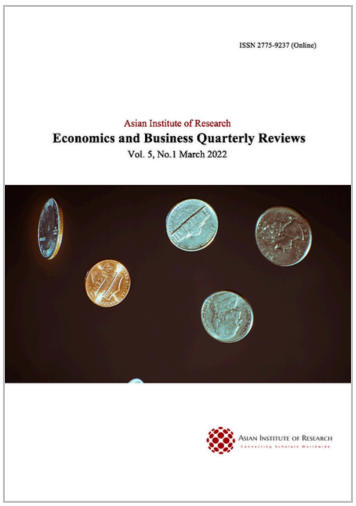

Economics and Business Quarterly Reviews

Jiang, J., & Deng, G. (2022). International Tourism Revenue Projections for Guilin in the Context of COVID-19. *Economics and Business Quarterly Reviews*, 5(1), 52-59.

ISSN 2775-9237

DOI: 10.31014/aior.1992.05.01.404

The online version of this article can be found at:
https://www.asianinstituteofresearch.org/

Published by:
The Asian Institute of Research

The Asian Institute of Research *Journal of Economics and Business* is a peer-reviewed International Journal. The journal covers scholarly articles in the fields of Economics and Business, which includes, but is not limited to, Business Economics (Micro and Macro), Finance, Management, Marketing, Business Law, Entrepreneurship, Behavioral and Health Economics, Government Taxation and Regulations, Financial Markets, International Economics, Investment, and Economic Development. As the journal is Open Access, it ensures high visibility and the increase of citations for all research articles published. The *Journal of Economics and Business* aims to facilitate scholarly work on recent theoretical and practical aspects of Economics and Business.

ASIAN INSTITUTE OF RESEARCH
Connecting Scholars Worldwide

ASIAN INSTITUTE OF RESEARCH
Connecting Scholars Worldwide

The Asian Institute of Research
Economics and Business Quarterly Reviews
Vol.5, No.1, 2022: 52-59
ISSN 2775-9237

DOI: 10.31014/aior.1992.05.01.404

International Tourism Revenue Projections for Guilin in the Context of COVID-19

Jingming Jiang[1], Guangming Deng[2]

[1] College of Science, Guilin University of Technology, Guilin, China. Email: 824512842@qq.com
[2] Institute of Applied Statistics, Guilin University of Technology, Guilin, China. Email: dgm@glut.edu.cn

Correspondence: Guangming Deng, College of Science, Guilin University of Technology, Guilin, Guangxi, 541006, China. Tel: 18076761871. E-mail: dgm@glut.edu.cn

Abstract
This paper collects relevant international tourism revenue data for Guilin from 2004 to 2020 for analysis and modelling using three algorithms, namely multiple linear regression model, support vector machine and random forest, to explore the variables affecting international tourism revenue in Guilin and to make model predictions for international tourism revenue from 2019 to 2020. The empirical evidence shows that the multiple linear regression model predicts the best results, especially the accurate prediction of the sharp decline in international tourism revenue when the new crown pneumonia epidemic spreads in 2020 , which can provide some scientific basis tourism development planning of Guilin city in the future.

Keywords: International Tourism Revenue, Spearman Correlation Analysis, Multiple Regression Model, Support Vector Machine, Random Forest, COVID-19

1. Introduction

Guilin has been known for centuries as "the best place in the world for Guilin landscapes," and many international tourists come here every year. The tourism industry has also contributed to the development of Guilin's economy. However, in the past two years, The global outbreak of COVID-19 has caused a downturn in the national and international tourism industry, resulting in huge losses(Xia & Feng, 2020) to tourism enterprises, associated industries and related workers. Feng Zhenglong(2021) proposed a strategy for the revitalisation of Guilin's tourism economy based on the challenges encountered by the Guilin tourism economy under the normal prevention and control of the epidemic, Zhou Jiuhe(2019) et al. constructed a hierarchical structure model and judgment matrix of the factors influencing tourism revenue in Guilin based on hierarchical analysis to determine the trends of the main influencing factors of tourism revenue in Guilin over the past 15 years, and Li Hui(2013) et al. constructed a prediction model of tourism demand in Guilin based on grey system theory, which can be applied to the forecast of tourism development in Guilin. Therefore, in order to better cope with this sudden epidemic, the relevant tourism departments must reasonably compare the tourism revenue forecast towards with the actual revenue from 2019-2020 and select the optimal forecast model to support the tourism international revenue forecast in the coming years. In this paper, the international tourism revenue and its various influencing factors for the 2004-2020 years

