Statistical Research on the Influencing Factors of Zhengzhou Metro Passenger Flow

Xinna Liu¹,∗

¹School of transportation, Shanghai Maritime University
* Corresponding author: 202030610163@stu.shmtu.edu.cn

Abstract. Subway passenger flow refers to the number of people entering the subway per unit time, which is an important indicator reflecting the popularity and value of the subway. This paper selects Zhengzhou Metro Line 1, which is currently the most mature and has the most complete passenger flow data in Zhengzhou City. First, using SPSS software to perform principal component analysis on the original data. Second, using MATLAB software to further process the data processed by principal component analysis. Finally, through the establishment of a gray correlation model, the results of Lanzhai Station on Zhengzhou Metro Line 1 are obtained that its import and export passenger flow has the most significant impact on the passenger flow of Zhengzhou Metro Line 1.

1. Introduction

With the steady development of China's economy and the continuous improvement of urban rail transit infrastructure, the subway transportation system has become a very important part of the rail transit system. The subway is a significant symbol of the integration of a city into the modern transportation of an international metropolis. It is not only a demonstration of the strength of a country's national strength and technological level, but also the most ideal transportation method to solve the traffic tension in a metropolis.

Therefore, a statistical study of the factors affecting the passenger flow of Zhengzhou Metro can be carried out on the Zhengzhou Metro Line 1 to understand the relevant factors affecting the passenger flow of Zhengzhou Metro and to plan the subway lines of Zhengzhou more scientifically.

2. Journals reviewed

Principal component analysis [1] relies on the data given, and the accuracy of the data has a great influence on the results of the analysis. The grey forecasting method [2] is based on the index rate and cannot consider the randomness of the system. The time series method [3] in the linear forecasting method, when encountering large changes in the outside world, such as changes in national policies, forecasts based on data that have occurred in the past, there will often be large deviations. The Kalman filter model [4] and the traffic congestion prediction based on the improved support vector machine model [5] have the disadvantage of relying too much on historical data or the amount of calculation is particularly cumbersome and complicated. Based on the method of principal component analysis, this paper proposes a grey relational prediction model for rail transit passenger flow. First, use SPSS software to perform principal component analysis on the original data, and then use matlab software to further analyze the data, so as to get the main sites that affect the passenger flow.
3. Theory and method

Step 1. Standardization of indicator data: Select the passenger flow data of the 15 stations of Zhengzhou Metro Line 1 from 7 am to 23:00 on a certain day. Calculate the difference between the number of importers and exporters at each site every two hours to get relevant data. Import the data into the SPSS software, and get as shown in Table 1:

Table 1. Data standardization

|   | X1  | X2  | X3  | X4  | X5  | X6  | X7  | X8  | X9  | X10 | X11 | X12 | X13  | X14  | X15  |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|
|   | 47.00 | -76.00 | 172.00 | 104.00 | 58.00 | -218.00 | -3.00 | 277.00 | 492.00 | -12.00 | -5.00 |
|   | -2.00  | 54.00 | 33.00 | -88.00 | 242.00 | -15.00 | -108.00 | -41.00 | -339.00 | 228.00 | 113.00 |
|   | -83.00 | 67.00 | 30.00 | 66.00 | 524.00 | -109.00 | 387.00 | 101.00 | -259.00 | 529.00 | 80.00 |
|   | -242.00 | 187.00 | -101.00 | -425.00 | -337.00 | 245.00 | -699.00 | -195.00 | -41.00 | -130.00 | 1590.00 |
|   | 464.00 | -136.0 | 549.0 | -421.0 | 347.0 | 347.0 | 533.0 | -926.00 | 103.00 | -151.00 | 173.00 |
|   | 347.0 | -114.0 | 191.0 | -1078.0 | -182.0 | 209.0 | -244.0 | 228.00 | 2255.00 | -257.00 | -336.00 |
|   | -686.0 | -167.0 | -53.0 | -705.0 | -222.0 | -222.0 | 294.0 | -1027.00 | 677.00 | -2557.00 | -335.00 |

Step 2. Data processing: Taking the difference of the import and export flow of 15 stations as the independent variable, the following results are obtained through factor analysis:

Table 2. Common factor variance

|   | X1  | X2  | X3  | X4  | X5  | X6  | X7  | X8  | X9  | X10 | X11 | X12 | X13 | X14 | X15  |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| initial | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| extract | 0.926 | 0.944 | 0.928 | 0.965 | 0.949 | 0.822 | 0.917 | 0.958 | 0.710 | 0.904 | 0.678 | 0.926 | 0.925 | 0.902 |

It can be seen from Table 2 that the common variances of these 15 variables are all greater than 0.5, and the common variances of 14 variables exceed 0.8, so the four common factors extracted can well reflect the main information of the original variables.

