

# Hybrid travel time estimation model for public transit buses using limited datasets

Ashwini Bukanakere Prakash<sup>1</sup>, Ranganathaiah Sumathi<sup>1</sup>, Honnudike Satyanarayana Sudhira<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Siddaganga Institute of Technology Tumakuru, Karnataka, India

<sup>2</sup>Gubbi Labs, Tumakuru, Karnataka, India

## Article Info

### Article history:

Received Sep 22, 2022

Revised Jan 12, 2023

Accepted Mar 10, 2023

### Keywords:

Bus travel time prediction

Dynamic model

Gradient boosting regression trees

Hybrid model

Machine learning

Passenger information system

## ABSTRACT

A reliable transit service can motivate commuters to switch their traveling mode from private to public. Providing necessary information to passengers will reduce the uncertainties encountered during their travel and improve service reliability. This article addresses the challenge of predicting dynamic travel times in urban areas where real-time traffic flow information is unavailable. In this perspective, a hybrid travel time estimation model (HTTEM) is proposed to predict the dynamic travel time using the predicted travel times of the machine learning model and the preceding trip details. The proposed model is validated using the location data of public transit buses of, Tumakuru, India. From the numerical results through error metrics, it is found that HTTEM improves the prediction accuracy, finally, it is concluded that the proposed model is suitable for estimating travel time in urban areas with heterogeneous traffic and limited traffic infrastructure.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Ashwini Bukanakere Prakash

Department of Computer Science and Engineering, Siddaganga Institute of Technology

Tumakuru, Karnataka, India

Email: ashvinibp@sit.ac.in

## 1. INTRODUCTION

Population growth [1] and migration of rural population towards cities have directly influenced the growth of vehicle population [2] in urban areas leading to congestion and increased travel time for commuters. Using public transit services such as buses can reduce the aforesaid problems of commuting in tier-2 cities (in countries like India) with limited mass transit options. To pull commuters toward public transit, it is necessary to provide reliable transit service [3]. Advanced passenger information system (APIS), a part of the intelligent transport system (ITS) [4] aims to provide dynamic information to passengers about the schedule, arrival, and travel times making the public transit service more reliable. Providing real-time information to passengers requires the integration of several static data such as spatial characteristics of the road network, schedule of buses, headway, and real-time data sources such as live location information of the buses, traffic flow, incidents, and congestion-related information. In cities with limited ITS infrastructure the aforementioned real-time data sources are rarely available but, recently the public transit buses in many cities are equipped with global positioning system (GPS) [5] that provides the dynamic location of the buses, speed, and heading.

The travel time naturally varies depending on the space and time such as the day of the week, (weekday or weekend), time of the day, the land use pattern (LUP) of the route, the number of signalized and non-signalized intersections, and so on. The LUP is correlated to traffic density, commercial/business areas exhibit high traffic volumes leading to congestion and more travel time. Also, the stochastic behavior of the traffic affects the travel time of buses. Delays at intersections [6] due to queuing, non-optimized signals [7], and mixed traffic lanes lead to an excess travel time for buses. Additionally, the bus driver's behavior [8] in

scenarios such as early start, on-time start, and delayed start of the trips will make the prediction process more complicated. Public transit travel time is predicted in two scales, long [9] and short-term [10]. If the travel time is forecasted more than 60 minutes ahead of the current time, it is a long-term, else a short-term prediction. Long-term prediction aid the decisions for operations planning of the buses [11], while short-term assists the passenger information systems [12], bus routing, fine-tuning schedules, and identifying bus bunching [13].

In the existing works, the bus travel times are predicted using manually collected data [14], GPS logs of the buses [15], automatic passenger counters [16], and mobile phone footprints [17]. GPS-based location data is the most common data source used for bus travel time prediction. From naïve prediction models such as historic averages [18], statistical models like Kalman Filtering [18], time series models such as autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA) [19]. Machine learning (ML) models such as artificial neural networks (ANN) to deep networks models [20], several authors have researched on the application of these models for short and long term travel time predictions. Considering high variances in the travel time authors in [21] have implemented the support vector regression (SVR) model in Chennai, India. Spatial and temporal SVR model is implemented to predict dynamic travel time using only GPS data of buses. Authors in [22] have used probe vehicle data, segment data, and weather data in Charlotte, North Carolina for short-term travel time prediction. Decision trees (DT), random forest regression (RFR), extreme gradient boosting (XGBoost), and long short-term memory (LSTM) models are compared and it is found that RFR outperforms other models. In [23] non-linear and linear models are compared for predicting the bus travel time using location data of the public transit buses and authors found that non-linear models DT, RFR, gradient boosting regression trees (GBRT) and k-nearest neighbors (KNN) outperformed the linear models; Linear regression, SVR, least absolute shrinkage and selection operator and ridge regression models, highlighting the non-linear behavior of travel times. The study was conducted in a tier-2 city (Tumakuru), India. A study [24] in Mumbai, India estimated the time of arrival of buses at bus stops and intersections using parametric hazard models based on Cox regression and accelerated failure time (AFT). Log-logistic and Weibull distributions are fitted to estimate the arrival time of buses at bus stops and intersections, and the authors found that the AFT model performed better with a 10% reduction in prediction variation.

