Traffic Flow Prediction Model Based on Multivariate Time Series and Pattern Mining in Terminal Area

ZHU Weiqi[1], *CHEN Haiyan*[1,3*], *LIU Li*[1], *YUAN Ligang*[2], *TIAN Wen*[2]

1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China;
2. College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China;
3. Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, P. R. China

(Received 20 June 2023; revised 7 October 2023; accepted 15 October 2023)

Abstract: To improve the accuracy of traffic flow prediction under different weather scenarios in the terminal area, a terminal area traffic flow prediction model fusing multivariate time series and pattern mining (MTSPM) is proposed. Firstly, a multivariate time series-based traffic flow prediction model for terminal areas is presented where the traffic demand, weather, and strategy of terminal areas are fused to optimize the traffic flow prediction by a deep learning model CNN-GRUA, here CNN is the convolutional neural network and GRUA denotes the gated recurrent unit (GRU) model with attention mechanism. Secondly, a time series bag-of-pattern (BOP) representation based on trend segmentation symbolization, TSSBOP, is designed for univariate time series prediction model to mine the intrinsic patterns in the traffic flow series through trend-based segmentation, symbolization, and pattern representation. Finally, the final traffic flow prediction values are obtained by weighted fusion based on the prediction accuracy on the validation set of the two models. The comparison experiments on the historical data set of the Guangzhou terminal area show that the proposed time series representation TSSBOP can effectively mine the patterns in the original time series, and the proposed traffic flow prediction model MTSPM can significantly enhance the performance of traffic flow prediction under different weather scenarios in the terminal area.

Key words: traffic flow prediction; multivariate time series; time series representation; pattern mining; deep learning

CLC number: TN925 **Document code:** A **Article ID:** 1005-1120(2023)05-0595-12

0 Introduction

Terminal areas are the most crucial and complex parts of the air transportation system, which makes the study of traffic conditions in terminal areas particularly important. Traffic flow prediction in a terminal area aims to accurately predict the air traffic flow within the terminal area over a period of time in the future. Accurate traffic flow prediction is vital for the management of airlines and airports as it enables decision-makers to develop more reasonable flight plans, optimize resource allocation, and efficiently dispatch personnel, which can improve operational efficiency and cost control. However, since traffic flow in a terminal area is influenced by various factors, including flight schedules, seasonal changes, passenger behavior, and unexpected events such as weather changes and flight delays, the prediction faces several challenges. The interaction and variability of these factors make it even more difficult to achieve satisfactory predictions. Therefore, it is crucial to select appropriate models and algorithms for traffic flow prediction, considering the characteristics of flight operations and the effects of various factors.

In previous studies on traffic flow prediction, researchers primarily relied on non-machine learning

*Corresponding author, E-mail address: chenhaiyan@nuaa.edu.cn.

How to cite this article: ZHU Weiqi, CHEN Haiyan, LIU Li, et al. Traffic flow prediction model based on multivariate time series and pattern mining in terminal area[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2023, 40(5): 595-606.

http://dx.doi.org/10.16356/j.1005-1120.2023.05.008

methods, such as moving average and exponential smoothing, which are based on historical data for statistical prediction[1-2]. However, these methods often incorporate subjective ideas and lack objectivity. With the advancement of machine learning, some researchers tried to use machine learning technologies for traffic flow prediction, including support vector machine (SVM), random forest (RF), contrast learning (CL), active learning (AL), and neural networks (NN)[3-4]. These algorithms can effectively handle nonlinear problems but require substantial amounts of data and appropriate feature engineering. Researchers also predicted the traffic flow by treating traffic flow as an individual time series and disregarding the impacts of other factors on traffic flow. In recent years, a few researchers have started to study the traffic flow forecasting problem from a multivariate time series aspect and achieved good results[5]. However, these methods lack the mining of the intrinsic patterns of the time series and thus fail to capture the trends information and deeper features of the time series.

In this paper, we attempt to construct a traffic flow prediction model based on multivariate time series and pattering mining (MTSPM) to tackle the complexity and high dimensionality of the terminal area traffic flow data. First, a multivariate time series fusion model is designed to use the related features for prediction. Meanwhile, a time series bag-of-pattern (BOP) representation based on trend segmentation symbolization, TSSBOP, is designed to extract the intrinsic patterns from the original series and use them for prediction. Finally, the two prediction results are fused and output. By considering both the effects of multivariate time series features and the intrinsic patterns of the original series, our method can give more accurate traffic flow prediction in the terminal area.

The main contributions of this paper are summarized as follows:

(1) A multivariate time series fusion model is designed to predict the traffic flow in the terminal area based on multiple influencing factors.

(2) TSSBOP representation method is proposed to mine intrinsic patterns from the original series.

(3) A traffic flow prediction model fusing multivariate time series and pattern mining is proposed, and its effectiveness and superiority are validated on the historical traffic dataset of the Guangzhou terminal area.

