Smart City Traffic Data Analysis and Prediction Based on Weighted K-means Clustering Algorithm

Lei Li
School of Computer and Mathematics, Harbin Finance University, Harbin, 150030, China

***Abstracts*—Urban traffic congestion is becoming a more serious issue as urbanization picks up speed. This study improved the conventional K-means method to create a new traffic flow prediction algorithm that can more accurately estimate the city's traffic flow. Firstly, the traditional K-means algorithm is given different weights by weighting, so as to analyze the traffic congestion in five urban areas of Chengdu by changing the weight values, and based on this, a traffic flow prediction model is further designed by combining with Holt's exponential smoothing algorithm. The findings showed that the weighted K-means method is capable of accurately identifying the patterns of traffic congestion in Chengdu's five urban regions and the prediction model combined with Holt's exponential smoothing algorithm had a better prediction performance. Under the environmental conditions of high traffic flow, when the time was close to 12:00, the designed model was able to obtain a prediction value of 9.81 pcu/h, which was consistent with the actual situation. This shows that this study not only provides new ideas and methods for traffic management in smart cities but also provides a reference value for the design of traffic prediction models.**

Keywords*—K-means; smart cities; traffic flow; prediction; holt; weight

I. Introduction

In the current context of rapid urbanization, with the continuous growth of urban population and motor vehicles, the urban transportation system is facing great pressure. Traffic congestion (TC) not only affects people's daily travel and increases travel costs, but also negatively affects the sustainable development of cities. Especially in China's new first-tier cities, TC has become a prominent social problem, which not only consumes a lot of time and resources but also exacerbates environmental pollution and poses a challenge to economic development and social stability [1-2]. Therefore, one of the most important problems in modern urban management is figuring out how to efficiently manage and optimize urban traffic while also increasing the effectiveness of the traffic system.

The development of modern information technology, especially the application of Internet of Things, cloud computing, and big data analytics, provides new solutions for traffic management in smart city (SC). In the context of SC construction, in-depth study of urban traffic data (TD) using big data analytics to predict and mitigate TC is of great significance for improving urban traffic management and realizing the intelligence and efficiency of the traffic system [3-4]. However, traditional traffic flow forecasting (TFF) methods often fail to fully take into account the spatio-temporal characteristics and complexity of TD, resulting in limited accuracy and usefulness of the prediction results [5-6]. Although existing research has developed a variety of traffic flow prediction models, most still rely on traditional algorithms such as single time series analysis or basic machine learning methods. These traditional methods often ignore the spatial-temporal characteristics of traffic data when dealing with complex urban traffic data, which leads to the limitation of accuracy and practicability of prediction. In addition, few existing studies take into account the importance of handling outliers and weight adjustment in traffic data, which further limits the effectiveness of the model in practical applications.

In order to solve this problem, a smart city traffic data analysis and prediction method based on weighted K-means clustering (K-means) is proposed. Taking Chengdu as an example, this paper first analyzes the traffic flow data from 2020 to 2023 by using the weighted K-means algorithm, and explores the traffic congestion types and traffic congestion coefficients in five urban areas of Chengdu. On this basis, the traffic flow prediction model is innovatively built by combining Holt algorithm, aiming to further improve the accuracy and practicability of traffic flow prediction. The improved traffic flow prediction method has significant advantages in dealing with complex traffic data with space-time dependence. Through comparative experiments, it is proved that this method not only improves the model's ability to recognize traffic congestion patterns, but also significantly improves the prediction accuracy by combining with Holt exponential smoothing algorithm. Therefore, this study not only fills the gap of existing research but also provides a more accurate and practical forecasting tool for urban traffic management, which has important theoretical and practical application value.

The study is organized into five sections: an analytical review of the relevant research work is included in Section II, and a quick introduction to the entire book is provided in Section I. Section III is the optimization design of the prediction method, Section IV is the testing of the algorithm performance. Discussion and conclusion is given in Section V and Section VI respectively.

II. Related Works

K-means (KM) is a sort of clustering technique that is currently frequently utilized for data analysis in many different industries. The authors Nguyen et al. introduced a novel KM modification designed to tackle the difficulties associated with categorical data clustering. Furthermore, the

study's findings demonstrated that, in comparison to the current algorithms, the new method's performance has a greater degree of reliability [7]. Daviran et al. combined a harmonic search, artificial bee colony meta-heuristic optimization algorithm with KM with the aim of solving one of the challenges of unsupervised clustering methods for mapping of mineral exploration potentials. The results of the study showed that the methodology used was able to pick appropriate clustering centroids and bring together objects in the same geospatial space for analysis [8]. A credit rating indicator system was developed by Chen et al. for online lending platforms. It consists of two qualitative and twelve quantitative indications that are representative of Chinese culture. The rotational component matrix's loadings were further refined into the online lending platform operation scale factor, capital dispersion factor, security factor, and profitability factor after factor analysis techniques decreased the dimensionality of the 14 indicators. Ultimately, KM was employed to group the component scores of every online lending platform in order to get the credit rating outcomes. The empirical findings demonstrated that, in comparison to online loan eye and online loan house, the suggested KM-based credit rating approach can more accurately offer credit ratings and effectively alert problematic platforms [9].

With the rapid development of cities, urban TFF becomes more and more important, and building a reasonable TFF model can not only warn the congestion pattern in advance, but also can be beneficial to the construction of urban road network (RN). Sun et al. proposed a method combining the K-means algorithm (KMA) and gated recurrent units for building short-term TFF models to cope with the effects of different TF patterns on the prediction results. The results indicated that the model takes into account the diversity of TF patterns, improves the prediction accuracy, and solves the short-term TFF problem more effectively than a single gated cyclic unit network, stacked self-encoder, random forest, and support vector machine regression [10]. Wang et al. proposed a multi-scale adaptive spatio-temporal prediction model, named AST-InceptionNet, aiming to solve the TFF problem in intelligent transportation systems with this model. The model effectively discovered potential spatio-temporal patterns by combining global and local map features, using the Inception part to integrate multi-scale spatio-temporal features. Experimental results revealed the satisfactory performance of AST-InceptionNet [11]. Huo et al. suggested a hierarchical TFF network that combines a newly designed long-term temporal Transformer network with a spatio-temporal GCN in order to address the over smoothing issue related to graph convolutional network (GCN)-based TFF approaches. The effectiveness and robustness of the suggested strategy were shown by the experimental findings on three publicly accessible TF datasets [12].

