Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Prediction of trends in bus travel time using spatial patterns

Hima Elsa Shaji[a], Arun K. Tangirala[b], Lelitha Vanajakshi[c]*

[a]*Graduate student, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, 600036, India.*
[b]*Professor, Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai, 600036, India.*
[c]*Associate Professor, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, 600036, India.*

Abstract

Studying patterns in traffic data is a basic analysis to understand the system. In this study, a large amount of bus travel data collected using vehicle tracking devices is analyzed for patterns. Travel time, in general, follows both spatial and temporal patterns. Spatial patterns are expected because travel times in particular sections on a roadway can be following similar patterns. For example, sections with a bus stop in it may show similar patterns due to stopping at the bus stops. The present study explores the use of data-driven approaches, primarily clustering, to identify the spatial patterns in bus travel times. Discrete Wavelet Transform (DWT) is used to extract trends from the travel time measurements. Two popular clustering algorithms - *k*-means and hierarchical clustering algorithms are used in this study to identify the spatial patterns and group sections with similar characteristics. Once the spatial patterns are obtained, the historic database is searched to identify similar cluster patterns and travel time trends are predicted using Pattern Sequence-based Forecasting (PSF) algorithm. The performance of the proposed algorithm for the prediction of travel time trends of trips occurring during peak and off-peak hours of a day was then compared based on prediction errors.

Keywords: Regressor identification; *k*-means clustering; hierarchical clustering; trend prediction; Pattern Sequence-based Forecasting;

1. Introduction

India has the second largest road network in the world with more than 200 million registered vehicles on the road as per the 2011 census of India (Chandramouli, 2012). This huge growth of vehicles has resulted in increased congestion and pollution. One possible solution to these problems can be to improve the quality of public

* Corresponding author. Tel.: +91 - 44 - 2257 4291.
E-mail address: lelitha@iitm.ac.in

transportation and attract more travelers to that. One major concern about the public transportation system is the lack of reliability and associated uncertainties. Providing real-time bus arrival information to users can be a way to address this, which requires accurate prediction of bus travel times.

Travel time is one of the most preferred traffic system performance measures, as it can be easily perceived by the users as well as by the operators and planners (Caulfied *et al.*, 2009). It is defined as the total time for a vehicle to travel from one point to another over a specified route, taking into account the stops, queues, and delays at intersections (Zhu *et al.*, 2009). The availability of accurate and reliable travel time information helps the users to select the shortest route between a given pair of origin and destination and thus optimize their trip travel times. It is also essential for the development of efficient traffic control and management strategies (Hough *et al.*, 2002; Vanajakshi *et al*., 2009; Yu *et al*., 2011).

Bus travel time prediction models reported in the literature can be broadly grouped into naïve models (Hong *et al*., 2006; Yiming *et al*., 2010), regression-based models (Bertini *et al*., 2004; Patnaik *et al*., 2004; Tétreault *et al*., 2010) , time series models (Rashidi *et al*., 2015; Williams *et al*., 2003), state space models (Wall *et al*., 1999; Vanajakshi *et al*., 2009; Kumar *et al*., 2017) and those based on machine learning approaches (Chien *et al*., 2002; Pan *et al.*, 2012; Bin *et al*., 2006; Yu *et al*., 2011). The accuracy of any of these prediction methods depends on the input variables provided to it. However, in all these studies, there lacks an automated method to analyze the data to identify the correct input, referred as regressors hereafter, to be used to estimate and/or predict bus travel times. For example, Kumar *et al.* (2013) predicted travel times by identifying significant regressors using pattern analysis by considering each day of the week separately. The patterns were identified offline and were not dynamic in nature. However, these patterns may not be static and may vary depending on the day, time, location etc. Thus, there is a need to identify correct regressors for the prediction algorithm by considering the natural groupings in the data and also the high variability in the system, ideally with an automated approach to group the travel time data in real time. Clustering is a data-driven technique which can be used to identify the natural groups within a data set. Once such groups are identified, the characteristics of the group can be used for better travel time prediction.

From the review of the literature, it was seen that few studies have been reported which took the aid of clustering in the prediction of traffic parameters. The use of partitioning clustering (Elhenawy *et al*., 2014; Ladino *et al*., 2016; Nath *et al*., 2010), hierarchical clustering (Weijermars, 2007; Chiou *et al*., 2014) and model-based clustering (Liu *et al*., 2017; Van Der Voort *et al*., 1996) have been reported in the prediction of traffic parameters. First, the respective clustering algorithm was used to group the data and then separate models were built on each of the cluster formed. However, these studies were limited to homogenous traffic conditions.

