Rail Transit Stations Classification Based on Spectral Clustering

ABSTRACT

To identify the function and positioning of urban rail stations, and provide further guidance for design and construction, a classification method based on spectral clustering algorithm is established. Firstly, based on the principles of comprehensiveness and robustness, 5 initial indicators were selected, including total entry count, total exit count, entrances count, bus connecting lines count, and metro connecting lines count. Secondly, we normalize the original data by Z-score method and extract two main clustering factors through principal component analysis. Finally, we propose a station classification model based on spectral clustering algorithm. The effectiveness of the proposed method is verified in Hangzhou Metro System. The K-means cluster algorithm and spectral cluster methods are employed. The results show that the proposed model can successfully identify the types of urban rail transit stations, clarify the function and orientation of each station.

Keywords: Urban rail transit; spectral clustering; station classification; cluster analysis

1. INTRODUCTION

The urban rail transit system plays an important role in optimizing the layout of urban functions, satisfying people's travel demands, mitigating the urban traffic congestion, and promoting economic and social development. As key nodes of the system, the urban rail stations could not only serve transportation functions, but also provide public space for social activities. Because of the difference between various types of stations' function, it's essential to classify the stations and discuss their roles in improving the operational efficiency of public transport, promoting differentiated management of stations, and forecasting passenger flow at new routes.

In the determination of clustering indicators, Scholars have listed two categories. One is to classify the station based on the characteristics of the connected node, by using the type of station connection mode, the number of track lines, the main function of the station service area, and the type of station land development as classification indicators[1]. Another is to classify the station based on the intention of the passengers and the station passenger flow as classification indicators[2]. Both two methods can reflect certain characteristics of rail stations, but they also lack the description of other characteristics.

In the selection of clustering methods for urban rail stations, most of the previous studies used traditional clustering methods, such as K-Means clustering[3], hierarchical clustering[4], and two-step clustering[5]. Li S et al.[6] constructed the index of the mutual influence between subway and land use, and divided Guangzhou metro stations into 6 categories by systematic clustering method; Deng P et al.[7] applied K-Means clustering to classify rail transit stations; Rao C et al.[8] analyzed the land development conditions around rail transit stations, and classified Hangzhou Metro Line 1 stations through Analytic Hierarchy Process (AHP). When using conventional clustering methods to classify stations, there are often

cases where local classification is too fine and one station is isolated as one class[9]. Therefore, a comprehensive and reliable station classification method is urgently needed.

As a clustering method based on the graph theory, spectral clustering algorithms have been widely used in various fields in recent years, such as image segmentation[10][11], data clustering[12][13], etc. Hu X et al[12] used a spectral clustering algorithm to classify the natural products in Polygonum multiflorum according to the molecular complexity, structure, and content. It provides a valid method for clustering plant-derived natural products. Yao Z et al[13] analyzed the classification characteristics of interstation passenger flow and obtained the time distribution types of passenger flow between urban rail transit stations using a spectral clustering method. The silhouette coefficient and Davies-Bouldin index are used to demonstrate that the proposed spectral clustering method has better classification results compared with other methods such as k-means.

Aiming at the above problems, this paper applied the spectral clustering method to establish a classification model of urban rail transit stations. Experimental verification analysis was carried out by the Hangzhou Metro System data.

2. RAIL TRANSIT STATION CLUSTERING ANALYSIS METHOD

2.1 Clustering Indicator Analysis

The selection of clustering indicators should be based on the following principles[14]:

(1) Each indicator can reflect a certain characteristic of the rail transit station, and all the indicators reflect all the node attributes and site attributes of the rail station as comprehensively as possible.
(2) Each indicator should be quantifiable.
(3) Values of each indicator are easy to obtain and collect.

Combined with the above indicator selection principles, 5 indicators are selected, total entry count, total exit count, entrances count, bus connecting lines count, metro connecting lines count. The initial cluster indicators are shown in Table 1.

