Front. Eng. Manag. 2025, 12(4): 952–970
https://doi.org/10.1007/s42524-024-3148-1

RESEARCH ARTICLE

Shengyou WANG, Chunjiao DONG, Chunfu SHAO, Sida LUO, Jie ZHANG, Meng MENG

Traffic state estimation incorporating heterogeneous vehicle composition: A high-dimensional fuzzy model

© The Author(s) 2024. This article is published with open access at link.springer.com and journal.hep.com.cn

Abstract Accurate traffic state estimations (TSEs) within road networks are crucial for enhancing intelligent transportation systems and developing effective traffic management strategies. Traditional TSE methods often assume homogeneous traffic, where all vehicles are considered identical, which does not accurately reflect the complexities of real traffic conditions that often exhibit heterogeneous characteristics. In this study, we address the limitations of conventional models by introducing a novel TSE model designed for precise estimations of heterogeneous traffic flows. We develop a comprehensive traffic feature index system tailored for heterogeneous traffic that includes four elements: basic traffic parameters, heterogeneous vehicle speeds, heterogeneous vehicle flows, and mixed flow rates. This system aids in capturing the unique traffic characteristics of different vehicle types. Our proposed high-dimensional fuzzy TSE model, termed HiF-TSE, integrates three main processes: feature selection, which eliminates redundant traffic features using Spearman correlation coefficients; dimension reduction, which utilizes the T-distributed stochastic neighbor embedding machine learning algorithm to reduce high-dimensional traffic feature data; and FCM clustering, which applies the fuzzy C-means algorithm to classify the simplified data into distinct clusters. The HiF-TSE model significantly reduces computational demands and enhances efficiency in TSE processing. We validate our model through a real-world case study, demonstrating its ability to adapt to variations in vehicle type compositions within heterogeneous traffic and accurately represent the actual traffic state.

Keywords traffic state estimation, heterogeneous traffic, T-distributed stochastic neighbor embedding algorithm, Fuzzy C-means machine learning algorithm

1 Introduction

The rapid expansion of the economy has led to increased levels of transportation, particularly on highways, resulting in a range of challenges, including traffic congestion, accidents, emissions, and noise pollution (Huang et al., 2021). These challenges are expected to worsen in major cities, necessitating the urgent improvement of highway network traffic management efficiency (Gao et al., 2023; Ling et al., 2022). A key requirement for enhancing this efficiency is the precise determination of the traffic state of the transportation network, commonly referred to as traffic state estimation (TSE), which is the key focus of this research.

The TSE problem is primarily addressed using machine learning and deep learning methodologies because it falls under the domain of unsupervised learning (Tian et al., 2022; Zhang et al., 2024). These approaches focus on clustering traffic states by identifying key traffic factors. For example, Cheng et al. (2020) employed the fuzzy c-means (FCM) clustering algorithm, a machine learning technique, for traffic state estimation. This method incorporates important factors such as traffic flow, traffic speed, and occupancy. However, it is crucial to note that

Received Dec. 24, 2023; revised Mar. 25, 2024; accepted Apr. 17, 2024

Shengyou WANG
School of Traffic Management, People's Public Security University of China, Beijing 100038, China

Chunjiao DONG, Sida LUO
Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Chunfu SHAO
School of Traffic and Transportation Engineering, Xinjiang University, Urumqi 830046, China

Jie ZHANG
Business School, University of Bristol, Bristol, BS8 1SD, UK

Meng MENG (✉)
School of Management, University of Bath, Bath, BA2 7AY, UK
E-mail: mm3042@bath.ac.uk

This research was supported by the National Key R&D Program of China (Grant No. 2023YFB4302702) and the Fundamental Research Funds for the Central Universities (Grant No. 2023JKF02ZK08).

these factors have mainly been applied to homogeneous traffic, specifically single vehicle types such as cars (Bhaskar et al., 2014; Zheng and Su, 2016). In reality, highway traffic is heterogeneous and consists of various vehicle types, including trucks (Portilla et al., 2020). Trucks, in particular, have lower handling stability than passenger cars, resulting in higher casualty rates in traffic accidents (Li et al., 2016). Due to their greater weight, trucks generally have slower speeds than passenger cars, especially in specific road sections, where speeds can drop to 30 km/h or even lower (Lyu et al., 2022). These slow-moving trucks create mobile bottlenecks on the road (Krampe and Junge, 2021), significantly reducing the running speeds of multiple road sections and degrading the overall road service level (Gashaw et al., 2018; Jamshidnejad et al., 2019). Furthermore, the delays experienced by passenger cars increase, which is counterproductive to the original purpose of highway construction (Hoogendoorn and Bovy, 2000; Liu et al., 2017; Hyun et al., 2021). Consequently, highway traffic comprises interactions between different vehicle types and their respective characteristics (Ruan et al., 2021), a scenario that should be reflected in a mathematical model for TSE.

In addressing these research gaps, our contributions are twofold. (i) We constructed a TSE traffic feature index system for heterogeneous traffic. This system is designed to reflect the diverse traffic characteristics of different vehicle types and includes indicators such as the traffic speed, traffic flow, and mixed flow rate of heterogeneous vehicles. Together, these components form a comprehensive high-dimensional traffic index system. (ii) We established a high-dimensional fuzzy TSE model (HiF-TSE) for heterogeneous traffic. This model incorporates feature selection, T-distributed stochastic neighbor embedding (TSNE) dimension reduction, and FCM clustering. The proposed HiF-TSE model offers the advantages of reducing computational complexity and providing improved accuracy in identifying traffic states. By integrating feature selection, TSNE dimension reduction, and FCM clustering, this model effectively captures the nuances of heterogeneous traffic dynamics and enables more precise and reliable traffic state estimations.

The remainder of this paper is structured as follows. Section 2 presents a comprehensive literature review, setting the context for our study and highlighting previous work in this area. Section 3 contains the main methodology proposed in this study, including the traffic feature index system and the new HiF-TSE model for heterogeneous traffic. Section 4 describes the data used in the case study, and Section 5 analyses the results from the model. A discussion is illustrated in Section 6, which contains the traffic state estimation performance. Section 7 summarizes the conclusion and suggests the future research direction.

2 Literature review

The TSE estimates the overall traffic operating state based on a scientific and rational traffic feature index system (Tian et al., 2022; Han et al., 2022). Greenshields (1935) proposed a fundamental diagram (FD) that depicts the relationship between traffic features and the traffic state, which has since been widely accepted. Current research on TSE focuses on three main directions: constructing traffic feature index systems, establishing estimation models, and reducing model calculation complexity.

2.1 Construction of a traffic feature index system

In the past, scholars mainly relied on single traffic indicators, such as traffic flow, average speed, and density, to estimate traffic states (Al Mallah et al., 2017). However, more recent studies have gone beyond these traditional indicators and proposed new traffic indices. For example, Cheng et al. (2020) proposed a new classification indicator called the ample degree, which considers traffic flow, speed, and occupancy. However, since there are functional relationships between traffic flow, speed, and occupancy, the inclusion of this new indicator in the traffic feature index system can lead to redundancy. Redundancy in indicators can amplify traffic features and introduce errors in TSE. Moreover, redundant indicators can reduce model efficiency and increase model complexity and running time.

As mentioned in the introduction, little attention has been given to the impact of heterogeneous vehicles, particularly trucks, on TSE. However, trucks are the main vehicles used for highway transportation (Romo et al., 2014). Their large size and high center of mass result in poorer handling stability compared to passenger cars, leading to higher casualty rates and property losses in traffic accidents. Therefore, to develop effective measures to minimize the impact of trucks on expressways, it is crucial to understand the influence of heterogeneous traffic flow and speed on different traffic states.

2.2 Establishment of estimation models

Since TSE is an unsupervised problem, it is necessary to quantify TSE in a manner that is easily understandable by establishing a traffic feature index system. To address this challenge, scholars have utilized clustering algorithms belonging to unsupervised models to estimate traffic states (Celikoglu and Silgu, 2016). Clustering algorithms aim to group data based on the similarities and differences of data points, with similar points being categorized together. Clustering algorithms can be categorized as clustering center-based, density-based, or hybrid model-based. Clustering center-based algorithms, such as

k-means and FCM and their variations, primarily rely on distance for center point selection (Zhang et al., 2023). Density-based algorithms, such as DBSCAN, assign adjacent data points with a specific density to a category (Yu et al., 2019). Hybrid model-based algorithms utilize different distribution functions to allocate data into different categories.

Clustering center-based algorithms are widely employed in the field of TSE (Nidheesh et al., 2017). Lin et al. (2013) proposed an enhanced k-means algorithm for network traffic classification estimation, which outperformed the standard k-means method in terms of overall accuracy and square error. Zheng and Su (2016) utilized a Markov random field and total variation regularization to enhance TSE accuracy with noisy traffic data.

2.3 Reduction of model calculation complexity

The models mentioned above are primarily suitable for low-dimensional traffic features. This limitation was identified by Bai and Li (2016), who concluded that conventional clustering algorithms struggle to identify parameters when dealing with high-dimensional traffic feature clustering. To address this challenge, dimension reduction technology, such as principal component analysis (PCA), has been widely used. PCA captures relationships in high-dimensional data and converts problem parameters into a set of uncorrelated PC scores. Duan et al. (2021) developed an adaptive clustering method based on the PCA method to filter LiDAR point clouds. One advantage of this method is its low computational complexity.

Although PCA is commonly used for dimensionality reduction, it is a linear method and is not suitable for nonlinear traffic feature data. The TSNE algorithm, on the other hand, is a nonlinear dimensionality reduction algorithm based on the T distribution. This approach maximizes the internal structure of high-dimensional data through dimensionality reduction (Erfani et al., 2023). If the data points are close in the high-dimensional space, TSNE will also keep them close in the dimensionality-reduced space (Zhu et al., 2019). The TSNE can reduce the dimensionality of high-dimensional data to two or three dimensions and present it in a visual form (Zong et al., 2020). It is a popular algorithm internationally that was originally developed for image processing and has been widely used in hyperspectral image classification and bioinformatics with successful results. However, TSNE algorithms have not yet been applied in the field of TES.

2.4 Comments on previous work

To accurately represent the traffic state in real-world traffic situations involving heterogeneous traffic, there are three key aspects of analysis: constructing a traffic feature index system specifically for heterogeneous traffic state estimation, establishing a suitable TSE model for heterogeneous traffic, and reducing the computational complexity associated with the TSE model for heterogeneous traffic. All three aspects are highly important and currently represent significant research gaps.

