TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2018 Volume 33(2): 543–554
doi:10.3846/16484142.2017.1298055

BUS ARRIVAL TIME PREDICTION USING MIXED MULTI-ROUTE ARRIVAL TIME DATA AT PREVIOUS STOP

Xuedong Hua[1], Wei Wang[2], Yinhai Wang[3], Min Ren[4]

[1, 2]*Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing, China*
[3]*Dept of Civil and Environmental Engineering, University of Washington, Seattle, United States*
[4]*Business School, Hohai University, Nanjing, China*

Submitted 25 December 2014; resubmitted 27 July 2015; accepted 29 October 2015; published online 02 May 2017

Abstract. The primary objective of this paper is to develop models to predict bus arrival time at a target stop using actual multi-route bus arrival time data from previous stop as inputs. In order to mix and fully utilize the multiple routes bus arrival time data, the weighted average travel time and three Forgetting Factor Functions (FFFs) – F1, F2 and F3 – are introduced. Based on different combinations of input variables, five prediction models are proposed. Three widely used algorithms, i.e. Support Vector Machine (SVM), Artificial Neutral Network (ANN) and Linear Regression (LR), are tested to find the best for arrival time prediction. Bus location data of 11 road segments from Yichun (China), covering 12 bus stops and 16 routes, are collected to evaluate the performance of the proposed approaches. The results show that the newly introduced parameters, the weighted average travel time, can significantly improve the prediction accuracy: the prediction errors reduce by around 20%. The algorithm comparison demonstrates that the SVM and ANN outperform the LR. The FFFs can also affect the performance errors: F1 is more suitable for ANN algorithm, while F3 is better for SVM and LR algorithms. Besides, the virtual road concept in this paper can slightly improve the prediction accuracy and halve the time cost of predicted arrival time calculation.

Keywords: bus arrival time prediction; multiple routes; support vector machine (SVM); artificial neutral network (ANN); linear regression (LR); forgetting factor function (FFF).

Introduction

Intelligent Transportation System (ITS) is becoming increasingly popular. In the field of public transit system, a couple of new technologies, such as Automatic Passenger Collection (APC), Automatic Vehicle Location (AVL) and Automatic Vehicle Identification (AVI), are used to provide better service to bus rides and enhance the Level Of Service (LOS) of public transit system. With the help of these technologies, bus rides can easily get transit-related information in real time. However, some incidents as well as the other running vehicles along public transit routes will affect the bus operation and thus leading to inaccurate information. Nowadays, the agencies are trying to provide passengers with more accurate predicted bus arrival information at stops using new algorithms and more precise data.

In well-developed cities, the traffic conditions, traffic flow patterns and characters, and the spatial and temporal distribution of bus rides stay stable. With the help of the historical traffic data, it is not too hard to predict the bus arrival time. However, in some other cities with heavy mixed traffic flow, bicycles and pedestrians have a great influence on the transit system, thus making it difficult to accurately predict the arrival time of buses. Besides, multiple transit routes running along one road segment will also affect the accuracy of arrival time prediction. How to collect, choose and process traffic data to predict bus arrival time more accurately, especially under the condition of heavy mixed traffic flow and multiple transit routes?

The improvement of traffic detection technology makes it possible to get the speed and location of buses in real time. The mixed traffic flow status, including bicycle and pedestrian flow, can also be automatic collected. Although the data collected from the new detection technologies and devices can make the time prediction easier, it is really an extra cost to add new equipment or update the exist ones. Suppose that the scale of a square city is 100 km^2 with a well-developed transit network, and the average bus stop spacing is 500 m, so the total

Corresponding author: Xuedong Hua
E-mail: *seu_huaxuedong@seu.edu.cn; qdurgk@163.com*

number of bus stops is 400. If the cost of each stop to update devices for bus arrival time prediction (such as HD cameras to monitor the arrival of buses at specific points and loop detectors to get the traffic flow status) is $2500, the total money spent will up to $ 1 million. It is a big sum of money. How to predict bus arrival time with the existing detection devices to lower the money cost?

This paper addresses the following research questions:

- 'How to collect, choose and process data to improve the bus arrival time prediction performance?';
- 'Which prediction models and algorithms perform best for bus arrival time prediction?'.

The rest of this paper is structured as follows: first, the literature review of bus arrival time prediction is outlined in section 1. Then, a set of models with three algorithms for bus arrival time prediction is setup in section 2. Section 3 demonstrates model calibration and evaluation processes with bus location data from Yichun (China). Results analysis and discussion are shown in Section 4 and 5, whilst conclusions are given in the end.

1. Literature

Bus arrival time at a certain bus stop is indeterminate and hard to predict because buses are affected by the overall dynamics of the transportation system (Horning *et al.* 2009), where changes occur on both regular (e.g., peak traffic jams) and random (e.g., accidents, special events) bases. Compared with the schedule, the actual arrival time of buses at stops fluctuates. Although the latest ITS technologies have greatly improved the performance of transit reliability, there still exists a gap between the scheduled time and the actual bus arrival time.

In the past decades, scholars have made great efforts to improve the reliability of bus arrival time prediction. Plenty of new models and algorithms have been developed, which can be mainly classified into four categories: Artificial Neural Network (ANN), Support Vector Machine (SVM), Kalman Filter (KF) and Non-Parametric Regression (NPR) or Linear Regression (LR) model.

ANN is motivated by emulating the intelligent data processing ability of human brains. Its prominent advantage for solving complex non-linear problems makes ANN popular in travel time predicting (Van Lint *et al.* 2005; Van Hinsbergen *et al.* 2009). Chen *et al.* (2004) developed a dynamic model that integrated the ANN and KF algorithms and used bus location data collected by APC system. Ding, Chien (2000) and Chien *et al.* (2002) proposed a link-based and a stop-based ANN model separately to predict bus arrival times in real time. They compared the performance of these two models and found both had their own advantages. Similar to Chien's *et al.* (2000) study, Yu *et al.* (2010) presented a hybrid model and found his model generally provides better performance than conventional ANN method. More recently, Lin *et al.* (2013) and Khetarpaul *et al.* (2015) proposed hybrid arrival time prediction models combining ANN and clustering methods to capture the traffic fluctuations and determine the parameter inputs of different clusters more clearly.

Similar to ANN, SVM is also a learning algorithm, which can map the inner relationship between the inputs and outputs (Cristianini, Shawe-Taylor 2000). SVM is special as it can manage a complex system with the corrupted data. It can even be used in times of training data shortage. Yu *et al.* (2006) built a SVM model and examined the feasibility and applicability of SVM in bus travel time forecasting. Later, Zheng *et al.* (2012) developed a multiple-stop prediction model and attempted to predict bus arrival times of the following multiple stops. The obtained results proved that his model was powerful to predict multiple stops arrival times at one time.

KF is a linear recursive predictive update algorithm used to estimate the parameters of a process model (Shalaby, Farhan 2004). By using AVL and APC dynamic data, Shalaby and Farhan (2004) tried to provide real-time information on bus arrival and departure time to bus rides with a KF bus travel time model. With the AVL data, Dailey *et al.* (2001) and Cathey, Dailey (2003) presented a KF based algorithm to predict transit vehicle arrival time up to one hour in advance. The prediction results for hundreds of locations were made widely available on the Web (MyBus HTML). Using ANN to infer decision rules from historical GPS data, Zaki *et al.* (2013) presented a KF-based model that fused prediction calculations with current GPS measurements.

