Ruisen JIANG, Ph.D.[1]
E-mail: 2018022016@chd.edu.cn
Dawei HU, Ph.D.[1]
(Corresponding author)
E-mail: dwhu@chd.edu.cn
Steven I-Jy CHIEN, Ph.D.[1,2]
E-mail: i.jy.chien@njit.edu
Qian SUN, Ph.D.[1]
E-mail: 1208049121@qq.com
Xue WU, Ph.D.[1]
E-mail: 1322766995@qq.com
[1] School of Transportation Engineering
Chang'an University, Xi'an, 710064, China
[2] Department of Civil and Environmental Engineering
New Jersey Institute of Technology
Newark, New Jersey, USA

Traffic Planning
Original Scientific Paper
Submitted: 24 Nov. 2021
Accepted: 17 Mar. 2022

PREDICTING BUS TRAVEL TIME WITH HYBRID INCOMPLETE DATA – A DEEP LEARNING APPROACH

ABSTRACT

The application of predicting bus travel time with real-time information, including Global Positioning System (GPS) and Electronic Smart Card (ESC) data is effective to advance the level of service by reducing wait time and improving schedule adherence. However, missing information in the data stream is inevitable for various reasons, which may seriously affect prediction accuracy. To address this problem, this research proposes a Long Short-Term Memory (LSTM) model to predict bus travel time, considering incomplete data. To improve the model performance in terms of accuracy and efficiency, a Genetic Algorithm (GA) is developed and applied to optimise hyperparameters of the LSTM model. The model performance is assessed by simulation and real-world data. The results suggest that the proposed approach with hybrid data outperforms the approaches with ESC and GPS data individually. With GA, the proposed model outperforms the traditional one in terms of lower Root Mean Square Error (RMSE). The prediction accuracy with various combinations of ESC and GPS data is assessed. The results can serve as a guideline for transit agencies to deploy GPS devices in a bus fleet considering the market penetration of ESC.

KEYWORDS

bus travel time prediction; GPS data; electronic smart card data; long short-term memory model; genetic algorithm.

1. INTRODUCTION

The growing demand for freight and passenger transport needs to be met in order to prevent negative externalities. To maximise the total profit with the trend of growing freight demands, lots of studies proposed ships or truck scheduling models [1, 2]. For passenger transit agencies, adjusting the structure of the transportation system based on growing passenger flows is of great importance [3, 4]. Moreover, since passenger demand drastically increases, it will deteriorate the service quality and efficiency of bus transportation systems, especially during peak hours [5]. Therefore, developing a model to enhance the service quality is desirable to facilitate bus operation.

Maintaining schedule adherence and promoting the level of service are challenging for bus transit agencies. Poor schedule adherence increases passenger waiting time and reduces the attractiveness of bus service. Developing a sound approach to predict bus travel time in real-time is desirable to facilitate bus operation. However, bus travel time is stochastic and sometimes difficult to predict because of many factors, such as passenger boarding/alighting demand, traffic conditions and delays at intersections [6, 7], especially in peak periods. Providing reliable and accurate bus travel time would be an effective way to improve the level of service [8–10].

Bus arrival/departure times at stops reported by GPS data can be used to develop bus travel time prediction models [11]. The automated data-collection system in Xi'an Traffic Information Centre consists of ESC and GPS data, which is used to monitor passenger flow and the location of buses in the urban bus transit system in real-time [12, 13]. However,

the GPS information (e.g. latitude/ longitude coordinates etc.) is obtained in a fixed time interval (e.g. 15 seconds), which may affect the prediction accuracy and stability. Therefore, the ESC data can be applied to fill the gap and improve prediction accuracy. Zhou et al. [14] predicted bus travel time using hybrid data. The results suggested that the model using hybrid data outperformed those using GPS data or ESC data, individually. Previous studies assumed that all buses equipped GPS devices and all passengers used smart cards. However, in a practical urban network, missing information in data stream is unavoidable for various reasons. For example, many agencies do not equip GPS devices for an entire bus fleet, and in some particular situations the smart card records are unavailable (i.e. passengers who paid cash for ticket fares, stops without boarding passengers etc.) [15, 16]. These situations may jeopardise the performance of prediction accuracy and stability [17].

Bus travel time is dynamic and stochastic, affected by various factors such as passenger boarding/ alighting time, traffic conditions and delays at intersections [18]. Recurrent Neural Network (RNN) considers a sequence of data inputs and has been widely used in time sequence analysis. The LSTM is an advanced form of RNN, which is robust for predicting bus travel time [19, 20]. However, LSTM is characterised by a set of hyperparameters, which shall be effectively determined by a sound algorithm to yield the best performance.

The focus of this study is to develop an LSTM model for predicting bus travel time, considering incomplete data. A GA is developed and applied to calibrate the hyperparameters of the LSTM model. Our findings show that optimised parameters via GA can effectively enhance the model performance in terms of prediction accuracy. Moreover, the available proportion of GPS data and ESC data could affect the prediction accuracy. Transit agencies shall effectively select the type and amount of data to predict bus travel time, according to the penetration of smart card users as well as the fraction of buses equipped GPS devices. The following section discusses the review of previous studies on bus travel time prediction. Then, section 3 discusses the proposed approach to predict bus travel time using incomplete data. Section 4 focuses on discussing a case study in Xi'an, China. Section 5 reveals the results of the case study. Finally, section 6 summarises research findings and discusses future research.

2. LITERATURE REVIEW

Bus travel time is stochastic because of traffic conditions between stops, dwell time at stops and delays at intersections, which fluctuate spatially and temporally. Providing reliable and accurate bus travel time is one of the effective ways to enhance the level of service [8]. However, developing a sound model considering incomplete data is a challenging task. Dai and Mu [17] predicted bus travel time with various degrees of missing ESC data. It was found that prediction accuracy decreased as the amount of data missing increased. Previous studies demonstrated that using real-time information (e.g. GPS data and/or ESC data) can improve prediction accuracy. However, missing information in GPS data and ESC data streams is common for various reasons, such as system stability affected by data communication under different geographical and environmental restrictions. Only a few studies predicted bus travel time considering the impact of missing data [16].

Deep learning models can map complex relations between the input factors and bus travel time without an explicit function form [21], which can be classified into Support Vector Machine (SVM) models, Kalman Filtering (KF) models, and Artificial Neural Network (ANN) models.

Zhong et al. [22] developed an SVM model to predict bus travel time. Yang et al. [23] integrated SVM and GA to predict bus travel time. Later, Peng et al. [24] enhanced the GA-SVM model by a principal component analysis algorithm. The results suggested that the enhanced SVM model outperformed the traditional SVM model. Bus travel time on a particular path has time sequence characteristics (i.e. consecutive buses operate in a similar traffic condition) [25]. However, the SVM model has limitations to forecast time sequence information, such as bus travel time.

KF is capable of updating the state variable with new observations, which has been widely applied to predict time sequence information [26]. Chien and Kuchipudi [27] developed a KF model to predict bus travel time using GPS and Automatic Passenger Counters (APC) data. Jairam et al. [28] applied a KF model to predict bus travel time. The results suggested that the KF model outperformed the SVM and the historical mean prediction model. Due to the inherent limitations of the Markov property, the KF model deteriorated as the number of time steps increased [29].

