TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Article

Transportation Research Record
1–12

DOI: 10.1177/0361198118770175
journals.sagepub.com/home/trr
SAGE

# Performance Comparison of Bus Travel Time Prediction Models across Indian Cities

**Jairam R[1], B. Anil Kumar[2], Shriniwas S. Arkatkar[1], and Lelitha Vanajakshi[2]**

**Abstract**

Road traffic congestion has become a global worry in recent years. In many countries congestion is a major factor, causing noticeable loss to both economy and time. The rapid increase in vehicle ownership accompanied by slow growth of infrastructure has resulted in space constraints in almost all major cities in India. To mitigate this issue, authorities have shifted to more sustainable management solutions like Intelligent Transport System (ITS). Advanced Public Transportation System (APTS) is an important area in ITS which could considerably offset the growing ownership of private vehicles as public transport holds a noticeable mode share in several major cities in India. Getting access to real-time information about public transport would certainly attract more users. In this regard, this work aims at developing a reliable structure for predicting arrival/travel time of various public transport systems under heterogeneous traffic conditions existing in India. The data used for the study is collected from three cities—Surat, Mysore, and Chennai. The data is analyzed across space and time to extract patterns which are further utilized in prediction models. The models examined in this paper are *k*-NN classifier, Kalman Filter and Auto-Regressive Integrated Moving Average (ARIMA) techniques. The performance of each model is evaluated and compared to understand which methods are suitable for different cities with varying characteristics.

Mobility is the freedom and ability to travel as fast as possible from a particular origin to a desired destination. Over the last two decades, rapid urbanization and economic development has led to a noticeable increase in motorized vehicles. India, like many other countries, is suffering from this unwelcome change leading to slow moving traffic, hour-long queues and longer commutation times. This has adversely affected the mobility of commuters in the country. As more and more vehicles get added to the existing road network, traffic congestion has become a decisive factor for indexing the quality of life in a given city. In this regard, there is a need for better traffic operation of public transport, making it a prime choice for daily commuters. Public transport is important for ensuring sustainable development in the country (*1*). To attract more people toward public transport, it should provide quality service to passengers. Providing real-time bus arrival information to passengers can make bus transport user-friendly and enhance its competitiveness with other modes of transport (*2, 3*). With a schedule of predicted arrival times at each bus stop available via display boards or as mobile or web application, people can make timely plans for future trips. This needs a way to accurately predict the travel time of buses. This study is focused on this prediction application, which could attract more passengers to use public transport, which in turn can lead to less traffic congestion.

The revolution of technology has brought immense change in the quality and frequency of useful data. In the transportation engineering domain, this has resulted in several advancements, mainly under the umbrella of Intelligent Transport System (ITS). One example is city transport buses being deployed with Global Positioning System (GPS) devices, which generate real-time vehicle location data every 1–10 seconds. This massive amount of vehicle trajectory data can be used for real-time travel time prediction and fleet management. Regardless of the amount of data collected, the prediction accuracy depends significantly on the quality and significance of the input. Time-tagged location data, usually represented in the form of trajectories, bring a great potential for real-time prediction of the vehicle travel times. The travel condition of a bus may easily be affected by various internal and external factors, including accidents, weather, road construction, government policies and weather. Also, for GPS fitted vehicles, errors often exist in positional data acquisition because of interference by urban canopies and other reasons.

[1]Department of Civil Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat India
[2]Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India

**Corresponding Author:**
Address correspondence to Shriniwas S. Arkatkar: sarkatkar@ced.svnit.ac.in

In this study, the challenges mentioned above are considered, and three prediction models are proposed incorporating the vehicle-based GPS data collected from buses from three different cities in India—Surat, Mysore, and Chennai. The data collected for the study includes data from bus-only lane network of Bus Rapid Transit System (BRTS) operated by SITILINK Ltd., mixed traffic bus lane from Mysore operated by Karnataka State Road Transport Corporation (KSRTC), and from Metropolitan Transport Corporation (MTC) in Chennai.

In general, approaches on travel time estimation and prediction using vehicle-based data can be broadly classified into two main classes—data-driven methods and model-based methods. Some of the most important studies which were carried out over the past few decades in these areas are reviewed here.

The various reported techniques for bus travel time prediction include prediction using average speed techniques (*4*), step-wise linear regression techniques (*5*), time-varying coefficient (TVC) linear regression model techniques (*6*), time series analysis techniques (*7–8*) and filtering techniques (*3, 9*). With the help of faster computers and approachable coding languages, researchers have widely explored machine learning techniques in real-time arrival/travel time prediction. Among several artificial intelligence techniques, the most significant methods include Artificial Neural Network (ANN) (*10*), Support Vector Regression (SVR) (*11*), and *k*-Nearest Neighbor (*k*-NN) classifier (*12*). In most of these machine learning techniques, the model learns the travel patterns existing in the field by extracting critical parameters such as travel time, segment speed, traffic volume, and so forth from historical data.

Even though these machine learning techniques were cultured enough to understand the traffic behavior, it requires a significant amount of training time to prune the models. Thus, the application of these models to real-time prediction scenarios would be difficult. Recently, researchers have explored the concept of similarities in historical trajectories and present trajectory of a bus, for predicting travel time on a real-time basis (*13*). The studies showed that historical trajectory-based travel time prediction is highly accurate when a large amount of historical data is available. The historical data inherits the stochastic features of traffic for each hour of the day, thus enabling the model to accommodate for most possible changes in the traffic conditions.

Due to the inconsistent nature of traffic, the travel time will not be the same over different time periods, for the same network. Several studies have been carried out to describe this travel time variability with the help of several distribution functions (*14–17*).

From the literature review, most of the studies are concentrated locally on the application of each method on a particular system of public transport. The results observed are constrained to a particular set of data. Thus, the comparison of each method would not reflect its true potential. This study compares the performance of three prediction techniques using the same data from three different site conditions. Such a uniform comparison of various methods across various cities has not previously been reported for bus travel time prediction, especially under Indian traffic conditions.

## Study Area and Data

The data used in this study were collected using GPS devices fitted inside buses from the three cities. Each bus is equipped with a GPS device that records the location status of the bus along with its movement. These data are transferred to a central server every 10 seconds. Three different routes were considered for this study:

1. Bus only corridor of 9 Km in Surat including 16 stops between Udhna Dharwaja and Sachin GIDC
2. Mixed lane traffic of 7 Km in Mysore with 12 stops between City Bus Stand and J P Nagar
3. A 27 Km long route with mixed traffic and 15 stops between the source, Saidapet Bus Stand, and destination, Kelambakam Bus Stand

The routes mentioned above are shown in Figure 1. These three routes characterize most possible combinations of heterogeneous traffic and land use that would comprise a public transport corridor. Each city being different in its own demography and culture, the data collected from these would help to test the robustness of the models used for prediction. The variation in the results from each model in these routes would highlight the advantage of each model against the other. This could act as guidelines to authorities in selecting suitable prediction models when applying them in the field.

With the help of the positioning technology, the GPS device records the current status of the bus in terms of longitude, latitude and time stamp. Each bus has its own identification number. The routes traversed by the bus are recorded with a unique Trip ID. It also contains information regarding multiple trips made by that bus in that day. Each trip was named in the format *"<date>.<direction><trip start time>"*, where date corresponds to the date of data, start time is the time at which the trip is initiated, and direction implies whether the trip is onward or return. For identifying the start and end of the trip, longitude and latitude of origin and destination stops are used. When the bus reaches the destination point, velocity recorded by the device will turn to zero for a specific period, that is, for the duration in which the bus is stationary at the depot. This helps in terminating the trip while the same is cross-checked with the latitude and longitude.

