




Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: [www.jsoftcivil.com](http://www.jsoftcivil.com)



## Evaluation of Applicability and Accuracy of Bus Travel Time Prediction in High and Low Frequency Bus Routes Using Tree-Based ML Techniques

Seyed Mohammad Hossein Moosavi<sup>1</sup>, Mahdi Aghaabbasi<sup>2\*</sup>, Choon Wah Yuen<sup>3</sup>,  
Danial Jahed Armaghani<sup>4\*</sup> 

1. Research Fellow, Centre for Transportation Research, Faculty of Engineering, University of Malaya, 50603, Kuala Lumpur, Malaysia

2. Researcher, Transportation Institute, Chulalongkorn University, Bangkok, 10330, Thailand

3. Associate Professor, Department of Civil Engineering, Faculty of Engineering, University of Malaya, 50603, Kuala Lumpur, Malaysia

4. Centre of Tropical Geoengineering (GEOTROPIK), Institute of Smart Infrastructure and Innovative Engineering (ISIIC), Faculty of Civil Engineering, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia

\*Corresponding author: [mahdi.a@chula.ac.th](mailto:mahdi.a@chula.ac.th) (M.A.); [jadaniel@utm.my](mailto:jadaniel@utm.my) (D.J.A.)

 <https://doi.org/10.22115/SCCE.2023.356348.1503>

### ARTICLE INFO

Article history:

Received: 16 August 2022

Revised: 29 December 2022

Accepted: 31 January 2023

Keywords:

Machine learning;

Tree-based models;

High-frequency route;

Low-frequency service;

Travel time prediction.

### ABSTRACT

Prediction of bus travel time is a key component of an intelligent transportation system and has many benefits for both service users and providers. Although there is a rich literature on bus travel prediction, some limitations can still be observed. First, high-frequency and low-frequency bus routes have different characterizations in both operational and passenger behavior aspects. Therefore, it is highly expected that bus travel time prediction methods for different frequencies must have different outputs. Second, in the era of big data, applications of machine learning (ML) techniques in travel time prediction have significantly increased. However, there is no single ML model introduced in the literature that is the most accurate in predicting bus travel, especially with regard to bus service frequency. Consequently, the main objective of this study is to determine the most applicable route construction approach and most accurate tree-based ML technique for predicting bus travel time on high- and low-frequency bus routes. The following tree-based ML techniques were adopted in this study: chi-square automatic interaction detection (CHAID), random forest (RF), and gradient-boosted tree (GBT). According to the results, CHAID was selected as the most accurate model for predicting travel time on high-frequency routes, while GBT showed the best performance for low-frequency service. CHAID analysis identified distance between stops and terminal departure behavior as the most significant factors of travel time on high-frequency routes. Moreover, we introduced the "key stop-based" route construction method for the first time, which is an accurate, reliable, and applicable method.

How to cite this article: Moosavi SMH, Aghaabbasi M, Yuen CW, Jahed Armaghani D. Evaluation of Applicability and accuracy of bus travel time prediction in high and low frequency bus routes using tree-based ML techniques. *J Soft Comput Civ Eng* 2023;7(2):74–97. <https://doi.org/10.22115/scce.2023.356348.1503>

2588-2872/ © 2023 The Authors. Published by Pouyan Press.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



## **1. Introduction**

Due to rapid growth in economic development and urbanization in last two decades, most of the cities are facing with traffic congestion and air/noise pollution around the world. Moreover, people life style is changed and many physical and mental health issues have arisen due to sedentary lifestyle. Public transport sharing in Kuala Lumpur is almost 20% and traffic congestion, air pollution, road accidents and also obesity between residents due to inactive lifestyle are direct and indirect outcomes of using private transport [1,2]. Moreover, there is no doubt that COVID-19 has a massive impact on the world. One industry that has had the earliest significant and unique impact is the transport infrastructure industry. Organizations will not only need to adapt to this new reality, but it is also a challenge to plan on how this industry can evolve stronger going forward, by enabling new ways of working that are safer and more efficient [3]. A study by in Chicago found that COVID-19 has a large impact on transit ridership and there have been a strong association ridership and numbers of COVID-19 cases and deaths as well. Despite this matter, the pandemic readiness and response frameworks needs the involvement of transport agencies; the role of those organizations, through the provision of situation knowledge and analysis, information-sharing and monitoring (for instance, fast-acting rumors) recognize vital transport functions and particulate matter was highlighted by these organizations, in maintaining good and productive ties with all stakeholders (emergency services, public health services, vendors and end users). Furthermore, public transport is also an essential service or front liners, to provide best mobility in times of pandemics as an accessible service to health care facilities. Therefore, all organizations across the world will need to adapt and evolve on a global scale, as the status quo will leave them ill equipped to tackle new paradigms in the future.

Developing an Intelligent Public Transportation System (IPTS) can be the most effective and economic solution to overcome this problem. However, establishing a reliable and attractive IPTS in order to make it the primary travel mode for commuters is a great challenge for authorities and service providers [4].

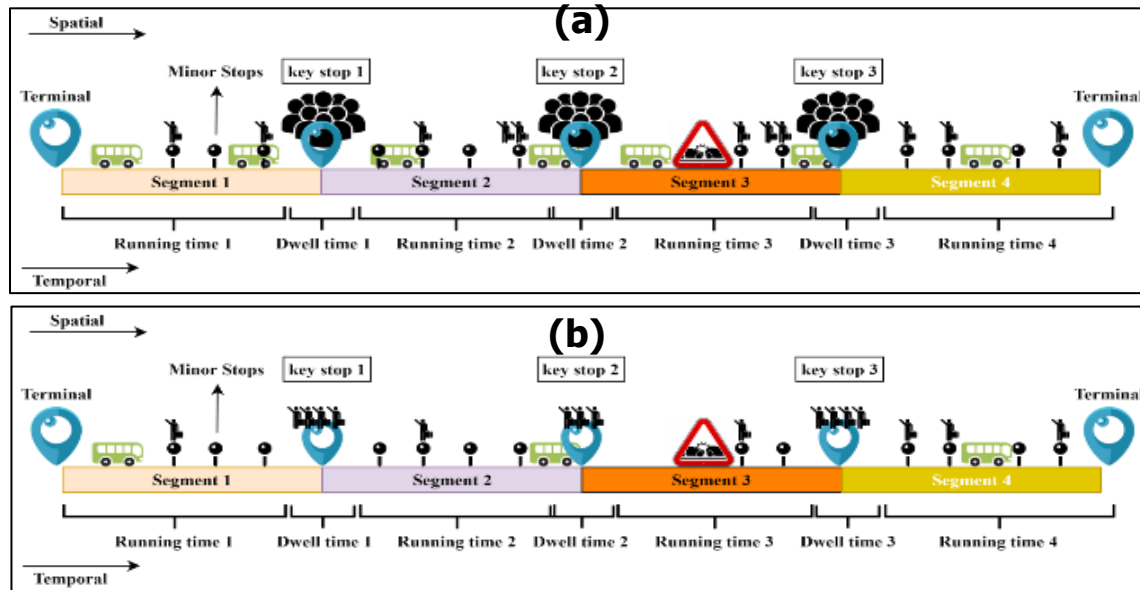
Travel time prediction accuracy is of great importance for both bus service providers and passengers. Since we are living in real time and big data era, providing reliable prediction of travel time is vital to maximize the advantages of these technologies [5]. Moreover, accurate prediction of travel time is essential in order to develop an Intelligent Transportation System. From perspective of passengers, providing accurate travel time/arrival time is one the most important indicator of a reliable and attractive bus service [6]. Implementation of emerging technologies in public transportation such as new automatic data collection and real-time tracking systems have created new era in transportation engineering and quality control. Big Data and Smart Data have provided new opportunities for service providers to enhance the reliability and attractiveness of bus service [7]. In public transportation sector, Automatic Data Collection Systems (ADCS), which record data every few seconds, are the best examples of Big Data sources. Accordingly, there is considerable number of researches on travel time prediction. However, significant gaps still can be observed.

Below is a comprehensive discussion on recognized gaps/limitations and main contributions of our proposed methods:

First, there are two types of bus service with respect to the service frequency. Bus service is considered as Low-frequency when schedule headway is more than 10 minutes, and High-frequency when headways are equal to or less than 10 minutes. Scheduled headway is not the only difference between high and low frequency bus services. According to literature, these two types of bus service have different characterization in both operational and passengers' behavior aspects [8–12]. Therefore, we expected that bus travel time prediction method and accuracy should be significantly related to bus service frequency. However, to the best of our knowledge, there is no study currently available that considers and compares travel time prediction in different bus service frequencies. Therefore, we selected two different bus routes with high and low frequency services and employed various machine learning techniques on each of them separately. The results were compared to clearly understand the most suitable and accurate approach for predicting bus travel time in high and low frequency bus routes.

Second, machine learning and Traffic theory-based approaches are popular methods among researches for predicting bus travel time [13,14]. After careful consideration of different methods, we concluded that Machine Learning is more appropriate approach for accurate prediction of travel time, in presence of Big Data (explanation on different methods is provided in section 2). However, according to available literature, it is not evident yet which Machine Learning method is the most accurate for predicting bus travel. Therefore, we designed this study to shed some light on this issue by conducting and comparing the most common tree-based machine learning methods. In addition, Chi-square automatic interaction detection (CHAID) method has strong capability to determine the relation between independent variables and target variable, which highly matches our needs to predict travel time. However, this machine learning method is neglected in previous studies. Therefore, we employed CHAID technique for the first time to predict bus travel time and compared the outputs with other machine learning models.

