Arrival Time Estimation and Visualization based on Bus Traffic Data

Hitomi IMAI[1], Kei HIROI[2], and Nobuo KAWAGUCHI[3]

[1] Graduate School of Engineering, Nagoya University, Aichi, Japan
imai@ucl.nuee.nagoya-u.ac.jp
[2] Institutes of Innovation for Future Society, Nagoya University, Aichi, Japan
k.hiroi@ucl.nuee.nagoya-u.ac.jp
[3] Graduate School of Engineering, Nagoya University, Aichi, Japan
kawaguti@nagoya-u.jp

Abstract. Bus transportation service is more influenced than other public transport by various factors such as traffic congestion, weather condition, number of passengers, traffic signals, and so on. These factors often cause delay and the users may feel inconvenience while waiting at the bus stop. Few previous studies analyzed the relationship between operation situations and multiple factors by visualization. Thus, we propose an arrival time estimation method and a visualization model. The arrival time estimation model dynamically updates the accuracy by estimating method using a combination of multiple regression model and Kalman filter. The visualization model analyzes relationships between delay and factors. The goal of this study is to realize a society where people can use the bus more comfortably.

Keywords: Bus arrival information system, Multiple regression model, Kalman filter, Visualization

1 Introduction

Many people use public transport bus service. According to a survey by the Ministry of Land, Infrastructure, Transport and Tourism[1], in Japan, about 12 million people use it every day. Recently, traffic data is collected by various systems[2], for example bus arrival information system. This system gets bus information using GPS: arrival or departure time, traveling locations (latitude and longitude), etc. Besides, the number of passengers and the behavior of the driver are also recorded. On the other hand, bus operation situations are more influenced than other public transport by traffic congestion[3], weather condition[4], number of passengers[5][6], traffic signals[7], and so on. These factors[8] are related to delay and the motion of buses changes complicatedly[9][10]. Many services inform the users of departure from a bus stop[11][12], but few services provide specific estimated arrival or delay times. Few previous studies analyzed the relationship between operation situations and multiple factors by visualization. The bus is delayed, and they may feel inconvenience while waiting at the

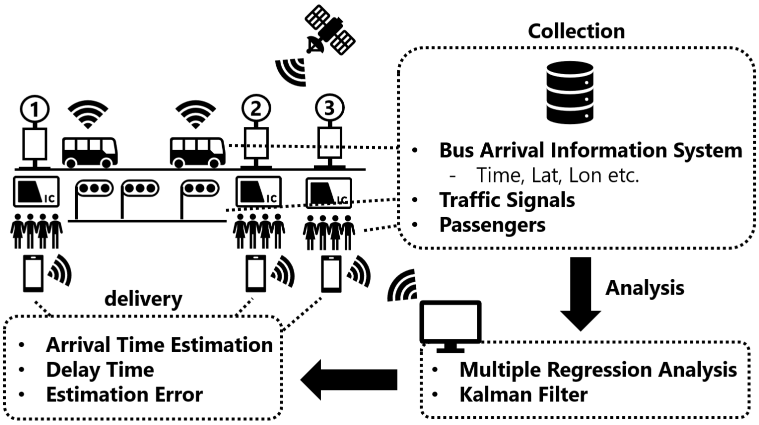


Fig. 1: Outline of the proposed system.

bus stop. Thus, we propose an arrival time estimation method and a visualization model. "EMRF (Extended Multiple Regression Filter) model": The arrival time estimation model dynamically updates the accuracy by estimating method using a combination of multiple regression model and Kalman filter. Multiple regression model estimates the trend in advance and Kalman filter updates the estimation to the optimal state based on it in Figure 1. As a feature of this method, the closer the bus is to the terminal station, the better the accuracy is improved. "Bus Tapestry": The visualization model analyzes relationships between delay and factors. It creates a heat map of operation situations (delay or premature) and adds bus stop positions, signal positions, and bus traffic data. We can visually find factors related to delay. The goal of this study is to realize a society where people can use the bus more comfortably.

2 Literature Review

2.1 Multiple Regression Analysis

Multiple regression analysis is a liner model and derives the dependent variable Y using multiple independent variables $X_i (i = 0, 1, \cdots)$ by Equation 1 as follows:

$$Y = a_0 + a_1 X_1 + \cdots + a_i X_i \tag{1}$$

where, a_i is a coefficient calculated for each independent variable. In the study by Patnaik et al[13], dependent variable was estimation time taken between bus stops. Independent variables were factors to influence delay, like a time required for timetable, a distance between bus stops, a number of passengers, the time to open and close the door, and so on. The estimation by this model was highly accurate. However, multiple regression analysis is a static estimation based on the past data and does not considers that passengers increase due to rainy weather and events holding near the bus top. Therefore, it difficult to respond to such a real-time changing environment and present the arrival estimated time to users.

