IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE (AI)-EMPOWERED INTELLIGENT TRANSPORTATION SYSTEMS

Received January 18, 2020, accepted February 10, 2020, date of publication February 27, 2020, date of current version March 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976574

A Bus Arrival Time Prediction Method Based on Position Calibration and LSTM

QINGWEN HAN[1,2], KE LIU[1], LINGQIU ZENG[3,4], GUANGYAN HE[3], LEI YE[1], AND FENGXI LI[1]

[1]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China
[2]Chongqing Key Laboratory of Space Information Network and Intelligent Information Fusion, Chongqing University, Chongqing 400044, China
[3]School of Computer Science, Chongqing University, Chongqing 400044, China
[4]Research and Development Centre of Transport Industry of Self-Driving Technology, Chongqing 400000, China

Corresponding author: Lingqiu Zeng (zenglq@cqu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 61601066, and in part by the Key Research and Development Projects of Chongqing Special Industries for Technological Innovation and Application Demonstration under Grant cstc2018jszx-cyzdX0064.

ABSTRACT Bus arrival time prediction not only provides convenience for passengers, but also helps to improve the efficiency of intelligent transportation system. Unfortunately, the low precision of bus-mounted GPS system, lack of real-time traffic information and poor performance of prediction model lead to low estimation accuracy - greatly influence bus service performance. Hence, in this paper, a GPS calibration method is put forward, while projection rules of specific road shapes are discussed. Moreover, two traffic factors, travel factor and dwelling factor, are defined to express real-time traffic state. Then, considering both historic data and real-time traffic condition, a hybrid dynamic BAT prediction factor, which achieves accuracy enhancement by taking into account traffic flow evaluation results and GPS position calibration, is defined. A LSTM training model is construct to realize BAT prediction. Experiment results demonstrate that our technique can provide a higher level of accuracy compared to methods based on traditional time-of-arrival techniques, especially in the accuracy of multi-stops BAT prediction.

INDEX TERMS Bus arrival time prediction, LSTM model, GPS data calibration.

I. INTRODUCTION

During the last decade, global warming has seriously harmed human survival environment and health security. The IEA (International Energy Agency) 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]. With low carbon consciousness and thorough popular feeling, more and more people are willing to use the public transportation, such as bus, subway. 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. Then, public transit companies devote to enhance the intelligence level of public transportation system and provide travel guide service Apps.Unfortunately, due to bad bus arrival time (BAT) prediction accuracy, these kinds Apps cannot satisfy passengers. Hence, how to improve BAT prediction, especially the accuracy of long-term BAT prediction, is the core of the work [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Chen.

The study found that there are many factors affecting BAT prediction accuracy. Although these factors are very complicated, from our perspectives, the main causes can be summed up three points: location error, real-time traffic feature and prediction error.

- Location error: Obviously, bus position error should have impact on BAT prediction results. Due to the cost restrain, location precision of bus on-board GPS location system is always low. Hence GPS data calibration is a valuable research issue that worth to explore.
- Real-time traffic feature: Most of BAT prediction methods use historic traffic data to estimate current traffic flow. However, the historical statistic dependence should cause slow response to real-time traffic condition - thus lead to prediction bias [4]. Then we believe that a bus data based real-time traffic evaluation approach is a topic worthy of in-depth study.
- Prediction error: Existing prediction algorithms are not as good as they should be, especially show in low multi-stops prediction accuracy. In our view, BAT prediction error may derive from unreasonable modelling.

Therefore, we focus on prediction model design, especially the model input and learning algorithm selection.

Hence, in this paper, our goal is to improve BAT prediction accuracy. Our study is developed in three aspect: (1) GPS calibration approach;(2) traffic factors definition, and (3) prediction model construction. Specifically, our research contributions are as follows:

- We separate the whole bus line into several road segments, namely link, and then propose a link-oriented projection rule to realize GPS calibration.
- A hybrid BAT evaluation method, which considers real-time traffic, historic trajectory, and bus stop dwell time [5], is proposed. Moreover, three factors, namely current driving segment (CDS) travel factor, stop-stop travel factor and bus stop dwelling factor, are defined as the weight factors of proposed weighted LSTM model. A new concept, CDS is defined to express real-time driving state.
- We propose a weighted LSTM models to realize multi-stops BAT prediction, while input and output of corresponding models are discussed.

The rest of this paper is organized as follows. Section II presents the related work, while the preprocessing of the raw dataset is given in Section III. In Section IV, a GPS calibration algorithm of GPS data is proposed. Then a prediction model of bus arrival time based on LTSM is proposed in Section V, while experimental results are presented and discussed in Section VI. At last, conclusions are given in Section VII.

II. RELATED WORK

BAT prediction approach is considered as an essential topic for smart city applications. In this section, we briefly review the related work, which can be grouped into two categories, i.e., map matching and prediction method.

A. MAP MATCHING

The performance of map matching approach could be denoted by four indicators, which are accuracy, integrity, continuity and availability. In past few years, to improve performance of all these four indicators, researchers focus on two topics, the one is road network matching, while the other is cost function definition.

In 2002, Li presented a GPS data calibration method [6], initiate a new line of bus arrival time (BAT) estimation. On the other hand, researchers work on map-matching approach in vehicle location and navigation system [7] and believe that a GIS (Geographic Information system/Geo-information system) based GPS data calibration method is useful in raw dataset pre-processing to get bus exact positions.

In 2012, P. S. Castro proposed a road network oriented map matching method [7], which laid the foundation to further study. The correlation feature of road network and vehicle cruising trajectory is discussed, while matching probability used as calibration criterion. To improve matching accuracy, [8] employed a trajectory segment method, which divides driving trajectory into a series road segments, then defined corresponding topological constraints, and proposed a spatial map matching algorithm. However, driving trajectory of social vehicle is un-certain. Road segment processing should lead to a high computation burden and further more influence the response speed. Fortunately, city buses travel on a regular route; which means their route task is solid and road segment scheme is relatively stable. In our previous work [9], we divide bus route into several links, and discuss the impact of dividing accuracy on map matching and BAT prediction. Hence, in this paper, we use road segment processing not only as the first step of map matching, but also the foundation of BAT prediction.

It is no doubt that GIS oriented method is the foundation of map matching. Rome for improvement remains. Researchers believe that map matching performance is influenced by a series factors, such as heading information [10], time information [11],driving speed [12], distant, historic data, azimuth of GPS [13], etc. Then different types of cost functions are defined. Literature [14] defined a weighting function, which considers two factors: a) the distance between GPS points and network arcs, and b) the area formed between GPS trajectory and path. In [15] the authors selected three factors, which are distance between the location point and the candidate road, the angle between GPS direction and road direction, and average distance, to construct cost function, while hierarchical fuzzy reasoning method is used to fulfill map matching. However, BAT prediction has critical real-time requirement, then complex methods, such as fuzzy reasoning, are not suitable to used here. In fact, to improve the real-time feature of map matching, driving trend, which should be derived from cruising direction, speed and previous instant position, is suitable for BAT prediction. Hence, here we use a projection strategy, which considers both driving speed and driving trend, to improve the performance of map matching.

