https://doi.org/10.1007/s40999-017-0210-y

RESEARCH PAPER

# Real Time Identification of Inputs for a BATP System Using Data Analytics

Rakesh Behera[1] · B. Anil Kumar[1] · Lelitha Vanajakshi[1]

Received: 2 January 2016 / Revised: 19 November 2016 / Accepted: 6 March 2017 / Published online: 5 April 2017

**Abstract** In recent times, bus arrival time prediction (BATP) systems are gaining more popularity in the field of advanced public transportation systems, a major functional area under intelligent transportation systems. BATP systems aim to predict bus arrival times at various bus stops and provide the same to passenger's pre-trip or while waiting at bus stops. A BATP system, which is accurate, is expected to attract more commuters to public transport, thus helping to reduce congestion. However, such accurate prediction of bus arrival still remains a challenge, especially under heterogeneous and lane-less traffic conditions such as the one existing in India. The uncertainty associated with such traffic is very high and hence the usual approach of prediction based on average speed will not be enough for accurate prediction. To make accurate predictions under such conditions, there is a need to identify correct inputs and suitable prediction methodology that can capture the variations in travel time. To accomplish the above goal, a robust framework relying on data analytics is proposed in this study. The spatial and temporal patterns in travel times were identified in real time by performing cluster analysis and the significant inputs thus identified were used for the prediction. The prediction algorithm used the Adaptive Kalman Filter approach, to take into account of the high variability in travel time. The proposed schemes were corroborated using real-world GPS data and the results obtained are very promising.

**Keywords** Input identification · Cluster analysis · Bus travel time prediction

## 1 Introduction

In today's busy society, information regarding arrival/travel time of vehicles is becoming more and more valuable. With the information of predicted arrival times of buses at each bus stop available via variable message signs (VMS) or as mobile or web application, people can make timely plans for their upcoming activities, which will also reduce the anxiety caused by uncertain delays. This necessitates a system that can inform the travelers about the current/expected future travel times of the concerned buses before they make their transit plans. This may also attract more passengers to use public transport, which in turn may reduce congestion. In the recent years, due to the advent of positioning and wireless communication technologies, wireless devices equipped with global positioning system (GPS) have been widely deployed on various private and public vehicles, generating massive amount of vehicle trajectory data, which can be used for fleet management [1] and other transportation applications. GPS based time-tagged location data, usually represented in the form of trajectories, brings a great potential for real-time prediction of the vehicle travel times. However, accurate prediction of bus arrival time is a challenging problem, especially under heterogeneous and lane-less traffic such as the one existing in India. The uncertainty associated with such traffic is very high and hence the usual approach of prediction based on average speed of bus and known distance to the bus stop will

✉ Lelitha Vanajakshi
lelitha@iitm.ac.in

Rakesh Behera
rakesh8248@gmail.com

B. Anil Kumar
raghava547@gmail.com

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

not be enough for accurate prediction. For accurate predictions under such conditions, there is a need to identify correct inputs that can capture the variations in travel time and use suitable prediction methodology that can take into account the variability.

Bus arrival time prediction methodologies reported in the recent past can be broadly classified as data driven and model based. Data driven techniques requires a good amount of database, whereas model based techniques requires a relatively limited data. However, irrespective of the amount of data required, one should use the most significant input for better prediction accuracy. Schweiger [2] suggested that the performance of prediction techniques in terms of their accuracy depends on the travel time patterns of the data collected. Identifying the most significant and effective input data and using them in prediction methods will improve their performance. With the availability of big amount of data from tracking devices, data analytics tools can be effectively used for these applications. In the present study, clustering analysis of historical travel time trajectories is used for identifying in real time the most significant input data that can be used for the prediction. The prediction was carried out using a model based approach with adaptive Kalman filtering (KF) as the estimation technique. Thus, this paper proposes a hybrid prediction framework to predict the travel time of buses by exploiting data analytics of historical trajectory data and an efficient state estimation technique capable of making precise estimations from the available travel time measurements.

The Sect. 2 section gives a brief overview of the literature in bus travel time prediction. The section on data analysis discusses the various analyses carried out with real-world bus trajectory data to explore patterns in travel time. Methodology section explains the schemes for similar trajectory selection and travel time prediction methodology. Section 7 presents the evaluation of the various schemes using real-world data. Finally, the study is concluded by summarizing the findings.

## 2 Literature Review

Many researchers have suggested different techniques to find out significant data to be used in the prediction. Existence of patterns in historical trajectories is one of the key reasons for the use of historical trajectories for travel time prediction. Zhang [3] proposed a pattern-based short-term traffic forecasting approach based on the integration of multi-phase traffic flow theory and time series analysis methods. Kwon et al. [4] used the data obtained from loop detectors to obtain day-to-day travel time trends to predict travel time using regression analysis. It was reported that there is a strong dependence between two successive vehicle travel times within a day. Lee et al. [5] used GPS data to analyze travel time patterns using historical travel time trajectories similar to the current trip. Jensen and Tie [6] explored the use of techniques that enable efficient trajectory data management. The trajectories were subsequently used by arrival time prediction algorithms that utilize techniques for efficient similarity search in historical database. Elhenawy et al. [7] developed a Genetic Programming algorithm to predict travel time by using k-means approach to partition the data in to similar clusters. Min and Winter [8] used spatio-temporal correlations for real-time traffic prediction. The techniques reported to predict travel/arrival times can be broadly classified into data driven and model based techniques. Data driven approaches predict travel time with the use of statistical relationships, which are derived from historical data (travel times, speeds, volumes, etc.). The most commonly reported data driven approaches in the literature include machine learning techniques, time series analysis, and historical averaging approaches.

Under data driven techniques, Patnaik et al. [9] used multivariate linear regression for bus arrival time estimation using automatic passenger counter (APC) data. Artificial neural network (ANN) is another widely used method under this category. Liu et al. [10] used neural networks to indirectly predict travel times using traffic volume and flow data. Afandizadeh [11] proposed a short-term traffic flow prediction approach based on an advanced multi-layer feed forward neural network (MLF) model synthesized using genetic algorithms (GA). Moghaddam et al. [12] used ANN to predict crash severity prediction in urban highways and identifying significant crash-related factors. Hinsbergen and Van lint [13] used an approach that combines neural networks with Bayesian Inference Theory for predicting travel times. Bansal et al. [14] developed freeway travel time prediction method by using back propagation neural networks. Fan and Gummu [15] compared three methods namely, ANN, KF and historical average for their performance in bus arrival prediction and concluded that ANN performs better than rest. Lin et al. [16] developed two ANN models namely Sub ANN and Hierarchical ANN to capture the variations in travel time over different days of the week. Real-time prediction using support vector regression (SVR) and support vector machine (SVM) has also been reported. Vanajakshi and Rilett [17] demonstrated the use of SVR for short-term travel time prediction when the nature of data is sparse and noisy. Yu et al. [18] compared four methods namely, ANNs, SVM, k-NN, and regression for their performance in arrival time prediction at bus stops with multiple routes and concluded that SVM performs better than the rest. However, both ANN and SVR are expensive in training for real-time updates. Guin [19] used a time series analysis approach called seasonal autoregressive integrated moving average (SARIMA) to predict travel times using historical travel time data.

