Performance Comparison of Data Driven and Less Data Demanding Techniques for Bus Travel Time Prediction

B. Anil Kumar[1], Vivek Kumar[2], Lelitha Vanajakshi[3*], Shankar C. Subramanian[4]

[1]Graduate Student, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, 600036, India
[2]Former Project Staff, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, INDIA, E-mail: viks.sai@gmail.com
[3]Associate Professor, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, 600036, India
[4]Associate Professor, Department of Engineering Design, Indian Institute of Technology Madras, Chennai, 600036, India

Abstract

Accurate travel time information of public transport will help operators to effectively manage and implement their operating strategies and help passengers by reducing the uncertainty about arrival time of buses at bus stops. The reliability of such information provided to passengers greatly depends on the prediction technique used, which in turn, depends on the quality of the input used in the prediction technique. In other words, identifying and using the correct input in the appropriate prediction technique is important. Prediction techniques can be data driven or less data intensive. The first part of this paper presents a systematic statistical approach for identifying the significant inputs for travel time prediction. The second part compares the performance of two popular prediction methods, one being the data driven Artificial Neural Network (ANN) method and the other being a model based approach using the Kalman Filter Technique (KFT) that is less data intensive, to predict bus travel time. The performances of both methods were evaluated using the data obtained from the field. It was found that ANN outperformed KFT in terms of prediction error, if a good database is available, and in case of limited data availability, KFT will be more advantageous.

*Keywords:*Bus travel time prediction; significant inputs; statistical analysis; Artificial neural network; Kalman filtering technique

1. Introduction

India is going through rapid changes in many fields including transportation that has led to several negative impacts such as congestion, pollution and delays. In order to reduce the ever increasing congestion problem in urban areas around the world, traffic

[*] Corresponding author: Lelitha Vanajakshi (lelitha@iitm.ac.in)

officials and traffic agencies are looking at Intelligent Transportation Systems (ITS) as one option. A significant growth in interest on Advanced Public Transportation Systems (APTS), a major functional area under ITS, has been witnessed in recent years. The main aim of APTS is to provide reliable public transport service in order to attract private vehicle travellers. A number of different preferential methods were introduced in many cities around the world to realize better service in public buses. These may include bus gate, bus lanes, bus priority signals, Bus Arrival Time Prediction (BATP) Systems, etc. Out of these, some methods need infrastructure expansion such as providing extra lanes for buses, where as other methods involve only a change in the operation and management strategies to increase the reliability of bus transit. One such example is the BATP system, which is superior when compared to other methods in terms of financial viewpoint and operational efficiency. BATP systems aim to predict bus arrival times at various bus stops and provide the same to passengers in real time. However, the information provided to passengers should be accurate, otherwise passengers may reject the system (Schweiger, 2003). This accuracy of information provided to passengers, greatly depends on the prediction technique used and the input data used for the same, and these issues form the focus of this study

Prediction techniques used for bus travel time/arrival time can be broadly classified into historical, data driven and model based techniques. Data-driven and historical approaches require a good amount of database, whereas model-based approaches have an advantage of requiring lesser data, making them suitable for real time implementation. In addition, irrespective of the amount of data required, one should use the best inputs to obtain better prediction accuracy. Thus, identifying the most significant inputs and incorporating the same in the prediction method will hopefully improve the prediction accuracy.

Machine learning techniques such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are commonly used to predict travel time because of their ability to solve complex non-linear relationships. ANN has proved to be one of the most effective tools for pattern recognition across different areas. Based on this, ANN was used as a prediction tool in this study. On the other hand, model-based techniques use models that can capture the dynamics of the system by establishing explicit mathematical relationships between variables. These along with a suitable estimation scheme such as Kalman Filtering Technique (KFT) are suitable for real time implementation and hence were chosen as another possible class of prediction methods. A comparison of these two methods was carried out in this study.

The most significant data that need to be used were identified by carrying out a pattern analysis of the data using a systematic statistical approach. The most significant inputs, thus identified, were used for the prediction techniques. The study used a spatial model in the sense that the evolution of travel time over space was characterized and analysed. Hence, a particular section of the road was divided into different subsections. In the ANN based approach, the network was trained separately for each subsection. In the model-based approach, the equations that characterize the evolution of travel time over space were used.

Thus, the objective of the present study was bus travel time prediction (BTTP) under Indian traffic conditions by using ANN and model-based approach using KFT by providing appropriate inputs to the models. The study compared the performance of these two approaches for bus travel time/arrival time prediction.

2. Literature Review

Since the present study focuses on machine-learning and model-based approaches to predict bus travel time, a few important related studies are discussed here. As a prominent approach for solving complex problems, ANNs have been gaining popularity in transportation. The ANN models identify the relationship between a dependant variable and a set of input variables by adjusting their parameters (Hagan, 1996). Previous studies demonstrated that ANNs have the potential to predict the traffic parameters such as traffic volumes, speeds and travel times on freeways and urban roads. Kalaputapu and Demetsky (1995) developed ANNs by using Automatic Vehicle Location (AVL) data to predict deviations in bus schedule. Based on the posted schedule and the output from ANNs, bus arrival time can be estimated. Chien *et al.*(2002) developed two ANN methods to predict bus travel time, based on link-based and path-based data. Park *et al.*(2002) predicted the same vehicle's travel time in the immediate future by developing a spectral basis artificial neural network (SNN). Jeong and Rilett (2004) developed a bus travel time prediction method based on ANN by taking arrival time, dwell time and schedule adherence as variables. Chen *et al.*(2004) developed a methodology to predict bus travel time using Automatic Passenger Counter (APC) data. The model consisted of two components, first an ANN model was trained with four input variables, day-of-week, time-of-day, weather, and segment characteristics to produce a base estimate of the travel time. Then, the dynamic algorithm was combined with the most recent information on bus location to predict arrival times to the next time intervals. Bin *et al.*(2006) developed an algorithm based on SVM to predict bus travel time by considering time-of-day and weather conditions as variables. They predicted the travel time based on the travel time of a current segment and the latest travel time in the next segment. Wang *et al.*(2009) used Support Vector Regression (SVR) taking departure time, bus travel time, parameters of traffic conditions along the bus route, and route specific parameters as variables to predict bus travel time. Yu *et al.* (2010) proposed a method based on SVM using the speeds of several previous buses as inputs. The results showed that the SVM outperformed the performance of historical average and ANN methods. Pan *et al.*(2012) proposed a self-learning prediction algorithm to predict bus travel time by using historic data. The historic data were trained by the back propagation neural network to predict average speed and travel time on links. All the studies discussed above dealt with homogeneous traffic conditions. Only a limited number of studies were reported using machine learning techniques for bus travel time prediction under heterogeneous traffic conditions. Ramakrishna *et al.*(2006) developed an ANN method and Multiple Linear Regression (MLR) method to predict bus travel time using limited Global Positioning Systems (GPS) based data. These models were applied to a case study bus route in the city of Chennai, India. They concluded that the ANN method was working better than the MLR method.