Though the prediction of travel time looks like a simple regression, the randomness of traffic density, weather conditions, congestion, and incidents demand sophisticated prediction models. Several researchers have developed short-term travel time prediction models using ML but few authors stress the necessity of hybrid models. However, most of the existing works are conducted in tier-1 cities with matured traffic infrastructure. But in tier-2 cities with limited infrastructure and data sources scenario, a dynamic model to predict travel time is yet to be explored. In this work, short-term travel times of buses are predicted using limited data sources, i.e., historical location data of buses, schedule, bus-stop information, and road geometry data. The objectives of the study are to predict/estimate the travel time of public transit buses at two spatial aggregation levels; route and segment levels in a tier-2 city with limited data sources as follows:

- a. Apply the GBRT machine learning model based on the historic data to predict the bus travel time.
- b. Propose the latest travel time estimation model (LTTEM) to estimate the dynamic bus travel time utilizing the travel time information of the preceding bus and historical data.
- c. Propose a hybrid travel time estimation model (HTTEM) that combines the results of the GBRT and LTTE models to improve the performance as compared to the individual models.

## 2. DATA

The location data of public transit buses of Tumakuru city during March 2021 is used for the modeling. The data includes the bus number, device-ID, date and time, latitude, longitude, speed, odometer, and location information. Route number 201: Tumakuru bus stand (TBS) to Kyathasandra (KYA) is chosen for model development and validation. The route under study is split into four segments based on the LUP. Segment 1 is from the TBS to Bhadramma choultry (BC) is a part of the central business district (CBD). Segment 2 is from BC to Shivakumara Swamiji circle (SSC) is an inner-city (IC) area, segment 3 from SSC to Batawadi (BW) is the inner suburban (ISU) region, and finally from BW to KYA segment 4, which is an outer suburban area (OSU). The map of the bus route is given in Figure 1, highlighting the bus stops and location of the split of segments. The route information is given in Table 1. Sample location data are presented in Table 2. The raw location data needs to be cleaned and pre-processed for analysis. For each trip start time, end time, segment speed, length traversed, and total travel time is extracted from the data at both the route level and the segment level. 500 upstream trips on weekdays between mornings 7:00 to evening 20:00 are used in the study. The LUP, the number of intersections, and the segment number are augmented to trip aggregates. The weather data is not considered as the study location has stable weather, and the climatic changes rarely impact the travel times during the study period.



Figure 1. Map of route number 201

Table 1. Route information

Parameters	S1	S2	S3	S4	Route
Origin – destination	TBS-BC	BC-SSC	SC-BW	BW-KYA	TBS-KYA
Length in kilometers	1.76	1	2.09	2.05	6.9
Bus stops	3	2	3	1	9
Signalized intersections	3	1	1	1	6
Land use pattern	CBD	IC	ISU	OSU	-

Table 2. Sample location data of route 201

Vehicleregno	Date_time	Location	Speed	Odometer	Latitude	Longitude
KA-06-F-0857	01-03-2021 07:29:58	TBS	13	12673.91	13.34319	77.09875
KA-06-F-0857	01-03-2021 07:30:08	TBS	22	12673.96	13.34275	77.09859
KA-06-F-0857	01-03-2021 07:30:18	Near by Karnataka state road transport corporation bus depot	23	12674.03	13.34216	77.09837
KA-06-F-0857	01-03-2021 07:30:28	Near by Karnataka state road transport corporation bus depot	8	12674.07	13.34181	77.09828
KA-06-F-0857	01-03-2021 07:30:38	Near by 4148, PH colony	21	12674.11	13.34145	77.09815

### 3. METHODS

Estimating the dynamic travel time needs the historical data of the trips, the up-to-the-minute location data of the buses, traffic on links, delays at intersections, and any incidents in the link. In most tier-2 cities, these data are rarely available, but developing models to estimate dynamic travel time is crucial in such locations. In this context, the travel time of buses is estimated using the limited data available by applying the methods discussed in the following subsections.

#### 3.1. Gradient boosting regression trees

In a study conducted [23] it is identified that non-linear models are suitable for predicting travel time in the current study location. Therefore, travel time data are modeled using a non-linear ML model, GBRT. As the name suggests it is an ensemble greedy model that works based on boosting approach. GBRT works on the notion that the possible next model, when combined with the preceding weak model (regression tree) will reduce the error in overall prediction, and every proceeding model tries to reduce the errors of the collective boosted ensemble of all preceding models. It is a widely used model in predicting travel times.