To summarize, there have been a series of researches conducted by many experts on the KMA and the TFF problem, but most of the researches use neural networks to build TFF models. In order to build the TFF model in a targeted way, this study takes the five urban areas of Chengdu City as an example, and builds the TFF model by improving the KMA, aiming to solve the TC problem of Chengdu City better.

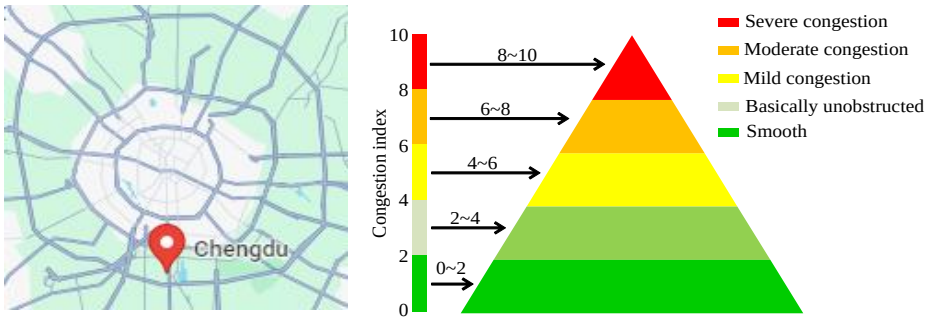
III. Traffic Data Analysis and Prediction Study of Smart City based on Clustering Algorithm

With the continuous growth of the number of residents and motorized vehicles in cities, the TC problem in large cities has attracted more and more attention. This study proposes a TC data analysis and congestion type identification method based on the WKM clustering algorithm, based on which a TFF model is constructed in combination with Holt, aiming to further improve the prediction effect of TF. Weighted K-means and Holt algorithm are selected for traffic flow prediction because traditional K-means have limited performance when dealing with high dynamic changes in traffic data, while weighted K-means can deal with this challenge more effectively through weight adjustment. At the same time, Holt algorithm can accurately capture the data trend and improve the prediction accuracy. In contrast, the commonly used neural network model may not perform well when the data is unstable or missing, while the proposed method combines the advantages of both and is suitable for dealing with complex urban traffic patterns, so as to provide more stable and reliable prediction results.

A. Design of Traffic Data Processing Method based on Temporal Clustering

Temporal clustering technique plays an important role in TFF of SC, which can not only analyze and process a large amount of TF data, but also improve the prediction accuracy and efficiency of traffic prediction models [13]. In this study, the TF data of Chengdu city was collected from 2020 to 2023 for clustering analysis as an example, in order to identify the TC patterns in different areas of Chengdu city and the changes of its spatial data. By the end of 2023, the resident population of Chengdu City will be about 7.16 million, and the number of motor vehicles has exceeded six million. Traffic congestion index (TCI) is an indicator used to measure the degree of TC in a city, which is usually calculated by analyzing data such as TF and vehicle speed. The RN structure of Chengdu city and the level of TCI are shown in Fig. 1.

Fig. 1(a) shows the RN structure of Chengdu City, which is a composite urban transportation system combining ring roads and radial RNs. Since 2006, Chengdu City has also used TCI as a key indicator to measure the traffic condition of urban roads and released real-time TC information to the public through various channels such as the Internet and WeChat public number. In Fig. 1(b), according to the value range of TCI, the TC situation can be divided into five different levels, which are smooth traffic, basic smooth traffic, light TC, moderate TC and severe TC. The urban RN's functioning and the TF's degree of smoothness can all be reflected in TCI over time; a higher value indicates a more severe degree of TC.



(a) Traffic network structure in Chengdu (b) Traffic congestion index pyramid structure

Fig. 1. Road network structure and traffic congestion index pyramid structure in Chengdu.

100,000 data were randomly selected from the TF data from 2020 to 2023 to be analyzed, and the 100,000 data collected included Jinjiang, Qingyang, Jinniu, Wuhou, and Chenghua districts. The collection interval of each data is 10 minutes, i.e., the whole day is divided into a total of 144 time segments to collect data. Five pieces of data were randomly selected from the 100,000 pieces of data collected for display, as shown in Table I.

Some of the data collected are given in Table I. In view of reasons such as mechanical equipment failures or operational errors, it is inevitable that the raw TCI data will contain omissions and anomalies. Therefore, appropriate preprocessing of these data is required before carrying out the data analysis work. Eq. (1) illustrates the computation procedure that is used to fill in the missing data using the linear interpolation method [14].

$$x_i = x_0 + \frac{i}{I+1} \times \left(x_{I+1} - x_0 \right) \quad \forall i = 1, 2, \cdots I \tag{1}$$

In Eq. (1), $i = 1, 2, \cdots I$ denotes a consecutive time period, x_i denotes the missing value in time period i, and x_0 denotes the congestion value recorded at time 0. x_{I+1} denotes the congestion value recorded at time period $I+1$.

The 2-sigma criterion is used in this study to deal with the anomalous TDs. Assuming that M represents the number of sampling points (SPs) per day, $M = 144$ is used since the interval between the data collection in this study is 10 minutes. Let N be the number of days of observation to get the data vector of TCI, and Eq. (2) depicts the expression.

$$X_n = \left(x_n^1, x_n^2, \cdots, x_n^M \right) \quad \forall n = 1, 2, \cdots, N \tag{2}$$

In Eq. (2), X_n denotes the data vector of TCI, $x_n^1, x_n^2, \cdots, x_n^M$ denotes the data vector of TCI under different observation days, respectively, and n denotes an arbitrary value of N. The mean value of TCI under multi-day observation time is further obtained from Eq. (2) as shown in Eq. (3).

$$\overline{X} = \left(\overline{x}^1, \overline{x}^2, \cdots, \overline{x}^M \right) = \left(\frac{1}{N} \sum_{n=1}^{N} x_n^1, \frac{1}{N} \sum_{n=1}^{N} x_n^2, \cdots, \frac{1}{N} \sum_{n=1}^{N} x_n^M \right) \tag{3}$$

In Eq. (3), $\overline{X}$ represents the average value of TCI under multi-day observation time. $\overline{x}^1, \overline{x}^2, \cdots, \overline{x}^M$ denotes the average value of TCI under different time periods, respectively. Based on Eq. (2) and Eq. (3) the formula for the remaining fluctuation on day n can be obtained as shown in Eq. (4).

