Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 3985302, 9 pages
https://doi.org/10.1155/2018/3985302

Research Article

Analyzing Capacity Utilization and Travel Patterns of Chinese High-Speed Trains: An Exploratory Data Mining Approach

Fanxiao Liu,[1,2] Zhanbo Sun ,[1] Peitong Zhang,[1] Qiyuan Peng,[1,2] and Qingjie Qiao[3]

[1]*School of Transportation and Logistics, Southwest Jiaotong University, China*
[2]*National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, China*
[3]*Beijing-Shanghai High-Speed Railway Co. Ltd, China*

Correspondence should be addressed to Zhanbo Sun; zhanbo.sun@home.swjtu.edu.cn

Received 23 May 2018; Revised 30 July 2018; Accepted 13 August 2018; Published 2 September 2018

Academic Editor: Zhi-Chun Li

Train capacity utilization (TCU), usually represented by passenger load factor (PLF), is a critical measure of effectiveness for rail operation. In literature, efforts are usually made to improve capacity utilization by optimizing rail operation and management strategies. Comparably little attention is paid to analyzing the factors that affect TCU and to understanding the behavioral patterns behind it. This paper applies exploratory data mining techniques to a 3-month long real world train operation data of the Beijing-Shanghai High-Speed Railway. Principal component analysis (PCA) is conducted to find the principal components that can efficiently represent the collected data. Clustering techniques are then applied to understand the unique characteristics that affect PLF and the travel pattern. The findings can be further used to guide train operation planning and facilitate better decision-making.

1. Introduction

Due to the vast land span and enormous transportation demand in China, railway transportation plays an increasingly vital role in China's economy. In general, Chinese high-speed rails are more preferable compared to other transportation modes, especially for long-distance trips. During the last five years, the railway passenger volume in China has been increasing with a yearly growth rate of 10%. According to the 2016 statistics, the Chinese railway passenger volume is 2.8 billion, which has increased 11% compared to 2015. Despite the continuous growth of railway transportation in China, it is found that the train capacity of some passenger lines is underutilized, especially during off-peak seasons. For example, the average passenger load factor of high-speed trains in China is around 60-70%. In extreme cases, the number is less than 40%. And this has motivated transportation researchers to develop methods to reduce such capacity waste. Optimizing train capacity utilization (TCU) is challenging. The challenges are mainly bifold: (i) the passenger travel patterns are highly stochastic and unpredictable; (ii) many factors may influence TCU, and the causalities are hard to be captured. To overcome these challenges, it has become an imperative task to find out the factors that affect TCU and to discover the behavioral patterns behind it.

Generally speaking, there are two approaches to understanding and improving train capacity utilization. One is model-based approach, which applies analytical models to study the effects of train operation and management strategies (e.g., timetabling and ticketing) on train capacity utilization. The second is data mining approach that empirically analyzes TCU and the interrelationship between TCU and the influential factors.

The model-based approach usually assumes that the causalities and quantitative relationships between rail passenger's choice and train operation/management factors are given. For example, pricing and ticketing are often considered as the main management strategies that directly affect TCU. For this, researchers have developed optimal pricing models for better train utilization and revenue generation. Zhang et al. [1] introduced a discriminative pricing method to improve TCU. You [2] formulated a constrained nonlinear integer programming model for railway seat allocation. Shibata et

al. [3], Park et al. [4], and Bao [5] developed seat class assignment models to increase the utilization rate of intercity railway. Wang et al. [6] studied the seat allocation problem to optimize TCU, with considerations of the passengers' random choice behaviors. Another portion of research targets improving TCU by optimizing train operational factors such as train scheduling and timetabling. Zha [7], Lan [8], and Shi et al. [9] developed train operation optimization models to maximize train capacity utilization. Bussieck et al. [10] proposed a novel method to optimize train operation plan by minimizing the number of transfer trips. Methods to improve TCU and revenue generation were also studied by Zhou et al. [11], Cadarso et al. [12], and Robenek et al. [13]. These studies usually assume passenger volume and trip-making decisions are known and fixed. Such assumptions, albeit idealistic, are quite common in literature mainly due to the lack of real world data (which is often true for rail transportation studies in China).

In contrast to the first approach, the empirical approach applies data mining techniques for pattern recognition and knowledge discovery from real world rail operational data. Although data mining approaches have been widely applied in many transportation applications (e.g., Zheng et al. [14]; Xie et al. [15]; Anand et al. [16]), such studies are rare in the field of railway transportation, mainly due to the lack of data. Only a handful of such examples are found in literature. For example, Xu et al. [17] used data mining techniques to analyze the time sequence and the spatial influence of trip making and presented a new approach for trip forecasting. Liu et al. [18] applied fuzzy clustering model to analyze passengers' travel behaviors and key factors relevant to the level of service. Zheng et al. [14] used a data mining approach to analyze train passenger flow and developed a model to forecast passenger volume. To the authors' understanding, no previous work has been done to analyze the influential factors of TCU.

The paper makes contributions in two aspects. (i) Exploratory data mining techniques are applied to a dataset that contains 3-month long real world train operational data of the Beijing-Shanghai High-Speed Railway. Such information is usually held by railway companies and is not available to the general public and the academia. (ii) The unique characteristics that affect PLF and the underlying behavioral patterns are discovered and further analyzed.

The rest of the paper is organized as follows. In Section 2, we briefly describe the data source used in the study. Section 3 presents the key methodologies used for data mining and knowledge discovery from train operation data. The experiment and numerical results are presented in Section 4, followed by the concluding remarks in Section 5.

2. Data Description

The Railway Passenger Transport Management Information System is an official rail operation and management system maintained by China Railway Corporation (CRC). The dataset used for this study was retrieved from the system, which contains 3-month rail operation information of the Beijing-Shanghai High-Speed Railway. This railway line is the most important transportation corridor connecting two largest cities of China. The rail-line has a total length of 1318 km and goes through 24 stations. These 24 stations can be further categorized based on their administrative levels, as shown in Table 1. In general, higher level indicates higher population and higher socioeconomic status. The dataset was further processed to extract 33 representative operational features. Descriptions of the features can be found in Table 2.

