IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Comparative Study on Prediction Algorithms for Power Grid System Access Failure Times

IOP Conf. Series: Earth and Environmental Science **252** (2019) 032183 doi:10.1088/1755-1315/252/3/032183

Comparative Study on Prediction Algorithms for Power Grid System Access Failure Times

Yi Yan*, Bo Li, Jianyi Xiao, Yunde Liang, Yanwei Shang and Kaidong Zhou

Information Center Guangdong Power Grid Company, China

*Corresponding author e-mail: yanyi@gdxx.csg.cn

Abstract. As functions of the power grid system gradually increase and its scale becomes larger, the frequency of access failure times will gradually increase. In order to prevent access failures and make countermeasures in advance, we perform model fitting analysis and fault risk prediction for the time series of power grid failure code. In this paper, we use the failure code 404 as the training data and use the SARIMA algorithm, Fbprophet algorithm, holt-winter algorithm, and GM algorithm respectively to construct a time series prediction model. According to the result of the model building and the calculating, we find that the root mean square error of SARIMA algorithm is 258.85, which is the lowest among these algorithms, and the root mean square errors of Prophet and holt-winter algorithm are 749.288 and 809.89, respectively. However, the root mean square error of GM algorithm reaches 1710.95, which is 6 times as many as the SARIMA algorithm. In conclusion, with algorithm analysis and the comparisons of these four algorithms, we recommend the SARIMA algorithm as a predictive model for the power grid system.

1. Introduction

In recent years, with the continuous development of computer and communication technology and applied to power grid management, the management efficiency of accessing the power grid system has been greatly improved, and the safe and stable operation of the access grid system has been guaranteed [1-5]. The serv2er logs a large amount of data when an HTTP request is made to multiple services on the grid. For example, the failure codes for regular failures are 404 and 500. The 404 failure code means no resources when accessing the page, 500 indicates page error; 400 for sudden failure (the server does not understand the syntax of the request), 401 (request for identity) Verification), 503 (the server is currently unavailable), 504 (not receiving the request from the upstream server in time), etc. When the scale of the power grid system becomes larger, the HTTP requests for multiple services of the power grid are also increasing. The fault codes for various access failures are massive, which requires a large amount of manual intervention for processing. This leads to the long processing and prediction time of the access fault, which affects the fault location time, which leads to prolonged fault events. The problem cannot be solved in time and cannot be prevented in advance, and the user experience is not good. In order to solve this problem, a set of intelligent fault prediction warning system with high accuracy and adaptability is needed, which can utilize these large amounts of data to adaptively predict and warn different types of faults.

2. Algorithm introduction

2.1. Data preparation and pre-processing

The raw data represents the number of failure codes 404 appearing in each hour from 00:00 on August 1, 2018, to 24:00 on August 28, 2018. A total of 672 data, of which 670 are valid data, and 2 data are missing. The average value of the raw data is 2864.23, the standard deviation is 1963.53, the minimum value is 163, and the maximum value is 6433.

We use the equalization method based on binning technology to supplement the missing data. The specific steps are as follows:

Divide all data into 24 boxes 24 hours a day, and put the data for the same time period into the same box every day;

Fill the missing data between the two valid values with an equal difference sequence and then round off;

The missing values at the beginning and the end are filled with adjacent valid values.

After the data is filled, 672 data can be obtained. The raw data and the complete data statistics are shown in Table 1.

Table 1. Statistical analysis of failure code 404 data

	Raw data	Missing value filled data
count	670	672
missing	2	0
mean	2684.23	2679.58
std	1963.53	1962.65
min	163	163
25%	934	933.25
50%	2223	2218
75%	4925	4921
max	6433	6433

A time series image of the complete data is shown in Figure 1.

