Abnormal Value Treatment and Seasonable Adjustment Method for Medium and Long-term Load Data

Xuxia Li[1], Yao Wang[2], Yingying Hu[3], Jia Li[4], Xing Tong[5], Renhui Liu

[1]Planning and Review Center, Economic and Electrical Research Institute of Shanxi Electrical Power Company of SGCC, TaiYuan, ShanXi, 030002, China

[2]Planning and Review Center, Economic and Electrical Research Institute of Shanxi Electrical Power Company of SGCC, TaiYuan, ShanXi, 030002, China

[3]Planning and Review Center, Economic and Electrical Research Institute of Shanxi Electrical Power Company of SGCC, TaiYuan, ShanXi, 030002, China

[4]Planning and Review Center, Economic and Electrical Research Institute of Shanxi Electrical Power Company of SGCC, TaiYuan, ShanXi, 030002, China

[5]R & D department, Shenzhen Orange AI Technology Co., Ltd., ShenZhen, GuangDong, 518000, China

[6]R & D department, Shenzhen Orange AI Technology Co., Ltd., ShenZhen, GuangDong, 518000, China

*Corresponding author's e-mail: 779656332@qq.com

Abstract. Since most of the data used for power demand early warning, forecasting and analysis is the original power data collected directly from the power system, the data cannot be directly applied to the specific analysis because of having two problems. First of all, because various errors will occur in the power data collecting and transmitting processes, some random factors may cause the power consumption data to fluctuate drastically in a short period. These errors or problem data that do not conform to the overall change rule of the power sequence may lead to wrong analysis results. Secondly, because the electricity sequence has obvious seasonal characteristics, and the inherent variation of the load sequence is often obscured by the seasonal variation factors, using the original data for analysis directly often fails to discover the inherent regularity of the electricity consumption data. Based on these two reasons, the original power data should be conducted with abnormal value processing and seasonal adjustment before the power demand early warning, forecasting and analysis.

1. Overview

Since most of the data used for power demand early warning, forecasting and analysis is the original power data collected directly from the power system [1-3], the data cannot be directly applied to the specific analysis because of having two problems. First of all, because various errors will occur in the power data collecting and transmitting processes, some random factors may cause the power consumption data to fluctuate drastically in a short period. These errors or problem data that do not conform to the overall change rule of the power sequence may lead to wrong analysis results. Secondly, because the electricity sequence has obvious seasonal characteristics, and the inherent variation of the load sequence is often obscured by the seasonal variation factors, using the original data for analysis

directly often fails to discover the inherent regularity of the electricity consumption data. Based on these two reasons, the original power data should be conducted with abnormal value processing and seasonal adjustment before the power demand early warning, forecasting and analysis.

This paper illustrates the identification and correction methods of four common anomalous data patterns and explains with specific algorithms in Section 2. Section 3 introduces several common seasonal adjustment methods, and analyzes the principle and method of X-12-ARIMA seasonal adjustment in details. Section 4 summarizes the whole paper.

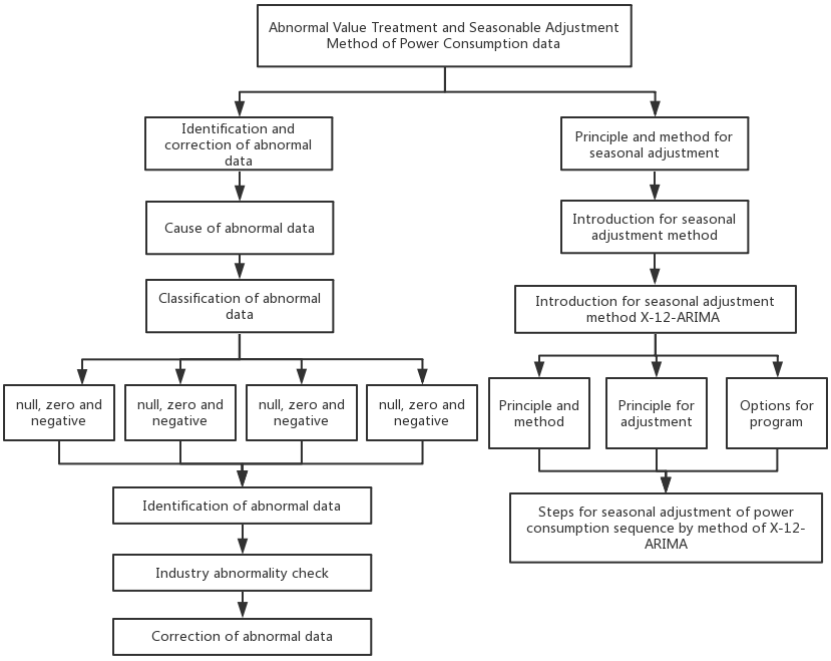

Figure 1. Content of this paper

2. Identification and correction of outliers

2.1. Cause and Classification of Outliers

Outlier data is a widely used concept. In order to avoid the confusion of concepts in classification, we first give the definition of outlier data from the perspective of power demand warning and prediction.