#### 3.2. Latest travel time estimation model

The travel time of buses is highly influenced by the real-time traffic situation. In a limited traffic-related infrastructure scenario, the information from the preceding trip can aid the estimation. Therefore, an analysis of the correlation between the preceding trips to the current trip within a 30-minute time frame is

conducted initially using (1). The level of variation in one feature due to variation in another is quantified through correlation analysis [25] method. The range of the correlation coefficient is -1 to +1, indicating perfect negative or positive correlation respectively.

$$x_t = \frac{\sum[(t_i - \mu t_i) * (t_k - \mu t_k)]}{\sqrt{\sum(t_i - \mu t_i) * \sum(t_k - \mu t_k)}} \quad (1)$$

In (1),  $t_i$  is the current trip travel time,  $t_k$  is the preceding trip travel time,  $\mu t_i$  and  $\mu t_k$  are the means of the current and preceding trip travel times and  $x_t$  is the correlation coefficient. Similarly, the correlation coefficient  $x_s$  of current trip travel speed to that of the preceding trip travel speed is estimated. The coefficients estimated are used in developing the proposed model. The Latest Travel Time  $LTT_{i,j}$  is estimated using the model in (2).

$$LTT_{i,j} = \begin{cases} \alpha_j * TT_{k,j} & | \text{if } RS_{k,j} < THRS_j \\ z_{1,j} * TT_{k,j} + z_{2,j} * TRT_{k,j} & | \text{otherwise} \end{cases} \quad (2)$$

The latest travel time  $LTT_{i,j}$  is estimated based on the traffic scenario during the previous trip in two traffic patterns.

Case 1: Congestion: If, the running speed of the segment during the previous trip falls below  $THRS_j$  (the threshold speed [26] of the segment  $j$ ), congestion in the link is presumed. The latest travel time in such a situation is calculated by adjusting the preceding trip's travel time  $TT_{k,j}$  by the defined correction rate  $\alpha_j$ .

Case 2: Congestion-free: If the running speed  $RS_{k,j}$  of the segment in the preceding trip is above the threshold range, then it is presumed that the link is congestion-free. In such cases, the  $LTT_{i,j}$  is estimated based on the weighted average of preceding trip travel times  $TT_{k,j}$  and  $TRT_{k,j}$  (preceding trip travel time based on speed  $RS_{k,j}$ ). The estimations for  $\alpha_j$ ,  $z_{1,j}$ ,  $z_{2,j}$ ,  $TRT_{k,j}$  and  $THRS_j$  given in (3) to (7).

The correction rate  $\alpha_j$  is estimated for the trips of *case 1* type in the historical data. Its value is in the range of  $0.8 \leq \alpha \leq 2$ . With the base value of 0.8, it is incremented in the steps of 0.05, up to 2. At each value of  $\alpha$ ,  $error_{i,j}$  is calculated using (3) where the travel time of the current trip is  $CTT_{i,j}$  and the preceding trip is  $TT_{k,j}$  in the archived data. The  $\alpha$  that yields the minimum  $error_{i,j}$  is the  $\alpha_j$  explored for segment  $j$ , and the respective  $\alpha_j$  will be used in (2).

$$error_{i,j} = \sum_{i=1}^n |TT_{k,j} * \alpha - CTT_{i,j}| \quad 0.8 \leq \alpha \leq 2 \quad (3)$$

The threshold for running speed  $THRS_j$  for each segment  $j$  is estimated based on historical data through a statistical approach  $1.5 * IQR$  rule as given in (4).  $Q1$  (Quartile 1), is the mean speed value below which 25% of data are present, and  $Q3$  (Quartile 3), is the mean speed value below which 75% of them are present when they are arranged in ascending order.  $IQR$  (InterQuartile Range) is the difference between  $Q1$  and  $Q3$ . Conventionally the constant 1.5 is multiplied by the  $IQR$  to estimate the lower and upper bound, the data points beyond this are considered outliers. In the current study, only the lower bound is used to predict congestion. If the speed of the preceding trip is an outlier (to the lower side), congestion is presumed, and the current travel time is increased  $\alpha_j$  times of the  $TT_{k,j}$

$$THRS_j = Q1(RS_j) - 1.5 * IQR(RS_j) \quad (4)$$

The  $z_{1,j}$ ,  $z_{2,j}$ , are the weights of the travel time and travel speed of the preceding trip respectively. In (5) and (6) the  $x_t$ ,  $x_s$  are the coefficients estimated using (1). The  $TRT_{k,j}$  is the travel time for segment  $j$  estimated based on the running speed  $RS_{k,j}$  of the preceding trip using (7), where  $d_j$  is the distance and  $RS_{k,j}$  is the preceding trip running speed of the respective segments. The  $Delay_j$  is the average delay at the intersection for segment  $j$  and it is added to the travel time estimated through running speed.  $Delay_j$  is determined based on field study and historic location data.

$$z_{1,j} = x_t / (x_t + x_s) \quad (5)$$

$$z_{2,j} = x_s / (x_t + x_s) \quad (6)$$

$$TRT_{k,j} = seconds(d_j / RS_{k,j}) + Delay_j \quad (7)$$

### 3.3. A hybrid model: hybrid travel time estimation model

The schematic of the implementation process of the hybrid model proposed is presented in Figure 2. The proposed HTTEM combines the predictions of the GBRT model and the LTTE model. The HTTEM is given in (8), where  $ETT_{i,j}$  is the estimated travel time of a trip  $i$  for segment  $j$ . In this a weighted average of predicted travel time  $PTT_{i,j}$  by GBRT and the latest travel time  $LTT_{i,j}$  by LTTE model, based on the weights  $(w_{1,j}, w_{2,j})$ . The weights are estimated by applying (10).