Considering model based approaches, Wall and Dailey (2009) developed an algorithm to predict bus travel time using the Kalman filter to track vehicle location and statistical estimation to predict bus travel time. Chen and Chien (2001) developed an algorithm to compare link-based and path-based travel time prediction methods using the Kalman filtering algorithm. They concluded that path-based prediction methods have better accuracy than link-based methods under normal traffic conditions. Chien and Kuchipudi (2003) also developed link-based and path-based travel time prediction methods to predict bus travel time. Their conclusion agreed with the earlier study that the path-based method worked better than the link-based prediction method. Cathey and Dailey (2003) used bus trips data collected on different days at same time of the day as inputs by using KFT that. The study compared the results with historical, regression and ANN models. Nanthawichit *et al.*(2003) used the KFT to estimate traffic parameters by integrating the data obtained from GPS equipped vehicles and loop detectors. Their study reported that the proposed method performed relatively better than other methods such as historical, regression and ANN methods. Shalaby and Farhan (2004) used a combination of Automatic Vehicle Location (AVL) and Automatic Passenger Counter (APC) data to predict bus travel time by using KFT. They developed methods to predict running time and dwell time of buses separately. Son *et al.* (2004) developed a method using KFT to predict travel time from bus stop to stop line at signalized intersections. Chu *et al.* (2005) developed a method to estimate section travel time by applying adaptive KFT. They integrated the data obtained from loop detectors and probe vehicles. They showed that the proposed algorithm performed better than the case where a single data source was used. Yang (2005) used the data obtained from GPS equipped vehicles to predict travel time by using the discrete KFT. Yu *et al.* (2011) developed several methods such as SVM, ANN, k-nearest neighbour algorithm and linear regression to predict bus travel time based on running time of multiple routes. Zhu *et al.*(2011) developed a method to predict bus travel time by considering bus stop delays and signal intersection delays associated with total travel times. Such studies reported from Indian traffic conditions are discussed next. Kumar and Vanajakshi (2014) developed a method to estimate the stream travel time from the using buses as probes.

There have been a few studies that had applied KFT for bus travel time prediction. Vanajakshi *et al.*(2009) proposed a model based method using space discretization approach to predict bus travel time. In space discretization, the route was spatially discretized into smaller subsections to predict travel time in the upcoming subsections by using previous buses data. The basic assumption in that approach is that the trip wise data were good enough for prediction and the model hypothesized a relation in travel time between neighbouring subsections. Padmanabhan *et al.*(2009) extended the above study by explicitly incorporating the dwell times into the model. However, due to constraints in data collection, the above studies considered just previous two buses data as inputs without considering patterns in travel time. Identifying the most significant trips and incorporating them in the analysis will definitely help in improving the prediction accuracy. Kumar and Vanajakshi (2012) analysed the weekly patterns and trip wise patterns in bus travel time data and reported a strong weekly pattern followed by a trip-wise pattern. However, they assumed the same travel time patterns for all days of the week, which may not be true.

From the above review, it can be observed that the input data for the prediction methods were taken without much analysis in most of the studies. It was observed that none of the studies analyzed the travel time pattern of all days of the week separately and under varying traffic conditions. It is important to do so, since patterns are not likely to be the same for all days of the week and under varying traffic conditions. For example, the travel time on weekends may follow a different pattern when compared to weekdays. The present study followed the methodology proposed by Kumar and Vanajakshi (2012) to identify the travel time data most suited as inputs and use them to develop an ANN model to predict bus arrival time. The performance of such a data driven technique is compared with a model based approach with lower data requirement. Also, the present study carried out a sensitivity analysis by training the ANN model with different amounts of data to identify when the data driven approaches would be beneficial over model based approaches

3. Data Collection and Analysis

GPS units are commonly used to collect data for real-time APTS applications. GPS can track vehicles continuously to provide their location details at desired time intervals. In the present study, data were collected from permanently fixed GPS units in Metropolitan Transport Corporation (MTC) buses in the city of Chennai, India. The route selected for the purpose of collecting data in the present study is 19B has a route length of around 30 km, which connects the Kelambakkam bus depot in the suburban part of the city to the Saidapet bus depot in the central part of the city. There are 20 bus stops and 13 signalised intersections in this route. The average time headway between two buses in this route was around 30 minutes. Figure 1 illustrates the study route with bus stop details and distances between the major bus stops are tabulated in Table 1. The selected route consists of urban roads with varying volumes, road width, traffic and land use characteristics such as residential commercial and institutional areas.

Table 1: Bus Stop Details.

S. No	Bus Stop Name	Distance between bus stops (in km)	Cumulative distance from the initial bus stop (in km)
1	Kelambakkam	0.00	0.00
2	Hindusthan Engineering College	2.51	2.51
3	SIPCOT	3.40	5.91
4	Navallur	1.61	7.52
5	Navalur Church	2.50	10.02
6	Semmaancheri	1.01	11.03
7	Kumaran Nagar	1.28	12.31
8	Shozhinganallur P.O. Office	1.43	13.74
9	Karapakkam	1.81	15.55
10	TCS	0.41	15.96
11	Mootachavadi	1.46	17.42
12	Mettupakkam	0.79	18.21
13	Thorapakkam	0.60	18.81
14	Tirumailai Nagar	1.25	20.06
15	Kanadachavadi	1.66	21.72
16	Lattice Bridge	1.73	23.45
17	WomensPoytechnic College	1.36	24.80
18	Madhya Kailash	1.01	25.82
19	Engineering College	0.82	26.64
20	Saidapet	3.30	29.94

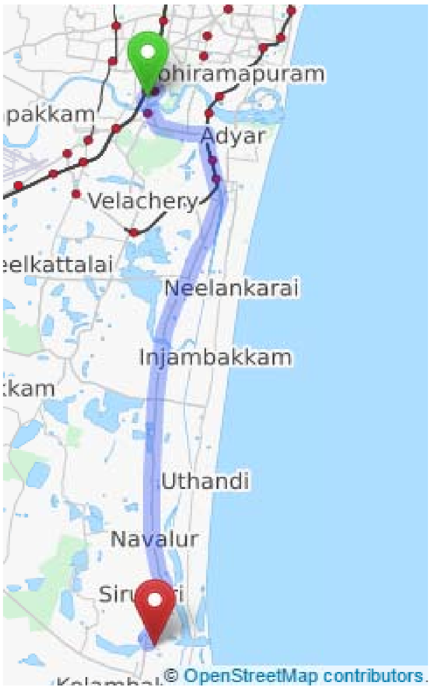

Figure: 1 19B Route.
Source: Open Street Maps.