In response, this paper examines the consideration of more realistic heterogeneous traffic flows and proposes a novel TSE model tailored for more accurate estimations. First, a comprehensive TSE traffic feature index system is established to specifically address the characteristics of various vehicle types. This system comprises four key elements: basic traffic parameters, the speed of heterogeneous vehicles, their traffic flow, and the mixed flow rate. Second, a high-dimensional fuzzy TSE model called HiF-TSE is developed, which builds upon this index system. The HiF-TSE model incorporates three critical processes: feature selection, which utilizes Spearman correlation coefficients to eliminate irrelevant traffic features; dimension reduction, which applies the T-distributed stochastic neighbor embedding algorithm to transform high-dimensional traffic data into a more manageable form; and FCM clustering, which employs the fuzzy C-means algorithm to efficiently segment the refined data into distinct groups. This advanced HiF-TSE model is notable for its ability to significantly reduce computational demands while enhancing TSE processing efficiency. To validate the effectiveness of the model, an empirical case study is conducted, demonstrating its practical application and efficacy in real-world traffic scenarios.

3 Methodology

3.1 Construction of the traffic feature index system

As previously stated, highway traffic is heterogeneous due to the presence of various vehicle types. To capture the unique traffic characteristics associated with each type, a nuanced heterogeneous traffic feature index system was developed that incorporates all vehicle types permitted on highways. This index system leverages the distinct traffic patterns of each vehicle type, as illustrated in Fig. 1. In accordance with the Chinese Automobile Classification Standard, trucks are classified as small, medium, large, or extralarge based on their size, while passenger cars are divided into small/medium passenger cars and large buses. To evaluate heterogeneous vehicles on a standardized basis, the concept of the passenger car equivalent (PCE) is employed. The traffic features within each category can be described as follows.

(1) The first category includes the basic traffic parameters commonly used in traditional TSE models (Cheng et al., 2020). These parameters include the percent time spent following, average headway, and time occupancy rate.

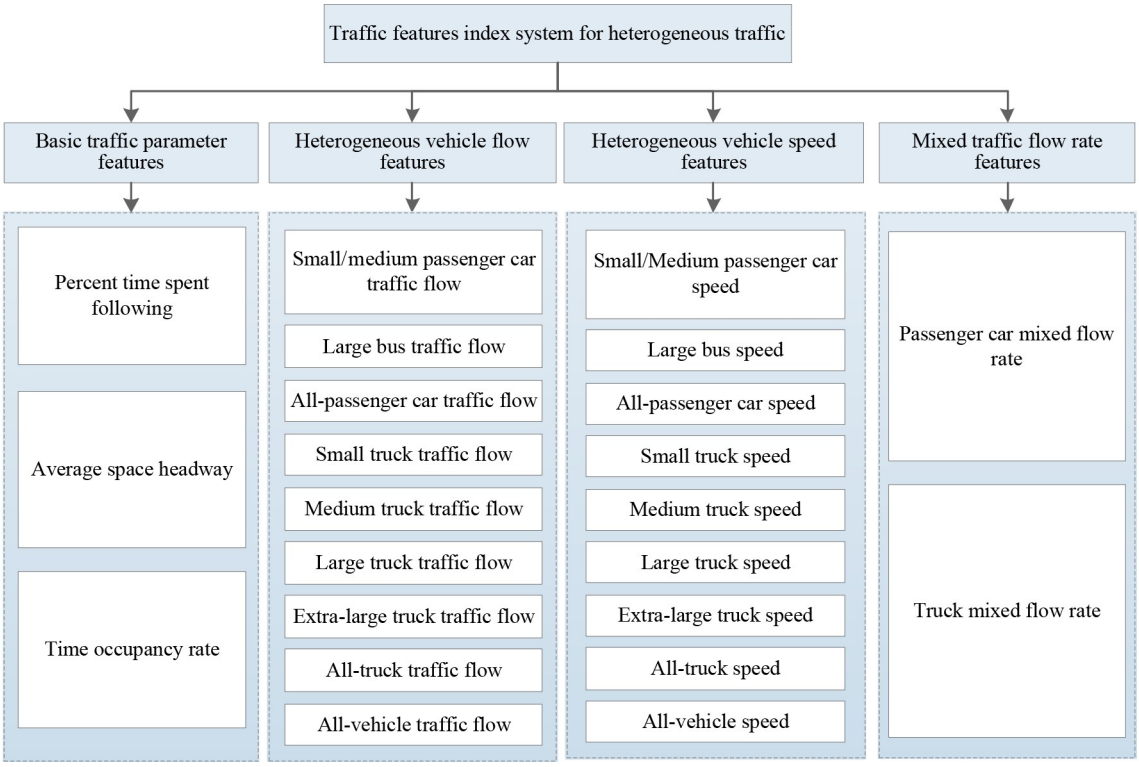


Fig. 1 Construction of the traffic feature index system for heterogeneous traffic.

(2) The second category consists of heterogeneous vehicle speeds, which are used to describe the speed characteristics of different types of vehicles. These include small, medium, large, and extralarge trucks; small/medium passenger cars; large buses; all trucks; all passenger cars; and all vehicles. Specifically, the all-truck speed refers to the average speed of small, medium, large, and extralarge trucks. Similarly, the all-passenger car speed refers to the average speed of small/medium cars and large buses. The all-vehicle speed refers to the average speed of all vehicle types.

(3) The third category addresses heterogeneous vehicle flow, which represents the flow characteristics of different types of vehicles. This category includes flows of small, medium, large, and extralarge trucks; small/medium passenger cars; large buses; all trucks; all passenger cars; and all vehicles. Among them, all-truck flow refers to the total flow of small, medium, large, and extralarge trucks. All-passenger car flow refers to the total flow of small/medium cars and large buses. All-vehicle flow refers to the total flow of all vehicle types.

(4) The fourth category pertains to the mixed flow rate, indicating the interaction between passenger cars and trucks. Specifically, it includes the truck mixed flow rate and the passenger car mixed flow rate. The truck mixed flow rate is the ratio of all-truck flow to all-vehicle flows. The passenger-car mixed flow rate refers to the ratio of all-passenger-car flow to all-vehicle flow.

3.2 HiF-TSE model

In this section, we propose a novel HiF-TSE model within the traffic feature index system for heterogeneous traffic. As illustrated in Fig. 1, the established traffic feature index system, applicable in traffic state estimation, includes high-dimensional data, posing significant challenges. Consequently, the proposed HiF-TSE model comprises three steps aimed at enhancing efficiency and reducing model complexity concurrently. The methodology can be outlined as follows:

(1) Initially, feature selection is conducted on high-dimensional traffic features to eliminate highly correlated collinear features, addressing the issue of feature redundancy.

(2) Subsequent to the initial processing, the second step involves dimensionality reduction of the high-dimensional traffic feature data, aiming to mitigate the curse of dimensionality.

(3) The final step entails determining the traffic state utilizing both the clustering model and the dimension-reduced data from the preceding step. To ensure stable cluster results, the bootstrap method (Banerjee and Monni, 2021) is employed to identify an optimal cluster center. Further elaboration of these three steps and the structure of the proposed HiF-TSE model are presented in Fig. 2.

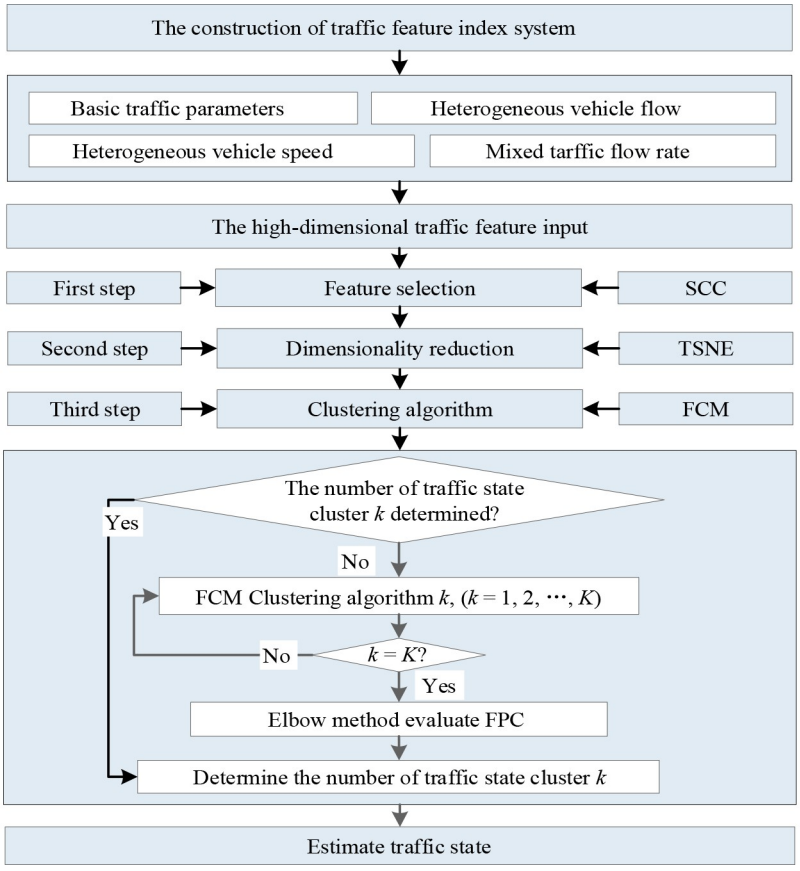


Fig. 2 The structure of the HiF-TSE model under heterogeneous traffic.

3.2.1 Feature selection

Feature selection is a pivotal data preprocessing step for mitigating model complexity by eliminating irrelevant features. The underlying principle of feature selection involves filtering out features associated with correlation coefficients exceeding a predefined threshold (Guan et al., 2021). Specifically, in the context of traffic features, these correlation coefficients are computed utilizing the Spearman correlation coefficient (SCC). Notably, SCC offers the advantage of not necessitating a specific data distribution; it employs the rank order of two features for linear correlation analysis, thus representing a nonparametric statistical approach. Given that the distribution of traffic features does not adhere to a normal distribution, the application of SCC is deemed suitable.

The high-dimensional traffic feature is defined as $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n)$, where $\boldsymbol{x}_i=(x_{i1},x_{i2},\ldots,x_{im})^{\mathrm{T}}$, where n represents the number of dimensions of the traffic feature. m represents the number of samples. $(\boldsymbol{F}_i,\boldsymbol{G}_i)$ is denoted as pairs of traffic features, $\boldsymbol{F}_i$, $\boldsymbol{G}_i\in\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$, $i=1, 2..., n$. The position of $\boldsymbol{F}_i$ in the ascending sequence is recorded as $\boldsymbol{E}_i$ $\boldsymbol{E}_i=(\boldsymbol{E}_{i1},\boldsymbol{E}_{i2},\ldots,\boldsymbol{E}_{im})^{\mathrm{T}}$, which is the rank. Similarly, $\boldsymbol{H}_i$ is defined as the rank of $\boldsymbol{G}_i$, $\boldsymbol{G}_i=(\boldsymbol{G}_{i1},\boldsymbol{G}_{i2},\ldots,\boldsymbol{G}_{im})^{\mathrm{T}}$. The SCC of $(\boldsymbol{F}_i,\boldsymbol{G}_i)$ is defined in Eq. (1).

$$\rho_{(F_i,G_i)}=1-\frac{6\sum_{j=1}^{m}(E_{ij}-H_{ij})^2}{m(m^2-1)}, \tag{1}$$

where $|\rho_{(F_i,G_i)}|\in[0,1]$. When a pair of traffic features are completely correlated, the $|\rho_{(F_i,G_i)}|$ value is 1. According to Piantadosi (2007), the correlation threshold is set to 0.90 in this paper. When $|\rho_{(F_i,G_i)}|\geqslant 0.90$, the redundant traffic features are filtered.

