

# Analysis of Local Characteristic Tourism based on Gray Prediction and BP Neural Networks

Wanyu Yang\*, Shuyang Wu, Zhonghua Sheng

School of Big Data and Statistics, Anhui University, Hefei, China

\*Corresponding author: 257228815@qq.com

## Abstract

**This study explores the strategy of utilizing online platforms to promote the recovery and development of the tourism industry against the background of the impact of the 2020 COVID-19 on the tourism industry. The development trend of Harbin's characteristic tourism industry was studied through college students' survey data, combined with multiple regression, gray prediction and BP neural network analysis methods. The results of the study show that combining network publicity with special tourism programs and attracting young people groups is the key, and suggestions for sustainable development are put forward. By helping the inheritance and innovation of local traditional culture, this study aims to promote the innovative development of tourism. This research framework provides theoretical and practical guidance for promoting the development of local tourism.**

## Keywords

**Multiple Regression Analysis; Gray Prediction; BP Neural Networks.**

## 1. Introduction

The COVID-19 has had a huge impact on the global tourism industry since its outbreak in 2020, especially across China, where the tourism industry has suffered unprecedented challenges [1-2]. In order to help the tourism industry recover and realize long-term development as soon as possible, this study uses the online platform as a tool to explore how to promote the recovery and development of the tourism industry by combining local characteristics and online publicity. Using university students as the research object, the study explores their attitudes and habits in tourism consumption, Internet use and the recognition of special tourism, as a basis for the research. Through the collection and analysis of questionnaire data, combined with multiple regression, gray prediction, and BP neural network model, we try to depict the development trend of Harbin's characteristic tourism. This study aims to provide theoretical guidance and practical suggestions for the sustainable development of local tourism and promote the inheritance and innovation of local traditional culture. Through this study, we hope to provide new ideas and strategic directions for the future development of local tourism.

## 2. Analysis of Influencing Factors based on Multiple Regression Modeling

In order to make a qualitative judgment on the future development trend of Harbin's cultural and tourism industry, we decided to use Harbin's tourism main economic indicators from 2012-2023 as the basic data for trend analysis. There are many factors affecting the tourism industry in Harbin, and the following are the indicators we used for this analysis:

Tourism income: Statistics on domestic tourism-related income, such as scenic spot tickets, hotel income, sales of tourism souvenirs, etc., can reflect the level of economic development of the tourism industry.

Number of tourism enterprises: Statistics on the total number of tourism-related enterprises, such as hotels, restaurants, travel agencies, etc., can reflect the market size of the tourism industry. Here we choose to express this indicator by counting the number of star class (five-star, four-star, three-star) hotels in the place.

Number of tourists: Here we mainly count the total number of domestic tourists received in the region, which can reflect the basic situation of tourism in the region.

Tourism publicity and promotion: Statistics on the input and effect of tourism publicity and promotion, such as advertising costs, online promotion effect, etc., can reflect the marketing effect of tourism. Here we choose the keyword search volume of "Harbin tourism" on Jitterbug platform to represent this indicator.

Tourist Satisfaction: Tourists' satisfaction with the tourism experience is collected through questionnaires, online evaluations and official website lookups, which can give an idea of the quality of tourism services in the place.

Convenience of transportation: The convenience of transportation in tourism can be reflected through the statistics of passenger transportation in Harbin, including flights, trains, coaches and so on.

## 2.1. Data Collection for Relevant Indicators

We have compiled statistics from the Harbin Bureau of Statistics on the tourism industry and the transportation industry and other industries to get Table.1.

