# Research on the Passenger Flow Distribution Mining of Urban Rail Transit Based on Clustering Algorithm: Taking Beijing as an Instance

Jiageng Ma[1, a], Bingxin Li[2, *], Yingqi Liu[3, b]

[1]Changchun University, School of Science, Changchun, China

[2]Minzu University of China, College of Science, Beijing, China

[3]China University of Mining and Technology, College of mathematics, China

*Corresponding Author Email: 1582841688@qq.com, [a] lixj8804@sohu.com, [b] 1932791790@qq.com

**Abstract.** Based on the principal component analysis and k-means cluster analysis, the paper classified and analyzed the 310 stations of the Beijing subway. According to data such as the number of people entering and leaving each station of the Beijing Metro, the line to which the station belongs, and whether it is a transfer station by period. We selected a total of 52-time variables to participate in the PCA analysis and finally got two principal components and named them principal component 1 and principal component 2. Subsequently, we clustered the Principal Component 1 and Principal Component 2 scores of 310 sites and the latitude and longitude of the sites in k-means clustering. According to the elbow method and the contour coefficient, the final clustering results into five clusters are classified. The five clusters of stations have strong regional characteristics, and the stations of each type are more concentrated. The results of principal component analysis and cluster analysis have passed the test and have good convincing power. Finally, we conduct an in-depth analysis of the two categories with the most stations. Cluster 4 contains the largest number of stations, located in the core area of Beijing's urban area, and the passenger flow is significantly higher than other areas. Cluster 1 contains the second largest number of stations, mainly including Fengtai District. The region's economy is underdeveloped, the passenger flow is small, and there are peak and trough periods. This paper performs cluster analysis on high-dimensional data. The principal component analysis is used to reduce the dimension of high-dimensional data and retain the time and space dimensions required for clustering. The final result was in line with expectations. It is improved the shortcomings of k-means clustering for high-dimensional data. This article analyzes the daily passenger flow in Beijing, and the analysis results may be random. In the subsequent analysis, more data can be added to make the analysis more accurate.

**Keywords:** Passenger flow distribution, Urban rail transit, Clustering algorithm, Data mining.

## 1. Introduction

The Beijing station is China's first rail transit line network to be constructed and officially put into operation. Since it is officially opened in 1971, the rail lines have been expanded. The number of subway stations has been growing. As of 2021, Beijing has more than 300 subway stations. As the capital of China and the first-tier city globally, its traffic quality has been widely paid attention by us. Therefore, we consider studying the passenger flow distribution of Beijing urban rail transit networks subway stations based on a high-dimensional clustering algorithm.

We have obtained the Beijing train operation map data and the O-D data of line network passengers (including only the public travel between 12:00 in each subway station).In this regard, we propose discussing each station's passenger flow distribution from the following dimensions: 1. Time dimension: the beginning of each station, the number of stops and exits every two minutes; 2. Space dimension: the coordinates of each station on the two-dimensional plane and the location distribution of the living zone near the subway station's point;3. spatial and temporal dimension: the relationship between the passenger flows distribution location and quantity of each station over time. Since there are more than 300 subway stations in Beijing, we have thousands or tens of thousands of dimensions

with only a time dimension. Reducing the dimension and making the lost data as small as possible will be the first challenge we encounter. Thus, we consider dimension reduction before clustering analysis.

The dimension reduction method is divided into linear and nonlinear methods, including the clustering of retaining local and global features, all of which are more complex, regardless of this method. In the linear method, the PCA method is general, and more adapted to the problem we discussed, so we chose the PCA dimension reduction method.

Cluster analysis refers to putting objects with the same properties and properties in a unified set grouped into a large class to study their common properties. Because objects of a class have similarities, saving complicated computational processes and drawing general conclusions faster. However, there are two main algorithms for clustering high-dimensional data clustering: subspace clustering (Subspace clustering) and similarity metric-based clustering (Similarity-Based Clustering). Since the problem we discuss is identifying classes with high similarity, one of the most common and simple methods is —K-means means using clustering based on the similarity metric. The K-means method has two disadvantages: 1. It must specify the number of clusters-K; 2. The initial cluster centre is random—the K-means method is a modified method applicable to this paper to obtain an improved K++ algorithm.

For simple calculation, this paper discusses the clustering analysis in the time dimension, condenses over 300 stations by steps, and obtains K(own assumptions) classes. The incorrect data was corrected, more perfect K categories were obtained, and the distribution characteristics of several main categories were obtained. Then, overall distribution characteristics were excavated for the passenger flow distribution of Beijing subway stations from the spatial and space-time dimensions.

## 2. Literature review

At present, experts at home and abroad have done a lot of research and application in both big data and rail transit, but for the combination of the two, using big data technology to conduct deeper data mining and research on rail transit is not comprehensive enough.

Wang and Zhao believed in their study that there is a long-term and sustainable interaction relationship between urban traffic and urban spatial evolution. They elaborated that the two realized common development and mutual promotion through the change of traffic accessibility[1]. Zhao, Yang, Liu obtained the residents along with the rail transit distribution and commuting characteristics data and analyzed the impact of the rail transit on urban living and employment space layout[2]. Zhang, li compare New York, London, and Tokyo for rail transit and urban area distribution show the example between different countries[3]. Pan tsunami and Ren analyzed the difference between different modes and proposed the interaction degree of node coupling and network coupling[4].

