Transportation Research Part C 19 (2011) 1157–1170

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Bus arrival time prediction at bus stop with multiple routes

Bin Yu [a,b,*], William H.K. Lam [a], Mei Lam Tam [a]

[a] Department of Civil and Structural Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, PR China
[b] Transportation Management College, Dalian Maritime University, Dalian, PR China

ARTICLE INFO

Article history:
Received 1 September 2010
Received in revised form 19 January 2011
Accepted 21 January 2011

Keywords:
Bus arrival time prediction
Multiple bus routes
Support vector machine
Artificial neural network
k nearest neighbours algorithm

ABSTRACT

Provision of accurate bus arrival information is vital to passengers for reducing their anxieties and waiting times at bus stop. This paper proposes models to predict bus arrival times at the same bus stop but with different routes. In the proposed models, bus running times of multiple routes are used for predicting the bus arrival time of each of these bus routes. Several methods, which include support vector machine (SVM), artificial neural network (ANN), *k* nearest neighbours algorithm (*k*-NN) and linear regression (LR), are adopted for the bus arrival time prediction. Observation surveys are conducted to collect bus running and arrival time data for validation of the proposed models. The results show that the proposed models are more accurate than the models based on the bus running times of single route. Moreover, it is found that the SVM model performs the best among the four proposed models for predicting the bus arrival times at bus stop with multiple routes.

1. Introduction

1.1. Backgrounds

Many applications of information systems and technologies, such as automatic vehicle location (AVL) or identification (AVI) systems and automatic passenger counters (APC), are now receiving increasing attentions in transportation management. They are considered as the key components of intelligent transportation systems (ITS). Recently, transit agencies also realize the operational benefits of ITS-related technology implementation. Based on these advanced technologies, transit agencies can acquire real-time bus information to reduce passenger journey time and improve management/service level. Thus, there is a growing interest in providing real-time bus arrival information for passengers using emerging electronic information and communication technologies. The availability of real-time bus information can help passengers efficiently schedule their departure time and make smart choices for their travel.

In practice particularly in transit-oriented cities like Hong Kong, it is very common to have several bus routes using the same bus stop in urban areas. In the bus stop with multiple routes, passengers would have several choices (different bus routes) to reach their destinations. Real-time bus information available at stop can be very helpful to passengers if they can know which bus will arrive first. Thus, there is a need for bus information at stop with multiple routes particularly in high density populated cities (e.g., Hong Kong) with large bus passenger demands. For example, there are nineteen bus routes using two bus stops (i.e. two separate bus bays and some stop boards at each bus bay) at the Cross Harbour Tunnel (North bound) in Hong Kong. Among these bus routes, there are some common lines passing several major urban areas on

* Corresponding author at: Department of Civil and Structural Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, PR China.
E-mail address: minlfish@yahoo.com.cn (B. Yu).

0968-090X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trc.2011.01.003

Hong Kong Island. There is always a long queue of passengers waiting for buses at the bus stop board of each route. Therefore, given the bus arrival time information at the bus stop with multiple routes, passengers who have multiple choices can choose the potentially suitable route for their journeys. With this real-time bus arrival time information, passengers can greatly reduce their waiting times and the efficiency of bus services can significantly be improved particularly at the major bus stops with high volume of passengers transferring from railway to bus modes.

Fig. 1 shows an example for illustrating the effort on providing bus arrival time information at bus stop with multiple routes. In the example, a passenger expects to go from Stop A to Stop B where there are three bus routes, route nos. 101, 102 and 103. The passenger can choose any one of these three bus routes to his destination at Stop B. If the passenger knows the bus arrival times of the next buses of the three bus routes at Stop A (e.g., 09:05, 09:04 and 09:02, respectively), he/she will wait for the next bus of the route no. 103 rather than the other two routes. Thus, his/her waiting time will be reduced.

Passengers, in general, are more interested in knowing the predicted arrival times of the following buses rather than the current locations of the buses. Thus, the accuracy of the prediction algorithm is very important in a successful bus information system. However, accurate prediction of bus arrival time is very difficult due to many stochastic variables involved (e.g., traffic conditions). Therefore, the deployment of bus arrival time prediction model is a challenging task.

1.2. Literature review

In the past decade, various sophisticated techniques and algorithms have been developed to predict bus travel time or arrival time by using AVL and/or APC data. These methods can be categorized as: artificial neural network (ANN) or Support vector machine (SVM) (Ding and Chien, 2000; Chien et al., 2002; Chen et al., 2004; Jeong and Rilett, 2004; van Lint et al., 2005; Vlahogianni et al., 2005; van Hinsbergen et al., 2009; Yu et al., 2006, 2010a,b), Non-parametric regression (NPR) model (Smith et al., 2002; Chang et al., 2010; Park et al., 2007; Zhang and Rice, 2003; You and Kim, 2000; Vlahogianni et al., 2006) and Kalman filter model (Wall and Dailey, 1999; Chien and Kuchipudi, 2003; Shalaby and Farhan, 2004; Chen et al., 2004; Yu et al., 2010a).

1.2.1. Artificial neural network/support vector machine models

ANN is motivated by emulating the intelligent data processing ability of human brains. ANN has been reported to be especially useful for finding solutions for complex non-linear problems. Chien et al. (2002) proposed two ANN-based models: the link-based ANN model and the stop-based ANN model, to predict bus arrival time. An adaptive algorithm was also developed to improve the performances of the ANN-based models. Their results showed that the link-based ANN model outperformed the stop-based ANN model for the prediction with a relatively small number of intersections. The results also indicated that the adaptive algorithm can improve the performances of the ANN-based models. Chen et al. (2004) proposed a dynamic algorithm that integrated the ANN model and a Kalman filter-based algorithm. Jeong and Rilett (2004) compared the performances of several methods: the historical data based model, the regression models, and the ANN model, for bus arrival time prediction. Their results showed that the ANN model outperformed the historical data based model and the regression model in terms of prediction accuracy. Van Lint et al. (2005) presented a freeway travel time prediction framework which combined state-space neural network with preprocessing strategies based on imputation. Their results indicated that a combination of these imputation procedures and the proposed model could be implemented a real-time application. Vlahogianni et al. (2005) presented a multilayered structural optimization strategy based on genetic algorithm, which was applied to both univariate and multivariate traffic flow data to evaluate the performance of the developed network. van Hinsbergen et al. (2009) used Bayesian inference theory to combine neural networks in a committee using. The proposed method had an evidence factor to act as a criterion of stopping the training and a tool to determine different neural networks. Their results showed that the proposed approach had a much higher accuracy.

SVM is a very specific type of learning algorithm characterized by the capacity control of the decision function, the use of the kernel functions, and the sparse solution (Cristianini and Shawe-Taylor, 2000; Vapnik, 1999, 2000). Yu et al. (2006, 2010b) developed the SVM-based models to predict bus arrival time. In the models, travel speeds of the preceding buses of the same bus route were used to estimate traffic conditions. Their results showed that the SVM-based models outperformed the ANN and historic mean prediction models, and SVM seemed to be a powerful alternative for bus arrival time prediction. With versatile parallel distributed structures and adaptive learning processes, ANN and SVM have recently been gaining popularity in bus travel/arrival time prediction.

