Time Series Student Project
Name: Lin, Tsung-Jen
Course: Summer 2016 Time Series

Title: Building the Air Pollution Forecasting Model

I. Introduction

As air pollution became a serious problem to human health and an important factor to global environment. For this issue, we are interest in the daily changes of the PM_{10} (Particulate matter < 10 micrometers in size) concentrations and would like to create a forecast model on it.
Data is collected in Chao-Chow Town and contains 1460 daily observations from September 1999 to August 2003. Also, data for forecasting includes 61 daily observations from September to October in 2003.

Time series plot reveals there is seasonal cycle with a period of one year. Hence, the main problem in our analysis is to estimate the seasonal effect. After resolving the effect, model could be built and started to forecast.

In the first stage, we adopt two methods, include Small Trend Method and Ordinary Least Square Method, to estimate the seasonal parameters. But residuals still with small variation in the model and it might be other seasonal effects. Therefore, we try to use spectral analysis to resolve it and outcomes shows the same conclusion. But we obtain a new model from previous by including the half-year-period seasonal parameter. After comparing the validity with these models, the model constructed by spectral analysis has minimal MSE and could be the best model for forecasting.

II. Data Transformation

Based on the time series plot, we found original data may exist the heteroscedasticity and could not be used. Hence, we try to apply logarithmic transformation to regenerate these data and hope to reduce the variation. After the data transformation, we obtain a stable data and further study are based on these data.

Figure 2-1and 2-2 reveal that there has no apparent trend, but still with strong seasonal effect. Therefore, we need to remove the seasonality when building the model.

Figure 2-1: The time series plot of the original data

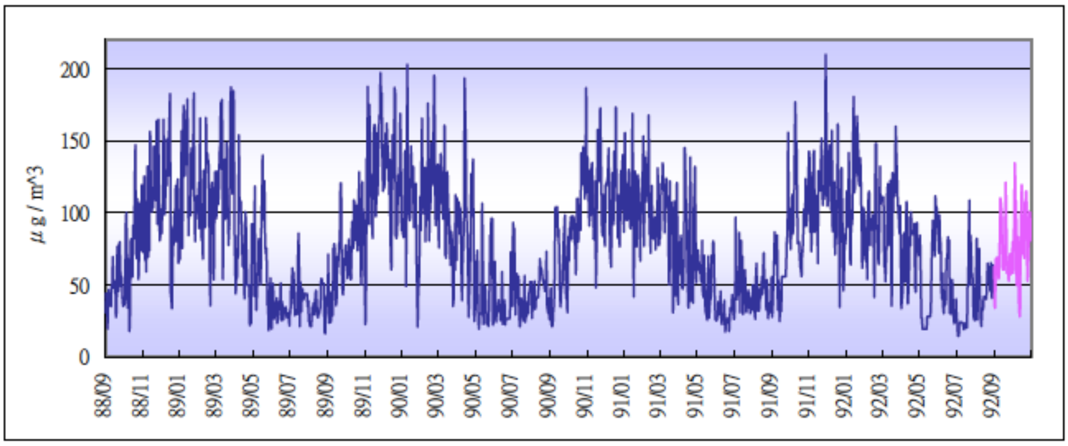
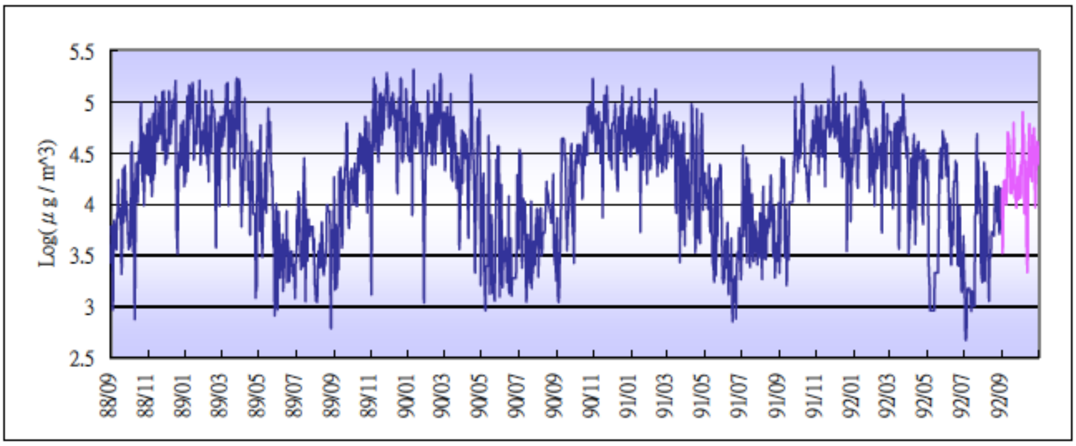


Figure 2-2: The time series plot of the transformed data



III. The elimination of seasonality

Method 1: Small Trend Method

We set the general model:

$$X_t = m_t + s_t + Y_t$$

Where:

X_t denotes the transformed series; m_t denotes the trend component; s_t denotes the seasonality component, and Y_t denotes the error term.

As the trend is small, it is reasonable to assume that trend is constant and denote m_i for the i^{th} year. With $\sum_{j=1}^{365} S_j = 0$, we use $\widehat{m_i} = \frac{1}{365}\sum_{j=1}^{365} x_{i,j}$ to be the unbiased estimator. While for S_j, j=1, 2, 3,…, 365, we use the estimates

$\widehat{S_j} = \frac{1}{4}\sum_{1}^{4}(x_{i,j} - \widehat{m_i})$ and satisfy the requirement that $\sum_{j=1}^{365} \hat{S}_j = 0$. The estimated error term for day j of the i^{th} year is

$\widehat{Y_{i,j}} = x_{i,j} - \widehat{m_i} - \widehat{s_j}$, i= 1,2,3,4; j= 1,2,…,365

The deseasonalized and detrended observations, $\widehat{Y_{i,j}} = x_{i,j} - \widehat{m_i} - \widehat{s_j}$, have no apparent seasonality or trend, and so the series of these observations is stationary.

We now proceed to resolve on residual analysis. The ACF plot of residuals represents an exponential decay, and the PACF plot shows that the partial autocorrelation is significant at lag 3. It suggests we fit the residuals with an AR(3) process. To set a model for $_tX$, let

$$X_t - m_t - s_t = \frac{\eta_t}{(1-\phi_1 B - \phi_2 B^2 - \phi_3 B^3)},\ \eta_t \sim WN(o, \sigma_\eta^2)$$

where $X_t - m_t - s_t$ is the stationary series.

The coefficients of the backward-shift operators are ϕ_1= 0.4637 , ϕ_2= 0.0192 and ϕ_3= 0.0841. But ϕ_2 is not significant, we exclude it from our model and then obtain following relationship :

$$X_t - m_t - s_t = \frac{\eta_t}{(1 - 0.4637B - 0.0841B^3)}$$

Then we need to check if η_t follows a white noise process. The ACF and PACF plots of η_t show that there is no apparent structure in the model, so we believe η_t follows a white noise process. On the other hand, the modified Ljung-Box test also concludes that $\{\eta_t\}$ is a white noise process.

After all we have the following model for X_t:

$$X_t = m_t + s_t + \frac{\eta_t}{(1-0.4637B-0.0841B^3)},\quad \eta_t \sim WN(o, \sigma_\eta^2)$$