are selected, and the three algorithms of multiple linear regression, support vector machine and random forest are used for regression modelling respectively, and the advantages and disadvantages of the three models are compared according to the three model evaluation indicators of mean square error (MSE), mean absolute error (MAE) and goodness of fit (R^2), and then the international tourism revenue for 2019-2020 is predicted, compared with the actual international tourism revenue for 2019-2020. Finally select the optimal prediction model--the multiple linear regression model, which can be applied to the tourism development forecasting of Guilin City.

2. Theoretical foundations

2.1 Model principles

2.1.1 Principle of multiple linear regression models

In real economic problems, a variable is often influenced by more than one variable, when it is necessary to use two or more influences as independent variables to explain the changes in the dependent variable, this is multiple regression also known as multiple regression, When there is a linear relationship between multiple independent variables and the dependent variable, the regression analysis performed is a multiple linear regression(Xu, 2021; Foreman, Hesse & Lundstrom, 2021). Let Y be the dependent variable and $x_1,\ldots,x_k$ the independent variable, the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \mu \qquad (1)$$

where k is the number of independent variables, β_0 is the regression constant and $\beta_1,\ldots,\beta_k$ is the regression coefficient.

2.1.2 Support vector machines

SVM(support vector machine) is simply a classifier that solves two types of problems: classification and regression(Zhang, 2000; Pontil & Verri, 1998). Regression and classification are essentially the same things in a sense.SVM Classification is about finding a plane in which the support vectors or all data (LSSVM) of two sets of classifiers are furthest from the classification plane; SVR regression is about finding a regression plane in which all data of a set are closest to that plane. The optimal hyperplane is not the "widest" as in SVM, but the one that minimizes the total deviation of all the sample points from the hyperplane. Regression is like finding the relationships inherent in a pile of data. Regardless of how many categories the pile of data consists of, a formula is obtained that fits this data and is able to find a new value when given a new coordinate value. So for SVR, it is finding a surface or a function that can be fitted to all the data (that is, all the data points, regardless of the category they belong to, that have the closest distance to that surface or function).

2.1.3 Random Forest

Random Forest (RF) is a new and emerging machine learning algorithm. It has a wide range of applications, from marketing to health care insurance, and can be used to model marketing simulations, count customer origins, retention and churn, and predict the risk of disease and patient susceptibility(Kapsiani & Howlin, 2021). Random Forest is an algorithm that integrates multiple trees through the idea of integrated learning, the basic unit of which is the decision tree, which is essentially a branch of machine learning - the Ensemble Learning approach. There are many ways to deal with over-fitting of decision trees, such as pruning techniques, but integration techniques can also deal with this problem. We use integration techniques to generate a variety of different decision trees and combine their predictions to reduce the probability of over-fitting. In order to generate a variety of decision tree models, we sample the data and use the bootstrap sampling technique. Suppose there is a n samples available for training, we by the bootstrap technique training set is sampled n times, each time to select a sample from a training set to record index and back into training set, so that we end up with single decision tree model, the training also has n samples, thus ensure the training every decision tree in data set is different, Thus, different forms of decision trees are generated.

2. 2 Introduction to model evaluation indicators

In order to facilitate comparison of the advantages and disadvantages of the three regression models constructed in this paper, three evaluation indicators commonly used in regression models were selected: mean squared error, mean absolute error and goodness of fit.

