Journal of Ambient Intelligence and Humanized Computing (2023) 14:9035–9048
https://doi.org/10.1007/s12652-022-04409-4

ORIGINAL RESEARCH

Data-driven traffic congestion patterns analysis: a case of Beijing

Xiang Li[1] · Jiao Gui[1] · Jiaming Liu[2]

Received: 22 September 2021 / Accepted: 12 September 2022 / Published online: 27 September 2022

Abstract
With the rapid increase of urban population and the number of motor vehicles, the traffic congestion problem is becoming more and more serious in megacities. This paper aims to identify the traffic congestion patterns and analyze their spatial–temporal variations by conducting cluster analysis for daily traffic congestion index curves. First, since the importance of sampling points in different time segments is not exactly the same, the coefficient of variation is taken to assign weight for improving K-means clustering algorithm. The improved weighted K-means clustering algorithm is proposed to identify the representative traffic congestion patterns. Second, the paired t-test method is used to analyze the spatial and temporal variations of traffic congestion patterns. Finally, case studies are conducted based on the real-life traffic congestion index data in Beijing, including over 670, 000 records covering six districts from January 1, 2017 to December 31, 2017. The results illustrate that traffic congestion patterns are both temporal dependent and spatial dependent, and the automobile license plate restriction has significant influence on the traffic congestion patterns. This study could be instructive for formulating specific traffic optimization and control schemes to alleviate congestion and promote the balance of traffic conditions.

Keywords Traffic congestion pattern · Traffic congestion index · Weighted K-means clustering · Paired *t*-test

1 Introduction

Traffic congestion is a phenomenon that the load of urban roads exceeds its specified capacity of traffic system, which especially occurs in commute peaks and poor weather conditions. Traffic congestion is a bane of urban life, especially in megacities, which significantly increases the travel cost for residents (Ke et al. 2020), causes more traffic accidents (Retallack et al. 2019), and makes traffic management extremely difficult (Praveen et al. 2021).

With a large number of data acquisition equipment densely distributed in the road network, it is possible to assess traffic characteristics by using the collected high-volume, real-time, and high-accuracy data from multiple and autonomous sources (Wu et al. 2014). In the case of traffic congestion, this information includes GPS data, map application data, data from massive sensors, and so on. Big data offers advantages over conventional data sources in terms of volume, velocity, variety, and veracity (Yaqoob et al. 2016). It can reveal some potential insights of smart cities after effective research and analysis (Chauhan et al. 2016). Therefore, it is highly important to analyze traffic congestion and its characteristics via big data technology for better traffic control and management.

Traffic congestion patterns refer to the data curves of traffic congestion index in one day with different curve characteristics (Zhao and Hu 2019). The grasp of urban traffic congestion patterns and their spatial–temporal evolution characteristics is instrumental to the accurate prediction of traffic situation and information provision for urban residents to optimize their daily travel decisions. From the macro perspective of urban management, it can provide the basis for road construction and city planning (Torkjazi et al. 2018). At the same time, understanding the evolution trend of urban traffic situation is helpful to judge and forecast the level and direction of the regional economic development (Li et al. 2019).

The analytical framework of this study is shown in Fig. 1. First, a linear interpolation method is used to fill in the missing values and a 2-sigma rule is used to identify and modify the outliers for data preprocessing. Second, an improved

✉ Xiang Li
lixiang@mail.buct.edu.cn

[1] School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China

[2] School of International Economics and Management, Beijing Technology and Business University, Beijing 100048, China

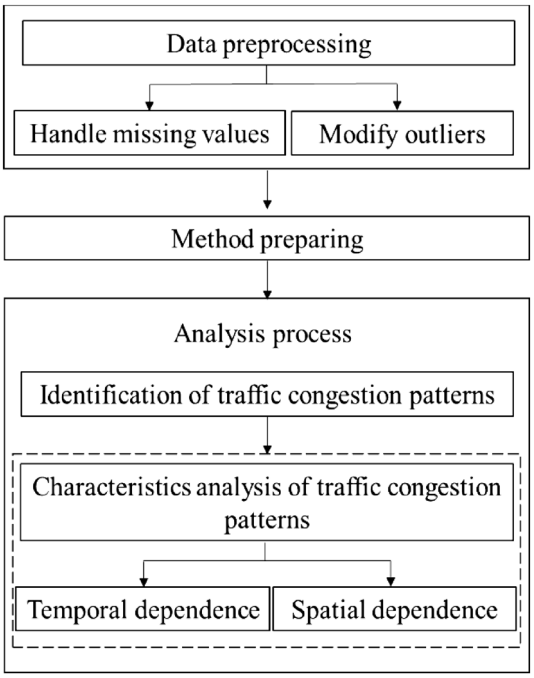

Fig. 1 The analytical framework of this work

weighted K-means clustering method is proposed to identify the traffic congestion patterns, which takes a weighting operation among the daily traffic congestion index data before conducting the clustering process. Finally, the spatial–temporal variations of traffic congestion patterns arising from the space difference, the time difference, and the automobile license plate restriction are analyzed.

The main contributions of this research are summarized as follows. By modifying the typical K-means clustering method, a novel clustering method on time series data is proposed to identify the traffic congestion patterns. Based on the real-life traffic congestion index data in Beijing, the paired t-test is carried out, it is revealed that the traffic congestion patterns are both spatial dependent (there are significant differences in the number and shape of traffic congestion patterns in different regions) and temporal dependent (the variations of dates and automobile license plate restriction both impact the traffic congestion patterns). This work strengthens the understanding on urban traffic congestion patterns and their spatial–temporal characteristics, which is helpful for the accurate prediction of traffic situation and the precise decision for traffic operations management.

The remainder of this work is organized as follows. Section 2 reviews the related researches. Section 3 describes the traffic congestion index data in Beijing. Section 4 introduces the data preprocessing process. Section 5 presents methodologies using in this research. Section 6 shows the identification results on traffic congestion patterns, and the spatial–temporal characteristics of traffic congestion patterns. Section 7 concludes the paper with a brief summary and gives some potential directions for future research.

2 Literature review

In recent years, scholars pay more and more attention to traffic congestion forecasting from the perspectives of traffic flow (Angayarkanni et al. 2021), traffic velocity (Jiang et al. 2021), delay time (Shelke et al. 2019), traffic cost (Tian et al. 2010), traffic congestion index (Wang et al. 2018), and so on. Here traffic congestion index is a comprehensive and integrated indicator, which is defined as a conceptual value that could synthetically reflect the traffic conditions. A higher traffic congestion index corresponds with the heavier traffic congestion. Su et al. (2019) considered total number of vehicles in the system varying over time and proposed a dynamic stochastic differential model to describe traffic flow based on the Markov chain theory. By using traffic flow data from the I-80 Freeway Dataset from the NGSIM program, it showed that the proposed approach provided more accurate predictions of traffic flow. Sanchez-Cambronero et al. (2017) took advantage of the plate scanning technique to propose an algorithm that minimizes the required number of registering devices and their location in order to identify vehicles candidates to compute and predict the travel times of a given set of routes (or sub routes). Wang et al. (2017) pointed out that the PageRank values can act as signals in predicting upcoming traffic congestions, and observed the aforementioned laws experimentally based on the trajectory data of 12,000 taxies in Beijing city for one month.