Model based approaches uses models based on the physics of the system and capture the dynamics of the system by establishing mathematical relationship between appropriate variables. Krishnan and Polak [20] explored recurring themes in traffic conditions and used k-Nearest Neighbors (k-NN) for indirectly predicting short term travel times using 15-min aggregate flow data. Esawey and Sayed [21] used a VISSIM micro-simulation model of down-town Vancouver to predict travel times using traffic volume and travel time data of nearby segments. Kalman Filtering (KF) is one of the most widely adopted estimation technique that is used in model based approaches. Son et al. [22] developed a method by using Kalman Filtering techniques to predict travel time from bus stop to stop line at signalized intersections. Shalaby and Farhan [23] used AVL and APC data to predict travel time of public transport vehicles by using the KFT. Model verification was done with the data extracted from VISSIM. Their method was compared with historical, regression and ANN methods. Chu et al. [24] used the adaptive Kalman filtering technique for travel time prediction. Nanthawichit et al. used GPS equipped probe vehicles and loop detectors data to estimate traffic parameters including travel time by using the Kalman Filtering technique.

Most of the above studies dealt with homogeneous traffic conditions. A few studies have been reported from heterogeneous traffic conditions such as the one existing in India. Vanajakshiet al [25] proposed a space discretization approach to predict bus travel time. In space discretization, the route was spatially discretized into smaller subsections to predict the travel time in upcoming subsections. Padmanabhan et al. [26] extended the above study by analyzing the dwell times explicitly. However, the above studies considered just two previous buses data as inputs without considering the patterns in travel time. It is well known that the travel time of the vehicles moving on fixed routes is not random. For example, there exist significant patterns in travel times for trips made around the same time of the day. Such patterns verify the possibility of using historical data of a certain segment to predict for the future traffic condition on the same segment. In the present study, the most significant input data that need to be used were identified by carrying out a clustering analysis using historical travel time trajectories. It can also be seen that there is a need to develop prediction techniques that can specifically take into account the high variability of travel time. This forms the basis for the development of the proposed framework with the objectives as:

1. To identify and explore the possible patterns in travel times by carrying out a clustering analysis using historical travel time trajectories.
2. To develop a prediction method that uses identified patterns that can capture the high variability in travel times.

## 3 Data Collection

### 3.1 Study Site and Description of Data

The raw GPS data used in this study were collected over a period of 4 months from the Metropolitan Transport Corporation (MTC) buses, running on one of the busiest routes in Chennai namely, 19B. Each bus is equipped with a GPS device that records the status of the bus every 10 s. Each data point consists of the GPS coordinates and the corresponding time-stamp. Location details of each bus stop in a selected route were collected and stored. The north-bound 19B route was chosen for analysis. This route has 15 stops, with the origin at the Kelambakkam bus depot and the last stop at Saidapet bus depot. It covers a distance of 29.4 km and the average trip duration from the origin to the destination, is more than 1 h. The average headway between vehicles in this route was around 30 min. A total of 2212 north-bound trajectories were collected in this route during the data collection period of 4 months. The route was divided into subsections of 200 m length for the present analysis. Figure 1 shows the study route with every 25th subsection marked.

## 4 Preliminary Analysis

### 4.1 Spatial Correlation in Travel Times

The raw GPS data were converted into the travel times using Haversine formula [27]. The transformed data were stored in a table format. Each row corresponds to a trajectory and the columns correspond to the segment-wise travel times in seconds. Correlations between the travel times on segments separated by a varying number of intermediate segments were calculated to check the possibility of using past segment travel times as a similarity measure to find historical trajectories. Pearson's correlation, which gives a measure of linear association between two variables, was calculated for pairs of segments. Given two variables $X$ and $Y$, with means $\bar{X}$ and $\bar{Y}$and the standard deviations $\sigma_x$ and $\sigma_y$ , correlation between them is computed as,

**Fig. 1** Route map of the study stretch. (Source: Openstreet Maps)

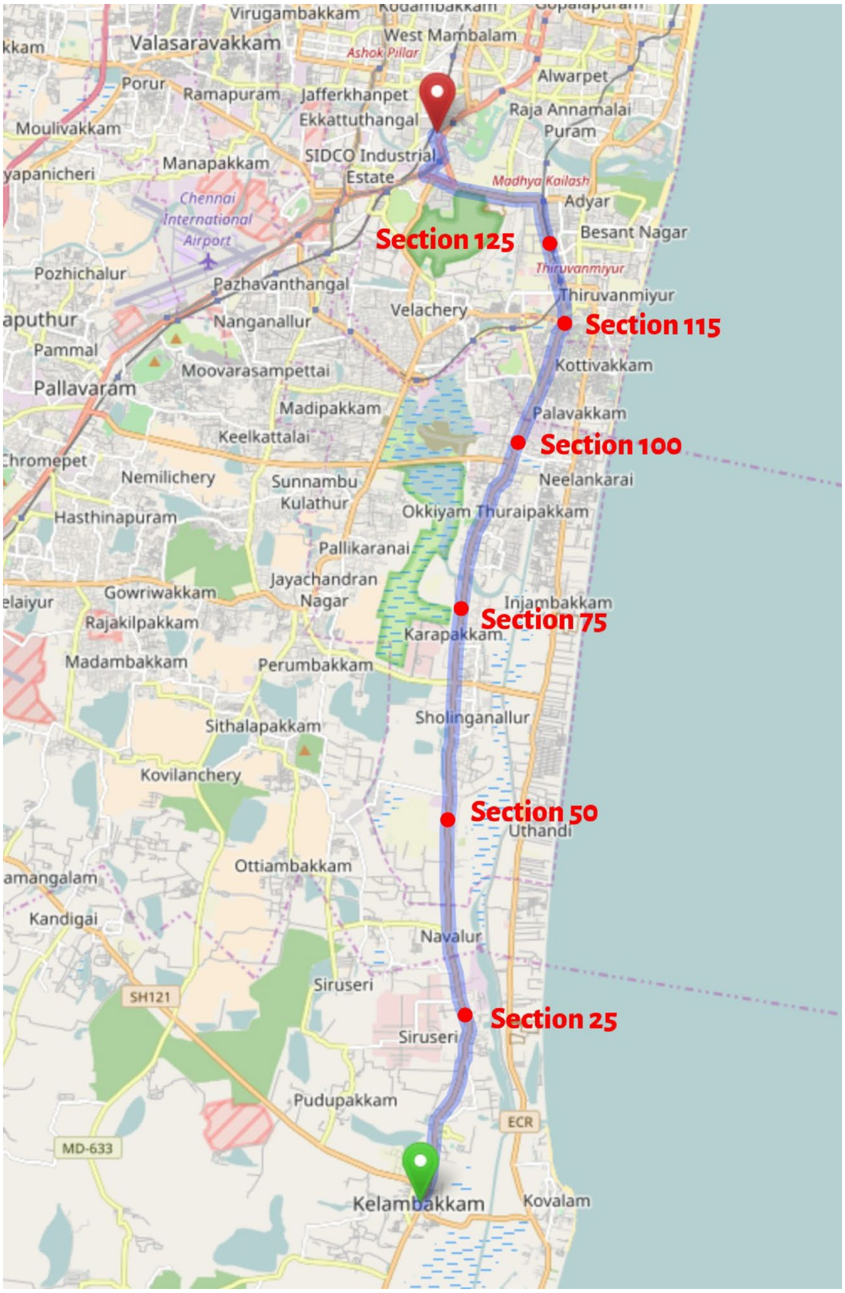
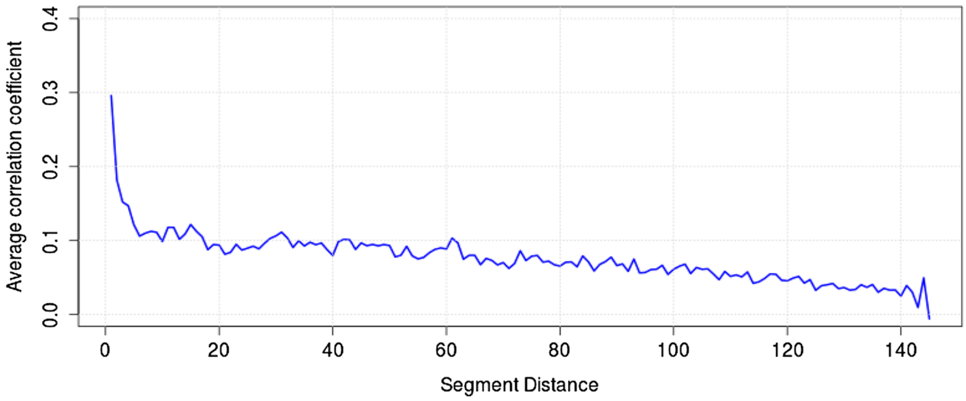


