Sunspot Prediction Based on The Adaptive ARIMA Model

Heng Zhao [1, #], Hao Chen [2, #], Linhao Qian [2, *, #]

[1] School of Computing and Artificial Intelligence, Wenzhou university, Wenzhou, China, 325000

[2] College of Sciences and Mathematics, Wenzhou university, Wenzhou, China, 325000

* Corresponding Author Email: 22211021222@stu.wzu.edu.cn

[#]These authors contributed equally.

Abstract. The sun is closely related to human beings, and any change in its activities may have a huge impact on human beings. The study and prediction of solar activities is particularly important for the survival and development of human beings. Humans often learn about solar activity by studying sunspots. In this paper, the study predicts the start time, end time, sunspot number, and area size of the next sunspot. In this paper, we predict the length and sunspot number and area of the next solar activity cycle based on the ARIMA time series model, the BP neural network model, and the trend model. In this paper, the ARIMA-BP neural network hybrid model is predicted that the next sunspot start time is 2031 and the end time is 2042. The trend model, the number of time and sunspot, predicts that the number of the next sunspot is 12462.58 and the area is 1640.68.

Keywords: ARIMA time series model, BP neural network model, ARIMA-BP neural network hybrid model, trend model.

1. Background

To illustrate the origin of solar activity, the following background is worth mentioning. As an astronomical phenomenon, sunspots are a dim region of the surface of the sun, usually associated with a strong magnetic field. Solar spot observations are crucial for understanding the solar cycle and the solar magnetic field [1]. The number and distribution of sunspots vary over time and are associated with periodic changes in solar activity [2].

Overall, the sunspot problem involves the study of the magnetic activity on the solar surface and its contribution to the solar activity cycle. By observing and recording the activities of sunspots, we can better understand the various behaviors of the sun and its effects on the Earth and other planets.

2. Research method

2.1. ARIMA model

2.1.1 Observation time series

First, the time series of the analyzed solar activity cycle were visually statistically described to understand its basic characteristics, including trends, seasonality, and periodicity [3].

2.1.2 Stability processing

The resulting time series is unstable and requires differential operation until stability is reached. Difference time is called difference order and is usually expressed by d.

2.1.3 Determine the model order

The ARIMA (p, d, g) model focuses on the selection of three parameters (p, d, g). In this problem, we use the Bayesian Information criterion (BIC) to select p and q. The Bayesian information criterion can give a simple approximate logarithmic model evidence, as follows:

$$\mathrm{BIC} = \mathrm{Accuracy(m)} - \frac{\mathrm{p}}{2}\log \mathrm{N} \tag{1}$$

2.1.4 Establish the ARIMA model

$$\varphi(B)(1-B)^{d}y_t = \theta(B)\varepsilon_t \quad (2)$$

The yt is the time series of historical observations, the d is the order of the difference, the p and the g are the order of the autoregressive model and the moving mean of previous observations, and the E is the sequence of independent identicously distributed white noise with mean and variance constant zero. The B is the lag operator, and the B satisfies the following expression:

$$B^n y^t = y_{t-n} \quad (3)$$

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p \quad (4)$$

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_a B^q \quad (5)$$

2.1.5 Model test

The fit of the model was used to test whether the sequence of the residues was white noise.

2.1.6 Testing by using the ARIMA model

Predict future data from the already fitted ARIMA models.

2.2. BP cerebellum model arithmetic computer

(1) Define a network structure

Determine the number of nodes in the input layer, usually with the same feature dimension. Determine the number of nodes and hidden layers, which requires selection based on the complexity of the problem and the characteristics of the data. The number of nodes determining the output layer is usually the same as the output dimension of the problem [4].

