Constructing Non-GPS-based ETA Models from Bus-stops Arrival and Departure History and Traffic Contexts

Abdul Wafi Haji Ismail[a], Veronica Anne Roddy[a], Mohammad Nur Adil Haji Harddy Hisysham[a] and Nur Izzati Nazatul Aqilah Junaidi[a]

[a]*School of Computing and Informatics, Universiti Teknologi Brunei, Jalan Tungku Link, Gadong BE1410, Brunei Darussalam*

ARTICLE INFO

Keywords:
non-GPS-based ETA models
predictive models
non-linear models
traffic-data analysis
decision tree
random forest
n-nearest neighbors
artificial neural network

ABSTRACT

In the absence of GPS data, bus stops' arrival and departure history and traffic environment context may be used in estimating ETA with good precision. In this paper, time stamps was used in the estimation of travel duration between bus stops. Employed to compute the duration between bus stops were three techniques, namely *statistical central tendencies*, *nonlinear regression techniques*, i.e., decision tree regressor, random forest regressor and k-nearest neighbors regressor, and an Artificial Neural Network (ANN) predictive model. Different characteristics of the traffic environment are captured by these techniques. An overall view of the ETA information is given by *statistical central tendencies* while context from the time stamps (e.g., weekdays vs. weekends) can be captured by *nonlinear regressors*. Additionally, predicting duration between bus stops based on previous adjacent duration spans implicitly captures temporal dynamics of traffic environment, and this information is exploited by the ANN predictive model. The feasibility of estimating ETA based on these two techniques was explained, evaluated and discussed.

1. Introduction

Access to accurate Estimated Time of Arrival (ETA) for public transportation users is helpful as it enables them to effectively plan their trips. As an example in the case of public buses, ETA can be approximated with good precision at operation time - that is, by using the position and speed of the bus - if the GPS information of the buses is available. It is common for local regulators to require bus operators to have GPS information for their fleets. For example, in Malaysia, bus operators are required by its government to install GPS for driving behaviors to be monitored on routes [1]. By making a small modification to the operator's data processing pipeline, ETA information could be made available at a relatively low cost to the public. In this paper, an alternative method to estimating ETA without the use of real-time GPS data is presented.

1.1. Non-GPS-based ETA

The main approach is:

1. At each bus stop, implement a low-cost bus arrival detector. This is achieved with the installation of a beacon on a bus, and Bluetooth Low Energy (BLE) [2] sensors at bus stops.
2. Create a look-up table of traveled duration between bus stops by exploiting historical data [3]. Subsequent bus stops' ETA can then be estimated in real-time - without the bus's GPS information - from the arrival time at the current bus stops as well as other traffic environment contextual information such as times of the day and days of the week.

Email addresses: M20230025@student.utb.edu.bn (A.W.H. Ismail); M20230096@student.utb.edu.bn (V.A. Roddy); M20230020@student.utb.edu.bn (M.N.A.H.H. Hisysham); M20230063@student.utb.edu.bn (N.I.N.A. Junaidi)

In this project, calculation of ETA data is made from particularly one bus service route's available historical arrival and departure information. Based of this temporal data, two stops' in-between travel duration can be calculated for any time of day and given day of a week. The knowledge distilled from this then serves as a representation of the travel duration between bus stops. Traditional descriptive statistics and machine learning techniques are then used to construct the look-up table models.

Of course, with this approach, traffic scene dynamics (e.g., congestion resulting from accidents and weather conditions) might not be fully captured. However, in a circumstance where having a GPS installed on a bus is not a requirement from a local transport regulatory body, the proposed approach can be an attractive solution in predicting ETA for a bus operator, especially considering its low operational and hardware costs as opposed to a GPS-based approach.

This project's approach is as such:

1. Document bus service data with the use of data analytic and visualization techniques, as well as clean and prepare benchmark datasets
2. Construct baseline models, namely descriptive statistical models, regression models and predictive models, which serve as benchmark models for the prepared datasets.

