Relationships between individual characteristics and real walking durations from departures to transit stations

Qi Chen[1], Siting Chen[2], 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

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

ABSTRACT

It is widely accepted that walking durations from departures to transit stations should relate to passengers' individual characteristics. However, since the walking duration is the reflection of the distribution of departure points, it may have no simple linear relationship with individual characteristics. Instead of examining the relationship between walking durations and individual characteristics directly, this study examines the feature distribution of individual characteristics over/under several thresholds of walking duration, trying to explore this relationship. The dataset used in this s rail transit trip survey is used as the dataset in this study which includes 4254 records. The features are estimated by using the random forests model. As a result, trip purposes have the most significant effect at any threshold of walking duration; the feature of peak hour also shows importance; the other features have different performance at different thresholds. The result shows a clear relationship between real walking durations and passengers' individual characteristics. Also, this result is expected to provide references 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 can influence the use of public transit, if considering from overall, they can be separated into two parts which are the convenience of public transit (environment factors) and the willingness of traveling by public transit (behavior factors). This study will focus on the behavior factors to examine the relationship between passengers' individual characteristics and walking durations from departures to transit stations.

At present, 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 by asking 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 individual characteristics generally should have different preferences of walking duration, even if they 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. Even though the surveyed walking durations should not be viewed as the direct reflection of passengers' willingness for walking durations, they can reflect the differences in the willingness of passengers who have different individual characteristics in an indirect way. On the basis of above-mentioned, this study attempts to give explanations on how individual characteristics influence the walking duration using the study case of Fukuoka subway. The expected result of this study can be viewed as a reflection of willingness for walking durations, and also is expected to provide references for helping determine the catchment area of transit stations.

LITERATURE REVIEW

This section reviews the literature on the issue of walking duration/distance to transit stations, some problems that still not well addressed are collated and summarized. The review is arranged into two parts, how to deal with the variable of walking duration (research object in this study), and how to estimate the feature of individual characteristics (independent variables in this study).

For the walking duration, the research object in this study, there is always a difficult point in obtaining the accuracy walking distance/duration 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/duration is not the reflection of how long people are willing to spend on walking to transit stations, but only the walking distance/duration between the departure and the transit station. For this problem, some studies chose a different perspective trying to explain the walking distance/duration by introducing one threshold of walking duration (Besser and Dannenberg 2005; McCormack et al. 2008). They examined the differences in the distribution of influencing factors over/under the specific threshold of walking duration. 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 example, it may lead to a great loss of information in the raw data, and it is also difficult to decide the threshold of walking duration. Moreover, another point for this issue is whether to choose walking distance or walking duration as the threshold. To date, there have been a lot of studies working on the relationship between walking distance and passengers' individual characteristics, some of them 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, passengers with the same individual characteristics may have the similar willingness of walking duration, but they generally have different willingness of walking distance due to different travel speed. This is also the reason why this study chooses the walking duration as the research object.

In the influencing factors for walking durations, passengers' individual 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), whereas, there are few studies having clearly verified the relationship between individual characteristics and walking distance/duration, 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 walk a longer 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 is for other categories of factors, such as the factors of transportation environment, land use, and willingness of passengers (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/duration can be influenced by some specific kinds of factors, such as socio-demographic characteristics, trip purposes, and built-environment, but the problem is how and how much the walking distance/duration can 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 walking distance/duration (research object) and the influencing factors (independent variables). 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 uses the thresholds of walking duration as the research object. For any given threshold of walking duration, the respondent who gives the answer greater than the given threshold can be viewed as this respondent can accept this threshold of walking duration. The relationship between walking durations and individual characteristics is explained by examining the feature distribution over/above the given thresholds of walking duration. For the analytical method, since various types of regression model has been verified not suitable for this issue, instead of finding the linear relevance between dependent and independent variables, this study introduces the approach of machine learning into this issue, and try to explain the relationship between walking durations and individual characteristics from the view of probability.