$$ETT_{i,j} = w_{1,j} * PTT_{i,j} + w_{2,j} * LTT_{i,j} \tag{8}$$

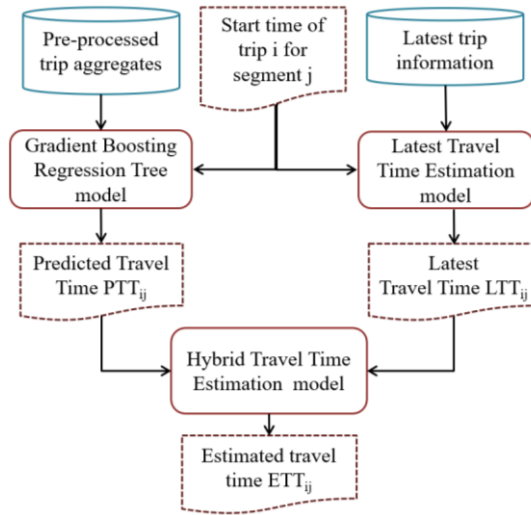


Figure 2. The schematic of the proposed model implementation process

The R-squared ( $R^2$ ) or the coefficient of determination is a statistical parameter that examines the linear relationship strength among the variables considered. The general equation for estimating the coefficient of determination measure is given in (9). Where  $y_{i,j}$  is the actual travel time of trip  $i$  for segment  $j$  which is the dependent variable,  $x_{i,j}$  is the set of independent variables namely the start time of the trip, day of the week, and the LUP, and  $f(x_{i,j})$  is the function that returns the predicted travel time.  $\mu y_i$  is the average of the actual travel time in the test data. The  $R^2$  value of the ML model is the  $w_{1,j}$  in (9), and the formula to estimate  $w_{2,j}$  is given in (10).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{i,j} - f(x_{i,j}))^2}{\sum_{i=1}^n (y_{i,j} - \mu y_i)^2} \tag{9}$$

$$w_{1,j} = R^2 \quad w_{2,j} = (1 - w_{1,j}) \tag{10}$$

The step-by-step procedure to dynamically estimate the travel time at the segment level is given in Algorithm 1. The steps to estimate the hybrid travel time at the route level are given in Algorithm 2. The dynamic travel time at the route level is estimated by summation of the  $ETT_{i,j}$  of each segment estimated dynamically.

**Algorithm 1: Hybrid travel time estimation procedure at segment level**

- Inputs
- Start time  $Start\_time_{ij}$ , for segment  $j$
  - Weights  $w_{1,j}$  and  $w_{2,j}$
  - The threshold running speed  $THRS_j$  for the segment  $j$
  - The correction ratio  $\alpha_j$ ,  $z_{1,j}$  and  $z_{2,j}$  of each segment  $j$
  - Preceding Trip information details:
    - a) start time  $Start\_time_{k,j}$
    - b) Running speed  $RS_{k,j}$
    - c) travel time  $TT_{k,j}$
- Output Estimated travel time  $ETT_{i,j}$  of a trip  $i$  for the segment  $j$  at  $Start\_time_{ij}$
- Step 1 At  $Start\_time_{ij}$  predict the travel time  $PTT_{i,j}$  using the machine model

Step 2 Check the traffic pattern if congestion is sensed go to step 4 else estimate the travel time  $TRT_{k,j}$  using the  $RS_{k,j}$ , the running speed of the preceding trip

Step 3 Estimate the latest travel time and go to step 5

$$LTT_{i,j} = z_{1,j} * TT_{k,j} + z_{2,j} * TRT_{k,j}$$

Step 4 Estimate the latest travel time

$$LTT_{i,j} = \alpha_j * TT_{k,j}$$

Step 5 Estimate travel time for segment j for the trip i

$$ETT_{i,j} = w_{1,j} * PTT_{i,j} + w_{2,j} * LTT_{i,j}$$

#### Algorithm 2: Dynamic travel time estimation procedure at route level

Input Start time  $Start\_time_i$  of the trip i

Number of segments  $n$

Preceding Trip information details of  $n$  segments along the route:

a) start time  $Start\_time_{k,j}$ , b) travel time  $TT_{k,j}$ , c) Running speed  $RS_{k,j}$

Output Estimated travel time for the route  $RTT_i$  with start time  $Start\_time_i$

Step 1 Initialize the start time of the initial segment  $x = Start\_time_{i,j}$  where  $j=1$

Step 2 Fore cast  $PTT_{x,j}$  for segment  $j$  using the GBRT model with the start time as  $x$

Step 3 Estimate the latest travel time

$$LTT_{i,j} = \left\{ \begin{array}{l} \alpha_j * TT_{k,j} \\ z_{1,j} * TT_{k,j} + z_{2,j} * TRT_{k,j} \end{array} \right\} \left| \begin{array}{l} \text{if } RS_{k,j} < THRS_j \\ \text{otherwise} \end{array} \right\}$$