Once the trip-wise data is identified, the distance between each GPS data pair is calculated using Haversine formula (*18*). The difference between consecutive time stamps gives the time difference ($\Delta t$) between each location data sent from the GPS device. Since segment-wise travel time is more

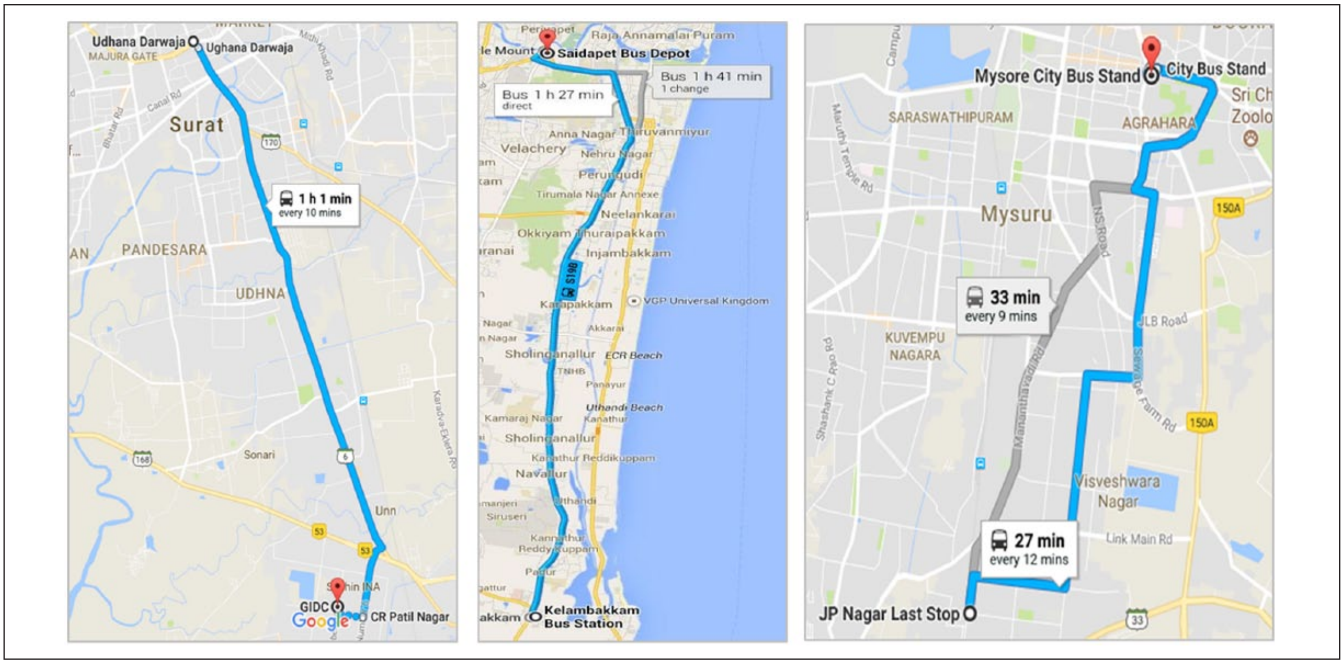


**Figure 1.** Study routes: (*a*) Surat (left), (*b*) Chennai (middle), and (*c*) Mysore (right).

practically useful to passengers than at a route level, the distances calculated between GPS time stamps are further converted to segment-wise travel time data. For a particular route, these segmental travel times were stored in a grid layout where each column represented a trip and each row represented a segment. Thus, a column consisted of the travel times on a particular trip for all the segments and a row consisted of the travel times on all the trips of the route for a particular segment. For this study, the routes were discretized into smaller segments of length 300 meters each. These segment-wise travel time data form the inputs to all the prediction methods which are detailed in the next section.

## Travel Time Pattern Analysis

Traffic in any roadway is expected to follow certain temporal and spatial patterns. Temporal patterns can include hourly pattern (peak and off-peak), daily pattern (weekday vs. weekend), weekly pattern (same days of the week having similar pattern) and so forth, whereas spatial patterns may be specific to certain sections or adjacent sections. The temporal variation in travel may be because of the change in land use during different time periods. Also, for different cities, the working hours might be different based on the population and industrial background of the city. To understand these influences in bus travel time, the data collected from three cities are studied to find the travel time variation over time. Identification of these patterns in the data will help in identifying the best inputs to be used for the prediction application.

To start with the temporal analysis, it is important to identify peak and off-peak hours of a day. For this, the travel time was plotted against each service hour of the day for each study area using box plots. Figure 2 illustrates the travel time variation over different hours of the day for Surat, Mysore and Chennai city. The peak hours from these plots were recorded as 9:00–11:00 a.m. and 6:00–8:00 p.m. as peak hours for Surat BRTS, 7:00–9:00 a.m. and 5:00–8:00 p.m. for Mysore City and 8:00–11:00 a.m. and 5:00–8:00 p.m. for Chennai City. The box plots depict the variation in the hourly travel for each route. The data from Chennai city was observed to have high variability compared with Surat BRTS and Mysore city data. The variation over the day for each city's datasets shows the challenges in predicting the travel behavior as it varies for different hours and for different cities. To understand the travel time variability, the present study explores statistical distributions.

Travel time patterns are often represented as trajectories, which give better understanding of travel time variation over space and time. Therefore, in this study, trajectory of each trip completed by the bus is plotted in a space–time field to understand the travel patterns. This helps to identify similar patterns for each case, using which temporal relations can be tested. For comparing the bus trajectories for temporal patterns, the hourly patterns (peak and off-peak hours) were plotted. These trajectory plots for peak and off-peak are shown in Figure 3 in the order Surat, Mysore and Chennai. The X-axis of the plots corresponds to each segment of the route, while Y-axis contains travel time in seconds. The plots show clear difference between peak and off-peak hour trips, with the peak hour trips occupying upper half of the spectrum. From Figure 3, it can be observed that the trips that happened in the peak hours are almost together at upper part

