Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **2**, May **2024**

Seasonal Tourism Demand Forecasting Based on Machine Learning in Big Data Environment

Jing Li*

Department of Tourism Management
Jinzhong University, Jinzhong 030619, P. R. China
fyt20170601@163.com

Bin Cao

Institute of Architecture
Universiti Sains Malaysia, 11800 Gelugor, Penang 999004, Malaysia
snjj521@126.com

*Corresponding author: Jing Li

Received November 24, 2023, revised January 26, 2024, accepted March 4, 2024.

Abstract. *In view of the serious imbalance between the actual amount of tourists and tourism resources, this article suggests a seasonal tourism demand forecast on the ground of machine learning in a big data environment. Firstly, the factors affecting the seasonality of tourism are analyzed, and three factors, tourism subject, tourism resource and tourism object, are obtained. The correlation degree of these three factors is analyzed by using grey series. Then, the dimensionality of the influencing factor data is decreased to a lower dimension through data dimensionality reduction technology to keep off overlapping problems. Secondly, genetic algorithm is adopted to make optimal the heavy and bias of BP neural network, taking dimensionality reduction factors as input variables, selecting and eliminating individuals according to fitness function, and finally finding the optimal solution to obtain the goal of accurate prediction. The simulation outcome indicates that the MAE, MAPE, RMSE and RMSPE of the seasonal tourism forecasting model designed on the ground of machine learning in this paper are 13.73, 2.49%, 18.82 and 3.68%, respectively, which are lower than the comparison model and greatly enhance the forecasting performance.*

Keywords: Machine learning, Seasonal tourism, Demand forecast, Neural network, Genetic algorithm

1. **Introduction.** As information technology developping, large-scale structured and unstructured data are constantly generated, forming big data and opening a new era [1]. Big data has the feature of large scale, many types, fast production speed and high value. The concentration of big data analysis and research in various area is how to tap out the potential value in big data. In recent years, tourism demand has been growing, and tourists' travel has generated a large amount of tourism data, and these multi-source heterogeneous data have formed a massive tourism big data. Tourism big data includes content actively uploaded and shared by tourists, as well as hidden data obtained by tourism management departments [2,3,4]. At present, the brisk growth of machine learning has gradually begun to analyze and apply complex data in various area, and great outcome has been obtained in application process. Applying machine learning to tourism big data analysis can greatly enhance the management level and economic benefits of the tourism industry. At present, user-generated content data is the main data type of tourism research, which

plays a significant role in tourist emotion analyse, tourism behavior analysis, tourism marketing and tourism recommendation. By analyzing temporal and spatial characteristics of tourist behavior, the task of tourism prediction can be realized. The data related to tourism activities can mainly be used as the forecast of tourism demand, and the future tourism demand can be forcasted by historical data.

1.1. **Related Work.** Nowadays, machine learning has been extensively adopted in the area of tourism demand forecasting with its strong forecasting ability (especially for nonlinear forecasting) [5]. For tourism demand forecasting, the overall performance of machine learning has been shown to outperform traditional time series models and econometric models. Machine learning used to predict the arrival of Japanese tourists in Hong Kong outperforms methods such as multiple regression, exponential smoothing and moving average [6]. Li et al. [7] used the improved BPNN model to forecast and analyze the amount of tourists in a certain place in China. Fritz et al. [8] established a time series model when predicting the number of air tourism in Florida, a local American state. Kaynak and Macaulay [9] used Delphi technique method and regional database to predict local tourism development. In terms of the study of Song and Li [10], time series analysis, econometric model and artificial intelligence is the most popular categories of forecasting methods [11]. Differential autoregressive mobility model (ARIMA) has been widely used in tourism demand forecasting with great performance [12, 13,14, 15]. In addition, Chaitip explores two different statistical models in the application of travel demand forecasting, where the ARIMA model does not always outperform the others, it performs well in traditional econometric models, but sometimes not as well as machine learning methods. For example, using Hong Kong's travel demand to be an instance, Cho [16] contrasted some normal visitor forecasting technology: exponential flating, univariate ARIMA, and he judged that the performance of artificial neural networks is superior to the another forecasting models.

Exponential smoothing (ES) and generalized autoregressive conditional heteroscedasticity (GARCH) are also commonly used as benchmark models for tourism forecasting [17, 18], and other econometric models have also been used in tourism demand forecasting models. For example, the Error correction model (ECM) [19]. Support vector machine (SVM) was first applied to tourism demand forecasting in the late 1990s. Empirical results show that compared with time series model and multiple regression model, SVM can obtain better prediction accuracy without special qualification, and is generally superior to time series model and multiple regression model [20, 21, 22]. Durbarry and Sinclair [23], Li et al. [24] used the Generalized dynamic factor model (GDFM) to predict the number of tourists in Beijing. Durbarry and Sinclair et al. adopted the market share analysis method to make an empirical analysis of tourism demand in France. Alegre and Pou [25] calculated the duration of tourists' stay in the Mediterranean Sea when modeling tourism demand and added it into the model as a supplementary explanatory variable. When predicting the amount of inbound tourists, Koc and Altinay [26] also divided the tourism market horizontally and analyzed the temporal and spatial changes of inbound tourists' consumption. Yu and Wang [27] used MATLAB tools to establish a BP neural network model to forecast a tourist volume in Beijing. Claveria and Torra [28] used three various models to forecast tourism demand, and found that ARMA model had the best fitting effect and the smallest model error. Hassani et al. [29] used the singular spectrum analysis model to forecast the tourists' amount in the United States, finally got the outcome that the prediction error of the model was small. Kim and Lee [30] took exchange rate and other tourism price characteristics as input variables when predicting inbound tourism demand in South Korea, and lastly found that exchange rate and relative price contributed the most to tourism demand, while the two proxy variables of transportation cost had no influence on tourism demand.