B. BAT PREDICTION

The core of BAT prediction is the model construction, which includes two key issues; the one is impact factors selection, while the other is prediction model definition.

Obviously, BAT prediction closely related to bus stop features. Hence, bus stop related factors, such as dwell time, passenger state, line number, etc. are considered as the basic parameters for BAT prediction. In [16], Amita *et al.* selected bus stop dwell time and passenger feature to establish a linear regression model to predict BAT. However, the simple linear relationship could not denote the complexity of real driving environment. Considering the complexity of driving environment, literature [17] proposed heterogeneous impact factor, which considers bus stop location, bus arrival time and departure time, line number of bus stop. It is true that all these factors could be used to express bus real driving state of bus stop, the authors did not consider the impact of real time traffic flow. Based on all above reasons, performance of bus stop oriented prediction method are far from satisfactory.

Literature [18] used traffic flow factor to construct prediction model to predict ETA (Expected Arrival Time) of social vehicle. Then we believe that an impact factor, which considers both bus stop features and traffic flow feature, could reasonable predict BAT.

And now, let us back to prediction models. In past decades,, researchers work on three types of BAT prediction models: (a)models based on historical data, (b)multilinear regression models, and (c)artificial neural network models [19]. Gong *et al.* [20] proposed a historical data oriented approach, in which historical statistic results is employed as an input of prediction model. It is no doubt that historical data is helpful in bus arrival time prediction. However, the historical statistic dependence should cause slow response to real-time traffic condition - thus lead to prediction bias. In [21], a set of regression models are used to estimate bus dwell times with data collected by automatic passenger counters installed on buses. However, the only drawback of this method is that it is difficult to establish a model including multiple variables. In [22], the authors employ SVM (support vector machine) model as prediction model. History bus data is used to realize SVM model training,while real-time features, such as bus real-time data, weather, time and data, are set as input of prediction model. Literature [23] proposed a hybrid model, which combines Fuzzy Logic and Neural Networks, and is effective in stated conditions. However, both two models are only applied for short term prediction. In [24], a DNN (deep neural network) model is used to realize BAT prediction and get high prediction accuracy. In 2019, Pang *et al.* [17] exploit the long-range dependencies among the multiple time steps for bus arrival prediction via recurrent neural network (RNN). Concretely, RNN with long short-term memory block is used to correct the prediction for a station by the correlated multiple passed stations. Compared to traditional neural networks, LSTM is a special RNN that remembers longer-term information. Hence, we believe that LSTM algorithm is suitable for dealing with problems related to time series. To enhance the prediction accuracy, [25] present a road segment issue and predict bus position according to path-section graph. Moreover, [26] present a revised road segment method, which considers both road section and bus stop. Time trend features of driving records and the spatial location information of bus stops are selected as input of LSTM model. Reference [27] set the passenger flow information and weather and other common features of each segment as LSTM input factors. Then, in this paper, selected bus record factors, weather factors, and location information are set as the inputs of LSTM networks, while Current driving segment (CDS) travel time, stop-stop travel time and bus stop dwell time are set as output of corresponding LSTM network. A hybrid BAT factor, which consider all three outputs of LSTM and traffic weight, is used to calculate BAT.

Then, in this paper, taking into account the characteristics of bus cruising and other influencing factors, a LSTM based method is proposed to realize multi-stops BAT.

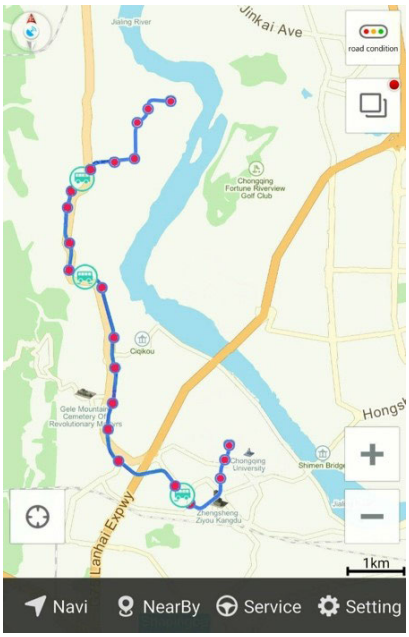


FIGURE 1. Digital map of Line 805.

III. PRELIMINARIES

A. RAW DATASET

In this paper, two datasets are employed in BAT prediction process. Bus dataset, which is provided by Chongqing Hengtong Bus Company, are used as main dataset, while weather data set are used as supplementary data set.

1) BUS RAW DATASET

Bus raw dataset contains driving records of 3000 buses in Chongqing, China. Each record is observed every two seconds and consists of 46 attributes, which are divided into three categories:

- Driving state: bus_ID, velocity, brake state, timestamp etc.
- Mechanical properties: battery state, gas pressure, motor speed, etc.
- GPS position: longitude, latitude, etc.

In this paper, 6 attributes - bus_ID, brake state, velocity, timestamp, longitude, latitude - are selected as the basic data to calibrate the bus route and predicted the BAT, while 1 year driving records of 30 buses on 805-bus line is employed as historic database for model training and testing. The six dataset attributes are used to construct bus raw dataset D, which is denoted as follows,

$$D = data[h][6] = \begin{bmatrix} Id_1 & V_1 & Br_1 & Lo_1 & La_1 & Tim_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Id_i & V_i & Br_i & Lo_i & La_i & Tim_i \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Id_h & V_h & Br_h & Lo_h & La_h & Tim_h \end{bmatrix}, \quad i = 1, 2, \dots, h \tag{1}$$

The detail digital map of 805-bus line is shown in FIGURE 1.

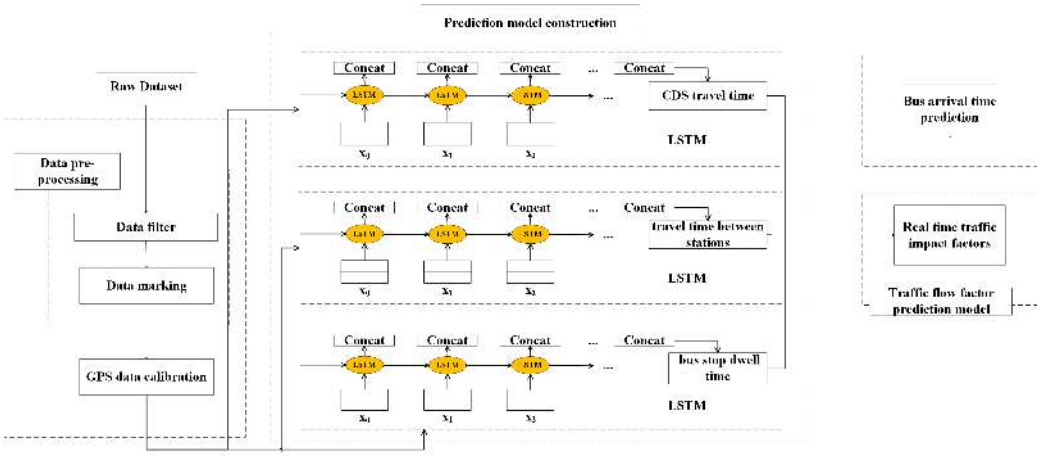


FIGURE 2. Basic processing procedure.