2.2.1 Mean Square Error:

Mean Squared Error (MSE), This statistical parameter is the mean value of the sum of the squares of the errors at the corresponding points of the predicted and original data and is given by the formula.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \overline{y_i})^2 \qquad (2)$$

Where y_i is the predicted value and $\overline{y_i}$ is the true value. the closer the MSE value is 0, the better the model selection and fit, and the more successful the data prediction.

2.2.2 Mean absolute error

Mean Absolute Error (MAE) is the average of the absolute values of the deviations of all individual observations from the arithmetic mean, with the formula.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \overline{y}_i\right| \qquad (3)$$

where y_i is the predicted value and $\overline{y_i}$ is the true value. The mean absolute error (MAE) is a better reflection of the actual predicted value error as the deviations are absolute and do not cancel out positive or negative. the closer the MAE value is 0, the better the model selection and fit.

2.2.3 Goodness of fit

R^2 is a measure of the amount of information not captured by the model as a proportion of the amount of information carried in the real label, and is given by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (4)$$

where $SST = SSR + SSE$, SST (total sum of squares) is the total sum of squares, SSR(regression sum of squares) is the regression sum of squares and SSE(error sum of squares) is the sum of squared residuals. The closer R^2 is to 1, the better, indicating that the introduced factors are sufficient to explain the variation of this data set.

3. Example analysis

3.1 Data sources

This paper collected the data of international tourism revenue of Guilin from 2004 to 2020 from the statistical Yearbook of Guangxi Zhuang Autonomous Region: international tourism revenue y , international tourist number x_1, gross regional product per capita x_2, The proportion of secondary industry in GDP x_3, The proportion of tertiary industry in GDP x_4,The park area x_5, Green coverage area x_6, Operation line network length x x_7, Taxis operate vehicles x_8, Annual mean temperature x_9, maximum temperature x_{10},Number of travel agencies x_{11}, Number of star hotels x_{12}, consumer price index x_{13}, number of libraries x_{14}, The amount of water supply x_{15}, (food) per capita possession x_{16}. The data were first standardized. The data were then analyzed by Spearman

correlation and autocorrelation to eliminate redundant variables and identify the variables with the most significant impact. Finally, the data were modelled and predicted using each of the three models, the evaluation indicators of the models were compared and analyzed to draw final conclusions.

3.2 Data processing

3.2.1 Spearman correlation analysis

The Spearman's correlation coefficient, also known as the Spearman's rank correlation(Samuel & Lyerly, 1952) coefficient is a measure of the correlation between two variables x and y has a value between -1 and 1. It is generally used to analyse the relationship between two continuous variables and is given by the formula

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (5)$$

where n is the number of samples and d represents the equivalence between the data x and y, ρ indicates the presence of different degrees of linear correlation between the two variables:

① 0.8<| ρ |<1, indicating that the two variables are extremely strongly correlated.

②0.6<| ρ |<0.8, indicating that the two variables are strongly correlated.

③0.4<| ρ |<0.6, indicating that the two variables are moderately correlated.

④ 0.2<| ρ |<0.4, indicating that the two variables are weakly correlated.

⑤ 0<| ρ |<0.2 means that the two variables are extremely weakly correlated or uncorrelated.

Spearman correlation analysis was performed on each independent variable and the dependent variable separately using R statistical software to obtain 16 Spearman correlation coefficient, and all independent variables with an absolute value of Spearman correlation coefficient greater 0.5 than were screened out, as shown in the figure1.