$$r_n = X_n - \overline{X} = \left(r_n^1, r_n^2, \cdots, r_n^M \right) \tag{4}$$

In Eq. (4), r_n denotes the residual volatility on day one. $r_n^1, r_n^2, \cdots, r_n^M$ denotes the value of residual volatility under different time periods, respectively. Using the sample standard deviation σ^m to represent the square root of $r_n^1, r_n^2, \cdots, r_n^M$, $m = 1, 2, \cdots, M$, the correction formula for outliers is obtained as shown in Eq. (5).

$$x_n^m = \begin{cases} \overline{x}^m + 2\sigma^m & r_n^m > 2\sigma^m \\ x_n^m & -2\sigma^m \le r_n^m \le 2\sigma^m \\ \overline{x}^m - 2\sigma^m & r_n^m < -2\sigma^m \end{cases} \tag{5}$$

In Eq. (5), x_n^m denotes the outlier.

TABLE I. EXAMPLES OF TRAFFIC DATA IN CHENGDU

ID	City center	Congestion index	Date	Time
23150	Jinjiang district,	7.6	2020.3.2	7:40~7:50
30264	Qingyang district	8.1	2020.9.24	11:50~12:00
41581	Jinniu district	6.5	2021.6.13	6:30~6:40
53294	Wuhou district	5.8	2022.8.15	21:10~21:20
63248	Chenghua district	6.3	2022.10.6	22:30~22:40

B. Traffic Flow Forecasting Based on Weighted K-Means-Holt

KM clustering is a widely used unsupervised learning algorithm that uses an iterative approach to partition the collection of SPs into subsets of classes, which has the advantages of being simple and easy to understand, computationally efficient, and suitable for handling large-scale datasets [15-16]. This study utilizes KM to complete the clustering of the TD SP collection. Additionally, Fig. 2 depicts its clustering procedure.

The clustering process of KMA is shown in Fig. 2. The initial clusters need to be selected first, followed by clustering the data objects and assigning them to the appropriate clusters. Over time, the size of the clusters is continuously adjusted to ensure that each object has the same category throughout the data set. The KMA is constantly repeated to generate the best clusters.

Assuming that there exists a subset of class K, $k=1,2,\cdots,K$. $C_1,C_2,\cdots,C_k$ denotes the set of SPs, the total bias of the set of SPs is minimized using the KM clustering algorithm, the process is shown in Eq. (6) [17-18].

$$\sum_{k=1}^{K}\sum_{X_n\in C_k}\sum_{m=1}^{M}\left(X_n-U_k\right)^2 \tag{6}$$

In Eq. (6), X_n is a sample data of M dimension, denoting the time series (TS) data with M SPs in a day. U_k is a vector of M dimension, denoting the clustering center of class k. The formula of U_k is shown in Eq. (7).

$$U_k=\frac{1}{|C_k|}\sum_{X_n\in C_k}x_n^m \tag{7}$$

The traditional KMA, although simple and efficient, faces several limitations when dealing with complex TD analysis and TFF tasks. To overcome these limitations, this research further proposes the WKM algorithm. By giving varying weights to distinct features, the WKM algorithm improves its ability to handle anomalous data and uneven feature relevance. This allows it to perform more accurately and efficiently in TD analysis and TFF. Eq. (8) displays the defined equation of the coefficient of variation, which is used to quantify the degree of dispersion of the collection of SPs.

$$CV_m=\frac{\sigma_m}{\overline{x}_m} \tag{8}$$

In Eq. (8), CV_m denotes the coefficient of variation at time period m. Based on the value of CV_m, a WKM clustering algorithm is further proposed and utilized to minimize the total weighted deviation of the clustering centers, which is shown in Eq. (9).

$$\sum_{k=1}^{K}\sum_{X_n\in C_k}\sum_{m=1}^{M}\left(CV_m X_n'-U_k'\right)^2 \tag{9}$$

In Eq. (9), X_n' denotes the TS data with M SPs in a day, $X_n'=\left(X_1',X_2',\cdots,X_N'\right)$. U_k' denotes the weighted clustering center of the first class, $U_k'=\left(U_1',U_2',\cdots,U_K'\right)$. The formula of U_k' is shown in Eq. (10).

$$U_k'=\frac{1}{|C_k|}\sum_{X_n\in C_k}CV_m x_n^m \tag{10}$$

To determine the best K-value, this study also used the contour coefficient to evaluate the clustering results related to the K-value until the best clustering result was selected as the final K-value. The expression of contour coefficient is shown in Eq. (11).

$$s\left(X_n'\right)=\frac{b_n-a_n}{\max\left\{a_n,b_n\right\}} \tag{11}$$

In Eq. (11), $s\left(X_n'\right)$ denotes the profile coefficient of X_n', and a_n denotes the average Euclidean distance (AED) between X_n' and other samples in the same group. b_n denotes the AED between X_n' and all the samples in its closest group. The average of the profile coefficients of all samples is the final profile coefficient, which is calculated as shown in Eq. (12).

$$S=\frac{s\left(X_1'\right)+s\left(X_2'\right)+\cdots+s\left(X_N'\right)}{N} \tag{12}$$

In Eq. (12), S denotes the final profile coefficient. According to Eq. (6) to Eq. (12) can be used to create the WKM clustering method's flowchart, which is depicted in Fig. 3.

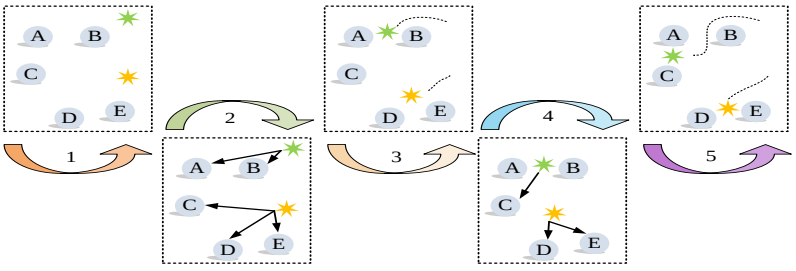


Fig. 2. K-means algorithm clustering process.