1.2. **Motivation and contribution.** From the above discussion, it can be seen that there is an imbalance between the dispersion of tourism imagination and the actual number of tourists, bringing about the waste of a wide amount of tourism resources. Therefore, accurate tourism demand forecasting is essential. In the current big data environment, this article suggests a seasonal tourism demand forecasting method on the ground of machine learning. The model first analyzes the incentives of seasonal tourism, and adopts the grey correlation analyse algorithm to solve the three factors affecting seasonal tourism. Then, the main factors of the seasonal tourism demand impact index are extracted through the main analytic hierarchy process. At last, genetic manipulation is used to generate a new BP neural network, select and eliminate individuals according to the fitness function, and finally find the optimal solution, so as to improve the prediction operation of the algorithm.

2. **Related theoretical analysis.**

2.1. **BP neural network.** It is also known as an error backward propagation neural network. From its structural characteristics, it is a feedforward structure composed of multiple nonlinear elements [31]. Its basic working principle is to reduce the error between the network neurons by comparing the output data with the actual data, and then adjusting the beginning weights and initial thresholds between the network neurons in terms of the comparison results. Neuron is the most basic building block, and the relationship between its input and output is as follows:

$$F = f\left(\sum_{j=1}^{S} Y_j H_j + a\right) \tag{1}$$

where F is neuron output, f is the transfer function, Y_j is neuron input, H_j is the system number before other neurons connected to this neuron; $a = H_0$ is the initial qualification value. According to the above settings $\overline{Y} = (y_1, y_2, ..., y_s)$, $\overline{H} = (h_1, h_2, \ldots, h_s)^T$, and $\overline{Y}.\overline{H} + a = m$, then there is $x = f(m)$.

In general, the model can be separated into three layers: input layer, hidden layer, and output layer. Contact with cross-flow between each other is realized through the initial values of neurons, and there is no communication between neurons in the layer. The network input is y_j, the output is x_s, and the input layer and the hidden layer are connected by weight h_{ji}. At the same time, the obscured layer is t_j, the threshold of the obscured layer is ϕ_i, and the threshold of the output layer is s_r. The hidden layer and the input layer are connected by weight g_{is}. Among them, the obscured layer t_j is obtained by summing the enter layer y_j and its corresponding weight h_{ji}, subtracting the obscured layer threshold ϕ, and then passing the function f. Then the hidden layer and the network input are expressed for Equation (2) and Equation (3) respectively.

$$t_i = f\left(\sum_{j=1}^{m} y_j h_{ji} - \phi_i\right) \tag{2}$$

$$x_s = v\left(\sum_{i=1}^{q} t_j g_{is} - s_r\right) \tag{3}$$

2.2. **Genetic algorithm.** It is a heuristic comb method on the ground of the process of biological evolution. Due to genetic algorithm is a global comb algorithm [32], it can effectively avoid local optimal challenge of the iterative process. Moreover, genetic algorithm takes the fitness function as the basis for judging whether the method found in the iterative procedure is optimal solution, without relying on other conditions, and has the characteristics of easy operation and parallel optimization. The most significant thing is that the current research and use of genetic algorithm has become mature, and then solve many problems in the procedure of using the algorithm, and offer greater convenience for users.

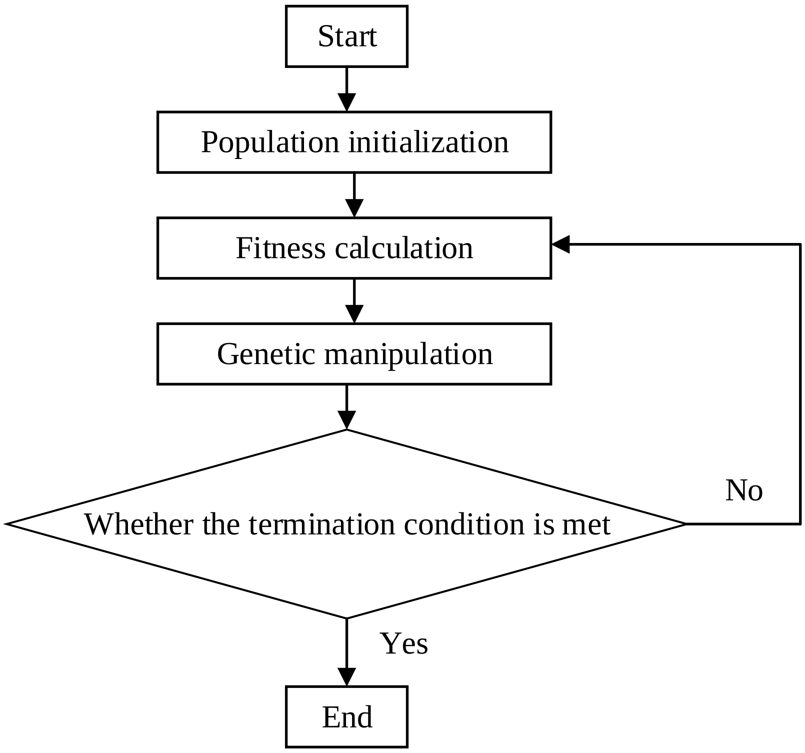
Start
Population initialization
Fitness calculation
Genetic manipulation
Whether the termination condition is met
No
Yes
End

Figure 1. Genetic algorithm flow chart

Figure 1 displays the basic flow chart of genetic algorithm, which is able to be concluded as follows:

(1) Initialization. Using Equation (4) to code the solving issue, the goal is to convert the problem to be solved into a language that the computer can recognize.

$$f_n(s) = -\frac{dR_n(s)}{ds} = -\frac{1}{R(n)}\frac{dR(n+1)}{ds} \tag{4}$$

where f represents the encoding function, s and n represent the sequence number and number of the population respectively, and R represents the real digitalization function.

(2) Individual evaluation. The cost of suitable routine is the basis to judge the merits of individuals in the population, and then provides a standard for a series of operations such as selection and crossover.