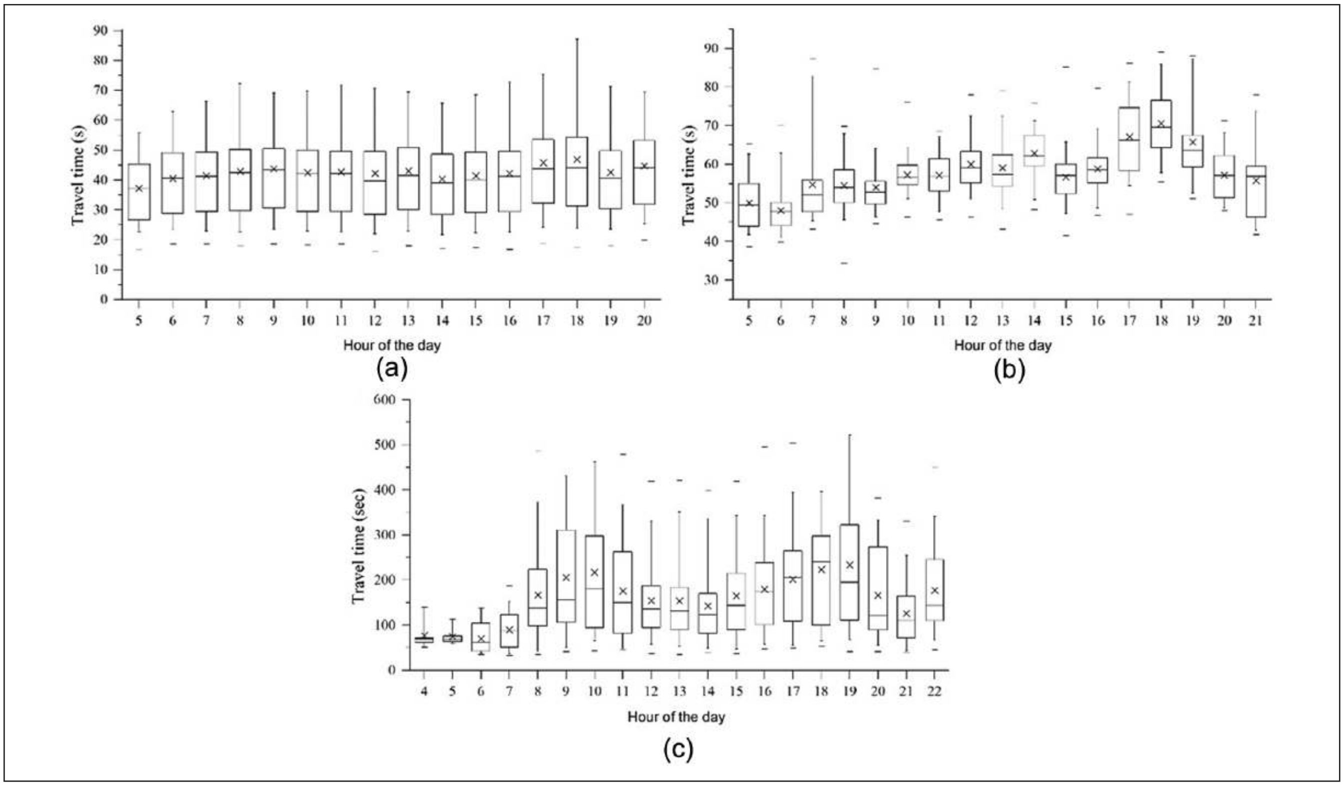


**Figure 2.** Hourly box plot for (*a*) Surat, (*b*) Mysore, and (*c*) Chennai.

of the figure (i.e., the trips are taking more time to cover the route) and the trips that happened in the off-peak hours are at lower part of the figure (i.e., these trips are taking relatively less time than peak hour trips to cover the route), showing a clear difference in variation in travel time. To confirm this, an alternative statistical analysis was also carried out by fitting the distributions to travel time data. As per the literature, travel time can be explained using three statistical distributions—Burr, Log normal and Generalized Extreme Value (GEV) (*14–17*). In this regard, distribution fitting has been carried out at three levels—peak, off-peak hour, and day of the week. GEV distribution is from a family of normal distribution and it consists of three parameters, namely shape parameter ($k$), location parameter ($s$) and size parameter ($\mu$). The parameters of GEV, particularly shape parameter, may be employed in identifying the variability of the travel time. Easy Fit Professional software (*19*) was used for distribution fitting in this study. The software runs the data through 60 probability distributions and provides the list of curves along with the best fitting curve using the Kolmogorov-Smirnov (KS) Test. Figure 4*a–c* shows the GEV plots for the peak and off-peak hour data for each city in the study.

Along with the hourly analysis, an additional analysis was carried out to identify the daily patterns in travel time data. The plots in Figure 4*d–f* show the daily pattern in travel time data. As Surat data contain BRTS data which is bus-only lane-based travel data, the variation is low, with the GEV closely spaced between each other. In Mysore, the travel time data is mixed traffic with lower vehicular population in the city, which reflects in the GEV PDF with noticeable difference in the peak of each curve. Moreover, the base of GEV which shows the variance in data for each curve is not extensively flat which shows that the data does not vary much within these hours. In the case of Chennai, the mixed traffic data and larger vehicular population in the city results in noticeable deviation between the peaks of each GEV PDF in addition to larger base.

From the preliminary analysis, the variation in travel time for each period is identified. The travel pattern varies for peak and off-peak hours and through the week it varies for each day. The plots help in identifying the dissimilarity in trajectories for each level of analysis, which is required in optimizing the historical data that is entered as input to each prediction model for easier and faster computation.

## Prediction Models

### k-*NN Classifier*

In this study, the historical trajectories are first classified based on the similarity in pattern with the input data. This was carried out using a non-parametric learning algorithm called $k$-NN classifier and the advantage with this method is that it does not make any assumptions on the underlying