Other methods including NPR and LR are not so popular, but they are quite simple in calibration and calculation (Park *et al.* 2007; Chang *et al.* 2010; Maiti *et at.* 2014). Patnaik *et al.* (2004) developed a set of regression models that estimate arrival times for buses traveling between two points along a route. Balasubramanian and Rao (2015) took cyclic variations in data into account, and proposed a NPR-based long-term bus arrival time prediction model.

Some details of previous studies that most related to bus arrival time prediction are listed in Table 1. As can be seen, ANN, SVM, and KF algorithms are generally applied when predicting the bus arrival time, while LR and NPR methods are less adopted. Some recent studies begin to integrate two or more algorithms together to make full use of goodness of each method. To the best of our knowledge, the existing literature is rarely found to predict bus arrival time using multiple routes data, except for the research done by Yu *et al.* (2011) – adopted all algorithms mentioned above separately to predict arrival time with multiple transit routes data. His study proves using multiple transit routes data can get a better prediction performance. However, although Yu made a progress in multiple transit routes data usage, some problems, such as whether multiple routes data are suitable for arrival time prediction under mixed traffic condition, how to determine the size of multiple routes data and how to use these data in a proper way, are still unclear and need to be discussed.

Table 1. Typical studies of bus arrival time prediction

Source	Model				Data from	
	ANN	SVM	KF	LR/NPR	Single route	Multiple routes
Chen *et al.* (2004)	✓				✓	
Chien *et al.* (2002)	✓				✓	
Ding, Chien (2000)	✓				✓	
Patnaik *et al.* (2004)				✓	✓	
Shalaby, Farhan (2004)			✓		✓	
Zheng *et al.* (2012)		✓	✓		✓	
Yu *et al.* (2011)	✓	✓	✓	✓		✓
Yu *et al.* (2010)		✓	✓		✓	
Yu *et al.* (2006)		✓			✓	
Cathey, Dailey (2003)			✓		✓	
Dailey *et al.* (2001)			✓		✓	
Lin *et al.* (2013)	✓				✓	
Khetarpaul *et al.* (2015)	✓				✓	
Maiti *et al.* (2014)				✓	✓	
Balasubramanian, Rao (2015)				✓	✓	
Zaki *et al.* (2013)			✓		✓	

2. Modelling

2.1. Detection Point Standardization

Currently, the data adopted for bus arrival time prediction are collected from some selected detection points. According to the locations of the detection points, there are 3 kinds of situations, as shown in Fig. 1:

- the detection point is between the target stop and the previous stop (location A);
- the detection point is just at the previous stop (location B);
- the detection point is located upstream of the previous stop (location C).

A further analyse and comparison are demonstrated as follows (situation 3 is similar to 2, so we do not analyse these two situations separately).

In situation 1, no other stops exist between location A and the target stop. When predicting the arrival time, only the bus running time needs to be taken into account. Bus arrival time prediction in this case can be calculated by:

$$\overline{T}^{target}_{k,arr} = T^{A}_{k,arr} + \overline{T}^{running}_{k,A-T}, \tag{1}$$

where: $\overline{T}^{target}_{k,arr}$ denotes the predicted arrival time at the target stop of bus k; $T^{A}_{k,arr}$ is the actual arrival time of bus k at location A; $\overline{T}^{running}_{k,A-T}$ denotes the predicted running time from location A to the target stop.

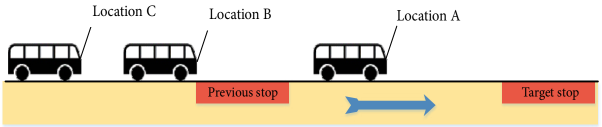

Fig. 1. An example of selected detection point

In situation 2, the bus is just arriving at the previous stop for boarding and alighting. Under this situation, not only the running time but the dwell time at previous stop should both be predicted. Arrival time of bus k can be obtained as follows:

$$\overline{T}^{target}_{k,arr} = T^{B}_{k,arr} + \overline{T}_{k,D} + \overline{T}^{running}_{k,B-T}, \tag{2}$$

where: $T^{B}_{k,arr}$ denotes the actual arrival time of bus k at location B; $\overline{T}_{k,D}$ is the estimated dwell time at the previous stop of bus k; $\overline{T}^{running}_{k,B-T}$ denotes the predicted running time between the location B and the target stop.

Normally, adding a prediction item will reduce the prediction accuracy, so situation 1 seems to be a wise choice. The previous studies also prove that situation 1 is more popular (Shalaby, Farhan 2004; Zheng *et al.* 2012; Yu *et al.* 2006, 2010, 2011). However, situation 2 has two unique advantages:

- compared with situation 1, choosing previous stop can fix the location of detection point. Although GPS-based systems (such as AVL) have been widely used to collect the bus locations, in order to guarantee the stability of data transfer and reduce the data size, it is a common practice to record the location data with a certain time interval. As a result, the recorded bus location cannot be exactly at the same location for each time. From this point of view, situation 2 is better due to its fixed detection points;
- it is possible to predict the bus arrival time at the next continuous multi-stop at one time. The actual arrival time at previous stop can be applied to predict the arrival time at the target stop, and the predicted arrival time then can be adopted to predict the arrival time at the next stop, with the

same prediction method. By doing this continuously, a series of predicted arrival times at downstream stops can be obtained.

In this paper, location B, the previous stop, is selected as the detection point. To simplify the prediction process, some transformations of the real road are made (shown in Fig. 2): the previous stop is removed and a virtual road is added instead. Virtual road is an imaginary road segment that the running time of buses at this unreal road is the same as the dwell time of buses at the previous stop. By doing this, the equation of bus arrival time prediction can be reduced to as similar as that of in situation 1:

$$\overline{T}_{k,arr}^{target} = T_{k,arr}^{previous} + \overline{T}_{k,travel}, \qquad (3)$$

where: $\overline{T}_{k,travel}$ denotes the predicted travel time at the total road segments (including the real road and the virtual road); $T_{k,arr}^{previous}$ is the actual arrival time of bus k at the start of the virtual road, $T_{k,arr}^{previous} = T_{k,arr}^{B}$.

Eq. (3) is the key formula in this paper to predict the bus arrival time.

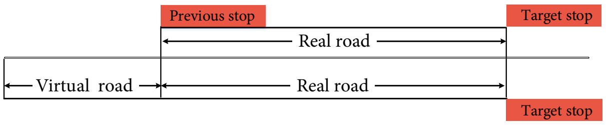

Fig. 2. Transformations of the real road

2.2. Framework of Prediction Model

In this paper, three factors (namely, input variables) are taken into account to predict the arrival time of bus k at a target stop, as follows:

- $T_{k+1,travel}^{same}$, total travel time (including the travel time at the real road and at the virtual road) of the preceding same route bus. Note that, the preceding same route bus means the bus from the same route of bus k, which is the closest to bus k and running at the downstream of the target stop.
- $T_{k+1,travel}^{diff}$, total travel time of the preceding different route bus. If only one bus route runs along the road segment, $T_{k+1,travel}^{diff}$ is equal to $T_{k+1,travel}^{same}$.
- $\overline{T}_{k,n}^{ave}$, the weighted average travel time of the preceding n buses. The preceding n buses denotes the closest n buses to bus k, which are just passed through the target stop.