Definition 1: From the perspective of early warning and prediction, if some data do not satisfy the general rule of the load curve, they will mislead the early warning and prediction results, then these data are called outlier data points.

Under this definition, not only the measurement and data transmission errors lead to the generation of abnormal data points, but also the abnormal fluctuation of the user-side load is also considered as an abnormal data point.

Abnormal data points in power demand early warning and forecasting are usually caused by the three aspects of data collection errors, distribution power fluctuations, and statistical system nuisance.

2.1.1. Data collection error

All abnormal data caused by secondary-side faults, including measurement failures, communication failures, and data processing system errors, are data collection errors. The abnormal data generated at this stage accounts for the highest proportion, and the interference to the early warning and prediction is also the largest, but the main mode of the abnormal data is also the most easily identified. The main modes are:

1. Zero data points and negative data points. The zero data point and the negative data point are outlier data with the highest proportion in the system, which is characterized by the data record of the point being null, zero or negative.

2. Similar phenomenon. A similar phenomenon refers to the situation that the electricity consumption of the whole society and all industries are the same at two different time points, which is mostly caused by the error of the data processing system.

2.1.2. Distribution power fluctuations

Abnormal data on power consumption may be caused by occasional irregular electricity use of the user side. For example, if a large-scale cultural performance is held in a few days in a certain place, the electricity consumption of the place will increase irregularly this month. The abnormal data pattern caused by this reason is: Outlier data points. Outliers are a few isolated points whose values differ greatly from other points. The reason for this phenomenon occurs is the electricity consumption behavior that causes it to happen is occasional. Once this accidental power consumption behavior ends, the anomaly ends immediately.

2.1.3. Statistical system nuisance

The data used in this paper refers to the industrial electricity consumption data in provincial administrative units. At this stage, there are significant differences in the statistical system of electricity consumption in different industries within the provincial administrative units, mainly reflected in the statistical starting date of electricity consumption of different industries is different. Some industries will change the statistical caliber in the process of statistics, which produces abnormal data in the other two modes:

1. Null data points. Industries with a late statistical start date are null at the point in time before the statistical start date.

2. Step phenomenon. The step phenomenon means that the electricity consumption is maintained at a certain level in the initial stage, and when it reaches a certain time point, it mutates to a certain value which is larger than the previous one, and thereafter maintains the value level up and down.

Combining the various anomaly data caused by the above three reasons, the abnormal data often appearing in the early warning and prediction of power demand can be summarized into four modes, namely: null, zero and negative data points, outlier data points, similar phenomena and step phenomena.

2.2. Identification of Abnormal Data

The identification methods of abnormal data in four different modes are different: null, zero and negative data points. The similar phenomenon can be directly judged according to the numerical characteristics of the data, while the outlier data points and step phenomena need to be Identification can only be made after the statistical characteristics of normal data are known. Judging from the range of data that needs to be judged, a similar phenomenon should be judged from the overall power consumption of various industries, while the judgment of other abnormal values only needs to pay attention to the electricity data of the industry.

In the identification process, corresponding to the electricity consumption matrix E, an abnormal data identification matrix I is generated, the size of which is the same as E. Where the first-dimension index of E is the industry serial number, the second-dimension index is the time serial number. The first-dimension indicator set of E is recorded as A, B is the second dimension indicator set of E. $\forall i \in A, j \in B$. When $E(i,j)$ is the correct data, $\mathrm{I}(i,j)$ is zero; when $E(i,j)$ is abnormal data, $\mathrm{I}(i,j)$ is identified as a non-zero value, the value size is related to the corresponding abnormal data type.

2.2.1. Null, zero and negative data points

$\forall i \in A, j \in B$, If $E(i,j)$ is null, zero, or negative, $\mathrm{I}(i,j)$ is marked as one, indicating a null, zero, or negative data point.