**Fig. 2** Average correlation coefficient versus the segment distance



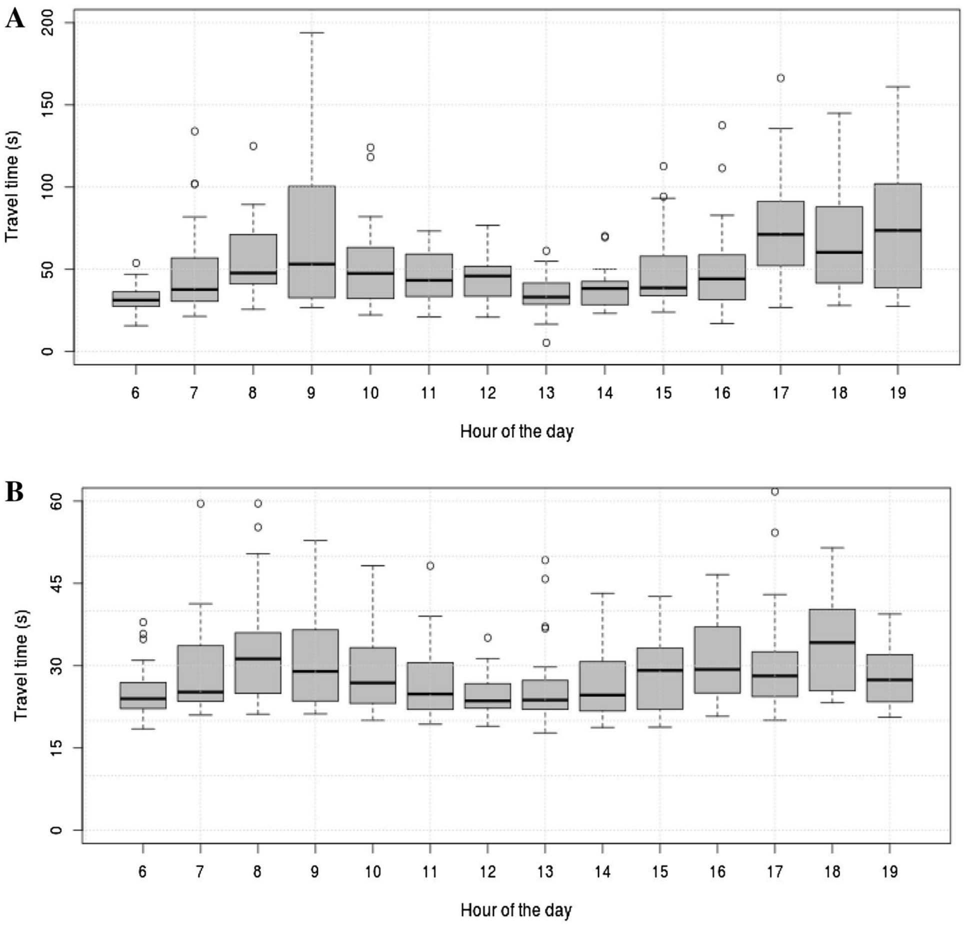


**Fig. 3** **a** Variation of travel times on Segment 28 across the hours of the day. **b** Variation of travel times on Segment 100 across the hours of the day

$$\gamma = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{(n-1)\sigma_X\sigma_Y}, \tag{1}$$

where, $n$ is the number of elements in $X$ and $Y$. The closer its value to 1, the stronger is the positive correlation [28]. Figure 2 shows the average Pearson's correlation coefficient with $Y$-axis showing the Pearson's correlation coefficient between two travel time arrays (historical travel times corresponding to two segments) and $X$-axis showing the number of segments between them, which is termed as the segment distance. From Fig. 2, it can be seen that when two segments are near to each other, the Pearson's correlation is higher than others. Therefore, a segment is more related to nearby segments than farther ones. Based on this, it can be concluded that segments closer to the one being analyzed are the most correlated ones and can be used as input for prediction.

### 4.2 Temporal Patterns in Travel Times

Travel times can be associated to temporal features as well. For example, the traffic conditions are usually the worst during the peak hours in the morning and evening. In weekdays, the travel times may be higher than those on weekends. The present study conducted analysis to identify these patterns, mainly the intra-day hourly pattern and day-wise pattern. To visualize the variation within a day, travel times were grouped into 14 time periods (corresponding to working hours of the day), each of 1-h duration. Figure 3a, b shows the hourly variations in travel times within a day for two representative segments namely, Segment 28 and Segment 100. From Fig. 3a, b, it can be seen that the travel times in the morning from 8 to 10 am and in the evening from 5 to 7 pm are relatively higher than others. Hence, these hours were considered as peak hours. In addition, it can be observed from the box plots that the peak hour trajectories have more variance than those in off-peak hours. Figure 4 shows a heat-map that represents the correlation matrix which was obtained by separating the historical travel times on Segment 28 into 14 bins (corresponding to the 14 working hours in a day) and calculating the Pearson's correlation coefficients among them. From Fig. 4, it can be observed that the diagonal squares are all white (correlation = 1) since these represent the correlation of one bin with itself. In addition, it can be observed that the

