Analytical Approach to Identify the Optimum Inputs for a Bus Travel Time Prediction Method

B. Anil Kumar, Snigdha Mothukuri, Lelitha Vanajakshi, and Shankar C. Subramanian

Even though new infrastructure is being developed to meet demand, increased urbanization and vehicle ownership have increased the congestion levels in Indian cities. Attracting more travelers to public transport is an option to reduce congestion but still remains a challenge, mainly because of the uncertainty of service. A reliable and accurate system for predicting vehicle arrival can help make public transportation more attractive. An accurate prediction method should be used to provide reliable information to passengers, and accuracy depends on the input data used. Therefore, identifying the optimum inputs and incorporating them in the prediction method become important. The optimum number of inputs required for best prediction performance was identified with an analytical approach. A model-based algorithm motivated by the Kalman filter was used to predict bus travel time with the use of GPS data. A case study was conducted on two selected bus routes in the city of Chennai, India, to evaluate the prediction accuracy of the proposed method. Results obtained from the algorithm were promising and showed the prediction accuracy to be ±5 min for a prediction window of 30 min in 92% of instances. The predicted travel time can be used to provide real-time bus arrival information to the public through various media, including web pages, mobile applications, and display boards.

Rapidly increasing vehicle congestion has been deteriorating the quality of life of people in urban areas of many developed and developing countries, including India. Caused mainly by rapid changes in urbanization, economy levels, vehicle ownership, and population growth, congestion leads to problems such as increased travel time, air pollution, and fuel use as well as decreased accessibility and mobility. To relieve congestion, demand-side and supply-side approaches can be adopted (*1, 2*). Supply-side approaches concentrate on infrastructure expansion (e.g., constructing more roads or adding more lanes to existing roads). Demand-side approaches concentrate on how to use the existing system and facilities more efficiently (e.g., congestion pricing).

Because infrastructure expansion cannot meet the growth in vehicle ownership, demand-side solutions such as better traffic operations and management must be explored. Intelligent transportation system technologies are gaining popularity for managing urban traffic, and one functional area is Advanced Public Transportation Systems (APTS) applications, which primarily use intelligent transportation system technologies to attract more travelers to public transportation. One attractive, popular application is predicting accurate bus travel times and providing that information (i.e., bus arrival times) to passengers. This application benefits passengers by reducing their waiting time at bus stops (and associated uncertainties) and allowing them to make reasonable travel arrangements when planning a trip. However, the information provided to passengers should be reliable to be effective. The present study concentrates on this area of predicting bus travel times for the development of accurate passenger information systems.

Commonly used prediction techniques can be classified primarily as model-based, instantaneous, and data-driven techniques. Model-based techniques require fewer data than data-driven and instantaneous approaches. However, regardless of the amount of data required, the significant input should be used optimally to improve prediction accuracy. Schweiger reports that the accuracy of prediction methods greatly depends on the input data (*3*). Thus, identifying the optimum input data and incorporating them in the prediction method hopefully would improve prediction accuracy. The present study used an analytical approach to identify the optimum number of inputs required for best performance and used the same in a mathematical model that was developed to predict bus travel time.

B. A. Kumar, S. Mothukuri, and L. Vanajakshi, Department of Civil Engineering, and S. C. Subramanian, Department of Engineering Design, Indian Institute of Technology, Madras, Chennai 600 036, India. Corresponding author: L. Vanajakshi, lelitha@iitm.ac.in.

Transportation Research Record: Journal of the Transportation Research Board, No. 2535, Transportation Research Board, Washington, D.C., 2015, pp. 25–34.
DOI: 10.3141/2535-03

LITERATURE REVIEW

The most widely used models for predicting bus travel time can be broadly classified in four categories: historic average, statistical, machine learning, and model based.

Historic Average Models

Historic average models use past travel times to predict future travel times (*4–6*). The main assumptions are that traffic patterns are cyclical and that the ratio between traffic parameters in a particular link remains constant over time. These models perform best under expected traffic conditions; performance suffers if the traffic pattern in the area of interest is not relatively stable.

Statistical Models

Statistical techniques, which include time series and regression methods, are popular for predicting bus travel time. Time series–based prediction methods assume that the future traffic patterns will be similar to observed historical traffic patterns. Their accuracy greatly depends on the correspondence between real-time and historical traffic patterns (*6*). Variations in the relationship between real-time and historical travel time data can significantly reduce prediction accuracy. These techniques require a large amount of reliable data.

Bhandari develops stochastic time series–based autoregressive and delay-propagation methods to predict bus arrival time by using 7 months of automatic vehicle location (AVL) data (*7*). Suwardo and Kamaruddin develop an autoregressive integrated moving average method to predict bus travel time from 1 year of AVL data collected from an expressway (*8*); this technique requires a large amount of reliable data for model development. In contrast, regression models forecast parameters with mathematical functions between the variable of interest and the parameters that affect it. Abdelfattah and Khan develop linear and nonlinear regression methods by using simulated data to predict bus travel time, delays, and the influence on prediction of various factors that affect delay (*9*). Patnaik et al., who develop a regression model to predict bus travel time by considering the number of stops, dwell times, the number of boarding and alighting passengers, and weather descriptors as independent variables, discover that weather is not an important input parameter for travel time prediction (*10*). Jeong and Rilett develop five regression methods to predict bus travel time by considering schedule adherence, distance, and dwell time as independent variables (*4*). Chang et al. use nearest neighbors nonparametric regression to predict bus travel time between bus stops (*11*). Ramakrishna et al. use 25 trips of GPS data to develop a multiple linear regression method for bus travel time prediction under heterogeneous traffic conditions; their results indicate that the other models outperform regression models (*5*).

Unlike historic average models, statistical models work satisfactorily under unexpected traffic conditions. Prediction accuracy depends on identifying and applying suitable variables. Thus, the applicability of regression models is limited because variables in the transportation field are highly intercorrelated.

Machine Learning Models

Machine learning approaches can address complex and nonlinear relationships between predictors. Two of the popular machine learning approaches for travel time prediction are artificial neural networks (ANNs) and support vector machines (SVMs). An ANN is a massively distributed processor made up of simple processors and has a natural tendency to store experimental knowledge to make it available for later use.

Chien et al. develop an enhanced ANN method to predict bus arrival time dynamically with two ANN models (one trained with link-based data and the other trained with stop-based data), implement the method, and validate results with simulated data (*6*). Chen et al. develop an ANN method to predict travel time between two points with automatic passenger counter data (*12*). In general, ANN is useful for prediction when mathematically formulating the relationship between inputs and outputs is difficult. However, it requires a large database and can be used to predict traffic parameters without explicitly addressing the traffic process.

