Traffic Speed Prediction Method Based on Spatiotemporal Sampling and LSTM

Jiazhao Zhang [1, *], Yuanjian Zhang [2], Xinyun Gao [3]

[1] College of Transportation Engineering, Chang'an University, Xi'an, China, 710018

[2] Department of Civil Engineering, Hangzhou City University, Hangzhou, China, 310015

[3] BJTU school of traffic and transportation, Beijing Jiaotong University, Beijing, China, 100000

* Corresponding Author Email: jiazhaozhang@126.com

Abstract. Accurate prediction of traffic speed is crucial for traffic management and planning. In order to solve the problems of low prediction efficiency of previous traffic speed prediction models and easy neglect of spatiotemporal characteristics, a traffic speed prediction method based on spatiotemporal sampling and LSTM model is proposed based on the 24-hour driving dataset of 4,000 taxis in Shanghai, and draw speed heat maps in different regions at different times to visualize the spatiotemporal characteristics of traffic speed. Experimental results show that the model has high prediction accuracy and good expansion potential.

Keywords: Traffic speed prediction, Spatiotemporal data, long short-term neural network.

1. Introduction

One of the major challenges facing modern cities is traffic congestion. Urbanization acceleration and transportation networks continuous expansion have had a negative impact on urban transportation efficiency, the environment, and people's lives quality. Therefore, traffic management, route planning, and traffic flow optimization require accurate traffic speed prediction. However, predicting traffic speed accurately has always been a challenging problem due to complex factors affecting traffic speed such as traffic volume, road conditions, and time changes.

In the past few decades, machine learning and deep learning rapid development have made Recurrent Neural Networks (RNNs) and their variations important tools in the field of traffic prediction [1]. Among them, Long Short-Term Memory (LSTM), as a special RNN architecture, has been widely used in sequence modeling and time series prediction tasks [2]. The LSTM network is known for its ability to capture long-term dependencies in time series data by utilizing memory units and gate mechanisms. This unique architecture has demonstrated impressive performance in the field of traffic prediction. However, in previous studies, the spatial distribution characteristics of traffic speed are often not considered sufficiently. Therefore, this paper comprehensively considers the space-time relationship, uses the LSTM model to predict the urban traffic speed, and verifies the influence of space-time characteristics on the accuracy of the traffic speed prediction model.

The paper is structured as follows: Section 2 reviews research related to traffic speed prediction. Section 3 describes the construction and training process of the LSTM model in detail. Section 4 describe the experiment parameters setting. Section 5 presents the experimental results and analysis. Finally, the experimental results are summarized and discussed, and prospects for future research are put forward.

2. Literature review

Traffic prediction involves a range of classification methods for various prediction tasks, including traffic flow, speed, arrival time, and demand. It can also be classified into peak and off-peak period prediction, as well as prediction under abnormal conditions. These categorizations highlight the need for real-time, accurate, and reliable traffic prediction models. Models for traffic prediction can be broadly divided into two categories: parameter-based and non-parameter-based methods [3].

2.1. Parametric approaches

The parameter method, also referred to the model-based method, is an approach where the model structure is pre-determined based on certain theoretical assumptions. The model parameters are then calibrated using historical data or combined with new observations to make predictions [4]. Traffic simulation models and time series models are mainly included in the parameter method.

Traffic simulation models are mathematical models used to simulate and analyze the behavior of transportation systems. Future traffic levels are estimated by simulating traffic to verify the applicability of traffic system design without using historical and real-time traffic data [5]. For example, the three-phase traffic theory proposed by Kerner explains actual spatiotemporal traffic patterns [6]. Although these theoretical models have helped to some extent in understanding traffic problems, they are often inefficient in analyzing large-scale transportation systems with massive real-time data [7].

2.2. Non-parametric approaches

Non-parametric methods have a flexible model structure and parameters that are not fixed. The most popular models for predicting traffic parameters are classical statistical models and artificial intelligence models [8]. The following are the more commonly used and classic methods for predicting traffic parameters.

2.2.1 Prediction methods based on statistical theory

Prediction methods for traffic flow based on statistical theory mainly include moving average model, time series model, and Kalman filtering model [9].

The historical average model is a basic and straightforward approach used in traffic prediction. It does not require complex data processing or model training and can quickly obtain results. The processing of missing values is also relatively simple, and missing values can be directly filled with historical average values. However, the historical average model ignores the correlation between data and cannot discover the connection between data. Therefore, the accuracy of the prediction results is not high. The time series model is a widely used and easily understandable approach in traffic prediction. It leverages historical data to capture patterns and trends, offering fast computations and accurate predictions. However, it primarily focuses on analyzing a single time series and may not account for complex relationships between variables. The choice of model constants is critical, and the model is best suited for short-term forecasting. To enhance its performance, combining the time series model with other approaches can provide more comprehensive insights and improved predictions.

