Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **3**, August **2025**

Long-Term Tourist Flow Prediction of Scenic Spots Based on Fuzzy Support Vector Machine

Dong-Yan Sun[1,*], Cheng-Ping Wang[2,3], Pei-Yu Wu[4]

[1]Chengdu Polytechnic, Chengdu 610041, P. R. China
wwppy123@163.com

[2]Chinese Language and Literature College, Southwest Minzu University, Chengdu 610041, P. R. China
[3]Minzu Language's Information Processing Lab, Chengdu 610041, P. R. China
wangchengping@126.com

[4]University of Malaya, Kuala Lumpur 50603, Malaysia
yb8859@163.com

*Corresponding author: Dong-Yan Sun

Received August 8, 2024, revised January 24, 2025, accepted May 6, 2025.

Abstract. *Traditional methods for predicting tourist flow in scenic spots are often difficult to effectively deal with long-term trends and complex nonlinear relationships, limiting their accuracy and reliability in practical applications. With the rapid development of the tourism industry and the arrival of the big data era, there is an increasing demand for being able to handle multidimensional data and make long-term accurate predictions. In order to solve these two problems, this paper proposes a long-term tourist flow prediction method for scenic spots based on fuzzy support vector machine. Firstly, the initial input indicators are dimensionally reduced by introducing sparse principal component analysis, which effectively reduces the model complexity and retains the most influential features. Second, a new periodic kernel function is designed, which can effectively capture the seasonal and periodic patterns in the data and improve the accuracy of long-term prediction. In addition, an improved design method for the affiliation function is proposed to optimise the sample weight allocation, which enhances the robustness and generalisation ability of the model. The experimental results show that, compared with the traditional fuzzy support vector machine model, the method proposed in this paper exhibits significant improvements in long-term tourist flow prediction, with the Mean Absolute Percentage Error (MAPE) reduced by 13.6% and the Root Mean Square Error (RMSE) reduced by 7.64%.*
Keywords: tourist flow prediction; fuzzy support vector machine; sparse principal component analysis; periodic kernel function; affiliation function

1. **Introduction.** Visitor flow forecasting is of strategic importance for scenic spot management and tourism development. Accurate long-term visitor flow forecasts can help scenic spot managers formulate reasonable development plans, optimise resource allocation, improve service quality, and provide a basis for the government to formulate relevant policies [1, 2, 3]. Especially in the current context of rapid development of tourism, scenic spots are faced with increasing demand for tourists and complex and changing external environment [4]. Long-term tourist flow prediction can not only help scenic spots to cope with possible peaks of passenger flow in advance, but also provide important references for infrastructure construction, environmental protection and sustainable development. Therefore, the development of accurate and reliable long-term tourist flow prediction

methods is crucial for improving the management level of scenic spots and promoting the healthy development of tourism [5].

As an intelligent algorithm combining the advantages of fuzzy logic and support vector machine, Fuzzy Support Vector Machine (FSVM) [6, 7] shows great potential for application in the field of tourist flow prediction. Compared with traditional prediction methods, FSVM can better deal with uncertainty and ambiguity in data, which is exactly in line with the data noise and uncertainty often encountered in tourist flow prediction. By introducing the affiliation function [8], FSVM can assign different importance to different samples, which improves the sensitivity of the model to critical data. In addition, FSVM has strong generalisation ability and robustness, and can effectively cope with seasonal fluctuations and the impact of unexpected events, which are common in the tourism industry. Therefore, applying FSVM to long-term tourist flow prediction in scenic spots is expected to significantly improve the accuracy and reliability of the prediction, and provide more powerful support for the management decision of scenic spots.

1.1. **Related work.** Visitor flow forecasting research has been widely concerned by academics and the industry, and the traditional methods mainly include time series analysis, regression analysis and neural network, etc. Cho [9] applied the ARIMA model to forecast the monthly visitor flow of a scenic spot, and the results showed that the method has high accuracy in short- term forecasting, but the grasp of the long-term trend is more deficient. Yu et al. [10] proposed a hybrid model combining grey prediction and linear regression for predicting annual tourist flow. Although this study improves the stability of the prediction, it fails to fully consider the multidimensional factors affecting the flow of tourists. In terms of neural network, Chen et al. [11] constructed a daily tourist flow prediction model based on BP neural network, which improved the prediction accuracy by introducing meteorological factors, but the interpretability of the model was poor. Zhang et al. [12] proposed a prediction method integrating genetic algorithm and wavelet neural network, which is excellent in dealing with nonlinear and non- stationary time series, but the computational complexity is high and not suitable for large-scale data processing. Overall, the traditional method performs better in short-term prediction, but still has deficiencies in long-term prediction, multifactor comprehensive analysis and model interpretability, which is difficult to meet the needs of modern scenic spot management.