Table 1 Description of Initial clustering indicators

ID	Indicator	Description
1	Total Entry Count	Total number of inbound passengers
2	Total Exit Count	Total number of outbound passengers
3	Entrances Count	Quantity of station entrances
4	Bus Connecting Lines Count	Quantity of connecting bus lines within 500m
5	Metro Connecting Lines Count	Quantity of connecting metro lines

2.2 Data Pre-processing

Before the cluster analysis, because the data units or orders of magnitude of each indicator may be different, it is necessary to standardize the original data to eliminate the influence of different data dimensions. In this paper, Z-score is selected to standardize the data.

x_{ij} is the j indicator's raw value of the i station, and x_{ij} is the standardized value. The calculation formula is shown in (2).

$$x_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j} \tag{2}$$

In which (2), $\overline{x_j}$ is the sample mean, s_j represents the standard deviation.

2.3 Feature Extraction Based on the PCA

5 initial clustering indicators were selected. Each of the feature value reflects the partial information of the rail transit stations in different aspects, so it is easy to have the problem of the multivariate collinearity. In order to avoid the situation of high computational cost and inaccurate clustering results caused by directly using the initial indicators for clustering, it is necessary to reduce the dimension of the features value and extract the common factors to replace the initial clustering indicators[15].

The Principal Component Analysis (PCA) use the idea of dimensionality reduction to condense multiple indicator variables into a few principal components by linear transformation. The principal components are independent of each other and could reflect most of the information of the original indicators.

2.4 Clustering Analysis

Clustering analysis could group objects into multiple classes. Objects in the same class have a high degree of similarity, and objects in different classes have a high degree of difference. And cluster analysis has the advantages of flexibility, robustness, distribution and self-organization. It can comprehensively use multiple indicator variables to perform careful and reasonable clustering of samples. Therefore, it is suitable for the classification of the rail transit stations.

When using the traditional clustering algorithm to classify rail transit stations, it often happens that the local classification is too fine to isolate a station as one class. Analyzing the principle of algorithm, it could be noticed that traditional clustering algorithm (such as K-means) directly classifies the characteristic matrix during the calculation, which makes the classification result susceptible to the influence of indicators with small differences[16].

Spectral clustering is a clustering algorithm based on the graph theory, and its essence is to transform the clustering problem into the optimal partition problem of the graph[17]. The spectral clustering algorithm calculates the eigenvalues and eigenvectors of the matrix first, and then selects the appropriate eigenvectors for clustering. Therefore, compared with traditional clustering algorithms, spectral clustering could adapt to more types of data distribution and work on arbitrary sample spaces.

2.4.1 Determine the Number of Clusters

Before clustering the sample data with the spectral clustering algorithm, the number of clusters should be determined first. In this paper, the number of clusters k is determined by the elbow method[18] and the silhouette coefficient method[19].

The core index of the elbow method is Sum of Squared Errors (SSE). The calculation formula is shown in (3).

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |^2 \tag{3}$$

In which (3), C_i represents cluster i ; p represents a sample in C_i ; m_i represents the mean of all samples in C_i ; SSE represents the clustering error of all samples, which

reflects the qualification of the clustering effect. The larger the value of k, the smaller the value of SSE, and the better the sample will be classified. Moreover, when k is less than the optional number of clusters, the decline speed of SSE is rapid. When k is greater than the optional number of clusters, the decline speed of SSE decreases significantly.

The core index of the silhouette coefficient method is the silhouette coefficient. The calculation formula is shown in (4).

$$S = \frac{b-a}{\max(a,b)} \tag{4}$$

In which (4), a is the degree of cohesion, which represents the average distance between X_i and other samples in the same cluster. b is the degree of separation, which represents the average distance between X_i and all samples in the nearest cluster. The calculation formula of the nearest cluster is shown in (5).

$$C_j = \arg\min_{C_k} \frac{1}{n} \sum_{p \in C_k} | p - X_i |^2 \tag{5}$$

In which (5), p is a sample in cluster C_k.

The average silhouette coefficient is the mean of the silhouette coefficients of all samples. The closer sample distance within the cluster and the farther the sample distance between clusters, the larger average silhouette coefficient and the better the sample will be classified.