DATA AND METHODS

Introduction for the study case of Fukuoka

The study case of Fukuoka, is the sixth largest city in Japan which has the largest population in Kyushu Island of Japan (more than 1.5 million). The dataset 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 (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.

Necessary hypothesis

Based on the ideas outlined above, this study proposes the following three necessary assumption.

- H1: The distribution of departures and destinations for people with different individual characteristics in Fukuoka, Japan has no significant features in spatial distribution.
- H2: The maximum acceptable walking duration of passengers with same individual characteristics should subject to the normal distribution.
- H3: The respondents are viewed as they can accept the walking duration that they answered.

Based on H1 and H2, if examining the threshold of walking duration t, set the proportion of k group in the whole surveyed sample as r_k, the proportion of people whose walking durations are under the given threshold of t minutes is marked as $r_k^{<t}$, the proportion of people whose walking durations are over the given threshold of t minutes is marked as $r_k^{>t}$. Now considering if individual characteristics have no significant correlation with maximum acceptable walking durations, there should be no significant differences among r_k, $r_k^{<t}$, and $r_k^{>t}$; otherwise, the three proportion should show significant differences, and the differences will show regular changes at different threshold t.

Based on H3, if someone gives the answer t minutes, it means this respondent can accept the walking duration of t minutes and any walking duration that less than t minutes. Indeed, this respondent perhaps can accept the walking duration over he answered, but in this study, we just concern about if he can accept the given threshold of the walking duration other than how long he can accept.

Description for the data

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 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 with null value or abnormal value. The procedure of data cleaning is shown in figure 2. For analyzing the relationship between walking duration and individual 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 passengers aged from 25 to 55 account for the major of the whole passengers, while schoolchildren aged under 15 take public transit very little. The distribution graph doesn't show significant peak values at any specific age group. Figure 4 shows the distribution of real walking durations to transit stations, it has a mean value of 8.32, and the standard deviation is 2. Notably, 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. Despite the existence of differences between survey and reality, as the second assumption proposed above, passengers are viewed that they can accept the walking duration what they answered, therefore, the peak values are thought available in this study.

Methods

For the issue of this study, the dependent object (walking duration to the transit station) is a continuous variable, while the independent objects (individual characteristics) 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 widely used 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, since what we concern with is the feature distribution of individual characteristics over/under the thresholds of walking durations, this issue can be converted into a binary classification problem. Giving a specific threshold of walking duration to go to a transit station, passengers walking over or under this threshold have different features of individual characteristics. That is, if the walking duration of a passenger is over a given threshold of walking duration, it can be considered that this given threshold of walking duration is an acceptable one (less than the maximum acceptable walking duration), 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 reflected by the proportion of walking longer than the given thresholds. This study will examine the relevance of passengers' individual characteristics and this proportion, the expression is given as Equation 1.

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

Where:

D is the dependent variable,

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

t^i is the maximum acceptable walking duration for travel individual i,

T is the given threshold of walking duration.

The essence of such issue can be viewed as a binary 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, the random forest model, to train and test the samples. Random forests model is an ensemble learning method mainly for classification and prediction. The random forests model is an extension and improvement for the decision tree model, it is operated 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 willingness of 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 is the probability of walking longer than the given threshold of walking duration for each passenger in terms of individual characteristics. The accuracy of this result is evaluated using the method of simple moving average by descending order of the predicted probabilities.

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, this study reclassifies it 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 the relationship between individual characteristics and walking durations. 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 individual characteristics. The walking duration more than 13 minutes is not accepted by most people even if they have different individual characteristics, and the threshold less than 5 minutes is also not sensitive to individual 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 walking longer than the given threshold of walking duration in terms of people with different characteristics. For estimating the model, the dataset is divided into two parts, 50% of the samples are used for fitting the model and obtaining the coefficients, the rest 50% are used for testing the ability of prediction. The prediction process is presented in Figure 5. For the random forest model, the prediction 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.

