

Prediction of Bus Arrival Time Using Real-Time on-Line Bus Locations

Chan-Tong Lam and Benjamin Ng

School of Applied Sciences
Macao Polytechnic Institute
Macao S.A.R., China
e-mail: {ctlam, bng}@ipm.edu.mo

Su Hou Leong

School of Applied Sciences
Macao Polytechnic Institute
Macao S.A.R., China
e-mail: P1504122@ipm.edu.mo

Abstract—Reliable and accurate prediction of bus arrival time is considered as one of the important services to attract people's choice of bus ridership. In this paper, we develop a simple yet accurate real-time bus arrival prediction system for a crowded small tourist city, like Macao, using accurate on-line bus locations provided by the government website. These accurate bus locations are freely available on-line, which are generated by dedicated sensors installed in bus stops and buses. We first proposed a link time model for storing all of the link times between adjacent bus stops on different bus routes in the entire network so that the trip time between any two of the bus stops in the network can be predicted in real-time. Three different simple models based on historical and real-time on-line bus locations, namely Simple Moving Average (SMA), Artificial Neural Network (ANN) and Hybrid Model, are proposed for the bus arrival prediction system, taking into account the real-time weather conditions. It was found that the Hybrid model perform the best among the three models. The average mean absolute percentage error (MAPE) for the Hybrid model is 17% and the average mean absolute error (MAE) and root mean square error (RMSE) is less than 1 minute. For future works, more advanced deep learning models with Kalman filtering can be evaluated, using on-line bus locations from more bus routes.

Keywords—bus arrival time prediction; real-time bus locations; simple moving average; artificial neural network; hybrid model

I. INTRODUCTION

Public transportation, such as metro trains, light rails or buses, is more efficient in carrying people around than private vehicles, especially in urban area with high traffic volume and population density. Macao is the world's most populated tourist city with a population density of about $21/\text{km}^2$ [1]. Moreover, Macao is receiving on average more than 35 million visitors per year, in addition to about 190,000 non-resident workers traveling to work daily from mainland [2]. Hence, public bus service plays an important role on reducing the traffic burden on the already congested traffic system.

In order to enhance the quality of bus services so as to attract more residents and tourists to access the public transportation services, the Macao Transport Bureau (DSAT) website provides accurate real-time on-line bus locations (by dedicated sensors installed in bus stops and buses) through

the “Bus Location” application [3]. It provides on-line real-time bus locations for bus riders, showing the bus is either at a bus stop or not at a bus stop. The time difference between the bus arriving time and the departing time at a particular bus stop can be obtained from the website. Accurate dwell time can also be obtained. The passengers have to first locate their starting bus stop (or intended starting bus stop) and then find the on-line location of the bus closest to their starting bus stop. Finally, the passengers have to estimate the bus arrival time, based on their riding experience and the number of bus stops their desired bus is stopping. However, the manual estimation of the bus arrival time is usually not accurate enough due to large variation of the arrival time, resulting from various unpredictable factors, such as traffic accidents, street constructions and weather conditions.

An accurate bus arrival time prediction system can motivate people's choice of bus ridership, by reducing the passengers' mean perceived waiting time [4]. In this paper, we develop a simple yet accurate enough real-time bus arrival prediction system for a crowded small city, using these accurate on-line bus locations provided by the DSAT website. The idea is to first collect accurate real-time bus location data from the DSAT website, then the data will be cleaned and used to develop suitable prediction model. Finally, the proposed system will be used to predict the bus arrival times between bus stops, namely the link times, which are then used to predict a particular trip time. Since Macao is a small tourist city with very high traffic density, traffic accidents and road maintenance often occur, resulting in larger variation of bus trip time. Moreover, the distance between adjacent bus stops is so small and the streets are so narrow and complex that even a low level of traffic congestion can greatly affect the accuracy of the bus arrival time prediction.