2.2 Kalman Filter

Kalman filter is a powerful mathematical tool that can estimate the future states of variables even without knowing the precise nature of the system modeled. In the study by Chen et al[14], the time required in the next interval was dynamically estimated based on the time required of the timetable and information accumulated form the starting station. Although Kalman filter can process information including errors and estimate dynamically, accurate estimation is difficult when a bus stop interval is characteristic.

3 Model Development

3.1 Arrival Time Estimation

EMRF model consists of multiple regression model and Kalman filter. Before departure, multiple regression model estimates changes in inputs, and after departure, Kalman filter estimates dynamically from the difference between the measured value and the estimated value based on the preliminary estimation.

First, we explain the multiple regression analysis in this study. Dependent variable is estimation time taken between bus stops[13]. Independent variables are factors to influence delay[15], like bus stop sections, delay in front of n stations, time zone, the day of the week, time required for timetable, and number of passengers. The delay is defined as the difference between the required time for timetable and the actual required time. For the time zone, Early Morning is defined until 7:00, Late Morning is defined from 8:00 to 10:00, Early Noon is defined from 10:00 to 13:00, Late Noon is defined as from 13:00 to 17:00, Evening is defined from 17:00 to 19:00, and Night is defined after 19:00. For the number of passengers, the number of people getting on the bus is compared with the number getting off, and the higher number is recorded.

Secondly, we describe our Kalman filter in Figure 2. Based on the estimation of the multiple regression model, Kalman filter estimates the required time for each bus stop interval. The end point is defined as bus stop N. At arbitrary bus stop k, EMRF model estimates the times required for bus stop intervals k − $(k + 1)$, n − $(k + 2)$, $\cdots$, ,k − N. We input values estimated by a multiple regression model as the initial state of the system. Specifically, for $k = 1$, a estimated value was calculated using past data. For $k > 1$, the model calculates the real time differences (delay or premature) using the estimated value for bus stop interval $(k-1) - N$ together with the timetable were used to calculate, and it inputs the results that are re-estimated using the multiple regression model. After the bus leaves the starting station, the model updates the system status and estimated value each time it arrives at the bus stop and repeats this motion until it reaches the end point. By repeating the update, it is possible to correct even if the estimated value and the actual measured value are different from each other. Additionally, as the bus approaches the end point, the accuracy of the estimation can be improved.

In general, the Kalman filter estimates the state of the system at time $(k+1)$ using the state equation based on the previous state by Equation 2 as follows:

$$x_{k+1,j} = \Phi_{k+1} x_{k,j} + u_k + Wk, j \tag{2}$$

where, $x_{k+1,j}$ is the state of the system at time $(k+1)$, Φ_{k+1} is the state-transition model, u_k is the state vector, and Wk, j is noise. We use a multiple regression model (Equation 1) instead of the state vector u_k in Equation 2. The relationship between an observation and a state variable is expressed by the observation equation of Equation 3.

$$z_k = H_k x_{k,j} + v_{k,j} \tag{3}$$

where, H_k is the observation model and $v_{k,j}$ is noise. We define the state variable $x_{k,j}$ as the estimated time $E_{k,j}$ and the real time required R_k in Equation 4 as follows:

$$x_{k,j} = (E_{k,j}, R_k) \tag{4}$$

Where, $T_{k,j}$ is the total value from arbitrary bus stop k to the bus stop j and R_k is the total value from the starting station to the bus stop k.

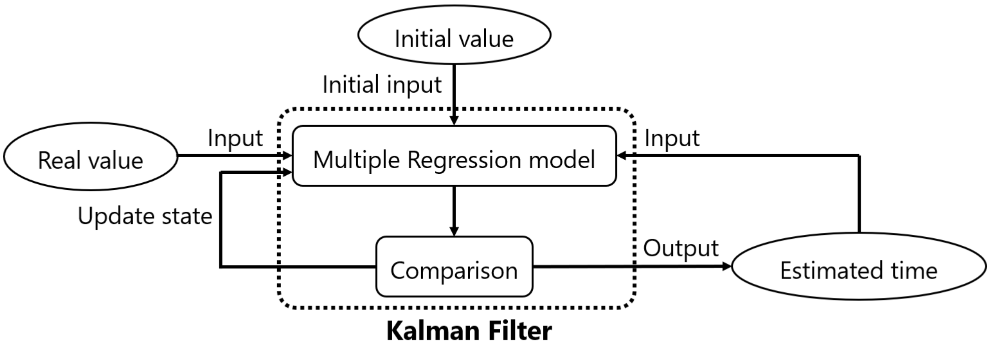
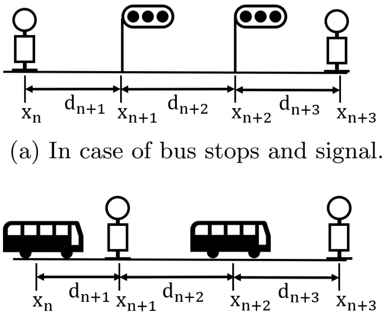


