International Journal of Geo-Information

MDPI

Article

Forecasting of Short-Term Daily Tourist Flow Based on Seasonal Clustering Method and PSO-LSSVM

Keqing Li *, Changyong Liang, Wenxing Lu, Chu Li, Shuping Zhao and Binyou Wang

School of Management, Hefei University of Technology, Hefei 230009, China; cyliang@hfut.edu.cn (C.L.); luwenxing@hfut.edu.cn (W.L.); 2019170692@mail.hfut.edu.cn (C.L.); zhaoshuping1753@hfut.edu.cn (S.Z.); 2013111070@mail.hfut.edu.cn (B.W.)

* Correspondence: lkqing1995@mail.hfut.edu.cn

Received: 12 October 2020; Accepted: 13 November 2020; Published: 13 November 2020

Abstract: The accurate prediction of tourist flow is essential to appropriately prepare tourist attractions and inform the decisions of tourism companies. However, tourist flow in scenic spots is a dynamic trend with daily changes, and specialized methods are necessary to measure it accurately. For this purpose, a tourist flow forecasting method is proposed in this research based on seasonal clustering. The experiment employs the K-means algorithm considering seasonal variations and the particle swarm optimization-least squares support vector machine (PSO-LSSVM) algorithm to forecast the tourist flow in scenic spots. The LSSVM is also used to compare the performance of the proposed model with that of the existing ones. Experiments based on a dataset comprising the daily tourist data for Mountain Huangshan during the period between 2014 and 2017 are conducted. Our results show that seasonal clustering is an effective method to improve tourist flow prediction, besides, the accuracy of daily tourist flow prediction is significantly improved by nearly 3 percent based on the hybrid optimized model combining seasonal clustering. Compared with other algorithms which provide predictions at monthly intervals, the method proposed in this research can provide more timely analysis and guide professionals in the tourism industry towards better daily management.

Keywords: seasonal clustering; short-term forecast; tourism flow forecast; optimization algorithm

1. Introduction

In recent years, owing to steady improvements in the standards of living, tourism has become an important part of leisure and lifestyle for people worldwide. According to data released by the World Travel Tourism Council, tourism was the third largest industry in the world in terms of the growth rate of Gross Domestic Product (GDP) in 2019. The growth rate of tourism was reportedly 3.5%, which was significantly greater than the global economic GDP growth rate of 2.5% [1]. In particular, the tourism industry created nearly 80 million jobs in China, accounting for 10.3% of the country's total labor force. At the same time, its output value was estimated to be 10.9 trillion Yuan, accounting for 11.3% of China's economy [1]. The rapid development of the world's tourism industry has promoted the vigorous development of China's own tourism industry. China's tourism industry has entered the stage of 'mass tourism', with people's willingness to travel constantly rising [2]. It is expected that the domestic tourism market will continue to thrive even in the post-epidemic era [3].

With the promotion of the economic improvement of the country and the region, the rapid development of tourism has also ushered in multiple problems pertaining to daily management services at tourist destinations, particularly at mountainous scenic spots, which play a pivotal role in Chinese tourism [4]. Their unique topography and landforms, extensive spatial range, poor natural conditions, and severe seasonal conditions make them inaccessible to personnel. In particular, the delivery of materials and resources, scheduling of arrangements for transportation, etc., pose

particular challenges to management services in mountainous environments [5]. The effects of these challenges are primarily reflected in delays in passenger flow. All tourist destinations experience heavy tourist seasons and off-seasons, resulting in a serious seasonal imbalance in the tourist flow [6]. During the tourism season, spots are often overcrowded. This causes traffic congestion, overextends hotel, catering, and personnel supplies, leads to the overutilization of tourism resources and the environment, and degrades the quality of service for tourists, reducing overall tourist satisfaction. On the other hand, the oversaturation of tourists in specific spots also poses a threat to their own personal safety [6]. For example, on 4 October 2014, due to a surge in the number of tourists during the Golden Week, the passenger capacity at the Three Gorges scenic spot in Yichang, Hubei, was insufficient, resulting in hundreds of tourists being stranded at the terminal. On 2 October 2013, several tourists were stuck at the entrance of Jiuzhaigou Valley because of overcrowding. On 26 October 2014, the traffic was almost paralyzed at the Beijing Xiangshan area, leading to thousands of people being stranded at the bus station. Furthermore, the Golden Week of Tourism has been witness to a series of security incidents which have resulted in a poor travel experience for tourists [7]. However, during the off-season, the number of tourists at destinations are considerably low, resulting in idle hotels and wasted resources, materials, personnel, etc. These considerations corroborate the significance of the accurate forecast of tourist flow in the tourism industry.

Tourist flow forecasting can be divided into two categories: long-term forecasting and short-term forecasting. Both have important implications, and the determination of an accurate trend can aid professionals in the tourism industry [8,9], particularly with respect to problems such as optimal allocation of resources and managerial staff [10].

The forecasting of tourist flow in tourist destinations is affected by several factors, including weather [11], climate [12], and temperature [13]. Tourism is inherently seasonal [14] as the constraints of time and climate create inevitably unbalanced tourist flows [15]. Both natural seasons and artificial seasons defined by holidays and other institutional factors play a part in the determination of tourist flow [16]. Thus, both factors must be considered during prediction attempts. To the best of our knowledge, scant attention has been paid to seasonality in previous works on this topic. For instance, Huang and Min established a seasonal autoregressive average model combined with a difference method to eliminate seasonal effects on tourist flow forecasting, and experimentally verified its effectiveness [17]. However, these studies have focused solely on the elimination of seasonal influences on the prediction of tourist flow by proposing seasonal index adjustments or by establishing a seasonal model. Few studies have considered the influence of the alternatives of natural seasons in the forecast of tourist flow.

Tourist flows exhibit complicated non-linear variations. This makes it difficult to identify a relationship between the tourist flow later and the current influencing variables based on simple mathematical models. In recent years, with the development of machine learning, nonlinear models have been widely used in short-term time series forecasting. For instance, artificial neural network (ANN)-based methods and support vector machines (SVM) have already been used in the forecasting of tourist flow [18,19]. However, neural network-based models lack a systematic procedure for model construction because of their flexibility. This necessitates multiple trials to identify the optimal parameters required to obtain a reliable neural model [20]. Compared with ANN, SVM is more capable of avoiding problems such as data overfitting and local minima while maintaining positive features such as robustness. Moreover, SVM is less complicated than ANN in terms of parameter selection [21]. The LSSVM is an upgraded version of SVM that was developed to improve the accuracy of the standard SVM [22]. Compared to SVM, it is capable of using equality constraints instead inequalities, enabling it to solve sets of linear equations instead of being restricting to quadratic programming [23]. However, the prediction accuracy of the LSSVM algorithm is significantly dependent on the selection of two specific parameters [24]. To address this drawback, certain optimization algorithms, including the genetic algorithm (GA) and the fruit fly optimization algorithm, are used to identify the optimal values of the LSSVM parameters to enhance its prediction accuracy [25,26]. Among those intelligence-based

optimization algorithms, PSO, proposed by Kennedy and Eberhart [27], has been widely used in optimization processes, model classification, machine learning, and neural network training [28] owing to its ease of implementation and its high coherence and coordination [29].

In addition to the development of such optimization algorithms, some studies have attempted to curate relevant information by analyzing comments on online forums. Certain researchers have used search engine data to forecast hotel demands [30,31] by designing a composite search index to forecast tourist flow [32]. Furthermore, Google Trends has been widely used to improve the performances of traditional models [10,33,34]. Related works have pointed out that combining different data sources and techniques can lead to higher accuracy [35]. Even price levels and web traffic have been as used as variables in certain studies [36]. User interactions on online forums have also been used to forecast tourist flows [37]. However, most of the methods are more suitable for long-term forecasting, rather than short-term forecasting.

