Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 7497094, 11 pages
https://doi.org/10.1155/2021/7497094

Research Article

An Affinity Propagation-Based Clustering Method for the Temporal Dynamics Management of High-Speed Railway Passenger Demand

Wenxian Wang,[1] **Tie Shi**,[2] **Yongxiang Zhang**,[3,4] **and Qian Zhu**[5]

[1]*School of Rail Transportation, Wuyi University, Jiangmen 529020, China*
[2]*China Railway Eryuan Engineering Group Co. Ltd, Chengdu 610031, Sichuan, China*
[3]*School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, Sichuan, China*
[4]*National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu 610031, Sichuan, China*
[5]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Wenxian Wang; wwx530@163.com

Received 27 May 2019; Revised 14 April 2021; Accepted 20 April 2021; Published 28 April 2021

Academic Editor: Nirajan Shiwakoti

The number of passengers in a high-speed railway line normally varies significantly by the time periods, such as the peak and nonpeak hours. A reasonable classification of railway operation time intervals is essential for an adaptive adjustment of the train schedule. However, the passenger flow intervals are usually classified manually based on experience, which is subjective and inaccurate. Based on the time samples of actual passenger demand data for 365 days, this paper proposes an affinity propagation (AP) algorithm to automatically classify the passenger flow intervals. Specifically, the AP algorithm first merges time samples into different categories together with the passenger transmit volume of the stations, which are used as descriptive variables. Furthermore, clustering validity indexes, such as Calinski–Harabasz, Hartigan, and In-Group Proportion, are employed to examine the clustering results, and reasonable passenger flow intervals are finally obtained. A case study of the Zhengzhou-Xi'an high-speed railway indicates that our proposed AP algorithm has the best performance. Moreover, based on the passenger flow interval classification results obtained using the AP algorithm, the train operation plan fits the passenger demand better. As a result, the indexes of passenger demand satisfaction rate, average train occupancy rate, and passenger flow rate are improved by 7.6%, 16.7%, and 14.1%, respectively, in 2014. In 2015, the above three indicators are improved by 5.7%, 18.4%, and 14.4%, respectively.

1. Introduction

With the extension and integration of China's high-speed railway network, it is becoming the preferred mode of travel. One of the most important tasks for high-speed railway passenger transport management is to adjust the line plan according to the characteristics of the fluctuations of the passenger flow over a year, so that the line plan is adapted to the passenger demand. Therefore, it is obvious that the annual passenger flow interval classification serves as the foundation for the adjustment of the line plan [1, 2]. The usual method used by the railway bureau is to classify annual passenger flow intervals according to the subjective experience of the engineering and technical personnel. However, the quality of this classification will largely depend on the experience of the personnel, and unreasonable classification results are likely to be generated since the characteristics of the seasonal fluctuation of passenger demand are usually not fully taken into consideration [3].

There have not been any directly relevant studies on the high-speed railway passenger flow interval classification problem. However, it is very similar to the time-of-day (TOD) interval identification problem when developing traffic signal timing plans. In the TOD interval identification problem, a full day is treated as a cycle that contains several time points, and time points with similar traffic attributes are classified into the same group, as shown in Figure 1. By using

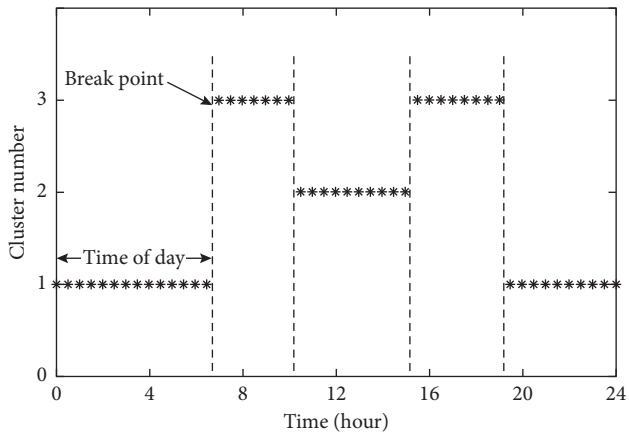

Figure 1: Interval classification and breakpoints of TOD.

the TOD interval identification, a day can be divided into several intervals and the traffic signal timing plans can be made on the basis of the dynamics of the traffic flow.

The difference between this strategy and the multiperiod control strategy of road intersections is that railway passenger transportation takes an annual cycle, divides the year into several periods with similar passenger flow according to the change of daily passenger flow, and then divides according to the average passenger flow of each period to prepare the train operation plans. In the actual operation of the railway, a manual decision-making method is adopted to solve the basic problem of passenger flow time division. The results of railway passenger flow time division in current research are relatively rare. The main contributions of this article are as follows:

(1) The paper takes a step forward in the production of an annual train operation plan. This method accounts for the dynamics and regularity of the annual passenger flow.

(2) A new affinity propagation (AP)-based classification method is proposed, which is applied to the time division of high-speed railway passenger flow.

(3) The effectiveness of the algorithm in dividing the annual passenger flow period is proven through the experiment of practical calculation examples of the Zhengzhou-Xi'an high-speed railway, and the superiority of the optimization method is demonstrated through the comparison with the actual situation.

The rest of this paper is organized as follows. In Section 3, we explain why passenger flow intervals need to be classified by analyzing the temporal dynamics of the passenger demand of a high-speed railway. Section 5 presents a comparative analysis of the adaptability evaluation indexes of the line plan based on the passenger flow interval classification results obtained by applying our proposed AP-based clustering method with the actual results obtained by the high-speed railway passenger transport management department.

2. Literature Review

Chinese railway transportation organization implements the basic principle of driving according to flow. Therefore, studying the characteristics and laws of railway passenger flow is of great significance for improving the adaptability of passenger flow of transportation organizations and the economic benefits of the railway. At present, there are many studies on laws of passenger flow in railway passenger demand management. From the perspective of research objects, the laws of flow can be roughly divided into the following four types:

(1) *Road Network Passenger Flow*. This type of research mainly focuses on the trend of high-speed rail passenger flow and is used to solve the social and economic problems of line network laying, price positioning, and high-speed rail development direction.

(2) *The Overall Passenger Flow of the Line*. This type of research mainly focuses on the spatial characteristics of the passenger flow and is used to analyze the number of train pairs and line utilization on a certain line.

