Chapter 5

Application to Air Pollution Data

5.1 Introduction

Ambient ozone pollution in urban areas constitutes a very important environmental problem. While the decreasing stratospheric ozone layer may lead to increased instances of skin cancer, high ambient ozone intensity has been shown to cause damage to the human respiratory system as well as to agricultural crops and trees (Lefohn and Runeckles (1987), Lippmann (1989)). Ground-level ozone has been studied quite intensively recently and many data analyses involving a number of statistical methods have been used. For example Cox and Chu (1992) and Smith and Huang (1993) studied daily maximum ozone concentration using extreme value theory while Guttorp, Meiring and Sampson (1994) used a space-time model to analyze ground-level ozone data.

In this thesis, we are interested in finding an efficient way to estimate the underlying periodic trend of ground-level ozone. Ozone measurements are correlated across time and between measuring sites (spatially), but here, we will analyze each site individually, so we will only consider correlation across time. In order to test the methods explained in the previous chapters, we will apply them to a data set of daily and monthly ozone levels and will compare the results with some well known and commonly used methods of analyses.

The pollutant of interest to us, O_3, is a *secondary* pollutant. There are two main kinds of air pollutants: *primary* and *secondary* pollutants. Secondary pollutants are produced by chemical reactions between pollutants and other constituents while primary pollutants are emitted directly by identifiable sources. O_3 is produced by oxidation of primary pollutants. This oxidation is driven by ultra-violet radiation coming from sunlight and comprises chemical reactions that are dependent on the temperature. Because the chemical reactions proceed while the air is being carried by winds, secondary pollutants are generally more widespread than primary pollutants. Therefore, secondary pollutants are sometimes refered to as *regional* pollutants. Due to the dependence of the governing chemical reaction on temperature, O_3 levels are high in early afternoon and midsummer and low overnight and in the winter. Also, ozone has a strong yearly pattern which is due to the fact that the creation of ozone is highly related to solar radiation.

5.2 Air Pollution Data

The data set used for the analysis includes the daily maximum of hourly levels of nitrogen dioxide, ozone (O_3), sulphur dioxide and the daily mean of nitrate and sulfate. These measurements were recorded in the southern portion of Ontario during the six year period from January 1 of 1983 to December 31 of 1988. Several monitoring networks in the province were used to collect the data, including the Environment Air Quality Monitoring network (OME), Air Pollution in Ontario Study (APIOS), and the Canadian Acid and Participation Monitoring Network (CAPMON). The reader is referred to Burnett et al (1994) for a complete description of the data set. Furthermore, some climatic data like temperature, humidity, pressure and wind speed were also available. Over all three monitoring networks, 37 monitoring locations (sites) are available, but not all sites monitored all five pollutants (24 sites monitored ozone levels). In the analysis to come, we assume that the variation caused by different networks is negligible, i.e., we consider that the three networks function in an equivalent manner and so we will not include network as a factor in the analysis. However, since only the ozone data are of interest to us, from now on, we will not consider the other four pollutants.

5.2.1 Daily Ozone Data

The daily ozone data cover a period of six years. Thus, for every site where the ozone levels were monitored, a complete daily ozone level time series would be made of a total of 2192 observations. However, some of the daily levels time series contain an excessive number of missing values. So, in order to control the quality of the data, all time series with more than one-third of their values missing are not included in the analysis. We assume that the probability of a time series having more than one-third of missing values is not related to those values. Therefore, we are not causing any bias by adopting such a strategy. Following that stategy, the number of monitoring sites for ozone level was reduced from 24 to 18. For those 18 sites, approximately eleven percent of the observations are missing. We assume that the missing data include randomly missing data and censored missing data, which occur when the actual air pollution level is below the detection limit of a monitor. Although penalized smoothing splines do not require equispaced data, the smoothing parameter selection technique MCV_l would not be clear with missing data points. Indeed, the value of l determines how many data points to leave out. Hence, if data points were missing, it would not be clear if we should leave out $2l + 1$ data points or $2l + 1$ days (or months). Therefore, in order to keep things simple, we use *imputation* to fill in missing data so that data points and days (months) have the same meaning. There exist many imputations methods like overall mean imputation, class mean imputation, hot-deck and cold-deck imputation, regression imputation, EM algorithm based imputation and multiple imputation. For more details on imputation methods, see Sarndel, Swensson and Wretman (1992) and Little and Rubin (1987). In this thesis, we used the technique of Sun (1994), which is based on multivariate normal theory. Sun analysed the same data set in a Bayesian interpolation setting. It is assumed that this imputation procedure will not cause serious bias since the percentage of missing data is small.

Daily ozone levels show a strong serial correlation along with a periodic yearly pattern. However, the variances of the residuals may not be constant, which results in a violation of the assumptions of this thesis. Indeed, by examining the top plot of Figure 5.1, we see that the plot of the residuals of the original data versus the fitted values at the arbitrary site #4 shows an obvious presence of heteroscedasticity. The fitted values were obtained using the BCV method. The lower plot in Figure 5.1 presents the residuals of the square root transformed data at the same site. There is an evident advantage to proceed to this tranformation as it gets rid of most of the heteroscedasticity in the residuals. Therefore, we are closer to satisfying the assumptions of homoscedasticity of the residuals. Also, the normal quantile-quantile plots shown in Figure 5.2 clearly demonstrate that the residuals obtained from the transformed data behave more normally than those obtained from the original data. A log transformation was also tried but it did not give satisfying results. Therefore, from now on, we will consider the square root transform of the daily O_3 levels. By examining the plot of the autocorrelation function (ACF) of the residuals of the transformed data (see the top plot of Figure 5.3), we can see that there is still some evidence of seasonality, hence there may be heteroscedasticity in the residuals. However, this seasonality is pretty small and is negligible in the estimation of the regression function. The partial autocorrelation function (PACF) plot of the residuals (see bottom plot of

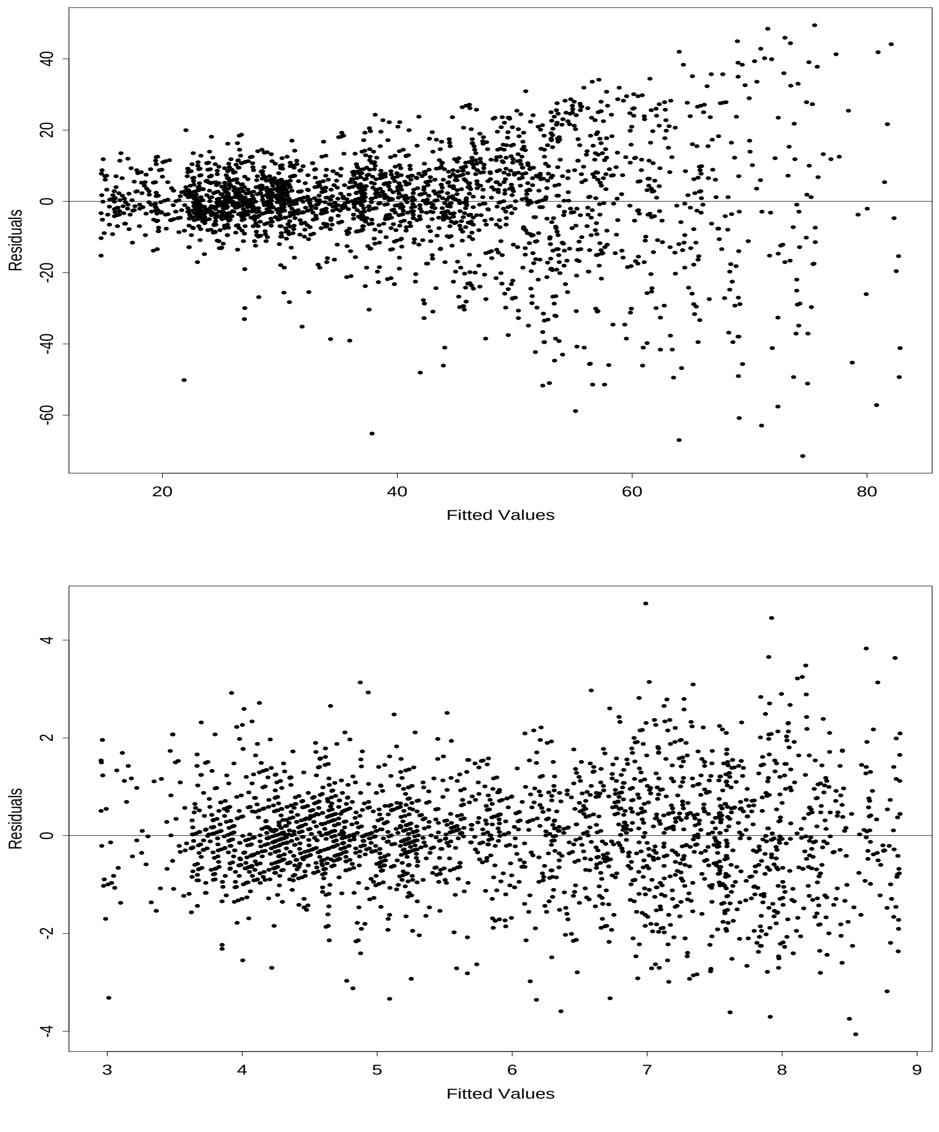

Figure 5.1: *Residuals of Original (top plot) and Square Root Transformed (bottom plot) Daily Levels of O_3 at Site #4.*

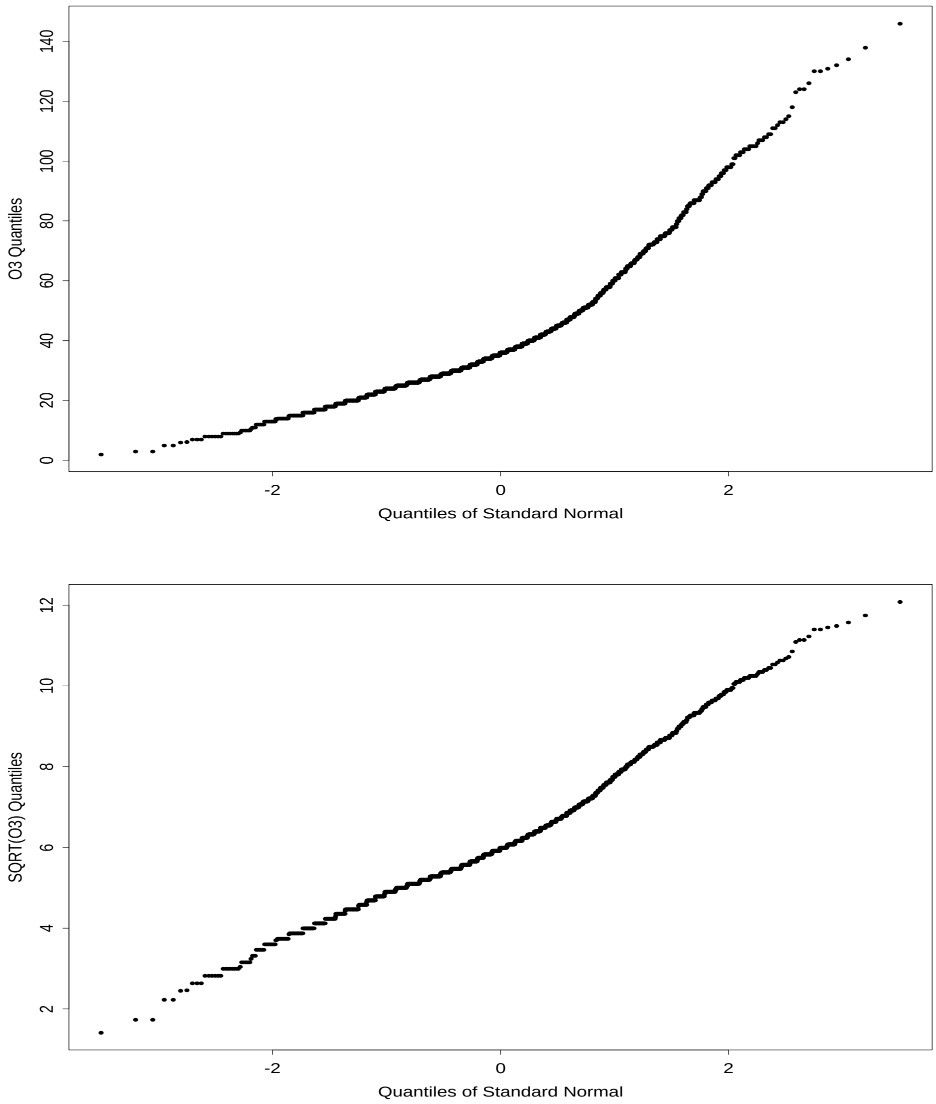

Figure 5.2: *Normal Quantile-Quantile Plots of Residuals for Original (top plot) and Square Root Transformed (bottom plot) Daily Levels of* O_3 *at Site #4.*

Figure 5.3), shows a significant serial correlation. In fact, the residuals of all sites follow an AR model with orders between 2 and 71 days. The example in Figure 5.3 shows a correlation following an AR(45) model.

5.2.2 Monthly Ozone Data

Monthly pollution levels are computed as the mean of observed daily levels in that month. Hence, since our data set covers a period of six years, a complete monthly ozone level time series would consist of 72 data points. By using the 18 sites that were kept for the daily data, we proceeded to compute the monthly data, using the original daily data. For those 18 sites, approximately two percent of observations ended up missing. However, since our monthly data have such a small proportion of missing data, the choice of an imputation method is in fact not critical as it might only cause negligible bias. So we filled in a value missing from the i^{th} site in the j^{th} month with the mean of the observed values of the same site in the j^{th} month of the other years. This method allows us to preserve the periodic property as we know that ozone has a strong seasonal pattern. For more detail on the imputation of the data set, please refer to Sun (1994).