II. RELATED WORKS

In general, there are three different approaches to obtain accurate real-time prediction of bus arrival times, including models based on historical data, statistical models (regression and Kalman filter), and machine learning models [5][6]. The models based on historical data (using average travel time and average speed obtained from Global Positioning System (GPS) [7]) generally require large number of data and assume a more consistent traffic patterns over the observed time windows. Similar to the historical data models, the applicability of the regression models is

limited due to highly variability of the transportation patterns, especially in urban areas with very high traffic density. One advantage of using Kalman filtering (KF) for bus arrival time prediction is its dynamic time estimation, given recent past data, although the estimation of statistical variance is required by the filtering process [8]-[10]. The KF method becomes unreliable if there is a huge difference in travel time between two consecutive time steps [11]. Machine learning models can be used to deal with non-linear relationships between predictors that can come up with a huge volume of information, especially for processing complicated and noisy data [5]. Popular machine learning models for bus arrival time prediction include Artificial Neural Network (ANN) [8][12][13][14], Support Vector Machine (SVM) models [8][15] and hybrid models [10][16].

Among different models, ANN, SVM and KF are generally used to predict the bus arrival time, while other models are less adopted [12]. Hybrid models are usually used to track the dynamic changes of the traffic conditions, in addition to the prediction of the arrival time based on ANN and SVM model. In [10], the authors combine an ANN model based on historical data and use KF to adjust the prediction based on real-time GPS measurements, taking into accounts the temporal and spatial variations of traffic conditions. In [16], the SVM model is used to predict the baseline travel times on the basic of historical trips, while the KF uses the latest bus arrival information, together with estimated baseline travel times, to predict arrival times at the next station. A complex three-stage mixed model, which consists of traffic delay jitter pattern training, KF and Markov historical transfer model for multi-step prediction, is proposed in [17].

The bus location data used for training of prediction models are mainly based on GPS locations obtained from so called Automatic Vehicle Location (AVL) system [7][8][10][11][12][15]. However, in order to guarantee the stability of data transfer and reduce the data size, it is a common practice to record the location data with a certain time interval, resulting in inaccurate bus locations [12]. Moreover, the accuracy of GPS data can be highly degraded in non-line-of-sight communication links, which often exist in a small and crowded city like Macao. Comparing with the general GPS-based bus location, the accuracy of bus location data provided by a sensor-based system is expected to be higher. Hence, we use these more accurate bus locations to develop an accurate bus arrival time prediction system, based on a hybrid model of Simple Moving Average (SMA) and an ANN model.

III. SYSTEM MODELLING

Fig. 1 depicts the system architecture for the proposed real-time bus arrival time prediction system using the publicly available on-line bus locations from the DSAT website. Since the website only shows accurately either the bus is at the stop or not at the stop, data cleansing is needed in order to obtain meaningful link times for prediction of bus arrival time. The most challenging part of the proposed system is the accuracy of real-time bus arrival prediction.

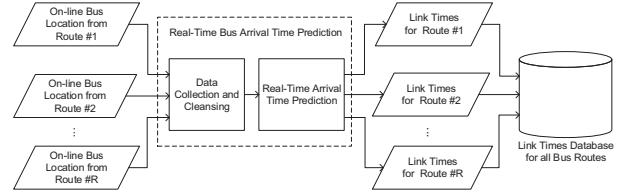


Figure 1. Architecture of real-time bus arrival time prediction system using on-line bus locations.

After obtaining the link times for all the stops on a bus route, these real-time link times for all routes are then stored in a database, from which the users can access to calculate the predicted trip time from and to any bus stops in the bus network. All the link times can be updated simultaneously as long as new observed arrival times are available.

