TOPSIS method of roller coaster evaluation system based on Mahalanobis distance

Dengfeng Liu¹, Haixin Yu², Yaqi Zhu³

¹,²,³School of Information Science and Engineering, Northeastern University, Northeastern University, Nanhui Street, Heping District, Shenyang, Liaoning, China
2511778502@qq.com¹, 1437083606@qq.com², 1311771780@qq.com³

Abstract. Aiming at the problem of roller coaster evaluation, this paper puts forward a method of objective evaluation of roller coaster TOPSIS method based on Mahalanobis distance. The descriptive numerical information of the existing 5124 roller coasters in the world is crawled by using the network crawler technology. According to the TOPSIS algorithm, “The World’s Top Ten Roller Coaster” list is generated and the roller coaster ranking based on roller coaster speed, height, length and inversion number is obtained. By introducing two indexes of information coverage and information difference rate, the roller coaster rating system is compared and discussed horizontally and vertically.

1. Introduction:
There are several roller coaster scoring sites on the web, and although some objective measures have been taken into account, they rely heavily on subjective input to determine the ranking of a particular roller coaster. In solving the problem of objective evaluation of roller coasters, only the roller coaster currently in operation is considered, and the family roller coaster and children's roller coaster are not considered in this paper, and the sled Snow roller coaster and mountain roller coaster are not considered in this paper. Horizontally and vertically, the list of “The World's Top Ten Roller Coaster” is generated, which has a good reference value.

2. the establishment and solution of the model:

2.1. the establishment of roller coaster evaluation system based on Mahalanobis distance

2.1.1. Data pre processing

The First Step: Data acquisition
Before generating a list, you need to have data on all the running roller coasters in the world, such as the speed, height, length and material of the roller coaster. After collecting data and analysis, the crawler was written using basic python capture tools such as the Python language and urllib2, from the Roller Coaster database (https://www.rcdb.com) and the roller coaster and amusement park database (https://www.ultimaterollercoaster.com/) crawls the data[1].

The second step: the numerical characterization of descriptive data
A large part of the data obtained is descriptive data, which is presented in the form of text and cannot be quantitatively analysed. In order to prepare for the subsequent quantitative analysis, it is necessary to quantify the text data that will be used later in this article such as geographical location and materials.
(1) Numerical value of geographical location

The difference in geographical location represents the level of consumption in each region to a certain extent. People's spending budgets will affect their intuitive experience of roller coasters, so we need to digitize the geographic location. The GDP and Gini Coefficient of a region can represent the level of consumption in the region, so GDP and Gini Coefficient are used as a numerical geolocation data.

(2) Numerical of materials

The roller coaster track is divided into wood and steel. By looking for data, it can be known that steel has a higher safety factor than wood. If the steel roller coaster and wood roller coaster safety factor is set to 1, the steel roller coaster safety factor weights of 0.6, wood roller coaster safety factor weight of 0.4 of the weights[^1].

Table 1. example of the web crawler section.

| Country/Region | Construction | Year/Date Opened | Length (feet) | Number of Inversions |
|----------------|--------------|-----------------|---------------|----------------------|
| France         | Steel        | 2005            | 400           | 0                    |
| Spain          | Steel        | 2006            | 450           | 2                    |
| Australia      | Steel        | 2013            | 630           | 3                    |
| France         | Wood         | 1989            | 3937          | 0                    |

The Third Step: the filling of missing data

It is easy to conclude that the data is missing a part, which will have an impact on the later research. Through the analysis, it can be concluded that for the missing data, the mean value is filled. This kind of data processing is helpful to increase the rationality and validity of the later research.

The Fourth Step: Normalization of descriptive data

The data units and sizes given in the questions are very different, which will cause problems in the analysis of the following. So we normalize the numerical data and the numerical text data. This will bring great convenience to the later research.

2.1.2. Establishment of index system

The comprehensive evaluation index of roller coaster is too many and will be repeated, and it is difficult to find the corresponding data. According to the comprehensive and accurate principle, combined with the selection of the first problem design scheme, we use these five indexes to establish the roller coaster evaluation system.
Figure 1. Five evaluation indicators of roller coaster.