As few research studies have been conducted to investigate short-term forecasting methods and substitutes to natural seasons in the forecasting process of tourist flows, we propose a seasonal clustering-based method, which can classify seasons based on their characteristics to address this shortcoming. We combine seasonal re-clustering and the PSO-LSSVM model and apply the combination for short-term daily tourist flow forecasting. The crucial hypothesis in this research is that seasonal clustering could improve tourist flow forecasting. Our results confirm the validity of the hybrid optimized model combining seasonal clustering and provide practically useful implications for management.

The remainder of this research is organized as follows. Section 2 presents the methods, including principles underlying the least squares support vector machine (LSSVM) and the particle swarm optimization (PSO) algorithms, and an illustration of the PSO-LSSVM procedure that considers seasonal clustering, and the experiments. Section 3 details their results. Section 4 is the discussion. Finally, Section 5 presents the conclusions, as well as the limitations and implications of this research.

2. Methods

2.1. Least Squares Support Vector Machine

The essential characteristic of LSSVM is that it is designed to utilize equality constraints and transform quadratic programming problems to problems of direct solution of quadratic equations. Consider a dataset (x_i, y_i), $x_i \in R^n$, $y \in R$, where x_i denotes the ith input item in an n-dimensional space and y_i denotes the output value corresponding to x_i, l is the total number of data points, $i = 1, 2, \cdots\cdots l$, and n is the number of dimensions of input variables. As a non-linear prediction model, the LSSVM model can be expressed as follows:

$$f(x) = w^T \phi(x) + b, \tag{1}$$

where w denotes the weight vector, b is the offset, and $\phi(x)$ represents a nonlinear transformation that maps the input data (x_i) into a high-dimensional feature space. According to the structure minimization principle, the optimization objective function of the LSSVM can be expressed as follows:

$$min \frac{1}{2} \|w\|^2 + \frac{1}{2} C \sum_{i=1}^{l} e_i^2, \tag{2}$$

$$s.t. w^T \phi(X_i) + b + e_i = y_i, i = 1, 2, \cdots\cdots,$$

where e_i denotes the error and C represents a positive penalty coefficient. A Lagrange multiplier, λ_i, is introduced to solve the optimization problem. Hence, Equation (2) can be transformed into the following form:

$$L(w, \lambda_i, b, e_i) = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{l} e_i^2 - \sum_{i=1}^{l} \lambda_i\Big(w^T\phi(x_i) + b + e_i - y_i\Big), \tag{3}$$

Next, the partial derivatives corresponding to each variable of Equation (3) are calculated:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} \lambda_i\phi(x_i) = 0 \Rightarrow w = \sum_{i=1}^{l} \lambda_i\phi(x_i), \tag{4}$$

$$\frac{\partial L}{\partial \lambda_i} = -\sum_{i=1}^{l}\Big(w^T\phi(x_i) + b + e_i - y_i\Big) = 0 \Rightarrow y_i = w^T\phi(x_i) + b + e_i, \tag{5}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} \lambda_i = 0, \tag{6}$$

$$\frac{\partial L}{\partial e_i} = \frac{1}{2} \times 2C\sum_{i=1}^{l} e_i - \sum_{i=1}^{l} \lambda_i = 0 \Rightarrow \lambda_i = Ce_i, \tag{7}$$

The variables, w and e_i, are then eliminated. This yields the following linear equation:

$$\begin{bmatrix} 0 \\ Y \end{bmatrix} = \begin{bmatrix} 0 & A^T \\ A^T & B + C^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \lambda \end{bmatrix}, \tag{8}$$

where $Y = (y_1, y_2, ..., y_l)$, $A = (1, 1, ..., 1)^T$, $B_{ij} = \phi(x_i)^T\phi\Big(x_j\Big)$, $\lambda = (\lambda_1, \lambda_2, ...\lambda_l)$, and I denotes the unit matrix. Hence, the LSSVM can be expressed as follows:

$$y = \sum_{i=1}^{l} \lambda_i K(x, x_i) + b, \tag{9}$$

where $K(x, x_i)$ denotes the kernel function of a feature space.

2.2. Particle Swarm Optimization

A PSO algorithm begins by initializing a random group of particles and obtains the optimal solution after performing several iterative searches. During each iteration, the particles update their positions and velocities based on individual and global extrema. Let us assume that there is a total of N particles that are initialized and scattered in a D-dimensional space. Further, assume that the position of the ith particle is $X_i = (x_{i1}, x_{i2}, \cdots\cdots, x_{iD})$, and that the current best position for the ith particle is $local\_x_i = (local\_x_{i1}, local\_x_{i2}, \cdots\cdots, local\_x_{iD})$, whereas the best position found by the entire swarm is $global\_x_i = (global\_x_{i1}, global\_x_{i2}, \cdots\cdots, global\_x_{iD})$. In such a scenario, the new position of a particle after t time-instants is obtained by adding the velocity vector $V_i = (v_{i1}, v_{i2}, \cdots\cdots, v_{iD})$ to its current position. This can be expressed as follows:

$$x_{iD}^{(t+1)} = x_{iD}^t + wP \times v_{iD}^{t+1}, \tag{10}$$

The velocity of any particle is updated using the following formula:

$$V_{iD}^{t+1} = wV \times V_{iD}^t + c_1 \times rand \times \Big(local_{x_{id}^t} - x_{id}^t\Big) + c_2 \times rand \times \Big(global_{x_{iD}^t} - x_{iD}^t\Big), \tag{11}$$

where c_1, c_2 denote the acceleration coefficients, wV, wP represent the elasticity coefficients with initial values equal to 1, *rand* denote two random numbers with uniform distributions in the range [0,1],

$local\_x_{id}^{t}$ is the best position identified by each individual particle, and $global\_x_{iD}^{t}$ is the best position identified by the global swarm.

2.3. Seasonal Clustering Approach

Several algorithms are used for clustering analysis, and they can be roughly divided into four categories [38]: (1) those based on cluster formation methodology, such as top-down, bottom-up, and analytical optimization techniques [39]; (2) those dependent on the cluster model obtained, such as stratification, centroids (e.g., K-means), distribution subspaces, and graph-based models; (3) those obtained via a membership function, which may be further subdivided into hard or soft clustering [40]; and (4) those that use groups to define the distinction between overlapping clusters and are less sensitive to noise because it becomes equally distributed among them [41].

The K-means clustering algorithm is a typical representative classification clustering algorithm due to its simplicity and effectiveness. It is particularly suitable for a simple clustering of big data. Considering that the primary characteristic of natural seasons is the change in weather [42], we attempt to analyze the correlation between climate-related factors and variations in daily tourist flow. The details of seasonal clustering are as follows.

Step 1: Analysis of the factors related to seasonal clustering.

Step 2: Input of the variables into the K-means algorithm to obtain the results of seasonal clustering.

2.4. Procedure of PSO-LSSVM Considering Seasonal Clustering

The present research primarily aims to prove that the use of seasonal clustering during the pre-processing of data is beneficial to the accurate prediction of daily tourist flow. Combined with historical tourist information, the PSO-LSSVM model is proposed to illustrate the positive impact of seasonal clustering on the prediction of tourist flow in tourist destinations. In the PSO-LSSVM model, the PSO algorithm is used as an optimization algorithm to optimize the regularization parameter (γ) and the kernel parameter (σ) of LSSVM. The considerations of seasonal clustering in PSO-LSSVM can summarized in the following steps.