Fig. 2 depicts the link times among different bus stops on different bus routes, where S_n^r in the circle represents the n th bus stop on the r th bus route. We denote the number of bus stops for the r th route as N_r . The link time between two adjacent bus stops is defined as the time difference between the arrival time at the current stop and the arrival time at the next bus stop. The notation $T_n^r[k_n]$ represents the observed link time between the $(n+1)$ th stop and n th stop on the r th bus route. Index k_n represents the most recent link times between the $(n+1)$ th stop and the n th stop stored in the database for bus arrival time prediction. Note that different number of observed link times for a particular link are stored for training of the ANN model for arrival time prediction. The predicted link time between the $(n+1)$ th stop and n th stop on the r th bus route is denoted as \tilde{T}_n^r .

Using the real-time predicted link time for all stops, the predicted trip time for the r th bus route departing from the α th bus stop and arriving at the β th bus stop can be estimated as,

$$\hat{T}_{\alpha\beta} = \sum_{i=\alpha}^{\beta-1} \tilde{T}_i^r, \quad r \in \{0, 1, \dots, R-1\} \quad (1)$$

where R is the total number of routes in the entire bus network, $\alpha < \beta$, $\alpha \in \{0, 1, \dots, N_r - 2\}$ and $\beta \in \{1, \dots, N_r - 1\}$. For the following discussions, we drop the route index r for convenience because we only consider bus arrival time prediction for a particular route. The above modelling of observed and predicted link times can also be used to predict the real-time trip times for a mixed number of bus routes.

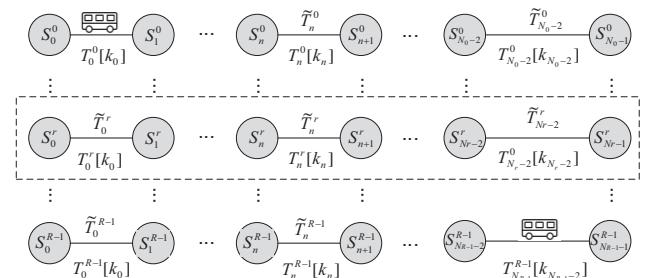


Figure 2. Modeling of observed and predicted link times.

IV. DATA COLLECTION

A. Real-time Bus Locations

Fig. 3 shows an example of the real-time locations provided by the DSAT government website [6]. The location of the bus is shown at two locations in the website, namely either at the bus stop (solid circle with label M16), e.g. the bus MU9003 or not at the bus stop, e.g. the bus MW6117. The bus riders have to estimate the arrival time based on their riding experience, traffic and the weather conditions. From the DSAT website, whenever the bus is at the bus stop, e.g. the bus MU9003 stopping at M16, the status of the approaching bus will be changed from 0 to 1 and then from 1 to 0 when the bus leaves the corresponding bus stop. By continuously checking the status of the buses, the real time locations of all the buses can be obtained. By comparing the arriving times between two adjacent bus stops, the link times for any particular bus can be obtained. Fig. 4 shows an example of data collected from the DSAT website for bus arrival time prediction, including the exact arrival times, day of week, weather conditions and Bus ID. Data cleansing is needed to extract the observed link times.



Figure 3. An example of real-time bus location.

1	StopID	Date	Time	Day of week	Weather	BusID
2	1	12/2/2019	16:42:35	1	1	MP5937
3	1	12/2/2019	17:08:34	1	1	MU9237
4	1	12/2/2019	17:40:03	1	1	MT7926
5	1	12/2/2019	17:45:08	1	1	MU9003
6	1	12/2/2019	17:52:02	1	1	MT7691
7	1	12/2/2019	18:57:03	1	1	MT7774
8	1	12/2/2019	19:26:13	1	1	MT7762
9	1	13/2/2019	16:14:35	2	1	MU9003
10	1	13/2/2019	16:21:12	2	1	MU9501

Figure 4. Example of bus arrival time with weather condition.