Fig. 2: Outline of the proposed method.

3.2 Visualization

Bus Tapestry creates a heat map of operation situations (delay or premature) and adds bus stop positions, signal positions, and bus traffic data. The vertical axis is the time zone, the horizontal axis is the distance from the starting to the ending bus stop and each position is the accumulated distance.



(a) In case of bus stops and signal.

(b) In case of bus.

Fig. 3: Definition of interval and total distance.

First, we explain how to calculate the data. Our visualization expresses operation situations by the difference between the actual time required and the required time. The value is positive when the bus is later than the timeline, and negative when the bus is earlier than the timeline. The route distance is the total value based on the bus stop and the location information of the signal (latitude and longitude) in Figure 3a. Our method calculate the interval distances d_{n+1} from the x_n to x_{n+1} using each position information. The total distance x_{n+1} is the sum of their values in Equation 5 as follows:

$$x_{n+1} = x_n + d_{n+1} \tag{5}$$

In this study, our method does not take account of the curve of the road, etc. We did not calculate from data by the bus arrival information system because there might be some error in it. Similarly, The travel distance of the bus is also the sum of the interval distances between the locations travelled in Figure 3b. For the travel distance, the location of the bus stop is the point where the departure information was recorded.

Secondly, we describe how to visualize the data. Our method visualizes operation situations and the signal position using our created data. The heat map represents operation situations (delay or premature) on the vertical axis is the time zone(hourly) and the horizontal axis is the distance(km). The horizontal axis expresses the distance between bus stops by putting the same data every 0.01 km within each bus stop interval. We add the signal position data (mileage) and make the number of traffic signals between routes visible. Consequently, we can analyze the influence of the signal between bus stops and operation situations of each time zone. Furthermore, we can evaluate operation situations in more detail by adding the location information of each bus to the visualized data.

4 Data Collection

In this study, data was collected from bus arrival information system. The recorded area is in Aichi Prefecture and it includes information of position, time, route, bus stop, and so on. This data was provided by Transportation Bureau City of Nagoya[11] and Meitetsu Bus Company Limited[12] through Location Information Service Research Agency(Lisra)[16]. The former data was recorded when arriving and departing the bus stop and when communicating every 30 seconds. The range of data collection was for December 13-22, 2014. It has 1030 buses, 3784 bus stops and 664 routes. The latter data was recorded only when departing the bus stop. The range of data collection was for July 1-15, 2016 and from January through October 2017. It has 710 buses, 1539 bus stops and 523 routes. We have number of passengers on Meitetsu Bus Company Limited. In addition, we point each position of traffic signals on the target bus routes.

5 Analysis of Results

5.1 R-squared by Multiple Regression Analysis

The results of a multiple regression analysis are shown in Figure 4. The data is from Meitetsu Bus Company Limited and the range of it is for March 1-31, 2017. For comparative purposes, R-squared for Nagoya City Bus data are collected in Table 4. The range of used data is for December 13-19, 2014. R-squared indicates how well independent variable accounts for the variability of another, dependent variable. The value of R-squared ranges from zero to one, with values closer to one indicating a lower degree of relative error. The highest R-squared value was 0.90. However, since coefficients had abnormally large values like 5.86×10^{11}, multiple regression analysis could not be performed adequately. This is because the route contains 25 bus stops, and as such there are too many independent variables. On the other hand, the smallest R-squared value is 0.46. It seems to be caused by irregular congestion in a bus stop interval. The average value of Meitetsu Bus was 0.69 and close to the average value of Nagoya City Bus (0.76). However, the value of Meitetsu Bus was slightly lower than that of the Nagoya City Bus Since The number of data on Meitetsu Bus was smaller than that on Nagoya City Bus. data of Nagoya City Bus was recorded when arriving and departing the bus stop and when communicating every 30 seconds, but data of Meitetsu Bus was recorded only when departing the bus stop. The relationship between R-squared and number of bus stops is shown in Figure 5. R-squared is 0.0053 and there was no correlation between that of Multiple Regression Analysis and number of bus stops in Figure 5.