(3) Genetic operator. Mainly includes selection, crossover, mutation and other operations, the basic principle is to optimize the individual gradually through the above series of operations. The cross operation and mutation operation of chromosome c_l and chromosome d_k at position i are shown in Equation (5) and Equation (6), respectively. Where

v and h are the current and maximum iterations respectively.

$$\begin{cases} c_{li} = c_{li}(1-d) + c_{ji}d \\ c_{ji} = c_{ji}(1-d) + c_{li}d \end{cases} \tag{5}$$

$$c_x = \begin{cases} c_x + (c_x - c_{\max})f(v), h \geq 0.5 \\ c_x + (c_{\min} - c_x)f(v), h < 0.5 \end{cases} \tag{6}$$

(4) Stop strategy. Under ideal conditions, through continuous iteration, the optimal solution will eventually be found, but due to limitations in various aspects, such as time and resources, it can be stopped when the result has reached the expected value or the iteration times have reached the pre-set value.

3. Correlation analysis of seasonal tourism influencing factors.

3.1. Analysis of seasonal tourism incentives. The reason of tourism seasonality relies on the tourism subject, the condition of tourism realization and the tourism object. As shown in Figure 1, tourism is a human behavior, the most concerned is as the main body of tourism activities: tourists. In terms of the motivation and main goal of tourism, tourists can be divided into: recreational tourists, cultural tourists, official tourists, family and personal affairs tourists, fitness tourists. However, there are obstacles between tourism need and tourism realization. Only by overcoming these obstacles, can the latent tourism need be transformed into the tourism demand that has practical significance for tourists and tourism commodity operators. From the perspective of external conditions, whether tourists travel depends on discretionary income, discretionary time, etc. From the perspective of space conditions, people's travel demand, especially the middle and long distance travel demand, is generally difficult to meet in peacetime, and only during a long holiday can travel, so the peak passenger flow is often relatively concentrated in time. On the ground of meeting the tourism conditions of tourists, then select the corresponding tourism objects, that is, tourism resources. The emergence of tourism seasonality is largely caused by the characteristics of tourism resources. Many sceneries display different beauty throughout the year; Some scenic spots have special seasonality, only in a certain season and time will present the best scenery, and finally appear the peak season and low season of tourism.

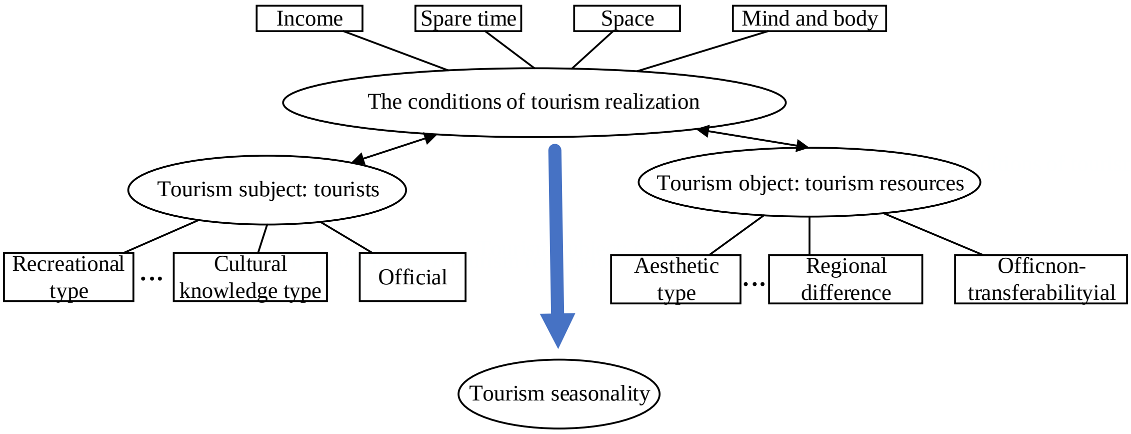

Figure 2. Seasonal factors affecting tourism

3.2. **Construction of factor correlation degree.** The basic idea of seasonal tourism correlation degree construction is to convert the data series among various factors into gray series respectively, and find out the most influential factors by comparing the correlation degree between various series as the input of machine learning. The computation process is as follows:

(1) Determine the analysis succession: determine the feature sequence (parent sequence) $\gamma(l)$ and the associated sequence (subsequence) $\chi_j(l)$.

(2) Dimensionless variable: this article uses the beginning cost of the series to carry out dimensionless processing, as shown in Equation (4).

$$x_j(l) = \frac{\chi_j(l)}{\chi_j(1)} \tag{7}$$

where $l = 1, 2, 3, \ldots, m$, $j = 1, 2, 3, \ldots, n$.

(3) Compute the correlation coefficient: the correlation coefficient of $\gamma(l)$ and $\chi_j(l)$ is as follows.

$$\phi_j(l) = \frac{\min_j \min_l |y(l) - x_j(l)| + \beta \max_j \max_l |y(l) - x_j(l)|}{|y(l) - x_j(l)| + \beta \max_j \max_l |y(l) - x_j(l)|} \tag{8}$$

record: $\Delta_j(l) = |y(l) - x_j(l)|$, then:

$$\phi_j(l) = \frac{\min_j \min_l \Delta_j(l) + \beta \max_j \max_l \Delta_j(l)}{\Delta_j(l) + \beta \max_j \max_l \Delta_j(l)} \tag{9}$$

where $\beta \in (0, \infty)$ is denoted the resolution coefficient, which is introduced to enhance the significance of the difference between ϕ. The greater β is, the greater the resolution is. In all, the cost peak of β is (0,1). When $\beta \le 0.5463$, the resolution is the best.

