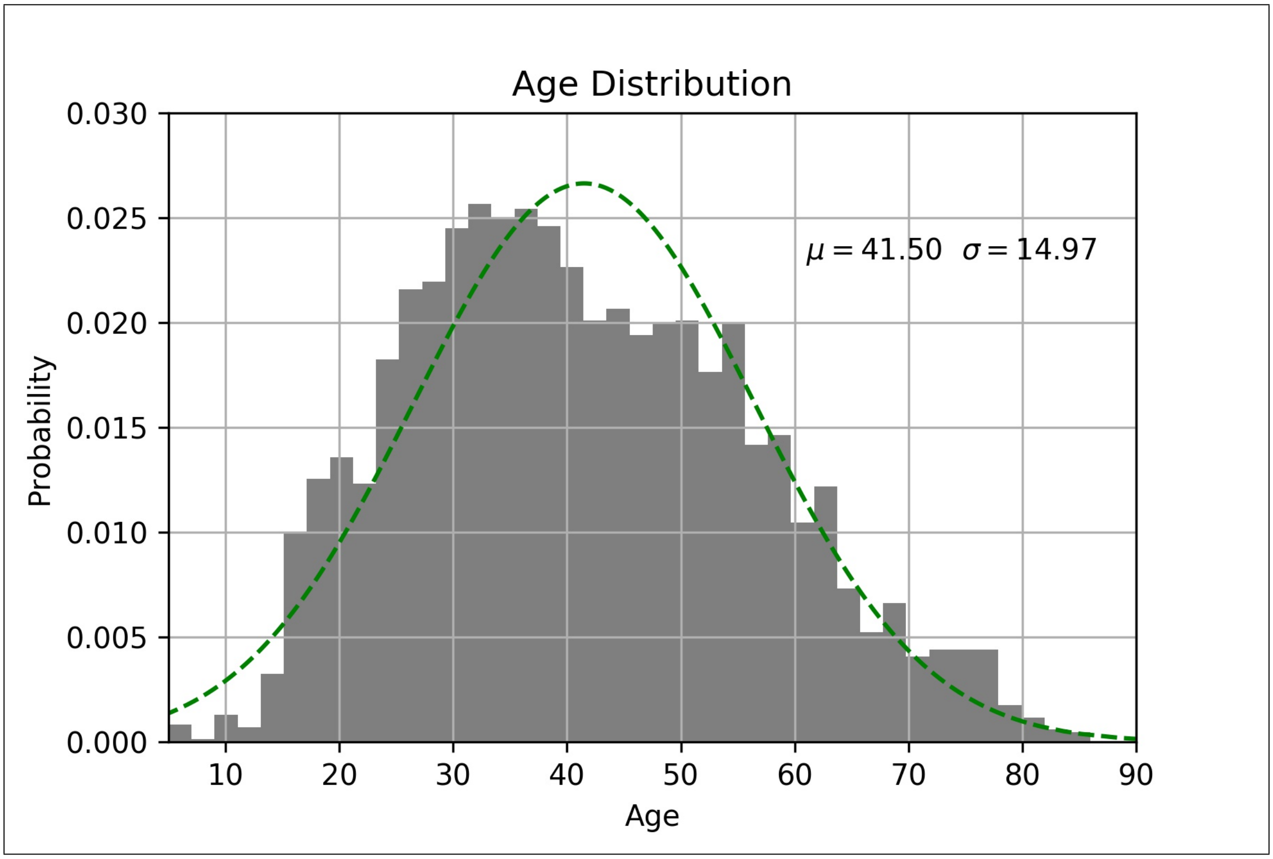


Fig. 4. Distribution of Walking Duration

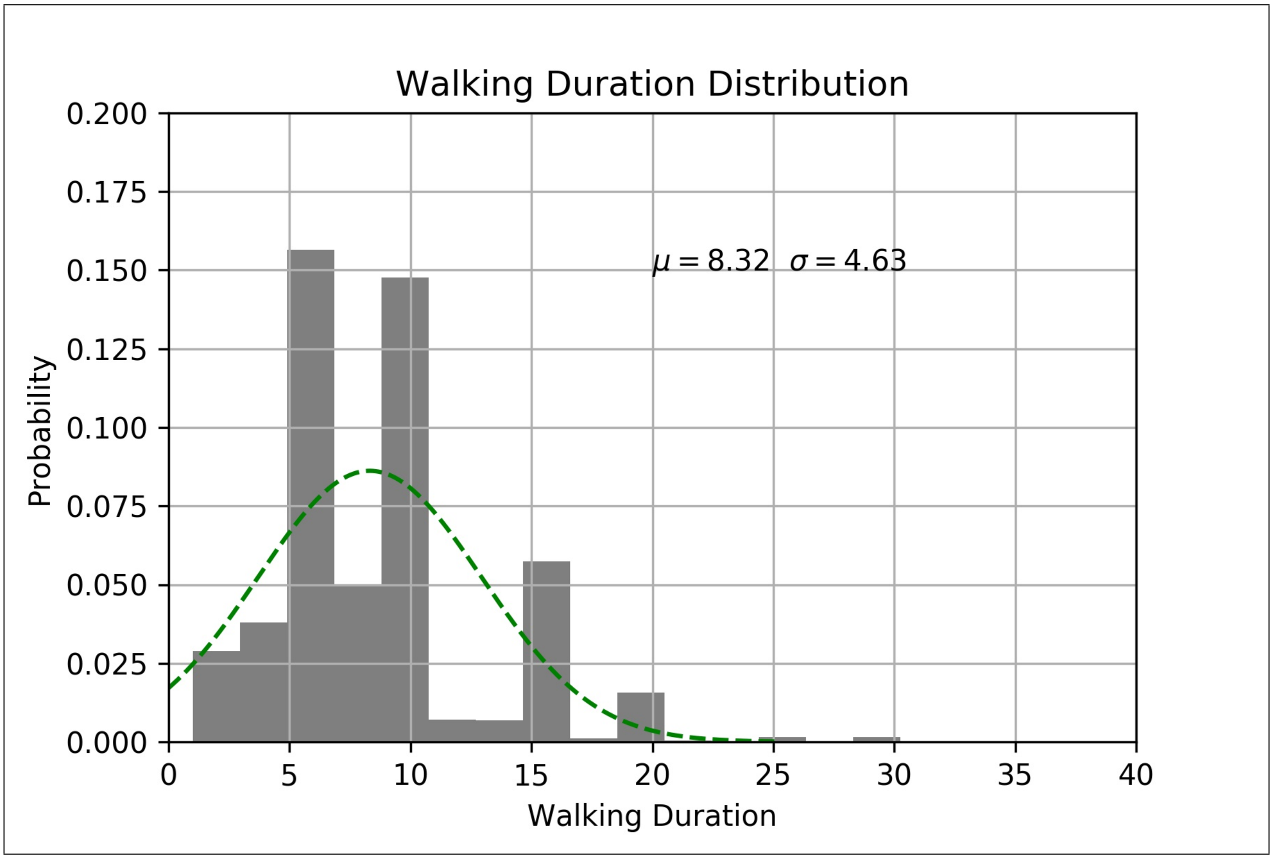


Fig. 5. Prediction Process in Random Forests Model

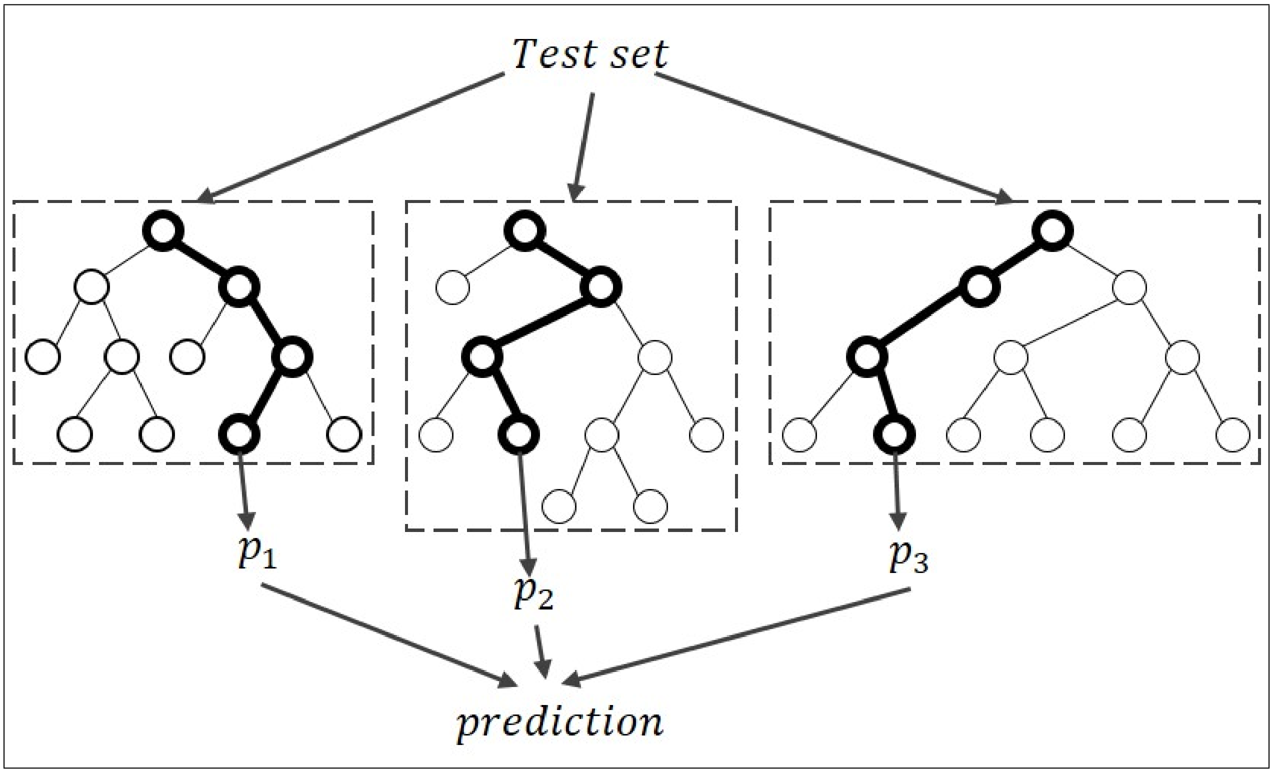