**Table 1.** Harbin city tourism data 2012-2023

| Year | Tourism revenue | Number of tourists | Tourist satisfaction | Number of star-rated hotels | Passenger traffic | Keyword search volume |
|------|-----------------|--------------------|----------------------|-----------------------------|-------------------|-----------------------|
| 2012 | 547.1           | 5052               | 74.56                | 76                          | 15618             | 0                     |
| 2013 | 662.4           | 5526               | 79.35                | 79                          | 13192             | 0                     |
| 2014 | 780.4           | 5990               | 74.04                | 81                          | 13986             | 0                     |
| 2015 | 900.9           | 6496               | 76.2                 | 66                          | 13158             | 0                     |
| 2016 | 1031.2          | 7040               | 74.53                | 67                          | 12523             | 0                     |
| 2017 | 1168.4          | 7688.9             | 75.9                 | 61                          | 11922             | 0                     |
| 2018 | 1369.1          | 8543.7             | 78.5                 | 56                          | 11210             | 0                     |
| 2019 | 1551.9          | 9544.2             | 80                   | 52                          | 11314             | 198                   |
| 2020 | 1099.9          | 7823.6             | 80.95                | 50                          | 4386              | 212                   |
| 2021 | 943             | 9184.9             | 82.47                | 44                          | 5196              | 293                   |
| 2022 | 499.2           | 5511.2             | 80.63                | 17                          | 2080.4            | 330                   |
| 2023 | 1692.5          | 13500              | 85.98                | 15                          | 4014.7            | 6622                  |

## 2.2. Multiple Regression Modeling

The total tourism revenue as the dependent variable  $y$ , time as the case label and other factors as independent variables are represented by the following characters:

$X_1$ : number of tourists,  $X_2$ : tourist satisfaction,  $X_3$ : number of star-rated hotels,  $X_4$ : passenger transportation,  $X_5$ : keyword search. The following multivariate functional equation is established:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \quad (1)$$

### 2.3. Multiple Stepwise Regression Model Solution

Stepwise regression can both avoid to a certain extent the entry of multicollinear variables into the regression equation and eliminate insignificant independent variables [3]. Table. 2 shows the results of the analysis obtained using stepwise regression:

**Table 2.** Variables entered/removed *a*

| Model | Input Variables | Excluded variable | Methodologies                                                                                              |
|-------|-----------------|-------------------|------------------------------------------------------------------------------------------------------------|
| 1     | $X_1$           | .                 | Step (condition: probability of F to be entered $\leq .050$ , probability of F to be removed $\geq .100$ ) |

Table. 2 shows that the input variable is  $X_1$  : number of tourists, where the condition for stepping is that the probability of the input  $F$  is less than 0.05, and the F-test is usually used for the overall significance test of the regression model. If the original hypothesis  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$  is true, it indicates that there is no linear relationship between  $y$  and  $X_1, X_2, X_3, X_4, X_5$  as a whole, in which case the F statistic is small. On the contrary, if the  $F$  statistic is larger than the critical value at the 0.05 level of significance, it means that the original hypothesis is false and  $X_1, X_2, X_3, X_4, X_5$  has a significant effect on  $y$ .

**Table 3.** ANOVA table

| Model |            | Square sum   | Degrees of freedom | Mean square | F      | Significance |
|-------|------------|--------------|--------------------|-------------|--------|--------------|
| 1     | regression | 12592717.008 | 1                  | 1259217.008 | 38.560 | 0.000b       |
|       | residual   | 326556.255   | 10                 | 32655.625   |        |              |
|       | total      | 1585773.262  | 11                 |             |        |              |

Table 3 shows the ANOVA results for each step in the regression fitting process. Significance is the probability that  $F$  is greater than the critical value of  $F$ . The ANOVA results show that when the model includes different independent variables, the probability of significance values is less than 0.001, i.e., there is sufficient evidence to reject the null hypothesis that the regression coefficients are all zero. Therefore, the final model should include these 2 independent variables with a good equation fit.

**Table 4.** Statistics and test results in each step of the regression process for  $y$

| Model |          | Unstandardized coefficient |                | Standardized coefficient | Covariance statistics |              |            |       |
|-------|----------|----------------------------|----------------|--------------------------|-----------------------|--------------|------------|-------|
|       |          | $B$                        | Standard error | Beta                     | $t$                   | Significance | Tolerances | VIF   |
| 1     | Constant | -77.219                    | 184.310        |                          | -0.419                | 0.684        |            |       |
|       | $X_1$    | 0.143                      | 0.023          | 0.891                    | 6.210                 | 0.000        | 1.000      | 1.000 |

From Table. 4 the final multiple regression equation can be derived:

$$y = 0.143X_1 - 77.219 \quad (2)$$

The results show that tourism revenue is ultimately related only to the number of tourists received, and because there is a covariance problem between the number of tourists and the

other four variables, we continue to use the number of tourists as the dependent variable  $X_1$  to perform a multivariate linear stepwise regression with the other four independent variables, and obtain the final results as shown in Table.5.