In addition, the existing literature has also carried out a large number of analyses on the traffic congestion characteristics. For example, by computing urban traffic evolution on temporal complex network with PageRank, Wang et al. (2017) found the congestion degree of a local region is not only affected by the traffic states of its neighboring regions but also those of the whole network. ShirMohammadi et al. (2020) analyzed the traffic density, congestion index and peak hours for the main network of Hamedan communication routes based on the collected data of speed performance, and simulated the relationship between traffic velocity and congestion index by using neural network and genetic algorithm. Kan et al. (2019) proposed a traffic feature analysis and classification approach to detect traffic congestion from taxis' GPS trajectories at the turn level. The case study in Wuhan supported the feasibility of this approach and proved that the proposed approach can sense traffic congestion at a lower cost compared with other approaches. Chen et al. (2021a, b) proposed a new categorization criterion to define traffic conditions as five levels based on speed performance index values, and applied the proposed criterion in a case study to investigate the daily curve of speed performance

Table 1 Comparative analysis of related works

Research direction	Literature	Data	Methodology
Traffic congestion forecasting	Angayarkanni et al. (2021)	Traffic flow data in California	Hybrid SVR-GWO-BES
	Jiang et al. (2021)	Traffic velocity data in Guizhou	Hybrid model named S-GCN-GRU-NN
	Shelke et al. (2019)	Delay time data in Network Simulator Version-2	Fuzzy priority based congestion-aware routing algorithm
	Tian et al. (2010)	Traffic speed data in simulation	Travel cost functions
	Wang et al. (2018)	Traffic congestion index data in Dalian city	Fuzzy mathematics
Analysis of traffic congestion characteristics	Wang et al. (2017)	Taxis' GPS trajectories data in Beijng	PageRank algorithm
	ShirMohammadi et al. (2020)	Speed performance data in Hamedan	Neural network and genetic algorithm
	Kan et al. (2019)	Taxis' GPS trajectories data in Wuhan	Traffic feature analysis and classification approach
	Chen et al. (2021a, b)	Floating car data in Beijing	Dissimilarity analysis
	Kim et al. (2019)	Traffic congestion data and activity-travel data in Los Angeles	Probability distribution function
Analysis of traffic congestion pattern	Wen et al. (2014)	TPI data in Beijing	Hierarchical clustering method
	Sun et al. (2019)	TPI data of five business circles in Qingdao	Hierarchical clustering algorithm
	Zhao et al. (2019)	Traffic congestion index data in Beijing	K-means cluster analysis
	Our work	Traffic congestion index data in Beijing	Weighted K-means clustering algorithm and paired t-test method

index data in Beijing. It was found that the curves vary significantly in shape on different days. Some research also detected traffic congestion characteristics from other perspectives, the results illustrated that there are significant differences across days (Kim et al. 2019).

Traffic congestion patterns refer to the data curves of traffic congestion index in one day with different curve characteristics (Zhao and Hu 2019). The existing literature mainly concerns on the traffic congestion forecasting and traffic congestion characteristics analysis, while the analysis on traffic congestion pattern is very rare. Wen et al. (2014) selected eight evaluation indices on traffic congestions in morning and evening peak hours, and then proposed a hierarchical clustering analysis method to divide the pattern characteristics of traffic congestions. The results revealed that weekdays included Normal Weekdays, Key Congested Weekdays, and Most Congested Weekdays. Sun et al. (2019) adopted hierarchical clustering algorithm to study the congestion patterns in Qingdao based on traffic performance index (TPI) data. The results showed that there were three categories of traffic congestion pattern: Workdays, Latter half of vacation (October 4th–8th), and Weekends and the beginning of vacation (October 1st–3rd). Based on the macro traffic congestion index data in Beijing, Zhao and Hu (2019) revealed that there were two typical traffic congestion patterns on weekdays by applying K-means cluster analysis, i.e., *weekday mode A* and *weekday mode B*. The former often appeared on Mondays and the main characteristic was the obvious morning peak and evening peak with similar congestion duration, while the latter often appeared on Fridays and the main characteristic was that the peak and duration of congestion in the evening were significantly higher than in the morning.

The above research has enlightening significance for the urban traffic management at the strategic level, but does not answer the following questions at the operational level: (1) whether the traffic congestion pattern is spatial dependent, should we carry out spatially differentiated traffic congestion management? (2) whether the traffic congestion pattern is temporal dependent, should we carry out temporally differentiated traffic congestion management? The motivation of this research is to answer these questions by using the micro traffic congestion index data, which could provide more valuable information for traffic management, planning and policy-making. Comparative analysis between the existing literature and this research is shown in Table 1.

3 Traffic congestion index data

Traffic congestion index is a conceptual value that can synthetically reflect the road traffic conditions (Zhao et al. 2019), which has been widely studied as an urban traffic

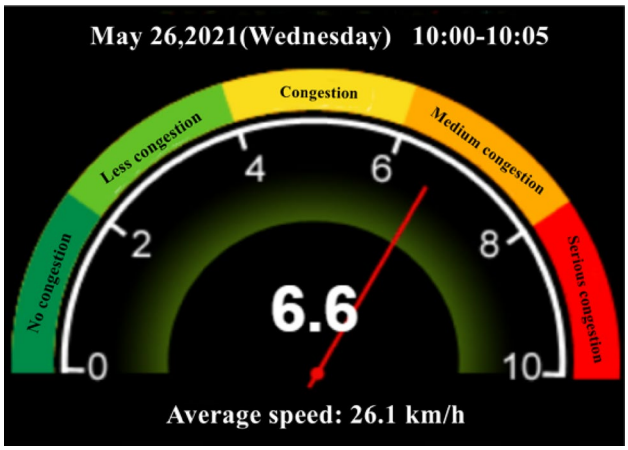

Fig.2 Real-time traffic congestion index in Beijing

situation indicator in literature (Wen et al. 2014; Sun et al. 2019; ShirMohammadi et al. 2020). As the capital of China, Beijing is a typical megacity with permanent residents 21.89 million and motor vehicles 6.57 million by the end of 2020. The road network structure of Beijing is a ring road system radiating urban districts. Since 2006, Beijing has established traffic congestion index as the core evaluation indicator of traffic conditions, and publishes the real-time traffic congestion index to the public through the Internet and APPs [1]. As shown in Fig. 2, the traffic conditions are divided into five grades as the traffic congestion index R ranges from 0 to 10, that is, no congestion ($0 \le R < 2$), less congestion ($2 \le R < 4$), congestion ($4 \le R < 6$), medium congestion ($6 \le R < 8$), and serious congestion ($8 \le R \le 10$). The higher the traffic congestion index, the heavier the traffic congestion (Wang et al. 2017). The traffic congestion index was 6.6 at time segment [10:00, 10:05) on May 26, 2021, which belongs to medium congestion. In this study, we collect the traffic congestion index data from January 1, 2017 to December 31, 2017, including over 670, 000 records covering six urban districts (Dongcheng, Xicheng, Chaoyang, Haidian, Fengtai, and Shijingshan). The sampling step of recorded data is five minutes, which means that the system records one piece of data each five minutes. As a result, the whole day (0:00–24:00) is partitioned into 288 time segments, and the length of each time segment is 5 min. An example of the recorded data is shown in Table 2.