Step 1. The natural seasons are clustered. The new natural season of the tourist destination combined with the spot's historical tourist data comprises a dataset. The original dataset is normalized and divided into training and test datasets.

Step 2. The parameters of the PSO algorithm, including population sizes, evolution times, and learning factors, are initialized.

Step 3. The swarm of particles is initialized with random individual velocities and positions.

Step 4. The various initialized parameters are fed into LSSVM, and then the fitness value of each particle is evaluated. In this research, the root mean squared error (RMSE) defined in the test dataset is used as the fitness function, as follows:

$$fitness = RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{12}$$

where n denotes the number data points in the dataset, and y_i and $\hat{y}_i$ represent the actual value and the estimated value, respectively. The local and global optima are then calculated following the fitness function.

Step 5. The velocity and position of each particle is updated using Equations (10) and (11).

Step 6. Steps 4 and 5 are repeated until the termination criterion is satisfied and the optimal values of the LSSVM parameters are obtained. The flow chart of the procedure of PSO-LSSVM is shown in Figure 1.

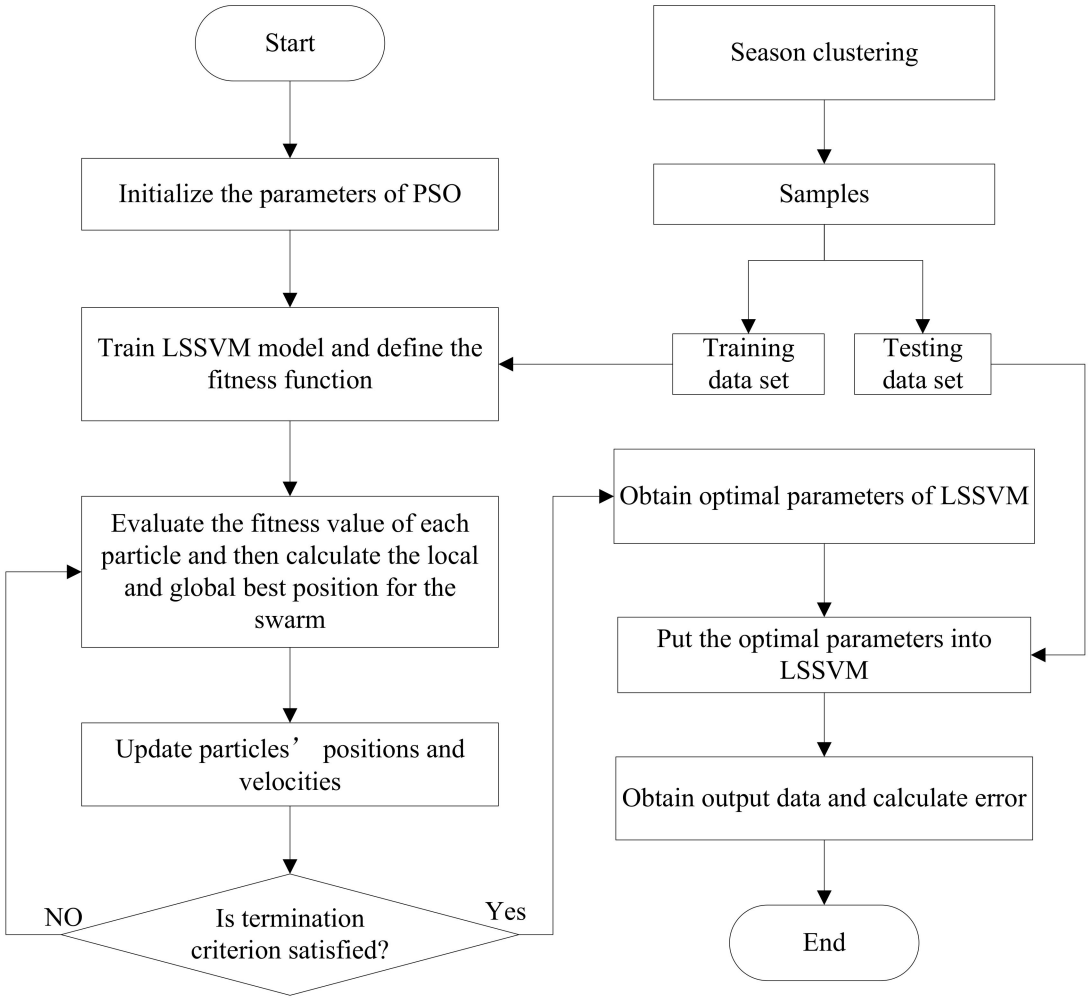

Figure 1. Flow chart of the procedure of the particle swarm optimization-least squares support vector machine (PSO-LSSVM).

In this research, to evaluate the forecasting accuracy, the mean absolute percentage error (MAPE) and RMSE are used as the evaluation criteria. It is evident that the values of MAPE or RMSE are inversely proportional to forecasting accuracy:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%, \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{14}$$

where $y_i, \hat{y}_i$ denote the actual and evaluated data, respectively, and n denotes the total number of data points in the test dataset. It should be noted that the RMSE indicator only considers the annual average in the last row of the table as a supporting indicator. Consequently, the MAPE indicator is more suitable for prediction of daily trends.

2.5. Data Preprocessing

To improve the accuracy of prediction, it is necessary to normalize the original sequence of input variables. The following normalized formula is adopted in this research:

$$u = \frac{(u_{\max} - u_{min}) \times (x_i - x_{min})}{x_{\max} - x_{min}} + u_{min} \quad (15)$$

where u denotes the normalized value with uniform distribution in the range [0,1]; and u_{max} and u_{min} are the upper and lower limits, respectively. In this research, it is assumed that u_{max} and u_{min}. x_i denotes the tourist flow on the ith day in the original one-year data series, and x_{min} and x_{max} denote the minimum and maximum values of the original sequence, respectively.

2.6. Data Collection and Correlation Analysis

To verify the feasibility of the proposed algorithm, the dataset of the daily tourist flow at Mountain Huangshan during the period of 2014 to 2017 is accessed, the tourist flow data comes from our cooperation project with Huangshan Management Committee. Besides, we investigated the spot's historical temperature and weather for this research; the temperature is measured in degrees Celsius and the weather is measured in different categories such as sunny, cloudy, heavy snow, moderate snow, and so on. The tourist flow dataset contains both original regular daily tourist flow data and original tourist flow data on holidays. Four types of data are included in the data set: X_1, the daily tourist flow on a particular day; X_2, the tourist flow volume on the same day in the previous week; X_3, the tourist flow volume on the same day of the previous year; and Y, the daily tourist flow on the subsequent day. Each type contains 1461 data points. The relationship between the historical tourist flow, which includes X_1, X_2, X_3, and the daily tourist flow of the subsequent day is primarily determined by the respective correlation coefficients—the correlation coefficients between pairs of data items are proportional to the suitability of the selected factors as inputs to the model.

Table 1 presents the correlation coefficients between X_1, X_2, X_3, and Y. As expected, X_1 is observed to be superior to the other factors. Consequently, X_1 is selected as the input variable in the proposed model.

Table 1. Correlation coefficients between the daily tourist flow of tomorrow and each element of the historical tourist flow.

