Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 12, No. 3, 2023

Prediction of Future Tourism Number in Hainan Region Based on Least Square Fitting

Jinghuai Zheng[1, a]

[1]Wuhan British-China School, Wuhan, 430030, China
[a]zhengjinghuai10@163.com

Abstract: Tourism is an important part of the city's income and its importance to the country is beyond doubt, not only contributing to the national economy and employment opportunities, but also promoting cultural exchanges and mutual understanding between people, and promoting social development and progress. Therefore, we should attach importance to the development of tourism, formulate reasonable policies and measures, and create a good tourism environment. Among them, predicting the total number of visitors in the short term in the next five years, so as to formulate more appropriate tourism-related policies and guidelines can not only better improve the overall income of the city, but also an effective way to improve people's living standards. In this paper, we calculate the correlation coefficient to predict the tourism flow of Hainan in the next five years, and apply the least square method to further improve the functional prediction image for this problem.

Keywords: Least square method, polynomial fitting, number of tourists, K-fold cross verification, root-mean-square error.

1. Introduction

1.1. Background introduction

In a country, the contribution of tourism to the economy is huge, and at the same time tourism is one of the fastest growing industries in the world, it directly and indirectly creates a large number of jobs, provides a source of livelihood and income, and brings stability and considerable revenue to the national Treasury. Among them, tourism also affects many related industries, such as hotels, catering, transportation, etc., which promotes the development of the entire national economy.

In addition, tourism is an important force driving social development and improving living standards. In order to attract tourists, the country needs to provide good infrastructure and services and introduce reasonable policies and subsidies, which include infrastructure construction, transportation improvement, environmental protection and other fields. At the same time, tourism also drives the development of related industries, such as handicrafts production, traditional cultural performances, etc., promoting the prosperity and inheritance of cultural industries.

Tourism for a city is not only a non-negligible income component, but also promotes cultural exchange and promotes social development and progress. [1]Therefore, we should attach importance to the development of tourism, among which the prediction of the number of tourists is an important step to effectively promote the development of tourism for a city. For different sizes of reception traffic, the policies required to be issued are also very different. For the data prediction of the flow of people, the establishment of the model and the elimination of errors are important factors that affect the accuracy of the estimation. In this process, it is particularly important to choose a suitable model and verification method for such data. For this problem, Hainan Province of China is taken as the research area of this paper, and the least square fitting method and k-fold cross-validation method are used to verify it.

2. Least Square Method

2.1. Basic Introduction

By minimizing the difference between the real target object and the fitted target object, the least square method uses the square of the multiplication to find the best function match of the data, that is, to find a line that minimizes the sum of squares of the difference of the ordinates of the observed data points on all the graphs, and also minimizes the variance. This difference can be measured by mean square error or least squares residual.[2] This method aims to find a set of coefficients that make the regression model best fit the given observational data, effectively eliminating possible errors in the experimental data.

Given a set of experimental data denoted D={(x1,y1), (x2, y2), (x3, y3),,, (xn, yn)}, find a function f(x)=ax+b such that the f(x) function fits D as best as possible and minimizes its squared variance Q, i.e:

$$Q = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \qquad (1)$$

2.2. Data Collection

Through the collection of Internet information, the existing tourism data of Hainan in recent years are summarized as follows. In this paper, the following data (Table 1) will be analyzed and processed by the least square fitting method and the data fitting and modeling function of MATLAB software. (Due to the outbreak of COVID-19 at the end of 2019, tourist arrivals in the region during 2020-2022 will not be included in this analysis)

Table 1.

Year	Total number of overnight visitors received(10,000)
2001	1124.76
2002	1254.96
2003	1234.1
2004	1402.89
2005	1516.47
2006	1605.02
2007	1845.51
2008	2060
2009	2250.33
2010	2587.34
2011	3001.34
2012	3320.37
2013	3672.51
2014	4789.08
2015	5336.52
2016	6023.6
2017	6745.01
2018	7627.39
2019	8311.2
2020	957.19
2021	814.39
2022	652.69

(Data source: Big Data Navigation)

2.3. Research Methods

First of all, the research theme and the conclusions to be verified are determined, and then the previous literature is retrieved through websites such as CNKI and scihub to find articles related to the research theme or method, and then sorted out and summarized. By using the least square method widely used in population prediction problems and the powerful calculation and modeling ability of matlab, the existing real data are quasi merged to generate function prediction point images. At last, the possible errors in the research process are eliminated by cross-validation method.

3. Model Establishment and Solution

3.1. Mathematical model through least square method

In this revised code, we use a quadratic polynomial n=2 to fit, and by setting the first column of the X-matrix to two we get a curvilinear function that minimizes the gap between the function and the observed value.

