I. Introduction

As the economy grows, air pollution has become a serious problem to human health. Its influence on people is usually immediate and violent. For this reason, we feel interests in the daily changes of the PM_{10} concentrations, and we would like to develop a model on it. (PM_{10}: Particulate matter < 10 micrometers in size)

Our data is gathered in Lin-Yuan(林園鄉), provided by the Graduate Institute of Environmental Engineering, National Taiwan University. The data for modeling contains 1460 daily observations from Sep. 01, 1999 to Aug. 31, 2003, and the data for forecasting contains 61 daily observations from Sep. 01 to Oct. 31, 2003.

The time series plot suggests a strong seasonal cycle with a period of one year, and therefore the main problem in our analysis is to estimate the seasonal effect. After the seasonal effect is found out, the model can be built, and then we are able to proceed to forecasting.

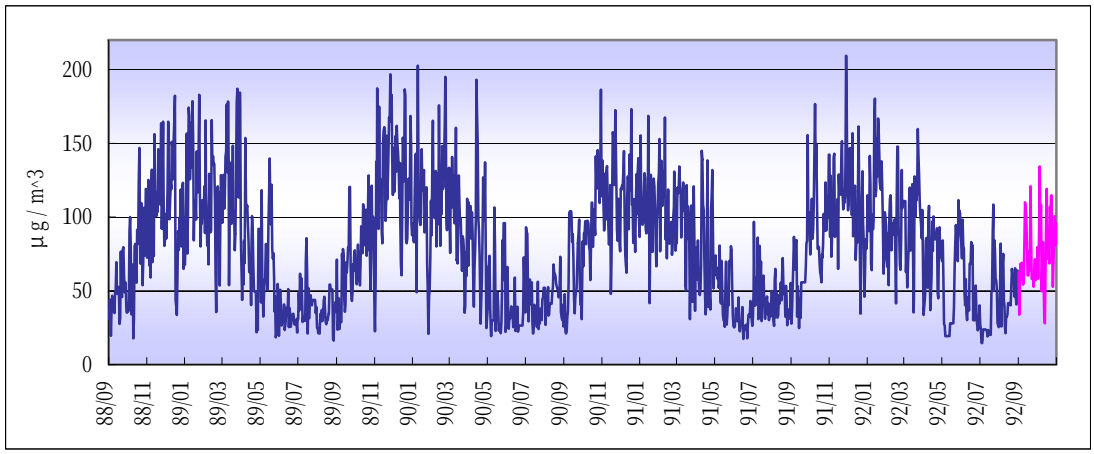
In the beginning we employ two methods —the Small Trend Method and the Ordinary Least Square Method—to estimate the seasonal component. Although by these methods we can build reasonable models, we find that there are still tiny cyclic variation in the residuals. We guess there maybe other seasonal effects on the series; thus, we turn to ask for help of spectral analysis, the consideration of the variance properties as a function of frequency. The result, being the same with our conjecture, shows that there is another origin of variance, say, the cycle with a period of a half year. With the half-year-period seasonal component being considered, we obtain a different model from the previous two.

A comparison among the three models is made with respect to their forecasting ability. After the comparison, we find that the model using spectral analysis has the minimal MSE, that is to say, in view of forecasting error, it works the best.

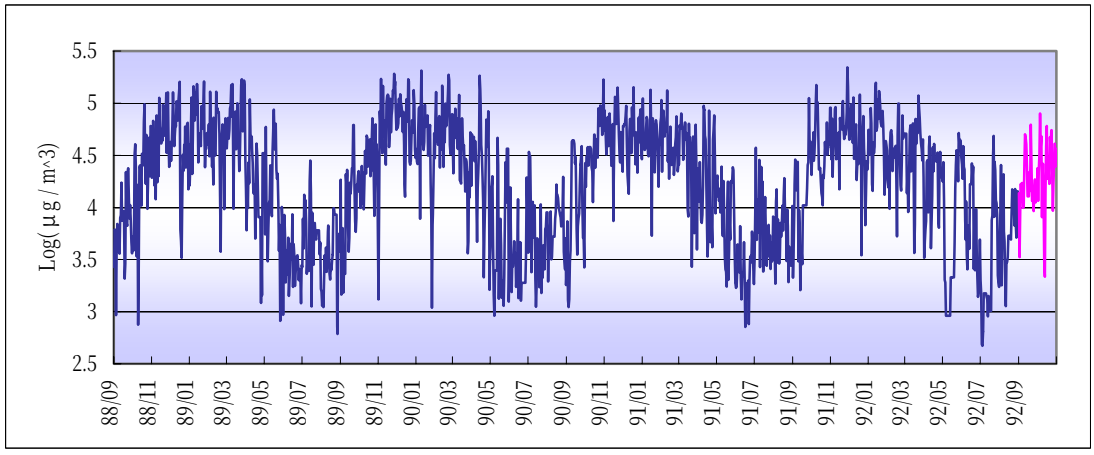
II. Data Transformation

From the time series plot, we find that there may be the problem of heteroscedasticity, which will cause inapplicability to our analysis. We therefore need to transform the data for fear of such problem. We apply a logarithmic transformation for variance stabilization. As we can see, the situation is improved after the logarithmic transformation. All our analysis will be based on the transformed data.

Figures 2-1and 2-2 show that there is no apparent trend; however, the seasonal effect is strong. Thus, the main task of our model-building is the removal of seasonality.



< 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

To make use of the 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.

Since the trend is small, it is not unreasonable to suppose that the trend is constant, say m_i, for the i^{th} year. Since $\sum_{j=1}^{365} s_j = 0$, we are led to the unbiased estimate

$$\hat{m}_i = \frac{1}{365}\sum_{j=1}^{365} x_{i,j}$$

while for s_j, j=1, 2,…,365 we have the estimates,

$$\hat{s}_j = \frac{1}{4}\sum_{i=1}^{4}(x_{i,j} - \hat{m}_i),$$

which automatically 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 of course

$$\hat{Y}_{i,j} = x_{i,j} - \hat{m}_i - \hat{s}_j, \qquad i = 1,2,3,4, \qquad j = 1,2,\ldots,365$$

We have estimated the seasonal component and the trend component. The deseasonalized and detrended observations, $\hat{Y}_{i,j} = x_{i,j} - \hat{m}_i - \hat{s}_j$, have no apparent seasonality or trend, and so the series of these observations is stationary.

We can now proceed to the work 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 X_t, let

$$X_t - m_t - s_t = (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)^{-1}\eta_t, \quad \eta_t \sim WN(0, \sigma_\eta^{\,2})$$

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

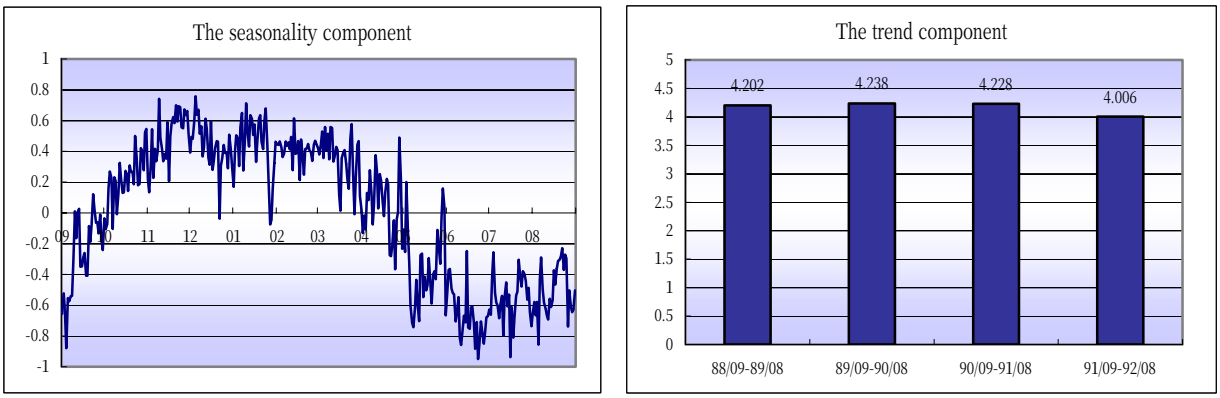
The coefficients of the backward-shift operators are $\phi_1 = 0.4637$, $\phi_2 = 0.0192$ and $\phi_3 = 0.0841$. However, ϕ_2 is not significant, we expel it from our model. We then obtain the following relationship :