2.2.2. Similar phenomenon

$\forall j_1 \in B$, if $\exists j_2 > j_1 \in B, \forall i \in A$ and $E(i, j_1) = E(i, j_2)$, then , $I(i, j_2)$ will be marked as 2 ($i \in A$), indicating the similar value data point.

2.2.3. Outlier data points

To identify outliers, first identify the anomaly increment value $\forall i \in A, j \in B$. It is assumed that:

$$\Delta E_i(j) = E(i, j+1) - E(i, j) \quad (1)$$

It represents the increment between two points in the power usage sequence. Since the various outliers in the original data have not been corrected in the identification phase, the statistics of the power usage sequence cannot be recognized. In order to identify the abnormal increment without any subjective or a priori information, it is necessary to use the statistical law of the sequence in the case where the probability distribution is unknown, and Chebyshev inequality provides a powerful tool for this work. According to Chebyshev's inequality [4], for arbitrary distribution of random variables, both meet.

$$P(|X - E(X)| < k\, D(X)) > \frac{1}{k^2} \quad (2)$$

Wherein, X represents a random variable, E(X) represents the mathematical expectation of the random variable X, and D is the variance of the random variable. k is a constant indicating that the random variable leaves the desired range. Taking k = 5, the probability that X falls within the interval $[E(X) - 5D(X), E(X) + 5D(X)]$ is about 96%. Then under the same standard, the probability that the random variable $\Delta E_i(j)$ falls within the interval $[E(\Delta E_i(j)) - 5D(\Delta E_i(j)), E(\Delta E_i(j)) + 5D(\Delta E_i(j))]$ is 96%, but since the actual power consumption increment is a random variable showing a certain normality, the probability that the load increment falls within the above interval should be More than 96%. Based on this, we believe that if the increment falls outside this interval, then this increment is an abnormal increment.

If the $\Delta E_i(j)$ and $\Delta E_i(j-1)$ is an abnormal increment, and both fall on both sides of the above interval, then $\mathrm{E}(i, j)$ is considered to be an outlier point, and $I(i, j)$ is 3, indicating the outlier data point.

2.2.4. Step phenomenon

The identification of the step phenomenon also needs to identify the anomaly increment according to the anomaly increment method described in former section. If for industry *i*, there is only an abnormal increment $\Delta E_i(j)$, then the industry *i* is considered to have a step phenomenon, set $I(i, j_k)$ as 4, $j_k = 1, 2, \ldots, j$ which means that all points in industry *i* are step outliers.

It should be noted that when a data point has been confirmed as an abnormal data point of a certain mode, the data point is no longer involved in the identification of other abnormal data patterns.

2.3. Correction of Abnormal Data

Before the correction of abnormal data, the industry data anomaly check is firstly performed. Set: $\lambda(i) = n_1(i)/n(i)$ as the industry anomaly degree, wherein $n_1(i)$ is the number of power consumption anomaly point in the industry *i*, the total number of electricity consumption data points for the industry I is $n(i)$. If it is $\lambda(i) < \lambda_{lim}$, it is considered that the abnormal degree of the industry *i* is within the normal range after the abnormality check, and the correction of abnormal point can be carried out. Otherwise, the industry *i* is considered to be an abnormal data industry, and the impact of industry *i* is not considered in the early warning and prediction analysis. Set λ_{lim} in this article.

When the industry electricity consumption data is checked by the abnormality, the linear interpolation correction is performed on the abnormal data of various modes uniformly, that is $\forall i \in A, j \in B$, if $I(i, j) \neq 0$, that is the linear interpolation correction is performed on the $E(i, j)$ data points.

2.4. Examples of Identification and Correction Results

The monthly electricity consumption data of various industries in A Province from 1999 to 2008 were processed by the above identification and correction method, and the following results were obtained.