The accuracy of results are evaluated by using the method of simple moving average. 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 the trend line in Figure 6 partly proved that this result can reflect the behavior of people with specific individual characteristics 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 in Table 3, trip purpose is the most important factor in determining the walking durations at all the 3 thresholds. The feature of peak hour is also significant in explaining the walking duration. People tend to walk a longer time 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 walk more than 8 minutes; the people whose age is between 45 to 64 can accept a 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 individual characteristics affected the walking durations. Therefore, for this kind of study, the assumptions proposed above are necessary. 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 acceptable walking duration for that group of people. Perhaps due to the reasons of discrepancy in investigation and the basic assumption, few existing studies can explain the relevant between walking duration and individual characteristics quantitatively correctly.

According to the achievement of previous studies, if examining the walking duration as a continuous variable, there is no linear relation between walking duration and individual characteristics. For this reason, 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 commonly acceptable walking durations range from 5 to 13 minutes, therefore, it can be inferred that people who prefer walking less than 5 minutes and who can accept walking longer than 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 walk to transit stations 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 a group of people, it shows a good predictive ability, especially 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. With this prediction, for example, if knowing the individual characteristics of residents in a particular area locating from the transit station T minutes, the general acceptability of walking to the station for the residents in this area should be predictive. Therefore, this prediction of willingness is expected to be used in planning the catchment area of transit stations or estimating the catchment area of existing transit stations.

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. and Rietveld, P. (2000). "How do people get to the railway station? The dutch experience." *Transportation Planning and Technology*, 23(3), 215–235.

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 walk longer than the given threshold of walking duration, while M is the opposite meaning.