**Table 5.** Statistics and test results in each step of the regression process for  $X_1$

| Model |          | Unstandardized coefficient |                | Standardized coefficient | Covariance statistics |              |            |       |
|-------|----------|----------------------------|----------------|--------------------------|-----------------------|--------------|------------|-------|
|       |          | $B$                        | Standard error | Beta                     | $t$                   | Significance | Tolerances | VIF   |
| 1     | Constant | 7027.223                   | 462.944        |                          | 15.179                | 0.000        |            |       |
|       | $X_1$    | 0.989                      | 0.241          | 0.792                    | 4.098                 | 0.002        | 1.000      | 1.000 |

The regression equation is:  $X_1 = 0.989X_5 + 7027.223$ , so we can assume that the number of keyword searches affects the number of tourists, and the number of tourists, in turn, affects the total tourism revenue.

### 3. Tourism Industry Development Forecast

In order to study the future development of the main factors affecting the development of Harbin's specialty tourism industry, we decided to use the gray prediction model to predict each factor one by one [4].

Gray forecasting is characterized by the fact that it does not require a large amount of historical data and is suitable for forecasting small samples and incomplete information. This method processes and predicts data by building a gray model, and its core model is the GM (1,1) model, i.e., a first-order univariate gray model. The model is capable of handling time series data, and through the processing of a small amount of known information, potential regularities can be mined, and thus future trends can be predicted.

#### 3.1. Gray Prediction Condition Judgment

First, in order to determine whether the raw data can be used in a gray prediction model, the raw data need to be subjected to a rank-ratio test to ensure that the data have a quasi-exponential pattern:

The sequence after accumulating  $r$  times is said to have a quasi-exponential law if

$x^r = \{x^r(1), x^r(2), \dots, x^r(n)\}$ , defines the rank ratio  $\sigma(k) = \frac{x^r(k)}{x^r(k-1)}, k = 2, 3, \dots, n$ , for  $\forall k, \sigma(k) \in [a, b]$ ,

and the interval length  $\delta = b - a < 0.5$ .

For the GM (1,1) model in, we only need to determine whether  $x^1 = \{x^1(1), x^1(2), \dots, x^1(n)\}$  has a quasi-exponential law. According to the above formula, the rank ratio of the sequence  $x^1$  is:

$$\sigma(k) = \frac{x^1(k)}{x^1(k-1)} = \frac{x^0(k) + x^1(k-1)}{x^1(k-1)} = \frac{x^0(k)}{x^1(k-1)} + 1 \quad (3)$$

Define  $\rho(k) = \frac{x^0(k)}{x^1(k-1)}$  to be the smooth ratio of the original sequence  $x^0$ , noting that,

$$\rho(k) = \frac{x^0(k)}{x^0(1) + x^0(2) + \dots + x^0(k-1)} \quad (4)$$

Assuming that  $x^0$  is a non-negative sequence, then as  $k$  increases, eventually  $\rho(k)$  will gradually approach 0. Therefore, to make  $x^1$  have a quasi-exponential law, i.e.,  $\forall k$ , and an interval length of  $\delta < 0.5$ , it is only necessary to ensure that  $\rho(k) \in (0, 0.5)$  is sufficient, when the sequence  $x^1$  has a rank ratio of  $\sigma(k) \in (1, 1.5)$ .

### 3.2. Empirical Analysis

Take the year-by-year data of the total tourism revenue of Harbin City from 2012 to 2023: [547.1, 662.4, 780.4, 900.9, 1031.2, 1168.4, 1369.1, 1551.9, 1099.9, 943, 499.2, 1692.45] as an example for the analysis, due to the impact of the epidemic, the total tourism revenues of the years 2019, 2021 and 2022, there are outliers in the total tourism revenue, so we remove these three outliers before conducting the gray prediction analysis.