1 Related Work

In the research field of traffic flow, many early researchers employed non-machine learning methods to make long-term or mid-long-term predictions under different weather scenarios. For example, Pan et al.[1] used a gray prediction algorithm to predict air traffic flow, taking into account the complexity, nonlinearity, and uncertainty of air traffic flow. Dmochowski et al.[2] analyzed the features of air traffic flow in terminal areas based on observational data. Liu et al.[6] combined a gray GM (2, 1) model, regression model, and system dynamics for strategic predicting of air traffic flow. However, most of these non-machine learning methods need to incorporate subjective ideas. Consequently, some researchers began to apply machine learning methods for traffic flow prediction. For example, Chen et al.[3] predicted traffic flow under similar scenarios by using active metric learning. Mao et al.[7] quantified the impact of convective weather on traffic capacity in the terminal area from the aspects of airspace, traffic, weather, and proposed a capacity prediction model based on random forest algorithm. Chen et al.[4] addressed different traffic flow prediction problems under different scenarios by comparative learning methods. Additionally, Hossain et al.[8] quantified the impact of convective weather on the terminal area and proposed a novel traffic flow prediction model based on the weather impact traffic index. Wang et al.[9] developed a stochastic dynamic model to predict traffic flow considering adverse weather conditions.

Machine learning methods always require large amounts of data and proper feature engineering. Dealing with high-dimensional large data sets can affect the training speed of the model, while unsuitable feature engineering can affect the accuracy of the prediction. Therefore, some researchers tried to use time series mining methods for traffic flow prediction. For example, Wang et al.[10] used an improved weighted first-order local method for air traffic flow prediction. Yang et al.[11] proposed a string segmentation algorithm based on regular expressions, combined with an aviation vertical profile model, and validated the traffic flow prediction results using real aeronautical fixed telecommunication network (AFTN) message data. After that, they applied a time series prediction model based on the echo state network to predict traffic flow. Song et al.[12] used the dynamic time warping (DTW) algorithm to measure traffic flow correlation between sectors, constructed a dataset based on correlation, and developed an long short-term memory (LSTM) network prediction model for traffic flow prediction under different input conditions. Although time series methods improved the prediction accuracy and efficiency, they primarily focused on individual traffic flow series without considering the influence of other factors on traffic flow. In recent years, researchers have attempted to address traffic flow prediction from the aspect of multivariate time series. They have achieved promising results by using deep learning models to incorporate multivariate features and predict traffic flow in the terminal area. For example, Lin et al.[13] proposed a multi-input deep learning model to predict traffic flow in the terminal area under convective weather. Yan et al.[14] developed a deep learning-based traffic flow prediction framework that can capture the spatial and temporal dependencies of historical traffic flows and predict inbound traffic flows. Peng et al.[15] proposed a multi-input deep learning model of traffic flow prediction for the terminal area under convective weather. They all aimed to improve the accuracy and stability of terminal area traffic flow prediction by extending the weather features that affect the traffic state. However, these multivariate time series methods ignore the learning of the intrinsic patterns of the series itself and therefore fail to capture the trends and deeper features of the time series.

In order to mine the intrinsic patterns of the time series, an effective representation of the time series is the most critical step. Zhang et al.[16] introduced a straightforward time series representation method called trend based symbolic aggregation approximation (TSAX), which can capture the trend features of each segment to enhance representation accuracy. Lin et al.[17] proposed a BOP representation that extracts substructures from the original series as high-level features of the time series. These substructures are transformed using SAX enabling dimensionality reduction of the entire series while preserving local features. Ruan et al.[18] proposed an enhanced symbolic representation known as TrSAX. Building upon the BOP representation, they combined SAX with the least squares method to describe the mean and trend information of the time series, which can further refine the substructures.

Inspired by all these researches, we attempt to study traffic flow prediction methods in terms of both multivariate time series and univariate time series, respectively. By exploring the intrinsic patterns and dependencies existing in the traffic flow series, a traffic flow prediction model fusing multivariate time series and pattern mining is constructed for the terminal area.

2 MTSPM Model

In this section, we will propose the traffic flow prediction model MTSPM. The workflow of MTSPM is presented in Fig.1. As the figure shows, the MTSPM model has four parts: Data processing, multivariate time series fusing, univariate time series pattern mining and weighted fusion outputting. The details of each part are described in the following sections.

2.1 Data processing

For accurate prediction of traffic flow in the terminal area under different weather scenarios, we first construct a traffic state dataset with weather, traffic demand, and control strategy as features and actual traffic flow as target values. Among them, the values of weather features can be calculated by ATMAP-CW from 3D weather avoidance area (WAF) data (containing both image and numerical forms) and regular weather report (METAR) da-