(4) Computation of related degree. There are many related degree values of the correlation coefficient. As the message is too dispersed for every contrast, it is essential for requiring the average value of the correlation degree value, concentrate the related coefficient of each indicator into one value, and express it as the number of related degrees among the contrast kinds and the cited series. The related equation is as follows.

$$Y_j = \frac{1}{m} \sum_{l=1}^{m} \phi_j(l), l = 1, 2, 3, \ldots, m \tag{10}$$

(5) Correlation degree ranking: The correlation degree of each factor is ranked from high to low, if, then the characteristic series is more similar to the correlation series.

4. **4. Seasonal tourism demand forecast based on machine learning.**

4.1. **4.1 Data dimensionality reduction of seasonal tourism demand factors.** On the basis of the above construction of the correlation degree of seasonal tourism factors, this paper mainly constructs a seasonal tourism demand prediction model based on machine learning and makes predictions. First, data dimensionality reduction technology is used to extract the main factors affecting seasonal tourism demand and reduce the dimensionality data of factors affecting seasonal tourism to a lower dimension. Then, the extracted main factors are taken as the enter of the BP network model. Aiming at these problems of unstable forecast outcome and low accuracy caused by defects such as random assignment of the weight threshold of BP neural network, genetic method is introduced to make optimal the weight and bias parameters of BP neural network, so as to improve the performance of seasonal tourism demand prediction. The model structure is shown in Figure 3.

Principal Component Analysis (PCA) is a usually used message dimensionality simplification technology that is able to convert high-dimensional message to a low-dimensional

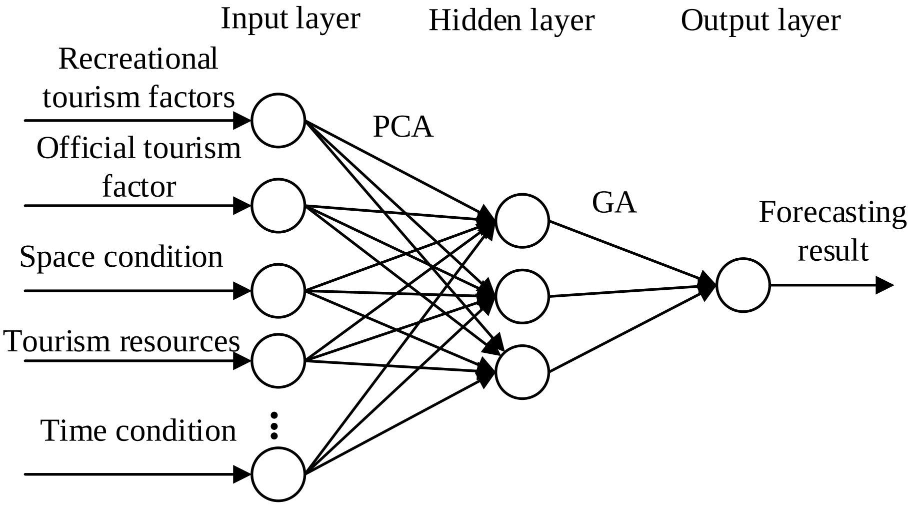

Figure 3. Model structure

representation while maximizing the retention of message in the data. The basic meaning of PCA is to convert the initial data into a set of novel variables, which are linear combinations of variables in the initial data and have certain special properties, that is, they are independent of each other and have the maximum variance in turn. Through PCA, this paper reduces the data of factors affecting seasonal tourism to a lower dimension, while retaining most of the information of the original data to achieve the purpose of data simplification. As a complex problem, the demand for seasonal tourism is analyzed based on the five main factors (recreational tourists, cultural tourists, official tourists, family and personal affairs tourists, fitness tourists) and three object factors in the previous section. Among the selected indicators, there is a certain information overlap problem to some extent. In order to reduce the correlation between the factors and make the prediction results of the model more effective and accurate, the index variables were reorganized, and the common factors were grouped into one class, and the principal component scores were re-extracted to forecast the seasonal tourism demand.

(1) Data preprocessing. The matrix Y of m indexes of the original q-dimensional random variable is standardized to obtain X. Normalized to: $x_{ji} = \frac{y_{ji}-\bar{y}_i}{t_i}$, where $X = \{x_{ji}\}$, $Y = \{y_{ji}\}$, $j = 1, 2, \ldots, q$.

(2) Discover the related coefficient matrix S. $S = \frac{X^T X}{m-1}$.

(3) Dealing with the feature equation of the correlation matrix S: $|S - \mu I_Q| = 0$, solve the equation to find the eigenvalues and eigenvectors of the coefficient matrix S. The eigenvalue μ_i, where $i = 1, 2, \ldots, q$, is obtained, and the eigenvalue is sorted from largest to smallest. The eigenvalue represents the influence of the principal component. The principal component that ultimately needs to be retained can be selected based on the feature root (< 1) or the cumulative variance explanation rate ($\geq 85\%$). At the same time, the utilization rate is generally represented by the cumulative variance explanation rate. Then the eigenvector $Sa = \mu_i a$ corresponding to the eigenvalue μ_i is solved according to the following equations.

$$\frac{\sum\limits_{i=1}^{n} \mu_i}{\sum\limits_{i=1}^{q} \mu_i} \geq 85\% \tag{11}$$

(4) Determine the principal component: $U_j = x_i^s a_j$. Where $x_1 = (x_{1i}, x_{2i}, \ldots, x_{mi})$, $U_j = (U_{j1}, U_{j2}, ..., U_{jq})$ $j = 1, 2, \ldots, n$, $i = 1, 2, \ldots, q$. For U, the first principal component, the second principal component, and so on, a total of n principal components are selected.