3.2.2 Dimension reduction

Following the filtration of redundant features, the retained traffic features still exhibit high-dimensional characteristics, necessitating reduction to alleviate the complexity and computational burden of the TSE model. This reduction is achieved through the utilization of the TSNE algorithm (Yang and Wang, 2020). TSNE, a nonlinear dimension reduction algorithm, is adept at transforming high-dimensional data into lower dimensions (Pezzotti et al., 2017). Notably, the TSNE excels in mapping high-dimensional features to low-dimensional space while preserving the probability distribution between features (Zhu et al., 2019; Zong et al., 2020).

After the feature selection step, the processed high-dimensional traffic feature is defined as $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_s)$,

$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^{\mathrm{T}}$, where s represents the number of dimensions of the processed high-dimensional traffic feature after feature selection, and the low-dimensional point after dimension reduction is defined as $\boldsymbol{Z} = (z_1, z_2, \ldots, z_r)$, $z_i = (z_{i1}, z_{i2}, \ldots, z_{im})^{\mathrm{T}}$, where r represents the number of dimensions after dimension reduction and $r < m < n$. The objective function of TSNE is to minimize the Kullback–Leibler (KL) divergence of the probability distribution distance between high-dimensional traffic features and low-dimensional points. The formula for minimizing KL is represented in Eq. (2).

$$\min KL(P\|Q) = \sum_i \sum_j p_{ij} \lg \frac{p_{ij}}{q_{ij}}, \tag{2}$$

where p_{ij} is the joint probability density of the high-dimensional traffic feature. q_{ij} is the joint probability density of the low-dimensional mapping point. The joint probability densities p_{ij} and q_{ij} are calculated using Eqs. (3)–(4):

$$p_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2 / 2\sigma^2\right)}, \tag{3}$$

$$q_{ij} = \frac{\left(1 + \|z_i - z_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|z_i - z_j\|^2\right)^{-1}}. \tag{4}$$

3.2.3 FCM clustering algorithm

After employing the TSNE algorithm to reduce high-dimensional traffic features, the resulting low-dimensional data are unsupervised and undergo further processing via a clustering approach. Notably, TSE represents an unsupervised problem that poses challenges in delineating boundaries between different traffic states. However, FCM has emerged as a clustering algorithm that is well suited for addressing this challenge, thus justifying its application in this study. A key advantage of the FCM algorithm lies in its ability to effectively handle ambiguous descriptions of traffic features within traffic states, thereby offering a more objective representation of the traffic state (Zhang et al., 2023). The FCM algorithm partitions traffic features into various clusters with a certain degree of membership probability. The objective function of the FCM algorithm is expressed in Eq. (5):

$$\min D = \sum_{k=1}^{K} \sum_{Z_i \in C_k} \omega_{ik}^{\alpha} \|Z_i - c_k\|^2, \tag{5}$$

where $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i}, \ldots, Z_{ri})$, $i = 1,2,...,m$, is the low-dimensional point after applying the TSNE to the high-dimensional traffic features. c_k is the cluster center of traffic state cluster k, $k = 1, 2,...,K$. $\boldsymbol{C}_k$ is the sample set of traffic state cluster k, $\boldsymbol{C}_k = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_l)$, $0 < l \leqslant m$. $\|\cdot\|$ represents the L^2 norm. ω_{ik}, $\omega_{ik} \in [0,1]$, represents the weight of the dimension reduction point $\boldsymbol{Z}_i$ belonging to the traffic state cluster k. ω_{ik} can also be expressed as the degree of membership. Each low-dimensional point will have a corresponding degree of membership, which is used to indicate the likelihood that the point belongs to the traffic state cluster. α, $\alpha \geqslant 1$ is a parameter that is used to control the fuzziness of the optimal solution of the FCM model. After FCM clustering, the clustered traffic state can be estimated by referring to the regular Greenshields FD.

4 Data

4.1 Data description

For this study, a small road network situated near the Beijing Capital International Airport was selected as the research area. The road network consists of six expressway sections. The research data were collected using loop coil detectors over a period of 14 days, from June 1st to June 14th, 2019. The collected data include basic information about traffic flow, speed, and occupancy rate. The traffic speed and flow were recorded for different types of vehicles. In addition, the collected data contain other essential information, such as the data and detector device ID. Detailed information about the collected data are presented in Table 1.

The description and information of the data values provided in Table 3 are derived from the traffic samples obtained through the loop coil detectors. Each sample comprises multiple feature values, with each feature having a different value type. Some feature values in the collected data, such as Section ID, Loop ID, and Time ID, are not utilized by the authors. Conversely, the feature set related to speed, flow, and occupancy is retained. Features such as large bus flow, large truck flow, and medium truck flow are converted into standard passenger car flow. According to the proposed traffic feature index system, certain traffic features cannot be directly obtained from the collected data, such as truck flow, passenger car flow, truck speed, passenger car speed, all vehicle flow, all vehicle speed, mixed flow rate of trucks, and mixed flow rate of passenger cars. These features need to be calculated using the regulations outlined in Section 3.1.

The number of lanes in different sections of the highway varies. For instance, Section HS1 consists of four one-way lanes, while HS3 has only three one-way lanes. Consequently, the traffic capacities differ. To avoid potential errors stemming from these variations, the authors converted the traffic features into an average value for a single lane.

4.2 Data analysis

To examine the necessity of the developed traffic feature

Table 1 Description and information of the data collected by the loop coil detector

Name	Description	Data type	Data information			
Section ID	Section number	Number	HS3	HS3	HS3	…
Loop ID	Device number	Char	212403121106	212403121106	212403121106	…
Date	Date	Date	2019/6/1	2019/6/1	2019/6/1	…
Time	Record time	Time	0:05	0:05	0:05	…
Time ID	Recording time period	Char	1	1	1	…
Number of lanes	one-way lanes	Number	3	3	3	…
Lane ID	Lane number	Char	11	12	13	…
PTSF	Percent time spent following	Number	98	93	92	…
ASHY	Average space headway	Number	321	213	212	…
TORT	Time occupancy rate	Number	34	43	35	…
STTF	Small truck traffic flow	Number	45	43	32	…
STSD	Small truck traffic flow	Number	45.6	41.3	32.3	…
…	…	…	…	…	…	…
Loop Fault	Device working status	Char	00	00	00	…

index system, the analysis focuses on the traffic flow and speed characteristics of trucks and passenger cars. The daily time distributions of the speed and traffic flow for trucks and passenger cars are illustrated in Fig. 3.

According to Fig. 3, the daily speed distribution of trucks is generally lower than that of passenger cars. The authors calculated the average speed for trucks, passenger cars, and all vehicles for a day, as presented in Table 2. The average daily speeds for trucks, passenger cars, and all vehicles are 43.20, 50.15, and 49.27 km/h, respectively. The average speeds of trucks are 6.95 and 6.07 km/h lower than those of passenger cars and all vehicles, respectively, indicating that passenger cars are 13.86% faster than trucks. The lower speed of the trucks may be attributed to their greater weight and body structure. Trucks have nonload-bearing bodies, which possess inferior integrity and more complex force conditions than passenger cars. Particularly on uphill sections, due to their heavy weight, trucks experience more strenuous climbing conditions, resulting in significantly lower speeds than those of passenger cars (Kong et al., 2018).

The daily time distributions of the traffic flows of trucks and passenger cars are shown in Fig. 4.

From Fig. 4, it is evident that the distribution of daily traffic flow for trucks differs from that for passenger cars. The daily traffic flow distribution of passenger cars displays distinct double peak characteristics, specifically during the morning peak from 8:00–10:00 and the evening peak from 17:00–19:00. On the other hand, the peak periods for truck traffic flow are from 20:00–23:00, around midnight, and from 2:00–5:00 in the early morning, while the corresponding traffic flow for passenger cars during those times is relatively low. Table 2 presents the basic mathematical characteristics of traffic flows at 5-min intervals for trucks, passenger cars, and all vehicles

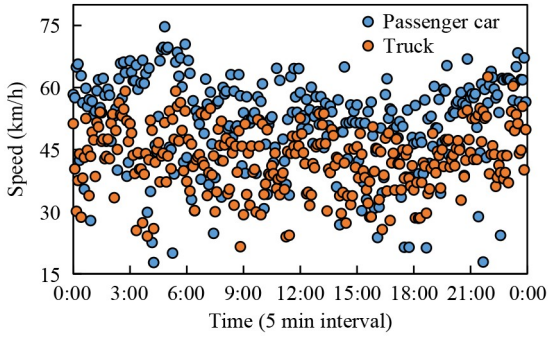


Fig. 3 The daily time distributions of the speeds of trucks and passenger cars.

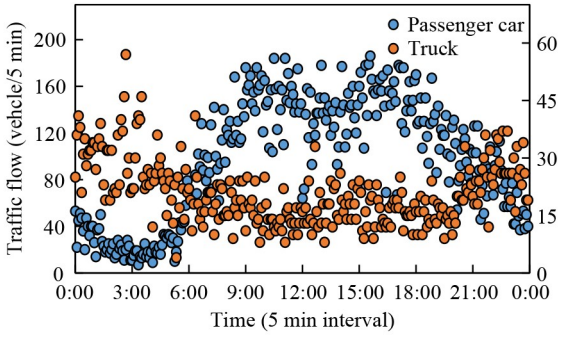


Fig. 4 The daily time distribution of the traffic flow of trucks and passenger cars.

throughout the day. The resultant average traffic flows for trucks, passenger cars, and all vehicles are 191.42, 234.13, and 425.55 pcu/h, respectively. The primary reason for the lower overall traffic flow of trucks compared to that of passenger cars is often attributable to local transportation policies. Local traffic management departments usually enforce restrictions on trucks driving

Table 2 Statistics of the traffic flow and speed of trucks and passenger cars

Type	Vehicle type	Minimum	Maximum	Mean	Variance	Skewness	Kurtosis
Traffic flow (pcu/h)	Passenger car	3	839	234.13	182.40	0.33	–0.87
	Truck	10	386	191.42	87.33	–0.24	–0.94
	All vehicles	15	970	425.55	242.05	–0.21	–1.33
Speed (km/h)	Passenger car	19.09	74.70	50.15	7.725	–0.23	–0.22
	Truck	22.00	63.00	43.20	8.30	0.96	1.12
	All vehicles	22.36	67.08	49.27	8.84	–0.83	0.49

Table 3 The SCC of traffic features

Traffic feature	Traffic feature	SCC
Small/medium passenger car traffic flow	All-passenger car traffic flow	0.99
Small/medium passenger car speed	All-passenger car speed	0.99
Passenger car mixed flow rate	Truck mixed flow rate	−1.00
All-passenger car speed	All-vehicle speed	0.90
All-vehicle traffic flow	Average space headway	0.90

on inner city roads from 7:00 am to 10:00 pm daily; consequently, trucks can only enter the city via highways at midnight and in the early morning.