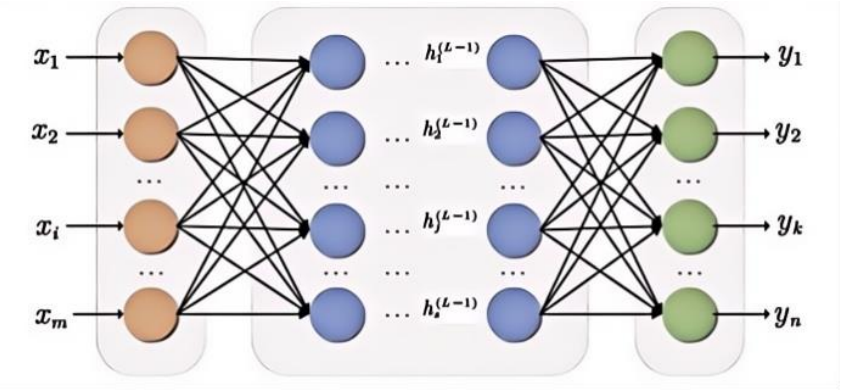


Figure 1. BP, the structure diagram of the cerebellar model

We can know that through Figure 1.

2.2.1 Initialize the weights and deviations

Call the training set, and the output is:

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}, x_i \in R^d, y_i \in R^l, \hat{y}_k = (\hat{y}_1^k, \hat{y}_2^k, \cdots, \hat{y}_l^k) \quad (6)$$

$$\hat{y}_j^k = f(\beta_j - \theta_j \quad (7)$$

$$\beta_j = \sum_{i=1}^{n} w_{ij} x_{ij} \quad (8)$$

Where, we are the connection weights form the i-th neuron to the j-th output.

Calculate the loss: Use the loss function to calculate the error between the model output and the actual label. Error in the memory network (x_k, y_k) is:

$$E_k = \frac{1}{2}\sum_{j=1}^{l}(\hat{y}_j^k - y_j^k)^2 \quad (9)$$

2.2.2 Backpropagation

When the neural network completes the forward calculation, the predicted value is reduced from the actual value to obtain the error value, and then the backpropagation adjusts the minimum weight value of the neural network [5].

2.2.3 Update parameters

Update the network parameters to reduce the loss of using gradient descent to other optimization algorithms.

2.2.4 Repeat training

Iterative update formula:

$$\Delta\omega_{hj} = \eta g_j b_h \tag{10}$$

$$\Delta\theta_j = -\eta g_j \tag{11}$$

$$g_j = -\frac{\partial E_k}{\partial \hat{y}_i^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} = -(\hat{y}_j^k - y_j^k) \cdot (\hat{y}_j^k)' \tag{12}$$

Where, b is the input data for that neuron. On this basis, the neural network constantly adjusts the weights and impulse weight values during the training process, so that the prediction error of the neural network is constantly close to 0.

2.3. RBF model

2.3.1 Data preparation and data standardization

Collecting and preparing datasets for training and testing. Make sure that the dataset contains the input features and the corresponding target variables. The input features are standardized to ensure that the training process of the neural network is more stable. Commonly used normalization methods include Z-score normalization or minimum-maximum normalization [6]. Using the Gaussian function as the radial basis function, the network output is expressed as follows:

$$y = \sum_{i=1}^{l} w_i exp(-\frac{1}{2\sigma^2}\|X_K - h_i\|)^2, i = 1,2,\cdots,m \tag{13}$$

$$X_K = (x_{k1}, x_{k2}, \cdots, x_{kn}) \tag{14}$$

So: x_k is the K-THK input sample, k=1,2, 3., K; K is the number of samples; h is the connection weight from the hidden layer to the output layer; m is the number of nodes in the hidden layer in the RBF neural network; and σ is the variance of a Gaussian function.

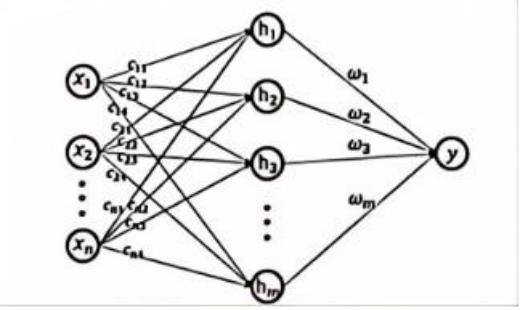


Figure 2. Structure diagram of the RBF model

We can know that through Figure 2.

2.3.2 RBF determination center

Select the location of the RBF neural network center. RBF neural network centers using iterative optimization of particle swarm optimization [7].