(5) Calculate the principal component load, and obtain the correlation coefficient between the corresponding principal component and the variable: $R_n = \sum_{i=1}^{m} \sum_{j=1}^{q} \mu_i a x_{ji}$

4.2. **Seasonal tourism demand forecast based on machine learning model.** Through the main analytic hierarchy process (PCA) in the above section, the main factors of the influencing indicators of seasonal tourism demand are extracted, and three principal component indicators (subject, object and resource) are obtained and used as input variables to forecast seasonal tourism demand. There are only three input variables and one output variable involved, so only one input layer, one hidden layer and one output layer are required. In this paper, genetic algorithm is used to search the parameter space of BP neural network to find the optimal weight and bias. In genetic algorithms, a BP neural network is treated as an individual whose weights and biases are encoded as genotypes and its performance is evaluated using a fitness function. Genetic manipulation is then used to generate a new BP neural network, select and eliminate individuals according to the fitness function, and finally find the optimal solution. The flow chart of optimizing BP neural network by genetic algorithm is indicated in Figure 4.

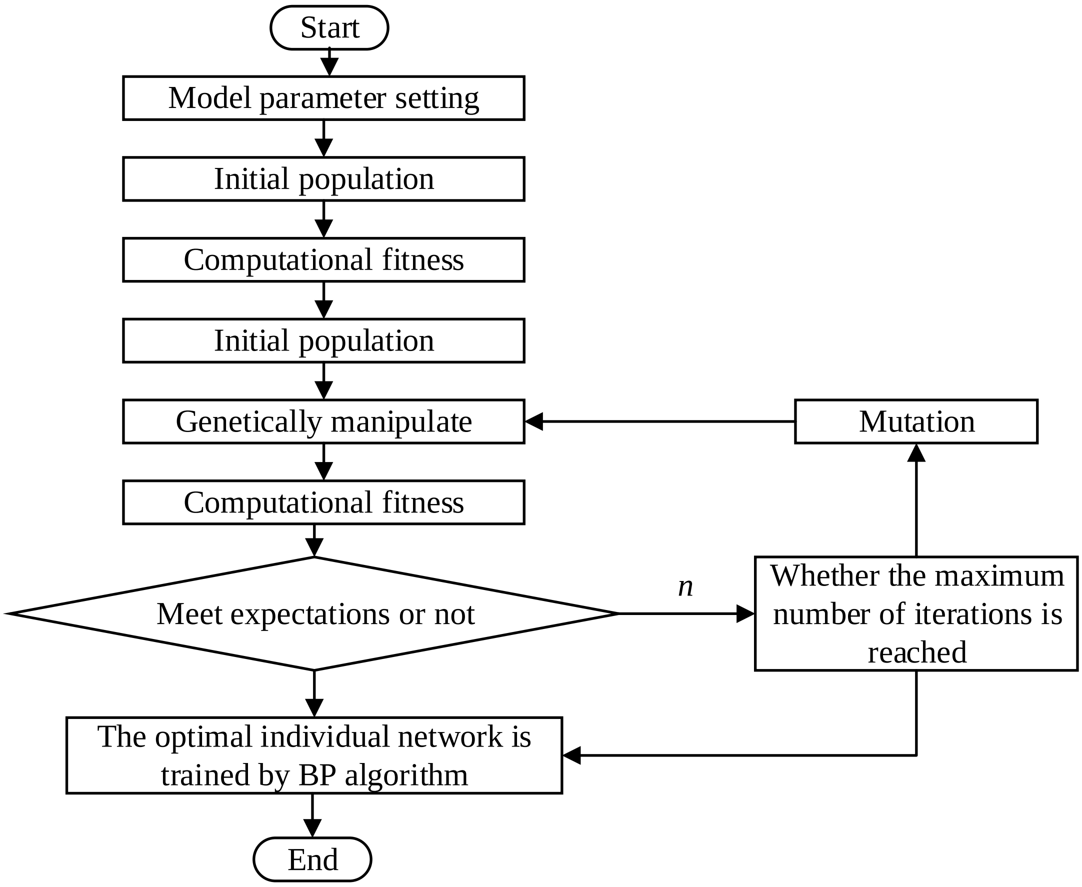

Figure 4. Optimization of BP neural network flow chart by genetic algorithm

(1) Conduct training and testing on the number of nodes in different hidden layers to find out the optimal neural network structure. The input value J_i of the i node in the hidden layer is calculated as shown in the following equation, where y_j represents the

position of the j-th neuron; H_j represents the weight matrix between the hidden layer j and the previous layer.

$$J_i = \sum_{j=1}^{n} H y_j + c_i \tag{12}$$

The calculation of the output value Q_i of the hidden layer i node is shown in the next equation. Here, ϕ represents the hidden layer excitation function.

$$Q_i = \phi(J_i) = \phi\left(\sum_{j=1}^{n} H_j y_j + c_i\right) \tag{13}$$

The input value J of the output layer is calculated as shown in the following equation.

$$J = \sum_{i=1}^{p} H_i Q_i + c = \sum_{i=1}^{p} H_i \phi\left(\sum_{j=1}^{n} H_j y_j + c_i\right) + c \tag{14}$$

(2) The optimal neural network is pre-trained for a certain amount of times, and the value range $[F_{\min}, F_{\max}]$ of weight and threshold is obtained, and calculate the individual fitness value F.

$$F = L(\sum_{j=1}^{m} cdt(x_j - Q_j)) \tag{15}$$

(3) Compute every peculiar evaluation affair and rank it. Network peculiar is able to be selected in terms of the following credibility value: $q_j = h \Big/ \sum\limits_{j=1}^{M} h_j$, where h_j is the adaptation value of individual j, which can be extented by the fault square, that is

$$h_j = \sum_{1}^{q} \sum_{1}^{l} (G_l - S_l)^2 \tag{16}$$

where $j = 1, 2, \ldots, M$ is the number of chromosomes, $l = 1, 2, 3, 4, 5$ is the number of output node layers, $q = 1, 2, 3, 4, 5$ is the amount of learning samples, and S_l is the target value.

(4) The cross operation of individuals v_j and v_{j+1} with probability Q_D generates new individuals v_j' and v_{j+1}', and the peculiar without cross function is directly copied.

(5) Use probability Q_N mutation to produce a new individual of v_i'.