Figur 3-1: The seasonality component S_t and the trend component m_t

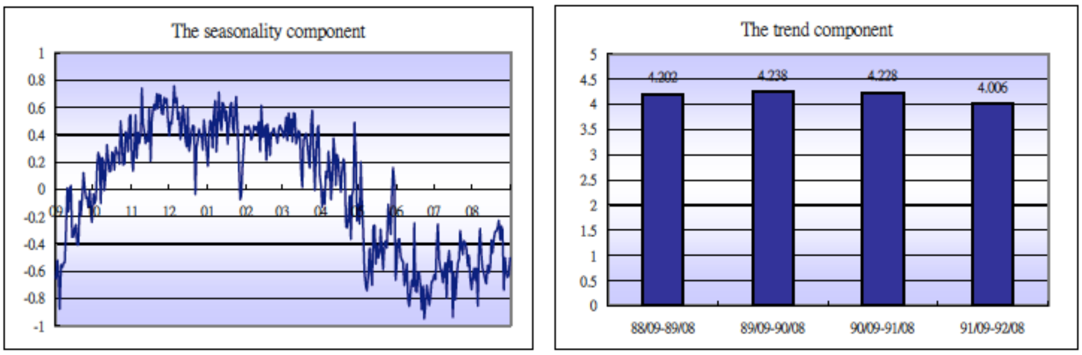


Figure 3-2: The detrended and deseasonalized observations

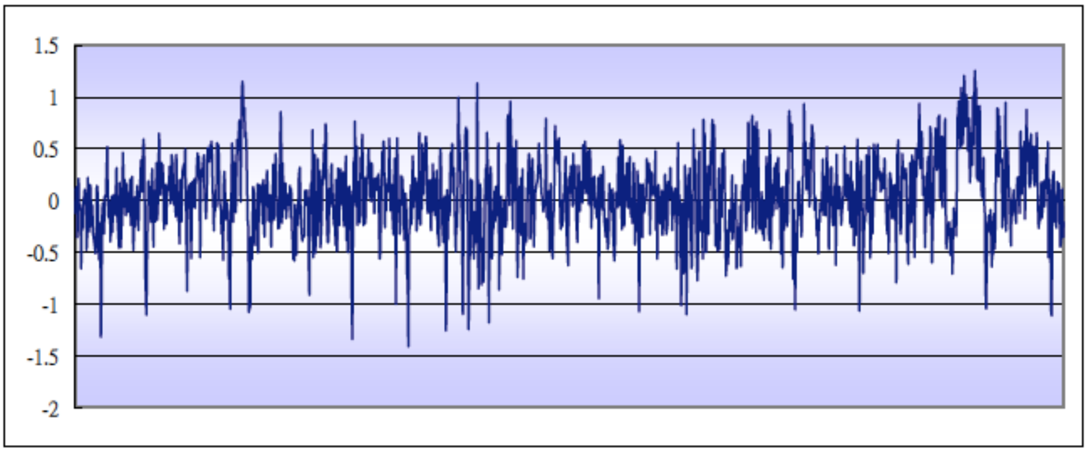


Figure 3-3: The ACF plot of the detrended and deseasonalized observations

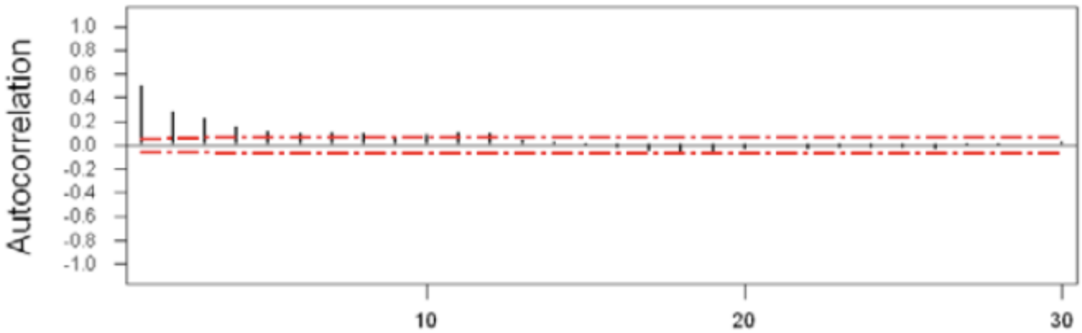


Figure 3-4: The PACF plot of the detrended and deseasonalized observations

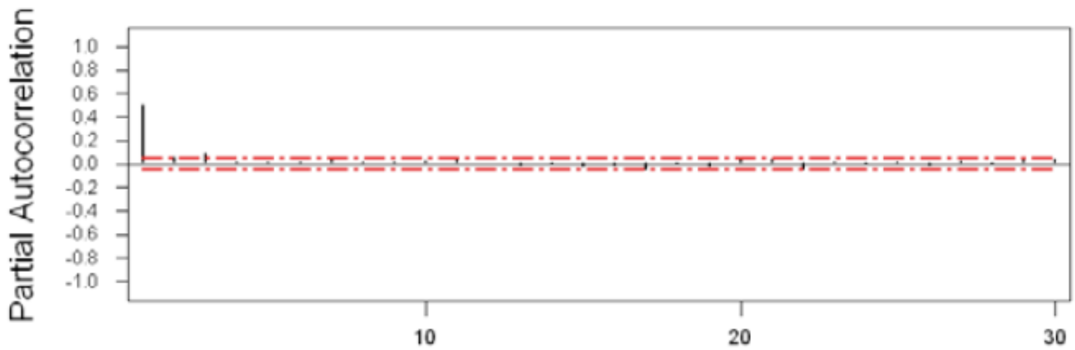


Table 3-1: Estimates of parameters of the AR(3) process

Table 3-1: Estimates of parameters of the AR(3) process

Type	Coef	SE Coef	T	P
AR 1	0.4637	0.0261	17.76	0.000
AR 2	0.0192	0.0288	0.67	0.505
AR 3	0.0841	0.0261	3.22	0.001
Number of observations: 1460				

Figure 3-5: The ACF plot of η_t

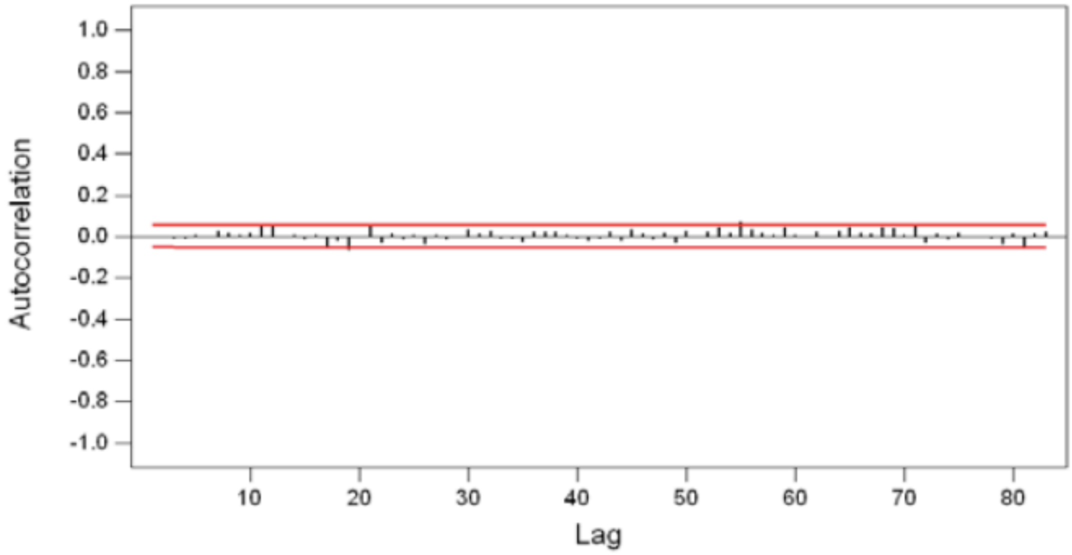


Figure 3-6: The PACF plot of η_t

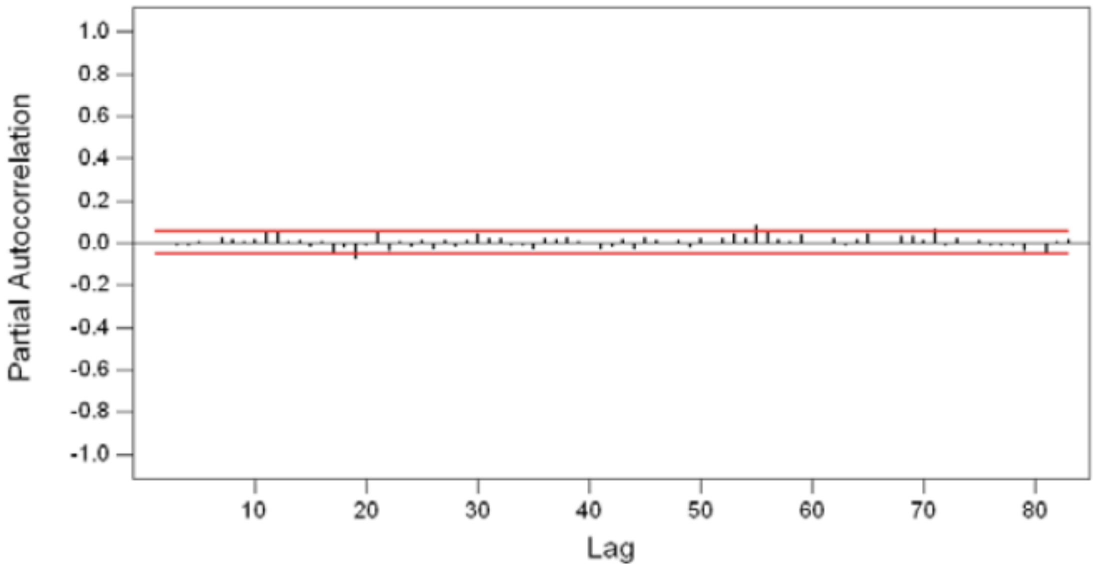


Table 3-2: Modified Ljung-Box Chi-Square statistic

Lag	12	24	36	48
Chi-Square	11.2	31.9	40.2	46.9
DF	9	21	33	45
P-Value	0.264	0.060	0.182	0.395

Method 2: OLS Method

As the regular cycle of the series, we try to model X_t with a cosine function. Observe behavior of the series we consider following model:

$$X_t = \mu + R \cdot \cos(\omega \cdot t + \theta) + \varepsilon_t,$$

where R denotes the amplitude, ω denotes the frequency, θ denotes the phase, and ε_t denotes the error term. Also, let $\bar{x} = \hat{u}$ be the estimator of μ.