The GPS data were collected every 5 seconds from 6 AM to 10 PM in the selected route. A total of 975 trips data were collected for a test period of 45 days. The collected GPS data included the ID of the GPS unit, time stamp and latitude and longitude at which the entry was made. The real-time communication of this data was made possible through General Packet Radio Service (GPRS) and the data received from each device were stored under a different file name. The raw data consisted of the location details of the bus for the entire duration for which it was being monitored. In the next step, the distance between two consecutive entries was calculated using the Haversine formula (Chamberlain, 2015), which gives the great circle distances between two points in a sphere from their location information. Thus, the processed data consisted of the travel times and the corresponding distance between consecutive locations for all the locations. The entire route was divided into 150 subsections of 100m length and the corresponding time taken to cover each subsection was calculated using the linear interpolation technique. The outliers in the data were removed by keeping the lower bound permissible value as 5th percentile travel time, and the higher bound permissible time as 95th percentile travel time for each 100 m subsection.

4. Methodology

The present study conducted a travel time pattern analysis by adopting the methodology from Kumar and Vanajakshi (2012). The analysis was carried out for each day of the week separately to identify the significant inputs to be used in the prediction. In the next step, travel time prediction methods that used these identified inputs were developed based on ANN and KFT. Performance comparison of ANN and KFT for the short-term prediction of bus travel time was also carried out.

4.1 Travel Time Pattern Analysis

To identify the relevant inputs, a pattern analysis was carried out by performing the Z-test for the mean of population of differences for paired samples at 5% level of significance. Out of the collected data, 28 days' data were used as input dataset for this analysis. The remaining data were used for the evaluation of the identified patterns and travel time prediction. The test compared each 100 m subsections' travel time of the output trip to the input trip to check whether the difference in the mean of the pair is zero or not. The check for within day pattern analyzes the travel time data of each trip with the previous trip(s) that happened on the same day. Trip-wise patterns can capture the traffic conditions such as incidents that happened on that day. The check for daily patterns analyses the significance of trips that happened around the same time period of the previous days to that of the current trip. In the present study, each trip was compared with the same time trips from the previous seven days' to analyze the daily and weekly patterns. For analyzing trip-wise patterns, each trip was compared with the previous five trips that happened within the same day. A basic assumption of the Z-test for the mean of a population of differences for paired samples data is that the differences of 100 m subsection travel time of the output trip and the input trip follow a normal distribution. Tests were carried out to find whether this assumption is true by using the statistical measure "skewness". Then, a ratio has been calculated between the number of times the claimed null hypothesis was accepted to the total number of times the hypothesis is tested. If the acceptance ratio was high, it was concluded that the input is significant in predicting the output trip. The pattern analysis results along with the acceptance ratios obtained for all days of the week are shown in Table 2.

Table 2: Pattern analysis results

Rank	ID	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Median
1	S1	t-1 (0.806)	d-7 (0.839)	d-7 (0.821)	t-1 (0.867)	d-7 (0.857)	d-7 (0.875)	d-7 (0.732)	0.839
2	S2	t-2 (0.806)	t-1 (0.806)	d-1 (0.821)	t-2 (0.822)	t-1 (0.844)	d-1 (0.857)	t-2 (0.722)	0.821
3	S3	d-7 (0.768)	t-2 (0.806)	t-1 (0.806)	d-7 (0.800)	d-1 (0.843)	d-2 (0.804)	d-1 (0.714)	0.804
4	S4	t-4 (0.750)	t-3 (0.806)	t-4 (0.806)	d-2 (0.804)	d-6 (0.800)	d-4 (0.786)	d-2 (0.696)	0.804
5	S5	t-3 (0.667)	d-2 (0.804)	d-5 (0.786)	d-1 (0.757)	d-2 (0.771)	t-1 (0.778)	d-4 (0.696)	0.771
6	S6	d-1 (0.643)	d-5 (0.084)	d-4 (0.768)	t-3 (0.756)	d-3 (0.700)	d-3 (0.750)	t-1 (0.694)	0.750
7	S7	t-5 (0.639)	d-4 (0.786)	t-5 (0.722)	d-6 (0.743)	t-2 (0.689)	t-2 (0.722)	t-3 (0.694)	0.722
8	S8	d-3 (0.429)	t-4 (0.778)	d-6 (0.696)	t-4 (0.733)	t-3 (0.667)	d-6 (0.661)	t-4 (0.694)	0.694
9	S9	d-5 (0.378)	t-5 (0.778)	t-3 (0.694)	d-4 (0.700)	t-5 (0.644)	t-3 (0.639)	d-3 (0.661)	0.661
10	S10	D-2 (0.357)	d-3 (0.732)	t-2 (0.667)	d-5 (0.671)	d-5 (0.629)	t-4 (0.639)	t-5 (0.639)	0.639
11	S11	d-4 (0.357)	d-6 (0.661)	d-3 (0.643)	t-5 (0.556)	t-4 (0.622)	t-5 (0.583)	d-6 (0.625)	0.622
12	S12	d-6 (0.232)	d-1 (0.286)	d-2 (0.482)	d-3 (0.386)	d-4 (0.386)	d-5 (0.357)	d-5 (0.589)	0.386

* (d-n) represents the same time trip n days ago and (t-n) represents the previous nth trip within the same day.

The results showed all the weekdays having a similar pattern and Sunday following a different pattern. Sunday showed a strong correlation to the previous trips on the same day followed by a weekly pattern.The median analysis of trip acceptance ratios was adopted in order to select the number of inputs for the KFT based algorithm and to develop the neural network. Here, the median of the acceptance ratio for all the seven days were analyzed, and the top 4 trips with values above 80% were selected as inputs.