2) WEATHER DATASET

Weather dataset is provided by social service department of Chongqing, China [28]. Weather features, such as temperature, air quality, are selected as input variable of LSTM network.

B. BASIC PROCESSING PROCEDURE

BAT prediction process includes three basic steps, which are data pre-processing, GPS data calibration and prediction model construction. Corresponding procedure is shown in FIGURE 2.

1) DATA PRE-PROCESSING

Data pre-processing includes two aspects, which are data filter and data marking.

Data filter aims to remove invalid data, such as duplicate data, error data, and long-term parking data.

Here we consider three kinds of invalid data.

- Long-term parking data. Due to the data collection system setting, GPS position of parking bus is set as (0, 0), and useless in BAT prediction process.
- GPS lost data. When the signal of GPS is lost, the data collection system should set bus speed as 0. These data points should be filtered either.
- Error data. Due to GPS error, raw dataset includes some unreasonable GPS points and large error points, which are defined as error location information, and should be filtered. Note here, there are two kinds of error location information. The unreasonable GPS points should be judged according to time relationship of two points, while the large error points should be discovered based on a defined error tolerance range. The process of error location information discovery is shown in FIGURE 3.

As shown in FIGURE 3, point G0, whose GPS position error out of error tolerance range, is considered as a error data point. On the other hand, position of previous time P1 is ahead of current time position P2. That is said, point G2, whose GPS position error in tolerable error range, is a time relation error

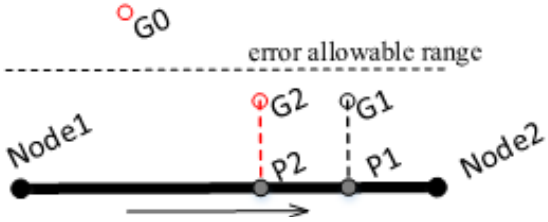


FIGURE 3. Data error point.

point. Data marking process adds corresponding time feature to raw dataset D. Here two kinds of features are considered. The one is day feature, which has two attributes; weekday and weekend. The other is time feature, which also has two attributes; peak time and off-peak time.

2) GPS DATA CALIBRATION

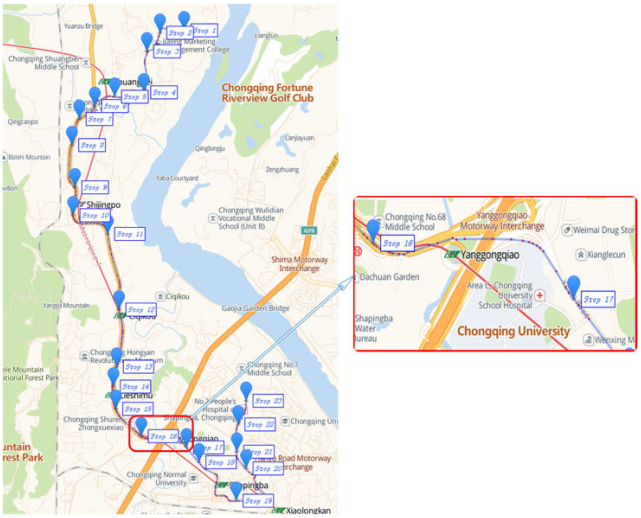
To improve the accuracy of BAT prediction, GPS data calibration process is done to calibrate bus position information.

3) PREDICTION MODEL CONSTRUCTION

The driving environment of bus is very complicated. Multi source information, such as road condition, weather condition, and traffic flow feature, etc. Therefore, the modeling process becomes more difficult and complicated.

Fortunately, some particular characteristics of bus make it possible for model simplification. City buses travel on a regular route; which means their route task is solid and the impact factor of driving environment is relatively stable during same time period. Moreover, the bus model of specific line is the same, and the modeling differences between buses could be ignored.

In this paper, three LSTM models are established to predict the CDS travel time, travel time between stops and bus stop dwell time. Then the three prediction times are weighted to the real-time traffic flow influence factors. Finally, add the weighted three prediction time to get the bus arrival time.



(a) Bus line (b)Road link

FIGURE 4. 805 bus line and road link.

IV. GPS CALIBRATION

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, while all subsequent processing is done based Amap.

A. ROAD SEGMENTATION

Definition 1: A path $P = (N, L)$ consists of a node set N and a link set L. Each element n in N is associated with a pair of position coordinates (x, y), which represents the spatial location of the node in object path. Each element l_k in L represent a link between node n_k and node n_{k-1}.

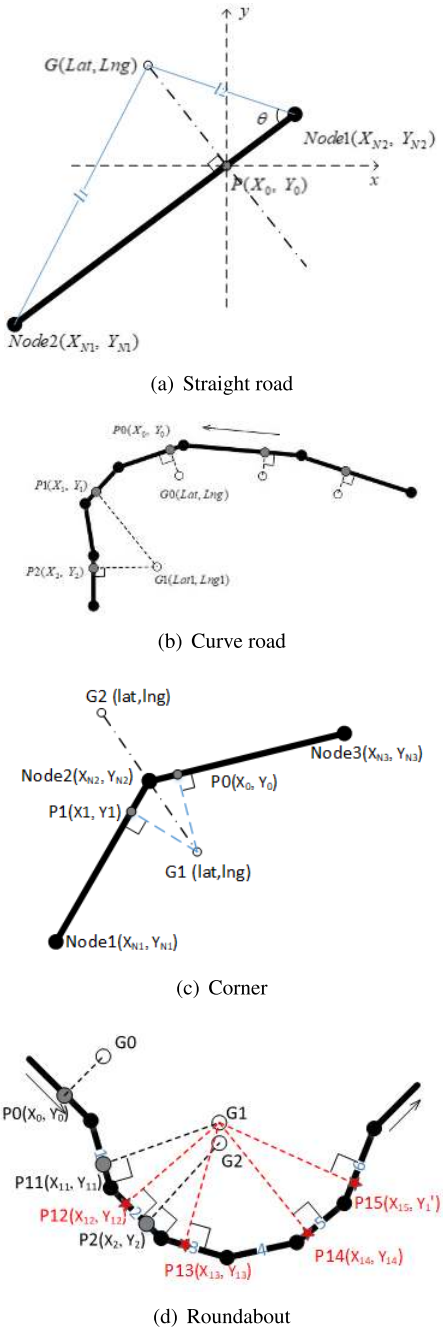
Then the whole bus path is divided into a series road segments. The length of link is set according to prediction accuracy requirement. The longer the road segment is, the lower the prediction accuracy. Here we set road link length as 20m. Corresponding equivalent routes are shown in the FIGURE 4.