Third, usually bus routes are too long that researchers divide them to shorter segments for analyzing and predicting travel time. It has been claimed that route construction methods increase the accuracy of travel time prediction by considering more accurate and detailed information. Linked-based and stop-based are two route construction approaches which have been used widely. Linked-based method constructs the route based on important intersection along the route, while Stop-based method divides the route based on bus stops. Recently, Ma et al. [15] proposed a segment-based route construction method which divided the route to transit and dwelling segments. Based on our findings, dividing bus route to transit (segment running time) and dwelling was first proposed by Milkovits [16]. However, Milkovits approach was much simpler and more applicable that many agencies are still using this method to analyzing bus service performance. He divided the bus route to segments based on “key-stops”, then analyzed the dwell times only at key stops and running times for segments between two key stop, as shown in Figure 1. Although, “key stop-based” route construction approach has been used for analyzing the performance of bus service, but studies which used this method for predicting bus travel time hardly can be found. As mentioned before, this approach is one of the most applicable methods with acceptable level of detail and considerations, which perfectly suits our objectives (applicability and accuracy). It is important to note that machine learning models have solved many transportation and civil engineering problems as well [17–33].



**Fig. 1.** Key stop-based route layout: a) High-frequency route, b) Low-frequency route.

This paper aims to determine the most applicable route construction approach and most accurate tree-based machine learning technique for predicting bus travel time in high and low frequency bus routes, separately. Accordingly, below is a list of contributions:

1. As discussed earlier, there are significant differences between high-frequency and low-frequency bus routes. To the best of author's knowledge, this is the first study to assess bus travel time prediction accuracy considering service frequencies.
2. Also, this is study to evaluate and compare the accuracy of tree-based ML techniques for predicting bus travel time. The literature on application of tree-based ML techniques in prediction of travel time is still shallow. This is the first study to analyze and compare these ML algorithms in this context. In addition, we employed CHAID technique for the first time to predict bus travel time and compared the outputs with other machine learning models.
- 3- Route construction methods increase the accuracy of travel time prediction by considering more accurate and detailed information. We proposed "key stop-based" route construction method for the first time, which is an accurate, reliable and applicable method.

The remaining of the paper is structured as follows: Section 2 presents a literature review on parametric, non-parametric methods for predicting bus travel time and factors influencing bus travel prediction. Literature review section followed by Methodology, Results and Discussion. The final section concludes the main findings and suggests future directions.

## 2. Related works

There are considerable numbers of study which evaluated and proposed travel time prediction models for bus service. Jairam et al. [13] categorized the approaches to predict bus travel time in to two main categories: Model-based approach and data-driven methods. He classified all machine learning models, linear regression, time series analysis and filter techniques under

category of Model-based approach. Ma et al. [15] in an inserting study classified travel time prediction model into six popular models: Historical mean method, Regression, K nearest neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Kalman filtering. In this study, we try to present a comprehensive discussion and review on all available travel time prediction models and approaches as summarized in Figure 2. In addition, brief discussion on advantages and disadvantages of each model is presented in this section.

Simple approaches are the easiest and fastest approaches for prediction of travel time. Instantaneous, historical average and hybrid models are common models in this category. Although these methods are simple, they have restrictive assumptions and weakness which make them unreliable. These methods are not recommended for travel time prediction and they are not in scope of our study.

Data-based approaches were widely used in the travel time prediction studies. This category includes various models with various applications. Basically, models in this approach evaluate and develop function between independent variables and target variable. This function is obtained from big data sources using regression models or machine learning methods, instead of historical average method. In a general categorization, data-based models can be divided into two main categories: Parametric and Non-parametric methods.

Traffic-based models have considerable advantages such as detailed information on traffic condition in typical and atypical situations and buses trajectory on a route during specific time of a day [34,35]. However, implementing these models requires deep knowledge of traffic theories and high mathematical and/or programming skills. Moreover, for predicting accurate travel time using this method, simulated/recreated traffic flow should be almost the same as real condition, which is very complicated and time consuming [14,36].

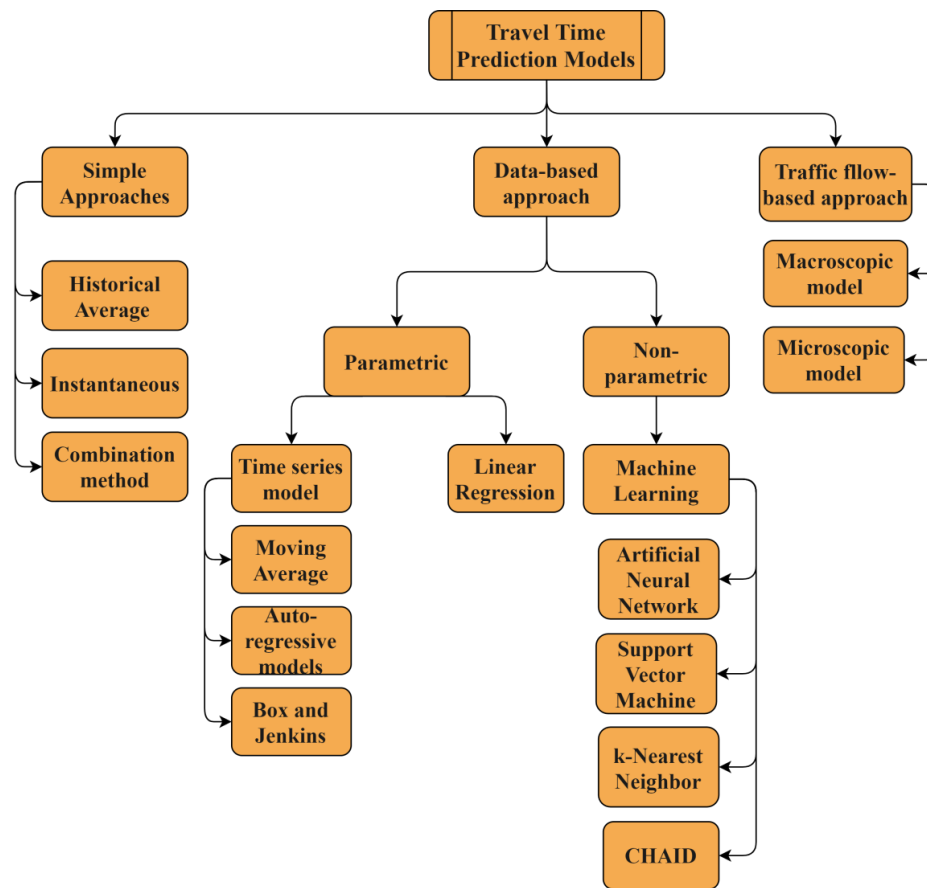
In an interesting researcher, authors studies and compared three different types of data fusion including: the artificial intelligence-based method, probability-based method and the evidence theory-based method [37]. Furthermore, Machine learning methods have been adopted widely in order to understanding complex behaviors. As an example, a combination of support vector machine (SVM) and the group method of data handling (GMDH)-type neural network and the grasshopper optimization algorithm (GOA)) were adopted to figure out the number of vehicles involved in an accident [32].

## 2.1. Parametric methods for travel time prediction

Regression techniques [38,39] and time series models [40,41] are most popular parametric for prediction of travel time. Travel time prediction using parametric methods, structure of function between target variable and independent variables must be fully predetermined. Generally speaking, regression models predict effect of various factors on travel time in form of an equation. In this regard, selected factors must have considerable impact on travel time and be independent of each other [42]. Bus travel time is affected by many factors such passenger behaviors, driver experience and Terminal Departure Devotions (TDD). Understanding and evaluating these factors can significantly increase the accuracy of travel time prediction. Therefore, we decided to discuss on this matter separately in section 2.3.

## 2.2. Non-parametric methods for travel time prediction

As explained above, in regression models selected factors for predicting travel time must be independent from each other. This is the main reason that researchers do not recommend these models for travel time prediction, because variables in traffic and transportation systems are deeply inter-correlated. Structure of Non-parametric models is not predetermined and must be obtained from the data. Non-parametric approaches such Machine Learning methods have been widely used in bus travel time prediction studies. Machine learning models have this capability to determine the non-linear relationships between independent variables and target variable, in a complex system with noise data. This can be the main reason of popularity of these models for prediction of bus travel time. Moreover, Machine learning models are able to accurately predict the bus travel time, without explicitly modeling and integrating traffic flow. More discussion on function of machine learning models will be presented in Methodology section.



**Fig. 2.** Overview of travel time prediction models.

Artificial Neural Network (ANN) is the most famous Machine Learning model for predicting the bus travel time [43–46]. ANN method is able to determine nonlinear complex relationships, which is suitable for bus routes and networks. Chien et al. [43] used ANN model to predict bus travel time for both stop-based and link-based route constructions methods, and both two ANN-based methods showed acceptable results. Gurmu and Fan [47] developed ANN-based model using time-tagged GPS data to predict bus travel time. They also proved that ANN-based models outperformed the regression and historical average models. Decision trees [38], local regression

[48] and Support Vector Machine (SVM) [49–52] are other non-parametric models which have used to predict bus travel time in the future. SVM approach is able to use kernel function and map the data sets into higher dimension to find the fittest linear relationship between input vectors and dependent variable [44,51,53,54].

