An Adaptive Bus Arrival Time Prediction Model

YU, Bo
Master student
Transportation Management College,
Dalian Maritime University, Dalian, 116026, China
E-mail: yubo198765@yahoo.com.cn

LU, Jing
Master student
Transportation Management College,
Dalian Maritime University, Dalian, 116026, China
E-mail: lj-ppx@yahoo.com.cn

YU, Bin
Associate Professor
Transportation Management College,
Dalian Maritime University, Dalian, 116026, China
E-mail: ybzhyb@163.com

YANG, Zhongzhen
Professor
Transportation Management College,
Dalian Maritime University, Dalian, 116026, China
E-mail: yangzhongzhen@263.net

Abstract: Effective prediction of bus arrival times is important to advanced traveler information systems. Here enhanced regression models, based on linear regression and an adaptive algorithm, are presented to predict the arrival times at stops. In the models, the linear regression models predict the baseline travel times between adjacent stops based on traffic conditions of the following segment; the adaptive algorithm uses the most recent bus arrival information, together with the estimated baseline arrival times from linear regression models, to predict arrival times at downstream stops. The adaptive method for bus arrival time prediction is examined with the data of bus No. 23 in Dalian City, China and at last some conclusions are given.

Keyword: *prediction, bus arrival times, linear regression, adaptive algorithm*

1. INTRODUCTION

Two components of intelligent transportation systems (ITS) are advanced traveler information systems (ATIS) and advanced public transportation systems (APTS). One of the key elements in APTS/ATIS is a model to predict transit vehicle arrival time with reasonable accuracy. The arrival time deviations of buses are usually caused by several stochastic factors (traffic congestion, ridership distribution, and weather condition). The resulting impact of these factors on the transit system comprises bunching between pairs of operating vehicles, increasing passengers waiting time, deterioration of schedule/headway adherence, unsmooth intermodal transfers, increasing the cost of operation and traffic delays, etc. The provision of timely and accurate transit travel time information can attract additional ridership and increase the satisfaction of transit users. In addition, transit operators can identify vehicles that fall behind schedule and react in a proactive way. Automatic Vehicle Location Systems, which is a part of ITS, have been adopted by many transit agencies and allows them to track their transit vehicles in real-time. While the provision of real-time information, such as bus location, is relatively straightforward, forecasting transit information, such as when a bus will arrive at a particular location, is significantly more complex (Jeong and Rilett, 2004). There has been a considerable amount of work on the subject of short-term traffic forecasting. The methodologies used in the previous work include Kalman filtering (Vythoulkas,1993; Okutani and Stephanedes, 1984; Cathey and Dailey, 2003), linear models (Kwon *et al*., 2000; Lin and

Bertini, 2004) and ARIMA models (Voort *et al*., 1996; Oda, 1990) simulation models (Hall *et al.*, 1980), neural network models (Chen *et al.*, 2004; Chien *et al.*, 2002; Park and Rilett, 1999), SVM (Yu *et al.*, 2006; Yu and Yang, 2009Wu *et al.*, 2004), among others.

The focus of this article is to use an adaptive regression model to predict bus arrival times in the near future. The adaptive regression model consists of two major elements: linear regression models and an adaptive algorithm. First, we model the relationship between the anticipated arrival times and an arrival time estimate using currently available data as being transiently approximately linear. Since the linear regression models are obtained based on historical data, it lacks the dynamic feature of adjusting predictions and their outputs are just taken as the baseline of predicted travel times. To account for the impact of stochastic factors during the running, an adaptive algorithm is developed to stepwisely reduce prediction errors.

The rest of the paper is organized as follows. We present the adaptive regression model for the travel time prediction problem in Section 2. Section 3 contains results and analyses including performance evaluation of the methodology. Lastly, the conclusions are presented in Section 4.

2. MODEL DEVELOPMENT

Arrival-time prediction depends on vehicle speed, traffic flow and occupancy, which are highly sensitive to weather conditions and time-of-day. These features make travel-time predictions very complex and difficult to reach optimal accuracy. The objectives of this research are to develop and apply models to predict bus arrival times at morning peak time (6:30 A.M. - 7:30 A.M) in weekday due to their popularity.

The main idea of the traffic forecasting is based on the fact that traffic behaviors possess both partially analogical and partially chaotic properties. In our models, linear regression models are expected to reconstruct the analogical motion from historical data, and an adaptive algorithm is used to reduce the errors caused by unanticipated factors are proposed.