First of all, the quasi-exponential law test, through the calculation of the index 1: the smooth ratio of less than 0.5 of the data accounted for 75%, index 2: in addition to the first two periods, the smooth ratio of less than 0.5 of the data accounted for 100%, in general, we believe that the index 1 is greater than 60%, the index 2 is greater than 90%, you can pass the test. As Figure 1 shown, except for the first two data, the smoothness of the latter data is less than 0.5, so we think that the total tourism revenue data passes the quasi-exponential law test, and we can continue to analyze it using the gray prediction method.

The raw data for making the GM (1,1) predictions are: [574.1 662.4 780.4 900.9 1031.2 1168.4 1369.1 1551.9 1692.45], the least squares fit yields a development coefficient of 0.13156, and the amount of gray interaction is 570.4074. as in Figure 3 shown, the data of the test group are closer to the real data distribution, indicating that the model parameters are accurately estimated, the data are appropriately processed, and the gray model is able to better capture the intrinsic laws of the data.

The forecast of total tourism revenue for the next two periods can be obtained from with the fitted values for the previous years and the forecasts for the next two periods, as shown in Table. 6 and Figure 2.

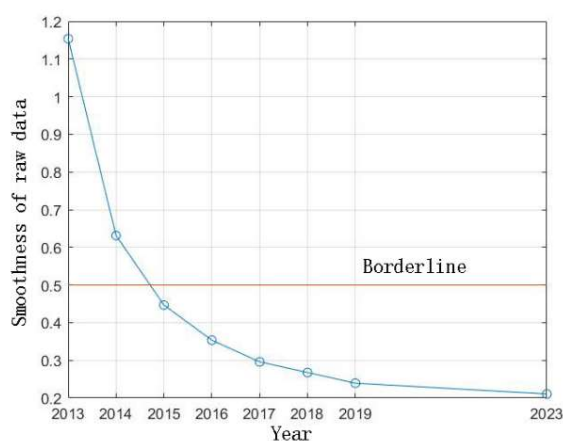


Figure 1. Smoothness graph

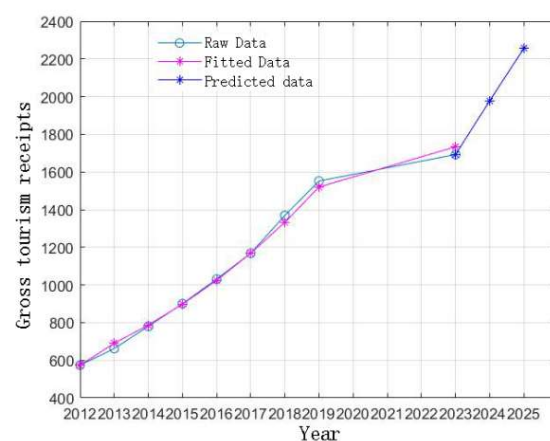


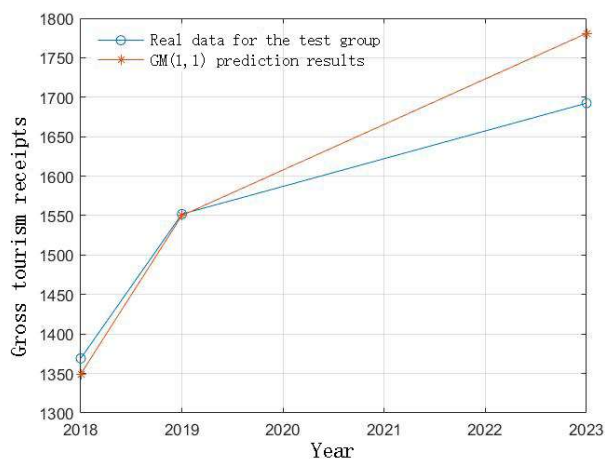
Figure 2. Predicted effect diagram

Table 6. Projections of gross tourism receipts for the next two periods

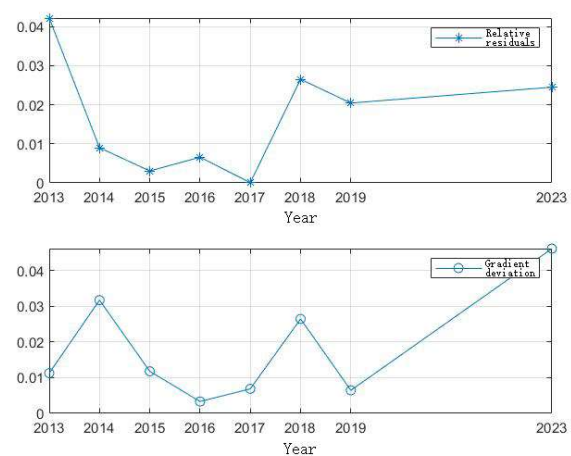
|                 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2023 | 2024 | 2025 |
|-----------------|------|------|------|------|------|------|------|------|------|------|------|
| Original value  | 547  | 662  | 780  | 900  | 1031 | 1168 | 1369 | 1551 | 1692 |      |      |
| Projected value | 574  | 690  | 787  | 898  | 1024 | 1168 | 1332 | 1520 | 1733 | 1977 | 2255 |

Mean relative residuals and mean rank deviation are two important metrics when assessing the effectiveness of gray predictions. Figure 4 shows the graph of relative residuals and rank deviation, from which it can be seen that the relative residuals in each period are within 0.04, and the rank deviation in each period is also within 0.05. The average relative residual reflects the relative error between the