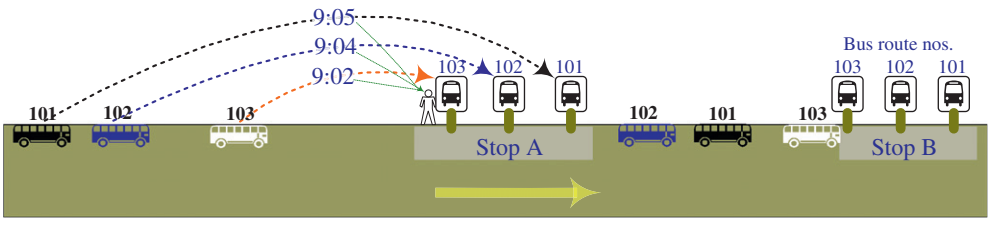

Fig. 1. An example for bus arrival time information at bus stop with multiple routes.

1.2.2. Non-parametric regression models

NPR is a relatively simple method for prediction without the need to estimate parameters. The NPR models are therefore more suitable for real-time applications. Zhang and Rice (2003) proposed a method to predict short-term freeway travel times using a linear model in which the coefficients varied as smooth functions of the departure time changed. Several varieties of the prediction procedures were presented and the results were encouraging. You and Kim (2000) developed a hybrid travel time forecasting model based on non-parametric regression for predicting link travel times in congested road networks.

Recently, k nearest neighbour (k-NN) is one of the most popular NPR methods, which has been widely applied in many fields (Smith et al., 2002; Chan et al., 2009; Tam and Lam, 2009). Chang et al. (2010) developed a bus travel time prediction model using k-NN. Their results showed that k-NN was an effective method to estimate bus travel time according to the prediction accuracy and computing time. Park et al. (2007) applied a non-parametric regression model for travel time prediction. Their model was implemented and evaluated using real-time transit data. k-NN methods have also been documented while a large database is desirable for increasing the prediction accuracy. However, the large sample size has significant implications on the timeliness of model execution (Smith et al., 2002).

1.2.3. Kalman filter models

Kalman filter is an efficient recursive procedure that estimates the future states of dependent variables. It is originated from the state-space representations in modern control theory. Chien and Kuchipudi (2003) developed a path-based model and a link-based model using Kalman filter to predict bus travel times. Their results showed that the link-based model would be more sensitive to travel time increment of the link with congestion or incident. Shalaby and Farhan (2004) discovered that the Kalman filter-based model outperformed the regression and neural network models in terms of accuracy. Cathey and Dailey (2003) proposed a prediction method for bus arrival/departure time which included three components, e.g., a track, a filter and a predictor. In the filter, Kalman filter was used to estimate the vehicle dynamical state. Chen et al. (2004) developed an enhanced algorithm based on Kalman filter to predict bus arrival time. Their results showed that the enhanced algorithm was effective for bus arrival time prediction compared with the standard ANN-based models.

In summary, many researches have been conducted on forecasting bus travel/arrival time for a single bus route, while few studies have investigated bus travel/arrival time prediction for multiple bus routes. In most previous researches, some factors related to road traffic (e.g., traffic speed and volume) were used to model bus travel/arrival time prediction. However, bus running is greatly different from that of other vehicles due to bus lane, bus stop and so on. To solve the problems, Yu et al. (2006, 2010b) applied the bus running times (speeds) of the preceding buses to predict the arrival time of the next bus. However, in their researches, the bus running times (speeds) of the same route were used to model the prediction. For the route segment passing through several bus routes, the bus arrival time prediction model can provide better prediction accuracy by integration of bus information of multiple routes.

1.3. Contributions

In general, bus operation is different for the buses arriving at stops with multiple routes or single route. Bus arrival time at stops with single route is mainly affected by traffic conditions between stops. However, besides traffic condition, bus arrival time at stops with multiple routes would also be influenced by buses of other bus routes. For example, limited capacity of bus stop may cause buses queue up at the bus stop. As a result, bus arrival delays would be increased. This is very common in urban areas of transit-oriented cities like Hong Kong, particularly during peak periods.

As to the predictions for single route or multiple routes, there is some difference in the development of model. Fig. 2 shows the difference between the predictions for single route and multiple routes. When the target bus n of the bus route no 101 arrives at Location A, assume the buses $k,\ldots, k+\mu,\ldots, k+\delta$, have gone through Bus Stop, i.e., the buses are the preceding buses. Since the running times of the preceding buses between Location A and Bus Stop can be obtained, the running times are used to estimate traffic condition and to construct the prediction model. For the prediction for single route, only the running time of bus (e.g., the bus $k+\mu$) on the same route is used to model the prediction. Thus, the performance of the

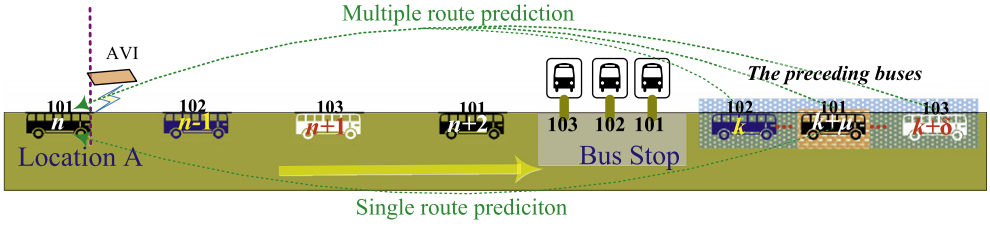

Fig. 2. The difference between the predictions of single route and multiple routes.

single route prediction model greatly relies on the timeliness and accuracy of the information of the preceding buses on the same route. However, except for buses on the same route, the bus information from other routes is also used for the multiple route prediction model. Compared with the bus information of the same route, multi-dimension bus information from multiple routes can provide more benefit (e.g., timeliness and reliability of the information), which will certainly improve prediction accuracy.

This paper seeks to make two contributions to the literature. Firstly, it attempts to develop the models to predict bus arrival time at the same bus stop but with multiple routes using real-world data. It is expected that the anxieties and waiting times of passengers can be reduced if passengers know when the next buses of multiple routes arrive at the bus stop. Bus running times of different routes are considered in the proposed models for predicting the bus arrival time. Secondly, the performances of several prediction methods, SVM, ANN, *k*-NN and linear regression (LR), are assessed and compared for forecasting bus running time with multiple routes. The performance comparison of several models can provide valuable insight for researchers as well as practitioners.

The structure of this paper is organized as follows: Section 2 provides the formulation of four proposed models for predicting bus arrival time at bus stop with multiple routes; Section 3 presents a case study together with results and analysis including performance evaluation of the four proposed models; and lastly, the conclusions are given in Section 4 together with suggestions for further study.

2. Methodologies

2.1. Prediction framework

Bus arrival time prediction at bus stop with multiple routes can be described in the following way: Given the bus arrival time of any bus route at a location, it is to predict the bus arrival time at the bus stop with multiple routes. Fig. 3 illustrates an example for the framework of the prediction in this study. When a bus (n) of any bus route (l) arrives at the Location A, the bus arrival time ($T^a_{l,n}$) can be recorded by some traffic data collection technologies (such as AVI). Then, the running time ($\hat{t}^{running}_{l,n}$) between the stop and the Location A is predicted by some methods. According to the arrival time of the bus at the Location A, the bus arrival time ($\hat{T}^s_{l,n}$) at the stop with multiple routes can be determined.

$$\hat{T}^s_{l,n} = T^a_{l,n} + \hat{t}^{running}_{l,n} \tag{1}$$

To predict bus running time in an accurate and timely manner, it is essential to determine the appropriate factors to estimate traffic conditions. Yu et al. (2006, 2010b) suggested that the running time(s) of the preceding bus(es) that has(ve) just reached the stop can be used to reflect the traffic conditions. Furthermore, they also pointed out the weighted average running times of several preceding buses could reduce the effect of accidents on the preceding buses. However, in the researches by Yu et al. (2006, 2010b), the bus running times of only the same bus route were applied to estimate traffic conditions. In fact, if integrating bus running times of different bus routes, the estimation accuracy of traffic conditions can be improved. In general, the most up-to-minute data can provide most reliable information for the prediction. Therefore, in order to take into account the timeliness of the bus running times of the preceding buses, the time headway between the target bus and the last preceding bus that has just reached the stop is considered in the proposed models.