In this study, a preliminary prediction methodology was developed based on Pattern Sequence-based Forecasting (PSF) algorithm. PSF algorithm which was introduced by Alvarez et al. (2011) has two major steps: clustering the travel time data using a suitable algorithm and then, prediction of travel time from the cluster label obtained. Shaji *et al*. (2018) used the PSF algorithm to predict the trends in bus travel times using the temporal patterns in bus travel time. The clustering + prediction framework developed was compared to the case when no clustering was done on the regressor dataset. It was seen that clustering prior to prediction helped to improve the prediction accuracy. In the current study, the prediction framework is further extended by considering the spatial patterns in travel times. The particular sections on a roadway can be expected to behave in a similar pattern. For example, the sections on a roadway without a bus stop or intersection may behave in a similar pattern. And this pattern may be different when compared to sections which have either a bus stop or an intersection in it. Based on this idea, a prediction algorithm has been developed which uses the spatial patterns in travel time data to predict the bus travel time trends.

Not many studies were reported under mixed traffic conditions which explored the use of an automated and dynamic regressor selection process, such as clustering, as a pre-tool for prediction. Majority of the studies that were discussed above were carried out under homogeneous and lane disciplined traffic conditions. Under mixed traffic conditions, travel time prediction becomes more complex and difficult with the entire width of the road occupied by vehicles of varying static and dynamic characteristics. The vehicle types range from animal-drawn carts to huge trailers, moving without following any lane discipline. These characteristics make the system more dynamic with high variability making predictions under such traffic conditions more challenging. Reported studies from such traffic conditions mostly assumed static patterns in travel times, which is highly unlikely. Such manual grouping of data using chronological factors may not be effective in undermining the patterns present in the data. The time-varying nature of travel time needs to be taken into account. From these identified gaps, the following objectives were laid out for the present study:

- Identify the spatial patterns in data using clustering algorithms.
- Predict the spatial trends in travel time using Pattern Sequence-based Forecasting technique.
- Compare the performance of clustering techniques in predicting the spatial trends.

2. Data Collection

The data used for the study were collected using the GPS units fixed on the Metropolitan Transport Corporation (MTC) buses in Chennai, the capital city of the state of Tamil Nadu, India. The 19B bus route was chosen as the study stretch. The route spans a length of 29.4 km and connects Kelambakkam, a suburban area of the city to Saidapet, a major commercial area of the city. The northbound 19B route was chosen for the analysis. The GPS data were collected every 5s from the GPS units fitted in these buses. A total of 1,071 trips were collected during a period of 90 days. The obtained GPS data includes date, timestamp, latitude, and longitude of the bus location. The distance between two consecutive GPS points was calculated using the Haversine formula (Chamberlain, 2014). For the purpose of analysis, the route was divided into sub-sections each of length 200m, leading to 140 sections. The 200 m section travel times were calculated using interpolation.

3. Methodology

A preliminary prediction methodology was developed based on a Pattern Sequence-based Forecasting (PSF) algorithm. The methodology has been explained in detail in section 3.3. This algorithm involves two main steps – clustering and prediction. The trip-wise data is initially clustered and cluster labels are obtained. Further, the prediction is carried out based on the cluster labels. Figure 1 shows the prediction framework used in this study.

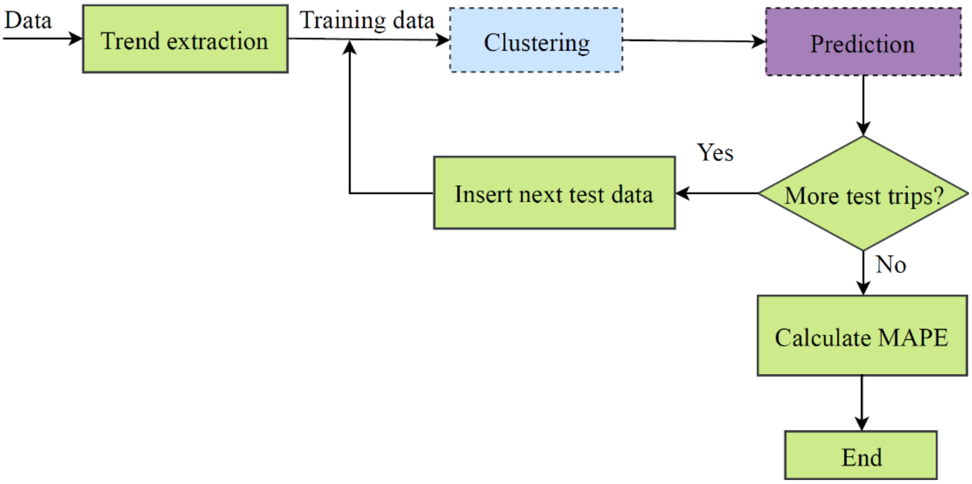

Fig. 1. Proposed methodology of PSF.

