Chapter 7

# On-board GPS (BTSC)

## Introduction

An increase in private-vehicle ownership is one of the primary reasons for pollution and greenhouse gas emissions. According to an article by the European Commission for Sustainable Urban Mobility [1], buses emit 20% less carbon monoxide, 10% fewer hydrocarbons, and 75% less nitrogen oxide per passenger mile compared to automobiles with a single occupant. Day-by-day increasing pollution shows a strong need to promote mass transit. However, as per the Singh [2], the share of public transport in work trips is only 18.1%. One of the key causes for such low share is the lack of comfort. Thus, to promote mass transit, basic conveniences are to be provided. Hence to enhance the comfort in the system, this work targets to improve travel time prediction using a GPS dataset. As shown in Figure 7-1 (a) and (b), passenger buses are mounted with GPS devices to track the speed, distance, elapsed time, and the real-time location in terms of latitude, and longitude. The dataset used for this study is collected from a stretch segmented based on the traffic interruptions at various points in that stretch. The data is prepared by mapping the GPS coordinates of the bus with the road junction to understand the location of the bus in transit and calculating the actual time the bus spent at various road segments. This information is inputted into various regression models such as Support Vector Regressor (Cortes, et al. [3]), Random Forest (Breiman [4]), Gradient Boosting (Friedman [5]) and Extreme Gradient Boosting (Chen, et al. [6]) to make predictions. The tracked information is also stored in the database for updating the model with new train data. We implement two models, one for the prediction of travel time of the whole stretch second for prediction at various links on the stretch. This information can be sent to the bus stop information display board to notify passengers about the current location of the bus and expected arrival time.

**Figure 7-1** (a) GPS Data Logger; (b) Study vehicle used for data collection

(a)

(b)

**Figure 7-2** Study route


### 1: Case Study

A study stretch of 26 km is selected near Hyderabad city on the four-lane divided national highway (NH-65), which is shown in Figure 7-2. The study section is divided into rural and sub-urban categories, which consist of 12 intersections, 13 mid-block openings, and 4 gentle curves. The data is collected for the passenger buses by mounting a high-end GPS data logger, as shown in Figure 7-1. The instrument captures the continuous positional coordinates, speed, distance, acceleration, heading, and slip angle along with the video data at a frequency of 10 Hz. The data collection was carried out for five weeks, ensuring clear and dry weather. Total 92 trip data were collected, comprising 69 travel hours and 2,116 kilometers.

### 2: Data preparation and Preliminary Analysis

The latitude and longitudinal information is used to track the

**Table 7-1** Descriptive statistics on travel time at various junction

| Time/segment | Ⅰ (1.9 km) | | Ⅱ (10.7 km) | | Ⅲ (6.2 km) | | Ⅳ (7.2 km) | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std. | Mean | Std. | Mean | Std. | Mean | Std. |
| Morning peak hours (8:00 am to 11:00 am) | 7.65 | 0.84 | 11.81 | 1.01 | 12.52 | 3.4 | 12.80 | 2.33 |
| Off-peak hours (11:00 am to 5:30 pm) | 5.88 | 3.49 | 12.01 | 0.69 | 9.11 | 1.72 | 11.35 | 2.03 |
| Evening peak hours (5:30 pm to 8:30 pm) | 5.24 | 2.65 | 12.93 | 1.37 | 11.39 | 2.28 | 13.22 | 1.85 |

**Table 7-2** Feature description for travel time prediction

| Parameter | Description |
|---|---|
| Segment id | Each segment is assigned with a unique id |
| Distance | Segment length |
| Direction | 0 and 1 for opposite directions |
| Time | Travel hours are divided into Morning/Afternoon/ Evening |
| Weekday/ weekend | 1 for the weekend, 0 for weekday |
| Prev_segment_tt | Travel time required for previous segment |