The above literature makes additional research and analysis on the distribution of rail transit and urban areas and discusses it from different aspects and perspectives. It is a problem for scholars at home and abroad to solve the deep combination of big data and data mining technology.

Zhang briefly summarized the way of urban rail transit data acquisition and analyzed the application of big data in urban rail transit networks[5]. In their study, Xu and others analyzed the characteristics of big data technology and the application prospects for the investigation into urban rail transit[6]. In order to further study its application prospects, Ran discussed the development trend of urban rail transit combined with the big data background and deeply analyzed the current situation and application of big data in the field of rail transit[7]. All the above scholars have a deep interest in the combination of urban rail transit and big data technology and show their bright development prospects and research space.

Ma used the key technology platform of big data in their research to build the architecture of railway big data applications and fully combine big data technology with the railway system[8]. Jiang explained that urban rail transit is unrealistic to rely only on human input to relieve the pressure in operation, management, and maintenance, and big data technology has broad application prospects

in rail operation and management[9]. Yu and Lv made full use of big data technology, using the density clustering method based on time series analysis and clustering algorithm[10].

The above literature combined big data technology with railway systems to promote the development of railway systems from different aspects. This paper also takes big data technology, based on a density clustering algorithm suitable for time series data, realizes the overall clustering of sites, completes the analysis of passenger flow distribution, and discusses its application to formulate daily and holiday passenger transport strategies.

## 3. Model and methodology

### 3.1. Principal Component Analysis(PCA)

PCA is a statistical analysis method of reducing the dimension of multiple variables affecting one thing to several major variables. The final selected principal components need to fully reflect the majority of the information, usually representing a linear combination of the original variables, and stipulating that the principal component variables are not related.

We selected PCA indicators as the number of stops every 15 minutes between 5:00-7:00 stations, stops every 5 minutes between 7:00-9:00 stations, the number of stops between 9:00-12:00 stations, the online station number, whether it is a medium transfer station (0-1 variable). The 52 temporal variables involved in the PCA analysis with data processing using python were expected to be in 2 - 3 principal components eventually. The main process is as follows.

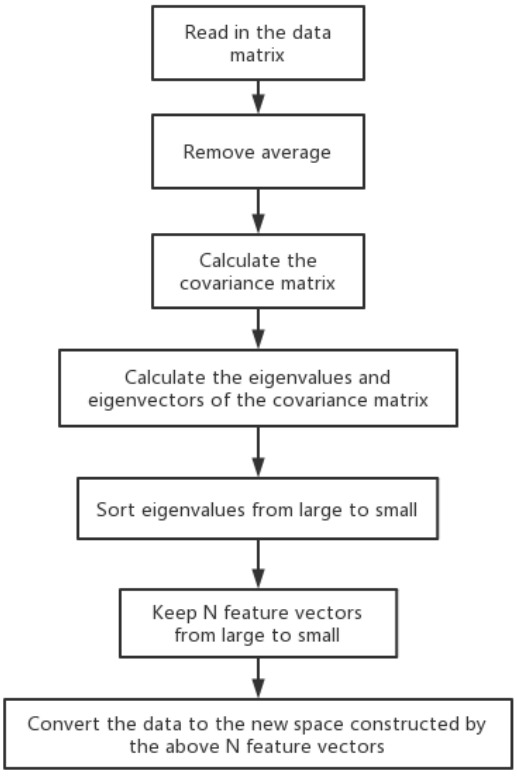



**Figure 1.** PCA flow chart

Choosing $k < p$ principal components to represent the original $p$ variables, only a small amount of information is lost. Then we have 2 assumes as follows.1. The principal component is only associated with the covariance array and assumes a strong correlation between the variables. 2. further discusses the no need for multivariate normality assumptions. Then we define formula (1) form matrix as shown in Table 1:

$$x = (x_1, x_2, \cdots, x_p)' \quad E(x) = \mu \quad V(x) = \Sigma \tag{1}$$

**Table 1.** variable matrix

| | *Variable*1 | *Variable*2 | ⋯ | *Variablek* | ⋯ | *Variablep* |
|---|---|---|---|---|---|---|
| *Item*1: | $x_{11}$ | $x_{12}$ | ⋯ | $x_{1k}$ | ⋯ | $x_{1p}$ |
| *Item*2: | $x_{21}$ | $x_{22}$ | ⋯ | $x_{2k}$ | ⋯ | $x_{2p}$ |
| | ⋮ | ⋮ | | ⋮ | | ⋮ |
| *Itemj*: | $x_{j1}$ | $x_{j2}$ | ⋯ | $x_{jk}$ | ⋯ | $x_{jp}$ |
| | ⋮ | ⋮ | | ⋮ | | ⋮ |
| *Itemn*: | $x_{n1}$ | $x_{n2}$ | ⋯ | $x_{nk}$ | ⋯ | $x_{np}$ |

