

A LSTM Based Bus Arrival Time Prediction Method

1st Lingqiu Zeng

¹State Key Laboratory of Vehicle NVH and Safety Technology

²Key Laboratory of Advanced Manufacture Technology for Automobile Parts (Chongqing University of Technology)

³College of Computer Science, Chongqing University

Chongqing, China

zenglq@cqu.edu.cn

2nd Guangyan He

College of Computer Science

Chongqing University

Chongqing, China

heguangyan@cqu.edu.cn

3rd Qingwen Han

College of Communication Engineering

Chongqing University

Chongqing, China

hqw@cqu.edu.cn

4th Lei Ye

College of Communication Engineering

Chongqing University

Chongqing, China

Yelei@cqu.edu.cn

5th Fengxi Li

College of Communication Engineering

Chongqing University

Chongqing, China

fengxilee@cqu.edu.cn

6th Lidong Chen

State Key Laboratory of Vehicle NVH and Safety Technology

EMC Testing Department of China Automotive Engineering Research

Chongqing, China

chenlidong@caeri.com.cn

Abstract—Bus arrival time prediction not only provides convenience for passengers, but also helps to improve the efficiency of intelligent transportation system. However, the low prediction accuracy becomes one of great puzzle. Considering both historic data and real-time traffic condition, in this paper, a new bus arrival time prediction method is proposed. A LSTM training model is used to get historic cruising speed, while two traffic factors are defined to illustrate real-time traffic state. Then a bus arrival time prediction is established based on speed values. Validation experiment results show that proposed method could predict the bus arrival time in special time span accurately.

Index Terms—Bus arrival time prediction long short-term memory traffic factor prediction model

I. INTRODUCTION

With the rapid development of the automobile industry, the number of various types of vehicles is increasing, which has put a lot of pressure on the environment and road traffic. The IEA survey found that 63.7% of the world's oil is used for transportation, and 35.3% of carbon dioxide emissions come from the use of oil [1]. Excessive carbon dioxide emissions can cause a series of environmental problems such as greenhouse effect, global warming and accelerated species extinction. The development of intelligent transportation systems can effectively alleviate traffic pressure, reduce pollution and reduce energy consumption. Optimizing public transport and making people prefer to use public transport instead of private cars should be an important part of developing an intelligent

transportation system. However, the uncertainty of the bus arrival time and the uncertainty of the time it takes to get to the destination by bus are two important factors that prevent people from taking the bus. Accurate bus arrival times could encourage people to take more environmentally friendly buses instead of driving cars. Therefore, bus arrival time prediction is an important part of intelligent transportation system.

For the prediction of bus arrival time (BAT), scholars from various countries have done a lot of research. Literature [2] combines real-time GPS data with historical data from the bus transport system in Dublin, Ireland to predict bus arrival time. The predicted time is the real-time delay time provided by GPS plus the historical average arrival time. A kernel regression model is used to represent the dependence between bus arrival time and location updates. In [3], the kalman filtering algorithm is used to obtain the bus arrival time dynamic prediction model. In recent years, researchers have applied machine learning and deep learning to bus arrival time prediction modeling and made progress. Machine learning and deep learning have shown a strong advantage in the field of time prediction. In [4], various methods and multiple bus lines are used to establish a bus arrival time prediction model. Literature [5] proposed the bus arrival time prediction model based on MapReduce combining clustering with neural network. This prediction model is faster and more accurate than the traditional BP neural network prediction model. Chen [6] et al. proposed a bus

arrival time prediction model based on support vector machine (SVM) algorithm and Kalman filter algorithm. The model uses SVM preliminary predictions and then dynamically adjusts the prediction time using the Kalman filter algorithm. Experiments show that this model has higher prediction accuracy. Compared to traditional neural networks, LSTM is a special RNN that remembers longer-term information, then LSTM algorithm is suitable for dealing with problems related to time series. Taking into account the characteristics of bus operations and other influencing factors, LSTM algorithm is chosen in this paper to build bus arrival time prediction model. Overall architecture of proposed methodology is shown in Fig. 1. Here three data pre-processing steps, which are data classification, data filtering and data segment, are involved, and a LSTM model is used to predict the bus arriving time.

The rest of this paper is organized as follows: Section II is preliminaries. Proposed methodology is presented in Section III. Section IV gives the related experiments results, while the conclusion and future works are described in Section V.