3.1. *Extraction of spatial trends in travel time*

Wavelet-based trend extraction was reported to be a more efficient trend extraction technique as they can reconstruct the signal without loss of much information (Dash *et al.*, 2004). Hence, in this study, spatial trends in bus travel time was extracted using Discrete Wavelet Transforms (DWT). For this, the trip wise travel time data is first decomposed into wavelets at different levels of resolution. The wavelet coefficients at the lower levels of resolution were then thresholded using soft policy. In the soft policy, the wavelet coefficients lesser than the threshold value are set to zero, whereas the threshold value is subtracted from those wavelet coefficients which are greater than the threshold. The levels of thresholding were chosen such that the trend extracted captures at least 70% of the signal. After thresholding, the signal is reconstructed. This reconstructed signal now contains the spatial trends extracted, which will be referred to as regressors in the study. Figure 2 shows the measured travel times and spatial travel time trends extracted using DWT for a sample trip for all the sections.

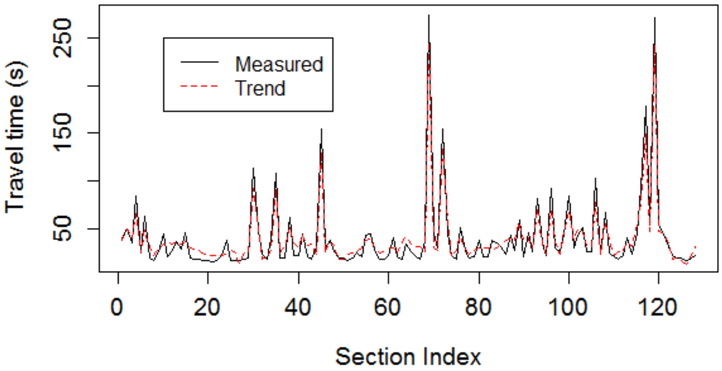

Fig. 2. Measured travel times and travel time trends for a sample trip.

3.2. *Clustering*

The next step in the proposed methodology is the identification of suitable clustering algorithms which can be used to group the regressors for the prediction algorithm. *k*-means and hierarchical clustering algorithms are two important clustering techniques used to group highly varying data. Hence, these clustering algorithms were chosen for further analysis. The spatial trends in travel time for each trip were used as inputs for the clustering algorithms. Each trip- wise travel time trend data were clustered using *k*-means and hierarchical clustering algorithms separately and were used as inputs for the prediction algorithm.

The optimum number of clusters (*k*) for each clustering algorithm needs to be found out, which depends on the clusterability of the data set. The optimum number of clusters is decided based such that objects within a cluster are similar to each other and are dissimilar to the objects in the other cluster. For both *k*-means and hierarchical clustering algorithms, the optimum number of clusters (*k*) was decided using elbow method (Thorndike, 1953). The elbow method gives you the optimum number of clusters (*k*) such that the total intra cluster variation is minimized. For each value of cluster number *K*, the sum of the squared distance between each member of the cluster and its centroid is calculated as given by equations 1 and 2. A graph is then plotted between the total within-cluster sum of squares (*W)* and the number of clusters (*K*) as shown in figure 3.

$$W_k = \sum_{r=1}^{K} \frac{1}{n_r} D_r \, , \tag{1}$$

$$D_r = \sum_{i=1}^{n_r - 1} \sum_{j=1}^{n_r} \left\| d_i - d_j \right\|_2 \quad , \tag{2}$$

where *K*- number of clusters, n_r – number of points in cluster *r* and D_r - the sum of distances between all points in a cluster. Figure 3 shows sample plots between the total within-cluster sum of squares (*W)* and number of clusters (*K*) for a sample trip using *k*-means. As the number of clusters increase, the value of *W* decreases. *k* is chosen as that value beyond which an increase in the number of clusters does not cause a significant reduction in *W*. From the figure, the optimal number of clusters (k_{kmeans}) for the sample trip was chosen as 4 for *k*-means. The same procedure was repeated to find the optimum number of clusters ($k_{hierarchical}$) in the case of hierarchical clustering.

