

Research Article

Yunna Wang*

Intellectualization of the urban and rural bus: The arrival time prediction method

<https://doi.org/10.1515/jisys-2021-0017>

received January 30, 2021; accepted March 13, 2021

Abstract: To improve the intelligence of urban and rural buses, it is necessary to realize the accurate prediction of bus arrival time. This paper first introduced urban and rural buses. Then, the arrival time prediction was divided into two parts: road travel time and stop time, and they were predicted by the support vector regression method and *k*-nearest neighbor (KNN) method. A section of a bus route in Pingdingshan city of Henan province was taken as an example for analysis. The results showed that the method designed in this study had better accuracy, and the result was closer to the actual value, with a maximum error of 84 s, a minimum error of 10 s, an average error of 42.5 s, and an average relative error of 5.74%, which could meet the needs of passengers. The results verify the reliability of the designed method in predicting the arrival time of urban and rural buses, which can be popularized and applied in practice.

Keywords: urban and rural bus, intelligent bus, arrival time prediction, relative error

1 Introduction

With the development of economy, China's urbanization process has accelerated [1], the links between cities and towns, towns, and townships have become more and more close, and the number of floating people and the floating scope are also increasing, which also stimulates the increase in transportation demand. Although the current urban and rural bus can meet the unified needs of people to some extent, there is still a gap with the urban bus, such as the imperfect infrastructure, the lack of unified planning, the difficulty in passenger transfer, and the long waiting time. Optimizing urban and rural buses is not only a means to narrow the gap between urban and rural areas but also a very important means to accelerate urbanization and promote rural development. The integration of urban and rural buses is to realize the extension of urban buses to rural areas and the integration of urban buses with rural areas to meet people's travel needs. The urban and rural bus is an important link to connect urban and rural economy, which can solve the chaos of the urban and rural passenger transport market. Moreover, it also plays a good role in solving the problem of urban traffic congestion. As a means of transportation in and out of the city, it has advantages. Urban and rural buses have longer routes than urban buses and have an urgent demand for vehicle information feedback. However, at present, the investment of intelligent facilities in urban and rural buses is significantly insufficient compared with urban buses; therefore, passengers cannot get timely feedback of vehicle arrival time at the bus station, leading to low convenience degree and low passenger satisfaction. Therefore, intelligent means are urgently needed to realize the automatization and informatization of buses [2] to further improve the convenience of buses [3] and passenger satisfaction. How to realize intelligent buses has been extensively discussed by researchers. The prediction of arrival time is an

* **Corresponding author: Yunna Wang**, Department of Urban and Rural Planning, School of Architecture and Urban Planning, Henan University of Urban Construction, Longxiang Avenue, Xincheng District, Pingdingshan, Henan 467036, China, e-mail: nyoa10@163.com

important part of intelligent buses. Currently, studies on the prediction of bus arrival time mainly concentrate on urban buses, and the methods used can be divided into three types. The first type of method is based on the spatiotemporal variation rules, i.e., predicting the future driving time according to historical data based on the spatiotemporal variation rules of bus driving time. Yang et al. [4] proposed a residual neural network-based method, which used two features about bus travel time and headway target section in both forward and reverse directions, and established the temporal-spatial information to predict the bus arrival time. They verified the effectiveness of the method through an experiment on Shenzhen No. 10 route. The second type of method is based on a variety of factors, such as weather, traffic flow, and road length. Based on the weather, working days/holidays, bus number, driver number, and other factors, Zhang [5] established a bus arrival time prediction model, took No. 1 bus in a city as the research subject to carry out the experimental verification, and obtained good prediction results. Sun et al. [6] designed a new system based on the real-time position data of the positioning system, average speeds of different line segments, the historical speed, and the temporal and spatial changes in traffic conditions. The case study on the actual bus found that the system had satisfactory accuracy. By combining the time vector-based method with the space vector-based method, Liu et al. [7] designed a comprehensive model based on long short-term memory and artificial neural network and verified the accuracy of the method through the entity dataset. The third type of method is a data fusion prediction method, which combines a variety of prediction methods to get better results. Currently, studies about the prediction of bus arrival time mainly took the urban bus as the subject and pay little attention to the urban and rural bus, and the prediction accuracy was also low. Based on the characteristics of the urban and rural public bus, this study predicted its arrival time. Aiming at the low prediction accuracy of the arrival time, this study divided the arrival time into two time periods, predicted the arrival time with two models, and carried out experiments on the actual bus operation data to verify the effectiveness of the method. This study makes some contributions to the better service of urban and rural buses.