predicted value and the actual value, and the smaller the value is, the higher the prediction accuracy is. The average relative residual of the model is 0.016547, which is a relatively small value, indicating that the model fits the original data better. The average level deviation is used to measure how well the prediction model grasps the dynamic characteristics of the original data series, and the average level deviation of the model is 0.017978, which is also a low value, indicating that the model performs well in reflecting the dynamic characteristics of the data.

In summary, based on the values of the average relative residuals and the average rank deviation, it can be concluded that the gray prediction model performs well in this scenario and the prediction is more accurate.



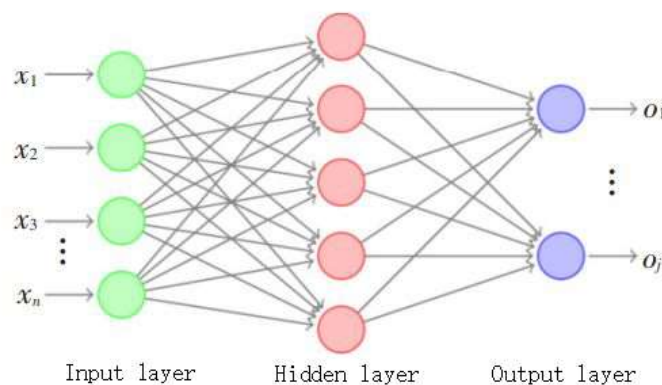
**Figure 3.** Comparison of test group and real data



**Figure 4.** Plot of relative residuals and level deviations

## 4. BP Neural Network Model

### 4.1. Construction of BP Neural Network Principle



**Figure 5.** BP neural network propagation diagram

BP (Back Propagation) neural network is a multi-layer feed-forward neural network trained according to the error back propagation algorithm [5]. It is able to realize various functions

such as classification and regression by learning the mapping relationship between inputs and outputs. As shown in Figure 5, a BP neural network usually consists of an input layer, one or more hidden layers, and an output layer. Each layer consists of a number of neurons, and the neurons between layers are connected by weights.

## 4.2. Solving for Predicted Values

According to the optimal choice of software, we carried out a neural network model with 5 hidden layers:

We first projected the total revenue and found that the projections deviated significantly from the actual results, believing that this result may have been affected by the epidemic, resulting in inaccurate projections.

Therefore, we chose to forecast the total revenue by time series prediction of the number of domestic tourists received, tourism satisfaction, the number of star-rated hotel enterprises, passenger transportation and keyword search index, and combine the prediction results of each index with the previous multivariate prediction model to get the predicted value of the total revenue, as shown in Figure 6 and Figure 7, which is predicted to be 154,363,000,000,000 Yuan in 2024.