Kalman filtering is an intelligent method that combines observed data and estimated data by iteratively managing errors in a closed loop, ensuring that errors remain within a certain range. However, its effectiveness is limited to linear process models and measurement models, and it falls short in achieving optimal estimation results in nonlinear scenarios.

2.2.2 Artificial Neural Network

Artificial neural networks are information processing systems designed to mimic the structure and function of the human brain. Neural network approaches allow for samples with significant missing or distorted data. There are various types of neural networks, and different neural network models can be considered based on the characteristics of the research subject [12]. The advent of deep neural network models in recent years has greatly advanced the application of artificial intelligence methods in traffic prediction. Deep learning-based neural network models have demonstrated their ability to capture the complex dynamics inherent in traffic data, yielding promising results in the field of traffic prediction. As a result, they have garnered significant research interest and attention.

In 2014, Huang et al. proposed a network architecture based on the new generation of artificial neural networks called Deep Belief Networks (DBN) and regression models for traffic flow prediction [13]. The efficacy of the model in capturing the stochastic properties of traffic data and enhancing

the accuracy of traffic prediction was verified through validation on multiple datasets. In a similar vein, Lv et al. proposed the utilization of a Stacked Autoencoder (SAE) model to effectively extract the essential features from traffic data, resulting in successful short-term traffic flow prediction [14]. However, this model has drawbacks such as a simple network structure, slow convergence speed, and susceptibility to overfitting.

To tackle the challenge of capturing temporal dependencies, Recurrent Neural Networks (RNN) and their variations, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have gained significant popularity in traffic prediction models. Wu et al. introduced a feature fusion architecture for short-term traffic flow prediction. This architecture combines Convolutional Neural Networks (CNN) with LSTM, enabling the consideration of both temporal and spatial dependencies in traffic flow data [15]. Other research studies have explored the utilization of advanced models such as Convolutional Fusion Long Short-Term Memory Network (FCL-Net) and Spatial-Temporal Relationship Convolutional Network (SRCN) to enhance the accuracy and effectiveness of traffic flow prediction [16]. However, the CNN introduced in existing methods is primarily suitable for images and regular grids in Euclidean space, making it challenging to adequately represent the spatial dependencies of complex topological structures in traffic networks.

In summary, significant progress has been made in research on LSTM for predicting traffic speed in recent years. However, existing research still has the following shortcomings in predicting traffic speed using LSTM:

(1) Lack of spatiotemporal correlation when data is lacking: Many existing methods only consider historical speed data and ignore spatiotemporal correlation of traffic speed. This makes it difficult for models to fully capture dynamic features and complexity of traffic flow.

(2) Limited prediction accuracy: Large errors are observed in some existing methods when predicting traffic speed. This may be because the model does not handle complex situations such as peaks and congestion in traffic flow properly, resulting in a decrease in prediction accuracy.

(3) Complex data preprocessing: Many studies require complex data preprocessing such as filling missing values, data smoothing and normalization. This increases the complexity and computational cost of research.

To address these issues, this study uses spatiotemporal sampling data and LSTM for traffic speed prediction:

(1) Spatiotemporal correlation is fully considered by combining temporal data with spatial features so that the model can better capture spatiotemporal dynamic features of traffic flow.

(2) Based on GPS data from Shanghai taxis, this study aims to use LSTM models to predict traffic flow in different regions and display predictive results by drawing speed heat maps. The study focuses on rush hour periods in the morning peak period, evening peak period, and two flat periods to reveal spatiotemporal changes in traffic flow.

3. Methodology

A traffic speed prediction method based on spatiotemporal sampling and LSTM is established. The overall framework is divided into three modules: data preprocessing module, model training module, and model evaluation and prediction module, as shown in Figure 1.

3.1. Data preprocessing module

3.1.1 Data cleaning

The purpose of data cleaning is to ensure the quality, accuracy, and consistency of the data, thereby providing a reliable data foundation to support subsequent analysis, modeling, and prediction processes. The data source of this article is GPS data of Shanghai taxis. The cleaning steps include removing data that is not in the research area, deleting useless data such as azimuth angle and passenger status fields from the source dataset. The cleaning steps are as follows:

(1) Data that is not in the research area is removed by defining the research scope of Shanghai and using the TransBigData package in Python to remove data outside the research scope. The longitude range of the research area is (120.8666, 122.2), and the latitude range is (30.6667, 31.8833).

(2) Useless data such as azimuth angle and passenger status fields are deleted from the source dataset.

(3) Abnormal data was deleted using the box plot method shown in Figure 2. The box plot method was used to identify potential outliers by drawing quantiles and quartile ranges of data. Data whose speed value had been 0 for more than 15 minutes was also deleted as it was considered that the vehicle did not belong to the driving state on the road.