ANN models have been widely applied to predict bus travel time. Jeong and Rilett [30] used ANN to predict bus arrival time using Automatic Vehicle

Location (AVL) systems. The results suggested that ANN outperformed SVM. RNN is an advanced form of ANN, which is capable of forecasting bus travel time [31]. Many deep learning methods, such as LSTM, Gated Recurrent Unit (GRU), Bi-directional Long Short Term Memory (Bi-LSTM), Bi-Gated Recurrent Unit (Bi-GRU) and Graph Convolution Network (GCN) are similar to the RNN structure, which has been widely applied to predict time sequence information. LSTM is an advanced RNN structure, which is robust for predicting bus travel time [19, 20]. Liu et al. [32] developed an LSTM model to predict bus travel time using GPS data and found that LSTM outperformed RNN. GRU is proposed as a simpler alternative to LSTM. GRU is a simpler alternative to LSTM but characterised by fewer hyperparameters, the performance of which can be improved by enhancing its training process [33]. Zhai et al. [34] developed a GRU model to predict vehicle speed and found that GRU outperformed a convolutional neural network (CNN) model. However, GRU simplifies the LSTM by reducing long-term time series feedback, which has limitations to forecasting stochastic information [35]. Bi-LSTM and Bi-GRU are the advanced LSTM and GRU structures, respectively, which train data from two directions. Xue et al. [36] developed a Bi-LSTM model to predict expressway traffic flow, which exhibited higher prediction accuracy during off-peak hours. Shu et al. [37] developed a Bi-GRU model to predict short-term traffic flow and found that Bi-GRU outperformed GRU. However, the training of Bi-LSTM and Bi-GRU is computationally expensive [37]. The travel time of a bus is highly correlated to the travel times of its leading buses (e.g. accidents and traffic jams). Thus, training from two directions is unnecessary. GCN is used to extract the features of topology structure in time sequence information, which is used to predict bus travel time, speed and passenger flow. However, it is computationally expensive [38].

LSTM is characterised by a set of hyperparameters, which shall be effectively determined to yield the best prediction performance using GA and Evolutionary Algorithm (EA). Zhao and Zhang [39] proposed a hybrid model including a learning-based algorithm and EA to solve multi-objective optimisation problems. With a GA, Dulebenets [40] proposed an improved GA to optimise the truck schedule. Liu et al. [41] proposed an EA with an angle-based selection strategy and a shift-based density estimation strategy to optimise multi-objective problems. Pasha et al. [42] proposed an EA to optimise a supply chain problem and found that EA outperforms the other metaheuristic algorithms (i.e. Variable Neighbourhood Search, Tabu Search, and Simulated Annealing). D'Angelo et al. [43] proposed a hybrid deep learning model using GA and a decision-tree model to distinguish between meningitis etiologies using standard and clinical datasets.

3. METHODOLOGY

The objective of this study is to develop a model to predict bus travel time from stop *i* to all downstream stops *j* considering missing data. To discuss the model development, the trajectories of buses on a general route are illustrated in *Figure 1*, where *k*

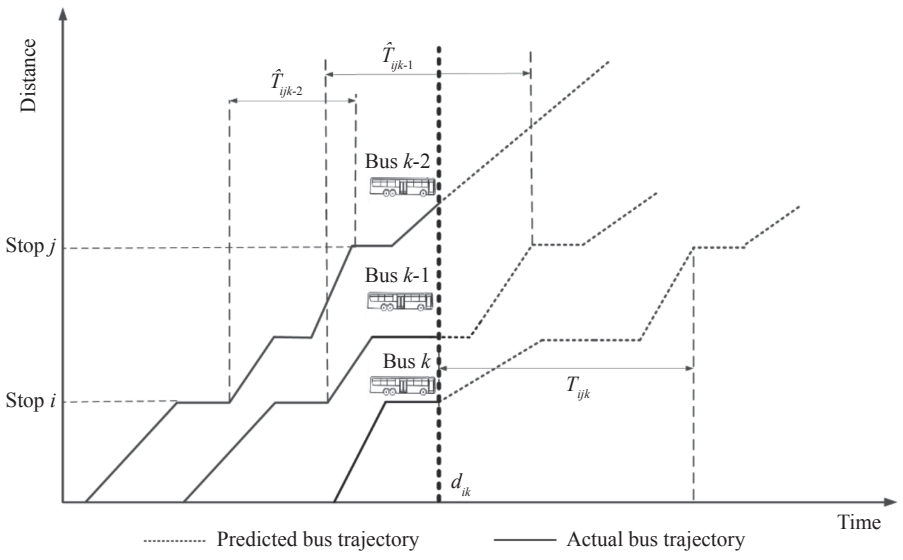


Figure 1 – Bus trajectories on a general route

represents the index of buses, i and j are indices of stops. Note that bus k-1 is immediately dispatched before bus k.

The travel time of bus k from stop i to j denoted as T_{ijk} can be determined by *Equation 1* [13].

$$T_{ijk} = a_{jk} - d_{ik} \tag{1}$$

where a_{jk} is the arrival time of bus k at stop j, and d_{ik} is the departure time of bus k at stop i.

T_{ijk} can be predicted by historic data, such as travel time of bus k-1 and bus k-2 denoted as $\hat{T}_{ijk-1}$ and $\hat{T}_{ijk-2}$ respectively in *Figure 1*.

The arrival and departure time of bus k at stop j reported by GPS data (i.e. a_{Gjk} and d_{Gjk}, respectively) can be determined by bus speed and the coordinates of stop locations. On the other hand, the arrival time of bus k at stop i reported by ESC data (i.e. a_{Eik}) is determined by the time of the first passenger boarding from stop i. The proposed model cannot obtain the departure time of bus k at stop i (d_{Eik}) using real-time ESC data (i.e. cannot judge who is the last boarding passenger in stop i immediately). The way to approximate d_{Eik} is a_{Eik} plus the average dwell time $\overline{D}_i$. Thus,

$$d_{Eik} = a_{Eik} + \overline{D}_i \tag{2}$$

Considering the availability of the ESC and GPS data, four scenarios are considered to obtain actual arrival/departure time information. Stops with passengers who use smart card, at the same time, receive accurate GPS signals, the actual arrival time a_{ik} is determined by the earlier arrival time, and the actual departure time d_{ik} is the later departure time reported from ESC and GPS data. Thus,

$$\begin{cases} a_{ik} = \min\{a_{Eik}, a_{Gik}\} \\ d_{ik} = \max\{d_{Eik}, d_{Gik}\} \end{cases} \tag{3}$$

where a_{Eik} and a_{Gik} are the arrival times of bus k at stop i obtained by ESC and GPS data respectively, while d_{Eik} and d_{Gik} are the departure times from stop i obtained by ESC and GPS data, respectively.