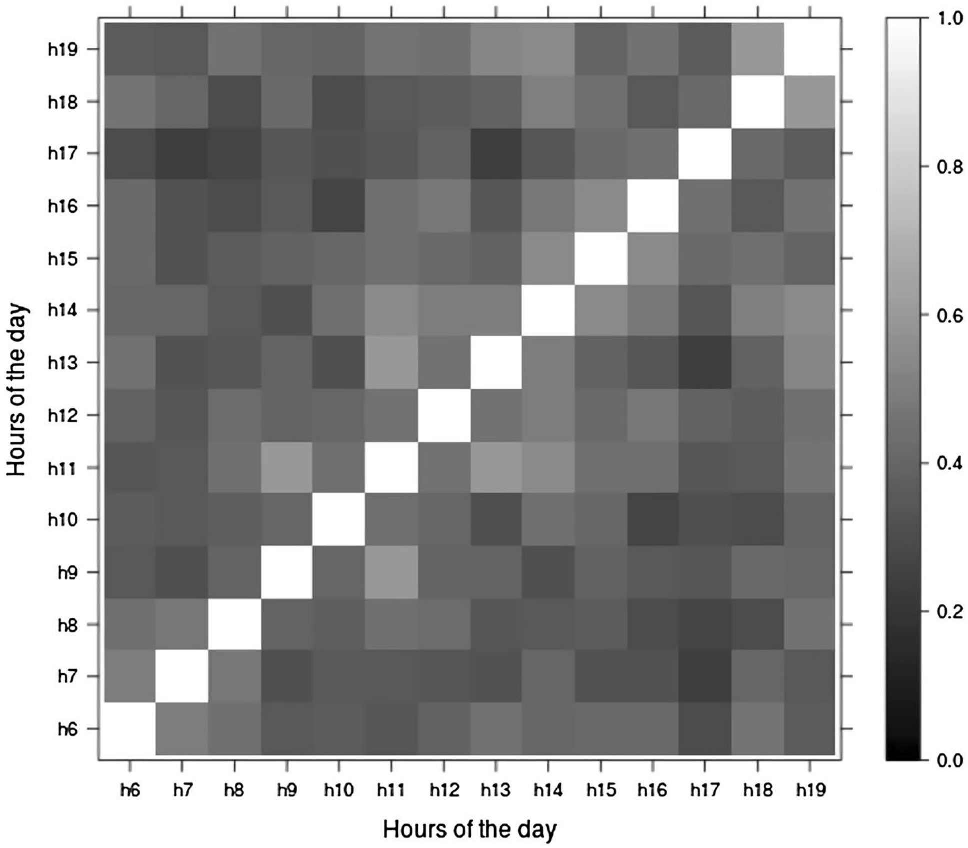


**Fig. 4** Correlations between travel times occurring in different hours of the day

squares closer to the diagonal are whiter than those away from the diagonal representing the historical travel times which occurred temporally closer (within a radius of 1–2 h) to each other are more correlated. Based on this the temporal neighborhood features were included for the selection of similar historical trajectories.

Travel times are not only correlated to the hour they happen, but also to the day on which they happen. To visualize and verify the correlations between travel times and the day, the days were first classified into two classes namely, weekday and weekend. Figure 5a shows the space–time trajectories of the trips that happened on weekdays and weekend. From Fig. 5a, it can be seen that travel times in weekdays have higher variance than those in weekends. Thus, the assumption of taking weekday/weekend as a discriminative feature for trajectory selection is valid. A similar analysis across different days of the week is shown in Fig. 5b and it can be seen that they are not distinctly different from each other and hence they were not separately analyzed.

From the analyses, it can be concluded that several patterns exist in the travel times of buses moving on the same route. In the present case, the weekday/weekend pattern and the intra-day hourly pattern are the most significant. Based on these patterns, schemes based on the temporal features of the trajectories were proposed, and is discussed in sections below. From the correlation analysis, it was concluded that the correlation between closely spaced segments is significant. Based on this, the past segments features were included while selecting similar historical trajectories as discussed in the next section.

## 5 Cluster Analysis

Based on the conclusions derived from the pattern analysis presented in the previous section, similar trajectory selection were adopted on the corresponding segments. This method is limited by the low efficiency that is caused by searching for similar travel times based on each segment, especially when the number of historical trajectories is large. To overcome this, travel times were grouped into clusters and match the current travel time to the cluster averages to find the appropriate cluster. Lee et al. [5] compared two robust clustering algorithms namely, K-means and V-clustering and reported V-clustering to be performing better. Based on this, V-clustering algorithm is adopted in this study. The V-clustering algorithm [29] allocates

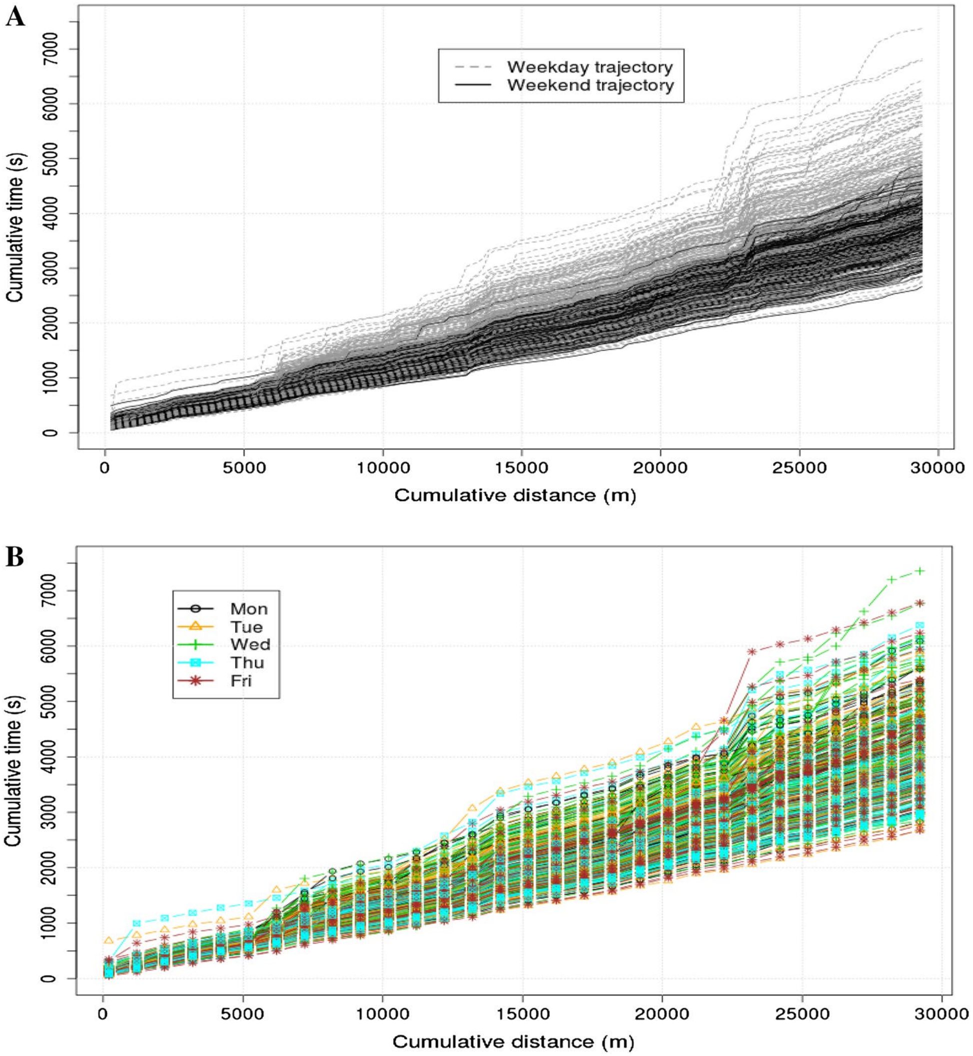


**Fig. 5** **a** Comparison between weekday and weekend trips. **b** Comparison between the weekdays

a sorted list of one-dimensional data into clusters. In this algorithm, a list of values is first sorted. Then it is split into clusters in an iterative manner. At each iteration, the list is split into two parts and the weighted average variance (WAV) is calculated for the resulting child lists. An optimum split is found out that minimizes the WAV of the resulting child lists. The WAV for a split at the *i*th element of the list is defined as