After data cleaning, about 20% of the original data was deleted. This ensured the integrity of the data information while improving the quality of the data.

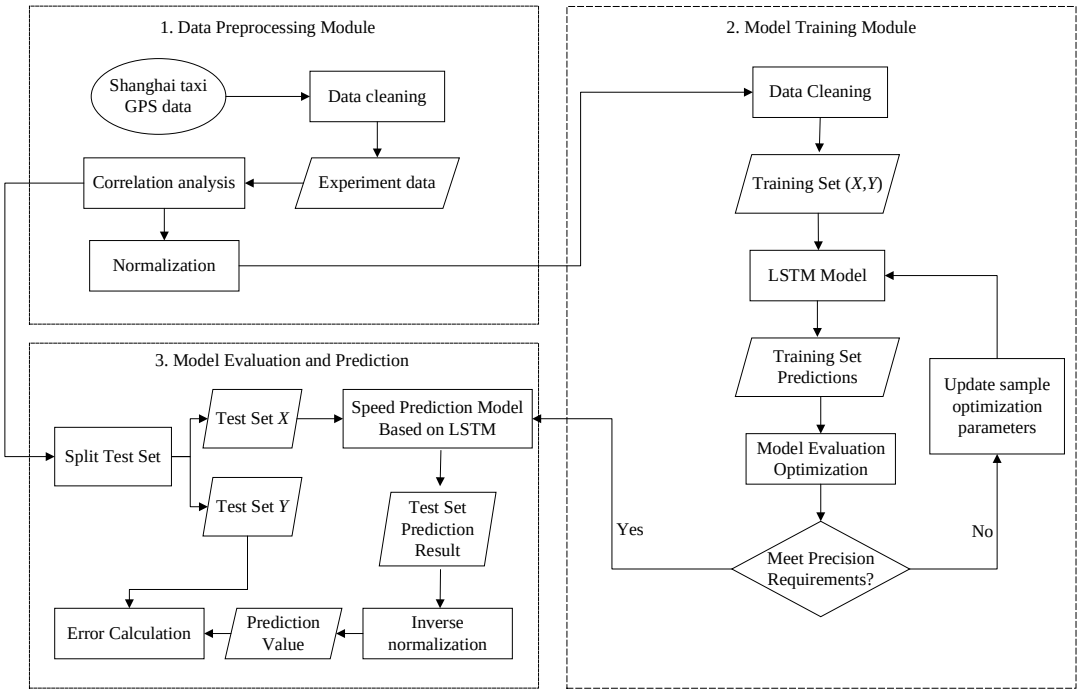


Figure. 1 Article Framework

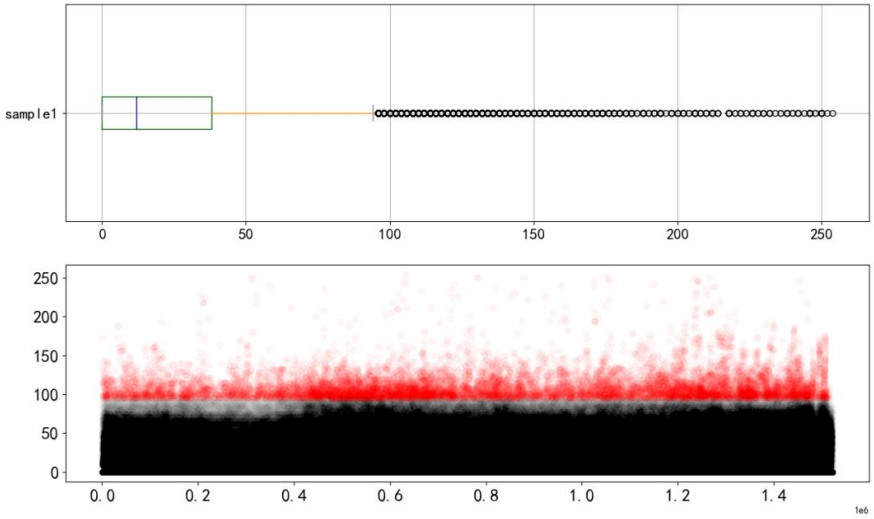


Figure. 2 Box plot method for outlier detection

3.1.2 Normalization processing

Data normalization is essential to eliminate the influence of varying data magnitudes on prediction results. It standardizes the values and accounts for speed differences among different regions,

enabling accurate analysis and pattern recognition in prediction models. The calculation formula for normalization is as follows:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Where: x' is the standardized value after processing.

x is the value before standardization;

$x_{\max}$. $x_{\min}$ are the maximum value and minimum value in the data.

3.1.3 Grid division

To facilitate geographical spatial calculations, it is necessary to convert the map into a grid format. Since Shanghai's longitude range falls between (120.8666, 122.2) and its latitude range falls between (30.6667, 31.8833), the geographical area of Shanghai was divided into a grid system. Each grid was assigned a unique number to simplify the calculation process.