For the monthly data, the fitted values used in computing the residuals were also obtained via the BCV method. Although monthly ozone levels have a weaker correlation structure than the daily ozone levels, they have an obvious seasonal pattern. This is shown in figure 5.4 as the ACF plot shows a slight seasonal pattern while the PACF plot suggests an AR model of order 9 for

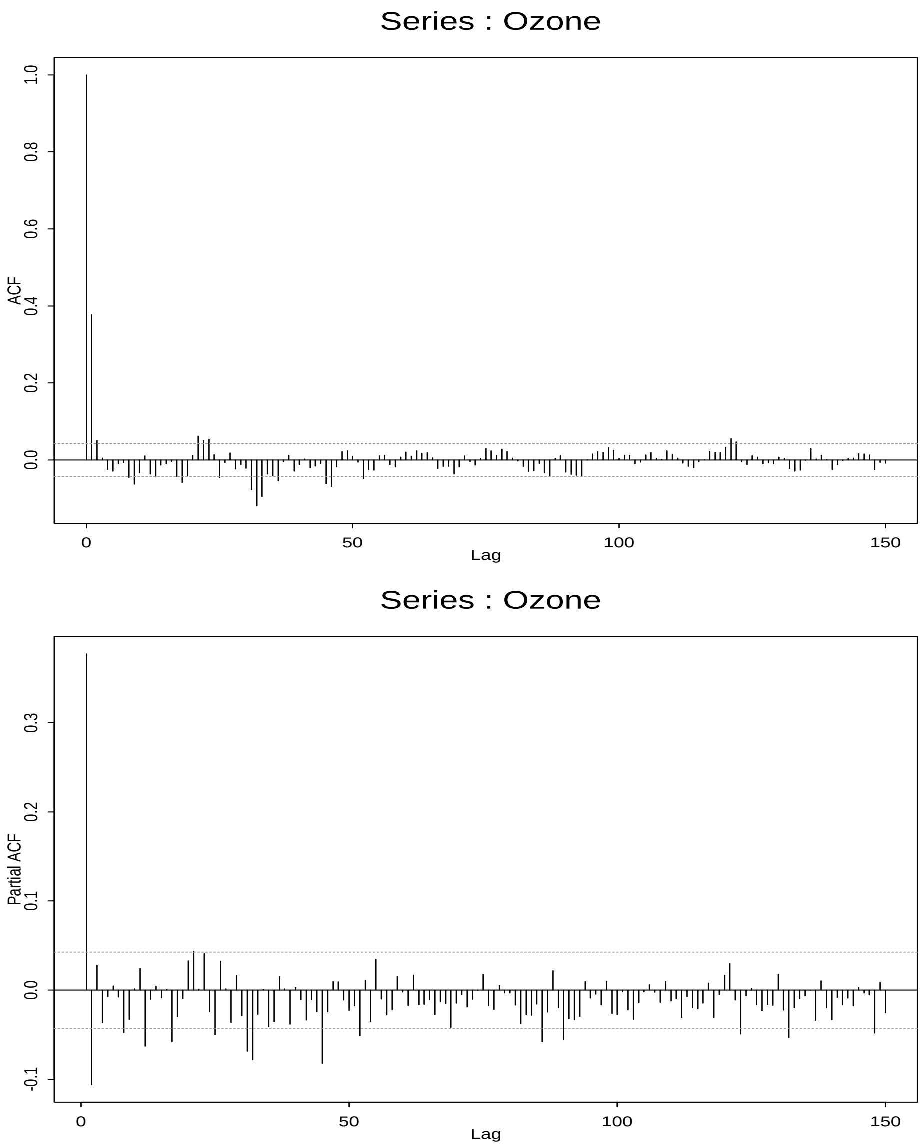

Figure 5.3: *Plot for Autocorrelation and Partial Autocorrelation of Square Root Transformed Daily O_3 Levels at Site #4.*

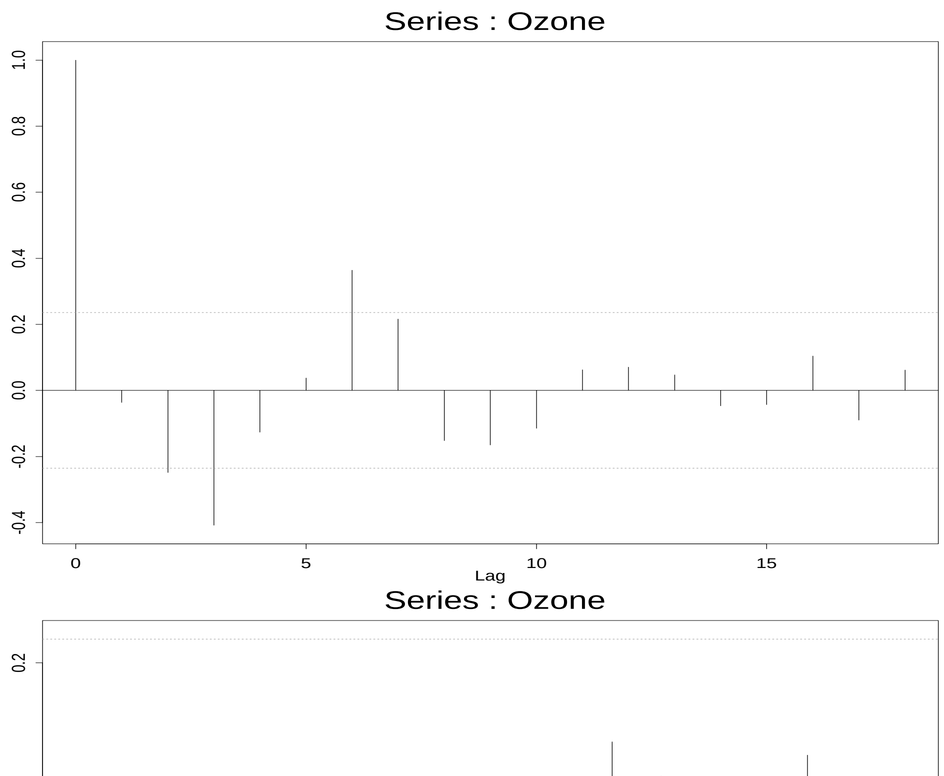

Figure 5.4: *Plot for Autocorrelation and Partial Autocorrelation of Monthly O_3 Levels at Site #4.*

site #4. Overall, PACF indicates that the order of the AR model for each of the sites varies from one to fifteen months while the ACF conserves the same periodic pattern. The seasonal pattern shown in the ACF plot could be evidence of heteroscedasticity in the residuals. However, as seen in Figure 5.5, the residuals of the monthly O_3 levels are better distributed than the ones coming from the tranformed data. The same conclusion arises from looking at Figure 5.6, the normal quantile-quantile plots. Therefore, since the residuals of the untransformed data seem more homoscedastic, we will carry out the analysis with the untransformed monthly O_3 levels data.

5.3 Methods of Analysis

To perform the analyses of the data set, we used all six methods employed in the simulation study. We used the same penalty as in Chapter 4. Since ozone is known to have a strong seasonal pattern and a yearly cycle, we respectively assumed $\omega = 2\pi/365$ and $\omega = 2\pi/12$ for the period of the daily and monthly ozone levels. However, with the NLS method, the period was estimated. The histogram of the estimated ω's for the daily and monthly data will be presented shortly. Due to ozone's yearly cycle and to the fact that each time series is over a period of six years, the number of blocks chosen for the BCV method was six where each block represents a calendar year. We therefore assume that ozone levels on a given site are roughly independent from year to year, with the years being replicates.

The monthly data were analyzed for all 18 stations using each of the

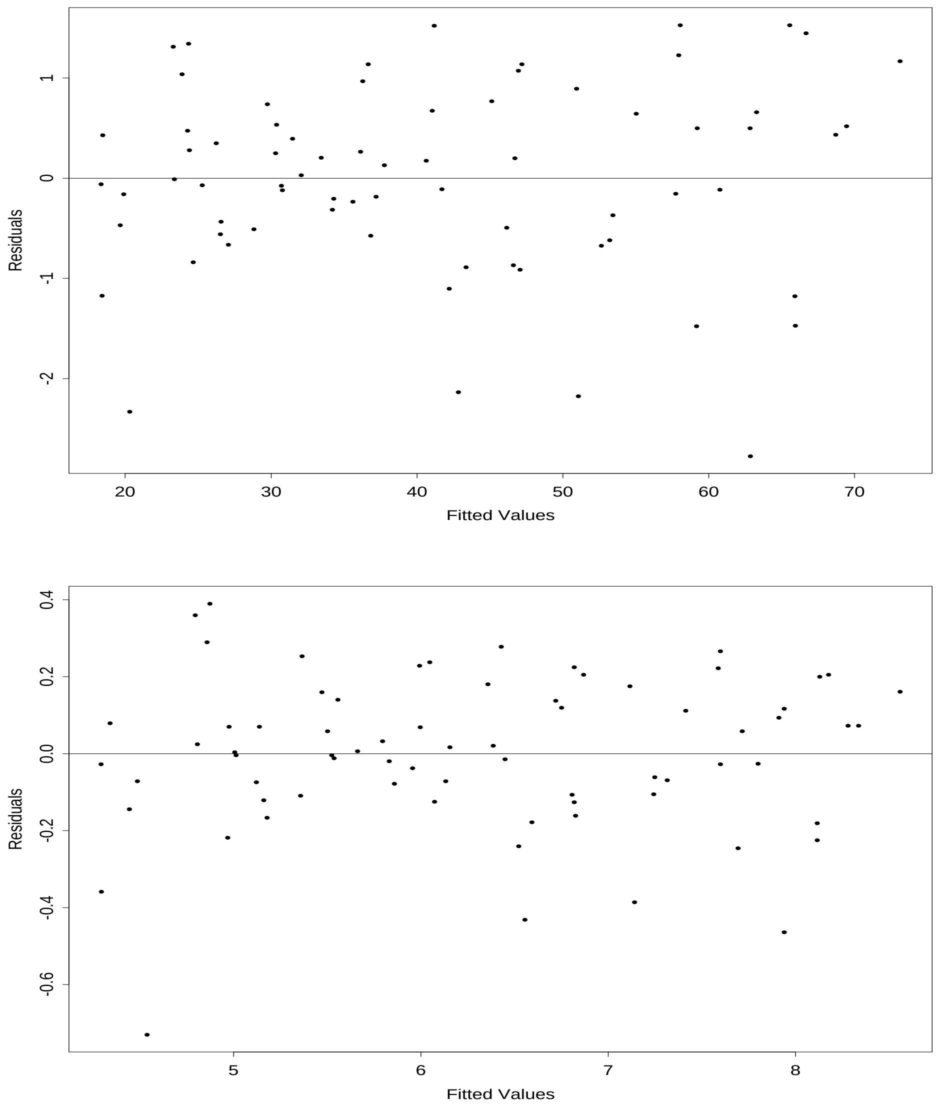

Figure 5.5: *Residuals of Original (top plot) and Square Root Transformed (bottom plot) Monthly Levels of O_3 at Site #4.*

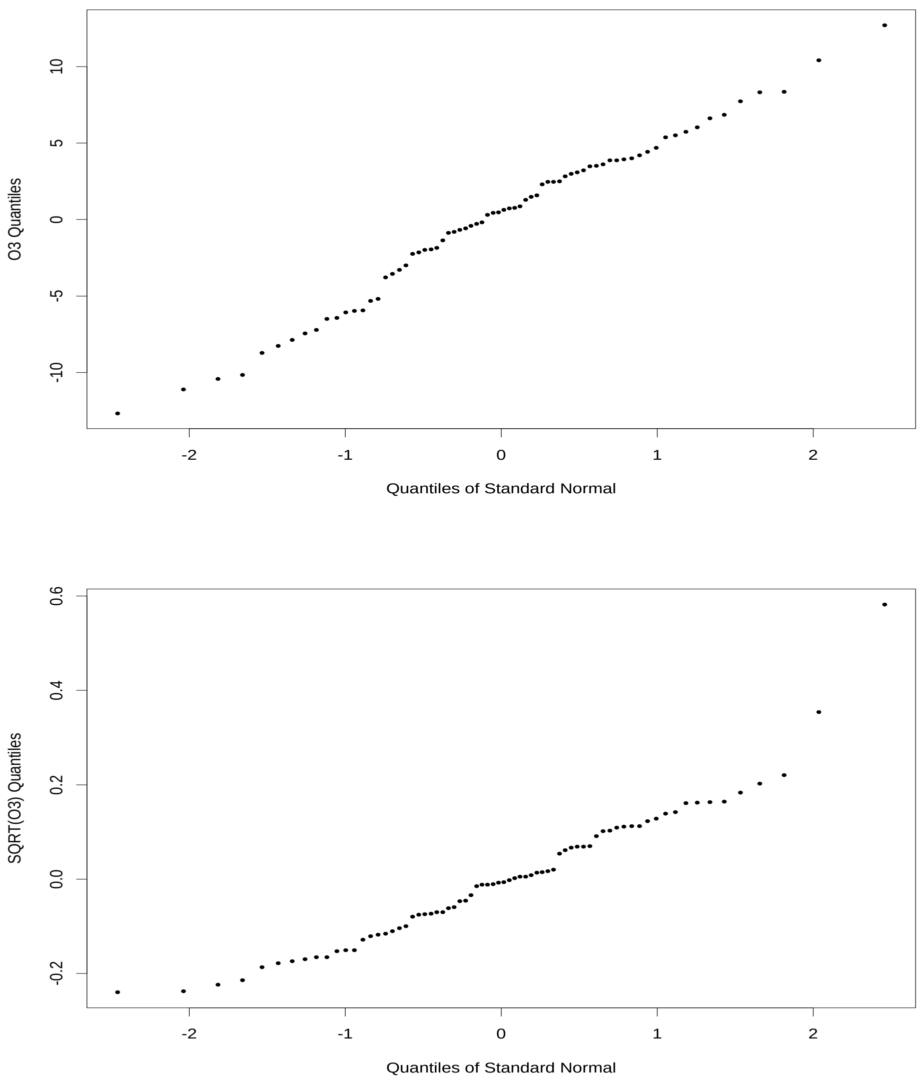

Figure 5.6: *Normal Quantile-Quantile Plots for Original and Square Root Transformed Monthly levels of O_3 at Site #4.*

methods. The daily data were analyzed for all sites using NLS, LS, IND and BCV. However, because of the large size of the daily data set (2192 data points per site), using MCV_l is very time consuming. Therefore, due to time restrictions, we only applied MCV_l to three randomly chosen sites for the values of $l = 1$ and $l = \hat{p}$, where $\hat{p}$ is the estimated order of the AR model of the residuals of the chosen site. The randomly selected sites were $\#2, \#7$ and $\#17$ and the residuals of these sites respectively follow an AR model of order 51, 41 and 37. Furthermore, due to computational restrictions of the memory usage by the Splus software, we could not apply the time series method to the daily ozone level data.

Since we are analysing a real data set, the true underlying regression function is unknown, thus we can not compute the MSE summary statistic. Therefore, quantifying the performance of each of the methods, in terms of estimating the trend, is slightly more complex than with the simulations. However, a good estimation method would be one that would not track the data too closely, hence not using too many parameters. Furthermore, an estimate using too few parameters would not be good either as it would oversmooth the data.

