Modeling, analysis and prediction of air particulate data

Abstract

The study of air quality is very important for people's health. The report uses airborne particulate data from 1970 to 1980 to model and forecast. Two suitable models ARIMA (0,1,1) and ARIMA (1,1,1) were established by analyzing the ACF and PACF graphs of the data. Subsequent model hypothesis checks showed that ARIMA (1,1,1) was a more suitable model, and subsequent air particulate data were predicted on this basis. The data themselves were also subjected to spectral analysis. The conclusion is that ARIMA (1,1,1) is a more suitable model with parameter estimates ma1 = -0.5162 and ar1 = -0.2444.

Keywords: Air quality, Time series, ARIMA model, Spectral analysis.

Introduction

Air quality is a point of great concern in people's lives, because air quality has a great impact on people's life and health. Especially in the current COVID-19 situation, most people choose to wear masks to prevent transmission of the virus. Good air quality can make people's lives a lot more comfortable, so there are a lot of studies on air quality and all kinds of diseases, both physical and psychological. Therefore, on the basis of previous studies, we can use statistical methods to study air quality data. There was a study 50 years ago that looked at the association between air pollution and mortality in Los Angeles, which recorded weekly particulate counts for the 1970s and 1980s. The data can be accessed through the R package

“astsa” and “data(part)”. Using this data, we will conduct modeling research on it and predict future data according to the established model.

Statistical Methods

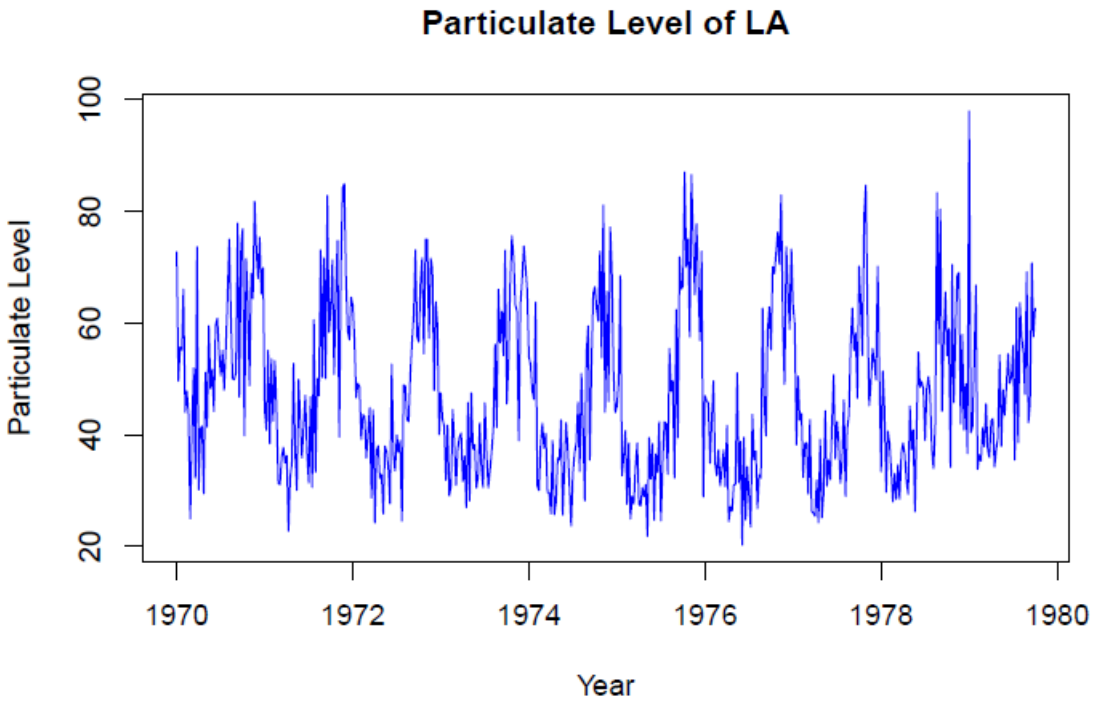

Fig.1 The plot of weekly data of particulate level of LA from 1970 to 1980.

Figure 1 presents the weekly data of particulate level of LA from 1970 to 1980. It can be seen from the figure that the data of the size of air particles presents an obvious periodicity. The data of each year has approximately equal mean and variance, but alternately presents a higher mean (about 70) and a lower mean (about 40) year by year. Therefore the raw data can not be considered as stationary as periodic significant different means exist. However, overall the mean data seems to stay the same. Therefore, differenced data was used to do further research.