2.4.2 Spectral Clustering

According to the different spectral decomposition forms of the Laplace matrix, there are varieties of implementation method of spectral clustering algorithm. The process of different methods could be summarized as the following steps[20][21]:

Step 1: construct a graph $G = (V, E)$ according to the data set $X = \{x_1, x_2, \cdots, x_n\}$, the vertex V_i of the graph represents the data x_i, and the weight w_{ij} of edge E_{ij} represents the similarity between x_i and x_j.

Step 2: construct the similarity matrix $W = \left[w_{ij} \right]_{n \times n}$ according to the Gaussian Kernel Function $w_{ij} = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2})$. Obviously, $w_{ij} = w_{ji}$, and elements on the diagonal of W are 0.

Step 3: calculate the degree matrix $D = diag(d_1, d_2, \cdots, d_n)$, where $d_i = \sum_{j=1}^{n} w_{ij}$.

Step 4: calculate the Laplace matrix.

Step 5: calculate the first k eigenvalues of the Laplace matrix L_{rw} and their corresponding eigenvectors (in ascending order). Then use the k eigenvectors to construct a matrix $V = [v_1, v_2, ..., v_k] \in R^{n \times k}$.

Step 6: treat each row of matrix V as a point, and perform clustering with K-means. If row i is classified into cluster j, the corresponding point belongs to cluster j.

In this paper, we select three commonly used spectral clustering algorithms for classification. When we obtain the Laplace matrix by formula (6), it is the unnormalized spectral clustering algorithm. When we obtain the Laplace matrix by formula (7), then normalize the matrix V into matrix U and turn to step 6, it is the NJW algorithm[22]. When we obtain the Laplace matrix by formula (8), then normalize the matrix V into matrix U and turn to step 6, it is the MS algorithm[23].

$$L = D - W \tag{6}$$

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = \text{I-}D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{7}$$

$$L_{rw} = D^{-1} L = \text{I-}D^{-1} W \tag{8}$$

3. HANGZHOU RAIL TRANSIT STATIONS CLUSTERING EXAMPLE

The passenger flow data used in this paper is the daily entry and exit statistics of the Automatic Rail Fare Collection System (AFC). The sampling time is from December 22, 2020 to December 31, 2020. After data selection, there are a total of 116 rail transit stations in Hangzhou.

3.1 The Test of KMO and Bartlett

The KMO and the Bartlett test were carried out to examine the structural validity. The results are shown in Table 2. The KMO test coefficient is 0.652, the significance is 0.000, and the structural validity is high. The PCA method can be used for data processing.

Table 2 The Test of KMO and Bartlett

KMO Test	Bartlett Test		
	Approximate chi-square	df	Sig.
0.652	552.015	10	0.000

3.2 Feature Extraction

The initial feature values were subjected to the principal component analysis, and the feature values of each principal component and their contribution rates were calculated. The results are shown in Table 3. Two principal component indicators are extracted from the 5 initial feature values reflecting the characteristics of the rail transit stations. These 2 indicators explain that 77.861% of the original information, indicating that the original information has been well extracted, so these 2 principal component indicators can be selected for further analysis.

Table 3 The Index Feature Value and Contribution Rate

Components ID	Feature Value	Contribution Rate/%	Accumulation
1	2.667	53.350	53.350
2	1.226	24.511	77.861
3	0.613	12.266	90.127
4	0.445	8.909	99.036
5	0.048	0.964	100.000

In order to analyze and describe the station classification result obtained by clustering, it is necessary to determine the meaning of each principal component after the factor extraction. The component matrix is shown in Table 4.

Table 4 Component Matrix

Initial Indicators	Components	
	1	2
Total Entry Count	**0.915**	−0.285
Total Exit Count	**0.910**	−0.302
Entrances Count	**0.765**	−0.127
Bus Connecting Lines Count	0.375	**0.793**
Metro Connecting Lines Count	0.527	**0.638**

It can be seen from Table 4 that the main component 1 has a larger load value in total entry count, total exit count, and the quantities of entrances and exits, which is 0.765~0.915. Therefore, the principal component 1 reflects the flow characteristics of stations. The principal component 2 has a larger load value on bus connecting lines count and metro connecting lines count, which is 0.638~0.793. Therefore, the principal 2 reflects the connection characteristics of the stations.