Stops without passengers who use smart card, at the same time, receive accurate GPS signals, and the actual bus arrival/departure time refers to GPS data. Thus,

$$\begin{cases} a_{ik} = a_{Gik} \\ d_{ik} = d_{Gik} \end{cases} \tag{4}$$

Stops with passengers who use smart card, at the same time, cannot receive accurate GPS signals, and the actual bus arrival/departure time refers to ESC data. Thus,

$$\begin{cases} a_{ik} = a_{Eik} \\ d_{ik} = d_{Eik} \end{cases} \tag{5}$$

Stops without passengers who use smart card, at the same time, cannot receive accurate GPS signals, and the bus travel time from stop i to downstream stops refers to the travel time of the immediate leading bus.

The proposed LSTM model is developed with hybrid data, and a GA is applied to optimise its hyperparameters. In *Figure 2*, the model consists of a set of cells (C_{ijk}) and is used to predict travel time (T_{ijk}) of bus k between stops i and j.

Cell C_{ijk} consists of an input layer with travel times from stop i to j of some previous buses. The number of inputs (i.e. previous buses) depends on a hyperparameter of the LSTM model (i.e. lag sizes). Lag size (L) refers to the number of input parameters (consecutive travel times) given as input of the

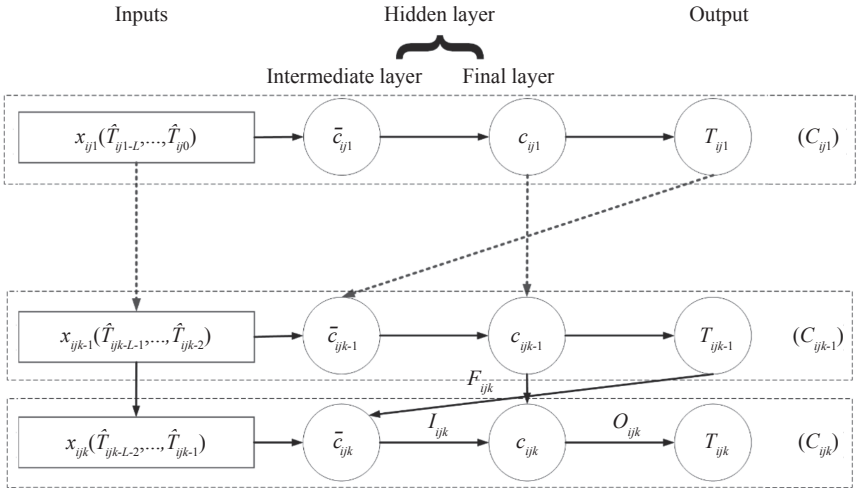


Figure 2 – Configuration of the LSTM model

model, the input parameters of LSTM x_{ijk} consists of $\hat{T}_{ijk-1} \ldots \hat{T}_{ijk-L}$ (e.g. if L=1, x_{ijk} is $\hat{T}_{ijk-1}$, if L=2, x_{ijk} consists of $\hat{T}_{ijk-1}$ and $\hat{T}_{ijk-2}$. One hidden layer consists of an intermediate layer and a final layer. One output layer includes the predicted travel time of bus k between stops i and j (T_{ijk}).

In input layer, if a preceding bus has arrived at stop j, $\hat{T}_{ijk}$ can be determined by *Equation 1*. Note that the number of preceding buses is determined by the lag sizes of the LSTM model.

The node of an intermediate layer in cell C_{ijk} denoted as $\bar{c}_{ijk}$ is determined by *Equation 6* [16].

$$\bar{c}_{ijk} = \sigma(\omega_{cij} \cdot [\hat{T}_{ijk-1}, x_{ijk}] + b_{cij}) \tag{6}$$

where b_{cij} and w_{cij} represent a bias and a weight matrix associated with the cells of the intermediate layer, respectively, which are adjusted in the LSTM model. σ represents a sigmoid function.

The output of a hidden layer in cell C_{ijk} denoted as c_{ijk}, is determined by $\bar{c}_{ijk}$ and c_{ijk-1} with *Equations 7–9* [16].

$$c_{ijk} = F_{ijk} c_{ijk-1} + I_{ijk} \bar{c}_{ijk} \tag{7}$$

$$F_{ijk} = \sigma(\omega_{Fij} \cdot [\hat{T}_{ijk-1}, x_{ijk}] + b_{Fij}) \tag{8}$$

$$I_{ijk} = \sigma(\omega_{Iij} \cdot [\hat{T}_{ijk-1}, x_{ijk}] + b_{Iij}) \tag{9}$$

where F_{ijk} and I_{ijk} are outputs of the forget gate and the input gate of cell C_{ijk}; b_{Fij} and w_{Fij} represent the bias and weight matrices of the forget gate; b_{Iij} and w_{Iij} represent the bias and weight matrices of the input gate, respectively.

The output of cell C_{ijk} is T_{ijk}, which can be determined by c_{ijk} with *Equations 10 and 11* [16].

$$O_{ijk} = \sigma(\omega_{Oij} \cdot [\hat{T}_{ijk-1}), x_{ijk}] + b_{Oij}) \tag{10}$$

$$T_{ijk} = O_{ijk} \cdot \tanh(c_{ijk}) \tag{11}$$

where O_{ijk} is the output of the output gate; tanh is a hyperbolic tangent function; b_{Oij} and w_{Oij} represent the bias and weight matrices of the output gate, respectively.

The proposed LSTM model is developed by using a module of the MATLAB Toolbox named lstmLayer. Batch sizes (B) and lag sizes (L) are two hyperparameters to be optimised for yielding the best performance in terms of accuracy [44].

Batch size (B) refers to how many cells are used in the LSTM model, which dictates the bias and weight matrices (e.g. b_{cij} and w_{cij}). Lag size (L) refers to the number of previous bus travel time given as input of the model.

The grid search algorithm was commonly applied to train hyperparameters B and L. However, the process is complicated and computationally expensive. In this study, we developed a GA to optimise B and L, which minimizes RMSE expressed by *Equation 12* [45].

$$RMSE = \sqrt{\sum_{k=1}^{N} |T_{ijk} - \hat{T}_{ijk}|^2 / N)} \tag{12}$$

where k is the index of buses; N is the number of samples.

To ensure that the hyperparameters always minimise RMSE, the fitness function of GA is the inverse of RMSE shown as *Equation 13*.

$$f = 1/RMSE \tag{13}$$

The GA starts by setting parameters and initialising a set of feasible hyperparameters (i.e. B and L). Then, the LSTM model employs these hyperparameters to predict travel time, followed by computing the RMSE of predicted travel time against actual travel time. Then, new solutions are produced through selection, crossover, mutation and fitness evaluation, until the terminating condition (i.e. max number of generations) is satisfied. The detailed description is given below and illustrated in *Figure 3*.

Step 1: Specify the parameters of GA, including population size (i.e. 20), crossover rate (i.e. 0.9), mutation rate (i.e. 0.1), and the maximum number of iterations (i.e. t_{max}=30), then generate the initial set of random feasible solutions and initialise the generation counter (i.e. t=0).

Step 2: Run the LSTM model with the hyperparameters suggested by the initial set of solutions and calculate the fitness function of each solution with *Equation 13*. After that, input the initial set of solutions and fitness function value of each solution to Step 3.