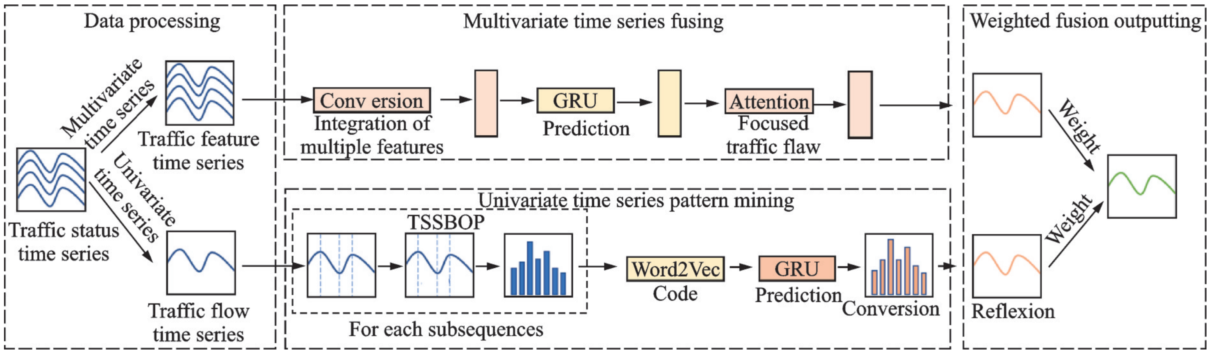

Fig.1 Workflow of the traffic flow prediction model MTSPM

ta, such as convective weather impact index, visibility impact index, wind speed, gust, rainfall, thunderstorm, wind direction, accumulated rain, etc. The values of traffic demand feature can be calculated from the flight plan and flight log data for actual flow, planned flow, tolerance, inbound delay, outbound delay, cancelled flight volume, normal flight volume, cabin volume and normal flight volume, etc. The control policy feature can be obtained from mile-in-trail (MIT) data by calculating the average crossing interval limit, policy release intensity, crossing interval limit effectiveness, whether the policy is active or not, and the traffic flow affected by the policy. The actual traffic flow can be counted directly from the executed flight plans. In this dataset, both features and targets are time series and they are synchronized in time in any sample.

2.2 Multivariate time series fusing

In the part of multivariate time series fusion, we first use convolutional neural network (CNN) model to fuse the input multivariate traffic flow feature time series, including weather series, traffic demand series and control strategy series, to extract the deep data features of each time series in the high-dimensional embedding space. Then, the computational results of CNN model are input to gated recurrent unit (GRU) model for traffic flow prediction, which can simplify the computational complexity and cost while maintaining accuracy.

However, GRU can only treat a time series as a unit during the computation, and cannot effectively handle long series or pay attention to the complex relationships existing in the time series. Therefore, GRU may not achieve the best results when dealing with sparse time series of traffic flow features. The attention mechanism is a technique widely used in deep learning to improve the performance of neural networks for processing time-series data. This mechanism helps the neural network to dynamically select and focus on the relevant parts of the input time series to extract useful information while reducing the reliance on irrelevant series, here we use GRUA to denote the GRU model with attention mechanism. Therefore, in this part, to improve the accuracy of traffic flow prediction, the attention mechanism is further utilized to focus on the traffic flow features and to form CNN-GRUA composite model for traffic flow prediction.

2.3 Univariate time series pattern mining

In the previous section, we used three traffic flow features weather, traffic demand and strategy for traffic flow prediction. However, the actual traffic flow itself is a time series with a strong cyclical trend. Therefore, in the univariate time series pattern mining part, we will study the traffic flow time series itself and mine the variation patterns and time-dependent structures existing in it to further predict the dynamic changes of the traffic flow series. In order to analyze and predict a time series, we should first segment and represent it effectively to retain the maximum amount of useful information in the time series. This is the general step of time series mining. To this purpose, a time series BOP representation based on trend segmentation symbolization, TSSBOP, is proposed in this paper to segment, represent and extract patterns from traffic flow time series. The process of TSSBOP is presented in the following sections.

2.3.1 Segmentation based on trend and sliding window

In order to preserve the trend information of the time series as much as possible in the segmentation process, we propose a dichotomous iterative breakpoint search method, as shown in Fig.2, where S_x ($x=1, i, j, m, q, n, k$) are points of the time series, D_i ($i=j, m$) are the distances between S_x and the line joining the start and end points, the black dashed line represents the original series and the red solid line represents the obtained segment.

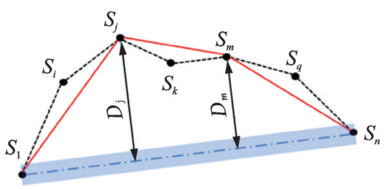

Fig.2 Dichotomous iterative search of breakpoints

First, we find the point with the largest trend change within the time series as the breakpoint, and divide the series into two subseries. By continuously dichotomizing and iterating in subseries, a number of breakpoints are obtained. The whole time series is traversed using the sliding window mechanism, and in each window, the segmentation strategy of breakpoints as the main and average segmentation as the secondary is adopted. $\omega-1$ breakpoints with the most trend information are selected to divide the series into ω segments. The adopted segmentation strategy makes the number of breakpoints in each subseries $\omega-1$. Let the number of breakpoints in each subseries be γ. If $\gamma \geqslant \omega-1$, $\omega-1$ breakpoints with the most trend information are selected from them to make ω segments. If $\lfloor(\omega-1)/2\rfloor+1 \leqslant \gamma < \omega-1$, the segments with the series length in the first $(\omega-1-\gamma)$ are bisected to form ω segments. If $\gamma < \lfloor(\omega-1)/2\rfloor+1$, then ω segments are divided directly by average length, where $\lfloor(\omega-1)/2\rfloor+1$ means rounding down to $(\omega-1)/2$.

Therefore, when there are few breakpoints, the overall segmentation process approximates the average segmentation and when there are enough breakpoints, the segmentation is performed by the iterative dichotomous trend segmentation. It can be seen that the segmentation strategy balances effectiveness and efficiency.

2.3.2 Symbolic representation based on trend and mean

After segmenting, we need to represent the segments in another way. In this paper, we design a symbolic representation based on trend and mean of a segment, so that each segment can be represented by two symbols, the trend and the mean, respectively.

For the part of trend, we directly use the slope value of each segment, which takes a range of $(-\infty, +\infty)$. This range can be divided into five non-overlapping regions and each region corresponds to a trend level and a capital letter, as shown in Table 1.