In recent years, the research on long-term tourist flow prediction mainly focuses on the application of machine learning algorithms and the fusion of multi-source data in two directions. Support vector machine (SVM), as a powerful machine learning tool, has been widely used in this field. Shabri [13] proposed a long-term tourist flow prediction model based on LSSVM and particle swarm optimisation, which improved the prediction accuracy by optimising the kernel function parameters, but the ability to handle outliers still needs to be improved. Paolanti et al. [14] combined deep learning with SVM to construct a hierarchical prediction framework, which made significant progress in the treatment of seasonal fluctuations, but the complexity of the model limited its promotion in practical applications. In terms of multi-source data fusion, Xiong et al. [15] proposed a long-term prediction model that fuses web search data and social media sentiment analysis, which improves the timeliness of prediction, but the effect of data noise is still significant. Despite the progress made in these studies, the following problems still exist: (1) the effectiveness of feature selection is insufficient, which leads to too high dimensionality of the model inputs; (2) insufficient consideration of periodic patterns in the data; and (3) irrational allocation of sample weights, which affects the model's generalisation ability. To address these issues, the introduction of Sparse Principal Component Analysis (sPCA) can effectively reduce the feature dimensions, the design of periodic kernel function helps

to capture the periodic features in the data, while the reasonable design of the affiliation function can optimise the sample weight allocation. The combination of these methods is expected to significantly improve the accuracy and reliability of long-term tourist flow prediction.

1.2. **Motivation and contribution.** Existing methods for forecasting long-term tourist flows in scenic spots still have limitations in dealing with complex non-linear relationships and uncertainties. Traditional time series analysis and regression methods are difficult to effectively capture long-term trends and seasonal fluctuations in tourism data. And although machine learning methods perform well in dealing with nonlinear relationships, they often suffer from poor prediction accuracy due to inappropriate feature selection and irrational allocation of sample weights. In addition, existing prediction models usually ignore cyclical patterns in the data, which is particularly important in long-term prediction. In order to solve the above problems, this paper proposes a long-term tourist flow prediction method for scenic spots based on FSVM, which combines sPCA and periodic kernel function to significantly improve the accuracy and reliability of prediction.

The main innovations and contributions of this work include:
(1) To address the feature selection problem, this paper introduces the sPCA method to downscale the initial input indicators. This method not only effectively reduces the complexity of the model, but also retains the most influential features for prediction, thus improving the prediction accuracy and computational efficiency of the model.
(2) In order to better capture the cyclical patterns in tourism data, a new cyclical kernel function is designed in this paper. This kernel function can effectively identify and utilise seasonal and cyclical features in the data, significantly improving the performance of the model in long-term forecasting.
(3) For the problem of sample weight allocation, this paper proposes an improved design method for the affiliation function. By reasonably allocating the importance of different samples, the method effectively improves the sensitivity of the model to the key data, and at the same time enhances the robustness and generalisation ability of the model.

2. Relevant technologies.

2.1. **Fuzzy support vector machine.** FSVM [16, 17] is a method that combines the advantages of fuzzy logic and SVM for classification and regression prediction. In contrast to standard support vector machines, FSVM allows sample points to have varying degrees of affiliation around the classification boundary, which can be used to reflect the importance of the sample or how "fuzzy" it is to the classification boundary. In the task of long-term tourist flow prediction, FSVM can effectively deal with the noise and uncertainty in the data, and improve the robustness and prediction accuracy of the model.

FSVM adjusts the contribution of sample points to the classification boundary by introducing an affiliation degree. Each sample point is assigned an affiliation value μ_i, which reflects the degree of "fuzziness" of the sample point to the classification boundary. The larger the affiliation value, the higher the contribution of the sample point to the classification boundary. The objective of FSVM is to minimise the classification interval while maximising the sum of the affiliations on both sides of the classification boundary [18], and its optimisation problem can be expressed as follows:

$$\min_{w,b,\xi} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \mu_i \xi_i \right\} \tag{1}$$

$$\text{s.t. } y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \ \xi_i \geq 0$$

where w is the normal vector of the classification hyperplane, b is the bias term, ξ_i is the non-negative slack variable, C is the penalty coefficient, μ_i is the affiliation value of the i-th sample, n is the number of samples, y_i is the category label of the i-th sample (either $+1$ or -1), x_i is the feature vector of the i-th sample.

The affiliation value μ_i can be determined based on the distance of the sample points from the classification boundary [19]. A common approach is to use a Gaussian function as the affiliation function:

$$\mu_i = \exp\left(-\frac{\|x_i - x_c\|^2}{2\sigma^2}\right) \tag{2}$$

where x_c is the centroid of the classification boundary and σ is the standard deviation of the Gaussian function. This way of definition ensures that sample points close to the classification boundary have higher affiliation values, while sample points far from the classification boundary have lower affiliation values [20, 21].

In solving the above optimisation problem, the Lagrange multiplier method is usually used to convert to a dyadic problem to simplify the computation. The optimisation problem in dyadic form is:

$$\max_{\alpha_i}\left\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j\right\} \tag{3}$$

$$\text{s.t. } 0 \le \alpha_i \le C\mu_i, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

where α_i is the Lagrange multiplier [22].

The final decision function is:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n}\alpha_i y_i x_i^\top x + b\right) \tag{4}$$

With the above equation, we can see how FSVM can optimise the standard Support Vector Machine model by introducing affiliation values and deal with data containing noise.

2.2. **Sparse principal component analysis.** sPCA [23, 24] is an improved principal component analysis method that reduces the number of features extracted in principal component analysis by introducing sparsity, thus simplifying the model and improving the interpretability. sPCA is particularly suitable for dealing with datasets containing a large number of variables, and in tourist traffic prediction, it helps us to filter out the features from the raw data that have the most influence on the prediction.