Compared with further buses, the closer preceding buses share more similar traffic information with bus k. To enhance the contribution of the closer buses, Forgetting Factor Function (FFF) is proposed. If $T_{k+j}^{previous}$ denotes the actual arrival time of the j^{th}-closest preceding bus at the start of the virtual road, $h_{k,k+j}$ denotes the time headway between the target bus k and the j^{th}-closest preceding bus, one can calculate FFF with the following equations:

$$f_1\left(h_{k,k+j},n\right) = \frac{1/h_{k,k+j}}{\sum_{i=1}^{n} 1/h_{k,k+i}};$$

$$f_2\left(h_{k,k+j},n\right) = \frac{1/\sqrt{h_{k,k+j}}}{\sum_{i=1}^{n} 1/\sqrt{h_{k,k+i}}};$$

$$f_3\left(h_{k,k+j},n\right) = \frac{1/\ln h_{k,k+j}}{\sum_{i=1}^{n} 1/\ln h_{k,k+i}}. \qquad (4)$$

By introducing FFF, one can calculate the weighted average travel time as follows:

$$\overline{T}_{k,n}^{ave} = \sum_{i=1}^{n} f\left(h_{k,k+i},n\right) T_{k+i,travel}, \qquad (5)$$

where: $T_{k+i,travel}$ is the actual total travel time of the i^{th}-closest preceding bus at the total road segments.

The prediction travel time $\overline{T}_{k,travel}$ in Eq. (3) is then generalized as a function of the above three factors:

$$\overline{T}_{k,travel} = f\left(T_{k+1,travel}^{same}, T_{k+1,travel}^{diff}, \overline{T}_{k,n}^{ave}\right). \qquad (6)$$

Substituting Eq. (6) into Eq. (3), the bus arrival time prediction model proposed in this paper can be obtained:

$$\overline{T}_{k,arr}^{target} = T_{k,arr}^{previous} + f\left(T_{k+1,travel}^{same}, T_{k+1,travel}^{diff}, \overline{T}_{k,n}^{ave}\right). \qquad (7)$$

2.3. Prediction Algorithm and Model Setup

To acquire the predicted bus travel time in the proposed model, SVM and ANN algorithms are applied. These two algorithms share the principles of non-linear, distributed, parallel and local processing and adaptation, which make them suitable for bus arrival time prediction. To simplify the process of modelling and improve the reliability of results, two widely-used encapsulated toolboxes are adopted. LIBSVM, a powerful software package developed by Chang and Lin (1989) for SVM classification, regression and distribution estimation, is used for SVM algorithm modelling, while the ANN toolbox in *Matlab 2010a* is introduced to train neural networks. For prediction results comparison, the simplest algorithm, LR is also tested in this paper.

Five bus arrival time prediction models are deployed with different combination of input variables. Table 2 shows detailed information of each model. Note that, model 4 is with only one variable $T_{k+1,travel}^{same}$, which is the same to previous studies.

Table 2. Model setup with different input variables

Model No	Input variables		
	$T_{k+1,travel}^{same}$	$T_{k+1,travel}^{diff}$	$\overline{T}_{k,n}^{ave}$
1	✓	✓	✓
2	✓		✓
3		✓	✓
4	✓		
5			✓

3. Model Calibration and Evaluation

3.1. Data Collection

The five proposed arrival time prediction models are tested using the data collected from Yichun (China). Yichun, located in southeast China, is an inland city with an area of 18700 km^2 and a population of 5.5 million. The per capita GDP of Yichun is $ 1814 in 2011. According to a sampling survey in 2011, the total number of travel made by public transit system is 98000 times per day (accounting for 7.8% of the total travel times. Private car, bicycle, motorcycle and transit are the first four most widely used travel modes in Yichun). In Yichun, there are 29 transit routes which cover 100% of the downtown area and are all equipped with AVL system. The traffic control and management are very poor in the downtown area: during the peak hour traffic congestion is quite common in the downtown area (Fig. 3b), while in off peak time the traffic flow runs smoothly (Fig. 3a). The dramatic fluctuation of traffic volume and speed makes bus arrival time prediction more difficult.

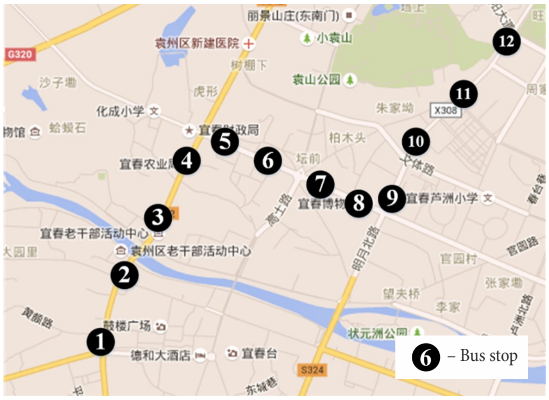

Fig. 4. Selected Bus stops for arrival time prediction

In the proposed prediction model, the bus arrival time at the start of the virtual road (namely, at the previous stop) is the only input. Therefore, the arrival time data from the AVL database and the location of every stops are picked up. Twelve stops are selected for arrival time prediction, shown in Fig. 4. Stop ①–④, ⑤–⑧ and ⑨–⑫ are located at Yichun Avenue (with heavy mixed traffic), Yuanshang Road (with moderate mixed traffic) and Mingyue Avenue (with moderate mixed traffic), respectively. The first eight stops are in old town area while stop ⑨–⑫ are in new build Central Business District (CBD) area. Bus arrival time data at these twelve stops from five weekdays (19–23 September 2011) are picked up by programming with *Microsoft Visual Basic 6.0*. Some more detailed information about the collected data and the bus routes are illustrated in Table 3 and Appendix.

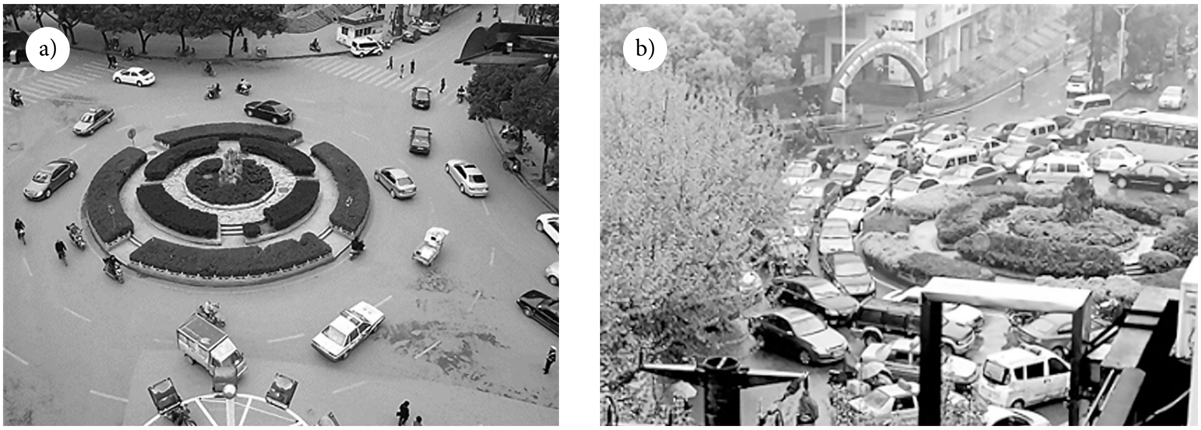

Fig. 3. Traffic condition during weekdays: a – in off peak hour; b – in peak hour

