Bus travel time prediction under high variability conditions

Kranthi Kumar Reddy, B. Anil Kumar and Lelitha Vanajakshi*

Department of Civil Engineering, Indian Institute of Technology Madras, Chennai 600 036, India

Bus travel times are prone to high variability, especially in countries that lack lane discipline and have heterogeneous vehicle profiles. This leads to negative impacts such as bus bunching, increase in passenger waiting time and cost of operation. One way to minimize these issues is to accurately predict bus travel times. To address this, the present study used a model-based approach by incorporating mean and variance in the formulation of the model. However, the accuracy of prediction did not improve significantly and hence a machine learning-based approach was considered. Support vector machines were used and prediction was done using ν-support vector regression with linear kernel function. The proposed scheme was implemented in Chennai using data collected from public transport buses fitted with global positioning system. The performance of the proposed method was analysed along the route, across subsections and at bus stops. Results show a clear improvement in performance under high variance conditions.

Keywords: Bus travel time, high variance conditions, prediction accuracy, support vector machines.

AUTOMATIC vehicle location (AVL) systems are being implemented in public transit systems in many Indian cities. The main benefit of using AVL systems is the availability of high quality and quantity of pertinent data such as vehicle locations, speed and travel times. Such information can be used to improve the reliability of transit passenger information system and transit management system, which can, in turn, improve the overall service quality. However, travel times in urban areas are prone to high degrees of variability due to the presence of signals, traffic congestion, geometric conditions of roads and weather conditions. This is particularly serious in the heterogeneous and lane-less traffic existing in countries such as India. Under such traffic conditions, various types of vehicles such as cars, buses, light and heavy motor vehicles, two-wheelers and bicycles share the road without any segregation for the various vehicle types. This leads to high levels of uncertainties and variability in traffic characteristics such as travel time. Furthermore, transit vehicles are frequently disturbed by congestions on service routes at different times of the day due to intersection delays, variations in demand, and excessive dwell times at bus stops. All these contribute to stochasticity, resulting in significant deviations in overall travel times. Problems such as bus bunching, increase in passenger waiting times, increase in the cost of operation, deterioration of schedule adherence, etc. result from such stochasticity, which could discourage passengers from using the transit system. One solution to address this problem is by providing information of bus arrival times and expected delays at all bus stops.

Developing models to predict bus travel times under such conditions is a difficult task. Prediction methods that work elsewhere in the world may be impractical for the aforementioned Indian traffic conditions. The following section reviews existing studies that have been carried out in the area of bus travel time predictions under both homogeneous and heterogeneous traffic conditions.

Literature review

There have been many studies on the prediction of travel times using techniques such as historical and real-time methods[1–4], statistical methods[5–7], machine learning methods[8–11] and model based methods[12–15]. Most of these studies were developed or tested for lane-disciplined and homogeneous traffic conditions. However, traffic conditions in many countries are different due to lack of lane discipline and heterogeneity. Very few studies are available for traffic under such conditions[16–20]. Even these studies do not focus on addressing the problem of high variability. The present study focuses on this problem and develops a system for bus travel time prediction (BTTP) under high variability conditions.

A review of the literature shows that Kalman filtering technique (KFT) and support vector machines (SVMs) are promising prediction tools to address the high variability problem. Dailey *et al.*[12] used a combination of AVL and historic database to predict travel time using KFT and statistical analysis. Cathey and Dailey[13] used bus travel time data as inputs to predict travel times using KFT that involved three components, viz. tracker, filter and predictor. Shalaby and Farhan[14] used a combination of AVL and automatic passenger count (APC) data to predict travel time using KFT. Nanthawichit *et al.*[15] used

*For correspondence. (e-mail: lelitha@iitm.ac.in) a combination of global positioning system (GPS) and loop detectors to estimate travel time using KFT.

All of the above studies were performed under homogeneous traffic conditions; only a limited number of studies have been reported for heterogeneous traffic. Vanajakshi *et al.*[16] proposed a model-based method using a space discretization approach to predict bus travel time. They used GPS data of previous two bus trips to predict next bus travel/arrival time using KFT. Padmanabhan *et al.*[17] extended the above study by explicitly analysing the dwell times. However, the above studies used data from two previous bus trips alone as inputs, without considering the patterns in travel time. Kumar and Vanajakshi[18] subsequently identified the most significant trips and incorporated them in the analysis[18]. The study analysed weekly patterns and trip-wise patterns in bus travel time data, and reported a strong weekly pattern followed by a trip-wise pattern. Vivek *et al.*[19] used GPS data to predict bus travel time using ANN, and the results were compared with those of space discretization methods. Kumar *et al.*[20] used a time discretization approach to predict bus travel time by considering temporal evolution in travel time. The results were compared with the space discretization approach[16], considering the evolution of travel time between spatial sections. It was shown that time discretization performed better than space discretization.

Bin *et al.*[8] used SVM to predict bus arrival time for four patterns, viz. peak traffic on sunny day (SP), off-peak traffic on sunny day (SO), peak traffic on rainy day (RP) and off-peak traffic on rainy day (RO). Results were reported to be promising compared to the ANN method. Wu *et al.*[9] used data from intelligent transportation web service project (ITWS) to predict travel time using SVM and showed that SVM gave better results than historic and real-time methods. Vanajakshi and Rilett[10,11] used SVM and ANN for short-term prediction of traffic parameters and reported SVM as a viable alternative to ANN. The study concluded that SVM would be a better choice for the prediction of travel time, when only a small amount of data is available for training, or when the training data have more variations[10]. Based on these reported advantages of SVM, especially when the variability is high, the present study explored its use for BTTP under Indian traffic conditions.

The main objective of the present study is to predict bus travel time/arrival time, paying special attention to the high variance problem. The first part of the study reformulates an existing model-based approach reported by Kumar *et al.*[20], to take high variance into account. The study reported high errors in sections with high variability. To address this issue, the problem was reformulated to explicitly incorporate the variance. This was attempted due to the possibility of incorporating the variance of the process disturbance and measurement noise into the Kalman filter formulation.

The second part uses SVM to address the same problem. From the literature, it was found that the SVM technique performs better than ANN and other standard techniques for prediction problems when the variability in data is high[10]. However, no significant studies have been reported on the use of SVM to predict bus travel times under Indian traffic conditions. Thus the present study uses the SVM technique for bus travel time prediction under Indian traffic conditions.

Data collection and preliminary analysis

GPS units are commonly used to collect data for applications that involve continuous tracking of vehicles and providing their location information at selected intervals. In the present study, data were collected using GPS units fixed in buses belonging to the Metropolitan Transport Corporation (MTC) in the city of Chennai, Tamil Nadu, India. The route selected for the study was 19B, which spans 30 km with varying land use, and traffic and geometric characteristics. It connects Saidapet, a major commercial area located in the southern part of the city, to Kelambakkam, a sub-urban area of the city. There are 20 bus stops and 13 intersections in this route. Table 1 gives the distance between the bus stops and cumulative distance from the initial bus stop.