2 The prediction model of urban and rural bus arrival time

2.1 Historical data-based prediction model

Accurate prediction of arrival time can reduce the waiting time of passengers and facilitate passengers to ride or transfer. It has a positive impact on improving passenger satisfaction and attracting passengers to travel by public transport. It is also the basis of scientific and reasonable dispatching of buses by bus enterprises. Generally, the forecast value of arrival time is displayed by the electronic station board in the station, but the electronic station board has not been widely distributed currently, and the accuracy of the forecast is not high. For urban and rural buses, the factors that affect the arrival time include (1) traffic congestion caused by emergencies; (2) speed slowdown caused by bad weather (fog, snow, etc.); (3) traffic congestion caused by heavy traffic during holidays; and (4) speed slowdown caused by rush hour.

At present, in the arrival time prediction, the most widely used method is the historical data-based prediction. It is assumed that the traffic mode is cyclic, and the parameters required for the prediction are presented in Table 1. Then, the prediction formulas are written as:

$$T_{it} = A_{i+1,t} - (A_{it} + W_{it}), \quad (1)$$

$$A_{jtk} = A_{Stk} + \sum_{i=M}^{N-1} T_{it} + \sum_{i=M}^{N-1} W_{it}. \quad (2)$$

The historical data-based prediction method is to average all the data of a period of historical operation to predict the arrival time of a bus. This method is simple in principle and easy to operate, but it needs a lot of historical data to support, showing poor real-time performance. The route of urban and rural buses is long, including not only rural roads with little congestion but also congested urban roads, which is very complex; therefore, this method is not suitable for the prediction of urban and rural buses.

Table 1: Parameters of the historical data-based prediction model

Parameter	Meaning
T	The number of periods divided
T_{it}	The time needed by a bus driving from station i to station $i + 1$
A_{it}	The average time needed by a bus driving from the first station to station i
$A_{i+1,t}$	The average time needed by a bus driving from the first station to station $i + 1$
W_{it}	The average waiting time at station i
A_{jtk}	The time needed by bus k driving from the first station to station j
A_{Stk}	The time needed by bus k driving from the first station to station j
S	The number of the current station
N	The number of the terminal station

Specifically speaking, the arrival time of a bus can be divided into two parts, one is the travel time of the bus on the road, and the other is the stop time of the bus at the station. However, most of the current research does not clearly distinguish them, leading to low accuracy of prediction. This study divided the arrival time prediction of a bus into two parts and calculated them respectively.

The travel time of a bus on the road refers to the time when the bus leaves station i and arrives station j . The time needed by a bus driving from station i to station $i + n$ is denoted as T . The stop time at station $i + n - 1$ is denoted as t_{i+n-1} . The travel time on the road is denoted as $T - t_{i+1} - t_{i+2} - \dots - t_{i+n-1}$. The travel time on the road is predicted by the support vector regression (SVR) method, and the bus stop time is predicted by the k -nearest neighbor (KNN) algorithm.

2.2 Travel time prediction model

Support vector machine (SVM) is a method of data classification [8]. SVM can mine rules from limited data and predict the unknown phenomena. Its learning ability and generalization ability are superior to neural networks. SVM can be used for predicting the driving time of road segments. In SVM, the classification hyperplane can be expressed as:

$$g(x) = wx + b, \quad (3)$$

where x refers to the sample vector. The class interval can be expressed as:

$$d(w, b) = \frac{2}{\|w\|}. \quad (4)$$

The process of solving the optimal hyperplane is equal to the process of solving $\frac{2}{\|w\|}$, i.e.,