Table 3. Detailed information of each road segment

Road segment No	From stop	To stop	Length [m]	Bus routes along the segment	No of collected data
1	①	②	495	1–2–8–17–116–118	2581
2	②	③	500	1–2–8–10–17	2369
3	③	④	420	1–2–3–8–10–17	2703
4	④	⑤	340	2–3–10	1081
5	⑤	⑥	430	2–3–6–7–10	2109
6	⑥	⑦	460	2–10	767
7	⑦	⑧	330	2–5–10–21	1336
8	⑧	⑨	320	2–5–10	666
9	⑨	⑩	290	2–5–7–12–20	1243
10	⑩	⑪	530	2–5–7–12–20	1193
11	⑪	⑫	520	5–7–11–12–13–20	1942

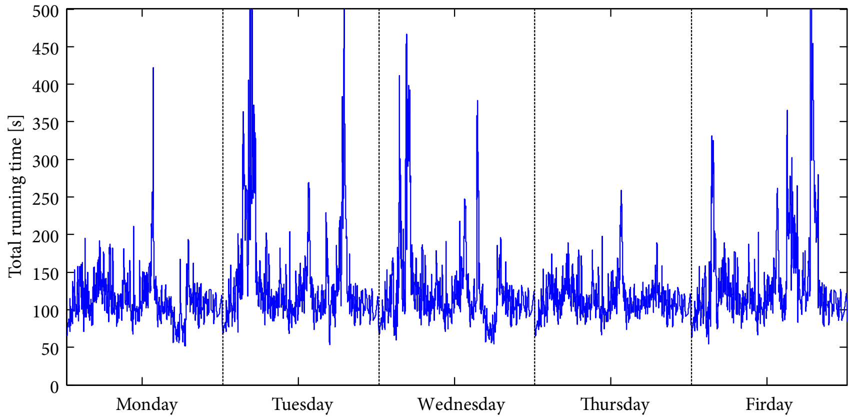

Fig. 5. Cyclical pattern of travel time from stop ① to ②

Fig. 5 displays bus travel time of all routes from stop ① to ② during the five consecutive typical weekdays. From Fig. 5, travel time shows a cyclical pattern over different weekdays. It is evident that bus travel time increases significantly in the morning peak, at noon and in afternoon peak. During the other hours, the bus travel time remains around normal values and fluctuates slightly. The bus travel time during weekdays is periodic and recurs every days. This cyclical pattern of bus travel time in Yichun is quite similar to that of speed and volume in freeways in previous studies (Zou *et al.* 2014; Zhang *et al.* 2014; Xia *et al.* 2011). Note that the cyclical travel time patterns are also observed at other 10 segments. Taken the cyclical pattern into account, it is feasible to predict bus arrival time with historical data.

3.2. Model Calibration and Evaluation

Model calibration is applied to find the optimal parameters for the proposed models that can best predicted the bus arrival time. In model calibration step, the historical bus arrival time data picked up of all the selected segments are divided into two parts: 80% are training data that are used for model calibration, and the rest 20% are adopted to test the prediction performance of each model.

For SVM algorithm, previous studies (Yu *et al.* 2006, 2010) suggested that Radial Basis Function (RBF) kernel was efficient for bus arrival time prediction. So in this paper, RBF is also selected as kernel for SVM. Since the calibration of parameters in SVM has a great influence on the accuracy of prediction, grid-search method is introduced to calibrate SVM. For ANN algorithm, the number of layers and hidden neurons have a great influence on the training speed of ANN as well as the outcomes. By enumeration, a four-layer ANN (an input layer, two hidden layers with 10 hidden neurons and 5 hidden neurons separately, and an output layer) are selected in this study.

To evaluation the performance of each model, three frequently-used performance measures, i.e. Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) are introduced. The equations for these performance measures are as follows:

$$MAE = \frac{\sum \left| \overline{T}_{k,travel} - T_{k,travel} \right|}{N};$$

$$MAPE = \frac{1}{N} \sum \frac{\left| \overline{T}_{k,travel} - T_{k,travel} \right|}{T_{k,travel}};$$

$$RMSE = \sqrt{\frac{\sum \left(\overline{T}_{k,travel} - T_{k,travel} \right)^2}{N-1}}, \qquad (8)$$

where: $T_{k,travel}$ is the actual total travel time.

4. Result Analysis

After model calibration and calculation, the predicted bus arrival time at eleven stops (from stop ② to ⑫) are obtained. The performance of each model and algorithm are evaluated by the values of MAE, MAPE and RMSE (Table 4) demonstrate the MAE, MAPE and RMSE of all models, algorithms and stops. Note that in the tables, the model names are made up with a letter and a number. The letter in upper case represents the algorithm, while the number indicates the model. L3, for example, means model 3 with LR algorithm. The following analysis and discussion of prediction performance are all based on Table 4.

4.1. Performance Comparison of Five Models

Although the performance of the proposed five models vary with respect to road segments and algorithms in Table 4, it is evident that different input variables can contribute to different prediction performance. Fig. 6 demonstrates the performance of the models in a simple way. The values of MAE, MAPE and RMSE in Fig. 6 are the average of those in Table 4 to eliminate the influence of road segments and algorithms.

From Fig. 6 it is easy to recognize the introduction of the weighted average travel time improve the performance of bus arrival time prediction. The conventional method, model 4, performs the poorest with the largest prediction errors. Compared with model 4, the other four models show a significant advantage. The perfor-

Table 4. MAE [s], MAPE [%] and RMSE [s]