4.2 Artificial Neural Network (ANN)

A neural network is a massively parallel distributed processor made up of simple processing units. ANNs basically replicates the intelligent data processing ability of human brains and are constructed with multiple layers of processing units, named as artificial neurons. There are three layers in an ANN namely the input layer, the hidden layer and the output layer. A basic unit of the connection is called as a neuron, which is connected by other neurons through synoptic weights. Based on the desired target value, the network will be trained, i.e., the weight matrix will be updated after each iteration of the algorithm. Thus, as the number of iterations increases, the predicted output matrix shifts close to that of the target value.

Since the problem under study was to predict the travel time pattern, a multi-layer feed-forward network with the Leveberg-Marquardt back propagation algorithm was used for training. This is a supervised learning algorithm, and is considered to be one of the fastest methods for moderate-sized feed-forward neural networks that may range up to several hundred connections (Simon, 1999). A hyperbolic tangent sigmoid function was used as the transfer function for both the hidden layer and the output layer. To make up the training set, ANN requires a good amount of database with the desired output and the corresponding input values. After training, the model will be simulated with a new set of input data to evaluate its performance. As mentioned earlier, in the present study, the four most correlated trips were chosen as inputs based on median analysis and used as input in the ANN analysis after normalising as below:

$$z_i = \frac{x_k(i) - \min_k(i)}{\max_k(i) - \min_k(i)}, \tag{1}$$

where $x_k(i)$ is the i^{th} element of column *'x'* in the dataset, and *$max_k(i)$ and $min_k(i)$* are the minimum value and the maximum value of that particular dataset respectively. Thus, the number of nodes in the input and output layers were 4 and 1 respectively. The optimum number of neurons required to train each section were identified by carrying out a heuristic analysis. Figure 2 shows the optimum number of neurons required for each subsection along with corresponding variance in travel time. From Figure 2, it can be observed that the subsections that were having a high variance of travel time required more number of neurons than the other subsections

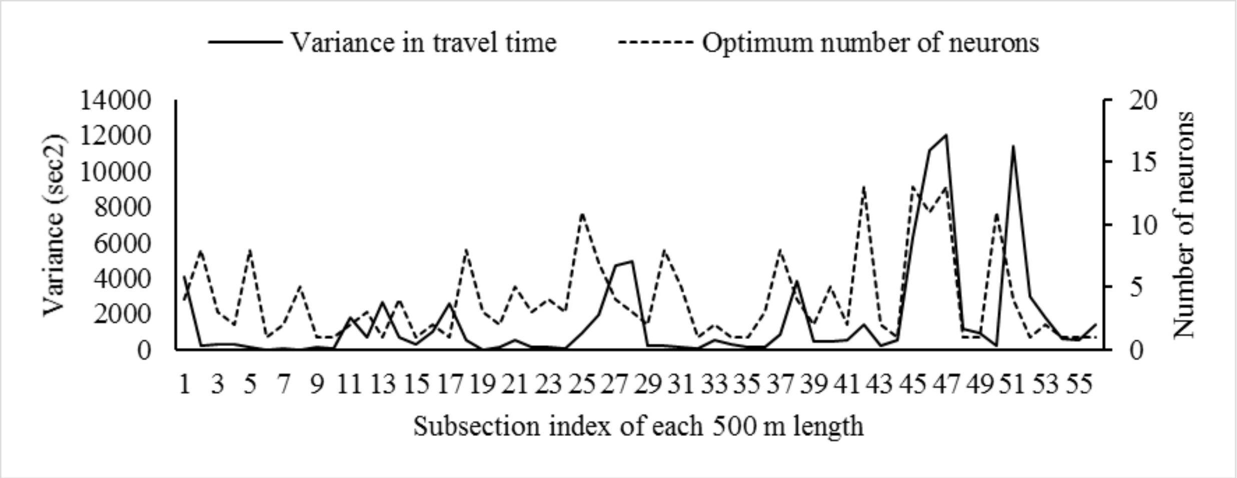

Figure2: Optimum number of neurons required for each subsection vs. Variance intravel time

For the present study, out of 35 days' data, 21 days' data were selected for training, 7 days' data were used for validation and remaining 7 days' data were kept for testing the performance. The training rule for the ANN used in the study was

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}\right]^{-1}\mathbf{J}^T\mathbf{e}, \tag{2}$$

where **J**is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights biases, and **e**is the vector of network errors. The Jacobian matrix can be computed through a standard back propagation technique (Simon, 1999). One of the reported problems in ANN training was over fitting. Over fitting reduces the generalizing capability of the network resulting in errors when tested with unseen data. To curb the problem of over fitting, regularization of the performance function was carried out. This was done by adding a term that consists of the mean of the sum of squares of the network weights and biases (MSW) to the typical performance function (*f*), Mean Square Error (MSE). Thus, the new performance function (f_m) became

$$f_m = \alpha\, M\mathrm{SE} + (1-\alpha)MSW\ . \tag{3}$$

In this study, the value of α was taken as 0.5. Since the weight matrix was randomly initialized for every run, multiple runs were carried out. The results obtained for all runs for selected subsections are shown in Figure 3. It can be observed from Figure 3 that there were variations in the MAPE for different random seed runs of ANN. Hence, from these runs, an ensemble averaging method was used to find out the final output. In this method, the output produced by the different trained networks are combined and linearly averaged. The motivation behind using this technique was the fact that differently trained networks (caused by random weights) converge to different local minima on the error surface, and the overall performance is improved by combining the outputs (Simon, 1999).

