Predicting number of tourists within peak seasons in Jiuzhai Valley

A combination model of linear regression and BP neural network

Qin Tianzhu
Southwestern University of Finance and Ecnomics,
Maple_Optboy@outlook.com

Pu Ziyi
Southwestern University of Finance and Ecnomics,
Puzy9188@gmail.com

Ye Juelin
Southwestern University of Finance and Ecnomics,
yeeejlin@gmail.com

Abstract

Since there exists a conflict between scenic reception capacity and peak season passenger flow, it is necessary to predict number of tourists in advance especially in peak seasons. The traditional approach is to forecast passenger flow according to regional GDP, consumption and local income level, which lacks accuracy and validity for short terms predictions. However, this article is able to forecast the number of visitors within one certain day in Jiuzhai Valley, and can greatly improve the management efficiency and resource utilization specifically in boom seasons. This is achieved by utilizing the Baidu Index, web crawler and text mining technology and implementing the combination of multivariate linear time series econometric model and BP neural network algorithm. In our experiment, the MSE value of the integrated model has been significantly decreased, indicating a more accurate predictive performance.

Keywords: Search Index, Passenger Flow, Neural Network, Text Mining

1. Introduction

Jiuzhai Valley attracts millions of tourists every year with its splendid scenery and natural Tibetan customs from all over the world and brings huge economic benefits to the local area. In 2016, the annual reception of overseas tourists exceeded 5 million passengers and tourism revenue reached 805 million yuan. However, there is an enormous gap in the seasonal passenger flow variation trend in Jiuzhai scenic area, which will easily lead to inappropriate management strategies and cause severe retention problems. For example, in October 2, 2013, thousands of tourists stranded in Jiuzhai Vally scenic spot, causing traffic lines collapse and several kilometers long congestion. Therefore, it is of significant importance to accurately predict the tourists' volume to improve the service level of the Jiuzhai tourism industry.

Previews work mainly focus on carrying out econometric methods in Index Prediction (Xin, Pan and Evans, 2015; Gawlik, Kabaria and Kaur, 2011), and the search words selected are mostly acquired by literature review (Huang, Zhang and Ding, 2013). This work takes all these into accounts and solve them in an innovative way.

With the popularization of Web media, the Internet is not only a platform to disseminate information but also an interaction platform for users to provide feedbacks via posting, votes, and search behaviors. This provides a valuable channel for us to predict the number of visitors in tourist attractions.

This article first uses the web crawler and text analysis techniques to find keywords people most tend to associate with and concerned about Jiuzhai Valley. After extracting the corresponding keywords from the Baidu search index and set them as independent variables, this work builds an initial multiple time series linear regression prediction model. In order to improve the forecast accuracy and validity especially for peak season, this work then corrects the residual error by applying BP neural network algorithm. By comparing the MSE index of both models, this work finds the integration model of linear and BP neural network modification parts obtains a better performance in predicting the seasonal visiting volume in Jiuzhai Valley.

The unique contribution of this study is as follows:

1. Utilize web crawler techniques and search index to outline the scenery search word profile.
2. Propose a combined model to realize this goal by merging linear time series econometric model and BP neural network algorithm.
3. This study focuses on the Jiuzhai Valley tourist flow but it is applicable to other scenic areas.

This paper aims to optimize the forecasting model of passenger flow through the search index of network keyword, and try to establish a more reliable model especially designed for Jiuzhaigou passenger flow prediction.

The remainder of this paper is organized as follows: Section 2 illustrates a review of the existing literature on relevant studies that provide theoretical reference for our prediction model. Section 3 presents the process of data crawling and processing, the establishment of tourist volume prediction linear model, non-linear residual improvement and its empirical results comparison. Section 4 outlines the evaluation of experimental model results. Conclusion summary and the implications of our work are identified in Section 5. Finally, the last section provides limitations, followed by suggestions for future research.

2 Related Work

This section reviews relevant studies in the areas of tourism demand and some work by Search Index. Besides, other data mining technologies related to tourism industry are also discussed.

2.1 Tourism Volume Prediction

With the rapid development of tourism, the research on the prediction of tourist volume has attracted wide attention. Scholars have built various qualitative and quantitative methods and models to predict the tourists' volume among different destinations. Liu and Li (2014) take insight into geography tourism flow in Yunnan Province. Du, Wang and Tu (2008) try to uncover the rule of tourism flow in holiday and establish the theoretical basis for the prediction. Tse (1999) takes a study based on cross-sectional data from 32 countries gathered between 1990 and 1995 which examined the interaction between tourist flows, expenditure and receipts, and domestic consumption. Among these tourism prediction research, the variables are simply economic indexes but lack the thoughts of socio-

economic behaviors. So far, user engagement in social media promotes over time which provides a cherished chance to take socio-economic into consideration.