For the sake of simplicity, "the preceding bus(es)" denotes the last bus(es) that has(ve) just reached the stop in the following sections. Assuming that n represents the target bus at the Location A, l represents the route no. of the bus n and L represents the set of bus routes. The factors considered in this study can be illustrated as follows.

(a) $t^l_{L,n}$ is the time headway between the target bus and the last preceding bus of any route among the route set L. The last preceding bus may be the same bus route or different bus route with the target bus. Thus, $t^l_{L,n}$ can be obtained by the following equation.

$$t^l_{L,n} = T^a_{l,n} - T^a_{L,k} \tag{2}$$

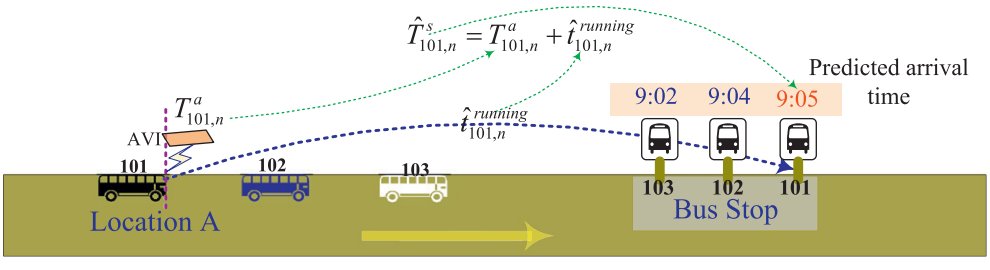

Fig. 3. An example for the prediction framework.

where $T^a_{l,n}$ represents the arrival time of the bus (n) of the bus route no. l at Location A. The bus k represents the last preceding bus of any route among the route set L. $T^a_{L,k}$ represents the arrival time of the bus (k) at Location A.

(b) $t^i_{l,n}$ is the time headway between the target bus and the last preceding bus of the same route no. l. Fig. 4 illustrates the difference between the two variables, $t^l_{L,n}$ and $t^i_{l,n}$.

$$t^i_{l,n} = T^a_{l,n} - T^a_{l,k+\mu} \tag{3}$$

where the bus $k+\mu$ represents the last preceding bus of the same bus route (l).

(c) $\bar{t}^r_{L,n}$ is the weighted average running time of several preceding buses (e.g., the bus k,..., $k+\mu$,..., $k+\delta$) of any routes among the route set L. In general, the last preceding bus to the target bus will contribute more to the weighted average running time than the further ones. A simple weighted method is to give each preceding bus a weight of the inverse of the time headway between the preceding bus and the target bus. Thus, $\bar{t}^r_{L,n}$ can be attained by the following equations.

$$\bar{t}^r_{L,n} = \sum_{j=1}^{\delta} \frac{1/(T^a_{l,n} - T^a_{L,j})}{\Gamma}(t^r_{L,j}) \tag{4}$$

$$\Gamma = \sum_{j=1}^{\delta} 1/(T^a_{l,n} - T^a_{L,j}) \tag{5}$$

where, $t^r_{L,j}$ is the running time between Location A and the stop of the jth preceding bus. Γ is the sum of the weight of each preceding bus. δ is the prediction horizon that is the number of the selected preceding buses.

(d) $t^r_{l,n}$ is the running time of the preceding bus (e.g., the bus $k+\mu$ in Fig. 3) of the same bus route No. l.

$$t^r_{l,n} = T^s_{l,k+\mu} - T^a_{l,k+\mu} \tag{6}$$

where $T^s_{l,k+\mu}$ is the arrival time of the bus ($k+\mu$) of the route No. l at the stop.

If $\hat{t}^{running}_{l,n}$ represents the prediction of bus running time between Location A and the stop, the prediction model for bus running times aims to generalize the relationship of the following form.

$$\hat{t}^{running}_{l,n} = f(t^l_{L,n}, t^i_{l,n}, \bar{t}^r_{L,n}, t^r_{l,n}) \tag{7}$$

To develop the bus running time prediction model, several techniques are employed in this study. With versatile parallel distributed structures and adaptive learning processes, ANN and SVM appear to be the suitable approaches for bus running prediction (Ding and Chien, 2000; Chien et al., 2002; Yu et al., 2006, 2010b). Thus, ANN and SVM are used to model bus running time prediction in this study. In addition, as simple and effective regression techniques, both k-NN and LR are also used for comparison.

2.2. Support vector machine models

SVM is a type of learning algorithms based on statistical learning theory, which can be adjusted to map the input–output relationship for the non-linear system. In addition, the solution of SVM is always unique and globally optimal since training SVM is equivalent to solving a linearly constrained quadratic programming problem. Therefore, SVM shows the strong resistance to the over-fitting problem and the high generalization performance. It is mainly because SVM can construct a mapping from one-dimensional input vector into high-dimensional space by the use of reproducing kernels. Here, Fig. 5 shows that the SVM-based model for the prediction of the bus arrival time at the stop with multiple routes.

2.3. Artificial neural network

ANN is a mathematical model by simulating the neural structure of the human brain. The ANN processes information by means of interaction between many neurons and the different links between neurons have been associated with weights. Based on the highly interconnected neural computing elements, ANN has the ability to model complex relationships

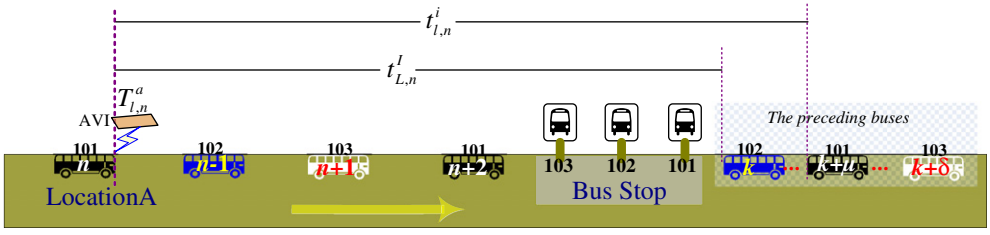

Fig. 4. The difference between variables $t^l_{L,n}$ and $t^i_{l,n}$.

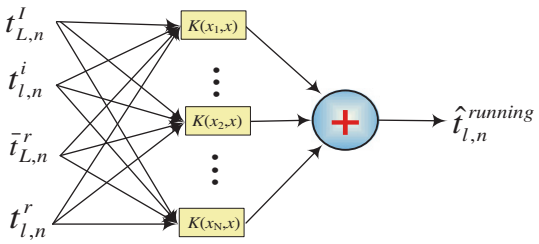

Fig. 5. Structure of the SVM model for the bus arrival time at the stop with multiple routes.

between inputs and outputs to find patterns in data. ANN includes two working phases, the learning phase and the recalling phase. During the learning phase, learning means using a set of observations are commonly used as a training signal in input and output layers. The recalling phase is performed by one pass using the weight obtained in the learning phase. ANN with three layers is chosen in this study as it is generally easy to use and can approximate almost any input/output relationships. The fully connected multilayer feed forward neural network with a back propagation (BP) algorithm has been applied successfully to deal with complex transportation systems (Huang and Ran, 2003). Hence, the neural network used to predict bus running time described as Fig. 6 in this study.

