Temporal data-driven short-term traffic prediction: Application and analysis of LSTM model

Xuange Huang [1, 4], Xinran Li [2]and Wenshuo Wang [3]

[1]School of Mathematical Sciences, Anhui University, Hefei, China
[2]School of Software Technology, Dalian University of Technology, Dalian, China
[3]School of Future Tech, South China University of Technology, Guangzhou, China

[4]hgpenguin@126.com

Abstract. With the development of the economy, the number of private cars has increased, and traffic congestion is becoming increasingly common. In order to prevent traffic congestion, short-term prediction of traffic flow is urgent. This article is based on the data of a Hangzhou elevated bridge from October 4th to 18th, 2015, with a timestamp of 5 minutes and a total of over 4000 pieces of data, to have the Long Short-Term Memory (LSTM) model to be trained and the following day's traffic flow to be predicted. The outcomes show that the model has high predictive capability and can reliably forecast future short-term transportation passenger flow, accurately reflecting the trend of data changes. Compared with the Auto-Regressive Moving Average (ARIMA) method with a Mean-Square Error (MSE) of 36, the LSTM model has a MSE of 22, which indicates a smaller MSE, indicating a more accurate prediction performance of LSTM. The LSTM model's prediction is beneficial for alleviating traffic congestion, providing great convenience for people's daily travel, and greatly reducing their travel time.

Keywords: LSTM, Traffic Flow, Short-Term Prediction, ARIMA.

1. Introduction

Short-term traffic prediction plays a crucial role in transportation planning and management. Accurate and reliable predictions of passenger flow patterns are essential for optimizing resource allocation, improving operational efficiency, and enhancing overall transportation systems. Traditional methods relying on field surveys and historical data analysis may not fully capture the dynamic and complex nature of passenger behavior. Therefore, there is a growing need for a range of advanced techniques to make more accurate and insightful short-term traffic predictions.

Nowadays, scholars nowadays enjoy using deep learning methods to achieve short-term traffic prediction. Traffic flow is influenced by many factors, deep learning has good applications in nonlinear functions. Consequently, deep learning can handle the complex characteristics of time series data [1]. Another approach is the use of ensemble models, such as Autoregressive Integrated Moving Average (ARIMA). Compared with neural network methods, the ARIMA method has a simpler principle. And since ARIMA adapts to most time series problems, it is also the most commonly used method [2]. Moreover, some researchers have explored the application of hybrid models, such as the combination of Markov Chain and Artificial Neural Network (ANN). The use of ANN has the following advantages, it can reset the relationship between input and output without excessive dependence on inputs [3].

Furthermore, approaches for time series analysis, such as Seasonal Autoregressive Integrated Moving Average (SARIMA) has been successfully employed to capture the seasonal patterns and trends in transportation data, leading to improved forecasting accuracy [4]. Other models such as Genetic Algorithm (GA) and Bayesian Networks (BN), which have also been explored in the literature, each offering unique advantages in the field of short-term transportation passenger flow prediction [5, 6].

Apart from the excellent methods above, LSTM has the advantages of good prediction performance and flexibility. The application of LSTM model in short-term traffic flow prediction effectively captures and models the complex temporal dependencies and patterns in transportation data, providing accurate and reliable predictions for efficient transportation planning and management. Thus, it outperforms the models above [7]. LSTM can leverage its memory cells and gating mechanism that selectively retain or forget information. In this case, it is able to capture and remember long-term patterns and dynamic changes in traffic flow. By training on historical traffic data, the model can learn the complex relationships between different variables and make predictions for future passenger flow [8]. It offers practical application value and decision support for selecting and determining the optimal intersection improvement scheme [9].

This paper applies LSTM model to short-term traffic prediction. Based on the predicted results, the government can formulate more rational transportation planning and policies to meet the urban transportation needs. Also, by anticipating potential congestion situations, the government can proactively implement measures to guide traffic flow, thereby improving traffic management efficiency.

2. Method

2.1. Data source

The dataset which comes from Kaggle incorporates data that were collected from April 4, 2015, to January 3, 2016 (https://doi.org/10.1016/j.neucom.2018.08.067). The traffic data was collected from multiple scattered locations on a viaduct road in Hangzhou. To be more specific, the dataset uses detectors to collect data every five minutes in the districts of Shangcheng, Gongshu, Xihu, and Binjiang [10]. The dataset includes attributes of license plate number, passage time, vehicle speed, specific time and traffic flow, which can well reflect the overall traffic operations. This dataset, with its time-intensive data characteristics, is well-suited for training models that prioritize accuracy, especially in the condition where this paper use it to train LSTM model. Moreover, the collection of various attributes by the detectors also ensures the accuracy of key data such as traffic count. The data set is shown in Table 1.