The average time headway between two consecutive vehicles in this route is about 15–30 min. Data were sent every 5 sec from 6 a.m. to 8 p.m., and data collected over 30 days were used in the study. The collected GPS data included the latitude and longitude information at fixed time intervals, time stamp corresponding to each entry and ID of the GPS units. Data were communicated in real time through general packet radio service (GPRS) and stored using sequential query language (SQL) database. Individual files were generated separately for each day. The distance travelled between two consecutive time intervals was then calculated using Haversine formulae[21], which provides the great circle distance between two points on a sphere from their latitudes and longitudes as

$$\text{Distance } (d) = 2r \arcsin$$

$$\left(\sqrt{\text{haversin}(\varphi_2 - \varphi_1) + \cos \varphi_1 \cos \varphi_2 \text{haversin}(\lambda_2 - \lambda_1)}\right), \quad (1)$$

where φ_1, φ_2 indicate the latitude of points 1 and 2; λ_1, λ_2 indicate the longitude of points 1 and 2, and r is the radius of the earth. Thus, the processed data comprised of the distance between consecutive locations of all the buses and corresponding time stamps. The entire road stretch was divided into subsections of 100 m length, and linear interpolation technique was adopted to calculate the time taken to cover each subsection. In the present study, the travel time variations were considered over time, similar to Kumar *et al.*[20]. The collected data were grouped into 14 time periods of one hour interval each, in order to visualize the travel time variation within a section over

Table 1. Distance between bus stops in 19B route

Bus stop	Distance between bus stops (km)	Cumulative distance from the initial bus stop (km)
Kelambakkam	0.00	0.00
Hindusthan Engineering College	2.51	2.51
SIPCOT	3.40	5.91
Navallur	1.61	7.52
Navalur Church	2.50	10.02
Semmaancheri	1.01	11.03
Kumaran Nagar	1.28	12.31
Shozhinganallur PO Office	1.43	03.74
Karapakkam	1.81	15.55
TCS	0.41	15.96
Mootachavadi	1.46	17.42
Mettupakkam	0.79	18.21
Thorapakkam	0.60	18.81
Tirumailai Nagar	1.25	20.06
Kanadachavadi	1.66	21.72
Lattice Bridge	1.73	23.45
Womens Poytechnic College	1.36	24.80
Madhya Kailash	1.01	25.82
Engineering College	0.82	26.64
Saidapet	3.30	29.94

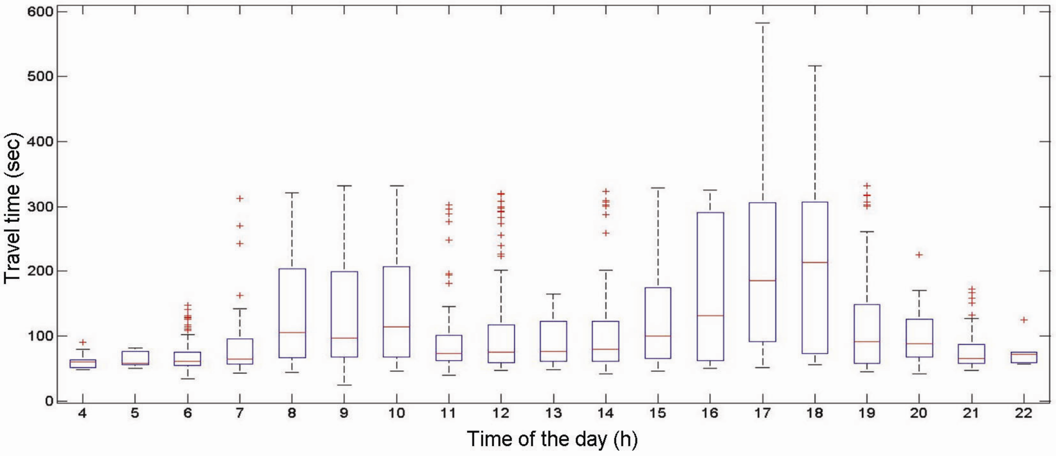


Figure 1. Travel-time variation in subsection 46 over a period of one week.

many days. Figure 1 shows the variations in travel times over a period of one week for a typical subsection, viz. subsection 46.

From Figure 1, it can be seen that travel times from 8 a.m. to 10 a.m. and 4 to 7 p.m. are relatively high, indicating peak hours. It can also be observed that the peak hours have more variance than the off-peak hours.

Methodology

Model-based approach

A promising method to predict bus travel/arrival times under heterogeneous traffic conditions is the time discretization-based model reported by Kumar *et al.*[20]. This is used as the base approach in the present study. Further analysis of the results reported in that study[20] showed that the performance of the method was lower for subsections with high mean and variance. In order to analyse the reasons for this behaviour, these critical subsections were located on a map. They were found to be mostly around intersections and bus stops, which leads to high mean and variance and in turn, to reduced performance. Hence, the first part of the study explicitly incorporated the variance into the model formulation to capture the high variance. In time discretization approach, the section was discretized into smaller subsections. The travel time of a bus in the upcoming time intervals was predicted using the data obtained from many earlier bus trips from the same

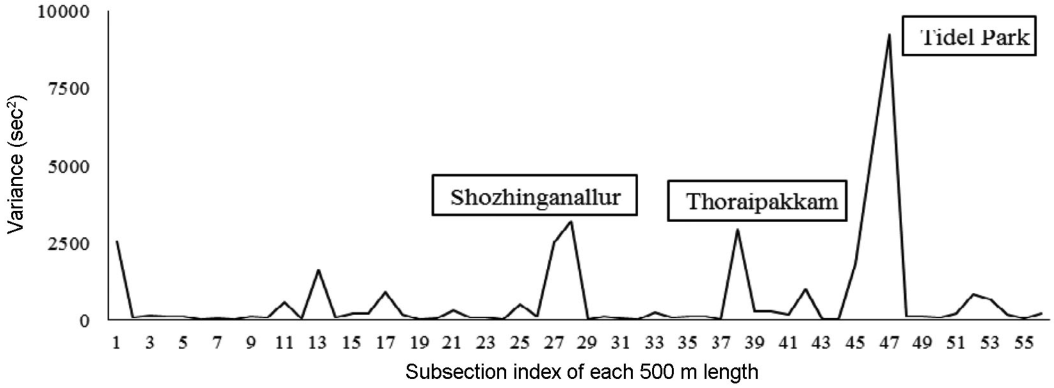


Figure 2. Variation in travel time across various subsections of the study route.

subsection. The model hypothesized a temporal relation in travel time and proposed a method to predict travel/arrival time.