3.2. Model training module

The LSTM unit mainly controls the flow of input and output data through forget gate, update gate and output gate to achieve information protection and control [17]. Its internal structure is shown in Figure 3. The forget gate determines what information will be discarded from the cell state, update gate determines what information needs to be updated in cell state, output gate determines what cell state needs to be output. Where *c. a. x. y* respectively represents long-term memory, activation value, input value, output value.

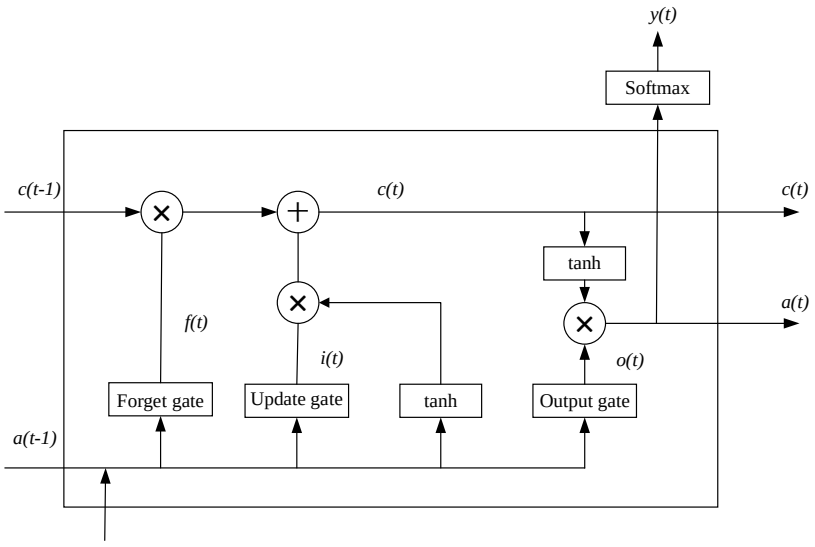


Figure. 3 LSTM Model

As a special RNN network structure, LSTM network overcomes the weakness of RNN network in long-term memory. The input vector sequence *x* is mapped to hidden vector sequence *h* by each long-short term memory unit through *T* iterations. As shown in equation below. i_t, f_t , o_t, c_t (t = 1,2,...,T) respectively represent input gate, forget gate, output gate, memory cell vector, and have same dimension with h_t .

$$f_t = \delta(\omega_f[h_{t-1}, x_t] + b_f) \tag{2}$$

$$i_t = \delta(\omega_i[h_{t-1}, x_t] + b_i) \tag{3}$$

$$c_t = f_t c_{t-1} + i_t(\tanh(\omega_c[h_{t-1}, x_t] + b_c)) \tag{4}$$

$$h_t = o_t \tanh(c_t) \tag{5}$$

$$o_t = \delta(\omega_0[h_{t-1}, x_t] + b_0) \tag{6}$$

Where δ represents sigmoid function; c_t, c_{t-1} represents memory cell state; h_{t-1}, x_t represents hidden state at time *t*-1; t represents time step input; $[h_{t-1}, x_t]$ represents concatenation between LSTM network *t*-1 hidden layer state h_{t-1} and x_t; ω_f. ω_i. ω_c. ω_0 are self-learning weight matrices of network; b_f. b_i. b_c. b_0 are bias terms of network.

3.3. Model Prediction and Evaluation Module

3.3.1 Model Prediction

(1) Data preparation: The data includes 100 grid files, and each file is a matrix with dimensions of $m \times n$, where *m* represents the amount of predicted data, and *n* represents the dimension of the data, including the time dimension and the speed dimension.

(2) Prediction: The training model is utilized to predict traffic speed values based on the provided input data, which subsequently generates a matrix.

(3) Inverse normalization: The output matrix is inverse normalized by the model to obtain predicted values.

3.3.2 Model Evaluation

This article uses RMSE (root mean square error), MAE (mean absolute error), and MAPE (mean absolute percentage error) to evaluate model prediction accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{8}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{9}$$

Among them, *n* represents the number of data in the test dataset, $\hat{y}_i$ represents predicted traffic speed values, and y_i represents actual traffic speed values.

4. Experimental Settings

4.1. Data preprocessing

The dataset used in this article comes from a public dataset of 4315 taxi GPSs in Shanghai on February 20th, 2007. The vehicle driving data has a sampling time interval of 1 minute and includes vehicle ID, sampling time, latitude and longitude, speed, azimuth angle and passenger status. The original data format is shown in Table 1.