(3) *Nodal Passenger Flow*. This type of research is mainly aimed at one or more nodes on a certain railway line. It is used to analyze the importance of nodes, railway stop schedule plan, node selection in train changing line, station node reconstruction, and expansion.

(4) *OD Passenger Flow*. This type of research includes OD passenger flow spatial characteristics and OD passenger flow time period characteristics. The former mainly focuses on the passenger flow conditions at the starting and ending points, the travel distance of the passenger flow, and the direction of passenger flow. The latter focuses on the preference and law of passenger flow time. The characteristics of passenger flow period can analyze the passenger flow more accurately and provide an important data basis for the preparation of a train operation plan.

The usual method to determine the TOD intervals is to draw the cumulative traffic flow count curve of a typical intersection for a day, and then those time points with significant fluctuation are chosen manually as the split points between two adjacent intervals; thus, the TOD is classified into intervals. Considering the nature of the traffic flow fluctuation, some scholars proposed theoretical methods to determine the TOD intervals using practical examples to demonstrate their methods. The proposed methods include heuristic search methods and clustering-based analysis methods [4, 5]. The heuristic search method defines the TOD interval identification problem as a mathematical optimization problem. Park et al. [6] introduced heuristic algorithms to determine the optimal TOD break points for traffic signal timing plans; this overcomes the disadvantages of clustering algorithms, which often generate infeasible TOD intervals. Abbas and Sharma [7] proposed a multiobjective evolutionary algorithm to solve

the TOD interval identification problem. However, the optimal number of break points cannot be determined due to the randomness of evolutionary algorithms. Park and Lee [8] designed a greedy search algorithm to obtain the optimal TOD break points that needs fewer evaluations than the genetic algorithm and is robust to various demand fluctuations. Lee et al. [9] introduced the transition cost between break points into the genetic algorithm, and the obtained results were better than those from the greedy search algorithm reported by Park and Lee [8].

Another method for determining the TOD intervals is the clustering-based analysis method. Hauser and Scherer [10] proposed the cluster analysis approach based on the concept of a high-resolution system state to deal with the TOD interval automatic identification problem. However, nonadjacent time points will be clustered into the same interval. Wang et al. [11] designed a nonhierarchical clustering algorithm, namely, the *K*-means method, to determine the TOD break points. However, the number of clusters needs to be specified manually in advance. Ratrout and Nedal [12] first combined the microsimulation method and the *K*-means algorithm to obtain the optimal TOD break points, and then an improved subtractive clustering-based *K*-means technique was proposed. Finally, it was proven that those two approaches can generate similar results. Liu et al. [13] considered the difference of traffic flow, averaged the traffic volume at the intersection in hours, and used the Webster method to calculate the signal duration to determine the discrimination threshold of the control period division. Shen et al. [14] proposed an improved *K*-means clustering algorithm which updated cluster center initialization and rules and applied it to the division of bus operation time. Zhao et al. [15] used the classical spectral clustering algorithm (Ng–Jordan–Weiss, NJW) to divide the time period and Synchro and SIMTraffic software to establish the optimal signal timing plan and simulation evaluation of the time period division results. Yao et al. [16] used the Ward minimum variance method to cluster the historical traffic flow data and the improved criterion of cube group as the cluster termination condition to determine the optimal number of plans and the best break points for multiplan control of traffic signals in TOD mode. Song et al. [17] introduced the dissimilarity matrix to determine the number of clusters, and the break points were determined based on nonintrusive data collection techniques. In addition, better signal timing plans can be made by that method. Moreira et al. [18] drew the annual vehicle travel time curve of the bus line, used dynamic time warping to evaluate the similarity of the travel time during different days, and clustered the date according to the travel time change curve to make the annual time period division. Li et al. [19] introduced an improved strategy of dynamic recursion on the basis of the ordered sample clustering algorithm to realize the division of road intersection signal time period. Song et al. [17] considered the coordination mechanism of passenger demand and travel time for the time division plan and proposed a bus operation time division model with the goal of minimizing the operating time cost of the bus fleet throughout the day. Bie et al. [20] gave a method for determining the number of transition periods under different timing schemes, established a calculation model for the total delay time of vehicles in adjacent intervals, and used the difference in total vehicle delay time as a threshold of the discrimination index to determine whether adjacent intervals are merged.

Previous studies related to TOD intervals have focused on road interactions, and the time-of-day is divided into N time points. Then, adjacent time points with similar traffic volumes are clustered into the same traffic interval.

However, there are great differences in the classification of operation intervals between high-speed railway passenger flow and road intersection intervals, which are reflected in the following aspects: (1) the departure times of railway passengers mainly depend on the train schedule, such that it is difficult to reflect the regularity of passenger flow throughout the day, and thus, the passenger flow intervals can only be classified with the length of study period equal to one year; (2) a sharply increase or decrease could be witnessed in the amount of passenger flow during the whole year, which is due to the existence of festivals, such as the Spring Festival and Tomb-Sweeping Day. As a result, the special festivals will be listed as independent time periods in the traditional clustering algorithm, which is inconsistent with the practical railway operation condition. Therefore, the classification approaches of road intersection intervals are not applicable to high-speed railway passenger demand classification.

By considering the characteristics of the fluctuation of the high-speed railway passenger flow, this paper proposes an AP-based clustering method to classify the high-speed railway passenger flow intervals accurately and automatically. First, time points in one year are taken as samples, and the number of passengers dispatched at each station along the high-speed railway line serves as variables that describe the samples. Second, the time points are clustered into groups to form the passenger flow intervals by using the AP-based clustering method. Finally, measures of effectiveness are employed to evaluate the clustering results so that the optimal number of clusters can be determined.

3. Temporal Dynamics of the Passenger Demand on a High-Speed Railway

The managers of a high-speed passenger railway need to change the line plan in accordance with the seasonal fluctuations in the passenger flow. However, the adjustment of a line plan is a systematic, complex, and massive task because the high-speed railway lines are too busy, and line plan is only adjusted a few times each year. In this situation, passenger flow intervals need to be classified reasonably so that the number of adjustments of the line plan can be determined accordingly, and the line plan can be always adaptable to the temporal dynamics of the passenger demand on the high-speed railway.