Table 1 Symbolic representation of trends

Slope value	Trend	Representation
$(-\infty, -1)$	Sharp decline	A
$(-1, -0.1)$	Slight decline	B
$(-0.1, 0.1)$	Slight fluctuation	C
$(0.1, 1)$	Slight rise	D
$(1, +\infty)$	Sharp rise	E

For the part of mean value, it simply describes the distribution of the values in a segment. For a given time series T and a subseries $x=[t_i, \cdots, t_{(i+l-1)}]$ in T, then the mean value of x can be calculated as follows.

$$\mu_x = \frac{\sum_{j=i}^{i+l-1} t_j}{l} \tag{1}$$

Since the normalized time series approximates a Gaussian distribution, it can be divided into multiple equal probability intervals, as shown in Table 2,

Table 2 Breakpoints for equal probability intervals of a Gaussian distribution

α	3	4	5	6	7	8	9	10
β_1	−0.43	−0.67	−0.84	−0.97	−1.07	−1.15	−1.22	−1.28
β_2	0.43	0	−0.25	−0.43	−0.57	−0.67	−0.76	−0.84
β_3		0.67	0.25	0	−0.18	−0.32	−0.43	−0.52
β_4			0.84	0.43	0.18	0	−0.14	−0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

where α is the number of breakpoints and β the interval boundary value of a Gaussian distribution. The values in the same interval are represented by the same symbol, so that the means of all the segments are mapped to different symbols according to the Gaussian distribution, as shown in Table 3.

Table 3 Symbolic representation of means

α	Mean range	Representation
3	$(-\infty,\beta_1),(\beta_1,\beta_2),(\beta_2,+\infty)$	a,b,c
4	$(-\infty,\beta_1),(\beta_1,\beta_2),\cdots,(\beta_3,+\infty)$	a,b,c,d
5	$(-\infty,\beta_1),(\beta_1,\beta_2),\cdots,(\beta_4,+\infty)$	a,b,c,d,e
6	$(-\infty,\beta_1),(\beta_1,\beta_2),\cdots,(\beta_5,+\infty)$	a,b,c,d,e,f
7	$(-\infty,\beta_1),(\beta_1,\beta_2),\cdots,(\beta_6,+\infty)$	a,b,c,d,e,f,g
8	$(-\infty,\beta_1),(\beta_1,\beta_2),\cdots,(\beta_7,+\infty)$	a,b,c,d,e,f,g,h

For each segment, it will finally be represented as multiple groups of words containing an uppercase letter and a lowercase letter in it. In this way, an original time series is symbolized as a string, retaining its key trend information and value distribution information, while the length has been greatly reduced.

2.3.3 Bagging of pattern

After symbolization, we can construct pattern "word lists" for the entire time series, where each word is a pattern. For example, for a symbolized time series $S=\{AaBaAc, AaBaAc, BcDaEb, DcDaEb, DcDaEb, DcDaEb, BcEaEb, AcDaEd\}$, by using digitally reduced counting, S will be reduced to $S'=\{AaBaAc, BcDaEb, DcDaEb, BcEaEb, AcDaEd\}$. Then, we construct the word list for each series as a histogram representation in the form: H: $AaBbAc=2, BaDaEb=1, DcDaEb=3, BcEaEb=1, AcDaEd=1$. For a given segment S_i, it is likely to be very similar to its neighboring subseries S_{i-1} and S_{i+1}, especially if S_i is in the smoothed region of the original time series. In this case, we may see many consecutive subseries mapping to the same word. To avoid storing these same words repeatedly, a counting method called number reduction is usually used. Specifically, for consecutive occurrences of the same word, only the first occurrence of the word is chosen to be recorded until a different word is encountered. The purpose of this method is to find out the patterns that occur in the series and their number of occurrences to make a histogram representation. After obtaining the histogram of each time series in the time series dataset, we constructed the word matrix $\boldsymbol{M}\left[\left(\varphi*\alpha\right)^{\omega}\right][s]$, where $\varphi=5$, α denotes the number of letters, ω the number of segments, and s the number of time series. Although the final number of matrices formed is huge, this matrix is sparse and most of the words are not mapped, so we can remove all the rows with value being of 0 to form the matrix $\boldsymbol{M}'[\cdot][s]$, thus ensuring that each word inside will appear once.

When we complete the above three steps, the various patterns present in an original time series are mined, which greatly reduces the computational complexity of subsequent time series prediction in terms of time and space while preserving the key information of the series.

2.3.4 Univariate time series predicting

To facilitate the input of the subsequent model, we first convert the obtained word matrix into an embedding vector encoding using the Word2Vec model. Word2Vec is a commonly used word embedding technique that maps words to a high-dimensional vector space and brings semantically similar words closer together in the vector space. Then, the vectors are fed into GRU model for prediction and the prediction results are converted into word matrix by the inverse operation of the Word2Vec model. Further, the word matrix of prediction results is converted into predicted values of the same form as the original series by the inverse operation of the TSS-BOP method to understand and analyze the prediction results intuitively.

2.4 Weighted fusion outputting

In order to effectively fuse the two predictions from the multivariate time series prediction model and the univariate time series prediction model, in this section, we propose a weighted fusion method based on the performance of the two models on the validation set. The data we used in this paper is the actual traffic data from Guangzhou terminal area. From the data, we can find the convective weather has significantly impact on the traffic flow but it does not frequently occur. The overall regularity of the data will make the prediction results relatively stable. Therefore, to capture the diversity in predictive accuracy as effectively as possible, we choose mean square error (MSE), which has a wider range of values, as the evaluation indicator. Meanwhile, in order to reduce the impact of traffic flow outliers caused by severe convective weather on prediction performance, we choose mean absolute error (MAE), which has a lower penalty for outliers as a complementary indicator. In summary, based on prior knowledge of actual traffic data from the terminal area, we use the combination of MSE and MAE for calculating the weights in the fusion process for the final predictions.