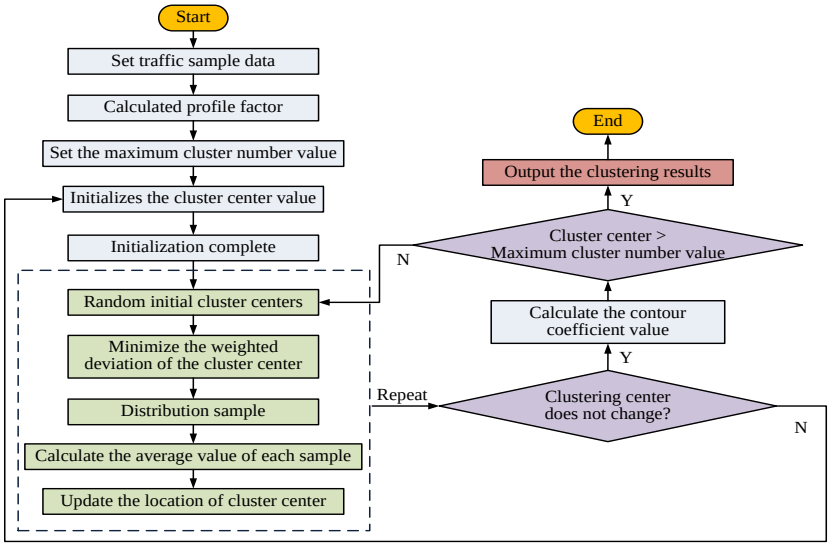


Fig. 3. Flow chart of weighted K-means algorithm running.

Fig. 3 shows the WKM algorithm's specific operation flow. A traffic sample data set is provided first, followed by the computation of the set's contour coefficients, the setting of a maximum number of clusters, and the initialization of the clustering center value. After initialization, the initial clustering centers are randomly selected from the traffic sample data set and the samples are assigned by minimizing the total weighted deviation of the clustering centers. Second, the cluster center's position is updated by computing the mean value of each class of samples. These two processes are continued until the cluster center stays constant. Finally, the contour coefficient value is calculated, and when the number of cluster centers at this point is greater than the maximum number of clusters value then the corresponding clustering results are output, otherwise the number of cluster centers is adjusted to reclustering.

The weighted K-means algorithm used in this study involves several key parameters, such as weight factor, initial selection of cluster center and number of iterations, which have a significant impact on the accuracy of prediction results and the convergence speed of the algorithm. In addition, the initial choice of cluster center has a decisive influence on the stability of the final result, and the number of iterations is directly related to the operational efficiency of the model. In addition to completing the analysis of TD using the WKM algorithm to identify different congestion patterns, it is also necessary to further build a TF warning model to provide real-time warnings of TF speeds to help alleviate TCs. The formula for data prediction at a certain time period in the future using Holt's exponential smoothing (ES) algorithm is shown in Eq. (13).

$$x_{m+h} = l_m + \left(\varphi + \varphi^2 + \cdots + \varphi^h\right)\theta_m \tag{13}$$

In Eq. (13), x_{m+h} denotes the predicted value in period $m+h$, φ denotes the damping coefficient, and h denotes the number of predicted dates. l_m denotes the horizontal smoothing equation, which usually denotes the primary ES value for period m. θ_m denotes the trend smoothing equation, which usually represents the quadratic ES value of the m period. The specific formula for l_m is shown in Eq. (14).

$$l_m = \alpha x_m + (1-\alpha)\left(l_{m-1} + \varphi\theta_{m-1}\right) \tag{14}$$

In Eq. (14), α denotes the horizontal smoothing coefficient, which takes values between 0 and 1. x_m denotes the observed value in period m. The specific formula for θ_m is shown in Eq. (15).

$$\theta_m = \beta\left(l_m - l_{m-1}\right) + (1-\beta)\varphi\theta_{m-1} \tag{15}$$

In Eq. (15), β denotes the trend smoothing coefficient, which also takes values ranging from 0 to 1. The TFF algorithm that combines the Holt ES algorithm with the WKM algorithm is denoted as WKM-Holt, and the operation flow of WKM and according to Eq. (13) to Eq. (15) can be obtained as shown in Fig. 4.

The flow chart of the operation of the WKM-Holt algorithm is given in Fig. 4. Firstly, a collection of historical traffic sample data needs to be given and the clustering is calculated according to the WKM algorithm, and secondly, the clustering samples and clustering center values are obtained and the temporal characteristics of the clustering results are summarized. Next, the historical traffic sample data set is used as a training set for the Holt ES model to obtain the level smoothing coefficients and trend smoothing coefficients that minimize the prediction error. Select a known moment of real-time sample data and use the prediction model to make predictions, match the predicted values and historical values and use the two-fold standard deviation solution for numerical anomaly warning. If the value is within a reasonable threshold then the data is normal, otherwise the prediction is alarmed, after detecting all the data to complete the prediction process of the WKM-Holt algorithm.