The above analysis highlights the heterogeneity between truck and passenger car traffic features and emphasizes the necessity of a constructed heterogeneous traffic index system for improving the accuracy of TSE solutions.

5 Results

5.1 Feature selection

As discussed in Section 3.2.1, the feature selection step is employed to address the issue of dimensionality by effectively reducing the complexity of learning tasks through filtering out irrelevant features. The calculation of correlations among traffic features utilizes the SCC. These correlations are then visually represented in the form of a heatmap, as indicated in Fig. 5, providing a clear representation of the relationships between various traffic features.

The heatmap allows for easy visual interpretation of the correlation between traffic features. Significant positive and negative correlations indicate the presence of collinear features among the traffic features. To enhance model efficiency, it is advisable to avoid redundant traffic features as much as possible.

The relationship between small/medium passenger car traffic flow and all-passenger car traffic flow, as depicted in Fig. 6(a), is linear and significantly correlated. This is likely because small/medium passenger cars make up the majority of passenger cars on highways. As a result, the traffic characteristics of all passenger car traffic flows are retained, while the traffic characteristics of small/medium passenger car traffic flows are disregarded as collinear features. Similarly, the traffic features of small/medium passenger car speed, passenger car mixed flow rate, and all-passenger car speed are retained, while all-passenger car speed, truck mixed flow rate, and all-vehicle speed are ignored as collinear features. The traffic characteristics of all-vehicle traffic flow and average space headway exhibit an exponential nonlinear relationship. The traffic characteristics of the average space headway are considered redundant. The relationships between these traffic features are illustrated in Figs. 6(b)–6(e).

5.2 Dimension reduction

After the feature selection step, which filtered out features with correlation coefficients greater than 0.9, five redundant traffic features were eliminated, reducing the original 23 dimensions to 18 dimensions. However, the remaining 18-dimensional traffic features still fall into the category of high-dimensional features, necessitating the use of dimensionality reduction methods to further reduce their dimensions.

There are two main methods for dimensionality reduction: PCA and t-distributed stochastic neighbor embedding (t-SNE). After standardization operations, the dimensionality reduction results of PCA and t-SNE were visually compared and are shown in Fig. 7.

From Fig. 7, it is evident that the dimensionality reduction results obtained from t-SNE are significantly different from those obtained from PCA. Using the t-SNE method, the dimensionality reduction results for each road section are uniformly distributed and exhibit distinct clusters. This approach is beneficial for estimating traffic states. Conversely, the dimensionality reduction results obtained

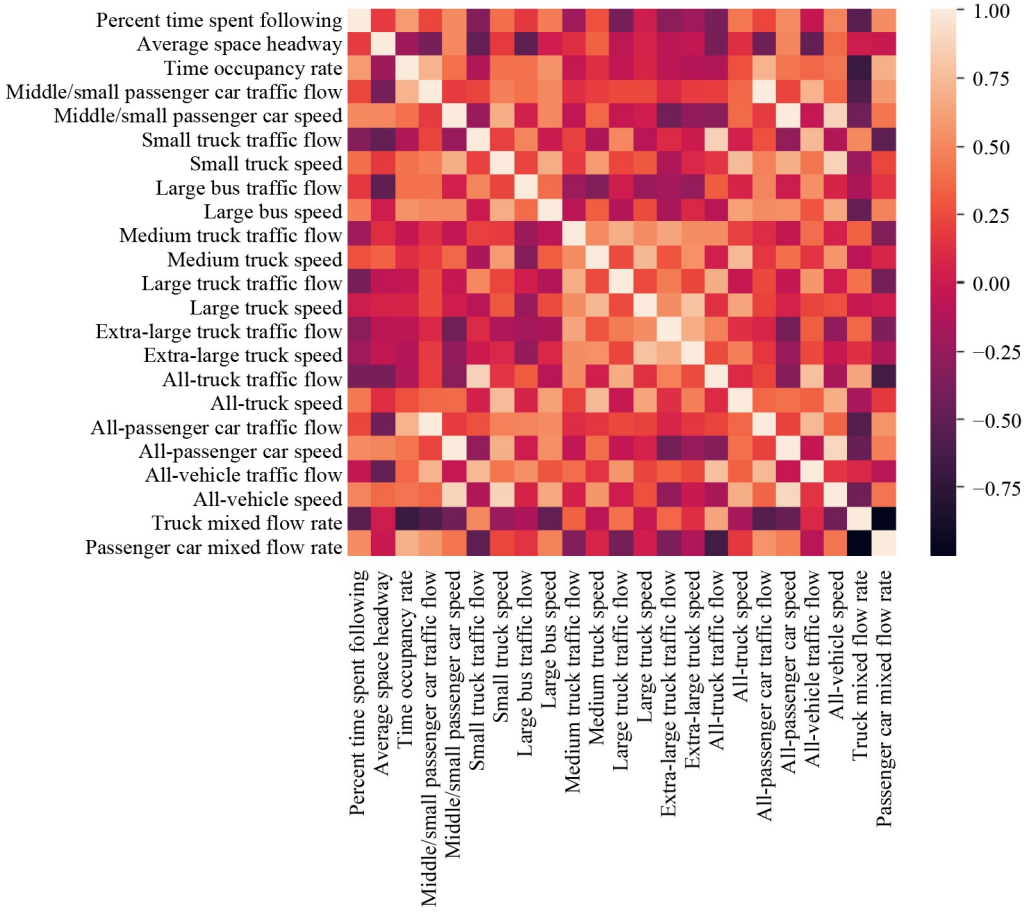


Fig. 5 The heatmap of the SCC for high-dimensional traffic features.

from the PCA method demonstrate a phenomenon of linear aggregation, with a discrete data distribution. The data distribution for each road section varies, and there are no prominent clustering characteristics. Therefore, the t-SNE dimensionality reduction method chosen in this paper is better suited for estimating traffic states.

In addition, the authors compared the absolute errors of TSNE dimension reduction with and without the feature selection process. The results, as shown in Table 4, provide insights into the error parameters and help determine whether the feature selection step contributes to improving the model's operating efficiency. KL divergence serves as the primary indicator of the dimension reduction goal of the TSNE method. A smaller KL divergence value indicates a smaller probability distribution distance between the high-dimensional traffic feature and the low-dimensional space, thus suggesting better dimension reduction performance. The mean σ^2, on the other hand, represents the mean of the Gaussian variance of all traffic feature data. A lower mean σ^2 signifies better performance in dimension reduction.

Table 4 displays the absolute error of TSNE dimension reduction with and without the feature selection process. It is evident that the results vary between the two methods. For instance, the feature selection process yields a mean value σ^2 of 0.120, which is 0.601 lower than the result obtained without the feature selection process. Similarly, the KL divergence is 57.883, reflecting a reduction of 7.245 compared to the result without feature selection. These findings indicate a reduction in the error of TSNE dimension reduction through the implementation of feature selection. Moreover, the dimension reduction coordinates also change after the feature selection process. These results underscore the importance of performing the feature selection process, as it contributes to improving the overall accuracy and efficiency of the proposed model.

5.3 FCM clustering

To commence FCM clustering, determining the optimal number of clusters, namely, the appropriate number of traffic state levels, is crucial. Previous studies often made assumptions about the number of traffic state classifications in advance, e.g., assuming three, four, or five levels (Xu et al., 2013). However, this approach is unreasonable due to the possibility of introducing human error into traffic state discrimination. As a result, in this study, the

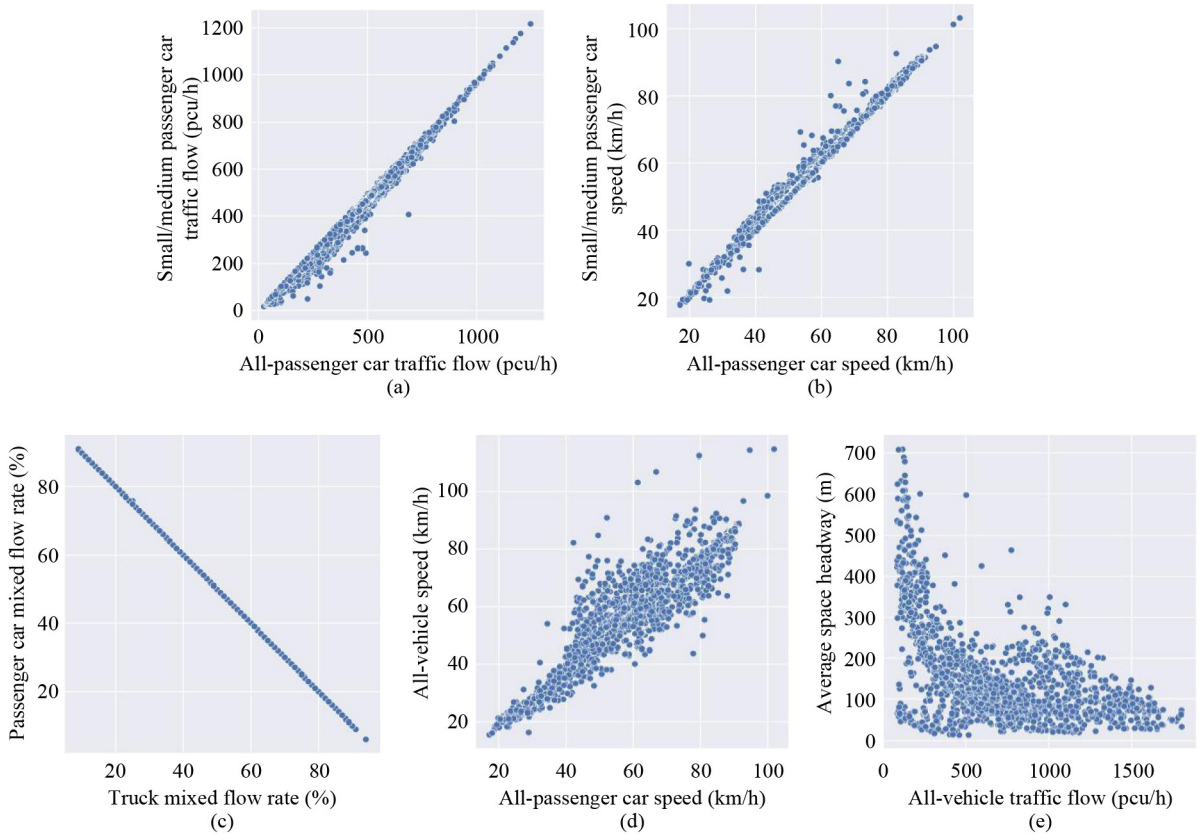


Fig. 6 The relationships between traffic features.