Table 1. Sample Dataset.

timestamp	traffic_count
2015/10/4 0:00:00	7
2015/10/4 0:05:00	6
2015/10/4 0:10:00	10
2015/10/4 0:15:00	4
2015/10/4 0:20:00	7
2015/10/4 0:25:00	3

2.2. Indicator selection and description

The dataset was originally collected for the study which was aimed to investigate the performance of the LSTM model in the case of losing certain parts of data. As a result, a few months of traffic flow data were missing in the dataset. Therefore, this paper selected the data from October 4th to 18th. After the selection, this paper obtained the timestamp and traffic count variables. Among the data of several roads, this paper chose one district data to train LSTM modal.

This paper estimated the MSE and RMSE to assess the predictive ability of the LSTM model. The average of the squared deviations between the predicted values and the actual values is used to calculate MSE, measuring the average error of the predictions. RMSE, on the other hand, is the square root of MSE, corresponding to the average deviation of the prediction errors. The MSE and RMSE numbers are both non-negative, and the closer they are to zero, the more accurate the prediction model is. These are the MSE and RMSE formulae.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_\iota}\right)^2 \tag{1}$$

$$RMSE = \sqrt{MSE} \tag{2}$$

Where Y_i is the real value and $\widehat{Y_\iota}$ is the predicted value.

2.3. Method introduction

LSTM model is a specialized type of recurrent neural network designed for processing and predicting time series data. It effectively addresses the limitations of traditional RNNs by modeling long-term dependencies and avoiding the problems of gradient vanishing and explosion. Particularly suitable for traffic flow forecasting, LSTM can capture intricate relationships, extract multiple patterns, and handle multivariate inputs. It adapts to various time scales and excels in forecasting at different time steps. Compared to the ARIMA model, the LSTM model offers several advantages for short-term passenger flow forecasting. Most importantly, LSTM model is capable of capturing complex temporal dependencies, allowing it to effectively model nonlinear patterns in passenger flow data. This enables better prediction accuracy in scenarios where traditional linear models like ARIMA may fall short.

The LSTM cell structure is shown in Figure 1. From Figure 1 people can tell that LSTM cells feature a unique structure with input, forget, and output gates, allowing them to effectively capture long-term dependencies and handle vanishing and exploding gradients. This enables accurate predictions and better modeling of complex temporal patterns, making LSTM suitable for various sequential data analysis tasks, which suits short-term traffic prediction well.

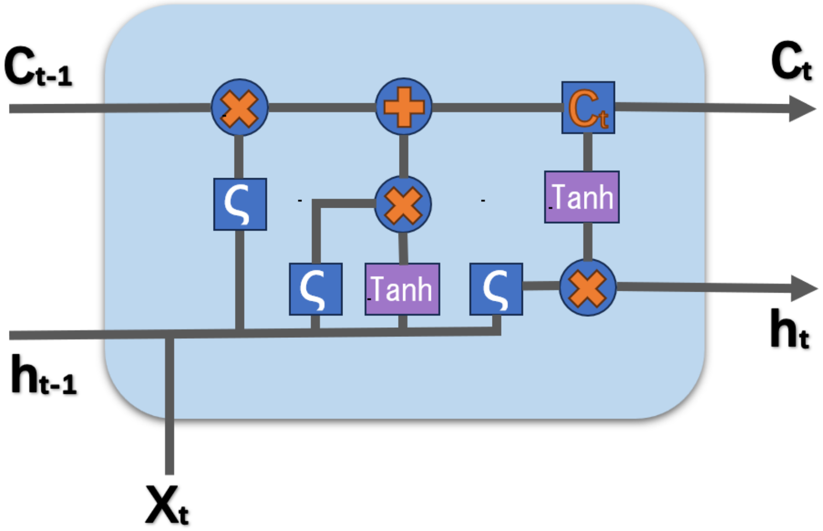

Figure 1. LSTM cell.

3. Results and Discussion

3.1. Descriptive analysis

This dataset records the traffic volume every five minutes from October 4th to October 18th, with the training set for the initial 14 days (4 through 17) and the test set for the 18th data. Figure 2 shows the changes in all traffic volumes. The graph has a certain periodicity.