2.3.3 Determine the width of the RBF

Determine the width of the RBF function and calculate the width of the RBF using the K-means clustering method.

Rights training:

Train the network through repeated iterative calculations. In the iterative process, the particle population updates its own speed and position with its own individual and global extremes. The iterative process is as follows:

$$v_{pd}{}^{(i+1)} = wv_{pd}{}^{(i)} + c_1 r_1 (P_{pd}{}^{(i)} - X_{pd}{}^{(i)}) + c_2 r_2 (P_{gd}{}^{(i)} - X_{id}{}^{(i)}) \quad (15)$$

$$X_{id}{}^{(m+1)} = X_{id}{}^{(m)} + v_{id}{}^{(m+1)} \quad (16)$$

$$\mathrm{d} = 1,2,\cdots,\mathrm{D};\ \mathrm{p} = 1,2,\cdots,\mathrm{n} \quad (17)$$

Among, PTHwi is the p particle; is the inertia weight; is the number of iterations; d is the dimension of the p particle; is the PX_{pd}THP_{pd}The position of the particle, is a single optimal position for the PTH particle. P is the position with the best SSE value of all particles.$v_{pd}c_1, c_2 r_1, r_2$ Is the velocity of the particle p; represents the acceleration factor, usually non-negative, is a random number distributed between [0,1].

The calculation formula for the linear decreasing inertia weight is:

$$\mathrm{w(k)} = \mathrm{w_{start}} - (\mathrm{w_{start}} - \mathrm{w_{end}})(\mathrm{Tmax_{max}}) \quad (18)$$

3. Solution and results

3.1. Data preprocessing

Since the resulting magnetic field data contain missing values, the missing values should be interpolated [8]. In this topic, the missing magnetic field data are interpolated by fitting the model. The general steps of the fit-based interpolation methods are as follows:

(1) Collect existing data: First, it is necessary to collect data sets containing missing data. Ensure sufficient samples and features in the dataset to perform nonlinear fitting.

(2) Model construction: select the appropriate model to adapt to the existing data. Select appropriate models based on the characteristics of the data and domain knowledge.

(3) Fit the existing data: Use the existing data to fit the selected nonlinear model. Here, methods such as regression analysis were used to estimate the parameters of the model.

(4) Inpolation of missing data: use the fitted non-linear model to predict the value of missing data. The eigenvalues in the existing data were replaced into the model to obtain the corresponding predicted values [9].

The model can be expressed as follows:

Using sequences, $\mathrm{f}(\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_{i-1}}, \mathrm{x_{i+1}}, \ldots, \mathrm{x_n})\mathrm{x} = \mathrm{x\,i\,f}(\mathrm{x})\mathrm{f}(\mathrm{x})$ Enter the input to fill in the missing values. In these cases, the function can be expressed as follows:

$$\mathrm{y} = \mathrm{f}(\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_{i-1}}, \mathrm{x_{i+1}}, \ldots, \mathrm{x_n}) + \mathcal{E} \quad (19)$$

The insertion effect is shown in the Figure 3:

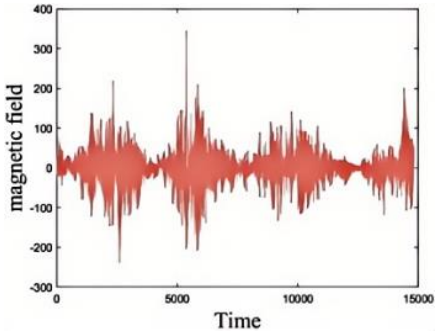


Figure 3. Sequence diagram in days.

Then, since the title requires it to be in months, the date should be converted to a sequential diagram in months, as shown Figure4:

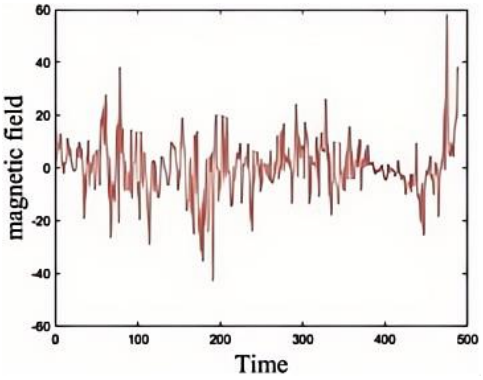


Figure 4. Sequence diagram in months.