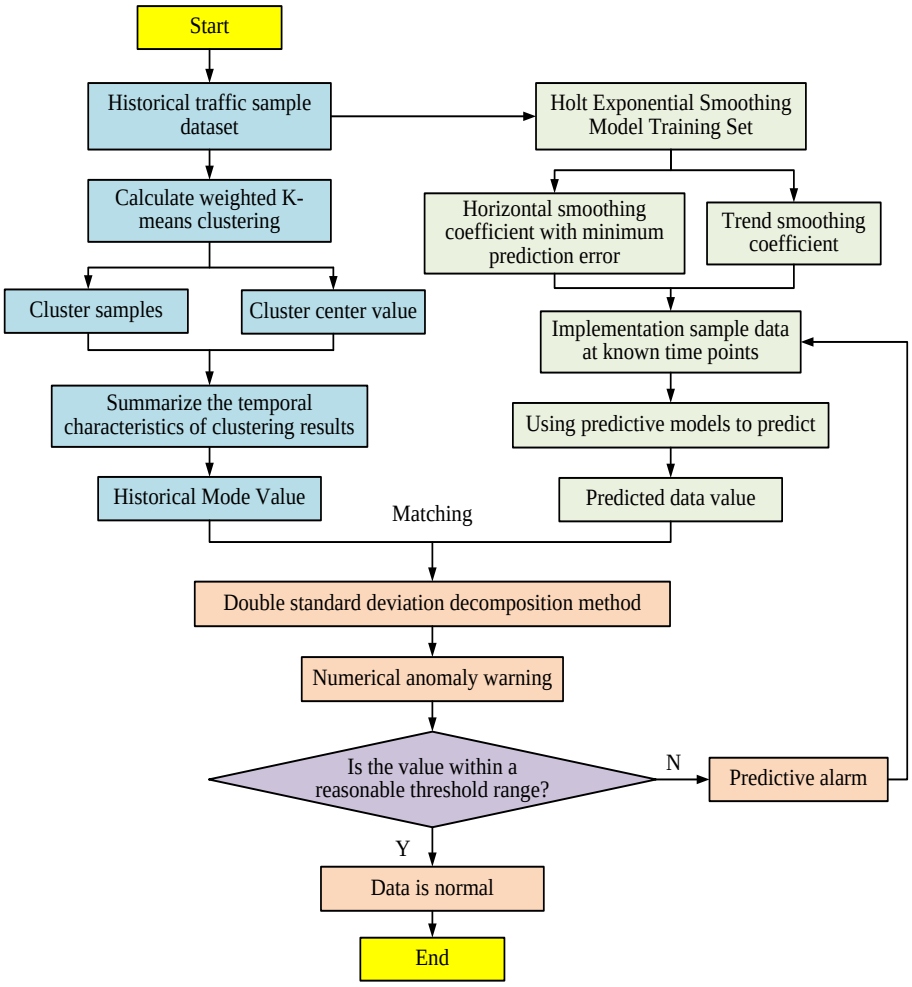


Fig. 4. Operation flow chart of weighted k-mean-holt algorithm.

IV. Performance Analysis of Clustering Algorithms and Application Effect Analysis

The capability of the WKM algorithm to analyze TD and identify TC patterns is tested through case studies, while the WKM-Holt prediction algorithm's prediction performance and the impact of its practical application are tested through case studies in the latter case.

A. Weighted K-Means based Traffic Congestion Pattern Recognition Results

The processed TF data of Chengdu City in subsection 2.1 is used as the experimental dataset for this case study, and 98,548 valid data are left after 100,000 data are preprocessed. The WKM method is used to cluster analyze the collected valid data, and the contour coefficients of the TF data of the five urban areas of Chengdu City under different number of clusters are obtained as shown in Table II.

TABLE II. Contour Coefficients of Five Urban Districts in Chengdu under Different Clustering Numbers

Number of clusters	District				
	Jinjiang district	Wuhou district	Jinniu district	Qingyang district	Chenghua district
2	0.31	0.24	0.23	0.34	0.17
3	0.35	0.32	0.20	0.31	0.15
4	0.32	0.28	0.18	0.29	0.11
5	0.28	0.27	0.19	0.26	0.08
6	0.23	0.24	0.17	0.25	0.06
7	0.18	0.21	0.15	0.19	0.05
8	0.16	0.17	0.12	0.17	0.03
9	0.13	0.13	0.13	0.18	0.04
10	0.11	0.10	0.08	0.14	0.05

Table II lists the contour coefficient values for Chengdu's five urban zones under various clustering numbers. When the clusters is 3, the contour coefficients of Jinjiang and Wuhou districts are able to reach the maximum value, which are 0.35 and 0.32, respectively. When the clusters is 2, the contour coefficients of Jinniu, Qingyang, and Chenghua districts are able to reach the maximum value, which are 0.23, 0.34, and 0.17, respectively. The TC modes of the five urban areas exhibit spatial correlation, as evidenced by the variations in the contour coefficients in Table II. Jinjiang and Wuhou districts, which are close to the main urban area, have three congestion patterns, while Jinniu, Qingyang and Chenghua districts, which are slightly away from the main urban area, have two congestion patterns. The detailed changes of TCI in the five urban areas under different congestion patterns are shown in Fig. 5.

The variation of TCI with different congestion patterns in five urban areas is given in Fig. 5. In Fig. 5, Mode 1, Mode 2, and Mode 3 represent three different congestion patterns, where Mode 1 has the best congestion, which usually occurs in the middle portion of weekdays. Mode 2 has moderate congestion and usually occurs at the beginning and end of the weekday, such as Mondays and Fridays. Mode 3 has the worst congestion and usually corresponds to holidays. Taking Fig. 5(a) of the two congestion modes as an example, the maximum TCIs of Jinniu district in Fig. 5(a) are 6.73 and 7.98 under mode 1 and mode 2, respectively. In addition, Wuhou district in Fig. 5(d) is selected from Fig. 5(d) and Fig. 5(e) for the analysis, and it is found that the maximum TCIs of Wuhou district are 7.81, 8.00, and 9.75 under modes 1, 2, and 3, respectively. By comparing the congestion indices of each district under different modes in Fig. 5, it can be seen that the congestion mode of each district corresponds to the number of its optimal clustering number, which shows that the congestion indices of each district are clustered in space. The results of exploring the effect of different linear numbers of motor vehicles on TCI in five urban areas are shown in Table III.

In Table III, when the restriction numbers are 4 and 9, the average congestion index at this time is 3.42, which is larger than the average congestion index under other restriction numbers. When the restriction numbers are 1 and 6, the average congestion index is the smallest, which is only 2.98. It can be seen that the number of license plates ending in 4 and 9 is small, while the number of license plates ending in 1 and 6 is large. The test findings in Table IV are produced by using a t-test to determine whether there were any significant differences in congestion between license plate numbers.