KFT was used as the estimation tool[22]. It can be used to estimate state variables, which are used to characterize system/processes, if the system equations can be represented in state space form. Implementation of the Kalman filter requires dynamic and statistic information of the system disturbances and measurement errors. It uses the model and system inputs to predict the *a priori* state estimate and uses the output measurements to obtain the *a posteriori* state estimate. Overall, it is a recursive algorithm, so that new measurements can be processed when they are obtained. It needs only the current instant state estimate and current input and output measurements to calculate the state estimate of the next instant. The inputs for such a prediction method were the travel time data from several previous buses in the section under consideration.

In the present study, the existing method was modified by incorporating mean and variance of travel time in each subsection separately, to capture the variability in travel time. The evolution of travel time between various travel time intervals in a given subsection is assumed to be

$$x(t+1) = a(t)x(t) + w(t), \tag{2}$$

where $a(t)$ is a parameter that relates the travel time in a given subsection over different trips, $x(t)$ is the time taken to travel for a given subsection at time interval, t and $w(t)$ is the associated process disturbance. The measurement process was assumed to be governed by

$$z(t) = x(t) + v(t), \tag{3}$$

where $z(t)$ is the measured travel time in a given subsection at time t, and $v(t)$ is the measurement noise. It was further assumed that both $v(t)$ and $w(t)$ are zero mean white Gaussian noise signals, with $Q(t)$ and $R(t)$ being the corresponding variances.

As can be seen from Figure 1, apart from the median travel time, the variance of travel times increases during peak hours. Figure 2 shows the change in variance spatially. It can be seen that there are selected sections where the variance is much higher than other sections. These may correspond to sections with bigger intersections or bus stops.

Implementation using model-based approach: In order to address the issue of high variability, the $Q(t)$ and $R(t)$ values, which represent the variance of the process disturbance and measurement noise in the time discretization method respectively, are updated using the latest available travel times in the same subsection as detailed in the steps below. Thus, the proposed scheme needs two sets of data for implementation, in which one set is used for the time update equations to calculate the parameter $a(t)$ and the other for the measurement update equations to generate the *a posteriori* estimate of travel time. The input data were taken as mentioned in Kumar *et al.*[20]. The steps followed were the same as those in the earlier study until step 3, as follows:

1. The route under consideration was divided into N subsections of equal length (100 m).
2. Let the length of the dataset 1 be g. The travel time data from dataset 1 were used to obtain the value of $a(t)$ through

$$a(t) = \frac{x(t+1)}{x(t)}, \quad t = 1,2,3,\ldots,(g-1). \tag{4}$$

3. Let $x_{\mathrm{TV}}(t)$ denote travel time of the test vehicles (TV) (the vehicle for which the travel time needs to be predicted) to cover a given subsection. It is assumed that

$$E[x_{\mathrm{TV}}(1)] = \hat{x}(1), \tag{5}$$

$$E[x_{\mathrm{TV}}(1) - \hat{x}(1)^2] = P(1), \tag{6}$$

where $\hat{x}(t)$ is the travel time estimate of a TV in the tth time interval. Step 4 incorporated the actual mean and variance in travel time as explained below.

4. For $t = 2, 3, 4,\ldots, (g-1)$, the following steps were performed:

a. The *a priori* travel time estimate was calculated using $\hat{x}^{-}(t+1) = a(t)\hat{x}^{+}(t)$, where the superscripts '–' and '+' denote the *a priori* and *a posteriori* estimates respectively.

b. The *a priori* error variance, P^{-} was calculated using

$$P^{-}(t+1) = a(t)P^{+}(t)a(t) + Q(t). \tag{7}$$

c. The Kalman gain, K was calculated using

$$K(t+1) = P^{-}(t+1)[P^{-}(t+1) + R(t+1)]^{-1}. \tag{8}$$

d. The values of $Q(t)$ and $R(t)$ were calculated as follows:

Using eqs (2) and (3), one can compute $Q(t)$ and $R(t)$ as

$$Q(t) = \frac{1}{S}\sum_{i=k-S}^{k}(w(i) - \bar{w}(k))^2, \tag{9}$$

$$R(k) = \frac{1}{S}\sum_{i=k-S}^{S}(v(i) - \bar{v}(k))^2, \tag{10}$$

where

$$\bar{w}(k) = \frac{\sum_{i=k-S}^{k} w(i)}{S}, \quad \bar{v}(k) = \frac{\sum_{i=k-S}^{k} v(i)}{S},$$

S is the number of previous buses travel time data used to compute the error covariance.

e. The *a posteriori* travel time estimate was calculated by considering the mean error obtained from the measurement noise and error variance using

$$\hat{x}^{+}(t+1) = \hat{x}^{-}(t+1) + K(t+1)$$
$$[z(t+1) - \hat{x}^{-}(t+1) - \bar{x}(t+1)], \tag{11}$$

$$P^{+}(t+1) = [1 - K(t+1)]P^{-}(t+1). \tag{12}$$

Thus, the objective here is to predict the travel/arrival time of TV using the travel times obtained from several previous vehicles in the given subsection. The above scheme was implemented in MATLAB. Figure 3 presents a sample plot showing a comparison of the predicted and measured travel times over the study stretch. For evaluating the effect of explicitly incorporating variance into the formulation (known as modified time discretization approach; MTDA), results obtained using the base method without incorporating variance (known as the base method) were also plotted in the figure.

The quantification of errors was done using mean absolute percentage error (MAPE) (eq. (13)) and the values obtained were 19.95% and 20.94% for the proposed and base approaches. This was calculated for multiple trips and the results are presented in Figure 4.

$$\text{MAPE} = \frac{\sum_{i=1}^{N}\frac{x_{\text{p}} - x_{\text{a}}}{x_{\text{a}}}}{N}, \tag{13}$$

where X_{p} is the predicted travel time obtained from the prediction algorithm to cover a given subsection and X_{a} is the corresponding travel time measured from the field.

From Figure 4, it can be observed that the predicted travel times closely match the actual travel times in both cases. However, it can be observed that incorporation of variance into the formulation did not show much difference in performance or led only to a slight improvement in performance, indicating the need for some other prediction method. The literature showed SVM to perform better under high variance situation[10], and hence this was used as detailed in the next section.

Support vector machines

SVMs are learning systems that use a hypothetical space of linear functions in a high-dimensional feature space, trained with a learning algorithm. Figure 5 explains the concept of a SVM. The main idea behind a SVM is that, for a given training sample, a hyper plane is constructed as the decision surface in such a way that the margin of separation between positive and negative examples is maximized[23]. This means that two sets of data points in the input space (which may be of dimension d) that are nonlinearly separable can be transferred into higher dimension (D) space using proper kernel function, and can thus be made linearly separable. The higher dimension should be greater than input dimensions ($D \gg d$) (ref. 23). In this higher dimension, a hyper plane can be constructed between the two sets of data points such that the margin of separation between the two is maximized. SVMs are based on the structural risk minimization (SRM) inductive principle, which seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and confidence level. The basic idea of support vector regression (SVR) is to map the data into high-dimensional feature space via nonlinear mapping and perform linear regression in this space. This linear regression in high-dimension space is equivalent to nonlinear regression in the low-dimension input space.