$$\min_{w,b} \frac{1}{2} \|w\|^2, \quad (5)$$

$$s \cdot t \quad y_i(wx_i + b) - 1 \geq 0. \quad (6)$$

Lagrange multiplier a is introduced to establish the Lagrange function. Then, the problem is transformed to:

$$\min_a \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N a_i, \quad (7)$$

$$s \cdot t \quad \sum_{i=1}^N a_i y_i = 0. \quad (8)$$

As SVM is mainly used for data classification, to predict the travel time, it is necessary to use the SVR algorithm [9]. Based on SVM, it is assumed that the input data is x_i and the output data is y_i . In the M -dimensional space, the classification hyperplane is written as:

$$f(x) = \sum_{i=1}^M w\phi(x) + b. \quad (9)$$

After the ϕ transformation of the input data, the insensitive loss function is set as:

$$L_\varepsilon(x_i, y_i, f) = \max\{0, |y_i - f(x_i)| - \varepsilon\}. \quad (10)$$

The SVR algorithm needs to find proper f to make $E(\varepsilon)$ minimum:

$$E(w) = \frac{1}{2}w^2 + C \left(\sum_{i=1}^1 L_\varepsilon(x_i, y_i, f) \right), \quad (11)$$

where w stands for the linear weight vector and C stands for the penalty factor. The above equation is solved by two relaxation vectors, δ_i and δ'_i , then,

$$\max \sum_{i=1} y_i(a'_i - a_i) + a_i - \frac{1}{2} \sum_{i=1} \sum_{j=1} (a_i - a'_i)(a_j - a'_j)K(x_i, y_j), \quad (12)$$

where a_i and a'_i are the corresponding Lagrange multipliers of δ_i and δ'_i and $K(x_i, y_j)$ is the kernel function. Finally, the output of the SVR algorithm can be written as:

$$f(x) = \sum_{i=1} (a_i - a'_i)K(x, x_i) + b, \quad (13)$$

i.e., the prediction result of the travel time of the bus on the road.

2.3 Stop time prediction model

The bus stop time at the station is affected by the number of people getting on and off the bus and the number of people on the bus when it is fully loaded. If a model is constructed based on the above factors, the support of a large amount of data is needed, and the generalization of the model is not strong. To predict the stop time, several data most similar to the current situation are found from the historical data, and the weighted average value is taken as the prediction result. The KNN algorithm [10] can realize it. The steps of predicting the stop time with the KNN algorithm are as follows.

- (1) In the training data, the nearest neighbor data are collected. k nearest-neighbor data were found out. The calculation formula of the distance is:

$$ED_i = \sqrt{\sum_{j=1}^{n-1} (T_j^i - T_j^m)^2}, \quad (14)$$

where i refers to the bus number, m refers to the bus number to be predicted, j refers to the station number, and T_j^i refers to the stop time of bus i at station j .

- (2) The weighted average value of k data is calculated:

$$T_n^m = \sum_{k=1}^K \left(T_n^k \times \frac{1/ED_k}{\sum_{k=1}^K \frac{1}{ED_k}} \right). \quad (15)$$

3 Case analysis

Xincheng District, Pingdingshan City, Henan, is located in the west of the area and is connected to the old city by urban roads and light rail. The transportation in the area is convenient, and it is the leading area of urban–rural integration. Therefore, the construction of intelligent urban and rural buses plays a positive role in developing the region. In this study, bus No. 16 was taken as an example. Bus No. 16 started from Xiangshan temple and ended at the Zunhuadian government, with 48 stations. The studied road segment was from Xiangshan temple to municipal building, with ten stations in total, as presented in Table 2. Global position system (GPS) operation data of bus No. 16 were obtained from Pingdingshan bus company, and the data from July 1 to July 31, 2020, were collected. A model was established based on the data of the first 30 days, and the data of the last day was used for verification. GPS data included bus line name, direction, date, coordinate, speed, etc. The missing data were supplemented by the adjacent data. The duplicate data were deleted. Then, the data were normalized. Moreover, the bus was surveyed by follow-up, and the data, such as the bus opening time, closing time, the number of passengers getting on and off, arrival time, and departure time, were recorded (Figure 1).