Then we consider the linear transformation as following formulas (2) and (3):

$$\begin{aligned} Y_1 &= a_1^{'}X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= a_2^{'}X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Y_p &= a_p^{'}X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{2}$$

$$Var(Y_i) = a_i^{'}\sum a_i \quad i = 1,2,\cdots,p$$

$$Cov(Y_i, Y_k) = a_i^{'}\sum a_k \;\; i,k = 1,2,\cdots,p \tag{3}$$

First of all, under $\|a_1\| = 1$ constraint, we find the vector $a_1$ that makes the $Var(Y_1)$ reach the maximum. Then, under $\|a_2\| = 1$ and $Cov(Y_1, Y_2) = 0$ constraints, we find the vector $a_2$ that makes the $Var(Y_2)$ reach the maximum. $Y_2$ is the second main component.At the third step, under $\|a_1\| = 1$ and $Cov(Y_i, Y_k) = 0 (k < i)$ constraints, we find the vector $a_i$ that makes the $Var(Y_i)$ reach the maximum. $Y_i$ is the i main ingredient.

Set $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ as the eigenvalues. $t_1 = (t_{1i}, t_{2i}, \cdots, t_{pi})' \; (i = 1,2,\cdots,p)$ is the corresponding set of orthogonal unit eigenvectors, then the first principal component is:

$$y_1 = t_{11}x_1 + t_{21}x_2 + \cdots + t_{p1}x_p = t_1^{'}x \tag{4}$$

$t_1$ is the substitute of $a_1$. $\lambda_1^2$ is the variance of $Y_1$ .Solving process of the second principal component is the same as above. The relation between the principal component vector and the original vector:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_4 \end{pmatrix} = \begin{pmatrix} t_1^{'}x \\ t_2^{'}x \\ \vdots \\ t_p^{'}x \end{pmatrix} = \begin{pmatrix} t_1^{'} \\ t_2^{'} \\ \vdots \\ t_p^{'} \end{pmatrix} x = T'x \tag{5}$$

Among them $T = (t_1, t_2, \cdots, t_p) = (t_{ik})$ is the orthogonal matrix.

### 3.2. K-means++ Algorithm

K-means method has several disadvantages: 1. Must specify the number of cluster clusters K;2. the initial cluster centre is random. Moreover, we could improve the K-means method in the following aspects:

**(1) For disadvantage 1 (cluster number K): Improve algorithm 1 core idea**

The algorithm can adjust the cluster centre number K: with two operations according to the actual situation of each category. The merge operation corresponds to reducing the number of cluster centres. Split operations, corresponding to increasing the number of cluster centres. First, we give input to the algorithm (input data and iterations are no longer introduced separately):

1. Expected number of cluster centres Ko: Although the number of cluster centres is variable during operation, the user still needs to specify a reference standard. In fact, the variation range of the number of cluster centres of this algorithm is also determined by Ko. Specifically, the final output range of the number of cluster centres is [Ko/2, 2Ko].

2. The minimum number of samples required by each class Nmin: used to judge whether the splitting operation can be performed when the samples contained in a certain class are highly dispersed. If the number of samples contained in a subcategory is less than Nmin after splitting, the splitting operation will not be performed on the category.

3. Maximum variance Sigma: used to measure the degree of dispersion of samples in a category. When the degree of dispersion of the sample exceeds this value, it is possible to perform a split operation (note that the conditions described in 2. must be met simultaneously).

4. The minimum allowable distance dmin between the two categories corresponding to the cluster centres: if the two categories are very close (that is, the distance between the two categories corresponding to the cluster centres is very small), the two categories need to be merged operate. The threshold of whether to merge is determined by dmin.

First, calculate the distance between the cluster centres of all current categories, represented by a matrix $D$ . $D(i,i)=0$ .Then the two classes( $D(i,j)<d_{\min}(i\neq j)$ ) need to be merged into a new class that's cluster centre location is: $m_{new}=\frac{1}{n_i+n_j}(n_i m_i+n_j m_j)$ .

$n_i$ and $n_j$ express the number of samples in these two categories. The above formula represents the number of samples in these two categories. The new cluster centre can be seen as a weighted summation of these two categories. If one of the classes contains a larger number of samples, the resultant new class is more biased.

Then the variance was calculated under each dimension and category for all samples, and the maximum variance was selected for all variances in each category $\sigma_{\max}$ .

**(2) Improve algorithm 2 for disadvantage 2 (the initial cluster centre is random)**

That is, the K initial clustering centres should be separated from each other as much as possible. Combined with the above two steps, we have the final method as shown in Figure 3.