Consider a set of training data points (x_1, y_1), (x_2, y_2), …, $(x_{\text{n}}, y_{\text{n}})$, where x_n is an n-dimensional input vector such as previous travel times of current segment, and y_{n} is the desired value. Let $\hat{y}_{\text{n}}$ be the predicted value such as travel

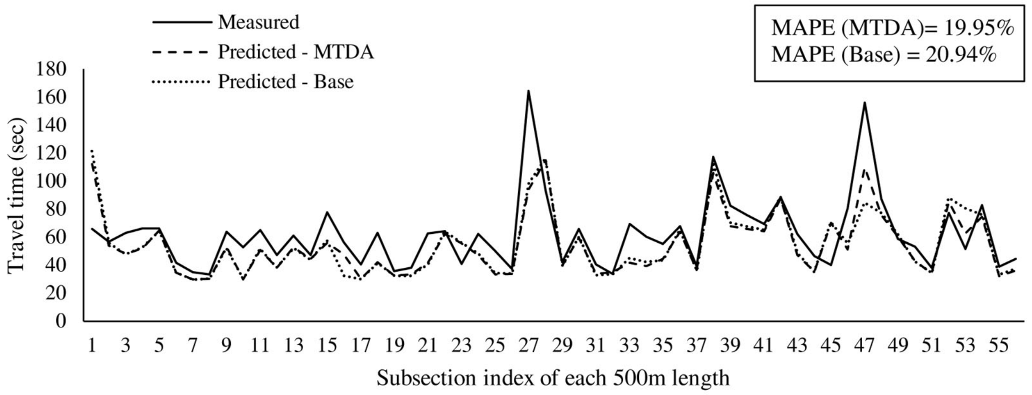


Figure 3. Comparison of predicted and actual travel times for a sample trip.

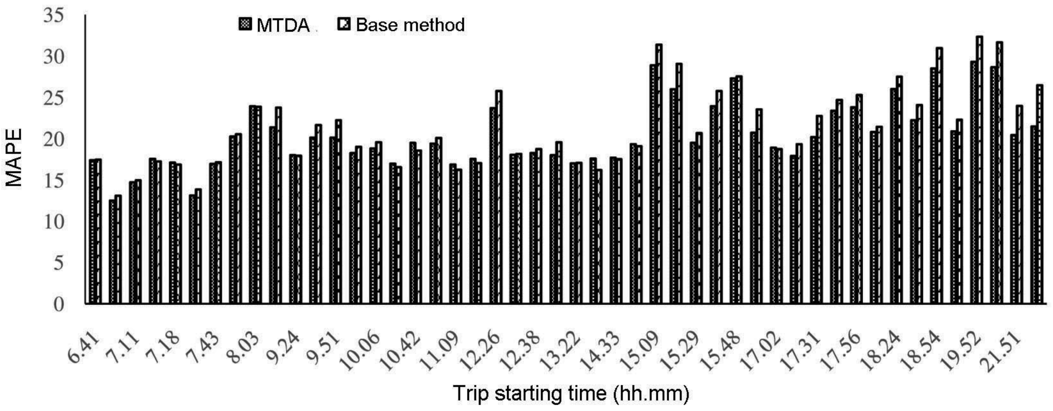


Figure 4. Errors obtained for various trips on a day using modified time discretization approach (MTDA) and base methods.

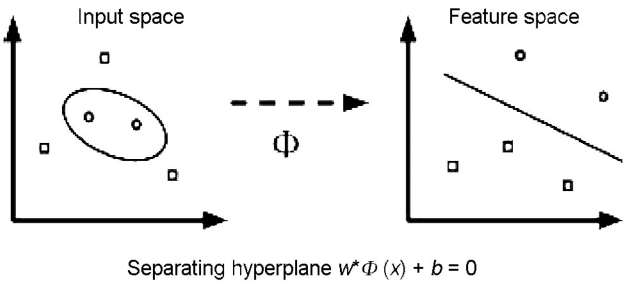


Figure 5. Basic idea of a support vector machine (SVM)[9].

time of next segment, and n is the number of training samples. Then, the output data vector can be in the form

$$y = f(x). \tag{14}$$

SVM approximates the function in eq. (14) using the following form

$$\hat{y}(\bar{x},\bar{\omega}) = \sum_{i=1}^{n} \omega_i \phi_i(x) + \omega_0 = \bar{\omega}^t \phi(x) + \omega_0, \tag{15}$$

where $\phi(x)$ represents the high-dimensional feature spaces that were nonlinearly mapped from the input space x. The coefficients ω_0, $\bar{\omega}$, etc. are estimated by solving a constrained optimization problem using Lagrangian multiplier method. The regression problem can be solved using ν-support vector regression (SVR) proposed by Scholkopf *et al.*[24]. The optimization equations for ν-SVR are

Cost function

$$\frac{1}{2}\| w \|^2 + c\nu\varepsilon + \frac{c}{N}\sum_{n=1}^{N}(\xi_n - \xi'_n), \tag{16}$$

Constraints:

$$y_n - \hat{y}_n = \varepsilon + \xi_n,$$

$$\hat{y}_n - y_n = \varepsilon + \xi'_n,$$

$$\varepsilon_n \geq 0,$$

$$\xi'_n \geq 0,$$

$$\varepsilon > 0. \tag{17}$$

In dual form, the above can be represented as

$$L_d(\bar{\alpha},\bar{\alpha}^1)=\sum_{n=1}^{N}\varepsilon_n(\alpha_n-\alpha'_n)$$

$$-\frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}(\alpha_m-\alpha'_m)(\alpha_n-\alpha'_n)k(x_m,x_n), \qquad (18)$$

with constraints as

$$\sum_{n=1}^{N}\varepsilon_n(\alpha_n-\alpha'_n)=0,$$

$$0<\alpha_{n1}\alpha'_n\le\frac{C}{N},$$

$$\sum_{n=1}^{N}(\alpha_n-\alpha'_n)\le cv, \qquad (19)$$

where C is the regularization constant, $k(\bar{X}_m,\bar{X}_n)$ is the kernel function used to transform the data into high-dimension feature space, and ε can be assumed as tube size and is the approximation accuracy placed on the training data points, as shown in Figure 6. If the predicted values are within the tube, the loss associated with that point is assumed as zero. Else, the loss is considered as the magnitude of the difference between the predicted value and radius ε of the tube. A large ε can depreciate the approximation accuracy placed on training points. In this study, LIBSVM tool box in MATLAB was used to predict the travel time for the next instances[25], and the kernel function used was a linear kernel of the form

$$K(x_i,x_j)=\gamma(x_i\cdot x_j)+coef, \qquad (20)$$

where K is linear kernel function, γ is width parameter.