The results of the analyses with regards to the PSE and the results concerning the number of parameters used by the different penalized methods are shown in the next two sections. However, the PSE scores are not really of interest for the purpose of this work. Remember that our goal is to estimate the underlying trend of the data. While the PSE score quantifies the prediction peformance of each estimator, it does not say anything about the ability of the methods to estimate the regression function. However, in general, a method that yields a small PSE score will be using a large number of parameters as the estimate tries to interpolate the data. Thus, generally a small PSE score would mean an estimate that is too "wiggly".

5.4 Results

5.4.1 Monthly O_3 Levels

Plots of the data and the fitted curves for sites #2, #7 and #17 are given at the end of this chapter as Figures 5.12-5.20. The plots of the fits for all remaining sites are shown in Appendix B.

Overall, MCV_l appears to be the method that performs the best. For the values of $l = 1, 2$ and 3 considered in this work, there was no definite pattern indicating an optimal l. However, MCV_l alleviates more and more the effect of the correlation as the value of l increases. Therefore, a data set with a stronger correlation structure should be analysed with a larger value of l since the correlation structure has more influence on the estimation of the regression function. Therefore, choosing a value of $l = \hat{p}$, where $\hat{p}$ is the estimate of the autoregressive order of the residuals, might be a good choice, that is, it might lead to an estimate that fits the data well. Indeed, we verified that rule with sites #2, #7 and #17, and for those three sites, the order of the autoregressive model is a good indication of the value of l to choose in order to get a good curve. Here, “good” is defined as the fit that subjectively appears to track the data closely enough without interpolation.

In the analysis of the monthly data, some values of l led to estimates using very parameters, thus yielding estimates that were really close to the parametric model. Therefore, due to their similarity with the parametric model, each of these estimates has roughly the same amplitude for all six yearly cycles. For some sites, these estimates are not fitting the data well since, at certain years, the estimated regression function lies above the highest points of the data. At the end of this section, we will present detailed analyses of three sites where such a problem arise.

BCV generally tracks the data too closely as it is driven by the similarity in the yearly cycles of each site. IND, as concluded in the simulation study, is highly influenced by the correlation structure and usually interpolates the data. The parametric methods, since they are restricted to a parametric family of models, are not flexible enough and they are unable to track the finer features of the data. Usually, LS, TS and NLS perform very similarly. However, in the case where the estimate of ω is not close to $2\pi/12$, NLS differs quite substantially from the other two parametric methods. Figure B.7 in the Appendix shows such an example at site #4. Indeed, the estimated period at that site is 0.3928, hence yielding an estimate with five cycles. Because the period was not well estimated, the estimate given by NLS is very different than that of LS or TS.

Now let us give summary information and a more detailed discussion of

PSE and the number of parameters used.

First of all, a paired t-test similar to the one used in Chapter 4 showed that NLS had a significantly higher mean PSE score than all the other methods as it is, on average, oversmoothing the data. We can also notice that the NLS

Table 5.1: *Monthly Levels of O_3 Mean Results.*

Method	PSE (SE)	log(PSE) (SE)
NLS	86.4 (21.3)	3.93 (0.25)
LS	33.3 (3.2)	3.43 (0.10)
TS	34.6 (3.1)	3.48 (0.09)
IND	8.2 (1.3)	1.81 (0.23)
BCV	11.1 (2.4)	2.16 (0.16)
$\mathbf{MCV}_1$	24.1 (2.7)	3.05 (0.14)
$\mathbf{MCV}_2$	25.6 (2.3)	3.17 (0.10)
$\mathbf{MCV}_3$	25.1 (1.8)	3.18 (0.07)

method is far more variable than any of the other methods used. The reason for such a poor performance and high variability can be explained by the fact that for six sites out of eighteen, NLS did not estimate ω very well. Indeed, Figure 5.7 shows clearly that for those six sites, the ratio of the estimated values of ω and $2\pi/12$ was around 0.8. This had the effect that NLS yielded, for those six sites, an estimate of the regression function with five cycles instead of six, thus the poor performance of NLS and its high variablility.

Table 5.2 gives the average effective number of parameters used by the different penalized methods. This table does not include the parametric methods as these use a fixed number of parameters each time. TS and LS each use three parameters while NLS uses four parameters as it is estimating the

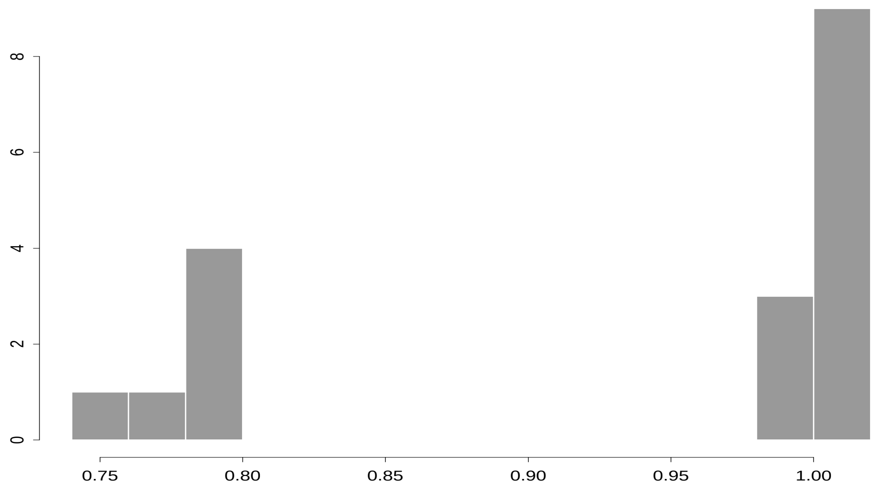

Figure 5.7: *Ratio of* $\hat{\omega}$ *and* $\omega = 2\pi/12$ *for Monthly Levels of* O_3.

period ω. The effective number of parameters used is very useful in quantifying the performance of the methods. Indeed, recall that a smooth estimate, close to the *true* regression function, should use a small number of parameters but not too small, while an estimate that would be too wiggly, hence trying to interpolate the data, would use a larger number of parameters.

First of all, we can see that IND and BCV use on average around 4 times more parameters than any other penalized method. This is confirmed by the average PSE of these two methods as they are significantly lower than any other methods used. For IND, the fact that it uses too many parameters is explained by the effect of the correlation in the data. Indeed, the correlation causes the estimate given by IND to quite often interpolate the data, hence yielding a low PSE score and equivalently, a high number of parameters used.

Table 5.2: *Number of Parameters Used for Monthly Levels of O_3.*

Method	**mean**	**median**	**S.E.**
IND	31.1	31.3	2.6
BCV	28.47	30.6	2.54
MCV$_1$	7.9	5.8	1.6
MCV$_2$	5.8	4.4	0.7
MCV$_3$	6.3	5.2	0.8

The reason why BCV is using a high number of parameters is due to the nature of the monthly data. Indeed, the yearly cycles of sixteen out of the eighteen site are very regular, hence the six groups used in BCV, for a given site, are very similar. This has the effect that BCV, for those sixteen sites, uses a small smoothing parameter which is equivalent to a high number of parameters. For the two remaining sites #7 and #17, BCV uses a large smoothing parameter which yields an estimate using approximately three parameters. Figure 5.8 shows, for sites #2, #7 and #17, the six sections that BCV uses to minimize its criterion. One can see that for site #2, the six sections appear to be very similar. Indeed, the beginning and the end of each section are very much at the same heights while the shape of the function in each section is pretty similar, that is, it is increasing in the first half of each section and then decreasing for the last second half. This explains why BCV uses, for site #2, 31.2 parameters and tends to interpolate the data. For sites #7 and #17, the reason why BCV uses so few parameters can be explained by looking at the top right hand-side plot and the lower plot of Figure 5.8. Indeed, for site #7, one can clearly see that within each of the six sections, the function starts and ends at different heights and does not behave similarly within each section. Therefore, BCV uses a large smoothing parameter to obtain an "average" curve that best fits the six different curves. The data from site #17 are similar but not as extreme. The number of parameters used by BCV is more variable than the MCV_l methods. This is so because for most sites, BCV yields an estimate using approximately 30 parameters while for the other two sites, it uses only three parameters.

Comparing the different values of l, MCV_1 uses significantly more parameters than both MCV_2 and MCV_3 while the last two do not significantly differ. This agrees with what was seen in the simulation study, that is, the average number of parameters used by MCV_1 is larger than the averages used by MCV_2 and MCV_3.

Table 5.3 shows, for all 18 sites, the number of parameters used by each method in addition to the estimated autoregressive order of the residuals calculated from the fitted values obtained with BCV. The orders were estimated using the *acf* function in Splus. As we can see, both IND and BCV use a large number of parameters. The number of parameters used by MCV_1 is quite variable as it is more sensitive to the correlation structure. Indeed, since it is using a value of $l = 1$, it does not eliminate all the effect of correlation. Also, it appears that the number of parameters used by MCV_1 tends to be higher than MCV_2.

Table 5.4 displays the correlation between the number of parameters used by the 5 penalized methods. The relationship is very similar to what

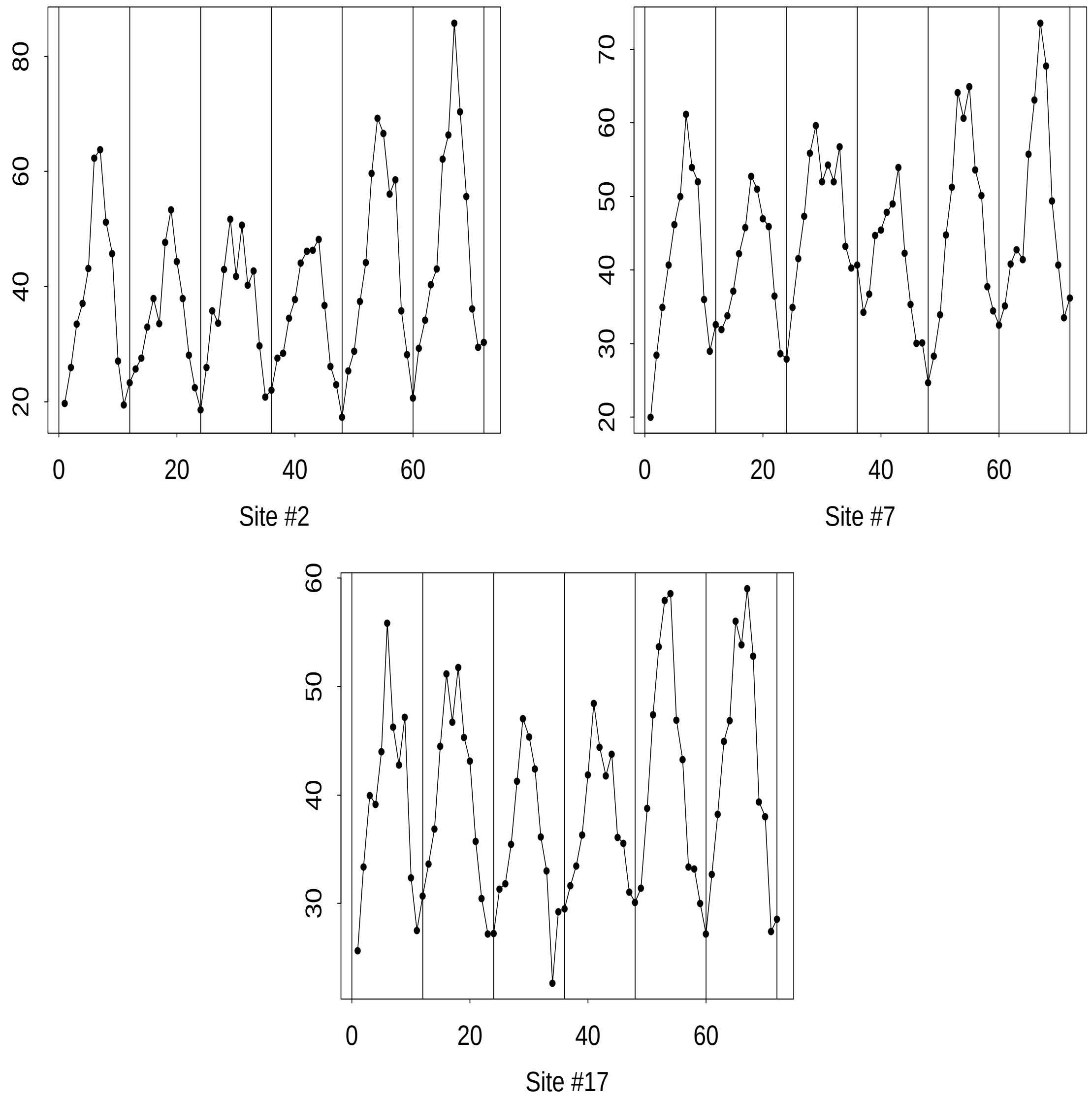

Figure 5.8: *Behaviour of BCV for Site #2 (top left), Site #7 (top right) and Site #17 (bottom).*

Table 5.3: *Number of Parameters used by each Method for the Monthly Data.* (a * represents a site that is studied in Chapter 5)

Site	**AR order**	**IND**	**BCV**	**MCV$_1$**	**MCV$_2$**	**MCV$_3$**
1	14	35.5	35.7	3.4	3.4	4.1
$*2$	12	29.8	31.1	10.9	10.6	10.1
3	14	44.0	27.4	4.1	4.9	7.9
4	15	32.5	35.7	9.3	9.0	12.2
5	13	59.2	30.1	31.0	3.9	4.2
6	14	28.0	29.4	3.9	3.5	3.4
$*7$	1	31.1	3.1	12.9	8.4	6.3
8	14	12.4	41.8	3.5	3.2	3.1
9	15	31.6	31.1	6.4	6.2	7.8
10	13	36.2	27.4	3.4	3.3	3.4
11	9	28.7	16.8	5.2	5.2	7.9
12	14	32.9	41.8	3.5	3.7	7.5
13	14	28.0	30.1	6.9	6.7	7.7
14	14	32.5	29.4	3.4	3.4	3.5
15	15	40.2	32.3	8.3	3.5	3.5
16	15	29.0	33.2	8.9	9.3	12.0
$*17$	1	14.8	3.0	13.8	12.6	3.1
18	13	13.9	33.2	3.9	3.7	3.1

was observed in the simulation study with the exception of the negative correlation between IND and MCV$_2$ which is very different from what was seen in the simulation study. However, if we examine closely this relationship (see Figure 5.9), we see that at two sites, namely sites #5 and #17 which are represented by an asterix in Figure 5.9, IND and MCV$_2$ respectively use an exceptionally high number of parameters. In fact, if we remove these two sites, the correlation goes from -0.268 to a positive value of 0.0468.