Step 3: The tournament selection is applied (i.e. randomly select two solutions from the initial set of solutions and then choose the one yielding the best fitness value). This process is repeated until the number of solutions is equal to the population size. Then, a new set of solutions is produced via uniform crossover (i.e. randomly selects portion genes of two parents, then exchanges these genes to produce new children) and random mutation operations (i.e. randomly selects two genes from a parent, then exchanges these genes to produce a new child). After that, the fitness function of each solution is calculated with *Equation 13*.

Step 4: Check if the maximum number of iterations is attained. If not, replace the initial set of solutions with a new set of solutions, update the generation counter and go to Step 2; otherwise, terminate the algorithm and report the best hyperparameters (i.e. the solution with the best fitness value in the set of solutions).

4. CASE STUDY

In the case study, we employ route 35 to test the proposed approach performance. Route 35 serves passengers in a central business district (CBD) in Xi'an, China, as shown in *Figure 4*. The study route is 11 km long and serves 21 stops with 3-minute headway during 7:00~9:00 and 18:00~20:00, and 7-minute headway in other periods. The average stop spacing is 0.55 km, and the fleet size is 23 buses. Each bus has 38 seats with two doors for boarding and alighting passengers individually. The GPS data and ESC data associated with the study route were obtained during weekdays in May 2019.

The ESC data consists of detection date, detection time, boarding stop ID and card ID as shown in *Table 1*, which is recorded and then fed into an

Table 1 – Sample ESC data

Detection Date	Detection Time	Boarding Stop ID	Card ID
5/4/2019	8:02:23	8	176821
5/4/2019	8:02:26	8	2819765
5/4/2019	8:05:13	9	543197
5/4/2019	8:05:14	9	2371283
5/4/2019	8:05:21	9	2715300
5/4/2019	8:07:58	10	351271
5/4/2019	8:08:04	10	765123

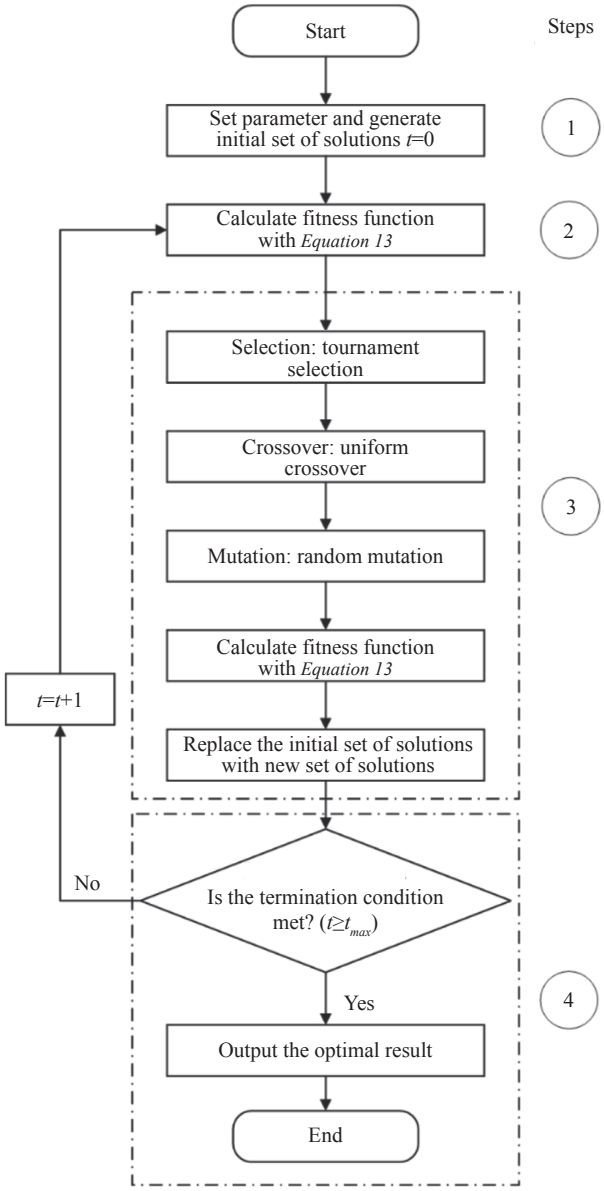


Figure 3 – GA algorithm for hyperparameters optimisation

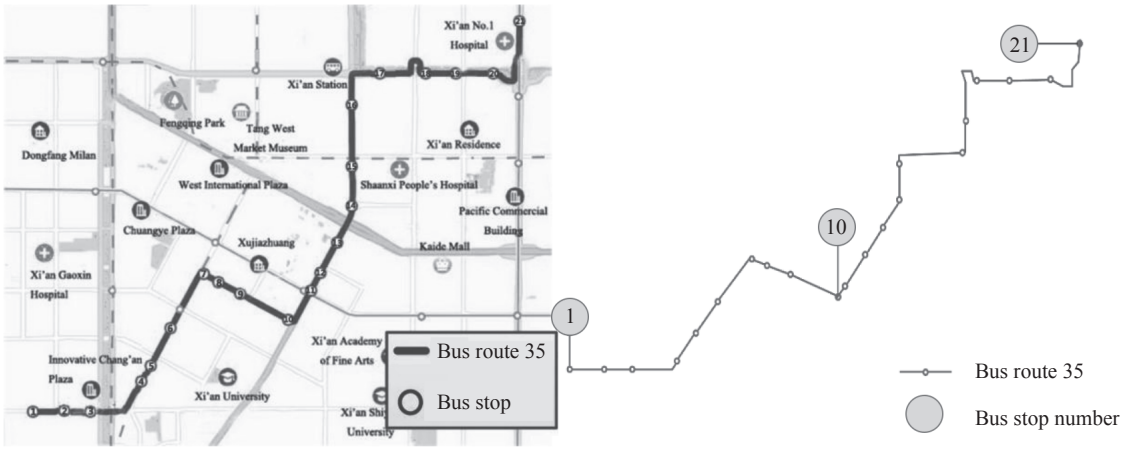


Figure 4 – Configuration of route 35

MS-Access database in Xi'an traffic information centre in real-time. There are 5,789 records for the weekdays in May 2019.

The GPS data consists of detection date, detection time, latitude, longitude and speed, which are reported on a 15-second interval basis as shown in *Table 2*. There are 355,467 records stored in the MS-Access database.

It is challenging to accurately predict bus travel time in peak hours because of the variation of traffic congestion and passenger demand [46]. Thus, the proposed approach performance is assessed using peak-hour data in the morning (i.e. 7:00~9:00). The statistics of the data are illustrated in *Table 3*, including the stop spacing, the average and standard deviation of link travel time and dwell time.