2. Related Works

Historical GPS data can be utilized in developing a model for forecasting travel times between bus stops [4, 5, 6]. However, various external factors such as accidents, weather conditions, seasonal influences on traffic, and others do impact vehicles' travel times. It thus becomes a challenge to

anticipate vehicles' travel times without real-time traffic information. This real-time traffic information can be extracted using sensors on road links such as inductive loops, making use of GPS data from vehicles, or crowd-sensed data.

Interpretation of observed traffic information can be performed using various techniques. Low-frequency probe vehicle data were used in computing descriptive statistical measures such as mean and covariance of transit times at the road link levels [7]. [8] used a linear regression model constructed from GPS-based taxicab origin-destination trip data to estimate the travel time for each road link. [9] on the other hand experimented with a *gradient-boosted regression tree* (an ensemble model), where [9] claimed that the technique improves the ETA estimation performance.

Information observed from other links in a citywide fashion may reveal useful traffic dynamic information. This information provides augmented data to local road links and can be utilized in improving the travel time estimations. For instance, [10] proposed a citywide real-time model estimating the travel time of any path based on the GPS trajectories of all vehicles in the system. In [11], travel times were predicted with the employment of spatial traffic evolution.

Real-time road traffic information can also be extracted from a crowd-sensed data approach [12]. [13] utilized a combination of both real-time road traffic information and contextual information regarding the road and transit systems learned from historical data to predict their effects on the public transit systems, such as bus delays.

3. Materials and Methods

3.1. Raw Data

The raw data obtained for this project is of the bus route operating in Johor Bahru, Malaysia. (see Figure 1).

The raw dataset is made up of the bus' vehicle ID as well as time stamps for the bus' arrival to and departure from each bus stop in the route. The bus service begins at around 6:00 and ends around before 00:00 midnight. This forms a sufficient data set for this project.

Figure 1 shows the map and bus stops along the examined route. It is worth noting that portions of data are corrupted. Cleaning of these records corrupted due to hardware failures present a challenge to this study. These errors produce missing entry and thus drastically reduces the available data from the already relatively small dataset.

3.2. Machine Learning Models for Bus ETA Estimation

In this project, historical arrival and departure information from existing bus service routes is exploited. In order to estimate ETA between bus stops for any service routes, statistical models and machine learning models are constructed from the available time stamp records to predict bus stop ETAs. The travel duration of arbitrary trips can be approximated from its connected bus stops along a given traveling path.

Three ETA estimation approaches are reported in this project:

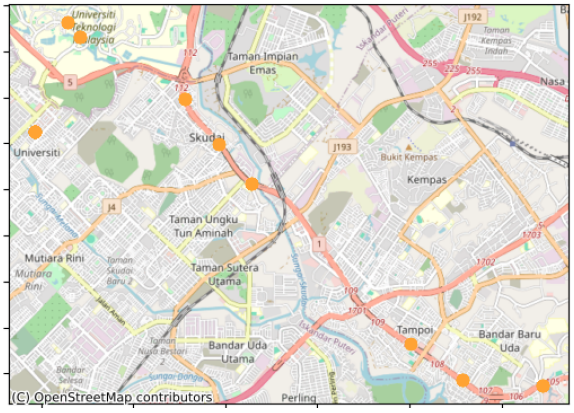

Figure 1: The dots indicate bus stops for the examined bus route

1. ETA from descriptive statistical measures, describing the ETA central tendency from all observed data.
2. ETA from nonlinear regression models, i.e., k-nearest neighbors regressor (KNN), decision tree regressor (DT), and random forest regressor (RF). These models are capable of describing ETA estimation with attention to the local context.
3. ETA predicted from a tri-gram predictive model. The model predicts the next ETA based on its two preceding busstops' in-between travel durations as well as local context of the traffic environment.

4. ETA Models Construction

From the raw data, duration spans between bus stops along the service route can be computed by using location and time information. Duration is simply the difference between the arrival time on a destination bus stop and departure time from an originating bus stop. With this method, the duration the bus stays idle at each bus stop - which could be from passengers or traffic - would also be excluded.