As mentioned before, the literature on application of tree-based ML techniques in bus travel time prediction is still shallow and studies in this context hardly can be found. Moreover, there is no literature available on performance of CHAID ML technique for bus travel time prediction.

### 2.3 Factor Affecting Bus Travel Time

High frequency bus routes are more sensitive to variations and trigger factors, due to higher passenger demand and shorter scheduled headway comparing to low frequency routes. In addition, in high-frequency bus routes, passengers tend to neglect the schedule and arrive at bus stops randomly. Therefore, providing accurate bus travel/arrival time for commuters can make a high-frequency bus service more reliable and attractive. It can be concluded, that predicting travel time in high-frequency bus routes is more challenging comparing to low frequency routes. Accordingly, to accurately predict the travel time we must clearly understand the impact of various factors on high and low frequencies bus service. Bus services are very unstable and they become easily irregular when an internal or external factor affects the service [55]. Woodhull [56], in a very interesting and fundamental study, divided the factors effecting the bus service regularity in to two main categories: external (exogenous) or internal (endogenous).

**Table 1**  
Factor Affecting Bus Travel Time Prediction Accuracy.

Factor	Description	Source
<b>Dwell time</b>	Time consumed by passengers alighting and boarding (sec)	APC, AFC
Boarding	Times consumed by passenger boardings (sec)	APC, AFC
Alighting	Times consumed by passenger alightings (sec)	APC, AFC
On-board load	On-board passengers more than 100% of bus capacity	APC, AFC
TDD	The delay from schedule departure time from the terminal in the studied segment (sec)	AVL
Driver Exp	Working experience of driver (in year)	Archive
Delay	the amount of service deviation from time table (sec)	AVL
AM/PM peak	A dummy variable that is equal to one if the run time is observed in peak hours zero otherwise: temporal variation of peak and off-peak	AVL, AFC
Lift	Use of wheelchair ramp for disable passengers (sec)	APC
<b>Running Time</b>	Time for travelling between two key stops or time points	AVL, AFC
Distance	Length of segment or actual distance between two key stops	Maps
No of stops	Actual number of stops between two key stops or time points	Maps
Boarding	The number of passengers boarding at the studied key stop or segment	APC, AFC
Alighting	The number of passengers alighting at the studied key stop or segment	APC, AFC
TDD	The delay from schedule departure time from the terminal	AVL
Delay	the amount of service deviation from time table (sec)	AVL, AFC
Driver experience	Working experience of driver (in year)	Archive
AM/PM peak	A dummy variable that is equal to one if the run time is observed in peak hours zero otherwise: temporal variation of peak and off-peak	AVL, AFC
Average load	Average onboard passengers during the studied run time	AFC, APC
Speed	Average speed of vehicle movement on segment	AVL

Following Woodhull study, many researches have been conducted on internal factors such as variation in passenger demand [57,58], Terminal Departure Deviation (TDD) [7], passenger boarding/alighting behaviours [59,60]. External causes of unreliability also have been evaluated by number of searchers: traffic congestion and accidents [61,62], impact of AM/PM peak hours [51,63,64] and adverse weather [65–68]. These factors are the main causes of inaccuracy and unreliability of travel time prediction models. Many researchers have evaluated the impact of these factors in prediction of bus travel time. However, the impact of these factors has not been included in many bus travel time prediction models due to unpredictable nature of these factors. Accordingly, after careful consideration of previous studies, we decided to include factors in Table 1, in order to predict the travel time in high and low frequency bus routes. Moreover, since applicability is one of the main objective of this study, all these factors were discussed with IT department of RapidKL Bus Company, to confirm the availability and accessibility of them.

### 3. Methods

This section presents an overview of the methodology of this study as shown in Figure 3. In the first step, the overview of data collection, route specifications and input acquisition will be discussed. Next, “key-stop-based” route construction approach for bus travel time prediction will be presented. Finally, Machine learning methods and output evaluation will be briefly described.

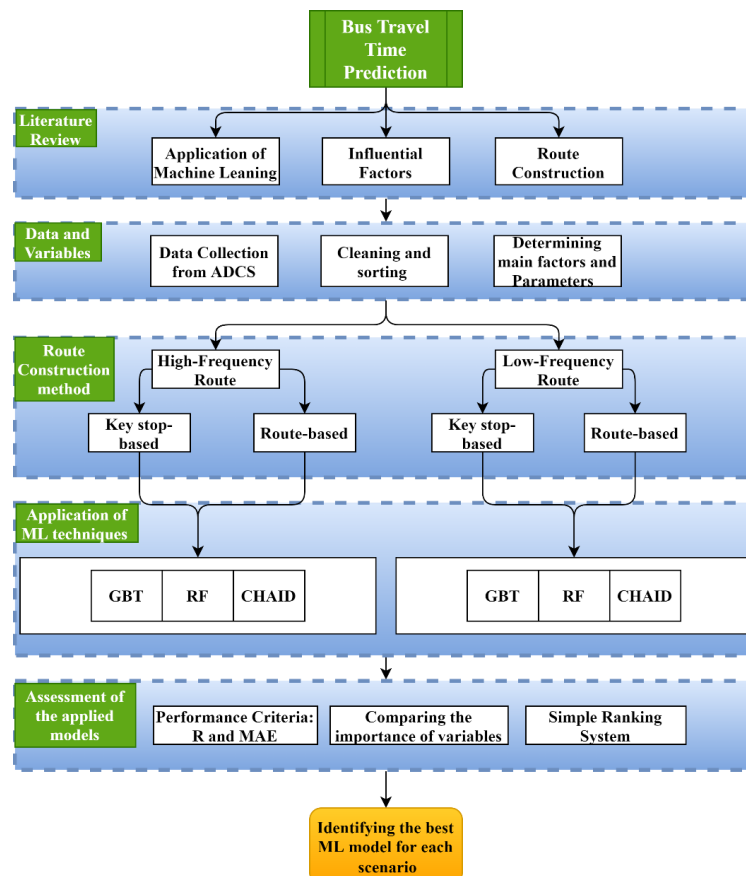
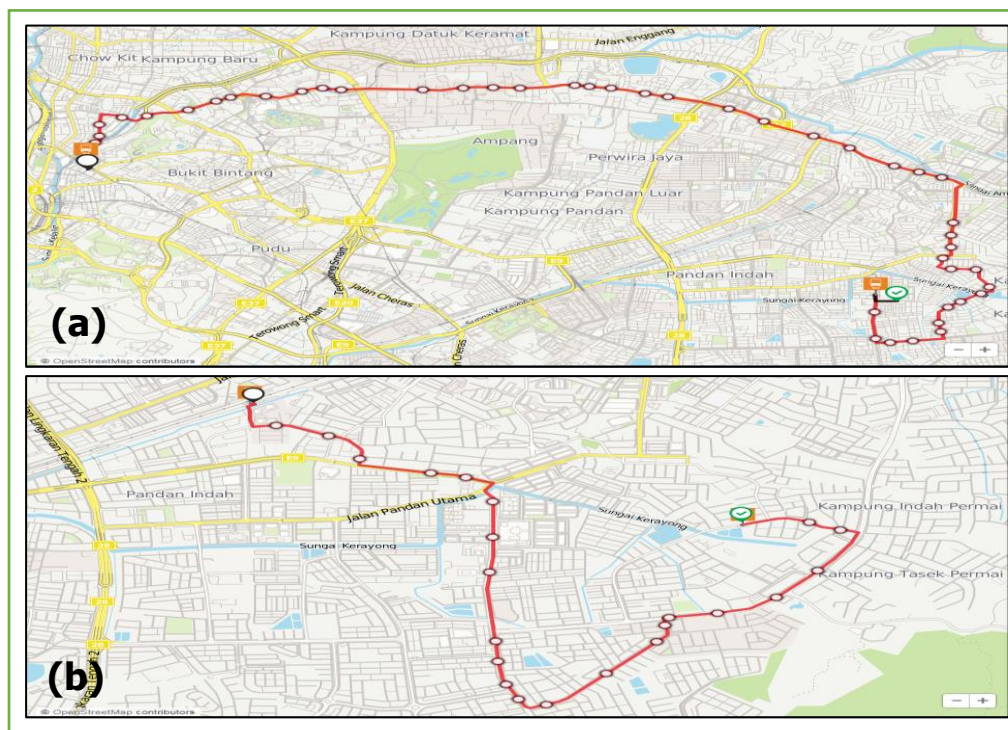


Fig. 3. Overview of methodology.



### 3.1. Data collection and input acquisition

The main source of data for this study was collected from ADCS which belongs to “RapidKL Bus Company”. RapidKL is a half-governmental and half-private public transport company which has been established to provide sustainable public transport service in Kuala Lumpur area (capital of Malaysia). AVL system records time-tagged bus location data every 5 seconds. Raw AVL data can be converted into departure times, arrival times, segment running times and finally the route travel times, using geofencing techniques. AFC system provides a rich data set on each transaction includes time, date, location (bus stop) and passenger specifications such as gender, age and even occupation. Opening/closing time of bus doors and number of passengers boarding and alighting are recorded by APC system. According to the objective of this study, a high frequency and one low frequency bus route were selected in two different zones of Kuala Lumpur (Figure 4).