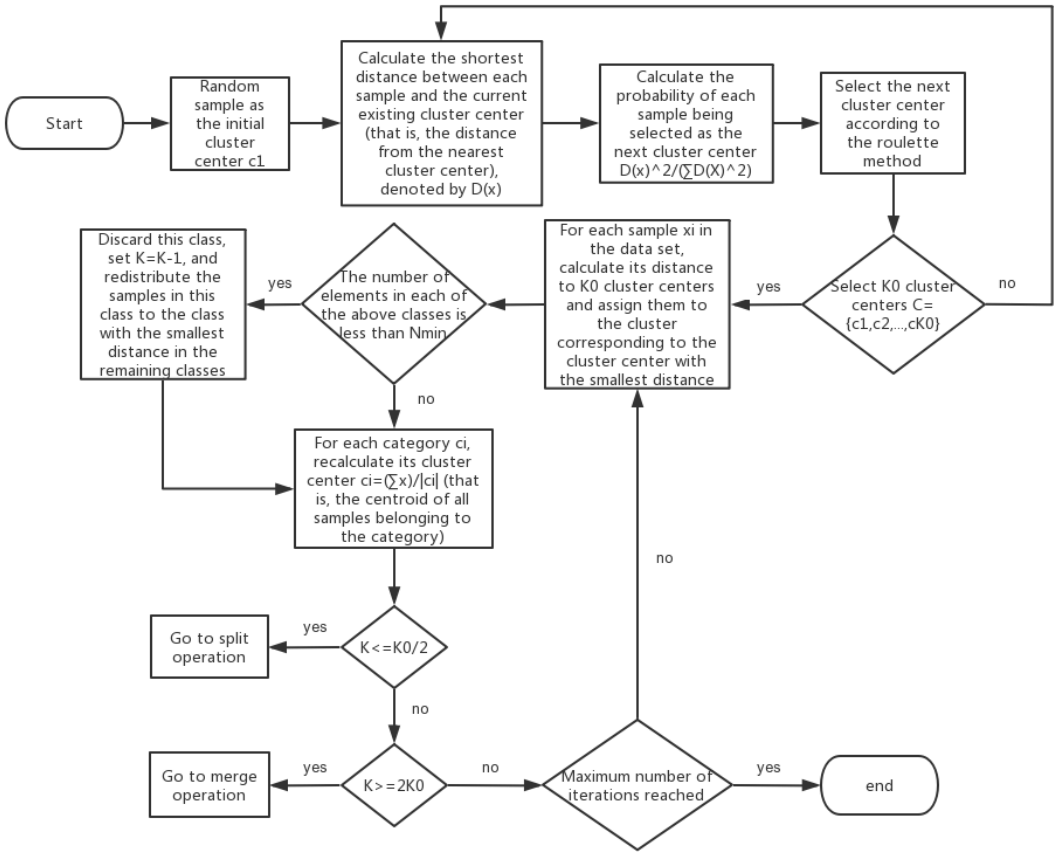



**Figure 2.** The k-means algorithm for combining the two

### 3.3. Elbow rule and contour coefficient

The calculation principle of elbow law is the cost function. The cost function is the sum of the degree of category distortion. If the members inside the class are more compact with each other, the distortion degree of the class is smaller. This paper defines the cost function by the sum of the squares of residuals (SSE). The specific functions are as follows:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{m^i} \left[ (x_j^i - x^i)^2 + (y_j^i - y^i)^2 + (\alpha_j^i - \alpha^i)^2 + (\beta_j^i - \beta^i)^2 \right] \cdots\cdots \quad (6)$$

Where, $k$ is the number of clusters, and the value is 1, 2, 3 $\cdots$ n(n is the number of original samples). After determining the value $k$, the improved $k-means$ algorithm is used to cluster it, and $k$ different clusters are obtained. The number of samples in the ith cluster is $m^i$. $x, y, \alpha$ and $\beta$ respectively represents the four dimensions after PCA dimensionality reduction. Taking the $x$ dimension as an example $x_j^i$ represents the $x$ dimension value of the $j$-th sample in the $i$-th cluster and $x^i$ represents the central sample value of the $x$ dimension of the $i$-th cluster.

Contour coefficient is an evaluation method of clustering effect, which combines cohesion and separation. According to the intracluster dissimilarity $a(j)$ and inter-cluster dissimilarity $b(j)$ of the sample $j$, the contour coefficient $S(j)$ of the sample $j$ is defined. If $S(j)$ is close to 1, it shows that the clustering of the sample $j$ is reasonable; The closer $S(j)$ is to -1 , it indicates that the sample $j$ clustering is unreasonable and should be classified into another cluster. If $S(j)$ is close to 0, it means that the sample $j$ is on the boundary of two clusters. The specific expression is as follows:

$$S(j)=\frac{b(j)-a(j)}{\max\{a(j),b(j)\}} \cdots\cdots (7)$$

$$a(j)=\sum_{q=1}^{m^{i_j}-1}\frac{\sqrt{(x_j^{i_j}-x_q^{i_j})^2+(y_j^{i_j}-y_q^{i_j})^2+(\alpha_j^{i_j}-\alpha_q^{i_j})^2+(\beta_j^{i_j}-\beta_q^{i_j})^2}}{m^i-1} \cdots\cdots (8)$$

$$b(j)=\min_{i\neq i_j}\left\{\sum_{q=1}^{m^{i_j}}\frac{\sqrt{(x_j^{i_j}-x_q^{i})^2+(y_j^{i_j}-y_q^{i})^2+(\alpha_j^{i_j}-\alpha_q^{i})^2+(\beta_j^{i_j}-\beta_q^{i})^2}}{m^i}\right\} \cdots\cdots (9)$$

$$S=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m^{i_j}}S(j)}{n} \cdots\cdots (10)$$

Where $i_j$ represents the $i$-th cluster, to which belongs the $j$-th sample, and $m^{i_j}$ represents the total number of samples of the $i$-th cluster, which belongs to the $j$-th sample. Finally, the average contour coefficient $S$ under the cluster number $k$ is obtained, and the contour coefficient principle judges the clustering effect.