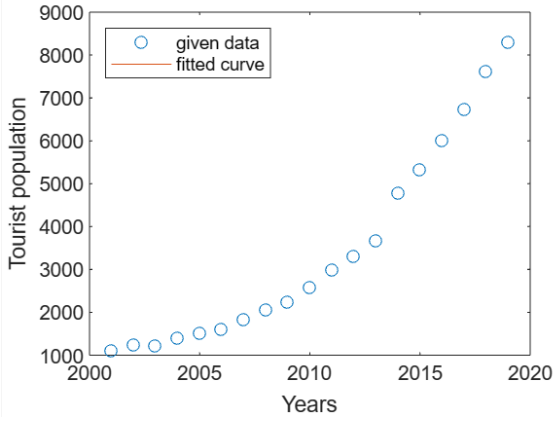

Figure 1.

Through MATLAB data processing and image generation above (figure 1), we can get the trend of tourist population in this region in recent years. It can be seen that the number of tourists received by this region has maintained an upward trend in recent years, and when the trend is relatively stable, the least square method can be used to fit the data and build the model including the prediction points

3.2. Generation and prediction of fitting curve

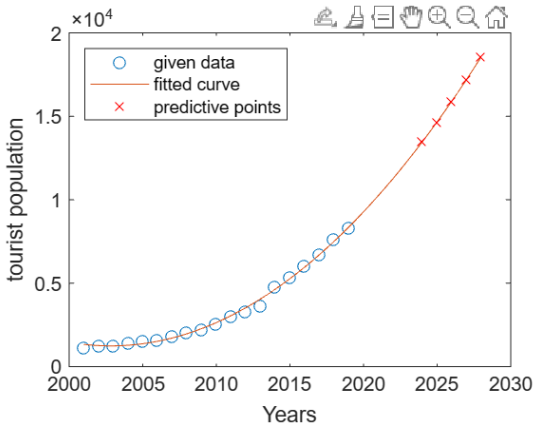

Figure 2.

The above figure 2 is a fitting curve image based on the observed data of the number of tourists from 2001 to 2019. Its code uses cubic polynomials for fitting, which can more accurately adapt the fitting curve and predicted value of the data. From the trend of the fitted curve in the figure, it can be clearly seen that the popularity of tourism in this region is continuously and steadily increasing.

Through the fitting calculation of this image, in the next five years (2024-2028), the number of tourists in the region is:

2024: Projected tourist population: 13,478.77 (million)

2025: Projected tourist population: 146,679.4 (million)
2026: Projected tourist population: 15,912.86 (million)
2027: Forecast tourist population: 17,213.63 (million)
2028: Forecast tourist population: 18,570.33 (million)

Of course, there are also many factors that affect the number of tourists, such as the economic situation, policy changes, technological development and so on. Here, this article is only for the future tourism number forecast reference.

3.3. Impact of national GDP and per capita GDP on the number of tourists

Table 2.

Years	Total GDP (trillion)	GDP per head(10000)
2001	11.009	0.87
2002	12.17	0.95
2003	13.74	1.07
2004	16.18	1.25
2005	18.73	1.44
2006	21.94	1.67
2007	27.01	2.05
2008	31.92	2.41
2009	34.85	2.62
2010	41.21	3.08
2011	48.79	3.63
2012	53.86	3.98
2013	59.3	4.35
2014	64.36	4.69
2015	68.89	4.99
2016	74.64	5.38
2017	83.2	5.96
2018	91.93	6.55
2019	98.65	7.01

(Data source: China Economy Data)

Along with China's rising GDP in (table 2), the disposable income of each Chinese is also rising. Among them, when people have a higher disposable income, travel is an effective way to improve the quality of life. From 2001 to 2019, more and more Chinese local tourists chose to travel in various holidays, which directly led to the continuous increase in the number of tourist receptions in the study area, and the total income of cities in the region also increased.

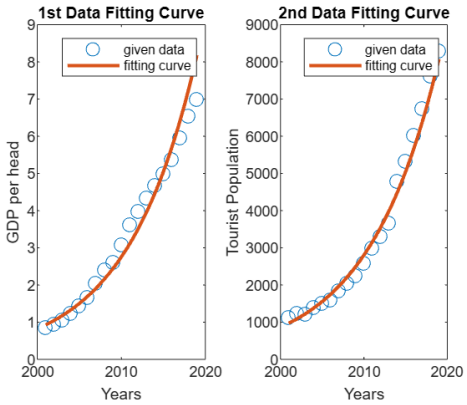

Figure 3.

The graph above (figure 3) shows the curve of per capita GDP and tourist arrivals in recent years.