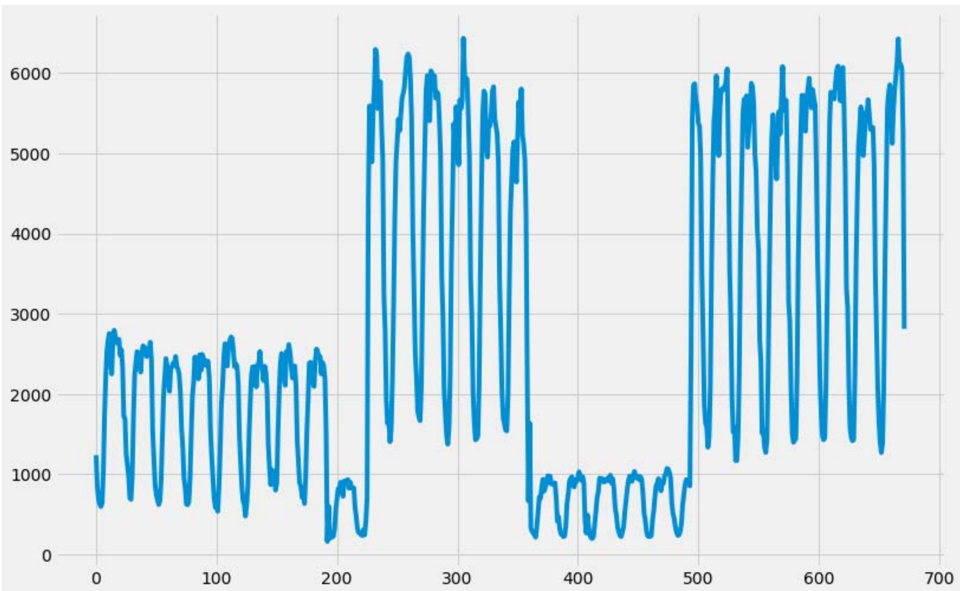

Figure 1. Time series diagram of failure code 404

The following report text will take the failure code 404 as an example, use different algorithm models for time series prediction, and perform effect analysis and comparison to select the best and most appropriate algorithm.

2.2. SARIMA algorithm principle and modelling steps

The full name of the ARIMA model is called the autoregressive moving average model and is denoted as ARIMA (p, d, q) [6]. The meaning is as follows: suppose a stochastic process contains d unit roots, which can be transformed into a stationary autoregressive moving average process after d difference, then the stochastic process is called the one-product (integral) autoregressive moving average process. The general form is

$$\Phi(L)\Delta^d x_t = \delta + \Theta(L)u_t$$

Where $x_t$is the original sequence, L is the backward shift operator, $\Delta^d = (1-L)^d$ is the d-order difference $\Phi(L) = 1 - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p, \Theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots \theta_p L^p, u_t$ zero mean white noise series

When the time series exhibits seasonal and linear trends, the stochastic seasonal model and model can be combined into a seasonal time series model, ie, a model to describe the time series, called SARIMA. The SARIMA model is a short-term prediction model. The core point is the processing of data. At the same time, the error generated by fitting the value is taken as the analysis factor. The outstanding advantage is that the accuracy of the short-term prediction result is high. The general form of the SARIMA model is expressed as

$$\Phi_p(L)A_p(L^T)(\Delta^d \Delta^T x_t) = \Theta(L)B_Q(L^T)u_t$$

Where T is the period of change of the seasonal sequence; L is the lag operator; $\Phi_p(L)$, $A_p(L^s)$are the non-seasonal and seasonal autoregressive polynomials; $\Theta(L)$, $B_Q(L^s)$ Representing non-seasonal and seasonal moving average polynomials; subscripts P , Q , p , q represent the maximum lag order of seasonal and non-seasonal autoregressive and moving average operators, respectively; d and D represent non-seasonal and seasonal difference times, respectively In practical applications, if the original sequence contains both trend and seasonality, it can be expressed as a Seasonal $\mathrm{ARIMA(p, d, q)} \times (\mathrm{P, D, Q, T})$model.

The main steps of the algorithm are as follows:

a) The sequence data x_t is obtained according to the time series scatter plot, autocorrelation function and partial autocorrelation function graph. The variance, trend and seasonal variation of the ADF unit root are used to identify the station's stationarity. The original sequence is converted to a smooth sequence by differential and seasonal differences. For the non-stationary time series, the d-order difference operation is first performed and turned into a stationary time series.

$$w_t = \Delta^d x_t \Delta_T^D x_t$$

Where w_t is a stationary sequence

b) The autocorrelation coefficient ACF and the partial autocorrelation coefficient PACF are obtained for the stationary time series w_t, and the best hierarchical p and the order q are obtained by analyzing the autocorrelation graph and the partial autocorrelation graph. $w_t \sim \mathrm{ARMA(p, q)}$, Model form is:

$$w_t = \varphi_1 w_{t-1} + \varphi_2 w_{t-2} + \cdots + \varphi_p w_{t-p} + \delta + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$

c) Using AIC as the evaluation index, iteratively change the parameters of the ARIMA model, obtain the ARIMA model parameters that make the AIC index optimal, and obtain the model. x_t~ARIMA(p,d,q)×(P,D,Q,T)

2.3. Prophet algorithm principle and modeling steps

The Prophet is an open source software released by Facebook's core data science team [7]. The basic model is as follows:

$$\mathrm{y(t) = g(t) + s(t) + h(t) + \epsilon t}$$

IOP Conf. Series: Earth and Environmental Science **252** (2019) 032183 doi:10.1088/1755-1315/252/3/032183

Here, the model divides the time series into three superpositions, where g (t) represents the growth function used to fit the aperiodic changes. S (t) is used to indicate periodic changes, such as weekly, yearly, seasonal, etc., h (t) represents changes caused by special reasons such as holidays, holidays, etc. Finally, εt is a noise term, and it is used to indicate that it is random and unpredictable. The fluctuations, we assume that εt is Gaussian.