2.2 Search Index Prediction

Technology of search engine growing over time, people found a certain relationship between the network search behavior and the actual events. Choi and Varian (2012) use Google Trends to predict the sales of motor and parts, unemployment benefits, number of visitors and even consumer confidence which are all applicable and valid in the econometric way. However, they just concern the query which is the right name of the predictor. Xin, Pan and Evans (2015) predict Hainan tourism volume by Baidu Search Index. Additionally, Huang, Zhang and Ding (2013) even research into one specific point, Fobidden City by Baidu Index. Xin, Pan and Evans (2015) use auto-regression moving average only. Gawlik, Kabaria and Kaur (2011) predict Hong Kong tourism by simply auto-regression.On the one hand they just implement original statistics model which dose not apply to high dimensional data's non-linear part. On the other hand, all of the previous work lack the explanation why the words in Baidu Index should be selected except for literature reviews, this work fills the gap by joining web crawler and text mining to look for the words related.

Some researchers take a further step by applying other data mining methods for search index prediction. The classification tree models are built to predict whether *Dengue* would happen in Guangzhou or Zhongshan (Liu et al., 2016). Whereas the CART analysis needs more features to avoid underfitting problem, which is still limited in Liu's work.

3 Model establishment and solution

This research tries to find the words related to *Jiuzhai Vally* by text mining first in order to get the scenic spot search word profile. Secondly, the time-series analysis is implemented to verify the linear part of the model. At last, the non-linear part is modified by BP neural network.

3.1 Model Overview

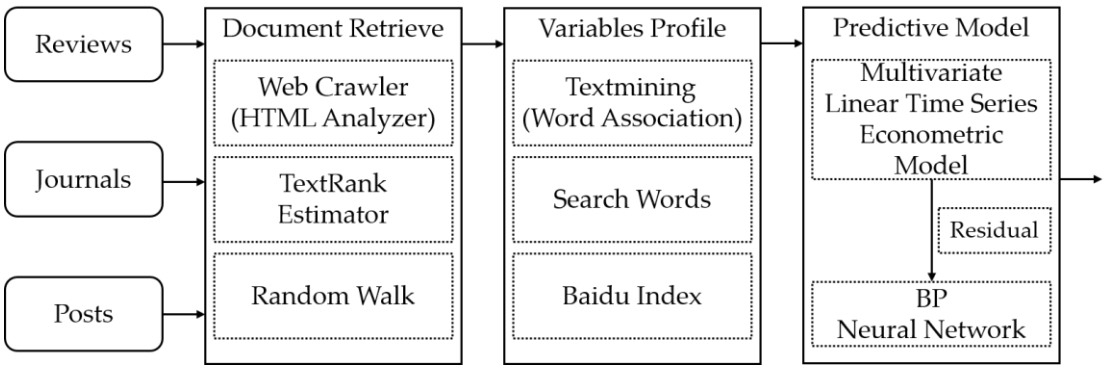


Figure 1. Model Overview

The Document Retrieved part including Web Crawler equipped with TextRank Estimator and Variables Profile will be explained in next section. Section 3.3 provides linear time series regression model construction process. And section 3.4 goes deep into the integration predictive model modified by BP neural network.

3.2 Search Word Profile construction

A simple way to forecast tourist volume by using search index is to query the search engine with just the word *Jiuzhai Valley* (Choi and Varian, 2011). But this will cause that only one variable could be use in the model. One way to increase forecasting accuracy is by incorporating more powerful predictors (Yang, Pan and Evans, 2014). This work looks for the relevant query by two steps, i.e., relevant documents retrieved and search words profile construction.

3.2.1 Relevant documents retrieved

What people write in a text with one word tends to be sought in search engine when enquiring that word (Choi, 2011; Yang, 2014). For example, words people would like to search with Jiuzhai Valley in the web search engine tend to occur frequently in the text concerned with Jiuzhai Valley. Thus, this work deploys a web crawler in the heuristic depth-first search way to grab the texts about Jiuzhai Valley on six Chinese largest travel social platform, i.e., Qunar, Ctrip (Li & Ma, 2013), Mafengwo, Zhihu, Sina and Baidu Tieba.