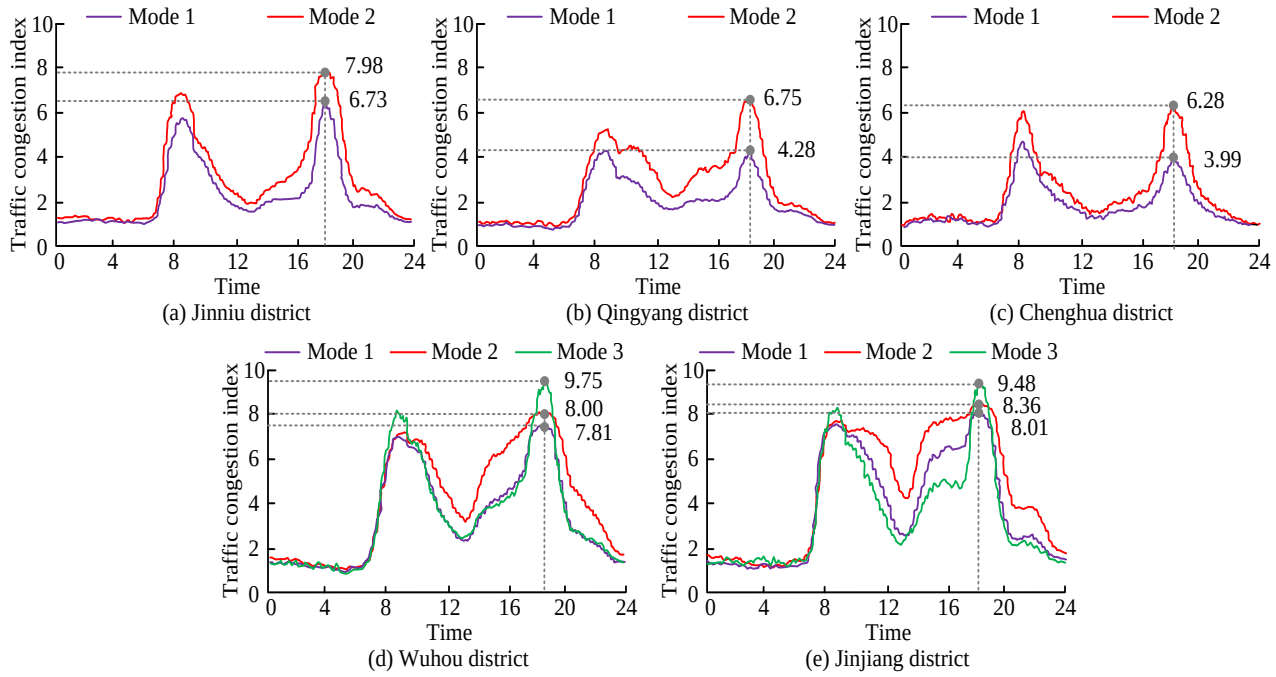


Fig. 5. Change of traffic congestion index in five urban areas of Chengdu.

TABLE III. STATISTICAL RESULTS OF TRAFFIC CONGESTION INDEX IN FIVE URBAN AREAS UNDER THE CONDITION OF VEHICLE LICENSE RESTRICTION

Motor vehicle restriction number	District					Average congestion index
	Wuhou district	Jinjiang district	Jinniu district	Qingyang district	Chenghua district	
0 and 5	4.21	4.27	2.45	2.31	2.41	3.13
1 and 6	4.01	3.89	2.39	2.33	2.28	2.98
2 and 7	4.37	4.29	2.48	2.69	1.92	3.15
3 and 8	4.15	3.89	2.36	2.42	2.28	3.02
4 and 9	4.48	4.32	2.86	2.79	2.65	3.42

In Table IV, when the restriction numbers are 4 and 9, at this time the restriction numbers are statistically significantly different from the other four groups of restriction numbers (P<0.05), which shows that the motor vehicle restriction policy has a certain significance on the TC pattern, and most of the cities can utilize the restriction strategy to alleviate the TC.

B. K-Means-Holt based Traffic Flow Forecasting Results

The 98,548 valid data are divided into training set and test set according to the ratio of 9:1, and Mini Batch K-Means Clustering (Mini-Batch-K-means), traditional KMA, and WKM algorithm are chosen as the comparison algorithms, and the prediction error performances of different algorithms under test set are obtained as shown in Fig. 6.

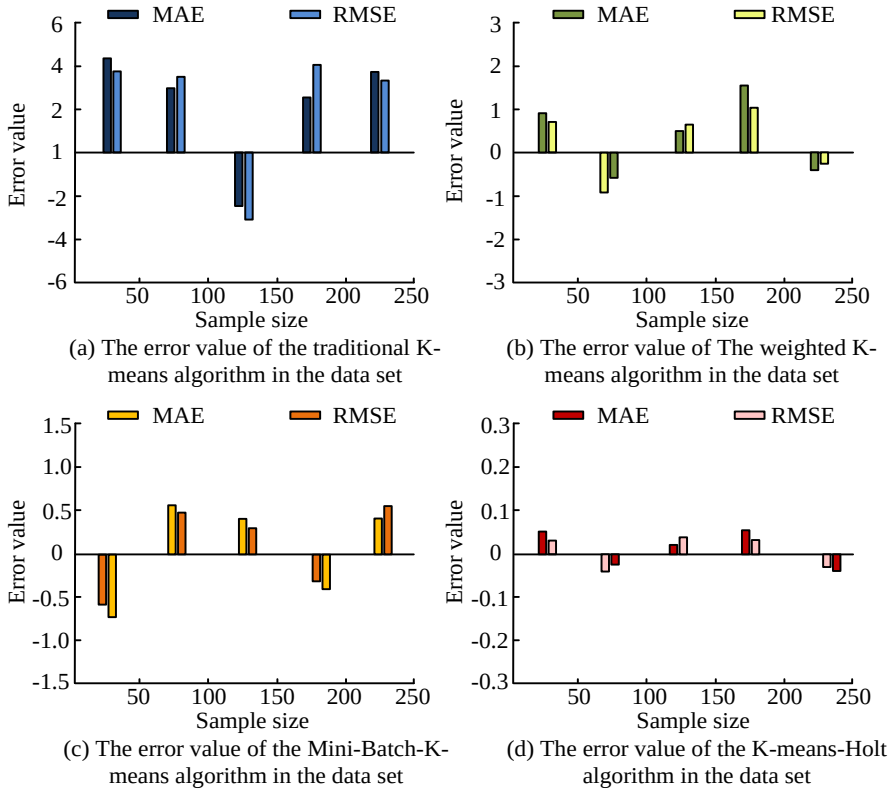
For KM, WKM, Mini-Batch-KM, and K-means-Holt (KMH) in the test set, the mean absolute error (MAE) and root mean square error (RMSE) are displayed in Fig. 6(a), (b), (c), and (d), respectively. Combined with Fig. 6, it can be noted that the error ranges of KM, WKM, Mini-Batch-KM, and KMH are -4~6, -1~2, -1~1, and -0.1~0.1, respectively, which shows that KMH performs best in terms of error. The prediction of KMH in different TF environments is shown in Fig. 7.

Fig. 7(a) and (b) display the KMH prediction findings for TF in high TF and low TF situations, respectively. The predicted values of KMH in both TF environments overlap well with the actual values. In both TF environments, when the time is close to 12:00, the TF is able to reach the peak value, which is 9.81 pcu/h and 9.75 pcu/h, respectively, and the prediction at this time basically coincides with the actual situation. The effect of KMH in the actual TFF is shown in Fig. 8.