4.1. Descriptive Statistic Baseline Models

Let $S = [s_1, \ldots, s_n]$ be a sequence of n bus stops on a service route where ETA information of a vehicle v traveling on this route is broadcasted to the public. With no available GPS data, the ETA of v arriving at s_i can be computed from historical traveled duration $d(s_i, s_{i+1})$ distilled from historical time stamp data along the service route as mentioned in the previous section.

The ETA a for the bus stop $s_{i+1}, \ldots, s_n$ can be estimated using the departure time at the current bus stop l_{s_i} and historical data of expected travel duration of the next bus stops, i.e., $a_{s_{i+1}} = l_{s_i} + d(s_i, s_{i+1})$.

$$\mathbb{E}[d(s_i, s_{i+1})] = \frac{1}{N} \sum_{j=1}^{N} (a_{s_{i+1}} - l_{s_i})_j \tag{1}$$

Equation (1) computes the mean value from all observations and can serve as a simple baseline model. Apart from

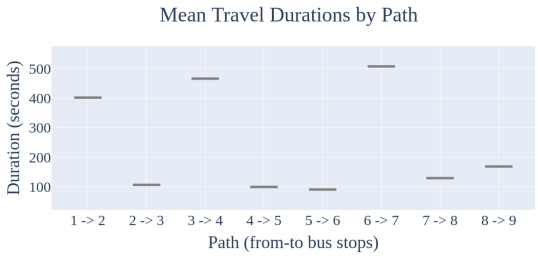

Figure 2: The mean duration $d(s_i, s_{i+1})$ between two consecutive bus stops along the examined service route

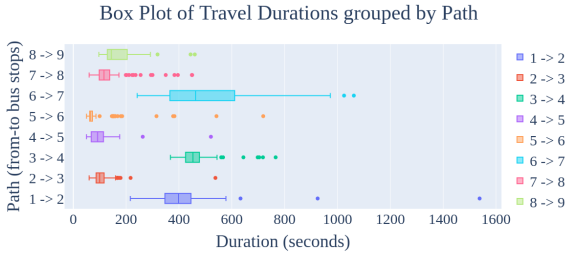

Figure 3: Box plot illustrating the median, upper and lower quartiles, upper and lower fences and maximum and minimums of duration $d(s_i, s_{i+1})$ grouped by paths

the mean value, other descriptive central tendency values such as the *mode* and *median* are other plausible candidates.

Figure 2 shows the mean duration spans of the examined service route computed for the month of October 2023, while Figure 3 shows the duration's median, upper and lower quartiles, maximum and minimum. This provides a descriptive summary but will not capture the dynamics of the traffic environments. The accuracy of descriptive baseline models is as good as the quality of the data employed to compute the central tendency. That is, the quality of descriptive ETA values, such as mode, median, mean, and standard deviations is according to the data.

4.2. Nonlinear Regression Models

Descriptive central tendencies such as mean, mode and median capture the global view from the whole data set but are not sufficiently effective in handling variations from local contexts e.g., traffic conditions from different days of a week or time of day. Other potential techniques are nonlinear regression models.

A typical parametric nonlinear regression model is expressed as

$$d(s_i, s_{i+1}) = f(X, \beta) + \epsilon \tag{2}$$

where β is a nonlinear parameter to be computed, ϵ is the error term, and X denotes independent variables, which can be either uni-variate or multivariate of nonlinear function $f(\cdot, \cdot)$, reflecting traffic characteristics between bus stops s_i and s_{i+1}.