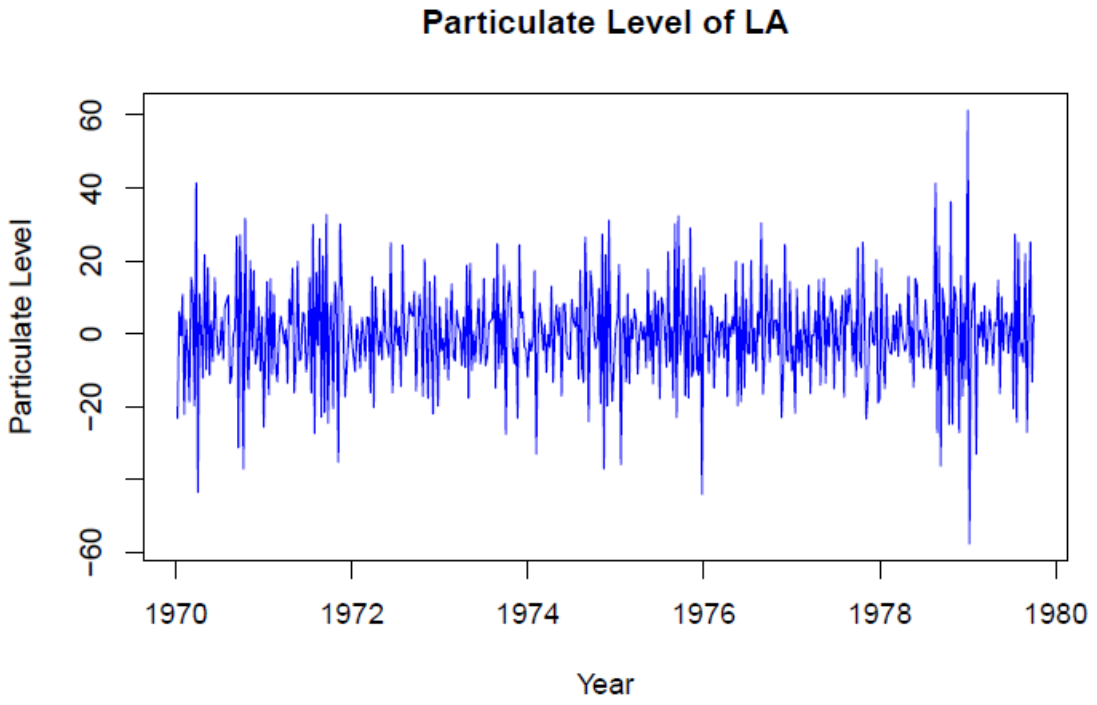

Fig.2 The plot of differenced weekly data of particulate level of LA from 1970 to 1980.

Figure 2 is obtained by subtracting the data with the one before it, which is:

$$\delta X_t = X_t - X_{t-1}$$

From Figure 2 we can roughly assume that the data have a constant mean and the same variance throughout. Therefore this differenced data can be considered as stationary. Furthermore, ACF plots could be used to build models. According to Figure 3, the ACF is cutting off at lag 1, while the PACF seems tailing off. These are the signs of model ARIMA(0, 1, 1). However, we could propose another model

ARIMA(1, 1, 1) to avoid bias.