3.3 Determine the Number of Clusters

The relationship between the number of clusters k and the silhouette coefficient is shown in Figure 1, and the relationship between the number of clusters k and SSE is shown in Figure 2. According to Figure 1, when the value of silhouette coefficient is the largest, k is equal to 2, indicating that the optimal number of clusters is 2. However, it can be seen from Figure 2 that SSE is still very large when k is equal to 2. Considering Figure 1 and Figure 2 comprehensively, when the second largest silhouette coefficient is taken, k is equal to 7, and now SSE is already at a low level. Therefore, the optimal number of clusters should be 7.

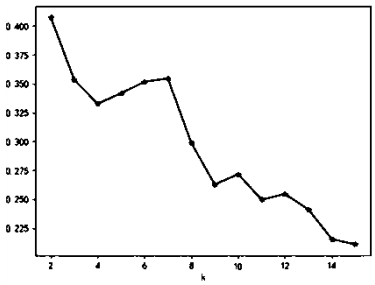

Figure 1 The Relationship between k and the Silhouette Coefficient

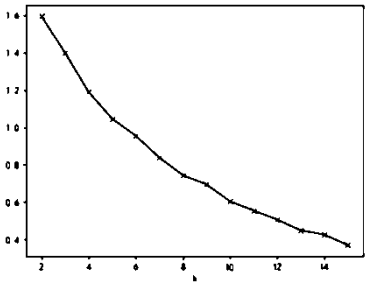

Figure 2 The Relationship between k and SSE

3.4 The Clustering Result

In this paper, four clustering methods, including K-means, unnormalized spectral clustering algorithm, NJW algorithm and MS algorithm, are used to analyze the rail transit stations described by 2 principal component factors, and the appropriate number of station classifications is 7 categories, and the specific classification details of 116 rail transit stations are obtained. The station classification result obtained by different spectral clustering algorithms are shown in Figure 3.

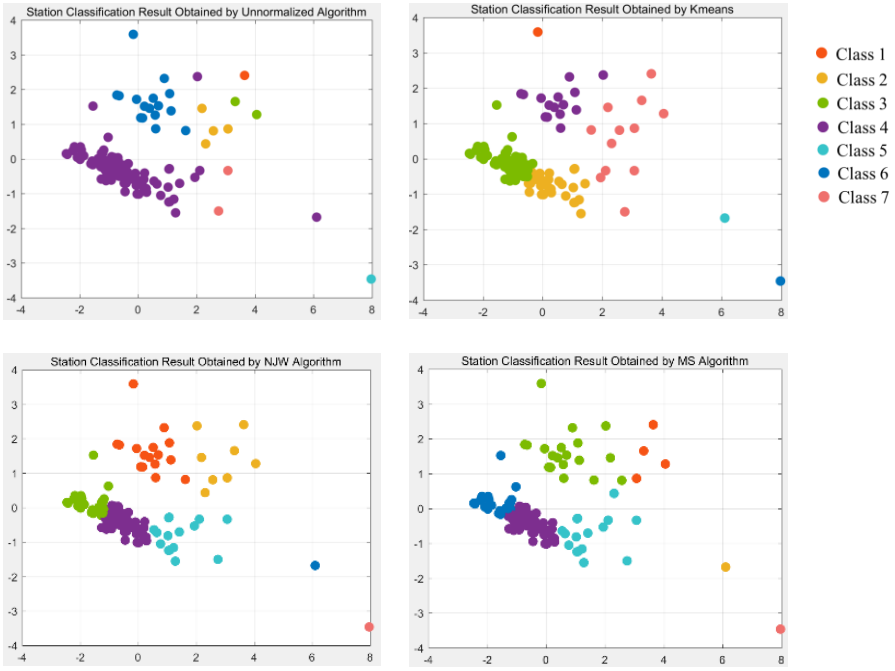

Figure 3 Station Classification Result Obtained by Different Clustering Methods

According to Figure 3, all the four clustering methods can classify the stations. Nevertheless, in the classification results of the unnormalized spectral clustering algorithm and K-means, some local classifications are so fine that some stations are isolated as one class. Additionally, the calculation speed of unnormalized spectral clustering is slow. Therefore, the classification results of the NJW algorithm and the MS algorithm are more in line with the requirements of station classification.