Step 4 Using the proposed Hybrid model calculate the travel time of the segment  $j$

$$ETT_{x,j} = w_{1,j} * PTT_{x,j} + w_{2,j} * LTT_{x,j}$$

$$ETT_j = ETT_{x,j}$$

Step 5 Estimate the start time of the next segment

$$x = x + ETT_{x,j}$$

$$j = j + 1$$

repeat steps 2 to 5 for each segment

Step 6 Estimate the route travel time at the start time

$$RTT_{sti} = \sum_{j=1}^k ETT_j$$

## 4. RESULTS AND DISCUSSIONS

For estimating the travel time using the proposed HTTEM, the correlation coefficients, the threshold running speed, the correction rate, and the R2 of the GBRT predictions are estimated using the models presented in the previous section, and the numerical results are presented in Table 3. A positive correlation is observed between the preceding to current trip travel times. The threshold speeds for segments vary from 12-25 km/hr indicating the impact of the LUP on travel time. The correction rate varies between 0.95 to 1.1.

Table 3: Coefficients and constants estimated in the study

	$x_t$	$x_s$	$(z_{1,j}, z_{2,j})$	$(w_{1,j}, w_{2,j})$	$\alpha_j$	$THRS_j$
Segment_1	0.52	0.65	(0.46,0.54)	(0.69,0.31)	0.95	12
Segment_2	0.69	0.52	(0.57,0.43)	(0.63,0.33)	1.05	15
Segment_3	0.50	0.46	(0.52,0.48)	(0.65,0.35)	1.1	18
Segment_4	0.49	0.58	(0.44,0.56)	(0.62,0.38)	1.0	25

### 4.1. Predictions at the route and segment levels

The predictions of the GBRT, LTTEM, and the proposed HTTEM against the actual travel time are presented in this section. Algorithm 1 is implemented and the travel times are estimated for each segment. The model predictions for a few samples at various start times of the day of all segments are presented in Figure 3. It is observed that the travel time estimates of the HTTEM are close to the actual as compared to LTTEM and

GBRT models emphasizing the suitability of a hybrid model. Algorithm 2 is implemented to estimate the travel time at the route level. The predictions of each model at the route level are presented in Figure 4. It is observed from the plots that, the predictions of the proposed HTTEM are superior as compared to GBRT and LTTEM highlighting the adaptability of the hybrid model at various times of the day.

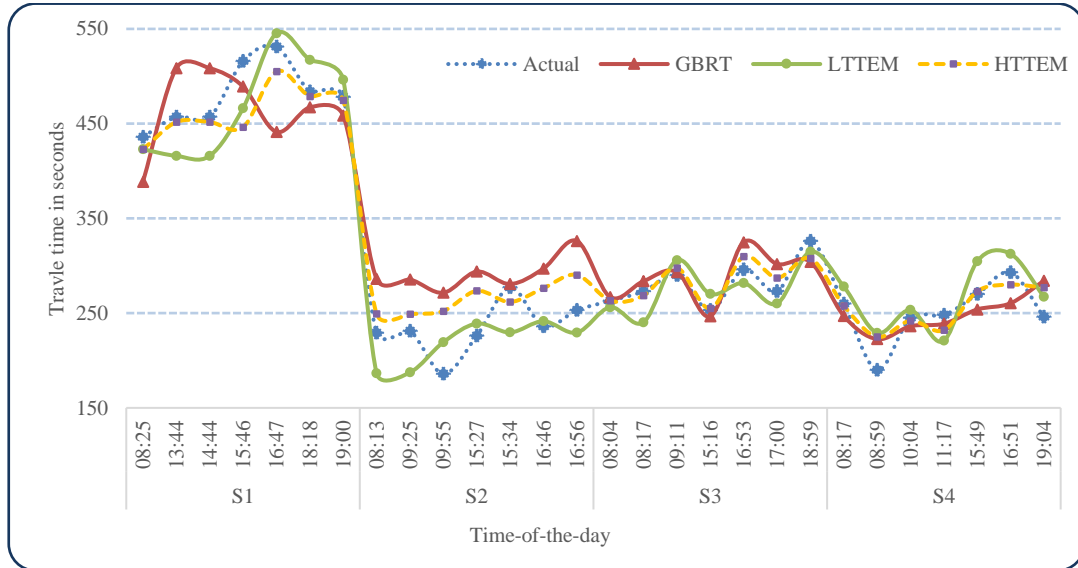


Figure 3. The actual and predicted travel times by GBRT and the proposed models for all segments

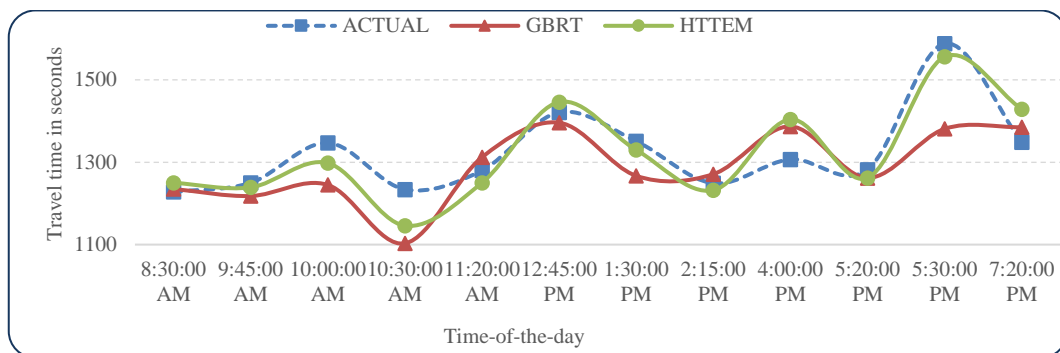


Figure 4. The actual and predicted travel times by GBRT and the proposed HTTEM at the route level