Table 2 – Samples of GPS data

Detection Date	Detection Time	Latitude	Longitude	Speed [km/h]
5/4/2019	8:02:00	108.946416	34.256137	0.02
5/4/2019	8:02:15	108.946381	34.256216	0.00
5/4/2019	8:02:30	108.946213	34.256072	0.00
5/4/2019	8:02:45	108.946357	34.256418	3.34
5/4/2019	8:03:00	108.946341	34.256146	8.45
5/4/2019	8:03:15	108.946137	34.256173	12.83
5/4/2019	8:03:30	108.946257	34.256164	15.02

Table 3 – Stop spacing and travel time of the study route

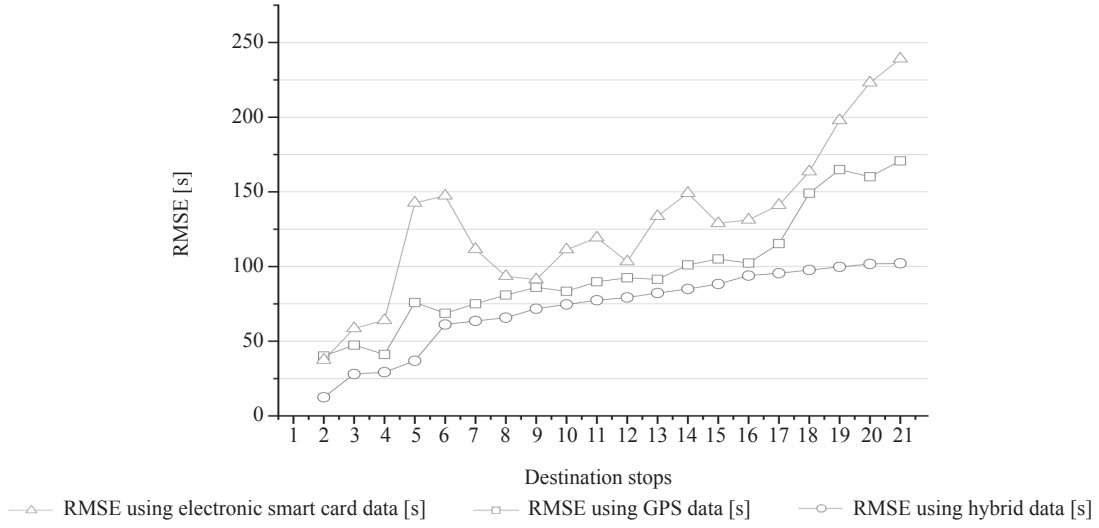
Stop No.	Stop spacing [m]	Average link travel time [s]	SD of link travel time [s]	Average dwell time [s]	SD of dwell time [s]
1	0	/	/	/	/
2	422	46	31	12	8
3	355	73	22	25	13
4	229	37	32	30	16
5	310	68	51	28	16
6	367	35	21	21	14
7	758	91	43	22	10
8	597	77	28	25	11
9	721	74	26	32	13
10	582	55	31	52	27
11	698	83	28	77	26
12	631	62	21	41	21
13	521	57	27	29	11
14	468	57	28	13	10
15	537	43	32	23	17
16	568	59	31	21	11
17	821	79	28	23	28
18	397	54	33	18	17
19	289	37	28	39	21
20	673	89	29	68	19
21	1179	179	65	/	/

Note: SD = standard deviation

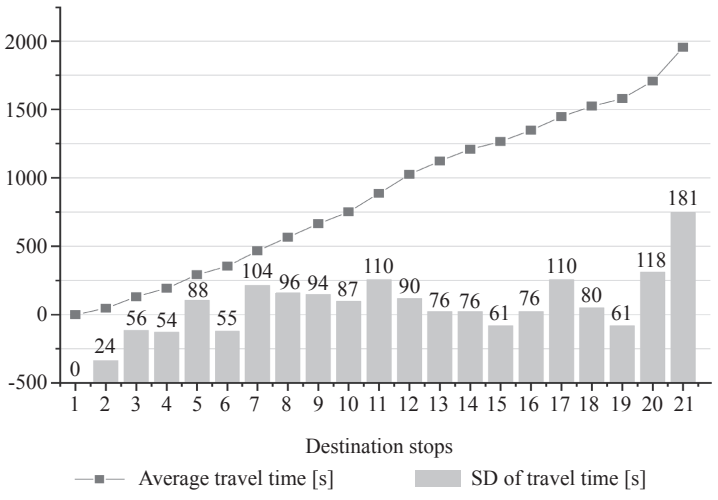
5. RESULTS AND DISCUSSION

The assessment was conducted by using different combinations of data sets including ESC data only, GPS data only and hybrid data (ESC and GPS data). We evaluated the prediction accuracy based on RMSE as shown in *Figure 5*, which illustrates the RMSE of bus travel time from stop 1 to all downstream stops (i.e. stops 2 through 21).

Figure 5 indicates that the RMSE increases with travel time. The RMSE with ESC data only is generally high and fluctuates significantly in stops with a high standard deviation of travel time, which can be attributed to the market penetration of smart cards (80% of passengers in the study route used smart cards). One can also observe that the RMSE of approaches using ESC and GPS data show a trend of increase along the average bus travel time, especially after stop 17. This can be attributed to the error of obtaining arrival time using ESC and GPS data, respectively. The model developed with hybrid data outperforms others in terms of lower RMSE; the average RMSE is reduced by 44.11% and 25.47%, respectively. Although its RMSE shows a trend of increase along the bus travel time, it seems quite stable with only a few hikes. This seems a significant benefit from adapting a deep learning approach, such as the LSTM model, to ensure higher prediction accuracy, because it can fine-tune the prediction result based on real-time data. With more accurate bus arrival information, passengers can therefore schedule their departure time to reduce waiting time at stops effectively.



a) RMSE with different data sets



b) Average and standard deviation of travel time

Figure 5 – RMSE with different data sets and average travel time (7:00-9:00)

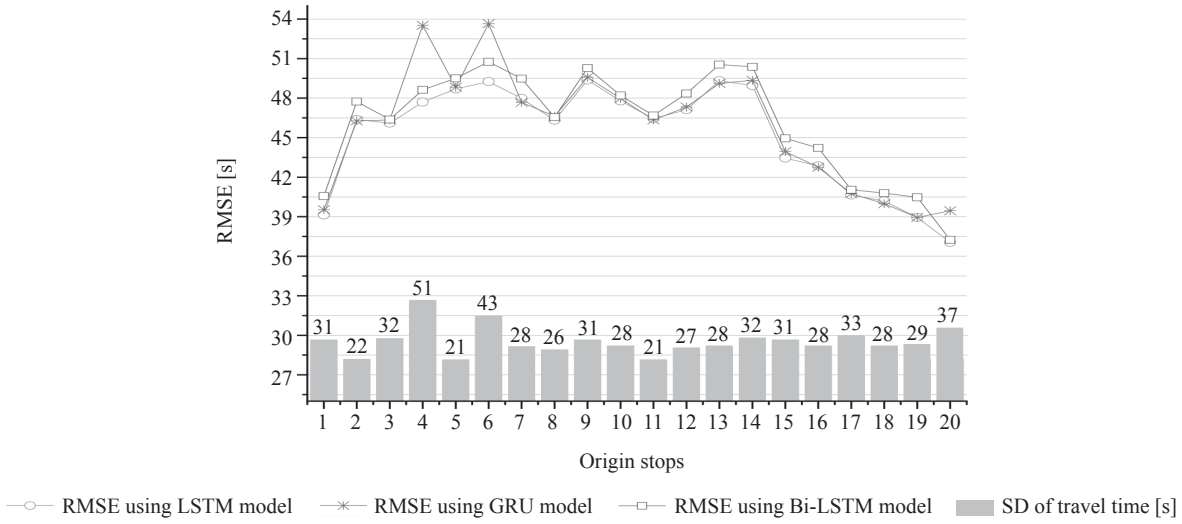


Figure 6 – RMSE with different models

The assessment is conducted by using different deep learning methods including LSTM, GRU and Bi-LSTM. We evaluate the prediction accuracy as shown in *Figure 6*. *Figure 6* illustrates the RMSE of bus travel time prediction model using different models from origin stop (stop *i*) to stop *i*+1 and standard deviation of travel time in this path.