Fig. 8(a) and 8(b) show the actual TF and the TF under KMH prediction at a certain time, respectively. In Fig. 8(a), the flow rate of the actual TF is mainly centered under 100 pcu/min, which is consistent with the TF under KMH prediction in Fig. 8(b), and it can be illustrated that KMH is able to effectively predict the TF.

TABLE IV. P-VALUE RESULTS OF T-TEST UNDER DIFFERENT VEHICLE LICENSE NUMBER RESTRICTION POLICIES

Motor vehicle restriction number	0 and 5	1 and 6	2 and 7	3 and 8	4 and 9
0 and 5	/	0.17	0.23	0.06	0.01
1 and 6	/	/	0.00	0.91	0.00
2 and 7	/	/	/	0.01	0.01
3 and 8	/	/	/	/	0.02
4 and 9	/	/	/	/	/



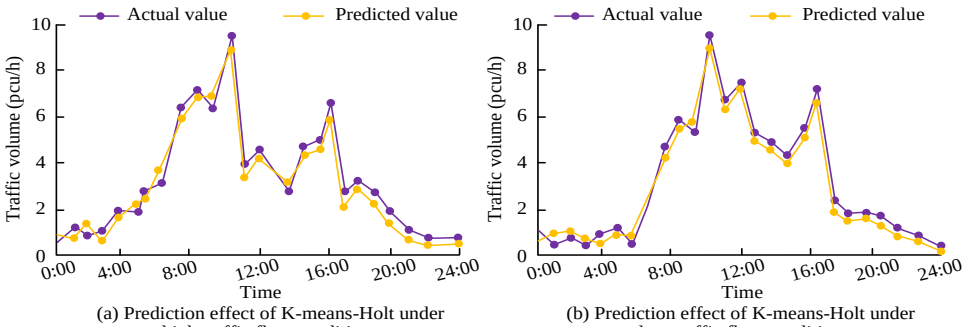
(a) The error value of the traditional K-means algorithm in the data set

(b) The error value of The weighted K-means algorithm in the data set

(c) The error value of the Mini-Batch-K-means algorithm in the data set

(d) The error value of the K-means-Holt algorithm in the data set

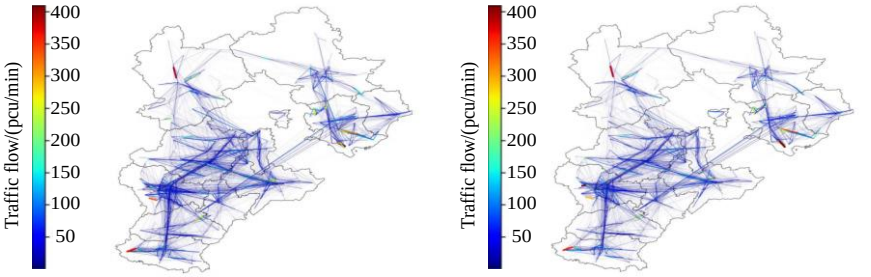
Fig. 6. Error performance of different prediction algorithms.



(a) Prediction effect of K-means-Holt under high traffic flow conditions

(b) Prediction effect of K-means-Holt under low traffic flow condition

Fig. 7. Prediction effect of K-means Holt under different traffic flow environments



(a) Actual traffic flow

(b) Traffic flow under K-mean-Holt prediction

Fig. 8. Prediction results of actual traffic flow by K-means-Holt.

In order to further evaluate the validity and scalability of the method proposed in this study, traffic data sets of Shanghai and Beijing were introduced to make predictions. Data sets collected from Chengdu, Beijing and Shanghai from January to March 2024 are recorded as data sets 1, 2 and 3 respectively. The collected data includes multi-dimensional information such as daily vehicle flow, speed and traffic density at different time periods. The three datasets cover all major urban areas of the three cities, totaling more than 500,000 data records. The prediction accuracy and prediction time of KMH in three types of data sets are shown in Table V.

TABLE V. PREDICTION EFFECT OF KMH IN THREE TYPES OF DATA SETS

Data set	Prediction accuracy	Prediction time
Data set 1	98.59%	1.05min
Data set 2	97.24%	1.21min
Data set 3	98.70%	1.18min

Table V shows the prediction effect of KMH in three types of data sets. As can be seen from Table V, the prediction accuracy of KMH in dataset 1, dataset 2 and dataset 3 is 98.59%, 97.24% and 98.70%, respectively, and the prediction time is 1.05min, 1.21min and 1.18min, respectively. It can be seen that the KMH designed in this research has a good forecasting effect on the traffic flow data of different cities, which can prove that the method has a good scalability.

To sum up, in order to cope with the growing challenges of urban traffic management, this study not only proposed the theoretical improvement of K-means algorithm, but also tested its practical application effect. Finally, the improved weighted K-means method was specially designed to cope with the dynamic and complex traffic patterns in five urban areas of Chengdu. The predictive model can not only accurately identify traffic congestion patterns, help urban planners and traffic management departments to take forward-looking measures, but also improve traffic flow and reduce traffic congestion. Finally, the traffic prediction model is deployed in the traffic control center of the city to predict the peak traffic flow and formulate more effective dispersal strategies.

V. DISCUSSION

In order to improve the accuracy and practicability of urban traffic prediction, this study designed a traffic flow prediction model combining weighted K-means and Holt algorithm. Compared with the short-term traffic prediction method proposed by Cheng et al in literature [19], although its method has excellent performance in spatial-temporal pattern mining, it may have limitations when dealing with extreme traffic conditions and unconventional data. By introducing a weighting mechanism, this study effectively improves the adaptability and accuracy of the model in processing high-dimensional data and complex network environments. The results show that the prediction error of this method is significantly lower than that of the traditional method, and the average error is reduced by 20%, especially in the traffic prediction of peak hours and holidays. In addition, Liao and Li proposed a traffic anomaly detection model using k-means and active learning methods in literature [20], which has good performance on multi-level data sets. However, the model still has room for improvement in real-time and computational efficiency. The prediction model combined with weighted K-means and Holt algorithm adopted in this study, while maintaining a high accuracy, significantly improves the computational efficiency, making the model more suitable for real-time large-scale traffic data processing. Through ablation test, it is found that the performance of this research method on multiple traffic data sets is better than that of the

comparison model, especially in complex traffic scenarios, such as urban holidays and special events, its accuracy and response speed are significantly improved. In addition, the potential of the model in practical applications is also explored. For example, in the application test in Chengdu, the accuracy of the model in predicting the traffic flow during the peak period reached 98.5%, and the system response time was as low as 0.2 seconds, which has important reference value for the traffic management department to implement traffic control and diversion during the peak period. Finally, the CPU time of the method in this study is significantly lower than that of the traditional model when completing the traffic prediction task, which further validates its application efficiency and practicability in the actual traffic system.