Segment \ Model	S1	S2	S3	S4	S5	A1	A2	A3	A4	A5	L1	L2	L3	L4	L5
MAE [s]															
1	25.02	25.50	25.54	33.57	26.12	13.89	14.23	12.04	33.18	13.02	23.71	23.70	23.93	39.09	23.92
2	31.07	31.11	31.34	37.62	31.88	30.60	32.55	32.33	40.25	32.68	30.69	30.57	32.26	43.16	32.15
3	19.63	19.75	19.84	24.62	20.98	19.76	19.64	20.24	24.49	20.41	21.20	21.21	22.46	27.38	22.52
4	26.74	26.48	26.49	32.54	26.49	27.38	28.67	29.52	34.34	30.18	28.30	28.38	28.38	37.07	28.47
5	12.65	12.63	12.68	14.31	12.58	12.60	12.57	12.33	14.20	13.07	13.52	13.52	13.94	16.85	13.99
6	51.08	51.52	51.52	56.58	54.32	53.61	53.78	52.22	55.34	54.71	55.92	55.92	55.92	64.87	56.66
7	10.24	10.17	10.30	11.50	10.23	10.63	10.56	10.90	11.89	11.92	11.45	11.48	11.65	13.94	11.65
8	24.14	27.24	26.28	28.97	30.98	26.86	33.33	29.75	32.62	30.64	33.72	34.55	35.04	38.02	35.49
9	11.43	11.53	13.11	11.76	13.18	10.88	11.66	12.69	11.93	13.09	12.69	12.69	14.71	14.09	14.86
10	31.12	31.40	32.26	37.34	32.85	33.38	33.52	33.11	39.01	32.04	35.84	35.82	37.26	47.75	37.21
11	26.80	26.74	26.87	32.69	26.75	27.58	27.48	27.29	32.15	27.88	27.85	27.84	28.71	37.08	28.80
MAPE [%]															
1	15.56	15.66	15.99	19.56	15.94	7.53	8.39	8.20	20.70	8.25	16.19	16.11	16.35	24.16	16.29
2	15.69	15.78	15.95	18.47	16.02	16.31	16.63	16.64	20.22	16.67	16.35	16.31	17.03	22.28	16.98
3	15.26	15.44	15.67	18.41	16.32	16.27	16.22	16.70	19.87	16.70	17.56	17.57	18.73	22.54	18.75
4	20.71	21.06	20.90	24.21	21.39	22.93	23.78	23.87	27.93	24.69	23.20	23.48	23.24	29.33	23.52
5	14.95	14.98	15.09	16.37	15.06	15.43	15.35	15.12	16.89	15.74	16.58	16.58	17.18	20.17	17.22
6	26.38	26.36	26.36	29.98	28.23	29.52	30.20	29.12	31.40	30.74	31.44	31.44	31.44	35.37	31.92
7	15.21	15.07	15.23	16.19	15.06	15.72	15.58	15.67	17.26	16.42	17.38	17.42	17.59	20.49	17.59
8	21.82	26.76	25.78	27.74	29.86	26.52	32.19	30.01	32.24	29.32	33.95	34.78	35.34	36.80	35.88
9	18.76	18.99	20.10	19.25	20.26	18.66	19.54	20.92	20.15	21.20	21.65	21.56	24.50	23.52	24.53
10	23.75	23.80	24.59	27.24	24.93	27.41	27.09	27.05	30.19	26.06	29.51	29.52	30.13	37.36	30.12
11	18.67	18.69	18.84	21.87	18.84	20.72	20.45	20.22	23.31	20.68	20.50	20.49	21.03	26.73	21.09
RMSE [s]															
1	52.28	54.06	52.50	67.82	58.24	55.50	41.75	27.60	66.65	31.27	41.41	41.70	41.42	73.52	41.76
2	54.24	54.52	53.42	66.74	56.40	48.76	57.35	57.83	75.20	58.64	49.11	48.72	51.49	69.20	50.59
3	37.87	37.84	36.68	46.73	39.13	36.45	36.22	36.77	44.46	37.71	39.23	39.19	41.70	48.14	41.40
4	50.90	49.46	49.95	61.81	49.48	50.96	52.65	55.56	63.77	56.17	50.04	50.19	50.11	63.27	50.24
5	20.76	20.48	20.37	24.65	19.88	20.46	20.23	19.29	23.89	21.42	20.19	20.19	20.37	26.87	20.31
6	93.48	93.80	93.80	92.67	94.72	93.49	93.88	91.06	87.04	92.27	92.61	92.61	92.61	99.43	92.72
7	18.72	18.63	18.89	21.62	18.81	19.75	19.74	21.71	21.84	26.60	19.69	19.74	19.80	23.79	19.80
8	42.57	44.93	43.51	46.90	47.42	43.79	54.20	44.95	54.20	48.80	50.05	50.34	51.02	56.34	51.01
9	17.92	18.08	20.68	18.58	20.77	16.37	17.71	19.35	18.34	19.83	17.97	17.98	20.98	20.10	21.11
10	48.99	48.61	49.04	60.78	49.54	49.45	51.09	49.35	61.38	47.41	51.88	51.69	52.76	69.81	52.55
11	40.09	40.01	40.20	52.31	40.13	39.65	39.28	39.09	49.52	39.82	40.97	40.97	41.44	55.95	41.47

mance improvements of the models with the weighted average travel time as input over the conventional model are all around 20%. Model 1, who has the most input variables, shows a tenuous advantage with respect to MAE and MAPE over model 2, 3 and 5. Although the MAE and MAPE of model 3 rank third among the five models, the RMSE of model 3 is the smallest, which indicates the number of larger prediction errors of model 3 is less than those of the other models. More remarkable, model 5 greatly overwhelms model 4 in terms of MAE, MAPE and RMSE, even if both of them are with only one input.

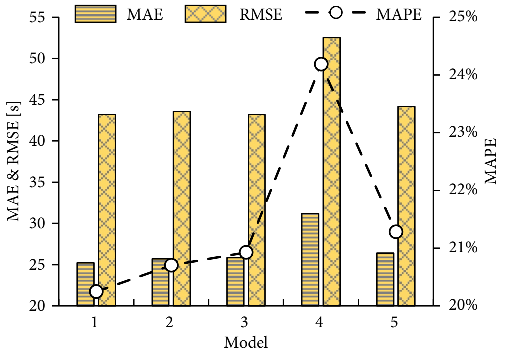

Fig. 6. Prediction performance of five models

Prediction results demonstrate that the introduction of the weighted average travel time can significant reduce the prediction errors, whatever the road segments and algorithms the models work with. Model 1 proves to be the optimal bus arrival time prediction model with respect to MAE, MAPE and RMSE. Model 5, which use the weighted average travel time as the only input, can be an alternative for arrival time prediction due to its fewest input and small errors. The most widely used prediction method, model 4, is the worst model.

4.2. Performance Comparison of Three Algorithms

Fig. 7 illustrates the performance of the three algorithms. It is obvious LR is the worst algorithm when compared with the other two algorithms from Fig. 7a. The prediction errors of LR algorithm is about 10% larger. Compared with ANN, SVM shows a tenuous advantage with respect to MAE and MAPE. The RMSE of ANN is the smallest among the three algorithms, indicating ANN algorithm is good at reducing the number of larger prediction errors.

As mentioned above, model 1 is the best prediction model. So in this subsection, the performance of the three algorithms with model 1 are demonstrated as well (in Fig. 7b). SVM with model 1 remains the best prediction algorithm while the performance of LR with model 1 is the poorest, which is consistent with the performance in Fig. 7a. However, instead of by ANN, the smallest RMSE is caused by LR algorithm with model 1. That means the prediction performance improvement caused by model 1 on LR algorithm is greater, although the SVM1 and ANN1 algorithms still preforms better as expected.

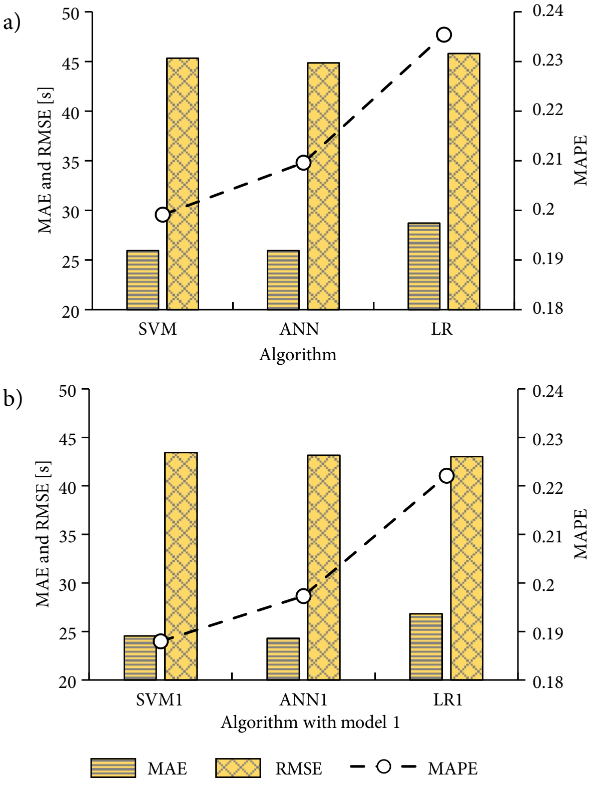

Fig. 7. Prediction performance of three algorithms:
a – algorithms with all models;
b – algorithms with only model 1

In summary, the two non-linear algorithms, SVM and ANN, outperform the LR algorithm. Although LR is still the worst bus arrival time prediction algorithm, the performance gap between LR and the two non-linear ones narrows when model 1 is introduced.