**Fig. 4.** Layout of selected routes for this study: (a) Route U32 (High-frequency), and (b) Route T350 (Low-Frequency).

Extracting, cleaning and integrating these data sets is the most challenging and critical step to building up a big and smart data source for analyzing and prediction accurate travel time. Bad quality of data (noisy and dirty data) can significantly affect the accuracy of travel time prediction models. Therefore, outlier detection [69] and missing data treatment techniques [70] were applied, before processing data to the next step. Figure 5 illustrates the overview of data collection and input acquisition which includes three main steps, as discussed above. In addition, Table 2 presents the initial descriptive analysis on input variables.

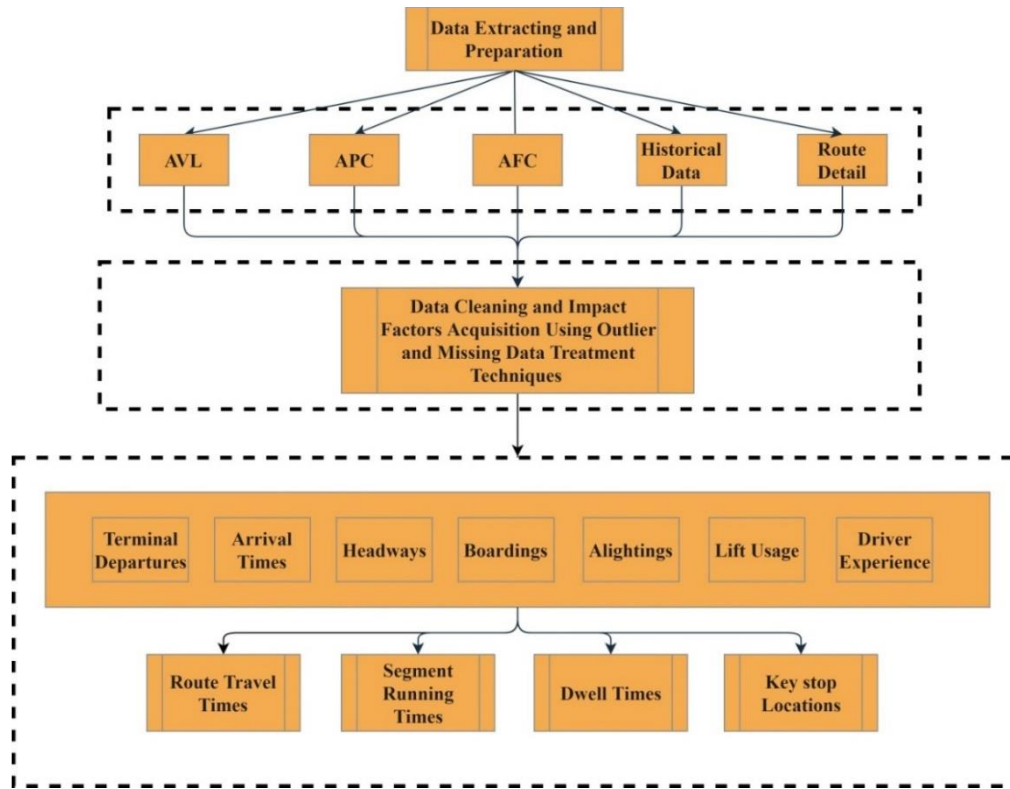


Fig. 5. Overview of data collection procedure and input acquisition.

### 3.2. Route construction approach

Route construction methods were comprehensively discussed in introduction section. As mentioned earlier, regarding to main objective of this study we decided to adopt Key stop-based route construction approach to predict the travel time. Dividing route to shorter segments based on key stops was proposed by Milkovits [16] and many researchers have used this approach to study the bus service performance. However, to the best of our knowledge, this study is the first attempts to adopt this approach for bus travel time prediction propose. In order to adopt this method, first step is to determine the key stops based on passenger demand and/or strategic locations such as interchange station. Afterward, route is divided to the running time and dwell time segments. As illustrated in Figure 1, running time segment is distance between two key stop (Terminals are considered as the first and last key stops). Dwell time is modeled separately only in key stops and dwell times related to minor stops are considered in running time model as total number of passenger boarding and alighting along a segment.

In high-frequency bus routes, there must be enough number of buses in service to maintain the 10 minute headways frequency. Therefore, it can be expected that accurate buses trajectory should be estimated only by relying on bus data (AVL, AFC and APC). Accordingly, we decided to examine travel time prediction without considering any specific route construction. In this context, the route is considered as one long segment and all the variables in Table 2 is included in the model for the whole route (such as speed, total number of passengers boarding and alighting for one trip along the route). As the summary, we predicted bus travel time under two scenarios: Key stop-based construction approach and route-based approach (no specific route construction).

**Table 2**

Tables of Variables.

Variable	Unit	Minimum	Maximum	Average	St. Deviation
<i>Dwell time</i>					
Load	-	13	51	33.06	8.72
Boarding	-	8	33	18.25	6.03
Alighting	-	1	40	19.32	9.46
*TDD	Sec	0	300	97.27	76.90
Delay	Sec	0	610	195.31	130.11
Driver experience	Year	1	13	6.20	3.84
Lift	Sec	0	1	0.11	0.32
AM/PM peak	-	-	-	-	-
<i>Running time</i>					
Distance	Km	1.65	5.7	3.42	1.66
No of stops	-	2	19	8.83	6.78
Boarding	-	2	36	13.55	8.67
Alighting	-	1	25	12.20	6.22
TDD	Sec	0	300	99.78	79.06
Delay	Sec	17	702	220.45	159.37
Driver experience	Year	1	13	6.23	3.86
AM/PM peak	-	-	-	-	-
Average load	-	12	44	25.31	7.45
Speed	Km/h	0	85	53.80	9.50
*TDD=Terminal Departure Deviation					

### 3.3. Machine learning techniques

This study compares the performance of various tree-based ML techniques to predict the bus Travel Time (TT) while they are applied on two route construction approaches under two different routes' frequency (as shown in Figure 3). As explained earlier, three ML techniques are used in this study in order to predicting the travel times, including Random Forest (RF), Gradient boosted trees (GB) and Chi-square Automatic Interaction Detection (CHAID). Kass [71] developed CHAID algorithm, which belongs to the decision tree-based (DT) models. This method is able to produce a non-binary tree structure. CHAID enjoys a series of Chi-Square tests for creating multiple sequential combinations, splits and finally a single DT. While some DT techniques such as CART are vulnerable to overfitting, the CHAID is able to prune automatically the tree which reduces the likelihood of overfitting. Besides, many rule-sets can be produced by the CHAID, and each rule may own a confidence level and accuracy.

GBT is a tree-based algorithm which is based on principle of boosting. It is a combination of models with high bias and low variance error with the purpose to lower down the bias and at the same time maintaining low variance. Boosting is the process where it learns several classifiers by altering the sample weight during each training process and these classifiers are combined linearly to enhance the performance of the classification, unlike other tree-based methods, deep trees and different training datasets are not used in boosting. The boosting trees construct shallow trees that are trained in the similar dataset but each tree is specialized in a specific feature of the

relationship between input and output. Successive shallow trees are trained in series with the objectives of (n)<sup>th</sup> tree is trained to reduce the prediction errors from the previous (n-1)<sup>th</sup> trees.

The objective of GBT is to form an additive model that minimises the loss function. The process of GBT method is as follow:

- 1) The model is beginning with a constant value that minimises the loss function.
- 2) At each iterative training process, the negative gradient of the loss function is estimated as the residual value in the current model.
- 3) New regression tree is trained to fit the current residual
- 4) Lastly, the final regression is combined with the previous model and the residual is updated.
- 5) The iteration in the algorithm is continued until the maximum number of iterations set by user is reached.

In short, GBT model improved previous poor performing data by constantly using regression tree to fit the residual. Random Forest technique was developed by Breiman [72] which was a combination classification technique.

### 3.4. Evaluation metrics

This present study used 10,000 records and adopted three advanced machine learning techniques to predict the bus travel time. The authors employed 70% of the observations as training set and 30% as testing set. As pointed out earlier, the predictions are built based on two main approaches of bus TT calculation. As mentioned earlier, two rout construction approaches were adopted to predict the TT: First approach calculates the bus TT using a sum of dwell time and running time and the second approach directly approximates the bus TT using some different variables. The results of these predictions are evaluated using two performance criterions, including mean absolute error (MAE) and linear correlation (R). Equations 1 and 2 present the MAE and R, respectively.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{im} - \hat{y}_{ip}| \quad (1)$$

$$R = \frac{\sqrt{\sum_{i=1}^n (y_{im} - \hat{y}_{ip})^2}}{\sqrt{\sum_{i=1}^n (y_{im} - \bar{y}_{im})^2}} \quad (2)$$

Where  $y_{im}$ ,  $\hat{y}_{ip}$ , and  $\bar{y}_{im}$  denote the measured, predicted and the mean of measured values, n signifies the total number of data.