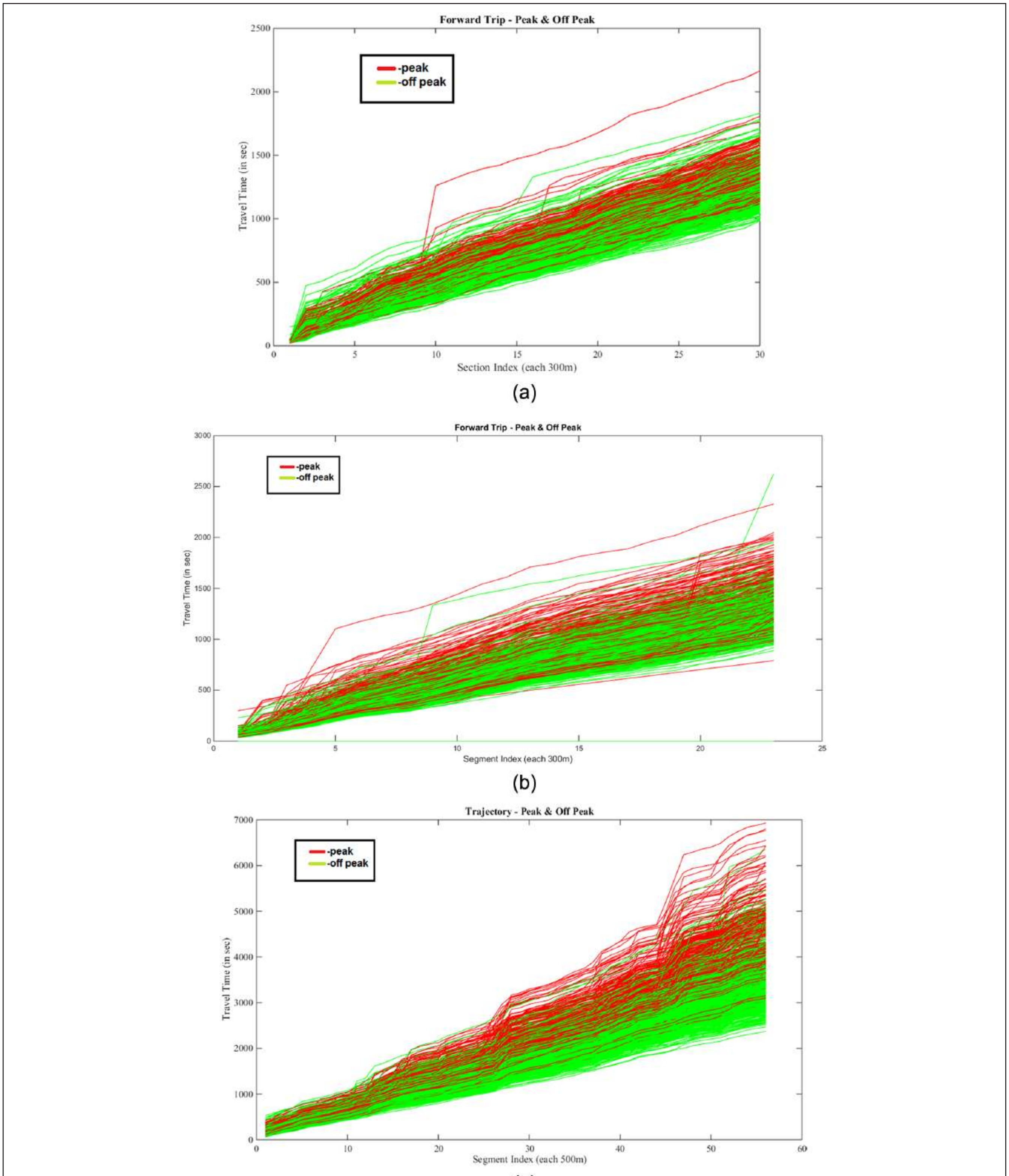


**Figure 3.** Hourly trajectory plot: (*a*) Surat, (*b*) Mysore, and (*c*) Chennai.

data, nor does it use any explicit training data to do any generalization, or it is very minimal. *k*-NN uses Euclidean distance between the input and historical data as a parameter for classifying the dataset, where *k* represents the number of neighbors. The algorithm finds the nearest neighbors based on Euclidean distance of each historical data. The class

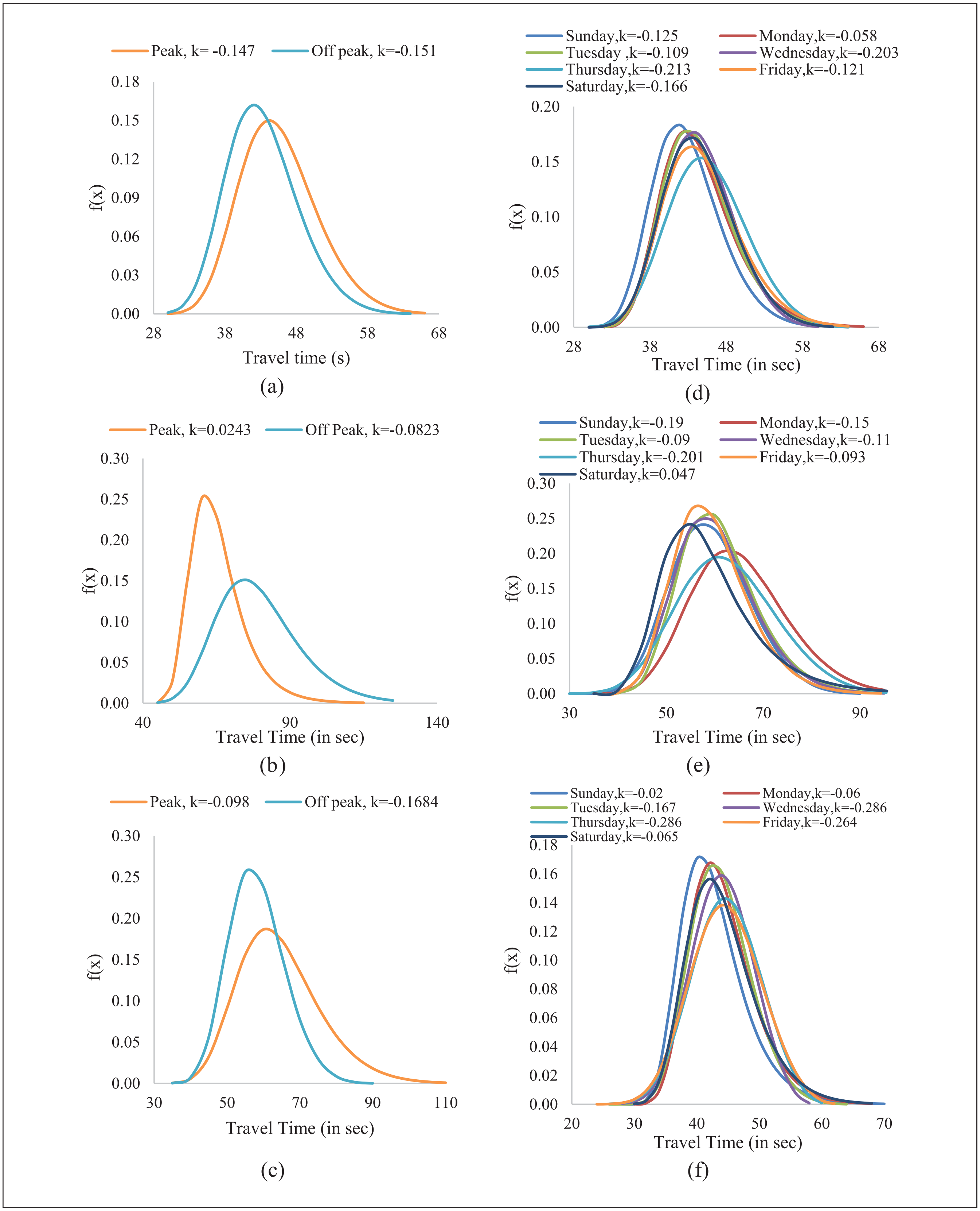


**Figure 4.** GEV PDF comparison for hour of the day in (*a*) Surat, (*b*) Mysore, and (*c*) Chennai, and day of the week in (*d*) Surat, (*e*) Mysore, and (*f*) Chennai.

**Table 1.** Sample $k$-NN Classifier Output for $n$ = 4 and $k$ = 10

| Nearest neighbor | $n$-4 | $n$-3 | $n$-2 | $n$-1 | $n$ | Euclidean distance |
|---|---|---|---|---|---|---|
| 1 | 45.26 | 63.69 | 33.37 | 70.86 | 69.90 | 26.53 |
| 2 | 30.46 | 60.47 | 36.17 | 53.07 | 56.90 | 27.77 |
| 3 | 44.85 | 56.64 | 29.07 | 59.91 | 61.00 | 28.69 |
| 4 | 38.29 | 90.31 | 29.28 | 53.08 | 82.46 | 29.03 |
| 5 | 35.78 | 73.64 | 44.85 | 56.64 | 64.60 | 29.07 |
| 6 | 28.69 | 70.66 | 38.29 | 90.31 | 87.02 | 29.28 |
| 7 | 32.36 | 57.61 | 33.27 | 70.00 | 66.32 | 29.78 |
| 8 | 35.84 | 50.35 | 34.47 | 55.11 | 51.68 | 30.46 |
| 9 | 26.53 | 46.53 | 33.06 | 55.38 | 52.46 | 30.51 |
| 10 | 29.28 | 53.08 | 29.03 | 97.64 | 87.73 | 30.80 |

which the output is closest to is determined by the majority voting on the class labels of its $k$-NNs (*20*).

The proposed prediction method being a real-time framework, the input to the classifier algorithm needs to be the patterns from the current trip. To identify such patterns, the travel time of previous '$n$' trips were taken as input. For example, if $n = 4$, for predicting the $p$th trip, the input to the algorithm will be the travel time of the ($p$-1)th, ($p$-2)th, ($p$-3)th and ($p$-4)th trips. The 'similar' trips from the historical data are then identified based on the Euclidean distance between the current trip travel time and the historical data, thus forming the cluster. Thus, there will be only one cluster formed containing '$k$' historical trajectories that are closely related or 'similar' to the given input. The average value of predicted travel times from this cluster is taken as the predicted value for the current input. The algorithm updates for each section with the new inputs of travel time for the same section from previous trips. This will accommodate for the changes in traffic in the field, making it change in real time. Sample results of $k$-NN classifier using $n = 4$ is shown in Table 1. Here, the parameters considered for the $k$-NN classifier algorithm are $l = 300$ m, $k = 10$ and $n = 4$. In this case, for predicting the $n$th trip, the previous four trip travel times are taken as input and the results for each combination are sorted with respect to the Euclidean distance, and $k$ nearest neighbors are considered for averaging.