Implementation using SVR: Data for a period of one month were used in this study. Of the data collected, 18 days data were used for training, 7 days data for cross-validation and the remaining were used to test the performance. Approximate entropy (ApEn) technique was used to identify the optimum inputs (number of previous subsections). ApEn is a technique used to quantify the amount of regularity and unpredictability of fluctuations in data over time[26]. Figure 7 shows ApEn obtained for the training data with respect to number of previous trips. This value of number of previous trips to be used as input may depend on the time headway of the buses. Since ApEn is the measure of uncertainty, if the time headway decreases, which means that there are more trips within a small interval of time, it will become easy to predict the next trip because uncertainty may decrease. Thus, with smaller headways, the number of previous trips required to predict the next trip may also decrease.

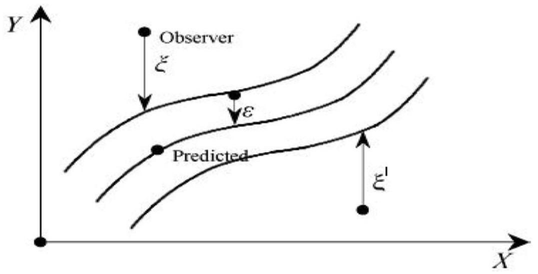


Figure 6. ε-Tube for support vector regression.

It can be seen from Figure 7 that when previous six or more trips are used as input, the uncertainty in the prediction is negligible. Hence, in the present study, data from previous six trips were used as input to predict the travel time of the next bus. Thus, input vector to SVR was a six-dimensional matrix with previous six trips' travel time, and output vector consisted of corresponding next trip travel time. Programs were written in MATLAB to generate input and output vectors for training, validation and testing. The four main unknown parameters of SVR, i.e. ν(Nu), γ (width parameter), C (cost/penalty parameter), and coef (coefficient), were obtained by trial and error.

Results and discussion

The results obtained were compared with the Kalman based proposed method (MTDA) presented in the previous section for a one-week period. In addition to MAPE, mean absolute error (MAE) was also used for comparison, which was calculated as

$$\text{MAE}=\frac{\sum_{i=1}^{N}X_{\text{s}}-X_{\text{TVM}}}{N}. \qquad (21)$$

Comparison of performance across subsections (high and low variance)

One of the main contributions of the present study is to identify a suitable prediction method for BTTP in

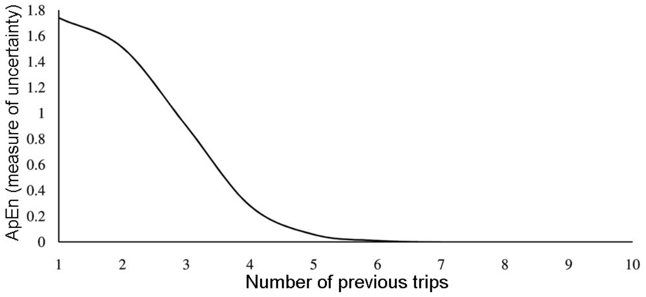


Figure 7. ApEn versus number of previous trips.

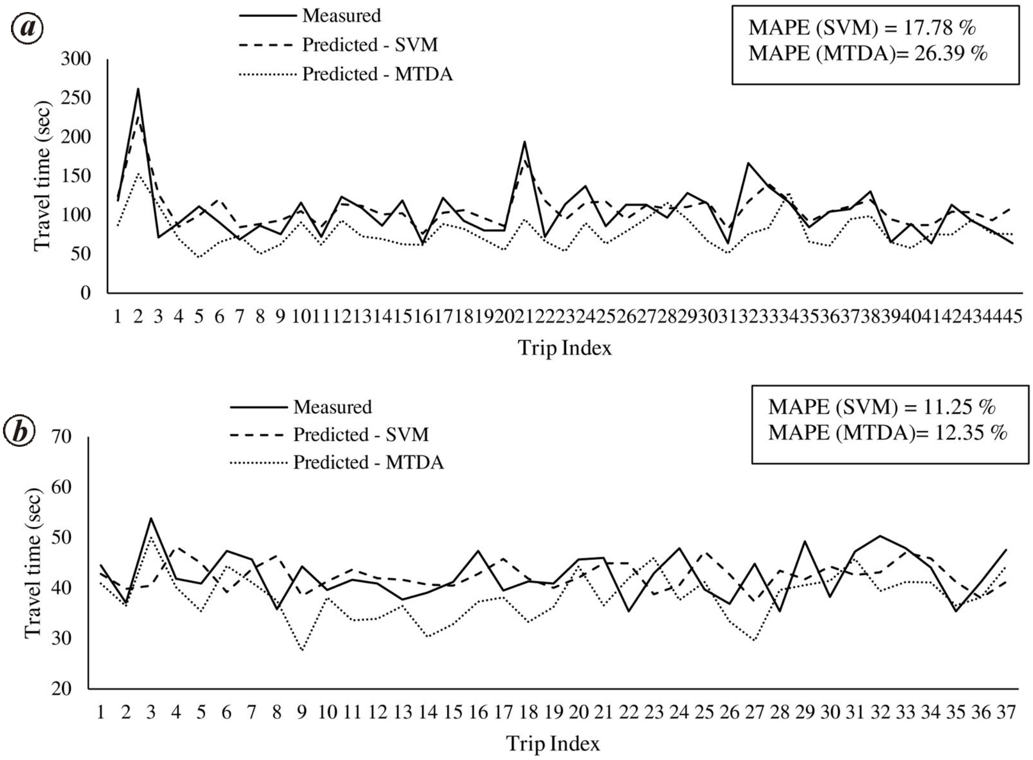


Figure 8. ***a***, Comparison of predicted and actual travel times for (***a***) a high variance section and (***b***) a low variance section.

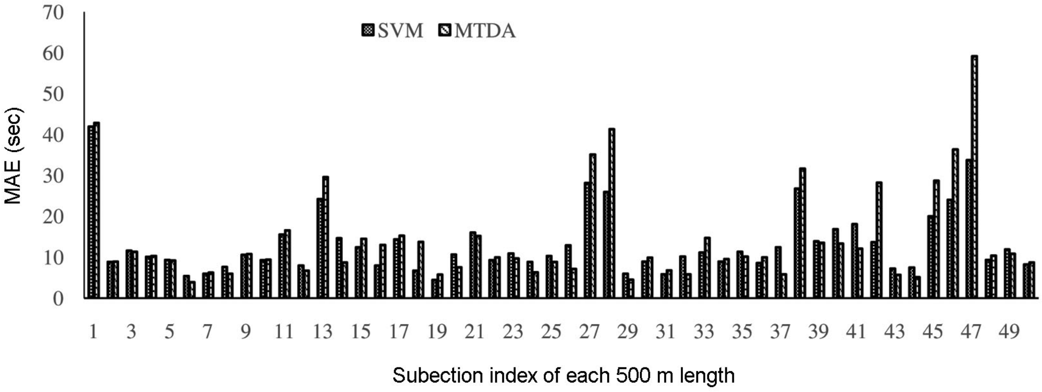


Figure 9. Comparison of deviation from actual travel time across subsections.

sections having high variability. Hence, a comparison of the performance of the SVM method was made with the MTDA for selected trips for each subsection in terms of MAPE and MAE. Figure 8 *a* and *b* shows sample results, where the predicted travel times from SVM and MTDA are shown against the actual travel times for the representative high variance and low variance sections, namely subsections 46 and 19. It can be seen that both SVM and MTDA are able to capture the variations comparably for low variance sections with a MAPE of 11.25% and 12.35%. On the other hand, in case of high variance sections, SVM was able to capture the variation better with an MAPE of 17.78% whereas MTDA had an error of 26.39%.