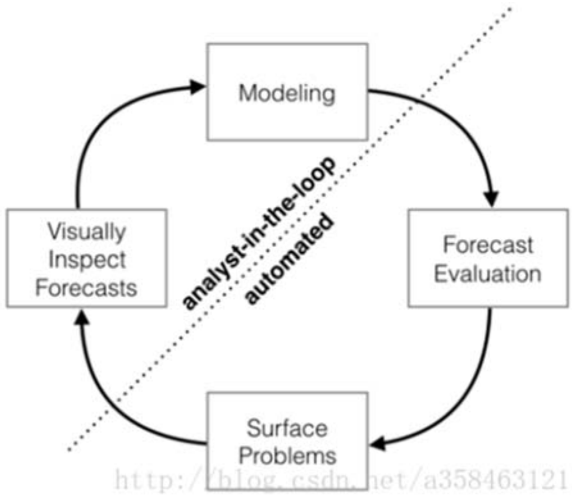

Figure 2. Prophet workflow

The workflow of the prophet is as shown in Figure 2. By integrating the two modules of modeling-evaluation, the rapid iterative optimization of the time series model is realized.

The publicity of the growth function is as follows:

$$g(t) = \frac{C(t)}{1 + \exp(-(k + \mathbf{a}(t)^\intercal \boldsymbol{\delta})(t - (b + \mathbf{a}(t)^\intercal \boldsymbol{\gamma})))}$$

Seasonal functions can use the Fourier series approximation:

$$s(t) = \sum_{n=-N}^{N} c_n e^{i \frac{2\pi n t}{P}}$$

The method of dealing with festivals is very simple. It is to set a dummy variable for the same holiday in the past and in the future, which can be expressed by the following formula. Where Di represents the i-th dummy variable, and if it belongs to this, it does not belong to 0.

$$h(t) = \sum_{i=1}^{L} \kappa_i \mathbf{1}(t \in D_i)$$

$$Z(t) = [\mathbf{1}(t \in D_1), \ldots, \mathbf{1}(t \in D_L)]$$

$$h(t) = Z(t)\boldsymbol{\kappa}. \quad \boldsymbol{\kappa} \sim \mathrm{Normal}(0, \nu).$$

2.4. Holt-winter algorithm principle and modeling steps

Exponential smoothing is a simple calculation scheme. According to different model parameters, the form of exponential smoothing can be divided into an exponential smoothing method, a second exponential smoothing method and a cubic exponential smoothing method. One of the exponential smoothing methods is for sequences without trend and seasonality, the second exponential smoothing method is for time series with the trend but no seasonal characteristics and the third exponential smoothing rule can predict time series with trend and seasonality. And "Holt-Winter" refers to three exponential smoothing [8].

The prediction result obtained by an exponential smoothing method is a straight line at any time. Not suitable for time series with the overall trend, if used to process sequences with overall trends, the smoothed values will lag behind the original data. The second exponential smoothing method preserves the smooth information and trend information so that the model can predict the time series with trends. The cubic exponential smoothing method is smoother than the second exponential, and a third amount is added to describe the seasonality. The equation for the cumulative seasonality is:

$$s_i = \alpha * (x_i - p_{i-k}) + (1-\alpha)(s_{i-1} + t_{i-1})$$
$$t_i = \beta * (s_i - s_{i-1}) + (1-\beta)t_{i-1}$$
$$p_i = \gamma(x_i - s_i) + (1-\gamma)p_{i-k}$$
$$x_{i+h} = s_i + h * t_i + p_{i-k+h}$$

The equation for the cumulative multiplication seasonality is:

$$s_i = \alpha * \frac{x_i}{p_{i-k}} + (1-\alpha)(s_{i-1} + t_{i-1})$$
$$t_i = \beta * (s_i - s_{i-1}) + (1-\beta)t_{i-1}$$
$$p_i = \gamma\frac{x_i}{s_i} + (1-\gamma)p_{i-k}$$
$$x_{i+h} = ()s_i + h * t_i)p_{i-k+h}$$

Where p_i is a periodic component representing the length of the period. X_ ((i+h)) is the equation for model prediction.