II. PRELIMINARIES

A. Raw dataset and coordinate transformation

Raw dataset, which includes driving records of 3000 buses in Chongqing, China, is provided by Cloud Control system of Chongqing Hengtong Bus Company. Each piece of data contains 46 variables, which could be divided into three categories: mechanical properties, operational status, and location information. The sample interval of data collection is 2 seconds. Typical variables in dataset are listed as follows.

- Mechanical properties: engine status, battery status, motor speed and gas pressure.
- Operating status: speed, gear, brake, time stamp, bus_ID.
- GPS location information: longitude, latitude.

In this paper, five attributes, which are bus_ID, speed, time stamp, longitude, and latitude, are selected as basic input features of prediction model. One year driving records of

30 buses for line 805 are used for model-training task. The bus route map of Line 805 is shown in the Fig.2. The GPS coordinate of raw dataset is WGS84, while Amap employs GCJ-02 coordinate system that is the standard of most mobile application. The coordinates obtained in WGS coordinate system have errors if they are used directly in Amap. Hence a coordinate transformation is implemented. In this paper, the transformation is done by the API provided by Amap.

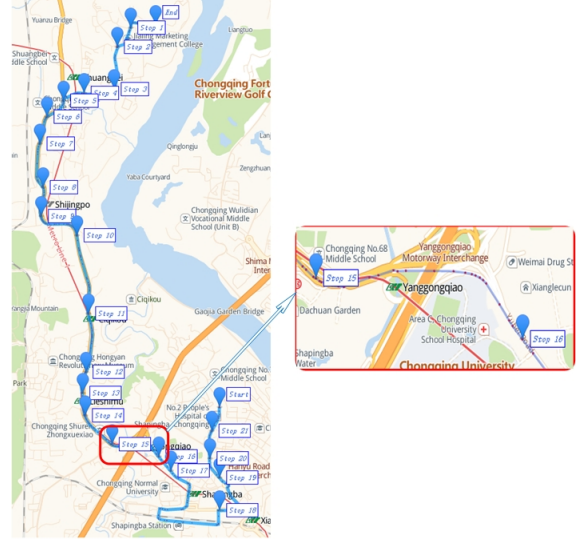


Fig. 2. Bus link.

B. Road section preprocessing and data division

In literature [7], the bus route is divided into a series of consecutive link, which could be expressed as Link set L. A pair of GPS geographical coordinates (longitude and latitude) are used to illustrate each link. Transformation from the original travelling data to the average driving speed data

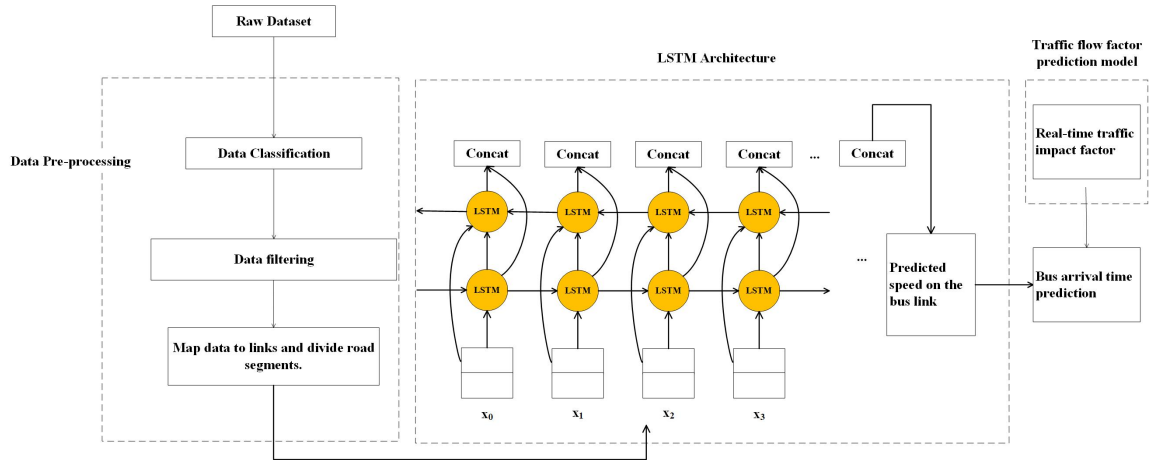


Fig. 1. Overall System Architecture.

within the special segment is executed to train the model based on LSTM-RNN.