2.1 Multiple Linear Regressions

Time-varying features germane to bus running are the key to travel time modeling. Among those features that may contribute to the variation of bus running, we select two variables, bus speeds on the current route segment, $v^c_{m,k-1\to k}$ and bus speeds on the next route segment, $v^n_{k\to k+1}$. $v^c_{k-1\to k}$ denotes the bus speeds of the current bus (bus m) on current segment (k-1→k) between stop k-1 and stop k (the k_{th} main stop), which are used to show the bus currently operational conditions, and $v^n_{k\to k+1}$ denotes bus speeds of the preceding bus (bus m -1) on following segment (k→k+1) between stop k and stop k+1, which are expected to estimate the traffic conditions of the following segment. However, it is difficult to directly measure the

two speeds. Thus, in this research, two approximate speeds, which are calculated by the lengths of the current segment/ the following segment and travel times between the adjacent stops of the current bus/ the proceeding bus, as a substitute for $v^{c}_{m,k-1\to k}$ and $v^{n}_{k\to k+1}$, respectively. Travel time is the difference between the arrival times at the adjacent stops, that is travel time includes dwelling time at one stop. The outputs of linear regression models are estimated travel speeds ($v'^{e}_{m,k\to k+1}$) of bus *m* on predicted route segment *k*→*k*+1. Integrating the lengths of the predicted route segment, the estimated travel times ($t'^{l}_{m,k\to k+1}$) on the predicted route segment from linear regression models are calculated. Then, the objectives of the linear regression models for bus arrival times are to generalize the relationship of the following form:

$$t'^{l}_{m,k\to k+1} = D_{k\to k+1} \times v'^{e}_{m,k\to k+1} = D_{k\to k+1} \times (C + C^{c} \times v^{c}_{m,k-1\to k} + C^{n} \times v^{n}_{k\to k+1}) \quad (1)$$

where, C,C^{c} and $C^{n}$are constant.$D_{k\to k+1}$ denotes the lengths of route segment *k*→*k*+1.

2.2 Enhanced Regression Models

Although the above two speeds have some "dynamic" feature, linear regression models developed previously are still based on historical data pool of bus trips. Certainly these regression models should be retrained regularly (e.g., daily or weekly - depending on the frequency of database update). Nevertheless, the regression models can't still adjust the prediction results with the most recent information from the current trip. The outputs from linear regression models only serve as the baseline estimate of travel times. To improve prediction accuracy, an adaptive algorithm, which was proposed by Chien *et al*. (2002), is adopted for adjusting prediction results in real time. Integrating arriving times of bus *m* at the current stop and the baseline travel times on the following segment ($t'^{l}_{m,k\to k+1}$) from linear regression models, the predicted arrival times $\hat{T}_{m,k+1}$ of bus *m* at stop *k*+1 by the enhanced regression models can be estimated by Eq. (2) as

$$\hat{T}_{m,k+1} = t^{a}_{m,k} + t'^{l}_{m,k\to k+1} + K_{m,k+1} e^{T}_{m-1,k+1} \quad (2)$$

where $t^{a}_{m,k}$ represents the measured arrival times for bus *m* at stop *k*. $e^{T}_{m-1,k+1}$ represents the prediction error of the enhanced regression model. $e^{l}_{m-1,k+1}$ represents the linear regression model. The prediction errors, $e^{T}_{m-1,k+1}, e^{l}_{m-1,k+1}$, can be determined when bus *m*-1 arrives at stop *k*+1.

$$e^{T}_{m-1,k+1} = t^{a}_{m-1,k+1} - \hat{T}_{m-1,k+1} \quad (3)$$

$$e^{l}_{m-1,k+1} = t^{a}_{m-1,k+1} - t^{a}_{m-1,k} - t'^{l}_{m-1,k\to k+1} \quad (4)$$

To reduce the prediction errors, the factors $K_{m,k+1}$ $(0<K_{m,k+1}\leq 1)$ is introduced to scale the adjustment of the prediction results from linear regression models. $K_{m,k+1}$ can be optimized by minimizing the covariant prediction errors $P^{T}_{m-1,k+1}$.