	X_1	X_2	X_3
Y	0.726	0.468	0.347

In addition, the severity of weather, weekday, and official holiday are also added to the model as dummy variables X_4, X_5 and X_6. $X_4 = \begin{cases} 1 \\ 0 \end{cases}$, where 1 represents severe weather, such as blizzard, heavy snow, moderate snow, heavy rain, thunderstorms, and showers, which would significantly affect people's willingness to travel, and 0 represents non-severe weather, such as sunny, cloudy, and drizzle. X_5 represents a matrix which represents the day of the week. $X_6 = \begin{cases} 1 \\ 0 \end{cases}$, where 1 represents an official holiday; 0 represents an ordinary day. The use of dummy variables is another difference between our research and previous ones. The incorporation of such factors allowed us to approach the problem of prediction from a more microscopic perspective.

2.7. Parameter Initialization and the Addition of Seasonal Factors

The initial parameters are set as follows, the size of the swarm is taken to be 30, maximum number of iterations is set as 300, and acceleration coefficients c_1 and c_2 are 2 and 2, respectively. To verify whether the ambient natural season affects the accuracy of prediction of the tourist flow on

the subsequent day, a binary virtual variable-based approach is introduced to represent the different seasons; $s_i = \begin{cases} 1 \\ 0 \end{cases}$ (1 represents the ith natural season i =1, 2, 3, 4).

3. Results

The results of the experiments above are shown in this section.

3.1. Analysis of Influence of Original Natural Season

This research aims to investigate the effect of seasonal changes on tourist flow on the subsequent day. The daily tourist flow at scenic destinations varies dramatically over the different seasons, primarily because of the differences in temperature. In this part, the year is assumed to be divided into four seasons following the meteorological department's scheme: spring (March, April, and May), summer (June, July, and August), autumn (September, October, and November), and winter (December, January, and February) [15]. Figure 2 illustrates the distribution of the daily tourist flow on the subsequent day at Mountain Huangshan over the period of 2014 to 2017.

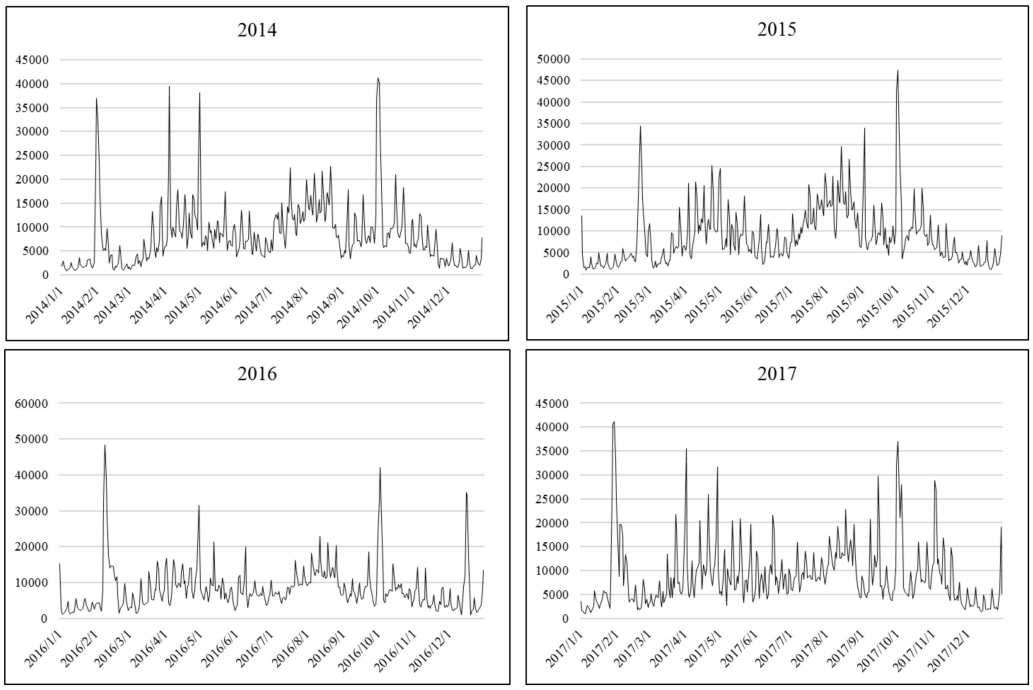

Figure 2. Daily tourist flow at Mountain Huangshan during 2014–2017.

It is clear from Figure 2 that due to the daily fluctuations in tourist flow, the distribution is complex and non-linear. Further, the daily tourist volume at Mountain Huangshan during the period from March to November is observed to remain high every year, whereas during December to January it appears to be consistently low. Further analysis of the data depicted in Figure 2 is presented in Tables 2 and 3.

Table 2. Total number of tourists during each season.

	Spring	Summer	Autumn	Winter
March 2014–February 2015	799,838	976,109	827,444	385,479
March 2015–February 2016	815,947	1,062,968	896,006	529,186
March 2016–February 2017	761,308	869,248	717,781	661,204

Table 3. Average number of tourists during each season.

	Spring	Summer	Autumn	Winter
March 2014–February 2015	8694	10,610	9093	4283
March 2015–February 2016	8869	11,554	9846	5815
March 2016–February 2017	8275	9448	7887	7346

Tables 2 and 3 reveal that the total tourist flow and the average tourist flow remain high during spring, summer, and autumn each year. It is further confirmed that the tourist flow is maximum during the summer and that it is the second highest during spring and autumn. The tourist volume in winter is significantly less than that during the other three seasons. Thus, it can be concluded that the tourist flows in different seasons are significantly different.

3.2. Predictions by the Models and Their Comparison before Seasonal Clustering

In this experiment, to satisfy the requirements of the model, the dataset is divided into a training dataset (2014–2016) and a test dataset (2017). To enhance the prediction accuracy, all the data are normalized using Equation (15) with a range of [0,1]:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{16}$$

where y denotes the normalized data, x denotes the original input data, and $x_{\max}$, $x_{\min}$ are the maximum and minimum values in the dataset, respectively.

Following that, the vectors $(X_1, X_4, X_5, X_6, S_1, S_2, S_3, S_4)$, including the natural seasons, are used as input variables in the predictive models, and the vectors (X_1, X_4, X_5, X_6), without considering seasonal factors, are used as input variables to the predictive models on a separate iteration for comparison purposes. Both the PSO-LSSVM algorithm and the LSSVM algorithm are adopted as predictive models for each of the two sets of input vectors. Table 4 presents the results of this experiment.

Table 4. Prediction results of PSO-LSSVM and LSSVM with different sets of parameters.

	(X_1, X_4, X_5, X_6)		$(X_1, X_4, X_5, X_6, S_1, S_2, S_3, S_4)$	
	LSSVM	**PSO-LSSVM**	**LSSVM**	**PSO-LSSVM**
January	42.52%	41.67%	38.65%	39.01%
February	40.00%	38.84%	34.82%	34.04%
March	29.88%	30.60%	36.81%	32.17%
April	34.04%	33.06%	31.38%	25.76%
May	42.74%	42.46%	43.06%	40.09%
June	28.44%	29.74%	30.14%	31.58%
July	10.07%	9.44%	9.10%	10.28%
August	13.44%	12.92%	13.93%	12.41%
September	26.73%	26.28%	27.05%	25.50%
October	20.72%	21.49%	22.07%	20.19%
November	24.85%	26.11%	28.18%	28.98%
December	34.28%	31.07%	24.47%	24.90%
Average MAPE	28.86%	28.53%	28.23%	27.08%
Average RMSE	4201	4214	4091	4030

(1) Table 4 reveals that the mean absolute percentage error corresponding to each month is not always better for the models that consider the seasonal factors than those of the models that do not. However, the average MAPE/RMSE scores of the two models are observed to be lower when they incorporate the seasonal factor within themselves. This establishes the fact that the ambient natural season is a factor that affects the accuracy of prediction.