1. Web Crawler Deployment

General webs can be divided into index pages and content pages (Guo et al. 2010; Kim and Kuljis, 2010). The former, say Qunar homepage, mainly provides other webs' entrances. So, in its HTML codes, sizes of text between *tags* are fairly uniform. While the latter is the pages containing essays or reviews about *Jiuzhai Valley*, and the sizes of text between *tags* diverse sharply among its HTML codes (Gao, Wu, & Zhang, 2016). The crawler tests the codes every time reaches a new page and grabs the text when meets the following condition:

$$\theta_i > \vartheta \quad i = 1, 2, 3, \ldots$$

Where θ_i is the standard deviation of sizes of text between tags, and ϑ is the a priori parameter to distinguish index page and content page.

Consequently, the larger-size text will be extracted for future analysis.

2. TextRank Estimator

After filtering index page by HTML analysis, texts related to other travel information still exists. Therefore, this work implements TextRank Estimator, which is a graph-based ranking model for text processing (Mihalcea & Tarau, 2004). Every text will be assigned a word that has the highest ranking which acts significantly and usually represents the topic of one text. Then, the estimator will match this word above to *Jiuzhai Valley* (In Chinese character). If they are not able to match up, this text will be abandoned. The mechanism descripted is as follows

$$ans_i = \begin{cases} 1 & \arg\max(TR_{ij}) = \text{'}Jiuzhai\ Vally\text{'} \\ 0 & \text{else} \end{cases} \qquad i \in \mathrm{I},\ j \in \mathrm{J}_i$$

Where ans is a binary variable, in which 1 means retaining the text while 0 means abandoning it. I is the text set and, J_i is the word set in text i. TR_{ij} means the TextRank of the j_{th} word in the i_{th} text.

3. Random Walk

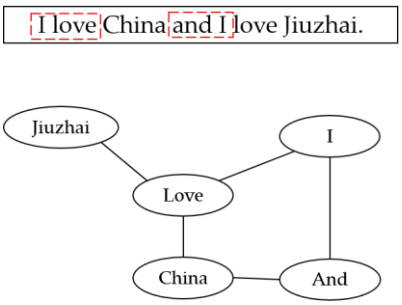
The crawler excavation depth should be restricted in consideration of the potential text's validity loss. Since Random Walk would make crawler more efficiency (Hassan, 2012), we use this algorithm to avoid sinking into dead circulation, i.e., a crawler should jump from Qunar to Ctrip in an a-prior probability model. This work controls the noise texts by cohering HTML Analysis and TextRank Estimator in random walk way which has rather stable results.

3.2.2 Construction search words profile

This work implements word association algorithm to the text dataset in a syntagmatic way. The word's likelihood to be converted from another word is defined as association between them, which is calculated by:

$$p(k \mid j) = (1 - s)\frac{C_{jk}}{\sum_{ij} C_{ji}}$$

Where C_{jk} describes the spatial distance between word j and word k. $\sum_{ij} C_{ji}$ is the spatial distance between word j and all its adjacent vertex. s is a damping variable that measures the degree that one word associates to others (Rahman et al., 2010; Raghavan & Tsaparas, 2002). As an example, shown in Figure 2, if two words are back to back, they construct a connect. The whole text would form an undirected graph and the spatial distance could be calculated.



	I	Love	China	And	Her
I		1	2	1	2
Love	1		1	2	1
China	2	1		1	2
And	1	2	1		3
Her	2	1	2	3	

Figure 2. An example of word association algorithm

After six iterations by this algorithm (Six Degree Separation Theory), every word in a text gets a score representing its association with *Jiuzhai Valley*.

Taking experiments to the whole text, the work gets an overall result. The alphabet format of the words associated with Jiuzhai Valley, which is also most likely to be sought when searching, are listed below. (See in Table 1)

Table 1. Explanatory Variables

Classification	Name	AlphaBet	Classification	Name	AlphaBet
SCENERY	X1	JiuZhaiGou	SERVICE	X12	JiuHuangJiChang
	X2	JiuZhaiGouJingQu		X13	JiuZhaiGouJiChang
	X3	JiuZhaiGouFengJing		X14	JiuZhaiGouMenPiao
	X4	JiuZhaiGouJingDian		X15	JiuZhaiGouLvYouJiaGe
	X5	ShuZhengGou		X16	JiuZhaiGouTianQi
	X6	RiZeGou		X17	JiuZhaiGouZhuSu
	X7	ChangHai		X18	JiuZhaiGouJiuDian
	X8	WuHuaHai		X19	JiuZhaiGouBinguan
	X9	JiuZhaiGouJingHai		X20	JiuZhaiTianTangZhouJiDaFanDian
	X10	NuoRiRangPuBu		X21	JiuZhaiGouGongLue
	X11	WuCaiChi		X22	JiuZhaiGouXingCheng
Total: 25				X23	JiuZhaiGouXianLu
				X24	JiuZhaiGouSanRiYou
			HUMANITY	X25	ZangMi