#### 4.2. Model performance

The error in the predictions of the GBRT model, LTTEM, and HTTEM are estimated based on the most commonly used regression error metrics [27] such as median absolute deviation (MAD), mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The errors are summarized in Table 4. It is observed from the errors that the HTTEM outperforms the LTTEM and GBRT models. It is also observed that percentage errors at coarse spatial aggregation i.e., route level is better than at segment level which is a fine aggregate. This indicates that at a higher spatial aggregation level the variations are normalized, and the predictions are better. There is also a possibility that the variations and the delays experienced at one segment by the buses are compensated by the driver during traversing [11] the other segments. For applications like route travel time predictions and timetable [28] generation, predicting travel times at route level spatial aggregation is needed and the proposed model is recommended.

At fine spatial aggregation levels, the impact of the variations is apparent, but this gives deeper insights into the travel times. For applications like bus arrival time prediction at bus stops and other passenger information systems, the predictions at the lower spatial levels are preferred for making precise predictions. Further, to optimize the travel times of a route, if the treatments or solutions are provided to individual segments

will have a better impact as the spatial properties are different for each segment. The predictions made at the segment level can be extended to bus stop levels in the future. Also, as a supplement to the error metrics to demonstrate the performance of the models, two new error scales were used.

- Percentage of trips with less than 60 seconds of error.
- Percentage of trips with less than 30 seconds of error.

The percentage of trips in both error scales is presented in the later columns of Table 4. The results depict that when the HTTEM is used on average 92.5% of trips have errors of less than a minute (60 seconds) at segment level and 66% at route level which are better as compared to GBRT with 84% at segments and 60% at the route level and LTTEM with 86% at the segment and 64% at the route level. Similarly, the errors of the HTTEM are less than 30 seconds during 78.5% of the trips at the segment level and 39% at the route level, respectively, whereas it is 61.5% and 30% for the GBRT model and 68% and 32% for LTTEM. From all the comparisons made in this section, it is evident that the proposed HTTEM outperforms the LTTEM and GBRT models. This also emphasizes the importance of the input data used for modeling. The proposed HTTEM and LTTEM perform better as compared to the GBRT model, as these models are developed based on recent trip information and historic data whereas the GBRT model uses only historic data. This highlights the need for dynamic data for short-term travel time prediction.

Table 4: Prediction errors comparison of the three models

Spatial entity	Error metrics	Error in seconds and percentage			Error scales	Percentage of trips in error scales		
		GBRT	LTTEM	HTTEM		GBRT	LTTEM	HTTEM
Route	MAD	49.0	47.0	48.0	≤ 60 seconds	60%	64%	66%
	MAE	64.0	52.0	52.0				
	RMSE	84.0	72.0	60.0	≤ 30 seconds	30%	32%	39%
	MAPE	5.1%	4.8%	4.0%				
Segment-1	MAD	16.5	17.7	13.7	≤ 60 seconds	82%	81%	88%
	MAE	22.9	18.8	18.4				
	RMSE	31.6	23.7	24.3	≤ 30 seconds	61%	60%	66%
	MAPE	5.4%	4.6%	4.5%				
Segment-2	MAD	27.4	24.1	19.5	≤ 60 seconds	91%	89%	93%
	MAE	32.4	24.5	21.6				
	RMSE	39.4	28.0	26.3	≤ 30 seconds	60%	58%	63%
	MAPE	13.5%	10.1%	9.1%				
Segment-3	MAD	13.4	17.0	13.7	≤ 60 seconds	73%	80%	93%
	MAE	14.9	20.2	13.7				
	RMSE	18.4	25.8	16.0	≤ 30 seconds	62%	85%	91%
	MAPE	5.2%	7.1%	4.9%				
Segment-4	MAD	16.5	19.0	14.6	≤ 60 seconds	89%	92%	96%
	MAE	18.4	20.1	15.6				
	RMSE	21.8	23.5	19.0	≤ 30 seconds	63%	70%	73%
	MAPE	8.0%	8.7%	6.9%				

The authors call attention to the need for real-time road traffic information through sensors, crowdsourced data of commuters [29], passenger counters, and bus passenger demand information to further optimize the dynamic travel time predictions. Multiple data sources can improve the model performance but integrating multiple data sources for real-time applications is again a challenge. Setting the infrastructure for dynamic traffic flow information across the entire road network is also economically not viable in a few countries. The crowdsourced data is a promising and economically viable supplementary data set. The current work is one such case where the regular bus trips are used as the probe vehicle trips and the details of the preceding trip as dynamic traffic information on the roads. Hence for such urban areas and tier-2 cities, the proposed model can be a feasible option with considerable performance to predict travel time.