Similar analyses were carried out for all sections and the results obtained are shown in Figure 8 in terms of deviation from actual travel time for both SVM and MTDA. It can be seen from Figure 9 that the performance of the proposed methods is comparable in low variance sections, whereas for high variance sections the error is much lower for SVM.

Comparison of performance across trips (peak and off-peak trips)

A comparison was also made for trips happening during different times of the day. Figure 10 *a* and *b* shows

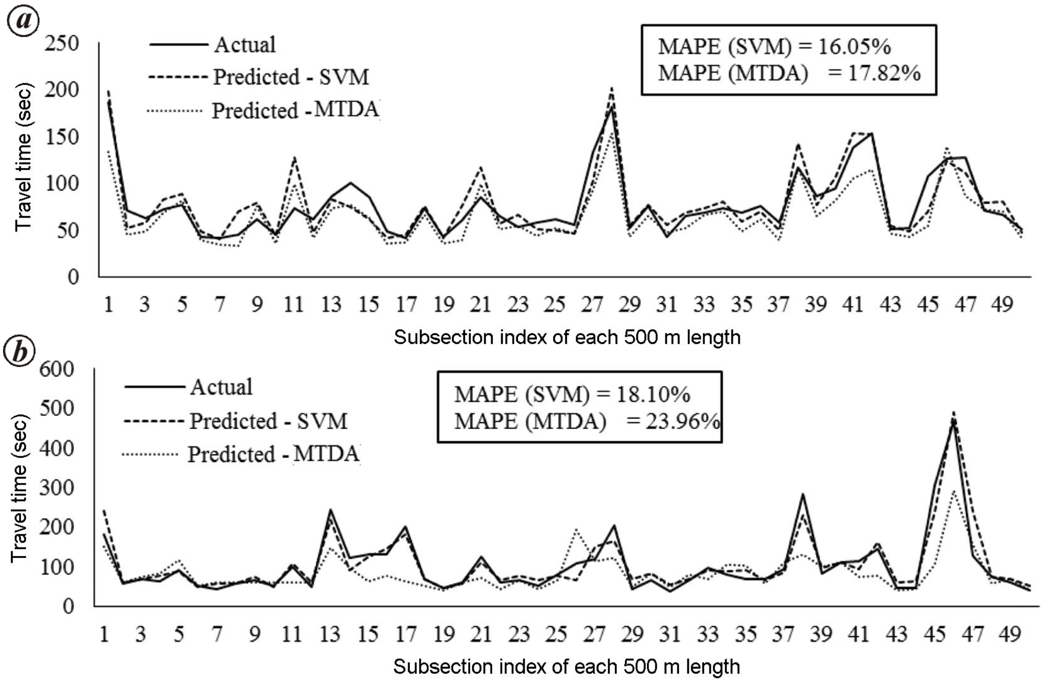


Figure 10. Predicted and observed travel times for (***a***) an off-peak trip and (***b***) a peak trip.

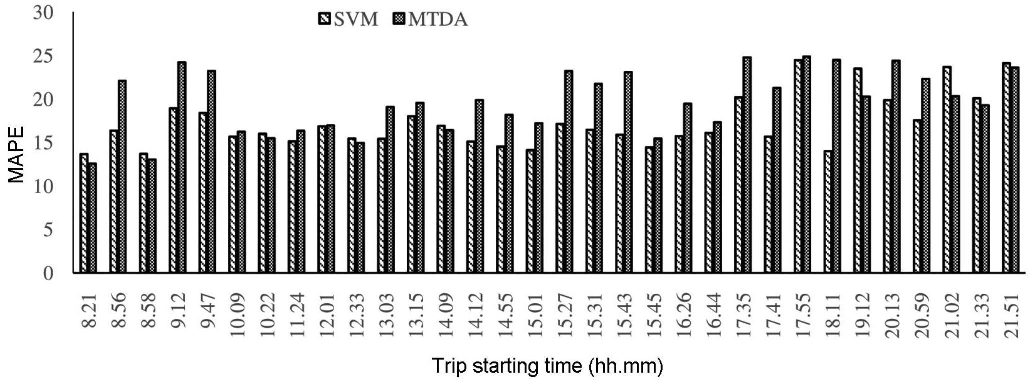


Figure 11. Comparison of performance between support vector machines (SVM) and Kalman filtering technique (KFT) for various trips of a sample day.

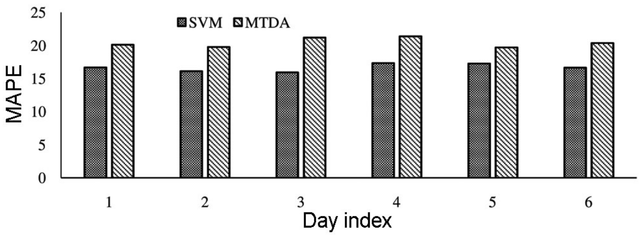


Figure 12. MAPE comparison between SVM and KFT for various days.

sample comparisons of the predicted and measured travel times over 500 m subsections for sample off-peak and peak trips respectively. From the figure, it can be observed that both SVM and MTDA perform comparably for off-peak trips, whereas for the peak trips SVM performs better and is able to capture the travel time variations more accurately. It is well known that the variability in travel time for trips during peak hours is more than those during off-peak hours. This could be a reason for the difference in performance during peak hours. Figure 11 shows the errors for all trips on a sample day, reinforcing the above result. Average performance comparison across days was also carried out. Figure 12 shows the results in terms of MAPE for the selected six days. It can be observed that SVM performs better than MTDA on an average scale as well.