List of Figures

Fig. 1. Distribution of Transit Stations

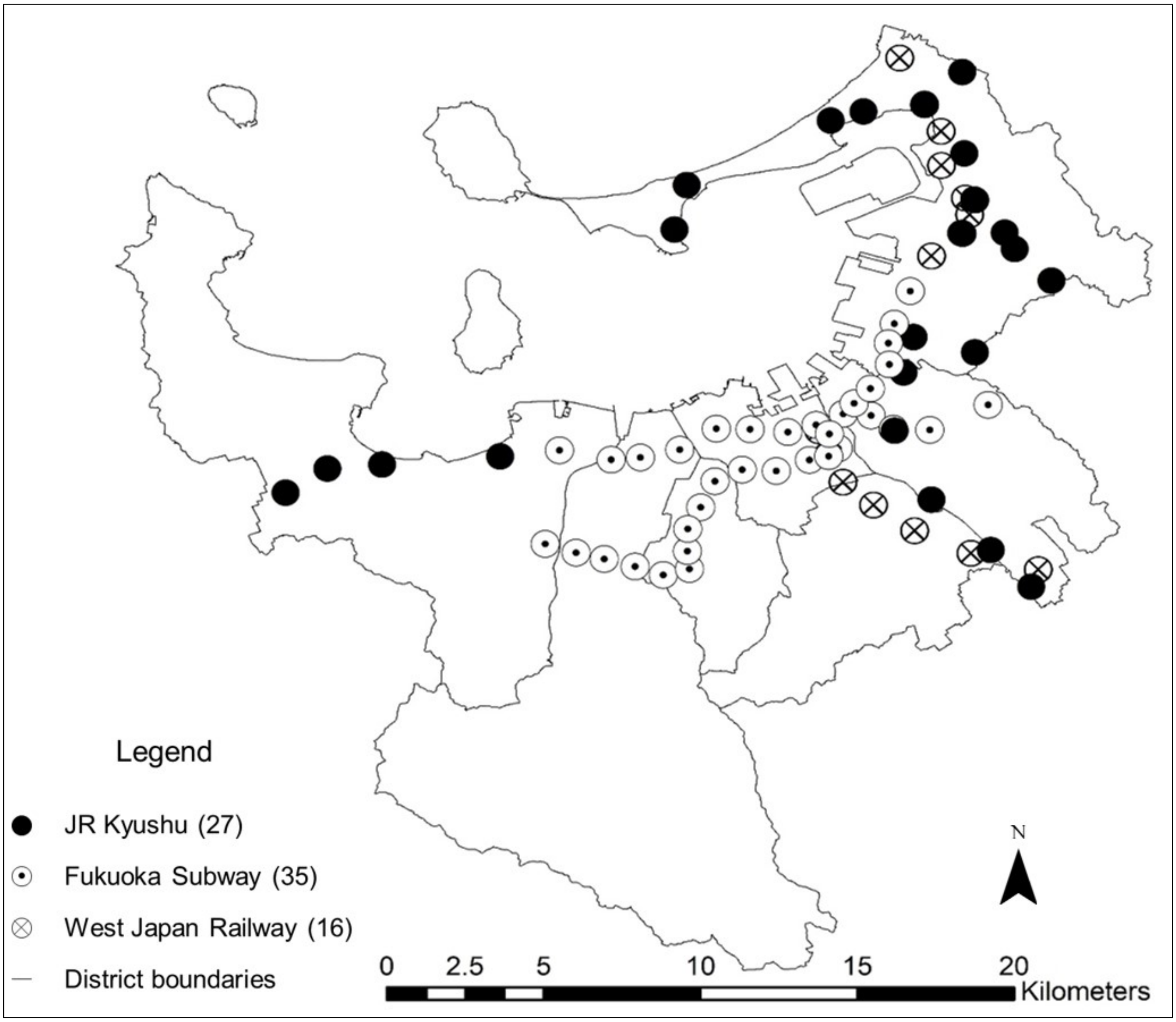

Fig. 2. Process of Data Cleaning