SVMs are supervised learning models with associated learning algorithms that can analyze data and recognize patterns in data that are used for classification and regression. Even though other machine learning approaches such as ANN have been extensively studied, few SVM applications are reported in the transportation field. Bin et al. propose an SVM method as a new neural network algorithm to predict bus travel time; they predict the arrival time from the travel time of a current segment and the latest travel time of the next segment (*13*). Chun-Hsin et al. compare the performances of support vector regression and other baseline predictors; results show that the support vector regression predictor can significantly reduce the relative mean and squared errors, but SVMs require a large amount of computation time (*14*).

Model-Based Models

Model-based approaches develop models that can capture system dynamics by establishing mathematical relationships between appropriate variables. One major advantage of the model-based approach is that, with limited data, the model can deliver the overall traffic state in both time and space domains of a system. Many model-based studies use estimation techniques such as Kalman filtering to estimate traffic parameters (e.g., density and travel time). The Kalman filtering technique (KFT) has an elegant mathematical representation and can be used effectively to accommodate traffic fluctuations with their time-dependent variables (*6*). It has been used extensively to estimate bus arrival times and is the basis for predicting future values or for improving the estimates of variables from earlier times (*11, 14–16*).

Wall and Dailey develop an algorithm to predict bus travel time by developing an algorithm that contains two components (tracking with KFT and prediction with statistical analysis); they use a combination of data obtained from AVL and a historical database (*17*). Cathey and Dailey use bus data collected on different days at the same time of day as inputs to predict bus travel time with KFT; their algorithm had three components (tracker, filter, and predictor), and they compared results from historic, regression, and ANN models (*15*). Shalaby and Farhan develop a method to predict bus travel time by using only 5 weekdays' worth of automatic passenger counter and AVL data on a specific bus route in downtown Toronto, Canada (*16*). They use the data obtained from VISSIM to validate results obtained from the prediction algorithm; the study compared the results with historic, regression, and ANN methods. Focusing mainly on comparing historical path–based and link-based travel time predictions, Chien and Kuchipudi develop a method based on Kalman filtering to predict bus travel time by using historical and real-time travel time data (*18*). Results show that for peak-hour travel time prediction, the historical path–based data performed better.

Chu et al. integrate the data obtained from loop detectors and probe vehicles to develop a method to predict bus travel time by applying the adaptive KFT (*19*). The algorithm was tested in a stretch of freeway by using the PARAMICS microscopic model. Nanthawichit et al. also integrate the data obtained from loop detectors and vehicles outfitted with GPS probes to obtain traffic parameters with KFT (*20*). This method performed better than historic, regression, and ANN methods. Son et al. develop a method that uses KFT to predict bus travel time from one bus stop to the stop line at the next signalized intersection (*21*). Yu et al. develop several methods (SVM, ANN, *k*–nearest neighbor algorithm, and linear regression) to predict bus travel time on the basis of the running time of multiple routes (*22*). Zhu et al. develop a method to predict bus travel time from bus stop delays and signalized intersection delays associated with total travel times; they consider traffic flow, signal delay, and dwell times as variables (*23*).

Most of the previous studies were carried out under homogeneous, lane-disciplined traffic conditions. In contrast, Indian traffic conditions are very different. Because traffic is heterogeneous in nature (a lack of lane discipline complicates the issue) and the growing number of vehicles leads to hectic traffic conditions, especially during peak hours, models for homogeneous traffic conditions may not work accurately under Indian traffic conditions. Therefore, models that can capture and study stochastic behavior must be developed. The present study is one step in this direction. In the first part of this study, the optimum number of inputs for the accurate prediction of travel time was determined. In the second part of the study, real-time travel time was predicted with the identified inputs.

DATA COLLECTION AND ANALYSIS

GPS devices commonly are used to collect data for real-time APTS applications. For the present study, Metropolitan Transport Commission buses in the city of Chennai, India, were fitted with GPS units to transmit real-time data. Two bus routes were selected. Route 19B connects Saidapet, a major area in the southern part of the city, with suburban Kelambakkam; it is about 30 km long, has 20 bus stops and 13 intersections, and has an average headway of 15 min. Route 5C connects two major areas—Tharamani bus depot in the southern part of the city and the Parry's Corner depot in the northern part—with nine major bus stops and 14 intersections; it is around 15 km long and has an average headway of 30 min. The two routes differ in geometric characteristics, volume levels, and land use characteristics (Figure 1). The road stretches are highly heterogeneous, with a mix of motorized vehicles (e.g., passenger cars, buses, trucks, auto rickshaws, and two-wheelers) and nonmotorized vehicles (e.g., cycles and animal-drawn carts).

GPS data, collected every 5 s from 6 a.m. to 8 p.m. over 15 days, include the GPS unit identification, a time stamp, and the latitude and longitude of the location at which the entry was made. Data were transmitted in real time via general packet radio service. The collected data were stored in a sequential query language database as separate files for each day. From GPS data, the distance between two consecutive entries was calculated with the Haversine formula (*25*), which gives the great circle distance (d) between two points on a sphere from their latitudes and longitudes:

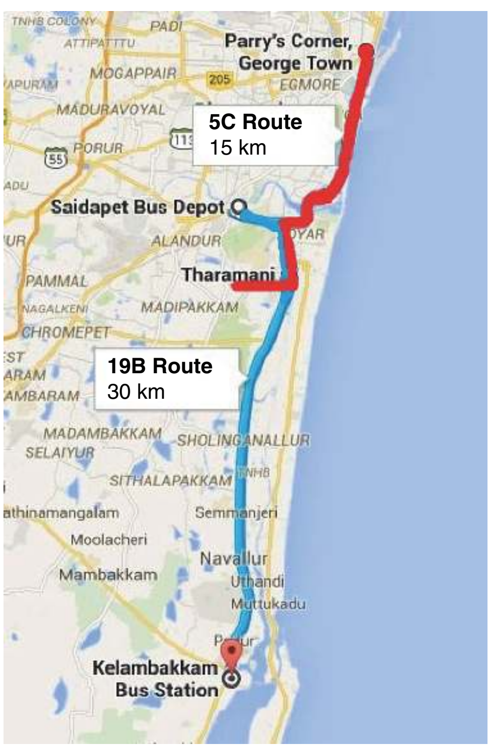

FIGURE 1 19B and 5C bus routes (*24*).

$$d = 2r \arcsin\left(\sqrt{\text{haversin}(\varphi_2 - \varphi_1) + \cos\varphi_1 \cos\varphi_2 \text{haversin}(\lambda_2 - \lambda_1)}\right) \tag{1}$$

where

r = radius of Earth (6,378.1 km);
φ_1, φ_2 = latitude of Points 1 and 2, respectively; and
λ_1, λ_2 = longitude of Points 1 and 2, respectively.