5.2 Accuracy Verification by Changing the Amount of Data

We verified how the change in the amount of data affects the accuracy of estimation by the multiple regression model, using the data from Meitetsu Bus

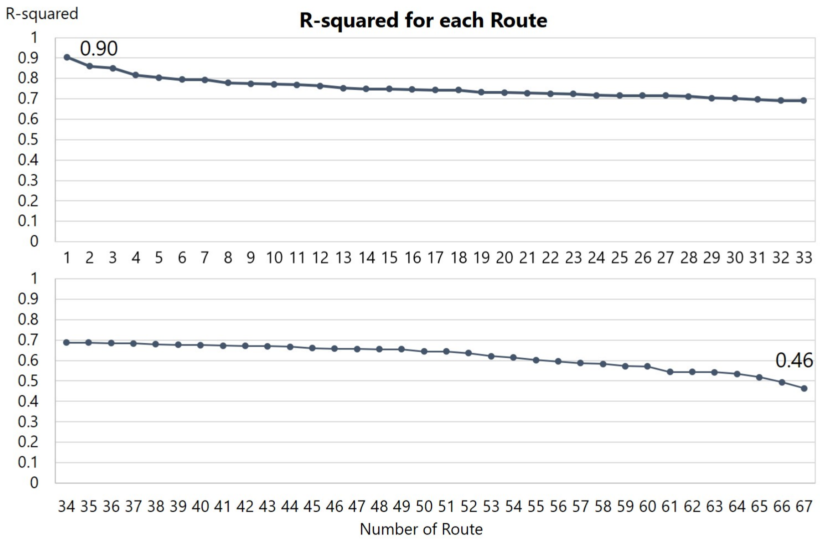


Fig. 4: R-squared for Meitetsu Bus.

Company Limited. The Compared data were data for 14 days (July 1-14, 2016), 101 days (July 1-14, 2016 and from January to March 2017), and 101 days excluding the abnormal values. The estimated date is July 15, 2016. We removed the abnormal values using the interquartile range. We calculated R-squared by comparing estimated and actual values for route 9 (Figure 4) in Table 2. Table 2 shows the value of R-squared increased as the number of data increased. Additionally, excluding the abnormal values further improved the accuracy of estimation.

Table 1: R-squared for Nagoya City Bus.

RouteID	R-squared
8415	0.79
8471	0.76
8784	0.69
8921	0.58
8939	0.80
8990	0.80
9014	0.80
9015	0.88
Average	0.76

Table 2: Variation of R-squared in route 9.

The type of data	R-squared
14 days	0.34
101 days	0.44
excluding	0.55

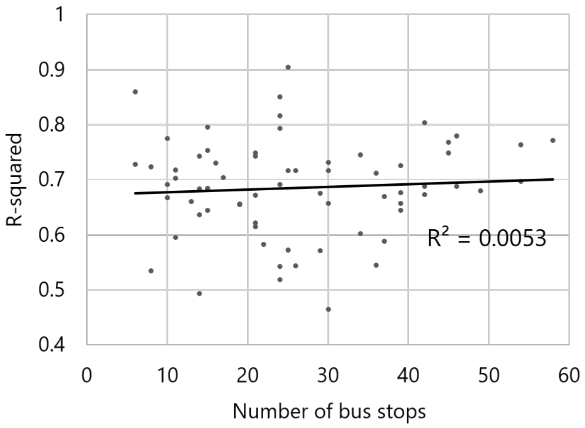
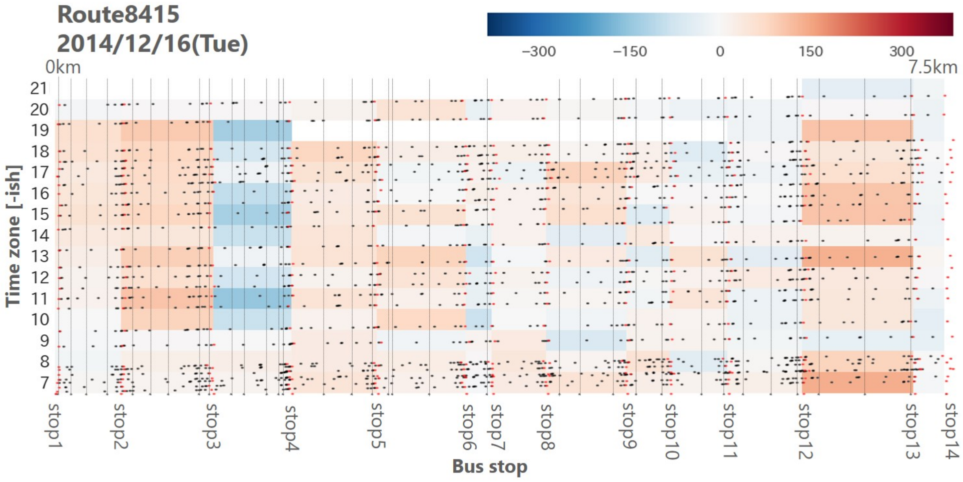