Introduction of each indicator[3].

1. Consumption level

Different groups of people represent different levels of consumption. When people start thinking about taking a roller coaster ride, they have a rough estimate of what it takes to ride a roller coaster in their hearts. The cost of riding a roller coaster in different regions is not the same, and there is even a big gap, which has a lot to do with the local consumption level. Therefore, the consumption level is selected as an index to objectively evaluate the roller coaster. The size of the index is related to the local GDP per capita and the Gini coefficient. The formula for calculating the indicator is as follows:

$$ZX = GDP \times \frac{1}{Gn}$$

2. Safety factors

The roller coaster is a thrilling amusement facility. There will be a lot of dangerous movements during the ride, so ensuring the safety of passengers is a factor that needs to be considered first. Therefore, the risk coefficient is selected as an index to objectively evaluate the roller coaster. After analysis, it can be found that the earlier the construction year of the roller coaster, the more serious the aging of the roller coaster, that is, the safety cannot be guaranteed. At the same time, the stronger the material, the higher the safety system. Therefore, the opening year and material of the roller coaster are important factors that determine the risk coefficient. The formula for calculating the indicator is as follows:

$$ZR = e^{c \times e^{-y}}$$

3. Comfort Level

People enjoy the pleasure of riding a roller coaster, and the comfort of the roller coaster will greatly affect people's mood. It is only when the roller coaster is highly comfortable that people can play more happily. Therefore, the comfort degree is chosen as an index to objectively evaluate the roller coaster. After analysis, it can be found that the length of the roller coaster ride has the greatest impact on the comfort of people on the roller coaster ride. The formula for calculating the indicator is as follows:

$$ZC = 1 + 0.2 \times du$$

Where $ZC$ indicates the value of comfort level, $l$ indicates the value of length of the roller coaster, and $du$ indicates the value of the duration of the roller coaster.

4. Stimulating factors
People who like or want to ride a roller coaster are hoping for a thrilling journey. If a roller coaster doesn't give people the thrill of adventure, the existence of this roller coaster will not be necessary. Therefore, the stimulus factor is selected as an index to objectively evaluate the roller coaster. After analysis, it can be found that the speed, G value, tilt degree, landing height, number of inversion and stimulating factor of roller coaster are proportional. The formula for calculating the indicator is as follows:

\[ ZT = sp \times gf \times va \times he \times iv \]

Which \( ZT \) indicates the value of factor thrill, \( sp \) indicates the speed of the roller coaster, \( gf \) indicates the roller's G Values, \( va \) indicates the value of the vertical angle of the roller coaster, \( he \) indicates the height value of the roller coaster, \( iv \) indicates the number of inversion times for the roller coaster.

5. Infrastructure

The infrastructure of a roller coaster can affect the passenger experience. Therefore, infrastructure is selected as an objective evaluation index of roller coaster. Roller coaster infrastructure indicators include the height, length and speed of the roller coaster. The formula for calculating this index is as follows:

\[ ZI = le \times sp \times he \]

\( ZI \) indicates the value of infrastructure, \( sp \) indicates the speed of roller coaster, \( le \) indicates the value of roller coaster length, and \( he \) indicates the height of roller coaster.

2.1.3. Determine indicator weights

Because of the correlation between the indexes, based on the principle of feasibility and accuracy, we use the combination empowerment method combined with the coefficient of variation method and the complex correlation coefficient method to judge the importance of the index\(^4\).

(1) Coefficient of variation method

If the actual value of an indicator can clearly distinguish the individual evaluated objects, the index is given a larger weight, and conversely, a smaller weight is given to the indicator, so it has a strong reliability\(^5\).