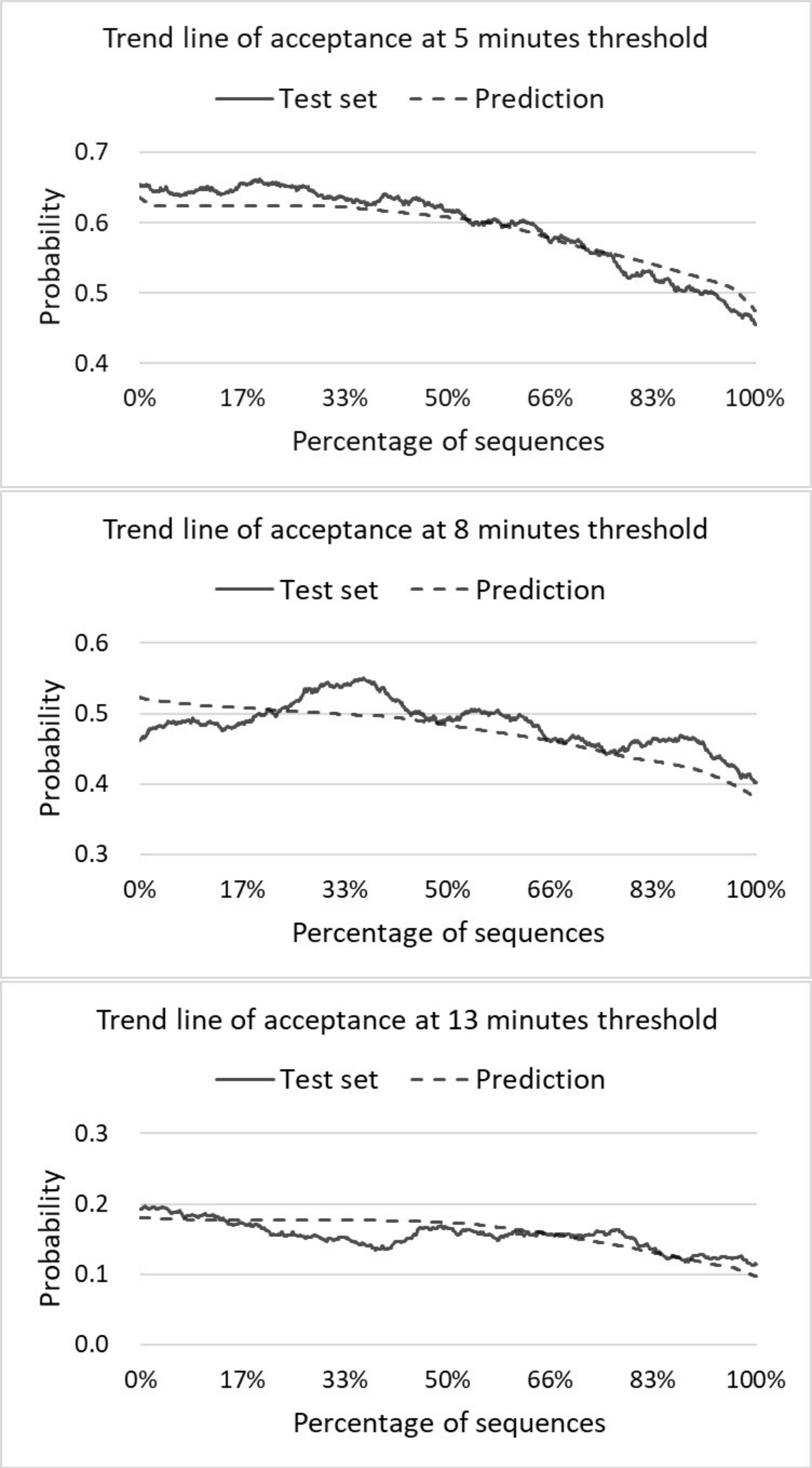


Fig. 6. Trend Line of Prediction and Test Set


Trend line of acceptance at 5 minutes threshold
Test set
Prediction
Probability
Percentage of sequences
Trend line of acceptance at 8 minutes threshold
Test set
Prediction
Probability
Percentage of sequences
Trend line of acceptance at 13 minutes threshold
Test set
Prediction
Probability
Percentage of sequences