Travel Route Planning Based on Genetic Algorithm and Greedy Algorithm

Zhiyi Wang*, Tong Zhou

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

*chn_sherry@nuaa.edu.cn

Abstract

Based on genetic algorithm and greedy algorithm, this study proposes a scientific tourism route planning model to maximize tourists' visiting experience and cost-effectiveness. Firstly, this paper pre-processed the data of 352 Chinese cities and their 35,200 attractions, and screened out the 50 most attractive cities by using principal component analysis and TOPSIS method, taking into account factors such as city size. Subsequently, tourists' travel route planning was optimized based on genetic algorithm. Finally, greedy algorithm was introduced to study the customized routes under different tourism needs, such as mountain view tourism routes, to optimize the multiple indicators involved. This study not only provides foreign tourists with personalized travel plans, but also successfully achieves a balance between economy and efficiency under the premise of guaranteeing tourists' reasonable time and cost.

Keywords

Principal Component Analysis (PCA); TOPSIS; Genetic Algorithm; Greedy Algorithm.

1. Introduction

This paper integrates a variety of computer algorithms and models, including Principal Component Analysis (PCA) [1], TOPSIS [2], Genetic Algorithm [3], and Greedy Algorithm [4], aiming to optimize the tourist routes and enhance the tourists' experience [5]. Firstly, for the 35200 attractions in 352 cities were scored, this paper utilized PCA for dimensionality reduction through data cleaning and integration to extract the key factors and ensure the reliability of the data. Then, the TOPSIS method was used to conduct a comprehensive evaluation of the cities, and the 50 destinations most favored by international tourists were selected. In addition, in response to the different needs of tourists, this paper designs multiple travel routes, using greedy algorithms and genetic algorithms to find the optimal balance between time and cost, ensuring that tourists visit as many cities and attractions as possible within a limited time. Through the comprehensive application of these algorithms, this paper not only improves the practicality of travel planning, but also provides tourists with personalized travel plans, which greatly enhances the travel experience and satisfaction [6].

2. Data Collection and Preprocessing

In this paper, 35,200 attractions in 352 cities are examined.

(1)Data merging

Data merge the collected tourist attractions of 352 cities and integrate them into a dataset that contains all attractions. The obtained dataset is viewed, not every city has 100 attractions, some cities have a small number of attractions, and only 226 attractions meet the conditions.

(2)Data Collection

As this paper needs to focus on the city, this paper collects 20 indicators for the subsequent evaluation of the city's urban scale, environmental protection, humanistic heritage, transportation system, as well as climatic characteristics, characteristics of the cuisine of these six factors. In this paper, the distance between cities is used, and to facilitate the calculation, the distance between two cities is approximated by straight line distance. The geographical distribution of the attraction cities is analyzed, and it is found that the attraction cities in the southeast are denser and those in the northwest are sparser.

(3)Data Cleaning

When processing the merged data, many missing values and outliers were found in the dataset. To simplify the data processing, this paper decided to delete those records with vacancies and zero scores in the rating columns based on reasonable grounds. Deleting these data had a minimal impact on the practicalities of this paper but was able to save significant time.

3. Analysis of Attraction Ratings

The results of data merging were first utilized for data analysis. The max function is used to find the highest rating value. The value obtained here is the highest rating (BS). The highest rating was obtained as 5. Then, from all the attraction data, all the attractions whose ratings reached this maximum score were filtered. After counting the number of these attractions, it was found that a total of 2,563 attractions received this highest rating.

Next, these attractions were grouped according to the city in which they were located. By counting the number of attractions in each category (i.e., each city), it was possible to obtain how many top-rated attractions each city had. Based on the count of the number of attractions, this section sorts the cities, prioritizing those with the most top-rated attractions. After the sorting was completed, the top 10 cities with the highest number of highest rated attractions were selected from the list. Figure 1 shows the results of the top 10 cities after sorting.