## 5. CONCLUSION

In this work, a HTTEM for dynamic travel time prediction is proposed. Sample trips of Tumakuru City Service, India during weekdays between mornings 7:00 to evening 20:00 are used for modeling. The location data are processed to trip aggregates at two spatial levels; route and segment, based on land use patterns. The GBRT is trained with historic data, to predict the travel time for the current start time of a trip at both route and segment levels. A LTTEM is developed based on preceding trip details and historical data. The proposed HTTEM dynamically adjusts the results of the GBRT model and the LTTEM. The predictions made by the model were assessed using error metrics MAD, MSE, RMSE, and MAPE and two additional error scales illustrating the percentage of trips that estimate the short-time travel time with less than 60 and 30 seconds of



error. The performance of the proposed HTTEM demonstrated better accuracy at the route and segment level as compared to the GBRT and LTTEM. The authors also emphasize supplementing the location data with crowdsourced data sources as an economically viable option to optimize the dynamic travel time predictions in the future. Overall, it is concluded that, in urban areas and tier-2 cities where there are limited traffic-related data sources, the proposed model that uses the preceding buses as probe vehicles can predict decent results.

## ACKNOWLEDGMENT

The location data of city service buses are given by Tumakuru Smart City Limited, Tumakuru, India. The authors are thankful to the Managing Director and the staff for their support.




## REFERENCES

- [1] H. S. Sudhira and K. V. Gururaja, "Population crunch in India: is it urban or still rural?," *Current Science Association*, vol. 103, no. 1, pp. 37–40, 2012.
- [2] "Road transport year book 2017-18 & 2018-19," *Government of India Ministry of Road Transport & Highways Transport Research Wing*, 2021, [Online]. Available: <https://morth.nic.in/sites/default/files/RTYB-2017-18-2018-19.pdf>.
- [3] F. Zheng, J. Li, H. van Zuylen, X. Liu, and H. Yang, "Urban travel time reliability at different traffic conditions," *Journal of Intelligent Transportation Systems*, vol. 22, no. 2, pp. 106–120, Dec. 2017, doi: 10.1080/15472450.2017.1412829.
- [4] A. Khadhir, B. Anil Kumar, and L. D. Vanajakshi, "Analysis of global positioning system based bus travel time data and its use for advanced public transportation system applications," *Journal of Intelligent Transportation Systems*, vol. 25, no. 1, pp. 58–76, Jan. 2021, doi: 10.1080/15472450.2020.1754818.
- [5] "Tumkur city bus service evaluation report," *Directorate of Urban Land Transport, Government of Karnataka, India*, 2013.
- [6] R. M. Savithamma, R. Sumathi, and H. S. Sudhira, "A comparative analysis of machine learning algorithms in design process of adaptive traffic signal control system," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 12054, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012054.
- [7] D. Desmira, M. A. Hamid, N. A. Bakar, M. Nurtanto, and S. Sunardi, "A smart traffic light using a microcontroller based on the fuzzy logic," *IAES International Journal of Artificial Intelligence (IJ)-(AI)*, vol. 11, no. 3, p. 809, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp809-818.
- [8] Y. Yang *et al.*, "Driving behavior analysis of city buses based on real-time GNSS traces and road information," *Sensors*, vol. 21, no. 3, p. 687, Jan. 2021, doi: 10.3390/s21030687.
- [9] C.-M. Chen, C.-C. Liang, and C.-P. Chu, "Long-term travel time prediction using gradient boosting," *Journal of Intelligent Transportation Systems*, vol. 24, no. 2, pp. 109–124, Mar. 2020, doi: 10.1080/15472450.2018.1542304.
- [10] P. He, G. Jiang, S.-K. Lam, and D. Tang, "Travel-time prediction of bus journey with multiple bus trips," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4192–4205, Nov. 2019, doi: 10.1109/tits.2018.2883342.
- [11] J. Wang and Y. Cao, "Operating time division for a bus route based on the recovery of GPS data," *Journal of Sensors*, vol. 2017, pp. 1–8, 2017, doi: 10.1155/2017/1321237.
- [12] A. Kvišis, A. Zacepins, V. Komasilovs, and M. Munizaga, "Bus arrival time prediction with limited data set using regression models," in *Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems*, 2018, pp. 643–647, doi: 10.5220/0006816306430647.
- [13] V. B. Santos, C. E. S. Pires, D. C. Nascimento, and A. R. M. de Queiroz, "A decision tree ensemble model for predicting bus bunching," *The Computer Journal*, vol. 65, no. 8, pp. 2044–2062, May 2021, doi: 10.1093/comjnl/bxab045.
- [14] M. A. P. Taylor, "Travel time variability-the case of two public modes," *Transportation Science*, vol. 16, no. 4, pp. 507–521, Nov. 1982, doi: 10.1287/trsc.16.4.507.
- [15] A. Chepuri, J. Ramakrishnan, S. Arkatkar, G. Joshi, and S. S. Pulgururtha, "Examining travel time reliability-based-performance indicators for bus routes using GPS - based bus trajectory data in India," *Journal of Transportation Engineering, Part A: Systems*, vol. 144, no. 5, May 2018, doi: 10.1061/jtepbs.0000109.
- [16] E. Wong and A. Khani, "Transit delay estimation using stop-level automated passenger count data," *Journal of Transportation Engineering, Part A: Systems*, vol. 144, no. 3, Mar. 2018, doi: 10.1061/jtepbs.0000118.
- [17] K. E. Zannat and C. F. Choudhury, "Emerging big data sources for public transport planning: a systematic review on current state of art and future research directions," *Journal of the Indian Institute of Science*, vol. 99, no. 4, pp. 601–619, Oct. 2019, doi: 10.1007/s41745-019-00125-9.
- [18] B. T. Thodi, B. R. Chilukuri, and L. Vanajakshi, "An analytical approach to real-time bus signal priority system for isolated intersections," *Journal of Intelligent Transportation Systems*, vol. 26, no. 2, pp. 145–167, Jan. 2021, doi: 10.1080/15472450.2020.1797504.
- [19] A. Comi, A. Nuzzolo, S. Brinchi, and R. Verghini, "Bus travel time variability: some experimental evidences," *Transportation Research Procedia*, vol. 27, pp. 101–108, 2017, doi: 10.1016/j.trpro.2017.12.072.
- [20] H. Zhang, H. Wu, W. Sun, and B. Zheng, "Deep travel: a neural network based travel time estimation model with auxiliary supervision," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul. 2018, pp. 3655–3661, doi: 10.24963/ijcai.2018/508.
- [21] A. K. Bachu, K. K. Reddy, and L. Vanajakshi, "Bus travel time prediction using support vector machines For high variance conditions," *Transport*, vol. 36, no. 3, pp. 221–234, Aug. 2021, doi: 10.3846/transport.2021.15220.
- [22] B. Qiu and W. (David) Fan, "Machine learning based short-term travel time prediction: numerical results and comparative analyses," *Sustainability*, vol. 13, no. 13, p. 7454, Jul. 2021, doi: 10.3390/su13137454.
- [23] B. P. Ashwini, R. Sumathi, and H. S. Sudhira, "Bus travel time prediction: a comparative study of linear and non-linear machine learning models," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 12053, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012053.
- [24] R. B. Sharmila, N. R. Velaga, and P. Choudhary, "Bus arrival time prediction and measure of uncertainties using survival models," *IET Intelligent Transport Systems*, vol. 14, no. 8, pp. 900–907, May 2020, doi: 10.1049/iet-its.2019.0584.
- [25] S. Senthilnathan, "Usefulness of correlation analysis," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3416918.
- [26] G. M. D'Este, R. Zito, and M. A. P. Taylor, "Using GPS to measure traffic system performance," *Computer - Aided Civil and*