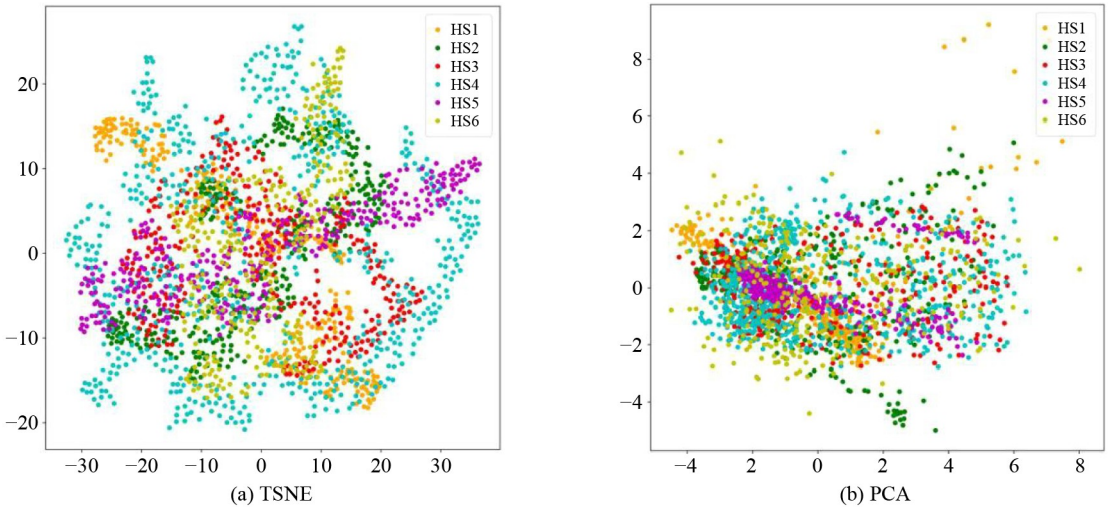


Fig. 7 Visualization of the dimension reduction results obtained by the TSNE (a) and PCA (b) methods.

clustering categories were tested individually, ranging from two to nine, to identify the optimal number of clusters. The FCM clustering results are presented in Fig. 8.

In Fig. 8, each 'series' represents a distinct cluster of traffic states. Additionally, the FPC is a metric used to evaluate the quality of clustering. It determines how well the traffic feature data are described by the proposed model (Puente et al., 2020). The FPC ranges from 0 to 1, where a value of 1 indicates the best clustering performance. The FPC can be used to select the number of clusters. A smaller FPC indicates greater convergence within each cluster. However, it is important to note that a smaller FPC does not always indicate better clustering performance. In extreme cases, treating each sample point as a cluster would result in a FPC of 0, but such a clustering result would have no practical meaning.

Table 4 The absolute error of TSNE dimension reduction with and without the feature selection process

Index	Feature selection process		Without a feature selection process	
Mean σ^2	0.120		0.721	
KL divergence*	57.883		65.128	
Absolute error	0.207		0.518	
Dimension reduction coordinate 6 examples	24.087	−10.892	20.611	−31.751
	49.020	−26.496	46.044	−18.944
	58.355	−14.556	56.537	−15.052
	65.884	30.708	34.352	42.341
	41.136	−6.210	39.974	−10.524
	66.298	30.222	35.007	41.990

Note: * The result of KL divergence is obtained after 250 iterations with early exaggeration.

Therefore, finding a balance between the number of clusters and the FPC value is necessary. The elbow method (Bahadur and Dhanalakshmi, 2020) can be used to determine this balance. According to the elbow method, as the number of clusters approaches the number of real clusters, the FPC rapidly decreases. However, when the number of clusters exceeds the number of real clusters, the FPC continues to decline but at a slower rate. Therefore, plotting the FPC value of the cluster center as a line graph helps identify the optimal number of clusters based on the inflection point. Figure 9 shows the trend of FPC values under different cluster numbers.

In Fig. 9, it can be observed that the FPC line graph has an inflection point highlighted by a blue circle when the number of clusters is five. According to the elbow method, the number of traffic state clusters in this study is determined to be five.

After determining the number of clusters, the authors aimed to ensure that the fuzzy clustering results were not affected by the sample data and that a stable fuzzy range was obtained. To achieve this goal, the bootstrap method was used to expand the sample size through data resampling (Banerjee and Monni, 2021). In statistics, the bootstrap method refers to resampling from the sample itself to infer the sample distribution. The fundamental idea of the bootstrap method is that the extracted sample contains all the necessary information; thus, the resampled samples can be considered to contain the same information. Moreover, the bootstrap method does not require external input and maintains its own stability in practical applications.

The resampled data are derived from the two-dimensional data after TSNE dimension reduction. When the sample size is insufficient, the bootstrap method is a reliable approach for enhancing the credibility of the statistical inference results. This approach involves creating a new sample by randomly selecting samples with replacements. The extent of variability mainly depends on the position of the cluster center. Once the cluster center position is determined, the variability also becomes established. Consequently, the stability range of the fuzzy clustering results largely relies on the cluster center.

In this study, the bootstrap method is employed to extract the same number of samples as the number of replacements and to recluster the new samples to obtain the cluster centers of the new samples. The bootstrap procedure is repeated 1000 times. The mean value of the cluster centers from 1000 bootstrap iterations is considered the final cluster center. An example of the cluster center position obtained by randomly sampling through 9 bootstrap operations is presented in Table 5.

Figure 10 shows the cluster center positions obtained after 9 bootstrap operations. From both Fig. 10 and Table 5, it can be observed that the cluster centers of multiple bootstrap samples are closely clustered, implying that the cluster centers are relatively stable. Compared to other methods, the bootstrap method produces more stable cluster centers. Furthermore, the fuzzy range of the traffic state category is more representative.

5.4 Determination of the traffic state

Since FCM clustering is an unsupervised model, determining the traffic state levels to which the clusters belong is still a question that needs to be answered. For instance, is cluster 3, as shown in Fig. 10, corresponding to a congested or unblocked traffic state? Which cluster pertains to the severely congested traffic state? To address these queries, the authors refer to the widely accepted relationship between traffic features and traffic state, as defined in the Greenshields FD, which is described in Section 2. The traffic features of all-vehicle speed and all-vehicle traffic flow are used to determine the traffic state of each cluster.

A box plot was generated to analyze the traffic characteristics within the five clusters. This statistical diagram is commonly employed to visualize clustering results and display data associated with decentralized materials (Singh and Saxena, 2021). Clusters 0 to 4 clearly differentiate the traffic features of all-vehicle speed and all-vehicle traffic flow. According to the Greenshields FD, a decrease in traffic speed indicates a transition from an uncrowded to a gradually crowded traffic state. As a result, the five clusters are sorted based on the values of the all-vehicle traffic speed box. The sorted clusters are then classified as smooth, basically smooth, basically smooth, moderate congestion, and mild congestion. The traffic states are rearranged in ascending order from unobstructed to congested, and the arranged box plots are presented in Fig. 11.

As depicted in Fig. 11(a), the maximum value of the traffic speed box corresponds to a smooth traffic state, while the traffic flow box (shown in Fig. 11(b)) is relatively low during this period. As the value of the traffic speed

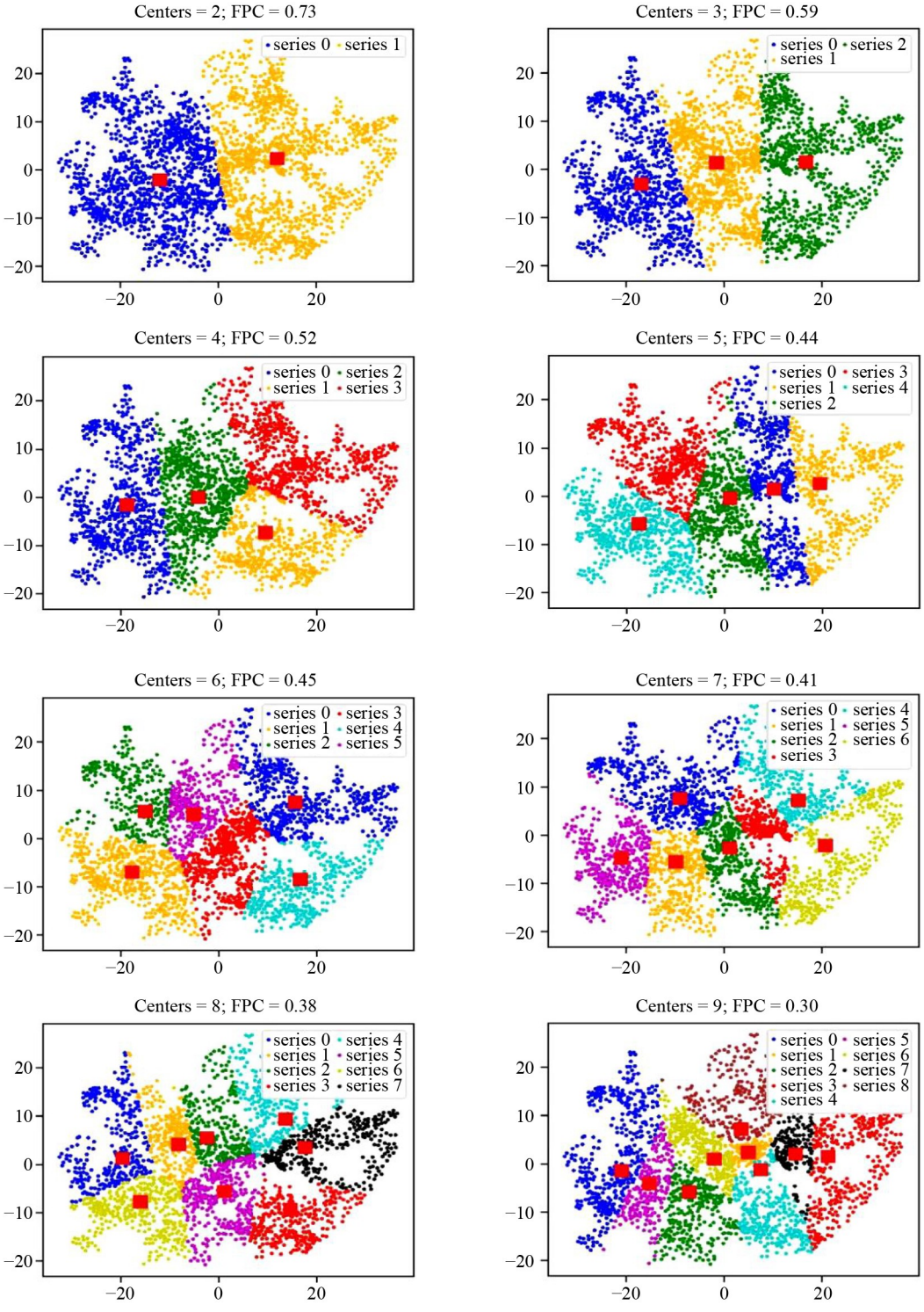


Fig. 8 The result of fuzzy clustering with the number of clusters from two to nine.

box gradually decreases, the traffic state becomes increasingly congested, and the flow undergoes a process of initial growth followed by a gradual decline. This observation aligns with the regular pattern in the Greenshields FD, highlighting the practical significance of the proposed HiF-TSE model for analyzing high-dimensional traffic features under heterogeneous traffic conditions.