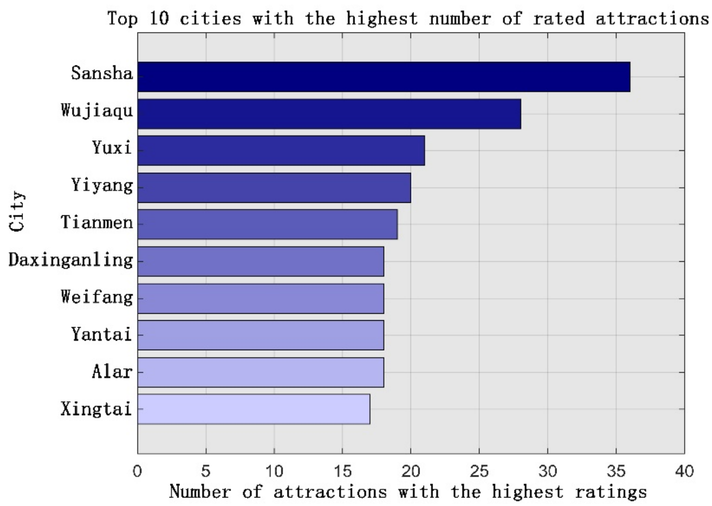

Figure 1. Cities with the highest number of top-rated attractions

Figure 1 shows that the top 10 cities with the greatest number of top-rated attractions are Sansha, Wujiaqu, Yuxi, Yiyang, Tianmen, Daxing'anling, Weifang, Yantai, Alar, and Xingtai.

4. Comprehensive City Evaluation

4.1. Data Dimensionality Reduction Based on Principal Component Analysis

This section collects and organizes a dataset containing 20 relevant metrics for 352 cities, which are classified into several main categories: city size, environmental protection, human heritage,

ISSN: 2688-8645

transportation system, climatic characteristics, and distinctive cuisine. These assessment criteria cover a wide range of aspects, including but not limited to air quality index (AQI), vegetation cover level (%), percentage of sewage purification (%), percentage of exhaust gas purification (%), percentage of garbage recycling and disposal (%), and number of historical sites (nos.). Since modeling directly with these high-dimensional data leads to excessive computational complexity and may trigger dimensional catastrophe, it is necessary to reduce the dimensionality of the data through dimensionality reduction techniques while retaining as much important information as possible.

(1)KMO test

Before choosing Principal Component Analysis (PCA), Kaiser-Meyer-Olkin (KMO) test was first conducted to determine whether the correlation between variables is suitable for performing factor analysis. The result of KMO test ranges from 0 to 1. A KMO value of more than 0.6 indicates that the data is suitable for PCA analysis. The specific steps are as follows:

1) Correlation coefficient matrix:

With p variables, calculate their correlation coefficient matrix $R = [r_{ij}]$.

2) Partial correlation coefficient matrix: Calculate the inverse of the correlation coefficient matrix $P = [a_{ij}]$.

3) Calculate the KMO statistic

For each pair of variables i and j, calculate the partial correlation coefficient r_{ij} and the correlation coefficient a_{ij}.

Calculate the square of the partial correlation coefficient and the square of the correlation coefficient: $\sum_{i\neq j} r_{ij}^2$ and $\sum_{i\neq j} a_{ij}^2$.

Calculate the KMO statistic:

$$KMO = \frac{\sum_{i\neq j} a_{ij}^2}{\sum_{i\neq j} a_{ij}^2 + \sum_{i\neq j} r_{ij}^2} \tag{1}$$

(2)Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was implemented for data dimensionality reduction on the set of indicators that passed the KMO test. PCA converts variables that were originally correlated into a new set of uncorrelated variables by linear transformation, and these variables are called principal components. They are ranked in decreasing order according to the magnitude of variance contribution to reveal the main trends in the dataset. The specific steps are as follows.

1) Data Standardization

Before performing PCA, the data are standardized so that the mean of each indicator is 0 and the standard deviation is 1, eliminating the influence of different scales on the analysis results. The standardization formula is:

$$Z = \frac{X-\mu}{\sigma} \tag{2}$$

Where Z is the normalized value, X is the original value, μ is the mean value, and σ is the standard deviation.