[1] [1] http://jtw.beijing.gov.cn/

Table 2 An example of traffic congestion index data in Beijing

ID	District	Congestion Index	Date	Time segment
58990400	Dongcheng	5.3	June 18, 2017	[7:00, 7:05)
58990401	Dongcheng	5.3	June 18, 2017	[7:05, 7:10)
58990402	Dongcheng	5.3	June 18, 2017	[7:10, 7:15)
58990403	Dongcheng	7.5	June 18, 2017	[7:15, 7:20)
58990404	Dongcheng	7.5	June 18, 2017	[7:20, 7:25)
58990405	Dongcheng	7.5	June 18, 2017	[7:25, 7:30)
58990406	Dongcheng	8.6	June 18, 2017	[7:30, 7:35)
58990407	Dongcheng	8.6	June 18, 2017	[7:35, 7:40)
58990408	Dongcheng	8.6	June 18, 2017	[7:40, 7:45)
58990409	Dongcheng	8.9	June 18, 2017	[7:45, 7:50)

Table 3 An example of missing value

ID	District	Congestion index	Date	Time segment
61094646	Xicheng	6.1	December 16, 2017	[7:30, 7:35)
61094585	Xicheng	7.7	December 16, 2017	[7:35, 7:40)
61094657	Xicheng	7.7	December 16, 2017	[7:40, 7:45)
61094676	Xicheng	7.7	December 16, 2017	[7:45, 7:50)
61094687	Xicheng	Nan	December 16, 2017	[7:50, 7:55)
61094668	Xicheng	8.1	December 16, 2017	[7:55, 8:00)
61094698	Xicheng	8.1	December 16, 2017	[8:00, 8:05)
61094709	Xicheng	8.2	December 16, 2017	[8:05, 8:10)
61094720	Xicheng	8.2	December 16, 2017	[8:10, 8:15)
61094731	Xicheng	8.2	December 16, 2017	[8:15, 8:20)

4 Data preprocessing

Due to mechanical failure or human error, there are missing values and outliers in the raw traffic congestion index data inevitably. As a result, data preprocessing is necessary before conducting the data analysis process. For the Beijing traffic congestion index data, the ratio of missing values is around 2.13% and among which more than 80% appear as single missing value. An example on the phenomena of missing value is shown in Table 3, in which the sample data with ID 61,094,687 takes congestion index value "Nan", indicating that the congestion index data at time segment [7:50, 7:55) is missing. An outlier in a dataset is an observation with value far away from other observations. In Table 4, the traffic congestion index between 3:45 a.m. and 3:50 a.m. in Shijingshan district is generally less than 1.5 from January 9, 2017 to January 18, 2017, while it suddenly rises to 9.8 on January 13, 2017, which could be considered as an outlier.

Table 4 An example of outlier

ID	District	Congestion Index	Date	Time segment
59016814	Shijingshan	1.2	January 9, 2017	[3:45, 3:50)
59019680	Shijingshan	0.4	January 10, 2017	[3:45, 3:50)
59022831	Shijingshan	0.8	January 11, 2017	[3:45, 3:50)
59026052	Shijingshan	0.8	January 12, 2017	[3:45, 3:50)
59029265	Shijingshan	9.8	January 13, 2017	[3:45, 3:50)
59032454	Shijingshan	1.1	January 14, 2017	[3:45, 3:50)
59035664	Shijingshan	0.6	January 15, 2017	[3:45, 3:50)
59038868	Shijingshan	1.2	January 16, 2017	[3:45, 3:50)
59042031	Shijingshan	1.2	January 17, 2017	[3:45, 3:50)
59045267	Shijingshan	1.0	January 18, 2017	[3:45, 3:50)

The missing values and outliers in the time series may distort the shape of traffic congestion patterns, therefore filling in the missing values and modifying the outliers should be performed first.

In literature, a great number of methods have been developed for filling in the missing values, in which linear interpolation method (Lu et al. 2003) is always used to tackle with the cases with small range of missing values, while empirical orthogonal function (Beckers et al. 2003), Gamma distribution function (Simolo et al. 2009), and autoregressive model (Kim et al. 2015) are more appropriate for dealing with the cases with large range of missing values. For the traffic congestion index data, since only single or small range of missing values are identified, the linear interpolation method is taken, which has been widely used in the preprocessing of transportation data analysis (Degen et al. 2007; Zhao et al. 2019; Sun et al. 2021). Assume that missing values are detected at successive time segment $i = 1, 2, \cdots, I$, while x_0 is the recorded congestion index at time segment $i = 0$ and x_{I+1} is observed congestion index at time segment $i = I + 1$. The linear interpolation method approximates the missing values x_i as follows

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

Taking Table 3 for example, there is only one missing value detected at time segment [7:50, 7:55). In this case, we have $I = 1$ and $x_1 = \left(x_0 + x_2\right)/2$. The filled congestion index at time segment [7:50, 7:55) should be $(7.7 + 8.1)/2 = 7.9$.

The detection and modification of outliers in time series are the key steps for data preprocessing. The existing methods mainly include 2-sigma rule (Li et al. 2015), 3-sigma rule (Klos et al. 2015), maximum likelihood estimation (Lee et al. 2006), Bayesian method (Kruschke et al. 2012), and multilevel model (Shi et al. 2008). For detecting the outliers sensitively and modifying them expediently, the 2-sigma rule is employed to handle the traffic congestion index data before feeding them into the clustering analysis algorithm, which has been widely used in transportation data analysis and achieved good performance (Li et al. 2015). Denote the daily time series data of traffic congestion index as an M-dimensional vector, where M is the amount of sampling points each day. For example, if the sampling step is 5 min, then there are 12 sampling points each hour and 288 sampling points each day. In this case, we have $M = 288$. Denote N as the number of observation days, x_n^m is observed congestion index at sampling time segment m on the n^{th} day. Then the traffic congestion index can be written as

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

The intraday trend $\overline{X}$ among these observation days, which represents the average value of daily traffic congestion index, can be formulated as

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

The residual fluctuations of the n^{th} day are

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

Finally, the sample standard deviation σ^m is calculated as the square root of $r_1^m, r_2^m, \cdots, r_N^m$ with $m = 1, 2, \cdots, M$. A point x_n^m is defined as an outlier if the absolute residual $\left|r_n^m\right|$ is greater than twice of the sample standard deviation σ^m.In this case, the observation value x_n^m is modified as $\bar{x}^m - 2\sigma^m$ or $\bar{x}^m + 2\sigma^m$. Otherwise, it is regarded as a regular point and its value should keep unchanged. The outlier detection and modification procedure is exhibited as follows:

$$x_n^m = \begin{cases} \bar{x}^m + 2\sigma^m, & \text{if } r_n^m > 2\sigma^m \\ x_n^m, & \text{if } -2\sigma^m \le r_n^m \le 2\sigma^m \\ \bar{x}^m - 2\sigma^m, & \text{if } r_n^m < -2\sigma^m \end{cases}, \ \forall\, n = 1, 2, \cdots, N,\ m = 1, 2, \cdots, M. \tag{5}$$

5 Research methodologies

Traffic congestion patterns refer to the data curves of traffic congestion index in one day with different curve characteristics (Zhao and Hu, 2019). In this section, an improved weighted K-means clustering method is proposed to identify

traffic congestion patterns, and paired *t*-test method is introduced to analyze the temporal and spatial dependence.

5.1 Improved weighted K-means clustering method

Time series data is multi-dimensional, dynamic and temporal-dependent. Although time series data is composed of multiple data samples connected by time points, it can also be expressed as a single object to be clustered in the form of column vector. Assume that $D = \{X_1, X_2, \cdots, X_N\}$ is a set of time series data where X_n represents a column vector. The target of time series clustering is to divide the given set into K different types of clusters represented as $C = \{C_1, C_2, \cdots, C_K\}$ in an unsupervised way, where C_k is defined as the k^{th} cluster and $D = \bigcup_{k=1}^{K} C_k$.