B. PROJECTION RULES

In this paper, we employ a link-oriented projection rule. Here three kinds of road shapes, which are straight road, curve road and roundabout, are considered. Corresponding projection process are shown in FIGURE 6, where $G(Lat, lng)$ is the position coordinates provided by on-board GPS system, $P(X, Y)$ is G's projection point on corresponding road link.

1) STRAIGHT ROAD

As shown in FIGURE 5(a), the link between Node1 and Node2 is a straight road. Then a simple vertical projection process is employed to project GPS position G to point P. The projection point P is unique.



(a) Straight road

(b) Curve road

(c) Corner

(d) Roundabout

FIGURE 5. Projection rules.

2) CURVE ROAD

For the curve road, the possible projection points are not unique. As shown in FIGURE 5(b), GPS position $G1$ has two possible projection points - $P1$ and $P2$. Hence an additional processing is needed. Here a bus travel distance factor d, which indicates distance between previous projection point $n-1$ and current projection point n, is defined. We can calculate the maximum driving distance of object bus as

$$d_{max} = \alpha \cdot \hat{v}_n \cdot t \tag{2}$$

where α is current traffic factor, which will be described in detail later. t is the time interval between point n and $n-1$.

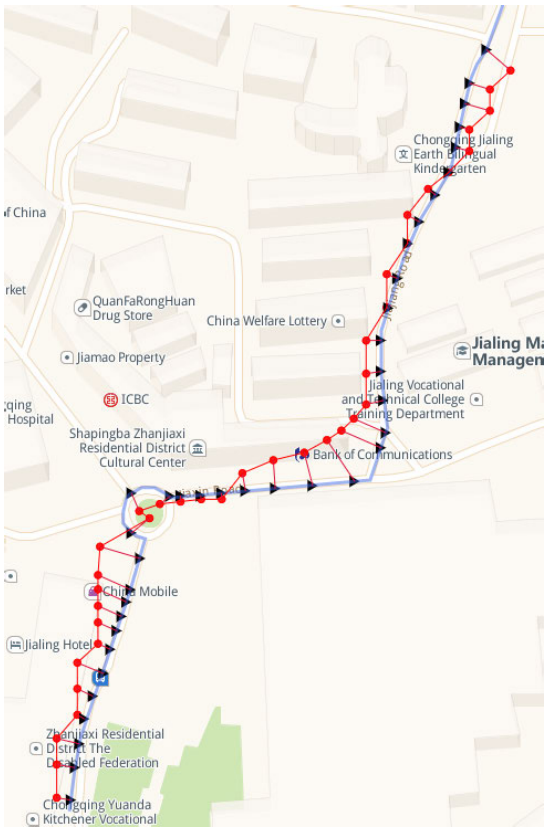


FIGURE 6. GPS data calibration results.

$\hat{v}_n$ is calculated as

$$\hat{v}_n = \frac{v_{n-1} + \bar{v}}{2} \tag{3}$$

where v_{n-1} is the bus speed of point n, while $\bar{v}$ is historic average speed of corresponding road segment.

Then a two steps procedure is employed here to select fit point from all projection points. Step 1:

1) Eliminating points that don't meet Equ.(2);
2) Comparing $Dis(G, P)$ of all remaining points, select the point with the shortest $Dis(G, P)$. $Dis(G, P)$ is the distance of point G and point P.

In addition, a special case-corner point, which is shown in FIGURE 5(c), must be considered here. Obviously, projection points of $G1$ - $P0$ and $P1$-are not fit point. Moreover, $G2$ could not obtain projection point according to Equ.(2) and vertical projection process. In this case, both $G1$ and $G2$ should be projected to the corner point.

3) ROUNDABOUT

As shown in FIGURE 5(d), two GPS points, $G1$ and $G2$, are close to each other, while $G1$ nears to the center of roundabout and have several project points, such as $P11(X_{11}, Y_{11})$, $P12(X_{12}, Y_{12})$, $P13(X_{13}, Y_{13})$, etc…Here the time order of GPS data and Equ.(2) is the useful information to project GPS points to fit position. The projection point of previous GPS point $G0$, $P0(X_0, Y_0)$, is selected as a reference point. Timestamp indicates that sample time of $G0$ is prior to that of $G1$. Hence we can infer that projection point of $G1$ is placed on link1 and illustrated as $P11(X_{11}, Y_{11})$.. Then the next GPS point $G2$ is projected to link2 and illustrated as $P2(X_2, Y_2)$.

C. CALIBRATION PROCESSING

According to the projection rules defined earlier, GPS calibration process is proposed here. Corresponding algorithm is given in Algorithm 1, while the calibration performance is shown in FIGURE 6.

Algorithm 1 GPS Data Calibration Algorithm

Input: Peak time dataset and Off-peak time dataset
Output: Filtered Dataset

```
set FilteredData={}
for all data in Peak time dataset and Off-peak time dataset do
    /* filter wrong data */
    if isValidGPS(data) == TRUE and inStudyRegion(data) == TRUE then
        /* filter out data with speed 0 */
        if data.speed!=0 then
            if isInRoute()==false then
                matchRunningRoute()
                obtainFitPoint()
            end if
            if isNoisePoint( )==false then
                if satisfyConstraint( )==false then
                    adjustFitPoint()
                end if
                addFilteredData(data)
            end if
        end if
    end if
end for
```

isValidGPS (data): Decide whether the GPS data is in the a valid GPS data
inStudyRegion(data): Decide whether the GPS data is in the study region
isInRoute( ): Determine whether the current point is on the Line
matchRunningRoute( ): Match the running route of the GPS point
obtainFitPoint( ): Obtain the fitting point on the current road segment
isNoisePoint( ): Whether it is a noise point
satisfyConstraint( ): Whether to meet the movement trend calibration
adjustFitPoint(): Adjust the fitting point
addFilteredData(data): Add data into FilteredData

As shown in FIGURE 6, all GPS data points are calibrated to the bus line.

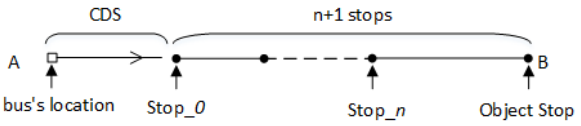


FIGURE 7. Bus route segments.