3.3 Linear Time Series Regression Model

The steps of linear model establishment are as follows:

3.3.1 Lag order confirmation

This work considers lagging variables due to the facts of un-synchronization between searching keywords behavior and the real traveling actions. Therefore, we do linear regression alone with the variable Y and the independent variable X of each lag order (limited maximum up to 15), and choose the most reasonable lag order when R-squared figures maintain the highest.

Through experimenting, we can see that the best lag orders of keywords are distributed from 0 to 7. For example, the lag order of the key word "Three-day tour in Jiuzhai Valley" is 4, and that means people are inclined to search information via Internet of that keyword around 4 days before their visit to Jiuzhai Valley.

3.3.2 Grainger causality test

This work uses Grainger causality test to verify the causal relationship between variables, and finds the keywords and the dependent variables are well explained in Grainger causality test. However, the independent variables X9 and X14 have no Granger causal relationship with Y. Hence, X9 and X14 will be removed from this model.

3.3.3 Stationarity test

For the model construction of time series, cointegration test is needed to determine whether long-term equilibrium relationship exactly exists.

3.3.4 Selection of autoregressive model

Although the visitor data varies dramatically influenced by holiday and vacations, we suppose a convergence of certain stable states in the number of tourists during closely adjacent periods. Accordingly, this work introduces the lag order of the explained variable into our linear model.

To validate this proposal, this work first does the regression with Y and lag variables of x1-x25, and gets the distributed lag linear model results both with and without adding the lag order of the dependent variable.

3.3.5 Cointegration Test

This work does cointegration test on the multivariate autoregressive model. Apart from the RESET test, this paper uses the data in 2016 to examine the predictive value with linear model and found the forecast deviation of tourists' number is larger in peak period seasons. Nevertheless, the booming season in popular spots are supposed to be paid close attention to by relevant departments. Therefore, in order to improve the effectiveness of practical applications of our model, this paper introduces the BP neural network to further revise and correct the prediction errors of the existing linear model, so as to achieve more accurate prediction results in the peak season.

3.4 BP model

The back-propagation algorithm (BP) is a well-known method of training a multilayer feedforward artificial neural networks (FFANNS). The algorithm repeats a two-phase cycle, namely propagation and weight update. When an input vector is presented to the network, it is propagated forward through the network layer by layer until it reaches the output layer.

This article sets errors in linear model during 2011-2015 as the training set, and errors in linear model in 2016 as the testing set. Input every 7 items and output the next errors. Then this article combines both linear and nonlinear parts, so as to gets more accurate prediction results in the peak season.

To fit the errors accurately, this article builds a BP neural network with one input layer, one output and one hidden layer.

3.4.1 Preprocessing data

The transfer function is the Sigmoid function in BP neural network. To improve the learning speed and sensitivity and avoid the saturation region of the function, this article first normalizes the input data. After that, we can inverse transform the output data to get the actual data set.

3.4.2 Choosing parameters of BP neural network

Considering input data of this model are continuous real numbers, and its output data are numbers between 0 and 1. This article sets training function as the Trainlm function, simulation network function as the Simuff function, the max iterations as 100, model error goal as e-7, and learning rate as 0.00001.

3.4.3 Choosing number of nodes in each layer

In general, the input nodes and output nodes equal actual needs, so this article applied 7 input nodes and 1 output node. According to the empirical formula $\sqrt{n+m+a}$ (n represents input nodes and m represents output nodes) and repeated experiments, this article finally applies 4 middle nodes for the hidden layer.

3.4.4 Training BP neural network and predicting the errors

This article sets errors r in 2011-2015 as the training set, and errors in 2016 as the testing set. We input every 7 items and output the next errors, and stop the training when testing set and training set show the same fitting trends.

4 Experimental evaluation

4.1 Data Statistics

This work first calculates the number of people entering Jiuzhai Valley from September to November in 2012 to 2015, and draws the corresponding line charts in Figure 3.It can be seen from Figure 3 that the fluctuation trends of daily entering tourists number in peak seasons (Sep. to Nov.) are basically similar within different years. Since predicting the annual surge in tourist volume is of great significance, we would go on data prediction of peak period by first establishing models with data from 2012 to 2015, and then testing the model performance and making further modifications.