(2) The annual mean absolute percentage error of the PSO-LSSVM model is observed to be better than that of the LSSVM model, which indicates that the PSO algorithm is an effective method to solve the optimization problem for the parameters in the LSSVM algorithm.

The prediction accuracies of PSO-LSSVM also demonstrate that the prediction errors corresponding to January, February, and May are relatively high when seasonal factors are not considered, and that the maximum prediction error is 42.46%. When the ambient natural season is considered, the high mean absolute percentage errors are, in particular, are observed to reduce by nearly 2.5%, even though the maximum prediction error remains high at 40.09%. This may be attributed to the fact that the daily tourist flow varies with the alternating seasons. Obtaining accurate forecasts simply based on the ambient natural seasonal factor is unrealistic. Hence, the pre-treatment of seasonal variation factor is necessary.

Therefore, PSO-LSSVM is verified to be an effective method for the accurate forecasting of daily tourist flow at tourist destinations. Further, the predictions verify that consideration of the ambient natural season reduces the prediction error by nearly 2%. However, given the differences in time and temperature, a simple incorporation of the seasonal factor cannot be expected to satisfactorily enhance the accuracy of forecasting. Hence, the pre-treatment of the seasonal variation factor is necessary.

3.3. Adjustment of Natural Seasons Based on K-Means

During the practical application of the predictive model, the climate changes from cold to warm or from warm to cold with the variation of seasons. In other words, the change of temperature within the same season might alter the trend of daily tourist flow at a destination, whereas the daily flow may be identical during successive months despite a season change between them if the difference in temperature is not palpable to tourists. Therefore, if the forecasting model considers the natural seasons directly, the accuracy of its predictions will be adversely affected. This leads to the necessity of pre-treating the seasonal variation factor.

Corresponding to each season, the daily tourist flow varies with the change of time and temperature. As is evident from the daily tourist data (from the cooperation project with Huangshan Management Committee) during the period from March, 2014 to February 2015 at Mountain Huangshan, the daily tourist volume varied in accordance with the maximum and minimum daily temperatures. Figure 3 illustrates the tourist flow over different seasons.

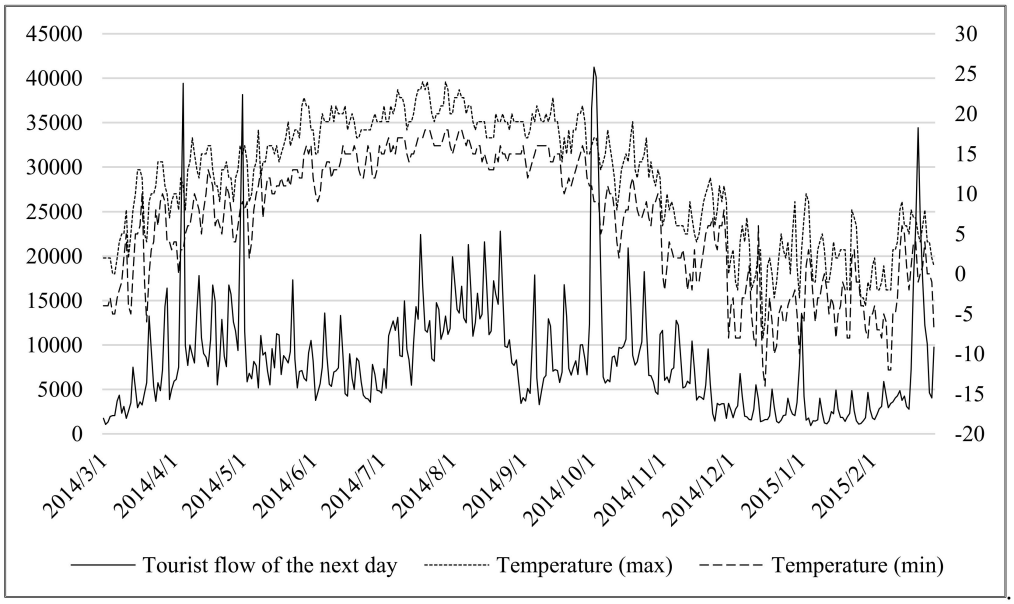

Figure 3. Daily tourist flow, along with the maximum and minimum daily temperatures, during March 2014 to February 2015.

As observed in the figure, the distribution of tourist flow over the four seasons exhibits an almost identical trend to that of daily temperatures, except for the sharp changes on four statutory holidays. Further conclusions can also be drawn from the data. During spring, the temperature in mountainous environments remains relatively low in early March, thereby lessening the daily tourist flow during that time. The data confirms that the daily number of tourists during this period is 2000 on average. With time, the temperature gradually rises as the climate becomes more comfortable. The climate becomes more suitable for travelling; thereby increasing the daily tourist flow at the mountain. Although summer is the hottest period of the year, the temperature at Mountain Huangshan stays consistent at 25 °C. Lu corroborated that Huangshan exhibits monsoon climate between June and August, which is quite conducive to travelling [14]. Moreover, the summer holidays are scheduled between July and August, during which people prefer to travel. Due to these factors, the daily tourist flow remains high during this period. In autumn, the overall temperature in mountainous destinations remains very comfortable during September and October, and the tourist flow remains high. However, the temperature starts to decrease in November, the number of people willing to visit the mountains lessens. Overall, in winter, the daily tourist flow at Mountain Huangshan remains low because of the low temperature. However, the tourist flow may exhibit increasing trends even in winter owing to the temporary rise in temperature, whereas during the majority of the season, the daily tourist flow exhibits the same distribution as the ambient temperature and humidity. Therefore, clustering the seasons at scenic tourist destinations according to the distribution of daily tourist flow is necessary.

Based on the analysis, the daily highest and lowest temperatures, the tourist flow of a particular day, and the time are selected as input variables. The K-means algorithm is adopted to adjust the natural season at the destination of Mountain Huangshan. Taking the data pertaining to 2014 as an example, the clustering results are shown in Figure 4.

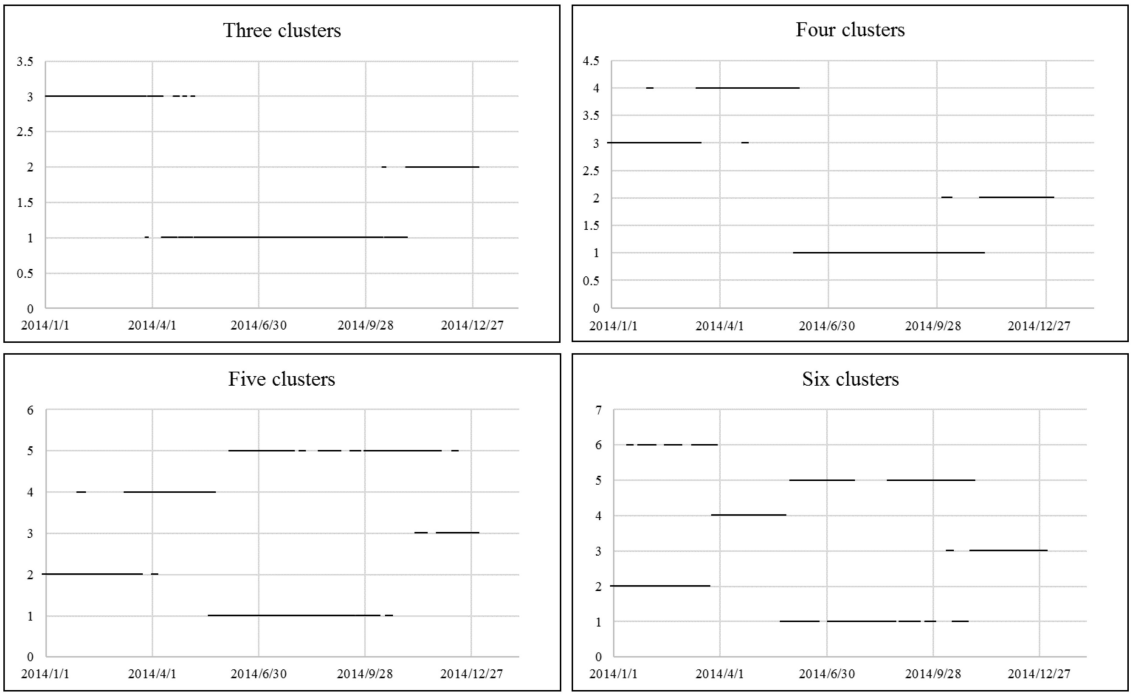

Figure 4. Results of seasonal clustering for the data pertaining to 2014.