As time stamp information is available from the obtained raw data, extra context C can be added in, which is based on the day of week and time of day. Expected ETA duration

	Regressors			ANN
Performance Metric	***Decision Tree***	***Random Forest***	***K-Nearest Neighbors***	***Tri-gram Predictor***
R^2 (μ)	0.7563	0.7966	0.8045	0.5492
R^2 (σ)	0.0396	0.0321	0.0324	0.0818
RMSE (μ)	95.37	85.40	82.50	122.05
RMSE (σ)	9.61	7.90	8.78	11.76
Parameters	DT: no depth limit for tree RF: employ 300 estimators KNN: K=7, *path* 5x weighted			
Dataset size	1209			497
Number of runs	40			20

Table 1
Summary of performance of each model from test datasets

spans are therefore expressed as

$$d(s_i, s_{i+1}, day), \text{or}$$
$$d(s_i, s_{i+1}, day, time)$$

Here, three regressors namely KNN regressor, DT regressor, and RF regressor are trained with data set $\{x^{(i)}, y^{(i)}\}$ where $x^{(i)}$ denotes the i-th training data sample and $y^{(i)}$ denotes the corresponding predicted value. More specifically

$$x = (busstop(i, j), day, time)$$

where $busstop(i, j) \in \{0, \ldots, 7\}$, $day \in \{0, \ldots, 6\}$, $time \in \{0, \ldots, 9\}$ and y is the duration between bus stops i and j, i.e. $d(s_i, s_j) \in \mathbb{R}$.

The encoding of $busstop(i, j)$ are enumerated for all bus stop pairs along the route. For example, the path from the first to second bus stop is represented by 0, second to third by 1, …, eighth to ninth by 7. The encoding of days in a week is simply an integer where 0, 1, …, 6 represents Sunday, Monday, to Saturday, respectively. Finally, the time of day is discretized and encoded into ten slots, where 0, …, 9 represent 6:00-8:00, 8:00-10:00, 10:00-12:00, 12:00-13:30, 13:30-16:00, 16:00-17:30, 17:30-19:30, 19:30-21:30, 21:30-23:15 and 23:13-24:00 respectively.

After cleaning the data obtained for the month of October 2023, there are 1209 records for the examined service route. The 70-30 repeated hold-out method was used for preparing training and testing datasets for constructing and evaluating all models.

Table 1 summarizes the performance of regression models through the root-mean-square error (RMSE) and the coefficient of determination (R^2) values, which are defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{3}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - y_i')}{\sum_{i=1}^{N}(y_i - \mathbb{E}[y])} \tag{4}$$