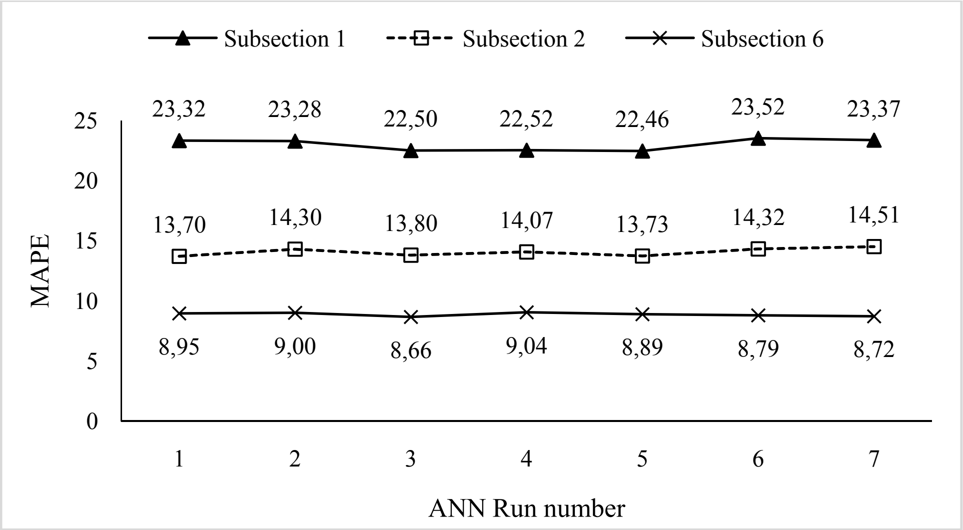

Figure 3: Variation in errors (MAPE) over runs for ANN in selected subsections

Next, a model-based approach using Kalman Filtering technique was used to predict the bus travel time and the details are presented in the next section.

4.3 Model Based Approach

A model based approach using KFT was used in this study for comparing the performance with ANN. The KFT (Kalman, 1960) can be used to estimate variables and parameters of systems that can be described using equations in the state space form. The implementation of the Kalman filter requires the information about system dynamics, statistical information of the system disturbances and measurement errors. It uses the model and the system inputs to predict the *a priori* state estimate and uses the output measurements to obtain the *a posteriori* state estimate. Overall, it is a recursive algorithm, and new measurements can be processed as and when they are obtained. KFT needs only the current instant's state estimate, the current input and the latest output measurements to calculate next instant's state estimate. The selected approach (Vanajakshi *et al.*, 2009) has minimal data requirement and the aim of this study is to find out whether the use of a large database and a corresponding data driven approach can improve the prediction accuracy significantly. The evolution of travel time between the various subsections was assumed as

$$x(k+1) = a(k)x(k) + w(k)\,, \tag{4}$$

where *a(k)*is a parameter that relates the travel time in the k^{th} subsection and the $(k+1)^{th}$ subsection, *x(k)* is the travel time to cover a given k^{th} subsection and *w(k)* is the associated process disturbance with the k^{th} subsection. The measurement process was assumed to be governed by

$$z(k) = x(k) + v(k)\,, \tag{5}$$

where *z(k*) is the measured travel time in the k^{th} subsection and *v(k)* is the measurement noise. It was further assumed that *w(k)* and *v(k)* are zero mean white Gaussian noise signals with *Q(k)* and *R(k)* being their corresponding variances. Thus, two sets of data are required to implement the KFT. One set of data was used in the time update equations to calculate the parameter *a(k)*, and the other in the measurement update equations to calculate the *a posteriori* estimate. In this study, the four most relevant trips identified from the pattern analysis as shown in Table 2 (represented as S1 to S4, with S1 being the most significant) were used in this regard. The average travel time of the trips S1 and S4 was used to calculate the parameter *a(k).* The average travel time of trips S2 and S3 was used to obtain the *a posteriori* estimate of the next vehicle (TV). The steps in the KFT algorithm are as follows:

1. The entire section from origin to destination was divided into *N* subsections of equal length.

2. The average travel time data obtained from S1and S4 were used to obtain the value of *a(k)* by

$$a(k)=\frac{x_{S1,S4}(k+1)}{x_{S1,S4}(k)},k=1,2,3.......(N-1). \qquad (6)$$

3. Let x_{TV} be the travel time taken by the test vehicle to cover the given subsection. For the first subsection, the TV travel time and the corresponding error variance was were taken as

$$E[x_{TV}(1)]=\hat{x}(1), \qquad (7)$$

$$E[x_{TV}(1)-\hat{x}(1)^2]=P(1). \qquad (8)$$

4. For *k=2, 3, 4…, (n-1),* the following steps were performed:
   i. The *a priori* estimate of the travel time was calculated by using

$$\hat{x}^-(k+1)=a(k)\hat{x}^-(k). \qquad (9)$$

   ii. The *a priori* error variance (P^-) was calculated by using

$$P^-(k+1)=a(k)P^+(k)a(k)+Q(k). \qquad (10)$$

   iii. The Kalman gain (*K*) was calculated by using

$$K(k+1)=P^-(k+1)[P^-(k+1)+R(k+1)]^{-1}. \qquad (11)$$

   iv. The *a posteriori* travel time estimate $(\hat{x}^+)$ and error variance (P^+) were calculated by using Equation 12 and Equation 13 respectively

$$\hat{x}^+(k+1)=\hat{x}^-(k+1)+K(k+1)[z(k+1)-\hat{x}^-(k+1)], \qquad (12)$$

$$P^+(k+1)=[1-K(k+1)]P^-(k+1). \quad (13)$$

The objective here was to predict the travel time of the TV in the $(k+1)^{\text{th}}$ subsection using the travel time obtained from the most significant inputs identified, when TV is in the k^{th} subsection. The above algorithm was implemented and the results obtained are discussed below.

5. Results

The results obtained from the implementation of the prediction methods in previous section were discussed in this section. The prediction accuracy was quantified using the Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) as

$$MAPE = \frac{\sum_{i=1}^{n} \frac{\left|x_p - x_a\right|}{x_a}}{N} * 100, \qquad (14)$$

$$MAE = \frac{\sum_{i=1}^{n} x_p - x_a}{N}, \qquad (15)$$

where x_p is the predicted travel time of TV to cover a given subsection and x_a is the corresponding travel time measured from the field. The performance evaluation of the proposed methods was carried out by comparing the predicted values with the actual values over a period of six days for various trips and subsections. Also, in the present study, a sensitivity analysis has been carried out, by training the ANN model with different amounts of datasets such as training the model with 1 weeks' data, 2 weeks' data and 3 weeks' data. The performance of ANN with different amounts of input data was compared with that of KFT that used data from the four most significant trips alone for a period of one week.

5.1 Model Based Approach

The performance of the ANN method that trained with different amounts of dataset was compared with KFT method that uses significant inputs from previous one-week data. Figure 4 shows the average MAPE obtained for all trips across test period. From Figure 4, it can be observed that the ANN trained using three weeks' data performed better than the model based approach, and is used in the rest of the analysis.