After adding the weather features from the weather dataset to the calibrated and preprocessed dataset, we get final dataset D', which have 14 features: bus_ID, brake state, velocity, timestamp, longitude, latitude, time, day, Maximum temperature, lowest temperature, AirQuality, Next stop, Current stop, Distance to next stop. Then the calibrated dataset D' is denoted as Equ.(4), shown at the bottom of this page.

V. BAT PREDICTION MODEL

This is the central part of this paper. As mentioned earlier, BAT prediction progress should address two problems. The one is impact factors selection, while the other is model construction.

A. IMPACT FACTORS SELECTION

In this part, we divide the bus route into a series segments based stops, as shown in FIGURE 7.

The bus driving distance between bus current position and object stop could be expressed as,

$$Dis(A, B) = Dis(A, Stop\_0) + \sum_{i=0}^{n} Dis(Stop\_i, Stop\_i + 1) \tag{5}$$

where shown $Dis(A, Stop\_0)$ is the distance between bus current position and stop 0, that is the length of CDS. $Dis(Stop\_i, Stop\_i+1)$ is the distance between bus stop i and $i+1$.

Then BAT of bus stop n could be considered as combination of three time periods, which are travel time between current position to next bus stop, bus stop dwelling time, and travel time between bus stop i and $i+1$. Hence, we define two factors - travel factor and dwelling factor, to denote BAT.

1) TRAVEL FACTOR

Definition 2: Current driving segment (CDS): the road segment between bus current location and next bus stop. It is well known that road traffic presents a time-vary characteristic. Then the travel time of CDS also presents time-vary feature, whose value varies around the mean value of historic travel time on object CDS.

Here we assume current time as t_0, and there are r bus traveled on CDS during $[t_0 - 30\ min, t_0]$, then corresponding travel factor is defined as,

$$\alpha_c = 1 + \frac{1}{r} \cdot \sum_{q=1}^{r} \frac{T'_{cq} - \bar{T}_c}{\bar{T}_c} \tag{6}$$

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

Likewise, we can obtain travel factor of all road segment between bus stops, for example, segment between $Stop\_k$ and $Stop\_k + 1$, and denote as $\alpha_\gamma(k, k+1)$.

2) DWELLING FACTOR

Bus stop dwelling time is influenced by a series factors, such as overcrowding feature, bus stop feature, passenger number, etc. However, there is no need to construct an impact factor based all these factors. Here we only consider historic dwelling time and current dwelling time.

Here we assume current time as t_0, and there are s bus dwell at object bus stop during $[t_0 - 30\ min, t_0]$, then corresponding dwelling factor is defined as,

$$\alpha_s = 1 + \frac{1}{s} \cdot \sum_{p=1}^{s} \frac{T'_{sp} - \bar{T}_s}{\bar{T}_s} \tag{7}$$

where $\bar{T}_s$ is the historic dwelling time and T'_{sp} is current dwelling time of bus p, $p \in s$.

3) BAT EVALUATION

Considering both travel factor and dwelling factor, BAT could be calculated as,

$$T = \alpha_c \bar{T}_c + \sum_{k=0}^{p} [\alpha_s(k)\bar{T}_s(k) + \alpha_\gamma(k, k+1)\bar{T}_\gamma(k, k+1)] \tag{8}$$

where $\bar{T}_c$ 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 kth stop. $\bar{T}_\gamma(k, k+1)$ is the average travel time between $Stop_k$ and $Stop_k + 1$. α_c, $\alpha_s(k)$, $\alpha_\gamma(k, k+1)$ and

$$D' = data'[l][14] = \begin{bmatrix} Id_1 & V_1 & Br_1 & Lo'_1 & La'_1 & Tim'_1 & Time_1 & Day_1 & MaxT_1 & MinT_1 & AirQ_1 & NextS_1 & CurrS_1 & Dis_1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ Id_i & V_i & Br_i & Lo'_i & La'_i & Tim'_i & Time_i & Day_i & MaxT_i & MinT_i & AirQ_i & NextS_i & CurrS_i & Dis_i \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ Id_l & V_l & Br_l & Lo'_l & La'_l & Tim'_l & Time_l & Day_l & MaxT_l & MinT_l & AirQ_l & NextS_l & CurrS_l & Dis_l \end{bmatrix}, \quad i = 1, 2, \dots, l \tag{4}$$

TABLE 1. Training set, validation set and testing set.

	CDS	Bus stop	Stop-stop
Total number of data	2100000	1320000	570000
Number of training data	1260000	792000	342000
Number of validation data	420000	264000	114000
Number of testing data	420000	264000	114000

are the CDS travel factor, stop-stop travel factor and bus stop dwelling factor.

B. MODEL CONSTRUCTION

1) DATA SET CLASSIFICATION

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 four million, while the data collection period is April 2016 to December 2016. The original data of the experiment is the driving data of 805 bus Stop2 to Stop19. The arrival time prediction model mainly consists of three factors, which are CDS travel time, bus stop dwell time and stop-stop travel time. Therefore, as shown in FIGURE 2, we build three LSTM models to predict these three factors separately. Hence, firstly, we should classify the calibrated dataset D' into three sub-datasets, which are employed in corresponding three LSTM model training, validation and testing. Then, input index of all three models should be explained subsequently.

According to calibrated GPS position information, dataset D' is divided into three sub-datasets, which are bus stop sub-dataset D_1'', stop-stop sub-dataset D_2'', and CDS sub-dataset D_3''. Corresponding classification rule is listed as follows,

Bus stop sub-dataset D_1'': The area of 10 meters radius from bus stops is considered as bus stop area. The bus records, whose calibrated GPS coordination belong to bus stop areas, fall into D_1''.

Stop-stop sub-dataset D_2'': The area between the adjacent boundary of two adjacent bus stop is defined as stop-stop area. The bus records, whose calibrated GPS coordination belong to bus stop areas, fall into D_2''. Note here, stop-stop area does not include the road segment between bus current position to next bus stop.

CDS sub-dataset D_3'': The bus records, whose calibrated GPS coordination belong to CDS areas, fall into D_3''.

Then, the three sub-datasets, D_1'', D_2'', and D_3'' are divided into their respective training set, validation set and testing set. Detail of dataset classification results is shown in TABLE 1.

As mentioned earlier, here we used three LSTM networks. Here we divide model input indicators into two categories; the one is common indicators, while the other is special indicators.

Common indicators include bus_ID, brake state, velocity, timestamp, longitude, latitude, time, day, the highest temperature, the lowest temperature, and air quality, which should be used as input indicator of all above three models.

TABLE 2. Parameter settings of prediction model.