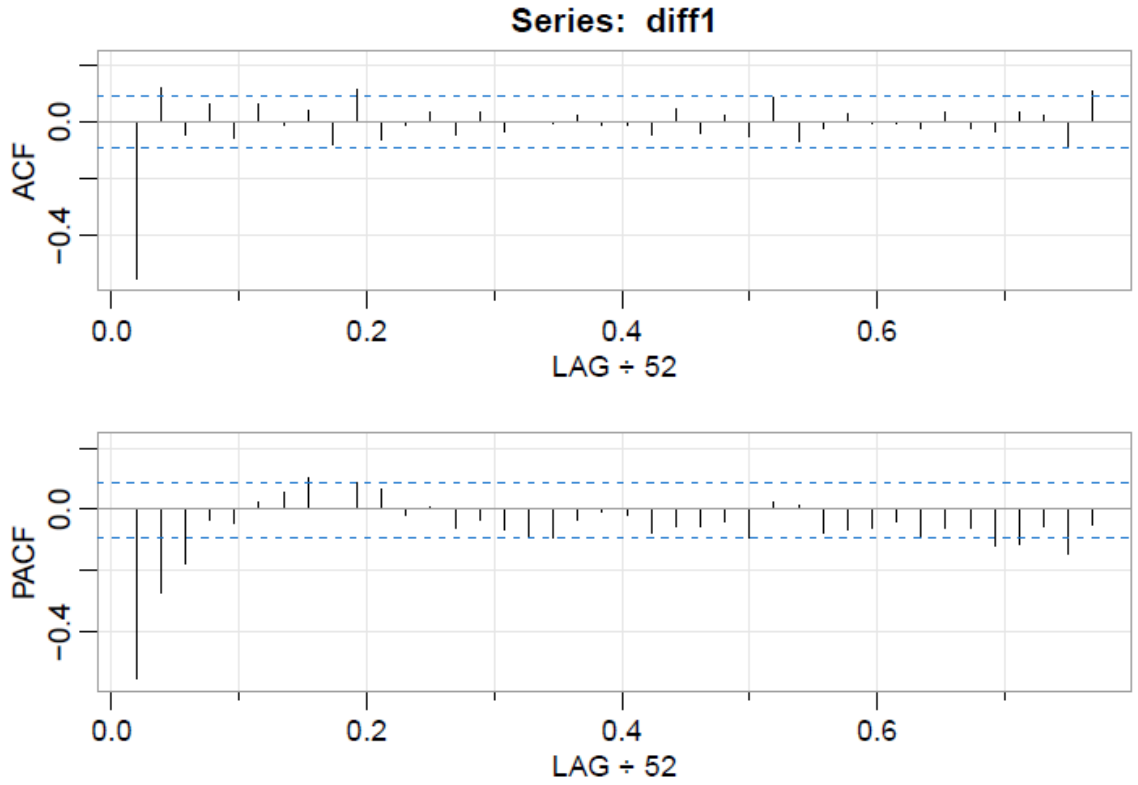

Fig.3 The ACF and PACF plot of the differenced data. Noting that the unit of lag is a year whereas the data is in weeks, therefore the lag is divided by 52.

Results

Now we have two models, ARIMA(0, 1, 1) and ARIMA(1, 1, 1). Further steps are needed to determine which one is the more appropriate model. First, for the model ARIMA(0, 1, 1):

	Estimate	Stand Error	P value
Ma1	-0.6410	0.0279	0.0000
Constant	-0.0005	0.1826	0.9979
AIC	7.719987		
AICc	7.720034		
BIC	7.745007		

Tab.1 The r calculation of estimates of model ARIMA(0,1,1).

We see from the above output for ARIMA(0, 1, 1) model, the p-value for the parameter estimate is very small, less than 0.0001, smaller than α = 0.05, showing that ARIMA(0,1,1) model parameters are statistically significant except the constant. Then for the model ARIMA(1, 1, 1):

	Estimate	Stand Error	P value
Ma1	-0.5162	0.0489	0.0000
Ar1	-0.2444	0.0594	0.0000
Constant	-0.0004	0.1945	0.9982
AIC	7.693015		
AICc	7.693109		
BIC	7.726376		

Tab.2 The r calculation of estimates of model ARIMA(1,1,1).

We see from the above output for ARIMA(1, 1, 1) model, the p-value for the parameter estimates are very small, both are less than 0.0001, smaller than α = 0.05, showing that ARIMA(1,1,1) model parameters are statistically significant except the constant. Both models seem reasonable according to their parameter estimates, therefore model diagnostics are needed. According to the diagnostic plots for ARIMA(0, 1, 1) and ARIMA(1, 1, 1). The standardized residuals show both have no obvious patterns. There are few outliers that exceed 2 standard deviations from the mean. The ACF Residuals plots show some significant spikes at lag 2, 4, 6 and 10 in ARIMA(0, 1, 1) model and at lag 6 in ARIMA(1,1,1) model, but they are not quite enough to be significant at the 5% level. Therefore the randomness assumption of both models has not been rejected. The residuals' normal Q-Q plots show that the assumption of normality is reasonable, except for the possible outliers. The residuals in the ARIMA(0, 1, 1) model are dependent. The p-values for Ljung-Box statistics for model ARIMA(1, 1,1) are better,some lags are above the significant level indicating