The length of road segment is set according to prediction accuracy requirement. The longer the road segment is, the lower the prediction accuracy. Here we set road segment length as 20m, 200m, and 400m respectively. Corresponding equivalent routes are shown in the Fig.3.

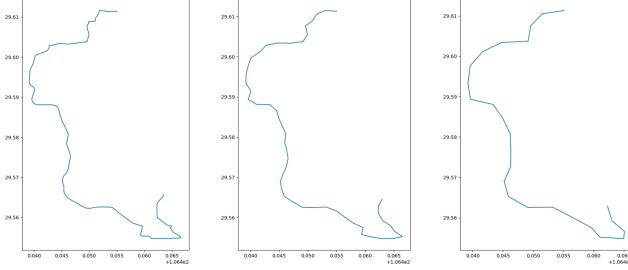


Fig. 3. The divisions of lengths of 20m, 200m, and 400m.

Moreover, to improve the prediction accuracy, corresponding dataset is divided into two categories, namely peak time set and off-peak time set.

C. Real-time traffic factor

Bus arrival time is influence by a series factors, such as road factor, traffic factor, weather factor and time factor, etc. Here we considered both historical experience value and real-time traffic state. The historical experience value could be obtained by historic data machine learning, while the real-time traffic flow could be expressed by real-time traffic factor. In this paper, two traffic impact factors, link factor and bus stop factor, are defined.

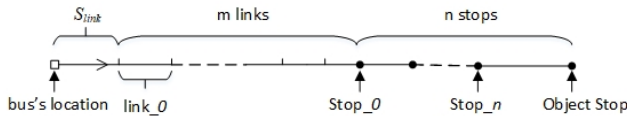


Fig. 4. Real-time traffic factor.

1) *Link factor*: Link factor is used to express the traffic state of current driving segment (CDS), which is the road segment from bus current location to the next stop point. For the time-vary property of road traffic condition, the buss travel speed on a specific CDS is also a time-vary value, which varies around the mean value of historic travel time on object CDS.

Assume current time is t_0 , During time period $[t_0 - 30min, t_0]$, there are n buses pass through current link. Then the traffic impact factor is defined as,

$$\alpha_c = 1 + \frac{1}{n} \cdot \sum_{q=1}^n \frac{(T'_{cq} - \bar{T}_c)}{\bar{T}_c} \quad (1)$$

where \bar{T}_c is the historical travel time of CDS, while T'_{cq} is current travel time of bus. $q, q \in \{n\}$. Likewise, we can

obtain traffic impact factor of all road segment between bus stops, such as segment between $stop_k$ and $stop_k + 1$, and denote as $\alpha_{\gamma(k,k+1)}$.

2) *Bus stop factor*: Another impact factor influencing travel time is bus stop dwell time.

Assume current time as t_0 , During time period $[t_0 - 30min, t_0]$, there are m buses, denote as set $\{m\}$, roll into object bus stop, we construct traffic impact factor of forward stops as,

$$\alpha_s = 1 + \frac{1}{m} \cdot \sum_{p=1}^m \frac{(T'_{sp} - \bar{T}_s)}{\bar{T}_s} \quad (2)$$

where \bar{T}_s is the time when the historical bus stops at the bus stop and T'_{sp} is bus stop dwell time. $p, p \in \{m\}$.

D. Bus arrival time prediction

Considering both link factor and bus stop factor, bus arrival time could be calculated as,

$$T = \frac{S_{link}}{V} + \alpha_c \bar{T}_c + \sum_{k=0}^p [\alpha_s(k) \bar{T}_s(k) + \alpha_{\gamma(k,k+1)} \bar{T}_{(k,k+1)}] \quad (3)$$

S_{link} is the distance between the current bus position and the end of the current link. Bus real-time speed V could be calculated as

$$V = \frac{d_1}{L} V_i + \frac{d_2}{L} \bar{V}_s \quad (4)$$

where d_1 is the distance between link start node and bus current location, while d_2 is the distance between bus current bus location and link end node. L is the link length. V_i is current bus speed and \bar{V}_s is historic experience average speed of the current road section. \bar{T}_c is the time through CDS. p indicates that there are p bus stops between the current location and the predicted target bus stops. $\bar{T}_s(k)$ is the predicted stop dwell time of the bus at the k th stop. $\bar{T}_{(k,k+1)}$ is the average travel time between $stop_k$ and $stop_k + 1$. α_c , $\alpha_s(k)$ and $\alpha_{\gamma(k,k+1)}$ are the CDS impact factor, stop factor and Inter-station impact factor respectively.