4.3. Performance Comparison under Different Traffic Conditions

Table 5 demonstrates the prediction performance under heavy and moderate mixed traffic. In the table, the values of prediction errors in the heavy mixed traffic column are the average predicted results of road segments 1–3, while the values in the moderate mixed traffic column are the average results of road segments 4–11.

Table 5. Performance under different traffic conditions

	Heavy mixed traffic	Moderate mixed traffic
MAE	26.24	27.11
MAPE	0.17	0.23
RMSE	48.87	44.00

It is evident the proposed models and algorithms performs well at both heavy and moderate mixed traffic: the values of MAE, MAPE and RMSE are very close. From MAE and MAPE, the models under heavy mixed traffic outperform under moderate mixed traffic. This is because in the old town area (namely, segments 1–3), few intersections are controlled by signal, while in Yuanshan Road and Mingyue Avenue, all the main intersections are signal controlled. This difference in traffic control makes the prediction performance under moderate mixed traffic is slight poorer than that under heavy mixed traffic. In addition, the lower value of RMSE under moderate mixed traffic indicates that the number of larger prediction errors under moderate mixed traffic is less, than those under heavy mixed traffic.

5. Discussion

5.1. Prediction Performance with Virtual Road vs. without Virtual Road

The concept of virtual road is proposed to predict bus dwell time at the previous stop and travel time at the real road segment together with only one prediction model. Although the introduction of virtual road can reduce the prediction calculation time by a half, the influence of virtual road on prediction preference should be further discussed.

Fig. 8a–f demonstrate the performance comparison between prediction with virtual road and without virtual road. One road segment with more bus routes, larger passenger demand and heavy mixed traffic (namely, segment 1, shown in Fig. 8a–c) and another segment with fewer bus routes, smaller passenger demand and moderate mixed traffic (segment 7, in Fig. 8d–f) are chosen to evaluate the influence of virtual road on bus arrival time prediction with different traffic conditions.

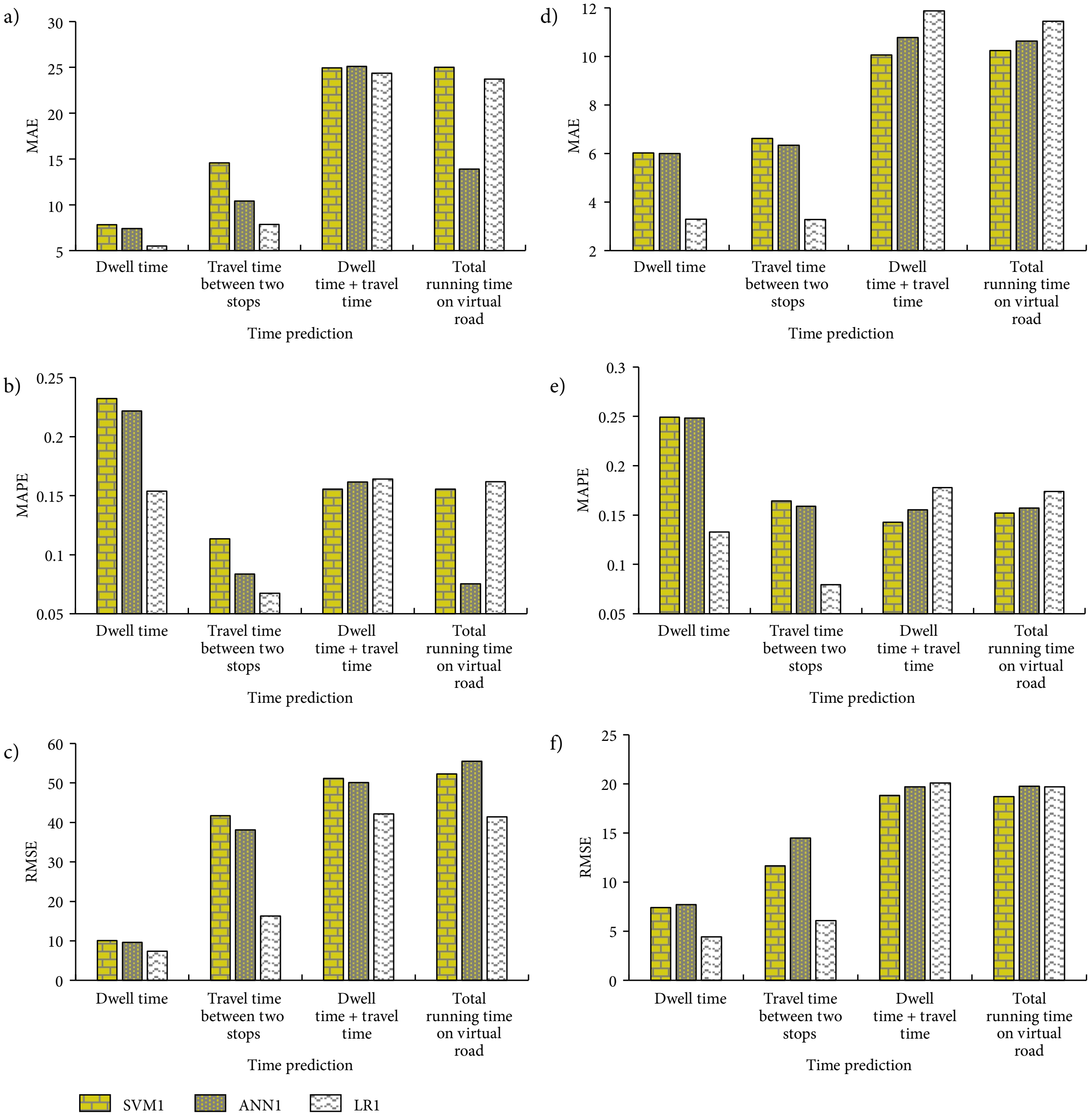

Fig. 8. Performance comparison with virtual road vs. without virtual road: a – MAE of road segment 1; b – MAPE of road segment 1; c – RMSE of road segment 1; d – MAE of road segment 7; e – MAPE of road segment 7; f – RMSE of road segment 7

Without virtual road, the predicted bus arrival time can be obtained by adding the predicted dwell time and predicted travel time. So the MAE and RMSE of predicted bus arrival time are larger than those of the predicted dwell time and travel time (Fig. 8a–f), while the MAPE performance of predicted bus arrival time is between that of travel time and dwell time. When the virtual road is introduced, the performance of bus arrival time prediction is slight better than or comparable to that without virtual road. The performance improvements of virtual road introduction are mostly around 3% for road segment 7, and 2% for road segment 1. More remarkable, for road segment 1, the largest improvement can over 53% (ANN1 in Fig. 8b). Besides the advantage of prediction error reduction, because dwell time and travel time are not required to be predicted separately, about a half of calculation time will be saved with virtual road.

The comparison results indicate the introduction of virtual road is beneficial. It does not result in prediction errors increasing at most times, and may even slightly reduce prediction errors. Besides, the time cost of predicted bus arrival time calculation will be halved as well.

5.2. Performance of Multi-Stop Arrival Time Prediction

As discussed in section 2, using the previous stop as the detection point can make it possible to predict bus arrival time at the next continuous multi-stop at one time with the same model. In this subsection, the performance of multi-stop bus arrival time prediction are evaluated. When buses arrive at stop ①, the arrival time at stop ②, stop ④, stop ⑥ and stop ⑧ are predicted using model 1. The prediction results are illustrated in Fig. 9.