Suppose there is (are) \( N \) object (s) of evaluation, using the above indicators \( x_1, x_2, \ldots, x_i (i = 1, 2, \ldots, 5) \) describes each object. The mean \( \bar{X}_i \) and variance \( S_i^2 \) of each indicator are obtained first\(^6\).

\[
\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}
\]

\[
S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{X}_i)^2
\]

The coefficient of variation of the index is:

\[
V_i = \frac{S_i}{\bar{X}_i} (i = 1, 2, \ldots, 5)
\]

Normalize \( V_i \) and we define objective weights \( \omega_i \):

\[
\omega_i^{(1)} = \frac{V_i}{\sum_{i=1}^{5} V_i}
\]
Table 2. Variation Coefficient method Weight.

| index | ZX   | ZR   | ZC   | ZT   | ZI   |
|-------|------|------|------|------|------|
| Weights | 0.124 | 0.138 | 0.205 | 0.420 | 0.113 |

(2) Complex correlation coefficient method

When the information of an evaluation index and other indexes is repeated more, it is considered that the index plays less role in the comprehensive evaluation and the weight is smaller. The index weight is related to its independence.

First, the correlation coefficient matrix of the five evaluation indicators is:

\[
R = \begin{bmatrix}
    r_{11} & r_{12} & \cdots & r_{15} \\
    r_{21} & r_{22} & \cdots & r_{25} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{51} & r_{52} & \cdots & r_{55}
\end{bmatrix}
\]

Then, the complex correlation coefficient between a certain index and other indexes is calculated, and the correlation coefficient matrix is decomposed:

\[
R = \begin{bmatrix}
    R_{n-1} & r_m \\
    r_m^T & 1
\end{bmatrix}
\]

Where \( R_{n-1} \) is the correlation coefficient matrix of other indicators,

\[ r_m = (r_{1m}, r_{2m}, r_{3m}, \ldots, r_{(m-1)m})^T \]

The complex correlation coefficient between indicators and other indicators is defined as:

\[ p_m = r_m^T R_{n-1}^{-1} r_m \]

The complex correlation coefficients are inverted and normalized to obtain the weights of each index.

\[ \omega^{(2)} = \frac{1}{\rho_{ij}} \sum_{j=1}^{n} \frac{1}{\rho_{ij}} \]

Table 3. weights of complex correlation coefficient method.

| indicators | ZX   | ZR   | ZC   | ZT   | ZI   |
|------------|------|------|------|------|------|
| weight     | 0.163 | 0.105 | 0.147 | 0.372 | 0.213 |

(3) The combination weighting method combined with the coefficient of variation method and the complex correlation coefficient method

The weights determined by the above two methods reflect the importance of each indicator in different aspects. You can combine the two approaches so that they have the best of both.

The final weight of indicators is determined as follows:

\[ w_i = \frac{w_{ij}^{(1)} w_{ij}^{(2)}}{\sum_{j=1}^{n} w_{ij}^{(1)} w_{ij}^{(2)}} (i = 1, 2, \ldots 5; j = 1, 2, \ldots n) \]
Table 4. combined weighting method.

| indicators | ZX  | ZR  | ZC  | ZT  | ZI  |
|------------|-----|-----|-----|-----|-----|
| weight     | 0.1435 | 0.1215 | 0.1760 | 0.3960 | 0.1630 |

2.1.4. Establishment and Solution of TOPSIS Comprehensive evaluation model based on Mahalanobis distance

According to the analysis, there is a certain degree of correlation between the indicators, while the traditional TOPSIS Method does not deal effectively with common problems. Therefore, we have established the evaluation method of TOPSIS based on Mahalanobis distance. The steps are as follows:

Step one: Build an evaluation matrix for sample data.

Let the decision matrix of this problem be \( R \):

\[
R = \begin{bmatrix}
    r_{11} & r_{12} & \cdots & r_{15} \\
    r_{21} & r_{22} & \cdots & r_{25} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{n1} & r_{n2} & \cdots & r_{n5} \\
\end{bmatrix}
\]

and \( R \) can form a normalized decision matrix \( Z' \) with elements of \( Z'_{ij} \), and:

\[
Z'_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^{n} r_{ij}^2}}, i = 1, 2, \ldots, 5, \quad j = 1, 2, \ldots, n;
\]

The \( r_{ij} \) is given by decision matrix \( R \) (Because of the huge amount of data, only five of the roller coasters were selected due to the limited layout). The \( r_{ij} \) is the evaluation value of the jth target for the ith object. The five indicators are: consumption level, stimulating factor, risk factor, comfort level, and infrastructure. And \( n \) evaluation targets are \( n \) roller coasters.