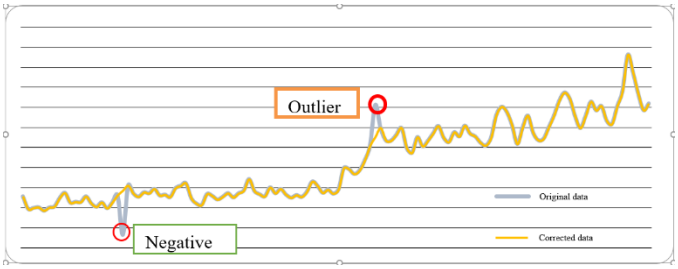

Figure 2. The correction results of negative value and the outlier of the monthly electricity consumption of the real estate industry in A Province from 1999 to 2008

The above anomaly data identification and correction method provides a reliable guarantee for early warning and prediction analysis of subsequent power demand. According to the prediction method, the monthly total social electricity consumption in A Province in 2008 was virtually predicted. Before and after the correction of abnormal data, the prediction accuracy was compared as follows:

Table 1. Comparison of relative error of monthly virtual power consumption before and after correction of abnormal data

Before Correction of abnormal data							Average relative error：**4.32%**					
Month	1	2	3	4	5	6	7	8	9	10	11	12
Error	6.9%	4.7%	2.8%	4.2%	2.4%	6.3%	2.5%	3.2%	2.4%	4.9%	3.4%	5.1%
After Before Correction of abnormal data							Average relative error： **2.30%**					
Month	1	2	3	4	5	6	7	8	9	10	11	12
Error	2.5%	3.1%	0.8%	1.9%	1.9%	2.4%	0.0%	3.2%	1.3%	3.5%	2.8%	1.5%

3. Principles and methods of seasonal adjustment of power consumption series

3.1. Introduction to Time Series Seasonal Adjustment Methods

The fluctuation of time series such as electricity consumption and above scale industrial added value has obvious periodic law with time. This phenomenon is called the seasonal effect. The seasonal adjustment of time series refers to the decomposition of time series with seasonal effects into components with obvious periodic changes with time and components that are basically independent of time changes according to certain mathematical methods.

In the usual seasonal adjustment algorithm, the monthly or quarterly time series data is considered to be composed of four components: long-term trend component (T), fluctuation cycle component (C), seasonal component (S) and irregular component (I). The long-term trend component represents the long-term trend characteristics of the time series. The fluctuation cycle component is a kind of boom change in a cycle of several years. In the study of time series, they reflect the basic changes in time series. The seasonal component is a cyclical change that occurs repeatedly every year, reflecting the cycle effect of 12 months or 4 quarters due to factors such as temperature, rainfall, and holidays. Irregular components, also known as random factors, residual fluctuations or noise, can be changed irregularly. This component is caused by accidental events such as strikes, accidents, earthquakes, bad weather, wars [5-6], etc.

The existing seasonal adjustment methods mainly include: moving average ratio method, TRAMO/SEATS method, X-11 method, X-12-ARIMA method, BV4 method and structural time series model[7-8] The literature [9]~[13] applied seasonal adjustment to the analysis of different macroeconomic indicators, and achieved satisfactory results and meaningful conclusions.

3.2. Introduction to the Principle and Method of X-12-ARIMA Seasonal Adjustment

In 1965, the famous US Census Bureau X-11 season adjustment program came out. It originated from the 1954 US Census Bureau's seasonal adjustment program "Model I". After more than ten years of development, it experienced 12 experimental versions of "Model II" and eventually formed X-11. The

Census Bureau X-12-ARIMA seasonal adjustment method is developed and based on the X-11 method and includes all the latest X-11-ARIMA and X-11-ARIMA/88 features, and has significant improvements in design of filters in the seasons and trends, results stability diagnostics, and ARIMA modeling capabilities and batch processing. Due to the excellent nature of the X-12-ARIMA seasonal adjustment, it has gradually become the default seasonal adjustment standard [14] adopted by statistical agencies around the world. The principle and method of seasonal adjustment of X-12-ARIMA will be briefly introduced below. The details of this method can be referred to literature [8] and [14].

The X-12-ARIMA program can be divided into two modules: regARIMA and enhanced X-11. regARIMA is used to preprocess data, including forward and backward continuation of sequences, detection of outliers and a priori adjustments of various effects, etc. The enhanced version X-11 is based on a seasonal adjustment of the moving average, and the final seasonal component, trend-cycle component, and irregular component are determined by three iterations of screening. At the end of the adjustment, X-12-ARIMA also gave a detailed diagnosis of the model, providing the necessary information for improving the model.