Traditional Principal Component Analysis (PCA) projects the raw data by linear transformation into a new coordinate system where the individual axes represent different directions of the data. These new axes are called principal components, and they are ordered by how much variance they explain. However, the principal components obtained by PCA are usually linear combinations of all the original variables, which can lead to problems of interpretability and computational complexity.

sPCA solves this problem by introducing sparsity constraints so that each principal component consists of only a small fraction of the original variables. This sparsity helps to improve the interpretability of the model and reduces the number of features, thus simplifying the model. sPCA's objective function can be expressed as:

$$\max_{a} a^\top S a \qquad \text{s.t.} \quad a^\top a = 1,\ \|a\|_0 \le k \tag{5}$$

where S is the sample covariance matrix, a is the unit-length principal component vector, $\|a\|_0$ denotes the number of non-zero elements in the vector a, and k is the pre-set sparsity limit, i.e., the maximum number of non-zero elements in each principal component [25].

In order to solve the above optimisation problem, sPCA usually employs the alternating direction method, LASSO regression or other optimisation algorithms to find sparse principal component vectors. One of the common strategies is to transform the original optimisation problem into the following form:

$$\max_a \left(a^\top Sa - \lambda\|a\|_1\right) \qquad \text{s.t.} \quad a^\top a = 1 \tag{6}$$

where λ is the regularisation parameter that controls the degree of sparsity. By adjusting the value of λ, the degree of sparsity of the resulting principal component vector can be controlled [26].

3. **Data preprocessing based on sPCA.**

3.1. **Introduction to Initial Input Indicators.**

In order to accurately predict the long-term visitor flow of scenic spots, we need to consider a series of factors that may affect the visitor flow. These factors usually include economic indicators, social indicators and demographic indicators. In this section, we will list these initial input indicators in detail and discuss their impacts on visitor flows.

Economic indicators reflect the economic status and level of development of the region and are essential for predicting tourist flows. In our study, the following economic indicators were considered:
(1) Gross Domestic Product (GDP): the overall scale of economic activity in the region, which has a significant impact on tourist flows.
(2) Gross retail sales of consumer goods: reflects the level of regional spending power and is closely related to tourist flows.
(3) Gross industrial output value: reflects the industrial strength of the region, which indirectly affects the attractiveness of tourism.
(4) The amount of investment in fixed assets of the whole society: it reflects the investment activity of the region and has a direct impact on the development of tourism.
(5) Value of agricultural and sideline products: some areas rely on agricultural products to attract tourists, and this indicator has a significant impact on the flow of tourists to a given area.

Social indicators reflect the social environment and cultural climate, which also have an impact on tourists' travelling decisions. The social indicators considered in this study include:
(1) Urban road area: reflects the infrastructure development of the area, and good transport conditions are conducive to the promotion of tourism.
(2) Urban greening coverage: a good ecological environment can attract more tourists, especially for destinations with a predominantly natural landscape.

Demographic indicators reflect the size and structure of a region's population and are critical to understanding visitor sources and tourism demand. The demographic indicators considered in this study include:
(1) Population size: Areas with larger populations usually have more potential tourists.
(2) Disposable income per capita: reflects the spending power of the population, with higher disposable income implying a stronger willingness to spend on tourism.

We define symbols for the above indicators as shown in Table 1.

Table 1. Definition of symbols for initial input indicators

Notation	Definition
GDP_t	GDP in period t
SCRT_t	Total retail sales of consumer goods in period t
IT_t	Gross industrial output in period t
IF_t	The amount of social investment in fixed assets in period t
AF_t	Value of agricultural output in period t
RA_t	Urban road area in period t
GC_t	Urban green coverage in period t
PN_t	The population size in period t
DI_t	Disposable income per capita in period t

With the introduction of the initial input metrics above, we have identified the key factors that influence visitor traffic. In the next section, we will use sPCA to filter out the most influential features for prediction.

3.2. **Introduction of Initial Input Indicators.** In order to effectively solve the problem of selecting the independent variables of the model, this study applies sPCA in multivariate statistical analysis to comprehensively analyse the nine initial indicators affecting the flow of tourists in scenic spots in order to establish a comprehensive evaluation expression, so as to identify the few major indicators that play a decisive role. This method not only greatly simplifies the data structure, but also effectively improves the analysis effect.

First, we standardised the raw data to eliminate inconsistencies in scale and differences in order of magnitude between indicators. The method of standardisation is as follows:

$$x_{ij}^* = \frac{x_{ij}}{X_j}, \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, 9) \tag{7}$$

$$\bar{X}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad (j = 1, 2, \ldots, 9) \tag{8}$$

We then compute the correlation coefficient matrix $R = (r_{ij})_{9\times 9}$, where r_{ij} $(i, j = 1, 2, \ldots, 9)$ is the correlation coefficient between the standardised variables x_i and x_j. The correlation coefficient matrix of the initial indicators is shown in Figure 1, and we can see that there is a strong correlation between the initial indicators [27], which provides a basis for dimensionality reduction using sPCA.

Next, we apply the sPCA algorithm and set appropriate sparse parameters to solve the sparse eigenvectors. By analysing the eigenvalues and eigenvectors, we calculate the contribution rate and cumulative contribution rate of each principal component, and select the principal component whose cumulative contribution rate reaches 85%, as shown in Table 2.