Table 1: City levels for the stations on the Beijing-Shanghai high speed railway.

ID	Station	Ab.	Level
s1	Beijing South	BJS	4
s2	Langfang	LF	2
s3	Tianjin West	TJW	4
s4	Tianjin South	TJS	4
s5	Cangzhou West	CZW	2
s6	Dezhou East	DZE	2
s7	Jinan West	JNW	3
s8	Taian	TA	2
s9	Qufu East	QFE	1
s10	Tengzhou East	TZE	1
s11	Zaozhuang	ZZ	1
s12	Xuzhou East	XZE	2
s13	Suzhou East	SZE	2
s14	Bengbu South	BBS	2
s15	Dingyuan	DY	1
s16	Chuzhou	CZ	1
s17	Nanjing South	NJS	3
s18	Zhenjiang South	ZJS	2
s19	Danyang North	DYN	1
s20	Changzhou North	CZN	1
s21	Wuxi East	WXE	2
s22	Suzhou North	SZN	2
s23	Kunshan South	KSS	1
s24	Shanghai Hongqiao	SHHQ	4

The operational features include passenger load factor (PLF) that directly indicates the capacity utilization of a train, date, ticketing strategy (TS), run duration (RDR), departure time (DT), train type (TT), number of stops (NS), run distance (RDI), stop schedule (SS), run speed (RS), and load coefficients (LCs) for all sections along the railway line. The authors are aware of other factors such as trip purposes and passenger social-economic status that could also affect TCU, but such information is not available from the CRC database. Since the ticket prices remain stable during the study period, pricing is not considered as an influential feature in the study.

In literature, PLF is used to assess TCU, and load coefficients are used to assess sectional capacity utilization. In this study, both PLF and load coefficients are considered as important features. Let C denote the train capacity (i.e., number of seats), D is the running distance, S is the number

Table 2: Extracted features.

No.	Feature	Variable type	Value	Notes
1	Passenger Load Factor (PLF)	Ratio	--	γ_{PLF}
2	Date (Date)	Ordinal	1~92	October 1st ~ December 31st
3	Ticketing Strategy (TS)	Nominal	0,1,2	0: Strategy for holiday seasons; 1: Strategy for weekends; 2: Strategy for weekdays
4	Run Duration (RDR)	Measurement	--	Train running time
5	Departure Time (DT)	Measurement	--	Departure time of each train
6	Train Type (TT)	Nominal	0,1	0: Normal trains; 1: Fast trains
7	Number of Stops (NS)	Measurement	--	Number of stops of a train
8	Run Distance (RDI)	Interval	--	Train running distance
9	Stop Scheme (SS)	Nominal	1~5	Higher SS indicates less frequent stops
10	Run Speed (RS)	Measurement	--	Train running speed
11~33	Load Coefficient (LC)	Ratio	--	l_k: load coefficient of the k-th section

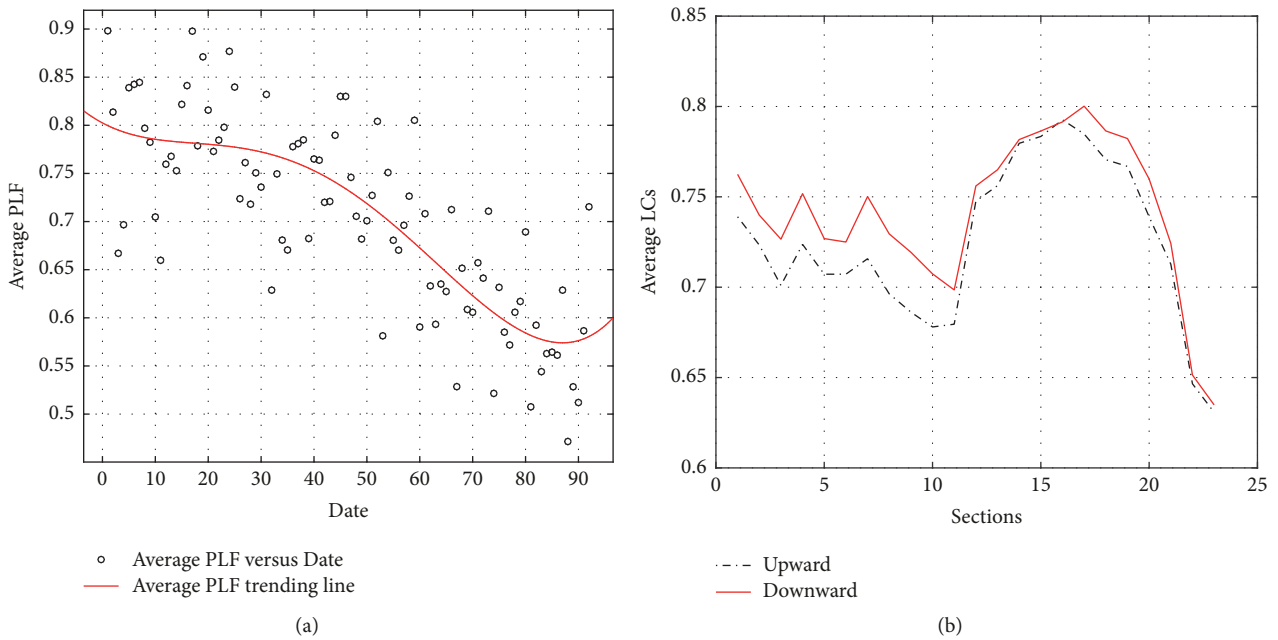


Figure 1: (a) Average PLF distribution; (b) average load coefficients.