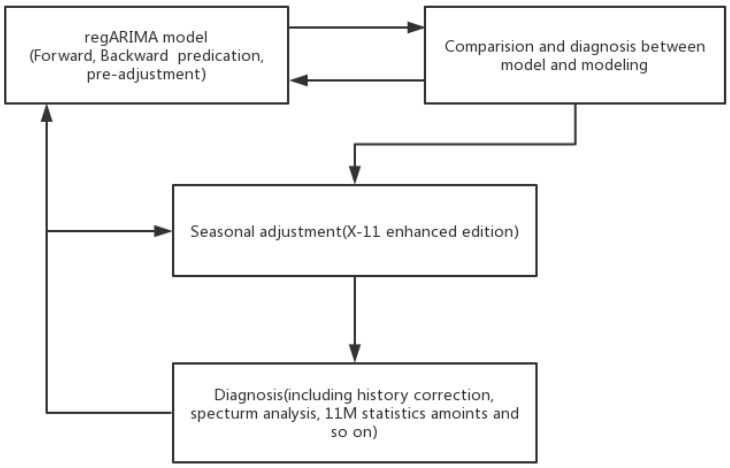

Figure 3. Basic flow chart of X-12-ARIMA seasonal adjustment procedure

Figure 3 is the basic flow of the X-12-ARIMA seasonal adjustment procedure, in which the solid arrow represents the flow of the program, the dashed line represents the actual process that needs to be experienced in the seasonal adjustment, and the best season adjustment of sequence is obtained by "adjustment-diagnosis-re-adjustment".

The pre-adjustment module regARIMA is mainly used to extend the time series, which is called the linear regression model with ARIMA time series error, and is an important innovation for ARIMA time series modelling. This method adds regression variables to the influencing factors such as outliers and calendar effects when establishing the ARIMA time series model, and automatically selects the significant effects and the best ARIMA model. Especially for some sequences with missing or outliers, regARIMA's coefficient estimation and prediction have certain robustness.

The enhanced X-11 module is used to decompose a monthly or quarterly time series into a trend-cycle component (TC), a seasonal component (S) and an irregular component (I). There are four types of decomposition models used:

1. Multiplicative model (M):

$$Y_t = TC_t \times S_t \times I_t \tag{3}$$

2. Additive model (A):

$$Y_t = TC_t + S_t + I_t \tag{4}$$

3. Pseudo additive model (PA):

$$Y_t = TC_t \times (S_t + I_t - 1) \tag{5}$$

4. Logarithmic addition model (LAD)：

$$Y_t = log(TC_t + S_t + I_t) \quad (6)$$

The multiplicative model is applied to sequences that maintain positive values and whose seasonal fluctuations also increase as the sequence level increases. Most macro-season time series apply to the multiplication model. At the core of the enhanced X-11 module is the X-11 computational prototype, which consists of three main phases, with repeated "filtering" of seasonal and trend components to obtain a final estimate of several components.

In terms of model diagnosis, X-12-ARIMA provides X-11-ARIMA's existing diagnostic tables and quality control statistics amount of $M1 \sim M11$. In addition, X-12-ARIMA also provides spectral estimation diagnosis of inspection season and trading day effect, translation interval for seasonal adjustment stability and historical correction diagnosis.

3.3. Principles and Options for Seasonal Adjustment of Power Consumption Series by X-12-ARIMA

X-12-ARIMA offers a variety of options for users in regARIMA modelling, model selection, calendar effect regression, model diagnosis, etc., so that we can optimize the adjustment of target sequence by adjusting these options. However, the flexibility of these options directly leads to the diversity of adjustment results. Therefore, before seasonal adjustment of the electricity consumption sequence, it is necessary to determine the principle of seasonal adjustment of electricity consumption. According to the characteristics of the electricity consumption sequence, the following two adjustment principles are chosen:

1. A priori information is combined with a posteriori information. The X-12-ARIMA algorithm gives many automatic options, such as automatic selection of seasonal and trending filters and automatic detection of many effects. In other words, the program will get some information in the target sequence based on the automatic "learning". From a view of methodological point, if some of the information in the sequence is known, the priori information is added "subjectively" in the seasonal adjustment, and the adjustment result obtained should be better than the "objective" adjustment effect.

2. Consider the impact of the calendar effect on electricity usage. The operating experience of the grid shows that there is a certain drop in electricity consumption during some holidays, and this drop will have a significant calendar effect in the monthly sequence of electricity usage. In the seasonal adjustment, reasonable consideration of the influence of the calendar effect can better discover the changing law of power demand.