B. Weather Conditions

Weather condition is one of the main factors that affect the bus arrival time because more commuters will be driving their private automobiles instead of motorcycles, which are popular transportation tool for daily works due to convenience. More than half of the 250,450 licensed motor vehicles are motorcycles [2]. People prefer motorcycles to automobiles during good weather conditions. Therefore, the weather data is very important to the accuracy of the bus

arrival time prediction. The Open Weather Map website [19] provides real-time weather conditions. The website also provides freely API service for the weather conditions in different formats, such as Java Script Notion (JSON), which are convenient for obtaining synchronized weather conditions with the bus arrival link times obtained through the DSAT website. Table I shows the mapping of weather conditions to numerical values for input to the proposed ANN architecture. The traffic for both weather ‘Clouds’ and ‘Clear’ should be normal to moderate, while the weather condition ‘Rain’ should result in congested traffic.

TABLE I. MAPPING OF WEATHER CONDITIONS

Weather Conditions	Numerical Value
Clouds	1
Clear	2
Rain	3

V. PREDICTION MODELS

A. Simple Moving Average Model

Intuitively, the bus arrival link time can be calculated by taking the simple moving average (SMA) of the previous bus arrival link times (historical data approach), given similar traffic conditions and flow of passengers for a certain period of time. For a particular bus route, the predicted n^{th} link time using the SMA technique, denoted as \tilde{T}_n^{SMA} , can be calculated as,

$$\tilde{T}_n^{\text{SMA}} = \frac{1}{M} \sum_{m=0}^{M-1} T_n[k_n - m], \quad (2)$$

where $T_n[k_n]$ is the most recent observed link time for the n^{th} link and M is the number of previous observed link times used for taking the SMA. The larger the value of M means the predicted link times depends on longer span of previous link times. For short distance (and hence link time) between adjacent bus stops, e.g. those Macao, it is expected the value of M will not be large due to fast traffic variations.

B. Artificial Neural Network Model

Fig. 5 shows a general architecture of the proposed multilayer artificial neural network, where x_i represents the i^{th} input to the input layer, w_i represents the weights for calculating the output to the nodes in the hidden layer, w_k represents the weights for calculating the output to the output layer and \tilde{T}_n represents the n^{th} predicted bus arrival link times. In general, the output to the j^{th} node in the hidden layer y_j can be calculated as the weighted sum of the input nodes, given as,

$$y_j = f(\sum x_i w_i + b_j) \quad (3)$$

where w_i is the weight for i^{th} link to the j^{th} node, $f(x)$ is the activation function, e.g. a Sigmoid function $f(x) = 1/(1 + e^{-x})$, which is common for feedforward neural network and b_j is the bias input. Similarly, the output of the hidden layer

for the n th link time \tilde{T}_n can be calculated as the weighted sum of the output of hidden nodes, given as,

$$\tilde{T}_n = f(\sum x_j w_k + b_n) \quad (4)$$

where w_k is the weight for j th link to the output node, $f(x)$ is the activation function for the output layer and b_n is the corresponding input bias. The network is trained using back-propagation algorithm [20]. It was found that the number of neurons in the hidden layer of a three layer model does not affect the prediction performance significantly in [11]. We chose an ANN architecture with three inputs, one hidden layers with five neurons and a single output of the predicted link time. The key to accurate prediction is to use the most recent link times, which are likely correlated with the current link time to be predicted. Due to the dynamic nature of our data, we only consider two most recent link times. Another input is the weather condition, which is also crucial indication of traffic congestion level.

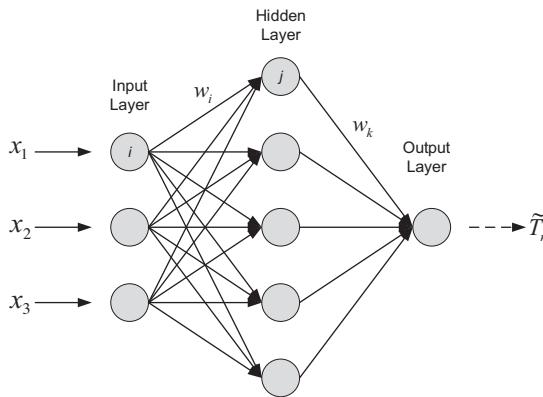


Figure 5. Structure of the proposed artificial neural networks.