$$K_{m,k+1}=\frac{1}{1+\frac{R^{l}_{m-1,k+1}}{P^{T}_{m-1,k+1}}} \tag{5}$$

where $P^{T}_{m-1,k+1}=E\left[(e^{T}_{m-1,k+1})^{2}\right]$ (6)

and $P^{l}_{m-1,k+1}=E\left[(e^{l}_{m-1,k+1})^{2}\right]$ (7)

represent the covariance of the prediction errors generated from enhanced regression models and linear regression models, and are formulated in Eqs. (6) and (7), respectively. $R^{l}_{m-1,k+1}$, formulated in Eq. (8), is the covariance for the random noise (independent and zero mean) observed from bus arrival times at stop *k*+1.

$$R^{l}_{m-1,k+1}=\frac{P^{l}_{m-1,k+1}}{K_{m-1,k+1}} \tag{8}$$

$R^{l}_{m-1,k+1}$ accounts for the effects of random drifting of enhanced regression models from the observed bus arrivals, which depends on the stochastic characteristics in specific transit systems. $P^{T}_{m-1,k+1}$, $P^{l}_{m-1,k+1}$ and $R^{l}_{m-1,k+1}$ are estimated and updated iteratively in real time to determine the optimal $K_{m,k+1}$ that helps the enhanced regression models to adapt to real-time situations by adjusting the prediction errors when a bus arrives at a stop. The benefit earned by applying the adaptive algorithm is in its computational efficiency; it can be integrated with linear regression models to enhance the prediction performance while adapting to a dynamic environment without exercising a lengthy retraining process.

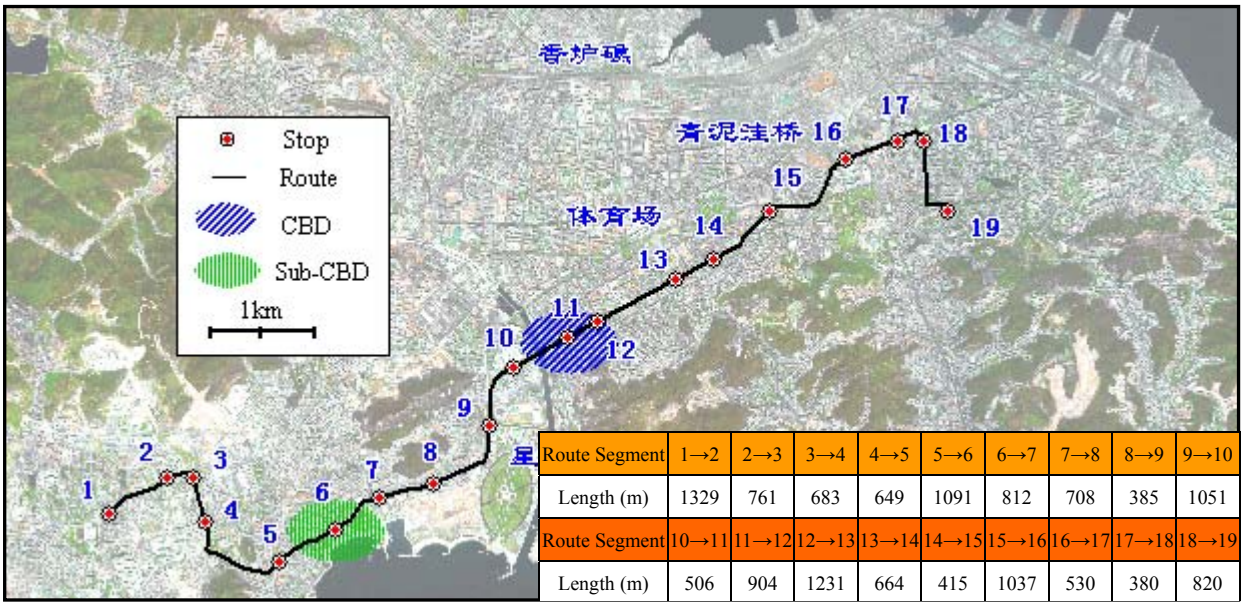
2.3 Applying Enhanced Regression Model in Bus Arrival Time Prediction