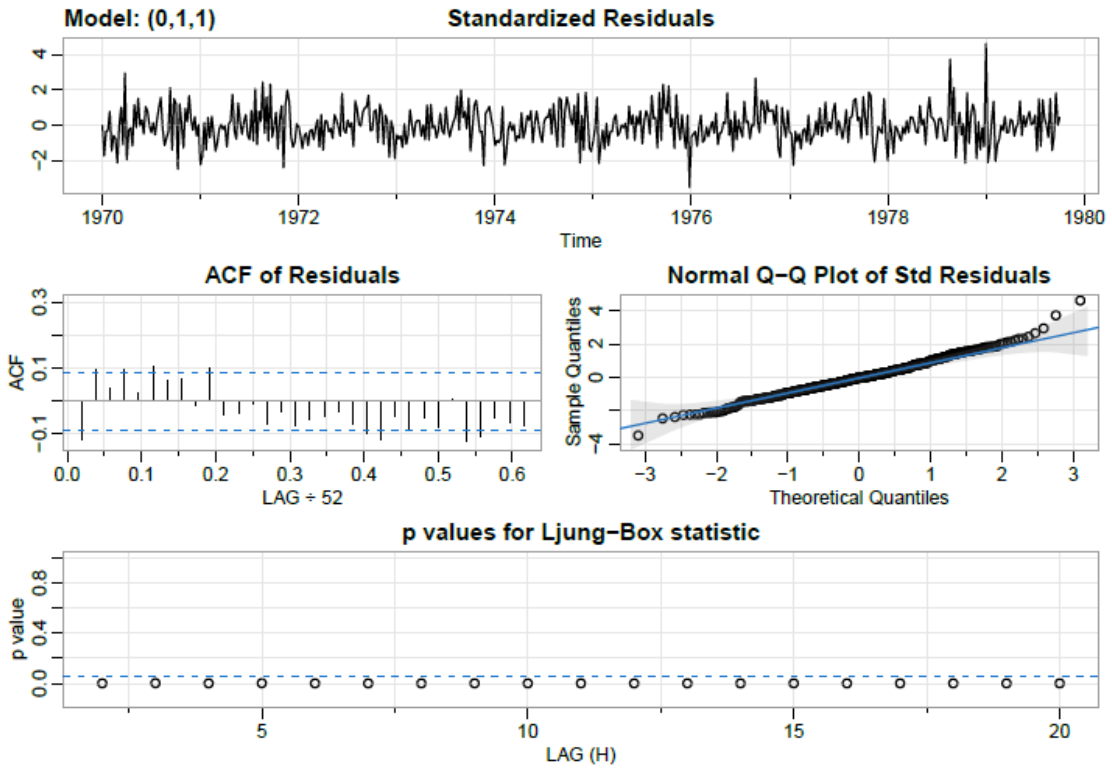

Fig.4 The diagnostics plots for model ARIMA(0,1,1)