(6) Insert novel peculiar into population Q and use Equation (17) to compute rating operation of novel peculiar.

$$Q = \phi\left(J\right) = \phi(\sum_{i=1}^{p} H_i \Phi(\sum_{j=1}^{n} H_j y_j + c_i) + c) \tag{17}$$

(7) If a satisfactory new individual is found, end, otherwise go to Step (4).

After all the required performance indicators are achieved, the optimized weight coefficient can be obtained by decoding the most individual in the final population.

5. Performance test and analysis.

5.1. **Prediction result analysis.** To assert the deed of the forecast model suggested in this article, Hong Kong, China is selected as the target city, and its monthly tourism volume is taken as the empirical research data to scientifically evaluate the accuracy and effectiveness of the designed forecast framework. In this article, monthly inbound visitor data for Hong Kong (January 2011 to August 2022) is collected from the WIND database (http://www.wind.com.cn/). From the collected data, we can see that the amount of tourists arriving in Hong Kong is seasonal and volatile, especially in February 2020, there was a "breaking point" phenomenon. For the purpose of verifying the performance of the prediction methods proposed in this article, the methods mentioned in literature [12, 16, 21] were also compared with this database, and all experiments were done on the Python platform of personal computer. For the convenience of analysis, reference [12] is denoted as FPTD, reference [16] as CTDA, reference [21] as CFTD, and the algorithm in this paper is denoted as STDF.

As can be seen from Figure 5, only the BP network is used in the tourist demand forecast of Hong Kong for each quarter from January 2020 to June 2022, marked as 1, 2, ... ,10. The fit degree of the early test set is not high, but the predicted number of tourists from June to August 2019 has a good consistency and high coincidence degree with the real amount of tourists. Moreover, when the amount of tourists enters the peak period, the divergence among the forecasted value and the true data becomes larger and larger, which is because the amount of tourists fluctuates greatly during the peak period. The use of historical data in the early learning period retained little information, coupled with the impact of holidays, so in the later period of the test set, the predicted outcome deviated from the actual results.