	CDS	Bus stop	Stop-stop
Input size	13	12	12
Output size	1	18	17
Cell_size	52	52	52
Batch_size	80	80	80
Time step	10	5	5
LR	0.006	0.006	0.006
Activation function	Tanh	Tanh	Tanh
Optimizer	Adam	Adam	Adam
Training epochs	50	50	50

Special indicators vary from network to network. CDS oriented network includes two special indicators, which are named as ''next stop'' and ''distance to next stop''. The special indicator of bus stop oriented network is ''current stop'', while that of stop-stop oriented network is ''next stop''.

2) MODEL CONSTRUCTION AND TRAINING

The basic network structure of all three LSTM models is same: the input matrix is fed to the LSTM layer with 52 neurons (cell_size = 52), and then the output of LSTM layer enters the dense layer, which forecasts based on the characteristic information at the output of LSTM layer and changes the output size. We choose the mean square error as the loss function and Tanh as the activation function. After a lot of experiments, it is found that using 0.006 as the learning rate can improve the performance of the model (LR=0.006). In reference [27], compared with other optimization algorithms, Adam algorithm has better accuracy characteristics, and this paper also chooses this algorithm. In our experiment design, the tensor board is used to visualize the complex training process of neural network.

The three models have different inputs and outputs.

For CDS oriented network, each training instance is a continuous 10 records, that is to say, it contains the information of the bus in the previous 20 seconds. Each record contains 13 features. The input layer dimension of the network is set to 13 (input size = 13) and the dependency length of the network is set to 10 (time step = 10). The output is the time to the next stop, so the output layer dimension is set to 1 (output size = 1).

For bus stop oriented network and stop-stop oriented network, each training instance is a continuous 5 records. Each record contains 12 features. The input layer dimension of these network is set to 12 (input size = 12), and the dependency length of the model is set to 5 (time step = 5). The output of stop oriented network is the bus stop dwell time of all stops, the output layer dimension is set to 18 (output size = 18). While the output of stop-stop oriented network is the

travel time between stop-stop, so the output layer dimension is set to 17 (output size = 17).

Detail of corresponding parameter setting is listed in TABLE 2.

3) PREDICTION EFFECT EVALUATION

In this paper, Mean Square Error (MSE), Root Mean Square Error (RMSE),Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the quality of the LSTM prediction. Their specific definitions are as follows,

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} [T(i) - \bar{T}(i)]^2 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} [T(i) - \bar{T}(i)]^2} \tag{10}$$

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |T(i) - \bar{T}(i)| \tag{11}$$

$$MAPE = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{|T(i) - \bar{T}(i)|}{T(i)} \tag{12}$$

Here $T(i)$ and $\bar{T}(i)$ represent the actual arriving time and the predicted arriving time, respectively. n is the bus stop number, while i is the serial number of bus stop.

VI. EXPERIMENT RESULTS ANALYSIS

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.

Experimental results analysis focus on two main points. The one is the performance of model training, while the other is efficient of comprehensive BAT prediction.

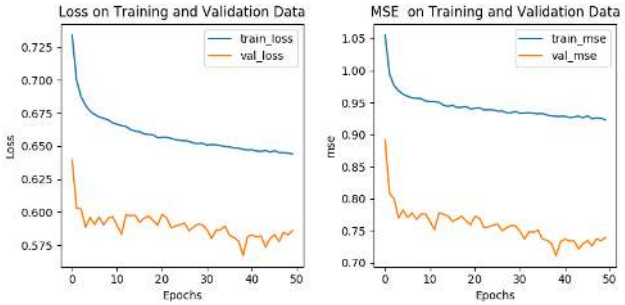
A. MODEL TRAINING

As mentioned earlier, here we employ three LSTM network to predict three indicators, which are CDS travel time, stop-stop travel time and bus stop dwelling time. Here all LSTM network should be trained solely.

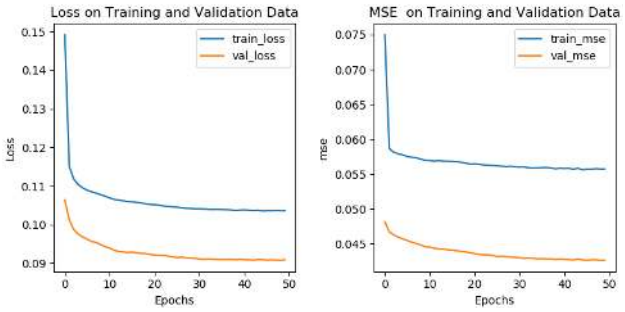
All training process is done based on TensorFlow 1.14 platform with i7-8700K 3.70 GHZ CPU and 32 GB memory.

Training performance are shown in FIGURE 8, while corresponding parameters are listed in TABLE 3. Here we employ MAE as loss function, while MSE as evaluation function. In FIGURE 8, label "train" corresponds to training dataset, and label "val" corresponds to validation dataset.

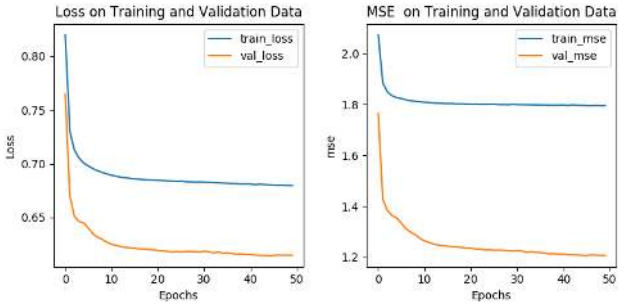
As shown in FIGURE 8, all three models present good convergence behavior. As shown in TABLE 4, MAPE value of stop-stop network and bus stop network is about 0.04, while that of CDS network is less than 0.03. As for MAE value, the highest of the three, that of stop-stop network, 0.61768min, is acceptable in our opinion.



(a) CDS



(b) Bus stop



(c) Stop-stop

FIGURE 8. Loss functions and evaluation function.

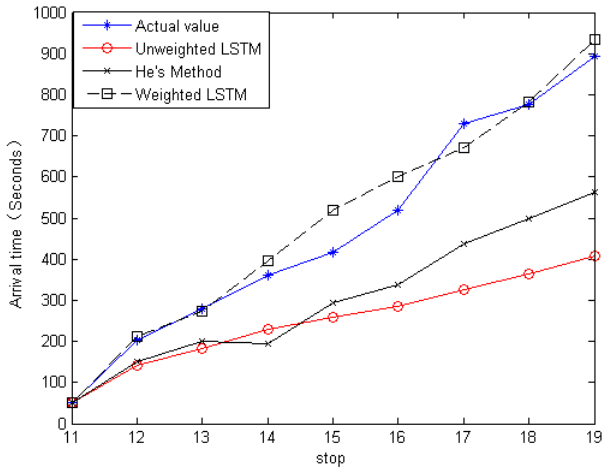
TABLE 3. MAE and RMSE of each model.