These parameters are estimated by OLS method and results are:

$$\widehat{X_t} = 4.2262 + 0.5927\cos(0.0172t - 2.1862)$$

where 0.0172 = 2π / 365.

Figure 3-7 shows a stationary process for the error term $\varepsilon_t = \widehat{X_t} - X$. The ACF plot of errors shows an exponential decay and a partial autocorrelation is significant at lag 3. It suggests that we fit the errors with an AR(3) process. Let $X_t - \widehat{X_t} = \frac{e_t}{(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)}$ is the stationary series, $e_t \sim WN(0, \sigma_e^2)$.

After model refinement, final model for X_t is

$$X_t = 4.2262 + 0.5927\times \cos(0.0172t - 2.1862) + \frac{e_t}{1 - 0.678B + 0.0539B^2 - 0.0771B^3}$$

Figure 3-7: The series ε_t

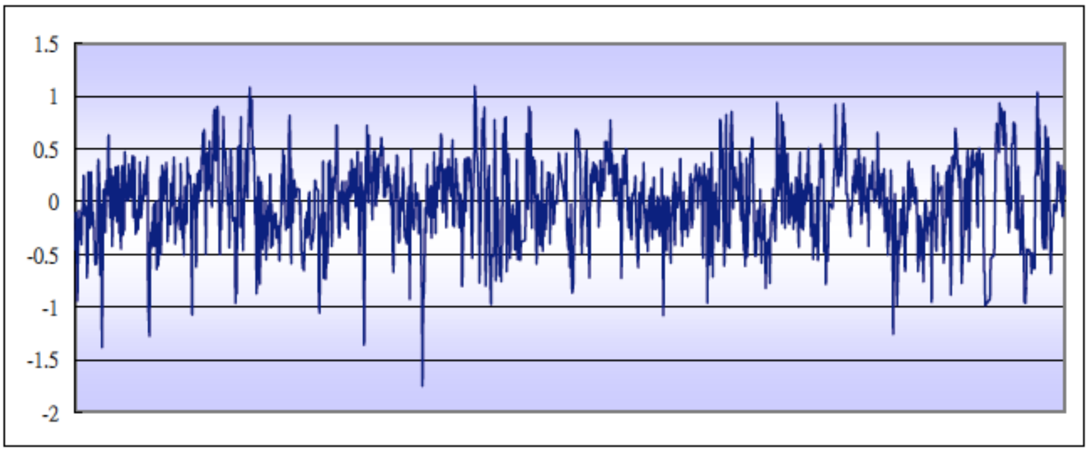


Figure 3-8: The ACF plot of ε_t

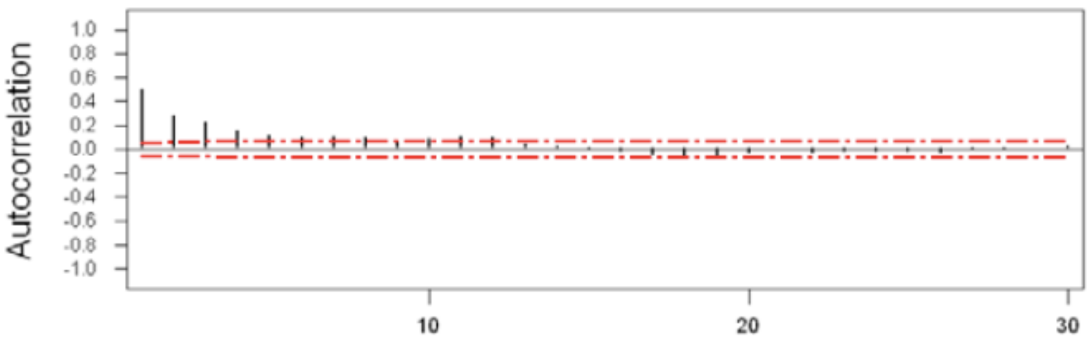


Figure 3-9: The PACF plot of ε_t

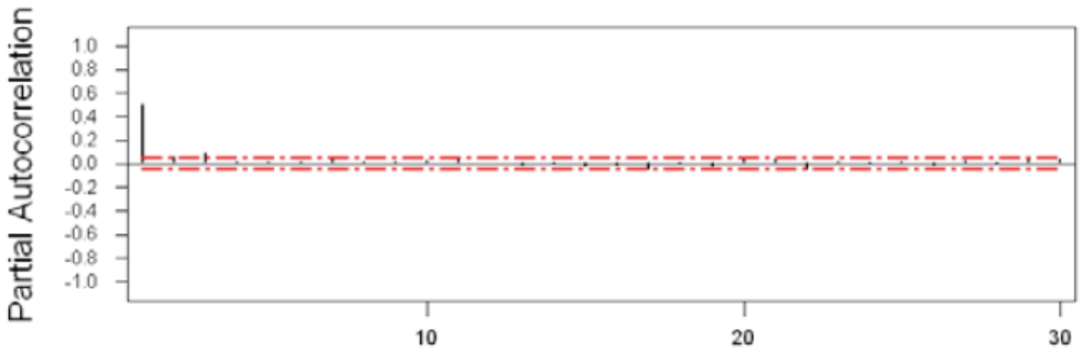


Table 3-3: Estimates of parameters of the AR(3) process

Type	Coef	SE Coef	T	P
AR 1	0.6078	0.0261	23.27	0.000
AR 2	-0.0539	0.0306	-1.76	0.078
AR 3	0.0771	0.0261	2.95	0.003
Number of observations: 1460				