Comparison of performance across bus stop sections

The proposed method was also evaluated by analysing the deviation of the actual travel time from the predicted

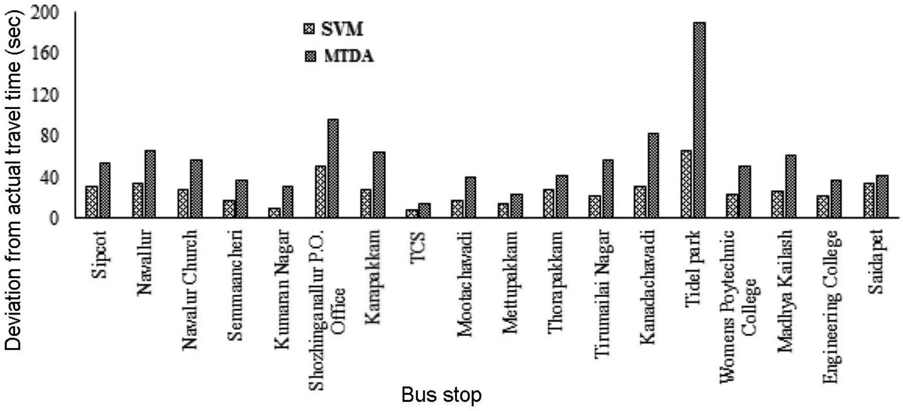


Figure 13. Deviation from actual travel time at different bus stops.

travel time (expressed as deviation from actual clock time), which is what users feel. Here, it is important to know the acceptable deviation from the user perspective. Earlier studies reported that 5 min prediction accuracy is acceptable for a bus with 1.5 h journey time[5]. In another study[27], it was reported that the passengers will have ± 5 min tolerance if 88% of the predicted times are within 5 min of the actual travel time. The TriMet Transit Tracker System in Portland reported that passengers have a tolerance of up to a minimum of 2 min and a maximum of 4.5 min of waiting time at bus stops[28]. Based on these, an accuracy of ± 5 min may be considered as the acceptable error limit from the user perspective. Figure 13 shows the deviation from actual travel time at different bus stops observed in this study. From this figure, it can be observed that the deviation is up to ± 90 sec and ± 195 sec at various bus stops along the 19B bus route for SVM and MTDA respectively. It can also be seen that the SVM method is able to predict the arrival time more efficiently than MTDA, and is within the tolerance limit.

Field implementation

The present study focuses on the development of a prediction method that can form the basis of a traveller information system, which will be useful to passengers. The predicted travel/arrival times can be shared with the passengers on real-time basis, and users can get the information through display boards placed at bus stops, inside the bus itself or through web portals or mobile applications.

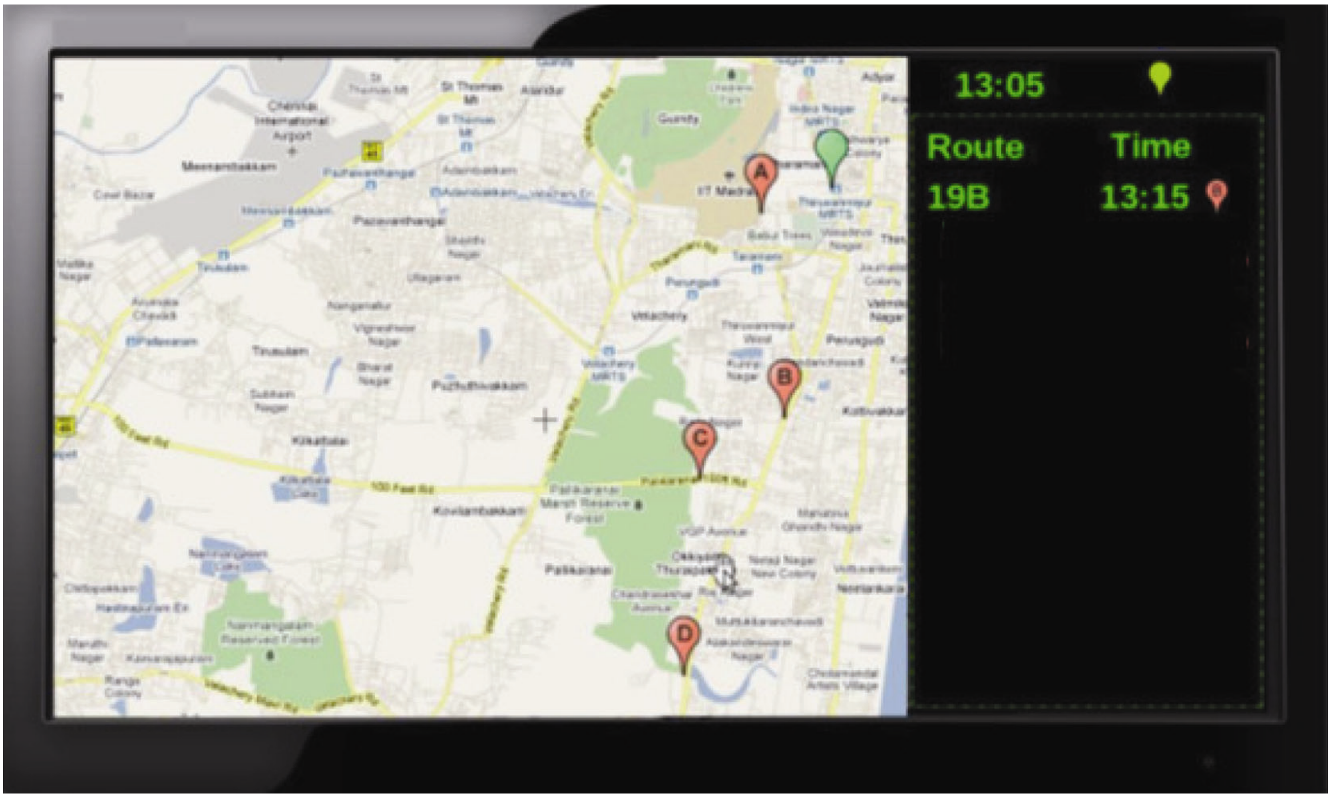
A commuter can potentially check the arrival details from the comfort of his/her office or home through the website, and can reach the bus stop close to the predicted arrival time. A Google Maps-based website has been developed (Figure 14), which can provide such information. The same can be provided through mobile applications or display boards for access them from bus stops or en-route. Such applications can be interactive and provide the users with the current location of the bus and its expected arrival time at any chosen bus stop.

Conclusion and scope for future work

The heterogeneity and lack of lane discipline makes Indian traffic highly varying and hence prediction methods that have hitherto been developed for homogeneous and lane-disciplined traffic conditions may not be applicable here. The present study is an attempt at developing a real-time bus-arrival prediction system that pays special attention to this high variance. The high variability was first addressed by explicitly incorporating mean and variance into an existing time discretization-based model. However, improvement in performance was not significant. Hence, a prediction methodology using SVR with linear kernel function in LIBSVM was developed and was implemented in MATLAB. ApEn technique was used to arrive at the optimum amount of data required to predict the next trip, and it was observed that data from previous six trips optimally predict the following trip.

Analysis was carried out using GPS data collected from buses running along route 19B in Chennai. Results showed comparable performance by SVM and the proposed MTDA for trips during off-peak hours. However, for trips during peak hours, SVM was able to capture the travel time variations better than MTDA. Overall, the proposed method showed better performance than MTDA using KFT.

The main challenge in using this approach is the requirement of a sufficiently large dataset and the need to separately tune the parameters for each section. The proposed method can be implemented in real-time for advanced public transportation systems applications on a large scale. The predicted travel times can be communicated to travellers through variable message sign boards or kiosks at bus stops, as well as through websites or mobile applications for pre-trip and en-route planning.