The clustering results of the HiF-TSE model provide the fuzzy interval range for each traffic feature in each traffic state, as presented in Table 6.

From Table 6, it is evident that as the traffic state shifts from smooth to severe congestion, the upper and lower limits of flow traffic features initially increase and then decrease. For example, in the case of large bus traffic

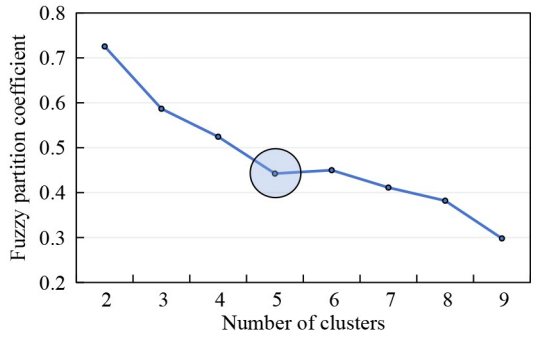


Fig. 9 The trend of FPC values under different cluster numbers.

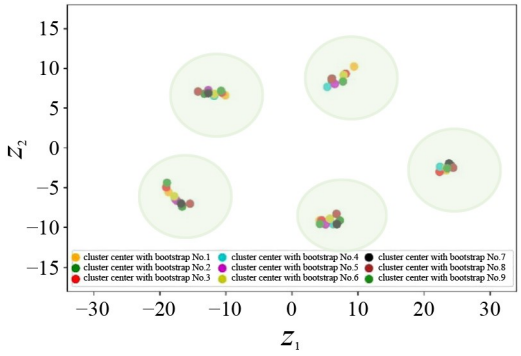
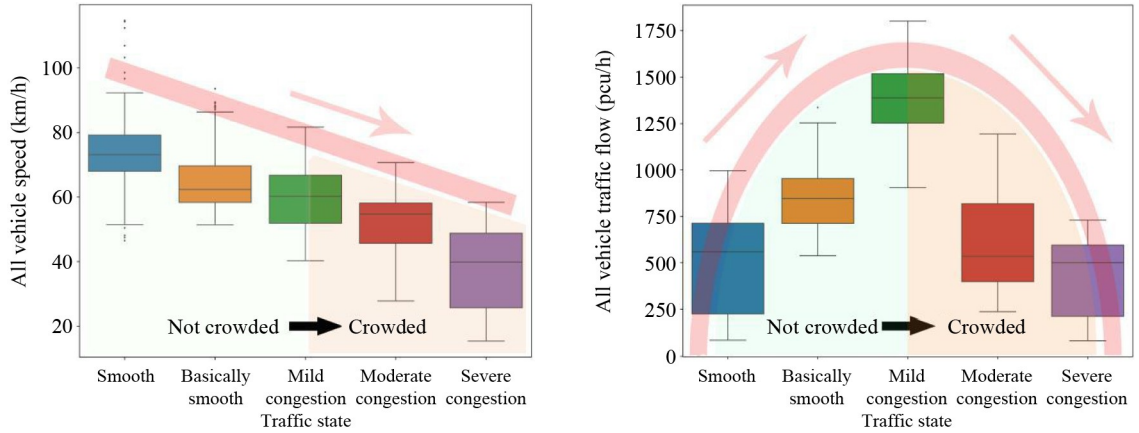


Fig. 10 The cluster center position obtained after 9 bootstrap operations.

Table 5 The cluster center position obtained by the bootstrap operation

Cluster center	Bootstrap 1		Bootstrap 2		Bootstrap 3		Bootstrap 4		Bootstrap 5	
	z_1	z_2	z_1	z_2	z_1	z_2	z_1	z_2	z_1	z_2
Cluster 0	4.22	−9.57	6.73	−8.31	6.80	−9.60	5.58	−8.90	5.06	−9.62
Cluster 1	7.71	8.36	5.99	8.69	6.04	8.72	8.14	9.45	6.47	8.05
Cluster 2	−18.92	−4.40	−15.44	−7.03	−16.79	−7.00	−18.35	−5.47	−17.51	−6.60
Cluster 3	23.49	−2.58	24.39	−2.50	23.77	−1.97	23.73	−2.79	23.42	−2.52
Cluster 4	−10.73	7.17	−14.22	7.10	−12.68	6.90	−11.00	6.69	−12.68	7.25

Cluster center	Bootstrap 6		Bootstrap 7		Bootstrap 8		Bootstrap 9		Final position	
	z_1	z_2	z_1	z_2	z_1	z_2	z_1	z_2	z_1	z_2
Cluster 0	6.23	−9.60	4.47	−9.12	7.23	−9.14	4.19	−9.12	5.61	−9.22
Cluster 1	5.33	7.67	8.11	9.34	6.05	8.39	9.39	10.25	7.02	8.77
Cluster 2	−17.67	−6.48	−19.03	−4.95	−16.63	−7.38	−18.69	−5.57	−17.67	−6.10
Cluster 3	22.40	−2.37	22.35	−3.00	24.06	−2.17	23.35	−2.82	23.44	−2.53
Cluster 4	−11.78	6.55	−10.62	6.97	−13.25	6.82	−10.08	6.60	−11.89	6.89



(a) All-vehicle speed (b) All-vehicle traffic flow

Fig. 11 Arranged box plot of the traffic state with traffic features.

flow (pcu/h), the upper limits indicating the transition from smooth to severe congestion are 0.00, 3.17, 8.25, 5.25, and 0.17, respectively. Initially, these limits increase before subsequently decreasing. The flow traffic features include the following: percent time spent following, time occupancy rate, small truck traffic flow, large bus traffic flow, medium truck traffic flow, large truck traffic flow, extralarge truck traffic flow, all-truck traffic flow, and all-vehicle traffic flow. Additionally, when the traffic state is unblocked, the upper and lower

Table 6 The fuzzy interval range of each traffic feature in each traffic state

Feature	Range	Smooth	Basically smooth	Mild congestion	Moderate congestion	Severe congestion
Percent time spent following (%)	Upper limit	39.97	42.25	49.29	36.48	23.67
	Lower limit	96.77	97.76	98.19	97.94	97.21
Time occupancy rate (%)	Upper limit	1.13	2.76	5.01	1.25	0.29
	Lower limit	26.73	28.42	32.02	22.88	19.21
Small truck traffic flow (pcu/h)	Upper limit	2.50	9.67	38.25	5.75	2.50
	Lower limit	527.75	638.75	736.00	643.50	500.75
Small truck speed (km/h)	Upper limit	49.08	42.32	32.45	16.77	13.04
	Lower limit	116.90	98.61	94.09	76.59	62.88
Large bus traffic flow (pcu/h)	Upper limit	0.00	3.17	8.25	5.25	0.17
	Lower limit	96.83	213.33	282.33	247.25	175.00
Large bus speed (km/h)	Upper limit	32.00	22.20	15.22	10.00	0.00
	Lower limit	93.25	85.39	73.52	69.93	65.79
Medium truck traffic flow (pcu/h)	Upper limit	0.00	5.33	33.50	7.33	3.00
	Lower limit	272.33	357.00	388.25	368.83	122.17
Medium truck speed (km/h)	Upper limit	34.43	26.49	20.69	12.74	11.50
	Lower limit	112.02	101.62	93.46	76.85	56.90
Large truck traffic flow(pcu/h)	Upper limit	0.75	1.25	11.75	0.50	0.00
	Lower limit	89.00	185.33	252.67	224.50	63.00
Large truck speed (km/h)	Upper limit	20.20	16.21	14.70	7.31	3.84
	Lower limit	74.90	65.31	59.67	47.69	35.80
Extralarge truck traffic flow (pcu/h)	Upper limit	0.00	5.65	17.17	6.25	0.00
	Lower limit	268.17	375.67	749.00	562.75	338.25
Extralarge truck speed (km/h)	Upper limit	0.00	0.00	0.00	0.00	0.00
	Lower limit	100.47	98.34	96.81	88.57	43.97
All-truck traffic flow (pcu/h)	Upper limit	35.17	75.50	274.17	48.50	35.83
	Lower limit	812.00	833.00	1163.50	850.25	563.00
All-truck speed (km/h)	Upper limit	44.94	31.24	30.60	19.73	13.06
	Lower limit	118.16	112.60	101.71	85.07	71.77
All-passenger car speed (pcu/h)	Upper limit	48.53	42.19	34.43	27.34	17.17
	Lower limit	101.89	87.98	77.90	75.41	72.95
All-vehicle traffic flow (pcu/h)	Upper limit	87.28	537.50	903.50	236.75	84.00
	Lower limit	995.56	1336.17	1801.50	1193.25	729.50
All-vehicle speed (km/h)	Upper limit	53.00	51.41	40.27	27.82	15.39
	Lower limit	114.66	93.58	81.66	70.70	58.35
All-truck mixed flow rate (%)	Upper limit	9.00	10.26	21.21	24.87	28.45
	Lower limit	67.25	79.49	88.27	90.76	92.02

limits of these features are at their lowest, whereas they are at their highest when the traffic state reaches mild congestion. It is important to note that when the traffic state reaches mild congestion, the upper and lower limits of these features are at their highest. Similarly, as the traffic becomes severely congested, the upper and lower limits of the aforementioned features gradually decrease from their highest values to their lowest values.

The trend observed in the upper and lower limits of the speed traffic features differs from that of the flow traffic features. The speed traffic features consist of the following: small truck speed, large bus speed, medium truck speed, large truck speed, extralarge truck speed, all-truck speed, all-passenger car speed, and all-vehicle speed. As the traffic becomes more crowded, the upper and lower limits of the speed traffic features gradually decrease.

This observation is consistent with the relationship between traffic state and speed stated in the Greenshields FD.

6 Discussion

6.1 Comparison with a conventional model

To validate the performance of the TSE for heterogeneous traffic, conventional methods were employed for comparison purposes. Since previous studies did not consider the various traffic features of heterogeneous vehicles, the all-vehicle flow and average speed were selected as the factors in the comparison model. By focusing on these two-dimensional traffic features, feature selection and dimension reduction operations were not necessary, allowing the direct application of the FCM algorithm to cluster traffic states. It should be noted that determining the optimal number of traffic state clusters under two-dimensional traffic features was a prerequisite. The elbow method was employed for the conventional model to identify the optimal number of clusters. The cluster results indicated that there are five levels of traffic states.