The authors also used a simple ranking system which sums the training and testing rankings of each model based on their evaluation criteria to achieve a cumulative performance ranking. This helped to conduct a more comprehensive comparison among the ML models and TT estimation approaches. In this ranking system, each value of R and MAE are ranked for each training and

testing datasets. Among the models developed, the model which has obtained the highest value of R and lowest value of MAE in each training and testing phases has received the ranking of four (because four models have been developed). In turn, the weakest performances have received the ranking of one. Then, testing and training rankings have been calculated and allowed the authors to calculate the cumulative ranking for each model. It is worth noting that the models that have equal value of R or MAE have assigned the same ranking. The calculation formula of the cumulative ranking is denoted in Equation 3.

$$\text{Model's cumulative ranking} = \sum_{\substack{1 \leq j \leq 3 \\ 1 \leq i \leq 3}} (\alpha_i + \beta_j) \quad (3)$$

Where, the  $\alpha$  denotes the training performance indicator;  $\beta$  denotes the testing performance indicator;  $i$  denotes training indicator number;  $j$  denotes testing indicator number;  $i=j=1$  represents  $R^2$ ;  $i=j=2$  represents MAE.

## 4. Results

### 4.1. Results of high frequency bus route

The results of the ranking calculation are presented in Tables 3 and 4 for route-based and key stop-based approaches, respectively. According to these results, for route-based approach, the CHAID and GBT achieved the highest training and testing rankings, respectively; however, the GBT obtained the most significant cumulative ranking. Concerning the key stop-based approach, the CHAID model achieved the highest training, testing, and in turn, cumulative ranking. As the CHAID model earned the highest cumulative approach, this model has been selected as the best model for high frequency bus route.

A comparison between route-based and key stop-based approaches shows that the GBT model obtained higher cumulative ranking within the route-based approach. On the other hand, CHAID model showed better performance within the key stop-based approach.

A comparison between the accuracy and error of the ML models developed based on the two approaches showed that the accuracy of the key stop-based approach in the training phase generally was higher than the accuracy of the route-based approach (except for RF). On the other hand, in the testing phase, GBT and RF models developed based on route-based approach had higher "R" compared to key stop-based approach. However, the error of models that created based on the route-based approach is typically less than the key stop-based approach for the training phase (except for CHAID). For the testing phase, the MAEs of all models within the key stop-based approach were higher than models developed based on route-based approach.

The importance score of variables in route-based and key stop-based models for high frequency service was estimated and shown in Figure 8. The motivation behind this analysis was to clearly understand which factors play a significant role in context of travel time prediction. For instance, Ma et al. [15] divided the bus route to dwelling and transit segments and then predicted dwelling and transit separately.

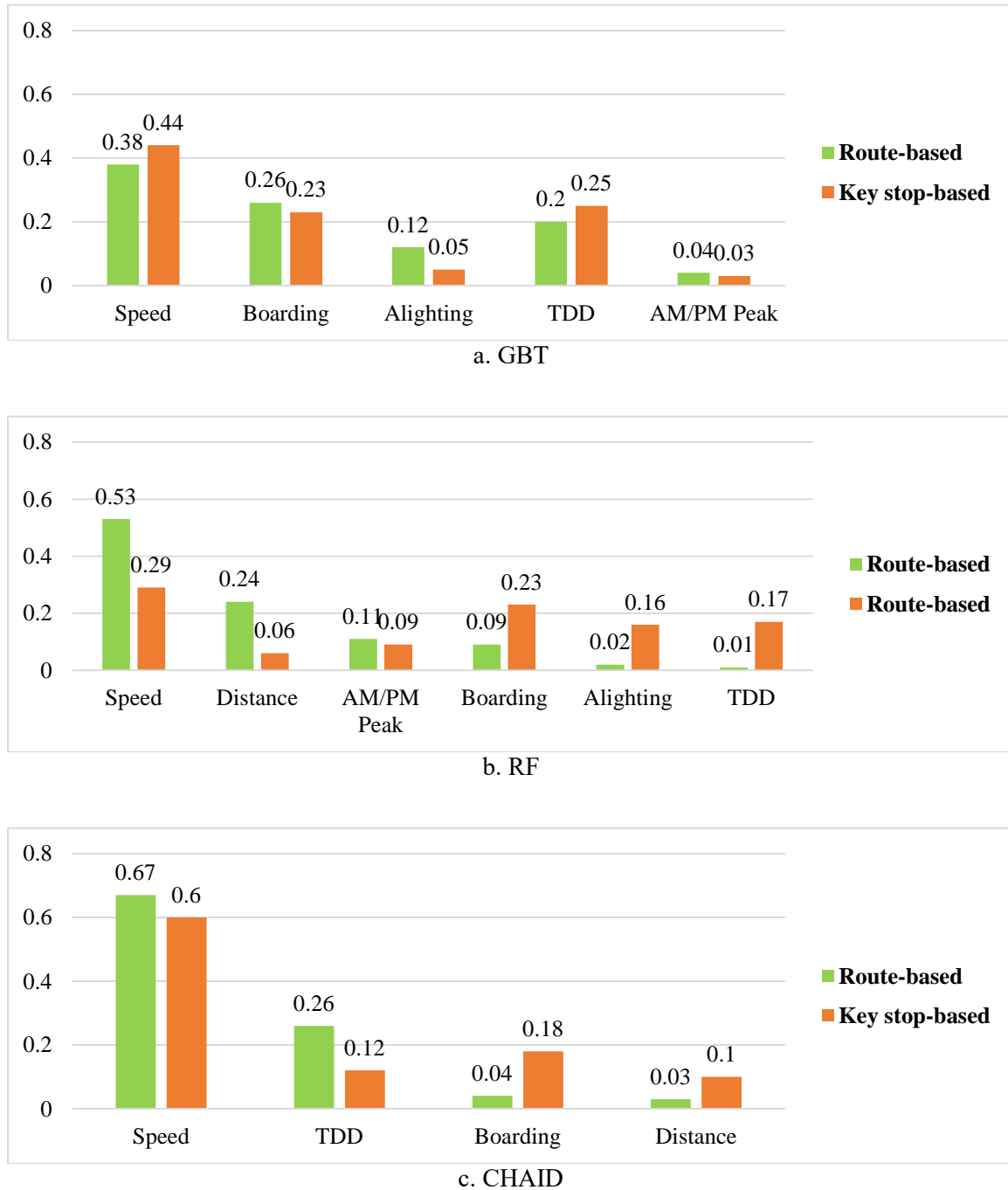
**Table 3**  
Ranking Calculation Results for Route-Based Approach in High-Frequency Route.

				GBT	CHAID	RF
Route based approach	R	TR	Value	0.93	0.94	0.89
			Rank	2	3	1
		TE	Value	0.90	0.77	0.84
			Rank	3	1	2
	MAE	TR	Value	21.28	20.75	80.92
			Rank	2	3	1
		TE	Value	32.79	66.74	98.79
			Rank	3	2	1
	<b>Training ranking</b>			<b>4</b>	<b>6</b>	<b>2</b>
	<b>Testing ranking</b>			<b>6</b>	<b>3</b>	<b>3</b>
	<b>Cumulative ranking</b>			<b>10</b>	<b>9</b>	<b>5</b>

**Table 4**  
Ranking Calculation Results for Key Stop-Based Approach in High-Frequency Route.

				GBT	CHAID	RF
Key-stop based approach	R	TR	Value	0.94	0.96	0.83
			Rank	2	3	1
		TE	Value	0.79	0.88	0.80
			Rank	1	3	2
	MAE	TR	Value	143.66	88.46	350.84
			Rank	2	3	1
		TE	Value	496.36	98.75	510.39
			Rank	2	3	1
	<b>Training ranking</b>			<b>4</b>	<b>6</b>	<b>2</b>
	<b>Testing ranking</b>			<b>3</b>	<b>6</b>	<b>3</b>
	<b>Cumulative ranking</b>			<b>7</b>	<b>12</b>	<b>5</b>

They considered boarding and speed as important impact factors for dwell and transit time, respectively. According to Figure 6, speed was identified as the most important variable by all three ML models for both route-based and key stop-based approaches. Moreover, boarding and alighting both play an important role in predicating dwell times, while TDD and distance between stops were recognized as impactful factors for predicting segment running times.



**Fig. 6.** Importance of variables for high frequency models.

#### 4.2. Results of low frequency approach

Tables 5 and 6 present the results of the ranking computation for route-based and key stop-based approaches, respectively.

**Table 5**

Ranking Calculation Results for Route-Based Approach in Low-Frequency Route.

				GBT	CHAID	RF
Route based approach	R	TR	Value	0.89	0.91	0.79
			Rank	2	3	1
		TE	Value	0.87	0.71	0.81
			Rank	3	1	2
	MAE	TR	Value	69.55	45.23	93.75
			Rank	2	3	1
		TE	Value	88.75	78.22	125.45
			Rank	2	3	1
	<b>Training ranking</b>			<b>4</b>	<b>6</b>	<b>2</b>
	<b>Testing ranking</b>			<b>5</b>	<b>4</b>	<b>3</b>
	<b>Cumulative ranking</b>			<b>9</b>	<b>10</b>	<b>5</b>

**Table 6**

Ranking Calculation Results for Key Stop-Based Approach in Low-Frequency Route.