Thus, the processed data consist of the travel times and the corresponding distances between consecutive locations of all buses. The entire section was divided into smaller 100-m subsections, and the time taken to cover each subsection was calculated with the linear interpolation technique.

METHOD

The prediction model for bus arrival time adopted in this study is based on that of Vanajakshi et al. (*26*), in which a spatially discretized evolution model was proposed to predict bus travel time with Kalman filtering (*27*). The entire route was discretized into smaller subsections, and the test vehicle travel time in the current subsection was obtained by using the travel time data of the two previous vehicles (denoted as PV1 and PV2). However, the basic assumption made in that model was that the travel time in a particular subsection is affected by the travel time only in the previous one subsection, which need not be true.

The present study theoretically analyzed the relaxation of this assumption by using many previous subsections ($k-1$, $k-2$, $k-3$, and so forth) and identified the optimum number of previous subsections that must be considered. In this case, the system equations consider the travel time from a series of subsections before the subsection under consideration (*27*):

$$x_{k+1} = \mathbf{a} \cdot \mathbf{x} + w \tag{2}$$

$$z_k = x_k + v \tag{3}$$

where

x_{k+1} = travel time taken to cover $(k+1)$th subsection,
$\mathbf{a}$ = vector that relates the travel time in $(k+1)$th subsection to travel time in previous n subsections [kth to $k-(n-1)$th],
z_k = measured travel time in kth subsection,
$\mathbf{x}$ = vector that represents the travel time in corresponding subsections, and
w, v = process disturbance and process noise, respectively, assumed to be characterized by normal distributions with zero mean their corresponding variances Q and R.

Also, for an nth-order model, $\mathbf{a} = [a_k, a_{k-1}, \ldots, a_{k-(n-1)}]^T$ and $\mathbf{x} = [x_k, x_{k-1}, \ldots, x_{k-(n-1)}]^T$.

The Kalman filtering equations may be time-update or measurement-update. Time-update equations use the model and system inputs to predict the estimate a priori; measurement-update equations use the output measurements to obtain the estimate a posteriori (*27*). Thus, two sets of data are required to implement the scheme presented earlier: one for the time-update equations and another for the measurement-update equations. The data obtained from the previous two buses (PV1 and PV2) were used in this study to calculate **a** for each subsection and estimate the travel time of the test vehicle a posteriori. Thus, to predict the bus travel time in the $(k + 1)$th subsection with the state equation, one must know the coefficient vector **a**, which varies dynamically from subsection to subsection. This estimation is discussed next.

To estimate coefficients of the state space equation (i.e., vector **a**), consider the state equation for an nth-order model (Equation 2) to predict the state variable in the $(k+1)$th subsection. In this, the coefficient corresponding to the $(k-m)$th subsection can be assumed to be

$$a_{k-m} = \frac{1}{n}\frac{x_{k+1}^{\text{PV1}}}{x_{k-m}^{\text{PV1}}} \qquad 0 \le m \le (n-1) \tag{4}$$

where

a_{k-m} = parameter that relates travel time in the $(k-m)$th subsection to travel time in the $(k+1)$th subsection,
n = order of the model,
x_{k+1}^{PV1} and x_{k-m}^{PV1} = subsection travel times obtained from PV1, and
m = number of previous subsections considered.

The concept of this relationship is explained later with a second-order linear model (i.e., the travel time in a particular subsection was assumed to depend on the travel times in the previous two subsections). Therefore, if $n = 2$ in Equation 2, then

$$x_{k+1} = \begin{bmatrix} a_k \\ a_{k-1} \end{bmatrix} \cdot \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} \tag{5}$$

$$x_{k+1} = a_k x_k + a_{k-1}x_{k-1}$$

This expression can be interpreted as a line in three-dimensional space passing through the origin and an arbitrary point (x_{k+1}, x_k, x_{k-1}). A line passing through a known point in space (x_0, y_0, z_0) and parallel to the vector with components (a, b, c) can be represented as the following:

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c} \tag{6}$$

If the components (a, b, c) correspond to a unit vector parallel to the line joining (0, 0, 0) and (x_0, y_0, z_0), then

$$a = \frac{x_0}{\sqrt{x_0^2 + y_0^2 + z_0^2}}$$

$$b = \frac{y_0}{\sqrt{x_0^2 + y_0^2 + z_0^2}}$$

$$c = \frac{z_0}{\sqrt{x_0^2 + y_0^2 + z_0^2}}$$

Simplifying the earlier equation gives

$$x = \left[\frac{a}{b} * (y - y_0) + x_0\right] = \frac{x_0}{y_0} * y \tag{7}$$

$$x = \left[\frac{a}{c} * (z - z_0) + x_0\right] = \frac{x_0}{z_0} * z \tag{8}$$

Because the point (x_0, y_0, z_0) is from the most recently available historical data here, the known point (x_0, y_0, z_0) can be written as $(x_{k+1}^{\text{PV1}}, x_k^{\text{PV1}}, x_{k-1}^{\text{PV1}})$. Therefore, the equations become

$$x_{k+1} = \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} * x_k \tag{9}$$

$$x_{k+1} = \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}} * x_{k-1} \tag{10}$$

Adding Equations 9 and 10 gives the following:

$$x_{k+1} = \frac{1}{2}\left[\frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} * x_k + \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}} * x_{k-1}\right] \tag{11}$$

which can be further simplified as

$$x_{k+1} = \left(\frac{1}{2} * \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}}\right) * x_k + \left(\frac{1}{2} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}}\right) * x_{k-1} \tag{12}$$

A comparison of Equations 5 and 12 yields

$$a_k = \frac{1}{2} * \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} \quad \text{and} \quad a_{k-1} = \frac{1}{2} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}} \tag{13}$$

when the order of the equation is 2 $(n = 2)$. Clearly, this relationship also is true for a first-order model. That is,

$$x_{k+1} = \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} * x_k$$

therefore

$$x_{k+1} = \left[\frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}}\right] \cdot [x_k] \qquad \text{for } n = 1 \tag{14}$$

If this relationship is assumed to be true for some order j (i.e., for $n = j$), then

$$x_{k+1} = \frac{x_{k+1}^{\text{PV1}}}{x_{k-m}^{\text{PV1}}} * x_{k-m} \qquad \text{for } 0 \le m \le j - 1 \tag{15}$$