2.5. *GM algorithm principle and modeling steps*

The GM (1, 1) model, the univariate first-order grey model, is the basic model of the grey system theory [9]. The principle is to accumulate the original sequence, make the generated sequence show a certain trend law, and establish a differential equation model for the generated sequence. The time response function prediction model is obtained by solving the differential equation to realize the prediction of the system. The modeling steps are as follows:

1) Accumulate

Remember that the original time series is: $X^{(0)} = \{X^{(0)}(1), X^{(0)}(2), \cdots, X^{(0)}(\mathrm{n})\}$ is a first-order accumulation to generate a new sequence y $X^{(1)}(\mathrm{k}) = \sum_{i=1}^{k} X^{(0)}(\mathrm{i}), k = 1, 2, \cdots, n.$

2) Construct a first-order differential equation

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = \mu$$

Among them, the parameters a, μ can be obtained by the least squares method.

To solve the differential equation, you can get the prediction model:

$$\hat{X}^{(1)}(k+1) = \left(X^{(0)}(1) - \frac{\mu}{a}\right)e^{-ak} + \frac{\mu}{a}.$$

3. Algorithm application

3.1. *SARIMA model establishment and prediction effect*

We take the data from the time point 504 to 624 as the training data, and the data at the time point 625 to 648 as the test comparison data. The training data and comparison data are shown in Figure 3.

Using Seasonal ARIMA$(\mathrm{p}, \mathrm{d}, \mathrm{q}) \times (\mathrm{P}, \mathrm{D}, \mathrm{Q}, \mathrm{T})$model, Where (p, d, q) is the aperiodic part, (P, D, Q, T) is the seasonal period part and T is the period. P and P are the autoregressive parts of the model, d and Dare the difference orders, and q and Q are the moving average parts of the model. The method of traversal search and the minimum AIC value are then used to determine the optimal combination of the two sets of parameters.

The corresponding model results are obtained and sorted according to the AIC value in ascending order. The final result and the predicted graph are as follows. The blue curve is the actual data, and the red curve is the prediction result:

$$(\mathrm{p}, \mathrm{d}, \mathrm{q}) \times (\mathrm{P}, \mathrm{D}, \mathrm{Q}, \mathrm{T}) = (4, 1, 0) \times (3, 1, 0, 24)$$
$$\mathrm{AIC} = 281.7753160202459$$