K-means clustering method uses iterative process to partition a collection of sampling points into subsets known as clusters (Li et al. 2012; Yang et al. 2018; Xu et al. 2020; Chen et al. 2021a, b). Assume that there are K clusters in the sample dataset, the target of K-means clustering is to minimize the total deviation

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

where $X_n = \left(x_n^1, x_n^2, \cdots, x_n^M\right)$ represents an M-dimension sample, and $U_k = \left(u_k^1, u_k^2, \cdots, u_k^M\right)$ is an M-dimension vector representing the center of cluster k, which is calculated as

$$u_k^m = \frac{1}{|C_k|} \sum_{X_n \in C_k} x_n^m, \ \forall m = 1, 2, \cdots, M, k = 1, 2, \cdots, K. \tag{7}$$

A cluster contains the cluster center and the data samples assigned to it. Each time a data sample is allocated, the cluster center will be recalculated according to the existing objects in the cluster. This process will be repeated until the termination condition is satisfied. The termination condition can be that the cluster centers keep unchanged or the sum of squares of errors is local minimum. Due to its good performance and computing efficiency, the K-means clustering algorithm has been widely used in the field of transportation data analysis (Zhao et al. (2019); Sun et al. 2021).

Based on the preprocessed time series data of traffic congestion index, an *improved weighted K-means clustering method* is proposed to identify traffic congestion patterns, which assigns differential weights among all M sampling time segments. Specifically, the sampling time segments with higher dispersion among daily congestion index are assigned with greater weights to strengthen their role in the clustering process. Conversely, the sampling time segments with lower dispersion among daily congestion index are assigned with smaller weights to weaken their influence. Here the *Coefficient of Variation* is taken (Arachchige et al. 2020) to measure the degree of dispersion, that is,

$$CV_m = \frac{\sigma_m}{\bar{x}_m}, \ \forall m = 1, 2, \cdots, M, \tag{8}$$

$$\bar{x}^m = \frac{1}{N} \sum_{n=1}^{N} x_n^m, \ \forall m = 1, 2, \cdots, M, \tag{9}$$

$$\sigma_m = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(x_n^m - \bar{x}^m\right)^2}, \ \forall m = 1, 2, \cdots, M, \tag{10}$$

where CV_m represents the degree of dispersion at sampling time segment m, $\bar{x}^m$ represents the sample mean of x_n^m at sampling time segment m, σ^m represents the sample standard deviation x_n^m at sampling time segment m. The coefficient of variation is an appropriate weight selection, which considers the stability and volatility of time series data at the same time.

Based on the value of CV_m, a weighted K-means clustering method is proposed to partition the time series data of traffic congestion index. The objective is to minimize the total weighted deviations to the cluster centers

$$\sum_{k=1}^{K} \sum_{X_n \in C_k} \sum_{m=1}^{M} \left(CV_m x_n^m - u_k^m\right)^2, \tag{11}$$

where $X_n = \left(x_n^1, x_n^2, \cdots, x_n^M\right)$ represents the time series data with M sampling points in a day, and $U_k = \left(u_k^1, u_k^2, \cdots, u_k^M\right)$ represents the weighted center of cluster k, which is defined as

$$u_k^m = \frac{1}{|C_k|} \sum_{X_n \in C_k} CV_m x_n^m, \ \forall m = 1, 2, \cdots, M. \tag{12}$$

For determining the best number of clusters, i.e., the value of K, *Silhouette Coefficient* (Rousseeuw, 1987) is taken to evaluate the clustering performance associated with each value of K, and the one that maximizes the clustering performance is selected. First, for each sample X_n, its Silhouette Coefficient is defined as

$$s\left(X_n\right) = \frac{b_n - a_n}{\max\left\{a_n, b_n\right\}}, \tag{13}$$

where a_n represents the average Euclidean distance between sample X_n and all the other samples in its cluster, and b_n represents the average Euclidean distance between sample X_n and all samples in its nearest cluster. Note that the Silhouette

Table 5 The optimal number of clusters at different districts

District	Silhouette Coefficient with different number of clusters									The optimal number of clusters
	2	3	4	5	6	7	8	9	10	
Haidian	0.26	0.21	0.22	0.18	0.19	0.18	0.16	0.15	0.17	2
Fengtai	0.35	0.19	0.21	0.22	0.18	0.15	0.15	0.14	0.13	2
Shijingshan	0.16	0.11	0.05	0.05	0.05	0.04	0.03	0.05	0.04	2
Chaoyang	0.32	0.34	0.25	0.21	0.16	0.17	0.17	0.16	0.15	3
Dongcheng	0.28	0.30	0.19	0.15	0.13	0.14	0.13	0.12	0.11	3
Xicheng	0.27	0.29	0.20	0.15	0.13	0.14	0.14	0.13	0.13	3

Coefficient works while the number of clusters is more than or equal to two, i.e., $K \geq 2$. Second, the Silhouette Coefficient for the whole dataset is defined as the mean Silhouette Coefficient among all samples, that is,

$$S = \frac{s(X_1) + s(X_2) + \cdots + s(X_N)}{N}, \tag{14}$$

which takes value in $[-1, 1]$. The closer the value is to 1, the better the clustering results.

Based on the above description, the general procedure for such a weighted K-means clustering method is summarized in Algorithm 1.

5.2 Paired *t*-test method

The paired *t*-test method is used to test whether the average difference between two set of paired sample data is zero. It can also be used in making observations on the same event under different conditions, in order to evaluate the influence of conditions on the event (Konietschke et al. 2014). The test is based on the difference between the values of a single pair denoted as $\{d_1, d_2, \cdots, d_L\}$, and the test statistic value t is calculated as

$$t = \frac{\sum_{l=1}^{L} d_l}{\sqrt{\frac{L\left(\sum_{l=1}^{L} d_l^2\right) - \left(\sum_{l=1}^{L} d_l\right)^2}{L-1}}}, \tag{15}$$

Algorithm 1: Weighted K-means clustering algorithm

1 Given a set of sample data $\{X_1, X_2, \cdots, X_N\}$, calculate the coefficient of variation CV_m with $m = 1, 2, \cdots, M$.

2 Initialize the largest number of clusters $K_{\max}$, and set $K = 2$.

3 Select K samples randomly from D as the initial clustering centers.

4 Assign all samples to clusters by minimizing the weighted deviations

$$\sum_{k=1}^{K} \sum_{X_n \in C_k} \sum_{m=1}^{M} \left(CV_m x_n^m - u_k^m\right)^2.$$

5 Recalculate the mean of samples in each cluster

$$u_k^m = \frac{1}{|C_k|} \sum_{X_n \in C_k} CV_m x_n^m, \ \forall\ m = 1, 2, \cdots, M,$$

and then update the cluster centroid to this new location.

6 Repeat Step 4 to Step 5 until the centroids stop moving.

7 Calculate the value of Silhouette Coefficient, and set $K = K + 1$.

8 Repeat Step 3 to Step 7 until $K > K_{\max}$. Return the best number of clusters and the corresponding clustering results.

