IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 4, 2020, accepted April 21, 2020, date of publication April 30, 2020, date of current version May 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991638

A Two-Dimensional Clustering Method for High-Speed Railway Trains in China Based on Train Characteristics and Operational Performance

LIANBO DENG, YUXIN CHEN, RUNFA WU, QING WANG, AND YIMING XU
Rail Data Research and Application Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China

Corresponding author: Lianbo Deng (lbdeng@csu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71471179 and Grant U1834209, in part by the Hunan Postgraduate Innovation Project under Grant 150110005, and in part by the Fundamental Research Funds for the Central Universities of Central South University under Grant 2019zzts272.

ABSTRACT At present, China has the world's longest high-speed railway (HSR) network, but most trains do not have clear market positioning and hierarchical standard. The current train hierarchies and adjustment is empirically set by HSR organisers. This paper applies a clustering method and cross-over analysis to study the classification of China's HSR trains and provides scientific operation suggestions. By adopting the timetable data and ticket booking data of Shanghai-Nanjing intercity railway, we establish a clustering index system consists of train characteristics indexes and operational performance indexes, then use t-distributed Stochastic Neighbour Embedding (t-SNE) to do dimensionality reduction for original data. We obtain the optimal clustering number by validity indexes and use k-means to cluster the HSR trains. After clustering, we use a cross-over analysis to illustrate the relationship between train characteristics, passenger demand and operational performance. We find there are two main types of train on Shanghai-Nanjing intercity HSR: trains departing on the hour that have good operational performance, trains with staggered stops and low-capacity based on the strategies of "low capacity, high density" that have better operational performance at peak time. After 19:00, due to the passenger demand decrease, train capacity can be reduced to avoid unnecessary waste. Short-distance trains could easily be replaced by long-distance trains with similar stop schedules and need to maintain a certain operation frequency. The proposed clustering method has universal applicability for Chinese HSR lines.

INDEX TERMS High-speed railway train, clustering analysis, train characteristics, operational performance, cross-over analysis.

I. INTRODUCTION

By the end of 2018, business mileage on Chinese railways had reached 131,000 km, of which high-speed railway (HSR) mileage accounted for 29,000 km. This means that China has built the largest HSR network in the world [1]. There are 3970.5 pairs of passenger trains per day; of these, multiple-unit trains make up 2775 pairs, or 69.8% of passenger trains, and these trains have become the main mode of railway passenger transport. With the expansion of HSR network and passenger transport market, trains operating on the line and passenger transport products are more complex and diversified, which raise new requirements for the operation and adjustment of HSR trains. Studying the classification of HSR trains and analysing the operational performance of trains with different train characteristics (stop schedule, travel speed, departure time, etc.), provide scientific operation suggestions for adjusting actual train planning and contribute great significance for improving both operational efficiency and social benefits.

At present, many countries with a certain scale of HSR, such as Japan, France and Germany, have formed

The associate editor coordinating the review of this manuscript and approving it for publication was Jesus Felez.

a well-established system for HSR train, which is differentiated by stop schedule and travel speed, to meet diversity of passenger demands, and improve the railway company's income. The Tokaido Shinkansen is the longest-running HSR line in Japan. There are three main train types operating on the line, namely, Nozomi, Hikari and Kodama. The main difference between them is the number of stops and the grades of stations. Nozomi only stop at large stations with high frequency. Hikari stop both large stations and some smaller stations. The first two types are trains with few stops and high travel speed in long mileage. Kodama stop at every station with low travel speed in small mileage, but can improve service accessibility between stations on the HSR line. These three train types are mainly differentiated by stop schedule, forming a hierarchical timetable with different travel speed for different passenger demand. Thus, organisers can design various ticket options to improve income.

China has some initial experience in train classification, but not clearly and specifically. Since the first HSR train line taken into operation in 2008, China has referred to its HSR trains called G-series, D-series and C-series trains. The first two series are based on their running speed (which correspond to the speed standard of HSR lines they mainly serve), while C-series serves intercity HSR lines. In order to avoid unnecessary waste of transport capacity, all trains actually operating on a HSR line have the same running speed, so the train on the HSR lines are divided by stop schedule and departure time. Currently, only portion of trains have obvious characteristics, such as those only stopping at provincial capitals and departing on the hour and those stopping at approximately every station. But most trains do not have clear market positioning and hierarchical standard. In general, the current hierarchies and adjustment of HSR trains is more like an empirical decision-making by the passenger transport organisers.

This paper attempts to apply clustering analysis method to fill the gap in HSR train classification. According to practical timetable data and ticket booking data, we cluster HSR trains based on train characteristics and operational performance, analyse actual characteristics of each cluster, evaluate their operational performance, and provide scientific operation suggestions for the design and adjustment of train hierarchies. The study can explore the coupling relationship between train characteristics and operational performance for HSR trains in China, then help to establish a hierarchical system of HSR train types that is suitable for the Chinese HSR network and passenger demand. Although due to the complexity of Chinese HSR network, the train types in different lines and regions are not the same. The research method proposed in this paper has universal applicability of Chinese HSR train.

The contribution of this paper is three aspects. (i) Based on actual timetable data and ticket booking data, we cluster HSR trains currently in service, which can reflect the actual characteristics of trains, their operational performance in passenger transport market, and the difference between them. The clustering research provides scientific operation suggestions for the design and adjustment of train hierarchies. In the past, the train hierarchies were mostly set up empirically by the organisers according to passenger transport market, and adjusted gradually in the operation process. (ii) We creatively take HSR trains as clustering research objects. According to a large number of actual ticket booking data of HSR system, we propose a clustering index system based on train characteristics and operational performance and a two-dimensional clustering method for HSR trains. (iii) We use a cross-over analysis to avoid the annihilation effect between indexes, reveal the relationship between train characteristics, passenger demand and operational performance, and then make reasonable operation suggestions for HSR system.

The remainder of the article is structured as follows. Section II describes related research work. Section III explains the building of an index system based on train characteristics and operational performance. Section IV presents a two-dimensional clustering method, using the t-SNE dimensionality reduction method, the k-means clustering method and validity indexes. Section V presents and analyses the results of clustering for Shanghai-Nanjing intercity HSR trains, based on train characteristics and operational performance, respectively. Section VI gives a cross-over analysis of Shanghai-Nanjing intercity HSR trains. Section VII presents the conclusion and discusses future work.

II. LITERATURE REVIEW

In this paper, we select two indexes of train characteristics and operational performance to cluster HSR trains, and put forward reasonable operation suggestions by analysing the relationship between them. Indexes of train characteristics are from the optimised contents that are generally concerned in researches of train timetable and train planning. Indexes of operational performance refer to the literatures that evaluate passenger service quality, compare different modes of transportation, and evaluate train timetable or train planning. In addition, clustering methods have been widely applied to passenger groups classification, transportation network nodes classification and traffic system evaluation. This section summarises existing literature on train characteristics, operational performance, and clustering methods used in the transportation field.

A. RESEARCH ON TRAIN CHARACTERISTICS

Many studies of train planning and train timetable construct models and algorithms to solve different optimisation objects (or multi-objectives). Claessens *et al.* [2] aimed to minimise train operating costs, and optimised the line, line types, routes, frequencies and train lengths (in the form of coaches). Ghoseiri *et al.* [3] reduced fuel consumption costs and shortened passenger journey time by optimising the origin and destination, paths and arrival and departure times. Bussieck *et al.* [4] adjusted the frequency of different train types within the railway network, such as InterCity (Express) trains (IC/ICE), InterRegio trains (IR), and connecting district town and commuter trains (CT), in order to meet passenger demand. Yue *et al.* [5] optimised train timetables based on travel routes, including the numbers of stops and stopping times. Kaspi and Raviv [6] input a pool of routes, the passenger demand, cycle time and safety and operational restrictions, and output an optimised timetable that can minimise passenger journey time and operational costs. Zhou and Zhong [7] proposed a generalised resource-constrained project scheduling formulation including segments, arrival and departure times, etc., and obtained a feasible timetable with guaranteed optimality for a single-track rail network. Repolho *et al.* [8] used a mixed-integer optimisation model to determine the optimal locations and numbers of trains on the railway line, and obtained an optimised stop schedule. To optimise the line planning, Fu *et al.* [9] according to stops, categorised trains into two classes: higher-classified trains and lower-classified trains, and gave priority to higher-classified trains in formulating and optimising a line planning.