Firstly, the data that falls outside the geographical scope of Shanghai, as shown in Figure 4, should be excluded. Additionally, any irrelevant information such as azimuth angle and passenger status should be removed to focus on the research objectives. The time data format should be converted into minutes, and the entire day's data collection time should be divided into time periods of 5 minutes each, resulting in 288 time periods. The research area should be divided into a grid network based on latitude and longitude, as depicted in Figure 5. The real-time location of taxis should be matched to the corresponding grids, and data with the same grid position and time period should be merged. The average speed should be calculated for each grid. All vehicles within the same grid should be

appended to a new table file, sorted according to the time period, and ultimately resulting in average speed data for the 288 time periods corresponding to each grid. With this data, further prediction can be conducted.

Table. 1 Original Data Format Description

Field	Unit	Description
Vehicle ID	—	Vehicle ID
Time	min	Time of data recording
Longitude	Degree	Longitude of vehicle
Latitude	Degree	Latitude of vehicle
Speed	km/h	Speed of vehicle
Angle	degree	Direction of vehicle travel
Passenger carrying status	—	Binary variable, '1' represents carrying passengers, '2' represents empty



Figure. 4 Study area

4.2. Model Parameter Setting

The parameter settings of this model are as follows:

(1) The training set and test set are divided in a time series ratio of 8:2.

(2) The Adam optimizer algorithm is used for gradient descent.

(3) The batch size is generally set at values such as 16, 32, 64, 128 or 256. If the value is too small, computing resources cannot be fully utilized; if it's too large it reduces update weight and bias speed. In this paper, we selected a batch size of 32.

(4) The loss function used here is mean square error (MSE).

(5) ReLU activation function was used.

(6) Time step was set at one.

(7) Epoch training number was set at fifty.

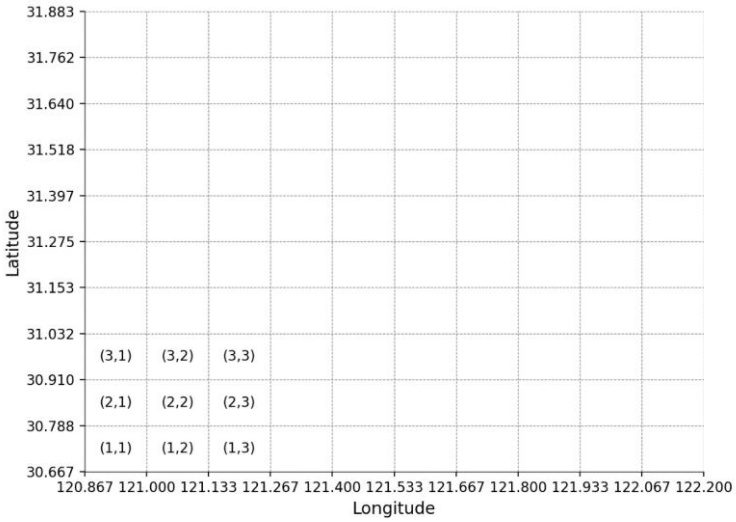
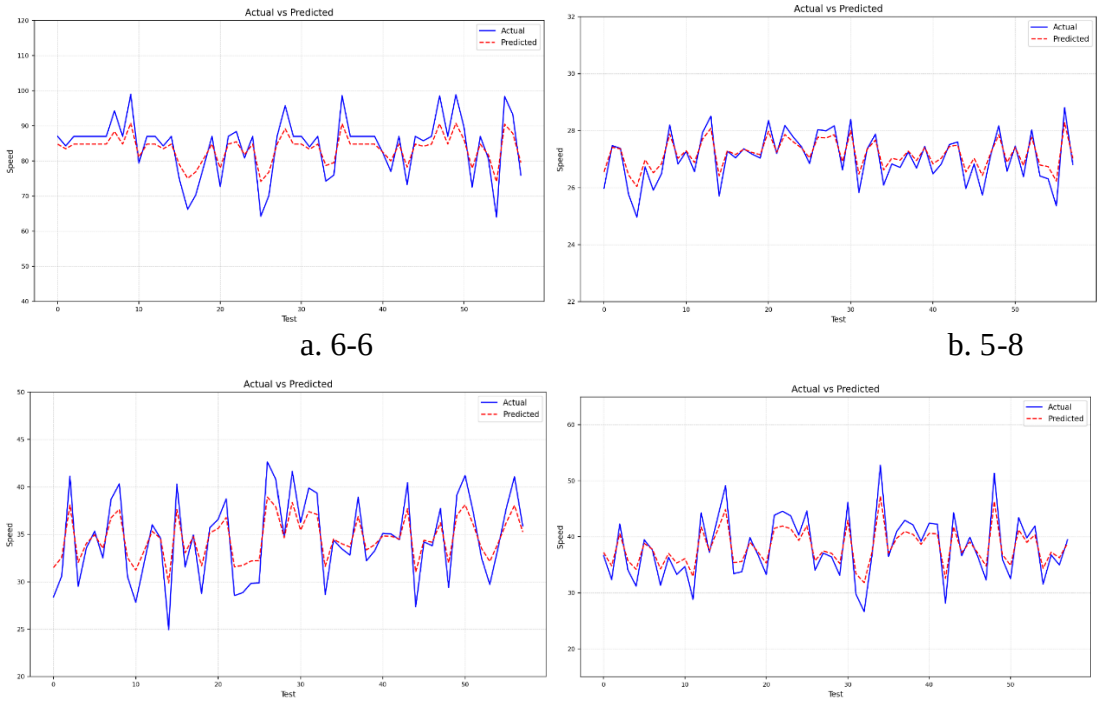