Let us now look closely at three specific sites and discuss the perfomances of the different estimation methods. We will use sites # 2,7 and 17 as

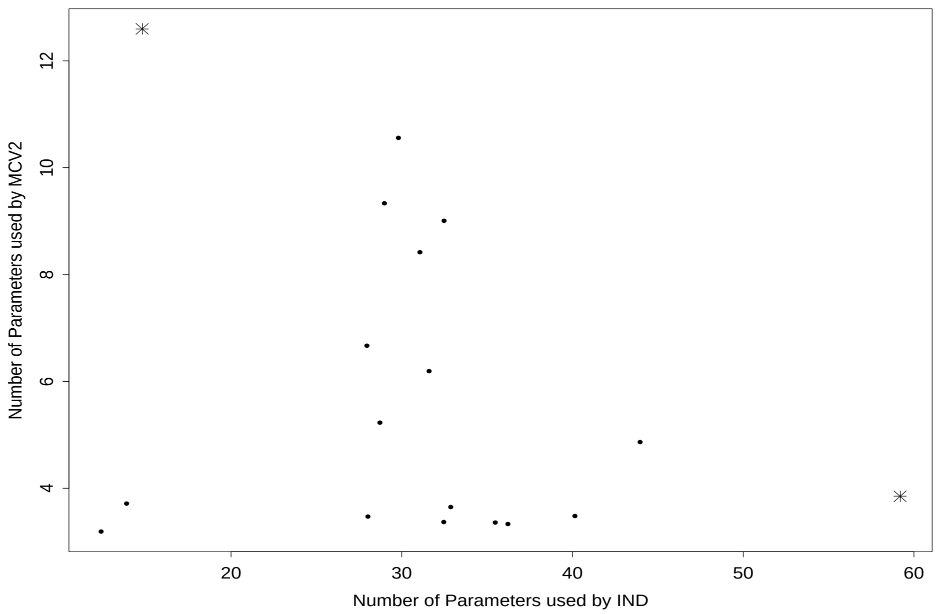

Figure 5.9: *Relationship between the Number of Parameters Used by IND and* MCV_2.

Table 5.4: *Correlation Between Number of Parameters Used for Monthly Levels of O_3 .*

METHODS	**BCV**	**MCV$_1$**	**MCV$_2$**	**MCV$_3$**
IND	0.111	0.497	-0.268	0.105
BCV	-	-0.290	-0.521	0.128
MCV$_1$	-	-	0.313	0.062
MCV$_2$	-	-	-	0.577

we want to look at the same sites that will be analysed more intensively for the daily O_3 levels data.

Site #2

Figure 5.12 displays the fit of the three parametric methods. While the three fits are very comparable, we can see that LS and NLS are practically identical. This is explained by the fact that the ratio of the estimated value of ω obtained by NLS and $2\pi/12$ is 0.997. Therefore, NLS and LS are basically the same method. However, in Figure 5.12 we see that the TS estimate has a slightly smaller amplitude than NLS and LS. That difference can be accounted for by the fact that TS takes the correlation into account while the two others methods do not.

Figure 5.13 shows the fits of the penalized methods that take correlation into account. We see that the MCV$_l$ fits are virtually identical for $l = 1, 2, 3$, and yield estimates with effective numbers of parameters of 10.9, 10.6 and 10.1 respectively. Since the AR model that fits the residuals has an order of 12 months, and in order to verify the rule of selection for l, we also show the estimate of MCV_{12}. This fit is different from the other MCV_l fits as it uses 20.9 parameters and it appears to fit the data better. Indeed, when looking at Figure 5.13, we see that at the peak of the fourth year (around the 42^{nd} month), all the MCV_l fits but MCV_{12} exceed the highest data point. Generally, MCV_{12} tracks the data more accurately while using a reasonable number of parameters.

The other method that takes correlation into account, BCV, uses 31.2 parameters and tends to interpolate the data. This result is due to the fact that the cycles of the data at site #2 are very similar, hence BCV uses a small smoothing parameter to minimize its criterion. See Figure 5.8 for a representation of the cycles used by BCV.

The most interesting plot is Figure 5.14 which shows the fits of IND, NLS and MCV_{12}. There, it becomes clear that the parametric methods are handicapped compared to the penalized methods. Indeed, NLS has to keep the same amplitude for each cycle (year) and therefore does not yield a good estimate when the peaks and valleys are not constant from year to year. On the other hand, IND and MCV_{12} yield a comparable fit. However, we have more confidence in the fit of MCV_{12} since it's taken correlation into account. Furthermore, MCV_{12} yields an estimate that interpolates the data less than IND.

Site #7

Figure 5.15 shows the plots of the parametric methods at site #7 and leads us to the same conclusions as for site #2. At site #7, the ratio of the estimated

ω and $2\pi/12$ is 1.006 which is once again very close to one. Therefore, here too, LS and NLS are basically the same method.

Figure 5.16, which shows the fits of the penalized methods that take correlation into account, is a little bit more interesting than the correspoding figure for site #2. Indeed, this time, as it is clearly shown in Figure 5.16, the fits of the MCV$_l$ methods are noticeably different. BCV used 3.01 parameters while MCV$_l$, $l = 1, 2, 3$, use respectively 12.9, 8.4 and 6.3 parameters. This can be explained by the fact that the correlation for the residuals of site #7 follow an AR model of order one with an estimated autocorrelation coefficient of 0.63. This setting is very similar to our simulation study when $\phi = 0.5$, and we do obtain very similar results indeed. To see the behaviour of BCV for site #7, please refer back to Figure 5.8.

The best fit among the MCV$_l$ methods appears to be MCV$_1$ as it is the only fit whose height does not exceed the data points at months 20 and 42. Indeed, it seems that MCV$_2$ and MCV$_3$ oversmooth the data. This agrees with the decision rule that leads to the choice of $l = \hat{p}$. Indeed, since the correlation structure is of short range, that is AR(1), MCV$_1$ takes into account a good amount of correlation.

Figure 5.17 shows once again that MCV$_1$ performs very well, while IND still interpolates the data too much and TS is unable to discern peaks of different amplitudes due to its parametric nature.

Site #17

In Figure 5.18, we see once more that the parametric methods perform similarly. The ratio of the estimated ω and $2\pi/12$ is 1.003, which means that the period estimated by NLS was very close to $2\pi/12$.

In Figure 5.19, we can see that MCV_3 and BCV perform very similarly. In fact, MCV_3 uses approximately 3.14 parameters while BCV uses 3.01 parameters. This means that those two methods behave very much like the parametric methods of Figure 5.18. BCV uses so few parameters since, as shown in Figure 5.8, the six cycles of the data are pretty different.

The two remaining methods in Figure 5.19, MCV_1 and MCV_2, use respectively 13.8 and 12.6 parameters and produce much better fits. The correlation for site #17 follows an AR(1), hence this can explain why MCV_3 oversmooths the data while MCV_1 and MCV_2 seem to do much better.

In Figure 5.20, we can see that IND and MCV_1 are very similar. The fact that IND does not try to interpolate the data is somehow surprising. However, this can be explained by the fact that the residuals at site #17, as mentioned above, are not highly correlated. Therefore, the effect of the correlation at that site is not very strong, hence the similarity in the estimates obtained by IND and MCV_1.

5.4.2 Daily O_3 Levels

Plots of the data and the fitted curves for sites #2, #7 and #17 are given at the end of this chapter as Figures 5.21-5.29. The plots of the fits for all remaining sites are shown in Appendix C.

Table 5.5: *Square Root Transformed Daily Levels of O_3 Mean Results.*

Method	PSE (SE)	log(PSE) (SE)
LS	2581.0 (89.5)	7.85 (0.04)
NLS	2568.1 (89.2)	7.84 (0.04)
IND	683.5 (40.1)	6.50 (0.06)
BCV	2330.0 (83.4)	7.74 (0.04)
$\mathbf{MCV}_1$	1996.3 (10.1)	7.60 (0.01)
$\mathbf{MCV}_{\hat{p}}$	2125.2 (22.3)	7.66 (0.01)

Overall, it appears that using $\mathrm{MCV}_{\hat{p}}$, with $\hat{p}$ the estimated order of the AR model of the residuals, results in smooth estimates that track the features of the data pretty efficiently and better than the other methods. Since the correlation structure is pretty strong, MCV_1 does not alleviate the correlation effect sufficiently, hence yielding estimates that are not smooth enough. BCV, as concluded in the simulation study and in the monthly data analysis, is too dependent on the cycles in the data and therefore is pretty unreliable. Finally, IND, which is highly influenced by the strong correlation structure in the residuals of the daily data, performs very poorly as it interpolates the data.

The results of the daily data analysis with regards to the PSE are shown in Table 5.5 while the results concerning the number of parameters used by BCV, IND, MCV_1 and $\mathrm{MCV}_{\hat{p}}$ are shown in Table 5.6. Recall that the summary statistics for MCV_1 and $\mathrm{MCV}_{\hat{p}}$ are based on only three sites: #2, #7 and #17.

First of all, we can see that NLS and LS do not significantly perform

Table 5.6: *Number of Parameters Used for Square Root Transformed Daily Levels of O_3.*

Method	**mean**	**median**	**S.E.**
IND	884.6	875.1	19.1
BCV	29.24	29.34	3.60
$\mathbf{MCV}_1$	71.72	72.10	2.61
$\mathbf{MCV}_{\hat{p}}$	26.37	26.10	3.18

differently. This can be explained by the fact that, as shown in Figure 5.10, the estimation of ω by NLS gave $\hat{\omega}$'s very close to the value used for LS, that is, $\omega = 2\pi/365$. Remember that in the case where $\hat{\omega} = 2\pi/365$, LS and NLS will yield exactly the same estimate as they are the same method. So since the ratio is close to one, the methods use a similar period and therefore perform quite comparably.

A good way of assessing the performance of the penalized methods, IND and BCV, is to look at the effective number of parameters that these methods use. IND uses significantly more parameters than BCV. This is explained by the fact that IND highly interpolates the data as it is strongly influenced by the correlation structure. Indeed, looking at the effective number of parameters used by IND in Table 5.6, we see that, on average, IND uses more than 880 parameters compared to slightly more than 29 for BCV. BCV uses surprisingly very few parameters. Remember that a daily ozone level time series has 2192 data points while each monthly time series only has 72 data points. Nevertheless, for the daily data, BCV uses approximately the same number of parameters as for the monthly data. This can be explained by the fact

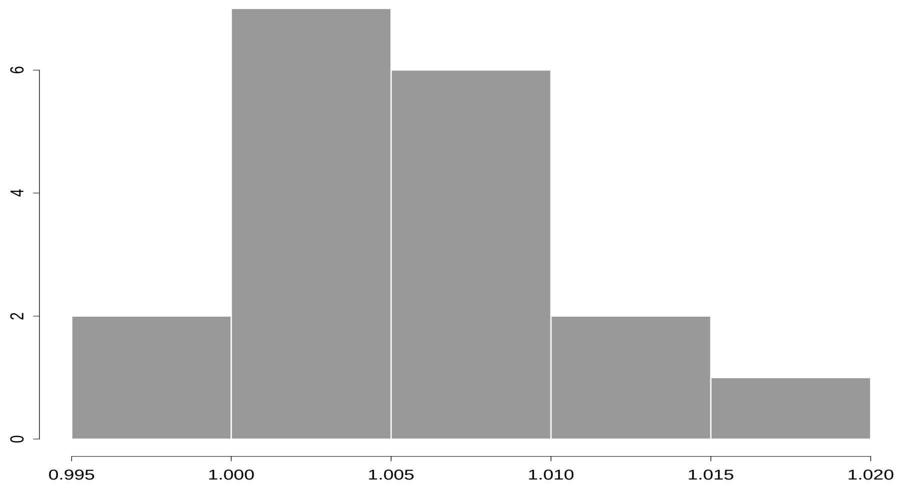

Figure 5.10: *Ratio $\hat{\omega}$ and $\omega = 2\pi/365$ for Square Root Transformed Daily Levels of O_3*

that with the monthly data, BCV tends to interpolate the data. On the other hand, in the daily data, the yearly cycles in a given site are generally very different from each other as they have lots of noise that must be smoothed for prediction. This causes BCV to use a large smoothing parameter to minimize the BCV criterion, hence using very few parameters.

Table 5.7 shows, for all the 18 sites of the daily data, the number of parameters used by each method in addition to the estimated autoregressive order of the residuals.

Let us now look more in depth at the fits of the differents methods when applied to the three randomly chosen sites. Figure 5.11 shows the fit of the IND method at site #2. Since IND performed similarly at all sites, we will

Table 5.7: *Number of Parameters used for the Daily Data.* (a * represents a site that was studied in Chapter 5)

Site	**AR order**	**IND**	**BCV**	**MCV$_1$**	**MCV$_{\hat{p}}$**
1	58	801.9	35.9	-	-
*2	51	859.2	29.3	76.0	32.0
3	2	771.0	27.8	-	-
4	45	960.5	29.3	-	-
5	33	890.0	29.3	-	-
6	26	817.9	3.0	-	-
*7	41	894.6	67.9	67.1	21.0
8	30	903.7	29.4	-	-
9	32	964.2	24.4	-	-
10	71	802.3	35.9	-	-
11	71	783.7	41.6	-	-
12	34	840.9	29.3	-	-
13	71	855.5	35.9	-	-
14	35	961.2	29.4	-	-
15	60	1093.6	35.9	-	-
16	37	866.9	3.0	-	-
*17	37	971.8	35.9	72.0	26.0
18	13	883.2	3.0	-	-

not show this method in the next figures since it only makes plots difficult to analyse.

Site #2

For this site, remember that the correlation structure follows an AR(51). Therefore, we computed the estimate of m via the MCV$_l$ method using $l = \{1, 51\}$.