Denoting MSE and MAE of the multivariate model as MSE_1 and MAE_1, and those of the univariate model as MSE_2 and MAE_2, the specific weight assignment formula is shown as

$$W_1 = \frac{\mathrm{MSE}_2 + \mathrm{MAE}_2}{\mathrm{MSE}_1 + \mathrm{MAE}_1 + \mathrm{MSE}_2 + \mathrm{MAE}_2} \tag{2}$$

$$W_2 = \frac{\mathrm{MSE}_1 + \mathrm{MAE}_1}{\mathrm{MSE}_1 + \mathrm{MAE}_1 + \mathrm{MSE}_2 + \mathrm{MAE}_2} \tag{3}$$

Finally, we can get the final predicted value of traffic flow as

$$Y = W_1 * Y_1 + W_2 * Y_2 \tag{4}$$

where Y_1 and Y_2 are the prediction results of the multivariate time series prediction model and the univariate time series prediction model, respectively.

2.5 MTSPM

The main flow of the proposed MTSPM model is shown in Algorithm 1.

Algorithm 1 MTSPM

Input: Dataset TS, traffic flow TS_i, epoch;

Output: Prediction Y; // Multivariate time series fusing

(1) Initialize the hyper parameters of CNN, time step, hyper parameters of GRU, network weight, query Q, key K and value's weight matrices V.

(2) For each epoch $\in \{1, 2, \cdots, N\}$ do

(3) Input dataset TS to multivariate TS prediction model.

(4) Call CNN to extract the multivariate features and save to O_l.

(5) Call GRU to capture contextual information about a series and output hidden state series Z_l.

(6) Call attention mechanism to learn weight assignments and output predicted value Y_2// Univariate time series pattern mining.

(7) Input traffic flow TS_i to univariate TS prediction model.

(8) For each sub in TS_i:

(9) Call segmentation to calculate trend points and save to TP.

(10) Call symbolization of TP for symbolic representation.

(11) Feature pattern = TSSBOP(TP)

(12) Call Word2Vec to encode Pattern and save to X

(13) Call GRU to predict and save to X'

(14) $Y_1' = \mathrm{Conversion}(X')$

(15) $Y_1 = \mathrm{Reflexion}(Y_1')$// Weighted fusion outputting

(16) $Y = W_1 * Y_1 + W_2 * Y_2$

3 Experiments and Analysis

3.1 Experimental setup

3.1.1 Design of experiments

In this section, the performance of the proposed MTSPM model is verified through several comparative experiments. The experiments are divided into four parts. Firstly, the ablation experiment is carried out to verify the effectiveness of the weighted fusion of the two models. Secondly, comparative experiments of five time series feature representations are carried out to verify the performance of the TSSBOP method. Then, the comparative experiment of seven deep learning models is carried out to verify the performance of CNN-GRUA model. All of the above experiments are conducted to predict traffic flow in the terminal area under convective weather scenarios. The fourth comparative experiment focuses on the performance of the proposed MTSPM model under no convective weather scenario. Here, a convective weather scenario is one in which weather phenomena such as thunderstorms, rainfall or strong winds occur. A no convective weather scenario is a scenario without these weather phenomena.

3.1.2 Implementation

All experiments are run on the Guangzhou terminal area traffic flow dataset, where the experiments in Sections 3.2, 3.3, and 3.4 are carried out on the traffic flow dataset with convective weather scenarios, while the experiments in Section 3.5 are carried out on the traffic flow dataset without convective weather scenarios. The dataset has 23 different features of weather, traffic demand, and strategy. The time series of each feature has 7 344 timestamps, each timestamp is 1 h. Since in the pre-tactical stage, the lead time of flow control release is in hours, 1 h flow prediction is more relevant to the actual meaning. We use 80% of the data for training, 10% for validation, and 10% for testing. We select five indicators, including MSE, MAE, root mean square error (RMSE), mean absolute percentage error (MAPE), and R^2 (Coefficient of determination), to evaluate the performance of the prediction models.

3.2 Ablation experiments

To simplify the representation, here we refer to the multivariate time series prediction model as Model 1 and the univariate time series prediction model as Model 2. To verify the validity of the weighted fusion of the prediction results of the two models, we compared the prediction results of Model 1, Model 2 and our model MTSPM (Model 1+Model 2), as shown in Table 4.

Table 4 Ablation experiment results

Model	MAE	MAPE	MSE	RMSE	R^2
Model 1	5.26	**0.14**	48.99	6.99	0.88
Model 2	5.07	0.13	45.02	6.70	**0.89**
MTSPM	**4.70**	**0.14**	**42.84**	**6.54**	**0.89**

From Table 4, it can be seen that MTSPM model has the lowest MAE, MAPE, MSE, RMSE, and the highest R^2 compared to Model 1 and Model 2, which means MTSPM has the best performance. Fig.3 shows the traffic flow predictions of three consecutive days. It can be seen from Fig.3 that the prediction of our MTSPM model is closer to the actual traffic flow, which proves that the weighted fusion model MTSPM can greatly improve the prediction accuracy by taking into account the influence of multiple features on traffic flow and the trend pattern of the traffic flow time series itself.