The enhanced regression models consist of linear regression models and an adaptive algorithm. Before application, the coefficients in linear regression models ought to be determined with historical data, and the well-regressed linear models can estimate the baseline travel times of each segment. For instance, we predict the arrival times for bus *m* at stop *k*+1. When bus *m* arrives at stop *k*, according to $v^{c}_{m,k-1\rightarrow k}$ and $v^{n}_{k\rightarrow k+1}$, the baseline travel times from stop *k* to stop *k*+1, $t'^{l}_{m,k\rightarrow k+1}$ is estimated by linear regression models. Simultaneously, the latest bus speeds on segment *k*-1→*k* is updated, i.e. $v^{n}_{k-1\rightarrow k}=v^{c}_{m,k-1\rightarrow k}$, which is used as the input variable of the following buses to predict. Then, according to the two observed arrival times, $t^{a}_{m-1,k+1}$ and $t^{a}_{m,k}$, the prediction errors of bus *m*-1 at stop *k*+1 and bus *m* at stop *k* are

calculated. Also, $K_{m,k+1}$ is determined. Thus, integrating the observed arrival times for bus m at stop k ($t^{a}_{m,k}$) and the baseline travel times from stop k to stop k+1 from linear regression model ($t'^{l}_{m,k\to k+1}$), the predicted arrival times ($\hat{T}_{m,k+1}$) of bus m at stop k+1 by the enhanced regression models can be determined. As the bus proceeds along its route, the prediction is updated whenever the most recent arrival information is obtained. The process is repeated till the bus reaches the final destination.

3. NUMERICAL TEST

The enhanced regression models for bus vehicle arrival time prediction have been tested with the data of transit route No. 23 in Dalian City, China. The transit route No. 23 goes from suburb to city centre with total of 19 stops and 14.5*km* per direction. In the numerical test, the parts of the transit route, 16 stops, and only the eastbound direction were studied. The part route and bus stops are as shown in Fig.1. The transit route is highly congested in the morning and afternoon peaks and the headway and the travel time during peak period are about 2.5minutes and 45 minutes. We describe the data used in the models first and then the results obtained.



Route Segment	1→2	2→3	3→4	4→5	5→6	6→7	7→8	8→9	9→10
Length (m)	1329	761	683	649	1091	812	708	385	1051
Route Segment	10→11	11→12	12→13	13→14	14→15	15→16	16→17	17→18	18→19
Length (m)	506	904	1231	664	415	1037	530	380	820

Fig.1 Configuration of transit route number 23

3.1 Data Collection and Processing

The collected data consist of the arrival times of the 16 stops in each individual trip at peak time during weekdays. There are 520 valid trips within this 1-month period, and there are 7,800 segment travel times in total. The desired data structure for enhanced regression models is a set of successive point records. However, some point records were missed during survey when buses skipped those points. To generate a set of successive point records, the bus arrival times at missing points were interpolated based on the available information at upstream and downstream points assuming the travel speed to be constant between two consecutive stops.

3.2 Model Identification

The training process utilizes three data sub-sets, respectively for training, cross-validation and testing. Then, about 10% samples of the data set are set aside as testing data. The remaining data are randomly assigned into two groups, the first with 300 trips for training and the second with the other data for cross validation. To improve the prediction precision, we develop a regression model for each segment. Thus, it is necessary to divide the training data into different sets according as route segments. The coefficients in the linear regression models are presented in Table 1.

It can be observed that the coefficients of some variables are negative in the linear regression models. This is not as expected as the predicted speed is proportional to the speeds of the previous vehicles on the predicted segment or the current segment. This can be attributed that the route goes through the central business district (CBD) of Dalian city. CBD traffic slows to a near-standstill in peak hours. This always induces the cases of bus bunching or large interval. The two cases alternate. The bus after the bunched buses has large interval with the previous bus and it will pick up more passengers than usual. Certainly, the more passengers will induce more dwelling times. Thus, the travel time of the current bus on the predicted segment is possibly large though the previous bus has high speed. However, the next bus of the current bus can pick up fewer passengers, which will lead to a higher speed. Therefore, the predicted speeds on the some segments vary inversely as the speeds of previous buses on the current segment or predicted segment.

Table 1 Coefficients of linear regression models

Segment		Coefficients	t Stat	R Square	Segment		Coefficients	t Stat	R Square
	C	19.21	37.17			C	22.33	4.06	
1	C^c	0	-	0.31	9	C^c	0.48	0.47	0.65
	C^n	0.14	22.96			C^n	-0.61	-1.20	
	C	11.65	4.93			C	8.79	2.91	
2	C^c	0.20	1.35	0.81	10	C^c	3.04	-2.35	0.72
	C^n	0.24	4.90			C^n	2.91	2.01	
	C	26.03	8.57			C	10.14	4.37	
3	C^c	0.05	1.06	0.73	11	C^c	2.01	2.43	0.55
	C^n	-0.41	-3.59			C^n	4.16	2.13	
	C	9.96	13.04			C	19.45	-1.25	
4	C^c	-0.15	-3.97	0.62	12	C^c	-0.26	3.83	0.61
	C^n	0.65	3.74			C^n	0.37	3.84	
5	C	7.55	3.55	0.77	13	C	20.11	4.08	0.55
	C^c	0.57	-2.43			C^c	-0.12	-2.09	