As can be seen from the figure, the linear trend of the disposable income of the people in the whole country and the number of receptors in the study area of this paper is very similar, with an upward curve.

4. Reduce Error and Verification

4.1. Part fitting and prediction

According to the data surveyed in this paper, from 2001 to 2010, the number of tourist arrivals in the region also showed a continuous upward trend, just like the overall data.

In order to verify the feasibility and accuracy of the calculation method mentioned in this paper, we first selected the data of the first 10 years from the data of the past 20 years as the known data, and used this set of data to carry out data fitting and least squares model establishment in Matlab,[3] and obtained the predicted value of the number of tourists from 2011 to 2013 from this set of data. They then compared the predicted value of the model with the actual number of tourist arrivals in the region from 2011 to 2013. It can be concluded that the model's ability to evaluate data and the model's ability to process data.

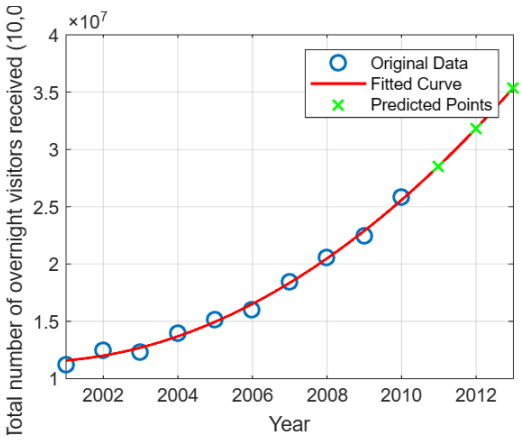

Figure 4.

From the model image (figure 4) above, it can be seen that the model predicts the number of tourists in the region between 2011 and 2013 as follows:

2011:25,871,900
2012:31,836,200
2013:35,385,200

The actual number of tourists from 2011 to 2013 is as follows:

2011:30,013,400
2012:33,203,700
2013:36,725,100

Through the comparison between the above predicted value and the real value, it can be concluded that the predicted value generated by the model in the prediction process is in good agreement with the real value. For example, in 2013, the real number of people is 36.7 million, while the number predicted by the model is 35,380 million, with only 3.8% error. In 2012, the actual number was 33.2 million, and the estimated number was 31.8 million, with a relative error of 4.3%; In mid-2011, the actual number was 30.1 million and the forecast was 25.87 million, with a relative error of 16%. Through calculation, it can be seen that the average error does not exceed 10%, and the error size is still acceptable in this paper.

4.2. k fold cross verification method

When we want to fit a model multiple times, we can use cross fitting to reduce the errors implied in the model.

When the amount of data is not large enough, K-fold cross-validation can be a good solution to this problem. For some shortcomings and shortcomings of simple cross-validation, K-fold cross-validation can be further improved and improved.

For example, simple cross-validation has the problem of insufficient utilization of data. When the user simply groups the collected observation data and verifies it according to the method, there is a strong correlation between the final result and the user's grouping method, which leads to the limitations of this method.

In order to solve the shortcomings of this simple cross-validation method, K-fold cross-validation divides the observation data sample into k subsets with equal memory size, and then completely traverses the contents of this k subset. [4]

In addition, the contents of the previous subset will be used as the verification set for each operation, and all the other samples will be used as the training set to verify and evaluate the model. Finally, the average value of k evaluation indexes is used as the final evaluation index of the model to verify. In general, the value of k is 10 in the K-fold verification method.

4.3. Holdout method

The holdout method divides the data into two parts, the first part is the training set and the second part is the test set. After the division is completed, the appropriate model is selected and then validated and evaluated

In the processing of this set of data, the data since 2001 is divided into two parts. In this paper, the first half of the data is used as the training set, and the second half of the data is used as the verification set. Polynomial regression method is used for data fitting and chart construction, like the figure 5 below.

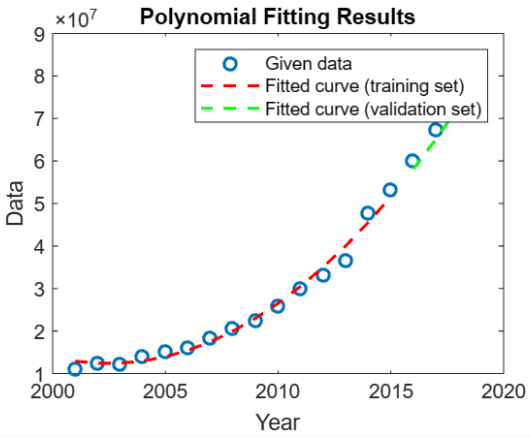

Figure 5.