Figure 3-10: The ACF plot of e_t

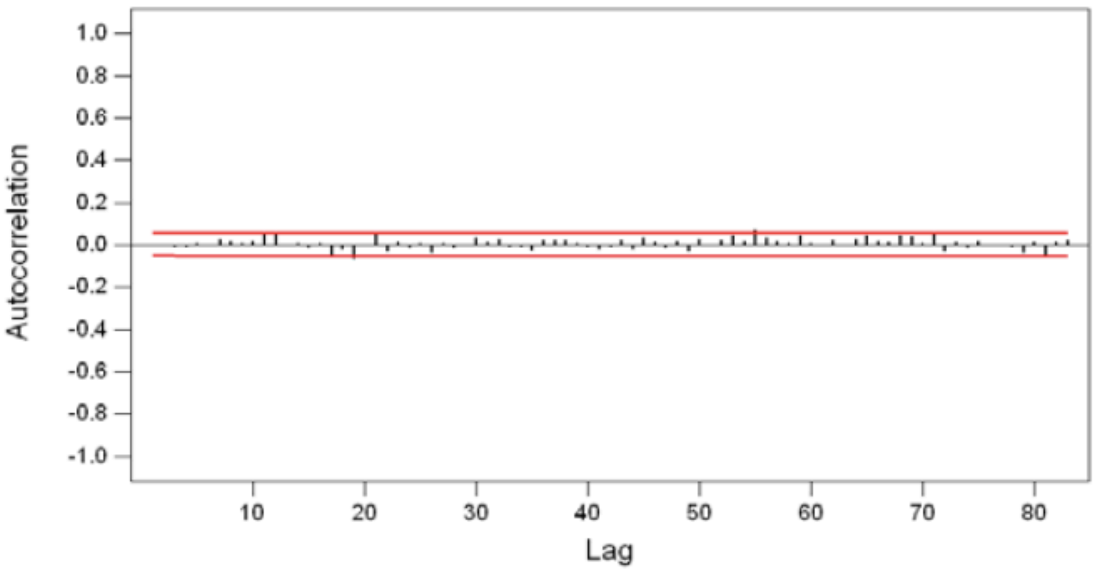


Figure 3-11: The PACF plot of e_t

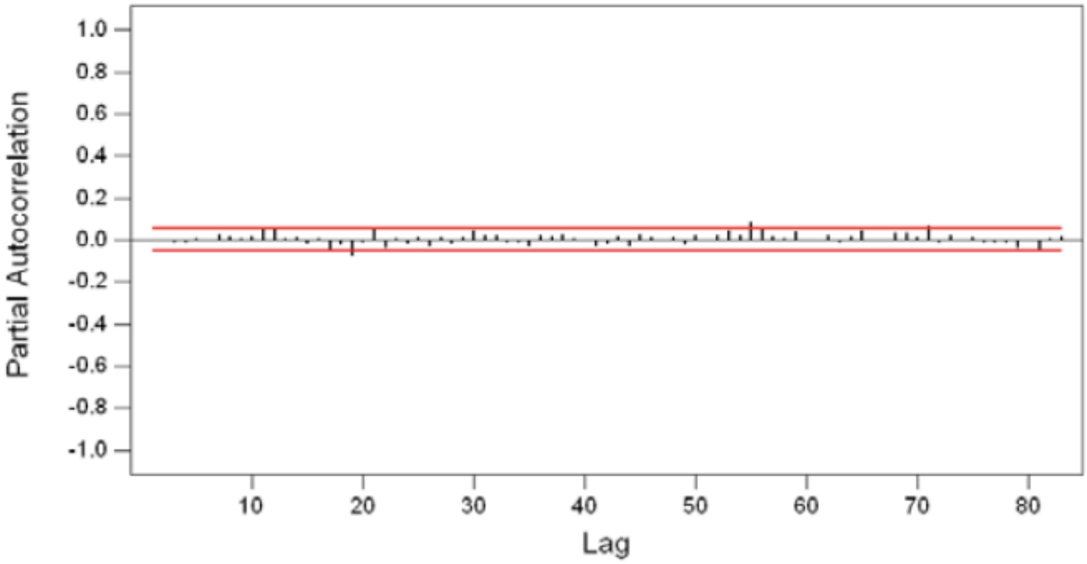


Figure 3-12: The series ε_t with figure 3-7 fit

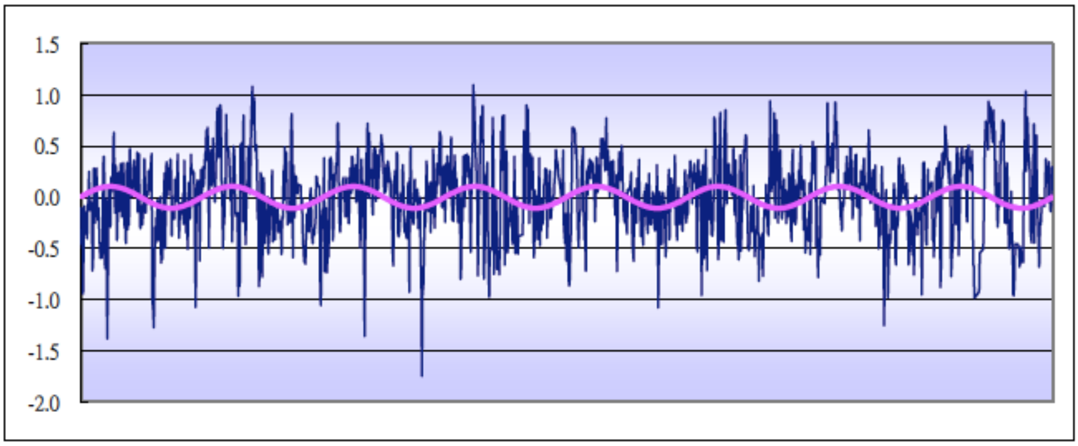


Table 3-4: Modified Ljung-Box Chi-Square statistic

Lag	12	24	36	48
Chi-Square	11.0	24.3	30.3	35.5
DF	9	21	33	45
P-Value	0.277	0.278	0.601	0.844

IV. Spectral Analysis

We estimated the seasonal parameters by above function. Though we set a reasonable model for X_t , from figure 3-12 we observed there still exists a tiny cycle with a period of about half a year. We think that the half-yearly cycle also has impact on PM_{10} concentrations, hence we apply spectral analysis to help us confirm our impact.

For $_tX$, consider the following Fourier transform decomposition

$$X_t = \frac{a_0}{2} + \sum_{k=1}^{m}[a_k cos(\omega_k t) + b_k sin(\omega_k t)],\text{ where } a_0 = 2\bar{X}$$

corresponds to the mean behavior, m denotes the number of frequencies in the Fourier Transform, ω_k denotes the Fourier frequencies = 2πk / n and n is the number of observations. The spectral density function shows the strength of the signal as a function of frequency, and the sum of the spectral density function over frequency equals the variance of the time series data. We only capture the most important origins of the variance and use them to estimate the seasonality.

Figure 4-1 and 4-2 show the periodogram for PM_{10} concentrations at Chao-Chow from September 1999 to August 2003. The signals at the yearly and half-yearly frequencies are easily visible. The largest peak visible in figure 4-1 occurs at a frequency of 0.01721 day^{-1}, or a period of 365 days; the second largest peak occurs at a frequency of 0.03443 day^{-1}, which is corresponding to the half-yearly pattern. We have the following model

$$X_t = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.0172t) + 0.004236\cos(0.03443t) + 0.1064\sin(0.03443t) + n_t$$

where n_t denotes the noise term including all other signals. $_tn$

Figure 4-3 shows no apparent trend or seasonality, which we believe that the series $\{n_t\}$ is stationary. The ACF plot of $\{n_t\}$ represents an exponential decay, and the PACF plot shows that the partial autocorrelation is significant only at lag 1. It suggests we fit the noise term with an AR(1) process. We set model for X_t be

$$X_t - \widehat{X_t} = \frac{\xi_t}{(1-\phi B)},\ \xi_t \sim WN\left(0, \sigma_\xi^2\right)$$

where $\hat{X} = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t) + 0.004236cos(0.03443t) + 0.1064sin(0.03443t)$, t = 0,1,2,...1459

Substituting ϕ= 0.5886 back into the model we obtain the relationship

$$X_t - \widehat{X_t} = \frac{\xi_t}{(1-0.5886B)}$$

Similar to previous analysis, we need to check if ξ_t follows a white noise process. The ACF and PACF plots of ξ_t show that there is no apparent structure in the model, so we believe that ξ_t follows a white noise process. The result of the modified Ljung-Box test supports the conclusion.

The final model for X_t is as the following:

$$\hat{X} = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t) +$$

$$0.004236cos(0.03443t) + 0.1064sin(0.03443t) + \frac{\xi_t}{1-0.5886B}, \text{ t = 0,1,2,...1459}$$