As shown in *Figure 6*, the proposed approach outperforms GRU and Bi-LSTM in prediction accuracy; the average RMSE is reduced by 0.86s and 0.95s, respectively. This is because the GRU is unstable in some paths with higher standard deviation of travel time (i.e. origin from stops 4 and 6). Since travel time is almost solely affected by leading buses in the research route, LSTM outperforms the model with Bi-LSTM in prediction accuracy.

To analyse how GA can improve the accuracy of the prediction approach, the performance is assessed by LSTM with GA and grid search using hybrid data set. In the approach without GA, the hyperparameters of each LSTM model are trained by a grid search algorithm. *Figure 7* compares the performance of the proposed approach with GA and LSTM with grid search from origin stop (stop *i*) to stop *i*+1.

As shown in *Figure 7*, the proposed approach outperforms the LSTM with grid search in prediction accuracy, and the average RMSE is reduced by 6.62%. This is because GA could stably find the optimal solution (i.e. hyperparameters for the proposed LSTM model). As a result, GA is important

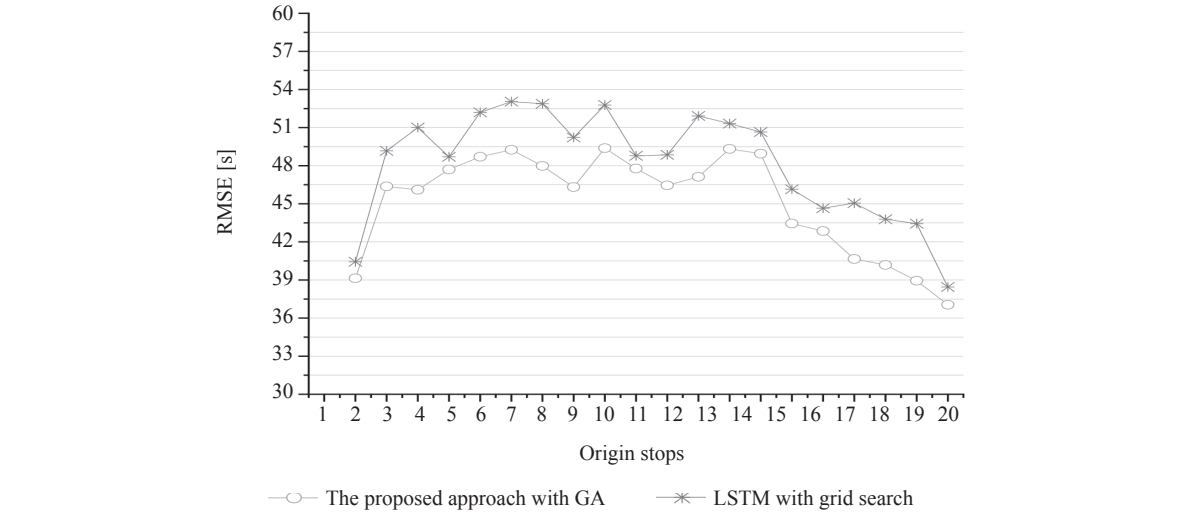


Figure 7 – RMSE of the approach with GA and LSTM with grid search

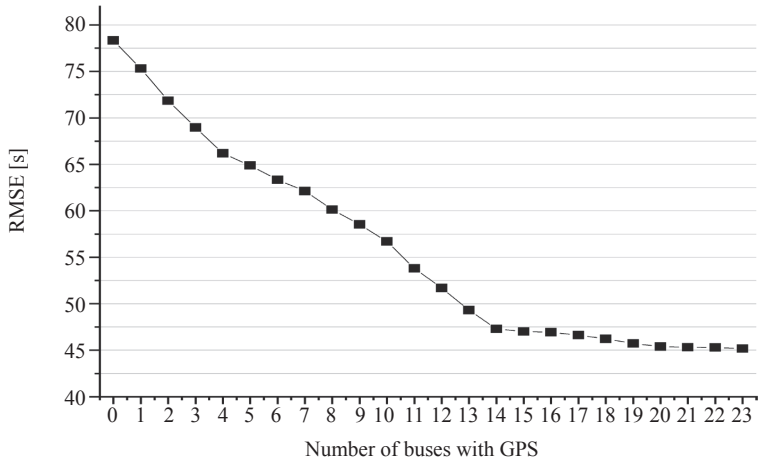
to train the hyperparameters for the LSTM model, which can improve the stability and accuracy of the proposed approach.

To test the overfitting of the model, we used the data obtained on odd days of route 35 during the weekdays in May 2019 as the train data set. We optimised the hyperparameters of the LSTM model using the train data set. Then, we used the same hyperparameters to predict bus travel time on even days of route 35 during the weekdays in May 2019 (i.e. test data set). The result shows that the average RMSE of the train data set and the test data set are 44.73s and 45.21s, respectively. The gap between the RMSE of the train data set and the test data set is small (i.e. 0.48s). Therefore, parameters obtained by GA are stable to predict bus travel time.

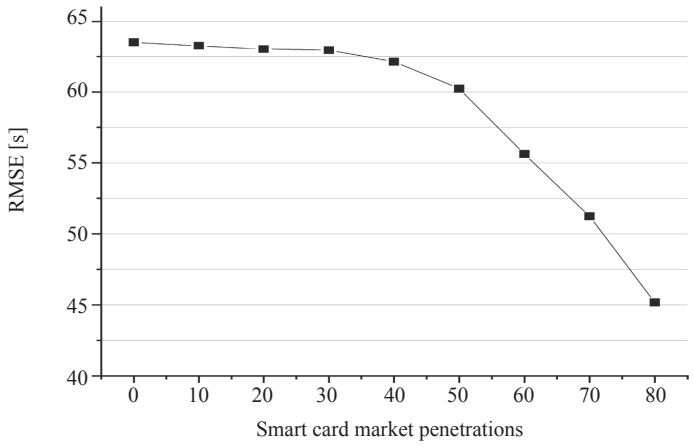
To analyse the impact of the available GPS data and ESC data (i.e. buses which equipped GPS devices and the proportions of passengers who paid cash for tickets), the proposed approach is developed using the hybrid data with various combinations of GPS and ESC data in *Figure 8*.

Figure 8a shows how RMSE changes when prediction is performed using hybrid data with different numbers of buses equipped GPS devices in the study fleet. As shown in *Figure 8a*, RMSE decreases significantly at the beginning (from 1 through 14 buses) as the number of buses equipped GPS devices increases, and then RMSE slightly decreases (from 14 through 23 buses). The agency may consider equipping 14 through 23 buses (i.e. more than 61% buses in the study fleet) subject to the budget constraint.