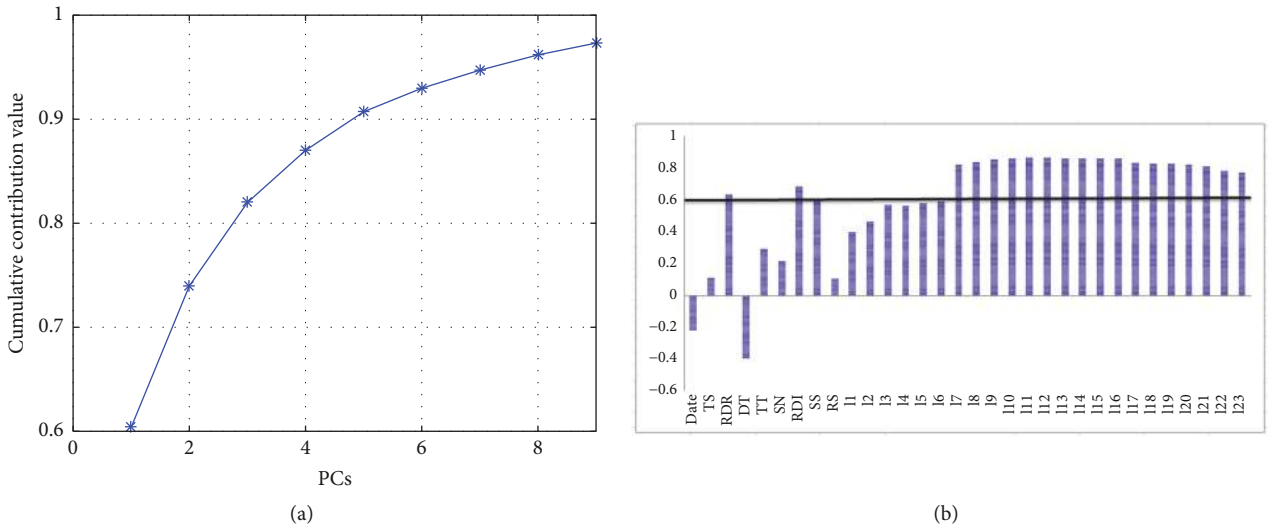


Figure 2: (a) Cumulative level of contribution; (b) correlations between PC1 and selected features.

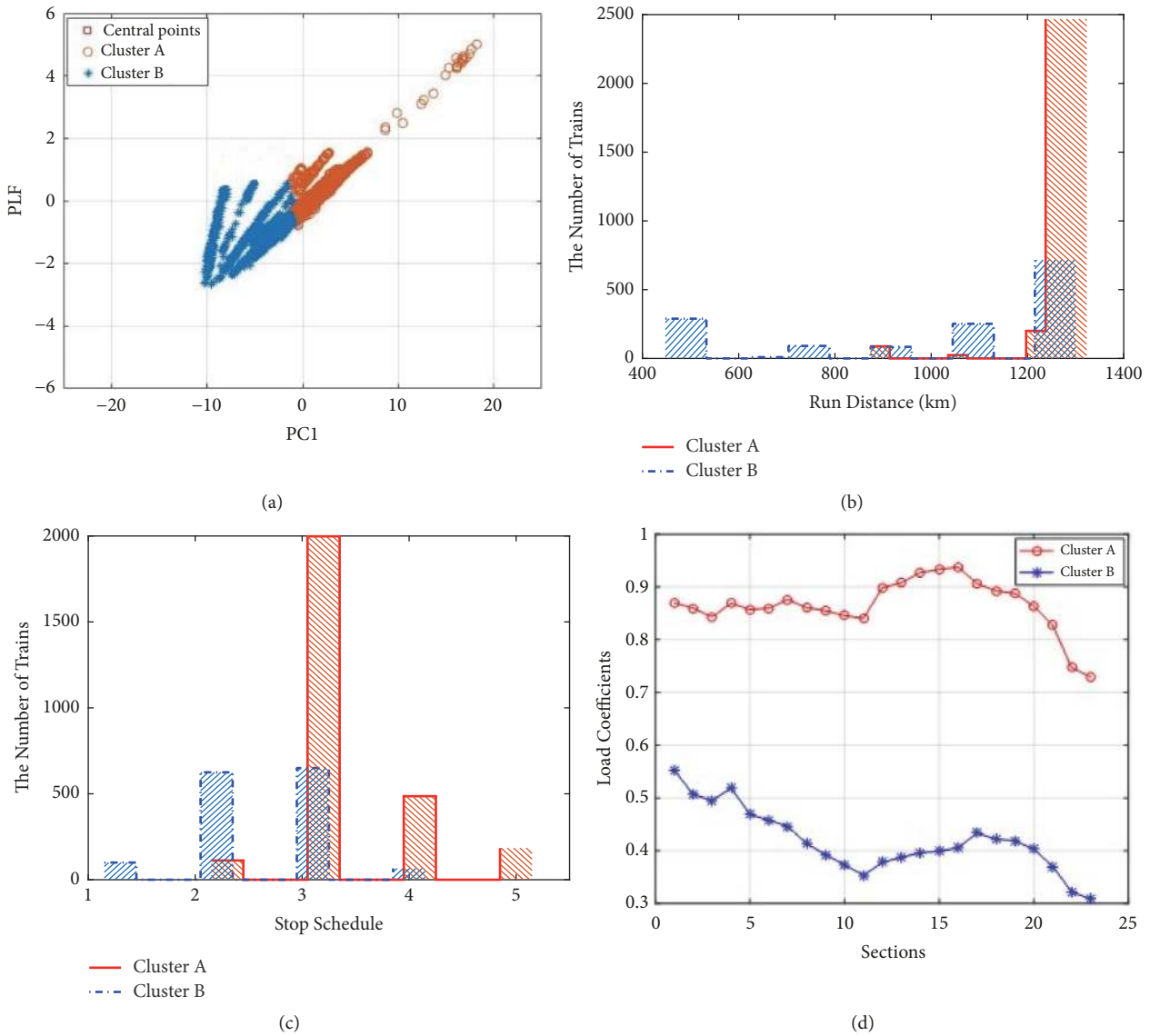


FIGURE 3: (a) The clustering result; (b) run distance distribution; (c) stop schedule distribution; (d) load coefficients of downward trains.