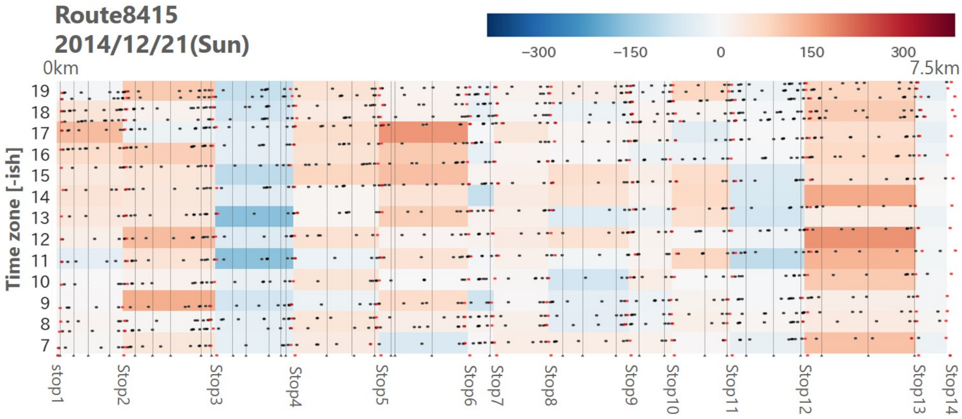
Fig. 5: The relationship between R-squared and number of bus stops.

5.3 Visualization

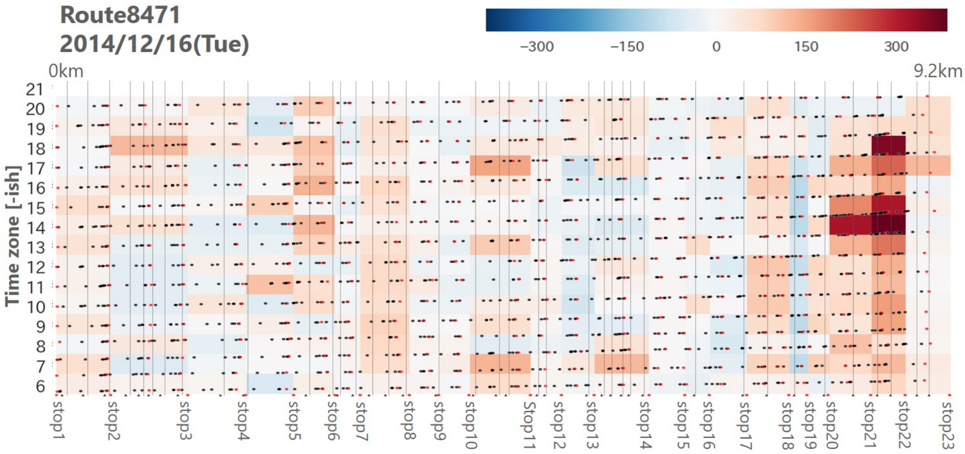
We used the data from Transportation Bureau City of Nagoya on December 16 and 21, 2014. The result of our visualization is presented in Figure 6. The black dots are the running position of the bus, the red dots are the bus stop position, and the dotted lines are the signal position. The white parts in the heat map are the time zone during which the bus was not running. Figure 6a and Figure 6b are other day of the same route and show a similar delay condition as a whole. However, in the range of bus stop 5 to bus stop 6, we find that the delay on Sunday is greater than those on Tuesdays from 16:00 to 17:00. Figure 6a and Figure 6c are the other route on the same day. They show the route 8471 has a large delay near three stations before the end point compared to the route 8415. Moreover, in Figure 6a, the signal between bus stop 2 and bus stop 3 does not significantly affect the delay because there are few points before it. On the other hand, the signal between stop11 and stop12 is likely to affect the delay because there are many points before it. Thus, we can find visually the relationship between delay and factors by our visualization.



(a) Visualization in route 8415(Tue).



(b) Visualization in route 8415(Sun).