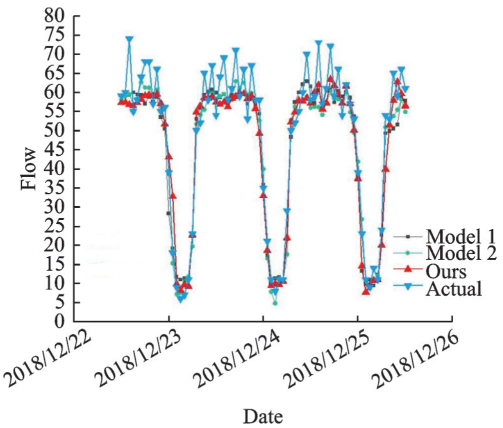

Fig.3 Ablation experiment results of consecutive days

3.3 Performance verification of TSSBOP

To verify the effectiveness of the proposed TSSBOP method, we compared it with other four time series representation methods, including SAX, ESAX, TSAX, and TrSAX, in Model 1, while using the same Model 2 and weighted fusing model. The experimental results are shown in Table 5.

Table 5 Comparison of five representation methods

Method	MAE	MAPE	MSE	RMSE	R^2
SAX	5.47	0.16	53.16	7.29	0.87
ESAX	5.11	**0.14**	47.45	6.89	0.88
TSAX	4.98	0.15	43.32	6.58	**0.89**
TrSAX	4.96	**0.14**	42.94	6.55	**0.89**
TSSBOP	**4.70**	**0.14**	**42.84**	**6.54**	**0.89**

From Table 5, we can see that although the improvement of TSSBOP on MAPE and R^2 is relatively small, there is a significant improvement on the remaining three indicators. Compared with other methods, TSSBOP can significantly downscale the features at the data level, so TSSBOP has the best representation ability. We present the predictions of three consecutive days, using five different representation methods in Fig.4. As shown in Fig.4, the predictions using TSSBOP method is closer to the actual flow than the other four representation methods, which proves that the proposed TSSBOP method can extract the features of the traffic flow series in terms of both trend and mean, and use the word matrix to mine the patterns existing in the time series, thus effectively improving the prediction performance.

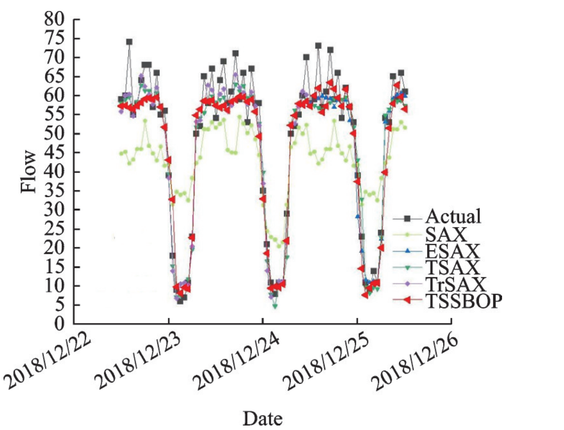

Fig.4 Prediction results of traffic flow by five representation methods

3.4 Performance verification of CNN-GRUA

To verify the effectiveness of Model 1 proposed in this paper, we compared the traffic flow prediction results of CNN-GRUA model with other six deep learning models, including CNN, LSTM, GRU, CNN-LSTM, CNN-GRU, and CNN-LSTMA, while using the same Model 2 and weighted fusing model. The experimental results are shown in Table 6.

Table 6 Comparison of seven deep learning models

Model	MAE	MAPE	MSE	RMSE	R^2
CNN	7.75	0.19	93.06	9.64	0.77
LSTM	6.32	0.16	63.89	7.99	0.84
GRU	6.16	0.16	63.22	7.95	0.84
CNN-LSTM	5.65	0.15	59.08	7.68	0.85
CNN-GRU	5.05	**0.14**	43.21	6.57	**0.89**
CNN-LSTMA	5.06	0.15	43.84	6.62	**0.89**
CNN-GRUA	**4.70**	**0.14**	**42.84**	**6.54**	**0.89**

From Table 6, it can be seen that CNN-GRUA model has the lowest MAE, MAPE, MSE and RMSE and the highest R^2 compared with the other six prediction models, which indicates that CNN-GRUA model has achieved the best performance in traffic flow prediction. We also present the predictions of three consecutive days, using seven different models in Fig.5. As shown in Fig.5, the prediction of CNN-GRUA model is closer to the actual traffic flow than the other six models, which proves that the proposed model for traffic flow prediction improve the performance by considering the long-term dependence and paying attention to traffic flow features during the deep learning process.

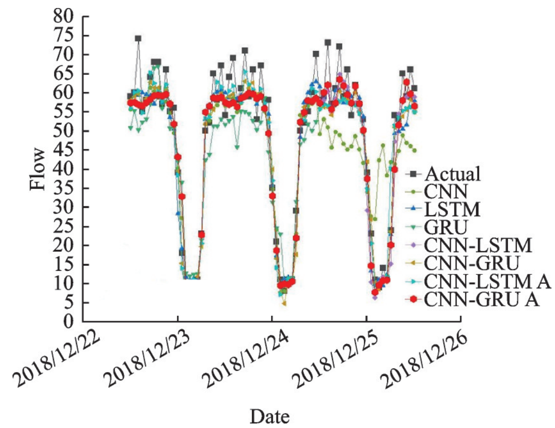

Fig.5 Prediction results of traffic flow by seven deep learning models

3.5 Prediction under no convective weather scenarios

The above experiments are all carried out on the terminal area traffic data under convective weather scenarios. In order to verify the performance of the proposed MTSPM model under no convective weather scenarios, we constructed the traffic dataset of Guangzhou terminal without convective weather, and carried out two comparative experiments. The experimental results are shown in Table 7 and Table 8. From Table 7 and Table 8, it can be seen