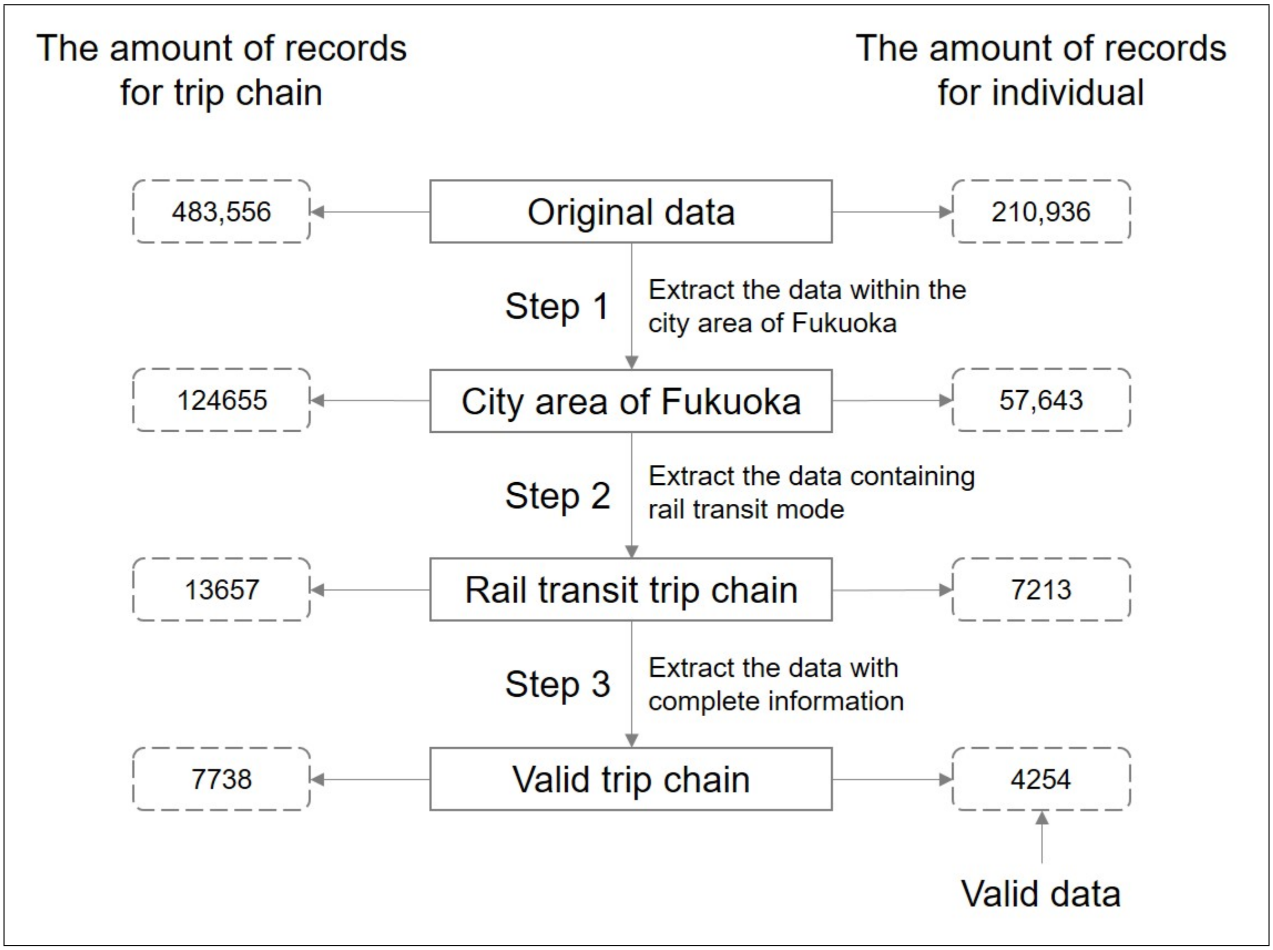

Fig. 3. The Age Distribution of Passengers

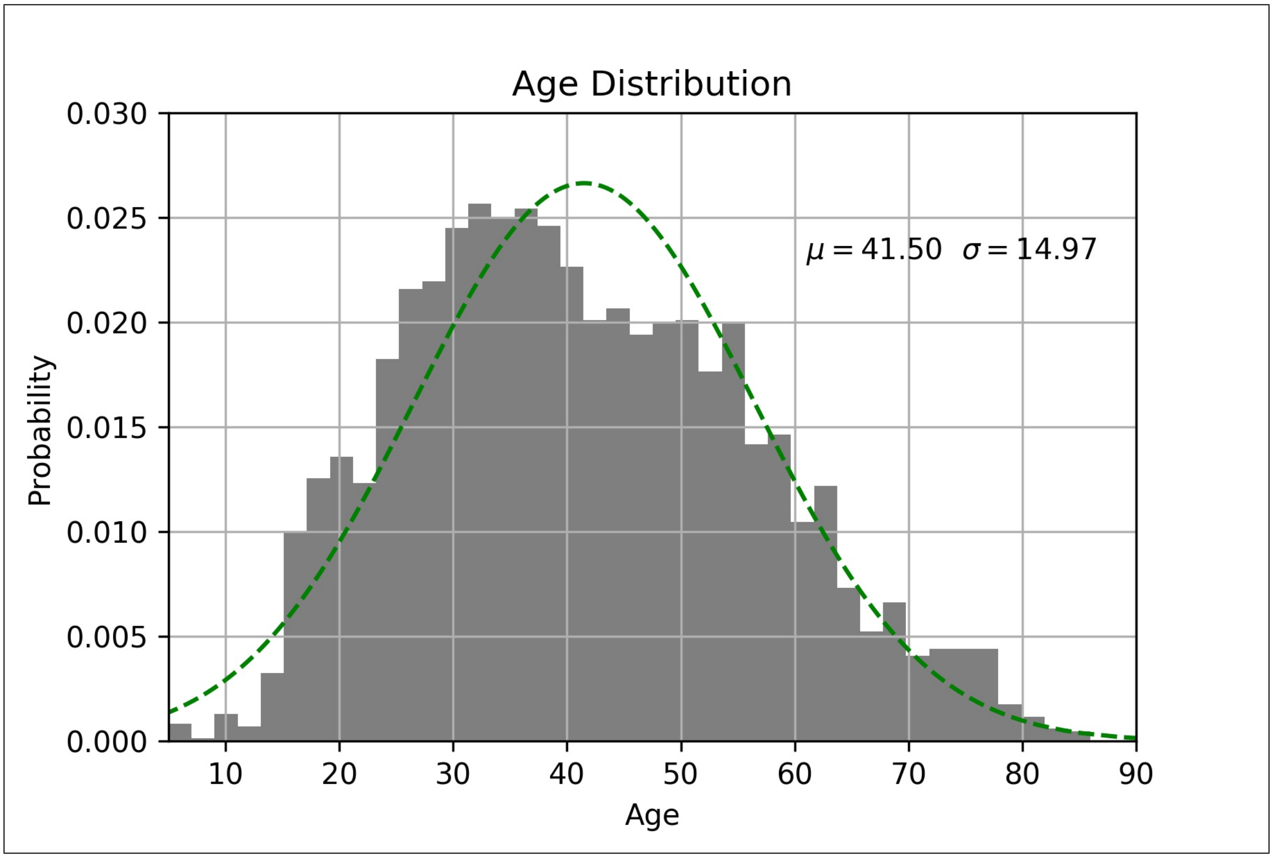