Table 3. Explanation of total variance

| ingredient | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
|------------|----|----|----|----|----|----|----|
| Initial eigenvalue | 6.774 | 3.543 | 1.880 | 1.226 | .825 | .506 | 246 |
| Variance% | 45.161 | 23.618 | 12.535 | 8.176 | 5.487 | 3.375 | 1.637 |
| accumulation % | 45.161 | 68.779 | 81.314 | 89.491 | 94.958 | 98.363 | 100.000 |
| Extract the sum of squares of the load | 6.774 | 3.543 | 1.880 | 1.226 | |
| Variance% | 45.161 | 23.618 | 12.535 | 8.176 | |
| accumulation % | 45.161 | 68.779 | 81.314 | 89.491 | |

In the principal component analysis method, the greater the variance interpretation rate, the stronger the interpretation ability, the more it can reflect the key factors of the original variables, and the more effective the extracted principal components. According to Table 3, the initial eigenvalue accumulation rate of the first four components is 89.491 %, meet the research conditions.

third step. Determine the principal components: According to Table 3, the cumulative contribution rate of the first four independent variables reaches 89.491%, so we can choose these four subway stations as the subway stations for the study. The four subway stations are Henan University of Technology Station, Zhengzhou University Station, and Lanzhou University Station Civic Center Station.
4. Establishment and analysis of the grey relational model of Zhengzhou Metro Line 1 stations

4.1 Model establishment and solution

1. Symbol description
   \(X_0\) represents the reference data column, \(X_1\) represents the row vector of Hegong University Station, \(X_2\) represents the row vector of Zhengda Station, \(X_3\) represents the row vector of Lanzhai Station, \(X_4\) represents the row vector of the Civic Center Station, and \(p\) represents Resolution coefficient, \(\delta_i(k)\) represents the correlation coefficient, and \(r_i\) is the correlation order.

2. Model establishment and solution
   (1) Form the original data into a 4*8 matrix:
      \[
      X_1, X_2, X_3, X_4 = \begin{bmatrix}
      47.0 & -2.0 & -83 & 242 & 64.5 & 347 & -430 & -686 \\
      -76 & -54 & 67.0 & 187.0 & 14.0 & 136 & -114 & -167 \\
      -23 & 33 & 30.0 & -331 & 643 & 549 & 191 & -53 \\
      -172 & -288 & -66 & -142 & -201 & -421 & -1078 & -705 
      \end{bmatrix}
      \]
   (2) Determine the reference data column
      Select the total number of people of each site's import and export difference in this day as the reference data column and record it as \(X_0\). After calculation, the reference sequence \(X_0\) can be obtained:
      \[
      X_0 = (120, -311, 80, -811, 1101, 611, -1431, -1611)
      \]
   (3) Dimensionless of index data
      There are usually the following methods for dimensionless index data:
      \[
      X = (x(1), x(2), \ldots, x(n))
      \]
      a) Initial value transformation, namely
         \[
         XD_1 = (x(1)d_1, x(2)d_1, \ldots, x(n)d_1)
         \]
         among them:
         \[
         x(k)d_1 = x(k) / x(1), x(1) \neq 0, k = 1, 2, \ldots, n
         \]
      b) Average transformation, namely,
         \[
         XD_2 = (x(1)d_2, x(2)d_2, \ldots, x(n)d_2)
         \]
         among them:
         \[
         x(k)d_2 = x(k) / \bar{x}, \bar{x} \neq 0, k = 1, 2, \ldots, n
         \]
      c) Minimization transformation, namely,
         \[
         XD_3 = (x(1)d_3, x(2)d_3, \ldots, x(n)d_3)
         \]
         among them:
         \[
         x(k)d_3 = x(k) / M, M \neq 0, k = 1, 2, \ldots, n
         \]
      d) Maximum value change, namely,
         \[
         XD_4 = (x(1)d_4, x(2)d_4, \ldots, x(n)d_4)
         \]
         among them:
         \[
         x(k)d_4 = x(k) / m, m \neq 0, k = 1, 2, \ldots, n
         \]
      e) Range change, namely,
         \[
         XD_5 = (x(1)d_5, x(2)d_5, \ldots, x(n)d_5)
         \]
         among them:
         \[
         x(k)d_5 = (x(k) - m) / (M - m), M - m \neq 0, k = 1, 2, \ldots, n
         \]
      The dimensionless data sequence forms the following matrix:
      \[
      X_0, X_1', X_2', X_3', X_4' = \begin{bmatrix}
      84.44 & 60 & -74.44 & -207.78 & -15.56 & -151.11 & 126.67 & 185.56 \\
      0.18 & -0.25 & -0.23 & 2.55 & -4.95 & -4.22 & -1.47 & 0.41 \\
      -0.48 & 0.8 & -0.18 & 1.18 & 0.56 & 1.17 & 2.99 & 1.96 
      \end{bmatrix}
      \]
      (4) Calculate the absolute difference between each evaluated object index sequence (comparison sequence) and the corresponding element of the reference sequence reference one by one, namely,
      \[
      |x_0(k) - x_i(k)| (k = 1 \ldots 8, i = 1 \ldots 4),
      \]
      Get the following matrix:
X₁, X₂, X₃, X₄ = 84.87 58.9 75.16 210.67 11.65 148.89 121.66 179.84
0.61 1.35 0.05 0.34 1.04 2.00 6.48 5.31
0.05 0.3 0.1 1.71 4.47 −3.39 2.02 3.76