				GBT	CHAID	RF
Key-stop based approach	R	TR	Value	0.88	0.91	0.78
			Rank	2	3	1
		TE	Value	0.68	0.88	0.70
			Rank	1	3	2
	MAE	TR	Value	197.23	56.75	484.77
			Rank	2	3	1
		TE	Value	532.77	212.46	574.23
			Rank	2	3	1
	<b>Training ranking</b>			<b>4</b>	<b>6</b>	<b>2</b>
	<b>Testing ranking</b>			<b>3</b>	<b>6</b>	<b>3</b>
	<b>Cumulative ranking</b>			<b>7</b>	<b>12</b>	<b>5</b>

According to these results, for route-based approach, the CHAID and GBT achieved the highest training and testing rankings, respectively; though, the CHAID achieved the most significant cumulative ranking. Regarding the key stop-based approach, the CHAID model achieved the highest training, testing, and in turn, cumulative ranking. Since the CHAID model received the greatest cumulative approach, this model has been nominated as the best model for low-frequency bus service.

The importance score of variables in route-based and key stop-based models for low-frequency service was estimated and shown in Figure 7. For route-based approach and RF and GBT models, distance was identified as the most important variable. In addition, for route-based



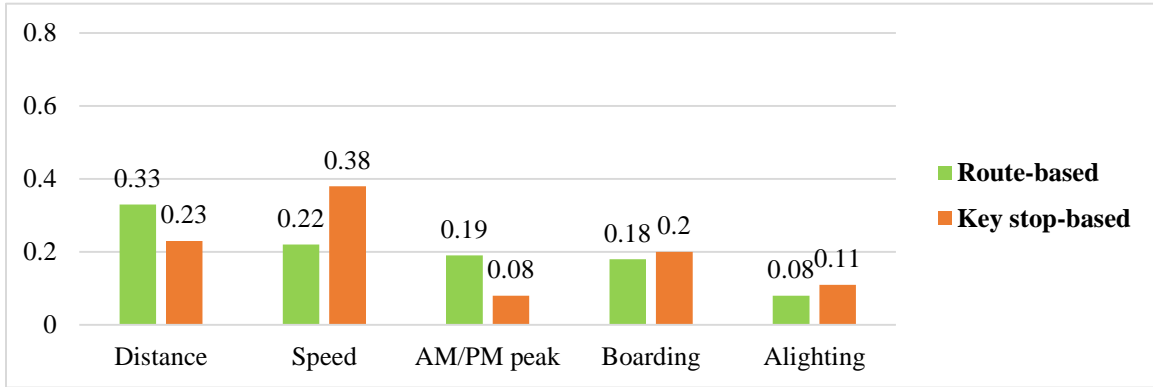
approach and CHAID model, speed was identified as the most important variable. For key-stop based approach and all the three ML models, speed was identified as the most important variable.

## **5. Discussion**

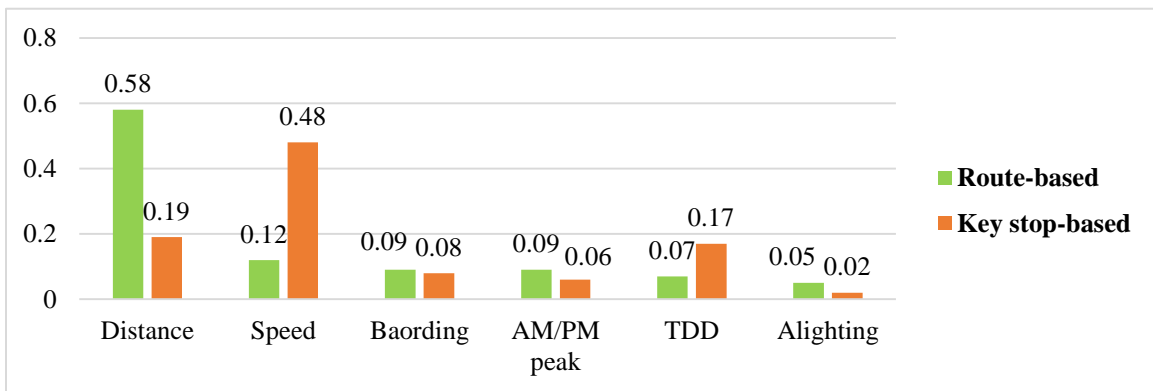
1- According to literature, high and low frequency bus routes have different characterizations and specifications. Passengers tend to neglect the schedule and arrive at bus stop randomly in high-frequency routes. Therefore, passengers put more value on real time information accuracy in high-frequency routes. From operational aspect, high-frequency bus routes (or routes during high-frequency operation) are dealing with short headways and high passenger demand. High-frequency bus services are more sensitive to variations and trigger factors (such as variation in demand, late departure from terminal and adverse weather) comparing to low-frequency service [12,16]. Consequently, we highly expected that accuracy of bus travel time prediction should be impacted by type of service frequency. Accordingly, this study was set out to investigate and compare the prediction of bus travel time using three different ML methods in high and low service frequencies, for the first time. AVL, APC and AFC data sets were used to conduct the analysis. Based on our findings, the accuracy of travel time prediction depends on the bus route frequency.

The results proved that the accuracy of bus travel time prediction is relatively higher in high-frequency bus route. The main reason for better results for high-frequency is the higher number of operating buses at route in a specific time (as shown in Figure 1). When there are more buses on route at a specific time, we have more accurate information about the traffic condition and vehicles' trajectory. In other words, this can be concluded that in high-frequency routes there is no need to simulate the traffic condition separately, since we have enough real time data of vehicles' movements. As an example, if any incident happens on the route (as shown in Figure 1, an accident on segment 3), there should be at least one bus on that segment to capture the slowness in the traffic movement and report it to following buses on the route very fast. Therefore, other buses are able to update their arrival times accordingly.

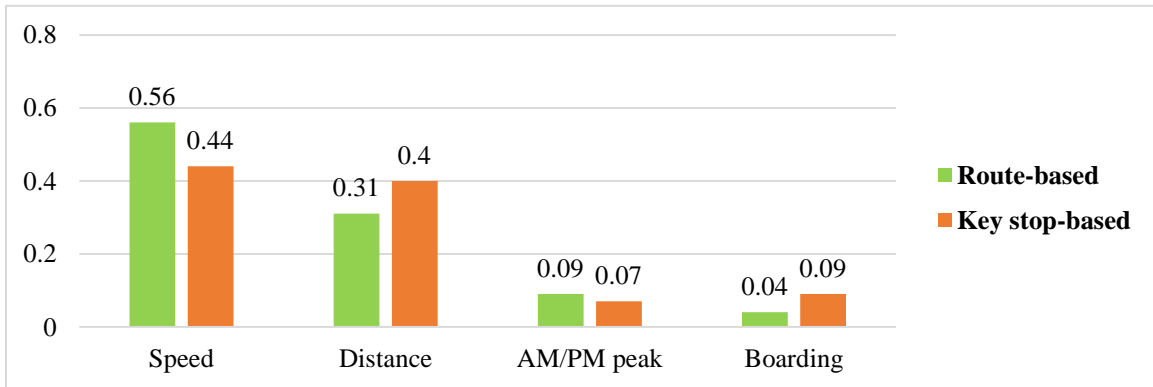
2- We employed the key stop-based route construction method for the first time for predicting the bus travel time. Our motivation for examining a new route graph method was the complexity and inapplicability (in some cases) of previous methods such as linked-based method. In key stop-based method, route is divided to two main temporal and spatial segments: Dwelling segments at key stops and running segments between two successive key stops (Figure 1). According to results, this can be concluded that key stop-based approach is a simple and accurate route construction method for predicting bus travel time. It is simple because in this method we only model dwell times at key stops and running times in segments between key stops. Key stop is an important stop with strategic location or/and high passenger demand. Milkovits [16] claimed that, in presence of Big Data, there is no need to consider each minor stop separately for estimating dwell times. However, it doesn't mean that we neglect the minor stops' dwelling times and they must be taken into account for predicting segment running time as total number of passengers boarding and alighting along the segment. This can be the main reason of high accuracy of this approach.



a. GBT



b. RF



c. CHAID

**Fig. 7.** Importance of variables for low frequency models.

In addition, we also examined a route-based method for prediction of bus travel time. This method, which is the simplest bus route graph, mostly can be used in the initial stages of designing a bus route and setting up an accurate schedule. Moreover, service providers need to accurately predict the bus travel time for many purposes such as designing new routes, planning future travels, scheduling or re-scheduling the current or future trips. In this case, service

providers mostly need to estimate travel time for whole route, instead of each segment. Therefore, we examined predicting travel time for whole route without any further route graph and construction, using various machine learning methods and considering factors in Table 1.

3- Nowadays, most of the bus companies have access to big and rich data sets by implementing new technologies in automatic data collection systems. ML techniques are the most suitable methods for predicting bus travel time by using these big and rich data sets. Therefore, ML methods have been widely used in this context. However, this was not evidenced which ML technique is the most appropriate one for predicting bus travel with respect to the bus service specification and frequency. Therefore, we designed this study to shed some light on this issue by conducting and comparing three well-known machine learning techniques, including GBT, RF and CHAID. While properties of CHAID method is highly fitted with requirements of bus travel time prediction, this method never been used before in this context. According to our output of our analysis, GBT can be selected as the best ML technique for predicting bus travel time in high-frequency service, while CHAID can be nominated as the most accurate ML method to predict the travel time in low-frequency bus service.