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

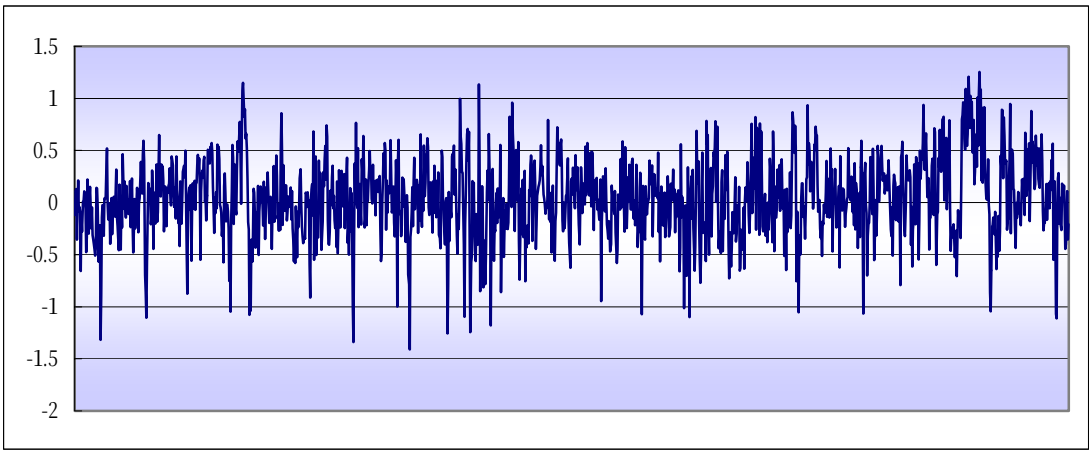
Next we are going to check if η_t follows a white noise process. The ACF and PACF plots of η_t show that there is no apparent structure in the model, so we believe that η_t follows a white noise process. On the other hand, the modified Ljung-Box test also concludes that $\{\eta_t\}$ is a white noise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

After all we have the following model for X_t :

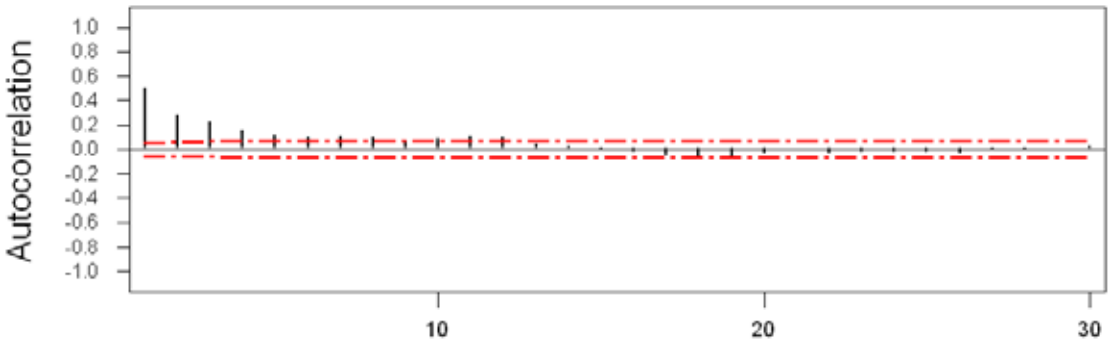
$$X_t = m_t + s_t + (1 - 0.4637B - 0.0841B^3)^{-1}\eta_t, \quad \eta_t \sim WN(0, \sigma_\eta^2)$$



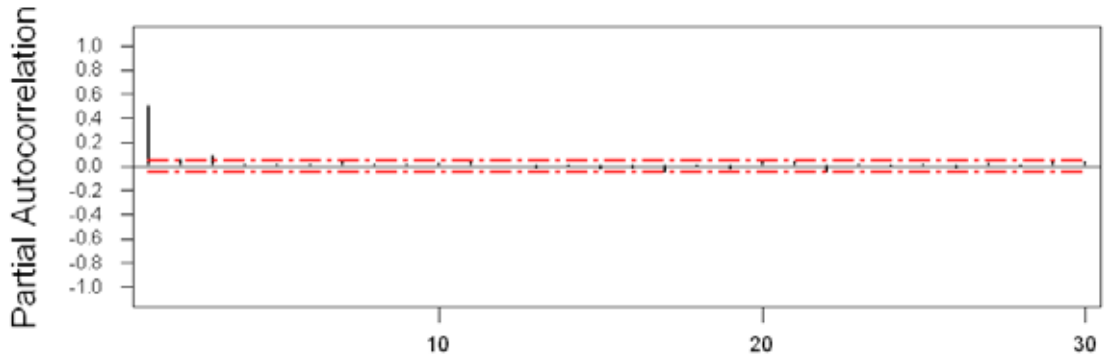
< Figure 3-1: The seasonality component s_t and the trend component m_t >



< Figure 3-2: The detrended and deseasonalized observations >



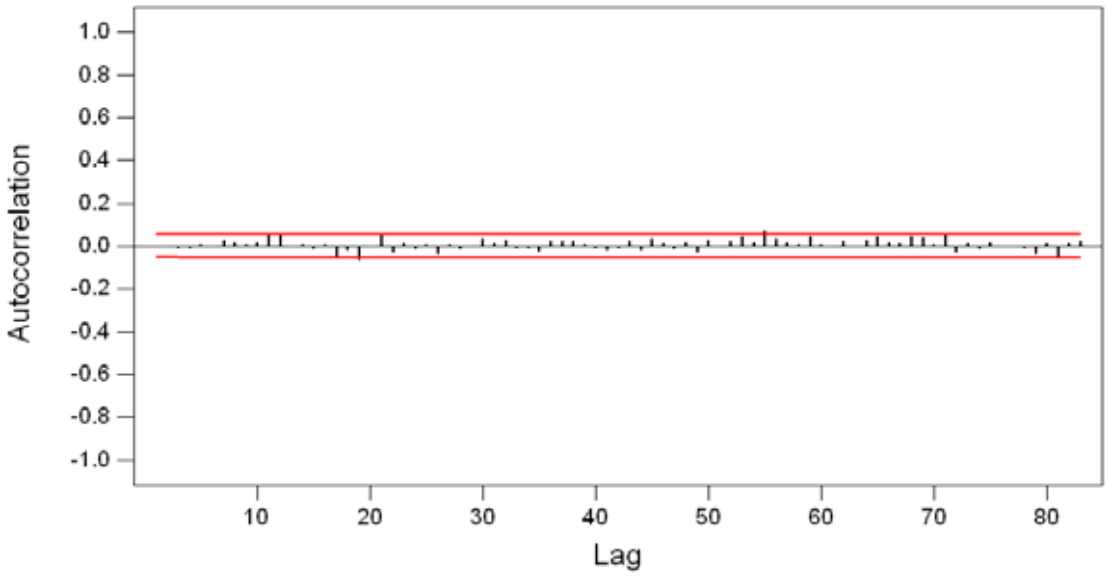
< Figure 3-3: The ACF plot of the detrended and deseasonalized observations >



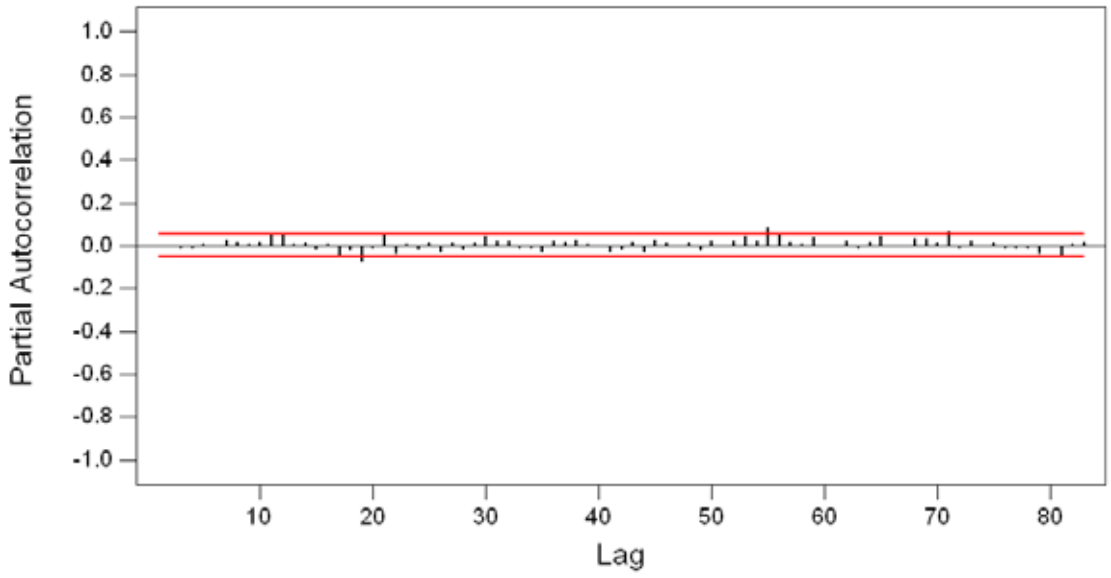
<Figure 3-4: The PACF plot of the detrended and deseasonalized observations >