Figure 5.21 shows that NLS and LS fit the periodic portion of the data very well but do not account for the increase in the amplitude of the last two

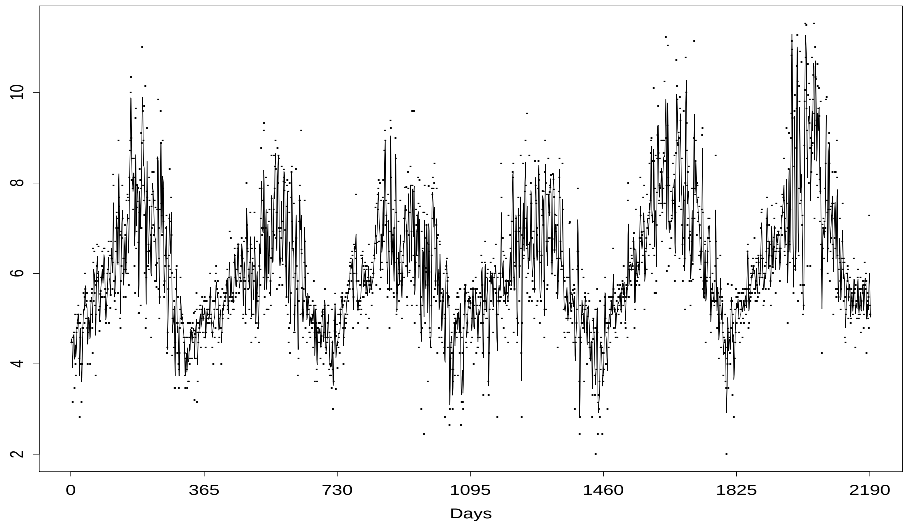

Figure 5.11: *Fit of the IND Method for the Square Root Transformed Daily Levels of O_3 at Site #2.*

peaks. NLS is clearly very similar to LS and this is supported by the fact that the ratio of the estimate period and $2\pi/365$ is 0.9955. Since LS and NLS are restricted to a parametric model in the family of $\{sin, cos\}$ functions, they cannot produce estimates with varying amplitudes. Therefore, they do not have the capacity to follow the peaks and the valleys of the data closely enough.

In Figure 5.22, we can see that MCV_1 is pretty "wiggly"as it uses 76.0 parameters. On the other hand, MCV_{51} and BCV behave similarly as the estimates use 32.0 and 29.3 parameters respectively. Both the MCV_{51} and BCV estimates are much smoother than the estimate using MCV_1 and they track the features of the data pretty well. MCV_{51} is smooth and uses fewer parameters since it really alleviates the effect of correlation by using a value of $l = 51$. BCV uses few parameters since the huge amount of noise requires a lot of smoothing in order to obtain a good prediction.

Comparing MCV_{51} and NLS (Figure 5.23) shows that MCV_{51}, while being a smooth function, has the capacity to follow more closely the fine features of the data like fluctuations in the peaks and the valleys. This is especially obvious by looking at the peaks located approximately at days 915, 1640 and 2000.

Site #7

Here, the correlation structure follows an AR(41). Therefore, we computed the estimate of m via the MCV_l method using $l = \{1, 41\}$.

As seen in the others sites, LS and NLS present almost identical fits (see Figure 5.24). This is confirmed by the ratio of the estimated period and $2\pi/365$ which is 1.002.

Figure 5.25 shows the fits of BCV, MCV_1 and MCV_{41} which use 67.9, 67.1 and 21.0 parameters, respectively. This explains why MCV_{41} is a lot smoother than the two other estimates. For reasons similar to those mentioned for site #2, MCV_{41} succeeds much better in diminishing the effect of the correlation on the estimate of the regression function. BCV is not as smooth since the six cycles for site #7 are more similar than in the previously analysed site.

In Figure 5.26 we present the comparison between the fits of MCV_{41} and NLS. While the two fits are pretty similar, once again, MCV_{41} offers more flexibility and allows its estimate to follow the fluctuations in the amplitude of the peaks.

Site #17

For site #17, the correlation structure follows an AR(37) so we computed the estimate of m via the MCV_l method using $l = \{1, 37\}$.

In Figure 5.27, we see that NLS and LS still perform almost exactly the same. Indeed, the ratio of $\hat{\omega}$ and $2\pi/365$ is 1.007. In Figure 5.28, we see that, since MCV_{37} uses 26.0 parameters, it is far smoother than MCV_1 which uses 72.0 parameters. BCV, using 35.9 parameters, is much smoother than MCV_1 but slightly more "bumpy" than MCV_{37}. By looking at Figure 5.29, we see that once again, even if NLS follows the main features of the data pretty well, it lacks flexibility compared to MCV_{37} in tracking the finer features of the data.

5.4.3 Conclusions

Overall, the method that performs best, both in the monthly and daily analyses is MCV_l. For the three sites studied in both the monthly and the daily data, $\mathrm{MCV}_{\hat{p}}$ provided the best estimate of trend, with $\hat{p}$ being the estimate of the order of the autoregressive model of the residuals. Indeed, by using $l = \hat{p}$, the estimate takes into account a good amount of correlation thus yielding a smooth estimate that still tracks the finer features of the data very well. We can therefore easily and accurately estimate the ozone level trend. BCV performs better in the daily data analysis, mainly due to the large amount of noise in that data that results in a smoother estimate. If the cycles are clearly defined, each with a lot of data points and a high degree of noise, BCV performs well but $\mathrm{MCV}_{\hat{p}}$ is still preferred. Otherwise, as with the monthly data and the simulation study, BCV is pretty unreliable. IND, as expected, performs very poorly for both the monthly and the daily data. Only when the correlation is negligeable does IND yield acceptable results. The parametric methods follow the periodic nature of the data quite faithfully but lack flexibility to track some changes in the amplitudes of the peaks and valleys.

In conclusion, the analysis of the ozone level data confirms what was seen in the simulation study, that is, that accounting for correlation via using MCV_l allows us to alleviate the effect of correlation while getting a very flexible estimate.