The set of time points in a year is denoted by $T = \{t_1, t_2, \ldots, t_n\}$, and the number of passengers dispatched at each station of the high-speed railway line at time point t is represented by $X(t) = \{x_1(t), x_2(t), \ldots, x_m(t)\}$, where $X(t)$

is the descriptive attribute. If the adjacent time points in the set T with similar attributes $X(t)$ are clustered into p disjoint nonempty subsets $\{T_1, T_2, \ldots, T_p\}$, then the time in a year is transformed into a finite number of continuous intervals; thus, the passenger flow intervals are classified. In this way, the passenger flow interval classification problem is indeed a clustering analysis problem, as shown in Figure 2.

According to the pattern of the fluctuation in the passenger flow, the following two key problems need to be dealt with while solving the high-speed railway passenger flow interval classification problem.

(1) The number of passenger flow intervals cannot be too small or too large. In the former situation, the temporal dynamics of passenger flow cannot be reflected properly. However, in the latter situation, there will be too much difficulty in implementing and adjusting the line plan.

(2) The time span of passenger flow intervals cannot be too short or too long. The former will make it difficult implement or adjust the line plan, and in the latter situation, the temporal dynamics of the line plan cannot be reflected properly.

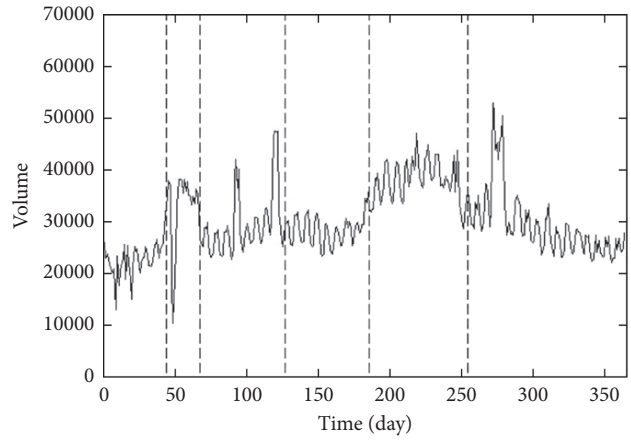

Figure 2: Description of passenger flow interval classification of high-speed railway.

4. Demand AP-Based Clustering Method

4.1. Affinity Propagation Method. AP is a state-of-the-art clustering algorithm developed by Brendan J. Frey and Delbert Dueck. It is based on the concept of "message passing" between data points. The AP algorithm has been chosen to partition the passenger flow period of this high-speed railway because it not only has a faster convergence speed when dealing with large-scale and complex data but can also avoid the clustering result being limited by the choice of the initial class representative point.

The AP algorithm performs clustering on a similarity matrix composed of sample data points. Like other clustering algorithms, its goal is to minimize the distance between each data point and its class representative point in the partition category to achieve a partition [6, 21]. The basic principle of this algorithm is introduced as follows:

(1) All the N samples in the dataset are regarded as candidate class representatives. The similarity between any two samples x_i and x_k is established, namely, by the attraction degree of each sample with other samples, and stored in the $N \times N$-dimensional similarity matrix.

(2) *Definition.* $s(i,k)$ indicates the similarity between samples x_k and x_i, that is, the quantization level of the suitability of sample x_k for being a class representative of x_i. When the algorithm is initialized, all the samples are assumed to have the same probability to be selected as a class representative, namely, assuming all $s(k,k)$ have the same value p. The similarity between two points is calculated by

$$s(i,k) = -\left\|x_i - x_k\right\|^2. \tag{1}$$

(3) Two important information parameters, the credibility matrix r and the availability matrix a, are employed by the algorithm: $r(i,k)$ is a credibility matrix that points from x_i to x_k, which indicates the degree of representativeness of sample x_k, that is, the degree to which it is suitable as a class representative of sample x_i; $a(i,k)$ is an availability matrix that points from x_i to x_k, which indicates the appropriateness of choosing x_k as the class representative of sample x_i. For any sample x_i, the sum of its credibility and availability with all other samples is calculated, and the sample x_k with the largest sum is selected as the class representative. The alternating update of the above two information quantities is the iterative process of the AP algorithm.

The update formula for the credibility matrix $r(i,k)$ is

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}. \tag{2}$$

The update formula for the availability matrix $a(i,k)$ is

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum_{i' \neq \{i,k\}} \max[0, r(i',k)]\right\}, & i \neq k, \\ \sum_{i' \neq k} \max[0, r(i',k)], & i = k. \end{cases} \tag{3}$$

4.2. Evaluation of the Validity of the Clustering. The validity of the clustering is used to quantify and evaluate the quality of the clustering results and determine the optimal partition of the dataset [22]. Clustering validity indexes are adopted to

evaluate which result generated by the clustering algorithm is optimal and the number of clusters corresponding to the optimal result is taken as the optimal clustering number. The output of the AP algorithm is a series of clustering results that contain different numbers of clusters; hence, the effectiveness of these clustering results needs to be evaluated. In this paper, the effectiveness indexes for evaluating the optimal clustering number include Calinski–Harabasz, Hartigan, and In-Group Proportion.

4.2.1. The Calinski–Harabasz Index. The Calinski–Harabasz (CH) index is a measure of the intraclass dispersion matrix and the interclass dispersion matrix for all samples, and the number of classes corresponding to the maximum value is taken as the optimal clustering number.

$$\mathrm{CH}(k) = \frac{\mathrm{tr}\, B(k)/(k-1)}{\mathrm{tr}\, W(k)/(n-k)}. \tag{4}$$

Here, k is the number of clusters, $\mathrm{tr}\, B(k)$ is the trace of the interclass dispersion matrix, and $\mathrm{tr}\, W(k)$ is the trace of the intraclass dispersion matrix.

4.2.2. The Hartigan Index. The Hartigan index [12] can be used in cases where the number of clusters is 1. The minimum number of classes with Hart ≤10 is the optimal clustering number.

$$\mathrm{Hart}(k) = \left(\frac{\mathrm{tr}\, W(k)}{\mathrm{tr}\, W(k+1)} - 1\right)(n-k-1). \tag{5}$$

4.2.3. In-Group Proportion Index. The In-Group Proportion (IGP) index [13] is used to measure whether the samples nearest to each sample in a class are in the same category. A cluster with a larger average IGP index has a better clustering quality, and the number of clusters corresponding to the maximum value is the optimal clustering number.

$$\mathrm{IGP}(u) = \frac{\#\left\{j|\mathrm{Class}(j) = \mathrm{Class}\left(j^N\right) = u\right\}}{\#\left\{j|\mathrm{Class}(j) = u\right\}}. \tag{6}$$

Here, u is the clustering standard, $\mathrm{Class}(j)$ is the standard of sample j, j^N is the sample nearest to sample j, and # denotes the cardinality of a set.