$$\text{wav}_i = \left|\frac{L_1^i}{L}\right| \text{var}\,(L_1^i) + \left|\frac{L_2^i}{L}\right| \text{var}\,(L_2^i), \tag{2}$$

where, $\left|L_1^i\right|$ and $\left|L_2^i\right|$ are the cardinalities of the resulting child lists for the *i*th split and var $(L_1^i)$ and var $(L_2^i)$ are their respective variances. The list is recursively partitioned so that the running time of the clustering algorithm for a segment with $M$ historical travel times becomes $O(\log M)$. Hence, the running time for the entire trajectory database is $O(N \log M)$[1], $N$ is the total number of segments on the route. The iteration is stopped when each cluster is left with a minimum number of travel times (or minimum number of trajectories, MNT), which is a tunable parameter (i.e. its value is decided to strike a balance between minimizing the errors in prediction and maximizing the computational speed). Each cluster for a segment is associated with a cluster average, i.e., the average of all the travel times in it. The selection of the values of various parameters of the clustering algorithm was made based on experimentation with real world trajectory data. The implementation was carried out as below.

Given a bus currently lies on the 4th segment i.e. $S_4$ (with the past segments $S_1$, $S_2$ and $S_3$), if the travel time on $S_3$ i.e.

[1] The running time of an algorithm, $f(x)$ (where $x$ is the input size) is said to be $O(g(x))$, which is read as $f(x)$ is big-oh of $g(x)$, if and only if there are constants $C$ and $n0$ such that $|f(x)| \leqslant \mathbf{C}|g(x)|$ whenever $x > n0$.

$t_3$ falls in a certain cluster for $S_3$, that cluster is taken as the matching cluster for $t_3$. All trajectories whose travel times on $S_3$ fall in the matching cluster are marked. Matching is usually done by finding the cluster for the particular segment, whose cluster average is closest to $t_3$ (which is the current trajectory's actual travel time on $S_3$). The same operation is applied using the actual travel times on $S_2$ and $S_1$ and then the trajectories which fall in the intersection of the matched clusters for the past three segments can be found. This method is known as segment filtering [5]. If for each of the three passed segments, the historical trajectories' travel times are similar to the current trajectory, they can be considered as "similar" to the current trajectory and can be used for prediction. However, the segment filtering method has a limitation when the number of historical trajectories is small and when the segments are smaller in length. In such a case, the sets of similar trajectories found for $S_1$, $S_{2,}$ and $S_3$ may not overlap at all (or have negligible intersection). To overcome this issue, the travel times for a fixed number of consecutive segments (which is a tunable parameter) were first aggregated and the clustering algorithm was applied. Then, the set of similar trajectories found for $S_3$ (using the aggregate travel time on $S_{3,}S_2$ and $S_1$) was used as the final set of similar trajectories, without applying segment filtering. For example, if the initial idea was to use three past segments in segment filtering, we aggregate the travel times for every triple of consecutive segments and then cluster the aggregated travel times for each such triple. As the bus moves from one segment to the next, this window of three past segments is maintained and the clusters for the triple are searched to find the match. In the next stage, schemes that use temporal features of the historical travel times were used to select the similar trajectories. Based on the findings from preliminary analysis, the weekday and weekend trips were separated. Then, the temporal neighborhood scheme is applied. For example, if the current trajectory occurs on a Wednesday (a weekday) and has a start time of 9:00 am, all the weekday trajectories that happened between 8:00 and 10:00 am (assuming a temporal neighborhood radius of 1 h) are returned by the temporal feature schemes. These trajectories are then passed to the past segment scheme which clusters the segment travel times and performs cluster matching using the past segment travel times (of the current trajectory) to find the final set of similar trajectories. This final set is then used by the prediction algorithm as discussed next.

## 6 Travel Time Prediction

The present study focused on a robust, short-term prediction technique based on Kalman Filter [30] which will take the identified similar trajectories from the above step as input to predict the next bus travel time. Vanajakshi et al. [25] is one of the earliest attempts that used KF with GPS probe vehicle data for short-term travel time prediction under heterogeneous traffic conditions. They used only previous two buses travel times to predict the travel time of the test vehicle (the ongoing trip). However, when the headways between the consecutive vehicles are more (approx. 1 h), the accuracy of this approach decreases. This is mainly due to the previous bus passing during an off-peak hour and the test vehicle passing in a peak hour, or vice versa. Since inputs to KF were fixed, this led to bigger errors during all transition periods. In the present study, this issue is addressed by making the inputs to KF dynamic in nature. As new similar trajectories are found using the cluster analysis, the inputs to the prediction algorithm are changed.

Travel time predictions were made using a model based approach based on the formulations given in Vanajakshi et al. [25]. The evolution of travel time between various segments is assumed to be governed by,

$$\Delta t_{i+1} = a_t \Delta t_{i+1} + w_t, \tag{3}$$

where $\Delta t_i$ is the travel time taken for covering $S_i$ (the $i$th subsection), $a_{\mathrm{i}}$ is a parameter that relates the travel time taken in $S_i$ to the travel time taken in $S_{i+1}$ and $w_i$ is the process disturbance associated with $S_i$. The measurement process was assumed to be governed by

$$z_i = \Delta t_i + v_i, \tag{4}$$

where $z_i$ is the measured travel time in $S_i$ and $v_i$ is the corresponding measurement noise. It was further assumed that $w_i$ and $v_i$ are zero mean white Gaussian noise signals with $Q_i$ and $R_i$ being corresponding variances. The prediction algorithm requires at least two trajectories as input in the form of segment-wise travel times. Trajectory which is more similar to the current one is called as base trajectory and the other one is called as correction trajectory. The data obtained from $T_{\mathrm{base}}$ is used to obtain the value of parameter $a_{\mathrm{i}}$ for each subsection. The data from $T_{\mathrm{corr}}$ were used in the prediction algorithm to obtain the estimate of the travel time of the test trajectory ($T_{\mathrm{test}}$). The steps involved in the prediction algorithm are:

1. The travel time data from $T_{\mathrm{base}}$ was used to obtain the value of $a_{\mathrm{i}}$,

$$a_i = \frac{\Delta t_{i+1}^{T_{\mathrm{base}}}}{\Delta t_i^{T_{\mathrm{base}}}}, i = 1, \ldots, (N-1), \tag{5}$$

where $\Delta t_i^{T_{\mathrm{base}}}$ is the travel time taken in $T_{\mathrm{base}}$ to cover $S_{\mathrm{i}}$.