The contents of train planning and train timetable, that express the characteristics of trains in different aspects, such as the optimisation objects, including stop schedule, origin and destination, capacity, the constraints and limits of optimisation objects, the optimised results, etc, can provide reference for our clustering index of train characteristics.

B. RESEARCH ON OPERATIONAL PERFORMANCE

Many literatures have adopted questionnaires combined with data analysis methods to evaluate train service quality and operational performance, or to select important factors that affect train service quality. Chou *et al.* [10] showed that passengers prioritised five attributes: car cleanliness, neat appearance of employee, employee service attitude, comfort of air conditioning, and on-time performance. Based on confirmatory factor analysis, Chou *et al.* [11] found that for the same service quality index, Taiwanese passengers were more concerned about the necessary facilities of infrastructure services, while Korean passengers paid more attention to frontline staff interaction. Nathanail [12] concluded that Hellenic Railway performed well in terms of safety, accuracy and service, but less well in terms of cleanliness and information provision to passengers.

Some studies have combined with market segmentation theory or have compared other transportation modes such as civil aviation, to make reasonable suggestions for the operation of trains. Dobruszkes [13] pointed out that if HSTs (high-speed trains) are to compete successfully with airplane, a number of factors come into play, including travel time, frequencies, fares, airline hubs, geographical structures of urban regions, etc. Wang *et al.* [14] showed that the operational performance of HSR trains is not only influenced by population densities and well-developed economies, but also by fares, frequency, travel distance, departure time and distance from airports/HSR station to the city centre.

Some research on evaluation of HSR train plan is for single train. Stoilova and Nikolova [15] evaluated transport plan with the transport satisfaction, average number of train stops, average distance travelled, average speed, reliability, availability of service with direct transport and transport capacity as indicators. Jiang *et al.* [16] evaluated train timetable by involving the number of alighting and boarding passengers, the train load factor, the number of passengers waiting for trains due to overcrowding in vehicles and the number of waiting passengers on the platform. Feng *et al.* [17] analysed the comprehensive effect of target speed, passenger capacity utilization rate and formulation of a HSR train on transport efficiency.

In the literatures that evaluate passenger service quality, compare different modes of transportation, and evaluate train timetable or train planning, the indexes reflecting the operation performance of HSR trains can provide guidance for this paper.

C. CLUSTERING METHODS USED IN THE TRANSPORTATION FIELD

The application of clustering analysis methods in the transportation field mainly focuses on three aspects. The first type of clustering analysis uses transportation objects such as passengers or goods as research objects. In the field of passenger transportation, some studies have put forward the target marketing strategies, considering the impact of passenger travel behaviour and other aspects in combination with market segmentation theory. Teichert *et al.* [18], Urban *et al.* [19] and Punel and Ermegun [20] conducted clustering analysis on groups of airline passengers, discussed various characteristics of different clusters, and proposed targeted marketing strategies. Lv *et al.* [21] and Liu *et al.* [22] obtained the critical characteristic variables, and used affinity propagation method and fuzzy C-means method, respectively to cluster HSR passengers. Li and Sun [23] divided goods into seven types to solve conflicts between the rapid increasing of cargo quantities and the customs limited supervision force, so that customs could focus on high risk level cargo.

The second type of clustering analysis involves transportation service nodes, which include ports, airports and railway stations. Cabral and Ramos [24] and Gianfranco *et al.* [25] clustered container ports to analyse the relationships between them. Vogel and Graham [26], Cui *et al.* [27] and Mayer [28] clustered airports, and explored the formation mechanism of airport groups and the relationships among airport categories, cargo types and geographical areas. Zhou *et al.* [29] and Zhao *et al.* [30] classified the stations in the HSR network according to the daily average volume of passenger traffic, laying the foundation for the design of a train planning.

The third type of clustering analysis aims to evaluate traffic safety, transportation service level, competitiveness, and related factors. Depaire *et al.* [31] examined the effectiveness of latent class clustering in identifying homogeneous types of traffic accident. Wei and Sun [32] combined an improved principal component analysis technique with clustering analysis to comprehensively evaluate regional traffic safety, while Yeo *et al.* [33] identified the components of competitiveness for ports and ranked container ports in Korea, China, and other countries.

On the one hand, clustering methods applied on railway field is mainly hierarchical clustering, which is not friendly to large sample and hard to visualise. On the other hand, the current clustering analysis research focuses on transportation network nodes classification, passenger groups classification according to passenger travel choice, while HSR trains classification keep not involved.

III. CONSTRUCTION OF A CLUSTERING SYSTEM FOR HSR TRAINS BASED ON TRAIN CHARACTERISTICS AND OPERATIONAL PERFORMANCE

Before we carry out a clustering analysis of trains, it is necessary to construct a scientific clustering index system. The rationality and comprehensiveness of the selected indexes directly affect the results of clustering analysis. Based on the contents of the train timetable and the characteristics of train passenger flow, the indexes are selected from the current statistical standards for railway passenger service in China, and a clustering index system for HSR trains is then established.

The train timetable includes origin and destination of train, travel distance, stop schedule, train grade and capacity. Let the HSR network be (V, E), where $V = \{v_1, v_2, \cdots, v_h\}$ is the set of stations and E is the set of segments. For each train $T \in \Omega$, the corresponding stop schedule is $V_T = \left(v_1, v_2, \cdots, v_{h(T)}\right)$, where $h(T)$ is the number of stations at which train T stops, including the origin and destination.

Let $RS(T)$ represent the set of origins and destinations (r, s) served by train T, namely $(r, s) \in RS(T)$. $Q_T(r, s)$ is the number of passengers on train T over (r, s), and $L_T(r, s)$ is the distance travelled by these passengers.

A. INDEXES BASED ON TRAIN CHARACTERISTICS

1) STOP SCHEDULE SCORE

Chinese railway stations are divided into six distinct grades according to passenger volume, freight volume, and technological volume. However, the grade of HSR stations is generally higher, and the function of the station grade in distinguishing the status of HSR stations is not apparent. In order to increase discrimination between HSR stations and describe the distribution of train stops, we divide HSR passenger stations within the study area $V = \{v_1, v_2, \cdots, v_h\}$ into τ grades according to the number of passengers dispatched by train. In view of [34], the importance of urban road network nodes in China, we take $\tau = 3$, that is, a central city node, a provincial capital city node, and a general county-level city node.

For any train T, the number of stops for grade j is m_j, the set of them is $M(T) = \left\{m_1, m_2, m_j \cdots, m_\tau\right\}$. $M(T)$ is used to describe the distribution of train stops. Thus, the stop schedule score for the train is:

$$P(T) = \frac{\sum_{j=1}^{\tau} \beta_j m_j}{h(T)} \tag{1}$$

where β_j represents the weight of different grades passenger stations. The train stop schedule score comprehensively reflects the stops made by the train and the grades of the stations, and trains with similar scores have strong similarities.

2) PASSING-STOPPING RATIO

For any train T, the passing-stopping ratio is the proportion of the number of its stops to the total number of stations passed through its route, (both of them exclude the origin and destination). It reflects the density of train stops, which is an important attribute of the train stop schedule.

$$\varphi(T) = \frac{h(T) - 2}{N(T)} \tag{2}$$

where $N(T)$ is the total number of stations along the route travelled by the train T, but not include the origin and destination.

3) OTHER INDEXES

We also choose other indexes to describe train characteristics, such as $w(T)$ to reflect the train travel distance, $t(T)$ the travel time, $v(T)$ the travel speed, and b the train capacity.

B. INDEXES BASED ON OPERATIONAL PERFORMANCE

1) TRAIN LOAD FACTOR

The train load factor refers to the ratio of the total number of kilometres travelled by passengers on the train to the number of kilometres travelled by all seats. It also indicates the ratio of passenger turnover to train quota turnover. This factor reflects whether seats are fully utilised

$$\gamma(T) = \frac{\sum_{(r,s)\in RS(T)} Q_T(r, s) \times L_T(r, s)}{b \times w(T)} \tag{3}$$

2) NUMBER OF PASSENGERS DISPATCHED

The number of passengers dispatched refers to the total number of passengers travelling on a train at a certain time.

$$Q(T) = \sum\nolimits_{(r,s)\in RS(T)} Q_T(r, s) \tag{4}$$

3) TRAIN PASSENGER TURNOVER

The train passenger turnover is the number of passenger kilometres completed by each train within a certain period, which can be expressed as the product of traffic volume and transportation distance. Train passenger turnover is one of the most important factors in a railway passenger traffic plan.

$$\Psi(T) = \sum\nolimits_{(r,s)\in RS(T)} Q_T(r, s) \times L_T(r, s) \tag{5}$$

4) FARE INCOME

This reflects the ticket revenue of the train, that is, the sum of the ticket revenue for the different types of passengers.

$$\Gamma(T) = \sum\nolimits_{(r,s)\in RS(T)} Q_T(r, s) \times p_T(r, s) \tag{6}$$

where $p_T(r, s)$ is the ticket price in the (r, s) section of train T. The higher the fare income, the better the economic benefit, meaning that more people are willing to take this train.