## *Auto-Regressive Integrated Moving Average (ARIMA)*

Box and Jenkins (*21*) had carried out thorough investigation on the analysis of time series (TS) and put forward well-structured classes of models, such as Auto Regressive (AR), Moving Average (MA), Auto Regressive Moving Average (ARMA), Auto Regressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA), and so forth for modeling TS data. In this study, the travel time is represented as segment-wise travel time. Since segment-wise travel time data are not TS as they do not happen at uniform intervals, an approach was adopted to take the total travel time for individual trips. Since each trip starts with respect to certain time of day, the total travel time can be considered as TS data.

An ARIMA model consists of four stages—the model identification, model estimation, diagnostic checking and forecasting. To start with, the data is checked for stationarity as the data set used in TS analysis is assumed to be stationary. If it is not stationary, stationarity is achieved by differencing the data. Figure 5, *a* and *b* illustrate a sample raw data and the same after stationarity. Once the degree of stationarity is defined, the order of AR and MA process is identified using autocorrelation function (ACF) and partial autocorrelation function (PACF). Figure 5, *c* and *d* show the PACF and ACF plot for a sample data. Correlation values range from −1 to +1. A value of +1 indicates that the two variables move together perfectly; a value of −1 indicates that they move in opposite directions. When building a TS model, it is important to include lag values that have large, positive autocorrelation values. Sometimes it is also useful to include those that have large negative autocorrelations. Thus, from Figure 5, *c* and *d*, the order of ACF is 1 while order of PACF can be between 1 and 4.

Once the model parameters are prepared, the models are run with route-level travel time as input, which is equal to the sum of travel time of all segments of the route. The forecasting accuracy in ARIMA is better for short term prediction (*4*). Thus, forecasting is carried out for each model with few steps at a time. The predicted values then feed back into the raw data as historical data. Since this study concentrates on predicting travel time at segment level, the predicted route level travel time are then decomposed to segment data. For decomposing, certain segment-specific coefficients are used, which were established using historical database. Travel time on each segment for a trip is divided by the corresponding total travel time, thus obtaining the segment-wise coefficients for the trip. These coefficients are formulated for each hour of the day for every route considered in this study. These segment-specific coefficients take into account the route travel time variations for each hour of the day, thus accommodating the dynamic nature of traffic which varies temporally and spatially.

## *Kalman Filtering Technique*

The third prediction approach explored in this study is using Kalman Filtering Technique (KFT). KFT uses the output from $k$-NN, that is, similar trajectories as input to predict the bus travel time. The implementation of KFT requires information regarding the system's dynamics, statistical information of the system disturbances and measurement errors. The evolution of travel time between various time intervals in a given subsection is assumed to be

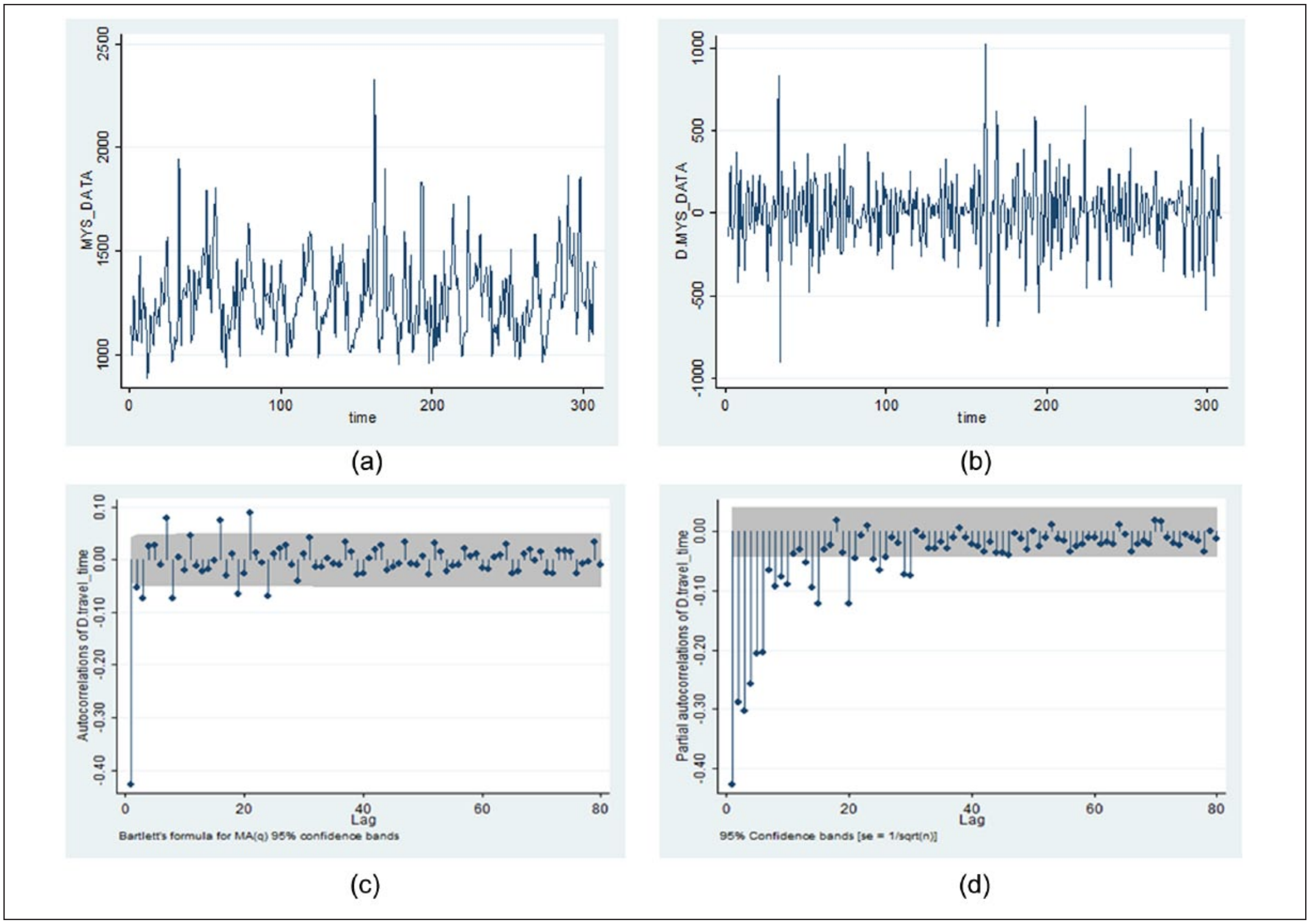


**Figure 5.** ARIMA Plots: (*a*) raw data, (*b*) stationary data, (*c*) ACF, and (*d*) PACF.

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

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

$$z(t) = x(t) + v(t), \tag{2}$$

where $z(t)$ is the measured travel time in a given subsection at time $t$ and $v(t)$ is the measurement noise. The algorithm requires two sets of data (S1 and S2) to implement the above scheme. Out of these two data sets, one set of data (S1) was used in the time update equations to calculate the parameter $a(t)$ and the other set of data (S2) was to be used in the measurement update equations to generate the *aposteriori* travel time estimate. So, the results obtained from *k*-NN classifying algorithm were arranged as two sets of data in order of preference from lower to higher (higher preference being the data that is nearer to the input pattern). The algorithm will be repeated for all N subsections. The objective here is to predict the travel time of the Test Vehicle (TV) by identifying significant travel time trajectories in a given subsection.