Therefore,

$$x_{k+1} = \begin{bmatrix} \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} \\ \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}} \\ \cdots \\ \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-(j-1)}^{\text{PV1}}} \end{bmatrix} \cdot \begin{bmatrix} x_k \\ x_{k-1} \\ \cdots \\ x_{k-(j-1)} \end{bmatrix} \qquad \text{for } n = j \tag{16}$$

When the same jth-order model is applied to estimate x_k [i.e., k is replaced with $(k-1)$] in Equation 16,

$$x_k = \begin{bmatrix} \frac{1}{j} * \frac{x_k^{\text{PV1}}}{x_{k-1}^{\text{PV1}}} \\ \frac{1}{j} * \frac{x_k^{\text{PV1}}}{x_{k-2}^{\text{PV1}}} \\ \cdots \\ \frac{1}{j} * \frac{x_k^{\text{PV1}}}{x_{k-j}^{\text{PV1}}} \end{bmatrix} \cdot \begin{bmatrix} x_{k-1} \\ x_{k-2} \\ \cdots \\ x_{k-j} \end{bmatrix} \tag{17}$$

This expression can be modified as

$$\frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} * x_k = \begin{bmatrix} \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}} \\ \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-2}^{\text{PV1}}} \\ \cdots \\ \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-j}^{\text{PV1}}} \end{bmatrix} \cdot \begin{bmatrix} x_{k-1} \\ x_k \\ \cdots \\ x_{k-j} \end{bmatrix} \tag{18}$$

and, further modified with Equation 16 as

$$\frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} * x_k = x_{k+1} - \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}} * x_k + \frac{1}{j} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-j}^{\text{PV1}}} * x_{k-j}$$

Further simplification gives

$$x_{k+1} = \frac{x_{k+1}^{\text{PV1}}}{x_{k-j}^{\text{PV1}}} * x_{k-j}$$

Therefore, the relationship also is true for order $j+1$ $(n = j+1)$:

$$x_{k+1} = \frac{x_{k+1}^{\text{PV1}}}{x_{k-m}^{\text{PV1}}} * x_{k-m} \qquad \text{for } 0 \le m \le j \tag{19}$$

Therefore, the state equation for the $(j+1)$th-order model becomes

$$x_{k+1} = \begin{bmatrix} \left(\frac{1}{j+1} * \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}}\right) \\ \left(\frac{1}{j+1} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}}\right) \\ \cdots \\ \left(\frac{1}{j+1} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-j}^{\text{PV1}}}\right) \end{bmatrix} \cdot \begin{bmatrix} x_k \\ x_{k-1} \\ \cdots \\ x_{k-j} \end{bmatrix} \qquad \text{for } n = j+1 \tag{20}$$

Thus, by mathematical induction, this relationship also is true for any order n:

$$x_{k+1} = \begin{bmatrix} \left(\frac{1}{n} * \frac{x_{k+1}^{\text{PV1}}}{x_k^{\text{PV1}}}\right) \\ \left(\frac{1}{n} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-1}^{\text{PV1}}}\right) \\ \cdots \\ \left(\frac{1}{n} * \frac{x_{k+1}^{\text{PV1}}}{x_{k-(n-1)}^{\text{PV1}}}\right) \end{bmatrix} \cdot \begin{bmatrix} x_k \\ x_{k-1} \\ \cdots \\ x_{k-(n-1)} \end{bmatrix} \tag{21}$$

After the coefficients are calculated, the a priori values and then the a posteriori values can be estimated. Consider a study route with a total of N subsections. To identify the optimum number of subsections required as input to predict the travel time for the next subsection, a thorough analysis was carried out by establishing a relationship between travel time of the subsection of interest and the previous many subsections ($G = 1, 2, 3, \ldots, (N-1)$), as shown in Equation 2 with a Gth-order model. The proposed algorithm was implemented by building a relationship between the travel times of the next subsections and the previous many subsections by increasing the number of previous number of subsections (G) one at a time until the prediction accuracy of that particular subsection and trip reaches an optimum value. The steps of the algorithm are listed in Figure 2.

RESULTS

Results obtained by implementing the earlier algorithm were compared with measured travel time data. The entire section was divided into N subsections of equal length (100 m). However, the data indicate a minimum distance of at least 500 m between consecutive bus stops. Therefore, the final comparison was made for travel times with 500-m subsections by adding the predicted travel times from the corresponding values for 100-m travel times.

The prediction was carried out over a 2-week period. Prediction accuracy was quantified as mean absolute percentage error (MAPE) and mean absolute error (MAE), mainly because the final impact was to be evaluated at selected bus stops (which are at varying distances). Thus, measures such as MAPE and MAE that can normalize the effect of such variations were preferred to evaluate the performance of the prediction method for uniform subsections (100 m long). MAPE and MAE can be calculated as

$$\text{MAPE} = \frac{\sum_{i=1}^{N} \frac{X_s - X_{\text{TVM}}}{X_{\text{TVM}}}}{N} \tag{22}$$

$$\text{MAE} = \frac{\sum_{i=1}^{N} X_s - X_{\text{TVM}}}{N} \tag{23}$$

where

X_s = predicted travel time obtained from the prediction algorithm to cover a given subsection,
X_{TVM} = corresponding travel time measured from the field, and
N = number of subsections under consideration.

Figure 3 shows the variation in MAPE by considering various numbers of previous subsections ($G = 1, 2, 3, \ldots, (N-1)$) to predict