TABLE 1. HSR train clustering index system.

Objective	Partition criterion	Variable
HSR train clustering index system	Train characteristics	Stop schedule score
		Passing-stopping ratio
		travel distances
		travel time
		travel speed
		Train capacity
	Operational performance	Train load factor
		Number of passengers dispatched
		Train passenger turnover
		Fare income
		Passenger average haul distance

5) PASSENGER AVERAGE HAUL DISTANCE

This refers to the average distance that each passenger is transported within a certain period, and is usually expressed as the ratio of passenger turnover to passenger volume. It is an important index used to analyse passenger behaviour characteristics and to reflect travel demand.

$$L(T) = \frac{\sum_{(r,s)\in RS(T)} Q_T(r,s) \times L_T(r,s)}{Q(T)} \tag{7}$$

In conclusion, the clustering index system of HSR trains based on train characteristics and operational performance is established, as shown in Table 1.

IV. TWO-DIMENSIONAL CLUSTERING METHOD OF HSR TRAINS

Using t-distributed Stochastic Neighbour Embedding (t-SNE) dimensionality reduction, we extract the features of the constructed HSR train clustering indexes. t-SNE dimension reduction uses a conditional probability to express the distance relationship, and is different from linear dimension reduction, which uses the Euclidean distance to describe the similarities between data. In high-dimensional space, t-SNE adopts a Gaussian distribution, while in low-dimensional space a t-distribution is used, which has a longer tail. As a result, t-SNE is one of only a few algorithms that consider both global and local structures at the same time, which is helpful when visualising data.

The k-means clustering algorithm is a classical partition-based clustering algorithm. Based on the proximity principle, data points are allocated to the nearest cluster to adjust the central position of each cluster until no further change is seen. Thus, by combining validity indexes to determine the k value, we use a two-dimensional clustering method for Shanghai-Nanjing intercity HSR trains, based on indexes of train characteristics and operational performance, and the k-means clustering algorithm.

Due to the obvious differences between the train characteristics and operational performance, we cluster them separately to acquire two types of train clusters, and carry out a cross-over analysis on these two clustering results. The advantages of this approach are: (i) it avoids the annihilation effect between the indexes of train characteristics and operational performance, i.e. the influence of one aspect or its index items will not be covered by another aspect or index; and (ii) using cross-over analysis, the results of clustering trains based on train characteristics can be investigated in terms of their operational performance. Hence, the service positioning of train clusters can be clarified, and suggestions for train clusters structure improvement can be put forward.

In addition, the data processing, analysis of example and visualisations in this paper are made by the programming of Python 3.7.

A. t-SNE DIMENSIONALITY REDUCTION

Stochastic Neighbour Embedding (SNE) is a manifold-based information theory learning method that was proposed by Hinton and Roweis [35]. Based on a Gaussian distribution, the Euclidean distance, which usually describes spatial data relations, is converted into a conditional probability of similarity. The aim is to minimise the distribution conditional probability difference between high and low dimensions, in order to maintain the data manifold structure when the data is mapped from high-dimensional space to low-dimensional space.

For any two high-dimensional data points x_i and x_j, SNE defines $p_{j|i}$ as the conditional probability that x_i has x_j as its neighbour:

$$p_{j|i} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i{}^2}\right)}{\sum_{k\neq i} \exp\left(\frac{-\|x_i - x_k\|^2}{2\sigma_i{}^2}\right)} \tag{8}$$

where σ_i is the Gaussian variance centred on data point x_i. Similarly, for high-dimensional data points x_i and x_j, the corresponding mapping points y_i and y_j have conditional probabilities $q_{j|i}$:

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k\neq i} \exp\left(-\|y_i - y_k\|^2\right)} \tag{9}$$

The purpose of SNE is to minimise the mismatch in the conditional probabilities between low-dimensional space and high-dimensional space. We use the Kullback-Leibler divergence, represented here as the cost function C, to measure this:

$$C = \sum_i KL\left(P_i \parallel Q_i\right) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{10}$$

where P_i is the conditional probability distribution of all other data points, while given a data point x_i; Q_i represents the conditional probability distribution of all other mapping points, while given a mapping point y_i. The best low-dimensional mapping can be obtained by using a gradient descent method to solve the cost function C.

However, SNE is hampered by optimising the cost function and solving the crowding problem. Hence, Maaten and Hinton [36] proposed a symmetric joint probability expression to replace the conditional probability in SNE:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \tag{11}$$

In low-dimensional space, they introduced a heavy-tailed distribution to describe the neighbour selection probability of low-dimensional mapping points:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{12}$$

The introduction of a t-distribution alleviates the crowding problem in the low-dimensional space. Numerous experimental results have shown that t-SNE gives better performance than many other dimensionality reduction methods for the same data sets, and can better maintain the neighbourhood structure of high-dimensional data. In this paper, we adopt t-SNE to reduce the dimensions of the train data for subsequent clustering and visual analysis.

B. K-MEANS CLUSTERING ALGORITHM

The k-means clustering algorithm is a classical approach, and is the most widely used clustering algorithm at present. It divides a data set containing n data points into k clusters, namely, $C_i\,(i = 1, 2, \cdots, k)$ so that the final result shows high similarity within clusters but low similarity between clusters. The specific steps of this algorithm are as follows:

(i) For a data set $X = \{x_1, x_2, \cdots, x_n\}$, randomly select k data points as the initial cluster centre $\mu_i\,(i = 1, 2, \cdots, k)$.

(ii) Using the Euclidean distance formula $J(C_i) = \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$, calculate the distance from each data point x_j to each cluster centre μ_i, and assign x_j to the nearest cluster.

(iii) Based on $\mu_i = \frac{1}{n_i}\sum_{x_j \in C_i} x_j$, recalculate each cluster centre μ_i, where n_i represents the number of data points in cluster C_i.

(iv) Repeat Step (ii) and Step (iii) until the locations of all cluster centres no longer change or the maximum number of iterations is reached.

C. VALIDITY INDEXES

Clustering validity indexes can help us to judge the quality of clustering algorithms. In general, these can be divided into internal validity indexes and external validity indexes. Internal validity indexes are mainly based on information of the data itself, while external validity indexes refer to the structure of known data sets, such as labels. Internal validity indexes are often used to identify the optimal k value for the same clustering method, and of these, the silhouette coefficient and DB index have shown high performance in terms of evaluation. In this paper, we use these two indexes to determine the k value.

1) SILHOUETTE COEFFICIENT

The silhouette coefficient reflects the closeness of each data point to its cluster and the separation between different clusters. In other words, the silhouette coefficient calculates the distance between each data point and other data points in the same cluster, as well as the distance between data points in different clusters, thereby evaluating the validity of the clustering results. The formula is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{13}$$

where $s(i)$ is the silhouette coefficient of data point i; $a(i)$ represents the mean distance from i to other data points within the same cluster a; and $b(i)$ represents the minimum of all mean distances between i and all data points in each cluster b (but not cluster a). The silhouette coefficient ranges between $[-1, 1]$: the closer its value to 1, the better the clustering result, while a value closer to -1 indicates that data points should be divided into other clusters.

We calculate silhouette coefficient for all data points, and obtain an average value, which is the overall silhouette coefficient of the current clustering results:

$$s_k = \frac{1}{n}\sum_{i \neq 1}^{n} s(i) \tag{14}$$

By modifying the k value, we can compare the silhouette indexes for different k values. In other words, we can select the best clustering number according to the closeness of data.