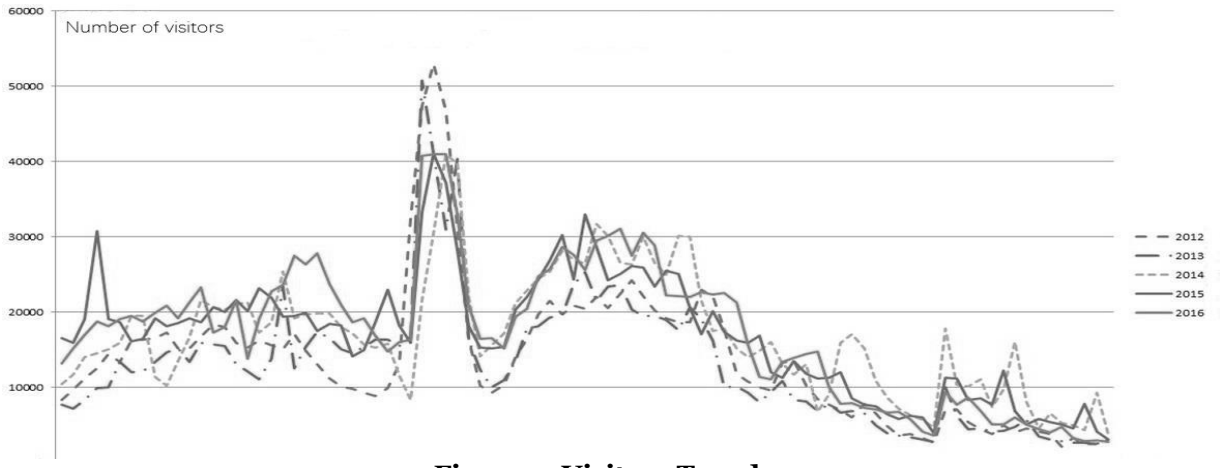


Figure 3. Visitors Trends

4.2 Results of linear model

The result of stationarity test (See in Table 2) shows this model satisfies the cointegration condition.

Table 2. Stationarity Test

Variable	Lag	Process	Stationary Order	Variable	Lag	Process	Stationary Order
x1	2		I(0)	x13	7		I(0)
x2	1		I(0)	x14	/	Delete	/
x3	0		I(0)	x15	3		I(0)
x4	2		I(0)	x16	2		I(0)
x5	4		I(0)	x17	3		I(0)
x6	4		I(0)	x18	4		I(0)
x7	3		I(0)	x19	3		I(0)
x8	0		I(0)	x20	4		I(0)
x9	/	Delete	/	x21	3		I(0)
x10	3		I(0)	x22	5		I(0)
x11	0		I(0)	x23	4		I(0)
x12	4		I(1)	x24	4		I(1)
				x25	3		I(0)

The distributed lag linear model result without adding the lag order of the dependent variable is shown in Table 3.

Table 3. Distributed Lag Linear Model

Variable	Lag	Coefficient	Variable	Lag	Coefficient
c		**-11818.62**	X13	7	**-14.50228**
XI	2	**0.172951**	X15	3	**-1.455483**
X2	1	**12.02738**	X16	2	**2.368873**
X3	0	**-7.36113**	X17	3	**7.012173**
X4	2	**19.67585**	X18	4	**9.672263**
X5	4	**5.92802**	X19	3	**10.00934**
X6	4	(1.404058)	X20	4	**7.628798**
X7	3	(1.791216)	X21	3	**12.84178**
X8	0	**18.70076**	X22	5	**0.091138**
X10	3	**-6.264255**	X23	4	(0.724518)
Xll	0	**1.669824**	X24	4	**2.796431**
X12	4	**6.68713**	X25	3	(5.607753)

From Tab.4, it can be obtained that the value of R-squared is 0.81, and DW is 0.82, which indicates a strong positive autocorrelation and severely undermines the basic classical assumptions of linear models.

$$Y = c + \sum_{i=1}^{25} \alpha_i X_i(Lag) + e_i$$

Then we have the results after adding the lag order of the dependent variable in Table 4.