III. PREDICTION MODEL

A. Test set and Training set

The path length of line 805, Chongqing, China, is about 11,000 meters. There are 26 bus stops along bus route. The amount of data is approximately 44,600, while the data collection period is April 2016 to December 2016. Five thousand of them were randomly selected as test sets, and the others were training sets. To improve prediction accuracy, we set road segment length as 20m.

As mentioned earlier, we divide raw dataset into two categories, which are peak time dataset and off-peak dataset. Here we define peak time periods as 7am-9am and 17pm-19pm. Then we got about 31,600 pieces off-peak data and about 13,000 peak time data.

The basic data set of the model is the actual time that passes through each section. The arrival time prediction model mainly consists of CDS travel time, bus stop dwell time and

travel time between stations. Therefore, LSTM model needs to predict these three parts of data separately, while the input data and parameter settings of the three models are different.

B. Model Construction and Training

The original data of the experiment is the driving data of 805 bus Stop2 to Stop19. The whole bus line is divided into 400 20-meter links. According to the GPS data calibration strategy, the data points are calibrated to the link. The input data and training data of LSTM model are the travel time of each link. The first layer of the model is the input layer, which needs to specify the input dimension, the number of data, the form of input data and so on. The input data of the model is 400 driving times in seconds.

In order to facilitate the input, 400 data are transformed into 20 rows and 20 columns matrix, which is the input matrix of the model. The input layer dimension of the model is set to 20 (INPUT_SIZE = 20); the hidden layer dimension is set to 100 (CELL_SIZE = 100), the number of layers is 2 (HIDDEN_LAYER = 2); and the output layer dimension is also set to 20 (OUTPUT_SIZE = 20). The activation function of the LSTM uses tanh function. After many experiments, it is found that using 0.006 as the learning rate (LR = 0.006) can improve the performance of the model. A loss function, mean squared error, is selected in this paper to illustrate prediction accuracy performance, while a gradient-based optimizer is used as essential component of compilation model. As mentioned in literature [8], comparing with other optimization algorithms, Adam algorithm performs a better accuracy characteristic. Hence, in this paper, we select Adam algorithm as basic algorithm for optimizer. The input of the model is time series on 20 links and the output is 20 time series on the next 20 links. The latter data series will be affected by the former data series. In order to enhance the dependency of the input sequence, the dependency length of the model is set to 20 (TIME_STEPS = 20). Finally, Tensor Board, a visualization tool, is used to visualize the complex training process of neural networks.

Table 1 lists the relevant parameter settings for each predictive model.

TABLE I
LSTM NETWORK PARAMETERS

	CDS travel	bus stop dwell	Interstation travel
INPUT_SIZE	20	1	1
OUTPUT_SIZE	20	1	1
CELL_SIZE	100	10	10
HIDDEN_LAYER	2	1	1
BATCH_SIZE	20	20	20
TIME_STEPS	20	3	3
LR	0.006	0.006	0.006

C. Prediction effect evaluation

In this paper, root mean square error (RMSE) and mean absolute error rate (MAPE) were used to evaluate the prediction performance of the model. Corresponding definitions are given as follows,

$$RMSE(T) = \sqrt{\frac{\sum_{i=1}^n (T(i) - T^*(i))^2}{n}} \quad (5)$$

$$MAPE(T) = \frac{1}{n} \sum_{i=1}^n \frac{|T(i) - T^*(i)|}{T(i)} \quad (6)$$

where $T(i)$ and $T^*(i)$ are the actual arriving time and predicted arriving time respectively, n is the bus stop number, while i is the serial number of bus stop.

IV. EXPERIMENT RESULTS

As mentioned earlier, here we only consider the path between link 2 and link 21. Hence, we set the time of start point for link2 as 0, which is considered as reference point for subsequently prediction process.

A. Experimental results and analysis

1) *Link travel time analysis*: This experiment combines RMSE and MAPE to analyze the link prediction time between Stop10 and Stop11, as shown in Fig.5.