We can observe that the cumulative variance contribution of the first three principal components (PC1, PC2, and PC3) reaches 88.6%, which exceeds the 85% threshold we set, so we choose these three principal components. In PC1, GDP, SCRT and DI have the highest loading coefficients of 0.42, 0.41 and 0.39 respectively. These three indicators also maintain relatively high loading coefficients in PC2 and PC3.

Finally, we construct the comprehensive evaluation function as follows:

$$F = \frac{\lambda_1}{\sum_{i=1}^{q} \lambda_i} F_1 + \frac{\lambda_2}{\sum_{i=1}^{q} \lambda_i} F_2 + \ldots + \frac{\lambda_q}{\sum_{i=1}^{q} \lambda_i} F_q \tag{9}$$

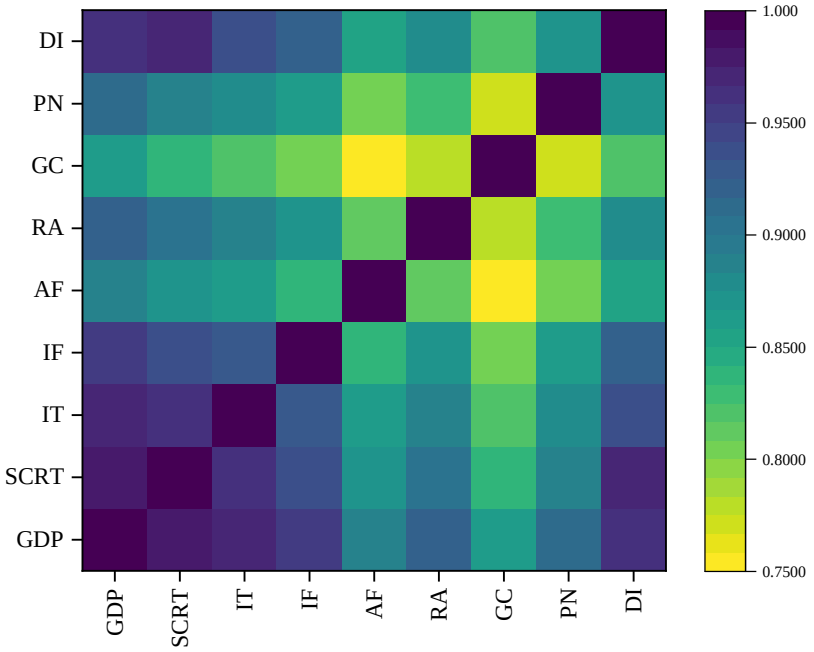

Figure 1. Matrix of correlation coefficients for initial indicators

Table 2. Principal Component Loadings and Variance Contributions

	PC1	PC2	PC3
GDP	0.42	0.15	0.03
SCRT	0.41	0.18	0.05
IT	0.38	0.12	0.07
IF	0.35	0.10	0.09
AF	0.25	0.08	0.12
RA	0.30	0.11	0.08
GC	0.22	0.09	0.11
PN	0.28	0.13	0.06
DI	0.39	0.20	0.04
Variance contribution (%)	68.5	12.3	7.8
Cumulative variance (%)	68.5	80.8	88.6

where λ_i is the eigenvalue corresponding to the selected q principal components and F_i is the corresponding principal component.

According to the size of the loading coefficient of the comprehensive evaluation function, we finally identified the following three main indicators as the input variables of the prediction model: GDP, SCRT and DI. These three indicators show the largest variance contribution and the highest loading coefficients in the results of the sPCA, and they reflect the level of the economic development of the region and the consumption capacity of the residents from different angles. Compared with the original nine indicators, these three indicators can capture the key factors affecting tourist flow more concisely and effectively, which is both statistically significant and consistent with the theory of tourism economics.

Through the application of the sPCA method, we effectively reduce the dimensionality of the data while retaining the interpretability of the original variables. This not only simplifies the process of constructing subsequent models, but also helps to improve the

efficiency and accuracy of the prediction model. In the next sections, we will use these three main indicators as input variables, combined with the improved FSVM algorithm, to construct a long-term tourist flow prediction model for scenic spots.

4. Tourist Flow Prediction Based on Improved FSVM.

4.1. Design of the Affiliation Function.

In the traditional support vector machine (SVM) model, all sample points are considered to have the same importance. However, in a real scenic spot visitor flow prediction problem, different sample points may have different importance levels. To solve this problem, we introduce FSVM to reflect the importance of each sample point during the training process by assigning an affiliation value to it.

Based on the three main input indicators identified in Section 3.2 (GDP, SCRT and DI), we designed a comprehensive affiliation function. The function takes into account the impact of these economic indicators on tourist flows and the position of the sample points in the time series.

Define the training sample set as $\{(x_i, y_i) \mid i = 1, 2, ..., n\}$, where x_i is the input vector and y_i is the corresponding tourist flow. We define the affiliation value $s_i \in (0, 1]$ for each sample point i and compute it as follows:

$$s_i = w_1 f_1(x_i) + w_2 f_2(t_i) + w_3 f_3(y_i) \tag{10}$$

where $f_1(x_i)$ is a function based on economic indicators, $f_2(t_i)$ is a function based on time, $f_3(y_i)$ is a function based on visitor flows, w_1, w_2, w_3 are the weighting coefficients, and $\sum_{j=1}^{3} w_j = 1$.