4.3. Assessment of the Adaptation to Passenger Demand. First, we establish a train operation plan and simulate a passenger flow distribution according to the average demand of passenger flow in each period after the passenger flow period was partitioned. The adaptability between the passenger demand and the train operation plan during each period is quantitatively evaluated and summarized using three indexes, namely, the satisfaction rate of passenger demand, the average attendance rate of the train, and the direct rate of passenger flow. The indexes are calculated as follows.

4.3.1. Passenger Demand Satisfaction Rate. Passenger demand satisfaction rate is mainly reflected in the passenger transport capacity and degree of satisfaction of the passenger demand provided by the train operation plan between each pair of passenger flow ODs of the high-speed railway and the related railway network. It can be expressed as the ratio of the passenger traffic volume to the total passenger demand that transport service is effectively obtained, which is constrained by the capacity of the available resources, especially train staff, under the conditions of the established train operation plan. The formula is as follows:

$$C_1 = \frac{\sum_w q'_w}{\sum_w q_w} \times 100\%, \tag{7}$$

where q_w indicates the total passenger flow between passenger flow OD pair w and q'_w indicates the total passenger volume transported by high-speed railway between passenger flow OD pair w.

4.3.2. Seat Occupancy Rate. Seat occupancy rate refers to the weighted ratio of the amount of passenger flow a train carries in its operating section to the total number of seats provided by the train, which is used to reflect the selection results of passengers in different passenger flow OD pairs for various types of high-speed trains. The average attendance rate of a train means the average rate of all train attendance rates in the evaluation range. The formula is as follows:

$$C_2 = \frac{1}{|h|} \cdot \sum_h \sum_{(i,j)} \frac{q_{ij}^h}{A_h \cdot |E_h|} \times 100\%, \tag{8}$$

where q_{ij}^h is the passenger flow carried by train h in the interval (i, j), A_h is the number of staff on train h, and E_h is the number of segments operated by train h.

4.3.3. The Direct Rate of Passenger Flow. The demand structure of passenger flow is made up of different demand directions, and each demand direction has a direct or transfer plan to reach the destination. The direct rate of passenger flow is expressed by the ratio of passenger flow directly to their destination without transfer to the total passenger flow in this direction between each point pair of passenger demand under the conditions of the established train operation plan and the structure of the passenger demand. It is calculated by

$$C_3 = \frac{\sum_w q_w^d}{\sum_w q_w^d + \sum_w \sum_e q_w^e}, \tag{9}$$

where q_w^d is the number of passengers who can reach their destination directly without the need of a transfer between passenger flow OD pair w and q_w^e is the number of passengers who have reached the destination by e transfers between passenger flow OD pair w.

4.4. Partitioning the High-Speed Railway Passenger Flow Periods. The partitioning of the high-speed railway passenger flow needs to reflect the variation of passenger demand with the seasons and holidays. For each different

direction of high-speed railway passenger flow, a railway station is the basis of the departure and another is the basis for the arrival for the passenger flow to realize a displacement. Therefore, representative large passenger stations on the high-speed railway lines have been the subject of research. According to the statistics on passenger volume during the year, time points with similar passenger flow statistics are grouped into the same category, and adjacent time points in the same category are seen as the same passenger flow period. Based on the basic principles of the above algorithm, the basic steps of partitioning the high-speed railway passenger flow into periods are as follows:

Step 1. Collect the information about the high-speed railway lines, stations, and volume of passengers sent at each time point from the stations along the railway lines. Set the maximum number of iterations of the algorithm to be $N_{\max}$.

To eliminate the differences of scales between variables, standardization was performed on each variable as follows:

$$Z - \text{score} = \frac{x - \overline{x}}{\sigma}, \tag{10}$$

where $Z-$ score is the standardized value, x is the daily volume of passengers sent between stations, $\overline{x}$ is the average number of passengers sent between all stations during the year, and σ is the standard deviation of the total number of passengers sent between all stations during the year.

The sample data vector in the cluster is

$$X(t) = \begin{bmatrix} x_{1,2}(t) & x_{1,3}(t) & \cdots & x_{1,n}(t) \\ & x_{2,3}(t) & \cdots & x_{2,n}(t) \\ & & \cdots & \cdots \\ & & & x_{n-1,n}(t) \end{bmatrix}, \tag{11}$$

where $X(t)$ is the passenger flow state vector of high-speed railway at time point t (day t) and $x_{i,j}(t)$ is the number of passengers sent from the i-th node to the j-th node of the high-speed railway at the time point t.

Step 2. Initialization: First, set the initial values of the credibility matrix $r(i,k)$ and the availability matrix $a(i,k)$ to 0.

Then, calculate the sample similarity matrix $s(i,k)$, using the Euclidean distance as the measure by formula (12). Set the diagonal element $s(k,k)$ to have the same median attractiveness value.

$$p = \frac{\sum_{i \neq j} s(i,k)}{N \cdot (N-1)}. \tag{12}$$

Here, N is the number of samples.

Step 3. Iteration:

(1)Update the usability and credibility using formulas (2) and (3).

(2) Set damping factor to eliminate the digital oscillations in the iteration.

$$\begin{cases} r_{\text{new}}(i,k) = \lambda \cdot r_{\text{old}}(i,k) + (1-\lambda) \cdot r(i,k), \\ a_{\text{new}}(i,k) = \lambda \cdot a_{\text{old}}(i,k) + (1-\lambda) \cdot a(i,k). \end{cases} \tag{13}$$

Here, $r_{\text{new}}(i,k)$ and $r_{\text{old}}(i,k)$ are the credibility matrices obtained from the previous update and this update, $a_{\text{new}}(i,k)$ and $a_{\text{old}}(i,k)$ are the availability matrices obtained from the previous update and this update, and $\lambda \in (0,1)$ is the damping factor, whose value is set to 0.9.

(3) Calculate the sum of the credibility and the availability of all the samples, and find the class center sample of each sample according to $\text{argmax}_k\{r(i,k) + a(i,k)\}$.