- Infrastructure Engineering*, vol. 14, no. 4, pp. 255–265, Jul. 1999, doi: 10.1111/0885-9507.00146.
- [27] A. Botchkarev, “Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio,” *SSRN Electronic Journal*, 2018, doi: 10.2139/ssrn.3177507.
- [28] W. Zhang, D. Xia, T. Liu, Y. Fu, and J. Ma, “Optimization of single-line bus timetables considering time-dependent travel times: a case study of Beijing, China,” *Computers & Industrial Engineering*, vol. 158, p. 107444, Aug. 2021, doi: 10.1016/j.cie.2021.107444.
- [29] R. R. Almassar and A. S. Girsang, “Detection of traffic congestion based on twitter using convolutional neural network model,” *IAES International Journal of Artificial Intelligence (IJ)-(AI)*, vol. 11, no. 4, p. 1448, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1448-1459.

## BIOGRAPHIES OF AUTHORS






**Ashwini Bukanakere Prakash**    is an Assistant Professor in the Department of Computer Science and Engineering (CSE) at Siddaganga Institute of Technology (SIT), Tumakuru, India. She completed her post-graduation from RVCE Bengaluru, Visveswaraya Technological University (VTU) in CSE and is presently pursuing a Ph.D. from VTU. Her research focuses on Machine Learning, Big Data, Intelligent Transportation, and Engineering Education. She has published 10+ papers in reputed conference proceedings and journals. She can be contacted at email: ashvinibp@sit.ac.in.



**Dr. Ranganathaiah Sumathi**    currently working as a Professor in the Department of CSE at SIT Tumakuru, India. She received her Ph.D. from Dr. M.G.R. Educational and Research Institute University, Chennai, India in Wireless Sensor Networks. She has published 45+ articles in reputed conference proceedings and journals. Her research interests are Computer Networks, Artificial Intelligence, Machine Learning, Big Data, Cloud Computing, and Virtualization. She can be contacted at email: rsumathi@sit.ac.in.



**Dr. Honnudike Satyanarayana Sudhira**    obtained his Ph.D. from the Indian Institute of Science, Bangalore for his thesis on “Studies on Urban Sprawl and Spatial Planning Support System for Bangalore, India” He has authored 70+ articles in reputed journals and conference proceedings. Currently, he is the Director of Gubbi labs. He has served as a referee for reputed international journals from publishers including Springer and Elsevier He can be contacted at email: sudhira@gubbilabs.in.