Table 2: Road section for study

Number	Name of station
1	Yangguanying
2	Yangguanying South
3	Xinxin community
4	No. 7 middle school
5	Henan university of urban construction
6	Municipal building (Xiangyun Park)
7	Xiangshun road crossing
8	Municipal approval service center
9	Municipal building (municipal intermediate court)
10	Municipal transportation bureau (China Guangfa Bank)

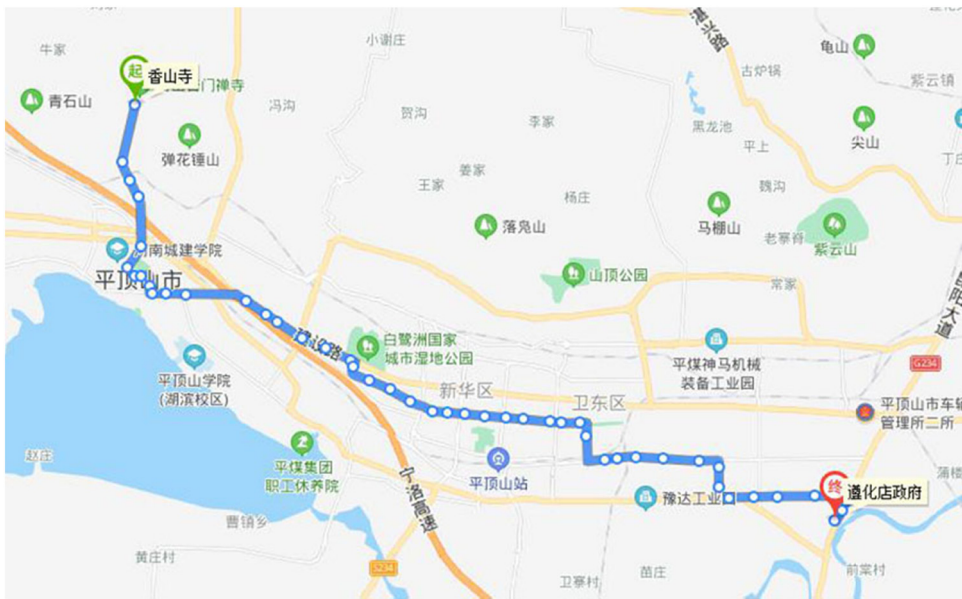


Figure 1: Route of bus No. 16 [11].

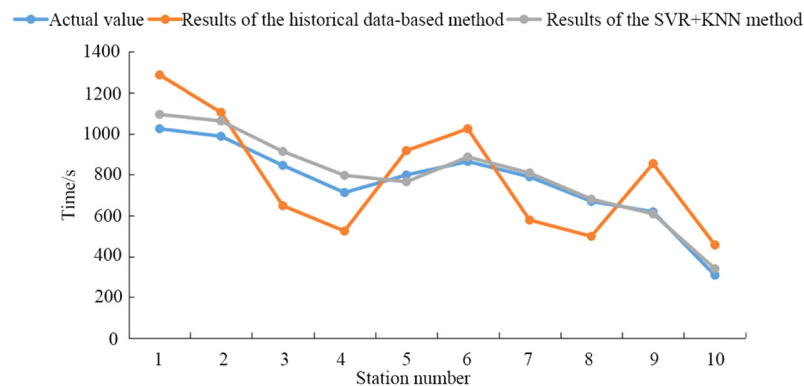
Table 3: Bus operation data on 1 day

Number	Station	Stop time (s)	Travel time (s)
1	Yangguanying	96	892
2	Yangguanying South	78	743
3	Xinxin community	64	656
4	No. 7 middle school	121	542
5	Henan university of urban construction	143	678
6	Municipal building (Xiangyun Park)	32	864
7	Xiangshun road crossing	54	715
8	Municipal approval service center	21	638
9	Municipal building (municipal intermediate court)	31	496
10	Municipal transportation bureau (China Guangfa Bank)	28	212

Taking July 1 as an example, the operation data of the bus are presented in Table 3.