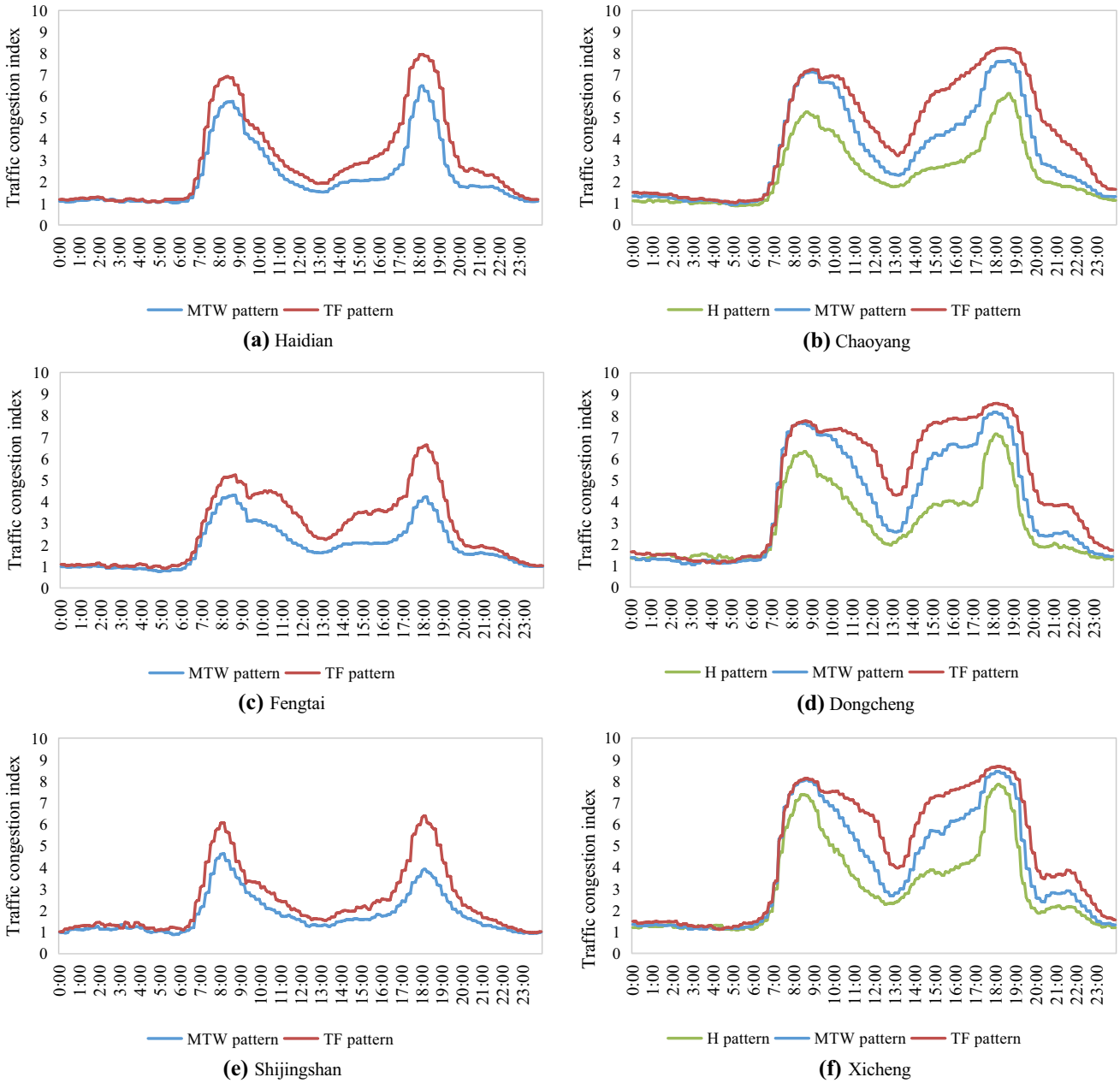

Fig. 3 The traffic congestion patterns at different districts

where L is the number of observations in a set of sample data.

If the two tailed P value that corresponds to the test statistic t with L-1 degrees of freedom is less than the chosen significance level (e.g., 0.10, 0.05, and 0.01), it indicates that the difference is significant between the two set of sample data (Hong et al. 2017).

6 Case studies

In this section, case studies are exhibited in details. First, traffic congestion patterns at different districts in Beijing are identified based on the proposed weighted K-means clustering method. Second, the temporal dependence of traffic congestion patterns is examined using paired t-test method.

Table 6 Statistical results of congestion patterns across weekdays at different districts

Weekday		Mon	Tues	Wed	Thur	Fri	Total
Haidian	MTW	42	32	34	10	2	120
	TF	4	16	16	40	48	124
Fengtai	MTW	45	44	42	24	21	176
	TF	1	4	8	26	29	68
Shijingshan	MTW	35	34	43	21	10	143
	TF	11	14	7	29	40	101
Chaoyang	H	8	7	11	9	4	39
	MTW	38	38	27	16	7	126
	TF	0	3	12	25	39	79
Dongcheng	H	11	4	9	5	5	34
	MTW	35	28	24	12	0	99
	TF	0	16	17	33	45	111
Xicheng	H	13	8	11	7	5	44
	MTW	33	33	31	12	6	115
	TF	0	7	8	31	39	85

Finally, the spatial dependence of traffic congestion patterns is tested by analyzing the indicator values of different traffic congestion patterns at different districts.

6.1 Identification of traffic congestion patterns

Based on the weighted K-means clustering method, this subsection identifies the traffic congestion patterns at different districts in Beijing. The Silhouette Coefficients with different number of clusters are shown in Table 5, which imply that the traffic congestion patterns are spatial dependent, the districts closer to the downtown (i.e., Chaoyang, Dongcheng, and Xicheng) have three categories of congestion patterns; while the districts far away from the downtown (i.e., Haidian, Fengtai, and Shijingshan) have two categories of congestion patterns.

First, taking the clustering results in Haidian district as an example. In Fig. 3a, there are two representative traffic congestion patterns on weekdays. The first pattern is less congested, which often appears at the first-half or middle of the weekdays (i.e., Monday, Tuesday, and Wednesday), while the second pattern is more congested, which generally appears at the last-half of the weekdays (i.e., Thursday and Friday). For simplicity, they are hereinafter respectively named as *MTW pattern* and *TF pattern*. The main difference between them occurs from 6:00 to 23:00, and they are relatively consistent from 0:00 to 6:00. The trends of these two curves are basically the same, namely the morning peaks appear around 8:00 and the evening peaks appear around 18:00, but TF pattern obviously takes higher values than MTW pattern. The characteristics of traffic congestion patterns in Fengtai district and Shijingshan district are similar to those in Haidian district which are shown in Fig. 3c, e.

Differently, there are three traffic congestion patterns on weekdays in Chaoyang district. In Fig. 3b, except the MTW pattern and TF pattern, there is a holiday pattern, which will be hereinafter named as *H pattern*. Compared with the MTW and TF patterns, the H pattern is the least congested, which often appears on working days within 3 days before and after holidays (e.g., the Spring Festival, the Mid-Autumn Festival, the National Day, the Dragon Boat Festival). Similarly, the trends of these three patterns are basically the same, but their values differ significantly. The characteristics of traffic congestion patterns in Dongcheng district and Xicheng district are similar to those in Chaoyang district which are shown in Fig. 3d, f.

Zhao and Hu (2019) revealed that there were two typical traffic congestion patterns on weekdays in Beijing by applying K-means cluster analysis, i.e., weekday mode A and weekday mode B. The former often appeared on Mondays while the latter often appeared on Fridays. As shown in Table 5, when the number of congestion pattern clusters is two at Chaoyang, the corresponding Silhouette Coefficient is 0.32, which does not reach its optimal value. Essentially, the Silhouette Coefficient reaches its optimal value 0.34 when the number of congestion pattern clusters is three at Chaoyang. Similar clustering results also occur in Dongcheng and Xicheng. Therefore, it is more reasonable to divide the traffic congestion patterns of Chaoyang, Xicheng, and Dongcheng into three categories rather than two categories.