3.5 The Clustering Result Category Feature Analysis

The characteristics of different types of stations are shown in Table 5.

Table 5 Characteristics of Different Types of Stations

ID	Type	Hub Characteristics	Location Characteristics	Scale Characteristics
1	Comprehensive transportation hub stations	●	◎	○
2	Urban center stations	◎	○	⊕
3	Urban sub-center stations	◎	⊕	●
4	Suburban stations	○	◎	○
5	Transportation connection stations	⊕	◎	⊕
6	Urban residential stations	○	⊕	⊕
7	Urban general stations	⊕	⊕	⊕

* ○＜⊕＜◎＜●

The characteristics of the stations in Class 1 are: the stations have strong hub characteristics, there are many buses and subway lines connected, the passenger traffic and other transportation methods are connected and transferred, and it is adjacent to a large comprehensive transportation hub, East Hangzhou Railway Station. Therefore, such stations are defined as comprehensive transportation hub stations.

The characteristics of the stations in Class 2 are: the characteristics of the station hub are strong. The number of connecting bus lines is large. And the traffic accessibility is high. Located in the central business district of the city, the surrounding land development intensity is high. and the average distance between stations is small. The scale of the stations is medium, but the passenger volume is large. Therefore, such stations are defined as urban center stations.

The characteristics of the stations in Class 3 are: the characteristics of the station hub are strong, the number of connecting bus lines is large, and the transportation connection is relatively convenient. With a large number of station entrances and exits, the station is large in scale. Therefore, such stations are defined as urban sub-center stations.

The characteristics of the stations in Class 4 are: the station hub characteristics are weak and traffic accessibility is relatively poor. The surrounding land development intensity is low

and it is far from the city center. The passenger volume of the station is small, and the scale of the station is small. Therefore, such stations are defined as suburban stations.

The characteristics of the stations in Class 5 are: the station hub features are moderate. The bus connection is relatively convenient, and the passenger volume is relatively large. And the stations are far apart. Therefore, such stations are defined as transportation connection stations.

The characteristics of the stations in Class 6 are: the station pivot feature is weak. The surrounding land has a high development intensity and a high degree of mixed construction, of which residential land accounts for a large proportion. Therefore, such stations are defined as urban residential stations.

The characteristics of the stations in Class 7 are: the characteristics of the station hub are moderate. And the passenger volume is not large. The surrounding land development intensity and building mix are high. The proportion of various types of land is relatively balanced. There are no obvious characteristics in all aspects. Therefore, such stations are defined as urban general stations.

4. CONCLUSION AND DISCUSSION

Taking Hangzhou Metro System as an example, in this paper we calculate the feature values of 5 initial indicators. The principal component analysis method is applied to carry out data dimensionality reduction on the feature values, and K-means clustering algorithm and three spectral clustering algorithms are selected to classify the stations. The stations are divided into 7 categories, namely the comprehensive transportation hub stations, urban center stations, urban sub-center stations, suburban stations, transportation connection stations, urban residential stations, and urban general stations. The results show that the model proposed in this paper can effectively identify the types of stations. The method is significant for the research of the different types of station passenger flow prediction and could provide useful reference for the urban rail transit management.

AUTHORS' CONTRIBUTIONS

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

REFERENCES

[1] Wang Y, Jin Y, Wu B. Definition and classification method of rail transit stations under TOD[J]. Traffic and Transportation, 2021, 37(02): 51-56.

[2] Tang L, Xu X. Optimization for operation scheme of express and local trains in suburban rail transit lines based on station classification and bi-level programming[J]. Journal of Rail Transport Planning & Management, 2022, 21: 100283.