2. Let $\Delta t_i^{T_{\mathrm{base}}}$ denote the travel time taken by $T_{\mathrm{test}}$ to cover $S_{\mathrm{i}}$. It is assumed that $E[\Delta t_1^{T_{\mathrm{base}}}$, and $E[\Delta t_1^{T_{\mathrm{base}}} - \Delta\hat{t}_1$, where $\Delta\hat{t}_1$ is the estimate of the travel time in $T_{\mathrm{test}}$ on $S_{\mathrm{i}}$.

3. For $i = 2\ldots, (N-1)$, the following steps are performed:

    a. The a priori estimate of the travel time is calculated by using

$$\Delta\hat{t}^{-}_{i+1} = a_i \Delta\hat{t}^{+}_{i}, \tag{6a}$$

where the superscript '−' indicates the a priori estimate and the superscript '+' indicates *a posteriori* estimate.

The a priori error variance was calculated by using

$$P^{+}_{i+1} = a_i P^{-}_{i} a_i + Q_i. \tag{6b}$$

The Kalman gain was calculated by using

$$K_{i+1} = \frac{P^{+}_{i+1}}{P^{+}_{i+1} + R_{i+1}}. \tag{6c}$$

The a posteriori travel time estimate and error variance were calculated using, respectively,

$$\Delta\hat{t}^{+}_{i+1} = \Delta\hat{t}^{-}_{i+1} + K_{i+1}[z_{i+1} - \Delta\hat{t}^{-}_{i+1} \tag{6d}$$

$$P^{+}_{i+1} = [1 - K_{i+1} \tag{6e}$$

To address the high variance of the travel time, the present study used an adaptiveKF, which takes into account the variability to a great extent. In the adaptive KF algorithm, the process disturbance $w_i$ associated with each segment $S_i$ was calculated using the actual travel times from the test trajectory as follows,

$$w_i = \Delta t^{T\text{test}}_{i+1} - a_i \Delta t^{T\text{test}}_{i}, \tag{7}$$

where, $\Delta t^{T\text{test}}_{i}$ is the actual travel time of the test vehicle on $S_i$, and the measurement noise $v_i$ associated with each segment $S_i$ was calculated as follows,

$$v_i = \Delta t^{T\text{base}}_{i} - \Delta t^{T\text{test}}_{i}, \tag{8}$$

Using these calculated values of $w_i$ and $v_i$, the variances $Q_i$ and $R_i$ (till the segment $S_i$) were calculated as,

$$Q_i = \frac{1}{i}\sum_{j=1}^{i}(w_j - \bar{w}_j)^2, \tag{9}$$

$$R_i = \frac{1}{i}\sum_{j=1}^{i}(v_j - \bar{v}_j)^2, \tag{10}$$

where,

$$\bar{w}_i = \frac{1}{i}\sum_{j=1}^{i} w_j, \tag{11}$$

$$\bar{v}_i = \frac{1}{i}\sum_{j=1}^{i} v_j. \tag{12}$$

Thus, the objective here is to predict the travel times of $T_{\text{test}}$ using the travel time data obtained from $T_{\text{base}}$ and $T_{\text{corr}}$. Corroborations were performed with actual GPS data collected from the MTC buses in Chennai and the results are discussed below.

## 7 Performance Evaluation

The results obtained from the implementation of the above algorithm are discussed in this section. Since the proposed algorithm uses dynamic inputs and updated the errors over each section, a comparison was carried out with the base method proposed by Vanajakshi et al. [25], which used the static inputs, i.e., the travel time data obtained from the previous two vehicles. For the purpose of evaluation, all the trips made during a period of 1 week were taken as test trips. The mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to quantify the prediction accuracy and were calculated as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{A}^{B_{\text{tar}}}_{i} - A^{B_{\text{tar}}}\right|, \tag{13}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{A}^{B_{\text{tar}}}_{i} - A^{B_{\text{tar}}}\right|}{A^{B_{\text{tar}}}} \times 100, \tag{14}$$

where, $\hat{A}^{B_{\text{tar}}}_{i}$ is the $i$th predicted arrival time at $B_{\text{tar}}$, $A^{B_{\text{tar}}}$ is the corresponding actual arrival time (calculated after the bus arrives at $B_{\text{tar}}$) and $n$ is the total number of arrival time predictions made.

Figure 6a shows a sample comparison of the predicted travel times and the measured travel times over every 600 m subsections for both methods. From the figure, it can be observed that the predicted travel times obtained from the proposed method were closely matching with the actual travel times than the base method. The corresponding MAPEs are also shown and it can be seen that the error is much less for the proposed method. Similar predictions were carried out for multiple trips and Fig. 6b shows results for all the trips of a sample day. From Fig. 6b, it can be observed that the proposed method is performing better than the base method for all trips and can be considered as a clear improvement for accurate bus arrival prediction. Figure 6c shows the performance comparison for multiple days. Here also, it can be observed that the proposed method is performing better than base method.

Since, one of the main interests of the present study was to predict the bus travel time and providing the