## 4. Results Analysis

As the capital of China, Beijing has a relatively dense flow of people, and the operation of its subway traffic has also attracted much attention. In order to analyze the operation of the Beijing subway, we found the O-D data of the entry and exit gates of 297 subway stations in Beijing between 0:00-12:00 in a certain month in Beijing. The data contains 1.0486 million pieces of data, that is, the time information of 1.0486 million passengers entering and exiting the gate, as well as the entry and exit of the passenger. We split the data from 5:00 to 12:00 in the morning and study the number of people entering each station at regular intervals, using this as the traffic data of the station, and then adding whether it is a transfer station and a total of 52 lines Dimension, PCA dimensionality reduction is performed.

### 4.1. PCA experiment results

PCA reduces highly correlated multiple data features to several main and uncorrelated composite variables. The left in Figure 4 is the gravel diagram of the eigenvalues of all components. We summarize the component with eigenvalues greater than one and draw the upper right Figure. The two components with the largest eigenvalues are defined as the principal component. Their cumulative variance contribution rate is greater than 85%.

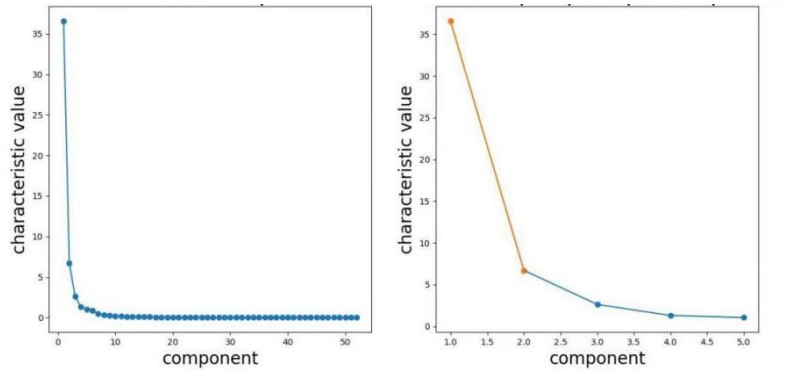



**Figure 3.** Gravel diagram of principal component analysis

We reduce the dimension of 51 data features into 2 principal components. We named it principal lpp (Number of people pitting every 15 minutes during low point period) and pp (Number of people pitting every 5 minutes during peak period). As shown in the Figure above, we drew the scatter plot of 310 sites about the two principal components. The scatter diagram shows that most stations are

concentrated in the lower-left corner of the image, but many stations are distributed in other image areas. The image shows no obvious correlation between the two principal components, which proves that it meets the expectation of principal component analysis.[11-18]

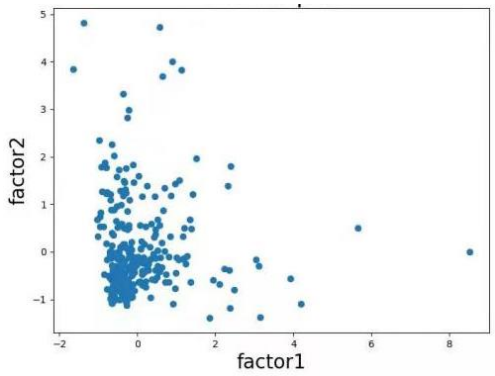



**Figure 4.** scatter plot of two principal components

According to principal component analysis, we reduce the dimension of 52 data features into two principal components. We named it lpp (Number of people pitting every 15 minutes during low point period) and pp (Number of people pitting every 5 minutes during peak period). As shown in the Figure above, we drew the scatter plot of 310 sites about the two principal components. The scatter diagram shows that most stations are concentrated in the lower-left corner of the image, but many stations are distributed in other image areas. The image shows no obvious correlation between the two principal components, which proves that it meets the expectation of principal component analysis.

After 52 dimensions are reduced, the p-value is less than 0.05 according to Barrett's sphere test, and the information of the original data contained in the two principal components is as high as 84.7%, indicating that these data are suitable for the principal component analysis.

### 4.2. K-means model simulation results

First, we tested the correlation of the four dimensions of the cluster analysis.