Figure. 5 Rasterized rendering

5. Result analysis

The speed of Shanghai traffic was predicted using the established LSTM model. Four grids were selected for analysis: the 3rd row and 7th column (3-7), the 5th row and 8th column (5-8), the 6th row and 6th column (6-6), and the 7th row and 8th column (7-8). Among them, 3-7 and 7-8 represent the peripheral areas of Shanghai, while 5-8 and 6-6 represent the central areas of Shanghai. The prediction results of the four regions are displayed as shown in Figure 6, and the curves of the loss function changes with the number of iterations are obtained as shown in Figure 7.



a. 6-6 b. 5-8

c. 7-8 d.3-7

Figure. 6 Actual and predicted lines

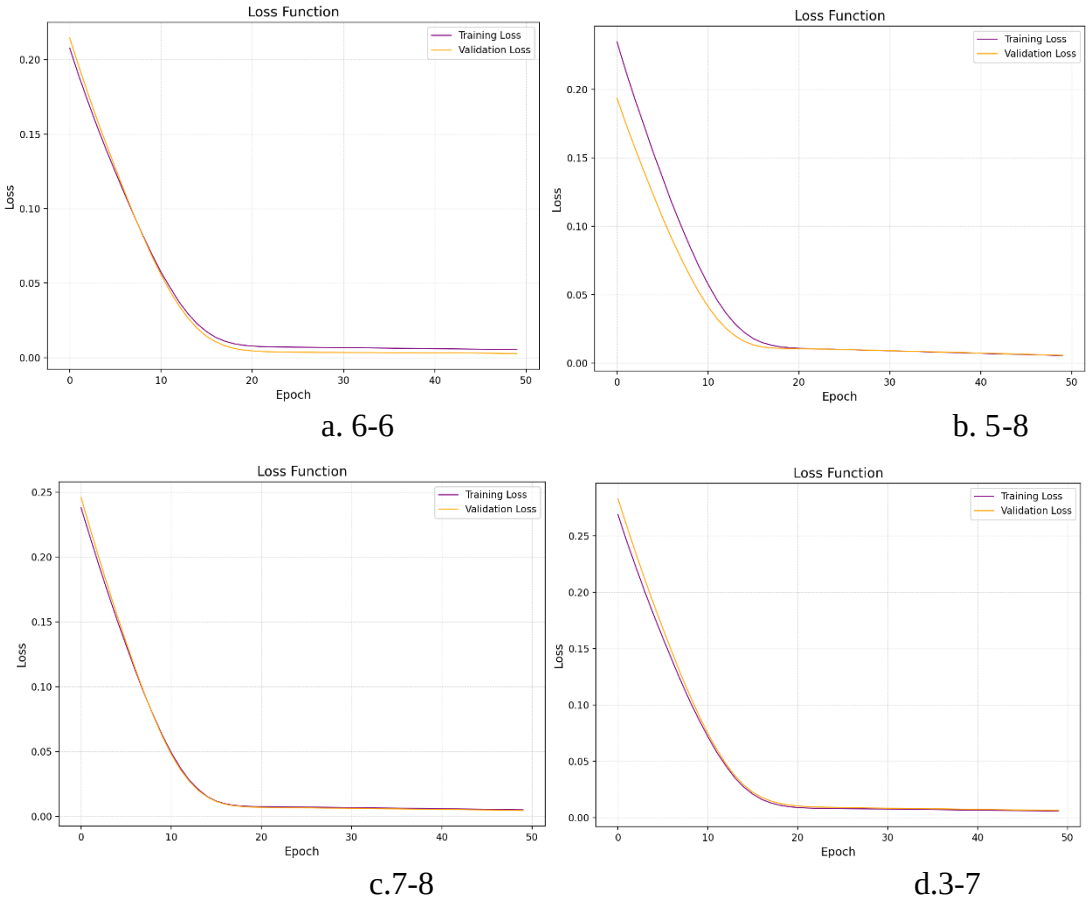


Figure. 7 Loss function

From the prediction curve, it is evident that the predicted speed closely matches the actual speed, regardless of the analyzed area. This observation suggests that the data processing was performed accurately, and the model exhibits a high level of prediction accuracy.