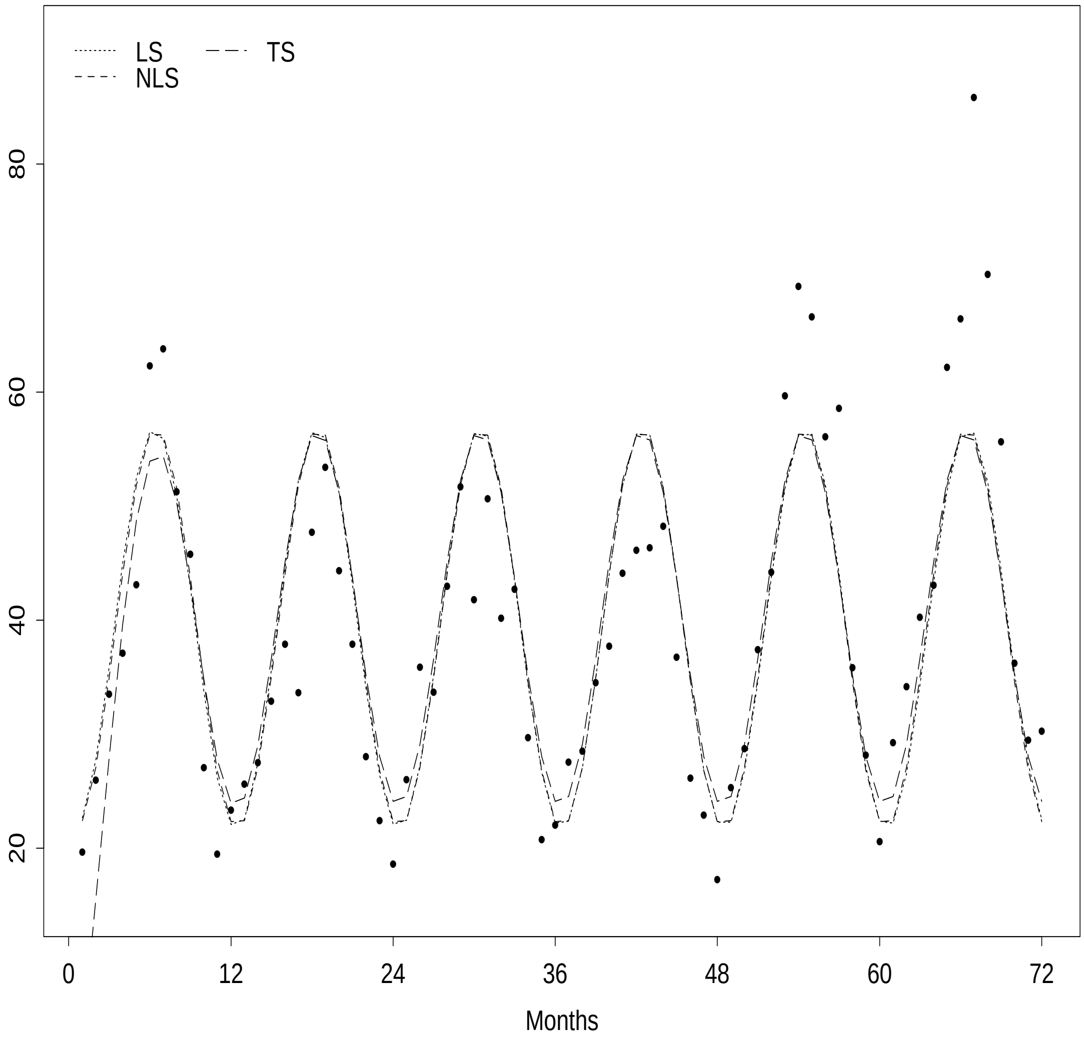

Figure 5.12: *Fits for Monthly Levels of O_3 at Site #2.*

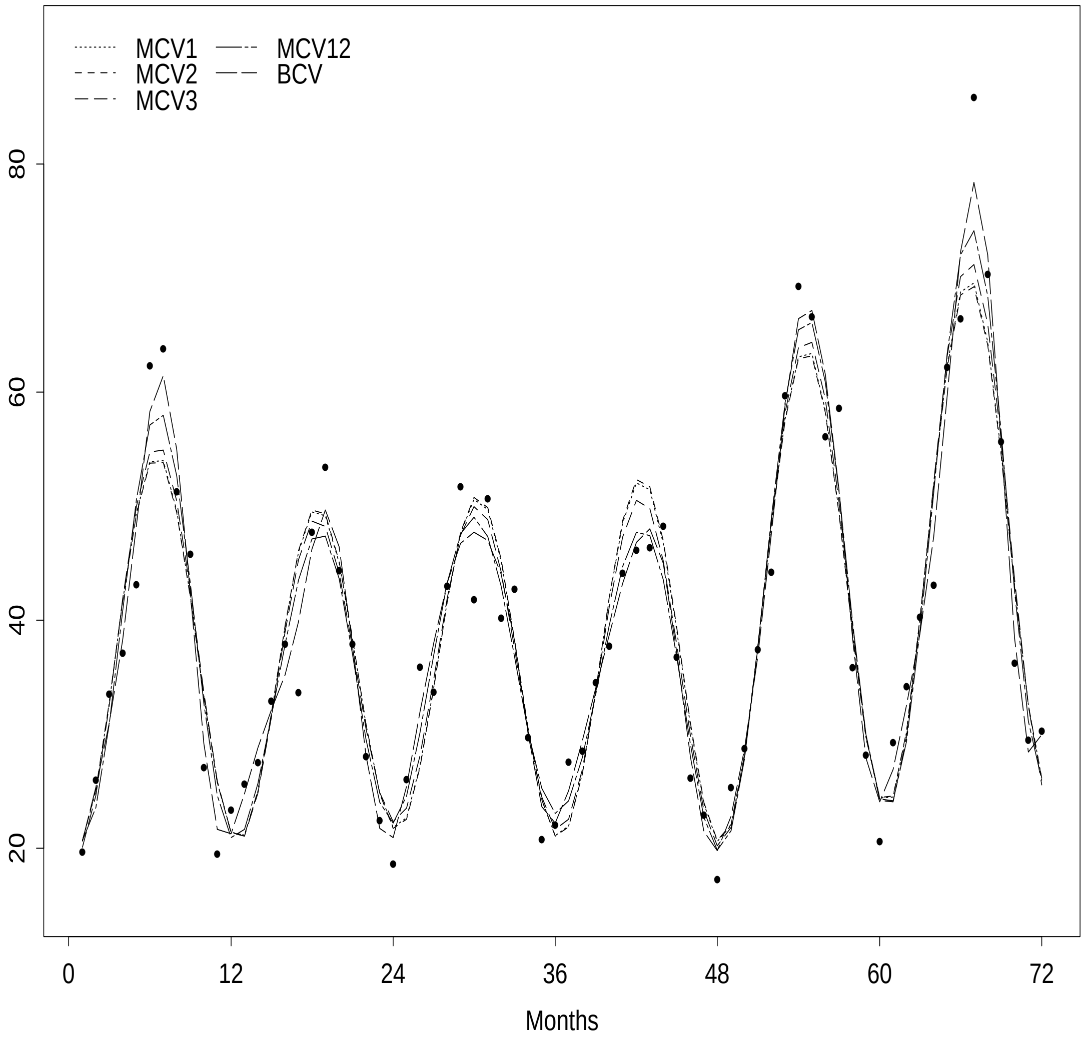

Figure 5.13: *Fits for Monthly Levels of O_3 at Site #2.*

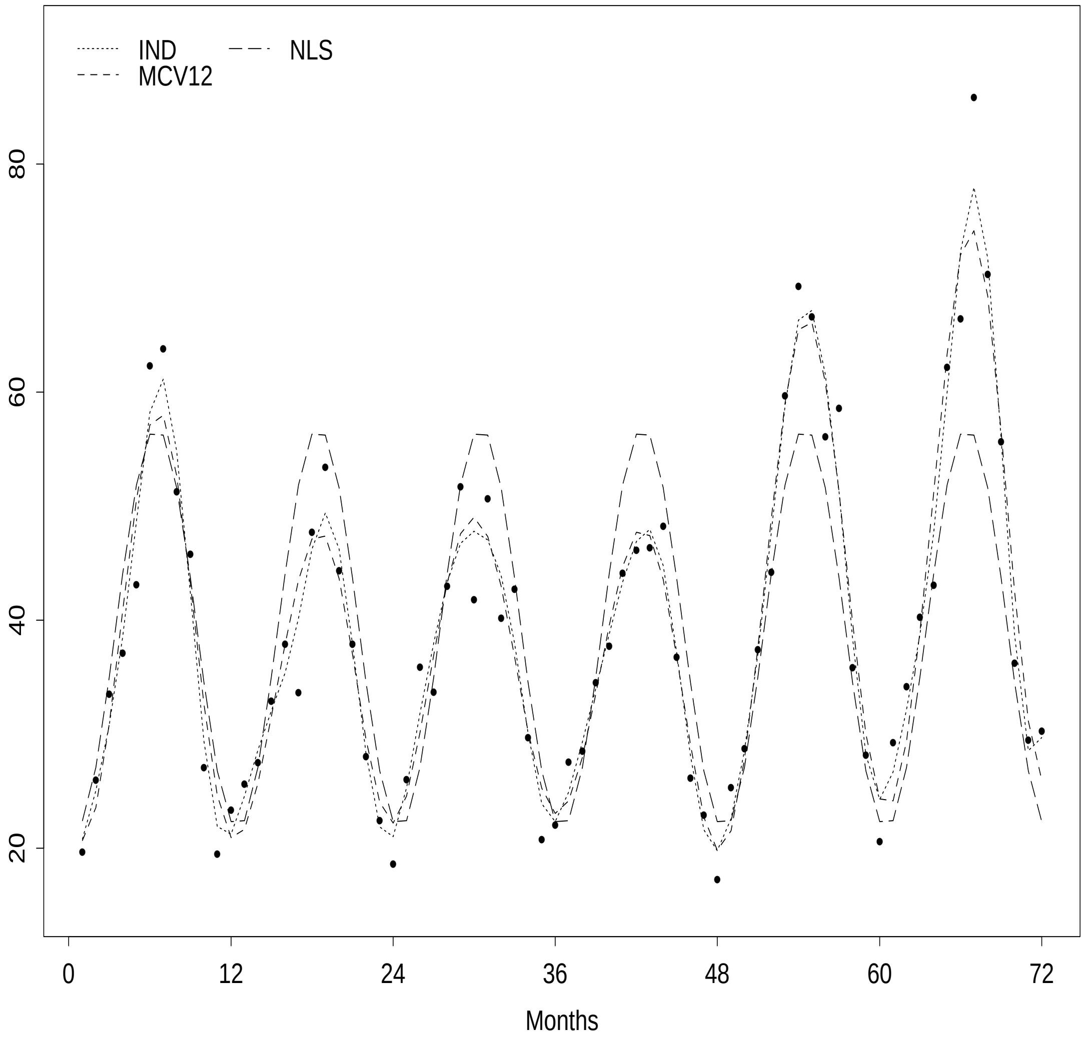

Figure 5.14: *Fits for Monthly Levels of O_3 at Site #2.*

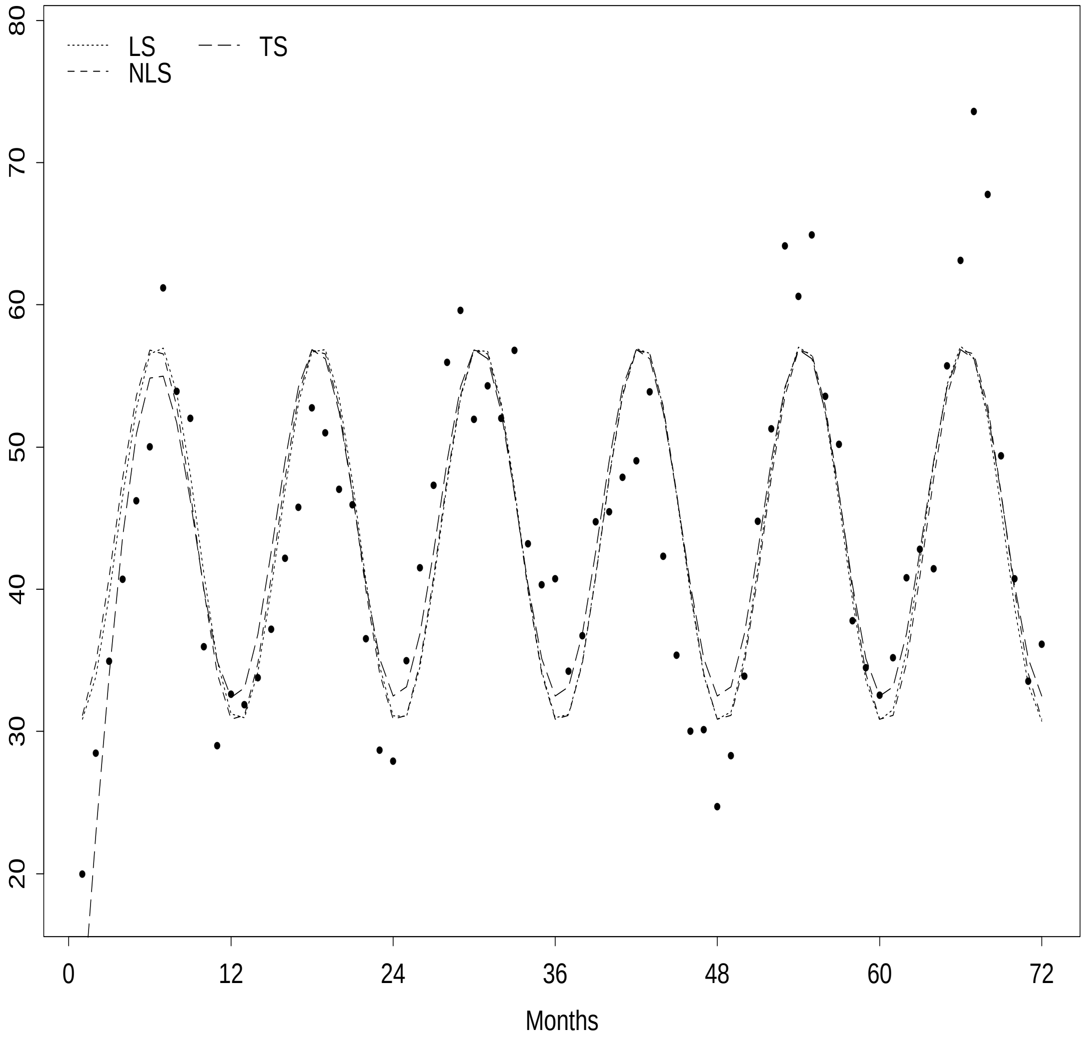

Figure 5.15: *Fits for Monthly Levels of O_3 at Site #7.*

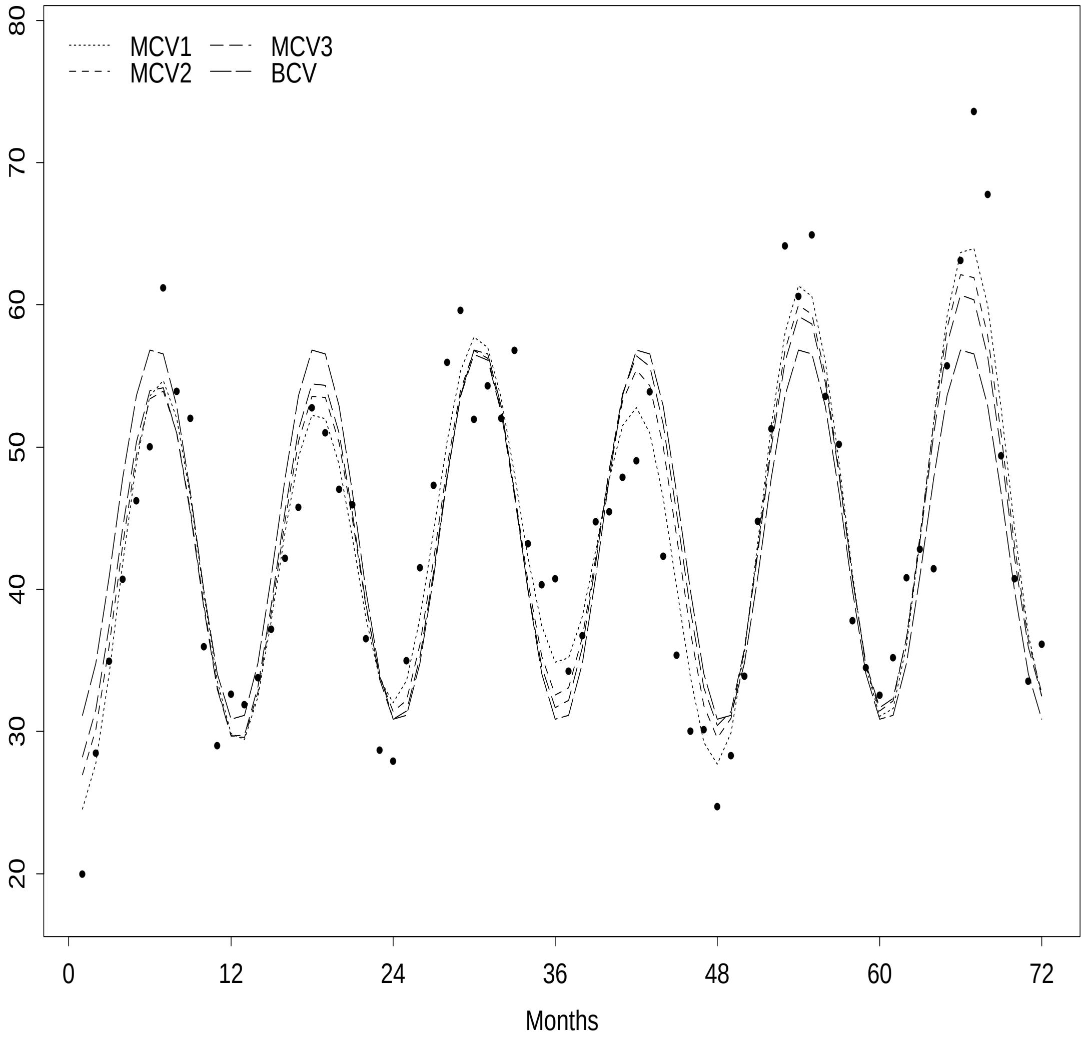

Figure 5.16: *Fits for Monthly Levels of O_3 at Site #7.*