Figure 8b shows how RMSE changes when the prediction is performed using hybrid data with different smart card market penetrations (i.e. the rate of available smart card data). In the case of 80% of ESC market penetration, the available ESC data ranges from 0% to 80%. It was found that RMSE



a) RMSE vs. number of buses with GPS



b) RMSE vs. smart card market penetrations

Figure 8 – RMSE with different percentages of hybrid data

decreases as the available ESC data increases. The available ESC data shall be higher than 50% and the significant accuracy of prediction can be expected.

6. CONCLUSION

In this study, real-world ESC and GPS data are used to develop the proposed LSTM model for predicting bus travel time with incomplete data. The results suggest that using hybrid data would outperform the approach with ESC data only and GPS data only; the RMSE is reduced by 44.11% and 25.47%, respectively. The results also suggest that LSTM outperforms GRU and Bi-LSTM; the RMSE is reduced by 0.86s and 0.95s, respectively.

To improve the accuracy of the approach, a GA was developed, which demonstrated itself effectively optimising the hyperparameters of the LSTM model. Its performance outperforms the traditional LSTM model that uses the hyperparameter determined by performing a grid search, in terms of lower RMSE (reduced by 6.62%).

The results also suggest that more than 61% and 50% of available GPS and ESC data, respectively, can be used to improve the performance of prediction accuracy. *Figures 8a and 8b* can serve as a guideline to equip GPS devices in a bus fleet subject to available ESC data and expected travel time prediction accuracy.

As immediate extensions of this study, the developed LSTM model can be enhanced by considering more realistic conditions. The avenues of future research may include the following: (i) considering link travel time uncertainties that may occur due to delay at intersections, weather characteristics and other factors; (ii) considering other routes which pass the same pair of OD (e.g. stops) to enrich the data set; (iii) improving the model performance based on other deep learning models (e.g. Bi-LSTM, Bi-GRU and GCN); and (iv) improving the performance of the prediction model by enhancing ESC data sets after the real-time data transmission technique has been improved.

ACKNOWLEDGEMENT

This paper was supported by the National Natural Science Foundation of Shaanxi, China, under grants 2020JQ-399 and 2021JZ-20. The authors also would like to thank the Xi'an public transport company and Xi'an traffic information centre for providing the data for this research.

姜瑞森，胡大伟，STEVEN I-JY CHIEN，孙倩，吴雪

利用混合不完整数据预测公交运行时间：一种深度学习方法

摘要

利用实时全球定位系统和电子智能卡数据预测公交运行时间可以减少乘客等待时间、提高公交到站准点率从而提高服务水平。然而数据传输过程中由于各种原因造成的信息缺失是不可避免的，这可能严重影响预测的准确性。为了解决这一问题，本研究提出了一种考虑不完整数据的长短时记忆模型以预测公交运行时间。为了提高模型的准确性和预测效率，本研究建立了遗传算法以优化长短时记忆模型的超参数。研究通过模拟数据和实际数据评估模型的性能。结果表明，使用混合数据的预测结果优于单独使用全球定位系统数据和电子智能卡数据的预测结果。结合遗传算法的预测模型预测结果降低了传统长短时记忆模型的均方根误差。研究评估了使用不同比例全球定位系统数据和电子智能卡数据的预测精度。研究结果可以结合电子智能卡的市场占有率帮助交通管理机构优化公交车队中部署全球定位系统设备的数量。

关键词

公交运行时间预测；*GP*S数据；电子智能卡数据；长短时记忆模型；遗传算法

REFERENCES

[1] Dulebenets MA. A Delayed Start Parallel Evolutionary Algorithm for just-in-time truck scheduling at a cross-docking facility. *International Journal of Production Economics*. 2019;212: 236-258. doi: 10.1016/j.ijpe.2019.02.017.

[2] Pasha J, et al. An integrated optimization method for tactical-level planning in liner shipping with heterogeneous ship fleet and environmental considerations. *Advanced Engineering Informatics*. 2021;48: 101299. doi: 10.1016/j.aei.2021.101299.

[3] Belokurov V, Spodarev R, Belokurov S. Determining passenger traffic as important factor in urban public transport system. *Transportation Research Procedia*. 2020;50: 52-58. doi: 10.1016/j.trpro.2020.10.007.

[4] Kağan Albayrak MB, Özcan İÇ, Dobruszkes F. The determinants of air passenger traffic at Turkish airports. *Journal of Air Transport Management*. 2020;86: 101818. doi: 10.1016/j.jairtraman.2020.101818.

[5] Enoch MP, et al. Future local passenger transport system scenarios and implications for policy and practice. *Transport Policy*. 2020;90: 52-67. doi: 10.1016/j.tranpol.2020.02.009.

[6] Zhao L, Chien S, Spasovic LN, Liu X. Modeling and optimizing urban bus transit considering headway variation for cost and service reliability analysis. *Transportation Planning & Technology*. 2018;41(7): 706-723. doi: 10.1080/03081060.2018.1504181.

[7] Zhao L, Chien S. Investigating the impact of stochastic

vehicle arrivals to optimal stop spacing and headway for a feeder bus route. *Journal of Advanced Transportation.* 2015;49(3): 341-357. doi: 10.1002/atr.1270.

[8] Chien S, Ding Y, Wei C. Dynamic bus arrival time prediction with artificial neural networks. *Journal of Transportation Engineering.* 2002;128(5): 429-438. doi: 10.1061/(ASCE)0733-947X(2002)128:5(429).

[9] Tilahun SL, Ong HC. Bus timetabling as a fuzzy multiobjective optimization problem using preference-based genetic algorithm. *Promet – Traffic&Transportation.* 2012;24(3): 183-191. doi: 10.7307/ptt.v24i3.311.

[10] Ma W, Lin N, Chen X, Zhang W. A robust optimization approach to public transit mobile real-time information. *Promet – Traffic&Transportation.* 2018;30(5): 501-512. doi: 10.7307/ptt.v30i5.2609.

[11] Mazloumi E, Rose G, Currie G, Moridpour S. Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Engineering Applications of Artificial Intelligence*. 2011;24(3): 534-542. doi: 10.1016/j.engappai.2010.11.004.

[12] Zhang J, et al. A real-time passenger flow estimation and prediction method for urban bus transit systems. *IEEE Transactions on Intelligent Transportation Systems*. 2017;18(11): 3168-3178. doi: 10.1109/TITS.2017.2686877.

[13] Ma J, et al. Bus travel time prediction with real-time traffic information. *Transportation Research Part C: Emerging Technologies*. 2019;105: 536-549. doi: 10.1016/j.trc.2019.06.008.

[14] Zhou Y, et al. Bus arrival time calculation model based on smart card data. *Transportation Research Part C: Emerging Technologies*. 2017;74: 81-96. doi: 10.1016/j.trc.2016.11.014.

[15] Dai Z, Ma X, Chen X. Bus travel time modeling using GPS probe and smart card data: A probabilistic approach considering link travel time and station dwell time. *Journal of Intelligent Transportation Systems*. 2019;23(2): 175-190. doi: 10.1080/15472450.2018.1470932.

[16] Petersen NC, Rodrigues F, and Pereira FC. Multi-output bus travel time prediction with convolutional LSTM neural network. *Expert Systems with Applications*. 2019;120(4): 426-435. doi: 10.1016/j.eswa.2018.11.028.