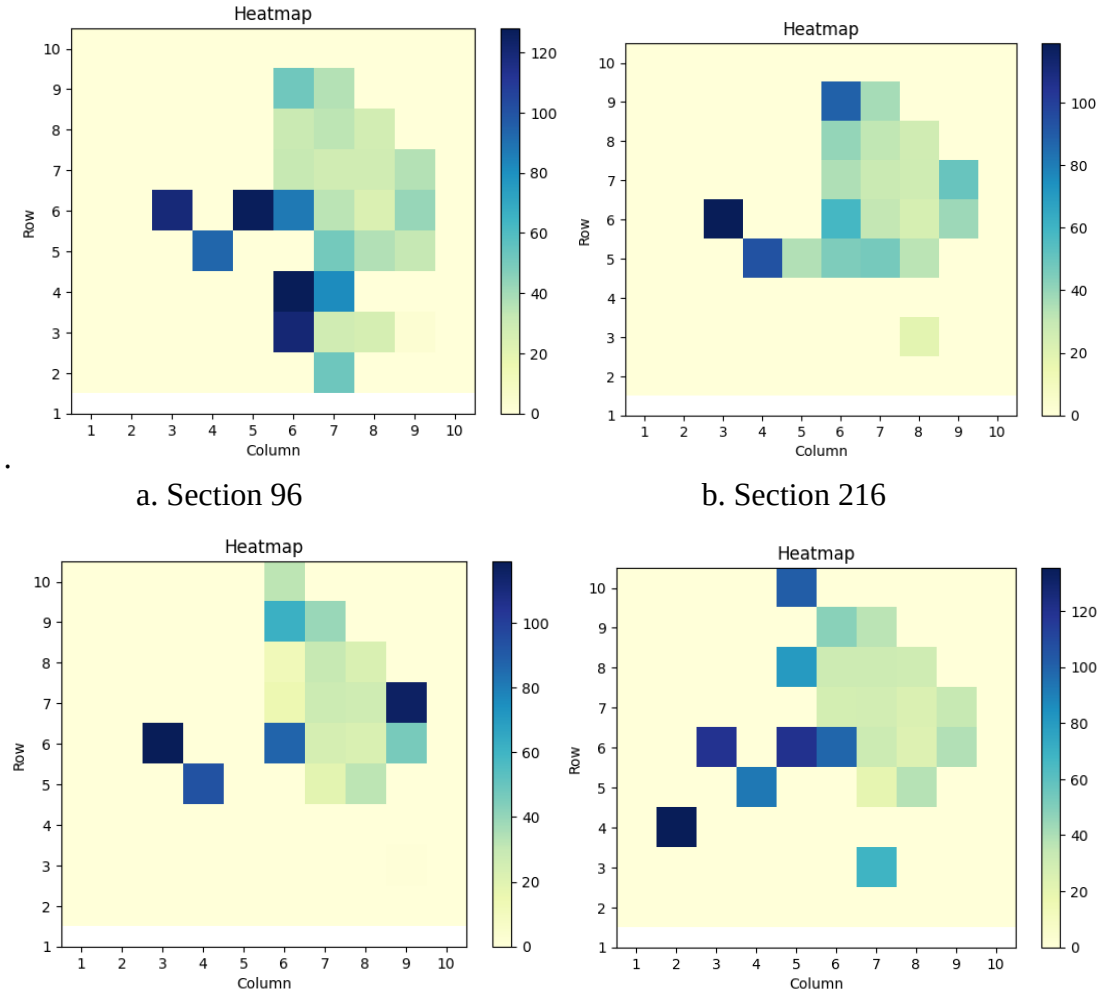
The low loss values obtained on both the training set and validation set indicate that the model has effectively learned from the data and can generalize well to unseen data. This suggests that the model is capable of capturing the underlying patterns and relationships in the training data, allowing it to make accurate predictions on new and unseen data.

According to the fitting results, evaluation indicators RMSE, MAE, and MAP were used to analyze speed distribution in different regions of Shanghai at different time periods, which are shown in Table 2. It can be seen that all indicators of location 5-8 are the smallest, while the RMSE and MAE of location 6-6 and the MAPE of location 7-8 are the biggest.

Table. 2 Model comparison result

Location	RMSE	MAE(%)	MAPE(%)
6-6	4.372	3.5911	4.391
5-8	0.374	0.295	1.106
7-8	2.120	1.755	5.255
3-7	2.307	1.904	5.149

Four time periods were selected for analysis: morning peak (8 am), evening peak (6 pm), flat peak (6 am and 3 pm) corresponding to Section 96, Section 216, Section 72, and Section180 respectively. Speed prediction was carried out and results were visualized to obtain a regional heat map as shown in Figure 8.



a. Section 96 b. Section 216

c. Section 72 d. Section 180

Figure. 8 Heat map of speed from different time period

It is important to note that an area with a value of 0 indicates incomplete data and cannot be predicted accurately. During peak hours, the speed of vehicles in the city center tends to be lower compared to the outskirts. In the early morning, the speed of vehicles in the city center shows a slight increase, while the speed on the outskirts remains relatively stable. These variations in speed across different areas provide valuable insights for efficient vehicle travel planning. By taking into account these differences, it is possible to guide people to travel more efficiently and optimize their routes accordingly.