(4) Update the current iteration number.

Step 4. Output the result: Judge whether the iterative process has reached the maximum number of iterations that have been set, that is, $n \leq N_{\max}$. If it is reached, then terminate the algorithm and output the partition results of all the categories of the time points; otherwise, return to step 3.

Step 5. Determination of the number of categories of time points: Calculate the three indexes of the partition results, namely, Calinski–Harabasz, Hartigan, and In-Group Proportion. Choose the optimal number of categories of time points and the corresponding category partition results.

Step 6. Determination of the passenger flow periods of the high-speed railway: Traverse all the partition categories of the cycle and compare each sample with each other. If the time points corresponding to the samples are adjacent (the beginning of the year and the end of the year are also considered as adjacent time periods), the samples are merged into one passenger flow period; otherwise, it is regarded as another passenger flow period.

5. Numerical Experiments

To test the effectiveness of our proposed AP-based clustering method, the Zhengzhou-Xi'an high-speed railway was taken as an example. It contains 10 high-speed railway stations and is approximately 505 kilometers long. There are 9 stations along the Zhengzhou-Xi'an high-speed railway, which are Xi'an North (No. 1), Weinan North (No. 2), Huashan North (No. 3), Lingbao West (No. 4), Sanmenxia South (No. 5), Shengchi South (No. 6), Luoyang Longmen (No. 7), Gongyi South (No. 8), and Zhengzhou East (No. 9). Among them,

the stations with the capacity of starting and ending trains are Xi'an North, Huashan North, Luoyang Longmen, and Zhengzhou East, as shown in Figure 3. In addition, Beijing West (No. 10), Shanghai (No. 11), Wuhan (No. 12), Guangzhou South (No. 13), Shenzhen North (No. 14), and Nanchang West (No. 15) are the starting and ending points of cross-line trains. The number of passengers dispatched at the Zhengzhou East railway station, Luoyang Longmen railway station, Sanmenxia South railway station, and Xi'an north railway station from the year of 2014 to 2015 has been collected as sample data. The collected data are further processed according to equations (10) and (11) so that the passenger flow intervals of the Zhengzhou-Xi'an high-speed railway can be classified.

5.1. Passenger Flow Interval Classification Results. The AP algorithm is first used to cluster the samples, and three validity assessment indexes, namely, Calinski–Harabasz, Hartigan, and In-Group Proportion, are used to evaluate the clustering results, as shown in Figure 4.

Figure 5 demonstrates the optimal number of clusters for the Zhengzhou-Xi'an high-speed railway line, which is based on the data collected from 2014 to 2015. Among them, the horizontal axis represents the number of days, that is, the sample in the annual passenger flow period division problem, which has a total of 365 days. The vertical axis represents the categories formed by the passenger flow period division, which contains a total of 5 categories. As shown in Figure 5(a), the ordinate value corresponding to the abscissa 1–24 takes the value 1; that is, the first day to the 24th day of 2014 are classified as category 1; the ordinate value corresponding to the abscissa 25–29 takes the value 2; that is, the 25th to 29th days of 2014 are classified into category 2, and the rest can be deduced by analogy.

By iterating over all the 5 clusters and comparing every two samples in the same cluster, the discontinuous intervals in the same cluster are split. The resulting classification of passenger flow intervals for the Zhengzhou-Xi'an high-speed railway line from the year 2014 to 2015 is presented in Table 1.

Table 1 shows that the 365 days of the year 2014 or 2015 are divided into 13 intervals for the Zhengzhou-Xi'an high-speed railway line, and the lengths of the time spans for interval 3, interval 6, interval 7, interval 8, and interval 12 are exactly the same in 2014 and 2015. However, the lengths of the time spans for other intervals are different because of the Chinese Spring Festival. The Spring Festival began on January 31st for the year of 2014, which is also the 31st day in the year, but in 2015, it started on February 19th, which is the 50th day in the year. In addition, the beginning of the time span for interval 2 is always 7 days before the Spring Festival, and the seasonal characteristics of the other intervals are obvious. The seasonal characteristics of all 13 intervals are summarized below.

The time span of interval 1 corresponding to the steady period of passenger flow stands between the Chinese New Year and the Chinese Spring Festival. The peak period of passenger flow during the Chinese Spring Festival is covered by the time span of interval 2, interval 3, and interval 4. The time span of interval 5 serves as the steady period of passenger flow between the Spring Festival and the Ching Ming Festival. The peak period of passenger flow during the Ching Ming Festival is covered by the time span of interval 6. Interval 7 has its time span corresponding to the steady period of passenger between the Ching Ming Festival and Labor Day. The time span of interval 8 covers the peak period of passenger flow during Labor Day. The steady period of passenger flow between Labor Day and the summer vacation is the time span of interval 9. The time span of interval 10 serves as the peak period of passenger flow during the summer vacation. The steady period of passenger flow between the summer vacation and the National Day is covered by the interval 11. The time span of interval 12 covers the peak period of passenger flow during the National Day. The time span of interval 13 serves as the steady period of passenger flow standing between the National Day and the Chinese New Year.

5.2. Correction of the Results of the Classification of the Passenger Flow Intervals. The time spans of interval 3, interval 6, and interval 8 are only one day long each, and it is inconvenient and difficult for the high-speed railway passenger transport management department to adjust the line plan. Therefore, the three intervals whose lengths are less than 7 days are merged with their adjacent intervals according to field work experience. The corrected classification of the passenger flow intervals for the Zhengzhou-Xi'an high-speed railway is listed in Table 2.

The classification of the passenger flow intervals in Table 2 can be summarized as follows. The time span of interval 1 corresponds to the steady period of passenger flow that stands between the Chinese New Year and the Spring Festival. The peak period of passenger flow during the Spring Festival is the time span of interval 2. The steady period of passenger flow between the Spring Festival and summer vacation is covered by the time span of interval 3. The time span of interval 4 covers the peak period of passenger flow during the summer vacation. The time span of interval 5 corresponds to the steady period of passenger flow between the summer vacation and the National Day. The time span of interval 6 serves as the peak period of passenger flow during the National Day. The steady period of passenger flow between Labor Day and the Chinese New Year is the time span of interval 7.