2) Calculating the covariance matrix

The standardized data form a covariance matrix S, which is used to measure the linear correlation between the variables. The covariance matrix is calculated as:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^T \tag{3}$$

Where n is the number of samples, X_i is the ith sample, and μ is the mean.

3) Eigenvalues and eigenvectors

The eigenvalues λ and eigenvectors v of the covariance matrix S are calculated.

The eigenvalues represent the variance of the principal components, and the eigenvectors represent the direction of the principal components.

4) Selection of principal components

Arranged in order of eigenvalues from largest to smallest, the first few principal components are selected so that their cumulative explained variance reaches 90% or more to ensure that the main information is retained.

The formula for calculating the cumulative explained variance is:

$$Cumulative\ Explanatory\ Variance = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{4}$$

Where k is the number of selected principal components and p is the number of original variables.

5) Transforming the data

The original data are projected onto the selected principal components to obtain the dimensionality reduced data matrix.

The new data matrix consists of the first k eigenvectors and each column represents a principal component. The transformed data matrix Y is calculated as:

$$Y = X \cdot W \tag{5}$$

Where X is the original data matrix after normalization and W is the matrix consisting of the first K eigenvectors.

The result part is shown in Figure 2.

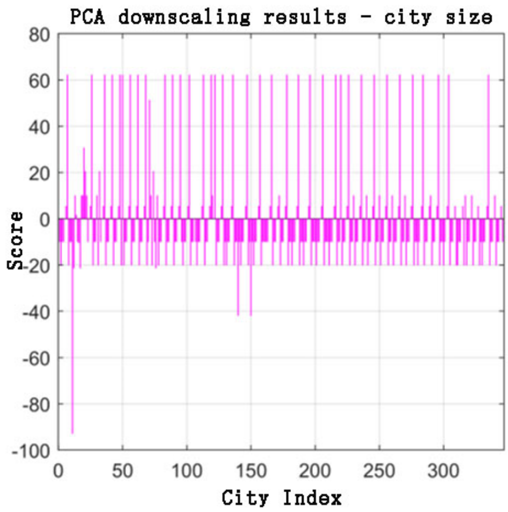

Figure 2. The result of dimensionality reduction of city size

4.2. TOPSIS Comprehensive Evaluation

TOPSIS (Approximate Ideal Solution Ranking Method) is a multi-criteria decision analysis technique that determines the optimal solution by measuring the distance of each candidate solution from the ideal solution and the negative ideal solution. TOPSIS is calculated as follows:

(1)Normalization

$$z_{ij} = \frac{x_{ij} - x_j}{\sigma_j} \tag{6}$$

Where, x_{ij} is the original data, x_j is the mean of the jth indicator, and σ_j is the standard deviation of the jth indicator.

(2)Entropy value calculation

$$H_j = -\frac{1}{log(n)}\sum_{i=1}^{n} p_{ij}\, log\left(p_{ij}\right) \tag{7}$$

Where, $p_{ij} = \frac{z_{ij}}{\sum_{i-1}^{n} z_{ij}}$, n is the number of samples.

(3)Weight calculation

$$w_j = \frac{1-H_j}{\sum_{j=1}^{m}(1-H_j)} \tag{8}$$

Where m is the number of indicators.

(4)Calculate the positive and negative ideal solutions

$$A^+ = \left\{max(V_{ij})|j \in J\right\}, \left\{min(V_{ij})|j \in J'\right\} \tag{9}$$

$$A^- = \left\{min(V_{ij})|j \in J\right\}, \left\{max(V_{ij})|j \in J'\right\} \tag{10}$$

Where J is a revenue-based metric and J' is a cost-based metric.