(1) Economic indicator function:

$$f_1(x_i) = \frac{1}{1 + \exp(-(\alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3}))} \tag{11}$$

where x_{i1}, x_{i2}, x_{i3} correspond to normalised GDP, SCRT and DI respectively, and $\alpha_1, \alpha_2, \alpha_3$ are the corresponding weights.

(2) Time function:

$$f_2(t_i) = 1 - \frac{t_{\max} - t_i}{t_{\max} - t_{\min}} \tag{12}$$

where t_i is the timestamp of the sample point; $t_{\max}$ and $t_{\min}$ are the maximum and minimum timestamps in the sample set, respectively. This function gives higher importance to recent data.

(3) Visitor flow function:

$$f_3(y_i) = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \tag{13}$$

where $y_{\max}$ and $y_{\min}$ are the maximum and minimum visitor flows in the sample set, respectively.

With this design, our affiliation function combines economic factors, time factors and historical tourist flows, allowing the model to better capture the importance of different sample points. Samples with higher economic indicators, more recent time points and larger tourist flows will receive higher affiliation values and play a greater role in model training.

4.2. **Noise Treatment.** Noise is a non-negligible problem in the long-term visitor flow prediction of scenic spots. Noise may come from multiple sources, such as data collection errors, the impact of unexpected events, or seasonal fluctuations. In order to improve the robustness and prediction accuracy of our model, we adopt the following noise processing strategy:

(1) **Wavelet transform denoising:**

We first apply wavelet transform to denoise the original time series data. The advantage of wavelet transform is that it is able to analyse the signal in both time and frequency domains simultaneously, which is suitable for dealing with non-stationary time series.

Let the original tourist flow time series be $\{y_t\}$, and we use the Discrete Wavelet Transform (DWT) to decompose it:

$$y_t = \sum_{j=1}^{J} D_j(t) + A_J(t) \tag{14}$$

where $D_j(t)$ is the detail coefficient of the jth layer and $A_J(t)$ is the approximation coefficient of the Jth layer.

We use the soft thresholding method for the detail coefficients:

$$\hat{D}_j(t) = \operatorname{sign}(D_j(t)) \cdot \max\left(|D_j(t)| - \lambda_j, 0\right) \tag{15}$$

where λ_j is the threshold for layer j, which can be determined by the VisuShrink method:

$$\lambda_j = \sigma_j \sqrt{2 \log N} \tag{16}$$

where σ_j is the standard deviation estimate of the jth layer of noise and N is the time series length.

After thresholding, we reconstruct to get the denoised time series $\{\hat{y}_t\}$.

(2) **Outlier Detection and Handling:**

To further improve the data quality, we use a local outlier factor (LOF)-based approach [28] to detect and handle outliers. The core idea of the LOF algorithm is to identify outliers by comparing the local densities of a sample point with other points in its neighbourhood. For each sample point x_i, we calculate its LOF value:

$$\mathrm{LOF}_k(x_i) = \frac{1}{|N_k(x_i)|} \sum_{x_j \in N_k(x_i)} \frac{\mathrm{lrd}_k(x_j)}{\mathrm{lrd}_k(x_i)} \tag{17}$$

where $N_k(x_i)$ is the set of k-nearest neighbours of x_i and $\mathrm{lrd}_k(x_i)$ is the local reachability density of x_i.

If $\mathrm{LOF}_k(x_i)$ is significantly greater than 1, we mark x_i as a potential outlier. For the detected anomalies, we use the local linear interpolation method to correct them:

$$\hat{y}_i = \frac{y_{i-1} + y_{i+1}}{2} \tag{18}$$

(3) **Integrated noise processing with FSVM:**

We further optimise the FSVM model by combining the above noise processing method with the affiliation function designed in Section 4.1. Specifically, we adjust the affiliation function so that it takes into account the results of the noise processing:

$$s_i' = s_i \cdot (1 - \alpha |\,\mathrm{LOF}_k(x_i) - 1|) \tag{19}$$

where s_i is the original affiliation value and α is a moderating parameter to control how much the LOF affects the affiliation.

In this way, we not only remove the high-frequency noise in the time series, but also identify and deal with the outliers, and at the same time introduce the sensitivity to noise

in the FSVM model. This comprehensive treatment can effectively improve the robustness of the model to noise, thus enhancing the accuracy of long-term tourist flow forecasts.

4.3. **Cyclical Kernel Function.** Tourist flows at scenic spots usually show obvious periodic features, which may include weekly, monthly or yearly cyclical patterns. To better capture these periodic features, we propose an improved periodic kernel function and integrate it into our FSVM model.

The periodic kernel function we designed is an improved version of the Radial Basis Function (RBF) based kernel, which combines several periodic components. It is defined as follows:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \cdot \prod_{k=1}^{M} \exp\left(-\frac{2\sin^2\left(\pi(t_i - t_j)/P_k\right)}{l_k^2}\right) \tag{20}$$

where x_i, x_j are the input vectors, t_i, t_j are the corresponding timestamps, σ is the bandwidth parameter of the RBF kernel, M is the number of cycles under consideration, P_k is the length of the kth cycle, and l_k is the length scale parameter of the kth cycle component.