Figure 4-1: The periodogram of the PM_{10} concentrations over frequency

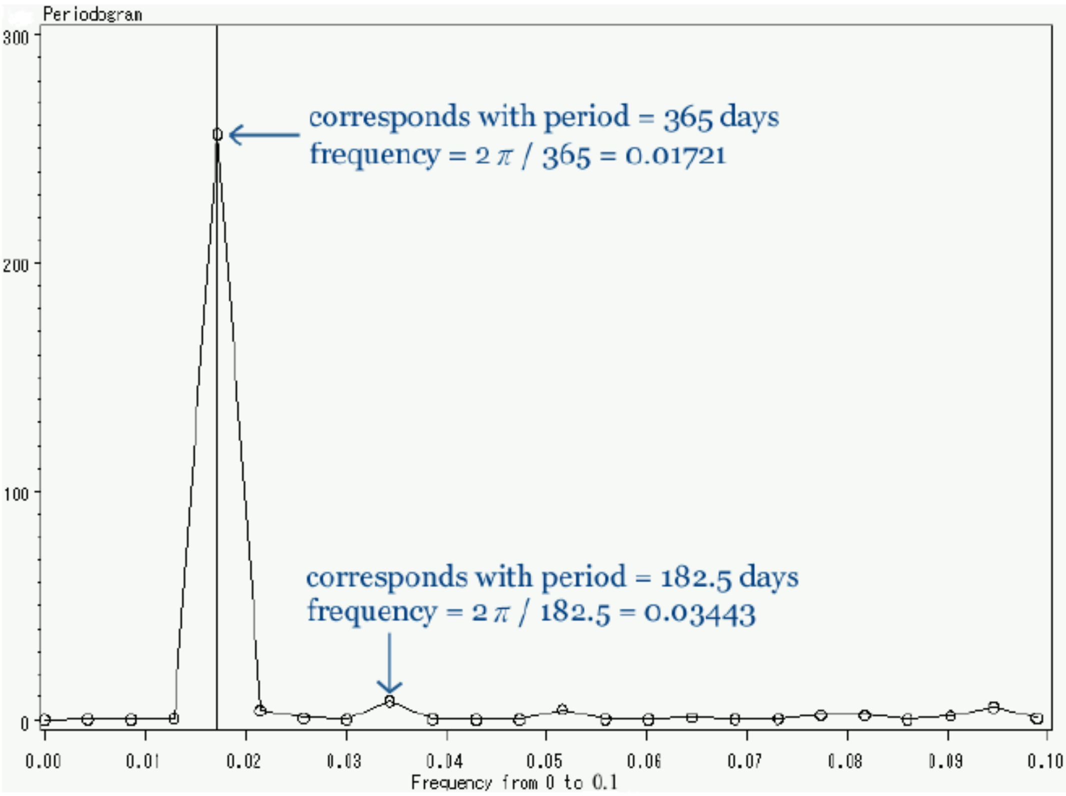


Figure 4-2: The periodogram of the PM_{10} concentrations over period

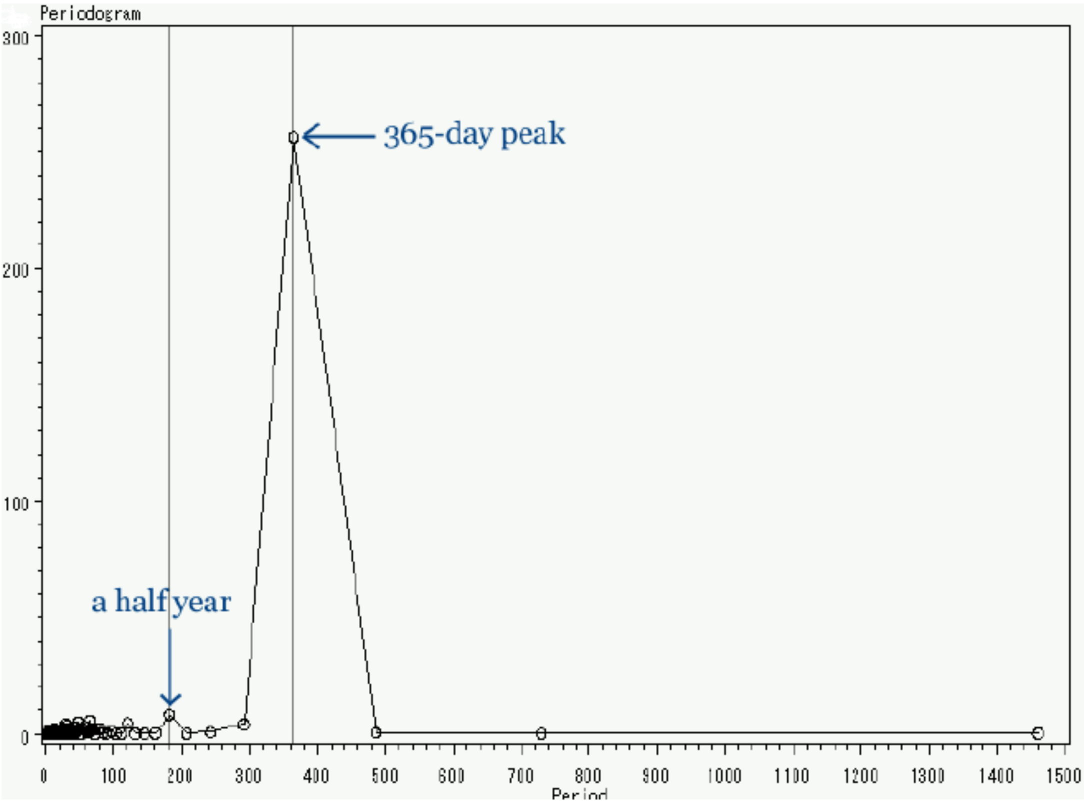


Figure 4-3: The series $\{n_t\}$

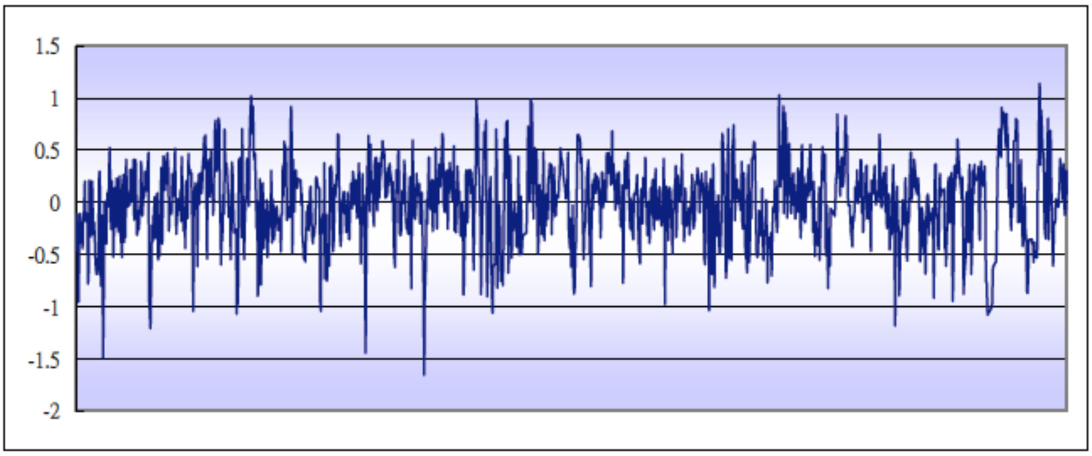


Figure 4-4: The ACF plot of n_t

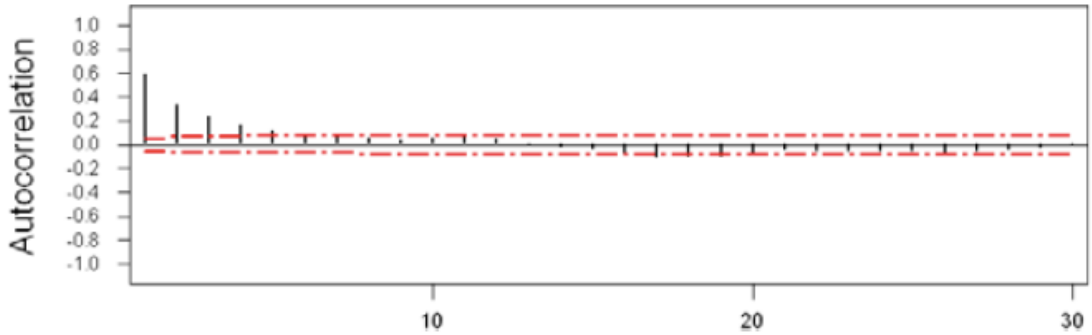


Figure 4-5: The PACF plot of n_t

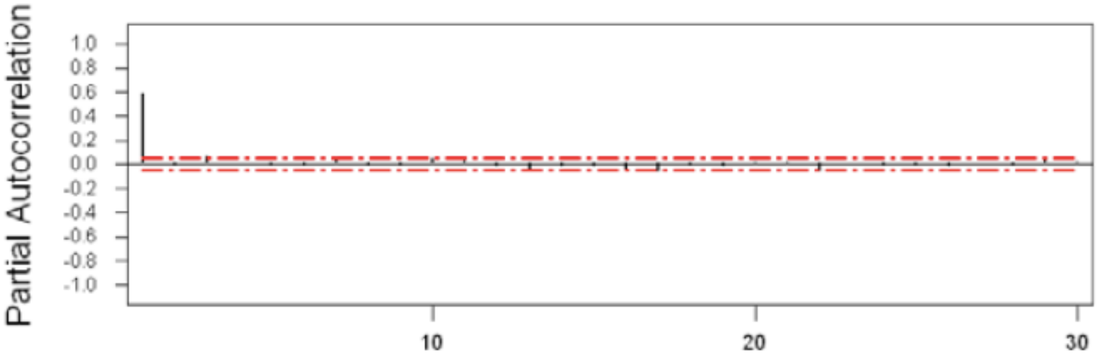


Table 4-1: Estimates of parameters of the AR(1) process

Type	Coef	SE Coef	T	P
AR 1	0.5886	0.0212	27.81	0.000

Number of observations: 1460

Figure 4-6: The ACF plot of ξ_t

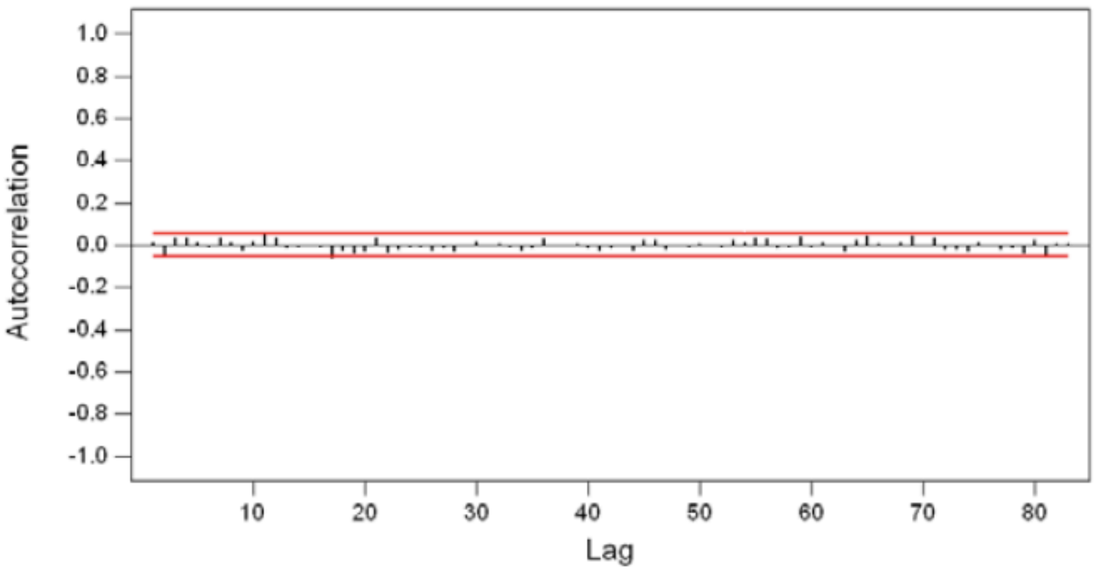


Figure 4-7: The PACF plot of ξ_t

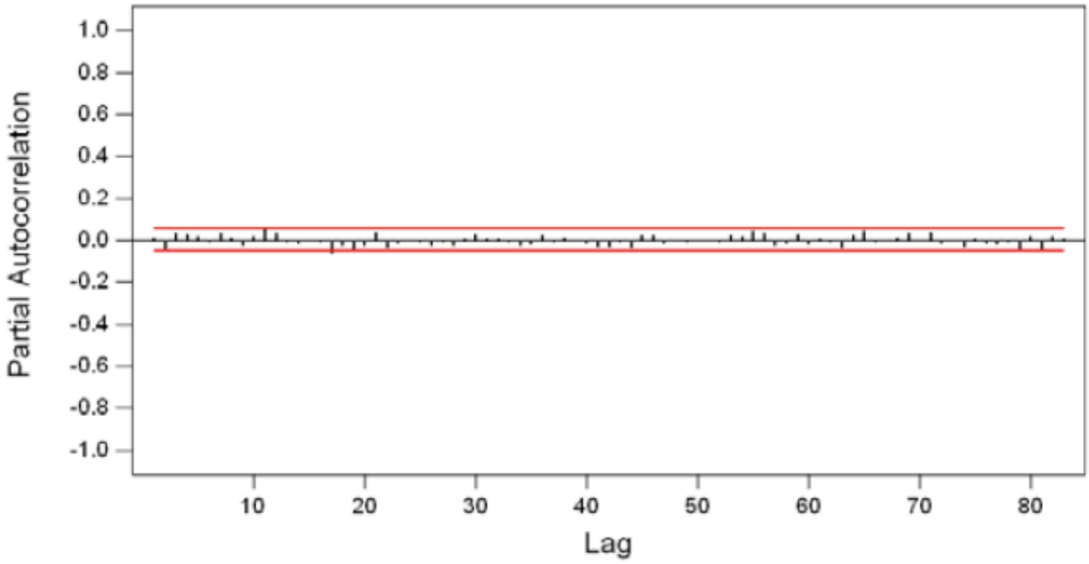


Table 4-2: Modified Ljung-Box Chi-Square statistic

Lag	12	24	36	48
Chi-Square	16.7	34.8	41.8	47.1
DF	11	23	35	47
P-Value	0.117	0.055	0.199	0.468

V. The Comparison among the Three Models

We have already built models for X_t and the difference among them lies on the estimations of seasonality. After modeling X_t, we next want to find out which one performs better. We make a comparison among these models at the aspect of forecasting ability. Before that, we are supposed to give the criterion for judging which model to be better in prediction. The criterion is based on the out-sample MSE and the number of outliers. The smaller the out-sample MSE, and the less the number of outliers, the better the model is.

We give one-step prediction to X_{t+1} and X_{t+2} respectively and then make a comparison based on the prediction results. As mentioned in the introduction, the data we use for prediction contains 61 observations from September to October in 2003.

1.Model Derived from Small Trend Method

The model is given by $X_t = m_t + S_t + \frac{\eta_t}{1-0.4637B-0.0841B^3}$, $\eta_t \sim WN(0, \sigma_\eta^2)$