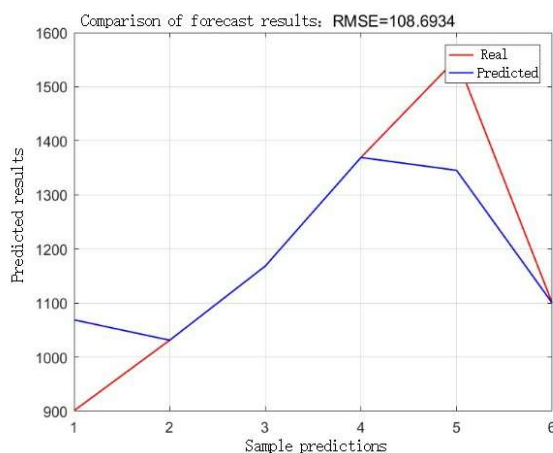


Figure 6. Gross tourism receipts training set chart

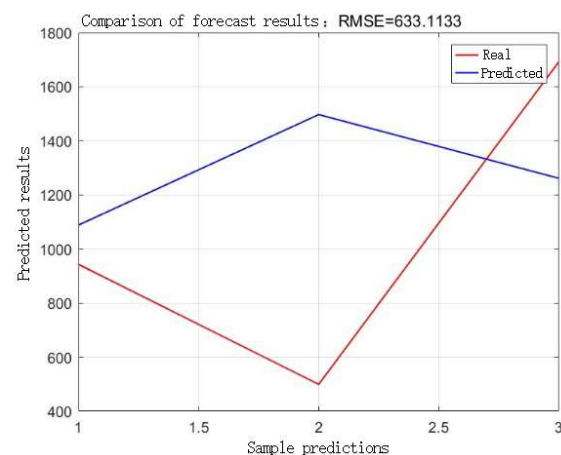


Figure 7. Test set chart of gross tourism receipts

## 4.3. Suggestions

Taking into account the content of the survey and analysis, we make suggestions based on the following aspects:

**Government:** Strengthening the collection and organization of relevant data to collect guiding information for future tourism development. Focus on online publicity and public opinion on the Internet to create a good image of the region. Vigorously explore, develop, and protect local cultural characteristics. Improving tourism infrastructure for young consumers. Emphasis on the training of relevant personnel and cooperation with universities.

**Corporate:** Keeping up with the market trend, grasping the local characteristics, and participating in the construction of the whole tourism industry chain. Actively assume social responsibility and take the initiative to optimize the tourism market environment.

**For the population:** Emphasize the development of the tourism industry and take the initiative to create a favorable tourism environment. Proactive participation in the preservation and promotion of local distinctive cultural industries.

## 5. Conclusion

Based on the tourism consumption habits, network use habits and cognition of characteristic tourism of college student groups, this study explores the development trend of Harbin's characteristic tourism industry through questionnaire survey and various analysis methods by combining network platform publicity with the development of local characteristic tourism projects. The results show that combining network publicity and local characteristic tourism projects is an effective way to promote the recovery and development of tourism. Through multiple regression analysis, gray prediction, and BP neural network model, we predicted the future development trend of Harbin's characteristic tourism industry, and proposed sustainable development suggestions for the government, enterprises and local residents. The conclusions of this study show that mining local characteristics, promoting online publicity, fostering local culture, and attracting young audiences are the keys to promoting the development of specialty tourism. We hope that these research results will provide reference for the sustainable development of tourism in different places and promote local economic prosperity and cultural heritage.

## References

- [1] Uğur N G, Akbiyık A. Impacts of COVID-19 on global tourism industry: A cross-regional comparison[J]. *Tourism management perspectives*, 2020, 36: 100744.
- [2] Hoque A, Shikha F A, Hasanat M W, et al. The effect of Coronavirus (COVID-19) in the tourism industry in China[J]. *Asian Journal of Multidisciplinary Studies*, 2020, 3(1): 52-58.
- [3] Wubetie H T. Application of variable selection and dimension reduction on predictors of MSE's development[J]. *Journal of Big Data*, 2019, 6: 1-44.
- [4] Zhang M, Zhao H, Fan L, et al. Dynamic modulus prediction model and analysis of factors influencing asphalt mixtures using gray relational analysis methods[J]. *Journal of Materials Research and Technology*, 2022, 19: 1312-1321.
- [5] Panda S, Panda G. Fast and improved backpropagation learning of multi-layer artificial neural network using adaptive activation function[J]. *Expert Systems*, 2020, 37(5): e12555.