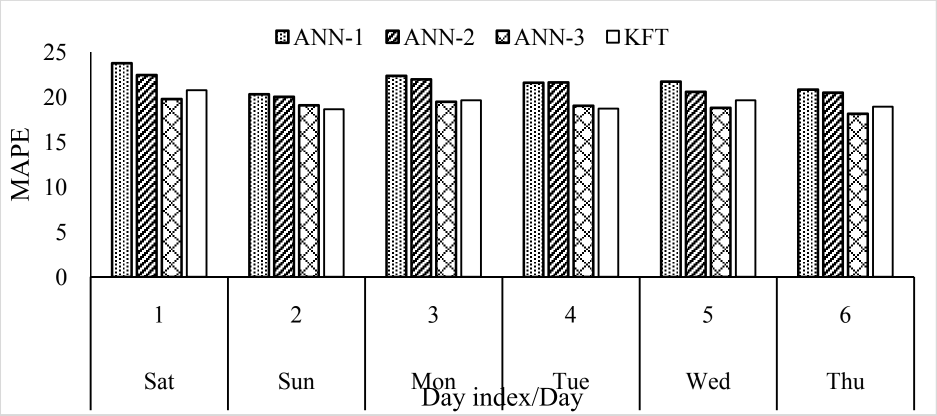

Figure4: MAPE comparison for ANN (with different training amounts) and KFT.

Figure 5 shows the performance of the ANN method that was trained with previous 3 weeks' data and the KFT method for a sample trip. It can be observed that the ANN trained with 3 weeks' data was able to capture the variations better than the KFT method. A similar analysis has been carried out for all trips that happened in a representative day and the results are presented in Figure 6. From Figure 6, it can be observed that ANN was able to perform better than the KFT method in the majority of the cases.

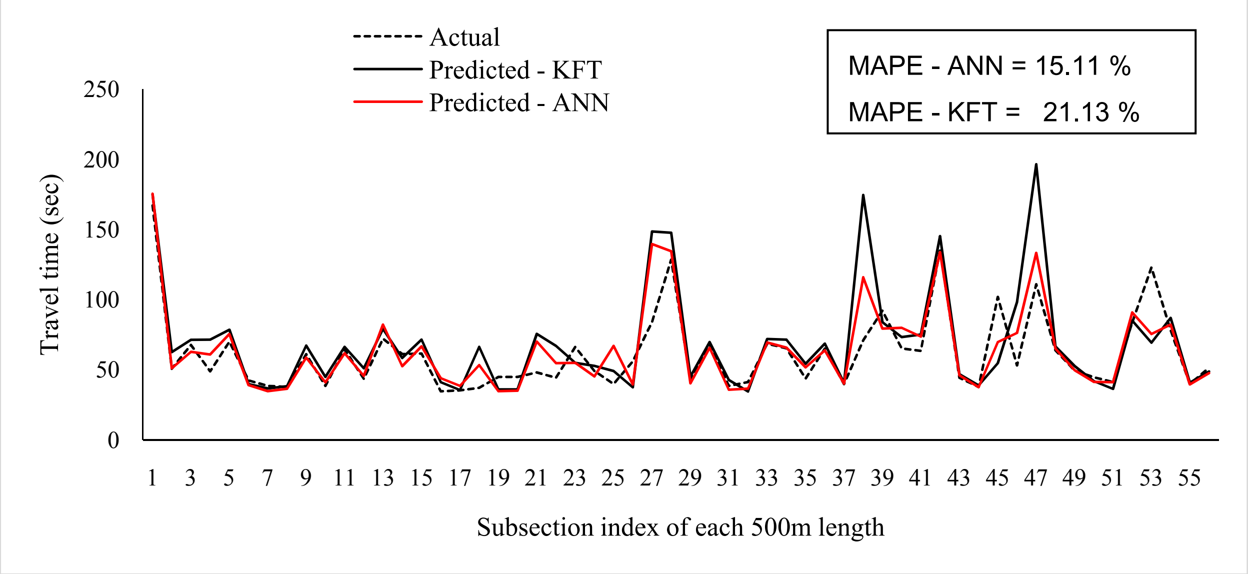

Figure5: Predicted and observed travel times for a sample trip during test period

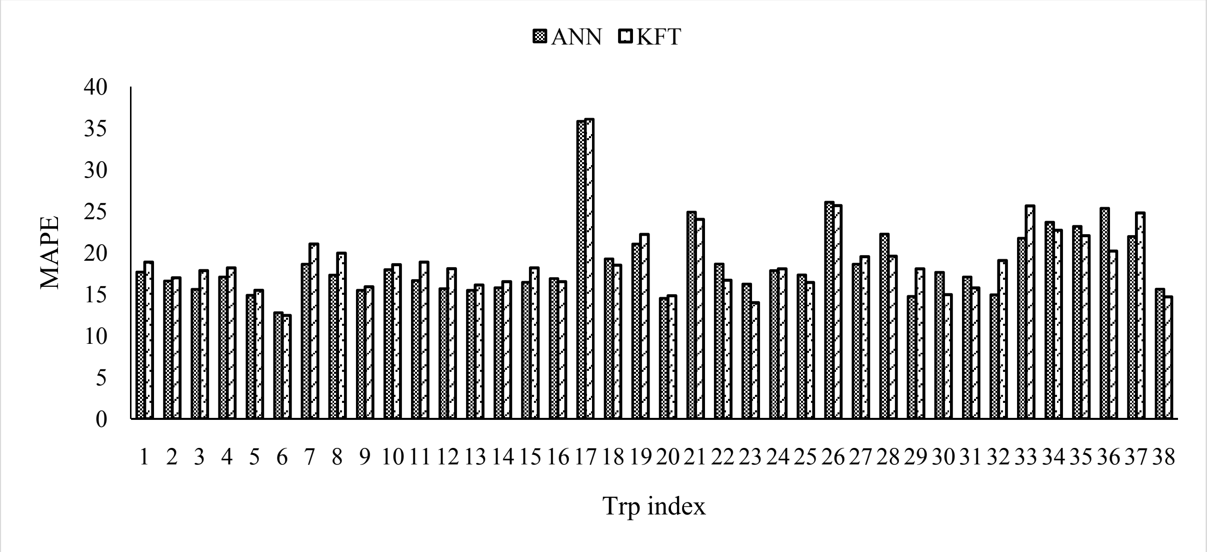

Figure6: Performance comparison over trips that happened on a sample day

5.2 Model Based Approach

The performance of the ANN method and the KFT were compared for each subsection also in terms of MAPE and MAE. Figure 7 shows the sample result obtained for a sample subsection, where the predicted travel times from ANN and KFT are shown against actual travel time. From Figure 7, it can be observed that ANN was able to capture the travel time better with an MAPE of 18.72% whereas as KFT had an error of 23.57%.