**Table 7-3** Feature importance and rank

| Parameter | Rank | Importance |
|---|---|---|
| Distance | 1 | 58.9% |
| Prev segment_tt | 2 | 15.22% |
| Time | 3 | 10.2% |
| Segment id | 4 | 7.3% |
| Weekday weekend | 5 | 4.27% |
| Direction | 6 | 3.97% |

bus over the selected route and at various junctions on the route. To account the effect of different traffic conditions on bus travel time, the study stretch is divided into segments characterized by important junctions on the route, namely IITH Main Road (Ⅰ), Isnapur (Ⅱ), Patancheru (Ⅲ), and BHEL-X-Road (Ⅳ) (Figure 7-2). The parameters used in the models (Table 7-1) are selected based on the travel time variations observed. The notable effect of time of travel can be discerned from the statistics thus, the time of travel is considered a parameter for predicting travel time. Travel time also varies concerning the day of travel being weekday or weekend. As there is no significant trend observed on different weekdays, a parameter called weekday/weekend is used to differentiate between weekdays and weekends. The direction of travel is also considered as a parameter to account the travel time variations due to changes in traffic conditions with respect to the direction of travel. For link travel time, the study stretch is segmented, and a unique ID is assigned to each segment. Along with the selected parameters, the segment ID, distance, and the time required to travel the previous.

## Model Evaluation

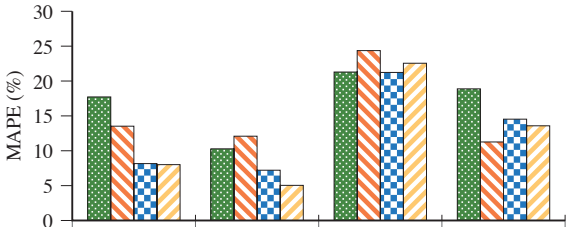
Input parameters/predictor variables have very high importance on the accuracy of the model. Table 7-2 and Table 7-3 shows the importance and ranking of the features used as input parameters. In the link travel time prediction, the feature importance of distance is evident as it is directly proportional to travel time. However, it is interesting to see the impact of travel time of the previous segment on the current segment. Other than these, time of travel also seemed to be an important feature.

As mentioned earlier, the error function used to evaluate the accuracy of models is MAPE. We also use RMSE to evaluate the performance of models. The comparative analysis of the four methods discussed in earlier sections is shown in Figure 7-3 and Figure 7-4. Figure 7-3 shows the performance of models in all the segments. In segments (links) Ⅰ, Ⅱ, and Ⅳ, we can observe a clear improvement in accuracy using XGBoost. In segment Ⅲ, we can observe that SVR and GB outperform RF and XGBoost, while XGBoost gives a better overall performance, as shown in Figure 7-4.

## Conclusion

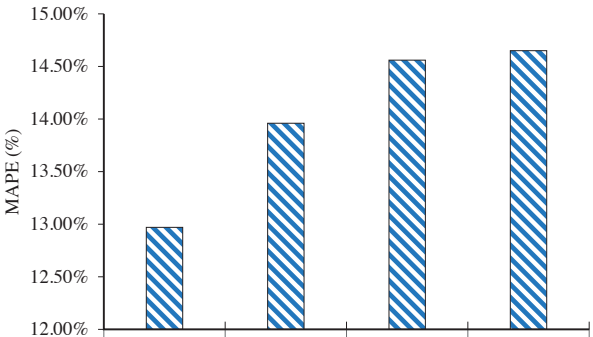
As mentioned in the methodology section, ensemble methods have the power to combine several weak learners to form one strong learner. Gradient Boosting and Random Forest work based upon this principle and thus are stronger than single models like Support Vector Regressor. Using the principles of Gradient Boosting, XGBoost is built with additional custom regularization terms, which makes it more powerful. As per the current literature, less amount of work has been done using the ensemble models for the prediction of the arrival time of buses, especially for Indian traffic conditions. Some literature talks about the comparison of Random Forest and Gradient Boosting for travel time prediction on time series data (Zhang, et al. [7], Gupta et al., [8]). As per the mentioned studies, Gradient Boosting and Random Forest both performed well in predicting travel time. However, Gradient Boosting was able to understand the complex relationship between the parameters giving an overall superior performance to Random Forest. As per the best knowledge of the authors, Extreme