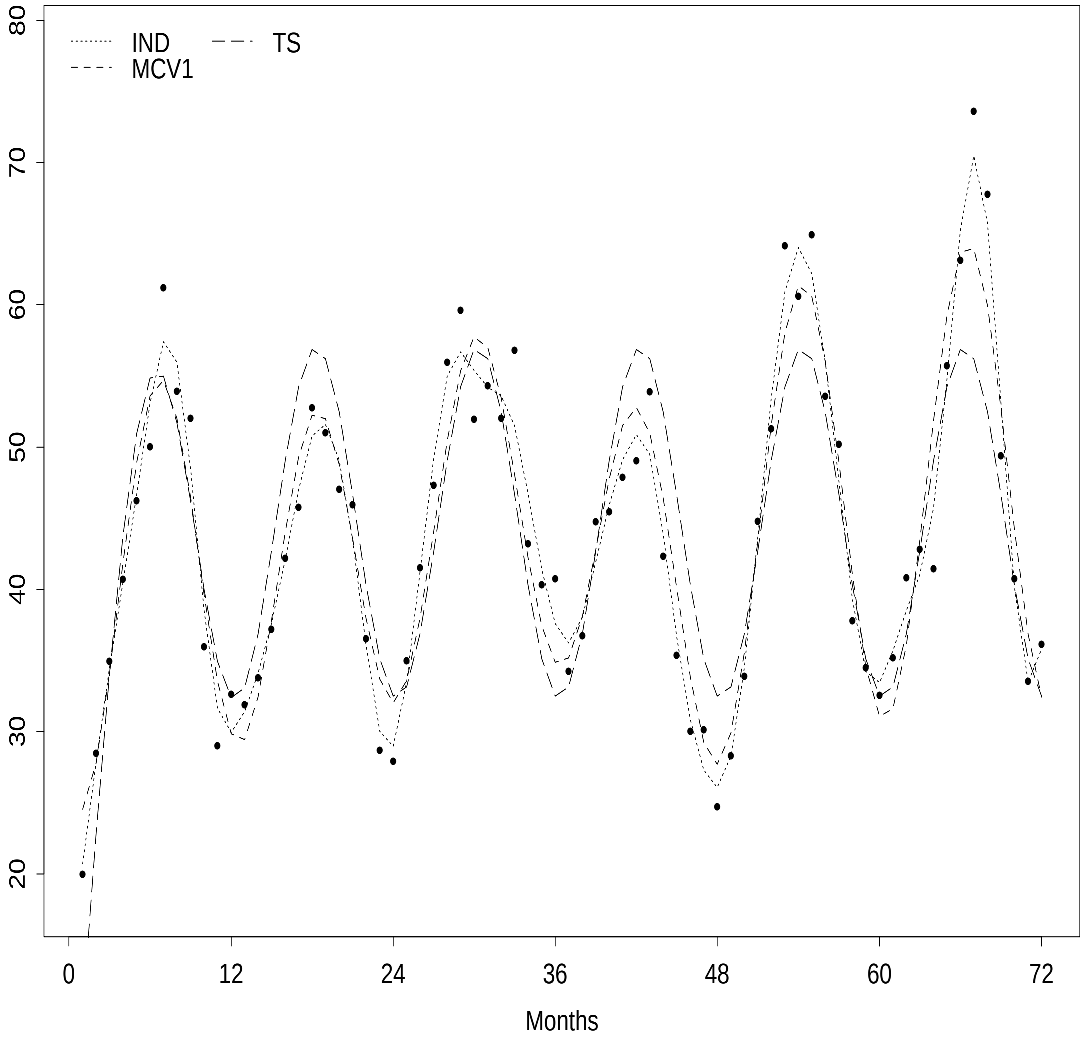

Figure 5.17: *Fits for Monthly Levels of O_3 at Site #7.*

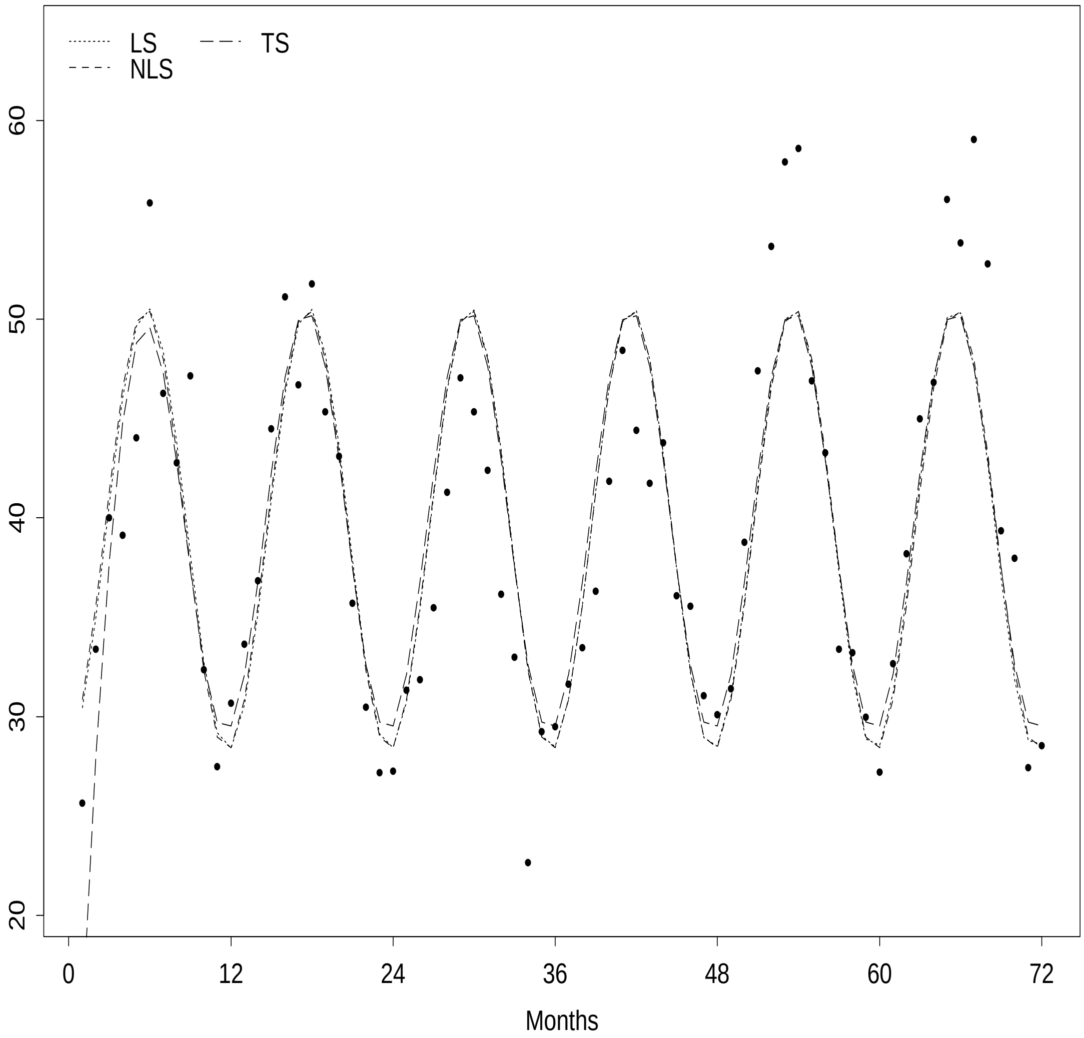

Figure 5.18: *Fits for Monthly Levels of O_3 at Site #17.*

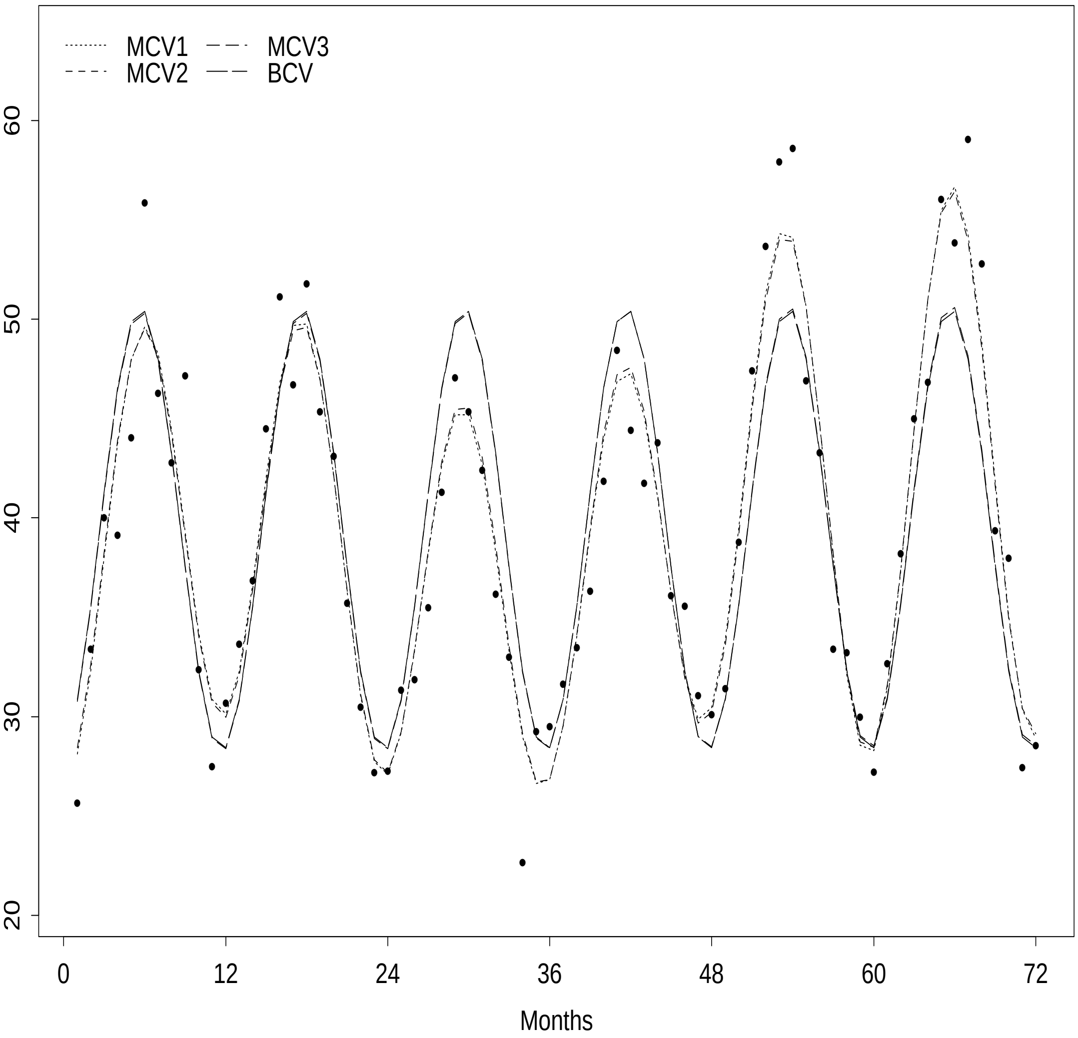

Figure 5.19: *Fits for Monthly Levels of O_3 at Site #17.*

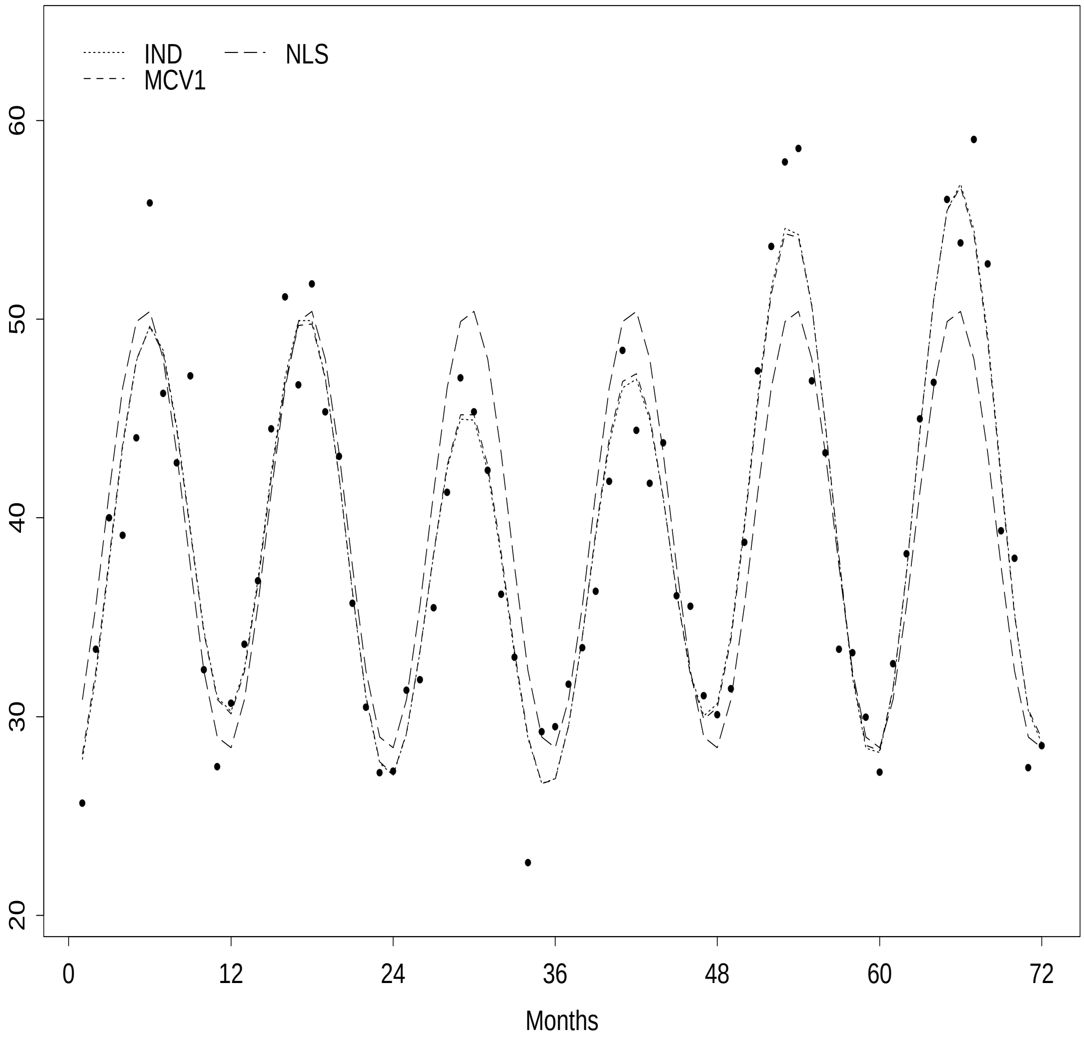

Figure 5.20: *Fits for Monthly Levels of O_3 at Site #17.*

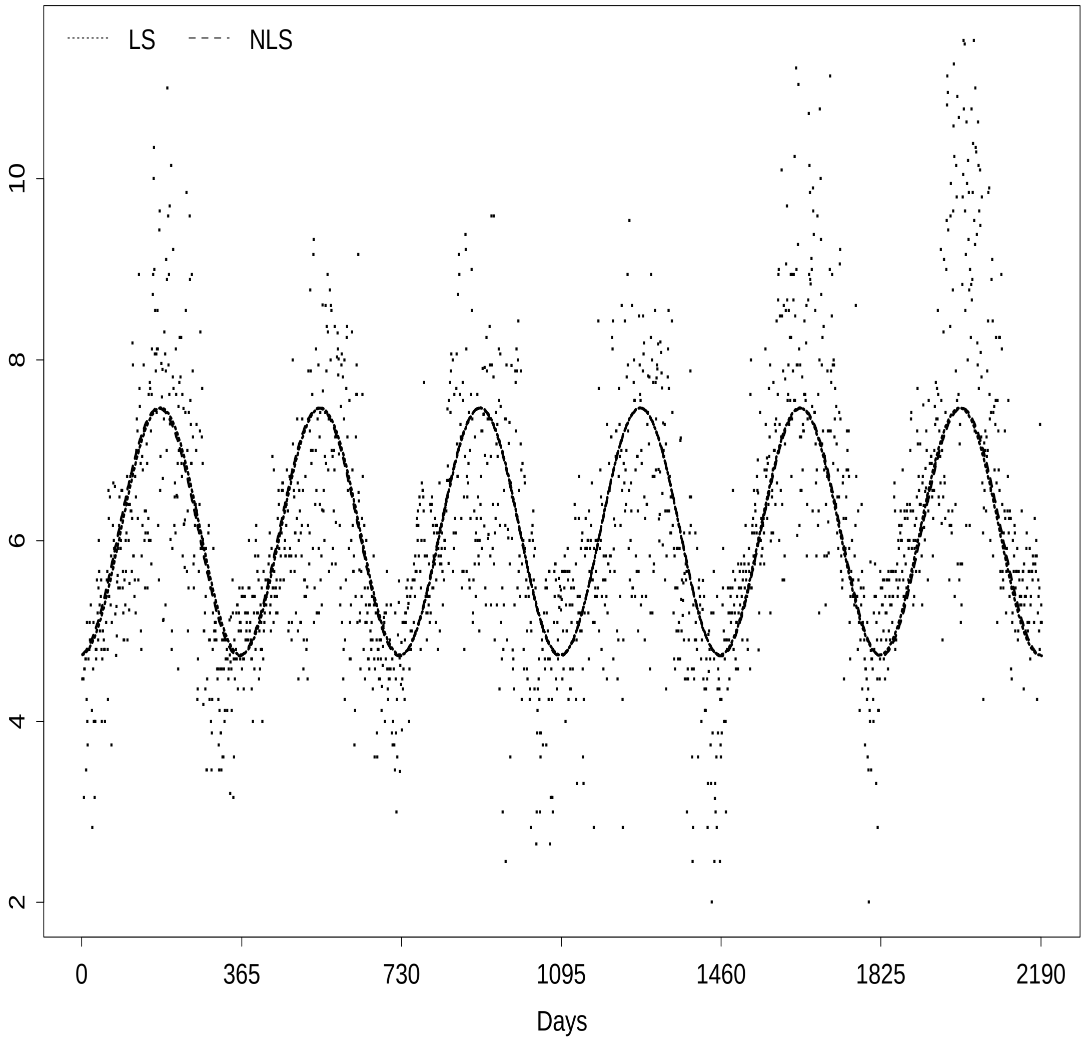

Figure 5.21: *Fits for Square Root Transformed Daily Levels of O_3 at Site #2.*