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

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

Due to the regular cycle of the series, we plan to model X_t with a cosine function. Observing the behavior of the series we consider the following form:

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

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

The parameters are estimated by OLS method. The result is:

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

where 0.0172 = 2 π / 365.

Figure 3-7 depicts a stationary process, the error term $\varepsilon_t = \widehat{X_t} - X_t$. The ACF plot of errors represents an exponential decay, and the PACF plot shows that the partial autocorrelation is significant at lag 3. It suggests we fit the errors with an AR(3) process. To set a model for X_t, let

$$X_t - \widehat{X_t} = (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)^{-1} e_t, \quad e_t \sim WN(0, \sigma_e{}^2)$$

where $X_t - \widehat{X_t}$ is the stationary series.

The coefficients of the backward-shift operators are $\phi_1 = 0.6078$, $\phi_2 = -0.0539$ and $\phi_3 = 0.0771$, which are all significant under significant level α = 0.1. Thus, we have the following relationship:

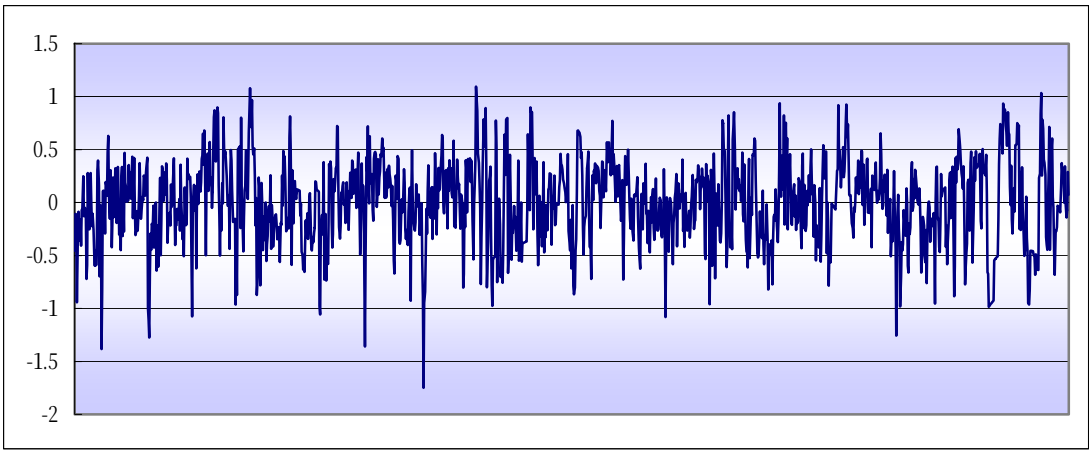
$$X_t - \widehat{X_t} = (1 - 0.6078B + 0.0539B^2 - 0.0771B^3)^{-1} e_t$$

Next we are going to check if e_t follows a white noise process. The ACF and PACF plots of e_t show that there is no apparent structure in the model, so we believe that e_t follows a white noise process. Also, the modified Ljung-Box test gives the same conclusion.