**Figure 7-3** Results for travel time (70%–30% train-test division)




| | I | II | III | IV |
|---|---|---|---|---|
| SVR | 17.72 | 10.28 | 21.289 | 18.88 |
| RF | 13.524 | 12.088 | 24.364 | 11.261 |
| GB | 8.16 | 7.21 | 21.231 | 14.53 |
| XGBoost | 8.01 | 5.038 | 22.55 | 13.58 |

**Figure 7-4** Results for travel time prediction (10-fold cross-validation)




| | XGBoost | GB | RF | SVR |
|---|---|---|---|---|
| MAPE | 12.97% | 13.96% | 14.56% | 14.65% |

Gradient Boosting has not been evaluated in the present studies on GPS data for bus arrival time prediction. In this study, we explored XGBoost to model travel time on GPS data using various parameters such as weekday, time of travel, road segment on which the bus is running, the direction of running, and travel time of the previous segment, etc. to implicitly learn the traffic patterns. Incorporating the mentioned parameters, the Extreme Gradient Boosting is found to predict significantly better than other benchmark models such as Gradient Boosting Machine, Random Forest, and Support Vector Regression with MAPE between 5.03% to 22.55% for link travel time prediction and 7.35% for prediction at a complete stretch. One issue with Extreme Gradient Boosting is related to optimization of parameters. As discussed in the model-optimization section, the results given by the model are significantly affected by the parameters. The tradeoff between time for training and accuracy of the model is to be considered while selecting the parameters. To summarize, XGBoost has potential to work considerably well in predicting travel time on highways. In recent times when numerous data collection sources are available, it is not trivial to find a specific model that works well on every kind of data. However, ensemble models have capability to combine multiple diverse models to form a strong model. Powerful model like XGBoost can prove very influential in such cases. In this study, the indirect parameters like weekday and time are considered along with time taken on previous segments to address spatial correlation. Exploring the impact of temporal correlation on prediction accuracy by considering time series data along with spatial correlation is the future scope of this work.

## References

[1] European Court of Auditors: Sustainable Urban Mobility in the EU: No Substantial Improvement Is Possible Without Member States' Commitment. https://www.eca.europa.eu/Lists/ECADocuments/SR20_06/SR_Sustainable_Urban_Mobility_EN.pdf, 2020.
[2] Singh, J.: City Public Transportation Development in India. https://www.intelligenttransport.com/transport-articles/21458/city-public-transportation-india/, Retrieved from 2016.
[3] Cortes, C., and V., Vapnik: Support Vector Machine. Machine learning, Vol. 20, No.3, pp.273–297, 1995.
[4] Breiman, L.: Random Forests. Machine learning, Vol.45, No.1, pp.5–32, 2001.
[5] Friedman, J. H.: Stochastic Gradient Boosting. Computational Statistics and Data Analysis, Vol. 38, No.4, pp.367–378, 2002.
[6] Chen, T., & C., Guestrin: XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM KDD International Conference on Knowledge Discovery and Data Mining, pp.785–794, 2016.
[7] Zhang, Y., and A., Haghani: A Gradient Boosting Method to Improve Travel Time Prediction. Transportation Research Part C: Emerging Technologies, Vol.58, pp.308–324, 2015.
[8] Gupta, B., S. Awasthi, R. Gupta, L. Ram, P. Kumar, B. R. Prasad, and S. Agarwal: Taxi Travel Time Prediction Using Ensemble-based Random Forest and Gradient Boosting Model. In Advances in Big Data and Cloud Computing, Springer, Singapore, pp.63–78, 2018.