The prediction performance of the model was evaluated by the relative error, $y = \frac{|x - x'|}{x}$, where x is the actual value and x' is the predicted value. The time of the bus arriving at the next station was predicted using the SVR method and KNN methods designed in this study. Moreover, the same data were also predicted using the historical data-based prediction method. The results are shown in Figure 2.

It was seen from Figure 2 that there was a big gap between the results obtained by the historical data-based prediction method and the actual situation, and the results of the SVR + KNN method were closer to the actual values. The prediction error was calculated, and the results are presented in Table 4.

**Figure 2:** Comparison of prediction results.**Table 4:** Comparison of prediction errors

Number	Prediction error of historical data (s)	Prediction error of the SVR + KNN method (s)
1	263	70
2	117	75
3	197	68
4	188	84
5	119	33
6	160	22
7	211	19
8	171	11
9	236	10
10	149	33

After calculation, the maximum error of the results obtained by the historical data-based prediction method was 263 s, the minimum error was 117 s, and the average error was 181.1 s, whereas the maximum error of the results obtained by the SVR + KNN method was 84 s, the minimum error was 10 s, and the average error was 42.5 s, which were significantly smaller than the former. The relative errors of the two methods were calculated, and the results are shown in Figure 3.

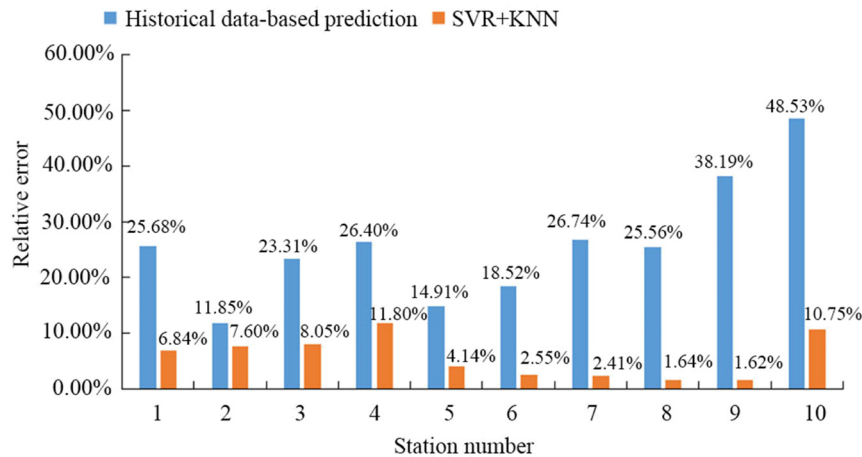


Figure 3: Comparison of relative errors.

It was seen from Figure 3 that the relative error of the SVR + KNN method was significantly smaller. The maximum relative error of the historical data-based prediction method was 48.53%, the minimum relative error was 11.85%, and the average error was 25.97%. However, the maximum relative error of the SVR + KNN method was 11.8%, the minimum error was 1.62%, and the average error was 5.74%. Based on the above results, it was found that the SVR + KNN method could predict the arrival time of urban and rural buses, showing a favorable accuracy.

To further verify the performance of the method, the arrival time of five successive buses was predicted. The road segment was the same as above. The average values of three errors of the prediction based on the historical data and the SVR + KNN method are presented in Table 5.

Table 5: The comparison of prediction errors of five successive buses

Bus number	The error of the prediction based on the historical data/s	The error of the prediction based on the SVR + KNN method/s
1	183	43
2	189	46
3	201	51
4	193	48
5	224	52
Average value	198	48

It was found from Table 5 that the error of the prediction of the arrival time based on the historical data had some fluctuations, the maximum error was 224 s, and the average value was 198 s; the error of the prediction of the arrival time based on the SVR + KNN method kept stable, about 50 s, and the average value was 48 s, which was 150 s less than the former. The results demonstrated that the SVR + KNN method had a good performance in predicting the arrival time of buses.