The results of classifying the passenger flow intervals in Table 2 can be used as the foundation for the evaluation and adjustment of the line plan, and the adaptability of the line plan can be evaluated according to the predicted passenger demand in each interval. In addition, the line plan needs to be adjusted if the evaluation results are not ideal.

5.3. Comparative Analysis of the Adaptability of the Line Plan to Passenger Demand. To illustrate how the line plan obtained using our proposed passenger flow interval classification is more adaptable, the adaptability evaluation indexes for the Zhengzhou-Xi'an high-speed railway from the year of

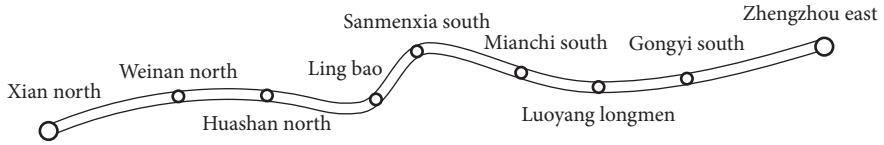

Figure 3: Zhengzhou-Xi'an high-speed railway.

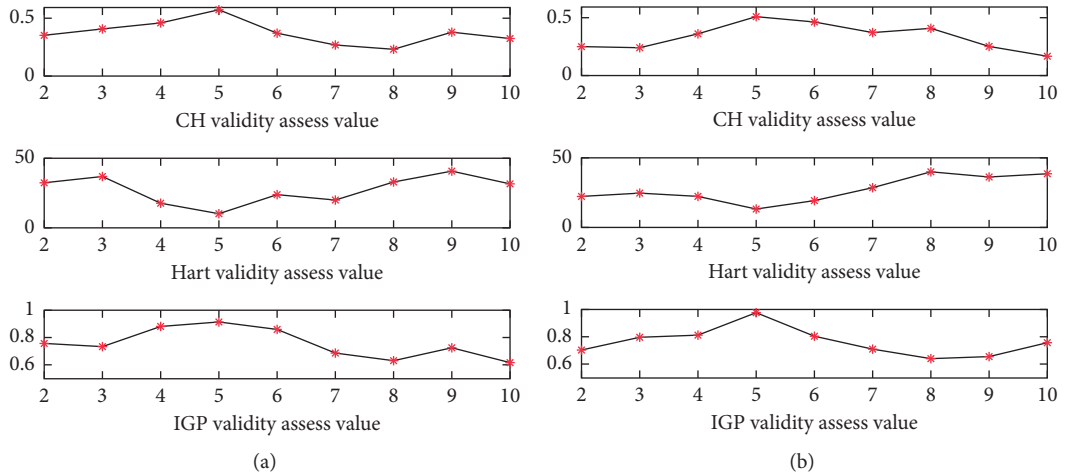

Figure 4: Three validity assessment indexes illustration: (a) validity indexes in 2014; (b) validity indexes in 2015.

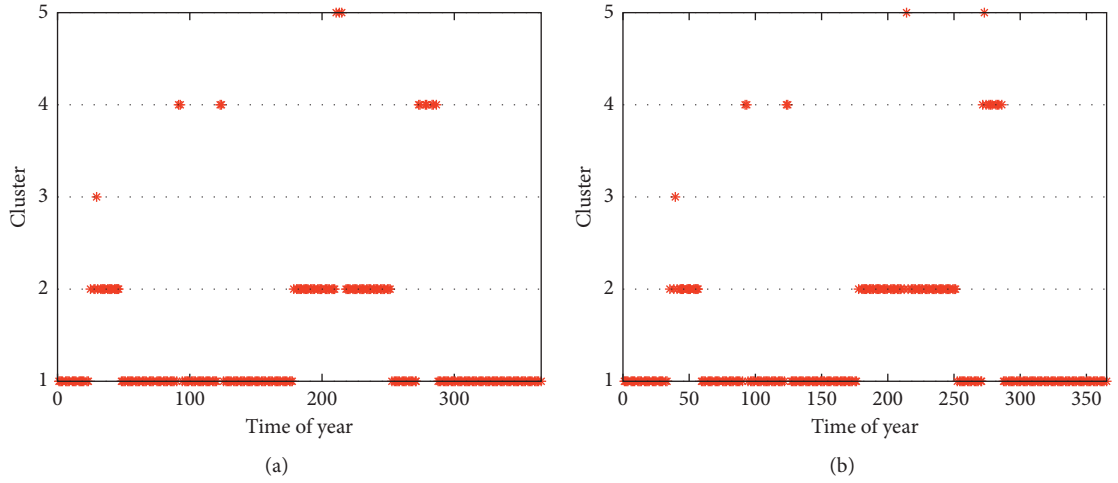

Figure 5: Passenger operation time interval classification results: (a) classification results in 2014; (b) classification results in 2015.

2014 to 2015 on the coordination between the line plan and passenger demand are calculated accordingly. In addition, the adaptability evaluation indexes of the line plan are also calculated using the passenger flow interval classification results actually used by the high-speed railway passenger transport management department. According to the passenger flow OD data of the Zhengzhou-Xi'an high-speed railway in 2014 and 2015 obtained by the Training Center for R&D of National Train Operation Diagram, first the average passenger flow of each period is calculated on the basis of the division of passenger flow periods in the previous paragraph, and the passenger flow is allocated to each train in simulation according to the actual train operation plan used in that period. Then, the passenger demand matching rate (including passenger demand satisfaction rate, the average train occupancy rate, and the direct passenger flow rate) of each time period is calculated based on formulas (7) to (9), as shown in Figure 6. On this basis, the matching rate of each passenger flow period is weighted and summarized to get the full-year passenger demand overall matching rate. The calculated matching rate is compared with the actual statistics of the Zhengzhou-Xi'an high-speed railway in 2014 and 2015, which is shown in Table 3.

Table 3 demonstrates that the passenger flow interval classification results obtained by applying our proposed AP-based clustering method can achieve better adaptability

Table 1: Passenger flow intervals.

Classification results	2014 Time span	2015 Time span
Interval 1	1–24	1–43
Interval 2	25–29	44–48
Interval 3	30	49
Interval 4	31–47	50–67
Interval 5	48–91	68–91
Interval 6	92–95	92–95
Interval 7	96–122	96–122
Interval 8	123–126	123–126
Interval 9	127–179	127–186
Interval 10	180–251	187–254
Interval 11	252–271	255–271
Interval 12	272–286	272–285
Interval 13	287–365	286–365

Table 2: Amended classification of the passenger flow intervals.