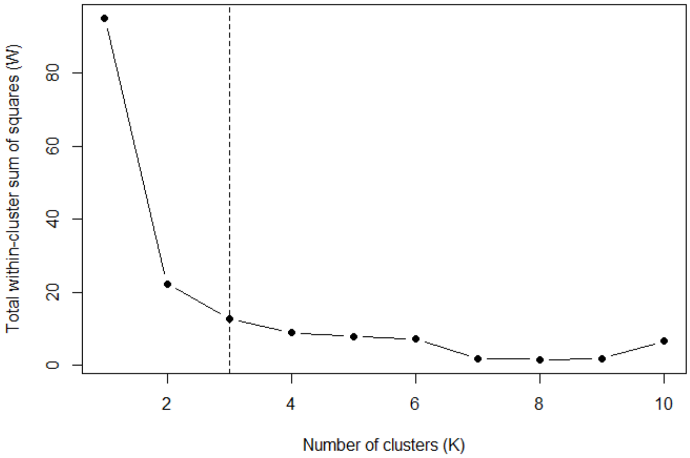

Fig. 3. Optimum number of clusters for a sample trip using *k*-means clustering.

3.3. *Pattern Sequence-based Forecasting (PSF)*

Given the travel time trends of a trip up to section *d*, our aim is to predict the travel time trend of the next section *d+1* for the same trip. Let *Y(i)* be the travel time trend of the trip for the i^{th} section and $C(i)$ Є{1,…,*k*} be the cluster label for the i^{th} section, where *k* is the number of clusters. Let S_w^i denote the sequence of cluster labels of the trip travel time trend for *W* consecutive sections, from section *i* backward as shown in equation 3.

$$S_W^i = [C_{i-W+1}, C_{i-W+2}, ..., C_i]. \tag{3}$$

Then, the prediction algorithm searches for the sequence of labels of length S_W^d in the database to obtain the subsequence set ES_{d+1} as:

$$ES_{d+1} = \{j \text{ such that } S_W^j = S_W^d\}. \tag{4}$$

In case, no matching pattern is found, the value of *W* is reduced by 1 and the pattern search is again initiated. Once the pattern is identified, the predicted value of travel time trend for the next section *d+1* is obtained by averaging the travel time trends of the members in the set ES_{d+1} as shown in equation 5.

$$\hat{Y}(d+1) = \frac{1}{size(ES_{d+1})} \sum_{j \in ES_{d+1}} Y(j+1), \tag{5}$$

where $size(ES_{d+1})$ is the number of elements in the set ES_{d+1}. The process is repeated until all the test sections are predicted.

The number of previous section labels to be considered (*W*) for PSF algorithm is another parameter that needs to be determined. The PSF algorithm searches the label sequence of *W* previous sections within the training data set for making a prediction. The value of *W* should be chosen such that the forecasting errors, as given in equation 6 are minimized.

$$\sum_{d \in ts} \left\| \hat{Y}(d+1) - Y(d+1) \right\|, \tag{6}$$

where $\hat{Y}(d+1)$ is the predicted travel time trend of the*(d+1)th* section and *Y(d+1)* is the measured travel time trend of the *(d+1)th* section and *ts* denotes the training dataset. However, to find the value of $\hat{Y}(d+1)$, the members in the set ES_{d+1} should be known, which again depends on the value of *W*. Hence the value of *W* was found using cross-validation. Cross-validation is done on a rolling basis in this study (Arlot *et al.*, 2010). The Mean Absolute Percentage Error for cross-validation (MAPE$_{cv}$) was calculated for every fold, varying *W* as shown in equation 7. The average value of MAPE$_{cv}$ (e_j) is then calculated for each window size by averaging across all the 10 folds as shown. *W* is selected as that value which minimizes the average value of MAPE$_{cv}$ as shown in equation 8.

$$e_j = \frac{1}{10} \sum_{i=1}^{10} MAPE_{cv} \{W = j\}, \tag{7}$$

$$W = \arg\min \{e_j\}, \tag{8}$$

where *j* varies from 1 to W_{max}. Figure 4 below shows the sample plot of rolling cross-validation performed for travel time trends of a sample trip using *k*-means. The values of *W* are chosen as 2 for the sample trip. The process is repeated until all the test trips are predicted.