C. Hybrid Model

Fig. 6 shows the structure of the proposed hybrid model using the MA and ANN model, where $T_n[k_n]$ is the most recent link time between the n th and $(n+1)$ th stop, while $T_n[k_n - m]$ is the m th previous link time relative to $T_n[k_n]$, M is the number of previous samples taken for the SMA model and W is the numerical input for weather condition.

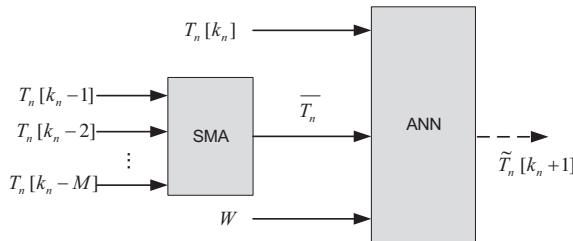


Figure 6. Structure of proposed hybrid model using SMA and ANN.

D. Performance Measure

The following three performance measures are used in our study: Mean Absolute Error (MAE), Mean Absolute Error Percentage (MAPE) and Root Mean Square Error (RMSE), given as [8][13]

$$MAE = \frac{1}{N} \sum_{n=1}^N |T_{\text{observed}}[n] - T_{\text{predicted}}[n]| \quad (5)$$

$$MAPE = \frac{1}{N} \sum_{n=1}^N \frac{|T_{\text{observed}}[n] - T_{\text{predicted}}[n]|}{T_{\text{observed}}[n]} \times 100\% \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^N (T_{\text{observed}}[n] - T_{\text{predicted}}[n])^2}{N}} \quad (7)$$

where $T_{\text{observed}}[n]$ represents the n th true observed link time, $T_{\text{predicted}}[n]$ represents the n th predicted link time and N is the total number of predicted link times.

VI. RESULTS AND PERFORMANCE EVALUATION

We evaluate the performance of the proposed prediction models by collecting the bus arrival link times on two bus routes (Route#3 and Route#102X), as shown in Fig. 7. Route#3, with 17 stops, runs in the main and the busiest streets in the Macao Peninsula, while Route#102X, with 12 stops, runs between Macao Peninsula and Taipa Island, crossing the Pte. Gov. Nobre de Cavalho Bridge. Table II shows the number of link times obtained for each of the bus stop during the busiest hours from 16:00 to 20:00.



Figure 7. Bus routes for data collection.

TABLE II. NUMBER OF DATA FOR MODEL TRAINING AND VALIDATION

	Stop	1	2	3	4	5	6	7	8
#3	Data	398	901	962	921	1040	1045	818	579
	Stop	9	10	11	12	13	14	15	16
	Data	903	1113	1084	931	851	1069	955	472
	Stop	1	2	3	4	5	6	7	8
#102X	Data	397	948	927	896	897	888	863	900
	Stop	9	10	11	12	13	14	15	16
	Data	915	932	901	---	---	---	---	---
	Stop	1	2	3	4	5	6	7	8

A. Effect of Time Period on Mean Arrival Link Times

Fig. 8 shows the average link times on Route#3 obtained during different time periods during weekdays. As expected, the busiest time period lies between 16:00 to 20:00, among which Stop#1 and #8 gives relative larger mean arrival times. Note that Macao is a tourist city, meaning that there will always be a large number of bus riders during non-busy times. Table III shows the number of data collected on Route#3 for obtaining the mean link time during different time periods.

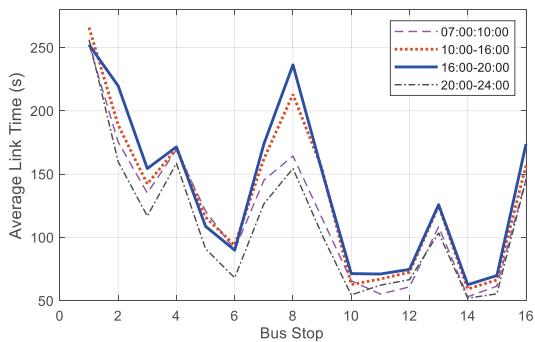


Figure 8. Average link times during different time periods.