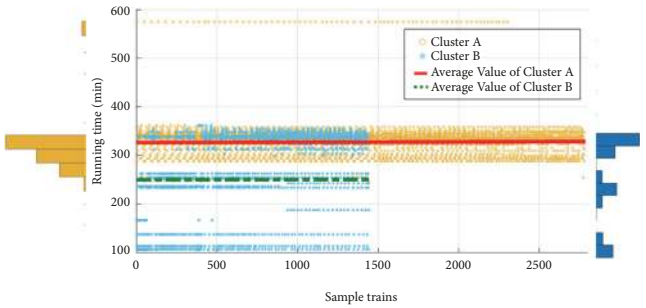


FIGURE 4: Travel time distributions for both clusters.

of stations. PLF can be expressed as (1). Similar definition can be found in Bao et al. [19, 20].

$$\gamma_{PLF} = \frac{\sum_{i=1}^{S-1} \sum_{j=i+1}^{S} n_{ij} \cdot d_{ij}}{C \cdot D} \quad (1)$$

Here n_{ij} and d_{ij} indicate the passenger OD volume and the section length between stations i and j, respectively. Since passenger OD is not available from the dataset, equivalently, we can use the sectional passenger volumes (v_k) to calculate PLF, as in

$$\gamma_{PLF} = \frac{\sum_{i=1}^{S-1} \sum_{j=i+1}^{S} n_{ij} d_{ij}}{C \cdot D} = \frac{\sum_{k=1}^{S-1} v_k d_k}{C \cdot D} \quad (2)$$

Note that the load coefficient of section k is known as $l_k = v_k/C$ according to [21]. Therefore, we can derive the following relationship between PLF and the sectional load coefficients, as in

$$\gamma_{PLF} = \frac{\sum_{k=1}^{S-1} l_k d_k}{D} \quad (3)$$

In Figure 1, we first show the aggregated statistics of the collected data. Figure 1(a) shows the PLF distribution and the

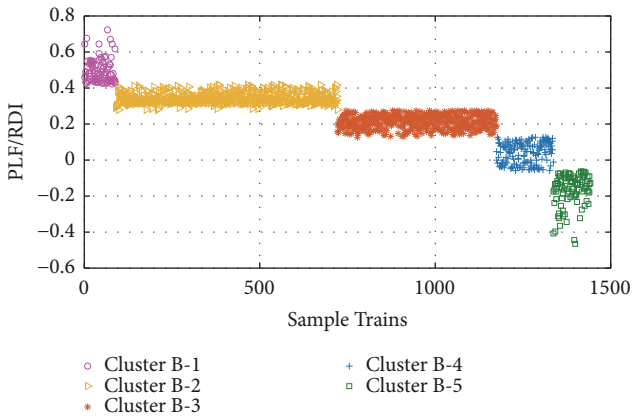


Figure 5: Clustering result using PLF/RDI as the only feature.

trend line. Figure 1(b) shows the average load coefficients of the upward and downward trains. It can be found that the average PLF decreases during the whole study period and the travel pattern may be characterized by two segment trips including s1(BJS)-s12(XZE) and s12(XZE)-s24(SHHQ).

3. Methodology

In the context of statistical analysis and data mining, exploratory data analysis (EDA) is a process of detective work that does not require a predetermined hypothesis to be tested. Rather, the role of EDA is to explore data in as many ways as possible, until a plausible "story" of the data is unearthed. Formal definitions of EDA and exploratory data mining can be found in Tukey [22] and Yu [23]. In this section, exploratory data mining approaches are applied to gain insights of the structure of the data and the underlying travel patterns. First, principal component analysis (PCA) is used to select the most salient features (called principal components) to represent the train operation data. Secondly, we use clustering techniques to discover the intrinsic relationship between TCU and the principal components.

3.1. Principal Component Analysis. PCA is a commonly used technique for dimensionality reduction and feature selection [24]. Here we use PCA to seek a low-rank approximation of the train operational data. In this step, the original 33 train operation features are transformed into a smaller set of new variables called principal components (PCs), which by concept retains similar amount of variation present in the original dataset. PCs are uncorrelated variables, ordered by their variance from the largest variance to the lowest one.

Suppose a zero-centered feature matrix $X = \{X_1, X_2, \ldots, X_N\}^{\mathrm{T}}$ contains $N = 133$ sample trains (called data points) and p=33 features marked as $\{x_1, x_2, \ldots, x_p\}$. $\sum = \mathrm{var}(X)$ is the $p \times p$ variance-covariance matrix. Denote λ_i and e_i as the ranked eigenvalues and associated eigenvectors of $\sum$, where $i = 1, \ldots, p$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. The goal of PCA is to determine a new set of representative variables Y_i, each considered as a linear combination of the original features, as in

$$\begin{aligned} Y_1 &= w_1^T X = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p \\ Y_2 &= w_2^T X = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2p}x_p \\ &\cdots \\ Y_p &= w_p^T X = w_{p1}x_1 + w_{p2}x_2 + \cdots + w_{pp}x_p \end{aligned} \tag{4}$$

and

$$\begin{aligned} \mathrm{var}(Y_i) &= w_i^T \sum w_i, \\ \mathrm{cov}(Y_i, Y_k) &= w_i^T \sum w_k \end{aligned} \tag{5}$$

where $w_1, w_2, \ldots, w_p$ are coefficients of the linear transformations and $\|w_i\|^2 = w_i^T w_i = 1$. By maximizing the variance of variables Y_i, it can be easily shown that $w_i = e_i$, $Y_i = e_i^T X$ and $\mathrm{var}(Y_i) = \lambda_i$. Variables Y_i are referred to as PCs. Further define the level of contribution as $\sum_{i=1}^{\tilde{p}} \lambda_i / \sum_{i=1}^{p} \lambda_i$, $\tilde{p} \leq p$, which represents the percentage of variation explained by the selected PCs. Therefore we can get a reasonable representation of the original data (e.g., with 80% level of contribution) with only a few PCs. Correlation analysis could be conducted to see the correlations between the PCs and the original features.