2) DAVIES-BOULDIN INDEX (DB INDEX)

The Davies-Bouldin index is defined as a value reflecting the intra-cluster similarity and inter-cluster separation in order to estimate the clustering results:

$$DB = \frac{1}{k}\sum_{i,j=1}^{k} \max_{i \neq j}\left\{\frac{\hat{d}_i + \hat{d}_j}{\hat{d}_{i,j}}\right\} \tag{15}$$

where $\hat{d}_i$ is the average distance from each data point in cluster i to the centre of cluster i; $\hat{d}_j$ represents the average distance from each data point in cluster j to the centre of cluster j; and $\hat{d}_{i,j}$ represents the Euclidean distance between the centres of clusters i and j. The smaller the value of the DB index, the smaller the intra-cluster distance, and thus the larger the inter-cluster distance, the better the clustering result.

V. ANALYSIS OF CLUSTERING RESULTS FOR SHANGHAI-NANJING INTERCITY HSR TRAINS

A. DATA SOURCE

The Shanghai-Nanjing Intercity Railway is 301 kilometres long and runs through 23 stations between Shanghai-Hongqiao and Nanjing South. This line connects the Yangtze river delta, which is the most densely populated, most productive and with the strongest economic growth in China, shortens the distance between Nanjing and Shanghai, optimises the passenger transport pattern, promotes the rapid development of regional economic integration, and occupies an important position in HSR network.

This paper selects operational data from Shanghai-Nanjing Intercity Railway for 31st July 2017. Since the day is not a statutory holiday, and similar to most operational days throughout the year, the clustering results will be objective.

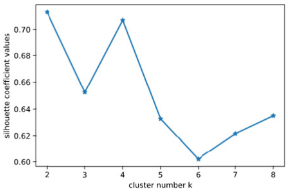


FIGURE 1. Silhouette coefficient.

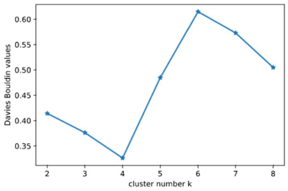


FIGURE 2. Davies-Bouldin index.

Secondly, data for 31st July 2017 is the most detailed and complete, ensuring the reliability for subsequent analysis.

The data was from the train timetable data and passenger ticket booking records: (i) A total of 245 train timetable data, including train ID, origin station, destination station, intermediate stop stations, train path, departure time at each stop station, arrival time at each stop station, travel distance, etc. (ii) A total of 14696 passenger ticket booking records, including train ID, ticket booking date, passenger boarding station, passenger alighting station, departure time at boarding station, arrival time at alighting station, travel distance, category of seat, category of ticket, ticket number, fare income, etc. In order to ensure the quality and integrity of the data, all trains in a day, a total of 111, were used as a sample to conduct quantitative clustering based on train characteristics and operational performance.

For such high dimensional data with a large amount of information and complex structure, the clustering algorithm is applicable, and the results obtained in the paper is reasonable and reliable.

B. ANALYSIS OF CLUSTERING RESULTS OF TRAIN CHARACTERISTICS FOR SHANGHAI-NANJING INTERCITY HSR TRAINS

1) CLUSTERING RESLUTS OF TRAIN CHARACTERISTICS

a: DETERMINATION OF k VALUE

The silhouette coefficient diagram and DB index diagram for Shanghai-Nanjing intercity HSR trains based on train characteristics are shown below.

Fig. 1 shows that when $k = 2$, the silhouette coefficient is the highest, followed by $k = 4$. However, as can be seen from Fig. 2, when $k = 4$, the DB index is the lowest, Hence, we use $k = 4$.

b: ANALYSIS OF k-MEANS CLUSTERING RESULTS

By following the above steps using t-SNE analysis, the k-means clustering algorithm and the optimum value $k = 4$, we obtain the clustering result of Shanghai-Nanjing intercity HSR trains based on train characteristics, as shown in Fig. 3.

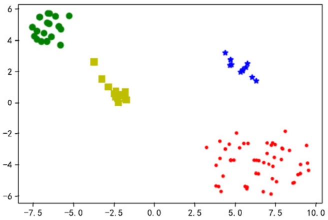


FIGURE 3. Visualisation of clustering results based on train characteristics.

Fig. 3 shows that the trains are divided into four clusters. The first cluster contains 13 trains in blue denoted as C1, the second 29 in green denoted as C2, the third 55 in red denoted as C3, and the fourth 14 in yellow denoted as C4. The specific clustering results are given in Table 2.

2) ANALYSIS OF CLUSTERING RESULTS BASED ON TRAIN CHARACTERISTICS

In this section, we analyse the clustering results of Shanghai-Nanjing intercity HSR trains based on train characteristics. The values and distributions of clustering indexes (including stop schedule score, passing-stopping ratio, travel distance, travel time, travel speed, and train capacity) for each cluster are shown in Table 3 and Fig. 4, respectively.

From Table 3 and Fig. 4, cluster C1 contains trains with a short travel distance, low capacity and low travel speed. Cluster C2 contains trains with high travel speed and high capacity, which only stop at large cities and depart on the hour to offer direct access between those cities. As can be seen from Table 3, the average travel speed of the trains in cluster C2 is 180.6 km/h, which is significantly larger than for the other clusters. In addition, the variable coefficient of the clustering indexes in C2 is smaller than for the other clusters, i.e. the trains in C2 are homogeneous benchmark trains, and the distribution is concentrated. Clusters C3 and C4 contain trains with staggered stops at different grades of stations, and both have relatively high travel speeds; the main difference between these clusters is train capacity. C3 contains low-capacity trains and has the highest proportion of trains, i.e. trains with staggered stops, low capacity and relatively high travel speed make up the primary type of Shanghai-Nanjing intercity HSR trains (49.5% of all trains). C4 contains large-capacity trains, and the other clustering indexes have similar values to those of C3. The specific characteristics of trains in each cluster are further analysed below.

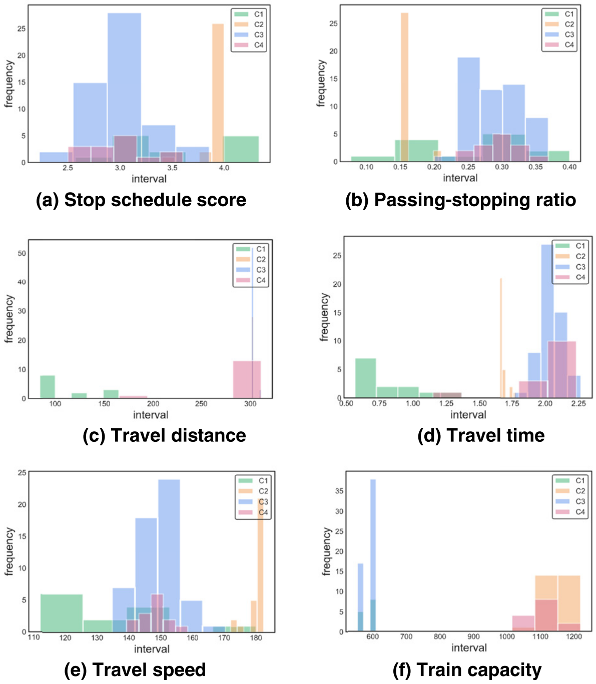
1) Cluster C1, all of them operate between Changzhou, Wuxi, Suzhou, and Shanghai. The travel distance is between 84 and 165 km (i.e. a short travel distance). The capacities of these trains are below 610 (low capacity), and the trains stop at low-grade stations, i.e. trains in C1 mainly serve passengers travelling between low-grade stations. The average travel speed is 134.4 km/h, the slowest of the four clusters.
2) Cluster C2, trains in C2 are benchmark trains operated by the railway company. These trains operate

TABLE 2. Clustering results based on train characteristics.

Cluster	Train number										Total
C1	G7033	G7128	G7201	G7202	G7205	G7206	G7207	G7208	G7209	G7210	13
	G7212	G7213	G7214								
C2	G7001	G7002	G7003	G7004	G7005	G7006	G7007	G7008	G7009	G7010	29
	G7011	G7012	G7013	G7014	G7015	G7016	G7017	G7018	G7019	G7020	
	G7021	G7022	G7023	G7024	G7025	G7026	G7027	G7028	G7096		
C3	G7030	G7031	G7032	G7034	G7035	G7037	G7038	G7039	G7040	G7044	55
	G7045	G7046	G7047	G7048	G7049	G7050	G7051	G7052	G7054	G7055	
	G7056	G7057	G7058	G7059	G7061	G7062	G7063	G7065	G7067	G7068	
	G7069	G7070	G7099	G7100	G7101	G7102	G7103	G7104	G7105	G7106	
	G7108	G7109	G7110	G7112	G7113	G7115	G7116	G7117	G7118	G7120	
	G7121	G7123	G7124	G7125	G7127						
C4	G7029	G7036	G7042	G7043	G7053	G7064	G7066	G7107	G7111	G7114	14
	G7119	G7122	G7126	G7211							

TABLE 3. Value statistics of clustering indexes for each cluster based on train characteristics.