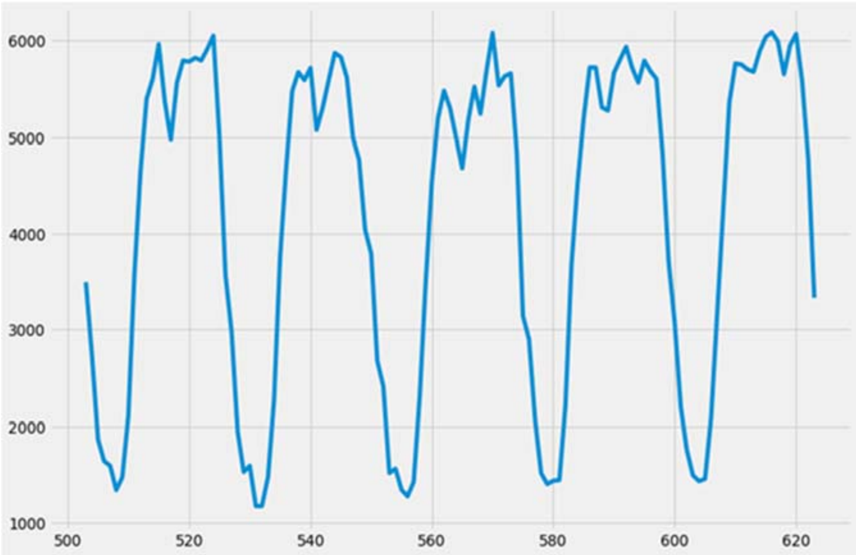

Figure 3. Failure code 404 training data

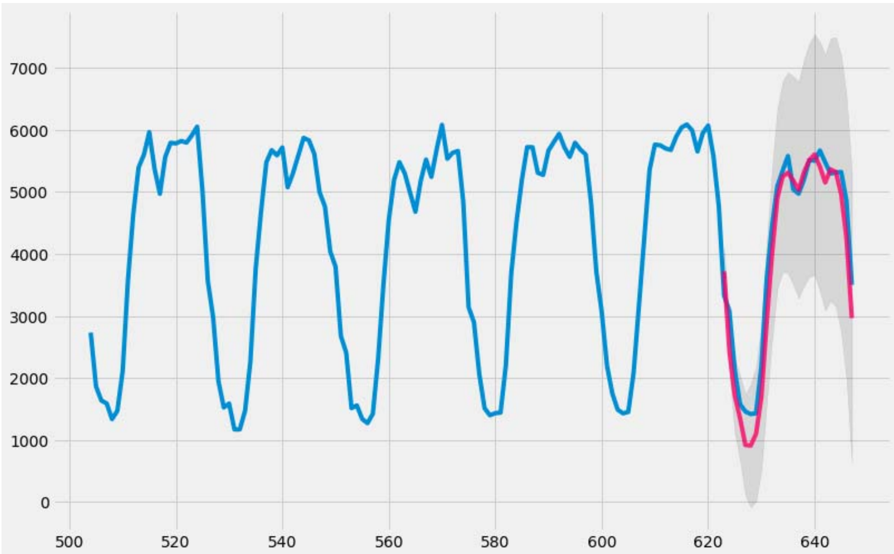

Figure 4. Failure code 404 prediction results

The root mean square error of the predicted value and the actual comparison value is 258.85, and the correct rate of the actual value in the predicted confidence interval is as high as 95.8%, and only one point is not within the confidence interval of the predicted value. It can be seen from the prediction results that the comparison data is basically consistent with the trend of the predicted values, indicating that the model is reasonable.

3.2. Fbpropher model establishment and prediction effect

Similarly, we take the data from the time points 504 to 624 as training data and the data at time points 625 to 648 as test comparison data. After repeated iterations, select year seasonality=15, the training effect of the prophet model is shown in Fig. 5, and the prediction data is shown in Table 2.

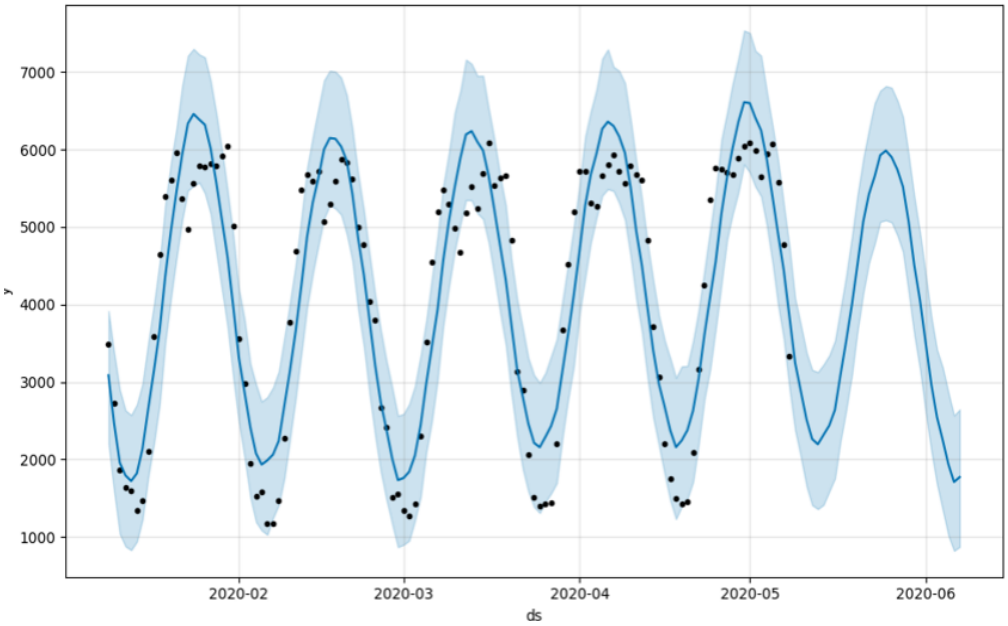

Figure 5. The training effect of the prophet model

Table 2. Prediction results of the prophet model

	actual value	Predictive value	Upper limit	Lower limit
0	3079	3244.872	4157.296	2384.562
1	2131	2878.934	3733.675	2053.695
2	1578	2517.56	3368.847	1600.265
3	1459	2263.074	3100.486	1378.233
4	1418	2195.565	3101.354	1292.857
5	1432	2314.715	3235.674	1504.789
6	2190	2436.1	3375.35	1651.515
7	3594	2635.959	3498.688	1779.867
8	4449	3095.465	4020.678	2200.653
9	5090	3532.527	4432.869	2635.018
10	5341	3996.906	4945.324	3104.193
11	5578	4521.577	5457.476	3607.103
12	5040	5067.64	5922.299	4151.767
13	4971	5423.113	6312.049	4486.235
14	5197	5649.479	6503.994	4766.61
15	5513	5926.873	6808.631	5070.599
16	5502	5985.907	6848.467	5130.801
17	5667	5901.895	6808.601	5038.834
18	5475	5744.491	6586.268	4839.588
19	5295	5520.145	6342.186	4661.403
20	5312	5067.884	5985.036	4141.497
21	5321	4502.488	5287.002	3607.122
22	4850	4056.163	4950.146	3118.174
23	3500	3506.952	4327.757	2575.487
24	2817	2969.865	3866.141	2141.913