2.4. k nearest neighbours algorithm

k-NN is a method for classifying objects based on closest observations in a feature space. *k*-NN method is one of the simplest machine learning algorithms. In *k*-NN method, the Euclidean distance is usually used to determine the distance between the input state and the historical data in the feature space. On the basis of the Euclidean distance, the *k* nearest neighbours with the least distance to the input state can be determined. To weight the contributions of each neighbour, a common distance-based scheme is adopted to compute the weight of each neighbour. The forecasts can then be obtained by taking the weighted average of the observations from the *k* nearest neighbours.

In traditional *k*-NN method, standard Euclidean distance is used to match the *k* nearest neighbours in the feature space. This means that each independent variable has the same importance to the input state. For the prediction of bus running time, the equal weight of the independent variables is unreasonable. In this study, a weighed distance (d_j) is introduced to assign higher weight to the more important independent variable.

$$\hat{t}^{running}_{l,n} = \sum_{j=1}^{k} \frac{1/d_j}{D}(t^r_{L,j}) \tag{8}$$

$$d_j = \sqrt{\frac{\lambda_1 \times (t^I_{L,n} - t^I_{L,k})^2 + \lambda_2 \times (t^i_{l,n} - t^i_{l,k})^2 + \lambda_3 \times (\bar{t}^r_{L,n} - \bar{t}^r_{L,k})^2 + \lambda_1 \times (t^r_{l,n} - t^r_{l,k})^2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}} \tag{9}$$

$$D = \sum_{j=1}^{k} \frac{1}{d_j} \tag{10}$$

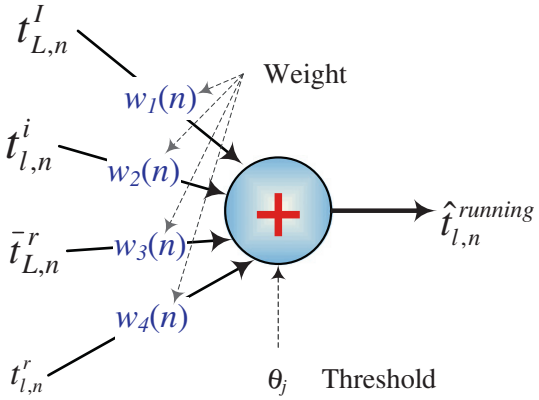

Fig. 6. Structure of the ANN model for the bus arrival time at the stop with multiple routes.

where d_j represents the weighted distance between the jth nearest neighbour and the input state. D represents the sum of the weighted distance of the k nearest neighbours. λ_1, λ_2, λ_3 and λ_4 represent the weights of the variables. The values of λ_1, λ_2, λ_3 and λ_4 are equal to the correlation coefficients between each independent variable and the dependent variable in this study.

2.5. Linear regression

Linear regression (LR) is the first type of regression analysis and used extensively in practical applications. For bus running time prediction, linear regression method is to model the relationship between the estimated bus running time (dependent variable) and the impact factors (independent variables).

Because the relationship between the estimated bus running time and the information of the preceding buses is very sophisticated, linear regression is extended with some interaction to make LR a more valid comparison with other models in this study. Here, the logarithm of the data set $\{\hat{t}_{l,n}^{running} : t_{L,n}^{l}, t_{l,n}^{i}, \bar{t}_{L,n}^{r}, t_{l,n}^{r}\}$ is taken, and then a new data set $\{\ln(\hat{t}_{l,n}^{running}) : \ln(t_{L,n}^{l}), \ln(t_{l,n}^{i}), \ln(\bar{t}_{L,n}^{r}), \ln(t_{l,n}^{r})\}$ is obtained. For the bus arrival time of the stop with multiple routes, linear regression model assumes that the relationship between the dependent variable and the independent variables (after the logarithm transformation) is approximately linear. Then, the approximate relationship is modeled as follow:

$$\ln(\hat{t}_{l,n}^{running}) = \beta_1 \times \ln(t_{L,n}^{l}) + \beta_2 \times \ln(t_{l,n}^{i}) + \beta_3 \times \ln(\bar{t}_{L,n}^{r}) + \beta_4 \times \ln(t_{l,n}^{r}) \tag{11}$$

where, β_1, β_2, β_3 and β_4 are coefficients that are related to the effects of impact factors on bus running time.

3. Case study

In this section, the proposed several models for predicting bus arrival time at bus stop with multiple routes have been evaluated by the real-world data in Hong Kong. Hong Kong has a highly developed and sophisticated bus route network that comprises about 700 bus routes. Over 90% of the daily journeys are on public transport, making it the highest rate in the world. In Hong Kong, real-time travel information system (RTIS) provides area-wide traffic information in the whole network (Tam and Lam, 2008). In RTIS, real-time traffic data (Autotoll tag records) are collected by AVI technology. The Autotoll tag records are initially used for electronic toll collection in Hong Kong. Almost all the buses have been installed with Autotoll tags for toll collection automatically.

Bus stop near the entrance of the Cross Harbour Tunnel (CHT) (North bound) in Kowloon Central urban area is selected for testing the proposed models. This bus stop is chosen because there are many bus routes with large passenger demands on harbour crossing every day. According to the locations of Autotoll tag readers, two directions, the west direction from Chatham Road North (CRN) to the CHT and the east direction from Ping Chi Street (PCS) to the CHT, are chosen. The locations of the bus stop near the CHT and the Autotoll tag readers are illustrated in Fig. 7, respectively. There are eight bus routes, operating along the west direction (through CRN), which include route nos. 102, 103, 104, 110, 112, 117, 118 and 171. Bus routes, operating along the east direction (through PCS), include route nos. 101, 107, 108, 109, 111 and 116. The distances from CRN and PCS to the CHT bus stop are about 0.62 km and 0.72 km, respectively.

3.1. Performance measures

The prediction results are evaluated in terms of the performances of three measures; namely, the mean absolute error (MAE_l), the mean absolute percentage error ($MAPE_l$) and the root mean square error ($RMSE_l$) of the route no. l. The three terms can judge the difference between the observed and the predicted running time in different aspects.

$$MAE_l = \frac{\sum |t_{l,n}^{running} - \hat{t}_{l,n}^{running}|}{N} \tag{12}$$

$$MAPE_l = \frac{1}{N} \sum \frac{|t_{l,n}^{running} - \hat{t}_{l,n}^{running}|}{t_{l,n}^{running}} \times 100\% \tag{13}$$

$$RMSE_l = \sqrt{\frac{\sum (t_{l,n}^{running} - \hat{t}_{l,n}^{running})^2}{N-1}} \tag{14}$$

where $t_{l,n}^{running}$ is the observed running time of the bus n of the route no. l. $\hat{t}_{l,n}^{running}$ is the predicted running time of the bus n of the route no. l. N is the number of the buses which have been observed.