	C1			C2			C3			C4		
	Range	Mean	Variable coefficient	Range	Mean	Variable coefficient	Range	Mean	Variable coefficient	Range	Mean	Variable coefficient
Stop schedule	2.57–4.33	3.50	0.1428	3.40–4	3.97	0.0291	2.22–3.86	3.02	0.0968	2.50–3.60	2.99	0.1034
Passing-stopping ratio	0.08–0.40	0.24	0.4099	0.15–0.21	0.16	0.0836	0.20–0.37	0.30	0.1306	0.23–0.37	0.29	0.1126
Travel distance	84–165	109.20	0.3110	301–311	301.30	0.0061	301–311	301.50	0.0075	165–311	294.90	0.1232
Travel time	0.57–1.37	0.81	0.2792	1.65–1.78	1.67	0.0203	1.77–2.27	2.03	0.0460	1.15–2.23	2.00	0.1246
Travel speed	112–180	134.40	0.1335	172–182.42	180.60	0.0166	134.80–170.40	149.00	0.0458	139.30–158.40	147.70	0.0316
Train capacity	554–610	588.90	0.0453	1015–1220	1158.90	0.0530	554–610	593.20	0.0424	1015–1220	1099.70	0.0590
Proportion		11.7%			26.1%			49.5%			12.6%	



(a) Stop schedule score

(b) Passing-stopping ratio

(c) Travel distance

(d) Travel time

(e) Travel speed

(f) Train capacity

FIGURE 4. Value distribution map for each index item based on train characteristics.

between Nanjing and Shanghai, and only stop at three high-grade stations (Suzhou, Wuxi, and Changzhou), and their average travel speed is 180.6 km/h, clearly higher than the other clusters. Trains in C2 depart from the origin station every hour on the hour, from 7:00 to 21:00. They are large-capacity trains that serve commuter passengers between big cities.

3) Cluster C3, it is almost half the number of HSR trains operating on the Shanghai-Nanjing intercity line. As in C2, the trains in C3 also operate between Shanghai and Nanjing, but the origin and destination stations have various combinations, i.e. they run between Shanghai, Shanghai-Hongqiao, Nanjing, and Nanjing South. Trains in C3 have staggered stops at different grades of stations, including high-grade stations such as Suzhou, Wuxi and Changzhou, and low-grade stations such as Kunshan South and Qishuyan. The average travel speed is therefore lower than in C2 but higher than in C1. Cluster C3 contains the primary type of train running on the Shanghai-Nanjing intercity HSR line. The low capacity and high operational frequency meet the strong demand for punctuality from intercity passengers.

4) Cluster C4, it contains trains with staggered stops but with large capacity; the other characteristics of these trains are similar to those in C3. They operate in rush hours to meet passenger demand, and are mutually complementary with the trains in C3.

C. ANALYSIS OF CLUSTERING RESULTS OF OPERATIONAL PERFORMANCE FOR SHANGHAI-NANJING INTERCITY HSR TRAINS

1) CLUSTERING RESLUTS OF OPERATIONAL PERFORMANCE

a: DETERMINATION OF *k* VALUE

The silhouette coefficient diagram and DB index diagram for Shanghai-Nanjing intercity HSR trains based on operational performance are shown below.

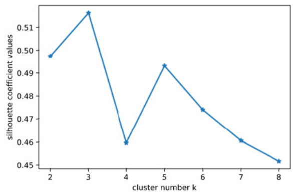


FIGURE 5. Silhouette coefficient.

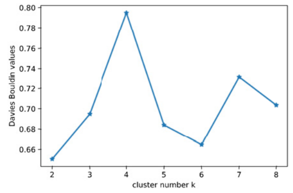


FIGURE 6. Davies-Bouldin index.

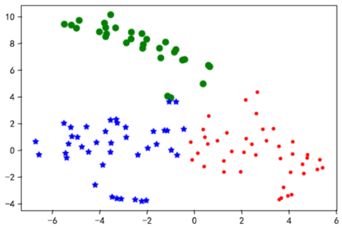


FIGURE 7. Visualisation of clustering results based on operational performance.

Fig. 5 shows that the clustering result is best for $k = 3$, followed by $k = 5$, and Fig. 6 shows that $k = 6$ is best, followed by $k = 3$ or 5. Hence, we choose $k = 3$.

b: ANALYSIS OF k-MEANS CLUSTERING RESULTS

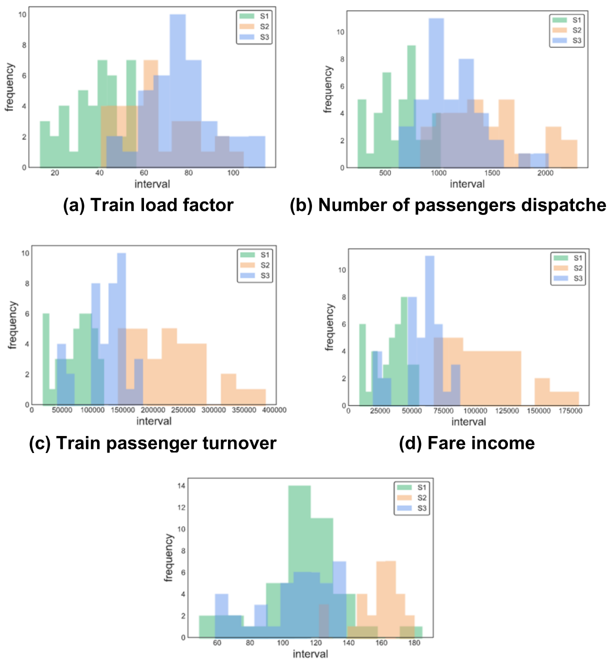
Based on t-SNE analysis, the k-means clustering algorithm, and the optimum value $k = 3$, we obtain the clustering results of Shanghai-Nanjing intercity HSR trains based on operational performance.

Trains are divided based on operational performance into three clusters. The first cluster contains 41 trains in red denoted as S1, the second 30 in green denoted as S2, and the third 40 in blue denoted as S3. Detailed results are given in Table 4 below.

2) ANALYSIS OF CLUSTERING RESULTS BASED ON OPERATIONAL PERFORMANCE

In this section, we cluster the Shanghai-Nanjing intercity HSR trains based on operational performance (where the clustering indexes include train load factor, number of passengers dispatched, train passenger turnover, fare income, and passenger average haul distance). The values and distributions of the clustering indexes are shown in Table 5 and Fig. 8, respectively.

From Table 5 and Fig. 8, we can see that the train load factor, number of passengers dispatched, and passenger turnover in cluster S1 are obviously lower than for the other clusters.



(a) Train load factor

(b) Number of passengers dispatched

(c) Train passenger turnover

(d) Fare income

(e) Passenger average haul distance

FIGURE 8. Value distribution map for each index item based on operational performance.

Trains in clusters S2 and S3 both have excellent operational performance, and some of these trains are at capacity over the whole journey. Trains in S2 mainly serve long-distance passengers, while those in S3 serve short-distance passengers. Although the average number of passengers dispatched for S2 and S3 are close, but the average value of the passenger turnover is larger for S2 than for S3. A more detailed analysis is given below.

1) Cluster S1, their train load factor, number of passengers dispatched and passenger turnover are lower than in the other clusters, and the average train load factor is only 38.43%. A low dispatched passenger number is the most important reason for the low load factor (although the passenger average haul distance is 111.9 km, which is not the lowest of the clusters).
2) Cluster S2, trains in S2 show good operational performance; the average load factor is 65.5%, and more than 13.3% of these trains have load factors above 90%. These trains serve long-distance passengers, so S2 has the highest average values for number of passengers dispatched and passenger average haul distance of 1461 and 157.2 km, respectively.
3) Cluster S3, trains in S3 have the highest average load factor of up to 76.64%, but a low passenger average haul distance of only 107.38 km. This demonstrates that there is high demand for short-distance passengers on the Shanghai-Nanjing intercity HSR line, which is mainly served by trains in S3. The operating company could therefore consider adding short-distance trains or

TABLE 4. Clustering results based on operational performance.