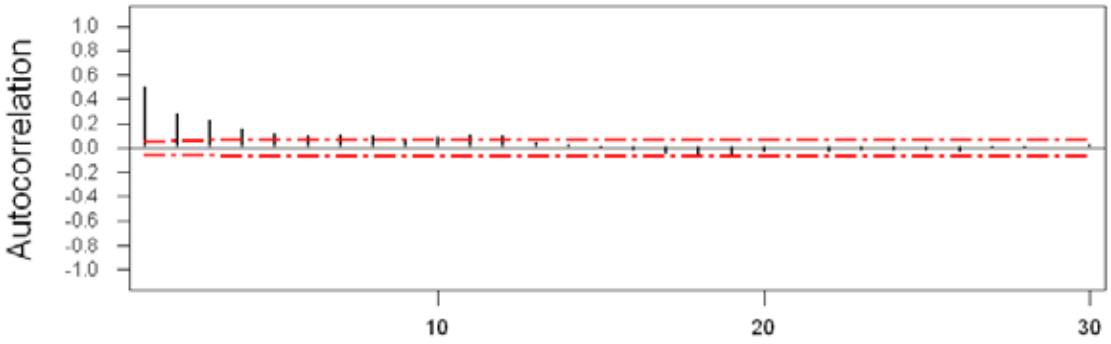
Finally, the model for X_t is

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

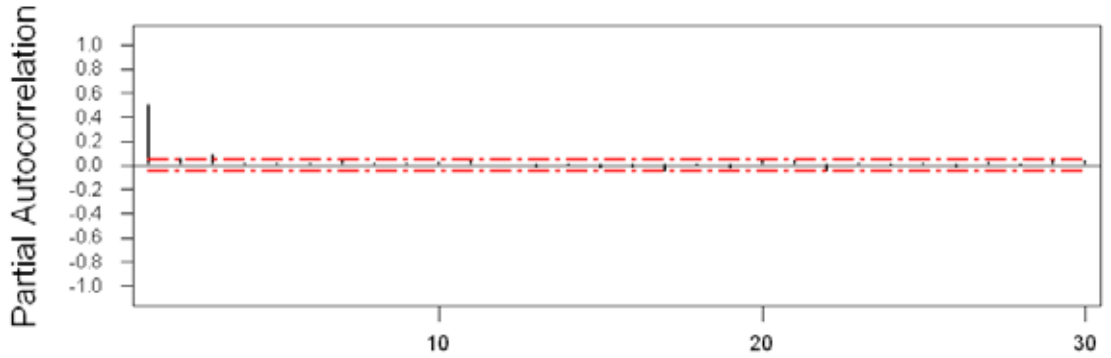
$$e_t \sim WN(0, \sigma_e{}^2)$$



< *Figure 3-7(a): The series* $\{\varepsilon_t\}$>



< *Figure 3-8: The ACF plot of* ε_t >

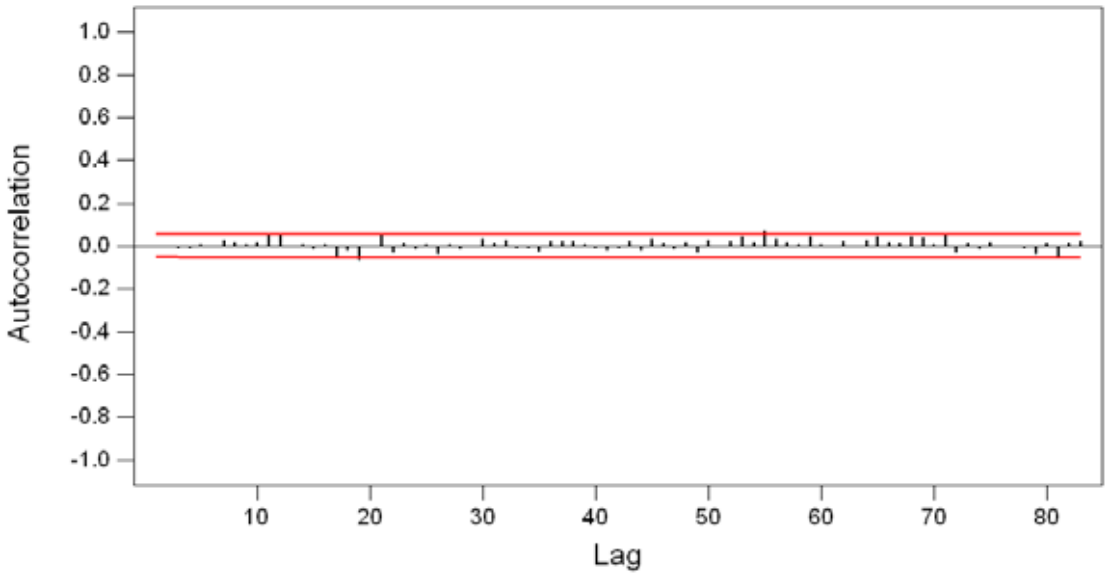


< *Figure 3-9: The PACF plot of* ε_t >

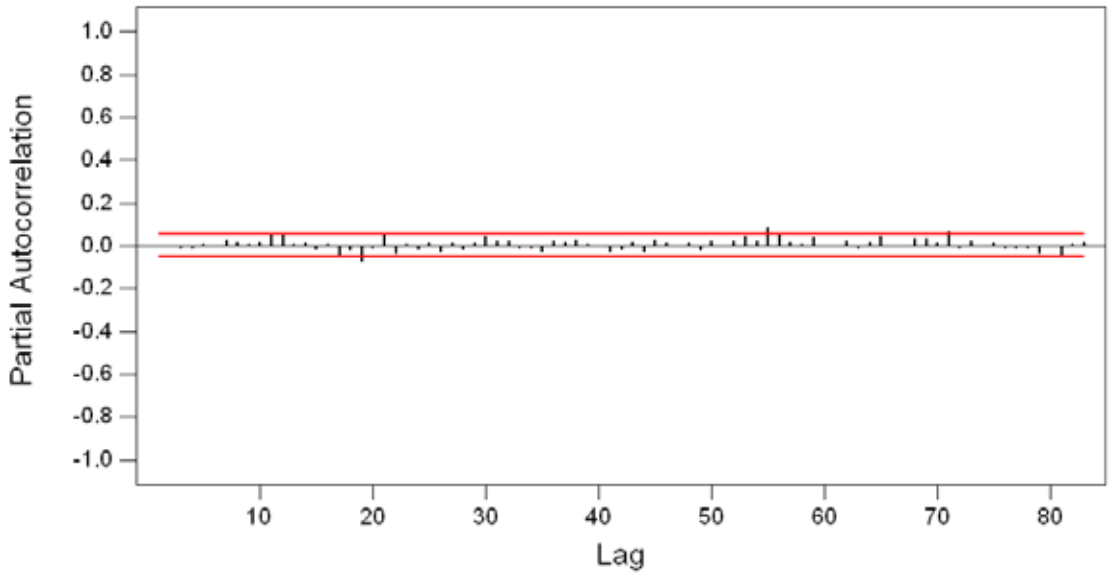
Table 3-3: Estimates of parameters of the AR(3) process

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 >

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 have estimated the seasonal component by a least-square fit using cosine function. Although by which we set a reasonable model for X_t, from figure 3-7(b) we observe that there is still a tiny cycle with a period of about half a year.

We think that the half-yearly cycle also has influence on the PM_{10} concentrations, so we employ spectral analysis, the consideration of the variance properties as a function of frequency, to help us confirm our inference.

For X_t, consider the following Fourier transform decomposition

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

where $a_0 = 2\overline{X}$ corresponds to the mean behavior, m denotes the number of frequencies in the Fourier Transform, ω_k denotes the Fourier frequencies = $2\pi k / n$ (n: the number of observations).

The spectral density function indicates 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 Lin-Yuan from Sep. 01, 1999 to Aug. 31, 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

$$X_t = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t)$$
$$+0.004236\cos(0.03443t) + 0.1064\sin(0.03443t) + n_t, \quad t = 0,1,2,...1459$$

where n_t denotes the noise term including all other signals.

Figure 4-3 shows no apparent trend or seasonality, which makes 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. To set a model for X_t, let

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

where $\widehat{X_t} = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t)$

$$+0.004236\cos(0.03443t) + 0.1064\sin(0.03443t), \quad t = 0,1,2,...1459$$

Substituting $\phi = 0.5886$ back into the model we obtain the relationship

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

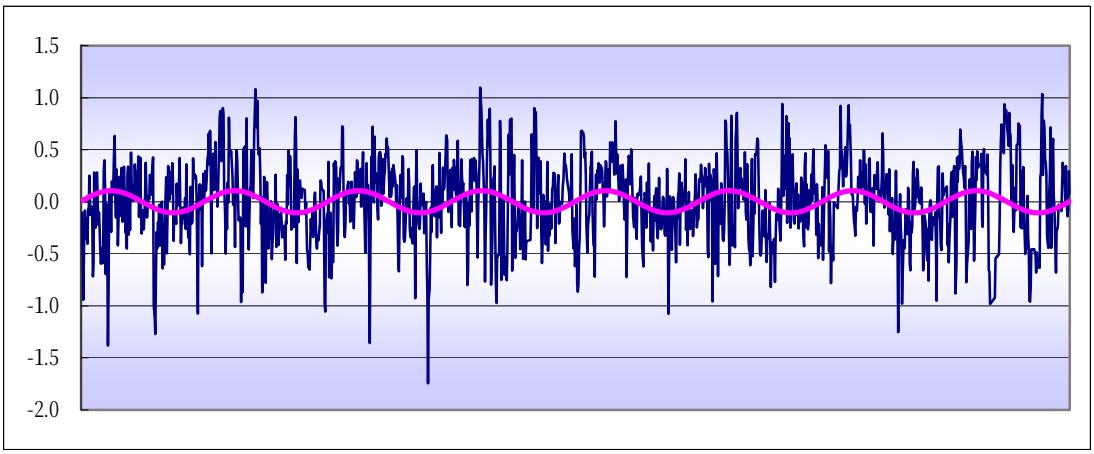
Similar to previous analysis, we need to check if ξ_t follows a white noise process. The ACF and PACF plots of ξ_t show that there is no apparent structure in the model, so we believe that ξ_t follows a white noise process. The result of the modified Ljung-Box test supports the conclusion.

The model for X_t is eventually as the following:

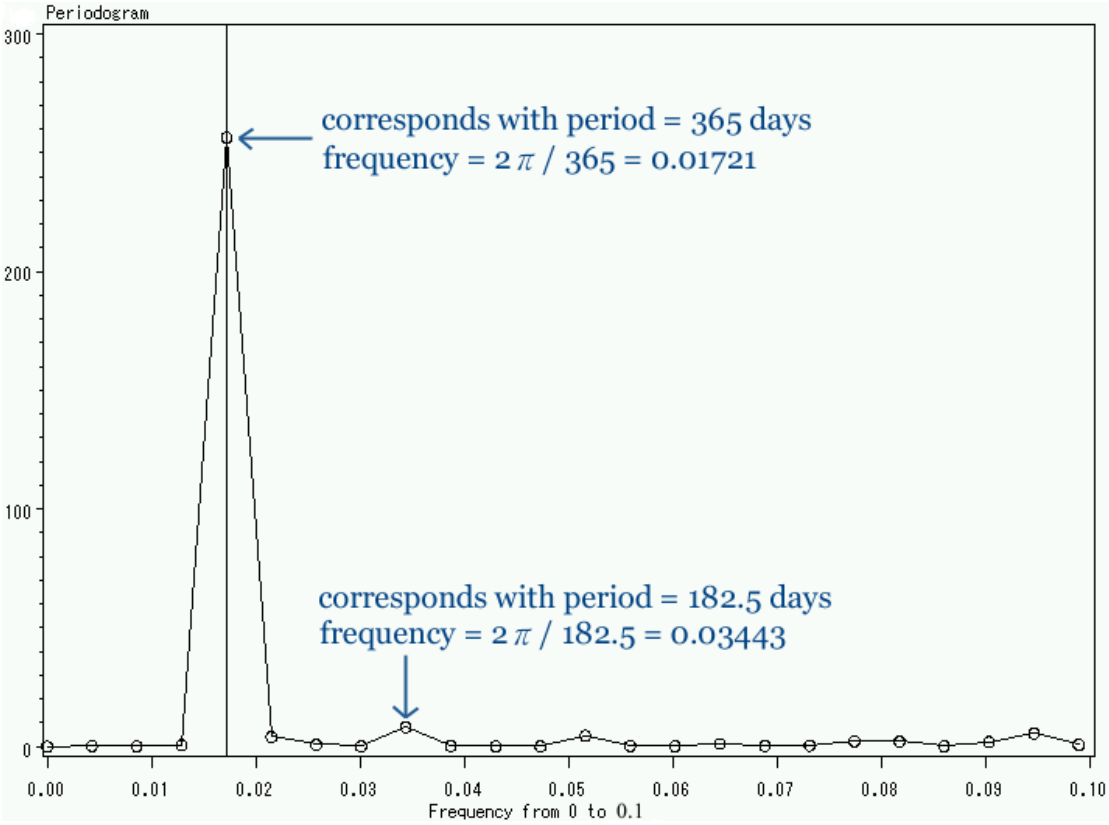
$$X_t = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t)$$