According to the data, the decision matrix is:

\[
R = \begin{bmatrix}
    0.2545 & 2.5085 & 1.6891 & 0.9345 & 0.9778 \\
    0.2365 & 2.5085 & 1.0310 & 0.9345 & 0.9778 \\
    0.1255 & 21.4125 & 1.1905 & 0.9814 & 0.9545 \\
    0.0676 & 7.1654 & 0.2090 & 0.9864 & 0.4524 \\
    0.2540 & 5.1263 & 2.0000 & 0.9153 & 1.5213 \\
\end{bmatrix}
\]

Step two: Normalize the matrix \( R \). Its formula is as follows:

\[
Z'_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^{n} r_{ij}^2}}, i = 1, 2, \ldots, 5, \quad j = 1, 2, \ldots, n;
\]

Substitute data into the above formula to get its normalized decision matrix.
\[
Z' = \begin{bmatrix}
1.2382 & 0.0046 & 0.2817 & 0.2068 & 0.2459 \\
1.1506 & 0.0046 & 0.3521 & 0.2068 & 0.2459 \\
0.6106 & 0.0390 & 0.1408 & 0.2172 & 0.2409 \\
0.3289 & 0.0131 & 0.3521 & 0.2183 & 0.1014 \\
1.2358 & 0.0093 & 0.0704 & 0.2026 & 0.2535
\end{bmatrix}
\]

Step three: The weight matrix \( W = (w_1, w_2, w_3, w_4, w_5) \) is multiplied on the right by the evaluation matrix \( R \) to obtain the value matrix.

Step four: Determine positive ideal solution \( S_i^+ \) and negative ideal solution \( S_i^- \).

Because in the scheme, as for the ideal solution, the larger the index value is, the better the value of each index; the smaller the negative solution is, the better the value of each index, so the ideal solution and the negative ideal solution are obtained. The ideal solution and the negative ideal solution are respectively:

\[
S^+ = (1.2382 \ 0.03090 \ 0.3521 \ 0.2183 \ 0.2535)
\]

\[
S^- = (0.3289 \ 0.0046 \ 0.0704 \ 0.2026 \ 0.1014)
\]

Step five: Calculate the distance \( S_{i_j}^+ \) from the ith evaluation roller coaster to the positive ideal solution \( S^+ \) and the distance \( S_{i_j}^- \) to the negative ideal solution \( S^- \). \( Z_i \) refers to the ith evaluation roller coaster, \( \Sigma \) is the covariance matrix of \( Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}, Z_{i5} \). And \( \Sigma^{-1} \) is the inverse matrix of \( \Sigma \). The formulas are as follows:

\[
S_{i_j}^+ = d(Z_i - Z^+)) = [(Z_{ij} - Z_{j}^+)^T \Sigma^{-1} (Z_{ij} - Z_{j}^+)]^{1/2}
\]

\[
S_{i_j}^- = d(Z_i - Z^-)) = [(Z_{ij} - Z_{j}^-)^T \Sigma^{-1} (Z_{ij} - Z_{j}^-)]^{1/2}
\]

Step six: Calculate the relative proximity.

The relative proximity of an arbitrary feasible solution to an ideal solution is defined as:

\[
C_{ij}^* = \frac{S_{i_j}^-}{S_{i_j}^- + S_{i_j}^+}
\]

the bigger the value is, the better the value is.