Cluster	Train number										Total
S1	G7027	G7028	G7038	G7040	G7042	G7044	G7047	G7048	G7051	G7056	41
	G7058	G7059	G7062	G7064	G7066	G7067	G7068	G7069	G7070	G7096	
	G7099	G7108	G7113	G7115	G7116	G7118	G7119	G7120	G7121	G7122	
	G7123	G7124	G7125	G7126	G7127	G7128	G7206	G7207	G7210	G7212	
	G7214										
S2	G7001	G7002	G7003	G7004	G7005	G7006	G7007	G7008	G7009	G7010	30
	G7011	G7012	G7013	G7014	G7015	G7016	G7017	G7018	G7019	G7020	
	G7021	G7022	G7023	G7024	G7025	G7026	G7029	G7036	G7043	G7114	
S3	G7030	G7031	G7032	G7033	G7034	G7035	G7037	G7039	G7045	G7046	40
	G7049	G7050	G7052	G7053	G7054	G7055	G7057	G7061	G7063	G7065	
	G7100	G7101	G7102	G7103	G7104	G7105	G7106	G7107	G7109	G7110	
	G7111	G7112	G7117	G7201	G7202	G7205	G7208	G7209	G7211	G7213	

TABLE 5. Value statistics of clustering indexes for each cluster based on operational performance.

	S1			S2			S3		
	Range	Mean	Variable coefficient	Range	Mean	Variable coefficient	Range	Mean	Variable coefficient
Train load factor	13.31–56.28	38.43	0.3115	40.59–104.55	65.50	0.2560	43.08–114.08	76.64	0.2017
Number of passengers dispatched	244–1017	624.83	0.3338	833–2294	1461.17	0.2651	628–2030	1132.78	0.2442
Train passenger turnover	18352–118656	71285.05	0.3991	141873–384798	229607.70	0.2647	41918–183170	122516.60	0.2868
Fare income	8676–55358	33384.60	0.4005	67610–181107	108254.32	0.2637	19597–87800	57163.33	0.2890
Passenger average haul distance	48.97–184.80	111.86	0.2283	122.02–180.07	157.24	0.0912	58.69–138.40	107.38	0.2132
Proportion		36.9%			27.0%			36.0%	

trains that stop at each station, or improving transport capacity. As we can see from Table 5, S3 has significantly lower average values of passenger turnover and passenger haul distance than S2, but higher load factors.

VI. ANALYSIS OF TWO-DIMENSIONAL CLUSTERING RESULTS FOR SHANGHAI-NANJING INTERCITY HSR TRAINS

A. INTEGRAL CLUSTERING RESULTS

With all indexes for integral one-dimensional clustering, we get four clusters which is shown by four different colours in Table 6. The number of trains and specific trains in four clusters of integral one-dimensional clustering are basically the same as the clustering result according to train characteristics. Due to mutual annihilation and interference effects between the two indexes, the relationship among train characteristics, passenger flow and operational performance cannot be clearly obtained. Therefore, this paper uses a two-dimensional clustering method based on train characteristics and operational performance to analyse the further detailed situation.

B. CROSS-OVER ANALYSIS OF CLUSTERING RESULTS FOR HSR TRAINS

There are some drawbacks in the results of integral one-dimensional clustering. In this section, we use a cross-over analysis to demonstrate the coupling relationship between these two aspects and put forward some suggestions for future operations.

TABLE 6. Cross-over and integral analysis.

<table>
<tr><th>Cluster</th><th>S1</th><th colspan="2">S2</th><th colspan="2">S3</th></tr>
<tr><td>C1</td><td>6</td><td colspan="2">0</td><td colspan="2">7</td></tr>
<tr><td>C2</td><td>3</td><td colspan="2">26</td><td colspan="2">0</td></tr>
<tr><td>C3</td><td>26</td><td colspan="2">0</td><td colspan="2">29</td></tr>
<tr><td>C4</td><td>6</td><td>3</td><td>1</td><td>1</td><td>3</td></tr>
</table>

Note: The table represents the results of cross-over analysis, while the colour represents the results of integral analysis

TABLE 7. Cross-over analysis.

Cluster	S1	S2	S3
C1	6	0	7
C2	3	26	0
C3	26	0	29
C4	6	4	4

We divide all of the trains into twelve clusters via a cross-over analysis, but since some of these contain no trains, the final total was nine clusters. The numbers of trains are shown in Table 7. Between them, clusters C2S2, C3S1 and C3S3 contain over 73% of the trains. Trains in C1 are clustered into C1S1 and C1S3. Most trains in C2 belong to C2S2, and show good operational performance. Cluster C4 has a relatively uniform distribution between C4S1, C4S2 and C4S3.

1) PERFORMANCE OF TRAINS WITH A SHORT TRAVEL DISTANCE

From the analysis in Section V.B, we know that C1 contains low-speed, low-capacity, and short-distance intercity trains, which is planned to serve short-distance passengers. Trains

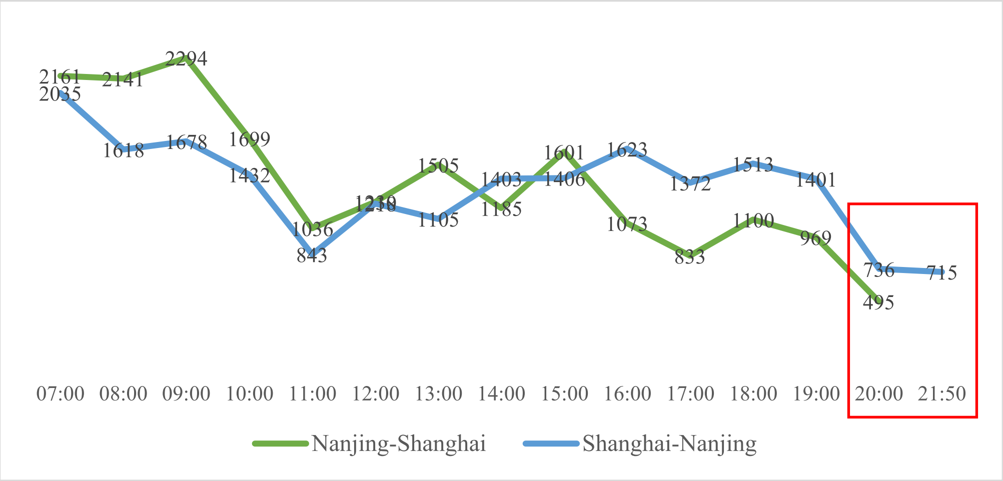


FIGURE 9. Relationship between number of passengers dispatched and departure time between Nanjing-Shanghai direction and Shanghai-Nanjing direction for cluster C2S2.

in C1S3 generally depart at morning peak times and there are more stops on the way, which matches the characteristics of intercity commuters, so their operational performance is better. Most trains in C1S1 leave after 21:00. These trains show relatively poor performance, since although their travel speeds are relatively fast, the passenger load factors are less than 43% and the number of passengers dispatched by these trains are less than 500. In the interests of both short-distance passengers and railway companies, it is desirable for railway companies to reduce the number of trains in C1S1 and keep the trains with more stops in C1S3.

2) PERFORMANCE OF TRAINS DEPARTING ON THE HOUR

Trains in Cluster C2 which depart on the hour on the Shanghai-Nanjing intercity line, are divided into two clusters. Trains in C2S2 show good operational performance, while trains in C2S1 do not. We will discuss these below.

C2S2 contains 26 trains, all of which operate between Nanjing and Shanghai, and only stop at three high-grade stations (Suzhou, Wuxi, and Changzhou) with good operational performance. More than 34.6% of the trains have load factors of over 75%. Train G7021 has minimum number of passengers dispatched of 833, which exceeds 65.9% of all of the trains. There are two reason: (i) The stations where the trains stop are large cities along the Shanghai-Nanjing intercity HSR route. there is high passenger demand for travel between major cities. (ii) Passengers prefer to choose these trains because of their few stops and high speed.

Fig. 9 illustrates the numbers of passengers dispatched in different directions for various train departure times, and we can identify commuter patterns based on these numbers. Passenger demand in both directions shows a declining trend from morning to evening, and passenger demand at morning peak times is higher than at evening peak times. However, passenger demand in the Shanghai-Nanjing direction shows higher fluctuations, while demand in the Nanjing-Shanghai direction is relatively flat. Moreover, the passenger flow from Nanjing to Shanghai is higher than that from Shanghai to Nanjing before 13:00, but it is opposite after 15:00. It is closely related to the characteristics of daily commute, residents take HSR trains from residence to workplace in the morning and return home from workplace in the evening.