In summary, by combining weighted K-means and Holt algorithm, this study proposes an efficient and accurate urban traffic flow prediction model. This not only provides new ideas and technical means for future urban traffic management but also has a positive impact on improving the overall efficiency and responsiveness of the urban traffic systems.

VI. Conclusion

The KMH algorithm was created in this work to complete the TFF task in an effort to enhance the performance of the existing TFF model even further. The results of the study indicated that the TDs of the five urban areas in Chengdu were analyzed as examples, and it was found that when the clustering number of the WKM algorithm was 3, the contour coefficients of Jinjiang and Wuhou districts reached the maximum values of 0.35 and 0.32, respectively. When the number of clusters was 2, the contour coefficients of Jinniu District, Qingyang District, and Chenghua District reached the maximum values of 0.23, 0.34, and 0.17, respectively, and at this time, the number of clusters just corresponded to the type of TC in each urban area, which indicated the spatial correlation of the TC patterns of the five urban areas. In addition, the average congestion index under different motor vehicle restriction numbers was also counted, and it was found that when the tail numbers were 4 and 9, the average congestion index was the largest, which was 3.42, and there was a statistically significant difference between this group of tail numbers and the other four groups of tail numbers ($P<0.05$). Finally, the TFF performance of the KMH algorithm was tested, and it was found that the prediction error of the KMH algorithm was as low as in the range of -0.1 to 0.1, and the TF under the prediction of the algorithm was basically the same as the actual situation. In summary, it can be concluded that the designed WKM algorithm can well analyze the clustering of TDs in space for the five urban areas, while the KMH algorithm is able to carry out accurate TFFs. Although the designed prediction method has a better performance, it should be followed up with a test of the method's prediction for TFs of other cities as a way of proving that the method has a better generalizability.

Acknowledgment

The research is supported by: School Level, Harbin Finance University of Jinyuan Scholar Support Program, (No. 900204).

References

[1] Annas M, Wahab S N. Data Mining Methods: K-Means Clustering Algorithms. International Journal of Cyber and IT Service Management, 2023, 3(1): 40-47.

[2] Purohit J, Dave R. Leveraging Deep Learning Techniques to Obtain Efficacious Segmentation Results. Archives of Advanced Engineering Science, 2023, 1(1): 11-26.

[3] Tian Z, Zhang S. Application of big data optimized clustering algorithm in cloud computing environment in traffic accident forecast. Peer-to-Peer Networking and Applications, 2021, 14(4): 2511-2523.

[4] Chen C, Liu Z, Wan S, Luan J, Pei Q. Traffic flow prediction based on deep learning in internet of vehicles. IEEE transactions on intelligent transportation systems, 2020, 22(6): 3776-3789.

[5] Li X, Gui J, Liu J. Data-driven traffic congestion patterns analysis: A case of Beijing. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(7): 9035-9048.

[6] Fernando C L, Yoshii T, Tsubota T. Combining the Deep Neural Network with the K-Means for Traffic Accident Prediction. International Journal of Computer and Systems Engineering, 2023, 17(1): 1-8.

[7] Nguyen T H T, Dinh D T, Sriboonchitta S, Huynh V N. A method for k-means-like clustering of categorical data. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(11): 15011-15021.

[8] Daviran M, Ghezelbash R, Niknezhad M, Maghsoudi A, Ghaeminejad H. Hybridizing K-means clustering algorithm with harmony search and artificial bee colony optimizers for intelligence mineral prospectivity mapping. Earth Science Informatics, 2023, 16(3): 2143-2165.

[9] Chen R, Wang S, Zhu Z, Yu J, Dang C. Credit ratings of Chinese online loan platforms based on factor scores and K-means clustering algorithm. Journal of Management Science and Engineering, 2023, 8(3): 287-304.

[10] Sun Z, Hu Y, Li W, Feng S, Pei L. Prediction model for short-term traffic flow based on a K-means-gated recurrent unit combination. IET Intelligent Transport Systems, 2022, 16(5): 675-690.

[11] Wang Y, Jing C, Huang W, Jin S, Lv X. Adaptive spatiotemporal inceptionnet for traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3882-3907.

[12] Huo G, Zhang Y, Wang B, Gao J, Hu Y, Yin B. Hierarchical spatio–temporal graph convolutional networks and transformer network for traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3855-3867.

[13] Li H, Yang S, Song Y, Luo Y, Li J, Zhou T. Spatial dynamic graph convolutional network for traffic flow forecasting. Applied Intelligence, 2023, 53(12): 14986-14998.

[14] Liu J, Kang Y, Li H, Wang H, Yang X. STGHTN: Spatial-temporal gated hybrid transformer network for traffic flow forecasting. Applied Intelligence, 2023, 53(10): 12472-12488.

[15] Doğan E. Short-term traffic flow prediction using artificial intelligence with periodic clustering and elected set. Promet-Traffic & Transportation, 2020, 32(1): 65-78.

[16] Wang Y, Jing C, Huang W, Jin S, Lv X. Adaptive spatiotemporal inceptionnet for traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3882-3907.

[17] Li H, Yang S, Song Y, Luo Y, Li J, Zhou T. Spatial dynamic graph convolutional network for traffic flow forecasting. Applied Intelligence, 2023, 53(12): 14986-14998.

[18] Liu J, Kang Y, Li H, Wang H, Yang X. STGHTN: Spatial-temporal gated hybrid transformer network for traffic flow forecasting. Applied Intelligence, 2023, 53(10): 12472-12488.

[19] Cheng S, Lu F, Peng P. Short-term traffic forecasting by mining the non-stationarity of spatiotemporal patterns. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(10): 6365-6383.

[20] Liao N, Li X. Traffic anomaly detection model using k-means and active learning method. International Journal of Fuzzy Systems, 2022, 24(5): 2264-2282.