	C^n	0.28	2.74			C^n	0.31	1.73	
	C	17.62	-1.90			C	14.13	2.82	
6	C^c	0.20	0.69	0.48	14	C^c	0.50	-3.69	0.69
	C^n	0.09	4.04			C^n	0.10	1.51	
	C	24.27	3.11			C	17.81	5.61	
7	C^c	-0.28	-1.42	0.57	15	C^c	0.04	2.43	0.83
	C^n	0.07	0.81			C^n	0.42	1.89	
	C	14.33	7.15						
8	C^c	0.33	3.30	0.66					
	C^n	0.21	2.38						

3.3 Results

3.3.1 Linear Regression Models Output

To validate the independent variables in linear regression models, the variations between linear regression model output and actual travel time are compared. The prediction accuracy is evaluated by computing the root mean squared error (RSME) of each bus route segment, which can be obtained from:

$$\mathrm{RMSE} = \frac{1}{J}\sum_{j=1}^{J}(t^{a}_{m-1,k+1} - \hat{T}_{m-1,k+1})^2 \tag{9}$$

where J is the number of test samples.

Then, the various linear regression models are implemented based on the same data sets, in which the parameters are shown in Table 2. The results are shown in Fig.2. It can be observed that the performance of the linear regression models with two independent variables is best. This can be attributed that the models with two variables simultaneously consider the traffic conditions on the current segment and the following segment. Also, the operational conditions of the following segment have more impact than the ones of current segment. In addition, we can see that the RMSEs of the prediction result for the second route segment is highest. It is mainly attributed that the passenger flows are large and the roads are narrow in the area. This may cause a variety of external influences, which affected travel times of transit vehicle on the links or dwell times at the stops and increased the RMSEs of the prediction results for the route segments. Except for the second route segment, it can be observed that the RMSEs of the prediction results for the tenth and twelfth route segments are also large. This may be due to the two route segments locates on the central business district (CBD) of Dalian city. CBD traffic slows to a near-standstill in peak hours, which caused transit vehicle arrivals to deviate from the schedule and the vehicle arrival times not to be accurately predicted.

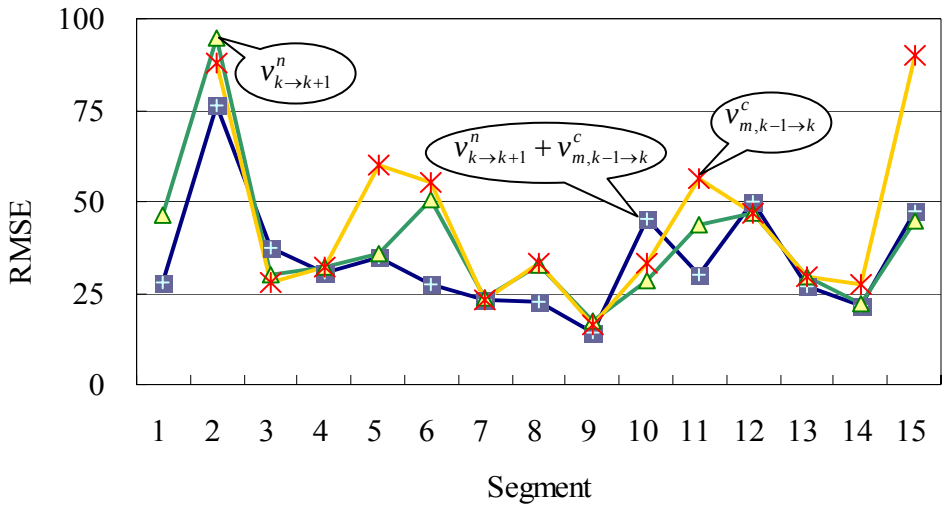


Fig.2 Comparative analysis of in various regression models

Table 2 Coefficients of other linear regression models

Linear regression models with the speed on the current segment															
Coefficients	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
C	20.50	10.46	11.16	20.27	17.50	16.41	22.94	7.29	21.88	14.36	20.32	22.04	21.19	7.19	15.53
C^c	0	0.27	0.68	-0.25	0.25	0.33	-0.05	0.48	0.08	0.87	0.15	-0.22	-0.2	0.77	0.45
Linear regression models with the speed on the next segment															
Coefficients	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
C	18.01	8.76	22.40	15.83	18.12	17.29	22.80	11.30	15.66	16.02	19.09	18.73	17.97	11.64	14.95
C^n	0.14	0.3	-0.11	0.42	0.33	0.3	-0.08	0.57	0.38	0.69	0.14	0.49	0.26	0.31	0.37