3.2. Data collection and processing

To obtain the actual bus running and arrival time data, the video surveys near the CHT bus stop have been carried out in typical weekdays on 11–12 May 2010 (Tuesday to Wednesday) and 8 June 2010 (Tuesday) during the morning peak

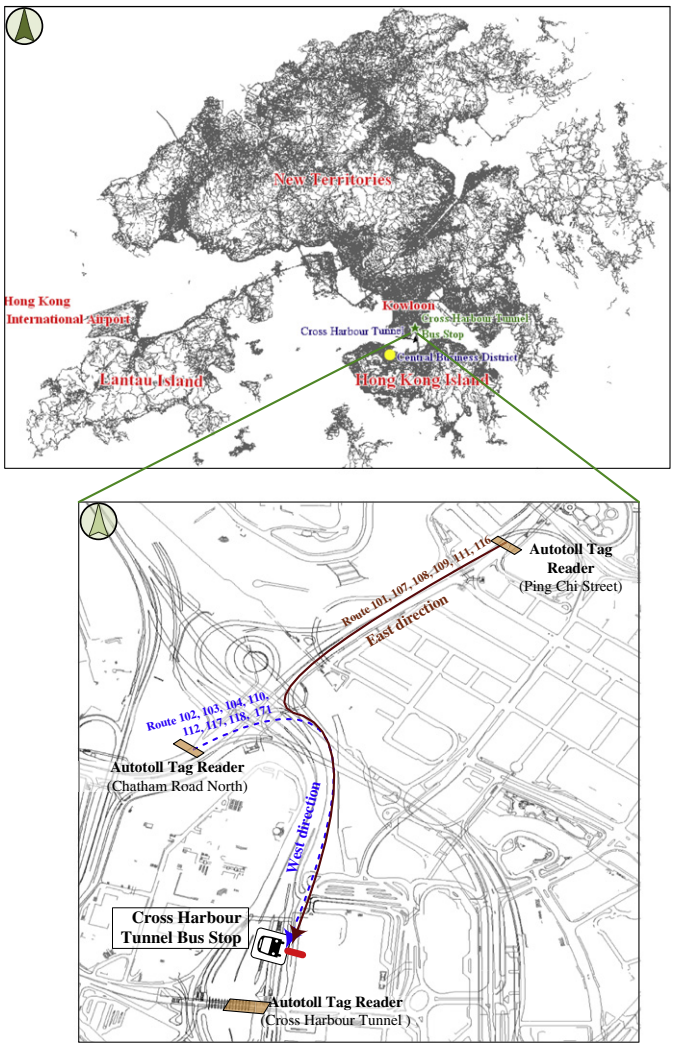

Fig. 7. Locations of the survey paths and the Autotoll tag readers.

(08:00–10:00). In the surveys, we have recorded the route no. and license plate of each bus passing through the CHT bus stop. Then, by a manual license plate matching with the Autotoll tags records, the actual bus arrival time at the CHT bus stop and the time passing through the Autotoll tag readers (CRN and PCS) can be acquired. The data recorded by the Autotoll tag readers at the toll-booths of the CHT have been used to check the license plate reading and the time.

A data filtering algorithm (Tam and Lam, 2008) was then applied to the observations collected from the surveys in order to filter out the outliers. The numbers of the valid observations on the 3 days are 237, 228 and 224, respectively. We divide these routes into two groups according to the different directions (CRN and PCS to CHT stop). Table 1 shows the number of the valid observations of each route at each day, and the main descriptive statistics for the collected travel time. From Table 1, the average bus travel time of the west direction (from CRN to the CHT bus stop) is obviously more than that of the east direction (from PCS to the CHT bus stop). The bus running times of the east direction varied from 170 to 485s and the average time is around 291s. The bus running times of the west direction are from 275 to 662s and the average time is around 449s. RMSEs of the east and west directions are 57.5 and 67.3s, respectively.

3.3. Model identifications

Before model identifications, the parameter δ of the weighted average running time should be determined. By sensitivity tests, bus running times of three preceding buses are used to calculate the weighted average running time in this study, i.e., $\delta = 3$. In model identifications, the observations are first classified by the bus route no. and the inputs of the prediction models are computed. Then, the observations on 11 May 2010 are set aside as testing data. The observations on 12 May and 8 June 2010 are selected as training data to calibrate the prediction models. To have the same basis of comparison, the same training and verification sets are used for all models.

Table 1
Sample sizes of each route and descriptive statistics for the collected data.

Route no.		Sample sizes (vehicle)			Descriptive statistics			
		11-May	12-May	8-June	Min (s)	Max (s)	Ave (s)	RMSE
PCS–CHT bus stop	101	32	29	26	197	424	300.40	61.29
	107	13	12	16	171	485	307.63	68.34
	108	11	9	8	229	436	292.72	45.43
	109	10	9	10	193	409	267.62	56.61
	111	29	30	34	170	456	291.58	58.46
	116	23	22	21	176	432	287.90	55.15
CRN–CHT bus stop	102	20	17	15	327	662	455.14	77.82
	103	11	8	8	360	584	442.62	60.22
	104	27	30	29	315	621	432.40	61.01
	110	7	8	6	369	616	472.39	72.52
	112	14	16	12	302	629	448.73	70.85
	117	5	6	7	352	562	442.06	63.98
	118	18	14	17	275	637	457.60	76.94
	171	17	18	15	308	639	438.43	66.89

Table 2
SVM and ANN models with different input parameters.

Model	Input parameters				Average MAE (s)	
	Bus time interval among the route set ($t^l_{L,n}$)	Bus time interval of the same route ($t^i_{l,n}$)	Weighted average bus running time among the route set ($\bar{t}^r_{L,n}$)	Bus running time of the same route ($t^r_{l,n}$)	SVM	ANN
1	√	√	√		32.82	**34.34**
2		√	√	√	33.37	38.69
3	√	√		√	34.34	36.34
4	√		√	√	35.21	37.01
5	√	√	√	√	**30.39**	35.02
6		√		√	37.76	42.22

3.3.1. Support vector machine models

The previous researches (Yu et al., 2006, 2010b) suggested that radial basis function (RBF) kernel was efficient for bus running time prediction. Thus, RBF kernel function is used for the SVM model in this study. To determine the inputs of the SVM model, sensitivity tests have been conducted. The SVM models with different input variables listed in Table 2 have been calibrated based on the data collected. Table 2 also showed that the average MAE of the prediction for all the routes using the SVM models with different parameters. The prediction errors of each route are shown in Fig. 8.

The models 1–5 integrate bus information of multiple routes to predict the bus arrival time at the bus stop. The model 6 is a standard method using bus information of a single route for the prediction. Obviously, the performance of the single route prediction model is the worst among the six models. This indicates that integrating bus information of multiple routes can improve the accuracy of the arrival time prediction at the bus stop with multiple routes. It is mainly because that bus information of multiple routes can reduce the effect of accidents on the preceding buses that may be the same or different route with the target bus. Furthermore, Fig. 8 also shows that the SVM model 5 can almost provide the best prediction accuracy for each route. Thus, all the four variables $t^l_{L,n}, t^i_{l,n}, \bar{t}^r_{L,n}, t^r_{l,n}$ are considered as the inputs of the SVM model in this study.

Before applying SVM, there are two parameters, C and ε, which are first determined. Parameter C is to determine the trade-off between the model complexity and the degree in the optimization equation. Parameter ε controls the width of the ε-insensitive zone which is used to fit the training data. Referring to the application of bus running time prediction from Yu et al. (2006, 2010b), it is recommended to the constraints of the two parameters which respectively attribute to the range $C \in [2^{-5}, 2^5]$ and $\varepsilon \in [0.1, 0.3]$.