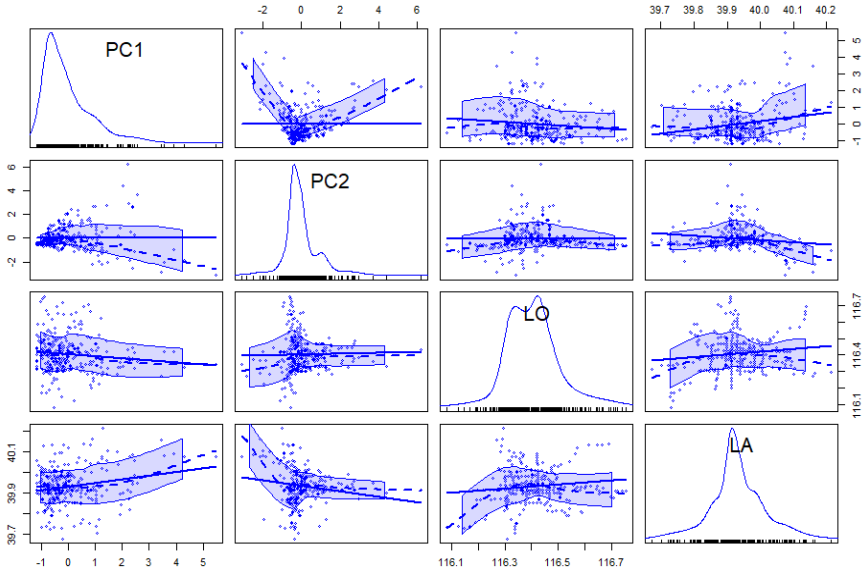



**Figure 5.** scatter plot matrix

In order to test the correlation of the four dimensions of clustering, the scatter matrix is drawn and shown in the Figure. At the same time, we fit a straight line for each scatter diagram to represent the correlation between the two dimensions. From the image, the correlation between each dimension is not obvious. The cost function of the different clusters is calculated to obtain the diagram of the cost function. As shown in the Figure above, the degree of distortion (cost function) is greatly improved, so consider selection as the number of clusters k=5.

### 4.3. Visualization of the clustering results

The following Figure shows the cluster heatmap explaining the clustering results. K-means clustering standardizes the four dimensions of each station to obtain a total feature value of each station and assigns each station to the nearest cluster centre by the size of the total feature value. The middle part of the clustering heatmap is filled with different colours, and the colours represent the standardized value of each site in the four dimensions. Darker colours have negative values, and lighter colours have positive values. The left part of the clustering heatmap is a dendrogram of the cluster population, which shows the clustering process of each station.

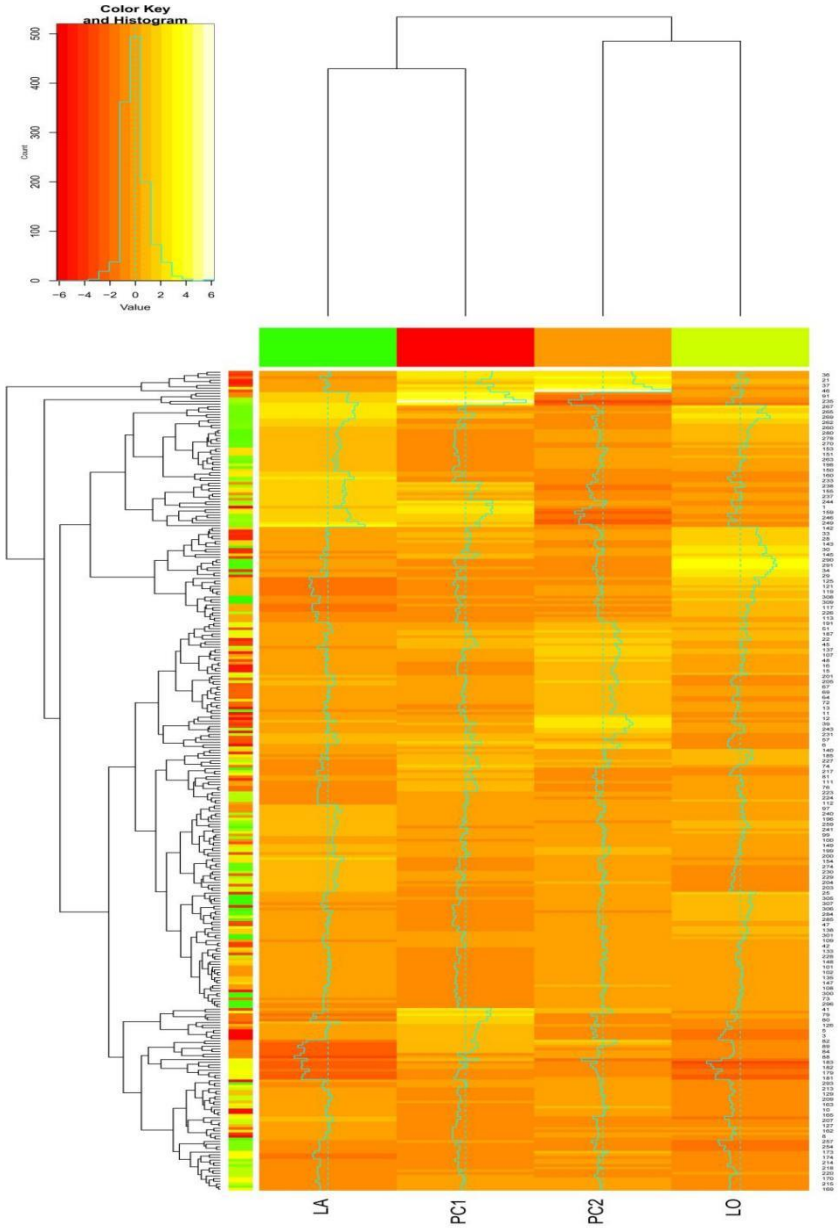



**Figure 6.** cluster heatmap of the results

In order to better show the clustering process of various clusters, we represent the dendrogram of each cluster of station separately, as shown in the following Figure:

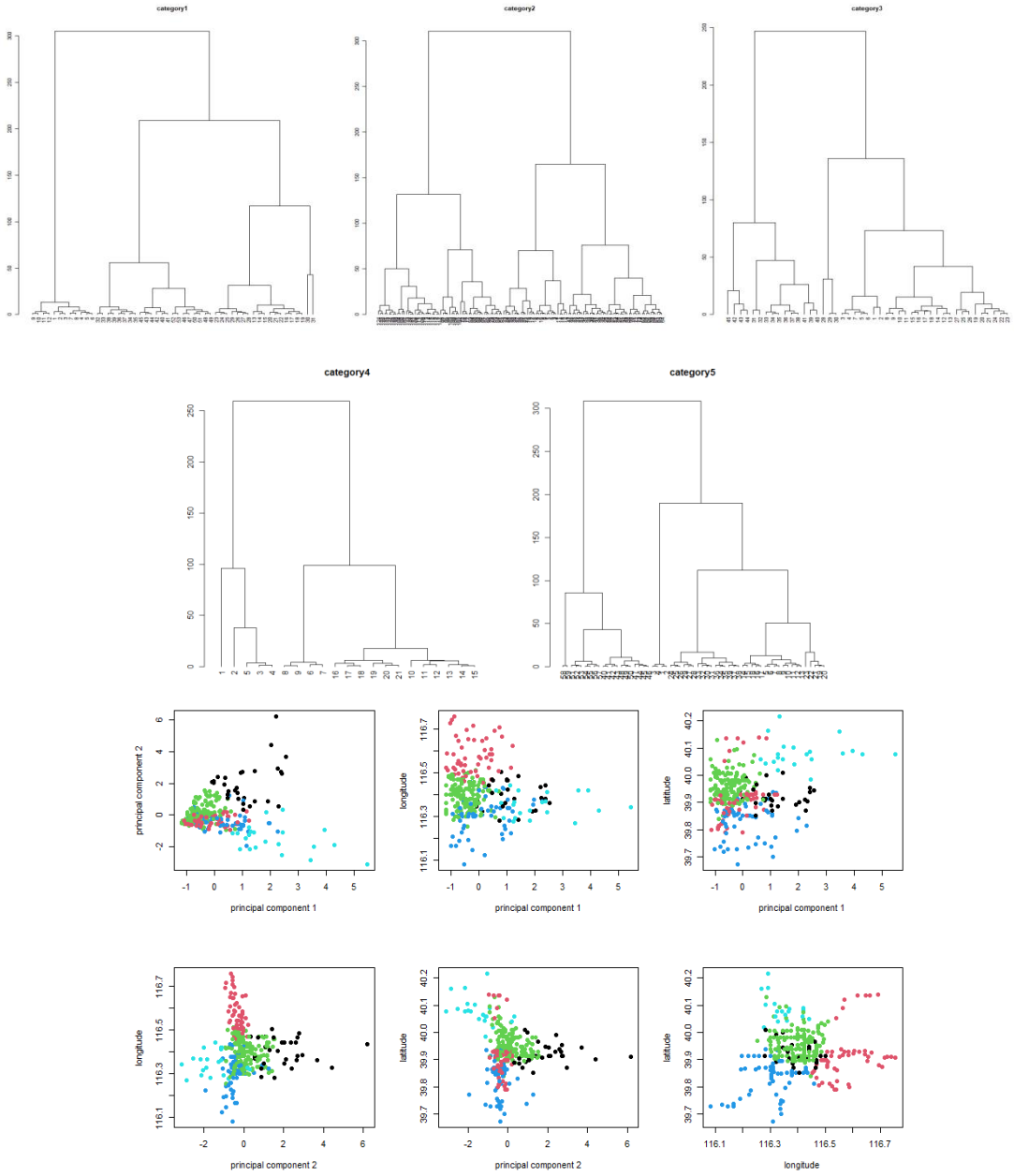



**Figure 7.** 2-dimensional feature scatter plot

Through the dendrogram of various stations, the station numbers contained in each cluster can be clearly indicated. The above Figure is a two-dimensional feature scatter plot between each dimension, with different clusters representing different clusters.

According to the clustering results, the cluster with the second largest number of sites is mainly in the southwest corner of Beijing, namely Fengtai District. A separate analysis of Fengtai District shows that the economy of Fengtai District is not as developed as that of the city centre. The passenger flow of the station is relatively large in the morning peak, but the passenger flow is small at other times. Our suggestion is to make the stations in the area more dispersed, increase the number of trains, and reduce the time interval between each subway trip.