TABLE III. NUMBER OF DATA COLLECTED FOR MEAN LINK TIME

Time Period	Data	Time Period	Data
07:00-10:00	331	16:00-20:00	429
10:00-16:00	629	20:00-24:00	415

Fig. 9 show the box plot for the arrival link time during the busiest time period 16:00-20:00. It can be seen that the data is not evenly distributed among different stops and that the variance is different. Hence, different weights for the ANN networks for different stops are expected.

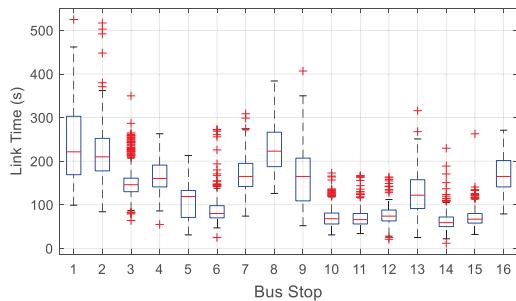


Figure 9. Box plot for arrival link time during period (16:00–20:00).

B. Performance Evaluation

Fig. 10 and Fig. 11 show the MAE, RMSE and MAPE performance using the link times obtained from bus Route #3 and #102X, respectively, with the number of data shown in Table II. There are three models, namely SMA, ANN, and Hybrid, for performance evaluation. For SMA, experimental

results show that three most recent link times give satisfactory performance. In both Fig.11 and 12, we choose $M = 3$ for the SMA performance evaluation. For ANN and Hybrid, 50% of the total data is used for training, while the other 50% is used for validation and testing of the models. For ANN models, the three inputs are two recent link time $T_n[k_n]$ and $T_n[k_n - 1]$ and the numerical weather condition W . For the Hybrid model, $M = 3$ (see Fig. 6) is chosen for SMA calculation. In general, the proposed Hybrid model performs better than the other two models for both bus routes evaluated, as expected. Noticed that the performance for link times at different stops varies, resulting from the varying nature of the traffic conditions at different locations of the stops.

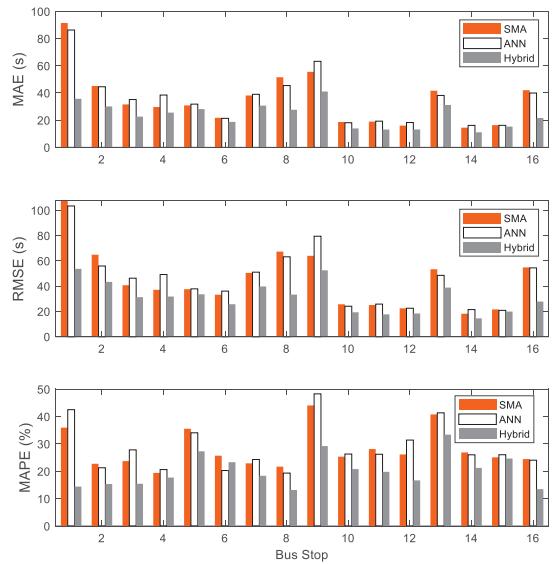


Figure 10. Performance results for Route#3.