As identifying the parameters in SVM, grid-search is used to pick up the optimal parameter values. Thus, for the bus running time prediction, the two parameters (C, ε) are selected as (2, 0.1).

3.3.2. Artificial neural network

A standard three-layer ANN is used to construct the prediction model for bus running time in this study. Similar to the SVM model, the input parameters of the ANN model are determined based on the results of the sensitivity tests. The combinations of input parameters of the ANN models are the same as the ones of the sensitivity tests of the SVM model. The ANN

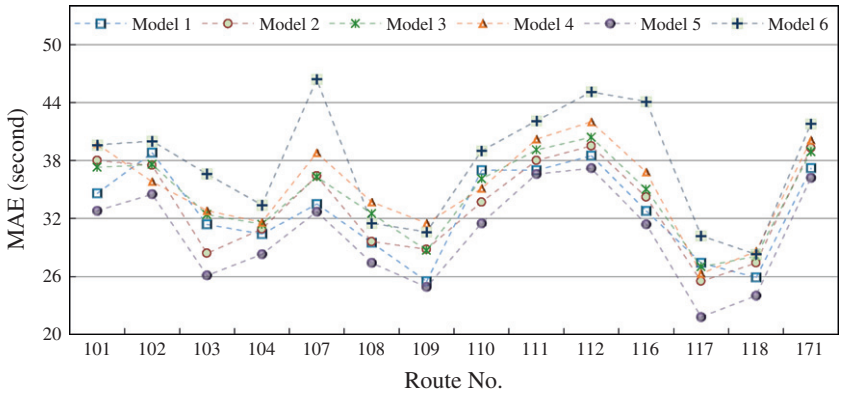

Fig. 8. Prediction errors of the SVM models with different input parameters.

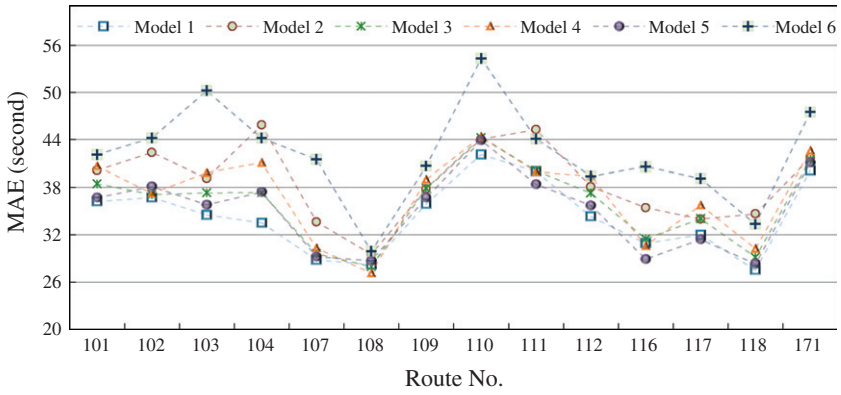

Fig. 9. Prediction errors of the ANN models with different input parameters.

models with different input parameters shown in Table 2 have also been trained by the BP algorithm. Fig. 9 shows the prediction errors using the six ANN models with different input parameters.

Similar to the comparison of the SVM models, the prediction error of the ANN model 6 (the single route prediction model) is the largest among the six ANN models. When comparing the average MAEs of the five ANN models (models 1 to 5) with bus information of multiple routes in Table 2, the ANN model 1 is the best model. However, it can also be observed from Table 2 and Fig. 9 that the prediction errors of the ANN models 1 and 5 are almost the same. To be consistence with the SVM model, the ANN model 5 with $\{t^l_{L,n}, t^i_{l,n}, \bar{t}^r_{L,n}, t^r_{l,n}\}$ is used for bus running time prediction in this study.

After determining the inputs of the ANN model, a scaled conjugate gradient algorithm (Moller, 1993) is used to train the ANN model. The number of hidden neurons is attained as five in this study. Thus, the final ANN model in this study is the ANN model with three-layer and five hidden neurons for bus running time prediction.

3.3.3. k nearest neighbours algorithm

In the *k*-NN model, the *k* nearest matches are determined among all the observations by the weighted distances. When computing the weighted distances, the parameters λ_1, λ_2, λ_3 and λ_4 are decided by the correlation coefficient of each input variable with the bus running time. In this study, $\lambda_1 = 0.268$, $\lambda_2 = 0.117$, $\lambda_3 = 0.217$ and $\lambda_4 = 0.348$, are calibrated respectively. To investigate the suitable value of *k* for the *k*-NN method in this study, sensitivity tests with different values of *k* in the range of 1–4 have been conducted. Fig. 10 shows that the prediction errors of the *k*-NN models with different value of *k*. Obviously, when *k* is set as 1, the prediction error is the largest among the four models. It is because that traffic conditions change rapidly and dynamically and accidents on the preceding bus will greatly affect the prediction accuracy. When *k* reaches 3 or 4, the prediction performance of the model is relatively better than that of the model with $k = 1$ or $k = 2$. It was also found that the difference of the prediction errors between the model with $k = 3$ or $k = 4$ is about 3%. Considering the complex of computation and the prediction accuracy together, $k = 3$ is selected for the *k*-NN model in this study.

3.3.4. Linear regression

Based on the preliminary analysis, it was found that time headways between the target bus and the last preceding bus of any route and of the same route ($t^l_{L,n}$ and $t^i_{l,n}$) are insignificant at 5% level for almost all the routes. Thus, these two variables

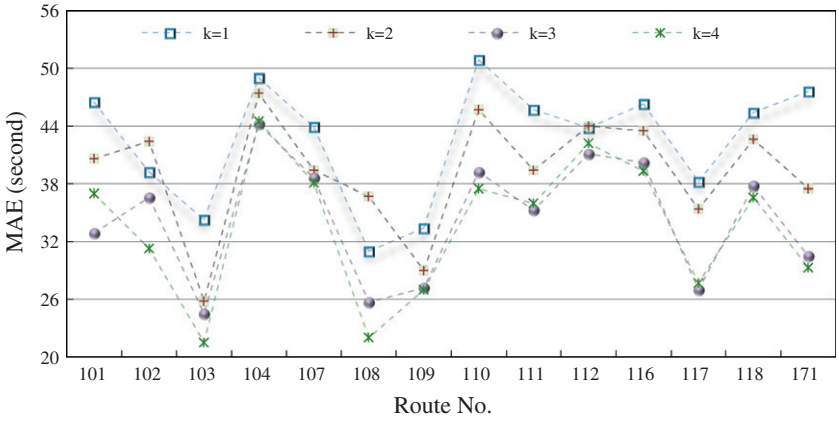

Fig. 10. Prediction errors of the *k*-NN models with different values of *k*.

Table 3
Coefficients of the independent variables in the LR model.

Route no.	$\ln(\bar{t}^r_{L,n})$	$\ln(t^r_{l,n})$	R^2	Route No.	$\ln(\bar{t}^r_{L,n})$	$\ln(t^r_{l,n})$	R^2
101	−0.210	1.222	0.82	103	−0.151	1.158	0.69
107	0.205	0.810*	0.74	104	0.353	0.646**	0.83
108	0.492	0.506	0.73	110	0.310	0.701**	0.79
109	1.122**	−0.123	0.82	112	0.269	0.735*	0.71
111	0.156	0.851*	0.67	117	0.446	0.557	0.84
116	0.143	0.856	0.77	118	0.176	0.831*	0.70
102	0.046	0.968*	0.81	171	0.073	0.931	0.83

* Significant at 0.05 level.
** Significant at 0.01 level.