3.2. Solve the sunspot period.

Based on the characteristics of the problem and the data analysis obtained, it was decided to use an adaptive hybrid ARMI-BP neural network model to predict the onset and duration of the current and next solar activity [10].

Model idea: The better the performance of past predictions, the higher the weight of future predictions, and the greater the contribution to the predicted value.

Prediction value record: for the actual observed value: $y_i = \{y_1, y_2, y_3, \ldots . y_t\}$

The predicted value of the algorithm k is recorded as:

$$\hat{y}_{k,i} = \{\hat{y}_{k,1}, \hat{y}_{k,2}, \hat{y}_{k,3}, \cdots, \hat{y}_{k,i}\} \tag{20}$$

Error analysis: The prediction error can be recorded as:

$$\mathrm{wmape_{k,id}} = \frac{\sum|\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{k,i}}|}{\sum \mathrm{y_i}} \tag{21}$$

The total error of Algorithm k is expressed as:

$$\mathrm{wmape_k} = \sum_{\mathrm{id}} \mathrm{wmape_{k,id}} \tag{22}$$

The weight calculation of an algorithm: In a hybrid algorithm, the better the performance of an algorithm in the test set, the higher the weights. The weight of algorithm k in the hybrid algorithm can be recorded as:

$$\omega_k = \frac{1/mape_k}{\sum_k (1/wmape_k)} \tag{23}$$

among,the ω_k is the weight of algorithm k in the hybrid algorithm.

Model building: combining ARIMA algorithm with BP neural network algorithm. The calculation formula can be expressed as follows:

$$\hat{y}_i = \sum_k \omega_k \hat{y}_k, i \tag{24}$$

Running the model produces the following results in Figure5 and Figure6:

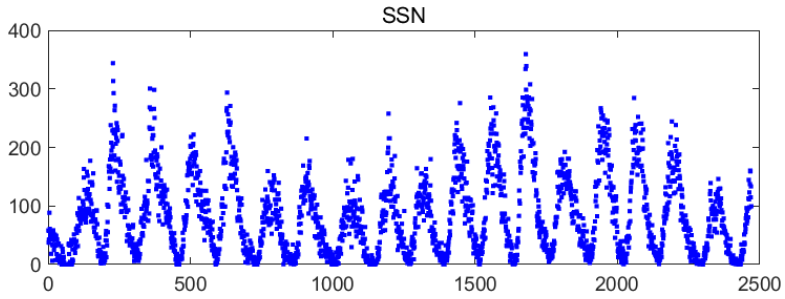


Figure 5. Sunspot number data, extraction table

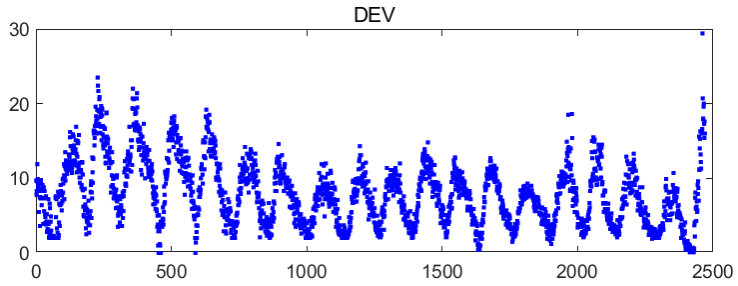


Figure 6. Standard deviation rainfall time relationship table

The prediction results of the sunspot change are also obtained in Figure7:

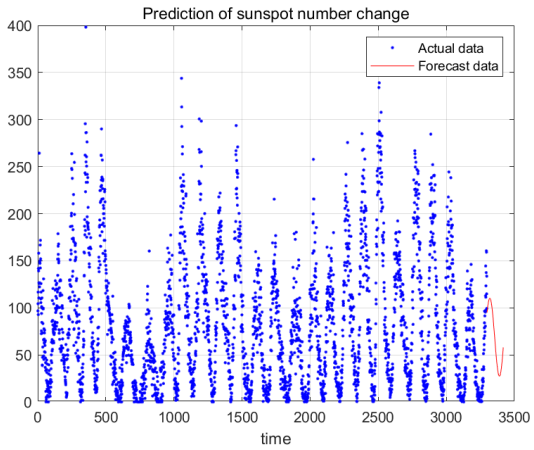


Figure 7. Actual and predicted number of sunspots

3.3. Solve the sunspot and the area.

Establishment of the trend extraction model: after collecting data for preprocessing and extracting key features, select the appropriate model according to the significance of the problem.