	CDS	Bus stop	Stop-stop
MAE(Minute)	0.56	0.10	0.62
RMSE	0.84	0.21	1.15
MAPE(%)	2.85	4.37	4.09

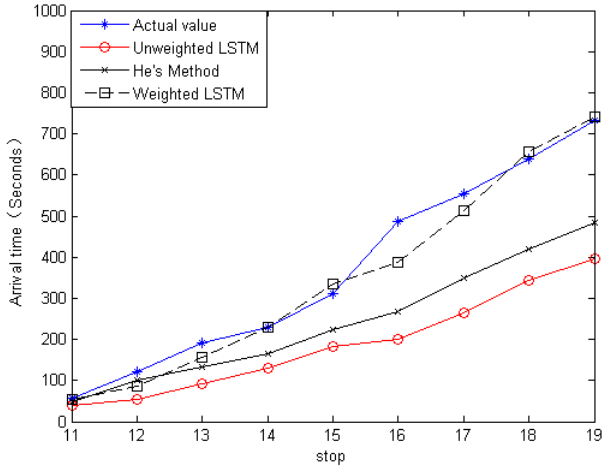
B. BASELINE METHODS

Based on the prediction value of all above three indicators, we should calculate BAT prediction value according to Equ. (8). To verify to validation of proposed method, the following four kinds of journey travel time prediction approaches are used as contrast.

- Historical Average (HA):The historical average method is to use the historical journey records as the basic data to



(a) Peak time



(b) Off-peak time

FIGURE 9. Prediction performance.

obtain the arrival time of current vehicles by calculating the average value.

- Support Vector Regression (SVR) [22]: Support Vector Machines (SVM) is useful to solve classical binary classification problem, while SVR is employed in literature [22] to predict BAT.
- Unweighted LSTM: For unweighted method, BAT value is calculated from a simple summation of above three LSTM outputs, as shown in Equ.(13).
- He's Method [26]: He's method is a historical data-oriented approach, which uses LSTM model to predict bus travel time.

$$T = \bar{T}_c + \sum_{k=0}^{p} [\bar{T}_s(k) + \bar{T}_\gamma(k, k+1)] \tag{13}$$

C. OVERALL PERFORMANCE

Overall performance of all above four kinds approach and proposed method are listed in table 4. As shown in Table 4, both MAE and MAPE performance of proposed method are better than that of other four. Moreover, MAPR values of HA and SVR are higher than 10%. Hence, the following comparison experiments just consider about three kinds approach, unweighted LSTM, He's Method and proposed weighted LSTM

TABLE 4. Comparison of overall performance.

	MAE(Minute)	MAPE(%)
HA	6.98	12.85
SVR	5.10	10.15
Unweighted LSTM	4.73	8.74
He's Method	3.49	6.46
Weighted LSTM	1.62	4.89

TABLE 5. BAT prediction error.

	Unweighted LSTM		Weighted LSTM		He's Method	
	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)
Peak	10.98	12.15	4.38	6.17	7.16	8.68
Off-peak	7.15	8.55	2.36	4.48	6.45	5.38

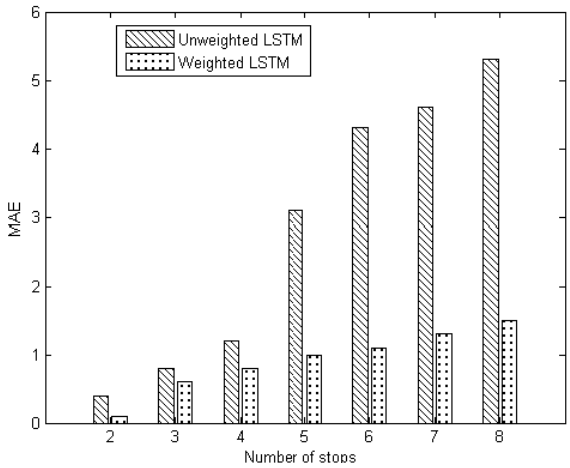


FIGURE 10. BAT prediction results with multiple bus stops.

As shown in FIGURE 9, the prediction arrival times of proposed method are in good agreement with actual values. Corresponding parameters are listed in TABLE 5.

As shown in Table 5, the MAPE value of peak time is about 0.06, while that of off-peak time is less than 0.05. Meanwhile, RMSE value of peak time is higher than that of off-peak time. Performance of multiple stops prediction is shown in FIGURE 10. Comparing with unweighted issue, prediction error of proposed method is relatively low. For example, BAT prediction value of 8th stop is about 1 minute and 40 seconds away from the actual value, which is considered as an acceptable waiting time for most of passengers.

VII. CONCLUSION

In this paper, a BAT prediction method is proposed. Firstly, to ensure prediction accuracy, a GPS calibration strategy is proposed to calibrate bus GPS points to bus line. Then, to demonstrate BAT, several novel definitions are presented, and travel factor and dwelling factors are used to measure real-time traffic state. At last, taking into account the characteristics of bus cruising and other influencing factors, three LSTM models are constructed to predict CDS travel time, stop-stop travel time and bus stop dwell time are predicted separately. A traffic weighted based BAT factor is defined to calculate multi-stop BAT values. Experiments results show that proposed method perform good in both peak time and off-peak time multi-stops BAT prediction. In our future work, other influence factors, such as rain, snow, road construction, and group characteristic of pedestrian, should be considered in model construction. Moreover, corresponding bus transfer times, transfer point choice and the travel cost should be taken into account.

REFERENCES

[1] C. Rolim, P. Baptista, G. Duarte, T. Farias, and J. Pereira, "Real-time feedback impacts on eco-driving behavior and influential variables in fuel consumption in a lisbon urban bus operator," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3061–3071, Nov. 2017.

[2] L. Moreira-Matias, J. Mendes-Moreira, J. F. de Sousa, and J. Gama, "Improving mass transit operations by using AVL-based systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1636–1653, Aug. 2015.

[3] D. W. Seng, J. W. Peng, J. Chen, and N. Zheng, "Bus arrival time prediction based on static and dynamic algorithms," *Adv. Transp. Stud.*, vol. 1, pp. 47–54, Nov. 2015.

[4] C. Chen, Y. Ding, Z. Wang, J. Zhao, B. Guo, and D. Zhang, "VTracer: When online vehicle trajectory compression meets mobile edge computing," *IEEE Syst. J.*, to be published.

[5] L. Moreira-Matias, O. Cats, J. Gama, J. Mendes-Moreira, and J. F. de Sousa, "An online learning approach to eliminate bus bunching in real-time," *Appl. Soft Comput.*, vol. 47, pp. 460–482, Oct. 2016.