The correct value of the actual value in the predicted confidence interval is as high as 68%, and the mean square error can be found to be 749.288. We compare the predicted results with the test data, as shown in Figure 6. It can be seen from the image that although the basic trend is consistent, the degree of fitting compared with the SARIMA algorithm is less than ideal, and it is greatly affected by seasonal parameters, which is not very suitable in this project.

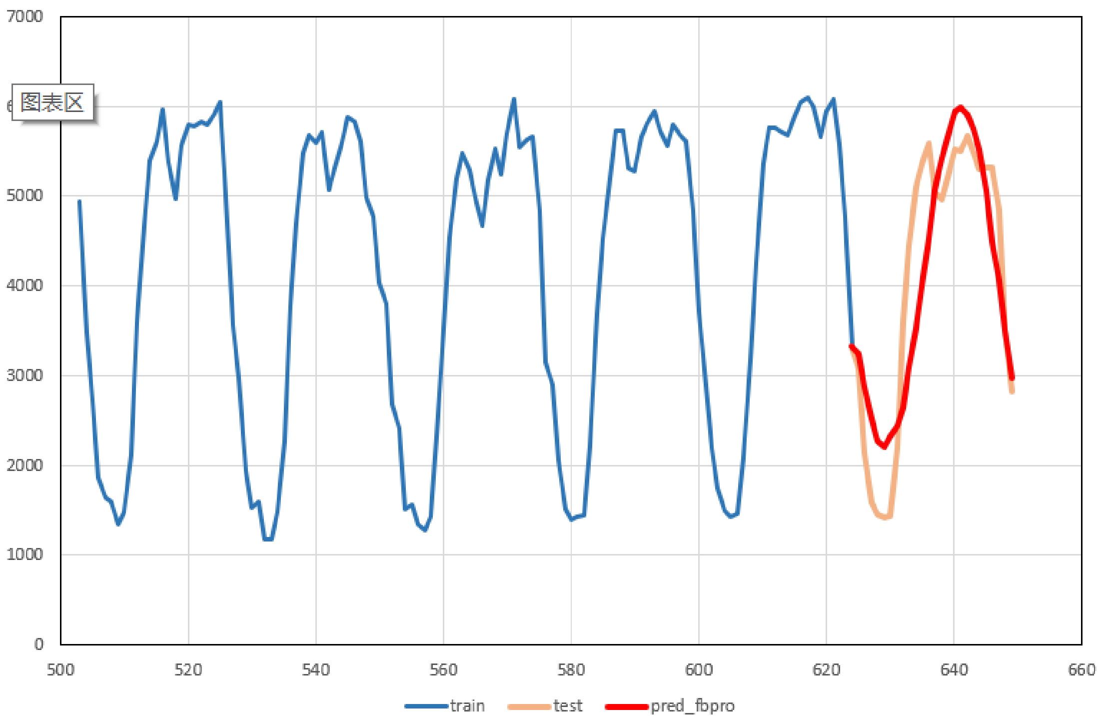

Figure 6. Comparison of prediction data and test data of the prophet model

3.3. *Holt-winter model establishment and prediction effect*

We use the cumulative seasonal Holt-winters algorithm for model training. The predicted results are shown in Figure 7.

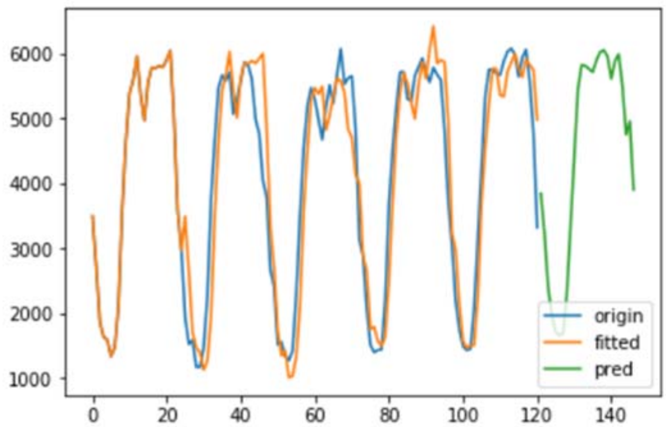

Figure 7. Holt-winters forecast results

We compare the predicted results with the real results, as shown in Figure 8.