Figure 5-1: True values VS Fitted values — The Small Trend Method

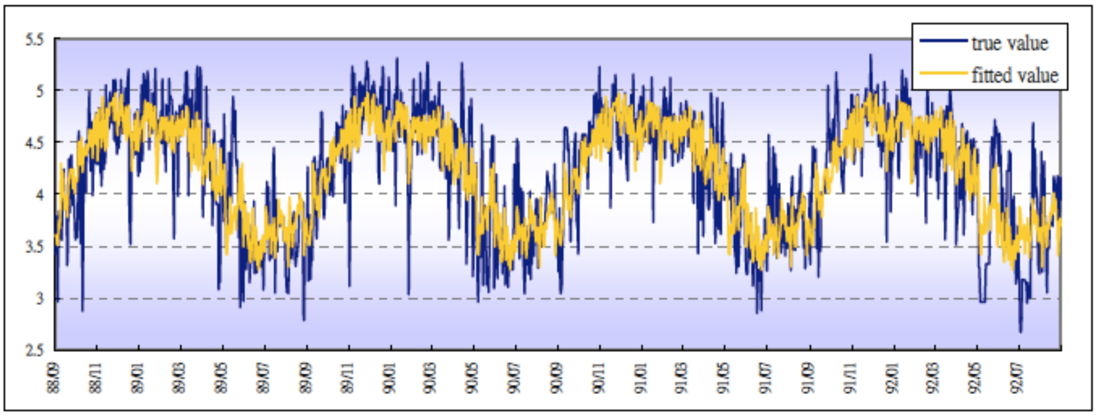


Figure 5-2: Results of the prediction for X_{t+1} — The Small Trend Method

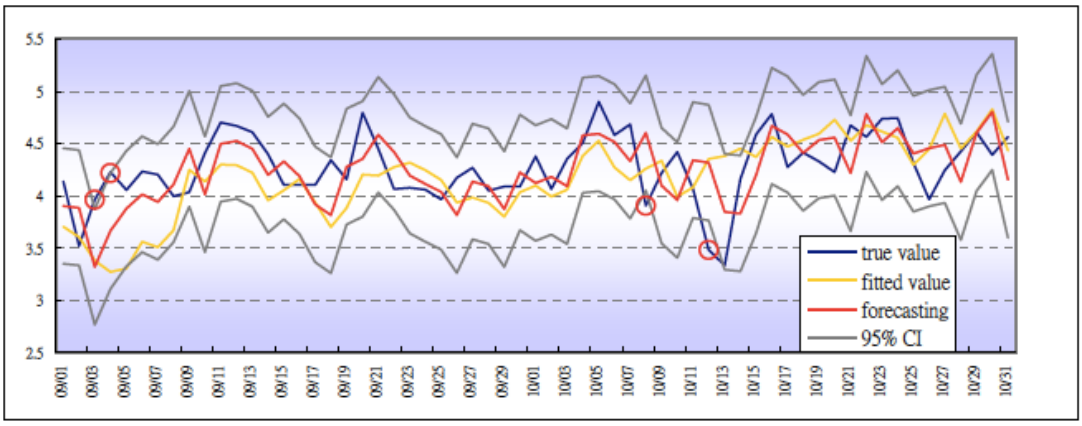


Figure 5-3: Results of the prediction for X_{t+2} — The Small Trend Method

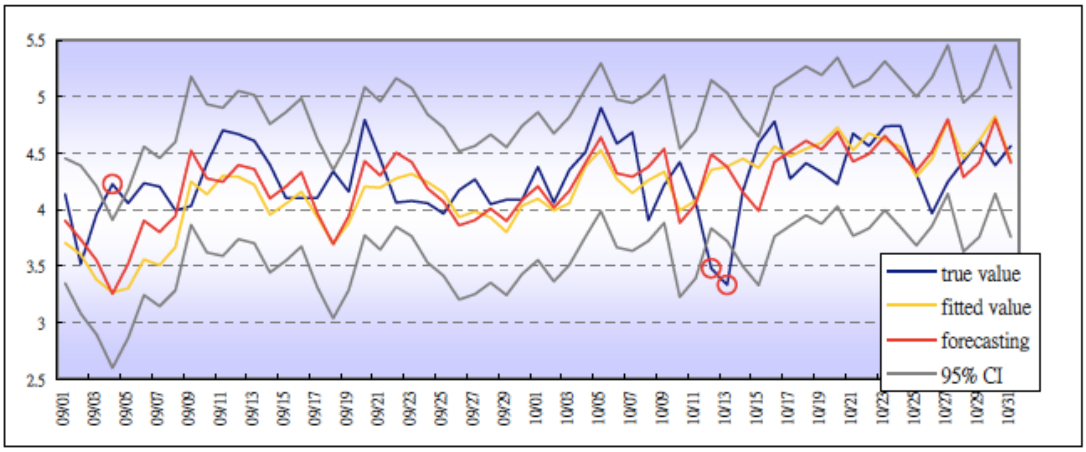


Table 5-1: Prediction results — The Small Trend Method

To be predicted	X_{t+1}	X_{t+2}
SSE	6.1749	8.6619
DF used	369	369
MSE	0.1065	0.1493
Average 95% CI width	1.1067	1.3079
Number of Outliers	4	3

2.Model Derived from OLS Method

$$X_t = 4.2262 + 0.5927cos(0.0172t - 2.1862) + \frac{e_t}{(1-0.6078B+0.0539B^2-0.0771B^3)}$$

Figure 5-4: True values VS Fitted values — The OLS Method

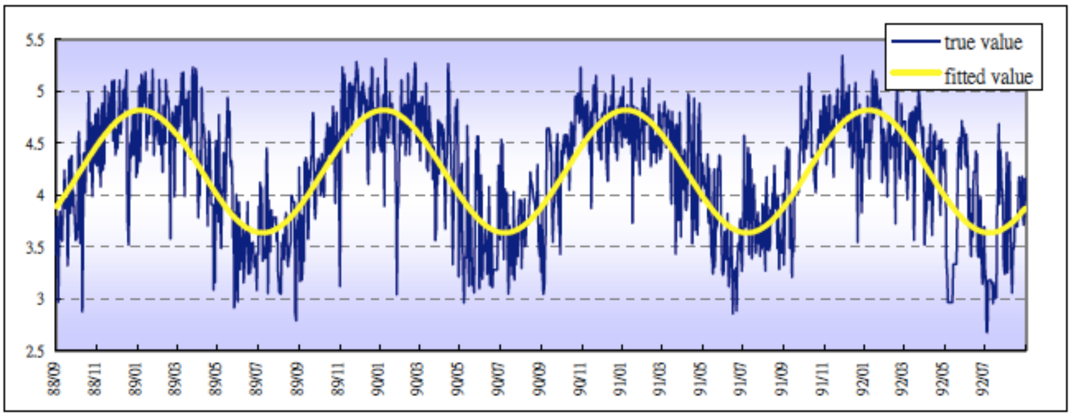


Figure 5-5: Results of the prediction for X_{t+1} — The OLS Method

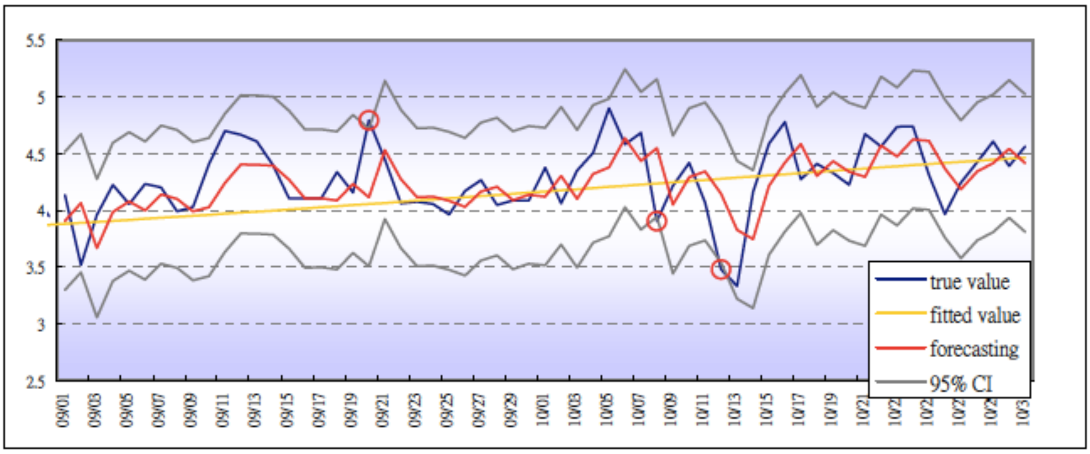


Figure 5-6: Results of the prediction for X_{t+2} — The OLS Method

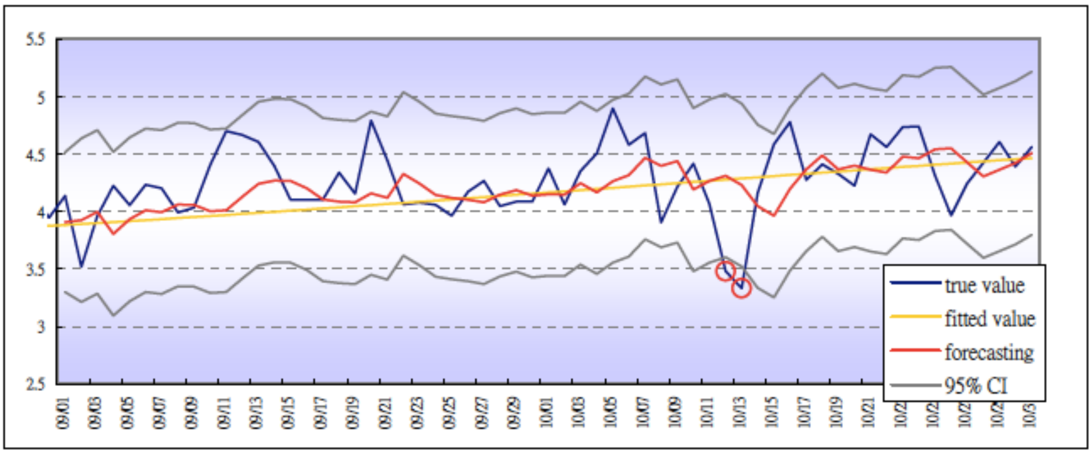


Table 5-2: Prediction results — The OLS Method

To be predicted	X_{t+1}	X_{t+2}
SSE	4.6279	6.5546
DF used	3	5
MSE	0.0798	0.1130
Average 95% CI width	1.2148	1.4216
Number of Outliers	3	2

3.Model Derived from Spectral Analysis

The model is given by

$$X_t = 4.226 - 0.34338cos(0.01721t) + 0.4831sin(0.01721t) + 0.004236\cos(0.03443t) + 0.1064\sin(0.03443t) - \frac{\xi_t}{1 - 0.5886B}$$