6.2 Temporal dependence of traffic congestion patterns

Now the temporal dependence of traffic congestion patterns can be examined. As described in Sect. 4.3, the clustering results of traffic congestion pattern are consistent with the

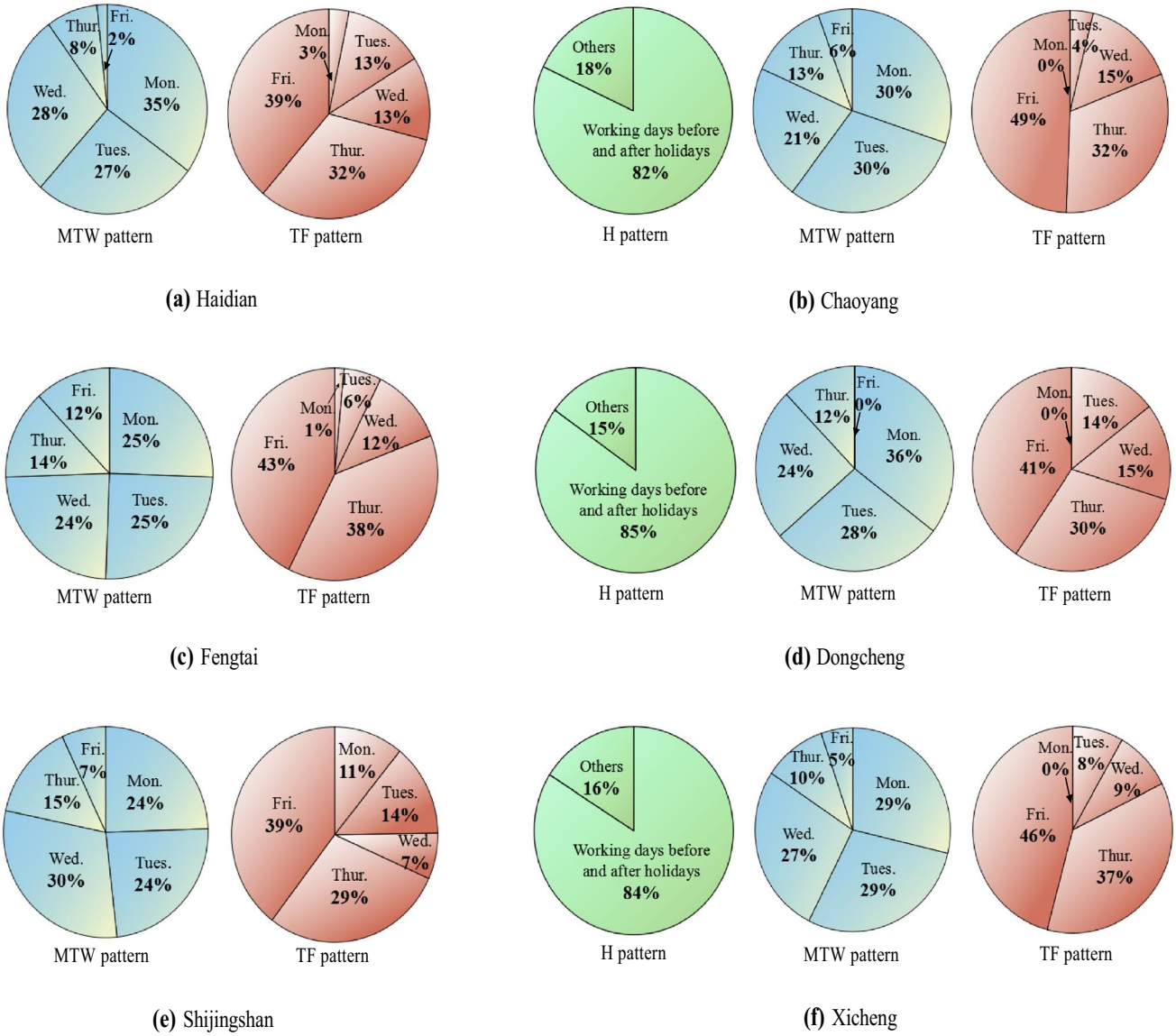

Fig. 4 Proportions of Mon., Tue., Wed., Thur., and Fri. across different congestion patterns

variation of dates, which is also how each pattern is named. Taking Haidian district as an example (See Table 6 and Fig. 4a), the MTW pattern includes 42 Mondays, 32 Tuesdays, 34 Wednesdays, 10 Thursdays and 2 Fridays, the total proportion of Mondays, Tuesdays and Wednesdays is 90%, and the total proportion of Thursdays and Fridays is only 10%; while the TF pattern includes 4 Mondays, 16 Tuesdays, 16 Wednesdays, 40 Thursdays and 48 Fridays, the proportion of Thursdays and Fridays is 71%, and the proportion of Mondays, Tuesdays and Wednesdays is 29%.

Similarly, in Chaoyang district (See Table 6 and Fig. 4b), the MTW pattern includes 38 Mondays, 38 Tuesdays, 27 Wednesdays, 16 Thursdays and 7 Fridays, the total proportion of Mondays, Tuesdays and Wednesdays is 81%, and the total proportion of Thursdays and Fridays is only 19%; while the TF pattern includes 0 Mondays, 3 Tuesdays, 12 Wednesdays, 25 Thursdays and 39 Fridays, the total proportion of Thursdays and Fridays is 81%, and the total proportion of Mondays, Tuesdays and Wednesdays is only 19%; the H pattern is evenly distributed from Monday to Friday, but the most significant distribution characteristic for this pattern is that it includes 32 working days within 3 days before and after holidays, and it accounts for 82% of the total number of days in H pattern.

Based on the above analysis results, it is concluded that the variation of dates greatly impacts the congestion

Table 7 Statistical results of congestion patterns across no-drive days

Restriction number		0 and 5	1 and 6	2 and 7	3 and 8	4 and 9	Total
Haidian	MTW	27	28	23	27	15	120
	TF	21	22	27	21	33	124
Fengtai	MTW	37	44	34	43	18	176
	TF	11	6	16	5	30	68
Shijingshan	MTW	32	30	27	30	24	143
	TF	16	20	23	18	24	101
Chaoyang	H	10	10	5	9	5	39
	MTW	24	26	31	30	15	126
	TF	14	14	14	9	28	79
Dongcheng	H	5	11	6	9	3	34
	MTW	19	23	24	20	13	99
	TF	24	16	20	19	32	111
Xicheng	H	10	13	6	10	5	44
	MTW	18	24	29	30	14	115
	TF	20	13	15	8	29	85

Table 8 The congestion degree at all six districts

	0 and 5	1 and 6	2 and 7	3 and 8	4 and 9
Haidian	2.62	2.62	2.69	2.62	2.79
Fengtai	2.14	2.05	2.21	2.04	2.45
Shijingshan	2.01	2.05	2.08	2.03	2.11
Chaoyang	3.46	3.46	3.55	3.39	3.80
Dongcheng	4.20	3.94	4.10	4.04	4.38
Xicheng	4.06	3.91	4.05	3.88	4.29
Average	3.08	3.00	3.11	3.00	3.30

Table 9 Paired *t*-test results among different restriction scenarios

	0 and 5	1 and 6	2 and 7	3 and 8	4 and 9
0 and 5	–	0.16	0.32	0.07	0.00
1 and 6	–	–	0.00	0.85	0.00
2 and 7	–	–	–	0.00	0.01
3 and 8	–	–	–	–	0.00
4 and 9	–	–	–	–	–

patterns, but can not completely explain the impact on traffic congestion patterns. Therefore, it could be inferred that there are other factors affecting the congestion patterns.