Fig. 4. Distribution of Walking Duration

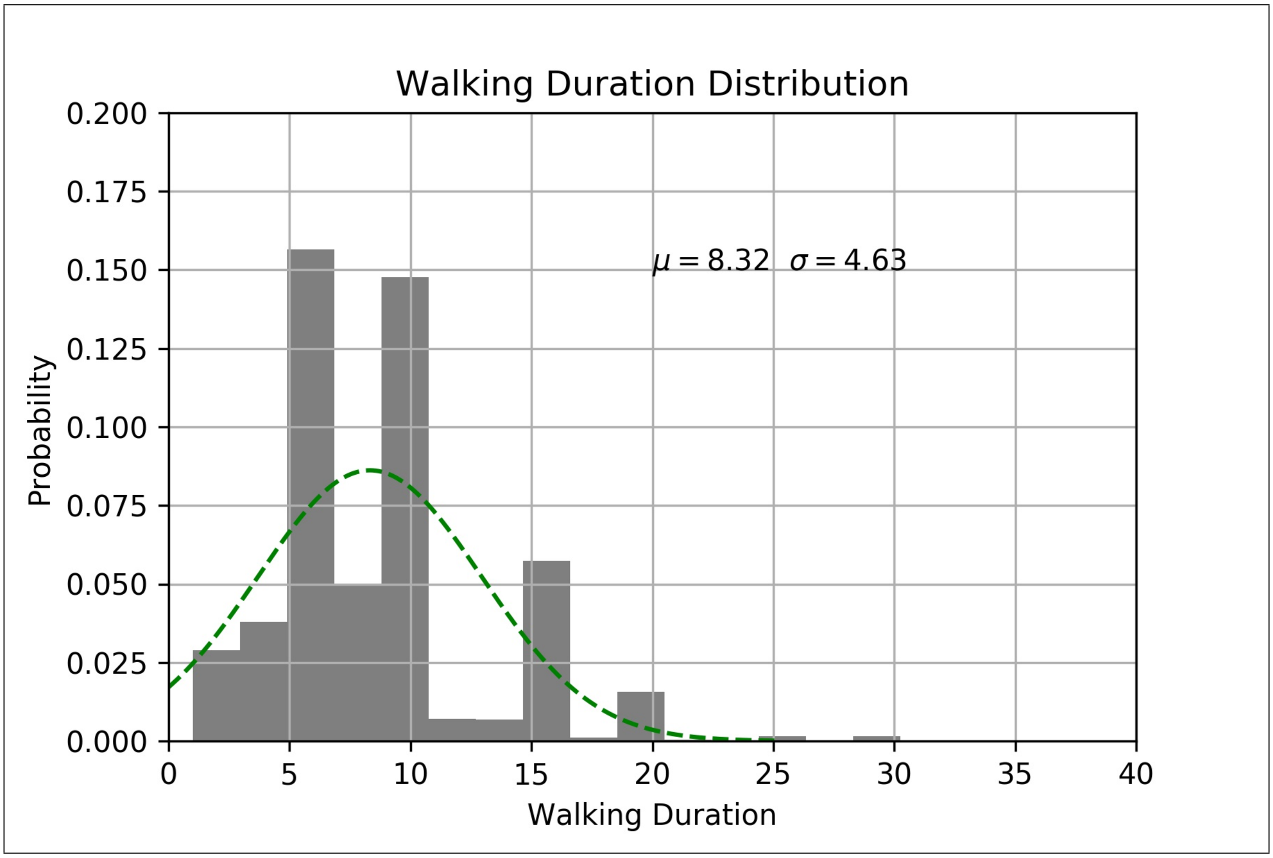