4 Discussion

The current research methods about intelligent buses can be roughly divided into three directions: first is for bus enterprises, mainly studying the scheduling, operation, and assessment of buses [12]; second is for government departments, mainly studying the planning of bus routes and the analysis of passenger flow distribution [13]; and the third is for passengers, such as arrival query [14]. The technologies used in arrival query include geographic information system [15], data mining [16], intelligent terminal [17], machine learning [18], etc. Based on the GPS data of buses, Wei and Sun [19] designed a bus scheduling planning model. The lower model coordinated the travel time by transfers between buses and between buses and subways, and the upper model allocated the bus to reduce the operation cost. Zhang et al. [20] studied the prediction of bus passenger flow, adopted the GM (1, 1) model, and verified the correctness of the method through experiments. To help passengers get information about the shortest route, Brata et al. [21] designed an efficient Android application that could guide users to the nearest bus station through the camera based on augmented reality. Suman and Bolia [22] proposed that passengers could be motivated to travel by bus through direct bus services and developed two mathematical formulas to optimize the bus network. Through the experiment on MATLAB, the method was applied to Delhi, India, and they found that the network replanned by the algorithm saved 8% of the travel time.

This study mainly analyzed the prediction of the arrival time of urban and rural buses. To further improve the accuracy of prediction, the bus arrival time was divided. In the current research, there are many different methods to divide the bus travel time, for example, divide it into the stop time and travel time, or divide it into deceleration arrival, vehicle parking, and acceleration departure. This study selected the former method and chose different prediction methods based on the characteristics of two periods. The experimental results demonstrated that the SVR + KNN method had a better performance in predicting the arrival time of bus on the same road segment compared to the historical data-based prediction method. The prediction error of the SVR + KNN method was significantly smaller than that of the historical data-based prediction method, and the average errors of the two methods were 181.1 and 42.5 s, respectively, showing a significant difference; the comparison of the relative error in Figure 3 also showed the same result. Finally, from the point of view of the stability of the algorithms, the prediction error of the SVR + KNN method was always controlled at about 50 s, which was superior to that of the historical data-based prediction method.

5 Conclusion

According to the characteristics of long routes and great prediction difficulty, the arrival time of the urban and rural bus was divided into two sections and predicted by the SVR + KNN method. The experiment was carried out on a section of the route of bus No. 16. The results showed that the prediction errors of the SVR + KNN method were smaller, within 100 s, compared to the historical data-based prediction method; the maximum error, average error, and average relative error of the SVR + KNN method were 84 s, 42.5 s, and 5.74%, respectively. Therefore, the SVR + KNN method can satisfy the demands of passengers and can be promoted and applied in practice.

Funding: No funding received.

Conflict of interest: Author states no conflict of interest.