When predicting the arrival time at the second stop, the prediction errors are acceptable: MAE, MAPE and RMSE are 15.59S, 7.76% and 74.98S, respectively (ANN1). When the target stops are further, all the indicators increase. MAE, MAPE and RMSE of the predicted arrival time at the furthest target stop (stop 8) dramatically increase to 391.61 s, 23.15%, 1393.21 s (ANN1). Although the change of percentage error from 7.76% to 23.15% seems not big, around 6.5 min ahead or behind of the actual bus arrival time cannot satisfy passengers.

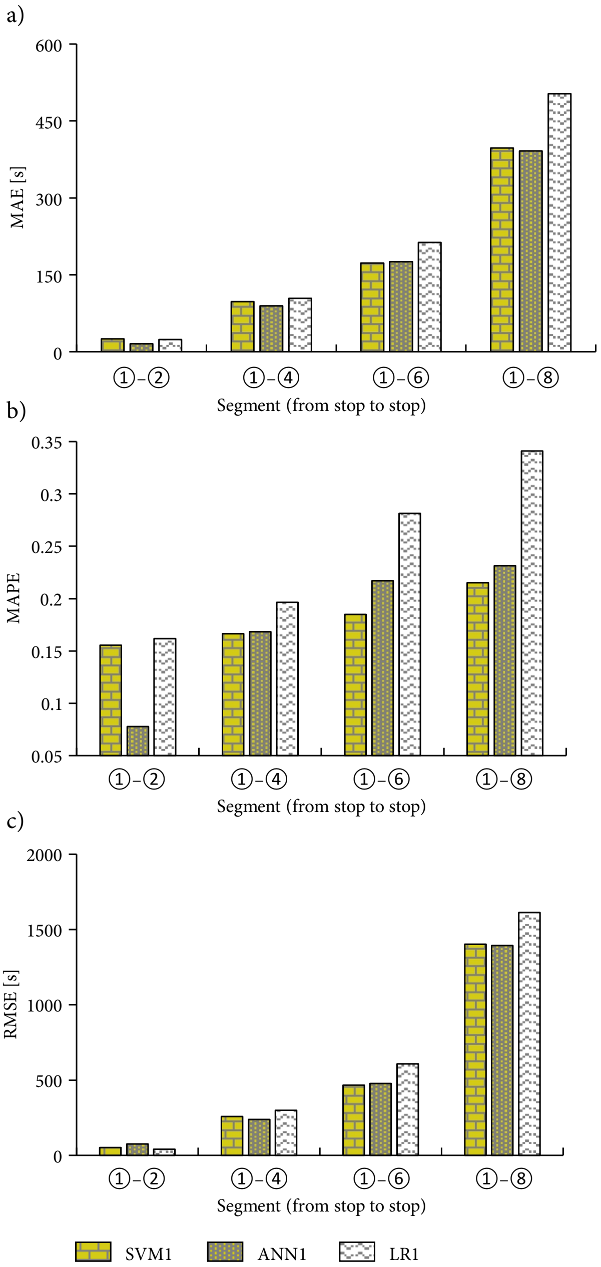

Fig. 9. Performance of multi-stop arrival time prediction: a – performance in MAE; b – performance in MAPE; c – performance in RMSE

The poor performance of predicted arrival time at some further stops make it to be nothing but only a referenced value for passengers in practise. To make it more meaningful to bus rides, the predicted time should be revised and updated continuously. When bus arrives at the first stop, the arrival time at all the downstream stops can be predicted, no matter accuracy or not. Then, when the bus arrives at the next stop, the new arrival time can be applied to the prediction model, and the arrival time at all the downstream stops can be re-predicted and updated. By doing this consistently, the predicted bus arrival time at each stop can be announced and updated more accurately and in real time.

5.3. Influence of FFF

The introduction of the weighted average travel time and FFF greatly improve the performance of bus arrival time prediction. In this part, some further discussion about the influence of FFF on each algorithm with model 1 is conducted. Table 6 shows the prediction errors at road segment 1 with different FFF as well as different values of parameter *n*.

It can be observed the influence of different FFFs is slight: the differences of prediction errors among F1, F2 and F2 are within 2%, and varies from algorithms to algorithms. F3 is more suitable for SVM and LR, while F1 can reduce the prediction errors for ANN. The value of parameter *n* can affect the prediction performance as well. The smallest MAPEs of SVM1 are with parameter $n = 2$, While the smallest MAPEs of ANN1 and LR1 are with parameter $n = 3$ and parameter $n = 4$, respectively.

In summary, the best FFFs for SVM1, ANN1 and LR1 in this study are F3 with parameter $n = 2$, F1 with parameter $n = 3$ and F3 with parameter $n = 4$, respectively.

Conclusions

This paper tried to improve bus arrival time prediction performance using actual bus arrival time data from multiple bus routes. $\overline{T}_{k,n}^{ave}$, the weighted average travel time, was proposed as a new input to mix and fully utilize the travel time date from multiple routes. In order to tell the different effect of contribution of preceding buses on the weighted average travel time calculation, three FFFs were introduced. Based on different combinations of input variables, five prediction models are proposed. Three widely used algorithms, i.e. SVM, ANN and LR, were tested to find the best for bus arrival time prediction. To test and evaluate the performance of the proposed approaches, AVL data from Yichun city, covering 11 road segments, 12 stops and 16 bus routes, were collected. The results demonstrate that the introduction of weighted average travel time can significantly improve the prediction accuracy. The proposed model 1, which contains the most information, outperforms the other four models. The prediction improvement of model 1 over the conventional model 4 is around 20%. The non-

Table 6. Prediction errors of FFFs with parameters *n*

FFF	*n*	SVM1			ANN1			LR1		
		MAE [s]	MAPE [%]	RMSE [s]	MAE [s]	MAPE [%]	RMSE [s]	MAE [s]	MAPE [%]	RMSE [s]
F1	4	17.55	13.69	22.64	17.32	13.79	22.38	**18.29**	**14.38**	**23.68**
	3	17.64	13.68	22.96	**17.18**	**13.68**	**21.95**	18.443	14.47	23.88
	2	**17.37**	**13.55**	**22.41**	17.97	14.28	22.73	18.58	14.59	24.04
F2	4	17.62	13.79	22.73	17.44	13.98	22.45	**18.18**	**14.33**	**23.57**
	3	17.78	13.80	23.05	**17.19**	**13.60**	**22.07**	18.35	14.42	23.81
	2	**17.54**	**13.62**	**22.92**	17.68	13.97	22.88	18.53	14.55	24.01
F3	4	17.48	13.70	22.60	17.68	14.15	22.85	**18.10**	**14.30**	**23.47**
	3	17.77	13.83	23.02	**17.51**	**13.86**	**22.61**	18.29	14.38	23.76
	2	**17.48**	**13.57**	**22.86**	17.70	14.08	22.70	18.51	14.53	23.99

linear algorithms, SVM and ANN are better than LR, for their good prediction accuracy.

Both of the two newly-introduced items in arrival time prediction models, namely virtual road and FFFs, are beneficial. The virtual road concept can slightly improve the prediction accuracy and halve the time cost of predicted arrival time calculation. FFFs can also affect the performance of each algorithms: F1 is more suitable for ANN algorithm, while F3 is better for SVM and LR algorithms.

The contribution of this paper is the development of the models to predict the bus arrival time with data from multiple routes. By the models, one can obtain the predicted bus arrival time with higher accuracy. Compared with the conventional methods, the extra work needs to be done is to mix bus arrival time data from multiple routes together.