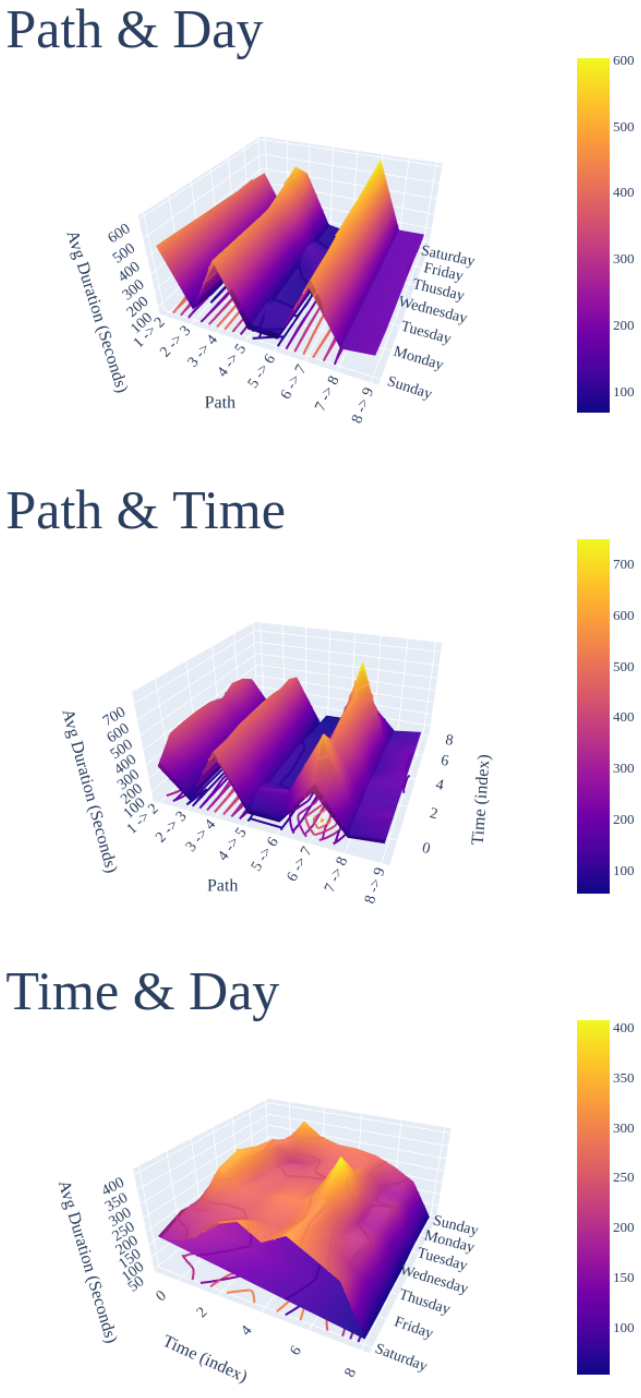

Figure 4: Surface plots illustrating the strength of inluence each variable ($path(s_i, s_i+)$, day, $time$) has on $d(s_i, s_{i+1})$. Evidently $path(s_i, s_i+)$'s is the strongest, which leads to the decision of the increasing its weight in the KNN regressor by 5-fold.

respectively, where y_i is the true value, $\mathbb{E}[y]$ is the expected value, and y_i' is the predicted value. Figure 5 displays the true values (dot) and predicted values (cross) of all regressors. All regressors, i.e., decision tree, random forest and KNN regressors, deliver comparable performance with the same test dataset.

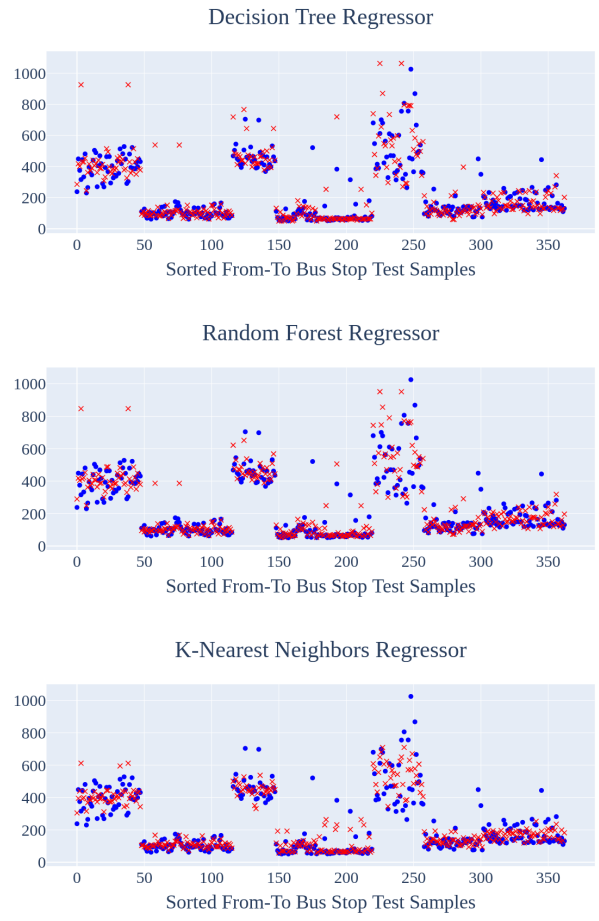

Figure 5: Scatter plots between bus stop data points (x-axis) and duration $d(s_i, s_{i+1})$ (y-axis) of the examined service route. Actual values (blue dot) and predicted values (red cross) are plotted in the same plot for ease of comparison. The plots are output from decision tree regressor, random forest regressor and KNN regressor, respectively. Good correlations between the test and the train data are observed