3.3.2 Enhanced Regression Model Output

The adaptive algorithm is designed to fine-tune the linear regression model outputs. Whenever the bus reaches a stop, the arrival time prediction from linear regression models can be adjusted by the newly obtained arrival information. To evaluate the performance of the adaptive algorithm, the enhanced regression models are compared with the linear regression models and the schedule. Figure 3 shows the comparison of RMSEs of the three methods.

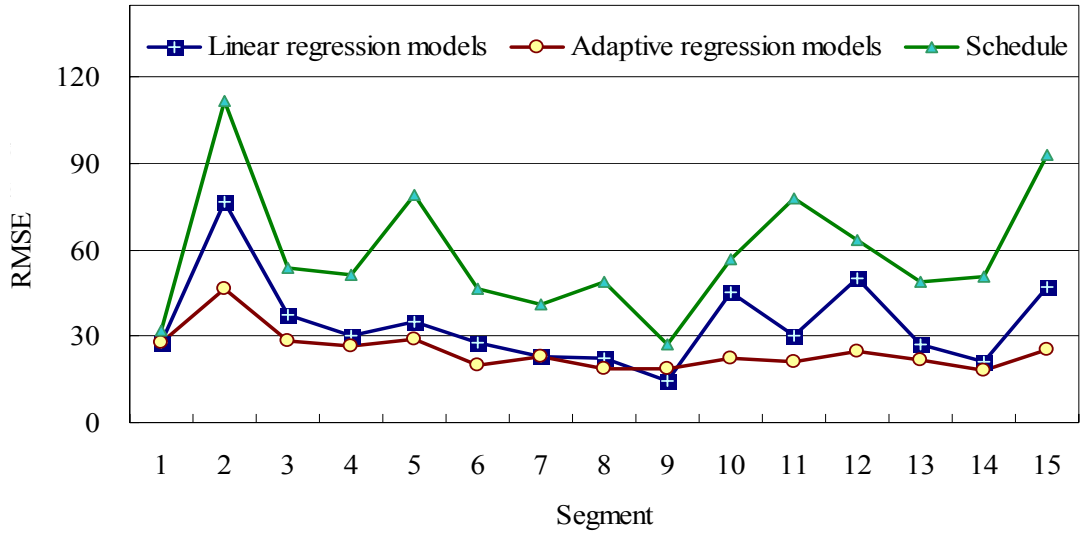


Fig.3 Comparison of three methods

It can be observed that in all cases, the linear regression models for all route segments are smaller than those of the timetable. The difference could be as large as 115, approximately. This indicates that, compared to the timetable, the linear regression models generally provide better indication of the bus arrival time between two adjacent stops.

Then, compared to the ones of linear regression models, the RMSEs predicted by the enhanced regression models for more route segments are smaller. However, we can see that the ninth segment has higher RMSE in enhanced regression model than the linear regression model. This can be attributed that the previous errors are accumulated in the enhanced regression models. Thus, for some segments the enhanced regression models can yield higher errors.

In addition, RMSEs of linear regression model outputs show a more fluctuant trend. Nevertheless, the enhanced linear regression model outputs do not show such a trend, in fact, its prediction error is generally stable with only few small scale changes. This is just as expected as the enhanced regression models have incorporated the latest bus arrival information in to the prediction and stepwisely reduced the errors of outputs. Especially in CBD, it is obvious that the performance of the enhanced regression model is best in three methods. So the enhanced linear regression model seems to be a powerful alternative for bus arrival time prediction.