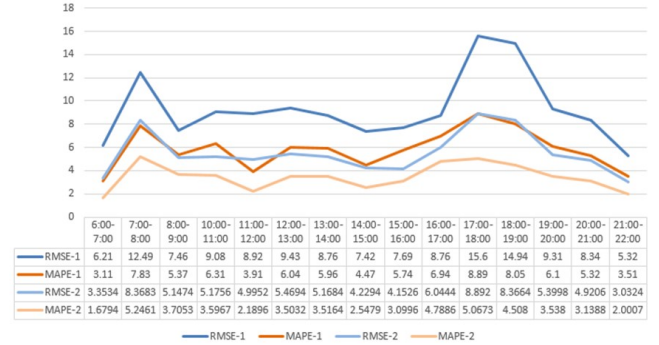


Fig. 5. Link time operation comparison analysis.

RMSE-1 is the root mean square error of the output value of the learning model compared with the real value. RMSE-2 is the root mean square error of output value of learning model weighted real-time traffic flow. MAPE-1 and MAPE-2 are similar. The results show that the prediction accuracy can be improved by weighting real-time traffic factors.

2) *Arrival time prediction*: In the previous section, we analyzed the unweighted prediction time of each link, and the predicted time weighted by the real-time traffic flow factor. In this section, we focus on the inter-station travel time between 7 and 20 o'clock, and other time periods are similar. Arrival time prediction outputs are shown in Figs. 6 to 9. Corresponding parameters are listed in table 2. The performance of proposed method is contrasted with literature [7]'s output.

As shown in Figs. 6 to 9, the prediction arrival time coincide with actual arriving time. Moreover, both RMSE and MAPE performance of proposed method are better than that of literature [7]. As shown in Table2, the MAPE value of morning and evening peak is about 0.06, while that of noon off-peak is less than 0.05. Meanwhile, the speed prediction RMSE value

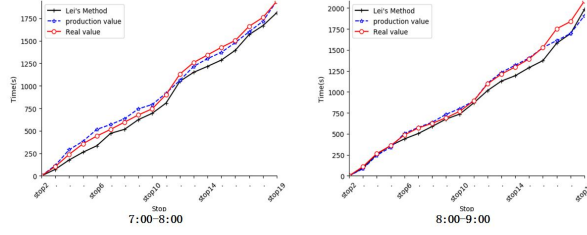


Fig. 6. Bus arrival time prediction(7:00-9:00).

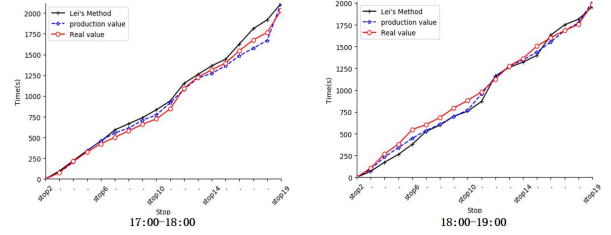


Fig. 9. Bus arrival time prediction(17:00-19:00).

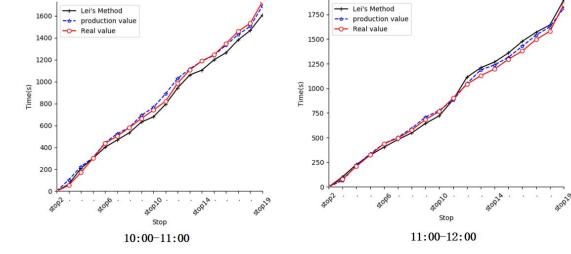


Fig. 7. Bus arrival time prediction(10:00-12:00).

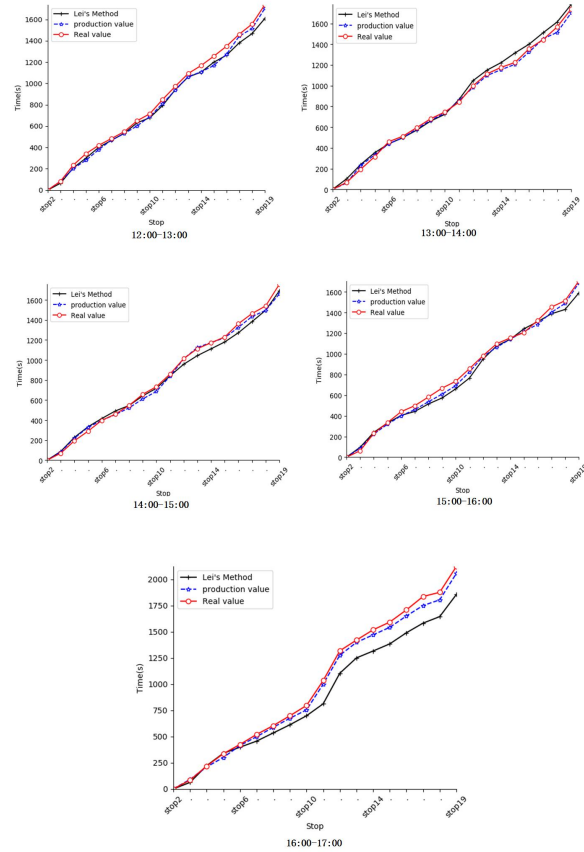


Fig. 8. Bus arrival time prediction(12:00-17:00).