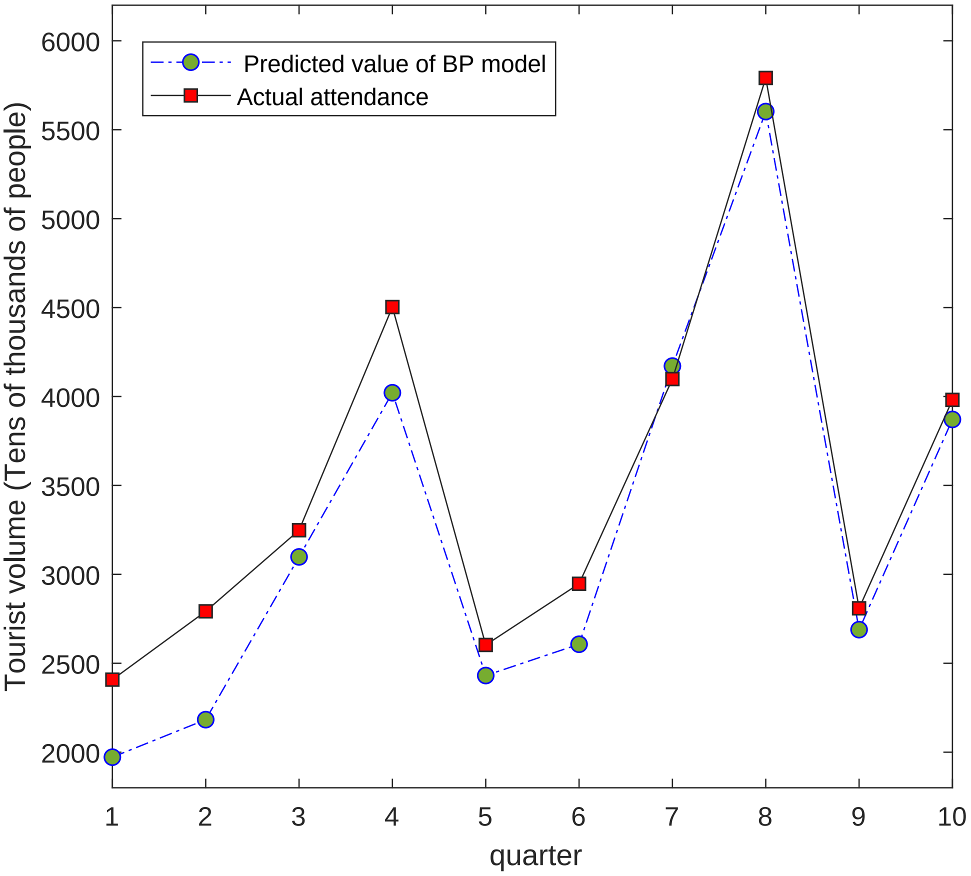

Figure 5. Prediction curve based on BP model

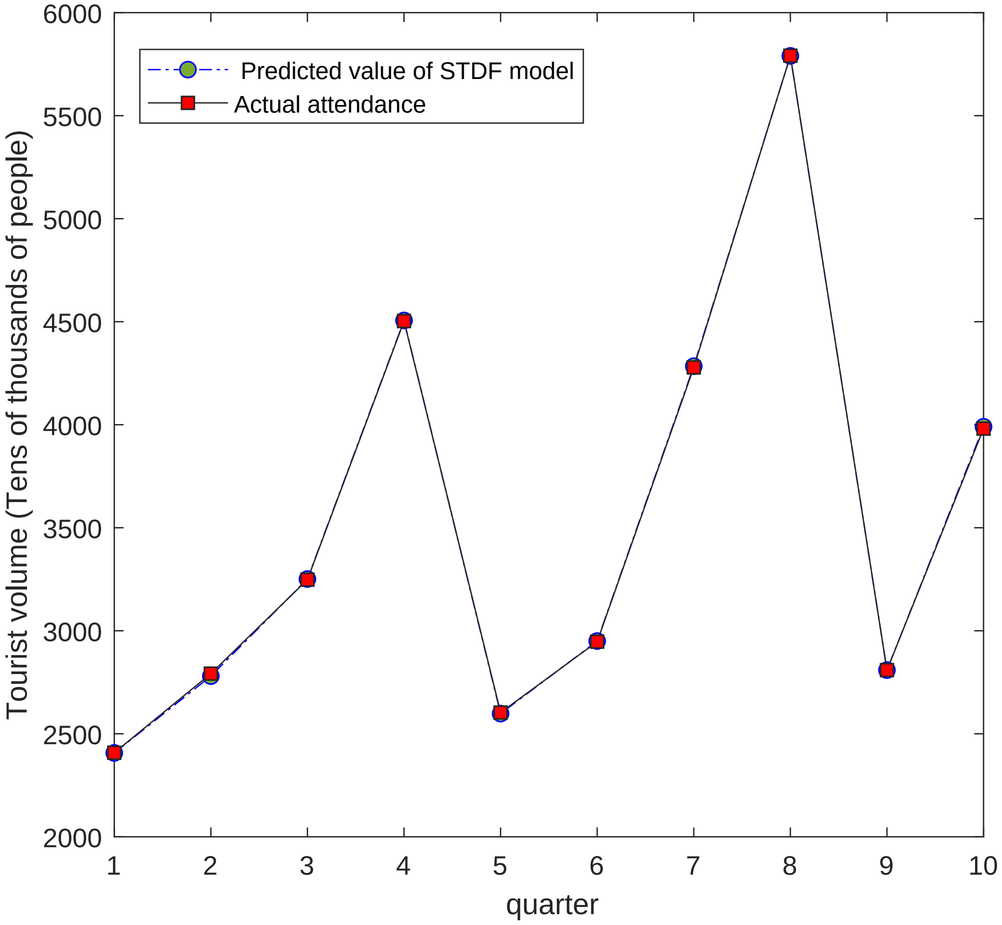

Figure 6. Prediction curve based on STDF model

The steps of establishing the STDF model on the ground of adding the genetic probability value are similar to those above, as shown in Figure 6. On the ground of BP network, STDF model was established by introducing genetic factors. Comparing the forecasted value and the actual tourist volume, it can be intuitively seen from the two line charts that the forecasted value of the STDF model is closer to the actual tourist volume line and almost coincides, which indicates that the forecast mistake is small. The Internet search data information in this period can effectively predict the number of Hong Kong tourists in advance and effectively monitor the number of future tourists with timeliness.

5.2. **Comparative analysis of predictive performance.** For the goal of evaluating and compare the prediction accuracy of the methods, four evaluation criteria commonly found in literature were adopted, including mean absolute error (MAE), mean absolute percentage error (mean absolute percentage error), and mean absolute percentage error (MAE). MAPE, root mean square error (RMSE), and root mean square percentage error (RMSPE), as indicated in Table 1.

As can be seen from Table 1, the CFTD model with the introduction of genetic factors has higher prediction accuracy than the CTDA model, and MAE, MAPE, RMSE and RMSPE all decline, among which MAPE drops to 4.54% and RMSE to 26.17. Compared with the single model CTDA, the FPTD model STDF with the introduction of BP neural network model has a significant improvement in the horizontal accuracy, especially when all weights are used as predictors, the prediction accuracy is further improved. MAE, MAPE, RMSE and RMSPE decreased to 28.74, 6.39%, 34.91 and 7.82%, respectively.

Table 1. Comparison of model prediction performance.

Model	MAE	MAPE(%)	RMSE	RMSPE
CTDA	32.96	7.48	38.19	9.54
FPTD	28.74	6.39	34.91	7.82
CFTD	21.97	4.54	26.17	5.23
STDF	13.73	2.49	18.82	3.68

Finally, the seasonal tourism demand forecasting model based on machine learning presented in this study showed the best forecasting effect, with MAE, MAPE, RMSE and RMSPE of 13.73, 2.49%, 18.82 and 3.68%, respectively. The empirical study found that when tourism subjects, conditions of tourism realization and tourism objects are all taken as input variables to predict tourist flow, the horizontal prediction errors are all lower than other benchmark models. Therefore, combining the factors affecting seasonal tourism with machine learning to predict the number of tourists can improve the prediction accuracy.

Table 2. Comparison of the relative improvement value RI

Model	Reference model	RI(%)
CTDA	Snaive	24.58
FPTD		18.43
CFTD		11.71
STDF		3.94

Through the Relative Improvement value RI in Table 2, we can further intuitively see the improvement effect of the STDF model suggested in this paper on MAPE compared with the Snaive model. RI is an indicator that compares the relative size of the improvement effect of different schemes on a certain indicator. RI is often used to assess the extent to which different regimens in an experiment or trial improve on an indicator. The greater the value of RI, the greater the improvement effect. Compared with Snaive model, CTDA model, FPTD model, CFTD model and STDF model are significantly improved, and the relative improvement values of MAPE of CTDA model, FPTD model, CFTD model and STDF model are 24.58%, 18.43%, 11.71% and 3.94% respectively. In particular, compared with CTDA model, FPTD model and CFTD model, the prediction effect of STDF model has been improved, and the relative improvement value of MAPE of STDF model is within 10%, which further indicates that STDF model shows the best prediction effect.

6. **Conclusion.** Aiming at the low forecasting performance of existing tourism forecasting models, this paper proposes a seasonal tourism demand forecasting model based on machine learning under big data environment. The model first analyzes the relevant factors affecting the seasonality of tourism, and analyzes their correlation degree. Then, the high-dimensional factor data is transformed into low-dimensional representation, and at the same time, the information in the data is preserved to the maximum extent, and the re-extracted principal component score is used to forecast the seasonal tourism demand. Secondly, genetic algorithm is adopted to optimize BP neural network to find the optimal solution in large-scale search problems. Finally, the experimental outcome indicates that the suggested method can greatly enahance the MAE, MAPE, RMSE and RMSPE of the prediction model, and can be excellently applied to seasonal tourism prediction. Because genetic algorithm can only optimize the weight and bias of one BP network each time,

the training speed of genetically optimized BP neural network model is relatively slow. In the future, we will consider using other more efficient optimization algorithms to replace genetic optimization algorithms, such as gradient-based optimization algorithms.