4.4. Root Mean Square Error (RMSE)

Root mean square error (RMSE), also known as the standard error, is often used to describe the difference between the real data and the predicted results of the model. The smaller the number of the value, the smaller the difference between the predicted value and the real value.

It measures the average size of the prediction error of the model. It is sensitive to abnormal data in the same set of linear data (such as values strongly different from other data sizes), so the root-mean-square error can well reflect the precision and stability of the measurement. The calculation method is to find the sum of squares of the difference between the real value and the predicted value, then divide this data by the sample size, and finally take the square root. The calculation formula is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{obs,i} - X_{model,i}\right)^2}{n}} \quad (2)$$

At the end of K-fold cross-validation, a root-mean-square error value (RMSE) is usually generated. Because in the K-fold cross-validation, the data will be divided into k subsets first, and then the simulation and training of the k subsets are traversed. In each validation, one of the k subsets will be used as the validation set, and the other subsets will be used as the training set, and finally the exact value evaluation results of k models will be obtained. In each validation, the difference between the predicted result and the true value can be used to calculate the RMSE value, and the average of k root-mean-square error values can be calculated as an effective indicator for the final model performance evaluation.

Therefore, the method of generating a root-mean-square error value through K-fold cross-validation can successfully and effectively measure the fitting ability of the model and help us judge the prediction level and error size of the model.

When K-fold cross-validation is performed on this set of data, it is also particularly important to find a suitable k value. Generally speaking, the value of k is 5 to 10, but the specific choice should be determined according to the size of the value. When the known data is small, a smaller value of k is more appropriate for this set of data; Conversely, when the known data is large, the value of k will be as large as possible, in order to balance the size of the data set and its variance (table 3).

Table 3.

K	RMSE
5	0.60704
10	0.4532
15	0.40799
20	0.37157

According to the powerful modeling and computing power of matlab, when k=5 in the above formula, the root-mean-square error value is 0.60704; When k=10, the root-mean-square error is 0.4532. The error value is kept within an acceptable range for the size of the data in this sample.

4.5. Comparison of two verification methods

K-fold cross-validation method can make better use of all known data and improve the accuracy of model evaluation by traversing the training set and test set of each group of partitions. At the same time, cross-validation can also better prevent the occurrence of overfitting during verification, which also represents that the method is relatively more accurate and stable.

However, K-fold cross-validation method also has some shortcomings, such as consuming a relatively large amount of time and resources, and when the data sample size is small,

this method may have some data bias. Due to the small number of collapsible samples, cross-validation may not provide particularly accurate verification data.

Compared with K-fold cross-validation method, holdout method is a simpler and more direct data validation method. First, the holdout method only divides the data into two mutually exclusive subsets to verify and evaluate the data. Second, since the holdout method only preserves a portion of the data set as the test set, it

It can better avoid the problem of data deviation caused by the limitation of data quantity.

However, holdout method still has some problems, such as low data utilization and difficult parameter tuning.

In general, holdout method and K-fold cross-validation method each have their own advantages and have their own areas of expertise. In this paper, after the evaluation and verification of these two verification methods, the fitting model becomes more efficient and accurate.

5. Research Conclusions

In this study, Hainan Province of China is taken as the research area, and the number of tourist arrivals in this area since the beginning of the 20th century is fitted and verified based on the calculation and modeling ability of matlab. Among them, the least square method is better than other methods in the case of not so large data to fit, the principle is to find a relatively simple mathematical model to describe the trend and direction of the data, without considering too much complex statistical methods.

But at the same time, the least squares fitting method will still be less ideal for data processing in some scenarios, such as overfitting or underfitting. Therefore, in order to avoid the occurrence of such events to the greatest extent, this paper adopts K-fold cross-validation method to eliminate such possible errors. Through the verification of this set of data, it can be concluded that the gap between the fitted curve and the known data has always remained in a relatively small range, which also means that the number of tourists to the study area in the next five years will steadily increase along with its previous trend, and the increase will show an acceleration growth state.

References

[1] Cheng, P. (2022). Accuracy analysis of data fitting and data interpolation in missing information supplement surveying. Mapping and Spatial Geographic Information, 12: 127-129.

[2] Huang, W. (2020) Prediction of the number of tourists in Weifang City based on GM (1,1) model. Advances in Applied Mathematics, 9(6).

[3] Zhao, H., Fang, W. (2019). Predicting tourist number of Guangdong Province based on fractal Auto-regressive model. Statistics and Application, 8(1).

[4] Gong, H., Chen, J., Xiong, W., et al. (2022). Estimation of forest stock using local sample optimal K-value KNN model. Journal of Northeast Forestry University, 11: 52-56.