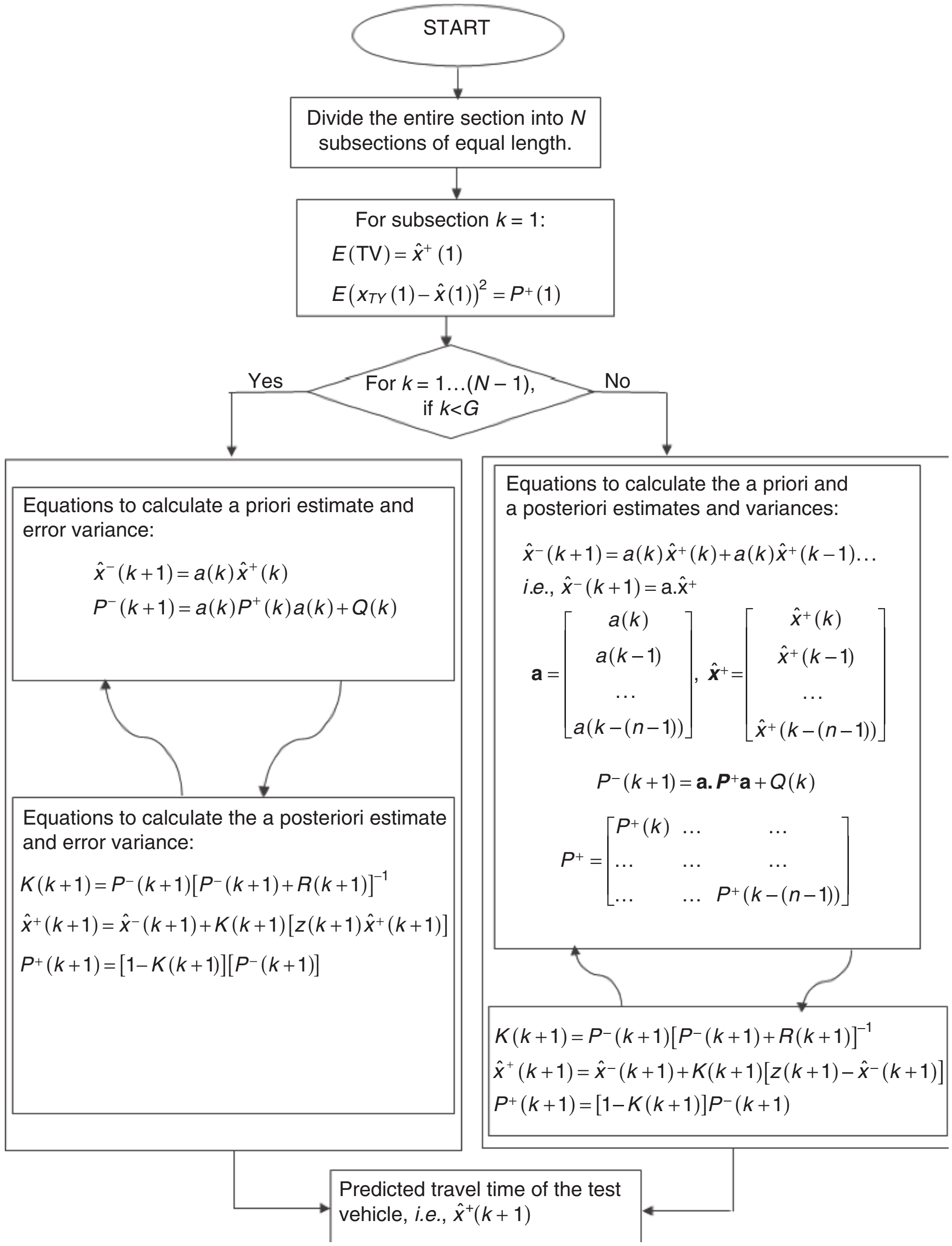

FIGURE 2 Flowchart representing the proposed method (E = expected travel time; TV = test vehicle; $\hat{x}$ = travel time; K = Kalman gain; P = error variance; z = latest available measurement from the field; − and + in superscript = a priori and a posteriori estimates, respectively).

the next subsection travel time for a sample test period of 1 week for Routes 19B and 5C. Results indicate that MAPE decreases with increases in the number of previous subsections. However, the error reduction was marginal after two subsections, and the error increased in a few cases, too. Therefore, the optimum number of previous subsections was determined to be two, and this number ($G = 2$) was used for further analysis. Figure 4 compares a sample of predicted and measured travel times over 500-m subsections with the corresponding MAPE values for sample trips on the two study routes. Predicted values closely match the measured data.

Figure 5 shows the error reduction for peak and off-peak trips while considering two subsections instead of one. Results show reduced error in most subsections because of the use of data from two previous subsections instead of one. Errors are reduced up to 8% for off-peak trips (Figure 5*a*) and up to 5% for peak trips (Figure 5*b*) when data from the two previous subsections are used as inputs instead of from only one previous subsection.

Because the main interest of the present study was to predict bus travel time and provide the arrival information of the same bus to the next bus stops, a performance was evaluated by comparing the

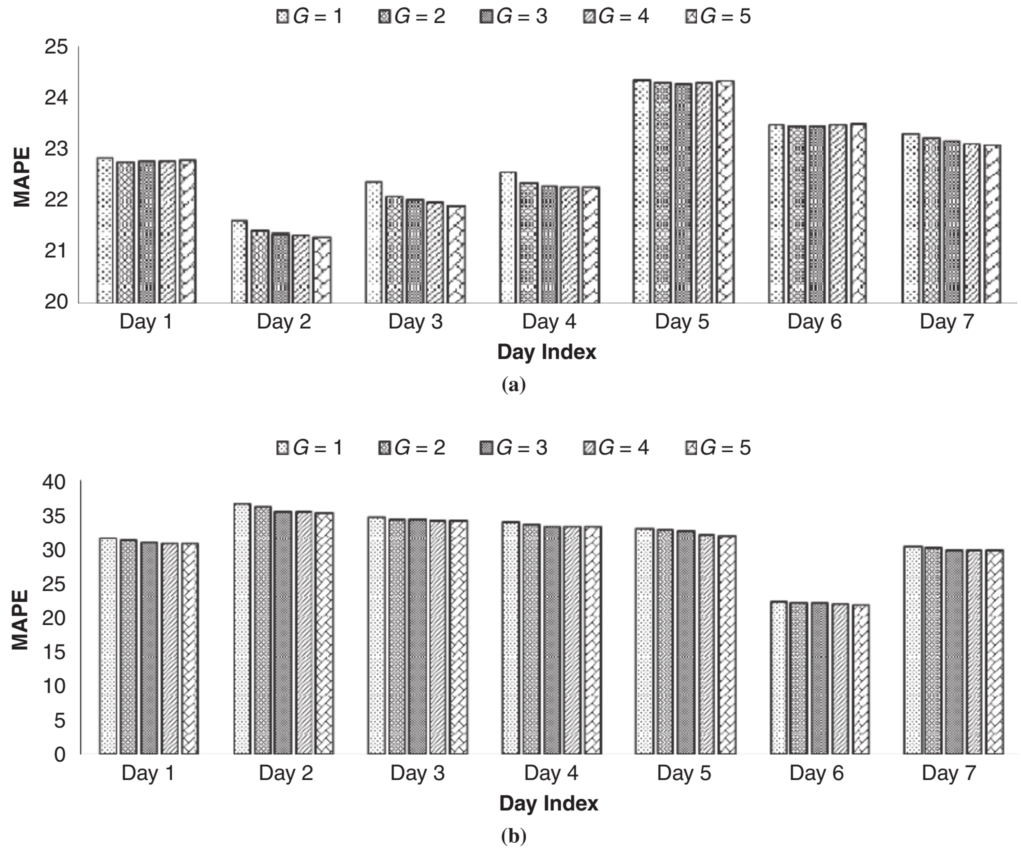

FIGURE 3 MAPE values for all trips with various combinations of G on (*a*) Route 19B and (*b*) Route 5C.