Automobile license plate restriction (ALPR) sets out rules that restrict automobile travel at particular date. For example, driving can be restricted based on vehicle license plate numbers for private cars. In details, vehicles with license numbers ending in 0 or 5 are prohibited from driving on Mondays; vehicles with license numbers ending in 1 or 6 are prohibited from driving on Tuesdays; vehicles with license numbers ending in 2 or 7 are prohibited from driving on Wednesdays; vehicles with license numbers ending in 3 or 8 are prohibited from driving on Thursdays; and vehicles with license numbers ending in 4 or 9 are prohibited from driving on Fridays. Generally speaking, the ALPR rules are updated quarterly and there are no driving restrictions on weekends. In China, ALPR is commonly implemented as a measure to reduce traffic congestion in megacities, e.g., Beijing, Tianjin, Guangzhou, Chengdu.

The influence of ALPR on the traffic congestion patterns is analyzed in this subsection. The statistical results about the congestion patterns across no-drive days at all six districts are shown in Table 7, which indicate that in TF pattern, the highest proportion is the days when the restriction numbers ending in 4 and 9. The number of days with restriction numbers ending in 4 and 9 increases from MTW pattern to TF pattern at Haidian, Fengtai, and Shijingshan, and increases from H pattern to MTW pattern, and then to TF pattern at Chaoyang, Dongcheng, and Xicheng.

For each district q with restriction scenario p, *Congestion Degree* is defined as the sum of the mean congestion index of each congestion pattern multiplied by the proportion of days. The higher the congestion degree, the more serious the congestion in the restriction scenario. If w_{pq} denotes the congestion degree at district q with restriction scenario p, then we have

$$w_{pq} = \sum_{k=1}^{K} e_{pq}^{k} \lambda_{pq}^{k}, \tag{16}$$

where e_{pq}^{k} represents the mean congestion index of the k^{th} congestion pattern, and λ_{pq}^{k} represents the proportion of days of the k^{th} congestion pattern with $\lambda_{pq}^{1} + \lambda_{pq}^{2} + \cdots + \lambda_{pq}^{K} = 1$. In Table 8, the numbers in the last row indicate the average congestion degree among six districts in each restriction

Table 10 The indicator values of different traffic congestion patterns at different districts

District	Congestion patterns	Minimum congestion index	Maximum congestion index	Mean congestion index	Variance	Duration of congestion (min)
Haidian	MTW	1.04	6.49	2.32	2.14	230
	TF	1.02	7.95	3.00	3.93	345
Fengtai	MTW	0.77	4.31	1.96	1.01	75
	TF	0.91	6.63	2.74	2.56	365
Shijingshan	MTW	0.85	4.64	1.81	0.85	45
	TF	0.97	6.39	2.40	2.13	225
Chaoyang	H	0.89	6.12	2.46	2.09	270
	MTW	0.94	7.65	3.42	4.70	540
	TF	1.03	8.24	4.24	6.07	735
Dongcheng	H	1.25	7.15	2.96	2.89	355
	MTW	1.05	8.17	3.88	6.22	630
	TF	1.11	8.58	4.71	7.39	780
Xicheng	H	1.09	7.84	3.11	4.02	395
	MTW	1.11	8.45	3.93	6.28	615
	TF	1.11	8.68	4.66	7.51	755

scenario. It is shown that the restriction scenario (4, 9) results in the highest congestion degree, while the restriction scenario (1, 6) and (3, 8) lead to the lowest congestion degree. This is due to the Chinese people's taboo for the number 4, which makes the number of vehicles ending in 4 very limited compared with other numbers. Conversely, 6 and 8 are the lucky numbers for Chinese people, which make the quantity of vehicles ending in 6 or 8 very large.

Paired t-test is taken to evaluate the difference of the congestion degree at all six districts under different restriction scenarios. The test statistic for paired t-test between scenario p and scenario p' is calculated as

$$t_{pp'} = \frac{\sum_{q=1}^{6} \left| w_{pq} - w_{p'q} \right|}{\sqrt{1.2 \times \left(\sum_{q=1}^{6} \left| w_{pq} - w_{p'q} \right|^2 \right) - 0.2 \times \left(\sum_{q=1}^{6} \left| w_{pq} - w_{p'q} \right| \right)^2}}, \tag{17}$$

where w_{pq} denotes the congestion degree at district q with restriction scenario p, $w_{p'q}$ denotes the congestion degree at district q with restriction scenario p'.

The results of paired t-test are given by Table 9. With significance level 0.05, if the two tailed P value is less than 0.05, it can be concluded that the values of congestion degree are significantly different. In Table 9, it is noted that the restriction scenario (4, 9) is significantly different from all other four scenarios; there is significant difference between scenario (1, 6) and scenario (2, 7); there is significant difference between scenario (3, 8) and scenario (2, 7); there are no significant differences among other scenarios. Therefore, the ALPR policy has an important influence on traffic congestion patterns.

6.3 Spatial dependence of traffic congestion patterns

As we have shown in Fig. 3, traffic congestion patterns are spatial dependent, that is, different districts have different number of traffic congestion patterns. If the traffic congestion index takes value more than 4.0, it means that the traffic situation is congested (See Fig. 2). In Table 10, the minimum congestion index, the maximum congestion index, the mean congestion index, the variance of congestion index, and the duration of congestion associated with all congestion patterns across all districts are calculated respectively. It is found that the maximum value, mean value, variance, and congestion duration increase gradually when the congestion pattern changes from H and MTW to TF, while the minimum value keeps almost unchanged. Most importantly, the maximum value, mean value, variance, and congestion duration have significantly difference across districts, which illustrate again that congestion patterns are spatial dependent.

Note that the shapes of H/MTW/TF patterns across districts are also significantly different as shown in Fig. 3. As the district gets closer to the downtown, the valley between the morning peak and evening peak becomes more sharp, the peak value gets greater, and the congestion lasts longer, which could also be observed in Table 10. Interestingly, although Haidian and Chaoyang have the similar distance to the downtown, their traffic congestion patterns are extremely different, which is reflected in both the number of congestion patterns and the specific indicator values. This phenomenon can be explained by the functional differences of these two districts in Beijing: Chaoyang is an important business center and foreign affairs

with relatively active traffic, while Haidian is an education center with relatively light traffic congestion.

7 Conclusions

Alleviating traffic congestion has always been an important challenge for the sustainable development of megacities. Accurate understanding of traffic congestion patterns and its characteristics is helpful to formulate scientific congestion prevention measures. In this paper, a traffic congestion pattern analysis framework was constructed based on the congestion index data. First, an improved weighted K-means clustering method was proposed to identify the traffic congestion patterns. Second, based on the identified traffic congestion patterns, the spatial–temporal variations of traffic congestion patterns were analyzed. Case studies with real-life data illustrated that the traffic congestion patterns are both spatial dependent and temporal dependent, and the automobile license plate restriction has important influence on the traffic congestion patterns.

On the basis of the results in this study, several issues are deserving of future study. First, the traffic congestion index data used in this study are collected according to the administrative division of Beijing, more precise division should be carried out for obtaining more valuable information, such as the congestion pattern analysis for roads or blocks. Second, for the analysis of influencing factors about congestion patterns, more issues should be considered, such as weather conditions, geographical conditions, emergency events, and so on. Third, the congestion pattern analysis in this paper is based on the congestion index data in Beijing, the congestion patterns and characteristics at other cities should be conducted and compared with Beijing.