In the figure, we use the number 1 to represent cluster 0, number 2 for cluster 1, number 3 for cluster 2 and so on. As is evident, when the year is divided into three seasons, some sample points are clustered into very few clusters. When it is divided into five classes, some objects belong to more than one category. However, when it is divided into six classes, only a few objects belong to each class,

which is insufficient to form a category. Tables 5 and 6 present the specific clustering results for the cases of three and four classes.

Table 5. Clustering results for three classes.

	One	Two	Three
Month	January, February, March	April, May, June, July, August, September, October	November, December

Table 6. Clustering results for four classes.

	One	Two	Three	Four
Month	January, February 1 March–14 March	15 March–31 March, April, May	June, July, August, September, October	November, December

To facilitate the presentation of the clustering results, Figure 5 is designed, from which it can be concluded that when the year is divided into three categories, April, May, June, July, August, September, and October are clustered into a single category. However, during April to October, the temperature initially increases and then decreases, affecting the daily tourist flow accordingly. After repeated trials, the results confirm that a stable state is reached when the year is divided into four seasonal classes. The final result is also presented in Figure 5, in which January, February, and 1–14 March is taken to constitute one class. During this time, the temperature is relatively low, and the daily tourist flow remains almost identical throughout the period. However, in late March, the temperature begins to gradually increase, and the climate becomes more comfortable. Thus, the daily tourist flow at mountainous destinations during this time is similar to that of April and May. Therefore, early March is classified in the same category as January and February, whereas late March is now classified in the same category as April and May. Similarly, in June, July, and August, although the surface temperature is relatively high, the temperature in mountainous spots remains relatively low; and so, they are grouped together with September and October into a single class. Meanwhile, November and December define their own category.

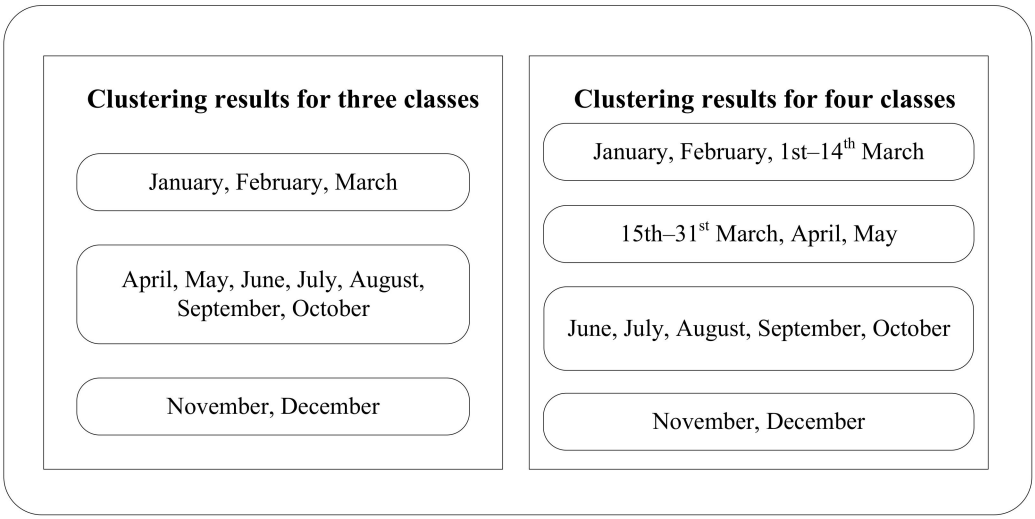

Figure 5. Clustering results of different quantity categories.

3.4. Predictions by Various Models and Their Comparison after Seasonal Clustering

To verify the effectiveness and feasibility of seasonal clustering, we use the vectors $(X_1, X_4, X_5, X_6, S_i)$ as input variables in the models, where $s_{i'} = \begin{cases} 1 \\ 0 \end{cases}$ (1 represents the ith natural season $i = 1, 2, 3, 4$) denotes the new natural seasons. In a separate experiment, we use the vectors representing

the originally defined natural seasons for comparison purposes. As before, both PSO-LSSVM and LSSVM are tested with respect to both sets of vectors. Tables 7 and 8 present the results of the predictions.

Table 7. Results of the predictions by PSO-LSSVM and LSSVM using three and four seasonal classes.

	$(X_1, X_4, X_5, X_6, S_1, S_2, S_3, S_4)$			
	Three		**Four**	
	LSSVM	**PSO-LSSVM**	**LSSVM**	**PSO-LSSVM**
January	39.59%	37.92%	36.31%	35.39%
February	33.07%	31.37%	34.17%	33.38%
March	31.26%	26.50%	30.90%	26.35%
April	35.31%	32.20%	31.47%	26.52%
May	48.07%	42.57%	44.44%	42.32%
June	28.78%	29.24%	30.12%	30.95%
July	8.80%	9.26%	9.15%	8.98%
August	15.20%	13.17%	14.67%	12.98%
September	29.09%	27.33%	29.23%	27.22%
October	21.72%	21.03%	22.11%	20.03%
November	25.01%	22.24%	24.89%	22.51%
December	26.70%	31.93%	27.29%	25.29%
Average MAPE	28.49%	27.00%	27.82%	25.91%
Average RMSE	4051	3977	3967	3798

Table 8. Results of the predictions by PSO-LSSVM and LSSVM under different definitions of seasons.

	Original		**New**	
	LSSVM	**PSO-LSSVM**	**LSSVM**	**PSO-LSSVM**
January	38.65%	39.01%	36.31%	35.39%
February	34.82%	34.04%	34.17%	33.38%
March	36.81%	32.17%	30.90%	26.35%
April	31.38%	25.76%	31.47%	26.52%
May	43.06%	40.09%	44.44%	42.32%
June	30.14%	31.58%	30.12%	30.95%
July	9.10%	10.28%	9.15%	8.98%
August	13.93%	12.41%	14.67%	12.98%
September	27.05%	25.50%	29.23%	27.22%
October	22.07%	20.19%	22.11%	20.03%
November	28.18%	28.98%	24.89%	22.51%
December	24.47%	24.90%	27.29%	25.29%
Average MAPE	28.23%	27.08%	27.82%	25.91%
Average RMSE	4091	4030	3967	3798

Table 7 illustrates that when the year is divided into four seasonal classes, the MAPE/RMSE scores of both models are better corresponding to each month than those when the year is divided into three seasonal classes. Further, the reasoning behind dividing the year into four seasonal categories has already been provided. Moreover, when the year is divided into four seasonal classes, the prediction accuracy of PSO-LSSVM is observed to be better than that of LSSVM, which establishes the feasibility of the proposed model.