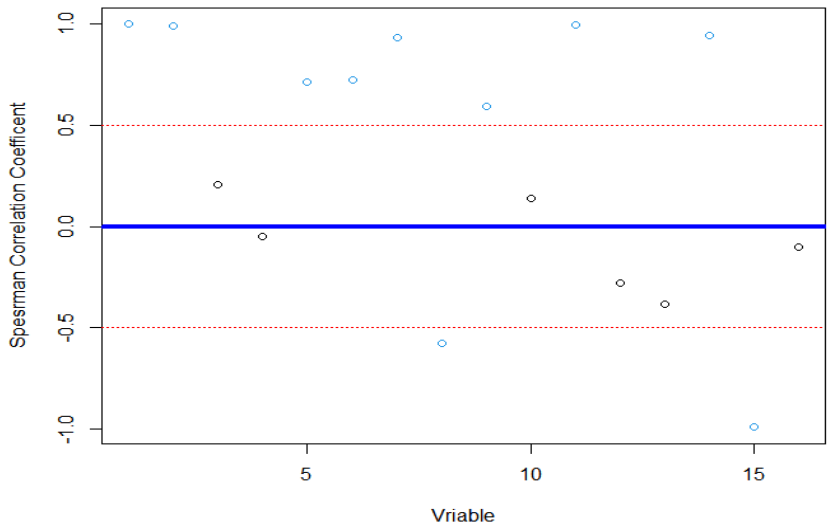

Figure 1: Spearman's correlation coefficient values for each Independent variable and the dependent variable

As can be seen from the graph1, Spearman correlation coefficients of six independent variables are concentrated between -0.5< ρ < 0.5,These independent variables are either weakly, very weakly or uncorrelated with international tourism income. The Spearman's correlation coefficients distributed between 0.5<| ρ |<1 specifically contain10 independent variable. This indicates10 that this independent variable is moderately to highly correlated with international tourism receipts. Therefore, the middle 6 variable with $x_3, x_4, x_{10}, x_{12}, x_{13}, x_{16}$ a low Spearman correlation coefficient is removed.

3.2.2 Autocorrelation analysis

This is because there is also a correlation between the10 remaining independent variables. If the absolute value of the correlation coefficient between the independent variables is greater 0.95 than,The two independent variables are highly correlated. Highly correlated independent variables will bring unnecessary calculation to the later data prediction, and may even lead to over-fitting.Therefore, it is necessary to analyze the autocorrelation of these 10 independent variables again. The R statistical software was used for pairwise correlation analysis of these 10 independent variables, and a correlation coefficient plot was drawn based on the correlation coefficient matrix, as shown in the figure 2. From the size and colour of the circles in the correlation coefficient plot, it can be seen that there is a high degree of correlation between the 10 individual independent variables and therefore it is necessary to propose redundant independent variables.

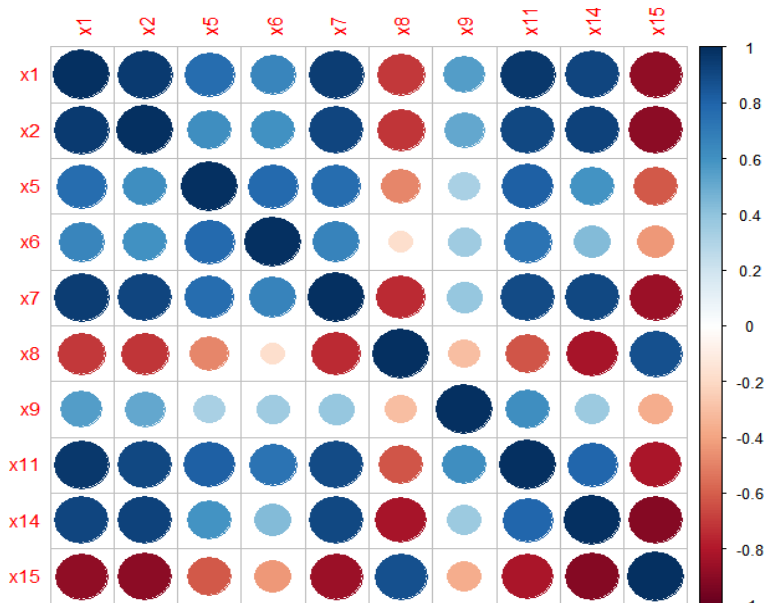

Figure 2: Correlation coefficient graph

As colour of the circles in Figure 2, $x_{11}, x_{15}$the size and colour of the circles are darker, indicating a strong correlation with the variables, so the independent x_{11}, x_{15}variables are removed.