$$+0.004236\cos(0.03443t) + 0.1064\sin(0.03443t) + (1-0.5886B)^{-1}\xi_t,$$

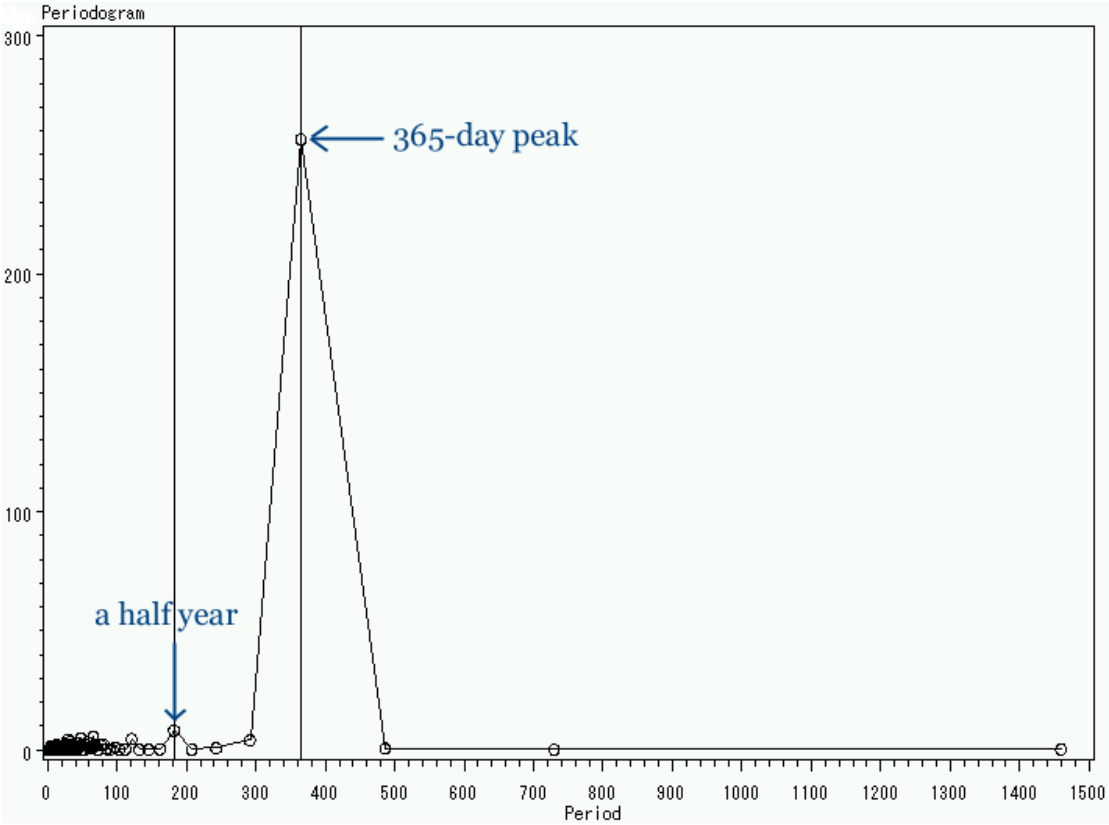
$$t = 0,1,...1459$$



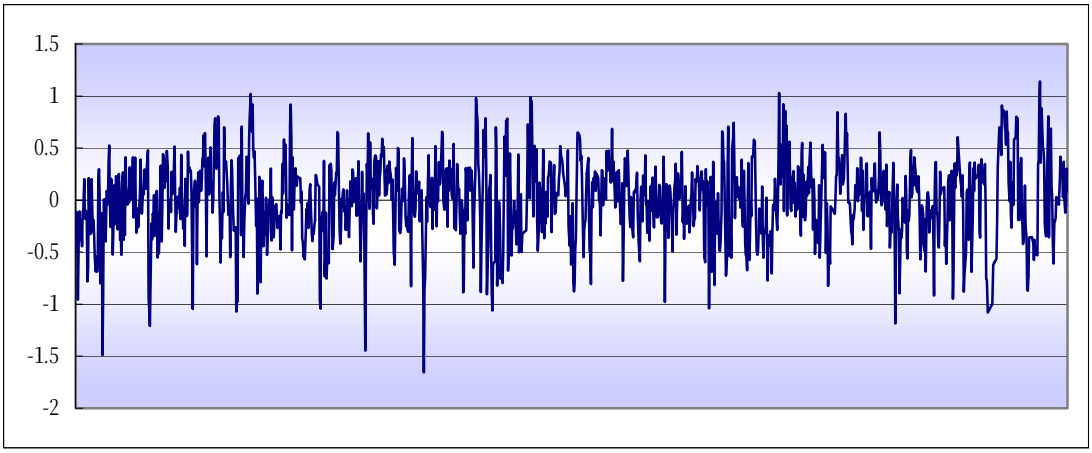
< Figure 3-7(b): The series $\{ \varepsilon_t \}$ with a fit >



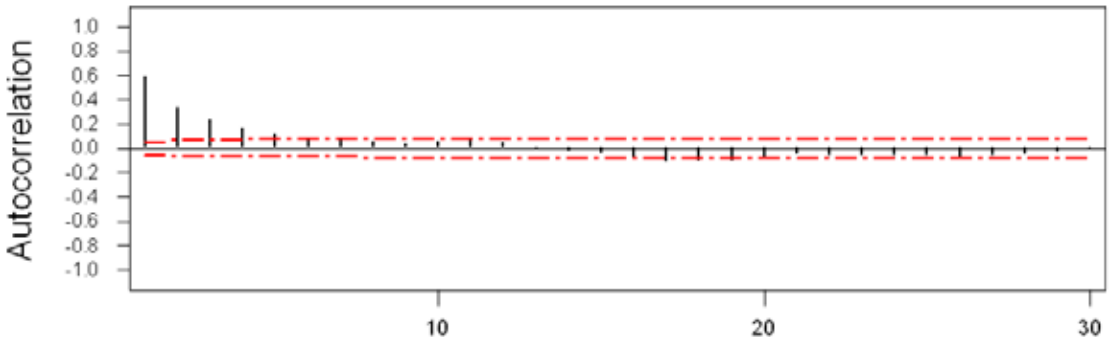
< Figure 4-1: The periodogram of the PM_{10} concentrations over frequency >



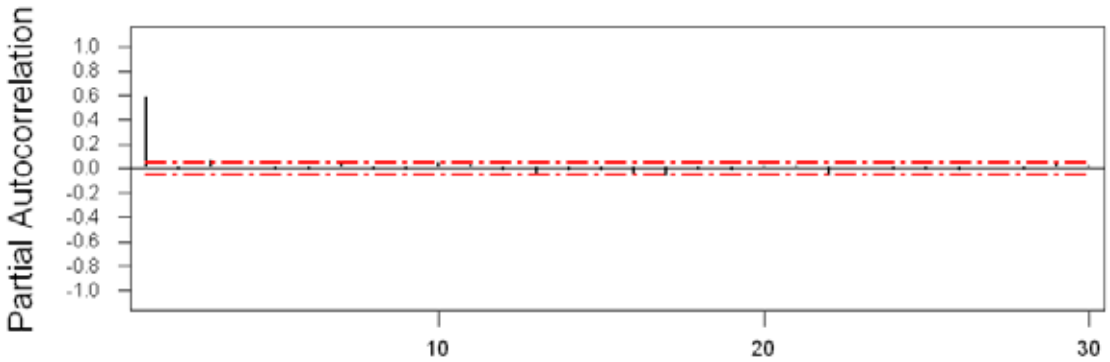
< Figure 4-2: The periodogram of the PM_{10} concentrations over period >