(5)Calculate the distance of each option from the positive and negative ideal solutions

$$D_i^+ = \sqrt{\sum_{j=1}^{n} w_j\left(v_{ij} - A_j^+\right)^2} \tag{11}$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} w_j\left(v_{ij} - A_j^-\right)^2} \tag{12}$$

(6)Calculate the relative proximity of each option

$$C_i^+ = \frac{D_i^-}{D_i^+ + D_i^-} \tag{13}$$

After completing the Principal Component Analysis (PCA) dimensionality reduction process for the indicators of each category, these dimensionality reduction data are integrated into the DataFrame, and the city information columns are added to it. Next, the entropy weighting method was utilized to determine the weights of each category, followed by the application of the TOPSIS method for the comprehensive evaluation of the cities as a means of calculating the overall score of each city. The formula for the overall score of each city is as follows:

$$S_i = \sum_{k=1}^{k} W_k Z_{i,k} \tag{14}$$

Finally, the 50 cities that are most desirable to foreign tourists are selected according to the order of the comprehensive score, which is sorted from high to low. Among them, the results of the top 25 cities are shown in Figure 3.

The scores of the 50 cities most desirable to foreign tourists in descending order are: Fuzhou, Beijing, Qionghai, Kokodara, Chuxiongzhou, Lingao, Chengdu, Yangzhou, Dingshi, Nanchong, Jiayuguan, Taizhou, Jincheng, Changde, Wujiaqu, Hangzhou, Qianjiang, Weifang, Xianning, Enshi, Shanwei, Huzhou, Shaoyang, Zhongshan, Guiyang, Lijiang, Huizhou, Tonghua, Anqing, Hohhot, Shuozhou, Xiong'an New Area. Hezhou, Danzhou, Panzhihua, Sanya, Jiyuan, Dali, Baotou, Xinzhou, Zhuhai, Fuzhou, Liangshan, Yunfu, Shizuishan, Bortala, Baicheng, Huangshan, Yuncheng, Ordos.

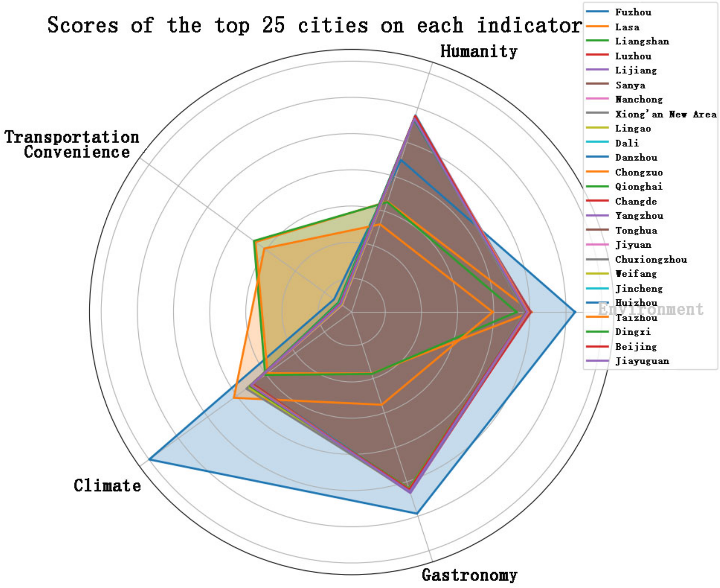

Figure 3. Radar chart of the scores of the top 25 cities on each indicator

5. Route Planning for Optimal Playing Experience

5.1. Data Optimization Processing

(1)Time data normalization

For the time interval a hour to b hour, calculate the average value of the time interval:

$$avg_time = \frac{a+b}{2} \tag{15}$$

For the time interval "day x to day y", convert to hours and average:

$$avg_time = \frac{(x\times 24)+(y\times 24)}{2} \tag{16}$$

Through the above process, the text format of all the play time is converted to hourly unit uniformly and recorded in the data table, which can facilitate the subsequent design of travel routes and the calculation of the total time consumed, tickets and transportation costs.