Classification results	2014 Time span	2015 Time span
Interval 1	1–24	1–43
Interval 2	25–47	44–67
Interval 3	48–181	68–184
Interval 4	182–251	185–254
Interval 5	252–271	255–271
Interval 6	272–286	272–285
Interval 7	287–365	286–365

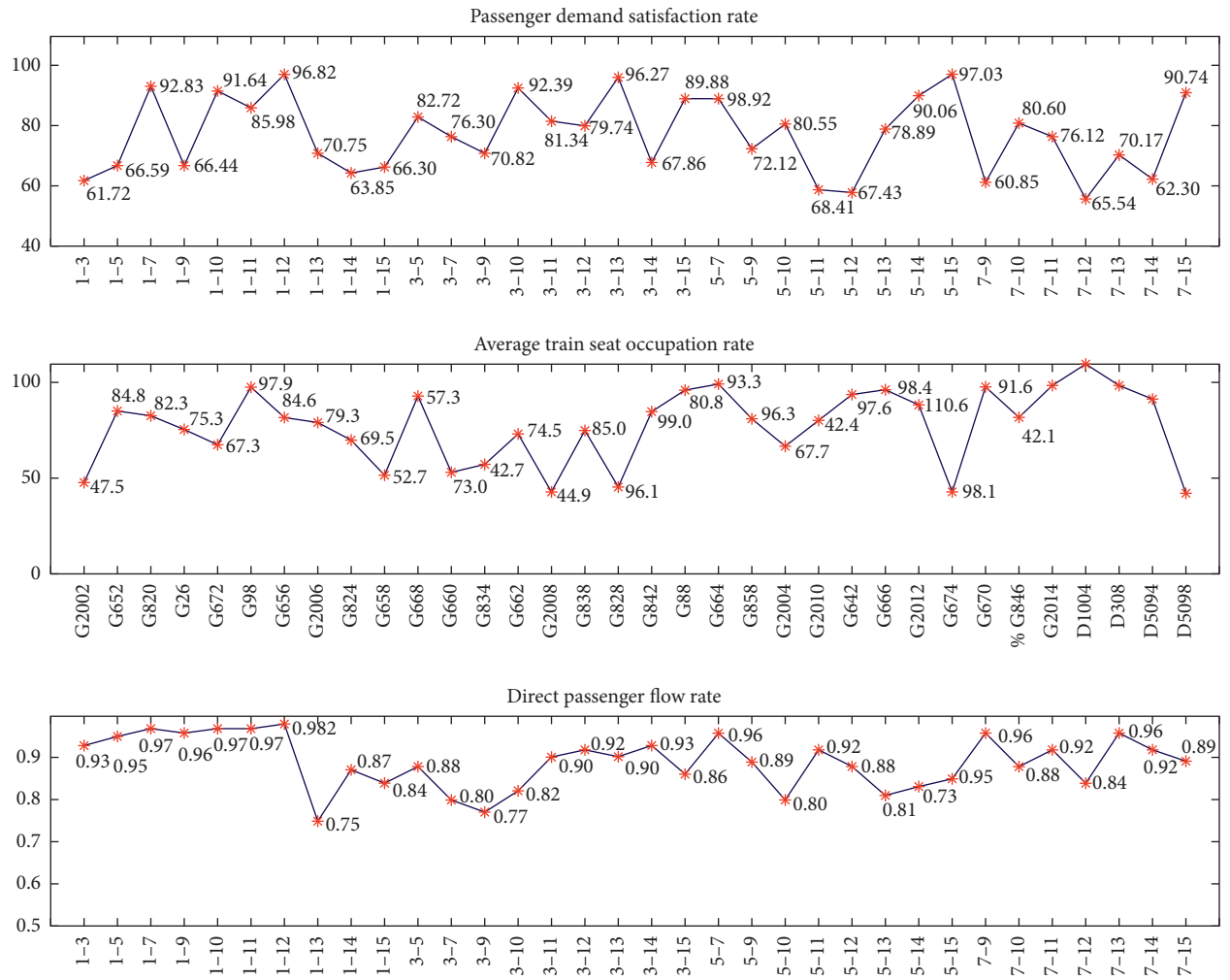

Figure 6: Related index value based on passenger flow intervals dividing simulation (by summer vacation in 2014).

Table 3: Comparison of our results with the actual situation.

	2014		2015	
	Actual	Our proposed method	Actual	Our proposed method
Passenger demand satisfaction rate	83.6%	91.2%	85.0%	90.7%
Average train seat occupation rate	45.4%	61.1%	42.9%	60.3%
Direct passenger flow rate	56.7%	70.8%	60.1%	74.5%
Number of adjustments in line plan	6	6	6	6

evaluation indexes, and the number of adjustments to the line plan remains unchanged. In contrast, the passenger demand satisfaction rate, average train seat occupation rate, and the direct passenger flow rate increase by 7.6%, 16.7%, and 14.1%, respectively, for 2014, and the corresponding values are 5.7%, 18.4%, and 14.4% for 2015.

6. Discussion

In the actual railway passenger transportation, the passenger transportation management department will adjust the train operation plan according to the characteristics of the annual passenger flow, which is specifically reflected in the "daily operation plan," "Spring Festival transportation operation plan," and "summer operation plan." The implemented essence of the above operation plan is the adjustment of the train operation plan based on the dynamic changes of the annual passenger flow, and the basis of the adjustment of the train operation plan lies in the reasonable division of the annual passenger flow period. Therefore, the study of the annual passenger flow period division based on AP has a strong social background and practical significance.

Based on the AP clustering algorithm proposed in this paper, the entire year of the high-speed railway is divided into 7 time periods which can be described as "flat peak period," "spring festival period," "flat peak period," "summer transport period," "flat peak period," "eleventh time period," and "flat peak period" according to time characteristics. On this basis, the train operation plan is prepared according to the average passenger flow of each time period, which can achieve the effect of adapting to the demand of passenger flow to the greatest extent.

7. Conclusions and Future Work

Our paper focuses on the problem of the classification of the passenger flow intervals for a high-speed railway so that the line plan of the high-speed railway can be adapted to the passenger demand. A novel AP-based clustering method is introduced to tackle the passenger flow interval classification problem using the data collected, concerning the number of passengers dispatched at each station along the high-speed railway line for about two or three years. In addition, three validity indexes, Calinski–Harabasz, Hartigan, and In-Group Proportion, are used to decide on the best number of clusters. Finally, field work experience is used to adjust the clustering results of our proposed method slightly to be more reasonable.