Although the operational performance of hourly trains is good, it is also affected by passenger flow at different times of a day. In Fig. 9, three trains leaving after 20:00 belong to C2S1. Because their departure times are late, they show poor operational performance. The load factors of these trains are all below 35%, and the numbers of passengers dispatched are 495, 736 and 715, respectively. Even so, it is necessary to continue running these benchmark trains to provide a convenient service for passengers. The train capacity could be reduced to improve efficiency.

3) PERFORMANCE OF TRAINS WITH STAGGERED STOPS

Clusters C3 and C4 contain trains with staggered stops at different grades of stations, and both have relatively high travel speeds. The main difference between these clusters is train capacity: C3 contains low-capacity trains, while C4 contains large-capacity trains.

Cluster C3 has the largest number of trains in all clusters, with the characteristics of ‘‘low capacity, high density’’. It helps to operate intercity trains of transit type, and meet passenger demand for different departure times. The operational performance of C3 is related to passenger flow. C3 is clustered into C3S1 and C3S3, in roughly equal proportions. Trains in C3S3 show good operational performance. The average number of passengers dispatched for C3S3 is 1132. The average passenger load factor is 74.13%. In contrast, the average number of passengers dispatched in C3S1 is 659, only half that of C3S3. Its average load factor is also very low at only 42.7%. They have basically same train characteristics,

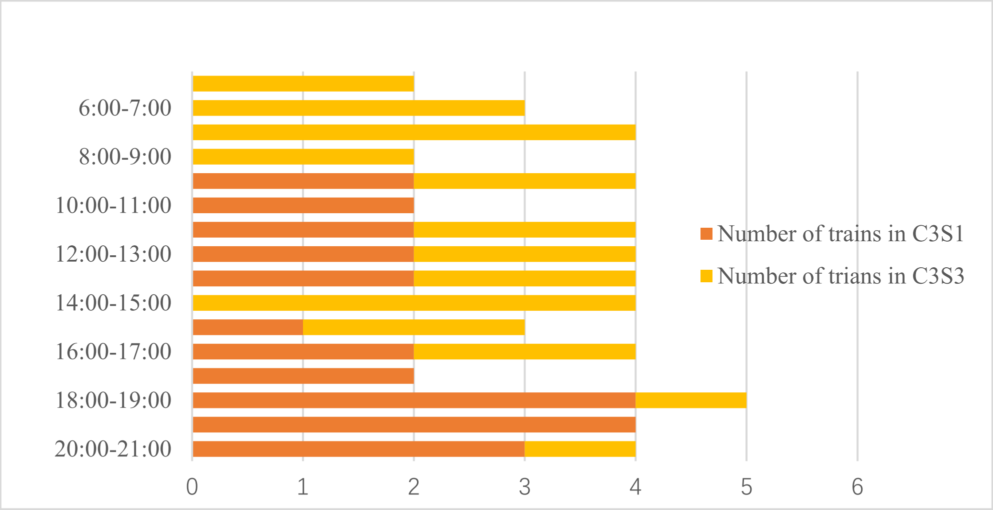


FIGURE 10. Relationship between the numbers of trains and departure times in C3S1 and C3S3.

but their operational performance is quite different. We will further analyse the reasons.

The relationship between the numbers of trains and departure times in C3S1 and C3S3 is shown in Fig. 10. It is obvious that C3S3 contains more trains with good operational performance, which operate at peak period, before 9:00 and between 14:00 and 16:00. Due to the decline in passenger flow, trains perform poorly and are clustered to C3S1. Hence, the operational performance of trains is closely related to the characteristics of passenger flow. In addition, travel demand in the Nanjing-Shanghai direction is significantly higher than that in the Shanghai-Nanjing direction. Trains operating in the Nanjing-Shanghai direction have better performance.

Considering the influence of peak and off-peak time and running direction on operational performance, while continuing to meet the passenger requirements, we suggest to adjust the number and operating time of trains in C3.

The distribution of cluster C4 is similar to those of C3, and trains in C4 shunt much short-haul traffic. Trains operating in Nanjing-Shanghai direction before 19:00 perform better. Therefore, we suggest to reduce the train capacity, flexibly increase the number of trains operating in the Nanjing-Shanghai direction at morning peak times, which is in line with ‘‘low capacity, high density’’ strategy of intercity HSR trains.

VII. CONCLUSION

In this paper, we adopt a clustering method and cross-over analysis based on train timetable data and ticket booking data from the Chinese HSR system to study HSR trains, and put forward appropriate practical suggestions. Taking Shanghai-Nanjing intercity trains as an example, we construct a two-dimensional clustering method based on train characteristics and operational performance. We then extract the key characteristic variables using t-SNE, and obtain two results through k-means clustering method. Finally, we apply a cross-over analysis to the results for train characteristics and operational performance, and obtain the coupling relationship in terms of two factors. The main conclusions are as follows.

Based on the train characteristics, Shanghai-Nanjing intercity trains can be clustered into four types: trains with a short travel distance, trains only stopping at big cities and departing on the hour, trains with staggered stops and low-capacity and trains with staggered stops and high-capacity. Based on the operational performance, trains can be clustered into three clusters: trains with poor operational performance, trains serving long-distance passengers with good operational performance and trains serving short-distance passengers with good operational performance. The number of these clusters are similar. The clustering results for the two indexes mentioned above are subjected to a cross-over analysis, and of the generated sub-clusters, three clusters contain a large number of trains with obvious features, accounting for 73% of all trains.

The hourly trains operate between Nanjing and Shanghai, and only stop at high-grade stations. These trains travel fairly fast, and their capacity is large; they mainly attract commuting flow between large cities. This kind of trains has a high load factor and high profits, indicating that there is strong passenger demand for traveling between large cities along the Shanghai-Nanjing route. These trains are recommended for use as benchmark trains, and should continue to operate at appropriately increasing density.

There is a certain demand for short-distance passenger journeys on the Shanghai-Nanjing intercity line. It shows that short-distance trains with low capacity and high density can better serve intercity passengers who are sensitive to travel times. For the convenience of operational management and turnover, short-distance trains can also be replaced by long-distance trains with similar stop schedules. These two types of trains need to maintain a suitable operating frequency in order to meet passenger demand for different departure times.

After 19:00 on the Shanghai-Nanjing intercity line, the train passenger load factor decreases significantly. The benchmark trains should continue to operate to serve commuting passengers, but capacity should be reduced to avoid unnecessary waste. The capacity or operating frequency of other trains can also be reduced.

Due to the complexity of Chinese HSR network, different lines and regions have different passenger demand, and train types and the clustering results may be different. The passenger demand in HSR transport market is stable for a period of time, so the method proposed in our paper is universal.

In future research, we intend to combine train operational data from several years, obtain the relevant trends in train clustering results and compare the differences between them. This approach can provide reliable suggestions for establishing a structured system of HSR trains by railway company.

REFERENCES

[1] Ministry of Transport of the People's Republic of China. (2019). *Statistical Bulletin of the Development of the Transportation Industry in 2018*. [Online]. Available: http://xxgk.mot.gov.cn/jigou/zhghs/201904/t20190412_3186720.html

[2] M. T. Claessens, N. M. van Dijk, and P. J. Zwaneveld, "Cost optimal allocation of rail passenger lines," *Eur. J. Oper. Res.*, vol. 110, no. 3, pp. 474–489, Nov. 1998.

[3] K. Ghoseiri, F. Szidarovszky, and M. J. Asgharpour, "A multi-objective train scheduling model and solution," *Transp. Res. B, Methodol.*, vol. 38, no. 10, pp. 927–952, Dec. 2004.

[4] M. R. Bussieck, P. Kreuzer, and U. T. Zimmermann, "Optimal lines for railway systems," *Eur. J. Oper. Res.*, vol. 96, no. 1, pp. 54–63, Jan. 1997.

[5] Y. Yue, S. Wang, L. Zhou, L. Tong, and M. R. Saat, "Optimizing train stopping patterns and schedules for high-speed passenger rail corridors," *Transp. Res. C, Emerg. Technol.*, vol. 63, pp. 126–146, Feb. 2016.

[6] M. Kaspi and T. Raviv, "Service-oriented line planning and timetabling for passenger trains," *Transp. Sci.*, vol. 47, no. 3, pp. 295–311, Aug. 2013.