Acknowledgements This work was supported by grants from the National Natural Science Foundation of China (Nos. 71722007 & 71931001), the Funds for First-class Discipline Construction (XK1802-5), the Key Program of NSFC-FRQSC Joint Project (NSFC No. 72061127002 and FRQSC No. 295837), the Fundamental Research Funds for the Central Universities (buctrc201926).

References

Angayarkanni SA, Sivakumar R, Ramana Rao YV (2021) Hybrid grey wolf: Bald eagle search optimized support vector regression for traffic flow forecasting. J Ambient Intell Humaniz Comput 12:1293–1304

Arachchige CNPG, Prendergast LA, Staudte RG (2020) Robust analogs to the coefficient of variation. J Appl Stat. https://doi.org/10.1080/02664763.2020.1808599

Beckers JM, Rixen M (2003) EOF calculations and data filling from incomplete oceanographic datasets. J Atmos Oceanic Tech 20(12):1839–1856

Chauhan S, Agarwal N, Kar AK (2016) Addressing big data challenges in smart cities: a systematic literature review. Info 18(4):1–10

Chen YC, Chen YL, Lu JY (2021a) MK-means: detecting evolutionary communities in dynamic networks. Expert Syst Appl 176:114807

Chen YY, Chen C, Wu Q, Ma JM, Zhang GH, Milton J (2021b) Spatial-temporal traffic congestion identification and correlation extraction using floating car data. J Intell Transp Syst 25(3):263–280

Degen WLF (2007) Sharp error bounds for piecewise linear interpolation of planar curves. Computing 79:143–151

Hong YM, Lee YJ (2017) A general approach to testing volatility models in time series. J Manag Sci Eng 2(1):1–33

Jiang MR, Chen W, Li X (2021) S-GCN-GRU-NN: a novel hybrid model by combining a spatiotemporal graph convolutional network and a gated recurrent units neural network for short-term traffic speed forecasting. J Data, Inform Manag 3:1–20

Kan ZH, Tang LL, Kwan MP, Ren C, Liu D, Li QQ (2019) Traffic congestion analysis at the turn level using taxis' GPS trajectory data. Comput Environ Urban Syst 74:229–243

Ke JT, Yang H, Zheng ZF (2020) On ride-pooling and traffic congestion. Transp Res Part B: Methodol 142:213–231

Kim J, Kwan MP (2019) Beyond commuting: ignoring individuals' activity-travel patterns may lead to inaccurate assessments of their exposure to traffic congestion. Int J Environ Res Public Health 16(1):89

Kim J, Ryu JH (2015) Quantifying a threshold of missing values for gap filling processes in daily precipitation series. Water Resour Manag 29:4173–4184

Klos A, Bogusz J, Figurski M, Kosek W (2015) On the handling of outliers in the GNSS time series by means of the noise and probability analysis. Int Assoc Geod Symp 143:657–664

Konietschke F, Pauly M (2014) Bootstrapping and permuting paired *t*-test type statistics. Stat Comput 24:283–296

Kruschke JK, Aguinis H, Joo H (2012) The time has come: Bayesian methods for data analysis in the organizational sciences. Org Res Methods 15:722–752

Lee SY, Xia YM (2006) Maximum likelihood methods in treating outliers and symmetrically heavy-tailed distributions for nonlinear structural equation models with missing data. Psychometrika 71:565–585

Li X, Wong HS, Wu S (2012) A fuzzy minimax clustering model and its applications. Inf Sci 186:114–125

Li L, Su XN, Wang YW, Lin YT, Li ZH, Li YB (2015) Robust causal dependence mining in big data network and its application to traffic flow predictions. Transp Res Part C: Emerg Technol 58:292–307

Li YC, Xiong WT, Wang XP (2019) Does polycentric and compact development alleviate urban traffic congestion? A case study of 98 Chinese cities. Cities 88:100–111

Lu ZD, Hui YV (2003) L1 linear interpolator for missing values in time series. Ann Inst Stat Math 55:197–216

Praveen DS, Raj DP (2021) Smart traffic management system in metropolitan cities. J Ambient Intell Humaniz Comput 12:7529–7541

Retallack AE, Ostendorf B (2019) Current understanding of the effects of congestion on traffic accidents. Int J Environ Res Public Health 16(18):3400

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Sanchez-Cambronero S, Jimenez P, Rivas A, Gallego I (2017) Plate scanning tools to obtain travel times in traffic networks. J Intell Transp Syst 21(5):390–408

Shelke M, Malhotra A, Mahalle PN (2019) Fuzzy priority based intelligent traffic congestion control and emergency vehicle management using congestion-aware routing algorithm. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01523-8

Shi L, Chen GM (2008) Detection of outliers in multilevel models. J Stat Plan Inference 138:3189–3199

ShirMohammadi MM, Esmaeilpour M (2020) The traffic congestion analysis using traffic congestion index and artificial neural network in main streets of electronic city (case study: Hamedan city). Program Comput Softw 46:433–442

Simolo C, Brunetti M, Maugeri M, Nanni T (2009) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. Int J Climatol 30(10):1564–1576

Su Y, Sun W (2019) Dynamic differential models for studying traffic flow and density. J Ambient Intell Humaniz Comput 10:315–320

Sun QX, Sun YX, Sun L, Li Q, Zhao JL, Zhang Y, He H (2019) Research on traffic congestion characteristics of city business circles based on TPI data: the case of Qingdao, China. Physica A 534:122214

Sun QX, Zhang Y, Sun L, Li Q, Gao P, He H (2021) Spatial–temporal differences in operational performance of urban trunk roads based on TPI data: the case of Qingdao. Physica A 568:125696

Tian Q, Yang H, Huang HJ (2010) Novel travel cost functions based on morning peak commuting equilibrium. Oper Res Lett 38(3):195–200

Torkjazi M, Mirjafari PS, Poorzahedy H (2018) Reliability-based network flow estimation with day-to-day variation: a model validation on real large-scale urban networks. J Intell Transp Syst 22(2):121–143

Wang MJ, Yang S, Sun Y, Gao J (2017) Discovering urban mobility patterns with PageRank based traffic modeling and prediction. Physica A 485:23–34

Wang WX, Guo RJ, Yu J (2018) Research on road traffic congestion index based on comprehensive parameters: taking Dalian city as an example. Adv Mech Eng 10(6):1–8

Wen HM, Sun JP, Zhang X (2014) Study on traffic congestion patterns of large city in China taking Beijing as an example. Procedia Soc Behav Sci 138:482–491

Wu X, Zhu X, Wu GQ, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107

Xu SJ, Chan HK, Ch'ng E, Tan KH (2020) A comparison of forecasting methods for medical device demand using trend-based clustering scheme. J Data Inf Manag 2:85–94

Yang Y, Zhou JD, Li X (2018) Energy-efficient stochastic chance-constrained programming model for train timetable optimization. J Syst Eng 33(2):197–211

Yaqoob I, Hashem IAT, Gani A, Mokhtar S, Ahmed E, Anuar NB, Vasilakos AV (2016) Big data: from beginning to future. Int J Inf Manag 36(6):1231–1247

Zhao PJ, Hu HY (2019) Geographical patterns of traffic congestion in growing megacities: big data analytics from Beijing. Cities 92:164–174