3.2. Clustering Analysis. Fuzzy c-means clustering (FCM; see [25]) is then used to discover the interrelationship between the principal components (PCs) and the passenger load factor (PLF). The purpose of clustering is to put "similar" samples into the same group and to explore the patterns reflected by different groups. Let $\widetilde{X} = \{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_N\}^T$, $N = 133$, be the transformed train samples, each has q features; i.e., $\widetilde{X}_i = \{\widetilde{X}_i^1, \widetilde{X}_i^2, \ldots, \widetilde{X}_i^q\}, i = 1, 2, .., N$. FCM is used to divide these samples into C clusters; each cluster is characterized by its sample mean, called the centroid. The approach is a standard and widely used data mining approach and is proven to be effective for knowledge discovery from a high-dimensional dataset [26]. FCM does not require each data point to only belong to exactly one cluster; therefore it usually outperforms hard clustering methods (e.g., K-means) for overlapped dataset. The objective of FCM is to minimize the summation of weighted distance between each sample and the centroid of each cluster, as in formulation (6), i.e., to minimize the differences of the samples within the same cluster.

$$\min J_{FCM} = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| \widetilde{X}_i - C_j \right\|^2 \tag{6}$$

Here $m \in [1, \infty)$ is the fuzzy factor that determines the fuzzy weight of the clustering results; u_{ij} is the degree of membership of $\widetilde{X}_i$ in cluster j; and C_j is the centroid of cluster j, in the q-dimensional feature space. Note that the distance between each sample and each cluster centroid is measured

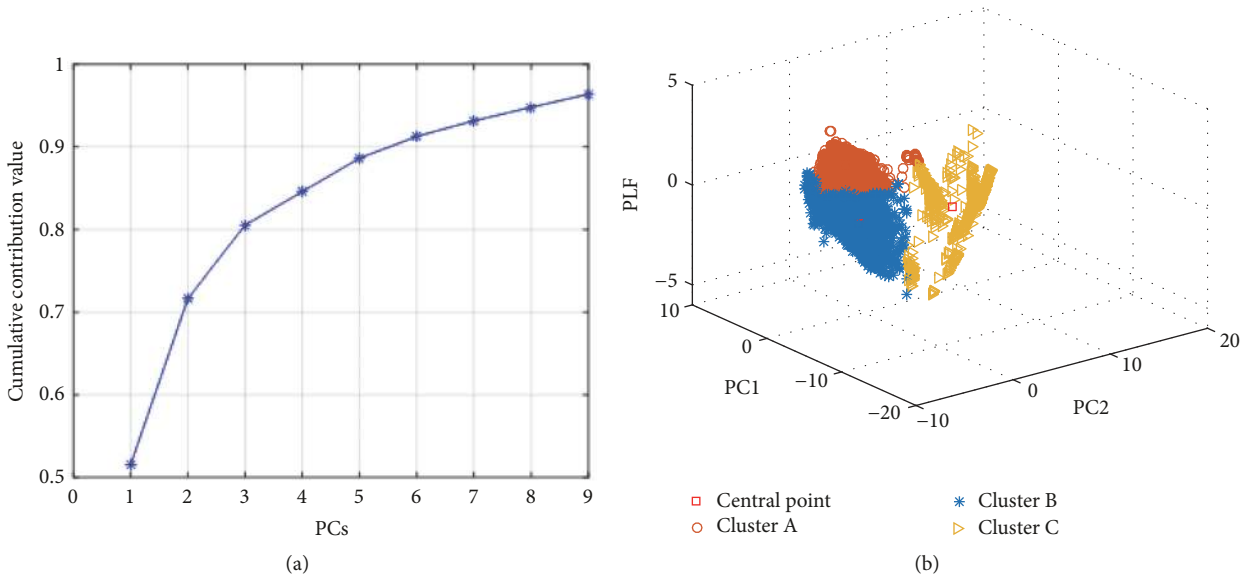


FIGURE 6: (a) Cumulative contribution of the PCs; (b) the clustering result.

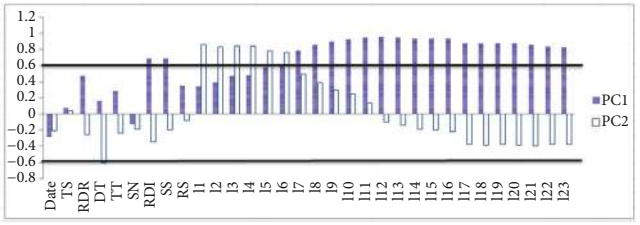


FIGURE 7: Correlations between PC1, PC2, and selected features.

by the Euclidean norm as in (7), where $\widetilde{X}_i^k$ represents the k-th feature of the i-th transformed sample and C_j^k denotes the location of centroid C_j at the k-th dimension.

$$\left\|\widetilde{X}_i - C_j\right\| = \sqrt{\sum_{k=1}^{q}\left(\widetilde{X}_i^k - C_j^k\right)^2} \tag{7}$$

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown in (6), with the updated degree of membership u_{ij} calculated using

$$u_{ij} = \left[\sum_{k=1}^{C}\left(\frac{\left\|\widetilde{X}_i - C_j\right\|}{\left\|\widetilde{X}_i - C_k\right\|}\right)^{2/(m-1)}\right]^{-1} \tag{8}$$

And the cluster centroid C_j can be updated using

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot \widetilde{X}_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{9}$$

The iterative algorithm terminates when $\|C^{(t+1)} - C^{(t)}\| \leq \varepsilon$, where ε is a stop criterion. $C^{(t)}$ is a $k \times q$ cluster centroid matrix, at iteration t. This procedure also at least converges to a local minimum point of J_{FCM}. It is noteworthy that the aforementioned procedure does not specify the number of clusters; the optimal number of clusters is determined based on the Xie-Beni coefficient [27] and Separation coefficient [28] in the experiment.