TABLE II
RMSE&MAPE PERFORMANCE OF PROPOSED METHOD

dataset	Time slot	Evaluation index			
		LTSM's Method		Lei's Method	
		RMSE	MAPE(%)	RMSE	MAPE(%)
off-peak	10:00-11:00	32.08	6.31	58.02	7.18
	11:00-12:00	29.92	3.91	53.77	5.97
	12:00-13:00	44.43	6.04	57.36	7.34
	13:00-14:00	23.76	3.60	43.62	6.81
	14:00-15:00	35.42	4.49	48.79	6.36
	15:00-16:00	33.32	5.74	59.77	8.43
peak	7:00-8:00	47.36	6.33	92.98	15.56
	8:00-9:00	67.46	5.37	87.65	7.66
	16:00-17:00	39.76	6.94	44.61	4.12
	17:00-18:00	43.64	8.31	80.10	14.95
	18:00-19:00	56.94	8.05	86.59	17.58

of evening peak is also higher than that of noon peak. That is said, performance of proposed method is influenced by real-time traffic state. Hence, in future work, we should improve the validity of real-time traffic factors.

CONCLUSION

Bus arrival time prediction is a key point for intelligent transportation system design. To improve the prediction accuracy, in this paper a LSTM base method is proposed. The key factors, which influence prediction accuracy, are discussed. Two kinds of real-time traffic related factors, which are link factor and bus stop factor, are defined. Then suggestion speed value of each link is obtained through training process. Two time sets, peak time set and off-peak time set, are considered in arrival time prediction. The experiment results show that proposed method presents high prediction accuracy. In our future works, the prediction time period should be shortened, while the GPS position calibration process should be considered.

ACKNOWLEDGMENT

This research is supported by National Nature Science Foundation of China, Project No. 61601066. Thanks for open research fund of State Key Laboratory of Vehicle NVH and Safety Technology Grant No: NVH SKL-201913. Thanks for the Key Laboratory of Advanced Manufacture Technology for Automobile Parts (Chongqing University of Technology), Ministry of Education, No. 2016KLMT01 and No.2017KLMT04. Thanks for Fundamental Research Funds for the Central Universities No.2018CDXYTX0009.

REFERENCES

- [1] Ozener, O., REAL DRIVING EMISSIONS AND FUEL CONSUMPTION CHARACTERISTICS OF ISTANBUL PUBLIC TRANSPORTATION. *Thermal Science*, 2017. 21(1): p. 655-667.
- [2] Sinn, M., et al. Predicting arrival times of buses using real-time GPS measurements. in *International IEEE Conference on Intelligent Transportation Systems*. 2012.
- [3] Mei, C., X. Liu, and J. Xia, Dynamic Prediction Method with Schedule Recovery Impact for Bus Arrival Time. *Transportation Research Record Journal of the Transportation Research Board*, 2005. 1923(1): p. 208-217.
- [4] Tero, A., et al., Rules for Biologically Inspired Adaptive Network Design. *Science*, 2011. 327(5964): p. 439-442.
- [5] Zhang, J., et al. Method of Predicting Bus Arrival Time Based on Map Reduce Combining Clustering with Neural Network. 2017.
- [6] Chen, X.M., H.B. Gong, and J.N. Wang, BRT Vehicle Travel Time Prediction Based on SVM and Kalman Filter. *Journal of Transportation Systems Engineering & Information Technology*, 2012. 12(4): p. 2934.
- [7] Jianmei, L., et al. A Bus Arrival Time Prediction Method Based on GPS Position and Real-Time Traffic Flow. in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*. 2017.
- [8] Li, P., M. Sun, and M. Pang, Bus Arrival Time Prediction Based on LSTM. *Proceedings of the 37th Chinese Control Conference*, 2018.