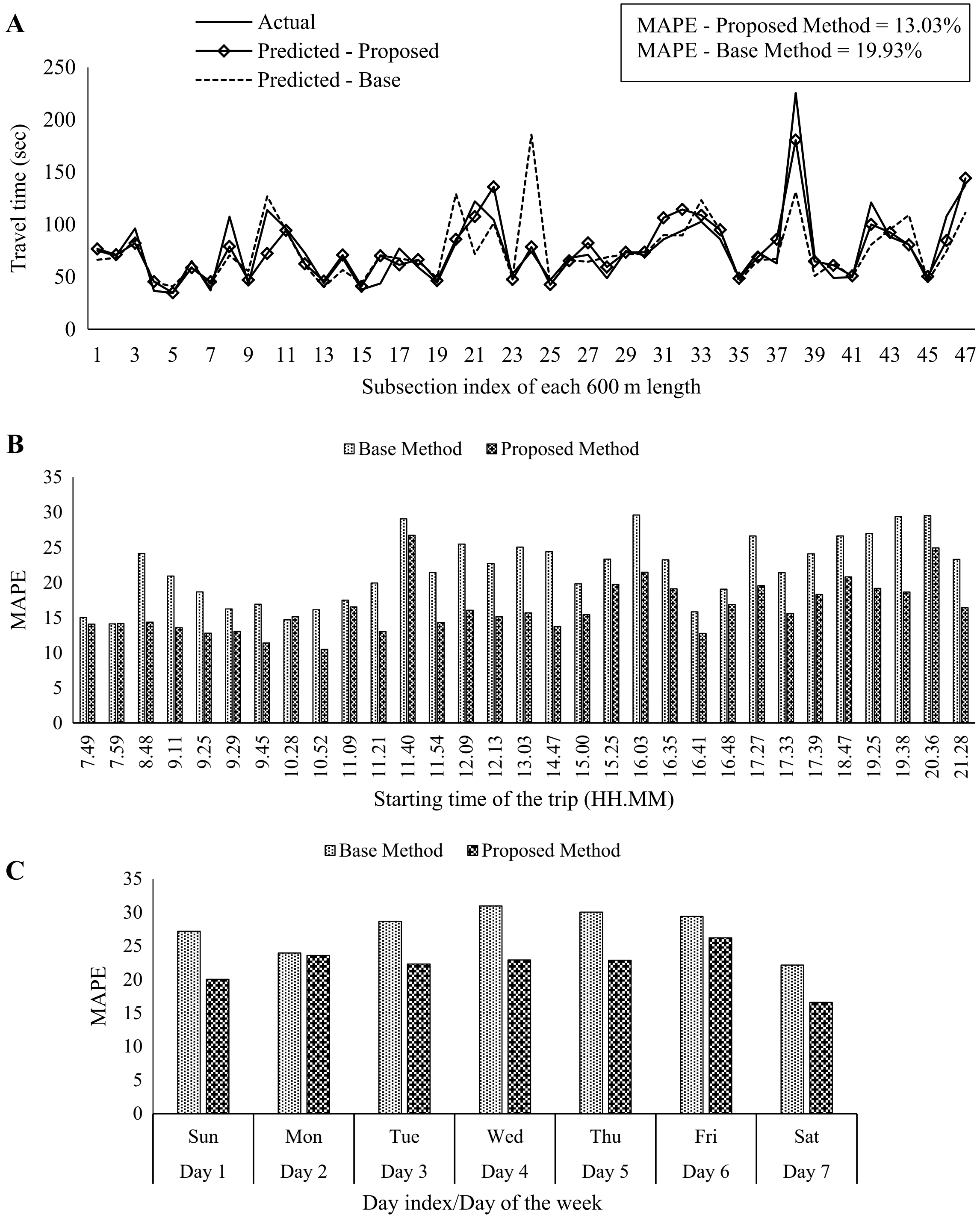


**Fig. 6** **a** Comparison of actual and predicted segment-wise travel times. **b** Performance comparison of proposed method and base KF for various tripson a sample day. **c** Performance comparison between proposed and base methods over several days

arrival information to passengers, it is also important how well in advance the prediction can be made to inform passengers. Previous studies [31–33] reported a tolerance level (deviation of the predicted arrival time from the actual arrival time) in the range of 2–5 min as acceptable to the users. Verma et al. [34] mentioned that in Indian cities, a majority of the public transit users are captive riders. In general, if the users are captive to public transport, their tolerance levels would be higher. Considering the factors such as captive riders and a lack of fixed time schedule for buses either at originating bus depots or at bus stops, a maximum value of ±5 min was

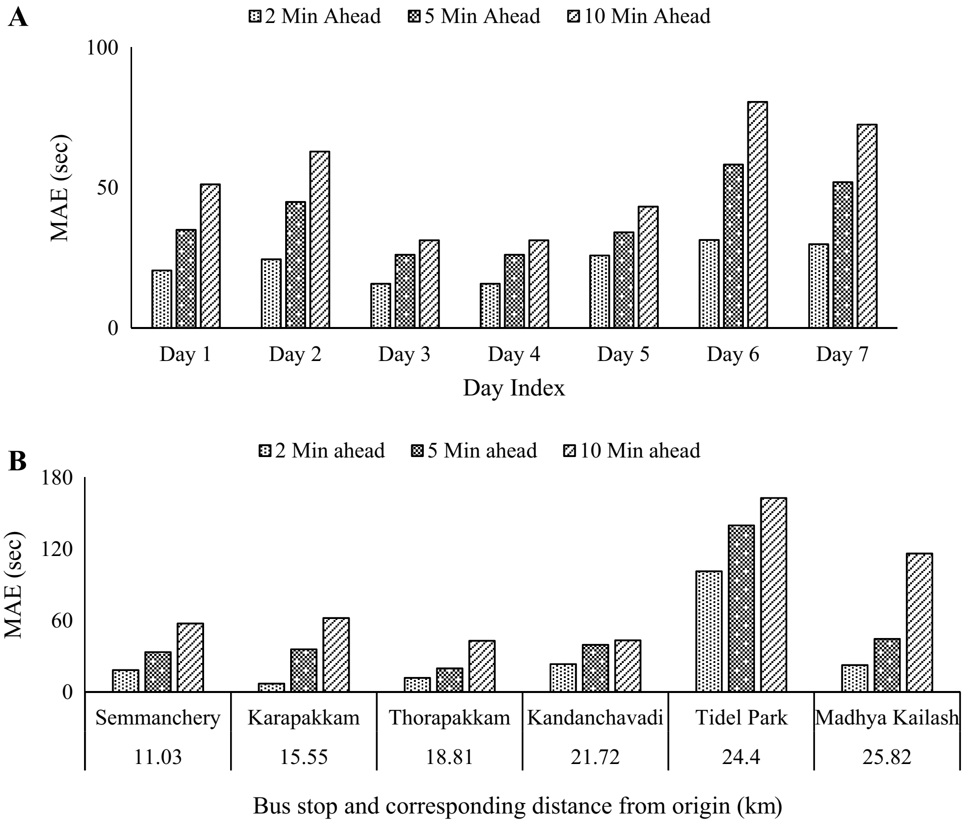


**Fig. 7** **a** MAE obtained for different prediction horizons at a selected bus stop on various days. **b** MAE obtained for different prediction horizons at different bus stops

taken as the cut-off for the tolerance level in this study. Analysis was carried out to check the prediction accuracy for varying prediction horizons, ranging from 2 to 10 min ahead. The analysis was carried out for various bus stops along the route, which are at varying distance from the origin. Since, the users feel the errors in terms of actual deviations; the errors were measured in terms of actual deviations in seconds and are shown in Fig. 7a for a selected bus stop for a test period of one week. From Fig. 7a, it can be observed that the maximum errors are around ±1.5 min for the considered bus stop up to a prediction horizon of 10 min. This can be considered as an acceptable performance based on the tolerance reported in the previous studies discussed above.

The above analysis was repeated for multiple days and multiple bus stops at varying distance from the origin and the average performance comparison is shown in Fig. 7b. From the results, it can be observed that the deviations are within the tolerance limits for all these bus stops. Thus, overall it can be seen that the proposed method with adaptive KF using suitable inputs from the cluster analysis has clear advantage over other methods. In addition, the errors in prediction for various prediction horizons and for bus stops at varying distances are evaluated and showed very good performance indicating this as a possible real time bus arrival prediction tool that can be used for field implementations.

## 8 Summary and Conclusions

Accurate bus arrival prediction becomes an extremely challenging task under the heterogeneous and lane less traffic conditions. These make the traffic more stochastic and complex and hence most of the existing solutions may not produce accurate results for the prediction of bus travel time. Hence, in the present study, a reliable framework was developed to predict bus travel time addressing some of these specific issues as detailed below.