The redundant independent variables were removed through Spearman correlation and autocorrelation coefficient analysis, and the final independent variables obtained were $x_1, x_2, x_5, x_6, x_7, x_8, x_9, x_{14}$. Next we build a model of Guilin's international tourism revenue with these eight independent variables for comparative analysis.

3.3 Selection of predictive models, building and analysis of results

When performing regression prediction analysis, it is often necessary to divide a known data set into a training set and a test set. The training set is used for the training of the model as well as its generation, while the test set is used to test the real prediction of the trained model to derive the accuracy of its prediction. In this paper, the international tourism revenue from 2004-2018 is used as the training set and the international tourism revenue from 2019-2020 is used as the test set.

We used three algorithms, namely multiple linear regression, support vector machine and random forest to build models for predicting international tourism income based on these eight independent variables.By analyzing the predictive effect of each model's MSE, MAE, R^2indicators, we selected the model with the best predictive effect.

3.3.1 Multiple linear regression

In multiple regression analysis, if the relationship between the dependent variable and multiple independent variables is linear, it is a multiple linear regression. Multiple linear regression is an extension of one-dimensional linear regression. Its basic principles and methods are similar to those of one-dimensional linear regression analysis, and the final regression equation obtained is:

$$y = -0.01547 + 0.70093\,x_1 + 0.25866\,x_2 + 0.17327\,x_5 - 0.10211\,x_6 + 0.18313\,x_7 + 0.06816\,x_8 + 0.06632\,x_9 - 0.24319\,x_{14} \quad (6)$$

Using the regression equation (6), the test set was tested and the results are shown in the figure3. As can be seen from the test plots, the multiple linear regression has very little prediction error for the sample points, with small gaps between many points and high overall accuracy of the model. the predicted data for 2019-2020 largely overlap with the actual data.

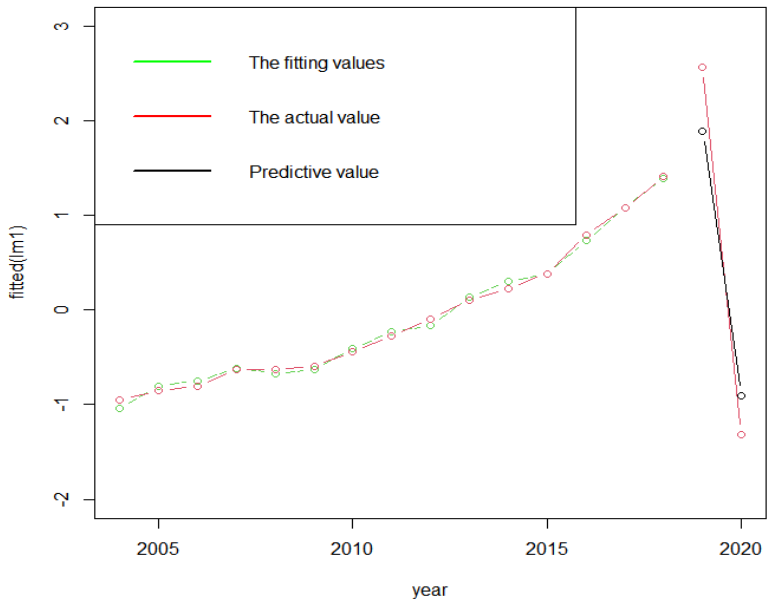

Figure 3: Multiple linear regression prediction graph

3.3.2 Support vector machines

The traditional machine learning approach of SVM was tested against making predictions of international tourism revenue, and the results are4 shown in the graph. As can be seen from the prediction graph, it illustrates the large prediction error of the support vector machine for the sample points, with large gaps between many points, resulting in a low overall accuracy of the model, especially for the predicted data for 2019-2020, which has a large gap with the actual data.