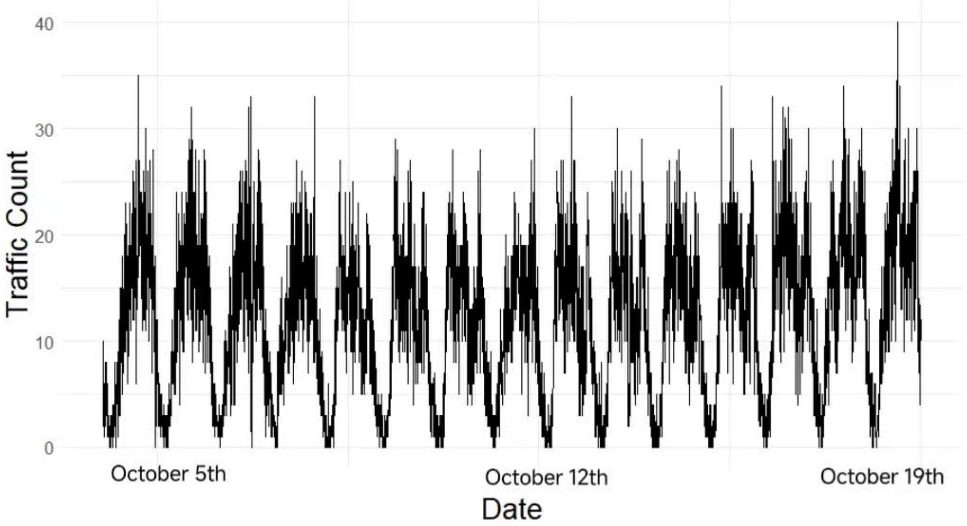

Figure 2. Time dependent graph of traffic flow on this road.

Randomly extract a day's worth of data from the dataset, as shown in Figure 3. The intensity of the hue indicates the volume of the passenger traffic, with darker colors indicating higher volumes. From the images, it is evident that the morning hours have the lowest passenger traffic, almost zero. Starting from 6am, the passenger flow gradually increases, slightly decreases after 8am, and then continues to increase until 12am. The passenger flow slightly decreased from 12:00 to 13:00, and continued to increase from 13:00 onwards. Afterwards, the traffic remained high until it gradually decreased at 22:00. From this data, it can be seen that it is consistent with the actual situation of people's daily lives.

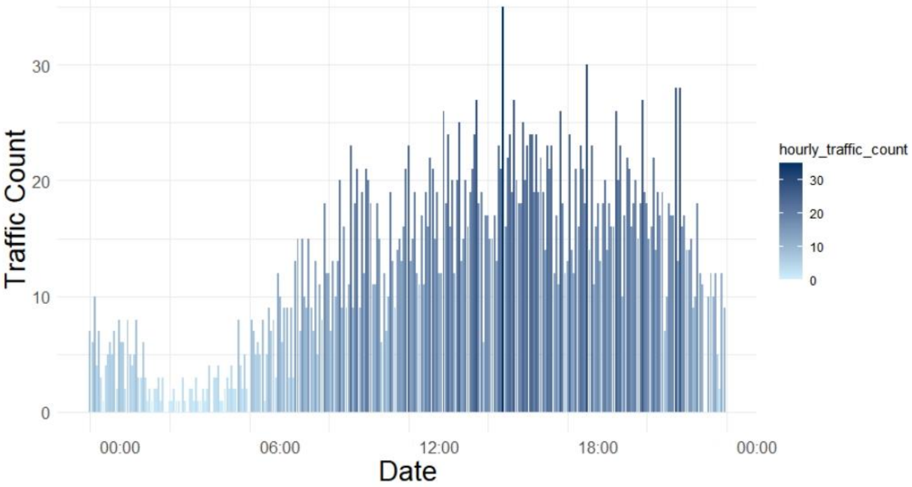

Figure 3. Data flow changes on a certain day.

After analyzing the data, the model was established using LSTM. The model's MSE and RMSE in the training set are 16.4455 and 4.0553, respectively. The model's MSE and RMSE in the test set are 22.1633 and 4.7078, respectively. Overall, the model has good performance indicators.

The resulting plot, as shown in Figure 4, depicts the data. To clarify, the area to the left of the vertical line represents the training set, while the area to the right represents the testing set. The original data is depicted as gray background points, the light blue represents the fitted values of the training data, and the dark blue represents the model's predictions on the testing set. It can be observed that the model's predicted points closely follow the fluctuations in the original data.