Ma et al. [15] argued that using bus GPS and smart card data are not enough for accurate prediction of bus travel time, since these data sets are not capable to reflect the real traffic condition and bus trajectories. Accordingly, he proposed a novel travel time prediction method based on combination of buses and taxis real time data. However, such hybrid methods (combination of two or more methods) have considerable limitations. Firstly, usually taxis' (hailing) GPS data is recorded by other private companies. Collecting data from these companies is the first challenge, since real time GPS data is considered as confidential data for most of the taxi and e-hailing companies. Secondly, even if we got access to taxis' real time GPS data, integrating taxis and buses data to predict the bus travel time is the second big challenge in real time prediction. Moreover, based on our findings, Ma et al. [15] argument is only applicable in low-frequency bus routes and could not be valid in high-frequency routes. Because in high-frequency routes there is always enough data of traffic condition and vehicle trajectories due to high number of operating buses, that we can accurately predict the travel time.

## **6. Conclusions**

Applicability and accuracy of different bus travel time prediction approaches were investigated in this study. First, there are considerable differences between high and low frequency bus routes for predicting the travel time. Therefore, in order to predict the bus travel time accurately, the frequency of bus service should be considered. Second, according to results, GBT can be selected as the best ML technique for predicting bus travel time in high-frequency service (with  $R=93\%$  and  $MAE=21.23$ ), while CHAID (with  $R=91\%$  and  $MAE=56$ ) can be nominated as the most accurate ML method to predict the travel time in low-frequency bus service. Moreover, bus travel time was predicted more accurate in high-frequency bus service ( $R$  in high frequency route is  $96\%$  and in low frequency is  $91\%$ ). Third, Key stop-based route construction approach is an accurate and reliable approach for predicting bus travel time, while this approach is much simpler and more applicable comparing to previous approaches. Finally, in term of the

importance of variables, both boarding and alighting should be considered for modeling bus dwell times. Moreover, speed (with 0.67 and 0.6 weight for route-based and key stop-based approach, respectively) and terminal departure deviations (with 0.26 and 0.12 weight for route-based and key stop-based approach, respectively) are significantly important variables for predicting bus travel times.

## Acknowledgments

The authors would like to acknowledge the Department of Civil Engineering, Faculty of Engineering, University of Malaya, for financial support under GPF009A-2019 grant. We would like to acknowledge all the experts and staffs in RapidKL and Pasarana Bus Company for providing data and information. In particular, authors would like to acknowledge the Centre for Transportation Research (CTR), Faculty of Engineering, University of Malaya and also Sustainable Urban Transport Research Centre (SUTRA), Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM) for providing research facilities.

## Funding

This research received no external funding.

## Conflicts of interest

The authors declare no conflict of interest.

## Authors contribution statement

Seyed Mohammad Hossein Moosavi: Conceptualization, Data curation, Writing – review & editing, Roles/Writing – original draft; Mahdi aghaabbasi: Data curation, Software; Formal analysis; Choon Wah Yuen: Project administration Formal analysis; Danial Jahed Armaghani: Writing – review & editing, Supervision, Validation.

## References

- [1] Nguyen-Phuoc DQ, Young W, Currie G, De Gruyter C. Traffic congestion relief associated with public transport: state-of-the-art. *Public Transp* 2020;1–27.
- [2] Mugion RG, Toni M, Raharjo H, Di Pietro L, Sebathu SP. Does the service quality of urban public transport enhance sustainable mobility? *J Clean Prod* 2018;174:1566–87.
- [3] Moosavi SMH, Yuen CW, Yap SP, Onn CC. Simulation-Based Sensitivity Analysis for Evaluating Factors Affecting Bus Service Reliability: A Big and Smart Data Implementation. *IEEE Access* 2020;8:201937–55.
- [4] Moosavi SMH, Choon Wah Y. Measuring Bus Running Time Variation during High-Frequency Operation Using Automatic Data Collection Systems. *Inst Transp Eng ITE J* 2020;90:45–9.
- [5] Bertsimas D, Delarue A, Jaillet P, Martin S. Travel time estimation in the age of big data. *Oper Res* 2019;67:498–515.
- [6] Tyndall J. Bus quality improvements and local commuter mode share. *Transp Res Part A Policy Pract* 2018;113:173–83.

- [7] Moosavi SMH, Ismail A, Yuen CW. Using simulation model as a tool for analyzing bus service reliability and implementing improvement strategies. *PLoS One* 2020;15:e0232799.
- [8] Bie Y, Xiong X, Yan Y, Qu X. Dynamic headway control for high-frequency bus line based on speed guidance and intersection signal adjustment. *Comput Civ Infrastruct Eng* 2020;35:4–25.
- [9] Badia H, Argote-Cabanero J, Daganzo CF. How network structure can boost and shape the demand for bus transit. *Transp Res Part A Policy Pract* 2017;103:83–94.
- [10] Gris  E, El-Geneidy A. Assessing operation and customer perception characteristics of high frequency local and limited-stop bus service in Vancouver, Canada. *Public Transp* 2020:1–16.
- [11] Anderson P, Daganzo CF. Effect of transit signal priority on bus service reliability. *Transp Res Part B Methodol* 2019.
- [12] Chen W, Yang C, Feng F, Chen Z. An improved model for headway-based bus service unreliability prevention with vehicle load capacity constraint at bus stops. *Discret Dyn Nat Soc* 2012;2012. <https://doi.org/10.1155/2012/313518>.
- [13] Jairam R, Kumar BA, Arkatkar SS, Vanajakshi L. Performance comparison of bus travel time prediction models across Indian cities. *Transp Res Rec* 2018;2672:87–98. <https://doi.org/10.1177/0361198118770175>.
- [14] Mori U, Mendiburu A,  lvarez M, Lozano JA. A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transp A Transp Sci* 2015;11:119–57. <https://doi.org/10.1080/23249935.2014.932469>.
- [15] Ma J, Chan J, Ristanoski G, Rajasegarar S, Leckie C. Bus travel time prediction with real-time traffic information. *Transp Res Part C Emerg Technol* 2019;105:536–49. <https://doi.org/10.1016/j.trc.2019.06.008>.
- [16] Milkovits MN. Simulating service reliability of a high frequency bus route using automatically collected data 2008.
- [17] Parsajoo M, Armaghani DJ, Mohammed AS, Khari M, Jahandari S. Tensile strength prediction of rock material using non-destructive tests: A comparative intelligent study. *Transp Geotech* 2021;31:100652. <https://doi.org/10.1016/J.TRGEO.2021.100652>.
- [18] Barkhordari M, Armaghani D, Mohammed A, Ulrikh D. Data-Driven Compressive Strength Prediction of Fly Ash Concrete Using Ensemble Learner Algorithms. *Buildings* 2022;12:132. <https://doi.org/10.3390/buildings12020132>.
- [19] Reza zadeh Eidgahee D, Jahangir H, Solatifar N, Fakharian P, Rezaeemanesh M. Data-driven estimation models of asphalt mixtures dynamic modulus using ANN, GP and combinatorial GMDH approaches. *Neural Comput Appl* 2022;34:17289–314. <https://doi.org/10.1007/s00521-022-07382-3>.
- [20] Ghanizadeh AR, Ghanizadeh A, Asteris PG, Fakharian P, Armaghani DJ. Developing bearing capacity model for geogrid-reinforced stone columns improved soft clay utilizing MARS-EBS hybrid method. *Transp Geotech* 2023;38:100906. <https://doi.org/10.1016/j.trgeo.2022.100906>.
- [21] Fakharian P, Reza zadeh Eidgahee D, Akbari M, Jahangir H, Ali Taeb A. Compressive strength prediction of hollow concrete masonry blocks using artificial intelligence algorithms. *Structures* 2023;47:1790–802. <https://doi.org/10.1016/j.istruc.2022.12.007>.
- [22] Mahmood W, Mohammed AS, Asteris PG, Kurda R, Armaghani DJ. Modeling Flexural and Compressive Strengths Behaviour of Cement-Grouted Sands Modified with Water Reducer Polymer. *Appl Sci* 2022;12:1016.
- [23] Tan WY, Lai SH, Teo FY, Armaghani DJ, Pavitra K, El-Shafie A. Three Steps towards Better Forecasting for Streamflow Deep Learning. *Appl Sci* 2022;12. <https://doi.org/10.3390/app122412567>.