(5) Confirm
\[
\min_{i} \min_{k} |x_{0}(k) - x_{i}(k)| = 0.01 \quad (1)
\]
\[
\max_{i} \max_{k} |x_{0}(k) - x_{i}(k)| = 0.61 \quad (2)
\]

among them: k = 1 ⋯ 8, i = 1 ⋯ 4.

(6) Calculate the correlation coefficient

Incorporate formula (3) to calculate the correlation coefficients between each comparison sequence and the corresponding elements of the participating sequence.
\[
\delta_{i}(k) = \frac{\min_{i} \min_{k} |x_{0}(k) - x_{i}(k)| + \rho \times \max_{i} \max_{k} |x_{0}(k) - x_{i}(k)|}{|x_{0}(k) - x_{i}(k)| + \rho \times \max_{i} \max_{k} |x_{0}(k) - x_{i}(k)|} \quad (3)
\]

among them: k = 1 ⋯ 8, i = 1 ⋯ 4. ρ is the resolution coefficient, 0 < ρ < 1. If ρ is smaller, the difference between correlation coefficients is larger, and the distinguishing ability is stronger; usually ρ = 0.5.

The correlation coefficient calculated by MATLAB is:

δ = 0.40 0.43 0.64 0.91 0.55 0.57 0.35 0.72
0.46 0.40 0.64 1.00 0.64 0.60 0.34 0.59
0.51 0.60 0.84 0.94 0.84 0.91 0.62 0.34
0.51 0.49 0.71 0.86 0.84 0.78 0.36 0.65

(7) Calculate the correlation order:
\[
r_{i} = \frac{1}{8} \sum_{k=1}^{8} \delta_{i}(k) \quad (4)
\]

Incorporating into equation (4) is calculated as: r₁ = 0.52, r₂ = 0.56, r₃ = 0.75, r₄ = 0.66.: Among them: r₁ is the proportion of people flow in Henan University of Technology Station, r₂ is the proportion of people flow in Zhengzhou University Station, r₃ is the proportion of people flow in Lanzhai Station, and r₄ is the proportion of people flow in Civic Center Station.

4.2. Analysis of model results
From the above results, we can see that the proportions of people flow at Henan University of Technology Station, Zhengzhou University Station, Lanzhai Station, and Civic Center Station are 0.52, 0.56, 0.75, and 0.66 respectively. Among them, the change in passenger flow at Lanzhai Station is the most obvious.

5. conclusion
By studying the stations that affect the passenger flow of Zhengzhou Metro Line 1, firstly, the relevant departments of Zhengzhou City can be more scientific in planning the subway route; secondly, for the subway company, it is more reasonable to analyze which station has a greater impact on the passenger flow Arrange staff to improve the work efficiency of subway staff and make travel more convenient for passengers.

In the future, as an important part of the urban public rail transit system, the development of the subway will definitely become more and more scientific. Therefore, collecting more accurate and multi-angle data and conducting more in-depth research will be the next important aspect.

References
[1] Zhao Qiang. Summary of principal component analysis methods [J]. Software engineering. 2016(6):1-3.
[2] Xue Wei. Statistical analysis methods and applications [M]. Beijing: University of Electronic Technology Press, 2013.
[3] Xie Leilei, Song Zhigang, He Xuhong. SPSS statistical analysis tutorial [M]. Beijing: People's Posts and Telecommunications Press, 2014.
[4] MATTEO S, FABIO D. Robustness of LSTM neural networks for multi-step forecasting of chaotic time series [J]. Chaos, Solitons and Fractals, 2020, 139: 110045.
[5] BRESSEL M, HILAIRET M, HISSEL D, et al. Extended Kalman filter for prognostic of proton exchange membrane fuel cell [J]. Applied Energy, 2016, 164: 220-227.