In addition, Figs. 12(a) and 12(b) display the box plots for all-vehicle speed and all-vehicle traffic flow under three different traffic state clusters. To illustrate the decreasing trend in the all-vehicle speed box plot, the traffic states are sorted in the order [0, 2, 1]. This sorting reveals that as the all-vehicle speed range gradually decreases, the all-vehicle traffic flow range gradually increases. This relationship between speed and flow, observed under the three clusters, differs from the results obtained using the proposed HiF-TSE model. The comparison indicates that the proposed HiF-TSE model, which is designed for heterogeneous traffic, provides a more accurate representation of traffic states than does the conventional model.

Figures 12(c) and 12(d) present the box plots for all-vehicle speed and all-vehicle traffic flow under a cluster number of five, using the 2D traffic feature data for further analysis. To illustrate the decreasing trend in the all-vehicle speed box plot, the cluster series are sorted in the order [3, 0, 1, 2, 4]. The overlap of the upper edge of the purple box with the red box, as highlighted by the blue circle, indicates that the traffic state estimation results obtained using the 2D traffic feature data cannot be effectively distinguished. Furthermore, there is a

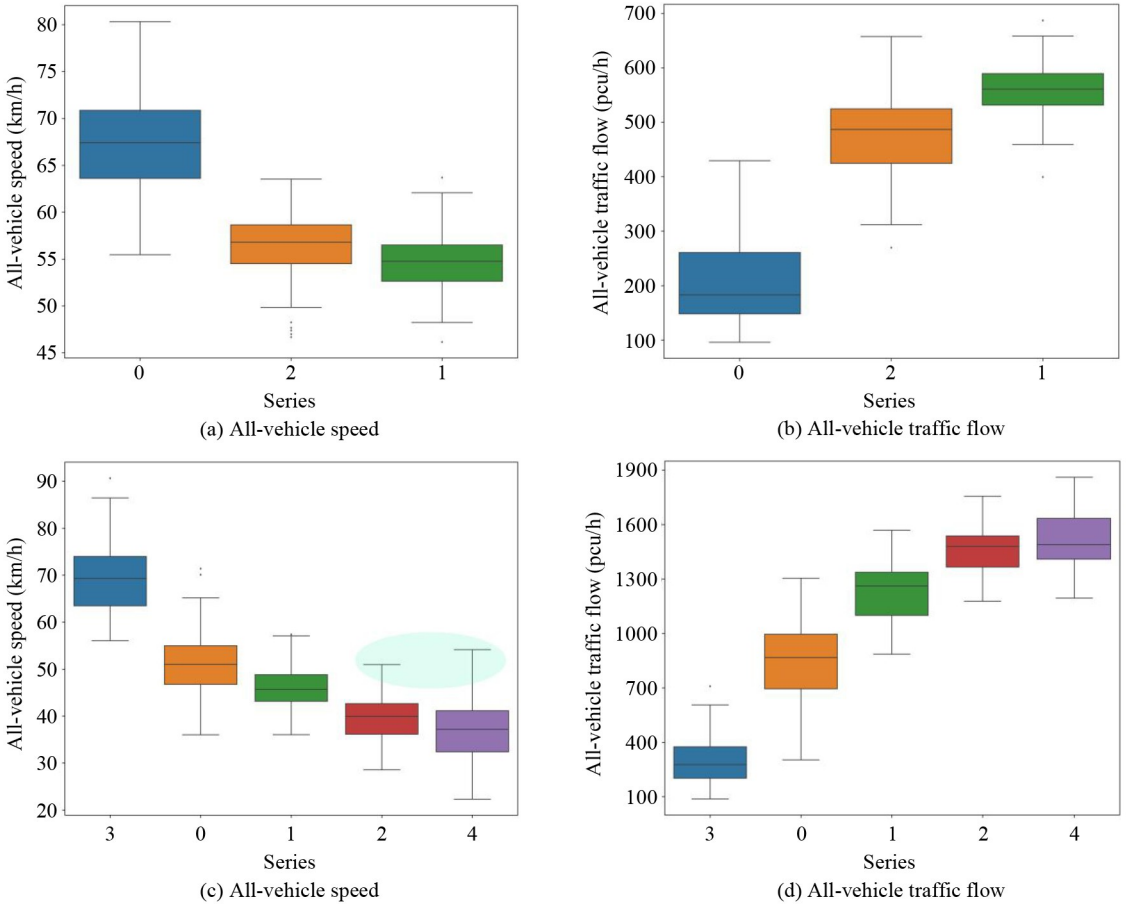


Fig. 12 Box plot of traffic features under 3/5 traffic state levels

monotonic relationship between the all-vehicle speed and all-vehicle traffic flow, similar to the trend observed for the three clusters. Regardless of whether the analysis is performed with three clusters determined by the elbow method or with five clusters, the TSE results obtained using simple 2D traffic features do not accurately reflect the traffic state. In contrast, the traffic features of heterogeneous vehicles proposed in this paper provide more accurate classification results.

6.2 Comparison with different vehicle compositions

Furthermore, the TSE results were compared for groups with similar all-vehicle flow values but with different compositions of heterogeneous traffic flow. Table 7 lists three such groupings along with their corresponding TSE results.

As depicted in Table 7, it is evident that the TSE results using the proposed HiF-TSE model are consistently smooth for No. 1-1, with an all-vehicle traffic flow value of 1621.50 pcu/h. Conversely, for Nos. 1-2, 1-3, 1-4, and 1-5, which have similar vehicle traffic flows to those of No. 1-1, the TSE results using the HiF-TSE model indicate mild congestion. Although the all-vehicle traffic flow values are similar, the other traffic features for Nos. 1-2, 1-3, 1-4, and 1-5 differ from those for No. 1-1. This suggests that even with similar all-vehicle traffic flows, the traffic state varies due to the different compositions of vehicles within heterogeneous traffic. Therefore, the stability of traffic flows is directly influenced by the composition of the vehicle types in heterogeneous traffic, highlighting the necessity of considering vehicle type compositions for accurate traffic state estimations.

Additionally, Table 7 reveals that the TSE result for No. 1-1 using the conventional model indicates moderate congestion, which contrasts with the TSE result using the HiF-TSE model. Furthermore, when applying the Greenshields model with an all-vehicle traffic flow of 1621.50 pcu/h and an all-vehicle speed of 80.83 km/h, the traffic state of No. 1-1 is determined to be “not crowded” rather than “moderate congestion.”

Moreover, despite the similar values of all-vehicle traffic flows, the proposed HiF-TSE model demonstrates greater sensitivity to changes in different vehicle type compositions. This sensitivity is evident in the varying TSE results between No. 1-1 and the other group members (Nos. 1-2, 1-3, 1-4, and 1-5). Conversely, the TSE results for the entire No. 1 group are consistently the same when determined by the conventional model, which indicates a lack of sensitivity to changes in traffic features. Furthermore, the results for groups two and three closely resemble those of group one.

To summarize, this section emphasizes that the traffic feature index system and the proposed HiF-TSE model exhibit greater sensitivity to differences in vehicle type compositions within heterogeneous traffic than does the conventional model. This strongly supports the belief that the proposed HiF-TSE model accurately reflects real traffic states.

7 Conclusions

Accurate TSEs play a crucial role in enhancing traffic management efficiency and informing traffic policies to alleviate congestion and enhance service quality. However, the accuracy of TSEs has been compromised due to the neglect of the heterogeneous nature of road traffic in previous studies. Our research addresses this gap by contributing to TSE studies in three main areas: first, by introducing a TSE traffic feature index system that accounts for heterogeneous traffic; second, by proposing a novel high-dimensional fuzzy TSE model, termed HiF-TSE, which effectively considers traffic heterogeneity; and third, by conducting a case study on a real road network to compare TSE results obtained using new and conventional methods.

Unlike previous studies that focused solely on a single vehicle type, the new TSE traffic feature index system embraces heterogeneous traffic and includes four feature categories: basic traffic parameters, average speed, traffic flow, and mixed flow rate of heterogeneous vehicles. This new index system can accurately depict the operational characteristics of multiple vehicle types.

The HiF-TSE model consists of three essential processes: feature selection, dimension reduction, and fuzzy clustering. The feature selection filters redundant traffic features using Spearman correlation coefficients. Dimension reduction employs the TSNE algorithm to reduce high-dimensional traffic feature data to two dimensions, thus enhancing the efficiency of subsequent clustering. Fuzzy clustering utilizes the FCM machine learning algorithm to partition two-dimensional data into several clusters. Together, these processes of the HiF-TSE model offer a significant advantage in reducing the computational load and improving the efficiency of TSE processing.

An analysis of the case study data demonstrates the accuracy of the HiF-TSE model, indicating that the optimal number of traffic state levels in TSEs is 5. This aligns well with the classic Greenshields fundamental model, which defines 5 traffic state levels: severe congestion, moderate congestion, mild congestion, unblocked, and smooth.

The HiF-TSE results were compared to those obtained using the conventional model, which utilizes an index designed for a single standard vehicle type. The comparison reveals that when the traffic feature data have the same values for overall traffic flow but different compositions of vehicle types, the TSE results differ. This discrepancy arises because the HiF-TSE model is more sensitive to changes in vehicle type compositions,

Table 7 Comparison of the TSE results of the HiF-TSE model and conventional model for heterogeneous traffic