This kernel function is designed with the following considerations in mind: (a) the RBF kernel part captures the overall similarity of the input features; (b) the periodicity part captures the multi-scale periodic patterns in the time series; and (c) by adjusting P_k and l_k, we have the flexibility to model different periodic features.

In order to determine the appropriate periodicity parameter P_k, the preprocessed tourist flow time series were Fourier transformed to obtain the spectrogram. The main periodic components are identified by analysing the peaks of the spectrogram. Typically, we consider annual cycles ($P_1 = 365$ days), seasonal cycles ($P_2 = 91$ days) and weekly cycles ($P_3 = 7$ days). The l_k parameters are fine-tuned for optimal performance using methods such as grid search or Bayesian optimisation.

Ultimately, we integrate the designed periodic kernel function into the objective function of the FSVM. The optimisation problem of the FSVM can be expressed as:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} s_i'\xi_i \tag{21}$$

subject to

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, ..., n \tag{22}$$

where $\phi(\cdot)$ is the feature mapping implicitly defined by our periodic kernel function and s_i' is the modified affiliation defined above.

By solving this optimisation problem, we obtain the decision function as:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \tag{23}$$

where α_i is the Lagrange multiplier.

In this way, our model is able to adaptively capture complex cyclical patterns in visitor flow data while accounting for the effects of noise and outliers, thereby improving the accuracy of long-term forecasts.

5. **Experimental Results and Analyses.**

5.1. **Experimental Setup.** In order to verify the effectiveness of our proposed long-term tourist flow prediction model for scenic spots based on sPCA and improved FSVM, we conducted a series of experimental simulations. We selected the daily visitor flow data of a well-known scenic spot from 1 January 2010 to 31 December 2023 as the experimental dataset. At the same time, we collected data on economic indicators such as GDP, total retail sales of consumer goods and per capita disposable income for the corresponding time period. The dataset was divided as follows:

Training set: 1 January 2010 to 31 December 2022.

Test set: 1 January 2023 to 31 December 2023.

In order to fully assess the performance of the model, we use the following evaluation metrics:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{24}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{25}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{26}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of the actual values, and n is the sample size.

To prove the superiority of our proposed model, we chose ARIMA, traditional SVM, Long Short-Term Memory Network (LSTM) and traditional FSVM for comparison. Based on the previous theoretical analyses and preliminary experimental results, the parameter settings of the proposed model are shown in Table 3 (these parameters are optimised on the validation set by grid search and 5-fold cross-validation).

Table 3. Parameterisation of the proposed model

Function	**Descriptions / Value**
sPCA	Number of principal components: 3
Affiliation function	Weighting factors: $w_1 = 0.4$, $w_2 = 0.3$, $w_3 = 0.3$
	Economic indicator weights: $\alpha_1 = 0.4$, $\alpha_2 = 0.3$, $\alpha_3 = 0.3$
Noise treatment	Wavelet: db4, decomposition layers: 4; LOF algorithm: $k = 5$;
	LOF Impact Parameter: $\alpha = 0.2$
Periodic kernel	RBF kernel bandwidth: $\sigma = 0.1$
	Number of cycles: $M = 3$; Cycle lengths: $P_1 = 365$, $P_2 = 91$, $P_3 = 7$
	Length scale parameters: $l_1 = 30$, $l_2 = 10$, $l_3 = 2$
FSVM	Penalty parameter: $C = 100$

All experiments were performed on a workstation configured with an Intel Core i7-10700K CPU, 32GB RAM, and an NVIDIA GeForce RTX 3080 GPU. The algorithms were implemented using Python 3.8, and the main dependent libraries include numpy, pandas, scikit-learn, pywavelets, and a custom FSVM implementation.

5.2. **Fitting Training.** In this section, we analyse in detail the model fitting performance on the training set. We compare the performance of our proposed Improved FSVM model with other comparison models on the training set. All models were trained using data from 1 January 2010 to 31 December 2021. Parameter settings were as described above and optimised by grid search and cross-validation for all models. Table 4 summarises the fitting performance of each model on the training set.

Table 4. Training results for fitting different models

Model	RMSE	MAPE (%)	R^2
ARIMA	1245.32	8.76	0.8654
Traditional SVM	1103.45	7.89	0.8912
LSTM	987.21	6.95	0.9134
Traditional FSVM	956.78	6.73	0.9201
Improved FSVM	823.56	5.87	0.9435

It can be seen that our proposed Improved FSVM model outperforms the other comparative models in all evaluation metrics. In terms of RMSE, our model achieves 823.56, which is about 14% lower than the second best conventional FSVM model. The MAPE of our model is 5.87%, which is significantly lower compared to other models. The R^2 of our model reaches 0.9435, which is close to 1, indicating that the model is able to explain most of the variation in the original data.

It is worth noting that although LSTM models usually perform well in complex time series forecasting tasks, our improved FSVM model still achieves better performance in this experiment. This may be due to the fact that our model better combines domain knowledge (e.g., the impact of economic indicators) and data characteristics (e.g., periodicity), rather than just relying on raw time series data.