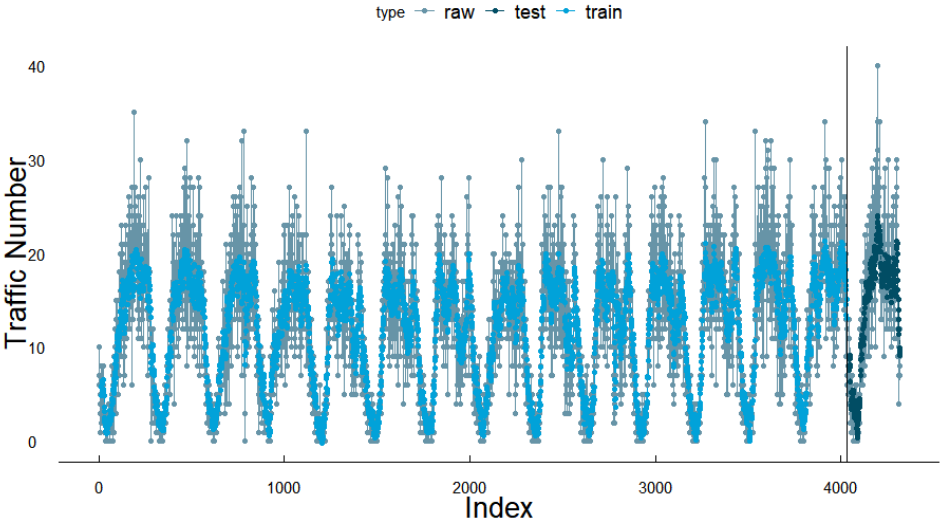

Figure 4. Fitted values of the training data, the predicted values of the testing data, and the original data.

Based on this model, predict the traffic flow every five minutes for the next day, October 19th. Figure 5 displays a comparison between the projected outcomes and the initial data. Which illustrates the prediction effect of this model is relatively good, and it can accurately predict the trend of traffic changes.

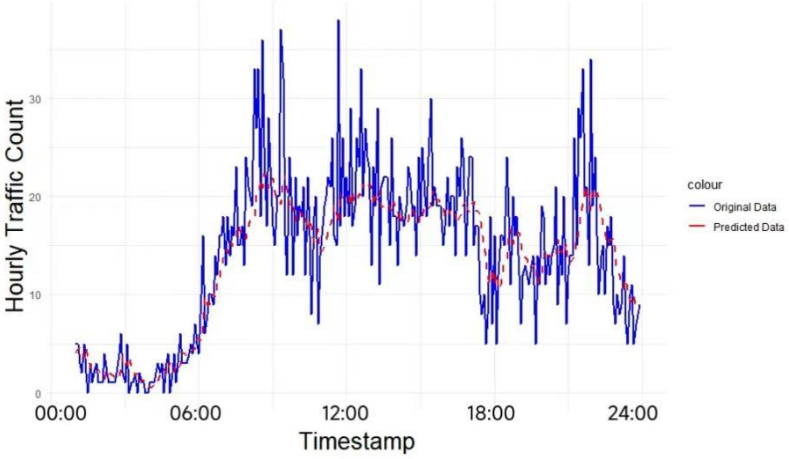

Figure 5. Comparison of raw data and predicted data every five minutes on October 19th.

3.2. Inferential analysis

ARIMA is a famous time series forecasting method proposed in 1970s. The fitting effect of ARIMA model and LSTM model is compared to verify whether the LSTM model is more effective than the traditional method. In this study, the final day serves as the test set while the prior two weeks serve as the training set. The training set's MSE and test set's RMSE are calculated separately.

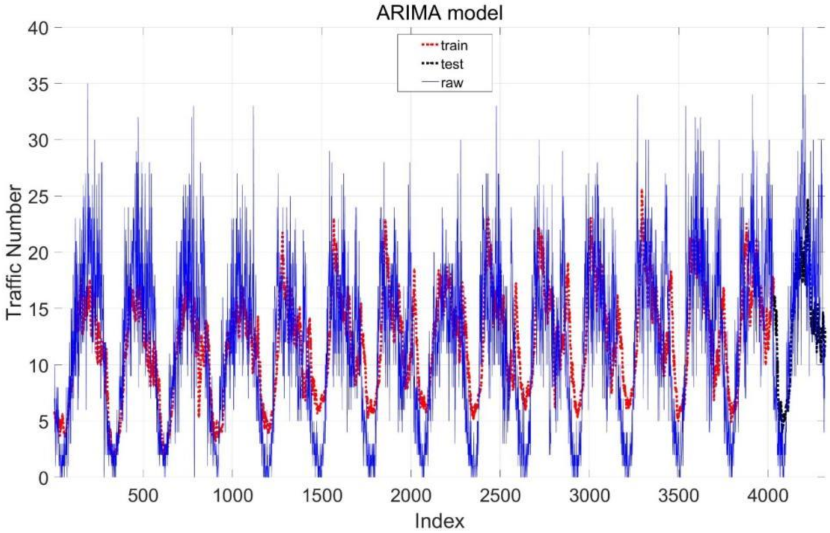

Figure 6. fitted values of the training data, the predicted values of the testing data, and the original data.