## Results

In this section, the performance of each model is evaluated and compared for the three cities. The prediction window is one week for each city which gives clear understanding of how the models perform for different days of the week. For a particular bus in an ongoing trip, the predicted segmented travel time ($TT_{pred}$) using prediction models are stored in the database. The observed segmented travel time ($TT_{obs}$) are collected from the historical data. For evaluating the accuracy of prediction, Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) were used as measures and are calculated as in Equations 3 and 4. The lower the errors, the more accurate are the predicted travel times and hence, the better the method.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{(\mathrm{TT_{obs}} - \mathrm{TT_{pred}})}{\mathrm{TT_{obs}}} \right| * 100, \tag{3}$$

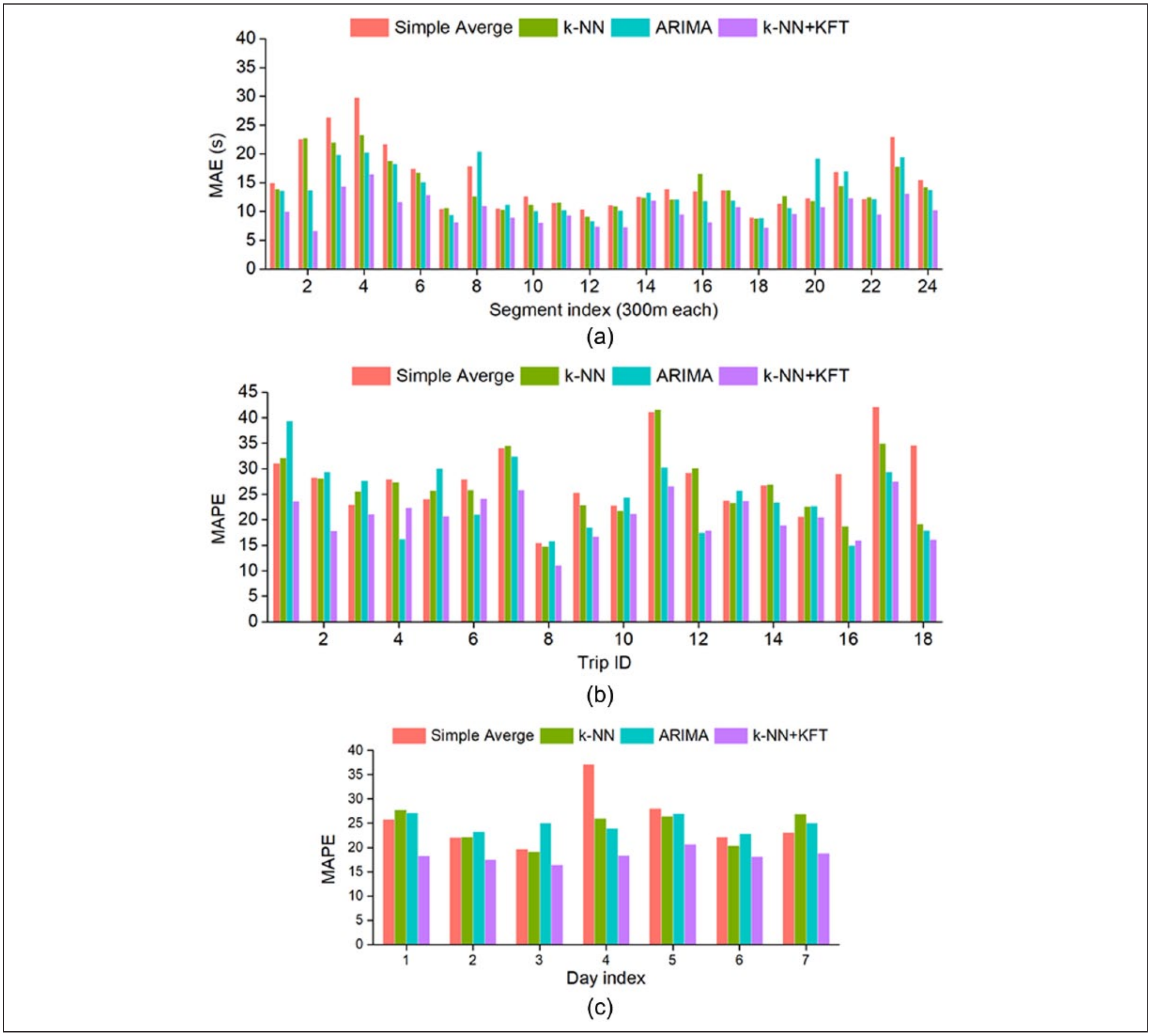


**Figure 6.** Performance comparison between models for Mysore city: (*a*) segment-wise, (*b*) trip-wise, and (*c*) day-wise.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\,\mathrm{TT_{obs}} - \mathrm{TT_{pred}}\,|, \tag{4}$$

Then, the performance evaluation was done at three levels—segment-wise, trip-wise, and day-wise. Performance of these methods was also compared with a base method, namely Simple Averaging Technique (SAT). This approach is carried out section-wise, where the average of the travel time of previous '*n*' trips was taken as the predicted travel time for that section.

The performance comparison for Mysore city is shown in Figure 6*a*–*c*. In Figure 6*a*, segment-wise average of 100 trips were taken and the *k*-NN + KFT performs relatively better than the other models with an average error of 10.21 seconds. The next most accurate model was found to be ARIMA with an average error of 13.79 seconds. The error for *k*-NN + KFT varied between 6.6 seconds and 16.48 seconds while for ARIMA it ranged between 8.4 seconds and 20.4 seconds. Figure 6*b* shows the accuracy of prediction for all trips in a sample day to understand how the models perform over a working day. Figure 6*c* shows a comprehensive summary of how the models perform over each day of the week. Here also, *k*-NN + KFT performs better than other models. It can also be seen that all the proposed methods performed better than the baseline SAT.

Figure 7*a*–*c*, show a similar comparison for Surat city BRTS corridor. Unlike in the previous case, the segment-wise comparison in Figure 7*a* shows that all proposed