Establishment of the trend separation model: use the trained model to separate the trend, get the trend part and the non-trend part, and visualize the separation trend, so as to better understand and explain the results of the model.

Countertrend model solution：Data preparation: Prepare the time series data, ensure the correct time order of the data, and conduct the necessary cleaning and preprocessing of the data [7] 。

Linear regression model: with time as the independent variable and sequence data as the dependent variable. The model can take the form of, where is the time series data, t is the time point, $\beta y_t = \beta_0 + \beta_1 \cdot t + \epsilon_t, \beta_0$ and β_1 is the regression coefficient, and it is the error term.

Model fitting: least squares were used to estimate the regression coefficient and fit the linear regression model.

Eliminate linear trend items:

$$y = \text{detrend}(x) \tag{25}$$

Eliminate polynomial trend items (the trend is a nonlinear curve):

$$p = \text{polyfit}(t, \text{signal}, 5) \tag{26}$$

$$x_{\text{trend}} = \text{polyval}(p, t) \tag{27}$$

$$y = \text{signal} - x_{\text{trend}} \tag{28}$$

Use the sgolay filters to eliminate the trend items (Figure8 and Figure 9) :

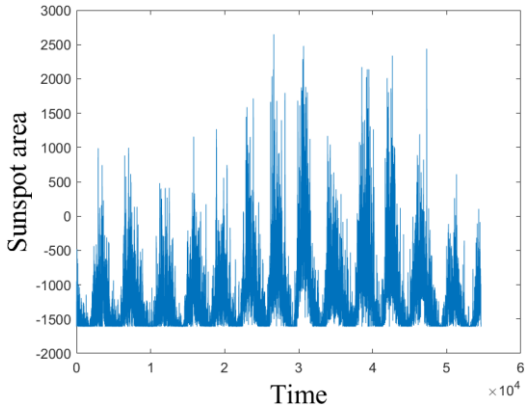


Figure 8. Post-trend fluctuations of the number of sunspots

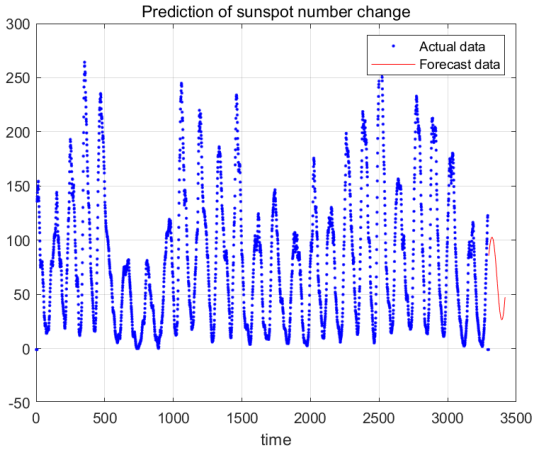


Figure 9. Trend fluctuations of sunspot regions

(1) End of decrease: subtract the trend from the original sequence to obtain the non-trend component. Non-trend component can be expressed as follows Figure10:

Figure 10. Plot of the relationship between time and the number of sunspots

$$\hat{\epsilon}_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 \cdot t) \tag{29}$$

(2) Results Analysis: The resulting non-trend sequence is analyzed to test whether the effect of the detrend is reached.