Table 7 Comparison of five representation methods without convective weather

Method	MAE	MAPE	MSE	RMSE	R^2
SAX	5.43	0.16	50.06	7.07	0.87
ESAX	4.76	0.13	38.90	6.23	0.90
TSAX	4.54	0.18	35.15	5.92	0.91
TrSAX	**4.18**	**0.11**	29.28	5.41	0.92
TSSBOP	4.19	0.12	**26.77**	**5.17**	**0.93**

Table 8 Comparison of seven deep learning models without convective weather

Model	MAE	MAPE	MSE	RMSE	R^2
CNN	7.49	0.88	84.88	9.21	0.78
LSTM	6.30	0.17	62.93	7.93	0.84
GRU	4.88	0.14	43.61	6.60	0.89
CNN-LSTM	4.42	**0.12**	31.23	5.58	0.92
CNN-GRU	4.31	**0.12**	29.83	5.46	0.92
CNN-LSTMA	4.31	**0.12**	29.83	5.46	0.92
CNN-GRUA	**4.19**	**0.12**	**26.77**	**5.17**	**0.93**

that the prediction performances of all the representation methods and the deep learning models on the dataset without convective weather are better than those under convective weather scenarios, both in MAE, MAPE, MSE, RMSE and R^2. This suggests that convective weather has an important impact on traffic flow in the terminal area. In a smooth weather scenario, the terminal area traffic flow has a stronger periodic pattern and thus is more easily predicted accurately.

We choose three consecutive days to demonstrate the prediction results of five different representation methods and seven different deep learning models, as shown in Fig.6 and Fig.7. It can still be concluded that the prediction results of the proposed TSSBOP method for pattern mining are closer to the real traffic flow for different time series representation methods. For different deep learning models, the prediction results of CNN-GRUA model are closer to the real traffic flow than the other models.

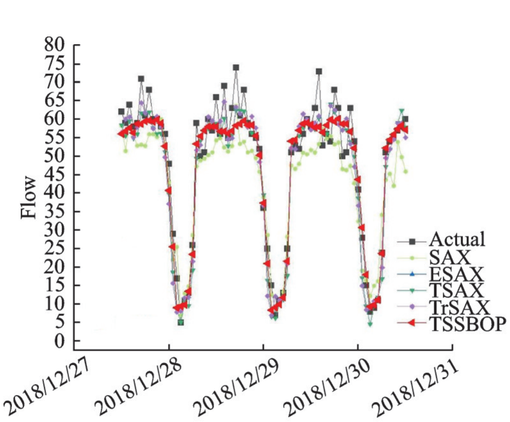

Fig.6 Prediction results of traffic flow by five representation methods without convective weather

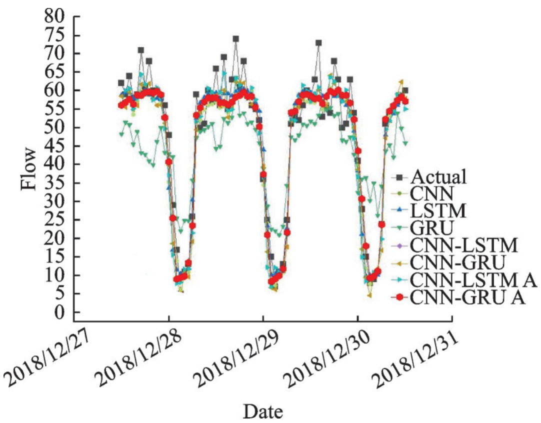

Fig.7 Prediction results of traffic flow by seven deep learning models without convective weather

4 Conclusions

To improve the accuracy of traffic flow prediction in terminal areas under different weather scenarios, this paper proposes a traffic flow prediction model based on MTSPM model. The model analyzes traffic flow in the terminal area from two perspectives: Multivariate time series and univariate time series. To improve accuracy and reduce computational complexity, GRU model is used to incorporate fused multivariate time series features. Additionally, the attention mechanism is introduced to focus on key traffic flow features and reduce reliance on irrelevant information. From the univariate time series perspective, this paper explores the mining of feature patterns within traffic flow series to improve prediction accuracy. A BOP representation method of time series based on trend segmentation symbolization is proposed to identify trend points in each subseries and discretizes the subseries into binary representations. Then, the discretized subseries is counted into a word matrix, and each word represents a feature pattern in the series. The experimental results on the historical dataset of the Guangzhou terminal area demonstrate the effectiveness of the proposed representation in mining feature patterns. Moreover, the multivariate time-series-based traffic flow prediction model for the terminal area significantly improves the accuracy of traffic flow prediction under different weather scenarios.

References

[1] PAN Zhiyi. Application of gray prediction method for air traffic flow[J]. Science and Technology Information, 2011, 374(18): 517.

[2] DMOCHOWSKI P A, SKORUPSKI J. Air traffic smoothness: A new look at the air traffic flow management[J]. Transportation Research Procedia, 2017, 28: 127-132.

[3] CHEN Haiyan, HOU Xiaye, YUAN Ligang, et al. Identification of similar air traffic scenes with active metric learning[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2021, 38(4): 625-633.

[4] CHEN Haiyan, LIU Zhenya, ZHOU Yi, et al. Recognition of similar weather scenarios in terminal area based on contrastive learning[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2022, 39(4): 425-433.