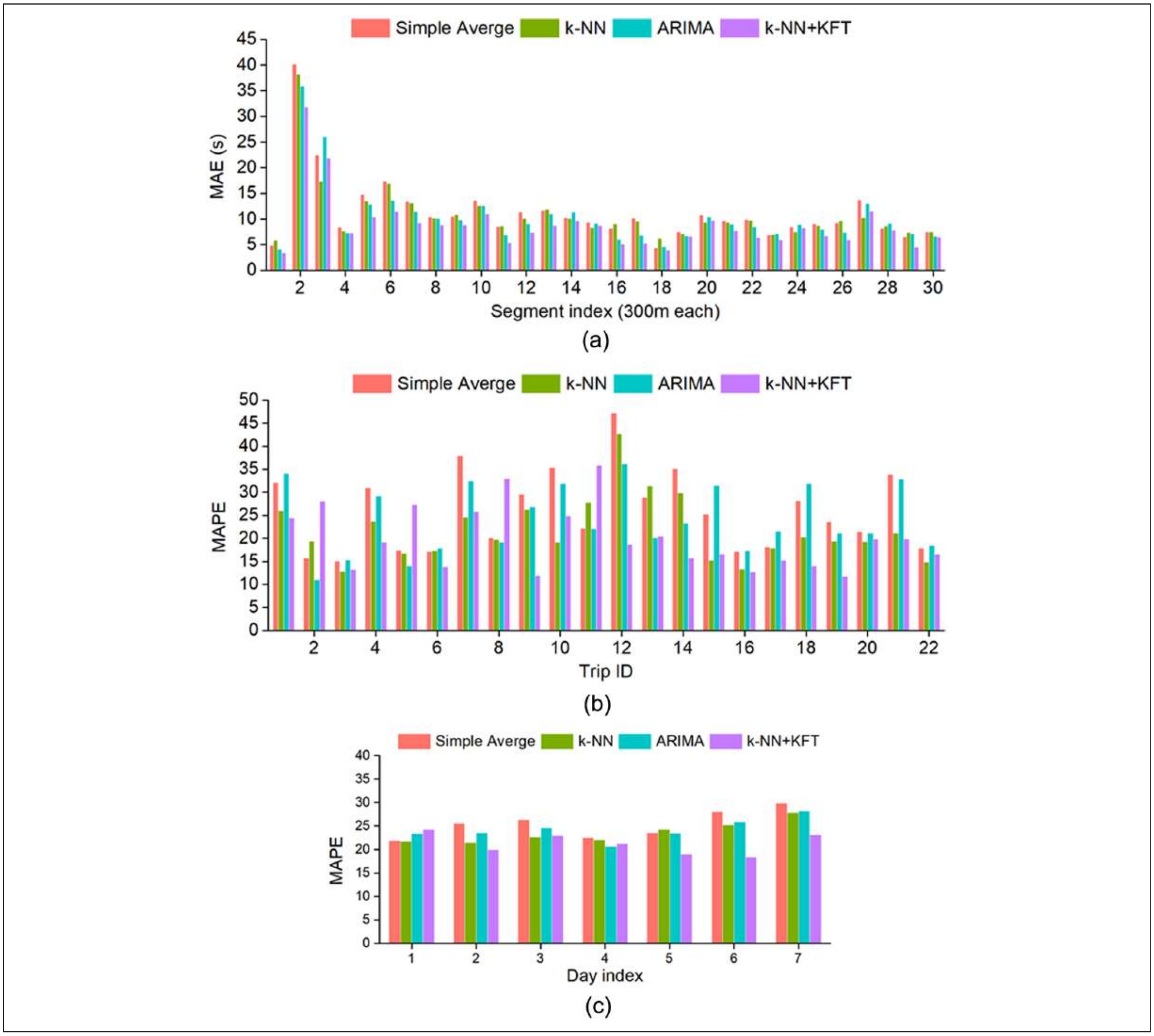


**Figure 7.** Performance comparison between models for Surat city: (*a*) segment-wise, (*b*) trip-wise, and (*c*) day-wise.

models performed comparably, and only slight improvements were observed when compared with the SAT. Segments 2 and 3 are the two high variance sections in this route, due to a major junction in the route where cross moving traffic interferes with the movement of BRTS. The performance of proposed models can be found to be much better in these sections compared with SAT. Although the segment-wise comparison shows only slight difference between the proposed models, Figure 7*c* gives better clarity on the performance through the prediction window. Except Sunday where the traffic would be generally low, the proposed models performed well for other days of the week. Thus, it can be concluded that for bus-only lane system, the sophisticated models might not be necessary. Overall, *k*-NN+KFT has the lowest error of 18.3% on Friday and the maximum error is on Sunday, with 24.21%.

In Chennai city, the performance comparison was carried out for 312 trips. The bus travel time data for Chennai were highly stochastic due to larger vehicle population observed in the metropolitan city. The heterogeneity in traffic composition was also observed to be highest in Chennai compared with other cities considered in this study. Figure 8*a*–*c* portrays the performance comparison of each prediction model carried out on Chennai city data.

In Figure 8*a*, out of 93 sections of the route, *k*-NN+KFT dominated majority of the sections with high accuracy ranging between 1.6 seconds and 58.65 seconds with a total average error of 9.13 seconds. It is followed by *k*-NN

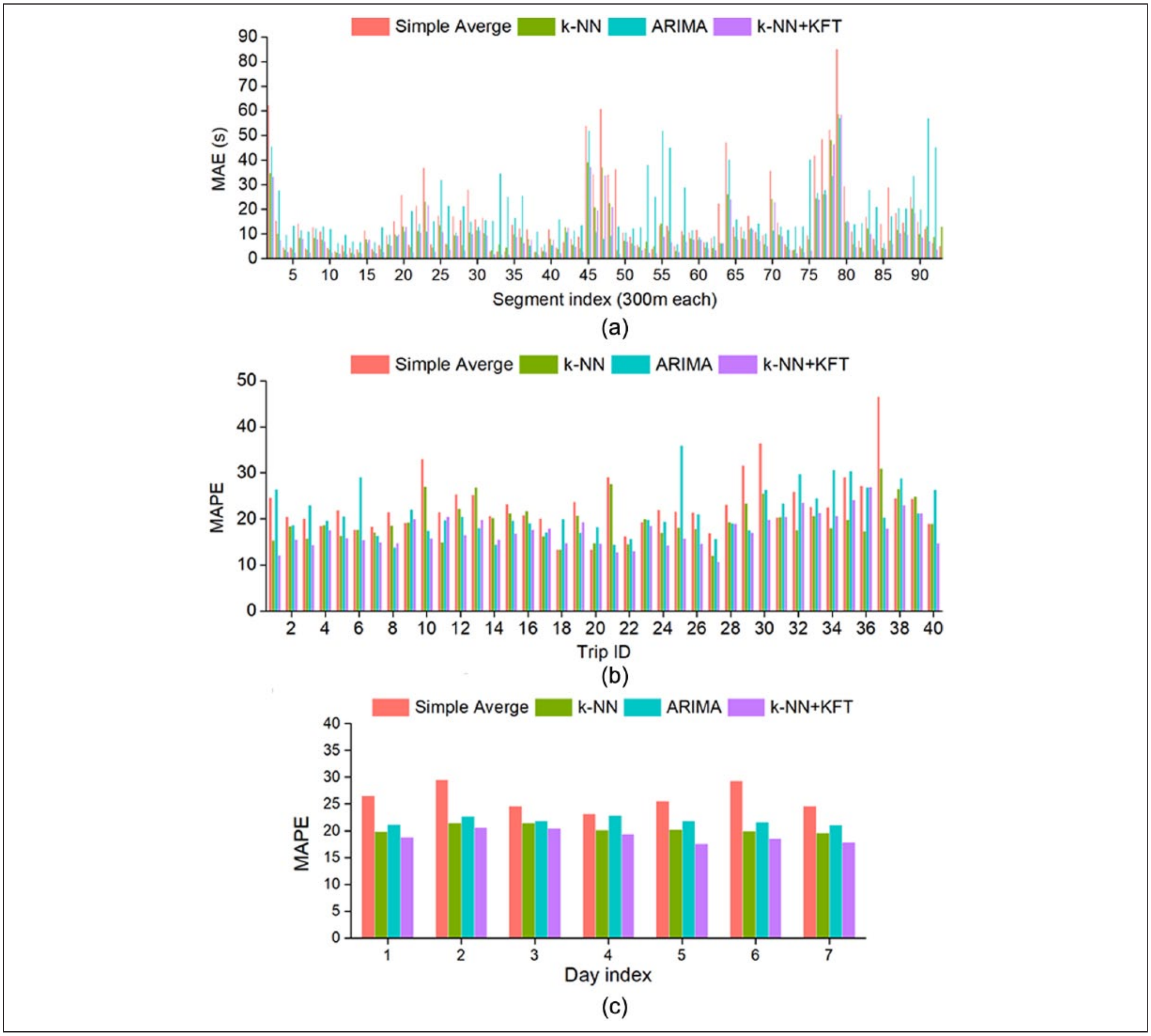


**Figure 8.** Performance comparison between models for Chennai city: (*a*) segment-wise, (*b*) trip-wise, and (*c*) day-wise.