Acknowledgment. This work was supported by Jinzhong University rural tourism Collaborative innovation center (No. SSKIZDKT2020166), Philosophy and Social Project of Higher Education Department of Shanxi Province (No. 2020W189), and Shanxi Province philosophy and social science planning project (No. 2022YY178).

REFERENCES

[1] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293-314, 2014.

[2] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.

[3] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, 2339, 2023.

[4] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, 1977, 2023.

[5] C.-M. Chen, S. Lv, J. Ning, and J. M.-T. Wu, "A Genetic Algorithm for the Waitable Time-Varying Multi-Depot Green Vehicle Routing Problem," *Symmetry*, vol. 15, no. 1, 124, 2023.

[6] K. Wang et al., "Uncovering Hidden Vulnerabilities in Convolutional Neural Networks through Graph-based Adversarial Robustness Evaluation," *Pattern Recognition*, vol. 143, 109745, 2023.

[7] S. Li, T. Chen, L. Wang, and C. Ming, "Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index," *Tourism Management*, vol. 68, pp. 116-126, 2018.

[8] R. G. Fritz, C. Brandon, and J. Xander, "Combining time-series and econometric forecast of tourism activity," *Annals of Tourism Research*, vol. 11, no. 2, pp. 219-229, 1984.

[9] E. Kaynak and J. A. Macaulay, "The Delphi technique in the measurement of tourism market potential: the case of Nova Scotia," *Tourism Management*, vol. 5, no. 2, pp. 87-101, 1984.

[10] H. Song and G. Li, "Tourism demand modelling and forecasting—A review of recent research," *Tourism Management*, vol. 29, no. 2, pp. 203-220, 2008.

[11] C. Goh and R. Law, "The methodological progress of tourism demand forecasting: A review of related literature," *Journal of Travel & Tourism Marketing*, vol. 28, no. 3, pp. 296-317, 2011.

[12] Z.-C. Li and D. Sheng, "Forecasting passenger travel demand for air and high-speed rail integration service: A case study of Beijing-Guangzhou corridor, China," *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 397-410, 2016.

[13] A. Jungmittag, "Combination of forecasts across estimation windows: An application to air travel demand," *Journal of Forecasting*, vol. 35, no. 4, pp. 373-380, 2016.

[14] P. F. Bangwayo-Skeete and R. W. Skeete, "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach," *Tourism Management*, vol. 46, pp. 454-464, 2015.

[15] G. Athanasopoulos et al., "The tourism forecasting competition," *International Journal of Forecasting*, vol. 27, no. 3, pp. 822-844, 2011.

[16] V. Cho, "A comparison of three different approaches to tourist arrival forecasting," *Tourism Management*, vol. 24, no. 3, pp. 323-330, 2003.

[17] R. Fildes, Y. Wei, and S. Ismail, "Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures," *International Journal of Forecasting*, vol. 27, no. 3, pp. 902-922, 2011.

[18] Y.-H. Liang, "Forecasting models for Taiwanese tourism demand after allowance for Mainland China tourists visiting Taiwan," *Computers & Industrial Engineering*, vol. 74, pp. 111-119, 2014.

[19] K. N. Lee, "Forecasting long-haul tourism demand for Hong Kong using error correction models," *Applied Economics*, vol. 43, no. 5, pp. 527-549, 2011.

[20] S. C. Kon and L. W. Turner, "Neural network forecasting of tourism demand," *Tourism Economics*, vol. 11, no. 3, pp. 301-328, 2005.

[21] O. Claveria and S. Torra, "Combination forecasts of tourism demand with machine learning models," *Applied Economics Letters*, vol. 23, no. 6, pp. 428-431, 2016.

[22] P.-F. Pai, K.-C. Hung, and K.-P. Lin, "Tourism demand forecasting using novel hybrid system," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3691-3702, 2014.

[23] R. Durbarry and M. T. Sinclair, "Market shares analysis: The case of French tourism demand," *Annals of Tourism Research*, vol. 30, no. 4, pp. 927-941, 2003.

[24] X. Li et al., "Forecasting tourism demand with composite search index," *Tourism Management*, vol. 59, pp. 57-66, 2017.

[25] J. Alegre and L. Pou, "The length of stay in the demand for tourism," *Tourism Management*, vol. 27, no. 6, pp. 1343-1355, 2006.

[26] E. Koc and G. Altinay, "An analysis of seasonality in monthly per person tourist spending in Turkish inbound tourism from a market segmentation perspective," *Tourism Management*, vol. 28, no. 1, pp. 227-237, 2007.

[27] Y. Yu and S. M. Wang, "The forecasting research of beijing tourism demand based on the BP neural network," *Applied Mechanics and Materials*, vol. 571, pp. 128-131, 2014.

[28] O. Claveria and S. Torra, "Forecasting tourism demand to Catalonia: Neural networks vs. time series models," *Economic Modelling*, vol. 36, pp. 220-228, 2014.

[29] H. Hassani et al., "Forecasting US tourist arrivals using optimal singular spectrum analysis," *Tourism Management*, vol. 46, pp. 322-335, 2015.

[30] J. Kim and C.-K. Lee, "Role of tourism price in attracting international tourists: The case of Japanese inbound tourism from South Korea," *Journal of Destination Marketing & Management*, vol. 6, no. 1, pp. 76-83, 2017.

[31] M. Buscema, "Back propagation neural networks," *Substance Use & Misuse*, vol. 33, no. 2, pp. 233-270, 1998.

[32] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, pp. 8091-8126, 2021.