6. Conclusion

The study used an LSTM model to predict traffic speed based on Shanghai taxi GPS data. Through experimental results and analysis, the following conclusions can be obtained:

(1) By comparing actual speed and predicted speed curves, it can be observed that the LSTM model can accurately capture dynamic characteristics of traffic data and has high prediction accuracy. This indicates that the LSTM model has good performance and potential in traffic prediction tasks.

(2) The LSTM model used in this study is a typical application of deep learning methods. It can fully utilize temporal characteristics and long-term dependencies of data. Compared with traditional methods, it can better capture regularities and trends of traffic data and improve prediction accuracy.

(3) Traffic flow was predicted in different regions and corresponding speed heat maps were drawn. Experimental results show that traffic speeds can be effectively predicted in different regions during morning peak, evening peak and two flat peaks, providing valuable information for traffic management and planning.

Although the research has achieved certain results, there are also some limitations. For example, different traffic datasets may affect prediction results. In addition, model structure and parameter selection can be further optimized and other deep learning models can be explored to improve prediction accuracy and effect.

In future research, it is suggested to explore how different levels of data aggregation affect the prediction performance. Another interesting avenue for investigation is to deepen the architecture of LSTM neural networks by adding multiple layers, which has the potential to enhance their learning capabilities.

References

[1] Madan R, Mangipudi P S. Predicting computer network traffic: a time series forecasting approach using DWT, ARIMA and RNN[C]//2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, 2018: 1-5.

[2] Shao Y, Zhao Y, Yu F, et al. The traffic flow prediction method using the incremental learning-based CNN-LTSM model: the solution of mobile application [J]. Mobile Information Systems, 2021, 2021: 1-16.

[3] Van Lint J W C, Van Hinsbergen C. Short-term traffic and travel time prediction models [J]. Artificial Intelligence Applications to Critical Transportation Issues, 2012, 22(1): 22-41.

[4] Khajeh Hosseini M, Talebpour A, Traffic Prediction using Time-Space Diagram: A Convolutional Neural Network Approach. Transport Res Rec 2019;2673:425-435.

[5] Nagy AM, Simon V, Survey on traffic prediction in smart cities. Pervasive Mob Comput 2018;50:148-163.

[6] Davis L C. Introduction to modern traffic flow theory and control: the long road to three-phase traffic theory[J]. Physics Today, 2010, 63(3): 53-53.

[7] Ma X, Tao Z, Wang Y, Yu H, Wang Y, Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies 2015;54:187-197.

[8] Akhtar M, Moridpour S. A review of traffic congestion prediction using artificial intelligence [J]. Journal of Advanced Transportation, 2021, 2021: 1-18.

[9] Mai T, Ghosh B, Wilson S. Short-term traffic-flow forecasting with auto-regressive moving average models[C]//Proceedings of the Institution of Civil Engineers-Transport. Thomas Telford Ltd, 2014, 167(4): 232-239.

[10] Yoo S R, Jeong J S, Jong J Y, et al. Forecast of marine traffic volume using time series model[C]//2013 International Conference on Fuzzy Theory and Its Applications (iFUZZY). IEEE, 2013: 122-126.

[11] Chow A H F, Li S, Szeto W Y, et al. Modelling urban traffic dynamics based upon the variational formulation of kinematic waves [J]. Transportmetrica B: Transport Dynamics, 2015, 3(3): 169-191.

[12] Karlaftis M G, Vlahogianni E I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights [J]. Transportation Research Part C: Emerging Technologies, 2011, 19(3): 387-399.

[13] Huang W, Song G, Hong H, et al. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks with Multitask Learning [J]. Intelligent Transportation Systems, IEEE Transactions on, 2014, 15(5): 2191-2201.

[14] Lv Y, Duan Y, Kang W, et al. Traffic Flow Prediction with Big Data: A Deep Learning Approach [J]. Intelligent Transportation Systems, IEEE Transactions on, 2015, 16(2): 865-873.

[15] Wu Y, Tan H. Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework [J]. Transportation Research Part C: Emerging Technologies, 2018, 90:166-180.

[16] Ke J, Zheng H, Yang H, et al. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach [J]. Transportation Research, 2017, 85:591-608.

[17] Zhong Chenhao, Wang Meng, Zhao Wei, et al. Short term traffic flow prediction at intersections based on LSTM Highway transportation technology, 2019, 35 (5): 119 – 123, 131.