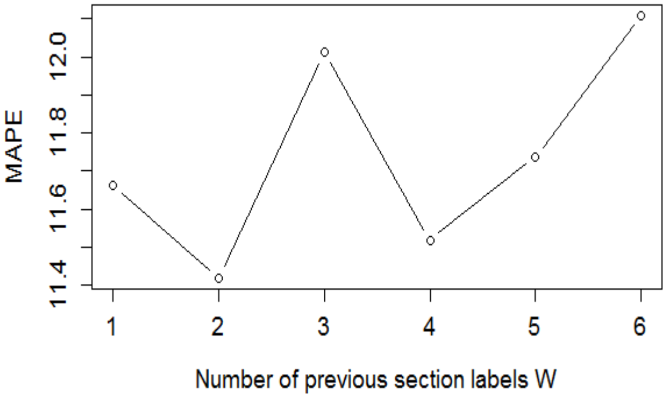

Fig. 4. Results of cross-validation for a sample trip using *k*-means clustering.

3.4 *Performance evaluation*

The PSF algorithm was implemented using the optimum number of clusters (*k*) and the number of previous trip labels (*W*) obtained. The performance accuracy of each of the clustering algorithm was quantified using Mean Absolute Percentage Error (MAPE) as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{\left|\hat{Y} - Y_m\right|}{Y_m} 100, \tag{9}$$

where $\widehat{Y}$ is the predicted value of travel time trend, Y_m is the corresponding measured travel time trend, and *n* is the number of observations in the test dataset. Figure 5(a) and 5(b) shows the measured and predicted travel time trends of a sample off-peak trip and peak trip respectively when clustering was done using *k*-means clustering algorithm prior to prediction. It can be seen that the predictions based on *k*-means clustering are able to capture the patterns in the case of both off-peak and peak trips.

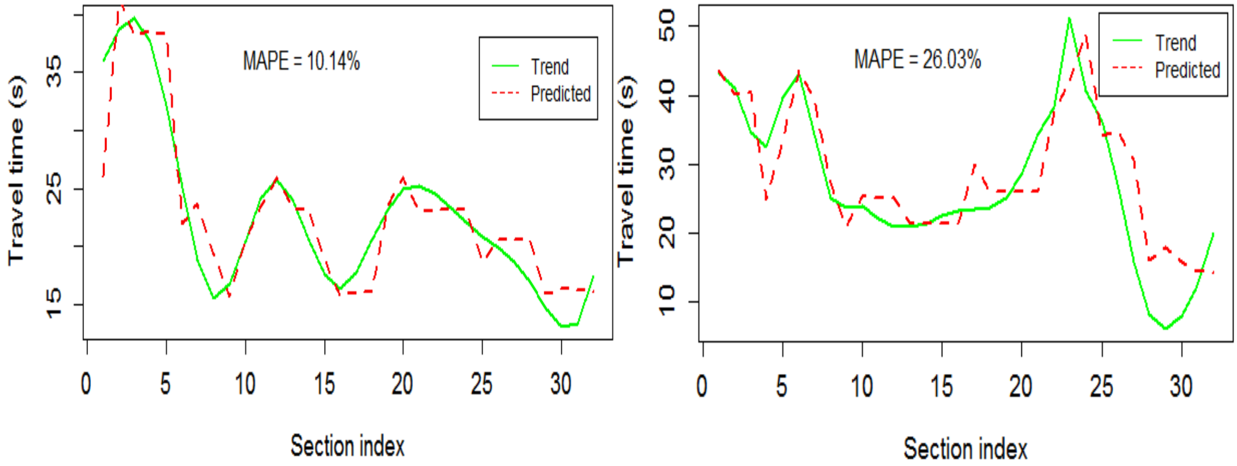

Fig. 5. Measured and predicted travel time trends of a peak trip using (a) *k*-means; (b) hierarchical.

Figure 6 shows the measured and predicted trend for sample trips which occurred during off-peak and peak time of the day when hierarchical clustering was used prior to prediction. It can be seen from the figures that in this case, predictions based on hierarchical clustering were not able to capture the patterns effectively as compared to the prediction based on *k*-means clustering. In the case of both off-peak and peak-trips, the predictions based on *k*-means clustering was able to better capture the variations in travel time trends. Also, the MAPE value was lower for predictions based on *k*-means clustering when compared to those based on hierarchical clustering. Hence, further analyses were carried out using *k*-means to group the regressor data prior to prediction.