4. Experiment and Numerical Results

We first separate the samples into downward trains and upward trains. PCA and clustering techniques are then applied to these two datasets. A few interesting findings are generated from the exploratory data analysis and they are discussed in this section.

4.1. Downward Trains. The downward trains represent trains travel from Beijing South (s1) to Shanghai Hongqiao (s24). PCA was firstly applied to the dataset. The cumulative level of contribution (with respect to PCs) is shown in Figure 2(a). It is found that PC1-PC3 account for more than 80% of the total variation. In Figure 2(b), it is shown that PC1 is strongly correlated (degree of correlation > 0.6) with a few features, including run duration (RDR), run distance (RDI), stop scheme (SS), and the sectional load coefficients l_7~l_{23}, from Jinan West (JNW) station to Shanghai Hongqiao (SHHQ) station. Some other features such as Date and Run Speed (RS) are not strongly correlated with PC1. This indicates that PC1 and the strongly correlated features account for the highest variation in the data.

In the following experiment, we use PC1 and PLF for fuzzy c-means clustering. Two optimal clusters are found, which are plotted in Figure 3(a). It can be observed that higher PLF is associated with higher PC1. Since PC1 is positively correlated with RDR, RDI, SS, and l_7~l_{23}, *it can be further inferred that longer run distance/travel time, higher level of stop scheme (i.e., fewer stops), and higher sectional loading*