(c) Visualization in route 8471(Tue).

Fig. 6: Result of visualization.

6 Evaluation of the Model

We used the data from Meitetsu Bus Company Limited for March 1-31, 2017 in route 9. The estimated date is January 31 (Tue), 2017. The model was created for 30 days, excluding the estimated date. For the estimated date, the number of passengers and delay in front of n stations were the average of 30 days.

6.1 Comparison by Estimation Errors

For schedule 10430, estimation errors by Multiple regression model and our model are presented in Figure 7. Schedule 10430 is a bus running from 20 to 21 hours. estimation errors are the difference between the estimated value and the actual value. It is a positive value when EMRF model estimates longer than the actual value, and it is a negative value when the model estimates shorter than the actual value. Figure 7 shows that estimation errors is smaller than that of Multiple regression model and EMRF model corrects the estimation.

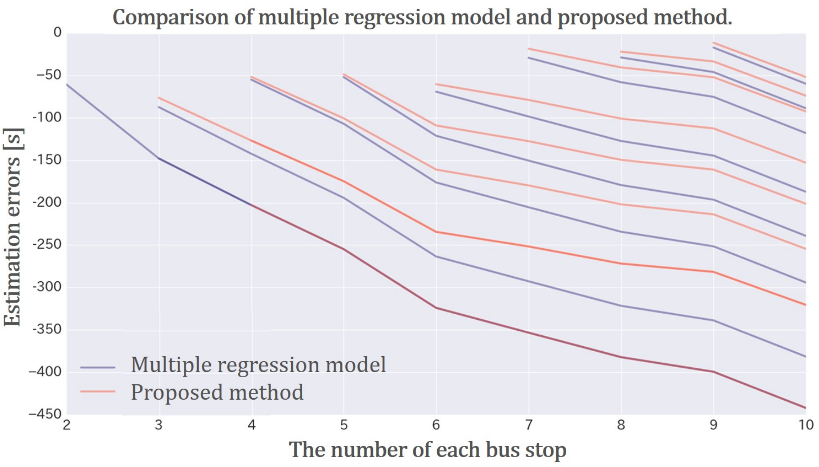


Fig. 7: Estimation error in route 9, Schedule: 10430.

6.2 Comparison by Values of RMSE

We evaluate the models using the value of RMSE (Root Mean Squared Error) in Equation 6 as follow:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad [s] \tag{6}$$

Where, N is the number of bus stops intervals, y_i is the actual value of i-th bus stops interval, and $\hat{y}_i$ is the estimated value of i-th bus stops interval. RMSE is an evaluation method that shows how the estimated value is separated from the actual value. The closer the value of RMSE is to 0, the more accurate the estimation. For schedule 10400, the value of RMSE by Multiple regression model and our model are presented in Figure 8. Schedule 10430 is a bus running from 10 to 11 hours. The estimated date varied from March 1 to 31. The horizontal axis is the bus stop estimated and the vertical axis is the value of RMSE. Figure 8 shows that the value of RMSE by our model is smaller than only by multiple regression model. Especially at bus stop 2, the estimation was largely corrected.

Similarly, for all schedules, the average value of RMSE are presented in Figure 9. This figure shows that the value of RMSE was smaller in our model even in the case of the average value by the data of one month. Therefore, it is assumed that our model can improve the accuracy of estimations.

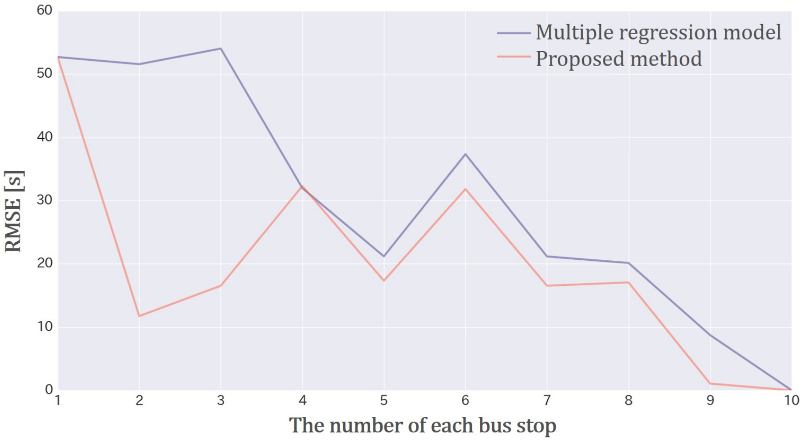


Fig. 8: The value of RMSE in route 9, Schedule: 10400.