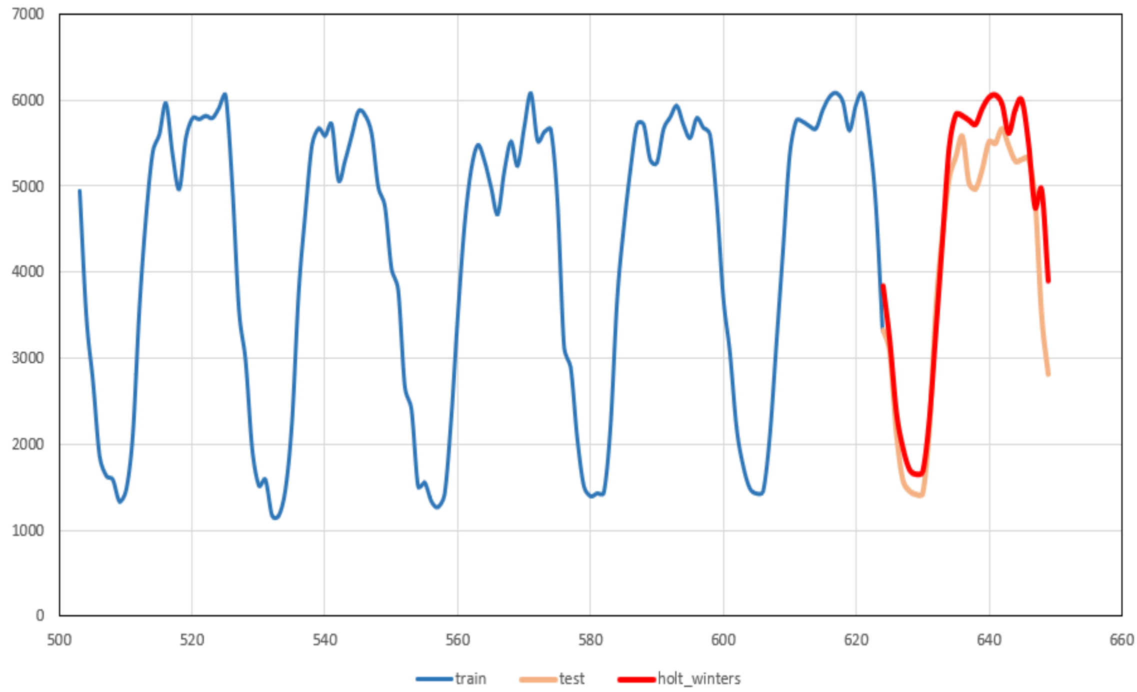

Figure 8. Comparison of prediction data and test data of the Holt-winters model

Holt-winters, as the mainstream predictive algorithm alongside ARIMA, basically follows the general trend. After calculation, its root mean square error reaches 809.89. In this project, the effect is less than that of SARIMA algorithm.

3.4. GM model establishment and prediction effect

The data in the period from 504 to 624 is trained as training data, and the time series prediction result of the GM (1, 1) gray prediction model is shown in FIG. It can be seen that since the GM model does not have periodicity, its prediction effect is very unsatisfactory, and the root mean square error reaches 1710.95, which is completely unsuitable for this project.