(3) Finally, the following predictions are obtained Figure 10:

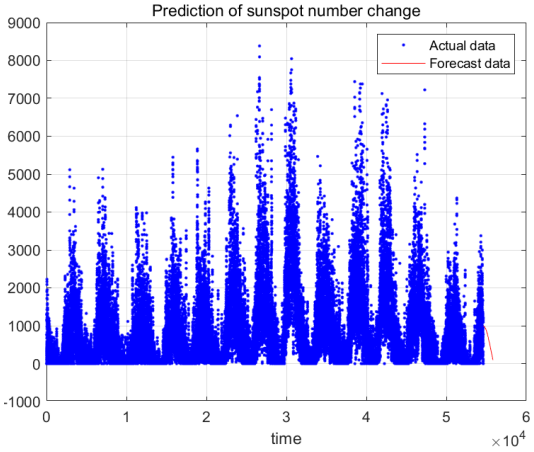


Figure 10. Time and sunspot area map

3.4. Analysis of the results

3.4.1 Solar Cycle forecast

In this problem, using the Fourier curve fitting together with the formula:

$$f(x) = a_0 + a_1 * \cos(x * w) + b_1 * \sin(x * w) \quad (30)$$

The obtained results are shown in Figure 11 and Figure 12:

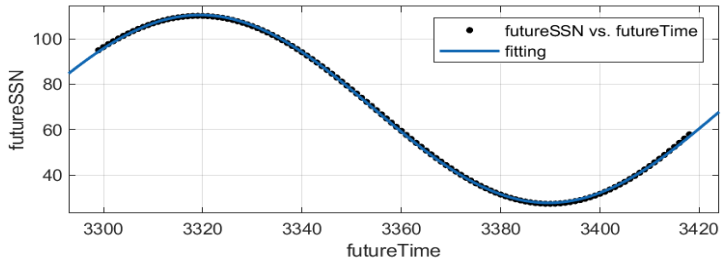


Figure 11. Fourier curve fit 1

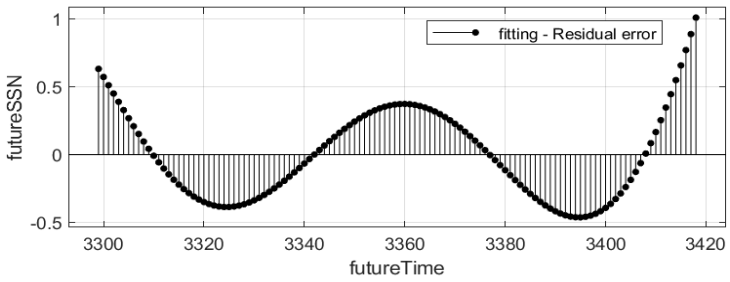


Figure 12. Fourier curve fit 2

The results presented in the following table are shown in Table1 and Table2:

Table 1. Coefficient and 95% confidence bounds

	graph	lower limit	superior limit
a_0	69.1124	69.0471	69.1777
a_1	-23.7068	-32.3046	-15.1090
b_1	-33.8336	-39.8831	-27.7840
w	0.0448	0.0447	0.0448

Table 2. Goodness of fit

	graph
SSE	13.4279
R^2	0.9999
law	116.0000
adjust R^2	0.9999
error of mean square	0.3402

3.4.2 Sunspot and area forecast

The data in this topic are polynomial curve fitted with the formula:

$$f(x) = p_1 * x + p_2 \tag{31}$$

The results are shown in the Figure13, Figure 14 and Table 3:

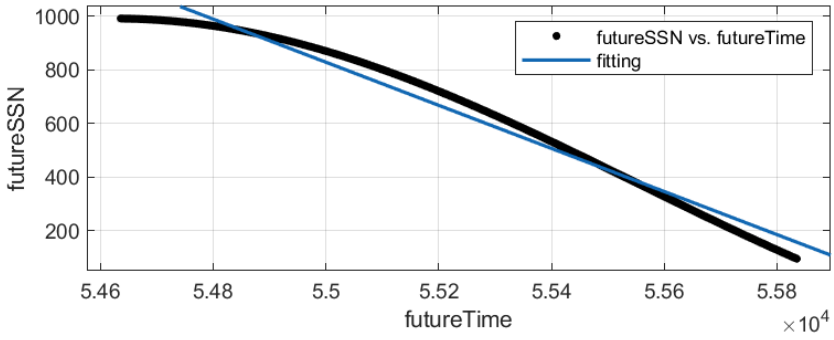


Figure 13. Polynomial curve fit 1

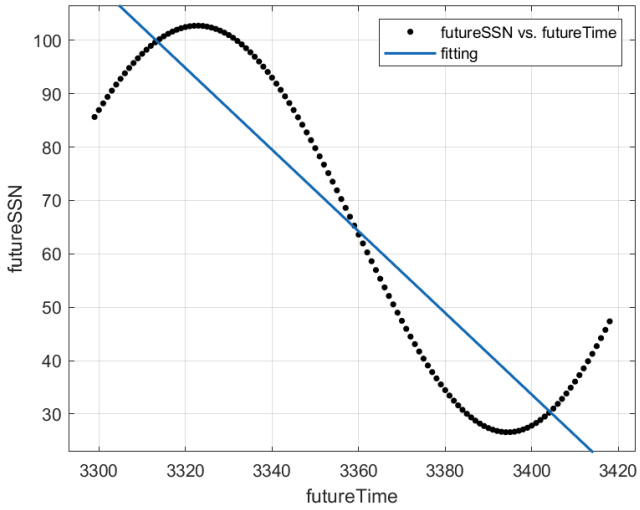


Figure 14. Polynomial curve fit 2.

Table 3. Fitness

	graph
SSE	1.6731e+04
R^2	0.8343
118	118
adjust R^2	0.8329
error of mean square	11.9073

4. Conclusion

The next solar cycle is projected to commence in 2031 and conclude in 2043. In the current solar cycle, the number of sunspots is 1144.21 with an area of 1532.79, while the next cycle is anticipated

to have 12462.58 sunspots with an area of 1640.68. According to Table 3, the model's R^2 is 0.8343, indicating a small error, high fit, and relatively successful prediction. The Degrees of Freedom for Error (DFE) is 118, which is high, suggesting good model performance. The adjusted R^2 is 0.8329, significantly higher than 0.6 and close to 1, indicating strong explanatory power of the independent variables on the dependent variable. The Root Mean Square Error (RMSE) is 11.9073, a value lower than other models, indicating a higher match between predicted and actual results. In conclusion, the model exhibits successful predictive performance for future solar activity cycles, characterized by high accuracy and explanatory power.

References

[1] Zhang person. Preliminary study on the relationship between sunspot activity cycle and annual precipitation in Beijing [J]. Journal of Capital Normal University (Natural Science edition), 1995 (01): 81-86.

[2] Guo Jijun, Guo Zhilong, Wang Ninglian. Study of sunspot activity cycles in the Earth's land-ocean temperature index since 1880. Plateau Meteorological System, 2020,39 (04): 851-858.

[3] Pan Chunhua, Sun Yan, Zhu Cun. Neural network and wavelet analysis of the periodic characteristics of the sunspot activity. Computer Technology and Development, 2016,26 (03): 158-161.

[4] Sun Wei, Wei Xiaoxiao, Liu Fuyao, Wang Jian, Xu Shuting, Yan Rui. Numerical simulation and prediction of the magnetic polarity index of the sunspot magnetic field. Progress in Geophysics 1-12.

[5] Jiang Tingting, Bi Shuben, Lu Ying, etc. The periodic characteristics of drought disasters in 1644-1911 and their response to sunspot activity [J]. Journal of Southwest Normal University (Natural Science Edition), 2018,43 (10): 56-65.

[6] Cheng Shu, Shi Yaolin, Zhang Huai. Prediction of sunspot changes based on the neural network [J]. Journal of the University of Chinese Academy of Sciences, 2022,39 (05): 615-626.

[7] P. B K, I. V K. The Lower Ionosphere and Tectonic Processes on Earth[J]. Geomagnetism and Aeronomy,2023,63(7):1038-1046.

[8] V. D E, V. A E. Investigation of Joy’s Law According to the Data of the Ussuriisk Astrophysical Observatory[J]. Geomagnetism and Aeronomy,2023,63(7):1007-1014.

[9] A. Y N, A. A O. Extreme Values of Sunspot Activity on a Long Time Scale[J]. Astronomy Letters,2023,49(7):421-429.

[10] Sunspot the Phoenix Sun [J]. NBA Special Issue, 2022, (20): 19 + 18.