[17] Dai D, Mu D. An algorithm for bus trajectory extraction based on incomplete data source. *Chinese Journal of Electronics*. 2012;021(004): 599-603.

[18] Kumar BA, Vanajakshi L, Subramanian SC. Bus travel time prediction using a time-space discretization approach. *Transportation Research Part C: Emerging Technologies.* 2017;79: 308-332. doi: 10.1016/j.trc.2017.04.002.

[19] Shabarek A, Chien S, Hadri S. Deep learning framework for freeway speed prediction in adverse weather. *Transportation Research Record.* 2020;2674(10): 28-41. doi: 10.1177/0361198120947421.

[20] He P, Jiang G, Lam SK, Tang D. Travel-time prediction of bus journey with multiple bus trips. *IEEE Transactions on Intelligent Transportation Systems*. 2018;20(11): 4192-4205. doi: 10.1109/TITS.2018.2883342.

[21] Serin F, Alisan Y, Kece A. Hybrid time series forecasting methods for travel time prediction. *Physica A: Statistical Mechanics and Its Applications*. 2021;579: 126134. doi: 10.1016/j.physa.2021.126134.

[22] Zhong S, Hu J, Ke S, Wang X. A hybrid model based on support vector machine for bus travel-time prediction. *Promet – Traffic&Transportation*. 2015;27(4): 291-300. doi: 10.7307/ptt.v27i4.1577.

[23] Yang M, Chen C, Wang L, Yang X. Bus arrival time prediction using support vector machine with genetic algorithm. *Neural Network World Journal*. 2016;26(3): 205-217. doi: 10.14311/NNW.2016.26.011.

[24] Peng Z, Jiang Y, Yang X, Zhao Z. Bus arrival time prediction based on PCA-GA-SVM. *Neural Network World Journal.* 2018;28(1): 87-104. doi: 10.14311/NNW.2018.28.005.

[25] Gal A, Mandelbaum A, Schnitzler F, Senderovich A. Traveling time prediction in scheduled transportation with journey segments. *Information Systems*. 2017;64: 266-280. doi: 10.1016/j.is.2015.12.001.

[26] Kumar BA, Vanajakshi L, Subramanian SC. Pattern-based time-discretized method for bus travel time prediction. *Journal of Transportation Engineering, Part A: Systems.* 2017;143(6): 04017012. doi: 10.1061/JTEPBS.0000029.

[27] Chien S, Kuchipudi MC. Dynamic travel time prediction with real-time and historic data. *Journal of Transportation Engineering.* 2003;129(6): 608-616. doi: 10.1061/(ASCE)0733-947X(2003)129:6(608).

[28] Jairam R, Kumar BA, Arkatkar SS, Vanajakshi L. Performance comparison of bus travel time prediction models across Indian Cities. *Transportation Research Record.* 2018;2672(31): 87-98. doi:10.1177/0361198118770175.

[29] Park D, Laurence RR. Forecasting freeway link travel times with a multilayer feedforward neural network. *Computer-Aided Civil and Infrastructure Engineering.* 1999;14(5): 357-367. doi: 10.1111/0885-9507.00154.

[30] Jeong R, Rilett LR. Prediction model of bus arrival time for real-time applications. *Transportation Research Record.* 2005;1927(1): 195-204. doi: 10.3141/1927-23.

[31] Pang J, Huang J, Du Y, Yu H. Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network. *IEEE Transactions on Intelligent Transportation Systems.* 2018;20(9): 3283-3293. doi: 10.1109/TITS.2018.2873747.

[32] Liu H, Xu H, Yu Y, Cai Z. Bus arrival time prediction based on LSTM and spatial-temporal feature vector. *IEEE Access.* 2020;8: 11917-11929. doi: 10.1109/ACCESS.2020.2965094.

[33] Irie K, Tüske Z, Alkhouli T, Schlüter R. LSTM, GRU, highway and a bit of attention: An empirical overview for language modeling in speech recognition. *Interspeech 2016*. 2016. p. 3519-3523. doi: 10.21437/Interspeech.2016-491.

[34] Zhai H, Cui L, Zhang W, Xu X. An improved deep spatial-temporal hybrid model for bus speed prediction. *Mathematical Problems in Engineering*. 2020(2): 1-11. doi: 10.1155/2020/2143921.

[35] Shen M, Xu Q, Wang K, Tu M. Short-term bus load forecasting method based on cnn-gru neural network. *Proceedings of Purple Mountain Forum 2019 - International Forum on Smart Grid Protection and Control.* Springer, Singapore; 2020. p. 711-722. doi: 10.1007/978-981-13-9783-7_58.

[36] Xue X, Jia Y, Wang S. Expressway traffic flow prediction

model based on Bi-LSTM neural networks. *IOP Conference Series: Earth and Environmental Science.* 2020;587(1): 012007. doi: 10.1088/1755-1315/587/1/012007.

[37] Shu W, Cai K, Xiong N. A short-term traffic flow prediction model based on an improved gate recurrent unit neural network. *IEEE Transactions on Intelligent Transportation Systems.* 2021;7: 1-12. doi: 10.1109/TITS.2021.3094659.

[38] Ding Y, et al. Interpretable spatio-temporal attention LSTM model for flood forecasting. *Neurocomputing.* 2020;403: 348-359. doi: 10.1016/j.neucom.2020.04.110.

[39] Zhao H, Zhang C. An online-learning-based evolutionary many-objective algorithm. *Information Sciences.* 2020;509: 1-21. doi: 10.1016/j.ins.2019.08.069.

[40] Dulebenets MA. An Adaptive Polyploid Memetic Algorithm for scheduling trucks at a cross-docking terminal. *Information Sciences.* 2021;565: 390-421. doi: 10.1016/j.ins.2021.02.039.

[41] Liu ZZ, Wang Y, Huang PQ. AnD: A many-objective evolutionary algorithm with angle-based selection and shift-based density estimation. *Information Sciences.* 2020;509: 400-419. doi: 10.1016/j.ins.2018.06.063.

[42] Pasha J, et al. An optimization model and solution algorithms for the vehicle routing problem with a "factory-in-a-box". *IEEE Access.* 2020;8: 134743-134763. doi: 10.1109/ACCESS.2020.3010176.

[43] D'Angelo G, Pilla R, Tascini C, Rampone S. A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees. *Soft Computing.* 2019;23(22): 11775-11791. doi: 10.1007/s00500-018-03729-y.

[44] Ergen T, Kozat S. Online training of LSTM networks in distributed systems for variable length data sequences. *IEEE Transactions on Neural Networks and Learning Systems.* 2017;99: 1-7. doi: 10.1109/TNNLS.2017.2770179.

[45] Davis RE. Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *Journal of Physical Oceanography.* 1976;6(3): 249-266. doi: 10.1175/1520-0485(1976)006<0249:POSSTA>2.0.CO;2.

[46] Kieu L, Bhaskar A, Chung E. Public transport travel-time variability definitions and monitoring. *Journal of Transportation Engineering.* 2015;141(1): 04014068. doi: 10.1061/(ASCE)TE.1943-5436.0000724.