Table 4. Revised Regression

Variable	Lag	Coefficient	Variable	Lag	Coefficient
c		**-2727.192**	X13	7	**-8.351559**
XI	2	(0.050518)	X15	3	(-0.092463)
X2	1	**7.395169**	X16	2	**1.408602**
X3	0	(-1.823561)	X17	3	**3.115465**
X4	2	**7.8632**	X18	4	(-2.718257)
X5	4	(2.859459)	X19	3	**5.96644**
X6	4	(4.208744)	X20	4	**5.063698**
X7	3	(0.047367)	X21	3	(2.783828)
X8	0	(0.300431)	X22	5	(2.351871)
X10	3	(-4.780066)	X23	4	(-3.395908)
Xll	0	(0.601248)	X24	4	(1.281687)
X12	4	(1.38586)	X25	3	(1.322331)
			Y	1	**0.647092**

It can be seen that the coefficient of resolution increases to 0.89, demonstrating a great improvement in model interpretability. At the same time, the DW statistics increases to 1.99 and is closer to the critical value of 2 without autocorrelation. Consequently, this model appears to be a better linear model.

To sum up, this work choose 24 independent variables consists of the lag orders of both keywords searching volume and the explained variable, and gets the autoregressive model equations

$$\begin{aligned} Y = & -2727.1920 + 0.0505X_{1,t-2} + 7.3952X_{2,t-1} - 1.8236X_{3,t} \\ & + 7.8632X_{4,t-2} + 2.8595X_{5,t-4} + 4.2087X_{6,t-4} + 0.0474X_{7,t-3} + 0.3004X_{8,t} \\ & - 4.7801X_{10,t-3} + 0.6012X_{11,t} + 1.3859X_{12,t-4} - 8.3516X_{13,t-7} - 0.0925X_{15,t-3} \\ & + 1.4086X_{16,t-2} + 3.1155X_{17,t-3} - 2.7183X_{18,t-4} + 5.9664X_{19,t-3} + 5.0637X_{20,t-4} \\ & + 2.7838X_{21,t-3} + 2.3519X_{22,t-5} - 3.3959X_{23,t-4} + 1.2817X_{24,t-4} + 1.3223X_{25,t-3} + 0.6471Y_{t-1} \end{aligned}$$

The Cointegration Test results show that P value is 0.00000 and is markedly less than the 0.05 significance level. Meanwhile, the t value is -12.49419 and is also less than the critical value given by MacKinnon table. Thus, this model satisfies with the cointegration conditions and there exists a long-term equilibrium relationship between variables.

After establishing the linear model, test on the model setting error is needed. This work uses Ramsey RESET test with Eviews and get the result shown in Table 5.

Table 5. Ramsey RESET

F-Statistics	Critical Value (1%)
0.0005	0.01

The test result rejects the original hypothesis, so the model setting error indeed significantly exists. According to the principle of Ramsey test, the residual regression results show the lack of the nonlinear part in the model. In addition, the comparison of prediction results and the actual value also shows the deficiency of the linear model (See in Figure 4).

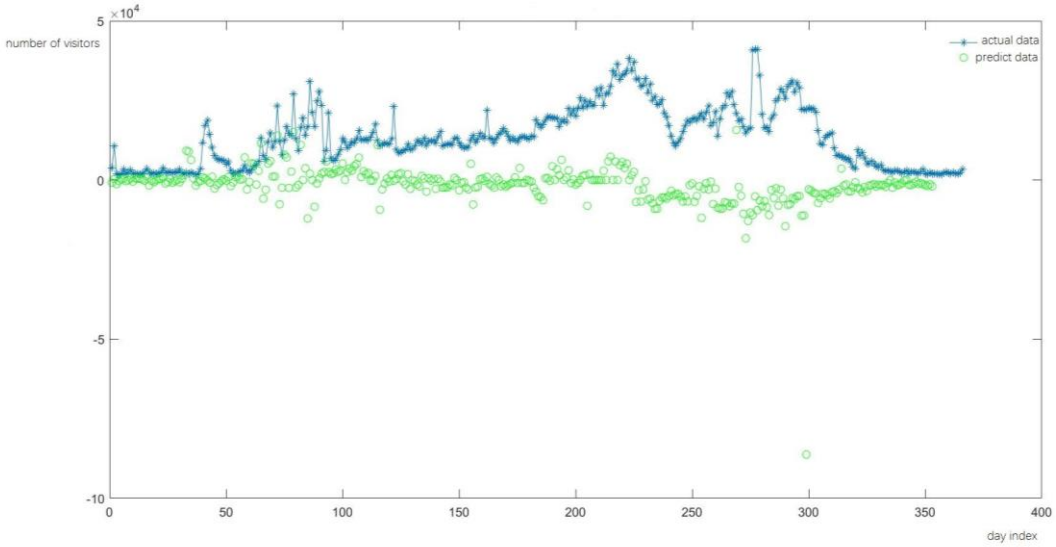


Figure 4. Linear Model Result

The value of Mean Squared Error (MSE) of this linear model is $2.3127 \acute{} 10^{10}$.