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



< *Figure 4-4: The ACF plot of* n_t >

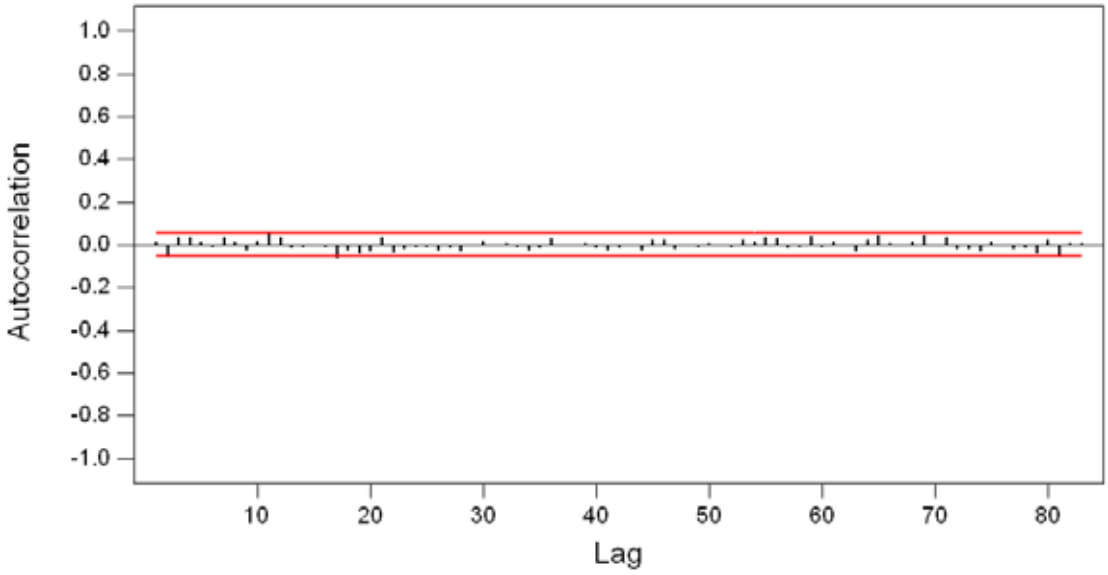


< *Figure 4-5: The PACF plot of* n_t >

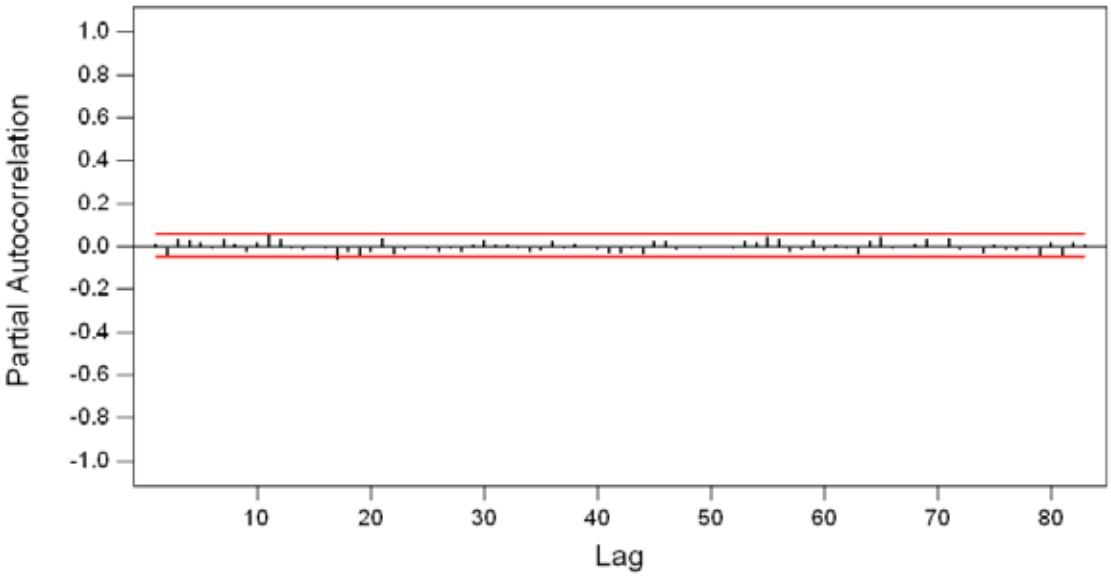
Table 4-1: Estimates of parameters of the AR(1) process

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

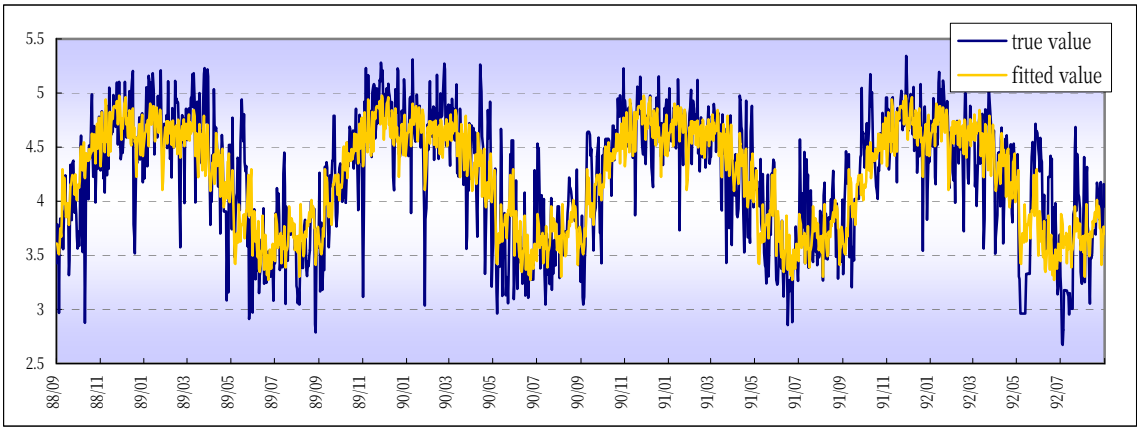
We have already built models for X_t, and the difference among them lies on the estimations of seasonality. After modeling X_t, we next want to find out which one performs better. We make a comparison among these models at the aspect of forecasting ability. Before that, we are supposed to give the criterion 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 Sep. 01 to Oct. 31, 2003.