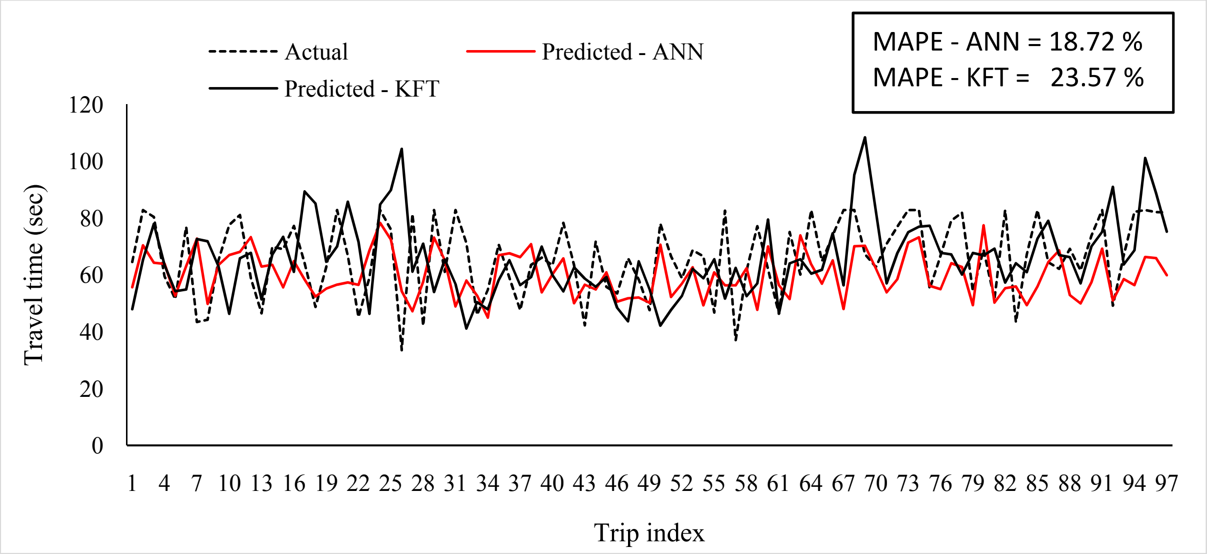

Figure7: Comparison of predicted and actual travel times for a selected subsection

A similar analysis has been carried out for all subsections and the average reduction in error while considering ANN over KFT during the test period is shown in terms of deviation from actual travel time in Figure 8. From Figure 8, it can be observed that the ANN was able to perform better in the majority of the subsections.

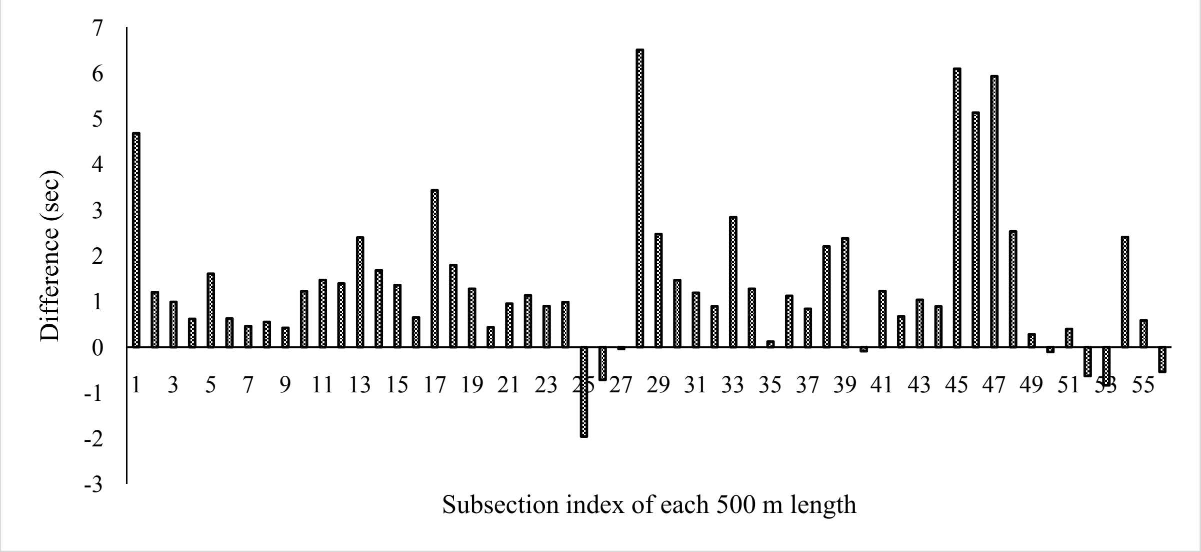

Figure8: Reduction in error by considering ANN rather than KFT for all subsections along the selected route

6. Conclusions

The main aim of Advanced Public Transportation System (APTS) is to attract passengers towards public transport, and thus help to reduce the congestion on urban roads. One of the popular applications of APTS is to provide information about bus travel time to passengers, which plays a key role in the successful implementation of any APTS applications. However, for this to happen in practice, the bus service should be made more attractive. The effectiveness of such information provided to passengers mainly depends on the prediction technique used and the input data used in the same. The present study compared the performance of two commonly used prediction techniques, one data driven and the second with minimal data requirement, for bus arrival prediction. The data driven technique selected is the Artificial Neural Network and a model based approach using the Kalman Filtering Technique with minimal data requirement of just two previous buses were used for less data demanding technique. To identify the most significant inputs, a travel time pattern analysis was carried out using statistical methods. While implementing the prediction algorithm using ANN, the optimum number of neurons required to train each subsection were identified by carrying out a heuristic analysis. Also, a sensitivity analysis was carried out with different size of training data sets for ANN, to identify when its performance becomes better than KFT. It was observed that ANN trained using data from more than two weeks performed better than KFT. However, if the data availability is less than two weeks, the performance of KFT was better. Thus, if one is constrained by database size, the model based approach using KFT can be a better option. Else, ANN would be able to provide more accuracy in prediction.