5.3. **Forecasting Results.** In this section, we present a detailed analysis of the prediction performance of each model on the test set (1 January 2023 to 31 December 2023). This analysis aims to assess the models' generalisation ability and long-term prediction accuracy, which is crucial for scenic management and decision making. The prediction performance of each model on the test set is shown in Figure 2.

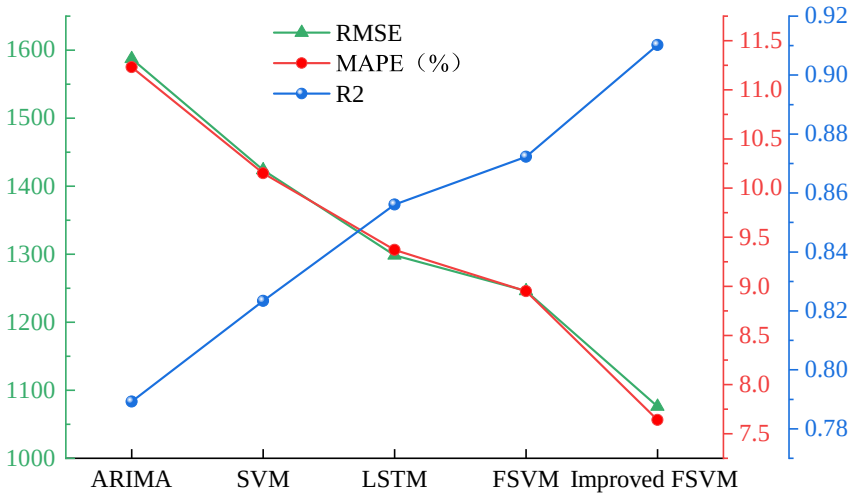

Figure 2. Comparison of predicted performance of different models on the test set

Overall, our proposed Improved FSVM outperforms other comparative models in all evaluation metrics, demonstrating the best predictive performance. Our model achieves an RMSE of 1076.32, which is about 13.6% lower than the second best conventional FSVM model. The MAPE of our model is 7.64%, and the R^2 of our model reaches 0.9102, significantly higher than other models.

The ARIMA model performs the worst, probably due to its difficulty in capturing complex nonlinear relationships and long-term dependencies. The traditional SVM improves

relative to ARIMA, but still does not handle the time series properties well. The LSTM shows better performance, thanks to its ability to capture long-term dependencies. The traditional FSVM is slightly better than the LSTM, indicating that the introduction of fuzzy membership does improve the robustness of the model. Our improved FSVM model significantly outperforms all other models, demonstrating the effectiveness of our proposed improvement strategy.

Possible reasons why our model performs well on the test set include:

- sPCA effectively selects the most relevant features and reduces the risk of overfitting.
- The improved design of the affiliation function takes into account a number of factors, allowing the model to capture the importance of the samples more accurately and improving the generalisation ability of the model.
- The noise handling strategy effectively removes noise and outliers from the data, improving the stability of the prediction.
- The periodic kernel function successfully captures the multi-scale periodicity of tourist flows, which helps in long-term forecasting.

It is worth noting that all models perform slightly worse on the test set than on the training set, which is a normal phenomenon reflecting the models' ability to generalise on unseen data. However, our model has the smallest performance drop on the test set, further demonstrating its superior generalisation ability.

6. **Conclusion.** In this paper, a long-term tourist flow prediction method for scenic spots based on Improved FSVM is proposed, which effectively solves the limitations of traditional prediction models in dealing with long-term trends and complex nonlinear relationships. By introducing sPCA, the model is able to extract the key features in the data more accurately, which significantly reduces the model complexity. In addition, the newly designed periodic kernel function enhances the ability to capture seasonal and periodic patterns in the data, ensuring the accuracy and reliability of the long-term prediction results. The following conclusions can be drawn from the experiments conducted on the actual scenic dataset:

- sPCA is able to effectively reduce the dimensionality of input features while retaining the most influential features, which improves the prediction accuracy and computational efficiency of the model.
- The newly designed periodic kernel function performs better in capturing seasonal and periodic patterns in long-term data compared to the traditional kernel function.
- The improved affiliation function design method optimises the sample weight allocation and significantly enhances the robustness and generalisation of the model.
- The FSVM strategy combining sPCA, periodic kernel function and improved affiliation function strikes the best balance between accuracy and reliability of long-term visitor flow prediction and is the optimal method recommended in this paper.

The experimental data in this paper mainly comes from a single scenic spot, and the limitation of the dataset may have affected the generalisation ability of the model. Future work should consider introducing more datasets from different types of scenic spots and different geographic locations to verify the effectiveness of the model in a wider range of application scenarios. In addition, combining deep learning techniques with FSVM can be further explored to improve the model's ability to handle large-scale high-dimensional data.

REFERENCES

[1] A. Attanasio, M. Maravalle, H. Muccini, F. Rossi, G. Scatena, and F. Tarquini, "Visitors flow management at Uffizi Gallery in Florence, Italy," Information Technology & Tourism, vol. 24, no. 3, pp. 409–434, 2022.

[2] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," Human-centric Computing and Information Sciences, vol. 9, 40, 2019.

[3] K. Madden, G. Lukoseviciute, E. Ramsey, T. Panagopoulos, and J. Condell, "Forecasting daily foot traffic in recreational trails using machine learning," Journal of Outdoor Recreation and Tourism, vol. 44, 100701, 2023.