Step seven: Sort by the value of \( C_{ij}^* \).

| Roller Coasters   | \( S_{i_j}^+ \) | \( S_{i_j}^- \) | \( C_{ij}^* \) | rank |
|------------------|-----------------|-----------------|-----------------|------|
| Kingda Ka        | 0.1113          | 0.9173          | 0.8918          | 4    |
| Extreme Rusher   | 0.2270          | 0.7832          | 0.7753          | 52   |
| Nemesis Inferno  | 0.2628          | 0.1162          | 0.3066          | 167  |
| Adrenaline Peak  | 0.3570          | 0.0796          | 0.1823          | 251  |
| Wild Thing       | 0.5489          | 0.8820          | 0.6164          | 123  |
2.2. Model of coincidence based on the scoring system and Euclidean distance

2.2.1. Ranking Results

According to our algorithm, the top ten of the world are discharged and then listed. By looking up the information, we found three Rating/ranking in the line system, listing the top ten of their website and the top ten of our own algorithms. The list is as follows:

| Rank | Results from TOPSIS                     |
|------|----------------------------------------|
| 1    | Steel Dragon 2000                      |
| 2    | Kingda Ka                              |
| 3    | Millennium Force                       |
| 4    | Top Thrill Dragster                    |
| 5    | Phantom's Revenge                      |
| 6    | Ten Ring Roller Coaster                |
| 7    | Tower of Terror II                     |
| 8    | Dodonpa                                |
| 9    | Superman The Escape                    |
| 10   | Goliath                                |

Table 6. Ranking of our algorithm generation.

| One               | Two                        | Three                              |
|-------------------|----------------------------|------------------------------------|
| Kingda ka         | Formula Rossa              | Kingda ka                          |
| Top Thrill Dragser| Kingda ka                  | Top Thrill Dragster                |
| Dodonpa           | Top Thrill Dragster        | Superman: Escape from Krypton      |
| Tower of Terror   | Dodonpa                    | Tower of Terror II                 |
| Superman The Escape| Superman.: Escape from Krypton | Red Force                         |
| Steel Dragon 2000 | Red Force                  | Fury 325                           |
| Millennium Force  | Tower of Terror II         | Steel Dragon 2000                  |
| Titan             | Ring Racer                 | Millennium Force                   |
| Goliath           | Steel Dragon 200           | Leviathan                          |
| Phantom's Revenge | Millennium Force           | Intimidator 305                    |

Table 7. Rankings generated by the online platform.

2.2.2. the establishment of the model of coincidence

The topic asks for a roller coaster rating/ranking result and description from our team's algorithm with at least two online system's roller coaster ratings/ranks compared and discussed. Since you want to compare multiple on-board rankings, we need a relative standard and correct roller coaster ranking, taking the top ten of this ranking, namely the list of the top ten roller coasters in the world, as a benchmark, and then let these multiple rankings and this benchmark ranking, calculating the degree of coincidence. The roller coaster ranking with the highest degree of coincidence is the best ranking.

We first build a standard correct roller coaster benchmark ranking and then calculate the coincidence of our ranking and the three online systems with the benchmark rankings, and finally we discuss the results and analysis. The construction of benchmark rankings.

**Step 1: Generate a total roller coaster list**

Our team's ranking and ranking for the four online systems are convincing, both of them are a list of the respective "World's Top Ten Roller Coasters" generated by certain standards and indicators, but may be due to different defects or deficiencies leading to incomplete rankings. Through analysis, you can
know that these five roller coaster rankings have a certain persuasiveness and reference, so you can know that the standard and correct roller coaster ranking also exists in these five rankings.

**Step 2: Implement a scoring system for the roller coaster in the list**

To generate a ranking from this roller coaster list, the fairest way is to implements a scoring system for each roller coaster of this list.

Through analysis, we can know that if a roller coaster ranks among the top in our algorithm and the ranking of the four websites, it is recognized as an excellent roller coaster, and it must be a roller coaster also among the best in the benchmark ranking. The same is true if a roller coaster is ranked lower or only one or two times in our team and the rankings of the four online systems, the roller coaster in the benchmark and the ranking will also be ranked lower. Through the above analysis, we have created the following scoring system.

\[
f_{ij} = \begin{cases} 
11 - m_{ij}, & 1 \leq m_{ij} \leq 10 \\
0, & 11 \leq m_{ij}
\end{cases}
\]

Among them, \( f_{ij} \) is the score of the jth roller coaster in the list on the ith public website, and \( m_{ij} \) is the ranking of the jth roller coaster in the list in the ith public website.

\[
F_j = \sum_{i=1}^{4} f_{ij}
\]

Where \( F_j \) is the total score of the jth roller coaster in the list.