4.3 Results of model modified by BP neural network

This article plots both BP neural network predict errors and actual errors in the same coordinate system (see in Figure 5). Obviously, the BP neural network fits the error of peak season accurately.

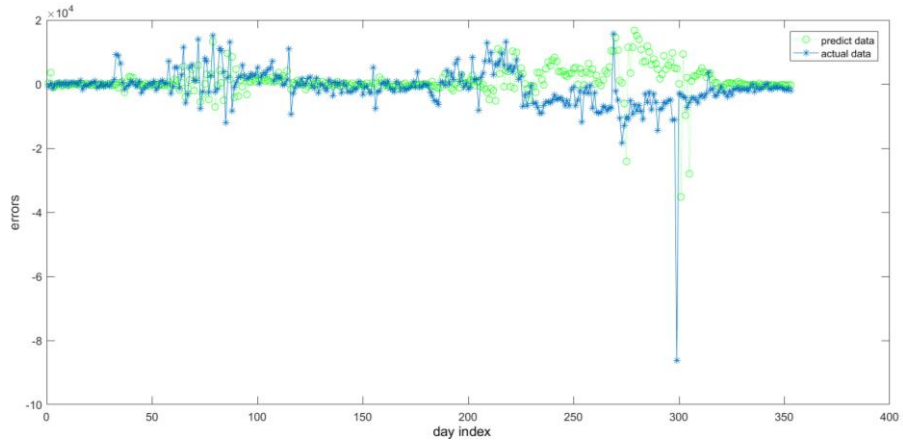


Figure 5. Fitting Error

4.4 Results of Integrated Model

Now we use BP neural network to modify errors in previous linear model.

This work combines both linear and nonlinear parts, and plot both our total predict result and actual data in the same coordinate system. Compare to figure 6, we can easily find that after the modification of BP neural network errors, the prediction results get closer to the actual data in the peak season.

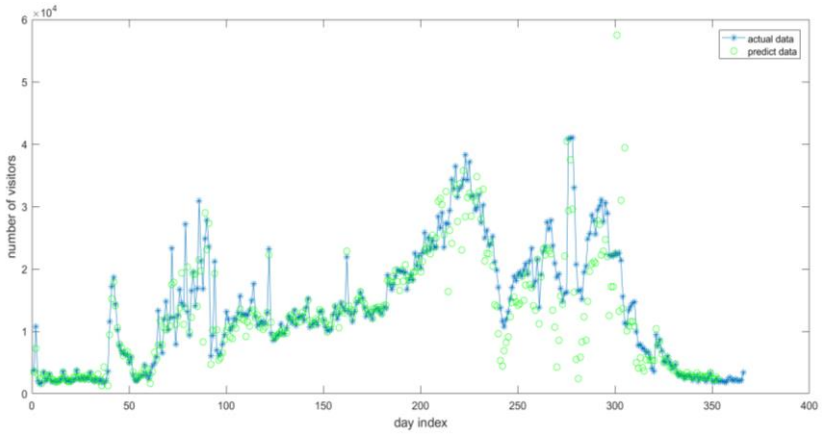


Figure 6 Model Results

The value of Mean Squared Error (MSE) of this model modified by BP neural network is 1.8238×10^{10}.

The MSE value of model modified by BP neural network is 1.8238×10^{10}, which is markedly smaller than that of the initial linear model (2.3127×10^{10}) and testifies higher model predicting accuracy and validity. Comparing the results, the integrated linear and BP neural network estimator enhances the prediction precision, and can provide some certain evidence for travel management decision.

5 Conclusion

Collective search and its trends reveal powerful insights and provide the ability to predict affairs related (Leung et al. 2013). Diverse methods in the preview literatures focus either on the ordinary linear relationship between them or the pure non-linear machine learning. This work merges both of them into one model concerning various possible influence structure.

Given that the explanatory variables should be correlated with *Jiuzhai Valley*, the internet crawler with precise target is implemented, in which this article grasps 233 texts around it from six largest platforms. Among them, the text mining procession picks up 25 Chinese phrases which are significantly correlated with *Jiuzhai Valley* (In Chinese Character). Acquiring their daily Baidu Index between 2012 to 2016, the multiple linear regression is employed to testify each variable and establish a prediction model. While this elementary models can well explain the linear part, the residual therewith cannot be ignored. Thus, this paper takes Back-Propagation Neural Network to explain the non-linear part.