Fig. 5. Prediction Process in Random Forests Model

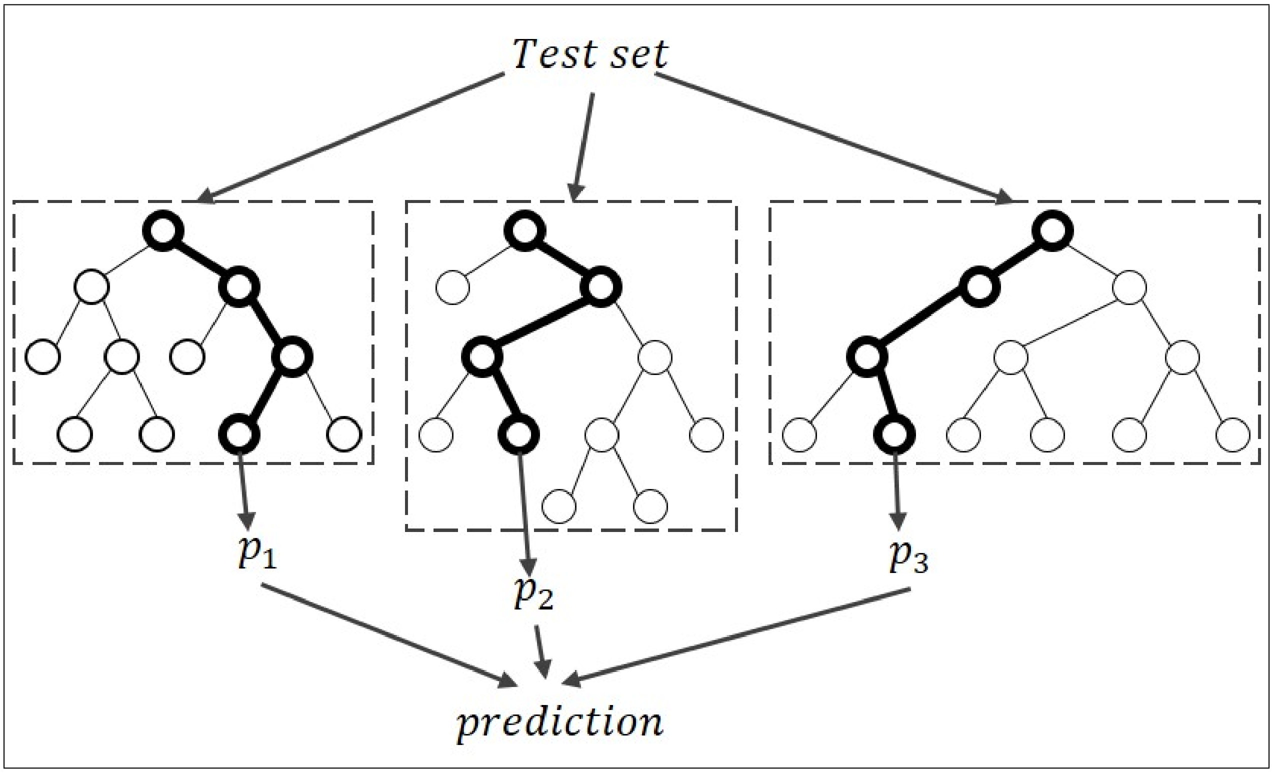

Fig. 6. Trend Line of Prediction and Test Set