4.3. Predictive Models

The dynamic of traffic environment, which ETA is dependent upon, can be revealed from previous duration spans observed from preceding bus stops of the same trip. In this project, similar to [14], two previous duration spans were employed in predicting the duration of the next stop. This is expressed as

$$d(s_i, s_{i+1}) \leftarrow \phi([d(s_{i-1}, s_i), d(s_{i-2}, s_{i-1})], C) \qquad (5)$$

where the predicted $d(s_i, s_{i+1})$ is computed from previous duration spans and a given context C. In [14], C would be comprised of only day of week and time of day. However, in our project, we decided to include, in addition to these, $busstop(i, j)$ as another perimeter variable. ϕ denotes the function constructed by training an artificial neural network (ANN), which has the inputs

$$x = (d(s_{i-1}, s_i), d(s_{i-2}, s_{i-1}), busstop(i, j), day, time)$$

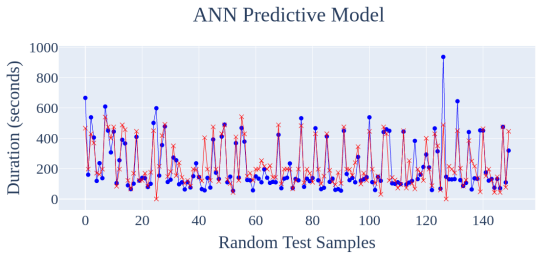

Figure 6: Comparison between predicted durations (red crosses) and actual values (blue dots) from the test dataset. The plot markers are connected to better illustrate the matching movements between the two.

where $d(\cdot,\cdot) \in \mathbb{R}, busstop(i,j) \in \{0,\ldots,7\}, day \in \{0,\ldots,6\}, time \in \{0,\ldots,9\}$, and the ANN's output would be duration $d(s_i, s_{i+1}) \in \mathbb{R}$. After further cleaning the existing data, discarding records with no valid preceding duration information, there are 497 eligible data samples. The 70-30 repeated holdout method is employed in preparing the data, with the resulting train dataset is further divided into equal portions of train and validation datasets. *Keras API* is used in constructing the ANN model, which is comprised of two hidden layers, five input nodes, and one output node. The hidden layers use the *Tanh* activation function, while the output node uses the *ReLU* activation function. The ANN is summarised as such:

```
Layer (type)                 Output Shape              Param #
=================================================================
input (Flatten)              (None, 5)                 0
dense (Dense)                (None, 256)               1536
dense_1 (Dense)              (None, 64)                16448
output (Dense)               (None, 1)                 65
=================================================================
Total params: 18049 (70.50 KB)
Trainable params: 18049 (70.50 KB)
Non-trainable params: 0 (0.00 Byte)
```

The performance of the trained ANN predictive model with the test dataset is also included in Table 1. Using two preceding durations in predicting the current travel duration evidently captures decently the temporal dynamics of the traffic environment. Figure 6 shows the comparison between the predicted durations (red crosses) and actual durations (blue dots).

4.4. Discussion

In this project, we employed machine learning models in order to estimate travel duration between bus stops based on historical temporal records. Data from the examined service route from October 2023 was employed in the constructing of three kinds of models, namely *statistical central tendency*, *regression models* and *predictive models*, which in different fashions suggest the duration between two bus stops.

The statistical central tendency approach suggests the expected $d(s_i, s_{i+1})$ values based on the whole dataset, i.e., the global picture; the regression models suggest estimated values based on nonlinear regression relationship between independent (i.e., $busstop(i,j)$, day of week, time of day) and dependent variable (i.e., $d(s_i, s_{i+1})$). These models provide a natural means to utilize contextual data into the model. Finally, the prediction model suggests estimated $d(s_i, s_{i+1})$ values based on two preceding duration spans (i.e., $[d(s_{i-1}, s_i), d(s_{i-2}, s_{i-1})]$) as well as the current path index ($busstop(i,j)$) and time information context. This model implicitly exploits dynamics of traffic environment via previous duration spans from the same trip. With these models, the ETA of all bus stops can be estimated once the buses start their service.