[6] L. Weigang, W. Koendjbiharie, R. C. de M Juca, Y. Yamashita, and A. Maciver, "Algorithms for estimating bus arrival times using GPS data," in *Proc. IEEE 5th Int. Conf. Intell. Transp. Syst.*, Sep. 2002, pp. 868–873.

[7] P. S. Castro, D. Zhang, and S. Li, "Urban traffic modelling and prediction using large scale taxi GPS traces," in *Proc. Int. Conf. Pervasive Comput.*, 2012, pp. 57–72.

[8] M. Liu and M. Li, "Research on floating car map matching algorithm," in *Proc. 25th Int. Conf. Geoinformatics*, Aug. 2017, pp. 1–5.

[9] L. Jianmei, C. Dongmei, L. FengXi, H. Qingwen, C. Siru, Z. Lingqiu, and C. Min, "A bus arrival time prediction method based on GPS position and real-time traffic flow," in *Proc. IEEE 15th Intl Conf Dependable, Autonomic Secure Comput., 15th Intl Conf Pervas. Intell. Comput., 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congress(DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 178–184.

[10] C. Chen, Y. Ding, X. Xie, S. Zhang, Z. Wang, and L. Feng, "TrajCompressor: An online Map-matching-based trajectory compression framework leveraging vehicle heading direction and change," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[11] W. Teng and Y. Wang, "Real-time map matching: A new algorithm integrating spatio-temporal proximity and improved weighted circle," *Open Geosci.*, vol. 11, no. 1, pp. 288–297, Jul. 2019.

[12] Y. Zhang, C. Guo, R. Niu, and Y. Mao, "A multi-weights map matching algorithm used in GPS system for vehicle navigation application," in *Proc. 12th Int. Conf. Signal Process. (ICSP)*, Oct. 2014, pp. 2375–2378.

[13] R. Kruger, G. Simeonov, F. Beck, and T. Ertl, "Visual interactive map matching," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 6, pp. 1881–1892, Jun. 2018.

[14] S. Romon, X. Bressaud, S. Lassarre, G. Saint Pierre, and L. Khoudour, "Map-matching algorithm for large databases," *J. Navigat.*, vol. 68, no. 5, pp. 971–988, Mar. 2015.

[15] J. Tang, S. Zhang, Y. Zou, and F. Liu, "An adaptive map-matching algorithm based on hierarchical fuzzy system from vehicular GPS data," *PLoS ONE*, vol. 12, no. 12, Dec. 2017, Art. no. e0188796.

[16] J. Amita, S. S. Jain, and P. K. Garg, "Prediction of bus travel time using ANN: A case study in delhi," *Transp. Res. Procedia*, vol. 17, pp. 263–272, 2016.

[17] J. Pang, J. Huang, Y. Du, H. Yu, Q. Huang, and B. Yin, "Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3283–3293, Sep. 2019.

[18] A. Hofleitner, R. Herring, and A. Bayen, "Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning," *Transp. Res. B, Methodol.*, vol. 46, no. 9, pp. 1097–1122, Nov. 2012.

[19] 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," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2034, no. 1, pp. 62–72, Jan. 2007.

[20] J. Gong, M. Liu, and S. Zhang, "Hybrid dynamic prediction model of bus arrival time based on weighted of historical and real-time GPS data," in *Proc. 25th Chin. Control Decis. Conf. (CCDC)*, May 2013, pp. 972–976.

[21] X. Jiang and X. Yang, "Regression-based models for bus dwell time," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 2858–2863.

[22] Y. Li, C. Huang, and J. Jiang, "Research of bus arrival prediction model based on GPS and SVM," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2018, pp. 575–579.

[23] S. Khetarpaul, S. K. Gupta, S. Malhotra, and L. V. Subramaniam, "Bus arrival time prediction using a modified amalgamation of fuzzy clustering and neural network on Spatio-temporal data," in *Australasian Database Conference* (Lecture Notes in Computer Science). Melbourne, Australia: Springer, 2015, pp. 142–154.

[24] W. Treethidtaphat, W. Pattara-Atikom, and S. Khaimook, "Bus arrival time prediction at any distance of bus route using deep neural network model," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 988–992.

[25] H. Xu and J. Ying, "Bus arrival time prediction with real-time and historic data," *Cluster Comput.*, vol. 20, no. 4, pp. 3099–3106, Jun. 2017.

[26] P. He, G. Jiang, S.-K. Lam, and D. Tang, "Travel-time prediction of bus journey with multiple bus trips," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4192–4205, Nov. 2019.

[27] Z. Huang, Q. Li, F. Li, and J. Xia, "A novel bus-dispatching model based on passenger flow and arrival time prediction," *IEEE Access*, vol. 7, pp. 106453–106465, 2019.

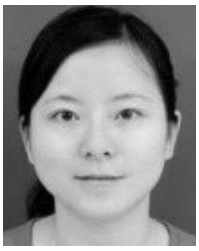
[28] *Historical Weather*, Chongqing, China. Accessed: Dec. 2016. [Online]. Available: http://tianqi.eastday.com/zhongqing_history/57516_201605.html?tdsourcetag=s_pctim_aiomsg

QINGWEN HAN was born in Chongqing, China, in 1969. She received the Ph.D. degree from Chongqing University, Chongqing, in 2010. She is currently an Associate Professor with the College of Microelectronics and Communication Engineering, Chongqing University. Her research interests include VANET, multicarrier communication systems, electro-optical signal processing, and time–frequency analysis and wavelets.

KE LIU was born in Sichuan, China. He received the B.S. degree from the College of Microelectronics and Communication Engineering, Chongqing University, China, in 2019, where he is currently pursuing the M.S. degree. His research is mainly in VANET.

LINGQIU ZENG was born in Chongqing, in 1975. He received the Ph.D. degree from Chongqing University, Chongqing, China, in 2009. He is currently an Associate Professor with the College of Computer Science, Chongqing University. His research interests include VANET, intelligent transportation, and data processing.

GUANGYAN HE was born in Chongqing, China. She received the B.S. degree from the School of Computer and Information Science, Southwest University, China, in 2016. She is currently pursuing the M.S. degree with the College of Computer Science, Chongqing University. Her research is mainly in VANET.



LEI YE received the B.Sc. degree from the Nanjing University of Post and Telecommunications, in 2001, and the Ph.D. degree from the University of York, U.K., in 2008. She is currently an Associate Professor with the College of Microelectronics and Communication Engineering, Chongqing University. Her research interests include wireless communications, especially modulation and coding, MIMO, and OFDM techniques.

FENGXI LI was born in Guizhou, China. He received the B.S. degree from the College of Electronic Information Engineering, Inner Mongolia Normal University, China, in 2016, and the M.S. degree from the College of Microelectronics and Communication Engineering, Chongqing University, China, in 2019. His research is mainly in VANET.