(2)Ticket Price Data Parsing

Use a regular expression to check whether the ticket price string contains a price number:

$$has\ price = regex\ match(ticket\ str, pattern) \tag{17}$$

If the string contains the word "free" and there is no price figure, then the ticket price is set to 0:

$$ticket\ price = \begin{cases} 0 & if\ has\ price\ is\ False\ and\ \text{free}\ in\ ticket\ str \\ parsed\ price & otherwise \end{cases} \tag{18}$$

Design the regular expression to match the adult ticket price pattern preferentially. If the first match fails, re-match to accommodate a different formulation of the ticket price format.

If the first regular expression finds no price information, reapply the expression to capture any price information in the string:

$$if\ tokens\ are\ empty: tokens = regex\ earch(ticket\ tr, price\ pattern) \tag{19}$$

Convert the parsed price to numeric format and store it. If price information is missing, it is recorded as NaN in the ticket price array:

$$\text{if tokens are not empty: ticket prices } [i] = \text{ convert to double (tokens[1])else: ticket prices } [i] = \text{ NaN} \tag{20}$$

5.2. Optimization Modeling

(1)Data preprocessing

From the data set, the highest rated attractions in each of the 50 "most desirable cities for foreign tourists" are filtered out.

Clean the attraction data, eliminate missing values or invalid data, and keep the key information such as attraction name, rating, ticket price and suggested playing time.

(2)HSR Travel Time and Cost Calculation

Haversine formula calculates the distance D between cities:

$$D = 2R \arcsin\left(\sqrt{sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + cos(\phi_1)\cdot cos(\phi_2)\cdot sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \tag{21}$$

Where $R = 6371\ km$ (mean radius of the earth), ϕ is latitude and λ is longitude.

HSR travel time T_{travel}:

$$T_{travel} = \frac{d}{v_{train}} \tag{22}$$

Assuming the average speed of HSR:

$$v_{train} = 300km/h \tag{23}$$

HSR cost C_{travel}:

$$C_{travel} = D \times p_{train} \tag{24}$$

If the cost per kilometer of high-speed rail $p_{train} = 0.47$ yuan.

(3)Heuristic algorithm establishment

Initialization parameters: set the initial city as Guangzhou, the total time limit $T_{limit} = 144$ hours, and the current time and cost as 0.

Iterate through the unvisited cities, for each city c_i.

Calculate the travel time and cost of HSR from current city $c_{current}$ to city c_i; select the highest rated attraction s_{best} in city c_i, get its travel time T_{visit} and ticket cost C_{ticket}; calculate the local travel time $T_{local} = 0.5$ hours; calculate the total time $T_{total} = T_{travel} + T_{local} + T_{visit}$; Calculate the total time including the rest time $T_{final} = 8$ hours; Select the best city: among the cities that meet the time limit, select the city with the highest overall experience score c_{best}.

(4)Update status

Update the current city as c_{best}; update the current time $T_{current} = T_{final}$; update the current cost $C_{current} = C_{current} + C_{travel} + C_{ticket}$; update the list of visited cities and attractions.

(5)Termination conditions

Repeat the above steps, if there is no suitable next stop or time runs out, then end the journey.

(6)Result Output

Output the tour route, visited attractions, total travel time, total cost and total number of attractions.

(7)Formula Summary

The total time includes rest time:

$$T_{final} = T_{current} + T_{total} + \left[\frac{T_{current}+T_{total}}{24}\right] \times 8 \tag{25}$$

Total cost:

$$C_{total} = C_{current} + C_{travel} + C_{ticket} \tag{26}$$

5.3. Minimum Cost Route Planning

The genetic algorithm is solved as follows:

(1)Initialize the population

Keep the paths generated by the heuristic algorithm, initialize the population by taking the paths of the heuristic algorithm as part of the population, and then randomly generate other paths to supplement the population.

(2)Calculate fitness, selection

Calculate the total cost and total time of a path, including the cost of traveling between cities, the cost of tickets, and the suggested playing time. Sort the population by the total cost of the path and filter out paths that exceed the effective play time limit.

(3)Crossover Variation

Combine two parent paths to generate a new child path using the Partial Matching Crossover (PMX) method. Randomly swap the locations of the two cities in the pathway to simulate genetic mutations.