## 5. Conclusions

Through the above analysis, it is understood that the characteristics of various stations are quite different. The Principal Component 1 scores of Cluster 2 and 3 stations are high, and the other three clusters are concentrated near 0. At the same time, the scores of principal component 2 of Cluster 3 stations are significantly higher.

Cluster 4 contains the largest number of stations among the five clusters accounting for 48.06%. It is mainly distributed within the Fifth Ring Road of Beijing, including Dongcheng, Xicheng, Chaoyang, and Haidian. Therefore, this area is the centre of Beijing's urban area, where very high medical care, education, and commerce. The scores of principal component 1 and principal component of such stations are also relatively concentrated, and the scores are both low. The stations in Cluster 4 had a large passenger flow. Especially after nine o'clock, the passenger flow of other clusters of stations has a clear downward trend, and the passenger flow of the four clusters of stations has a small decrease. There are large shopping malls and commercial streets near these stations, so it is necessary to receive many passengers every day. At the same time, there are more transfer stations in Cluster 4. Passengers need to transfer at stations and spend more time.

Cluster 1 is mainly distributed on the southwest side of the urban area, mainly including Fengtai District. The development of Fengtai District has been slow in the past ten years. There are fewer schools and large companies in Fengtai District than in Dongcheng, Haidian, and other areas. The traffic is not very developed, and there are many residential areas. On the contrary, there are few commercial streets. The characteristic of these clusters of stations is that the traffic volume of stations varies greatly over time. Only in the morning peak period is the passenger flow is large, and the passenger flow in other periods is relatively small. Cluster 1 contains the second-highest number of sites among the five categories.

This paper performs cluster analysis on high-dimensional data to analyze the daily passenger flow in Beijing. PCA is used to reduce the dimension of high-dimensional data and retain the time and space dimensions required for clustering. The results of the analysis may be random. In the subsequent analysis, more data can be added to make the analysis more accurate.

## References

[1] Wang Chuncai. Study on the Interaction Mechanism between Urban Traffic and Urban Space Development [D]. Beijing Jiao Tong University, 2007.

[2] Zhao Hui, Yang Jun, Liu Changping. takes Beijing as an example [J].Economic geography, 2011,31 (09): 1445-1451.

[3] Zhang Xiaodong, Li Shuang. World Urban Transportation System Comparison Study —— takes New York, London and Tokyo [J]. Beijing planning and construction, 2010 (06): 30-34.

[4] Pan Tsunami, Ren Chunyang. Spatial coupling relationship between rail transit and urban public activity center system takes Shanghai as an example [J]. Urban Planning Journal, 2005 (04): 76-82.

[5] Zhang Fina. Development trend of Urban rail transit in the context of big data [J]. The World of Digital Communications, 2020 (03): 149.

[6] Xu Yuping, Qin Gong, Zhang Zheng.The Big Data Application Study of the Urban Rail Transit Survey [J]. Railway Transport and Economy, 2015,37 (04): 78-81.

[7] Ran Jian. Development trend of Urban rail transit in the context of big data [J]. Low Carbon World, 2019,9 (08): 259-260.

[8] Ma Limei, Dan, Gao Zhiyuan, Li Huajie.Big Data Technology and Its Industry Application: Concept Framework Research Based on Railway Field [J]. Journal of Beijing Jiao Tong University (Social Science Edition), 2019,18 (03): 58-67.

[9] Jiang Ziwang. The Application of Big Data Technology in Urban Rail Transit Operation and Management [J]. Technological Innovation and Application, 2020 (05): 174-175.

[10] Yu Chengcheng, Lv Hongxia. Research on Passenger Marketing Strategy Based on Big Data Railway Passenger Portrait [J]. Journal of Railways, 2020,42 (08): 23-28.

[11] Huertas J A, Palacio A, Botero M, et al. Level of traffic stress-based classification: A clustering approach for Bogotá, Colombia[J]. Transportation Research Part D: Transport and Environment, 2020, 85: 102420.

[12] Sfyridis A, Agnolucci P. Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling[J]. Journal of Transport Geography, 2020, 83: 102658.

[13] Alisoltani N, Zargayouna M, Leclercq L. A sequential clustering method for the taxi-dispatching problem considering traffic dynamics[J]. IEEE Intelligent Transportation Systems Magazine, 2020, 12(4): 169-181.

[14] Ryu U, Wang J, Pak U, et al. A clustering based traffic flow prediction method with dynamic spatiotemporal correlation analysis[J]. Transportation, 2021: 1-38.

[15] Jin J, Wang W. Influential features PCA for high dimensional clustering[J]. The Annals of Statistics, 2016, 44(6): 2323-2359.

[16] Dey S, Das S, Mallipeddi R. The Sparse MinMax k-Means Algorithm for High-Dimensional Clustering [C]// IJCAI. 2020: 2103-2110.

[17] Dey S. Sparse min-max k-means algorithm for high-dimensional data clustering [J]. 2018.

[18] Paul D, Chakraborty S, Li D, et al. Principal Ellipsoid Analysis (PEA): Efficient non-linear dimension reduction & clustering [J]. arXiv preprint arXiv: 2008.07110, 2020.