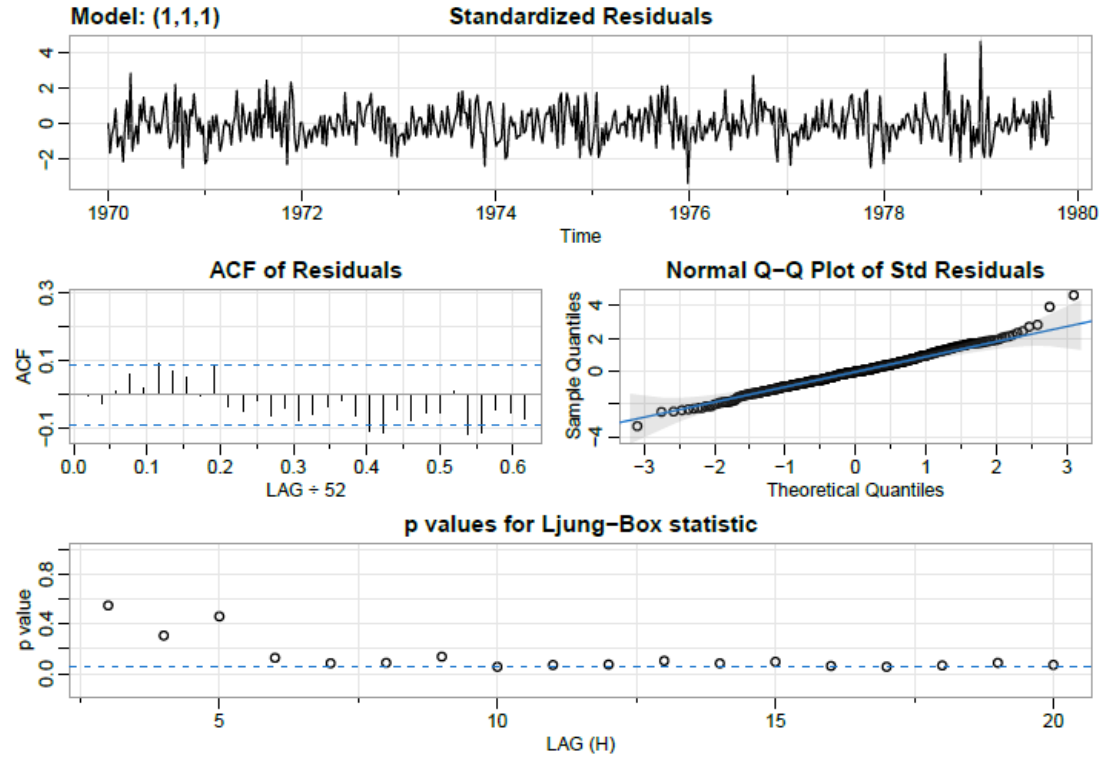

Fig.5 The diagnostics plots for model ARIMA(1, 1, 1).

that they are independent. Overall, the ARIMA(0, 1, 1) and ARIMA(1, 1, 1) models' residuals seem iid and normal with mean zero and constant variance. However, the ARIMA(0, 1, 1) model fails the Ljung-Box test. In addition, the AIC, AICc, and BIC are all smaller for the ARIMA(1, 1, 1) model. Therefore model ARIMA(1, 1, 1) is selected for prediction. Following are the plot with prediction and the details of prediction data.

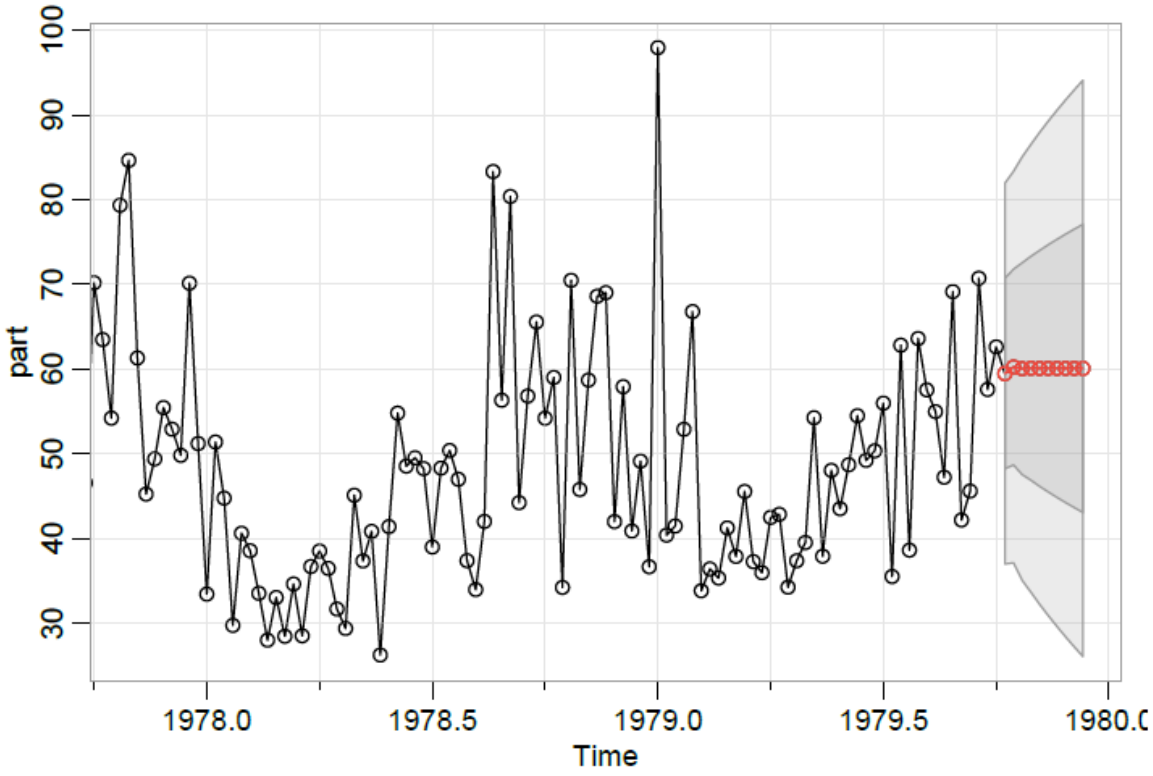

Fig.6 The plot of particulate level with 10 more predictions.