Figure 5-7: True values VS Fitted values — Spectral Analysis

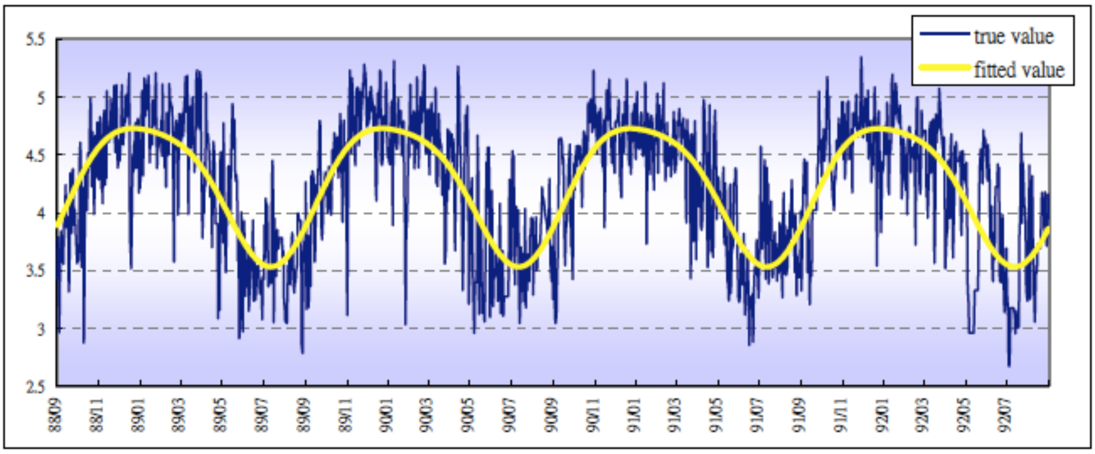


Figure 5-8: The OLS fit and Spectral Analysis fit

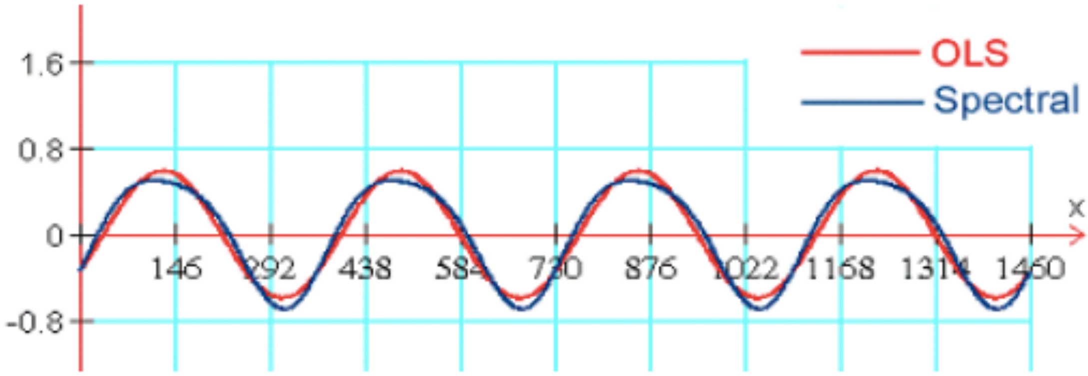


Figure 5-9: Results of the prediction for X_{t+1} — Spectral Analysis

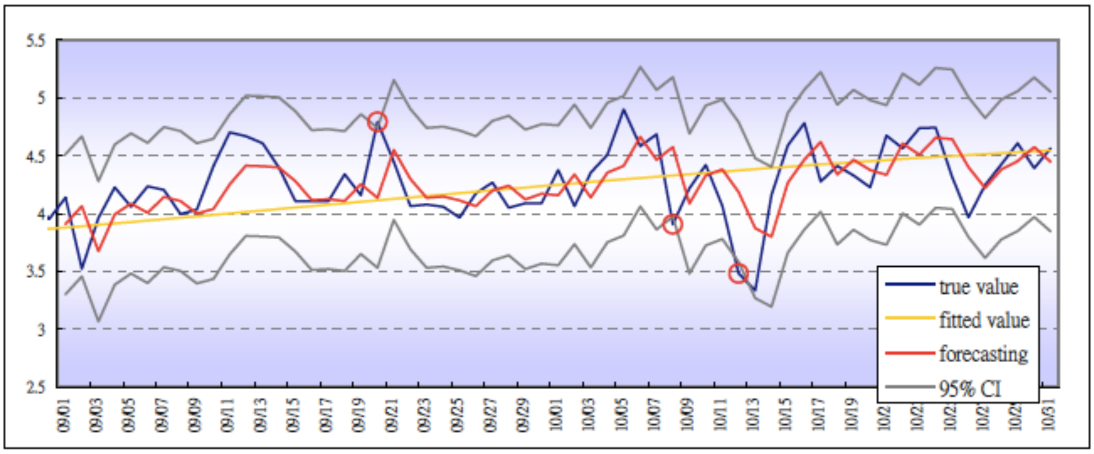


Figure 5-10: Results of the prediction for X_{t+2} — Spectral Analysis

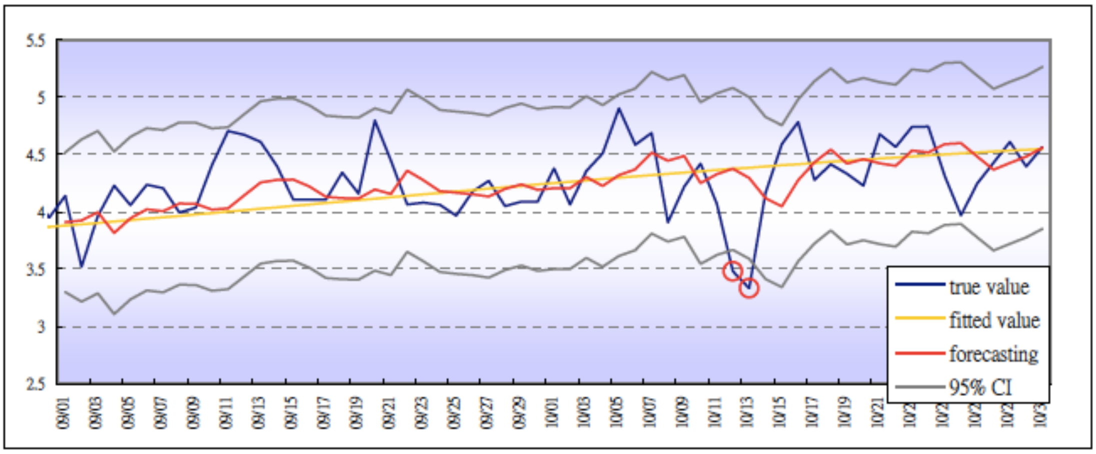


Table 5-3: Prediction results — Spectral Analysis

To be predicted	X_{t+1}	X_{t+2}
SSE	4.5734	6.4249
DF used	5	5
MSE	0.0762	0.1071
Average 95% CI width	1.2098	1.4103
Number of Outliers	3	2

Table 5-4: The Overall Prediction results

	Small Trend		OLS		Spectral Analysis	
To be predicted	X_{t+1}	X_{t+2}	X_{t+1}	X_{t+2}	X_{t+1}	X_{t+2}
SSE	6.1749	8.6619	4.62789	6.554555	4.5734	6.4249
DF used	369	369	3	3	5	5
MSE	0.1065	0.1493	0.079791	0.11301	0.0762	0.1071
Average 95% CI width	1.1067	1.3079	1.214798	1.421603	1.2098	1.4103
Number of Outliers	4	3	3	2	3	2

VI. Conclusion

Assume the model obtained from the small trend method be model 1, model 2 is from OLS method and model 3 is from spectral analysis.

Due to model 1 contains the average values of the past four years, it is easily affected by some extreme values. For this reason, predicted values of model 1 represent large fluctuations as we seen in figures 5-2 and 5-3. But due to easily fluctuations, it often makes errors in forecast.

For model 2 and model 3, fluctuations of the predicted values are smaller. The predicted value with model 2 is mainly changing with its past three observation values while the predicted value with model 3 is mainly varying with previous observation value. Hence, these two model with fewer errors than model 1.

In table 5-4, model 1 has the largest MSE and more outliers than other two models, which reveals model 1 may not be a good forecast model. At last, we build model 3 with the consideration of the half-year seasonality component, the MSE in model 3 is smaller than model 2. Also, the 95% CI of model 3 is narrower than that of model 2. Hence, we make a little improvement on our model by losing 2 degrees of freedom. At last we conclude model 3 is the best one from these model to forecast future outcome.