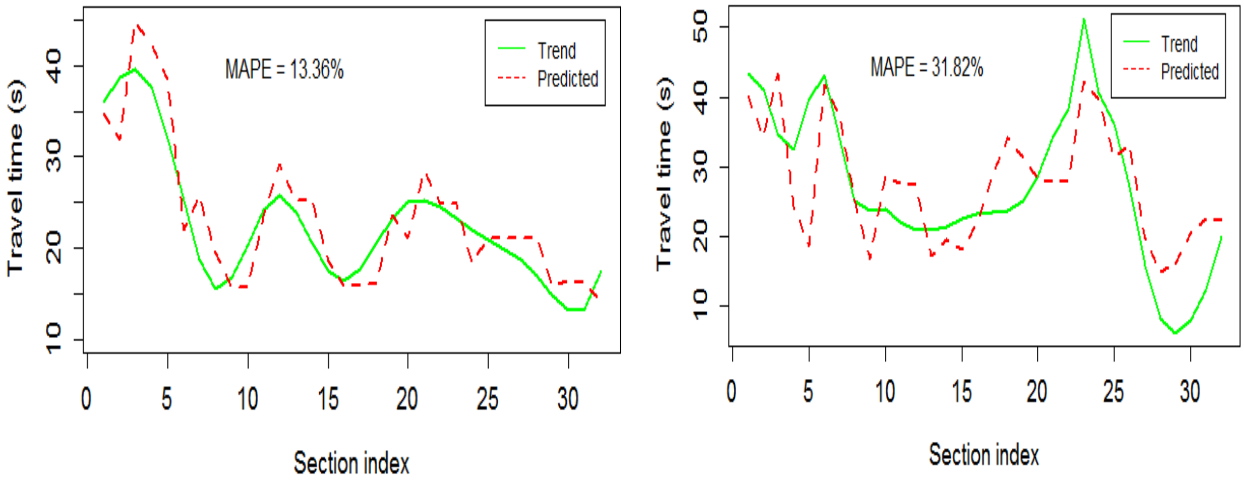

Fig. 6. Measured and predicted travel time trends of an off-peak trip using (a) *k*-means; (b) hierarchical.

From the clustering analysis, it was seen that the *k*-means clustering algorithm grouped the sections along the route mainly into 3 groups- sections with low variance in travel time, sections with medium variance in travel time and sections with a high variance in travel times. The performance of *k*-means clustering-based prediction algorithm to predict section travel time trends using spatial travel time trends was analyzed separately for these three groups. Figure 7(a), 7(b) and 7(c) show the performance of *k*-means based predictions for representative sections belonging to low variance, medium variance, and high variance travel time sections respectively. It can be seen that the MAPE value is lowest for the trend prediction in the case of low variance section, whereas the prediction errors are high for the high variance section. The performance of the developed spatial pattern based PSF algorithm remains poor for high variance section. In this case, the value of MAPE is high and also fails to capture the patterns in travel time trends of the high variance section. Hence, for high variance section, there is a need to develop more accurate prediction approaches, which can take into account the high variability associated with these high variance sections.