According to the above adjustment principle, combined with the specific characteristics of the power consumption time series, the options for seasonal adjustment of the electricity consumption in this paper are determined as follows:

1. Select the multiplicative model, that is $Y_t = TC_t \times S_t \times I_t$;

2. Corresponding to the multiplicative model, logarithmic transformation of the sequence in regARIMA and adjustment links;

3. Eliminate the effects of leap year factors through pre-adjustment;

4. Automatic detection of outliers in the regARIMA session;

5. Add user-defined regression variables to the regARIMA link to estimate the Spring Festival effect;

6. Automatically select the ARIMA model in ARIMA modelling. If multiple models are selected, choose the model with the best prediction expansion effect.

7. When extending the sequence with the regARIMA model, predict the value for the next 24 months.

8. Automatically select seasonal and trend filters in the X-11 section;

9. The first and second limits for the irregularity correction in the X-11 link are 1.5 and 2.5, respectively.

3.4. Steps for Seasonal Adjustment of Power Consumption Sequence by X-12-ARIMA

This article uses the X-12-ARIMA seasonal adjustment program embedded in the EViews software to seasonally adjust the electricity consumption data of various industries.

According to the adjustment options, write the X-12-ARIMA user readme file spring_adjustment.txt as follows:

```
power=0.000000
regression{
    variables=(lpyear)
    user=(spring)
    start=1999.1
    file="c:\EViews\springB5A15.txt"
    usertype=holiday
    print=(all)
    save=(hol otl)
}
automdl{
    file="c:\Program Files\eviews5\x12a.mdl"
    method=best
    identify=all
}

forecast{maxlead=24}
outlier{ }
x11{
    sigmalim = (1.5,2.5)
    save = ( D10 D11 D12 D13)
    savelog = (q,q2,fbl,fd8,msf)
}
```

Figure 4. User Readme file for Seasonal Adjustment of Power Consumption Sequence by X-12-ARIMA

Set the target adjustment sequence be a0. If the path of the above readme file is: "c:\EViews", the seasonal component, the de-seasonal component, the trend cyclic component, the irregular component, the holiday component, and the outlier component of the sequence a0 need to be saved after the seasonal adjustment. The following is entered in the EViews software command area.

That is, the seasonal adjustment of the a0 power consumption sequence is completed. By following this step, the electricity consumption data of all industries is batched to complete the seasonal adjustment of electricity consumption in all industries.

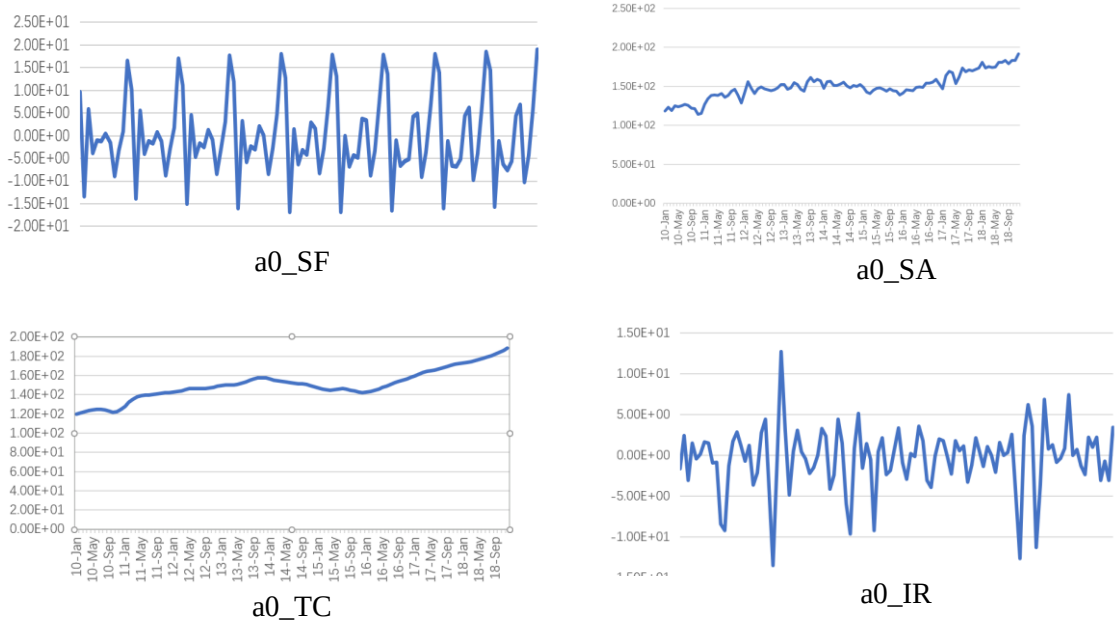

Figure 5. Seasonal adjustment results of monthly electricity consumption in A Province from 2010 to 2018

3.5. Analysis of the Results of Seasonal Adjustment of Power Consumption Series by X-12-ARIMA
Taking the monthly electricity consumption of the whole society in A Province from 2010 to 2018 as an example for the seasonal adjustment, and the following results were obtained.