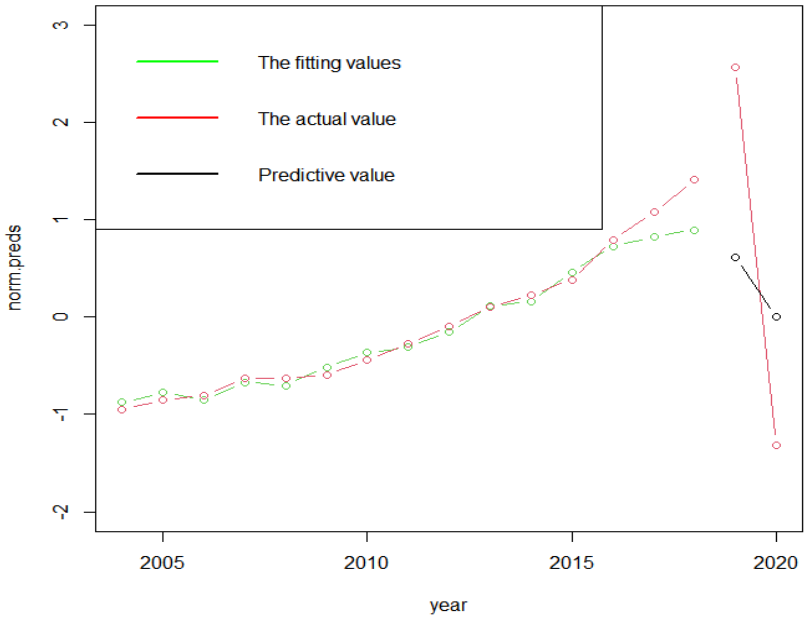

Figure 4: Support vector machine prediction graph

3.3.3 Random forests

Forecasts were made using the random forest model and the test results are shown in the figure5: as can be seen from the forecast graph, this illustrates the large forecast error of the support vector machine for the sample points, with large gaps between many points, resulting in a low overall accuracy of the model, especially for the forecast data for 2019-2020, with large errors, similar to the trend of the support vector machine forecast graph.

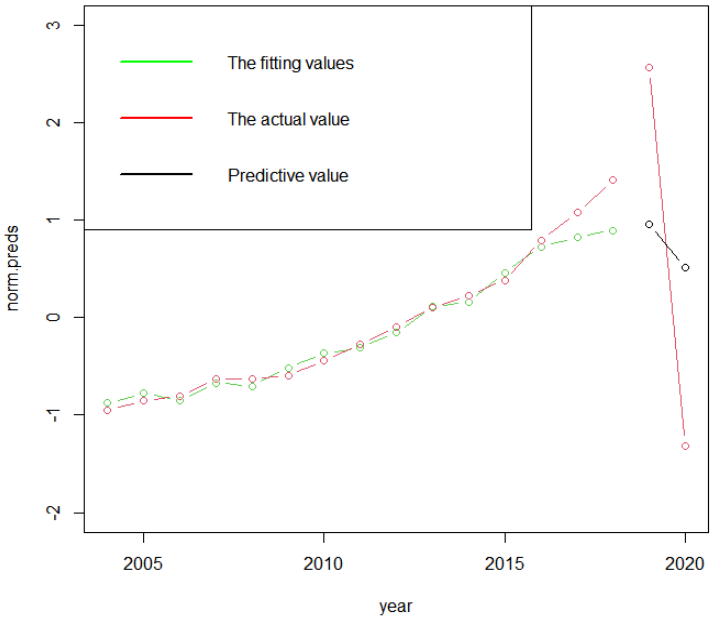

Figure 5: Random forest prediction graph

Although the graphical method is intuitive, it does not allow for precise conclusions to be drawn. Therefore, the evaluation metrics of the three models are combined in this paper. Table 1 shows the details of the error analysis for the model evaluation, using the three evaluation criteria of MSE, MAE and R^2. The training effects of the models are evaluated separately.