(4)Update population

Retain the best two individuals in the current population. Generate new offspring through selection, crossover and mutation, ensuring that the total time of the new individuals does not exceed the effective play time. Replace the old population with the newly generated one.

(5)Return to the optimal path

In each generation, select the individual with the highest fitness as the new population. Gradually find a better solution by iterating through multiple generations. Return the final optimal path and its corresponding total cost.

5.4. Model Solving

Through the above steps, gradually plan the optimal travel route for tourists within 144 hours, aiming to reduce the total expenditure on tickets and transportation, while ensuring that they visit as many cities as possible and enjoy the best overall experience.

Travel Route: Guangzhou→Zhongshan→Huizhou→Shanwei→Heyuan→Hengyang→Changde →Xianning→Ma'anshan.

Table 1 shows the total cost and time used for the trip and the number of attractions visited.

Table 1. Travel Route Parameter Table

Total tour time	Total cost	Number of cities/attractions
142 hours	1010.26	9

6. Mountain View Tour Route Planning

The data is first preprocessed and the model is initialized. This includes importing detailed information about the cities and mountain views and defining initial parameters such as total time limit, current clock, cost, host city, and a list of visited cities and attractions. The initial entry city is chosen as the city with the highest rated mountain view. Subsequently, a greedy algorithm is used to plan the travel route, filter the cities that have not been visited, evaluate the distance and cost of the HSR trip, select the highest rated mountain view with the lowest entrance fee to visit, and calculate the time required for the entire visit, including transportation, sightseeing, and rest time. Within the time limit, the city with the lowest cost is selected as the next visit. After each selection, the current status is updated with the city, time, cost and excursion history, and planning continues until time runs out or there is no suitable next stop. Finally, the complete trip plan, list of attractions, total time, total cost and total number of attractions will be displayed.

6.1. Model Construction

(1)City spacing

The semipositive vector formula is a formula used to calculate the great circle distance (i.e., the shortest distance measured along the surface of the earth) between two points on the earth, which considers the spherical shape of the earth and is therefore more accurate than the simple Euclidean formula in calculating distances on the earth. In this paper, we use the semipositive vector formula and the latitude and longitude of two cities to calculate city-to-city distances. Assuming that the latitude and longitude of city A and city B are (lat_1, lon_1) and (lat_2, lon_2) respectively, the distance d between them can be calculated by the following formula:

$$a = sin^2\left(\frac{\Delta lat}{2}\right) + cos(lat_1) \cdot cos(lat_2) \cdot sin^2\left(\frac{\Delta lon}{2}\right) \tag{27}$$

$$c = 2 \cdot atan2\left(\sqrt{a}, \sqrt{1-a}\right) \tag{28}$$

$$d = R \cdot c \tag{29}$$

Where:

$\Delta lat = lat_2 - lat_1$ is the latitude difference between two cities.

$\Delta lon = lon_2 - lon_1$ is the difference in longitude between the two cities.

R is the mean radius of the earth, which usually takes the value of 6371 kilometers.

$atan2$ is a two-parameter arctangent function, which provides a wider domain of definition than the ordinary arctangent function.

(2)Travel time and travel cost

Set the speed of HSR as $V_g$km/h, and take 300 km/h here, and set the cost of HSR as F_g yuan/km, and take 0.5 yuan/km here.The travel time t_g and travel cost C_g from the current city to the target city can be computed by the following equations can be computed by the following equations:

$$t_g = \frac{d}{V_g} \tag{30}$$

$$C_g = d \cdot F_g \tag{31}$$

Where d denotes the distance traveled by HSR.

(3)Play time and local catch-up time

The calculation of play time and local catch-up time consists of three parts: first, the catch-up time within the city is set to half an hour; second, the play time is determined according to the time required for specific attractions; and lastly, the total intra-city time is the cumulative sum of the travel time, the catch-up time within the city, and the play time.