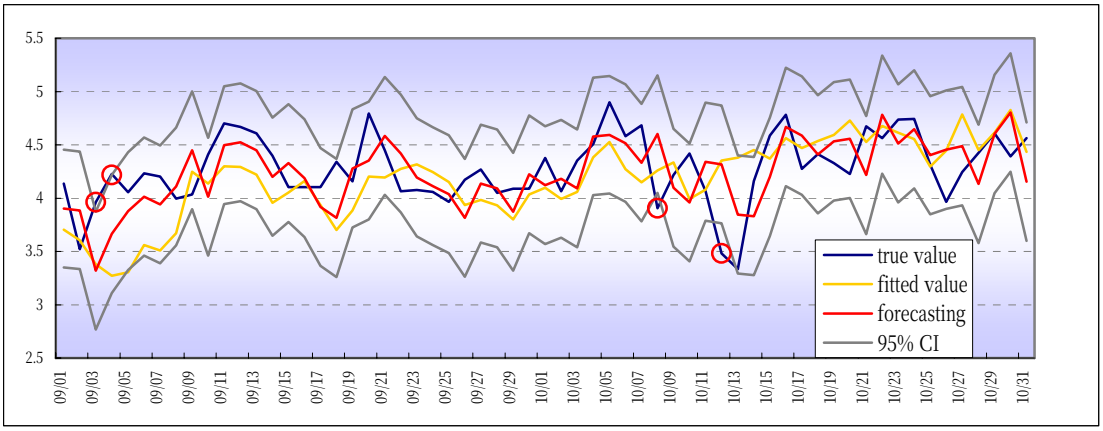
1.Model Derived from Small Trend Method

The model is given by

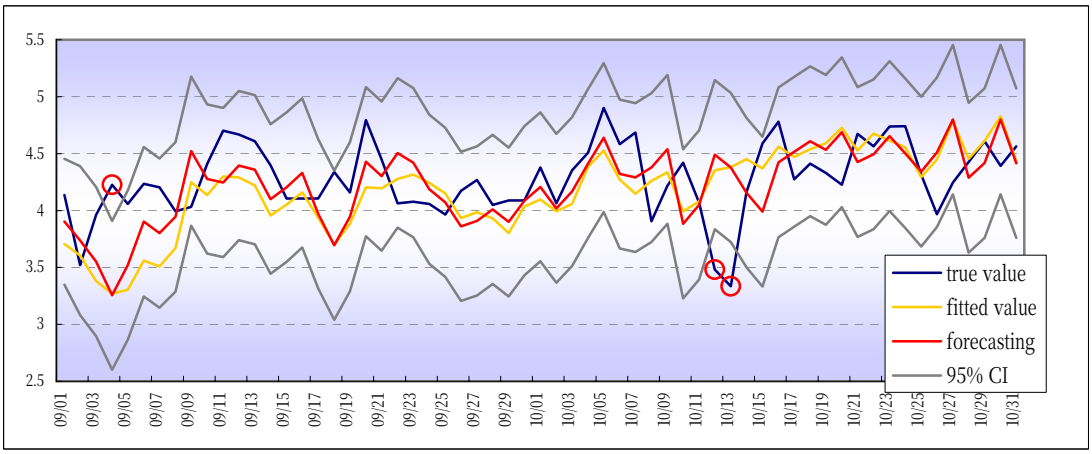
$$X_t = m_t + s_t + (1 - 0.4637B - 0.0841B^3)^{-1}\eta_t, \quad \eta_t \sim WN(0, \sigma_\eta^2)$$



< 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 >

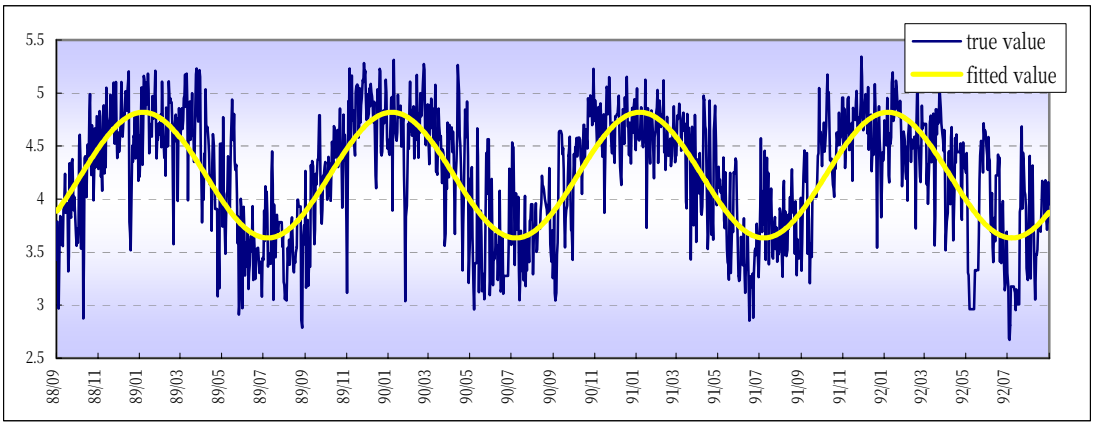
Table 5-1: Prediction results — The Small Trend Method

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

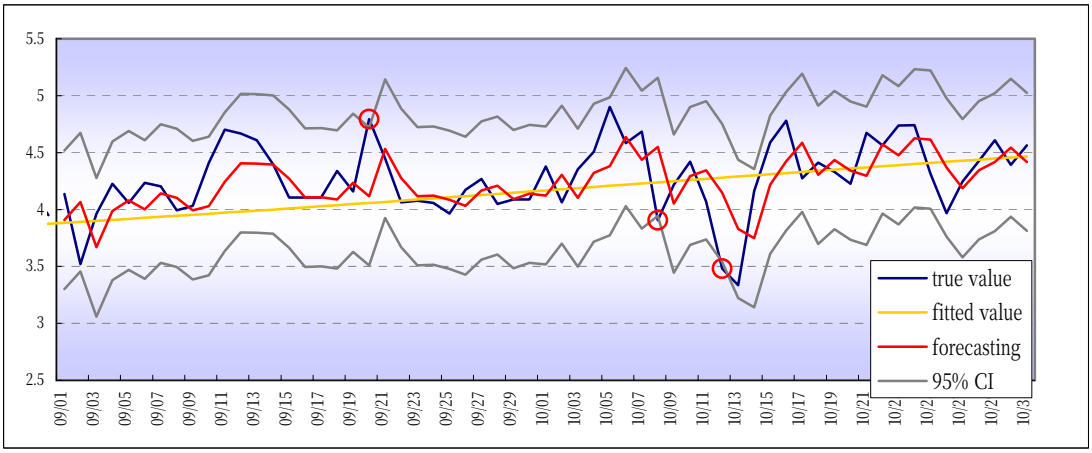
2.Model Derived from OLS Method

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

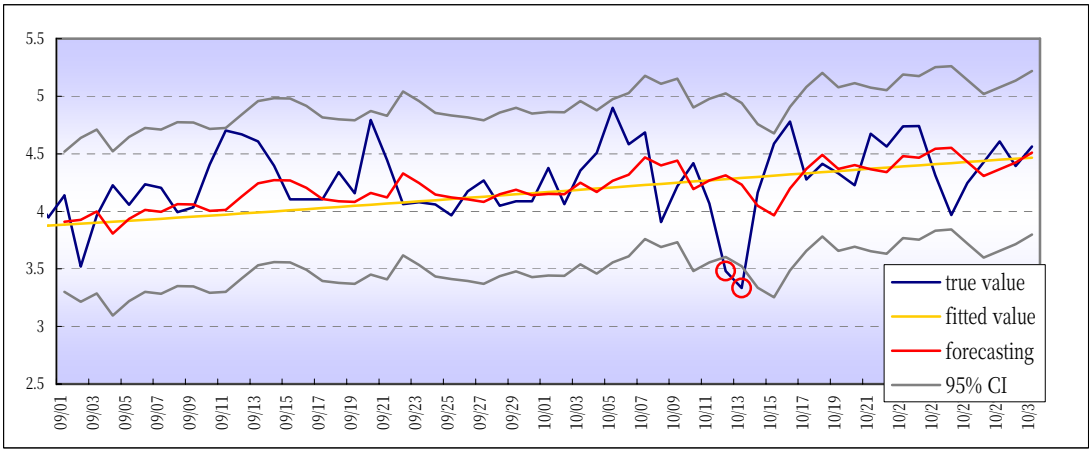
$$e_t \sim WN(0, \sigma_e^2)$$



< *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* >

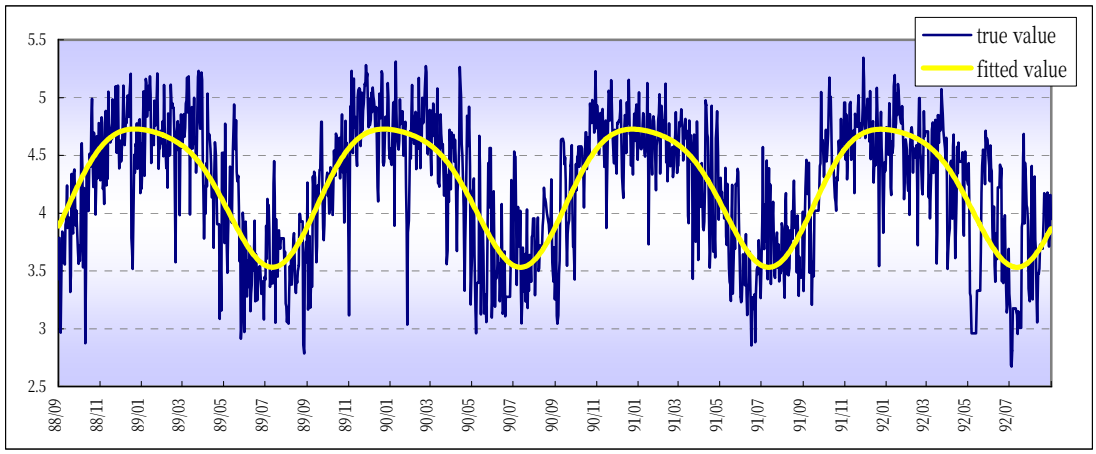
Table 5-2: Prediction results — The OLS Method