7 Examination of Presentation Method

We propose a method of presenting arrival time including estimation errors. Using the standard deviation, the estimated required time is calculated with some leeway, and presented to users with an accuracy of about 95% in Equation 7 as follows:

$$E' = E \pm 2SD(E - R) \tag{7}$$

where, E' is required time including estimation errors, E is estimated required time, and R is actual required time. Showing users the shortest arrival time

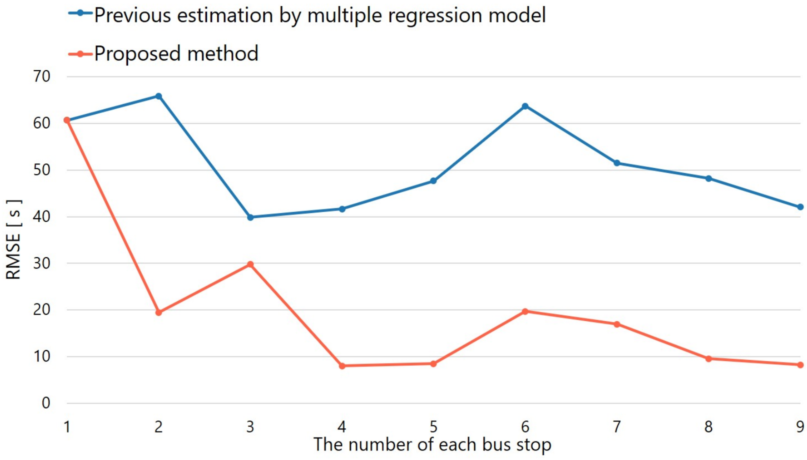


Fig. 9: The average value of RMSE.

allows them to broaden their choice of actions in Figure 10. For example, “ If this time the earliest possible, let’s go to a convenience store ”, or “ Since there is no need to hurry, let’s walk slowly ”, etc. Presenting the latest arrival time has the effect of alleviating the anxiety of, “ How long will I have to wait at the bus stop? ” Our method can also show the estimated arrival time at the destination stop and inform the users of it because our model is possible for all bus stop intervals. Presenting specific estimated arrival times in this way gives users a more accurate idea of operation situations, making it easier to act.

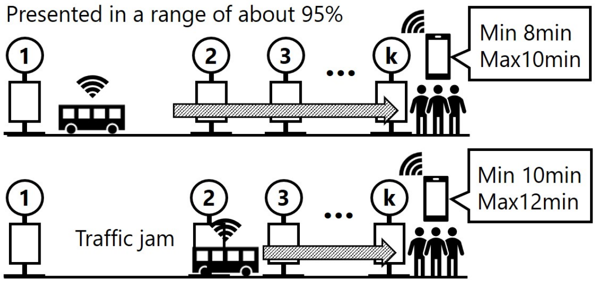


Fig. 10: How to present to users.

Moreover, in order to know the view of the users on presentation method of the estimated arrival time, we investigated using the questionnaires. This period was for January 29-30, 2018. We got responses of 184 people using SNS and the number of valid responses was 169 people. About how much to refer according to estimation errors, the users envaulted estimation errors in 5 stages : “Never”, “Hardly ever”, “Neutral”, “Some of the time”, “All of the time”. When the estimation errors is within 1 minute, “All of the time” accounted for about 90 % of the total. When the estimation errors is within 1-5 minutes, “All of the time”, “Some of the time” or “Neutral” accounted for about 90 % of the total. Therefore, it is assumed that the standard of estimation errors is within 5 minutes. Figure 11 shows the results of the questionnaire about the presentation method. There were 4 kinds of Sample screens : “Estimated delay time”, “Estimated arrival time”, “Estimated time remaining”, “Graphical presentation”. “Estimated arrival time” and “Estimated time remaining” each accounted for about 40 % of the total. Thus, we found that the users prefer display of arrival time than delay time. It is assumed that the users can use comfortably the application in which they can select presentation method because the answers were divided.

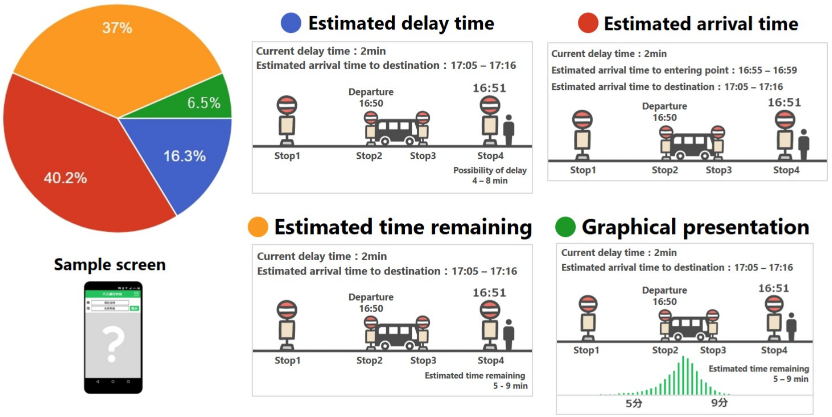


Fig. 11: The result of the questionnaire about the presentation method.