- [24] He B, Armaghani DJ, Lai SH. Assessment of tunnel blasting-induced overbreak: A novel metaheuristic-based random forest approach. *Tunn Undergr Sp Technol* 2023;133:104979. <https://doi.org/10.1016/j.tust.2022.104979>.
- [25] Armaghani DJ, Asteris PG, Fatemi SA, Hasanipanah M, Tarinejad R, Rashid ASA, et al. On the Use of Neuro-Swarm System to Forecast the Pile Settlement. *Appl Sci* 2020;10:1904.
- [26] Barkhordari, M., Armaghani, D., Asteris P. Structural Damage Identification Using Ensemble Deep Convolutional Neural Network Models. *C Model Eng Sci* 2022;doi: 10.32604/cmescs.2022.020840.
- [27] Hasanipanah M, Monjezi M, Shahnazar A, Armaghani DJ, Farazmand A. Feasibility of indirect determination of blast induced ground vibration based on support vector machine. *Measurement* 2015;75:289–97.
- [28] Koopialipour M, Asteris PG, Salih Mohammed A, Alexakis DE, Mamou A, Armaghani DJ. Introducing stacking machine learning approaches for the prediction of rock deformation. *Transp Geotech* 2022;34:100756. <https://doi.org/10.1016/j.trgeo.2022.100756>.
- [29] Asteris PG, Lourenço PB, Roussis PC, Adami CE, Armaghani DJ, Cavaleri L, et al. Revealing the nature of metakaolin-based concrete materials using artificial intelligence techniques. *Constr Build Mater* 2022;322:126500.
- [30] Guido G, Haghshenas SS, Haghshenas SS, Vitale A, Gallelli V, Astarita V. Development of a binary classification model to assess safety in transportation systems using GMDH-type neural network algorithm. *Sustain* 2020. <https://doi.org/10.3390/SU12176735>.
- [31] Guido G, Haghshenas SS, Haghshenas SS, Vitale A, Astarita V, Haghshenas AS. Feasibility of stochastic models for evaluation of potential factors for safety: A case study in southern Italy. *Sustain* 2020. <https://doi.org/10.3390/su12187541>.
- [32] Guido G, Haghshenas SS, Haghshenas SS, Vitale A, Astarita V, Park Y, et al. Evaluation of Contributing Factors Affecting Number of Vehicles Involved in Crashes Using Machine Learning Techniques in Rural Roads of Cosenza, Italy. *Saf* 2022, Vol 8, Page 28 2022;8:28. <https://doi.org/10.3390/SAFETY8020028>.
- [33] Ghanizadeh AR, Delaram A, Fakharian P, Armaghani DJ. Developing Predictive Models of Collapse Settlement and Coefficient of Stress Release of Sandy-Gravel Soil via Evolutionary Polynomial Regression. *Appl Sci* 2022;12:9986. <https://doi.org/10.3390/app12199986>.
- [34] Papageorgiou M, Papamichail I, Messmer A, Wang Y. Traffic simulation with METANET. *Fundam. traffic Simul.*, Springer; 2010, p. 399–430.
- [35] Li L, Chen X, Li Z, Zhang L. Freeway travel-time estimation based on temporal–spatial queueing model. *IEEE Trans Intell Transp Syst* 2013;14:1536–41.
- [36] Rahman MM, Wirasinghe SC, Kattan L. Analysis of bus travel time distributions for varying horizons and real-time applications. *Transp Res Part C Emerg Technol* 2018;86:453–66. <https://doi.org/10.1016/j.trc.2017.11.023>.
- [37] Guido G, Haghshenas SS, Vitale A, Astarita V. Challenges and Opportunities of Using Data Fusion Methods for Travel Time Estimation. 2022 8th Int Conf Control Decis Inf Technol CoDIT 2022 2022:587–92. <https://doi.org/10.1109/CODIT55151.2022.9804014>.
- [38] Nikovski D, Nishiuma N, Goto Y, Kumazawa H. Univariate short-term prediction of road travel times. *Proceedings. 2005 IEEE Intell. Transp. Syst. 2005.*, IEEE; 2005, p. 1074–9.
- [39] Du L, Peeta S, Kim YH. An adaptive information fusion model to predict the short-term link travel time distribution in dynamic traffic networks. *Transp Res Part B Methodol* 2012;46:235–52.
- [40] Hofleitner A, Herring R, Bayen A. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transp Res Part B Methodol* 2012;46:1097–122. <https://doi.org/10.1016/j.trb.2012.03.006>.
- [41] Yildirimoglu M, Ozbay K. Comparative evaluation of probe-based travel time prediction techniques under varying traffic conditions. 2012.

- [42] Moosavi SMH, Ma Z, Armaghani DJ, Aghaabbasi M, Ganggayah MD, Wah YC, et al. Understanding and Predicting the Usage of Shared Electric Scooter Services on University Campuses. *Appl Sci* 2022;12:9392.
- [43] Chien SI-J, Ding Y, Wei C. Dynamic bus arrival time prediction with artificial neural networks. *J Transp Eng* 2002;128:429–38.
- [44] Xu H, Ying J. Bus arrival time prediction with real-time and historic data. *Cluster Comput* 2017;20:3099–106.
- [45] Mazloumi E, Rose G, Currie G, Moridpour S. Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Eng Appl Artif Intell* 2011;24:534–42. <https://doi.org/10.1016/j.engappai.2010.11.004>.
- [46] Li R, Rose G. Incorporating uncertainty into short-term travel time predictions. *Transp Res Part C Emerg Technol* 2011;19:1006–18.
- [47] Gurmu ZK, Fan WD. Artificial neural network travel time prediction model for buses using only GPS data. *J Public Transp* 2014;17:3.
- [48] Simroth A, Zähle H. Travel time prediction using floating car data applied to logistics planning. *IEEE Trans Intell Transp Syst* 2010;12:243–53.
- [49] Wu C-H, Ho J-M, Lee D-T. Travel-time prediction with support vector regression. *IEEE Trans Intell Transp Syst* 2004;5:276–81.
- [50] Mendes-Moreira J, Jorge AM, de Sousa JF, Soares C. Comparing state-of-the-art regression methods for long term travel time prediction. *Intell Data Anal* 2012;16:427–49.
- [51] Yang M, Chen C, Wang L, Yan X, Zhou L. Bus arrival time prediction using support vector machine with genetic algorithm. *Neural Netw World* 2016;26:205.
- [52] Bin Y, Zhongzhen Y, Baozhen Y. Bus Arrival Time Prediction Using Support Vector Machines. *J Intell Transp Syst* 2006;10:151–8. <https://doi.org/10.1080/15472450600981009>.
- [53] Yu B, Lam WHK, Tam ML. Bus arrival time prediction at bus stop with multiple routes. *Transp Res Part C Emerg Technol* 2011;19:1157–70. <https://doi.org/10.1016/j.trc.2011.01.003>.
- [54] Yu B, Wang H, Shan W, Yao B. Prediction of bus travel time using random forests based on near neighbors. *Comput Civ Infrastruct Eng* 2018;33:333–50.
- [55] Zheng F, Van Zuylen H, Liu X. A methodological framework of travel time distribution estimation for urban signalized arterial roads. *Transp Sci* 2017;51:893–917.
- [56] Woodhull J. Issues in on-time performance of bus systems. Unpubl Manuscript Los Angeles, CA South Calif Rapid Transit Dist 1987.
- [57] Zhou Y, Yao L, Chen Y, Gong Y, Lai J. Bus arrival time calculation model based on smart card data. *Transp Res Part C Emerg Technol* 2017;74:81–96.
- [58] Ramakrishna Y, Ramakrishna P, Lakshmanan V, Sivanandan R. Use of GPS probe data and passenger data for prediction of bus transit travel time. *Transp. L. Use, Planning, Air Qual.*, 2008, p. 124–33.
- [59] Tirachini A. Estimation of travel time and the benefits of upgrading the fare payment technology in urban bus services. *Transp Res Part C Emerg Technol* 2013;30:239–56. <https://doi.org/10.1016/j.trc.2011.11.007>.
- [60] Shalaby A, Farhan A. Prediction model of bus arrival and departure times using AVL and APC data. *J Public Transp* 2004;7:3.
- [61] Domenichini L, Salerno G, Fanfani F, Bacchi M, Giaccherini A, Costalli L, et al. Travel time in case of accident prediction model. *Procedia-Social Behav Sci* 2012;53:1078–87.
- [62] Yildirimoglu M, Geroliminis N. Experienced travel time prediction for congested freeways. *Transp Res Part B Methodol* 2013;53:45–63.

- [63] Ma Z, Koutsopoulos HN, Ferreira L, Mesbah M. Estimation of trip travel time distribution using a generalized Markov chain approach. *Transp Res Part C Emerg Technol* 2017;74:1–21.
- [64] Kumar BA, Vanajakshi L, Subramanian SC. Bus travel time prediction using a time-space discretization approach. *Transp Res Part C Emerg Technol* 2017;79:308–32. <https://doi.org/10.1016/j.trc.2017.04.002>.
- [65] Zhou M, Wang D, Li Q, Yue Y, Tu W, Cao R. Impacts of weather on public transport ridership: Results from mining data from different sources. *Transp Res Part C Emerg Technol* 2017;75:17–29.
- [66] Wang Y, Bie Y, An Q. Impacts of winter weather on bus travel time in cold regions: Case study of Harbin, China. *J Transp Eng Part A Syst* 2018;144:5018001.
- [67] Miao Q, Welch EW, Sriraj PS. Extreme weather, public transport ridership and moderating effect of bus stop shelters. *J Transp Geogr* 2019;74:125–33.
- [68] Tao S, Corcoran J, Rowe F, Hickman M. To travel or not to travel: ‘Weather’ is the question. Modelling the effect of local weather conditions on bus ridership. *Transp Res Part C Emerg Technol* 2018;86:147–67.
- [69] You J, Kim TJ. Development and evaluation of a hybrid travel time forecasting model. *Transp Res Part C Emerg Technol* 2000;8:231–56.
- [70] Shen L, Hadi M. Practical approach for travel time estimation from point traffic detector data. *J Adv Transp* 2013;47:526–35.
- [71] Kass G V. An exploratory technique for investigating large quantities of categorical data. *J R Stat Soc Ser C (Applied Stat)* 1980;29:119–27.
- [72] Breiman L. Random forests. *Mach Learn* 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.