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

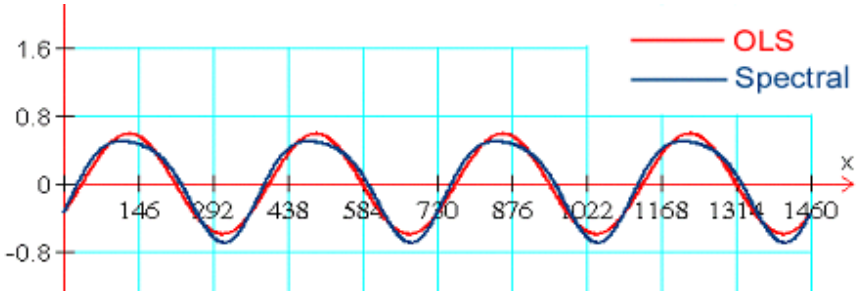
3.Model Derived from Spectral Analysis

The model is given by

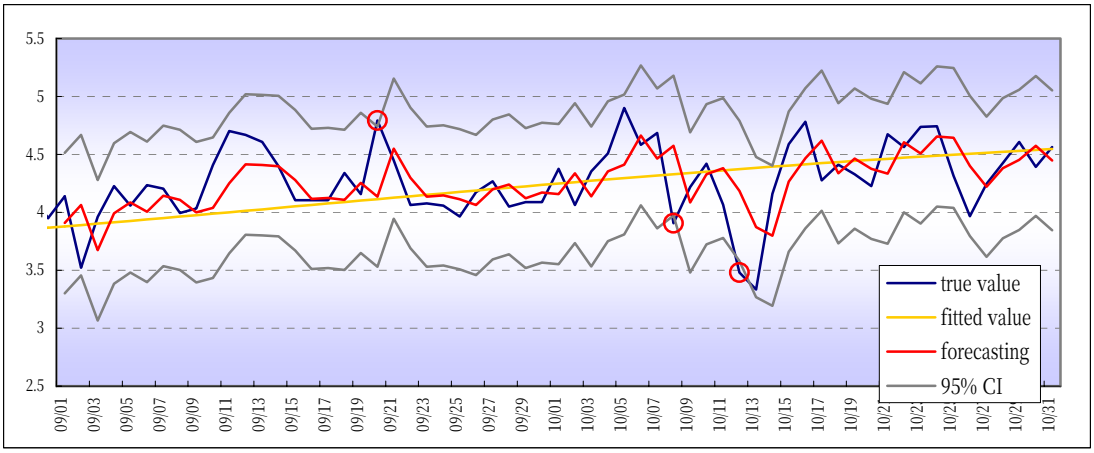
$$X_t = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t)$$
$$+0.004236\cos(0.03443t) + 0.1064\sin(0.03443t) + (1-0.5886B)^{-1}\xi_t,$$
$$t = 0,1,...,1459$$



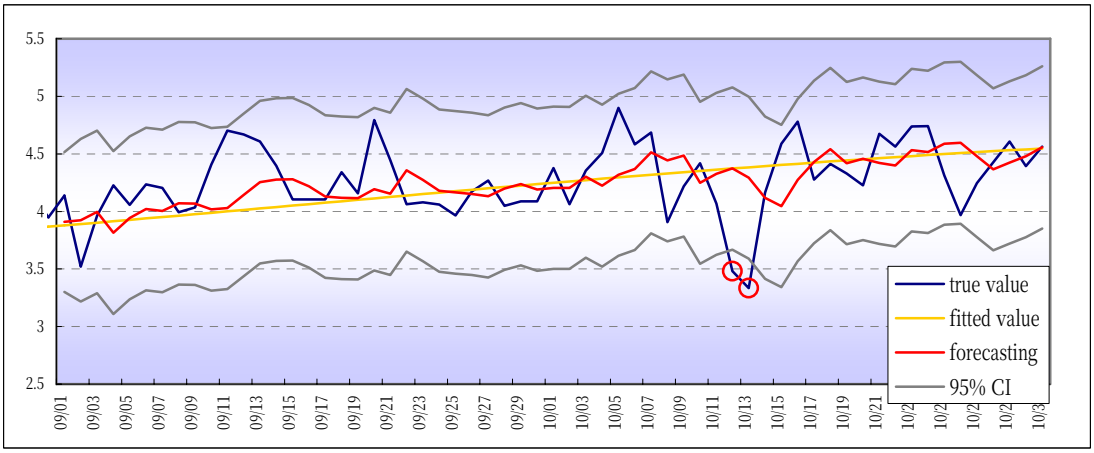
< Figure 5-7: True values VS Fitted values — Spectral Analysis >



< 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

For simplicity, let the model obtained from the small trend method be model 1; from OLS method, model 2; from spectral analysis, model 3.

Because model 1 contains the average values of the past four years, it is easily affected by some extreme values. For this reason, the predicted values of model 1 represent larger fluctuations as we can notice from figures 5-2 and 5-3; moreover, owing to the fluctuations, it easily makes errors when we forecast.

As for model 2 and model 3, the fluctuations of the predicted values are smaller. For model 2, the predicted value is mainly changing with its past three values while for model 3 the predicted value is mainly varying with its past one realization. Therefore they make fewer errors than model 1 when we forecast.

In table 5-4, when giving prediction, model 1 has the largest MSE and the most outliers, which shows the poorest forecasting ability. Finally, for we build model 3 with the consideration of the half-year seasonality component (the difference is shown in figure 5-8), the MSE of model 3 is smaller than that of model 2. Also, the 95% CI of model 3 is narrower than that of model 2. We therefore make a little improvement on our model by losing 2 degrees of freedom. But, as for the number of outliers, model 3 in fact has as many as model 2 has.

VII. More on the Topic

Lin-Yuan and Chao-Chow(潮州鄉) are both towns lying at southwest Taiwan; Lin-Yuan is near the seashore while Chao-Chow is near the mountains. The monsoon is blowing from the southwest to the northeast in spring and summer, but conversely in autumn and winter. The monsoon from the southwest can directly blow into inner Taiwan, however, the northeasterly monsoon would be blocked by Central Mountains. From the geographic view, we guess that the suspended particulate is moving from the southwest to the northeast all over the year. Therefore we wonder if the PM_{10} concentrations in Lin-Yuan could be a leading indicator of that in Chao-Chow.

We plan to build a transfer function model for Lin-Yuan and Chao-Chow. Let Y_t be the PM_{10} concentrations observed in Chao-Chow. The time series plot of Y_t holds a similar pattern to that of X_t, so we employ spectral analysis to estimate its seasonal component. In order to avoid the redundancy, we directly show the fitted model

$$Y_t = 4.2463 - 0.3912\cos(0.0172t) + 0.5289\sin(0.0172t)$$
$$-0.0033\cos(0.03443t) + 0.1477\sin(0.03443t) + \varsigma_t, \quad t = 0,1,2,...,1459$$

where ς_t denotes the error term.

We have already built a model for X_t with spectral analysis, which is given by

$$X_t = 4.226 - 0.34338\cos(0.01721t) + 0.4831\sin(0.01721t)$$
$$+0.004236\cos(0.03443t) + 0.1064\sin(0.03443t) + (1-0.5886B)^{-1}\xi_t,$$
$$t = 0,1,...,1459$$

We have to fit ς_t with AR(1) process with the same coefficient of X_t. This leads ς_t to equal $(1-0.5886B)^{-1}\upsilon_t$. To be concise, we rewrite $X_t = \widehat{X}_t + (1-0.5886B)^{-1}\xi_t$ and $Y_t = \widehat{Y}_t + (1-0.5886B)^{-1}\upsilon_t$ respectively.

However, the CCF of $\{\upsilon_t\}$ 、 $\{\xi_t\}$ shows no leading relationship. We guess this is because the distance between is not great enough to make a significant lag, that is, the impact of Lin-Yuan on Chao-Chow cannot last for more than one day.