classifier algorithm with an overall average error of 10.68 seconds, while the segment-wise error varied between 2.47 seconds and 58.85 seconds. Similar to the high variance sections of BRTS corridor, the *k*-NN+KFT model performs well in the sections 76–79, which were the high variance section in Chennai's 19B route. Figure 8*b* shows the comparison of prediction results for a sample day where *k*-NN+KFT performs better compared with other models. The day-wise comparison of prediction results can be seen in Figure 8*c*, which also shows that *k*-NN+KFT performs the best followed by *k*-NN classifier. The error for *k*-NN+KFT varies between 17.5% on Monday and 20.65% on Friday. Thus, *k*-NN+KFT has less error both spatially and temporally compared with other models proposed in the study.

## Conclusion

In this paper, the problem of real-time prediction of bus arrival/travel time using historical travel time data is studied. For implementing a prediction framework, the authorities should understand the limitations and advantages of the models and type of inputs required for the purpose. This study concentrates on showing how each model performs under different scenarios, namely bus-only lane and mixed traffic. The study explores three prediction models which used segment-wise bus travel time data as input. To start with, GPS data containing bus travel time information were collected from three cities—Surat, Mysore and Chennai. Surat has bus-only lane while Mysore and Chennai have mixed lane for buses. The data collected from these three cities were analyzed to test the robustness of the proposed prediction methods. First, a pattern analysis of bus travel time data was carried out. Analysis was done for various hours of the day and days of the week. The peak hours were identified and it was observed that peak and off-peak trips have distinctly different characteristics. The travel time data was next investigated for the statistical distributions and GEV was found to be the best fit. To compare the prediction accuracy of the models, MAPE and MAE were used as measures. The predicted values were compared with the observed travel times to understand the level of accuracy. Also, a baseline SAT was used in the comparison to identify the extent to which the models performed. For Surat BRTS data, the comparisons showed that, on the whole, all models performed similarly. Even though *k*-NN+KFT performed better in each case spatially and temporally, the difference in accuracy was small. This might be because BRTS are less influenced by external traffic except in intersections where the bus lane is exposed to mixed traffic. Since the *k*-NN, ARIMA and *k*-NN+KFT require computing power, the authorities can opt for the SAT in this case. However, in the case of mixed lane traffic, kNN+KFT performed better than the other models. Finally, the travel time information using the proposed methods can be expressed in terms of estimated arrival times for each bus and shared with the users through display monitors at bus stops, mobile application, and so forth.

## Acknowledgments

The authors acknowledge the Centre of Excellence in Urban Transport at IIT Madras; SITILINK Ltd., Surat; and Mitra ITS wing, KSRTC for the GPS data.

## References

1. Shimamoto, H., N. Murayama, A. Fujiwara, and J. Zhang. Evaluation of an Existing Bus Network Using a Transit Network Optimization Model: A Case Study of the Hiroshima City Bus Network. *Transportation*, Vol. 37, No. 5, 2010, pp. 801–823.
2. Meng, Q., and Z. Liu. Trial-and-Error Method for Congestion Pricing Scheme under Sideconstrained Probit-Based Stochastic User Equilibrium Conditions. *Transportation*, Vol. 38, No. 5, 2011, pp. 819–843.
3. Vanajakshi, L., S. C. Subramanian, and R. Sivanandan. Travel Time Prediction under Heterogeneous Traffic Conditions Using Global Positioning System Data from Buses. *IET Intelligent Transportation Systems*, Vol. 3, No. 1, 2009, pp. 1–9.
4. 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., Wireless and Mobile Networking Conference*, IEEE, Toulouse, France, 2011.
5. Kwon, J., B. Coifman, and P. Bickel. Day-to-Day Travel-Time Trends and Travel-Time Prediction from Loop-Detector Data. *Transportation Research Board: Journal of the Transportation Research Board*, 2007. 1717: 120–129.
6. Zhang, X., and J. A. Rice. Short-Term Travel Time Prediction. *Transportation Research Part C*, Vol. 11, 2003, pp. 187–210.
7. Al Deek, H. M., M. P. D'Angelo, and M. C. Wang. Travel Time Prediction with Non-Linear Time Series. *Proc., 5th International Conference on Applications of Advanced Technologies in Transportation Engineering*, Newport Beach, Calif., 1998, pp. 317–324.
8. Napiah, M., and I. Kamaruddin. ARIMA Models for Bus Travel Time Prediction. *Journal of the Institution of Engineers Malaysia*, Vol. 71, No. 2, 2010, 49–58.
9. 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. Presented at 82nd Annual Meeting of the Transportation Research Board,Washington D.C., 2003.
10. Liu, H., K. Zhang, R. He, and J. Li. A Neural Network Model for Travel Time Prediction. *Proc., IEEE International Conference on Intelligent Transportation Systems*, Shanghai, IEEE, 2009, pp. 752–756.
11. Vanajakshi, L., and L. Rilett. Support Vector Machine Technique for the Short Term Prediction of Travel Time. *Proc., IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, IEEE, 2007.
12. Robinson, S., and J. W. Polak. Modeling Urban Link Travel Time with Inductive Loop Detector Data by Using the k-NN Method. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. 1935: 47–56.
13. Lee, W. C., W. Si, L. J. Chen, and M. C. Chen. HTTP: A New Framework for Bus Travel Time Prediction Based on Historical Trajectories. *Proc., 20th International Conference on Advances in Geographic Information Systems*. California, USA. 2012.
14. Kieu, L. M., A. Bhaskar, and E. Chung. Public Transport Travel-Time Variability Definitions and Monitoring. *ASCE Journal of Transportation Engineering*, Vol. 141, No. 1, 2014, 04014068.
15. Lei, F., F. Wang, G. Lu, and J. Sun. A Travel Time Reliability Model of Urban Expressways with Varying Levels of Service. *Transportation Research Part C. Emerging Technologies*, Vol. 48, 2014, pp. 453–467.
16. Akhilesh. *Study of Travel Time Variability on Bus Routes in Mysore City*. M.Tech thesis report. Sardar Vallabhbhai National Institute of Technology, Gujarat, 2015.
17. Susilawati, M. A. P., S. Taylor, and Somenahalli, 2011. Modelling urban travel time variability with the Burr regression technique. *34th Australasian Transport Research Forum (ATRF)*, Adelaide, South Australia, Australia, 2011.
18. Chamberlain, R. G. *Great Circle Distance between Two Points*. http://www.movabletype.co.uk/scripts/gis-faq-5.1.html. Accessed February 14, 2016.
19. Easyfit 5.6. *Mathwave technologies*, 2004. http://www.mathwave.com
20. Ding, C., and X. He. K-Nearest Neighbor Consistency in Clustering. *Lawrence Berkeley National Laboratory*, Berkeley, Calif., 2004.
21. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*, 4th ed. John Wiley & Sons, Inc., Hoboken, N.J., 2013.

*The Standing Committee on Transportation in the Developing Countries (ABE90) peer-reviewed this paper (18-04897).*