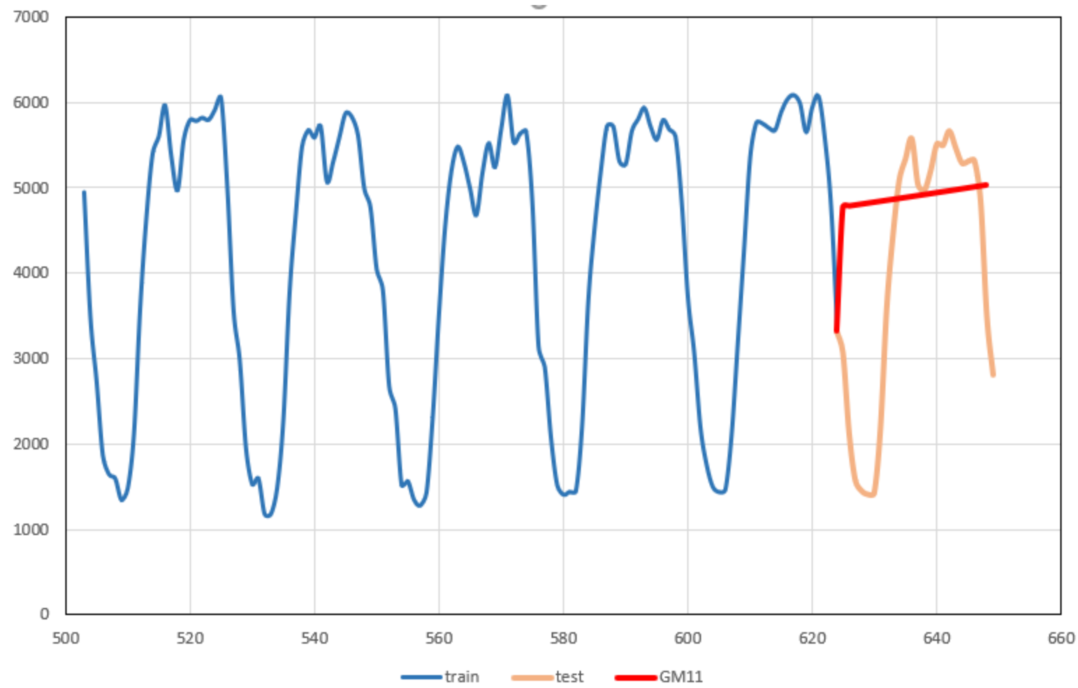

Figure 9 . Comparison of predicted data and test data of GM model

4. Algorithm comparison analysis and conclusion

In this paper, the training data is used for the period from 504 to 624. Four models are used for training and time series prediction, and the prediction results are compared with the real results to obtain the comparison data as shown the following table.

Table 3. Performance of four algorithms

	Root mean square error	Error interval accuracy
SARIMA	258.85	95.80%
prophet	749.288	68%
holt-winter	809.89	-
GM(1,1)	1710.95	-

Combined with the model training results, it can be seen that the GM(1,1) model has a good effect on the processing of the trend time series, while the time series of the processing cycle volatility is often not effective and is completely unsuitable for the needs of the project.

Due to good statistical properties, the SARIMA model and the Holt-Winters seasonal model are widely used. Among them, the SARIMA model has a developmental feature of comprehensive extraction sequences, especially for the sequence of complex interactions between long-term trends, seasonal effects and random fluctuations, while the Holt-Winters seasonal model is the process of decomposing sequences by exponential smoothing. Especially deterministic factors have a clear advantage and do not determine the sequence of weak information. Overall, the SARIMA model has a higher prediction accuracy than the Holt-Winters seasonal model. The reason may be that the SARIMA model can not only comprehensively and fully extract information such as long-term trends, periodicity and random fluctuations of the sequence, but also can determine the exact relationship between these factors, while the Holt-Winters seasonal model can only extract the certainty of the sequence. Information is wasted because of random fluctuation information.

Facebook's open source Prophet Prediction algorithm tends to build models based on past data, with reasonable logic behind it. It is basically a library for building time series data prediction models, rather than using traditional modeling methods such as ARIMA, which are fitted addition regression models. It can also be seen that the accuracy of the prediction effect is higher, but there is still a certain gap compared to the SARIMA algorithm.

In summary, the project will select SARIMA as the basic model for modeling prediction.

References

[1] Wang Fang, Gu Wei, Yuan Xiaodong, Zhou Gan, Wu Zhi and Lin Deqing. Power quality management platform for smart grid [J]. Electric Power Automation Equipment, 2012, 134-139.
[2] Xing Jia-lei, Yang Hong-geng, Chen Wen-bo, Zhao Rui. Open power network model management system under smart grid infrastructure [J]. Power System Protection and Control, 2010, 227 - 232.
[3] Fang X, Misra S, Xue G, et al. Smart grid-The new and improved power grid: A survey[J]. IEEE communications surveys & tutorials, 2012, 14 (4): 944 - 980.
[4] Fadaeenejad M, Saberian A M, Fadaee M, et al. The present and future of smart power grid in developing countries[J]. Renewable and Sustainable Energy Reviews, 2014, 29: 828-834.
[5] Tuballa M L, Abundo M L. A review of the development of Smart Grid technologies[J]. Renewable and Sustainable Energy Reviews, 2016, 59: 710-725.
[6] Box G E P, Jenkins G M, Reinsel G C, et al. Time series analysis: forecasting and control[M]. John Wiley & Sons, 2015.
[7] Sean J. Taylor, Ben Letham, Prophet: forecasting at scale[N\OL]. Facebook research, 2017-02-23. https://research.fb.com/prophet-forecasting-at-scale/.
[8] Chatfield, Chris. "The holt-winters forecasting procedure." Applied Statistics (1978): 264 - 279.
[9] Liu, Sifeng, and Yi Lin. Grey information: theory and practical applications. Springer Science & Business Media, 2006.