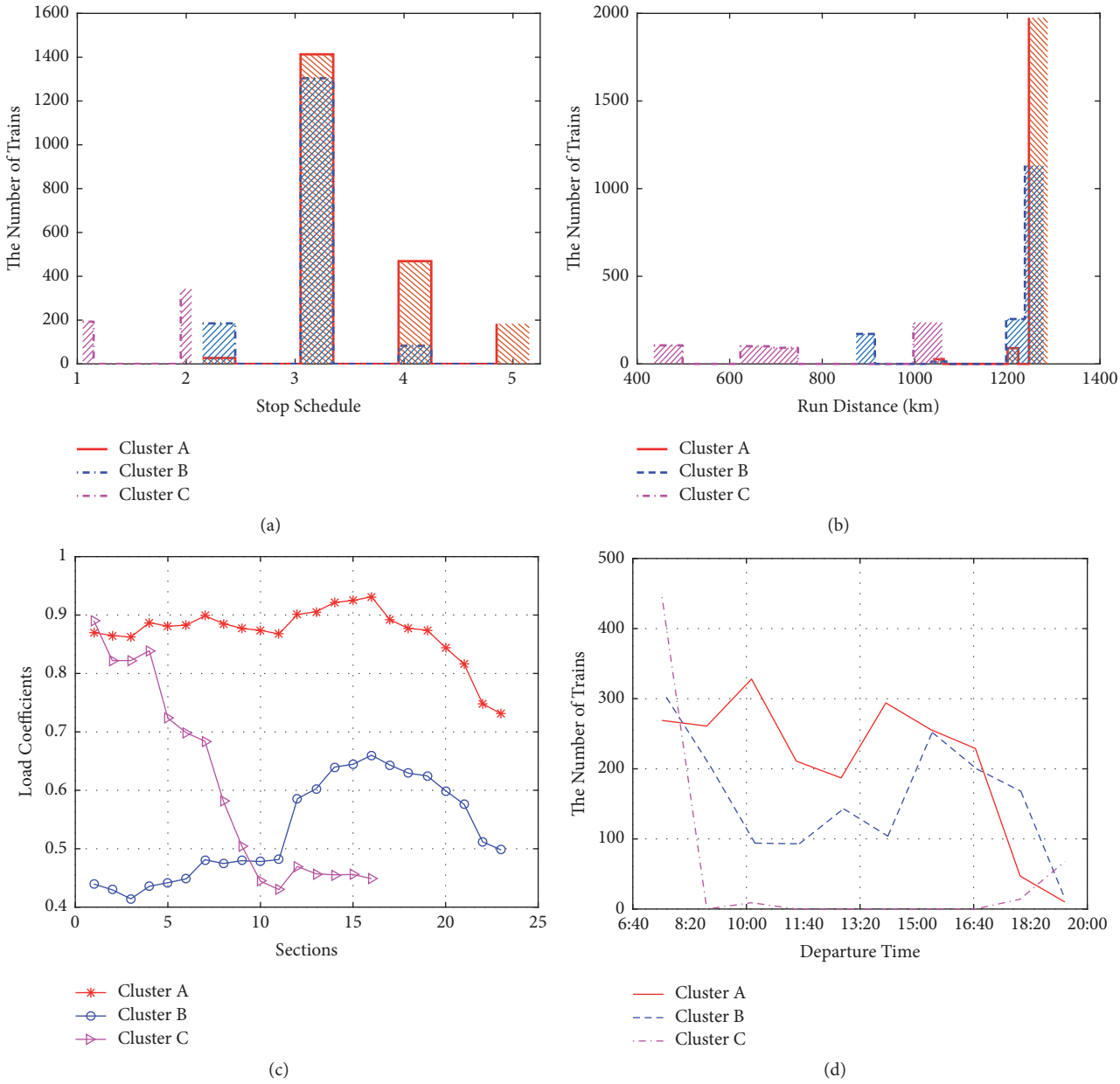


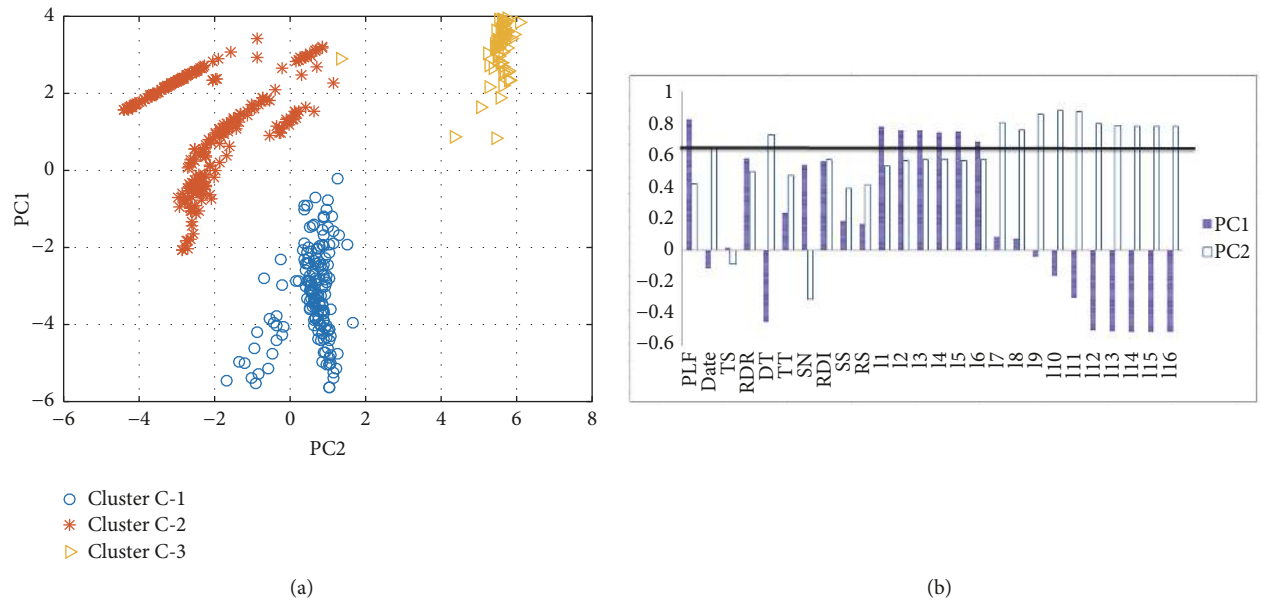
FIGURE 8: (a) Stop schedule distribution; (b) run distance distribution; (c) load coefficients of upward trains; (d) departure time distribution.

coefficients from JNW station to SZN station are associated with trains of higher PLF. Such inference can be validated by plotting the distributions of these original features for each cluster, as shown in Figures 3(b), 3(c), 3(d), and 4.

It is also noticed that cluster B in Figure 3(a) shows the multifurcated lines with different slopes, representing different rates of PLF to PC1. To further analyze the pattern, we used RDI as a surrogate of PC1 and applied the clustering model using PLF/RDI as the only feature. The results in Figure 5 have shown five clusters which correspond to the five linear lines shown in Figure 3(a). *The results imply that the marginal effect of RDI gradually decreases; i.e., changing short-distance trains to medium-distance trains seems to be more beneficial (in terms of the gain in PLF) compared to changing medium-distance trains to long-distance trains.* This finding can be used to guide train scheduling.

4.2. Upward Trains. The cumulative level of contribution of each PC is shown in Figure 6(a) for the upward trains from Shanghai Hongqiao (s24) to Beijing South (s1). We then conducted clustering analysis using PLF, PC1, and PC2. It is found that the optimal number of clusters is 3, as shown in Figure 6(b).

Figure 7 shows the original features that are strongly correlated with PC1 and PC2. In particular, it is found that l_{23}~l_7, RDI, and SS are strongly correlated with PC1; LCs (l_6~l_1) and departure time (DT) are strongly correlated with



(a) (b)

Figure 9: (a) The subclusters of cluster C; (b) correlations between PCs and selected features of cluster C.

PC2. For the upward trains, a few findings of TCU and passenger travel patterns can be put forward.

As observed in Figure 6(b), compared to the samples with lower PLF (cluster B), trains with higher PLF (cluster A) are associated with larger PC1, indicating that *higher SS (fewer stops), longer RDI, and higher LCs lead to better train capacity utilization.* This is further verified in Figures 8(a), 8(b), and 8(c). Such finding is consistent with the downward trains.

The result in Figure 6(b) also shows that a cluster C, separated from the other two clusters, has large variation in the dimension of PC2. By further analyzing the distributions of DT (a surrogate of PC2), it is found that cluster C is associated with the samples that have early departure time (as shown in Figure 8(d)) and go through fewer sections/shorter distance (as shown in Figures 8(b) and 8(c)). These samples correspond to the extra (temporal) short-distance trains that depart in the early morning. We then rerun the PCA and clustering models only for cluster C samples to further explore the patterns of these extra trains. The results are shown in Figure 9.

It is shown that PLF and LCs (l_6~l_1) are strongly correlated with PC1; date, DT, and LCs (l_{16}~l_7) are strongly correlated with PC2. As in Figure 9(a), cluster C-2 and cluster C-3 are in the higher region of PC1; cluster C-1 is in the lower region of PC1. It is found that early of this quarter and early DT are associated with higher PLF with greater LCs (l_6~l_1), as illustrated by cluster C-2; late of this quarter and relatively late DT also lead to the higher PLF with greater LCs. It is noteworthy that early of the quarter corresponds to the "Golden week" (Chinese national holiday) and late of the quarter is close to the New Year. *Therefore, the extra trains with early or late departure time are better utilized in the holidays seasons compared to those in other seasons.*

By scrutinizing Figure 8(c), it is found that the major trip attraction for cluster A trains is Beijing (as the load coefficient is high at section 1), and the major trip attraction for cluster B trains is the city of Xuzhou (XZE station), a medium-level city. Combining the patterns in Figures 8(a) and 8(c), it can be concluded that *passengers traveling to Beijing prefer to choose the trains with fewer stops, most likely due to their higher value of time.*

5. Concluding Remarks

This paper proposes an exploratory data mining approach to discover the influential features of TCU and understand the travel patterns using real world train operational data. Several interesting findings were reported in the paper, as summarized below.

(1) Run distance and stop scheme are found to be closely related to TCU. Per the specific dataset, trains with longer run distance and fewer stops result in higher TCU.