Acknowledgements

This work was supported by National Key Basic Research Program of China (No 2012CB725402), the State Key Program of National Natural Science of China (No 51338003) and Scientific Research Foundation of Graduate School of Southeast University.

References

Balasubramanian, P.; Rao, K. R. 2015. An adaptive long-term bus arrival time prediction model with cyclic variations, *Journal of Public Transportation* 18(1): 1–18. https://doi.org/10.5038/2375-0901.18.1.6

Cathey, F. W.; Dailey, D. J. 2003. A prescription for transit arrival/departure prediction using automatic vehicle location data, *Transportation Research Part C: Emerging Technologies* 11(3–4): 241–264. https://doi.org/10.1016/S0968-090X(03)00023-8

Chang, C.-C.; Lin, C.-J. 1989. *LIBSVM: a Library for Support Vector Machines*. Available from Internet: https://www.csie.ntu.edu.tw/~cjlin/libsvm

Chang, H.; Park, D.; Lee, S.; Lee, H.; Baek, S. 2010. Dynamic multi-interval bus travel time prediction using bus transit data, *Transportmetrica* 6(1): 19–38. https://doi.org/10.1080/18128600902929591

Chen, M.; Liu, X.; Xia, J.; Chien, S. I. 2004. A dynamic bus-arrival time prediction model based on APC data, *Computer-Aided Civil and Infrastructure Engineering* 19(5): 364–376. https://doi.org/10.1111/j.1467-8667.2004.00363.x

Chien, S. I.; Ding, Y.; Wei, C. 2002. Dynamic bus arrival time prediction with artificial neural networks, *Journal of Transportation Engineering* 128(5): 429–438. https://doi.org/10.1061/(ASCE)0733-947X(2002)128:5(429)

Cristianini, N.; Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press. 204 p.

Dailey, D.; Maclean, S.; Cathey, F.; Wall, Z. 2001. Transit vehicle arrival prediction: algorithm and large-scale implementation, *Transportation Research Record: Journal of the Transportation Research Board* 1771: 46–51. https://doi.org/10.3141/1771-06

Ding, Y.; Chien, S. I. 2000. *The Prediction of Transit Arrival Times Using Link-Based and Stop-Based Artificial Neural Networks*. New Jersey Institute of Technology, Newark, US. 5 p.

Horning, J.; El-Geneidy, A. M.; Hourdos, J. 2009. *Estimating Running Time and Demand for a Bus Rapid Transit Corridor*. Report No CTS 09-24. University of Minnesota, US. 69 p. Available from Internet: http://www.its.umn.edu/Publications/ResearchReports/reportdetail.html?id=1852

Khetarpaul, S.; Gupta, S. K.; Malhotra, S.; Subramaniam, L. V. 2015. Bus arrival time prediction using a modified amalgamation of fuzzy clustering and neural network on spatio-temporal data, *Lecture Notes in Computer Science* 9093: 142–154. https://doi.org/10.1007/978-3-319-19548-3_12

Lin, Y.; Yang, X.; Zou, N.; Jia, L. 2013. Real-time bus arrival time prediction: case study for Jinan, China, *Journal of Transportation Engineering* 139(11): 1133–1140. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000589

Maiti, S.; Pal, Arp.; Pal, Ari.; Chattopadhyay, T.; Mukherjee, A. 2014. Historical data based real time prediction of vehicle arrival time, in *2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, 8–11 October 2014, Qingdao, China, 1837–1842. 0https://doi.org/10.1109/ITSC.2014.6957960

Park, S. H.; Jeong, Y. J.; Kim, T. J. 2007. Transit travel time forecasts for location-based queries: implementation and evaluation, *Proceedings of the Eastern Asia Society for Transportation Studies* 6: 1–11. https://doi.org/10.11175/eastpro.2007.0.237.0

Patnaik, J.; Chien, S.; Bladikas, A. 2004. Estimation of bus arrival times using APC data, *Journal of Public Transportation* 7(1): 1–20. https://doi.org/10.5038/2375-0901.7.1.1

Shalaby, A.; Farhan, A. 2004. Prediction model of bus arrival and departure times using AVL and APC data, *Journal of Public Transportation* 7(1): 41–61. https://doi.org/10.5038/2375-0901.7.1.3

Van Hinsbergen, C. P.; Van Lint, J. W. C.; Van Zuylen, H. J. 2009. Bayesian committee of neural networks to predict travel times with confidence intervals, *Transportation Research Part C: Emerging Technologies* 17(5): 498–509. https://doi.org/10.1016/j.trc.2009.04.007

Van Lint, J. W. C.; Hoogendoorn, S. P.; Van Zuylen, H. J. 2005. Accurate freeway travel time prediction with state-space neural networks under missing data, *Transportation Research Part C: Emerging Technologies* 13(5–6): 347–369. https://doi.org/10.1016/j.trc.2005.03.001

Xia, J.; Chen, M.; Huang, W.; 2011. A multistep corridor travel-time prediction method using presence-type vehicle detector data, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 15(2): 104–113. https://doi.org/10.1080/15472450.2011.570114

Yu, B.; Lam, W. H. K.; Tam, M. L. 2011. Bus arrival time prediction at bus stop with multiple routes, *Transportation Research Part C: Emerging Technologies* 19(6): 1157–1170. https://doi.org/10.1016/j.trc.2011.01.003

Yu, B.; Yang, Z.; Chen, K.; Yao, B. 2006. Bus arrival time prediction using support vector machines, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 10(4): 151–158. https://doi.org/10.1080/15472450600981009

Yu, B.; Yang, Z.-Z.; Chen, K.; Yu, B. 2010. Hybrid model for prediction of bus arrival times at next station, *Journal of Advanced Transportation* 44(3): 193–204. https://doi.org/10.1002/atr.136

Zaki, M.; Ashour, I.; Zorkany, M.; Hesham, B. 2013. Online bus arrival time prediction using hybrid neural network and Kalman filter techniques, *International Journal of Modern Engineering Research* 3(4): 2035–2041.

Zhang, Ya.; Zhang, Yu.; Haghani, A. 2014. A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model, *Transportation Research Part C: Emerging Technologies* 43(1): 65–78. https://doi.org/10.1016/j.trc.2013.11.011

Zheng, C.-J.; Zhang, Y.-H.; Feng X.-J. 2012. Improved iterative prediction for multiple stop arrival time using a support vector machine, *Transport* 27(2): 158–164. https://doi.org/10.3846/16484142.2012.692710

Zou, Y.; Zhu, X.; Zhang, Y.; Zeng, X. 2014. A space–time diurnal method for short-term freeway travel time prediction, *Transportation Research Part C: Emerging Technologies* 43(1): 33–49. https://doi.org/10.1016/j.trc.2013.10.007

APPENDIX

Basic information about bus routes

Route No	Service hours	Headway [min]
1	6 am–8 pm	7
2	6 am–10 pm	4
3	6:30 am~8 pm	8
5	6 am–9 pm	5
6	6 am–8 pm	6
7	6 am–7 pm	8
8	6 am–8:30 pm	7
10	6:20 am–8 pm	
11	6 am–7:30 pm	7
12	6 am–6:30 pm	10
13	6 am–6:30 pm	20
17	6 am–7:30 pm	7
20	6 am–8 pm	8
21	6:40 am–6:30 pm	
116	6 am–6:30 pm	15–20
118	6 am–5 pm	40