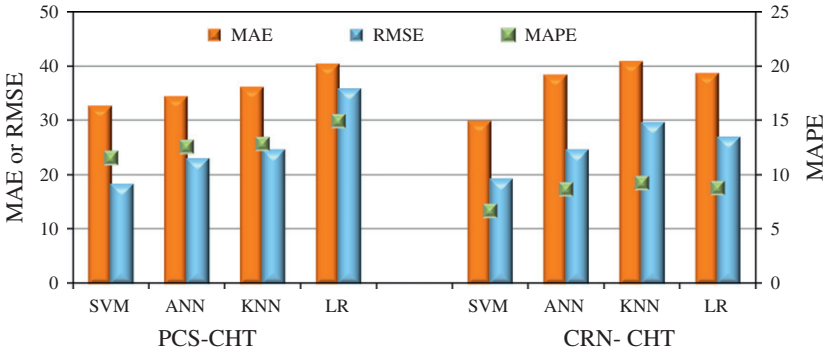

Fig. 11. Comparison of the performance among four methods.

were not adopted in the LR model. The coefficients of the independent variables in the resultant LR model for each of the bus route are shown in Table 3.

There is some difference in the coefficients of the two input variables in Table 3. $\ln(t^r_{l,n})$ shows higher influence on the predicted bus running time in most of the bus routes except the route no 109. This indicates that the running time of the preceding bus of the same route can provide more reliable information for the prediction. Furthermore, $\ln(t^r_{l,n})$ is significant at 5% level or 1% level for the route nos. 107, 111, 102, 104, 110, 112 and 118, while $\ln(\bar{t}^r_{L,n})$ is significant for the route No. 109.

3.4. Validation results

In this section, the bus arrival time at the CHT bus stop with multiple routes are forecasted by the SVM, ANN, *k*-NN and LR models. The average value of the MAE, the MAPE and the RMSE of the four models for all the bus routes are summarized in Fig. 11 and details are attached in Appendix A.

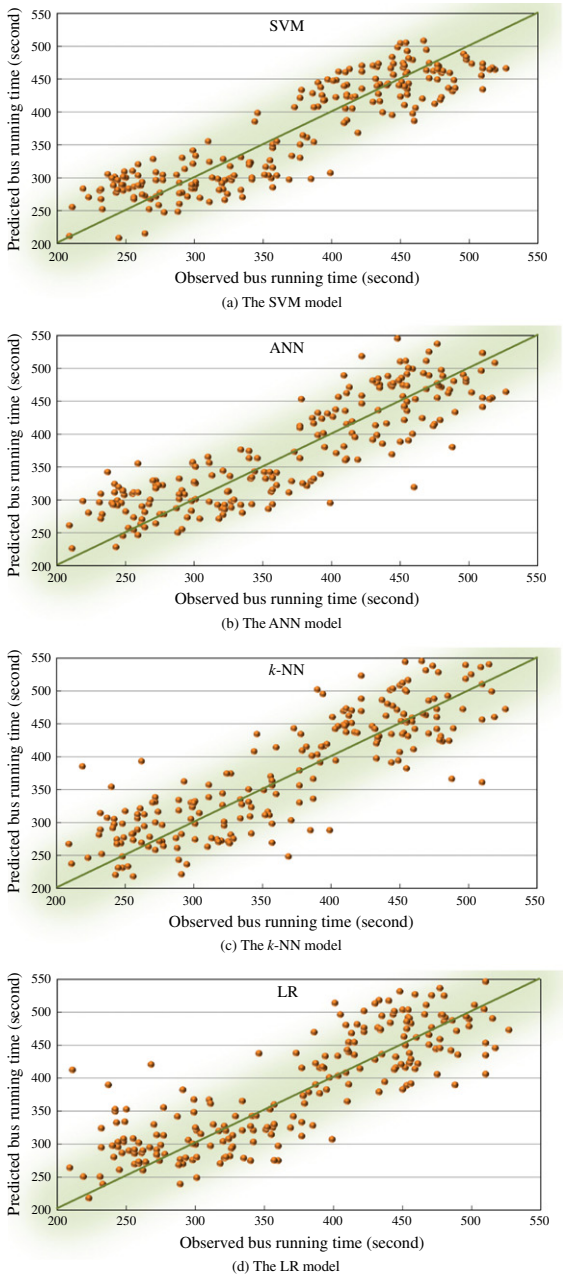

Fig. 12. Predictability of the four models on bus running time.

Table 4
The correlation coefficients (r) and the t-values for the four methods.

	SVM	ANN	k-NN	LR
r	0.9	0.87	0.85	0.84
t-value	−2.2	−1.98	1.57	1.54

Fig. 11 shows the comparison of the performance among four different methods from MAE, MAPE and RMSE, respectively. In Fig. 11, the horizontal axis is divided into two directions: PCS–CHT (the east direction) and CRN–CHT (the west direction). It can be seen that the SVM model has the best prediction performance among the four models in two directions. It is mainly due to that SVM implements the structural risk minimization principle and over-fitting is unlikely to occur with SVM.

Although the performance of the ANN model is worse than that of the SVM, the ANN model outperforms the *k*-NN and LR models. In summary, the performance of the LR model is worst among the four models. However, for the arrival time prediction for the west direction, the LR model is better than the *k*-NN model. From Appendix A, the average MAPEs of the SVM model are 11.5% and 6.69% for the east direction (from PCS to the CHT bus stop) and the west direction (from CRN to the CHT bus stop). For the fourteen bus routes, the MAPEs of the SVM model are from 4.49% to 13.23%, while the MAPEs of the ANN, *k*-NN and LR models are from 6.84% to 15.11%, from 6.94% to 16.89% and 6.78% to 24.99%, respectively. When comparing the maximum error of the prediction for each route using the different methods, it can be found that the maximum prediction error of the SVM model is the lowest except for the prediction for the route nos. 103 and 108. In summary, the SVM model has the best prediction performance among the four models for ten out of fourteen bus routes. Although the performance of the ANN model is slightly better than the one of the *k*-NN model, the *k*-NN model is still an alternative method for bus running time prediction due to its simple structure.

Fig. 12 and Table 4 show the bus running times predicted by the four models against the observed running times. It can be seen from the figure that the results of the SVM model are much closer to the observed data than the other three methods. The correlation coefficient (r) which reflects the accuracy of the bus running time prediction of the four methods is 0.90, 0.87, 0.85, 0.84, respectively. The coefficient implies that the proportion of the predicted bus running times from each model is well fitted with the observed bus running times. Also, it can be seen from the results of the *t*-tests of the four methods that only the SVM and ANN models passed the *t*-tests, whose *t*-values are larger than 1.96. This indicates that the SVM and ANN models are significant at 5% level. In summary, based on the validation results, the performance of the SVM model for bus arrival time prediction at the stop with multiple routes is shown to be satisfactory.

4. Conclusions

This paper investigated the bus arrival time prediction at bus stop with multiple routes. Bus running times of different routes were used to predict the bus arrival times by four proposed models, namely, SVM, ANN, *k*-NN and LR. In order to develop these four proposed models, bus running and arrival time data were collected from the observation surveys near the CHT in Kowloon urban area of Hong Kong. The results showed that the proposed models were more accurate than the models based on bus running times of single route. Moreover, the comparison results showed that the performance of the SVM model was the best among the four models for the bus arrival time prediction. It was also found that *k*-NN was an alternative method for the bus running time prediction compared with ANN. In summary, LR was the worst one among four models since its performance varied from the similarity between the current data and the preceding data, while compared with other three models, LR had the simplest structure.