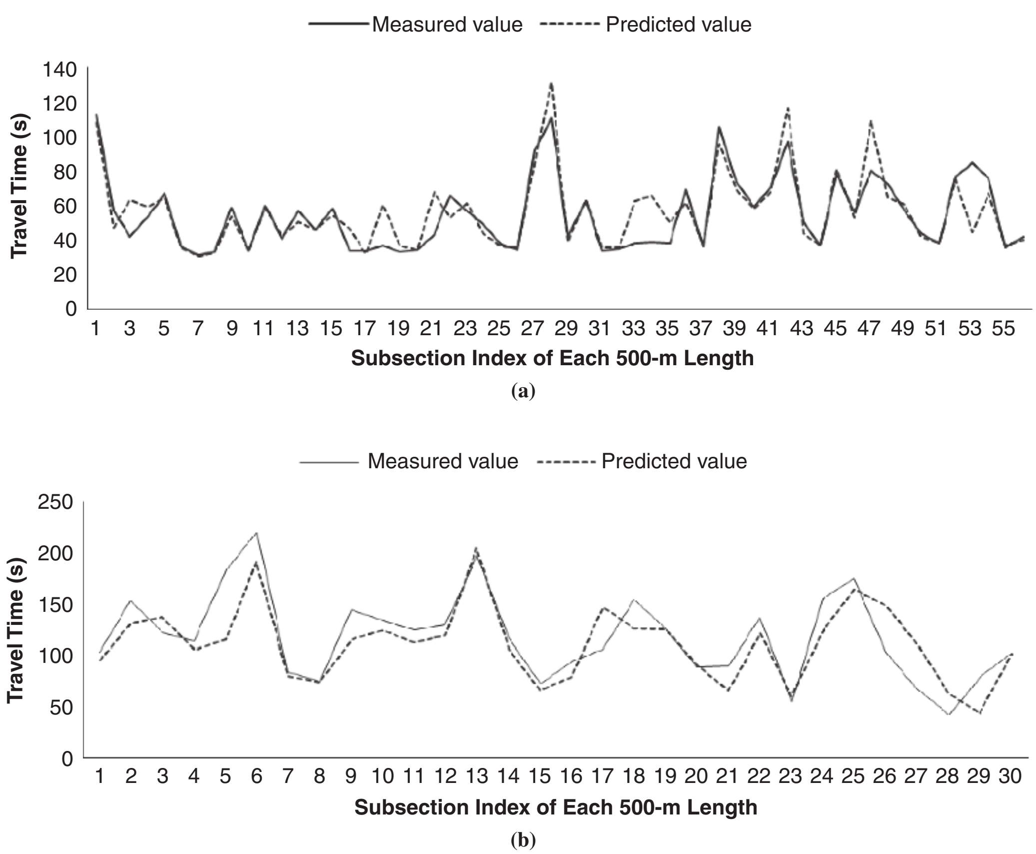

FIGURE 4 Actual and predicted travel times for sample trips over 500-m subsections on (*a*) Route 19B (MAPE = 13.38) and (*b*) Route 5C (MAPE = 17.33).

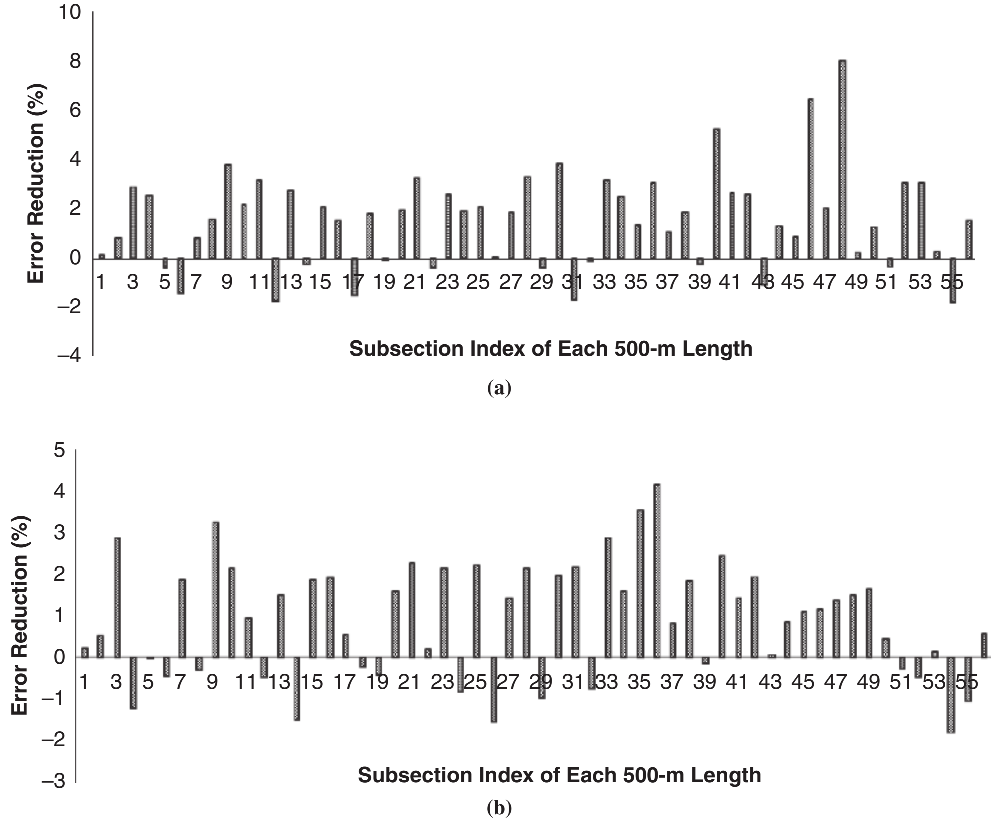

FIGURE 5 Error reduction by considering previous two subsections rather than previous one subsection for (*a*) off-peak trips and (*b*) peak trips.

deviation of the predicted arrival time and the observed arrival time of the bus at all bus stops during a 14-day test period on the selected study route. Data from the literature indicate that, for a bus with a 1.5-h journey, 5 min is an acceptable level of prediction accuracy (*7*). According to Warman, passengers will tolerate up to ±5 min if 88% of the predicted times are less than ±5 min (*28*). The TriMet Transit Tracker System reports that passengers tolerate a waiting time of 2 to 4.5 min at bus stops (*29*). Therefore, an accuracy of ±5 min is considered the acceptable error limit in this study. Figure 6 shows the frequency (as a percentage) of times when the deviation was less than ±1 min, less than ±2 min, less than ±3 min, less than ±4 min, and less than ±5 min for two selected bus stops on Route 19B (Karapakkam and Saidapet) by using the proposed approach. Results indicate that prediction error was ±5 min in 92% of the cases.