8 Conclusion

In this study, we proposed EMRF model and Bus Tapestry. EMRF model is a dynamic model for arrival time estimation combining multiple regression model and Kalman filter. We verified the accuracy of the estimation using R-squared and evaluated EMRF model by error or RMSE. The results showed that the average of estimation errors improved from 186 seconds to 17 seconds. We also presented estimated arrival time including estimation errors. We investigated using the questionnaires and got 184 answers to the presentation method. The results showed that the standard of estimation errors was within 5 minutes and the users prefer display of arrival time than delay time.

Bus Tapestry is a method of visualization for analyses operation situations. We can visually compare operation situations of other days or routes and find the same or difference features. Additionally, we can see the relationship between delay and signal in more detail. In the future, it may be possible to estimate abnormal values and use machine learning. Furthermore, to start an estimating service, it is necessary to conduct a demonstration experiment and collect opinions of users.

References

1. Ministry of land, infranstructure, transport and tourism. http://www.mlit.go.jp/en/index.html. January 2018.
2. Bo Yu, Jing Lu, Bin Yu, and Zhongzhen Yang. An adaptive bus arrival time prediction model. *the Easten Asia Society for Transportation Studies*, Vol. 8, pp. 1126 – 1136, 2010.
3. Michael L. Anderson. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *American Economic Review*, Vol. 104, No. 9, pp. 2763 – 2796, 2014.
4. Victor W. Stover and Edward D. McCormack. The impact of weather on bus ridership in pierce county, washington. *Journal of Public Transportation*, Vol. 15, No. 1, pp. 95 – 110, 2012.
5. Chen Zhang and Jing Teng. Bus dwell tme estimation and prediction : A study case in shanghai-china. *Procedia - Social and Behavioral Sciences*, Vol. 96, pp. 1329 – 1340, 2013.
6. Amer Shalaby and Ali Farhan. Prediction model of bus arrival and departure times using AVL and APC data. *Public Transportation*, Vol. 7, No. 1, pp. 41 – 61, 2004.
7. Chin-Woo Tan, Sungsu Park, Hongchao Liu, Qing Xu, and Peter Lau. Prediction of transit vehicle arrival time for signal priority control:algorithm and performance. *IEEE Transactions on Intelligent Transportation System*, Vol. 9, No. 4, pp. 688 – 696, 2008.
8. Andrei Iu and Adrian Friday. Statistical modelling and analysis of spare bus probe data in urban areas. *International IEEE Annual Conference on Intelligent Transportation Systems Madeira Island*, pp. 1256 – 1263, 2010.
9. Takashi Nagatani. Dynamical transitions to chaostic and periodic motions of two shuttle buses. *Physica A:Statistical Mechanics and its Applications*, Vol. 319, No. 1, pp. 568 – 578, 2003.
10. Takashi Nagatani. Chaos control and schedule of shuttle buses. *Physica A:Statistical Mechanics and its Applications*, Vol. 371, No. 2, pp. 683 – 691, 2006.
11. Toransportation bureau city of nagoya. http://www.kotsu.city.nagoya.jp/en/pc/OTHER/TRP0001448.htm. January 2018.
12. Meitetsu bus company limited. http://www.meitetsu-bus.co.jp/english/index.html. January 2018.
13. Jayakrishna, Patnaik Steven Chien, and Athanassios Bladikas. Estimation of bus arrival times using apc data. *Journal of Public Transportation*, Vol. 7, No. 1, pp. 1 – 20, 2004.
14. Mei Chen, Xiaobo Liu, and Jingxin Xia. Dynamic prediction method with schedule recovery impact for bus arrival time. *Transportation Research Record Journal of the Transportation Research Board*, pp. 208 – 217, 2005.
15. Alejandro Tirachini. Estimation of travel time and the benefits of upgrading the fare payment technology in urban bus services. *Transportation Research Part C: Emerging Technologies*, Vol. 30, pp. 239 – 256, 2013.
16. Location information service research agency. http://lisra.jp/en. January 2018.