(2) The marginal effect of travel distance decreases in terms of the gain in TCU. Making the short-distance trains into medium-distance trains is more beneficial compared to making medium-distance trains into long-distance trains.

(3) The extra (temporal) trains are better utilized during the holiday seasons, and the extra trains in off-peak seasons are not as well-utilized.

(4) Passengers to major cities prefer trains with fewer stops. Such behavioral pattern can be explained by their value of time.

These findings, albeit case-specific, have shown that the proposed approach is a useful tool for data mining and knowledge discovery from train operational data and it can

be utilized to facilitate smarter decision-making for train operation and management.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The first and the fourth authors are supported by the National Key Research and Development Program of China (Project No. 2017YFB1200701).

References

[1] X. M. Zhang, D. M. Zhao, and S. D. Wen, "Revenue Management of Railway tickets," *Railway Transport and Economy*, vol. 28, no. 7, pp. 7–9, 2006.

[2] P.-S. You, "An efficient computational approach for railway booking problems," *European Journal of Operational Research*, vol. 185, no. 2, pp. 811–824, 2008.

[3] J.-Y. Lee, J.-H. Chung, and B. Son, "Incident clearance time analysis for Korean freeways using structural equation model," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 8, pp. 1850–1863, 2010.

[4] C. Park and J. Seo, "Seat inventory control for sequential multiple flights with customer choice behavior," *Computers & Industrial Engineering*, vol. 61, no. 4, pp. 1189–1199, 2011.

[5] Y. Bao, *The theory and methods for railway seat inventory control*, Beijing Jiaotong University, 2014.

[6] X. Wang, H. Wang, and X. Zhang, "Stochastic seat allocation models for passenger rail transportation under customer choice," *Transportation Research Part E: Logistics and Transportation Review*, vol. 96, pp. 95–112, 2016.

[7] W. X. Zha and Z. Fu, "Research on the optimization method of through passenger train plan," *Journal of the China Railway Society*, vol. 22, no. 5, pp. 1–6, 2000.

[8] S. M. Lan, "Research on the passenger train plan for Beijing-Shanghai high-speed railway," *Railway Transport and Economy*, vol. 24, no. 5, pp. 31–34, 2002.

[9] F. Shi, L.-B. Deng, X.-H. Li, and Q.-G. Fang, "Research on passenger train plans for dedicated passenger traffic lines," *Journal of the China Railway Society*, vol. 26, no. 2, pp. 16–20, 2004.

[10] M. R. Bussieck, T. Lindner, and M. E. Lubbecke, "A fast algorithm for near cost optimal line plans," *Mathematical Methods of Operations Research*, vol. 59, no. 2, pp. 205–220, 2004.

[11] W.-L. Zhou, F. Shi, Y. Chen, and L.-B. Deng, "Method of integrated optimization of train operation plan and diagram for network of dedicated passenger lines," *Tiedao Xuebao/Journal of the China Railway Society*, vol. 33, no. 2, pp. 1–7, 2011.

[12] L. Cadarso, Á. Marín, J. L. Espinosa-Aranda, and R. García-Ródenas, "Train Scheduling in High Speed Railways: Considering Competitive Effects," *Procedia - Social and Behavioral Sciences*, vol. 162, pp. 51–60, 2014.

[13] T. Robenek, S. Sharif Azadeh, Y. Maknoon, and M. Bierlaire, "Hybrid cyclicity: Combining the benefits of cyclic and non-cyclic timetables," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 228–253, 2017.

[14] D. Zheng, Y. Wang, P. Z. Tang, and Y. P. Wu, "Application of data mining in the forecasting of railway passenger flow," *Advanced Materials Research*, vol. 834-836, pp. 958–961, 2013.

[15] X.-L. Xie and X.-F. Gu, "Research on data mining model of intelligent transportation based on granular computing," *International Journal of Security and Its Applications*, vol. 10, no. 7, pp. 281–286, 2016.

[16] S. Anand, P. Padmanabham, A. Govardhan, and R. H. Kulkarni, "An Extensive Review on Data Mining Methods and Clustering Models for Intelligent Transportation System," *Journal of Intelligent Systems*, vol. 27, no. 2, pp. 263–273, 2018.

[17] W. Xu, H. K. Huang, and Y. Qin, "Study of railway passenger flow forecasting method based on sptio-temporal data mining," *Journal of Northern Jiaotong University*, pp. 401–405, 2004.

[18] J. Liu and N. Zhang, "Empirical research of intercity high-speed rail passengers' travel behavior based on fuzzy clustering model," *Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology*, vol. 12, no. 6, pp. 100–105, 2012.

[19] Y. Bao, J. Liu, M.-S. Ma, and L.-Y. Meng, "Nested seat inventory control approach for high-speed trains," *Tiedao Xuebao/Journal of the China Railway Society*, vol. 36, no. 8, pp. 1–6, 2014.

[20] Y. Bao, J. Liu, M.-S. Ma, and L.-Y. Meng, "Seat inventory control methods for Chinese passenger railways," *Journal of Central South University*, vol. 21, no. 4, pp. 1672–1682, 2014.

[21] S. G. Arul, "Methodologies to monetize the variations in load factor and GHG emissions per passenger-mile of airlines," *Transportation Research Part D: Transport and Environment*, vol. 32, pp. 411–420, 2014.

[22] J. W. Tukey, *Exploratory data analysis*, Addison-Wesley, Boston, Massachusetts, USA, 1977.

[23] C. Ho Yu, "Exploratory data analysis in the context of data mining and resampling.," *International Journal of Psychological Research*, vol. 3, no. 1, p. 9, 2010.

[24] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[25] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[26] C. W. Wang and J. H. Jeng, "Image compression using PCA with clustering," *International Symposium on Intelligent Signal Processing & Communications Systems*, vol. 41, no. 11, pp. 458–462.

[27] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[28] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering," *Pattern Recognition*, vol. 32, no. 7, pp. 1089–1097, 1999.