[7] X. Zhou and M. Zhong, "Single-track train timetabling with guaranteed optimality: Branch-and-bound algorithms with enhanced lower bounds," *Transp. Res. B, Methodol.*, vol. 41, no. 3, pp. 320–341, Mar. 2007.

[8] H. M. Repolho, A. P. Antunes, and R. L. Church, "Optimal location of railway stations: The lisbon-porto high-speed rail line," *Transp. Sci.*, vol. 47, no. 3, pp. 330–343, Aug. 2013.

[9] H. Fu, L. Nie, L. Meng, B. R. Sperry, and Z. He, "A hierarchical line planning approach for a large-scale high speed rail network: The China case," *Transp. Res. A, Policy Pract.*, vol. 75, pp. 61–83, May 2015.

[10] P.-F. Chou, C.-S. Lu, and Y.-H. Chang, "Effects of service quality and customer satisfaction on customer loyalty in high-speed rail services in taiwan," *Transportmetrica A, Transp. Sci.*, vol. 10, no. 10, pp. 917–945, Nov. 2014.

[11] J.-S. Chou, C. Kim, P.-Y. Tsai, C.-P. Yeh, and H. Son, "Longitudinal assessment of high-speed rail service delivery, satisfaction and operations: A study of taiwan and korea systems," *KSCE J. Civil Eng.*, vol. 21, no. 6, pp. 2413–2428, Sep. 2017.

[12] E. Nathanail, "Measuring the quality of service for passengers on the hellenic railways," *Transp. Res. A, Policy Pract.*, vol. 42, no. 1, pp. 48–66, Jan. 2008.

[13] F. Dobruszkes, "High-speed rail and air transport competition in western europe: A supply-oriented perspective," *Transp. Policy*, pp. 870–879, Jun. 2011.

[14] J. Wang, J. Jiao, C. Du, and H. Hu, "Competition of spatial service hinterlands between high-speed rail and air transport in China: Present and future trends," *J. Geographical Sci.*, vol. 25, no. 9, pp. 1137–1152, Sep. 2015.

[15] S. Stoilova and R. Nikolova, "An application of AHP method for examining the transport plan of passenger trains in Bulgarian railway network," *Transp. Problems*, vol. 13, no. 1, pp. 37–48, 2018.

[16] Z. Jiang, C.-H. Hsu, D. Zhang, and X. Zou, "Evaluating rail transit timetable using big passengers' data," *J. Comput. Syst. Sci.*, vol. 82, no. 1, pp. 144–155, Feb. 2016.

[17] X. Feng, Y. Jie, and H. Liu, "Comprehensive effect of operational factors on transport efficiency of a high-speed railway train," *Tehnicki Vjesnik Tech. Gazette*, vol. 24, no. 2, pp. 497–502, Apr. 2017.

[18] T. Teichert, E. Shehu, and I. von Wartburg, "Customer segmentation revisited: The case of the airline industry," *Transp. Res. A, Policy Pract.*, vol. 42, no. 1, pp. 227–242, Jan. 2008.

[19] M. Urban, M. Klemm, K. O. Ploetner, and M. Hornung, "Airline categorisation by applying the business model canvas and clustering algorithms," *J. Air Transp. Manage.*, vol. 71, pp. 175–192, Aug. 2018.

[20] A. Punel and A. Ermagun, "Using Twitter network to detect market segments in the airline industry," *J. Air Transp. Manage.*, vol. 73, pp. 67–76, Oct. 2018.

[21] H. Lv, W. Wang, and S. Pu, "Classification of railway passengers based on cluster analysis," (in Chinese), *J. Transp. Syst. Eng. Inf. Technol.*, vol. 16, no. 1, pp. 129–134, Feb. 2016.

[22] F. Liu, Q. Peng, H. Liang, "High-speed railway passenger ticketing behavior characteristics based on PCA and clustering," (in Chinese), *J. Transp. Syst. Eng. Inf. Technol.*, vol. 17, no. 6, pp. 126–132, Dec. 2017.

[23] L. Yan-hai and S. lin-yan, "Study and applications of data mining to the structure risk analysis of customs declaration cargo," in *Proc. IEEE Int. Conf. E-Bus. Eng. (ICEBE)*, 2005, pp. 761–764.

[24] A. M. R. Cabral and F. D. S. Ramos, "Cluster analysis of the competitiveness of container ports in brazil," *Transp. Res. A, Policy Pract.*, vol. 69, pp. 423–431, Nov. 2014.

[25] F. Gianfranco, P. Claudia, S. Patrizia, and F. Paolo, "Port cooperation policies in the mediterranean basin: An experimental approach using cluster analysis," *Transp. Res. Procedia*, vol. 3, pp. 700–709, Oct. 2014.

[26] H.-A. Vogel and A. Graham, "Devising airport groupings for financial benchmarking," *J. Air Transp. Manage.*, vol. 30, pp. 32–38, Jul. 2013.

[27] Q. Cui, Y.-M. Wei, Y. Li, and W.-X. Li, "Exploring the differences in the airport competitiveness formation mechanism: Evidence from 45 chinese airports during 2010–2014," *Transportmetrica B, Transp. Dyn.*, vol. 5, no. 3, pp. 325–341, Jul. 2017.

[28] R. Mayer, "Airport classification based on cargo characteristics," *J. Transp. Geography*, vol. 54, pp. 53–65, Jun. 2016.

[29] P. F. Zhou, B. M. Han, and Q. Zhang, "High-speed railway passenger node classification method and train stops scheme," *Appl. Mech. Mater.*, vols. 505–506, pp. 632–636, Jan. 2014.

[30] C. Zhao, F. Liu, and X. Hai, "A new approach for hierarchical dividing to passenger nodes in passenger dedicated line," *J. Inf. Process. Syst.*, vol. 14, no. 3, pp. 694–708, Jan. 2018.

[31] B. Depaire, G. Wets, and K. Vanhoof, "Traffic accident segmentation by means of latent class clustering," *Accident Anal. Prevention*, vol. 40, no. 4, pp. 1257–1266, Jul. 2008.

[32] L. Wei and J. Sun, "The evaluation method of road traffic safety based on the combined principal component cluster analysis," in *Proc. Int. Conf. Bus. Manage. Electron. Inf.*, Guangdong, China, May 2011, pp. 103–107.

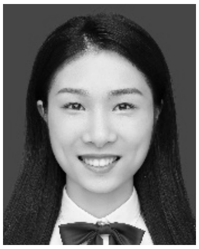
[33] G.-T. Yeo, M. Roe, and J. Dinwoodie, "Evaluating the competitiveness of container ports in korea and China," *Transp. Res. A, Policy Pract.*, vol. 42, no. 6, pp. 910–921, Jul. 2008.

[34] F. Shi, Z. Li, and S. Zhao, "Optimization method for train passing-stopping ratio under passenger travel demand agglomeration of high speed railway," (in Chinese), *China Railway Sci.*, vol. 38, no. 5, pp. 121–129, Sep. 2017.

[35] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 41, no. 4, 2003, pp. 833–840.

[36] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

LIANBO DENG received the Ph.D. degree from Central South University, Changsha, China, in 2007. He is currently a Full Professor in planning and management of traffic and transportation with Central South University. His research interests include the transportation organization of high-speed railway, optimization of train operation plan of high-speed railway, transfer choice of passenger flow, and passenger flow assignment of urban rail transit.



YUXIN CHEN received the B.E. degree in traffic and transportation from Central South University, Changsha, China, in 2017, where she is currently pursuing the M.Eng. degree in traffic and transportation engineering. Her research interests include the transfer choice of passenger flow and fare optimization for urban rail transit.

RUNFA WU received the B.E. degree in transportation engineering from Central South University, China, in 2016, where he is currently pursuing the Ph.D. degree in transportation engineering. His research interests include high-speed railway passenger assignment, urban transit timetable and rolling stock circulation optimization, and high-speed railway line planning.

QING WANG received the M.Eng. degree from Central South University, Changsha, China, in 2007, where she is currently pursuing the Ph.D. degree in traffic and transportation engineering. Her research interests include the optimization of train operation plan of high-speed railway and fare optimization for urban rail transit.

YIMING XU received the B.E. degree in traffic and transportation from the China University of Mining and Technology, Xuzhou, China, in 2018. She is currently pursuing the M.Eng. degree in traffic and transportation engineering with Central South University, Changsha, China. Her research interest includes performance evaluation for hub station.