Table 1: Model evaluation error analysis table

Algorithms	MSE	MAE	R^2
Multiple linear regression	0.0023	0.0413	0.9896
Support vector regression	0.0253	0.1018	0.9499
Random forest regression	0.0971	0.0971	0.8315

It can be seen from table 1 that the multiple linear regression model has the smallest MSE, MAE and the largest R^2, which belongs to the model with the best effect. MSE and MAE of support vector regression and random forest regression are basically equal, but the support vector regression of R^2 greater than random forest regression, which indicates that the error of these two models is about the same, but the model fitting effect of support vector machine regression is slightly better than that of random forest regression model.

As easily obtained from the graphs and tables above, firstly by building and forecasting the model, we chose the multiple linear regression model with the best results. As can be seen from the prediction graph Figure 3, the multiple linear regression model predicts an excellent match for the international tourism data from 2019-2020 and it can be seen that the predicted values are also on a sharp downward trend due to the outbreak of the COVID-19 in 2020, which led to a sharp decline in international tourism revenue in Guilin. Secondly the comparison of the methods of support vector regression and random forest regression reveals that traditional multiple linear regression is applicable to linear regression problems, while machine learning performs poorly and is applicable to non-linear regression problems with fewer samples in this study.

4. Conclusion

Through the comparison tests of the three models and the analysis of the model indicators, we found that the multiple linear regression model has a better prediction effect for the international tourism revenue data of Guilin, and its prediction effect reaches our ideal state, especially for the special year of 2020, which can provide some reference value for the prediction of international revenue of Guilin in the following years. The model is of some significance to both the Guilin Tourism Bureau and the relevant national authorities.

References

Xia,. J. C., & Feng, X. X.(2020). Impact of novel coronavirus pneumonia on tourism industry and Countermeasures. China Business and Market(03),3-10. doi:10.14089/j.cnki.cn11-3664/f.2020.03.001.

Feng, Z. L.(2021). Strategies for revitalizing Guilin's tourism economy. Cooperative economy and Technology, (23):18-20.doi:10.13665/j.cnki.hzjjykj.2021.23.006.

Zhou, J. H., & Lu, P. (2019). empirical study on the factors influencing tourism income in Guilin. Journal of Guangxi Normal College (Philosophy and Social Science Edition), 40(02):84-92. doi:10.16601/j.cnki.issn1002-5227.2019.02.015.

Li, H., Wang, Y., & Yin, H.(2013). Forecasting and analysis of tourism demand in Guilin based on grey system theory. Henan Science, 31(05):679-682. doi:10.13537/j.issn.1004-3918.2013.05.020.

Xu, Z. Shu. (2021). Predicting the impact of taxation on tax revenue in the "floor stall economy" - based on a multiple linear regression model. China Circulation Economy(09),165-168. doi:10.16834/j.cnki.issn1009-5292.2021.09.052.

Foreman, N., Hesse, A., & Lundstrom, C.(2021). Machine Learning Fails To Improve Marathon Time Prediction Compared To Multiple Linear Regression: 161. Medicine & Science in Sports & Exercise 53: doi:10.1249/01.MSS.0000759656.97049.16.

Zhang, X. G.(2000). On statistical learning theory and support vector machines. Acta Automatica Sinica, (01):36-46. doi:10.16383/j.aas.2000.01.005.

Pontil, M. & Verri, A.(1998). Properties of support vector machines. Neural computation, 10(4): doi: 10.1162/089976698300017575

Kapsiani, S. & Howlin, B. J.(2021).Random forest classification for predicting lifespan-extending chemical compounds.. Scientific reports(1), doi:10.1038/S41598-021-93070-6.

Samuel, B. & Lyerly.(1952).The average spearman rank correlation coefficient. Psychometrika(4), doi:10.1007/BF02288917.