7. Acknowledgement

The authors acknowledge the support for this study as part of the project CIE/10-11/168/IITM/LELI by Ministry of Urban Development, Government of India, through letter No. N-11025/30/2008-UCD.

References

Bin, Y., Zhinzhen, Y. and Baozhen, Y. (2006). “Bus arrival time prediction using support vector machines”, *Journal of Intelligent Transportation Systems*. 10(4), pp. 151-158.

Cathey, F.W. and Dailey, D.J. (2003). “A Prescription for Transit Arrival/Departure Prediction using AVL Data”, *Transportation Research Part C: Emerging Technologies*, 11, pp. 241-264.

Chamberlein. (2015). *Great circle distance between two points*. http://www.movable-type.co.uk/scripts/gis-faq-5.1.html. Accessed on June 20, 2015,

Chen, M. and Chien, S. (2001). “Dynamic freeway travel time prediction with probe vehicles data, link-based and path-based”, *Computer Aided Civil and Infrastructure Engineering*. 19, pp. 364-376.

Chen, M., Liu, X.B. and Xia, J.X. (2004). “A dynamic bus arrival time prediction model based on APC data”, *Computer Aided Civil and Infrastructure Engineering*, 19, pp. 364–376.

Chien, S.I.J. and Kuchipudi, C.M. (2003). “Dynamic travel time prediction with real-time and historic data”, *ASCE Journal of Transportation Engineering*, 129(6), pp. 608–616.

Chien, S.J., Ding, Y. and Wei, C. (2002). “Dynamic bus arrival time prediction with artificial neural networks”, *ASCE Journal of Transportation Engineering*, 128(5), pp. 429–438.

Chu, L., Oh J.S. and Recker, W. (2005). "Adaptive Kalman Filter Based Freeway Travel time Estimation", *Proceedings of Transportation Research Board, Transportation Research Board*, National Research Council, Washington, D.C., USA.

Hagan, M.T., Demuth, H.B. and Beale. M. (1996). *Neural network design*, PWS, Boston.

Jeong, R. and Rilett, L. (2004). "The prediction of bus travel time using AVL data", *Proceedings of 83rd Annual Meeting of the Transportation Research Board*. National Research Council, Washington D.C., USA.

Kalaputapu, R. and Demetsky, M.J. (1995). "Application of artificial neural networks and automatic vehicle location data for bus transit schedule behavior modelling", *Transportation Research Record*, 1497, pp. 44–52.

Kalman, R.E. (1960). "A New Approach to Linear Filtering and Prediction Problems",*Transaction of the ASME-Journal of Basic Engineering*, 82(1), pp. 35-45.

Kumar S. V. and Vanajakshi, L. (2014). "Urban Arterial Travel Time Estimation Using Buses as Probes", *Arabian Journal of Science and Engineering*. 39, pp. 7555–7567.

Kumar, S.V. and Vanajakshi. L. (2012). "Pattern identification based bus arrival time prediction", In *proceedings of the Institute of Civil Engineers-Transport*, paper 1200001.

Nanthawichit, C., Nakatsuji, T. and Suzuki, H. (2003). "Application of probe vehicle data for real-time traffic state estimation and short-term travel time prediction on a freeway", *Proceedings of 82nd Annual Meeting of the Transportation Research Board.* National Research Council, Washington D.C., USA.

Padmanabhan, R.P.S., Vanajakshi, L. and Subramanian, S.C. (2009). "Estimation of bus travel time incorporating dwell time for APTS applications", *Intelligent Vehicles Symposium*, 1931-0587.

Pan, J., Dai, X. and Xu, X. (2012). "A self-learning algorithm for predicting bus arrival time based on historical data model", *Proceedings of IEEE 2nd International Conference on Cloud Computing and Intelligent Systems*.

Park, D., Rilett, L. and Han, G. (2002). "Spectral basis neural networks for real-time travel time forecasting", *Journal of Transportation Research Board*. 125(6), pp. 515-523.

Ramakrishna, Y., Ramakrishna, P., Laxshmanan, V., Sivanandan, R. (2006) "Bus Travel Time Prediction Using GPS Data", In *9th International Conference on Geographic Information, Technology and Applications, Map India*, New Delhi, India. http://www.gisdevelopment.net/proceedings/mapindia/2006/student%20oral/mi06stu_84.htm.

Schweiger, C. (2003). "Real-Time Bus Arrival Information Systems", *Technical report*, *Transportation Research Board*, TCRP Synthesis 48, Washington D.C., USA.

Shalaby, A. and Farhan, A. (2004). "Bus travel time prediction for dynamic operations control and passenger information systems", *Proceedings of 83rd Annual Meeting of the Transportation Research Board*. National Research Council, Washington D.C., USA.

Simon, H. (1999). *Neural networks, A comprehensive foundation*. Pearson education group, Inc., New York.

Son, B., Kim, H.J., Shin, C.H. and Lee, S.K. (2004). "Bus Arrival Time Prediction Method for ITS Application", *Knowledge Based Intelligent Information and Engineering Systems*, Springer Berlin Heidelberg, pp. 88–94.

Vanajakshi, L., Subramanian, S.C. and Sivanandan, R. (2009). "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses", *IET intelligent Transportation Systems*. 3(1), pp. 1-9.

Wall, Z. and Dailey, D.J. (1999). "An algorithm for predicting the arrival time of mass transit vehicles using automatic vehicle location data", *Proceedings of 78th Annual Meeting of the Transportation Research Board*. National Research Council, Washington D.C., USA.

Wang, J., Chen, X. and Guo, S. (2009). "Bus travel time prediction with v-support vector regression", *IEEE Intelligent Transportation Systems*. 3(1), pp. 655-660.

Yang, J.S. (2005). "Travel time prediction using GPS test vehicle and Kalman filtering techniques", *American Control Conference*, Portland, Oregon, USA.

Yu, B., Lam, W.H.K. and Tam, M. (2011). "Bus arrival time prediction at bus stop with multiple routes", *Transportation Research Part - C*, 19(6), pp. 1157–1170.

Yu, B., Yang, Z.Z. and Wang, J. (2010). "Bus travel-time prediction based on bus speed", *Proc. of the Inst. of Civil Engineers Transport*, 163, pp. 3–7.

Zhu, T., Ma, F., Ma, T. and Li, C. (2011). "The prediction of bus arrival time using global positioning system data and dynamic traffic information", *Wireless and Mobile Networking Conference (WMNC), 4th Joint IFIP*. Beijing, China.