**Step 3: Generate a baseline ranking**

After calculation, you can get the scores of all the roller coasters in the list, sorting the roller coaster according to the size of the score. The top ten roller coasters are listed below, which is the list of benchmark rankings.

| Rank | Roller Coaster          |
|------|-------------------------|
| 1    | Kingda Ka               |
| 2    | Millennium Force        |
| 3    | Millennium Force        |
| 4    | Tower of Terror II      |
| 5    | Superman The Escape     |
| 6    | Goliath                 |
| 7    | Ten Ring Roller Coaster |
| 8    | Top Thill Dragster      |
| 9    | Dodonpa                 |
| 10   | T Express               |

2.2.3. calculation of coincidence degree

After analysis you can know if a roller coaster ranked in our team (and the other three online systems) is the same as the rankings in the benchmark ranking list, then our team (the other three online systems) will be credible. So when a roller coaster list has a higher similarity with a benchmark ranking, that is, the highest degree of coincidence a list has, the best the roller coaster ranking list is. We use Euclidean distance to calculate the coincidence among our team's rating/ranking list, the four online systems and benchmark ranking list. These six ranking lists are generated into six 10-dimensional vectors as follows:

\[
p_i = [g_{i1}, g_{i2}, \ldots, g_{i10}], (i = 1, 2, \ldots, 5)
\]
Where \( p_i \) is the rating/ranking list of our team and list of four online systems, which \( g_{ij} \) is the code name of the jth roller coaster.

\[
p = [g_{g_1}, g_{g_2}, ..., g_{g_{10}}]
\]

\( p \) is the benchmark ranking list, and \( g_{ij} \) is the code of the ith roller coaster in the benchmark ranking list.

The Euclidean distance is calculated as follows:

\[
d_i = \sqrt{\sum_{j=1}^{10} (g_{ij} - g_{ij})^2}, (i = 1, 2, ..., 5)
\]

Where \( d_i \) is the Euclidean distance of the ith ranking list and the benchmark ranking list.

\[
C_i = \frac{1}{1 + d_i}, (i = 1, 2, ..., 5)
\]

Where \( C_i \) is the coincidence of the ith ranking list and the benchmark ranking list. The greater the coincidence, the better the ith ranking list.

2.2.4. Analysis of results
We calculate the coincidence among the list of our team, the three online systems and the benchmark ranking list. The results obtained are as follows:

| Table 9. Coincidence results. |
| Online system ranking | TOPSIS | Most dangerous | Most scary | Most exciting |
|-------------------------|--------|----------------|------------|--------------|
| Coincidence degree      | 0.8540 | 0.4792         | 0.64436    | 0.5587       |

2.3. Conclusion
Combining the coefficient of variation method with the complex correlation coefficient method to determine the weight of the index, not only can separate the evaluated objects clearly but also can eliminate duplicate information between indicators. Combined with the improved rank-to-ratio comprehensive evaluation model, effectively solves the problem of multi-indicator decision making, making calculations simple and highly integrated with other methods. Used in the roller coaster evaluation model, each indicator is quantified separately so that it can accurately influence the indicator on the target, thus it has strong portability and can be applied to the solution of other indicator systems.

Also the TOPSIS comprehensive evaluation model based on Mahalanobis distance can be applied to the fields of social economy and the like.

Reference
[1] CNIC. Special Vehicle Market Development Research Report [R]. Beijing: China Internet Information Center, 2015.
[2] China Special Car Market Analysis Report[R]. Shanghai: Roland Berger Management Consulting Company, 2016
[3] Didi Media Research Institute and First Financial Business Data Center[EB/OL].
[4] Jiang Qiyuan, Xie Jinxing. Mathematical Model Fourth Edition [J]. Beijing: Higher Education Press,2011:82-130.
[5] Huang Fei. Comprehensive evaluation of the development level of smart grid[D].Beijing Jiaotong University, 2016.Comparison Library dissertation database
[6] Zhou Pin. MATLAB numerical analysis application tutorial. Beijing: Electronic Industry Press,2014.