Table 8 shows the results of the predictions by PSO-LSSVM and LSSVM under different definitions of seasons.

(1) As is evident from Table 8, although the adoption of seasonal clustering does not reduce the monthly mean absolute percentage error, it does reduce the annual mean absolute percentage error by nearly 1.5%. Additionally, the RMSE indicator also corroborates our conclusion. This establishes that seasonal clustering is effective in enhancing the prediction accuracy.

(2) The annual MAPE/RMSE score of the PSO-LSSVM model is observed to be better than that of LSSVM overall, as can be seen from Table 8, PSO-LSSVM model has a better performance than LSSVM in most of the months, the error was reduced by an average of nearly 1.5%. This corroborates our conclusion that the PSO-LSSVM model is an effective method to forecast daily tourist flow at scenic tourist destinations.

(3) The seasonal clustering that produces the best results classifies January, February, and 1–14 March into one group, November and December into another group, and April and May into yet another group.

By comparing the predictions by PSO-LSSVM, we corroborate that the mean absolute percentage error corresponding to March decreases significantly after the seasonal adjustment. Although the MAPE scores corresponding to April and May are a little higher than those before clustering, the MAPE scores of November, December, and March are lower than those before clustering, and the value of MAPE is observed to decrease throughout the year. Therefore, the method proposed in this research is effective, moreover, the RMSE indicator also corroborates the validity of the proposed method.

4. Discussion

The prediction of daily tourist flow at scenic destinations is essential to the tourism industry, and the accuracy of forecasting is highly significant for the optimal distribution of tourism resources [8,37,43]. Mountain Huangshan is a famous scenic spot in China, and its daily tourist volume is known to exhibit complex nonlinear characteristics and the historical tourist data exhibits various trends of fluctuation during different seasons [44]. This research considers the tourist flow data at Mountain Huangshan between 2014 and 2017 as a dataset and analyzes the variation of daily tourist volumes with respect to different seasons. On the one hand, particle swarm optimization is used to optimize the least squares support vector machine; on the other hand, we focus on rearranging the seasons by clustering algorithm. In response to results in our research, it can be pointed out that the prediction performance can be improved from two aspects: the predictor itself and the input of the algorithm. The experimental results above verify the correctness of our research that the effect of classical forecasting model can be optimized by seasonal adjustment and it has an inspiration and practical value for short-term daily tourist flow forecasting.

In summary, compared with the previous research, the differences and advantages of this research are as follows:

(1) Instead of forecasting tourists flow at monthly or yearly intervals, this research is conducted at a daily time interval, and this improvement can significantly increase the efficiency of prediction.

(2) The prediction performance of the hybrid model in this research is significantly improved via the proposed optimization algorithm, which can be seen from the Section 4.

(3) Seasonal adjustment and division were included into the forecasting model as factors in our research, and it proves to be an effective method to improve the predictive performance of the model. Meanwhile, previous research works rarely considered the question, as mentioned in Section 2.

The results of this research are helpful to tourism management, and the following practical implications can be provided in management:

(1) According to the results of seasonal clustering, managers can always adopt a different hybrid model instead of using the same model. Namely, it can improve the specificity of actual management.

(2) The accurate short-term daily tourist flow forecasting can help reduce the number of crowding incidents to improve the quality of tourists' experience.

(3) In terms of resource allocation management of scenic spots, the accurate tourist flow forecasting method presented in this research can reduce the waste of resources.

In general, this research has an inspiration for tourist flow forecasting. It fills the gap of tourist flow forecasting by introducing the idea of seasonal clustering, which proved to be effective. The results of this research can also provide some practical implications.

5. Conclusions

In this research, the ambient natural season is taken to be an essential factor in the prediction of the daily tourist flow on the subsequent day, and a hybrid optimized model is proposed. The experimental results corroborate that: (1) season is a factor that profoundly affects the accuracy of prediction of the daily tourist flow, which can be supported by evidence from Table 4; (2) seasonal adjustments improve the prediction accuracy effectively by nearly 3%. In particular, it is suitable for months that exhibit significant temperature variations, e.g., March. Evidence from Tables 7 and 8 can support it; (3) the superiority of PSO-LSSVM over LSSVM is also verified and it can be supported by evidence from Tables 4, 7, and 8. This is attributed to the role of the PSO method in the determination of optimal values of LSSVM parameters based on its excellent coherence coordination. Further, the effective adjustment of natural seasons based on the K-means algorithm is another important reason behind the superiority of PSO-LSSVM. Thus, based on the idea of seasonal adjustment, PSO-LSSVM combined with the K-means algorithm was established to be a convenient and feasible method for daily tourist volume forecasting. The experimental results in this research support this conclusion.

However, the proposed method still suffers from certain limitations which could be improved in future works. First, this research was conducted with a focus on the practical utility of the method, and the underlying theory merits further research. Second, certain factors such as weather could be considered in greater complexity than was considered in this research to further improve the prediction accuracy. In addition, the method of seasonal adjustment deserves further research.

In general, this research proves the reliability of improving the prediction effect based on seasonal adjustment, and the accuracy of short-term prediction of the daily tourist flow achieved by the proposed hybrid model is beneficial to professionals in the tourism industry, enabling them to reasonably allocate appropriate resources in advance. This research also contributes to the research on short-term forecasting, which is significant as most existing studies have focused on monthly or annual prediction.

Author Contributions: Conceptualization, Keqing Li and Chu Li; Funding acquisition, Changyong Liang; Investigation, Keqing Li; Methodology, Keqing Li and Chu Li; Project administration, Changyong Liang; Resources, Changyong Liang, Wenxing Lu and Shuping Zhao; Software, Binyou Wang; Supervision, Changyong Liang, Wenxing Lu and Shuping Zhao; Validation, Keqing Li; Visualization, Keqing Li; Writing–original draft, Keqing Li; Writing–review & editing, Keqing Li. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation, grant number 71771075, 71331002.

Acknowledgments: This study is part of a research work of the National Natural Science Foundation (71771075, 71331002).

Conflicts of Interest: The authors declare that there are no competing interests regarding the publication of this paper.

References

1. World Travel & Tourism Council. *Global Economic Impact & Trends 2020*; World Travel & Tourism Council: London, UK, 2020.
2. Yin, J.; Bi, Y.; Zheng, X.M.; Tsaur, R.C. Safety Forecasting and Early Warning of Highly Aggregated Tourist Crowds in China. *IEEE Access* **2019**, *7*, 119026–119040. [CrossRef]
3. China Tourism Academy. China Domestic Tourism Development Report 2020. Available online: http://www.ctaweb.org/html/2020-9/2020-9-14-13-2-83232.html (accessed on 1 October 2020).
4. Lu, W.; Wei, X. Spatio-temporal Distribution Pattern of Cable Car Passenger Flow in Panholidays: A Case Study of Huangshan Scenic Area. In Proceedings of the 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 26–29 June 2017; pp. 35–42.
5. Wang, Q.; Lu, L.; Yang, X.Z. Influencing factors of water resources security in water shortage mountain resorts. *J. Arid Land Resour. Environ.* **2014**, *28*, 48–53.
6. Yang, X.Z.; Wang, X. Tourism crowding characteristics and adjusting patterns of mountain scenic spots during special periods: A case study of Huangshan Mountain. *Geogr. Res.* **2019**, *38*, 961–970.