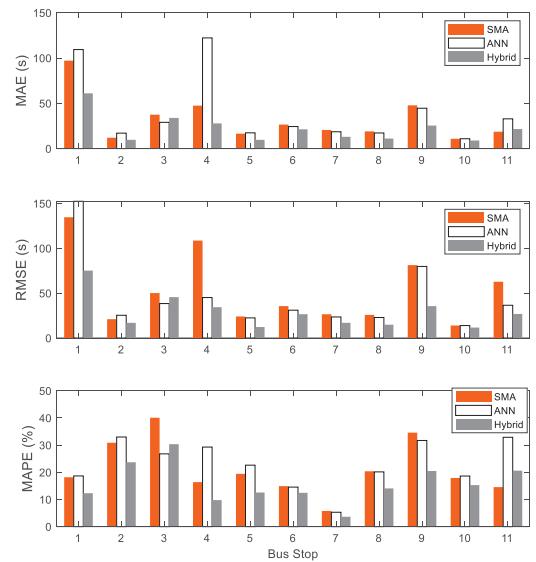


Figure 11. Performance results for Route#102X.

C. Performance Comparison among Models

Fig. 12 shows the comparison of average performance (over all predictions made for Route#3 and Route#102X during the busy time period from 16:00 to 20:00) among the three models. The Hybrid model gives the best performance, while the SMA perform the worst. The average MAPE for the Hybrid model is 17% and the average MAE and RMSE is less than 1 minute.

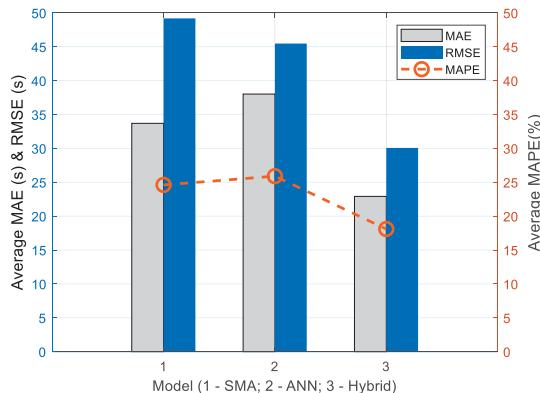


Figure 12. Performance comparison among different models.

VII. CONCLUSIONS

A simple yet accurate enough real-time bus arrival prediction system has been developed and evaluated, using accurate on-line bus locations provided by the DSAT website. An arrival link time based model has been proposed for storing the link times between adjacent bus stops, with which the trip time between any two bus stops in the entire bus network can be predicted in real-time. Three different simple models based on historical and real-time on-line data are proposed, namely Simple Moving Average, Artificial Neural Network, and Hybrid, taking into account the real-time weather conditions. It was found that the Hybrid model perform the best among the three models. The average MAPE for the Hybrid model is 17% and the average MAE and RMSE is less than 1 minute. The proposed system and model can be used to increase bus ridership in Macao. For future works, more advanced deep learning models (e.g. recurrent neural network) with Kalman filtering can be evaluated, using on-line locations from more bus routes.