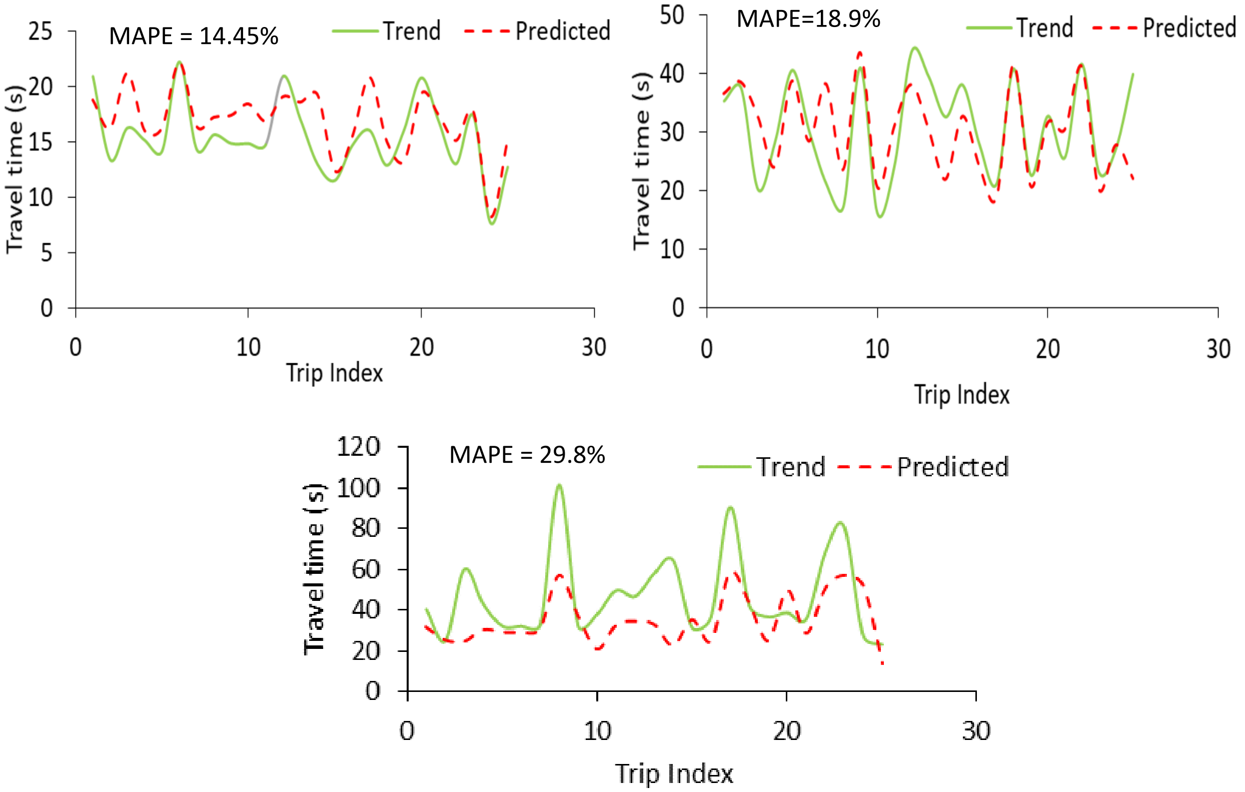

Fig. 7. Measured and predicted travel time trends using *k*-means based spatial PSF for (a) low variance section; (b) medium variance section; (c) high variance section.

4. Summary and Conclusions

The prediction of bus travel times is a widely researched area. The prediction of travel times becomes even more difficult under the mixed traffic conditions prevailing in India. Accurate prediction of bus arrival times is required to improve the quality of public transport. The proposed methodology predicts bus travel time trends using spatial patterns in bus travel time data. The following conclusions were made from the study:

- Travel time along the study stretch varies both temporally and spatially. The temporal patterns in travel time were analyzed in an earlier study. In this study, the spatial patterns in bus travel times were studied.
- A spatial PSF algorithm was developed to predict the travel time trends on sections downstream of the route by using spatial patterns of sections upstream of the route. A performance comparison was made between the

predictions based on *k*-means clustering and the predictions based on hierarchical clustering. It was seen that for both peak and off-peak trips, the predictions based on *k*-means worked better.

- The performance of the developed spatial PSF algorithm was analyzed for sections with low variance in travel times, medium variance in travel times and high variance in travel times separately. For both low and medium variance travel time sections, the predictions obtained using spatial PSF algorithm worked well. However, for high variance section, the performance of spatial PSF was not satisfactory. More efficient prediction algorithms need to be developed which takes into consideration the high variance of travel times in the high variance sections. Future extension of this work may include developing prediction algorithms by considering the spatiotemporal patterns in travel time.

5. Acknowledgment

The authors acknowledge the Centre of Excellence in Urban Transport, IIT Madras, for the data collection.

6. References

Alvarez, F. M., Troncoso, A., Riquelme, J. C., & Ruiz, J. S. A., 2011. Energy time series forecasting based on pattern sequence similarity. IEEE Transactions on Knowledge and Data Engineering, 23(8), 1230-1243.

Arlot, S., & Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

Bertini, R. L., & El-Geneidy, A. M., 2004. Modeling transit trip time using archived bus dispatch system data. *Journal of transportation engineering*, *130*(1), 56-67.

Bin, Y., Zhongzhen, Y., & Baozhen, Y., 2006. Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems, 10(4), 151-158.

Caulfield B. and O'Mahony M., 2009. A stated preference analysis of real-time public transit stop information. Public Transportation, 12 (3), 1-20.

Chamberlain, R.G. Great Circle Distance between Two Points. <http://www.movabletype .co.uk/scripts/gis-faq-5.1.html.> (Mar, 14, 2013).

Chandramouli, C., 2012. Census of India 2011-Houses Household Amenities and Assets: Transportation. Tech. rep., Government of India.