Prediction	Lower bound of 90%CI	Upper bound of 90%CI
59.45651	40.97575	77.93727
60.22659	41.22333	79.22985
60.03786	39.47314	80.60258
60.08344	38.35325	81.81362
60.07175	37.17162	82.97189

60.07406	36.07594	84.07218
60.07295	35.02142	85.12448
60.07268	34.01110	86.13425
60.07220	33.03809	87.10630
60.07177	32.09898	88.04456

Tab.3 The 10 predictions and its 90%CI based on model ARIMA(1,1,1).

Furthermore, spectrum analysis are done to the data and here is the result:

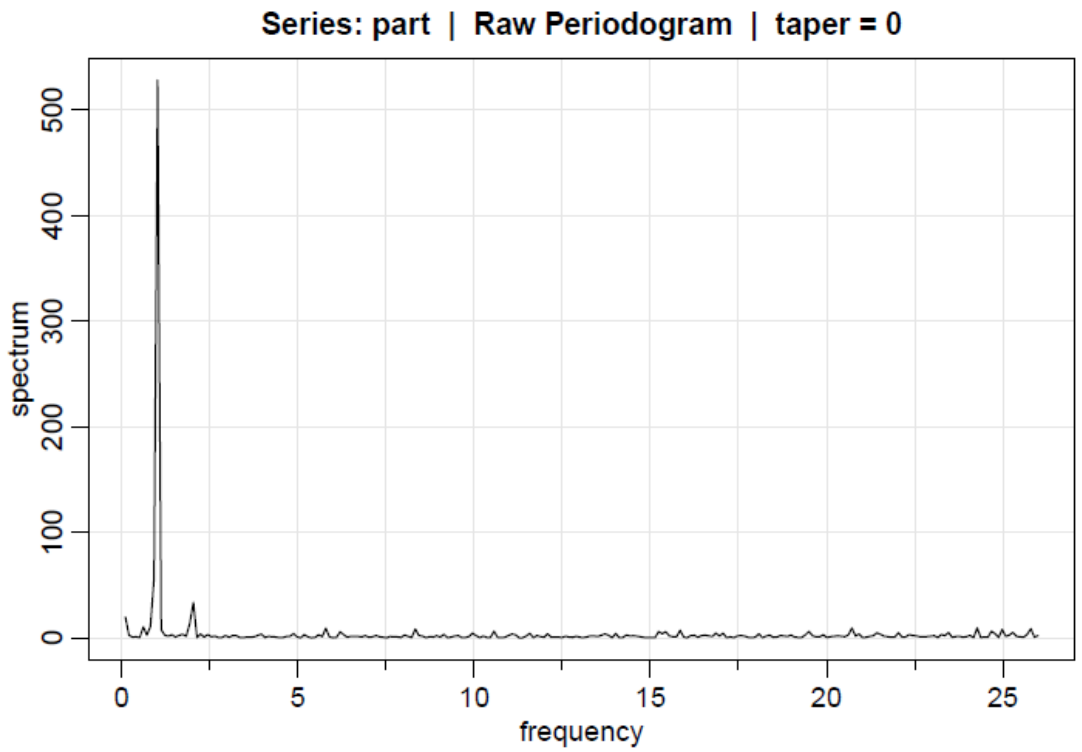

Fig.7 The spectrum analysis plot of the data.

Dominant Frequency	Spectrum	80% Lower bound	80% Upper bound
1.0156	528.5134	229.53045	5016.2378
0.9141	56.5113	24.54255	536.3613
2.0312	33.5363	14.56463	318.3005

Tab.4 The data of spectrum analysis.

According to Table 4, the first peak is significant since the periodogram ordinate is 528.5134, which is approximately larger than the 80%CI upper bound of the second

and third peak. In contrast, the second and third peak are not significant since their periodogram ordinates lie in the 80%CI of each other.

Discussion

After the differenced method used, the data seem better at consistent mean and variance compared to the raw data. However, there are still obvious periodic patterns,so probably other transformations like detrending the data first is better to create a stationary time series and establish a better model.