After verifying 23 of them with one-week lags in the long-run equilibrium, this study especially uses the MSE index to evaluate the accuracy and efficiency of the model performance. According to the comparison, we find the combination model achieves the best predictive performance.

This work makes material contributions in both theoretical and practical aspects.

First, the integration of Web Crawler, Text Mining Technology and Search Index are used to construct the scenic spot Search Keywords Profile.

Besides, previous works have focused either on machine learning or on statistical models, but their integration has been minimal. This research emphasis on the high efficiency from machine learning without compromising interpretability. This study innovatively combine the linear time series regression model and BP neural network algorithm to improve the accuracy of prediction validity in scenic spot especially in peak seasons.

Moreover, although this research only focus on predictive issues in Jiuzhai Valley, the above integrated model can be widely applied to other tourist attractions to give stimulating ideas to the tourist volume forecast in booming season and future development tactics through public opinion on the network.

6 Limitations and Prospects

There are still some limitations in this work due to time and device restrictions. We would like to revise them and enhance our work results.

The data grabbed may be affected by some commercial activities. Consequently, the data in Baidu may be virtually higher than reality. In the forum, the discussion of a scenic spot will change over time, and the time point of our choice may have an impact. Hence, the future work needs to take into account the different timeline to increase the robustness of the model.

The neural network has diverse types that aim at different type of data. It's inadequate that this article only does BP-NN testing on data sets, and there may be more advanced machine learning algorithms available. Therefore, future work may consider more algorithm merging or precise selection in models.

Besides, some other view point like Disney or Imperial Palace, which provides Flow recording equipment, might be involved in the future research.

References

Choi H, Varian H. *Predicting the Present with Google Trends*[J]. Economic Record, 2012, 88(Supplement):2–9.

Du J, Wang R, Tu X Y. The establishment of tourism holiday flow prediction model[C]// Intelligent Control and Automation, 2008. Wcica 2008. World Congress on. IEEE, 2008:1091-1095.

Gawlik E, Kabaria H, Kaur S. Predicting tourism trends with Google Insights[J]. 2011.

Gao Q, Wu P, Zhang J. Web page information extraction based on document object model and block distribution algorithm [J]. information theory and practice, 2016, 39 (4): 133-137.

Guo Y, Tang H, Song L, et al. ECON: An Approach to Extract Content from Web News Page[C]// Web Conference. IEEE, 2010:314-320.

Hassan S, Mihalcea R, Banea C. Random-Walk Term Weighting for Improved Text Classification[C]// International Conference on Semantic Computing. IEEE, 2012:242-249.

Huang X K, Zhang L F, Ding Y. Study on the predictive and relationship between tourist attractions and the Baidu Index: a case study of the Forbidden City[J]. Tourism Tribune, 2013, 28(11):93-100.

Liu K, Wang T, Yang Z, et al. Using Baidu Search Index to Predict Dengue Outbreak in China[J]. Scientific Reports, 2016, 6:38040.

Liu S Y, Wei L I. Review of the Researches on Tourism Economy Connection Based on the Perspective of Tourism Flow[J]. Journal of Chongqing Technology & Business University, 2014.

Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[J]. Unt Scholarly Works, 2004:404-411.

Pan B, Xiang Z, Tierney H, et al. Assessing the Dynamics of Search Results in Google[C]// Information and Communication Technologies in Tourism, Enter 2010, Proceedings of the International Conference in Lugano, Switzerland, February. DBLP, 2010:405-416.

Raghavan P, Tsaparas P. Mining Significant Associations in Large Scale Text Corpora[C]// IEEE International Conference on Data Mining, 2002. ICDM 2003. Proceedings. IEEE, 2002:402-409.

Rahman C M, Sohel F A, Naushad P, et al. Text Classification using the Concept of Association Rule of Data Mining[J]. 2010.

Tse R Y C. A simultaneous model of tourism flow, spending and receipts.[J]. Tourism Economics, 1999:251-260.

Xin Y, Pan B, Evans J A, et al. Forecasting Chinese tourist volume with search engine data[J]. Tourism Management, 2015, 46:386-397.

Yang X, Pan B, Evans J A, et al. Forecasting Chinese tourist volume with search engine data[J]. Tourism Management, 2015, 46:386-397.