Another important question was how far in advance a prediction could be made to inform passengers waiting at bus stops about bus arrival times. Results are presented in Figure 7 as MAPE and MAE at prediction horizons from 2 to 30 min. Table 1 shows the percentages of times the deviation was less than ±5 min for selected bus stops when the prediction was made 30 min ahead. Results indicate that predictions were within ±5 min for 92% of instances at the Saidapet bus stop, which is the farthest from the route origin. Also, errors are

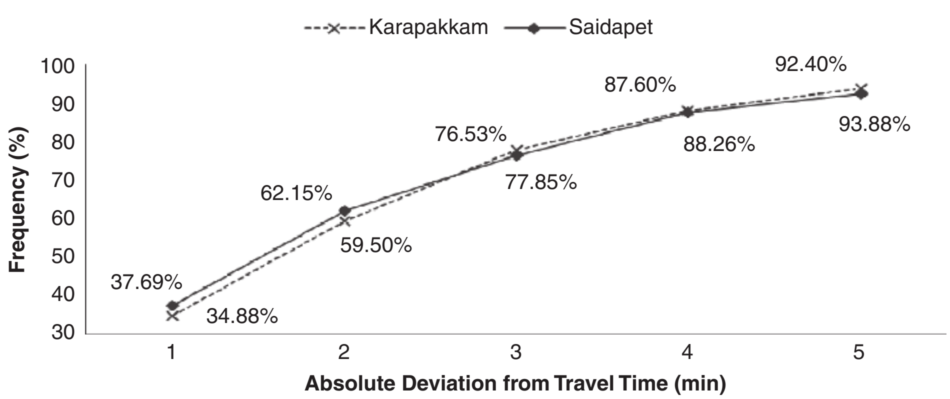

FIGURE 6 Results of predicted arrival time deviation at selected bus stops.

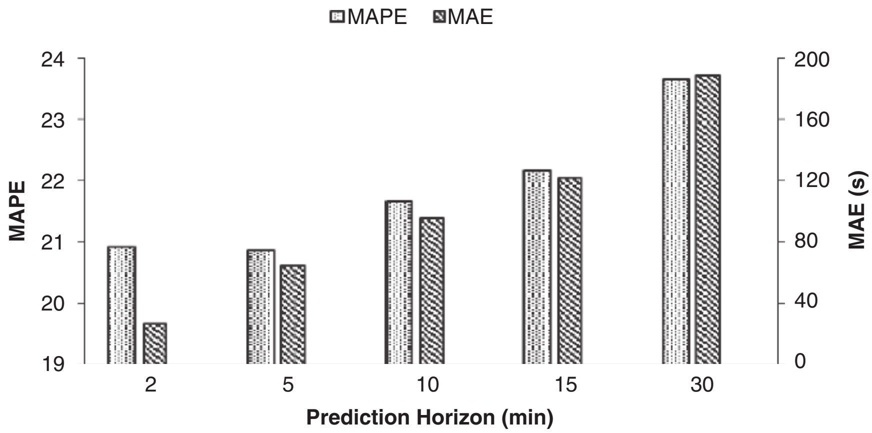

FIGURE 7 Average prediction accuracy as MAPE and MAE for varying prediction horizon.

TABLE 1 Performance of Predictions 30-Min Ahead for Selected Bus Stops

Bus Stop Number	Bus Stop Name	Cumulative Distance from Initial Bus Stop (km)	Percentage Within ±5-min Deviation	MAPE	MAE (s)
1	Semmancheri	11.03	97.36	24.88	140
2	Shozhinganallur	13.74	95.37	26.54	152
3	Karapakkam	15.55	93.88	26.65	159
4	Thoraipakkam	18.81	94.38	26.08	167
5	Lattice Bridge	23.45	93.55	26.84	166
6	Women's Polytechnic College	24.80	94.05	26.20	166
7	Madhya Kailash	25.82	93.22	27.25	177
8	Saidapet	29.94	92.40	26.66	170

still within 25% MAPE and ±5 min MAE, respectively, indicating acceptable performance for a prediction horizon up to 30 min.

SUMMARY AND CONCLUSIONS

The main aim of APTS is to attract passengers to public transportation and, in turn, reduce congestion on urban roads. One way to make public transportation more attractive is to provide accurate information about bus arrival times. Prediction accuracy depends mainly on optimal inputs and a suitable prediction method. This study developed a method for predicting travel time after identifying the optimal inputs. A KFT-motivated algorithm was used to predict bus travel time by using GPS-based data from public transport buses in Chennai, India. Results show that the error reduction was marginal when more than the previous two subsections were considered. Also, the error reduction when previous two subsections were considered was minimal compared with the case considering only one previous subsection. Therefore, for field implementation, the previous one subsection travel time may be used to reduce computational effort.

The present study used travel time data from the previous two buses alone to predict the travel time of the next bus, without considering historical and current time data. Thus, current traffic conditions, as well as repeating patterns in travel time (e.g., peak and off-peak traffic), may not be captured in the prediction process. The proposed method may be further improved by explicitly incorporating section-specific characteristics (e.g., bus stops and signals) and by automating prediction results.

The results obtained from the algorithm seem promising and can be used to implement an APTS application. The predicted travel time obtained from the proposed algorithm can be expressed as time remaining or actual clock time and can be displayed at bus stops, on buses, through web portals, or in SMS text messages.

ACKNOWLEDGMENT

The authors acknowledge support for this study as a part of the Centre of Excellence in Urban Transport project funded by the Ministry of Urban Development, Government of India.