The model's MSE and RMSE in the training set are 30.7257 and 5.5431, respectively. The model's MSE and RMSE in the test set are 36.8591 and 6.0712, respectively. Table 2 displays a comparison of the MSE and RMSE for the two models.

Table 2. Comparison between two models.

Data	Model	MSE	RMSE
Training set	ARIMA	30.7257	5.5431
	LSTM	16.4455	4.0553
Test set	ARIMA	36.8591	6.0712
	LSTM	22.1633	4.7078

It can be seen from the comparison that the MSE and RMSE of the ARIMA model are obviously larger than the LSTM model. As a result, when processing this dataset, the LSTM model performs better than the ARIMA model.

4. Conclusion

Based on the constructed LSTM model, this paper conducted a simulation analysis of the traffic flow in Hangzhou's roads. The outcomes demonstrate the great accuracy with which the LSTM model can forecast traffic flow.

In the next step, by comparing the results with those predicted by the ARIMA model, it is demonstrated that the LSTM model has higher feasibility and accuracy in predicting traffic flow. The results further demonstrate the advantages of the LSTM model over traditional models, indicating its potential applicability in various scenarios.

Based on the prediction results of LSTM model for short-term transportation passenger flow, traffic management departments can better predict the possible problems. If resource allocation and traffic guidance can be carried out in advance, social efficiency will be greatly improved and the waste of public resources will be reduced. Such a forecast also provides great convenience for people's travel, and they can better plan the travel time and route.

This paper solely relied on time and traffic flow to train the LSTM model and conduct simulation analysis. Next, to further improve the accuracy of this method, additional attributes such as weather conditions and the number of traffic lights can be incorporated. This will enable further exploration of the upgrade and application of the LSTM model in short-term transportation passenger flow prediction.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Liu Y, Liu Z and Jia R 2019 DeepPF: A deep learning based architecture for metro passenger flow prediction *Transp. Res. Part C: Emerg. Technol.* 101 pp 18-34

[2] Chafak T, Nur S, Cenk G and Okan O 2023 Short term load forecasting based on ARIMA and ANN approaches *Energy Rep.* 9 pp 550-557

[3] Promise I, Enwere, Encarnación Cervantes-Requena, Luis A, Camuñas-Mesa, José M and de la Rosa. 2023 Using ANNs to predict the evolution of spectrum occupancy in cognitive-radio systems. *Integration.* 93 pp 102070

[4] Feng T, Zheng Z, Xu J, Liu M, Li M, Jia H and Yu X 2022 The Comparative Analysis of SARIMA, Facebook Prophet, and LSTM for Road Traffic Injury Prediction in Northeast China *Public Health* 10 p 3389

[5] Zhang J, Qu S, Zhang Z and Cheng S 2022 Improved genetic algorithm optimized LSTM model and its application in short-term traffic flow prediction *PeerJ Comput Sci.* 8 p 1048

[6] Zhu Z, Xu M, Ke J T, Yang H and Chen X Q 2023 A Bayesian Clustering Ensemble Gaussian Process Model for Network-Wide Traffic Flow Clustering and Prediction *Transportation Research Part C: Emerging Technologies* 148 p 104032

[7] Cao B and Gao M T 2018 Short-Term Traffic Flow Prediction Based on LSTM *Modern Computer* 332 p 3

[8] Liu C 2022 Short-term traffic flow prediction based on LSTM and its variants *Transport Energy Conservation & Environmental Protection* 18 pp 99-105

[9] K. Sharath Kumar and M. Rama Bai 2023 LSTM based texture classification and defect detection in a fabric. *Measurement: Sensors* 26 p 100603

[10] Tian Y, Zhang K L, Li J Y, Lin X Y and Yang B L 2018 LSTM Based Traffic Flow Prediction with Missing Data *J. Neurocomputing* pp 297-305