1. The high variability in the traffic leads to lack of systematic patterns in the data and hence use of previous trips or days or weeks may not always work under such conditions. With big amount of tracking data available from buses, big data analytics tools can be adopted to identify similar trips under such conditions. The present study explored the use of one such tool namely the cluster analysis to identify the most significant inputs to be used in the prediction algorithms. This will help using the most relevant data as input for prediction, leading to better prediction accuracy. Another advantage is that, this analysis can be carried out in real time avoiding the need for any static pattern to be assumed.
2. To address the high variability in the system, an adaptive Kalman filtering approach was developed. The

proposed method was corroborated using real-world data. The results obtained were compared with a base method that used static inputs and base KF, and showed a clear improvement in performance.

3. Performance evaluations were carried out for multiple trips over multiple days as well as for multiple bus stops and for varying prediction horizons. The results obtained showed consistently good improvement in accuracy compared to the base prediction method using static inputs. The predicted values were checked for the passenger expected accuracy and were found to be within acceptable limits.

The proposed method can be implemented in real-time for advanced public transportation systems (APTS) applications on a large scale. The predicted travel times can be informed to the travelers through variable message sign (VMS) boards or kiosks at bus stops as well as through websites or mobile applications for pre-trip and en-route planning.

**Acknowledgements** The authors acknowledge the 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.

## References

1. Afandizadeh SH, Khaksar H, Kalantari N (2013) Bus fleet optimization using genetic algorithm a case study of Mashhad. IJCE Trans A Civil Eng Int J Civil Eng 11(1):43–52
2. Schweiger C (2000) Real-time bus arrival information systems. Technical report, Transportation Research Board, TCRP Synthesis 48, Washington DC
3. Zhang L (2014) Pattern-based short-term traffic forecasting for urban heterogeneous conditions. IJCE Trans A Civil Eng Int J Civil Eng 12(3):371–377
4. Kwon J, Coifman B, Bickel P (2007) Day-to-day travel-time trends and travel-time prediction from loop-detector data. In Transportation research board: Journal of the Transportation Research Board, No. 1717, Transportation Research Board, National Research Council, Washington, D.C., pp. 120–129
5. Lee WC, Si W, Chen LJ, Chen MC (2012) HTTP: a new framework for bus travel time prediction based on historical trajectories. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems
6. Jensen CS, Tie D (2008) TransDB: GPS data management with applications in collective transport. In: Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, Mobiquitous, ICST, Brussels
7. Elhanawy M, Chen H, Rakha HA (2014) Dynamic travel time prediction using genetic programming. In: Proc. of the Transportation Research Board 93rd Annual Meeting, Washington, DC
8. Min W, Wynter L (2011) Real-time road traffic prediction with spatio-temporal correlations. Transp Res Part C Emerg Technol 19:606–616
9. Patnaik J, Chein S, Bladihas A (2004) Estimation of bus arrival times using APC data. J Public Transp 7(1):1–20
10. Liu H, Zhang K, He R, Li J (2009) A neural network model for travel time prediction. In: IEEE International Conference on Intelligent Transportation Systems, Shanghai
11. Afandizadesh S, Kianffar J (2009) A hybrid neuro-genetic approach to short-term traffic volume prediction. IJCE Trans A Civil Eng Int J Civil Eng 7(1):41–48
12. Rezaie MF, Afandizadeh S, Ziyadi M (2011) Prediction of accident severity using artificial neural networks. Int J Civil Eng 9(1):41–48
13. Hinsbergen VCPJ, Van Lint JWC, Van Zuylen HJ (2009) Bayesian committee of neural networks to predict travel times with confidence intervals. Transp Res Part C Emerg Technol 17:498–509
14. Bansal P, Chen MC, Hsu CC (2015) A freeway travel time prediction and feature selection model integrating principal component analysis and neural networks. In Proc. of the transportation research board 94th annual meeting, Washington, DC
15. Fan W, Gummu ZK (2014) Dynamic travel time prediction models for buses using only GPS Data. In: Proc. of the transportation research board 93rd annual meeting, Washington, DC
16. Lin Y, Yang X, Zou N, Lei J (2013) Real-time bus arrival time prediction: an application to the case of Chinese cities. In: Proc. of the transportation research board 92nd annual meeting, Washington, DC
17. Vanajakshi L, Rilett L (2007) Support vector machine technique for the short term prediction of travel time. In: IEEE Intelligent Vehicles Symposium, Istanbul
18. Yu B, Lam WHK, Tam ML (2011) Bus arrival time prediction at bus stop with multiple routes. Transp Res Part C Emerg Technol 19:1157–1170
19. Guin A (2006) Travel time prediction using a seasonal autoregressive integrated moving average time series model. In: IEEE Intelligent Transportation Systems conference
20. Krishnan R, Polak J (2008) Short-term travel time prediction: an overview of methods and recurring themes. In: Proceedings of the Transportation Planning and Implementation Methodologies for Developing Countries Conference (TPMDC), Mumbai
21. Esawey Md, Sayed T (2011) Using buses as probes for neighbor links travel time estimation in an urban network. Transp Lett Int J Transp Res 3:279–292
22. Son B, Kim HJ, Shin CH, Lee SK (2004) Bus arrival time prediction method for ITS application. In Knowledge based intelligent information and engineering systems. Springer, Berlin Heidelberg, pp 88–94
23. Shalaby A, Farhan A (2004) Prediction models of bus arrival and departure times using AVL and APC data. J Public Trans 7(1):41–60
24. Chu L, Oh JS, Recker W (2005) Adaptive kalman filter based freeway travel time estimation. In Proceedings of transportation research board, transportation research board, National Research Council, Washington, DC
25. Vanajakshi L, Subramanian SC, Sivanandan R (2009) Travel time prediction under heterogeneous traffic conditions using GPS data from buses. IET J Intell Transp Syst 3(1):1–9
26. Padmanabhan RPS, Divakar K, Vanajakshi L, Subramanian SC (2009) Development of a real-time bus arrival prediction system for Indian traffic conditions. IET J Intell Transp Syst 4(3):189–200
27. Chamberlain RG (2014) Great circle distance between two points. http://www.movabletype.co.uk/scripts/gis-faq-5.1.html. Accessed 31 Mar 2014
28. Rees DG (2001) Essential statistics. Chapman & Hall/CRC Publishing, Inc., London

29. Yuan J, Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y (2010) T-drive: driving directions based on taxi trajectories. In: Proceedings of 18th SIGSPATIAL international conference on advances in geographic information systems
30. Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 82(1):35–45
31. Bhandari RR (2005) Bus Arrival Time Prediction using Stochastic Time Series and Markov Chains. In Ph. D dissertation, Department of Civil Engineering, New Jersey Institute of Technology. http://archives.njit.edu/vol01/etd/2000s/2005/njit-etd2005-038/njit-etd2005-038.pdf. Accessed 12 Jan 2013
32. Crout DT (2007) 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
33. Warman P (2003) Measure impacts of real-time control and information systems for bus services. Transport Direct, UK Department of Transport
34. Verma A, Sreenivasulu S, Dash N (2011) Achieving sustainable transportation system for Indian cities—problems and issues. Curr Sci 100(9):1328–1339