Rest schedules for visitors are an integral and important part of planning a travel itinerary to ensure that they are well rested. Within each 24-hour cycle, eight hours must be scheduled for rest. The current total time is calculated by adding the available time to the 8 hours of rest and then performing a 24-hour division to ensure that the total time is kept within one day.

(4)Iteration and Termination Conditions

In the process of selecting the next best city to visit, each unvisited city is filtered by first considering the highest rated mountain views and calculating the time and cost required to visit them. The next city is then selected based on the principle of minimizing total time and total cost. The evaluation criterion is the experience value, which is directly related to the rating of the mountain view. When selecting, it is important to ensure that the total time does not exceed the established time limit and that the total time plus the current time does not exceed the maximum time allowed. Once a city has been selected, status information is updated with the current city, time, and cost.

The trip will be terminated when a suitable next stop cannot be identified or when the travel time (144 hours) is exhausted.

6.2. Results Presentation

Using the heuristic greedy algorithm for analysis, this paper finalizes an efficient travel route covering five cities. The travel order of the five cities is Sanya→ Baoting→ Maanshan→ Shangrao→Nanchang. The length of stay in each city is 16.26 hours in Sanya, 6.94 hours in Baoting, 80.89 hours in Maanshan, 13.05 hours in Shangrao, and 12.55 hours in Nanchang. Within the time limit of 144 hours of touring, the foreign tourists enjoyed the highest rated mountain attractions while keeping the total touring time within 142.74 hours, with a total cost of RMB 1,283.18.

7. Conclusion

This study provides efficient and economical route choices for foreign tourists' travel planning in China through scientific data analysis and algorithm optimization. First, by cleaning and analyzing the rating data of 352 cities and 35,200 attractions, the attractions with the highest scores were identified, and the 50 most popular cities were further filtered. Subsequently, the tourism routes were optimized by combining factors such as city size, cultural characteristics, and transportation accessibility with a comprehensive assessment using principal component analysis and TOPSIS methods. In response to different tourists' needs, the study provided the best time-based tour experience routes and the lowest cost-based routes respectively, while personalized routes were designed for tourists who love mountain scenery, ensuring that tourists could visit as many attractions as possible and enjoy a rich travel experience within 144 hours. The study not only enhances tourist satisfaction, but also provides strong support for the development of China's tourism industry.

References

[1] Yan Guanzheng. Competitiveness analysis of Qingdao tourism based on principal component analysis[J]. Modern Business,2023, (17): 59-62.DOI: 10.14097/j.cnki.5392/2023.17.012.

[2] Yu Jiayan, Zhang Pei, Weng Lingyan. Research on development potential and countermeasures of rural tourism land based on entropy weight TOPSIS model[J]. Tourism Overview,2024, (20):38-40.

[3] XU Xiangrong. Research on tourism recommendation algorithm based on improved genetic algorithm and multi-source heterogeneous data[D]. Xi'an University of Technology, 2023.DOI: 10. 27398/d.cnki.gxalu.2023.001863.

[4] Yuan Jiangshu, Feng Zhenyu, Zhu Tianle, et al. Optimization design of tourism routes in Huangshan Mountain scenic area based on greedy algorithm[J]. Modern Commerce Industry,2020,41(20): 32-33.DOI: 10.19311/j.cnki.1672-3198.2020.20.016.

[5] Sun Haodong, Wang Rui, Liu Siyang, et al. Research on Tourism Route Planning Based on Multi-objective Optimization Model[C]//Professional Committee of Urban Transportation Planning, China Society of Urban Planning. Green Numerical Intelligence: Enhancing Quality and Efficiency - Proceedings of the 2024 Annual Conference on Urban Transportation Planning in China. Beijing Institute of Urban Planning and Design; Changsha University of Technology; Beijing Institute of Technology, 2024:13. DOI: 10.26914/c.cnkihy.2024.038591.

[6] Dong Kailing. Research on personalized tourism based on the integrated development of tourism and transportation[J]. Tourism Overview,2021, (10):123-125.