7. Feng, L. Research on Tourism Public Crisis Countermeasures Based on Big Data. In Proceedings of the 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 24–26 May 2019; pp. 1273–1279.
8. Song, H.; Li, G. Tourism demand modelling and forecasting—A review of recent research. *Tour. Manag.* **2008**, *29*, 203–220. [CrossRef]
9. Lim, C.; McAleer, M. Forecasting tourist arrivals. *Ann. Tour. Res.* **2001**, *28*, S0160–S7383. [CrossRef]
10. Prosper, F.; Bangwayo-Skeete, R.W.S. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tour. Manag.* **2015**, *46*, 454–464. [CrossRef]
11. Denstadli, J.M.; Jacobsen, J.K.S.; Lohmann, M. Tourist perceptions of summer weather in Scandinavia. *Ann. Tour. Res.* **2011**, *38*, 920–940. [CrossRef]
12. Gössling, S.; Scott, D.; Hall, C.M.; Ceron, J.P.; Dubois, G. Consumer behavior and demand response of tourists to climate change. *Ann. Tour. Res.* **2012**, *39*, 36–58. [CrossRef]
13. Pu, W.; Quan-sheng, G. An analysis of annual variation of tourist flows and climate change in Hainan Province. *Geogr. Res.* **2009**, *28*, 1078–1084. [CrossRef]
14. Lu, L.; Xuan, G.; Zhang, J. An approach to seasonality of tourist flows between coastland resorts and mountain resorts: Examples of Sanya, Beihai, Mt. Putuo, Mt. Huangshan and Mt. Jiuhua. *Acta Geogr. Sin.* **2002**, *57*, 731–740. [CrossRef]
15. Zheng, Q.; Chen, R.; Sun, J.S. Study on the Influencing Factors of Tourism Scenic Spots for Traffic—Taking Huangshan Scenic Area as an Example. *J. Bengbu Coll.* **2014**, *3*, 98–102. [CrossRef]
16. Chen, R.; Li, G. A Study on the Forecasting Method of AGA-SVR Modeled Holiday Tourist Flows Based on SEA. *Tour. Sci.* **2016**, *30*, 12–23. [CrossRef]
17. Huang, J.H.; Min, J.C.H. Earthquake devastation and recovery in tourism: The Taiwan case. *Tour. Manag.* **2002**, *23*, 145–154. [CrossRef]
18. Teixeira, J.P.; Fernandes, P.O. Tourism time series forecast with artificial neural networks. *Tékhne* **2014**, *12*, 26–36. [CrossRef]
19. Chen, K.Y.; Wang, C.H. Support vector regression with genetic algorithms in forecasting tourism demand. *Tour. Manag.* **2007**, *28*, 215–226. [CrossRef]
20. Palmer, A.; Montano, J.J.; Sesé, A. Designing an artificial neural network for forecasting tourism time series. *Tour. Manag.* **2006**, *27*, 781–790. [CrossRef]
21. Yan, X.; Chowdhury, N.A. Mid-term electricity market clearing price forecasting: A hybrid LSSVM and ARMAX approach. *Int. J. Electr. Power Energy Syst.* **2013**, *53*, 20–26. [CrossRef]
22. Sun, W.; Zhang, J. Forecasting Day Ahead Spot Electricity Prices Based on GASVM. Proceedings of 2008 International Conference on Internet Computing in Science and Engineering, Harbin, China, 28–29 January 2008; pp. 73–78.
23. Chen, Y.; Yang, Y.; Liu, C.; Li, C.; Li, L. A hybrid application algorithm based on the support vector machine and artificial intelligence: An example of electric load forecasting. *Appl. Math. Model.* **2015**, *39*, 2617–2632. [CrossRef]
24. Yuan, X.; Chen, C.; Yuan, Y.; Huang, Y.; Tan, Q. Short-term wind power prediction based on LSSVM–GSA model. *Energy Convers. Manag.* **2015**, *101*, 393–401. [CrossRef]
25. Zhu, X.; Xu, Q.; Tang, M. Comparison of two optimized machine learning models for predicting displacement of rainfall-induced landslide: A case study in Sichuan Province, China. *Eng. Geol.* **2017**, *218*, 213–222. [CrossRef]
26. Cong, Y.; Wang, J.; Li, X. Traffic Flow Forecasting by a Least Squares Support Vector Machine with a Fruit Fly Optimization Algorithm. *Procedia Eng.* **2016**, *137*. [CrossRef]
27. Eberhart, J.K.R. Particle swarm optimization. Proceedings of 1995 IEEE International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995.
28. Mohana, S.J.; Saroja, M.; Venkatachalam, M. Comparative analysis of swarm intelligence optimization techniques for cloud scheduling. *Int. J. Innov. Sci. Eng. Technol.* **2015**, *1*, 15–19.
29. Zeng, N.; Zhang, H.; Liu, W.; Liang, J.; Alsaadi, F.E. A switching delayed PSO optimized extreme learning machine for short-term load forecasting. *Neurocomputing* **2017**, *240*, 175–182. [CrossRef]
30. Yang, Y.; Pan, B.; Song, H. Predicting hotel demand using destination marketing organization's web traffic data. *J. Travel Res.* **2014**, *53*, 433–447. [CrossRef]

31. Pan, B.; Wu, D.C.; Song, H. Forecasting hotel room demand using search engine data. *J. Hosp. Tour. Technol.* **2012**, *3*, 196–210. [CrossRef]
32. Li, X.; Pan, B.; Law, R.; Huang, X. Forecasting tourism demand with composite search index. *Tour. Manag.* **2017**, *59*, 57–66. [CrossRef]
33. Artola, C.; Pinto, F.; de Pedraza García, P. Can internet searches forecast tourism inflows? *Int. J. Manpow.* **2015**, *36*, 103–116. [CrossRef]
34. Choi, H.; Varian, H. Predicting the Present with Google Trends. *Econ. Rec.* **2012**, *88*, 2–9. [CrossRef]
35. Shen, S.; Li, G.; Song, H. An Assessment of Combining Tourism Demand Forecasts over Different Time Horizons. *J. Travel Res.* **2008**, *47*, 197–207. [CrossRef]
36. Höpken, W.; Ernesti, D.; Fuchs, M.; Kronenberg, K.; Lexhagen, M. Big data as input for predicting tourist arrivals. *Ann. Tour. Res.* **2017**, *28*, 1070–1072. [CrossRef]
37. Colladon, A.F.; Guardabascio, B.; Innarella, R. Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decis. Support Syst.* **2019**, *123*, 113075. [CrossRef]
38. Oktar, Y.; Turkan, M. A Review of Sparsity-based Clustering Methods. *Signal Process.* **2018**, *148*, 20–30. [CrossRef]
39. Gordon, A.D. A Review of Hierarchical Classification. *J. R. Stat. Soc. Ser. A* **1987**, *150*, 119–137. [CrossRef]
40. Davé, A.B.R.N. Robust clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 29–59. [CrossRef]
41. Peña, M. Robust clustering methodology for multi-frequency acoustic data: A review of standardization, initialization and cluster geometry. *Fish. Res.* **2018**, *200*, 49–60. [CrossRef]
42. Lin, L. A Study on the seasonal changes in the tourism in mountain resorts—A case study of the Huangshan Mountain. *Geogr. Res.* **1994**, *4*, 50–58. [CrossRef]
43. Shao-Jiang, L.; Jia-Ying, C.; Zhi-Xue, L. A EMD-BP integrated model to forecast tourist number and applied to Jiuzhaigou. *J. Intell. Fuzzy Syst.* **2018**, *34*, 1045–1052. [CrossRef]
44. Li, K.; Lu, W.; Liang, C.; Wang, B. Intelligence in Tourism Management: A Hybrid FOA-BP Method on Daily Tourism Demand Forecasting with Web Search Data. *Mathematics* **2019**, *7*, 531. [CrossRef]