[5] LI Nan, JIAO Qingyu, ZHU Xinhua, et al. Prediction of departure aircraft taxi time based on deep learning[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37(2): 232-241.

[6] LIU Sichen. Research on strategic flow forecasting method based on grey linear regression combination model[C]//Proceedings of the Subforum on Aerospace Technology Innovation and Economic Transformation and Development in the Yangtze River Delta. Nanjing, China: Jiangsu Aerospace Society, 2012: 9. (in Chinese)

[7] MAO Limin, PENG Ying, LI Jie. Random-forest based terminal capacity prediction under convective weather[J]. Systems Engineering-Theory & Practice, 2021, 41(8): 2125-2136. (in Chinese)

[8] HOSSAIN M M, SAMEER A, DELAHAYE D. An evolutionary computational framework for capacity-safety trade-off in an air transportation network[J]. Chinese Journal of Aeronautics, 2019, 32(4): 999-1010.

[9] WANG S, CHU J, LI J, et al. Prediction of arrival flight operation strategies under convective weather based on trajectory clustering[J]. Aerospace, 2022, 9(4): 189.

[10] WANG Chao, ZHU Ming, ZHAO Yuandi. Air traffic flow prediction model based on improved weighted first-order localization method[J]. Journal of Southwest Jiaotong University, 2018, 53(1): 206-213. (in Chinese)

[11] YANG Yang, WANG Chao. Optimal prediction management of flight flow within the air traffic flow sector[J]. Computer Simulation, 2017, 34(9): 74-78. (in Chinese)

[12] SONG Weikai, ZHANG Honghai, WAN Junqiang, et al. Short-term prediction of air traffic flow based on dynamic time regularization and long- and short-term memory[J]. Science, Technology and Engineering, 2022, 22(25): 11256-11261. (in Chinese)

[13] LIN Yi, ZHANG Jianwei, LIU Hong. An algorithm for trajectory prediction of flight plan based on relative motion between positions[J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19(7): 905-916.

[14] YAN Z, YANG H, LI F, et al. A deep learning approach for short-term airport traffic flow prediction[J]. Aerospace, 2021, 9(1): 11.

[15] PENG Ying, WANG Hong, MAO Limin, et al. Terminal traffic flow prediction method under convective weather using deep learning approaches[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2021, 38(4): 634-645.

[16] ZHANG K, LI Y, CHAI Y, et al. Trend-based symbolic aggregate approximation for time series representation[C]//Proceedings of 2018 Chinese Control and Decision Conference (CCDC). Shenyang, China: IEEE, 2018: 2234-2240.

[17] LIN J, KHADE R, LI Y. Rotation-invariant similarity in time series using bag-of-patterns representation[J]. Journal of Intelligent Information Systems, 2012, 39(2): 287-315.

[18] RUAN H, HU X, XIAO J, et al. TrSAX—An improved time series symbolic representation for classification[J]. ISA Transactions, 2020, 100: 387-395.

Acknowledgements This work was supported by the National Key R&D Program of China (Nos.2022YFB2602403, 2022YFB2602401) and the National Natural Science Foundation of China (No.71971112).

Authors Mr. **ZHU Weiqi** received his B.S. degree in software engineering from Zijin College of Nanjing University of Science and Technology, in 2020. He is currently a postgraduate at the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include time series and data mining.

Dr. **CHEN Haiyan** received her M.S. and Ph.D. degrees in computer science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2005 and 2012, respectively. She is currently an associate professor at College of Computer Science and Technology, Nanjing University of

Aeronautics and Astronautics. Her research interests include machine learning and data mining.

Author contributions Mr. ZHU Weiqi designed the study, complied the models, conducted the analysis and wrote the manuscript. Dr. CHEN Haiyan contributed to data for the MTSPM model. Ms. LIU Li contributed to the processing of data. Dr. YUAN Ligang interpreted the results of MTSPM model. Dr. TIAN Wen contributed to the discussion and background of the study. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

基于多元时序和模式挖掘的终端区交通流预测

祝玮琦[1]，陈海燕[1,3]，刘 莉[1]，袁立罡[2]，田 文[2]

(1.南京航空航天大学计算机科学与技术学院，南京 211106，中国；
2.南京航空航天大学民航学院，南京 211106，中国；3.软件新技术与产业化协同创新中心，南京 210023，中国)

摘要：为了提高终端区不同气象场景下的交通流预测准确率，提出一种融合多元时序和模式挖掘(Multivariate time series and pattern mining, MTSPM)的终端区交通流预测模型。首先，给出了一种基于多元时间序列的终端区交通流预测模型，通过深度学习模型CNN-GRUA将终端区的交通需求、天气和策略特征进行融合并用于交通流预测；其次，针对交通流这一单变量时间序列，设计了一种基于趋势分段符号化的时间序列BOP(Bag-of-pattern)表示方法——TSSBOP，通过基于趋势的分段、符号化和模式表示来挖掘交通流序列中的内在模式；最后，根据两个模型在验证集上的预测精度进行加权融合，得到最终的终端区交通流预测值。在广州终端区的历史数据集上的对比实验表明，所提出的TSSBOP表示法能够有效挖掘出原始序列中的模式，所提出的基于MTSPM的终端区交通流预测模型能有效提高不同气象场景下的交通流预测性能。

关键词：交通流预测；多元时间序列；时间序列表示；模式挖掘；深度学习