[3] Xia X, Gai J. Classification of Urban Rail Transit Stations and Points and Analysis of Passenger Flow Characteristics based on K-Means Clustering Algorithm[J]. Modern Urban Transit, 2021(04): 112-118.

[4] Zhang L, Meng B, Yin Q. Classification of urban rail transit stations based on SAX[J]. Journal of Geo-information Science, 2016, 18(12):1597-1607.

[5] Li W, Zhou M, Dong H. Classifications of stations in urban rail transit based on the two-step cluster[J]. INTELLIGENT AUTOMATION AND SOFT COMPUTING, 2020, 26(3): 531-538.

[6] Li S, Peng J, Wu Z, et al. Exploring the Relationship between Urban Rail Transit & Land Use and Their Quantitative Measurement Model: A Case Study of Guangzhou[J]. Journal of Guangzhou University (Natural Science Edition), 2016, 15(03): 63-69+2.

[7] Deng P, Zheng C, Ma G, et al. Classification of Rail Transit Stations based on AFC Data Mining[J]. Journal of East China Jiaotong University, 2019, 36(02): 77-82.

[8] Rao C, Gao X. Study on the Land Development around Metro Stations by Typing: A Case Study of Hangzhou Metro Line 1[J]. Journal of Zhejiang University (Science Edition), 2020, 47(02): 231-243.

[9] Xu W, Zheng C, Ma G, et al. Urban Rail Transit Site Classification based on k-means Clustering[J]. Journal of Guizhou University (Natural Sciences), 2018, 35(06): 106-111.

[10] Angulakshmi M, Priya G G L. Walsh Hadamard transform for simple linear iterative clustering (SLIC) superpixel based spectral clustering of multimodal MRI brain tumor segmentation[J]. Irbm, 2019, 40(5): 253-262.

[11] Duan J, Chen L, Chen C L P. Multifocus image fusion with enhanced linear spectral clustering and fast depth map estimation[J]. Neurocomputing, 2018, 318: 43-54.

[12] Hu X, Yang J, Wei F, Ma S. Clustering analysis of natural products derived from Polygonum multiflorum Thunb. based on spectral clustering algorithm[J]. Chinese Journal of Pharmacovigilance, 2022, 19(04): 390-394.

[13] Yao Z, Gao G, Zheng H, et al. Time distribution types of passenger flow between urban rail transit stations based on spectral clustering[J]. Urban rapid rail transit, 2022, 35(2): 99-104.

[14] Gao X, Ni C. Logistics Development Quality Evaluation based on Grey Clustering Analysis[J]. Railway Transport and Economy, 2018, 40(04): 23-29.

[15] Zhao D, Weng F, Ma S. A Railway Passenger Service Quantity Evaluation based on Comprehensive Principle Components Analysis[J]. Railway Transport and Economy, 2020, 42(03): 18-23.

[16] Yu L, Li Y, Chen K. Using Spectral Clustering for Urban Rail Station Classification[J]. Journal of Transport Information and Safety, 2014, 32(01): 122-125+129.

[17] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[C]//Advances in neural information processing systems. 2004: 321-328.

[18] Bholowalia P, Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN[J]. International Journal of Computer Applications, 2014, 105(9).

[19] Aranganayagi S, Thangavel K. Clustering categorical data using silhouette coefficient

as a relocating measure[C]//International conference on computational intelligence and multimedia applications (ICCIMA 2007). IEEE, 2007, 2: 13-17.

[20] Meila M, Xu L. Multiway cuts and spectral clustering[Z]. University of Washington Technical Report, 2003: 442-447.

[21] Verma D, Meila M. A comparison of spectral clustering algorithms[J]. University of Washington Tech Rep UWCSE030501, 2003, 1: 1-18.

[22] Ng A Y, Jordan M I, Weiss Y. On spectral clustering analysis and an algorithm[C]. Advances in Neural Information Processing Systems. 2002: 849-856.

[23] Meila M, Shi J. Learning segmentation by random walks[J]. Advances in Neural Information Processing Systems, 2000, 13: 873-879.