References

- [1] Chen XF, Yao SM, Zhang LC. The theory and practice of urban-rural integration in china under the new urbanization. *Sci Geogr Sin.* 2016;36(2):188–95.
- [2] Ram S, Wang Y, Currim F, Dong F, Dantas E, Sabóia LA. SMARTBUS: a web application for smart urban mobility and transportation. *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*; 2016. p. 363–8.
- [3] Matyakubov M, Rustamova O. Development of smart city model: smart bus system model. *Acta Turin Polytech Univ Tashkent.* 2019;9(4):3.
- [4] Yang C, Ru XL, Hu B. Route temporal-spatial information based residual neural networks for bus arrival time prediction. *J Harbin Inst Tech (New Ser).* 2020;27(4):35–43.
- [5] Zhang H. Dynamic bus arrival forecast system based on time-varying parameter model. *J Mech Eng Res Dev.* 2017;40(1):164–73.
- [6] Sun DH, Luo H, Fu LP, Liu WN, Liao XY, Zhao M. Predicting bus arrival time on the basis of global positioning system. *Data, Transport Res Rec.* 2018;2034(2034):62–72.
- [7] Liu HJ, Xu HZ, Yan Y, Cai ZS, Sun TX, Li W. Bus arrival time prediction based on LSTM and spatial-temporal feature vector. *IEEE Access.* 2020;8:11917–29.
- [8] Awad M, Khanna R. Support vector machines for classification. Berkeley, CA: Efficient Learning Machines, Apress; 2015. p. 39–66.
- [9] Xu S, Zhao Y, Wang M, Shi X. Determination of rice root density from Vis–NIR spectroscopy by support vector machine regression and spectral variable selection techniques. *Catena.* 2017;157:12–23.
- [10] Ha D, Ahmed U, Pyun H, Lee CJ, Baek KH, Han CH. Multi-mode operation of principal component analysis with *k*-nearest neighbor algorithm to monitor compressors for liquefied natural gas mixed refrigerant processes. *Comput Chem Eng.* 2017;106(2):96–105.
- [11] [https://map.baidu.com/search/%E5%B9%B3%E9%A1%B6%E5%B1%B1%E5%B8%8216%E8%B7%AF%E5%85%AC%E4%BA%A4%E8%BD%A6/@12608708.523035714,3973129.8449999997,13.46z?querytype=s&da_src=shareurl&wd=%E5%B9%B3%E9%A1%B6%E5%B1%B1%E5%B8%8216%E8%B7%AF%E5%85%AC%E4%BA%A4%E8%BD%A6&c=213&src=0&wd2=%E5%B9%B3%E9%A1%B6%E5%B1%B1%E5%B8%82&pn=0&sug=1&l=13&b=\(12590130.665892858,3964584.0307142856;12627286.38017857,3981675.659285714\)&from=webmap&biz_forward=%7B%22scaler%22:1,%22styles%22:%22pl%22:7D&device_ratio=1](https://map.baidu.com/search/%E5%B9%B3%E9%A1%B6%E5%B1%B1%E5%B8%8216%E8%B7%AF%E5%85%AC%E4%BA%A4%E8%BD%A6/@12608708.523035714,3973129.8449999997,13.46z?querytype=s&da_src=shareurl&wd=%E5%B9%B3%E9%A1%B6%E5%B1%B1%E5%B8%8216%E8%B7%AF%E5%85%AC%E4%BA%A4%E8%BD%A6&c=213&src=0&wd2=%E5%B9%B3%E9%A1%B6%E5%B1%B1%E5%B8%82&pn=0&sug=1&l=13&b=(12590130.665892858,3964584.0307142856;12627286.38017857,3981675.659285714)&from=webmap&biz_forward=%7B%22scaler%22:1,%22styles%22:%22pl%22:7D&device_ratio=1)
- [12] Wang Y, Yan X, Zhou Y, Zhang WY. A feeder-bus dispatch planning model for emergency evacuation in urban rail transit corridors. *Plos One.* 2016;11(9):e0161644.
- [13] Kong X, Li M, Tang T, Tian K, Moreira-Matias L, Xia F. Shared subway shuttle bus route planning based on transport data analytics. *IEEE T Autom Sci Eng.* 2018;15:1507–20.
- [14] Cats O, Loutos G. Real-time bus arrival information system: an empirical evaluation. *J Intell Transport S.* 2016;20(2):138–51.
- [15] Derevitskiy I, Voloshin D, Mednikov L, Karbovskii V. Traffic estimation on full graph of transport network using GPS data of bus movements. *Proc Comput Sci.* 2016;101:207–16.
- [16] Liu Y, Liu C, Yuan NJ, Duan L, Fu Y, Xiong H, et al. Exploiting heterogeneous human mobility patterns for intelligent bus routing. *Knowl Inf Syst.* 2017;50:383–415.
- [17] Krishnan B, Reshmy AK. GPS based bus tracking system using Android app. *Int J Pharm Tech.* 2016;8(4):19839–45.
- [18] Moreira JM, Moreira-Matias L, Gama J, de Sousa JF. Validating the coverage of bus schedules: a machine learning approach. *Inform Sci.* 2015;293:299–313.
- [19] Wei M, Sun B. Bi-level programming model for multi-modal regional bus timetable and vehicle dispatch with stochastic travel time. *Cluster Comput.* 2017;20:401–11.
- [20] Zhang Z, Xu X, Wang Z. Application of grey prediction model to short-time passenger flow forecast. *AIP Conf Proc.* 2017;1839(1):1–6.
- [21] Brata KC, Liang D, Pramono SH. Location-based augmented reality information for bus route planning system. *Int J Electr Comput Eng.* 2015;5:142–9.
- [22] Suman H, Bolia N. Improvement in direct bus services through route planning. *Transp Policy.* 2019;81(3):263–74.