5. Conclusion

The project introduces follows methods employed by [14] in estimating ETA for public transportation without the use of real-time GPS data. That is, by employing machine learning models based on bus stops' arrival and departure history along the route. Three distinct approaches are utilized, namely statistical central tendencies, nonlinear regression techniques, and a predictive model that considers the previous duration spans on the same route. The study presents a comprehensive analysis of the feasibility of ETA estimation without real-time GPS data, and the proposed machine learning models demonstrate promising results. This work follows the contribution by [14] which include: (i) preprocessing bus service data using data analysis and visualization techniques, (ii) preparing benchmark datasets for further research, and (iii) constructing baseline models, including descriptive statistical models, regression models, and predictive models that serve as benchmarks for the prepared datasets.

6. Supporting Information

Items relevant to this project are

- the Jupyter Notebook script which all code written in performing the tasks here is contained.
- the raw dataset which these tasks are performed on
- a web-based Dashboard consisting of interactive graphs extracted from the notebook.

All of the above are uploaded and documented on this project's GitHub repository:

https://github.com/wafibismail/davis-busroutes

References

[1] Icop safety guideline and industrial code of practice., 2015. URL: https://www.apad.gov.my/sumber-maklumat1/garis-panduan/120-buku-.

[2] E. Mackensen, M. Lai, T. M. Wendt, Bluetooth low energy (ble) based wireless sensors, in: SENSORS, 2012 IEEE, IEEE, 2012, pp. 1–4.

[3] S. Gunady, S. L. Keoh, A non-gps based location tracking of public buses using bluetooth proximity beacons, in: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), IEEE, 2019, pp. 606–611.

[4] H. Yu, R. Xiao, Y. Du, Z. He, A bus-arrival time prediction model based on historical traffic patterns, in: 2013 International Conference on Computer Sciences and Applications, IEEE, 2013, pp. 345–349.

[5] X. Zhang, Z. Liu, Prediction of bus arrival time based on gps data: Taking no. 6 bus in huangdao district of qingdao city as an example, in: 2019 Chinese Control Conference (CCC), IEEE, 2019, pp. 8789–8794.

[6] L. Ye, P. Thiengburanathum, P. Thiengburanathum, A real-time bus arrival time prediction system based on spark framework and machine learning approaches: a case study in chiang mai, in: 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, IEEE, 2021, pp. 243–248.

[7] E. Jenelius, H. N. Koutsopoulos, Travel time estimation for urban road networks using low frequency probe vehicle data, Transportation Research Part B: Methodological 53 (2013) 64–81.

[8] X. Zhan, S. Hasan, S. V. Ukkusuri, C. Kamga, Urban link travel time estimation using large-scale taxi data with partial information, Transportation Research Part C: Emerging Technologies 33 (2013) 37–49.

[9] F. Zhang, X. Zhu, T. Hu, W. Guo, C. Chen, L. Liu, Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations, ISPRS International Journal of Geo-Information 5 (2016) 201.

[10] Y. Wang, Y. Zheng, Y. Xue, Travel time estimation of a path using sparse trajectories, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 25–34.

[11] H. Chen, H. A. Rakha, C. C. McGhee, Dynamic travel time prediction using pattern recognition, in: 20th World Congress on Intelligent Transportation Systems, TU Delft, 2013, pp. 1–17.

[12] J. Wan, J. Liu, Z. Shao, A. V. Vasilakos, M. Imran, K. Zhou, Mobile crowd sensing for traffic prediction in internet of vehicles, Sensors 16 (2016) 88.

[13] R. Barnes, S. Buthpitiya, J. Cook, A. Fabrikant, A. Tomkins, F. Xu, Bustr: Predicting bus travel times from real-time traffic, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3243–3251.

[14] S. Phon-Amnuaisuk, S.-J. Tan, K.-C. Khor, M.-L. Tham, Y.-W. Chong, S. Omar, N. F. B. Ibrahim, M. N. Yusoff, I. Mashud, Non-gps-based eta models constructed from historical gps data and traffic contexts, in: 2023 8th International Conference on Business and Industrial Research (ICBIR), IEEE, 2023, pp. 603–608.