[4] H. Song, and S. F. Witt, "Forecasting international tourist flows to Macau," Tourism Management, vol. 27, no. 2, pp. 214–224, 2006.

[5] J.-W. Bi, Y. Liu, and H. Li, "Daily tourism volume forecasting for tourist attractions," Annals of Tourism Research, vol. 83, 102923, 2020.

[6] C.-F. Lin, and S.-D. Wang, "Fuzzy support vector machines," IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 464–471, 2002.

[7] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," IEEE Transactions on Fuzzy Systems, vol. 13, no. 6, pp. 820–831, 2005.

[8] Q. Fan, Z. Wang, D. Li, D. Gao, and H. Zha, "Entropy-based fuzzy support vector machine for imbalanced datasets," Knowledge-Based Systems, vol. 115, pp. 87–99, 2017.

[9] V. Cho, "Tourism forecasting and its relationship with leading economic indicators," Journal of Hospitality & Tourism Research, vol. 25, no. 4, pp. 399–420, 2001.

[10] G. Yu, Z. Schwartz, and B. R. Humphreys, "Data patterns and the accuracy of annual tourism demand forecasts," Tourism Analysis, vol. 12, no. 1–2, pp. 15–26, 2007.

[11] C.-F. Chen, M.-C. Lai, and C.-C. Yeh, "Forecasting tourism demand based on empirical mode decomposition and neural network," Knowledge-Based Systems, vol. 26, pp. 281–287, 2012.

[12] Y. Zhang, G. Li, B. Muskat, and R. Law, "Tourism demand forecasting: A decomposed deep learning approach," Journal of Travel Research, vol. 60, no. 5, pp. 981–997, 2021.

[13] A. Shabri, "A hybrid of EEMD and LSSVM-PSO model for tourist demand forecasting," Indian Journal of Science and Technology, vol. 9, no. 36, pp. 1–6, 2016.

[14] M. Paolanti, A. Mancini, E. Frontoni, A. Felicetti, L. Marinelli, E. Marcheggiani, and R. Pierdicca, "Tourism destination management using sentiment analysis and geo-location information: a deep learning approach," Information Technology & Tourism, vol. 23, pp. 241–264, 2021.

[15] W. Xiong, M. Huang, B. Okumus, S. Chen, and F. Fan, "The predictive role of tourist-generated content on travel intentions: Emotional mechanisms as mediators," Asia Pacific Journal of Tourism Research, vol. 27, no. 5, pp. 443–456, 2022.

[16] S. Abe, "Fuzzy support vector machines for multilabel classification," Pattern Recognition, vol. 48, no. 6, pp. 2110–2117, 2015.

[17] A. Chaudhuri, and K. De, "Fuzzy support vector machine for bankruptcy prediction," Applied Soft Computing, vol. 11, no. 2, pp. 2472–2486, 2011.

[18] T.-Y. Wang, and H.-M. Chiang, "Fuzzy support vector machine for multi-class text categorization," Information Processing & Management, vol. 43, no. 4, pp. 914–929, 2007.

[19] Q. Xu, H. Zhou, Y. Wang, and J. Huang, "Fuzzy support vector machine for classification of EEG signals using wavelet-based features," Medical Engineering & Physics, vol. 31, no. 7, pp. 858–865, 2009.

[20] J. Liu, "Fuzzy support vector machine for imbalanced data with borderline noise," Fuzzy Sets and Systems, vol. 413, pp. 64–73, 2021.

[21] M. Nilashi, H. Ahmadi, A. A. Manaf, T. A. Rashid, S. Samad, L. Shahmoradi, N. Aljojo, and E. Akbari, "Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates," International Journal of Fuzzy Systems, vol. 22, pp. 1376–1388, 2020.

[22] J. Hao, S. Luo, and L. Pan, "Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes," Scientific Reports, vol. 12, no. 1, 9858, 2022.

[23] R. Drikvandi, and O. Lawal, "Sparse principal component analysis for natural language processing," Annals of Data Science, vol. 10, no. 1, pp. 25–41, 2023.

[24] J. Zhang, Y. Dai, Z. Feng, and L. Dong, "An enhanced temporal algorithm-coupled optimized adaptive sparse principal component analysis methodology for fault diagnosis of chemical processes," Process Safety and Environmental Protection, vol. 174, pp. 663–680, 2023.
[25] W. Huang, and K. Wei, "An extension of fast iterative shrinkage-thresholding algorithm to Riemannian optimization for sparse principal component analysis," Numerical Linear Algebra with Applications, vol. 29, no. 1, e2409, 2022.
[26] H. Robert Frost, "Eigenvectors from eigenvalues sparse principal component analysis," Journal of Computational and Graphical Statistics, vol. 31, no. 2, pp. 486–501, 2022.
[27] Y. Liu, X. Li, and X. Lin, "Evaluating carbon neutrality potential in China based on sparse principal component analysis," Energy Reports, vol. 9, pp. 163–174, 2023.
[28] Y. Zhao, M. A. Lindquist, and B. S. Caffo, "Sparse principal component based high-dimensional mediation analysis," Computational Statistics & Data Analysis, vol. 142, 106835, 2020.