No.	Percent time spent following (%)	Time occupancy rate (%)	Small truck traffic flow (pcu/h)	Small truck speed (km/h)	Large bus traffic flow (pcu/h)	Large bus speed (km/h)	Medium truck traffic flow (pcu/h)	Medium Truck speed (km/h)	Large truck traffic flow (pcu/h)	Large truck S\speed (km/h)	Extra large-truck traffic flow (pcu/h)	Extra large truck speed (km/h)	All truck traffic flow (pcu/h)	All truck speed (km/h)	All passenger car traffic flow (pcu/h)	All passenger car speed (km/h)	All vehicle traffic flow (pcu/h)	All vehicle speed (km/h)	All truck traffic flow rate (%)	Traffic state	
																				HiF-TSE model	Conventional model
1-1	73.10	8.80	326.83	82.97	34.17	56.38	77.17	80.16	55.33	73.19	214.33	82.30	673.67	98.98	947.83	67.92	1621.50	80.83	41.55	basically smooth	moderate congestion
1-2	66.93	12.44	459.00	64.08	50.50	49.63	247.00	60.49	96.00	48.12	59.67	60.37	861.67	59.09	769.50	46.14	1631.17	52.98	52.83	mild congestion	moderate congestion
1-3	66.58	8.41	364.83	70.58	22.00	71.05	51.50	71.83	31.83	53.92	38.17	67.95	486.33	70.19	1136.00	56.46	1622.33	60.57	29.98	mild congestion	moderate congestion
1-4	72.78	8.49	297.33	64.83	31.17	61.17	52.33	61.05	74.67	46.62	43.50	66.17	467.83	58.68	1168.33	67.41	1636.17	64.91	28.59	mild congestion	moderate congestion
1-5	65.07	10.25	359.50	68.09	35.83	65.25	75.33	64.89	63.50	54.15	64.50	67.14	562.83	66.19	1073.50	50.70	1636.33	56.03	34.40	mild congestion	moderate congestion
2-1	73.10	9.04	441.67	77.44	36.00	56.69	101.00	77.75	94.67	75.88	0.17	181.00	84.12	818.33	86.90	65.46	1799.83	75.21	4.67	basically smooth	moderate congestion
2-2	73.03	13.42	506.33	77.88	56.00	41.53	194.00	61.08	140.67	52.54	0.25	104.67	69.84	945.67	67.92	49.37	1767.83	59.29	3.95	mild congestion	moderate congestion
2-3	73.39	11.26	300.50	65.65	25.17	44.53	70.83	58.82	152.67	42.98	0.27	76.33	74.89	600.33	55.04	60.80	1778.67	58.86	4.21	mild congestion	moderate congestion
2-4	66.53	10.95	321.17	65.43	26.33	55.77	65.50	58.54	79.00	47.16	0.14	75.67	69.47	541.33	62.35	51.42	1743.67	54.82	3.98	mild congestion	moderate congestion
2-5	71.45	10.83	351.67	69.75	30.50	55.32	67.50	61.73	84.83	50.04	0.15	46.00	72.97	550.00	62.62	58.55	1797.67	59.79	4.06	mild congestion	moderate congestion
3-1	66.88	7.44	343.83	87.26	30.33	69.20	71.33	82.73	44.33	84.06	0.08	80.17	85.11	539.67	91.74	68.31	1602.17	76.21	5.31	basically smooth	moderate congestion
3-2	98.19	24.23	562.00	63.14	54.50	49.60	437.75	55.39	125.50	21.71	0.23	228.50	20.55	1353.75	51.35	55.19	1595.50	51.93	1.29	mild congestion	moderate congestion
3-3	64.76	10.97	290.83	62.51	34.67	56.10	72.17	60.76	79.33	53.51	0.14	130.50	60.71	572.83	65.55	49.69	1602.67	55.36	3.79	mild congestion	moderate congestion
3-4	57.33	11.80	392.17	71.29	30.00	56.06	170.17	69.39	171.50	58.34	0.31	178.83	73.68	912.67	69.62	46.00	1610.33	59.39	4.58	mild congestion	moderate congestion

particularly in heterogeneous traffic. Consequently, it accurately reflects actual traffic states more effectively.

Moreover, the proposed HiF-TSE model is not restricted to applications involving heterogeneous traffic. It can also be extended to emerging traffic, which is composed of both autonomous vehicles and human-driven vehicles. The different operational characteristics exhibited by these two vehicle types have been widely demonstrated. However, the variations in traffic states resulting from different compositions of autonomous and human-driven vehicles require further investigation, possibly through future studies.

Author contributions

The authors confirm their contribution to the paper as follows: study conception and design: SY W, CJ D, M M; data collection: CJ D; analysis and interpretation of results: SY W, CF S, SD L; draft manuscript preparation: SY W, J Z, M M. All authors reviewed the results and approved the final version of the manuscript.

Competing Interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Al Mallah R, Quintero A, Farooq B (2017). Distributed classification of urban congestion using VANET. IEEE Transactions on Intelligent Transportation Systems, 18(9): 2435–2442

Bai J, Li K (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. Review of Economics and Statistics, 98(2): 298–309

Bahadur S B, Dhanalakshmi R (2020). Building a fuzzy logic-based artificial neural network to uplift recommendation accuracy. The Computer Journal, 63(11): 1624–1632

Banerjee S, Monni S (2021). An orthogonally equivariant estimator of the covariance matrix in high dimensions and for small sample sizes. Journal of Statistical Planning and Inference, 213: 16–32

Bhaskar A, Tsubota T, Kieu L M, Chung E (2014). Urban traffic state estimation: Fusing point and zone based data. Transportation Research Part C, Emerging Technologies, 48: 120–142

Celikoglu H B, Silgu M A (2016). Extension of traffic flow pattern dynamic classification by a macroscopic model using multivariate clustering. Transportation Science, 50(3): 966–981

Cheng Z, Wang W, Lu J, Xing X (2020). Classifying the traffic state of urban expressways: A machine-learning approach. Transportation Research Part A, Policy and Practice, 137: 411–428

Duan Y, Yang C, Chen H, Yan W, Li H (2021). Low-complexity point cloud denoising for LiDAR by PCA-based dimension reduction. Optics Communications, 482: 126567

Erfani M, Baalousha M, Goharian E (2023). Unveiling elemental fingerprints: A comparative study of clustering methods for multi-element nanoparticle data. Science of the Total Environment, 905: 167176

Gao Z, Huang H, Guo J, Yang L, Wu J (2023). Future urban transport management. Frontiers of Engineering Management, 10(3): 534–539

Gashaw S, Goatin P, Härri J (2018). Modeling and analysis of mixed flow of cars and powered two wheelers. Transportation Research Part C, Emerging Technologies, 89: 148–167

Greenshields B D (1935). A study of traffic capacity. In: Proceedings of 14th Annual Meeting of Highway Research Board, HRB, Washington, D. C.. 14(1): 448–477

Guan D, Chen K, Han G, Huang S, Yuan W, Guizani M, Shu L (2021). A novel class noise detection method for high-dimensional data in industrial informatics. IEEE Transactions on Industrial Informatics, 17(3): 2181–2190

Han Y, Zhang M, Guo Y, Zhang L (2022). A streaming-data-driven method for freeway traffic state estimation using probe vehicle trajectory data. Physica A, 606: 128045

Hoogendoorn S P, Bovy P H L (2000). Continuum modeling of multiclass traffic flow. Transportation Research Part B: Methodological, 34(2): 123–146

Huang W, Lu C, Fang D (2021). Special issue: City and infrastructure engineering and management. Frontiers of Engineering Management, 8(1): 1–4

Hyun K, Mitra S K, Jeong K, Tok A (2021). Understanding the effects of vehicle platoons on crash type and severity. Accident Analysis and Prevention, 149: 105858

Jamshidnejad A, Lin S, Xi Y, De Schutter B (2019). Corrections to "integrated urban traffic control for the reduction of travel delays and emissions". IEEE Transactions on Intelligent Transportation Systems, 20(5): 1978–1983

Kong D, List G F, Guo X, Wu D (2018). Modeling vehicle car-following behavior in congested traffic conditions based on different vehicle combinations. Transportation Letters, 10(5): 280–293

Krampe J, Junge M (2021). Deriving functional safety (ISO 26262) S-parameters for vulnerable road users from national crash data. Accident Analysis and Prevention, 150: 105884

Li X, Li X, Xiao Y, Jia B (2016). Modeling mechanical restriction differences between car and heavy truck in two-lane cellular automata traffic flow model. Physica A, 451: 49–62

Lin S, De Schutter B, Xi Y, Hellendoorn H (2013). Integrated urban traffic control for the reduction of travel delays and emissions. IEEE Transactions on Intelligent Transportation Systems, 14(4): 1609–1619

Ling S, Ma S, Jia N (2022). Sustainable urban transportation development in China: A behavioral perspective. Frontiers of Engineering Management, 9(1): 16–30

Liu S, Hellendoorn H, De Schutter B (2017). Model predictive control for freeway networks based on multi-class traffic flow and emission models. IEEE Transactions on Intelligent Transportation Systems, 18(2): 306–320

Lyu Z, Hu X, Zhang F, Liu T, Cui Z (2022). Heterogeneous traffic flow characteristics on the highway with a climbing lane under different truck percentages: The framework of Kerner's three-phase traffic theory. Physica A, 587: 126471

Nidheesh N, Abdul Nazeer K A, Ameer P M (2017). An enhanced deterministic K-means clustering algorithm for cancer subtype prediction from gene expression data. Computers in Biology and Medicine, 91: 213–221

Pezzotti N, Lelieveldt B P F, Maaten L, Hollt T, Eisemann E, Vilanova A (2017). Approximated and user steerable *t*-SNE for progressive visual analytics. IEEE Transactions on Visualization and Computer Graphics, 23(7): 1739–1752

Piantadosi J, Howlett P, Boland J (2007). Matching the grade correlation coefficient using a copula with maximum disorder. Journal of Industrial and Management Optimization, 3(2): 305–312

Portilla C, Espinosa J, De Schutter B (2020). A multi-class urban traffic model considering heterogeneous vehicle composition: An extension of the S model. Transportation Research Part C, Emerging Technologies, 115: 102613

Puente C, Palacios R, González-Arechavala Y, Sánchez-Úbeda E F (2020). Non-intrusive load monitoring (NILM) for energy disaggregation using soft computing techniques. Energies, 13(12): 3117

Romo A, Hernandez S, Cheu R L (2014). Identifying precrash factors for cars and trucks on interstate highways: Mixed logit model approach. Journal of Transportation Engineering, 140(3): 04013016

Ruan T, Zhou L, Wang H (2021). Stability of heterogeneous traffic considering impacts of platoon management with multiple time delays. Physica A, 583: 126294

Singh T, Saxena N (2021). Chaotic sequence and opposition learning guided approach for data clustering. Pattern Analysis & Applications, 24(3): 1303–1317

Tian J, Song X, Tao P, Liang J (2022). Pattern-adaptive generative adversarial network with sparse data for traffic state estimation. Physica A, 608: 128254

Xu F, He Z, Sha Z, Sun W, Zhuang L (2013). Traffic state evaluation based on macroscopic fundamental diagram of urban road network. Procedia: Social and Behavioral Sciences, 96: 480–489

Yang F, Wang M (2020). A review of systematic evaluation and improvement in the big data environment. Frontiers of Engineering Management, 7(1): 27–46

Yu H, Chen L, Yao J, Wang X (2019). A three-way clustering method based on an improved DBSCAN algorithm. Physica A, 535: 122289

Zhang J, Mao S, Yang L, Ma W, Li S, Gao Z (2024). Physics-informed deep learning for traffic state estimation based on the traffic flow model and computational graph method. Information Fusion, 101: 101971

Zhang Y, Lu Z, Wang J, Chen L (2023). FCM-GCN-based upstream and downstream dependence model for air traffic flow networks. Knowledge-Based Systems, 260: 110135

Zheng Z, Su D (2016). Traffic state estimation through compressed sensing and markov random field. Transportation Research Part B: Methodological, 91: 525–554

Zhu W, Webb Z T, Mao K, Romagnoli J (2019). A deep learning approach for process data visualization using *t*-distributed stochastic neighbor embedding. Industrial & Engineering Chemistry Research, 58(22): 9564–9575

Zong W, Chow Y W, Susilo W (2020). Interactive three-dimensional visualization of network intrusion detection data for machine learning. Future Generation Computer Systems, 102: 292–306