REFERENCES

- [1] WORLDPOPULATIONREVIEW.com, “World Countries by Population Density 2019”, 2019. [Online]. Available: <http://worldpopulationreview.com/countries/countries-by-density/> [Accessed: 20- July- 2019].
- [2] DSEC.gov.mo, “2019 Macao in Fig.s”, 2019. [Online]. Available: https://www.dsec.gov.mo/getAttachment/12eb1a83-ecbf-4494-9f50-263804f033b3/E_MN_PUB_2019_Y.aspx?disposition=attachment. [Accessed: 19- July- 2017].
- [3] DSAT.gov.mo, “Bus Location Service”, 2019. [Online]. Available: <http://www.dsat.gov.mo/bus/site/busstopwaiting.aspx?lang=tc> [Accessed: 20- July- 2019].
- [4] R. Mishalani, M. McCord, and J. Wirtz, “Passenger Wait Time Perceptions at Bus Stops: Empirical Results and Impact on Evaluating Real-Time Bus Arrival Information,” *Journal of Public Transportation*, vol. 9, no. 2, 2006, pp. 89-106.
- [5] M. Altinkaya and M. Zontul, “Urban Bus Arrival Time Prediction: A Review of Computational Models,” *International Journal of Recent Technology and Engineering*, vol. 2, no. 4, 2013, pp. 164-169.
- [6] R. Chudhary, A. Khamparia, and A. Gahier, “Real Time Prediction of Bus Arrival Time: A Review,” *Proc. 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016)*, Oct. 2016, pp. 25-29, doi: 10.1109/NGCT.2016.7877384.
- [7] D. Sun, H. Luo, L. Fu, W. Liu, X. Liao and M. Zhao, “Predicting Bus Arrival Time on The Basis of Global Positioning System Data,” *Journal of the Transportation Research Board*, No. 2034, 2007, pp. 62-72, doi:10.3141/2034-08.
- [8] C. Bai, Z. Peng, Q. Lu and J. Sun, “Dynamic Bus Travel Time Prediction Models on Road with Multiple Bus Routes,” *Computational Intelligence and Neuroscience*, vol. 2015, pp. 1-9, doi:10.1155/2015/432389.
- [9] W. Fan, and Z. Gurmu, “Dynamic Travel Time Prediction Models for Buses Using Only GPS Data,” *International Journal of Transportation Science and Technology*, vol. 4, no. 4, 2015, pp. 353-366.
- [10] M. Zaki, I. Ashour, M. Zorkany and B. Hesham, “Online Bus Arrival Time Prediction Using Hybrid Neural Network and Kalman Filter Techniques,” *International Journal of Modern Engineering Research (IJMER)*, vol. 3, no. 4, 2013, pp. 2035-2041.
- [11] P. He, G. Jiang, S. Lam, and D Tang, “Travel-Time Prediction of Bus Journey With Multiple Bus Trips,” *IEEE Transactions on Intelligent Transportation Systems*, 2018, pp. 1-14, doi: 10.1109/TITS.2018.2883342
- [12] R. Jeong and R. Rilett, “Bus Arrival Time Prediction Using Artificial Neural Network Model,” *Proc. 2004 IEEE Intelligent Transportation Systems Conference*, October 2004, pp. 988-993.
- [13] J. Amita, J. Singh, G. Kumar, “Prediction of Bus Travel Time using Artificial Neural Network,” *International Journal for Traffic and Transport Engineering*, vol. 5, no. 4, 2015, pp. 410-424, doi: 10.7708/ijtte.2015.5(4).06.
- [14] X. Hua, W. Wang, Y. Wang, and M. Ren, “Bus Arrival Time Prediction Using Mixed Multi-Route Arrival Time Data at Previous Stop,” *Transport*, vol. 33, no. 2, 2018, pp. 543-554, doi: 10.3846/16484142.2017.1298055.
- [15] M. Yang, C. Chen, L. Wang, X. Yan and L. Zhou, “Bus Arrival Time Prediction Using Support Vector Machines with Genetic Algorithm,” *Neural Network World*, vol. 3, 2016, pp. 205-217, doi: 10.14311/NNW.2016.26.011.
- [16] B. Yu, Z. Yang, K. Chen and B. Yu, “Hybrid model for prediction of bus arrival times at next station,” *Journal of Advanced Transportation*, vol. 44, 2010, pp. 193-204, doi: 10.1002/atr.136.
- [17] J. Li, J. Gao, Y. Yang, and H. Wei, “Bus Arrival Time Prediction Based on Mixed Model,” *China Communications*, vol.14, no. 5, 2017, pp. 38-47, doi: 10.1109/CC.2017.7942193
- [18] T. Yin, G. Zhong, J. Zhang, S. He and B. Ran, “A Prediction Model of Bus Arrival Time at Stops with Multi-Routes,” *Transportation Research Procedia*, vol. 25, 2017, pp. 4623-4636, doi: 10.1016/j.trpro.2017.05.381.
- [19] OPENWEATHERMAP.com, “Open Weather Forecast”, 2019. [Online]. Available: <https://openweathermap.org/> [Accessed: 20- July- 2019].
- [20] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Pearson: Prentice Hall, 2009, pp.129–141.