REFERENCES

1. Houghton, J., J. Reiners, and C. Lim. *Intelligent Transport System: How Cities Can Improve Mobility.* IBM Institute for Business Value, Armonk, N.Y., 2009. ftp://ftp.software.ibm.com/common/ssi/pm/xb/n/gbe03232usen/GBE03232USEN.pdf. Accessed May 1, 2013.
2. Dewan, K. K., and I. Ahmad. Carpooling: A Step to Reduce Congestion (A Case Study of Delhi). *Engineering Letters,* Vol. 14, No. 1, 2007. http://www.engineeringletters.com/issues_v14/issue_1/EL_14_1_12.pdf. Accessed May 1, 2014.
3. Schweiger, C. L. *TCRP Synthesis 48: Real-Time Bus Arrival Information Systems.* Transportation Research Board of the National Academies, Washington, D.C., 2003.

4. Jeong, R., and L. R. Rilett. Bus Arrival Time Prediction Using Artificial Neural Network Model. *Proc., 7th International IEEE Conference on Intelligent Transportation Systems,* Oct. 3–6, Washington, D.C., 2004, pp. 988–993.
5. Ramakrishna, Y., P. Ramakrishna, V. Lakshmanan, and R. Sivanandan. Bus Travel Time Prediction Using GPS Data. *Proc., Map India 2006.* http://www.gisdevelopment.net/proceedings/mapindia/2006/student%20oral/mi06stu_84.htm.
6. Chien, S. I. J., Y. Ding, and C. Wei. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. *Journal of Transportation Engineering,* Vol. 128, No. 5, 2002, pp. 429–438.
7. Bhandari, R. R. *Bus Arrival Time Prediction Using Stochastic Time Series and Markov Chains.* PhD dissertation. New Jersey Institute of Technology, Newark, N.J., 2005. http://archives.njit.edu/vol01/etd/2000s/2005/njit-etd2005-038/njit-etd2005-038.pdf. Accessed Jan. 12, 2014.
8. Suwardo, M. N., and I. Kamaruddin. ARIMA Models for Bus Travel Time Prediction. *Journal of the Institute of Engineers Malaysia,* Vol. 71, No. 2, 2010, pp. 49–58.
9. Abdelfattah, A. M., and A. M. Khan. Models for Predicting Bus Delays. In *Transportation Research Record 1623,* TRB, National Research Council, Washington, D.C., 1998, pp. 8–15.
10. Patnaik, J., S. Chien, and A. Bladikas. Estimation of Bus Arrival Times Using APC Data. *Journal of Public Transportation,* Vol. 1, 2004, pp. 1–20.
11. Chang, H., D. Park, S. Lee, H. Lee, and S. Baek. Dynamic Multi-Interval Bus Travel Time Prediction Using Bus Transit Data. *Transportmetrica,* Vol. 6, No. 1, 2010, pp. 19–38.
12. Chen, M., X. B. Liu, and J. X. Xia. A Dynamic Bus Arrival Time Prediction Model Based on APC Data. *Computer-Aided Civil and Infrastructure Engineering,* Vol. 19, 2004, pp. 364–376.
13. Bin, Y., Y. Zhinzhen, and Y. Baozhen. Bus Arrival Time Prediction Using Support Vector Machines. *Journal of Intelligent Transportation Systems,* Vol. 10, No. 4, 2006, pp. 151–158.
14. Chun-Hsin, W., W. Chia-Chen, S. Da-Chum, C. Ming-Hua, and H. Jan-Ming. Travel Time Prediction with Support Vector Regression. *Proc., 2003 IEEE Conference on Intelligent Transportation Systems,* Shanghai, China, Oct. 12–15, 2003.
15. Cathey, F. W., and D. J. Dailey. A Prescription for Transit Arrival/Departure Prediction Using AVL Data. In *Transportation Research Part C,* Vol. 11, 2003, pp. 241–264.
16. Shalaby, A., and A. Farhan. Prediction Models of Bus Arrival and Departure Times Using AVL and APC Data. *Journal of Public Transportation,* Vol. 7, No. 1, 2004, pp. 41–60.
17. Wall, Z., and D. J. Dailey. An Algorithm for Predicting the Arrival Time of Mass Transit Vehicles Using Automatic Vehicle Location Data. Presented at 78th Annual Meeting of the Transportation Research Board, Washington, D.C., 1999.
18. Chien, S. I. J., and C. M. Kuchipudi. Dynamic Travel Time Prediction with Real-Time and Historic Data. *Journal of Transportation Engineering,* Vol. 129, No. 6, 2003, pp. 608–616.
19. Chu, L., J. S. Oh, and W. Recker. Adaptive Kalman Filter Based Freeway Travel Time Estimation. Presented at 84th Annual Meeting of the Transportation Research Board, Washington, D.C., 2005.
20. Nanthawichit, C., T. Nakatsuji, and H. Suzuki. Application of Probe-Vehicle Data for Real-Time Traffic-State Estimation and Short-Term Travel-Time Prediction on a Freeway. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1855,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 49–59.
21. Son, B., H. J. Kim, C. H. Shin, and S. K. Lee. Bus Arrival Time Prediction Method for ITS Application. In *Knowledge Based Intelligent Information and Engineering Systems.* Springer, Berlin, 2004, pp. 88–94.
22. Yu, B., W. H. K. Lam, and M. L. Tam. Bus Arrival Time Prediction at Bus Stop with Multiple Routes. *Transportation Research Part C,* Vol. 19, No. 6, 2011, pp. 1157–1170.
23. Zhu, T., F. Ma, T. Ma, and C. Li. The Prediction of Bus Arrival Time Using Global Positioning System Data and Dynamic Traffic Information. *Proc., 4th Joint IFIP Wireless and Mobile Networking Conference (WMNC),* Beijing, China, Oct. 26–28, 2011.
24. Google Maps. https://maps.google.co.in/. Accessed July 1, 2014.
25. Chamberlain, R. G. *Great Circle Distance Between Two Points.* http://www.movabletype.co.uk/scripts/gis-faq-5.1.html. Accessed Feb. 14, 2013.
26. Vanajakshi, L., S. C. Subramanian, and R. Sivanandan. Travel Time Prediction Under Heterogeneous Traffic Conditions Using GPS Data from Buses. *IET Journal on Intelligent Transportation Systems,* Vol. 3, No. 1, 2009, pp. 1–9.
27. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME—Journal of Basic Engineering,* Vol. 82, No. 1, 1960, pp. 35–45.
28. Warman, P. *Measured Impacts of Real-Time Control and Information Systems for Bus Services.* Transport Direct, U.K. Department for Transport, London, 2003.
29. Crout, D. T. Accuracy and Precision of the Transit Tracker System. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1992,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 93–100.

The Standing Committee on Transit Management and Performance peer-reviewed this paper.