The Zhengzhou-Xi'an high-speed railway has been taken as an example to illustrate the effectiveness of the proposed method. We first collect the number of passengers dispatched at four high-speed railway stations from the year of 2014 to 2015, and then our proposed method classifies the 365 days in a whole year into 13 continuous intervals. Finally, the 13 intervals are further integrated into 7 intervals according to the field work experience. Moreover, we compared our classification results with those used by the railway bureau, and the comparison shows that our method can improve the passenger demand satisfaction rate, average train seat occupation rate, and the direct passenger flow rate significantly without changing the number of adjustments to the line plan. In addition, our method can be performed in a fast computer-aided way, which is more objective and accurate.

This paper has proposed an effective AP-based clustering method to classify high-speed railway passenger flow intervals. The limitation of the research in this paper is that the proposed method is limited to the division of passenger flow time periods for a certain railway line. However, China's high-speed railway network is complicated, and different lines connect different regions with different economic development statuses and passenger flow distribution characteristics. As a result, the time span between the peak period and the flat peak period of each line in the same year is different. Therefore, how to divide the time period of high-speed railway passenger flow under the networked condition is a future research direction in this field.

Data Availability

The original passenger flow data used to support the findings of this study are currently under embargo. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research reported here was sponsored by the National Natural Science Foundation of China (61703351) and the Science and Technology Plan of the China Railway Corporation (2016X006-D).

References

[1] E. Barrena, D. Canca, L. C. Coelho, and G. Laporte, "Single-line rail rapid transit timetabling under dynamic passenger

demand," *Transportation Research Part B: Methodological*, vol. 70, pp. 134–150, 2014.

[2] J. Yin, A. D'ariano, Y. Wang, L. Yang, and T. Tang, "Timetable coordination in a rail transit network with time-dependent passenger demand," *European Journal of Operational Research*, in Press, 2021.

[3] J. Yin, T. Tang, L. Yang, Z. Gao, and B. Ran, "Energy-efficient metro train rescheduling with uncertain time-variant passenger demands: an approximate dynamic programming approach," *Transportation Research Part B: Methodological*, vol. 91, pp. 178–210, 2016.

[4] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis: communications in Statistics," *Communications in Statistics*, vol. 3, no. 1, 1974.

[5] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biology*, vol. 3, no. 7, p. 1, Article ID research0036, 2002.

[6] B. Park, D. H. Lee, and I. Yun, "Enhancement of time of day based traffic signal control," in *Proceedings of the IEEE International Conference on Systems*, Washington, DC, USA, October 2003.

[7] M. M. Abbas and A. Sharma, "Optimization of time of day plan scheduling using a multi-objective evolutionary algorithm," *Civil and Environmental Engineering*, vol. 20, 2005.

[8] B. Park and J. Lee, "A procedure for determining time-of-day break points for coordinated actuated traffic signal systems," *KSCE Journal of Civil Engineering*, vol. 12, no. 1, pp. 37–44, 2008.

[9] J. Lee, J. Kim, and B. Park, "A genetic algorithm-based procedure for determining optimal time-of-day break points for coordinated actuated traffic signal systems," *KSCE Journal of Civil Engineering*, vol. 15, no. 1, pp. 197–203, 2011.

[10] T. A. Hauser and W. T. Scherer, "Data mining tools for real-time traffic signal decision support & maintenance," *IEEE*, vol. 3, pp. 1471–1477, 2001.

[11] X. Wang, W. Cottrell, and S. Mu, "Using K-means clustering to identify time-of-day break points for traffic signal timing plans," in *Proceedings of the Intelligent Transportation Systems, IEEE, 2005*, Hong Kong, China, September 2005.

[12] N. T. Ratrout and T. Nedal, "Subtractive clustering-based K-means technique for determining optimum time-of-day breakpoints," *Journal of Computing in Civil Engineering*, vol. 25, no. 5, pp. 380–387, 2011.

[13] D. B. Liu, L. L. Dai, L. I. Ya, and Y. X. Wang, "Intersection traffic control period division method based on signal cycle calculation," *Journal of Jilin University (Engineering and Technology Edition*, vol. 43, 2013.

[14] Y. Shen, T. Zhang, and J. Xu, "Homogeneous bus running time bands analysis based on K-means algorithms," *Journal of Transportation Systems Engineering and Information Technology*, vol. 14, pp. 87–93, 2014.

[15] W. Zhao, D. Wang, and W. Zhu, "Optimization of time-of-day breakpoints based on improved NJW algorithm," *Journal of Zhejiang University (Engineering Science)*, vol. 48, pp. 2259–2265, 2014.

[16] J. Yao, J. Xu, and Y. Han, "TOD optimal control method of urban traffic based on clustering analysis," *Journal of Traffic and Transportation Engineering*, vol. 14, pp. 110–116, 2014.

[17] X. Song, W. Li, D. Ma, Y. Wu, and D. Ji, "An enhanced clustering-based method for determining time-of-day breakpoints through process optimization," *IEEE Access*, vol. 6, pp. 29241–29253, 2018.

[18] J. M. Moreira, L. Moreira Matias, J. Gama, and J. F. Sousa, "Validating the coverage of bus schedules: a machine learning approach," *Information Sciences*, vol. 293, no. 1, pp. 299–313, 2015.

[19] W. Li, F. Sun, X. Li, and D. Ma, "Time-of-day breakpoints for traffic signal control using dynamic recurrence order clustering," *Journal of ZheJiang University (Engineering Science)*, vol. 52, pp. 1150–1156, 2018.

[20] Y. Bie, X. Gong, and Z. Liu, "Time of day intervals partition for bus schedule using GPS data," *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 443–456, 2015.

[21] R. Guo and Y. Zhang, "Identifying time-of-day breakpoints based on nonintrusive data collection platforms," *Journal of Intelligent Transportation Systems*, vol. 18, no. 2, pp. 164–174, 2014.

[22] A. V. Kapp and R. Tibshirani, "Are clusters found in one dataset present in another dataset?" *Biostatistics*, vol. 8, no. 1, pp. 9–31, 2007.