Route	Time
19B	13:15

Figure 14. Bus stop information dissemination unit with map display.

The failure of the model-based approach with explicit error incorporation in improving prediction performance may be attributed to the fact that it used a simple linear dynamic model, which may not be capable of capturing the characteristics of the system well under such scenarios. The prediction performance may improve if an advanced nonlinear model is used, which can characterize the system more accurately, instead of the linear model used in this study. The performance of SVM method may be further improved by explicitly considering traffic-related variables such as driver characteristics (age, vision), vehicular characteristics (kilometres travelled, engine characteristics), weather information, etc.

1. Zhu, H., Zhu, Y., Li, M., Ni, L. M. and Ni, M., SEER: Metropolitan-scale traffic perception based on lossy sensory data. In INFOCOM, IEEE, Rio de janeiro, 2009; doi:10.1109/INFOCOM.2009.5061924.
2. Chen, G., Yang, X., Zhang, X. and Teng, J., Historical travel time based bus-arrival-time prediction model. In Proceedings of the 11th International Conference of Chinese Transportation Professionals (ICCTP), Nanjing, China, 2011.
3. Li, R. and Rose, G., Incorporating uncertainty into short-term travel time predictions. *Transp. Res. Part C*, 2011, **19**, 1006–1018.
4. Tiesyte, D. and Jensen, C. S., Assessing the predictability of scheduled-vehicle travel times. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle United States, 2009.
5. Bhandari, R. R., Bus arrival time prediction using Stochastic time series and Markov Chains. Ph D dissertation, Department of Civil Engineering, New Jersey Institute of Technology, USA, 2005.
6. Jeong, R. and Rilett, L. R., Bus arrival time prediction using artificial neural network model. In IEEE Intelligent Transportation Systems Conference, Washington, DC, 2004, pp. 988–993.
7. Chien, S. J., Ding, Y. and Wei, C., Dynamic bus arrival time prediction with artificial neural networks. *J. Transp. Eng.*, 2002, **128**(5), 429–438.
8. Bin, Y., Zhongzhen, Y. and Baozhen, Y., Bus arrival time prediction using support vector machines. *J. Intell. Transp. Syst.*, 2006, **10**(4), 151–158.
9. Wu, C. H., Su, D. C., Chang, J., Wei, C. C., Ho, J. M., Lin, K. J. and Lee, D., An advanced traveler information system with emerging network technologies. In Proceedings of the 6th Asia-Pacific Conference Intelligent Transportation Systems Forum, Taipei, Chinese–Taipei, 2004, pp. 230–231.
10. Vanajakshi, L. and Rilett, L. R., Support vector machine techniques for the short term prediction of travel time. In IEEE Intelligent Vehicles Symposium, Istanbul, 2007, pp. 600–605; doi:10.1109/IVS.2007.4290181.
11. Vanajakshi, L. and Rilett, L. R., A comparison of the performance of artificial neural networks and support vector machines for the prediction of speed. IEEE Intelligent Vehicle Symposium, 2004, pp. 194–199; doi:10.1109/IVS.2004.1336380.
12. Dailey, D. J., Maclean, S. D., Cathey, F. W. and Wall, Z. R., Transit vehicle arrival prediction: algorithms and large-scale implementation. *Transp. Res. Rec.: J. Transp. Res. Board*, 2001, **1771**, 46–51.
13. Cathey, F. and Dailey, D., A prescription for transit arrival/departure prediction using automatic vehicle location data. *Transp. Res. Part-C*, 2003, **11**(3), 241–264.

14. Shalaby, A. and Farhan, A., Bus travel time prediction for dynamic operations control and passenger information systems. In 83rd Annual Meeting of the Transportation Research Board, National Research Council, Washington DC, USA, 2004.
15. Nanthawichit, C., Nakatsuji, T. and Suzuki, H., Application of probe vehicle data for real-time traffic state estimation and short-term travel time prediction on a freeway. In Proceedings of the 82nd Annual Transportation Research Board Meeting, Washington DC, USA, 2003.
16. Vanajakshi, L., Subramanian, S. C. and Sivanandan, R., Travel time prediction under heterogeneous traffic conditions using GPS data from buses. *IET J. Intell. Transp. Syst.*, 2009, **3**(1), 1–9.
17. Padmanabhan, R. P. S., Divakar, K., Vanajakshi, L. and Subramanian, S. C., Development of a real-time bus arrival prediction system for Indian traffic conditions. *IET J. Intell. Transp. Syst.*, 2009, **4**(3), 189–200.
18. Kumar, S. V. and Vanajakshi, L., Pattern identification based bus arrival time prediction. *Proc. Inst. Civil Eng. – Transport*, 2014, **167**(3), 194–203.
19. Vivek, K., Kumar, B. A., Vanajakshi, L. and Subramanian, S. C., Comparison of model based and machine learning approaches for bus arrival time prediction. In 93rd Annual Meeting of the Transportation Research Board, Washington DC, USA, 2004.
20. Kumar, B. A., Vanajakshi, L. and Subramanian, S. C., Pattern based bus travel time prediction under heterogeneous traffic conditions. In 93rd Annual Meeting of the Transportation Research Board, Washington DC, USA, 2004.
21. Chamberlain, R. G., Great circle distance between two points; http://www.movabletype.co.uk/scripts/gis-faq-5.1.html (accessed on 14 March 2013).
22. Kalman, R. E., A new approach to linear filtering and prediction problems. *Trans. ASME – J. Basic Eng.*, 1960, **82**(1), 35–45.
23. Vapnik, V. N., An overview of statistical learning theory. *IEEE Trans. Neural Networks*, 1999, **10**(5), 988–999; doi:10.1109/72.788640.
24. Scholkopf, B., Smola, A. J., Williamson, R. C. and Bartlett, P. L., New support vector algorithms. *Neural Comput.*, 2006, **12**(5), 1207–1245.
25. Chang, Chung, C. and Lin, C. J., LIBSVM: a library for support vector machines. *ACM Trans. Int. Syst. Technol.*, 2011, **2**(3), 27.
26. Pincus, S. M., Approximate entropy as a measure of system complexity. *Cross Mark*, 1991, **88**(6), 2297–2301.
27. Warman, P., Measure impacts of real-time control and information systems for bus services. Transport Direct, Department of Transport, UK, 2003.
28. Crout, D. T., Accuracy and precision of TriMet's transit tracker system. In Proceedings of the 86th Annual Meeting, Transportation Research Board of the National Academics, Washington, DC, USA, 2007.

ACKNOWLEDGEMENTS. We acknowledge support for this study as a part of the sub-project CIE/10-11/168/IITM/LELI under the Centre of Excellence in Urban Transport project funded by the Ministry of Urban Development, Government of India, through letter No. N-11025/30/2008-UCD.

Received 20 April 2015; revised accepted 2 March 2016

doi: 10.18520/cs/v111/i4/700-711