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

Chiou, Y. C., Lan, L. W., & Tseng, C. M., 2014. A novel method to predict traffic features based on rolling self-structured traffic patterns. Journal of Intelligent Transportation Systems, 18(4), 352-366.

Dash, S., Maurya, M. R., Venkatasubramanian, V., & Rengaswamy, R., 2004. A novel interval-halving framework for automated identification of process trends. AIChE journal, 50(1), 149-162.

Elhenawy, M., Chen, H., & Rakha, H. A., 2014. Dynamic travel time prediction using data clustering and genetic programming. Transportation Research Part C: Emerging Technologies, 42, 82-98.

Hong, L., Dihua, S., 2006. Arrival time prediction method of GPS-based real-time public transport vehicles. In the first session of the Intelligent Transportation and Artificial Intelligence Conference Proceedings. Guanzhou.

Hough, J. A., Bahe, C., Murphy, M. L., & Swenson, J., 2002. Intelligent transportation systems: Helping public transit support welfare to work initiatives. Res. Report, North Dakota State University.

Kumar, B. A., Vanajakshi, L., & Subramanian, C., 2013. Pattern-Based Bus Travel Time Prediction under Heterogeneous Traffic Conditions. Transportation Research Record, Transportation Research Board, National Research Council, Washington, DC.

Kumar, B. A., Vanajakshi, L., & Subramanian, S. C., 2017. Bus travel time prediction using a time-space discretization approach. Transportation Research Part C: Emerging Technologies, 79, 308-332.

Ladino, A., Kibangou, A., Fourati, H., & de Wit, C. C., 2016. Travel time forecasting from clustered time series via optimal fusion strategy. In Control Conference (ECC), 2016 European (pp. 2234-2239). IEEE.

Liu, S., McGree, J., White, G., & Dale, W, 2017. Transport mode identification by clustering travel time data. ANZIAM Journal, 56, 95-116.

Nath, R. P. D., Lee, H. J., Chowdhury, N. K., & Chang, J. W., 2010. Modified K-means clustering for travel time prediction based on historical traffic data. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 511-521). Springer, Berlin, Heidelberg.

Pan, J., Dai, X., Xu, X., & Li, Y., 2012. A Self-learning algorithm for predicting bus arrival time based on historical data model. In Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on(Vol. 3, pp. 1112-1116). IEEE.

Patnaik, J., Chien, S., & Bladikas, A., 2004. Estimation of bus arrival times using APC data. Journal of public transportation, 7(1), 1.

Rashidi, S., & Ranjitkar, P., 2015. Estimation of bus dwell time using univariate time series models. Journal of Advanced Transportation, 49(1), 139-152.

Shaji, H. E., Tangirala, A. K., & Vanajakshi, L., 2018. Evaluation of Clustering Algorithms for the Prediction of Trends in Bus Travel Time, Paper presented at TRB 97th Annual Meeting, Washington DC, January 2018.

Tétreault, P. R., & El-Geneidy, A. M., 2010. Estimating bus run times for new limited-stop service using archived AVL and APC data. Transportation Research Part A: Policy and Practice, 44(6), 390-402.

Thorndike, R. L., 1953. Who belongs in the family?. Psychometrika, 18(4), 267-276.

Vanajakshi, L., Subramanian, S. C., & Sivanandan, R., 2009. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. IET intelligent transport systems, 3(1), 1-9.

Van Der Voort, M., Dougherty, M., & Watson, S., 1996. Combining Kohonen maps with ARIMA time series models to forecast traffic flow. Transportation Research Part C: Emerging Technologies, 4(5), 307-318.

Wall, Z. and Dailey, D. J., 1999. An algorithm for predicting the arrival time of mass transit vehicles using automatic vehicle location data, Paper presented at TRB 78th Annual Meeting, Washington DC, January 1999.

Weijermars, W. A. M., 2007. Analysis of urban traffic patterns using clustering. University of Twente.

Williams, B. M., & Hoel, L. A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of transportation engineering, 129(6), 664-672.

Yiming, L., 2010. Research on Location Technology and Arrival Time Prediction Algorithm for Public Transport Vehicles Based on CAN bus[D]. Tian Jin Electronic Information engineering College.

Yu, B., Lam, W. H., & Tam, M. L., 2011. Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C: Emerging Technologies, 19(6), 1157-1170.

Zhu, T., Kong, X., & Lv, W., 2009. Large-scale travel time prediction for urban arterial roads based on Kalman filter. In Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on (pp. 1-5). IEEE.