In this paper, only the bus data was used to estimate the current traffic conditions. Further study will consider more factors such as running times of other vehicles (by type) and traffic flow variation so as to enhance the performance of the proposed prediction models.

Acknowledgements

The work described in this paper was jointly supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region to the Hong Kong Polytechnic University (Project No. PolyU 5195/07E), an internal research grant from the Research Committee of The Hong Kong Polytechnic University (Project No. J-BB7Q), the special grade of the financial support from China Postdoctoral Science Foundation 201003611 and Humanities and Social Sciences Foundation from the Ministry of Education of China 10YJC630357.

Appendix A. The details of results of four methods

Route no.		PCS–CHT							CRN–CHT								
		101	107	108	109	111	116	AVE	102	103	104	110	112	117	118	171	AVE
MAE (s)	SVM	39.4	34.9	28.7	25.2	34.2	32.6	32.5	32.9	24.4	29.4	34.8	32.6	21.3	26.2	36.5	29.8
	ANN	38.7	29.5	29.5	35.0	41.3	32.0	34.3	38.8	35.4	32.0	58.9	35.7	31.0	29.0	45.5	38.3
	k-NN	40.6	34.6	28.8	28.6	46.9	36.6	36.0	47.9	30.2	40.3	60.0	32.4	38.6	36.4	40.5	40.8
	LR	47.1	41.2	21.8	59.7	40.6	31.6	40.3	45.5	31.2	39.4	32.4	31.2	53.3	34.0	41.4	38.6
MAPE (%)	SVM	13.2	11.7	9.7	11.3	12.3	11.0	11.6	7.2	5.5	7.0	7.7	7.1	4.5	6.2	8.3	6.7
	ANN	12.8	11.1	10.4	14.5	15.1	11.4	12.6	8.6	8.4	7.4	13.0	7.8	6.9	6.8	10.5	8.7
	k-NN	13.6	10.8	10.0	12.7	16.9	13.1	12.8	10.4	7.3	9.5	13.6	6.9	8.6	8.7	9.1	9.3
	LR	16.8	14.7	7.8	25.0	14.1	11.2	14.9	10.1	7.6	9.1	7.4	6.8	11.7	8.3	9.3	8.8

(*continued on next page*)

Appendix A (*continued*)

Route no.		PCS–CHT							CRN–CHT								
		101	107	108	109	111	116	AVE	102	103	104	110	112	117	118	171	AVE
Max (s)	SVM	91.7	68.4	55.7	49.0	87.2	72.3	70.7	75.7	62.3	57.5	73.8	62.7	43.7	52.3	61.0	61.1
	ANN	103.8	104.7	55.0	74.0	96.3	77.1	85.2	97.6	79.8	107.7	141.0	65.8	44.0	62.0	96.3	86.8
	k-NN	120.6	70.4	48.4	58.2	165.6	81.5	90.8	149.2	59.7	122.1	130.6	64.4	60.0	108.4	101.0	99.4
	LR	107.3	152.5	51.4	201.4	110.5	91.9	119.2	83.4	71.4	98.4	83.0	69.3	101.2	112.6	104.1	90.4
RMSE	SVM	20.2	14.2	16.2	17.2	21.9	19.7	18.2	20.6	21.2	16.9	25.1	19.1	15.8	16.5	18.3	19.2
	ANN	27.6	27.8	13.4	24.0	23.2	21.6	22.9	28.3	23.8	24.2	42.1	20.5	12.1	21.9	23.2	24.5
	k-NN	32.2	23.8	14.8	16.5	37.1	22.4	24.5	37.7	22.3	32.7	38.0	21.3	25.8	30.5	28.0	29.5
	LR	31.8	39.8	17.6	73.7	25.9	25.5	35.7	20.2	23.0	30.0	27.2	21.3	30.8	32.8	29.5	26.8

References

Cathey, F.W., Dailey, D.J., 2003. A prescription for transit arrival/departure prediction using automatic vehicle location data. Transportation Research Part C 11 (3), 241–264.

Chan, K.S., Lam, W.H.K., Tam, M.L., 2009. Real-time estimation of arterial travel times with spatial travel time covariance relationships. Transportation Research Record 2121, 102–109.

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.

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.

Chien, S.I.J., Kuchipudi, C.M., 2003. Dynamic travel time prediction with real-time and historic data. Journal of Transportation Engineering 129 (6), 608–616.

Chien, S.I.J., Ding, Y., Wei, C., 2002. Dynamic bus arrival time prediction with artificial neural networks. Journal of Transportation Engineering 128 (5), 429–438.

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

Ding, Y., Chien, S., 2000. The prediction of bus arrival times with link-based artificial neural networks. In: Proceedings of the Fifth Joint Conference on Information Sciences, Atlantic City, NJ.

Huang, S.H., Ran, B., 2003. An Application of Neural Network on Traffic Speed Prediction under Adverse Weather Condition. The 82nd Annual Meeting of the Transportation Research Board, Washington, DC.

Jeong, R., Rilett, L.R., 2004. Bus Arrival Time Prediction Using Artificial Neural Network Model. In: 7th International IEEE Conference on Intelligent Transportation Systems: (ITSC 2004), Washington DC.

Moller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6 (4), 523–533.

Park, S.H., Jeong, Y.J., Kim, T.J., 2007. Transit travel time forecasts for location-based queries: implementation and evaluation. Journal of the Eastern Asia Society for Transportation Studies 7, 1859–1869.

Shalaby, A., Farhan, A., 2004. Prediction models of bus arrival and departure times using AVL and APC data. Journal of Public Transportation 7 (1), 41–61.

Smith, B.L., Williams, B.M., Oswald, R.K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C 10 (4), 303–321.

Tam, M.L., Lam, W.H.K., 2008. Using automatic vehicle identification data for travel time estimation in Hong Kong. Transportmetrica 4 (3), 179–194.

Tam, M.L., Lam, W.H.K., 2009. Short-term Travel Time Prediction for Congested Urban Road Networks. The 88th Annual Meeting of the Transportation Research Board, Washington, DC.

van Hinsbergen, C.P.I.J., 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 17(5), 498–509.

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 13(5–6), 347–369.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Transactions on Neural Networks 10 (5), 988–999.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer, New York, NY, USA.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2005. Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. Transportation Research Part C 13 (3), 211–234.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2006. Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. Transportation Research Part C 14 (5), 351–367.

Wall, Z., Dailey, D.J., 1999. An algorithm for predicting the arrival time of mass transit vehicle using Automatic Vehicle Location data. The 78th Annual Meeting of the Transportation Research Board, Washington, DC.

You, J.S., Kim, T.J., 2000. Development and evaluation of a hybrid travel time forecasting model. Transportation Research Part C 8 (1–6), 231–256.

Yu, B., Yang, Z.Z., Yao, B.Z., 2006. Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems 10 (4), 151–158.

Yu, B., Yang, Z.Z., Chen, K., Yu, B., 2010a. A hybrid model for bus arrival time prediction. Journal of Advanced Transportation 44 (3), 193–204.

Yu, B., Yang, Z.Z., Wang, J., 2010b. Bus travel-time prediction based on bus speed. Proceedings of the Institution of Civil Engineers – Transport 163 (1), 3–7.

Zhang, X.Y., Rice, J.A., 2003. Short-term travel time prediction. Transportation Research Part C 11 (3–4), 187–210.