In Figure 5, a0_SF represents the seasonal factor component of the total social electricity consumption, a0_SA represents the deleted seasonal factor component of the total social electricity consumption, a0_TC represents the cyclical trend component of the total social electricity consumption, and a0_IR represents the irregular component of total social electricity consumption. In the seasonal adjustment diagnosis, the quality control statistics amount are all less than 1, indicating that it is acceptable for seasonal adjustment. It can also be seen from the figure that after seasonal adjustment, the cyclic trend component (TC) is relatively smooth, reflecting the growth trend of power consumption development, which can also indicate that the seasonal adjustment is successful. Since the electricity consumption component (SA) component with deleted seasonal factor only removes the seasonal factors with strong regularity in the original sequence, the information on the electricity consumption of the industry has almost no loss, and the overall regularity has been more obvious. Therefore, in the following papers, the SA component of the power consumption sequence is used for early warning prediction analysis.

4. summary

This paper introduces two data pre-processing tasks that must be performed before conducting power demand early warning and prediction studies. One is the identification and correction of abnormal data, and the other is the seasonal adjustment of the electricity consumption sequence. The identification and correction of abnormal data is attributed to the attribute of abnormal electricity consumption data used in this paper, and the identification and correction methods are separately studied according to different abnormal data patterns. The seasonal adjustment part of the electricity consumption sequence mainly introduced the classic X-12_ARIMA seasonal adjustment method in econometric analysis is, and this method is used to process the industry electricity consumption data, and the industry electricity consumption sequence is decomposed into seasonal component, trend cyclic component and irregular component.

This paper is the basis of the subsequent papers. The processing of abnormal data provides basic data protection for power demand early warning and prediction analysis, while the seasonal adjustment removes the interference of seasonal factors for this analysis, so that the electricity consumption law can be more obvious. Two data pre-processing tasks are necessary for the subsequent analysis and research.

Acknowledgement

Science and technology support project of State Grid Corporation of China, Contract number: SGSXJY00PSJS190003

References

[1] Findley, Monsell, Bell, Otto, Chen. (1998)New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program[J]. Journal of Business and Economic Statistics, vol.02, pp.127~152.
[2] Mbamalu G A N, EI-Hawary M E. (1993)Load Forecasting via Suboptimal Seasonal Autoregressive Models and Iteratively Reweighted Least Squares Estimation[J]. IEEE Tram on Power Systems, 8(1):343~348.
[3] Yang ZM, Li, W, Yan ZM. (2019)The Relationship between Temperature Change and Electric Power Demand —Evidence from China 's Urban Panel Data from 2000 to 2014, Journal of Beijing Institute of Technology (Social Sciences Edition). vol. 21(5), pp. 44-55
[4] Xiao TS. (2000)Stochastic mathematics.Beijing, Higher Education Press, in press.
[5] Chen JM. (2005)Theory and practice of macroeconomic statistical analysis, Economic Science Press, in press.

[6] Qiu D. (2001)National economic statistics, Northeast University of Finance and Economics Press, in press.
[7] Wang Y, Xu, Y, Wang J. (2005)Economic statistics, Mechanical Industry Press, in press.
[8] Peng XP. (2007)Analysis and design of medium and long term load forecasting and early warning system. North China Electric Power University.
[9] Li JL, Zhao ZQ. (2006)Management statistics, Tsinghua university press, in press.
[10] Jiang QY, Xin WX, Xie JX, Yang DH. (2005)College Mathematics Experiment, Tsinghua university press, in press.
[11] Ji, GL, (2005)Discussion and application of the construction method of the leading index system of China's macro economy, Jilin University.
[12] Fan L. (2006)A study on the leading index system of macro economy in Fujian Province. Fuzhou University.
[13] Ding WB. (2004)A study on the leading economic index of Beijing, Journal of Shanxi University of Finance and Economics, vol. 26(04), pp.38-44.
[14] Investigation and Statistics Department of the people's Bank of China, (2006)Seasonal adjustment of time series X-12-ARIMA -- principle and method, China Financial Press, in press.