4. CONCLUSIONS

One of the major stochastic characteristics in transit operations is that vehicle arrivals tend to deviate from the posted schedule. Poor schedule or headway adherence is undesirable for both users and operators, since it increases passenger waiting/transfer time, discourages passengers from using the transit system, and degrades the operation efficiency and productivity. To predict travel time, the consideration of traffic condition is essential. Considering the complexity and difficulties of traffic congestion, this research used the speeds (travel times) of preceding/current bus on segments to estimate traffic conditions of segments and

developed the prediction models. The enhanced regression model consists of two major elements: linear regression models and an adaptive algorithm. The linear regression models provide the baseline times of segments based on historical trip data. To account for the impact of unexpected delays during running, the adaptive algorithm uses the most recent bus arrival information, together with the estimated baseline travel times from linear regression models, to predict arrival times at downstream stops. Tests show that the linear regression models are quite powerful in modeling variations in arrival times along the service route and the adaptive algorithm can effectively integrate the latest bus information to predict accurate bus. Furthermore, if real-time data collected from traffic surveillance systems and transit monitoring systems are available, the enhanced regression models can be similarly developed to adapt to transit operations in a changeable environment. In addition, the methods can be applied online with the bus trip in progress because of its simplicity in calculation. Therefore, the methodology developed in this study can potentially be used for providing real-time bus arrival time prediction for each stop along the route.

ACKNOWLEDGMENTS

This research is financed by the National Science Foundation for Post-doctoral Scientists of China 20080440168 and the Doctoral Program Foundation for Young Scholar of Institutions of Higher Education of China through project 20070151013.

REFERENCES

Cathey, F.W. and Dailey, D.J. (2003); A prescription for transit arrival/departure prediction using automatic vehicle location data. **Transportation Research Part C 11**:241–264.

Chen, M., Liu, X.B., Xia, J.X. and Chien S.I. (2004); A Dynamic Bus-Arrival Time Prediction Model Based on APC Data, **Computer-Aided Civil and Infrastructure Engineering 19**:364-376.

Chien, I-Jy., Ding, Y. and Wei, C. (2002); Dynamic Bus Arrival Time Prediction with Artificial Neural Networks, **Journal of Transportation Engineering, ASCE, 128(5)**:429–38.

Hall, M.D., Vliet, D.V. and Willumsen, L.G. (1980); SATURN: A simulation assignment model for the evaluation of traffic management schemes. **Transportation Engineering Control, 21**:168-176.

Jeong, R. and Rilett, L.R. (2004); *Bus Arrival Time Prediction Using Artificial Neural Network Model*, IEEE Intelligent Transporlatlon Systems Conference, Washington, D.C., USA: 988-993.

Kwon, J., Coifman, B., and Bickel, P. (2000); "Day-to-Day Travel Time Trends and Travel Time Prediction from Loop Detector Data," **Transportation Research Record no. 1717**, **Transportation Research Board**:120-129.

Lin, W.H. and Bertini, R.L. (2004); Modeling schedule recovery processes in transit

operations for bus arrival time prediction, **Journal of Advanced Transportation, 38 (3)**: 347-365.

Oda, T. (1990); An Algorithm for Prediction of Travel Time Using Vehicle Sensor Data. Proceedings of the IEE 3rd International Conference on Road Traffic Control, London:40-44.

Okutani, I. and Stephanedes, Y. J. (1984); Dynamic Prediction Of Traffic Volume Through Kalman Filtering Theory. **Transp. Res.,part B 18 (1)**:1–11.

Park, D. and Rilett, L. R. (1999); Forecasting Freeway Link Travel Times With A Multilayer Feedforward Neural Network. **Computer-Aided Civil and Infrastructure Engineering, 14(5)**:357–367.

Voort, M. V. D., Dougherty, M. and Watson, S. (1996); Combining Kohonen maps with ARIMA time series models to forecast traffic flow, **Transportation Research Part C, 4**:307-318.

Vythoulkas, P.C. (1993); Alternative Approaches to Short Term Traffic Forecasting for Use in Driver Information Systems, **Transportation and Traffic Theory**:485-505.

Wu, C. H., Ho, J. M., and Lee, D. T. (2004); Travel-time prediction with support vector regression, **IEEE Transactions on Intelligent Transportation Systems, 5(4)**:276–281.

Yu, B. and Yang, Z.-Z. (2009); A Dynamic Holding Strategy in Public Transit Systems with Real-Time Information. **Appl Intell,31(1)**:69-80.

Yu, B., Yang, Z.Z. and Yao, B.Z. (2006); Bus Arrival Time Prediction Using Support Vector Machines. **Journal of Intelligent Transportation Systems, 10(4)**:151–158.