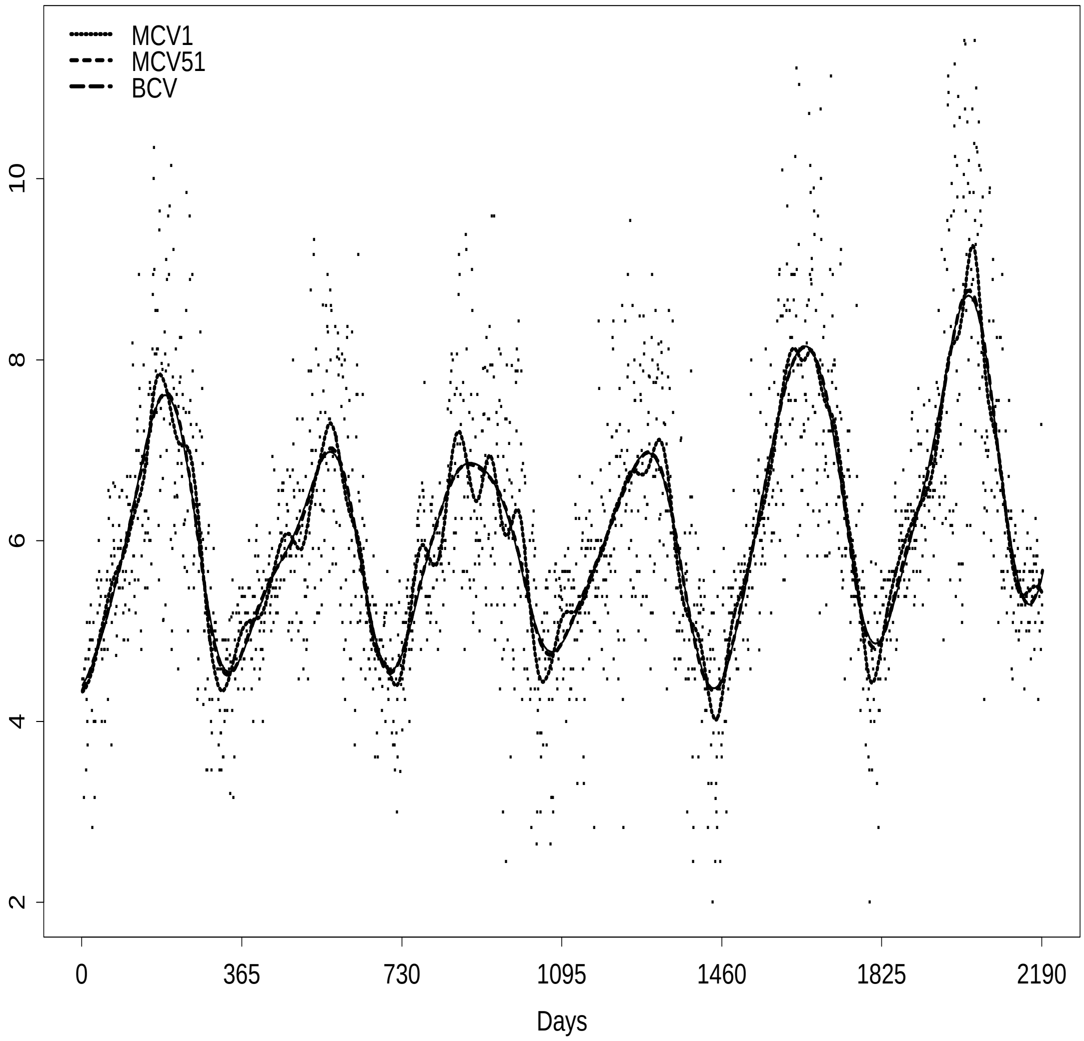

Figure 5.22: *Fits for Square Root Transformed Daily Levels of O_3 at Site #2.*

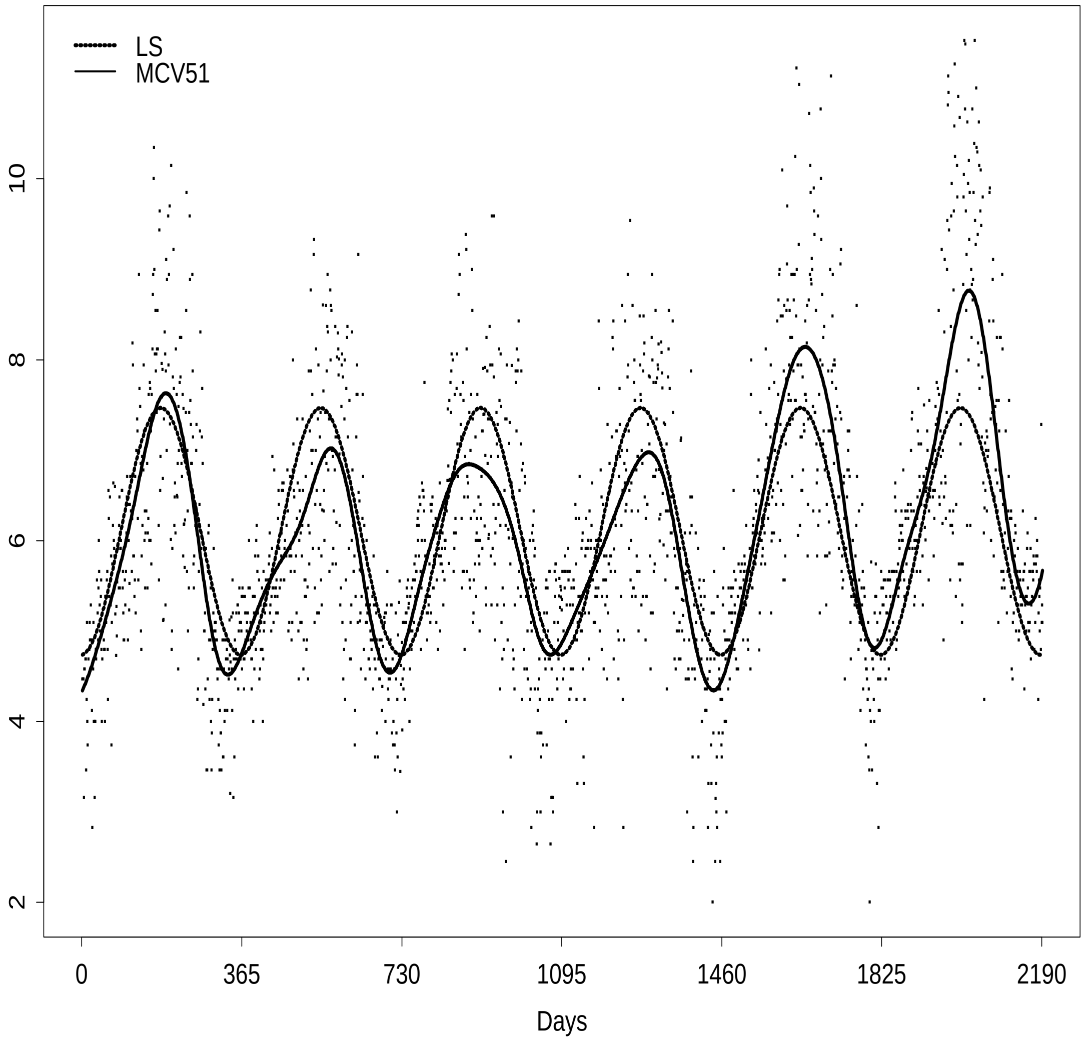

Figure 5.23: *Fits for Square Root Transformed Daily Levels of O_3 at Site #2.*

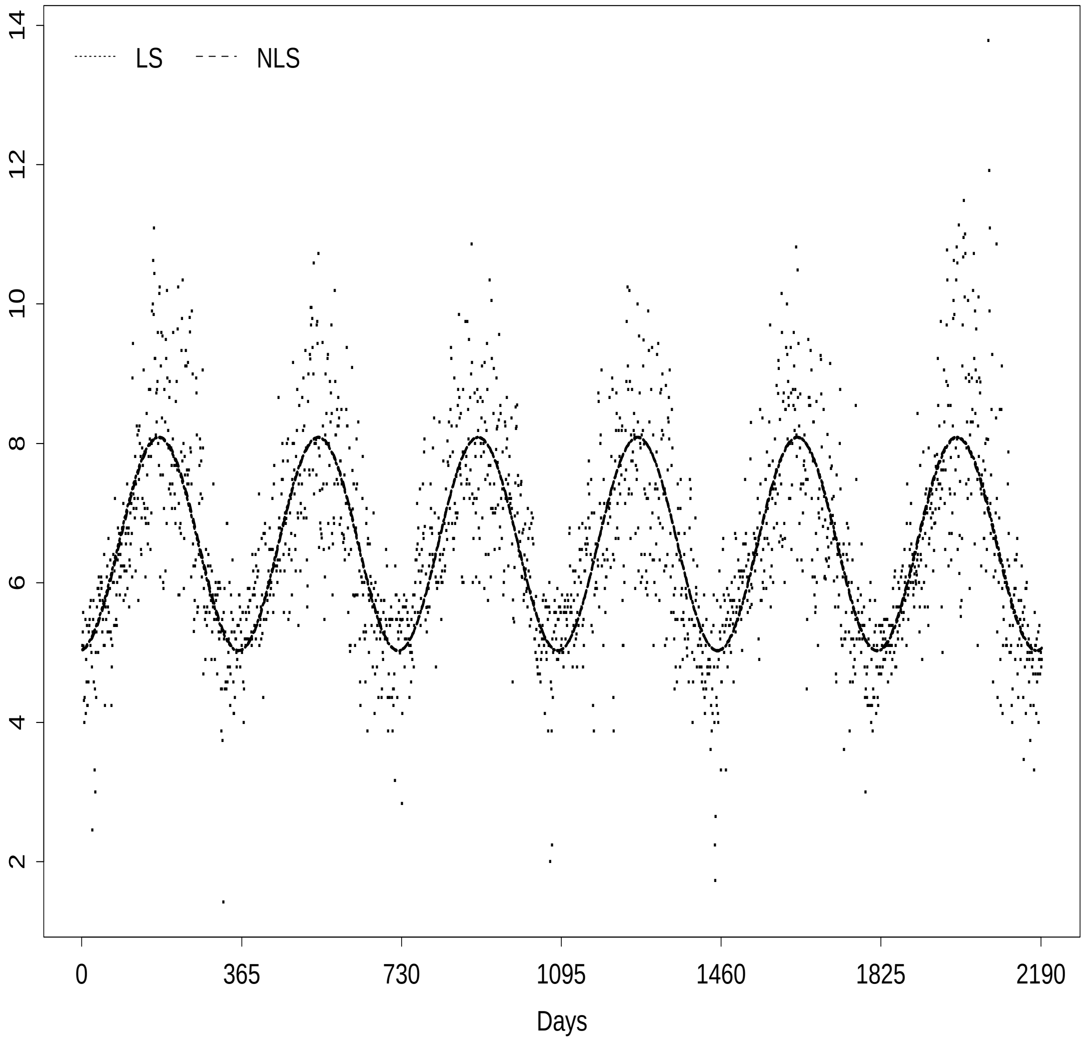

Figure 5.24: *Fits for Square Root Transformed Daily Levels of* O_3 *at Site #7.*

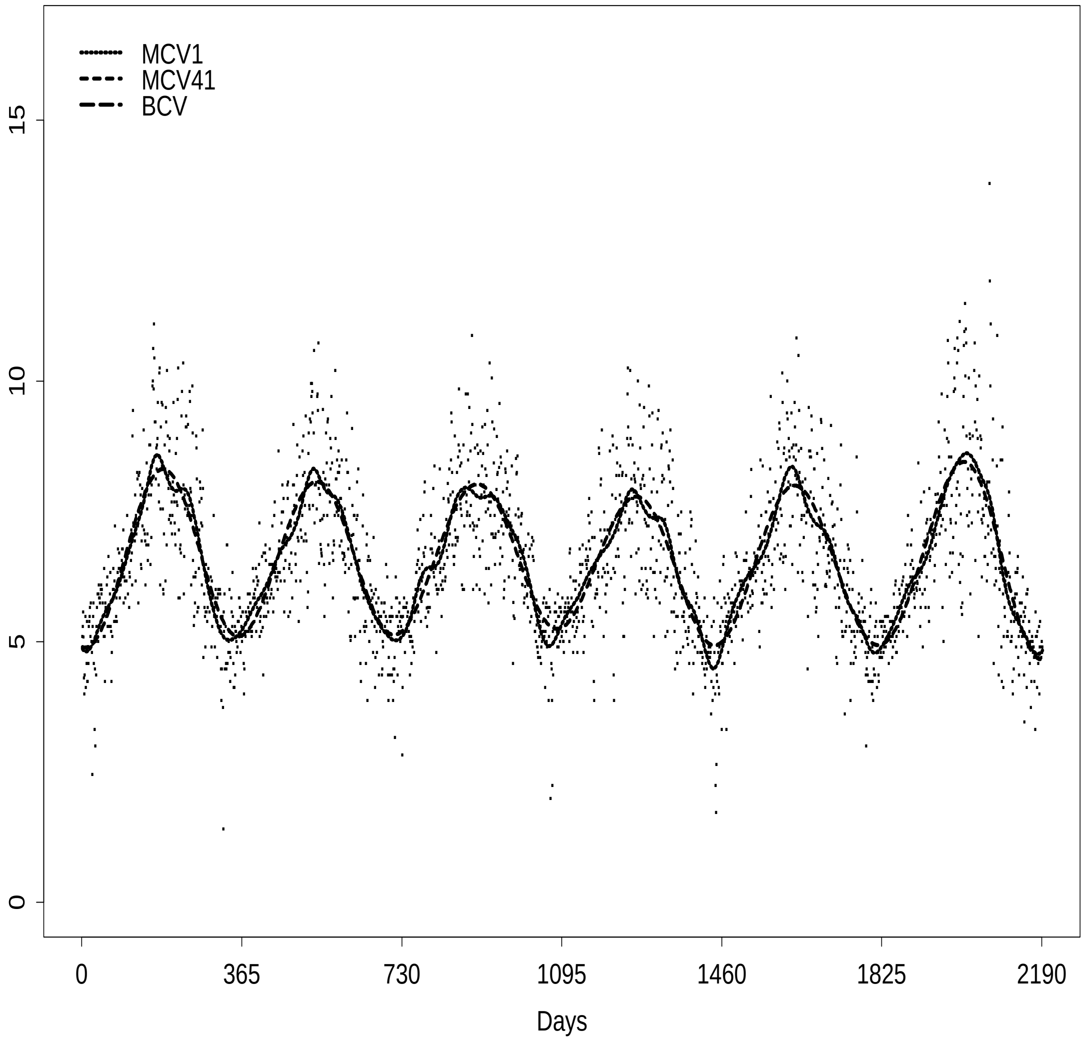

Figure 5.25: *Fits for Square Root Transformed Daily Levels of O_3 at Site #7.*

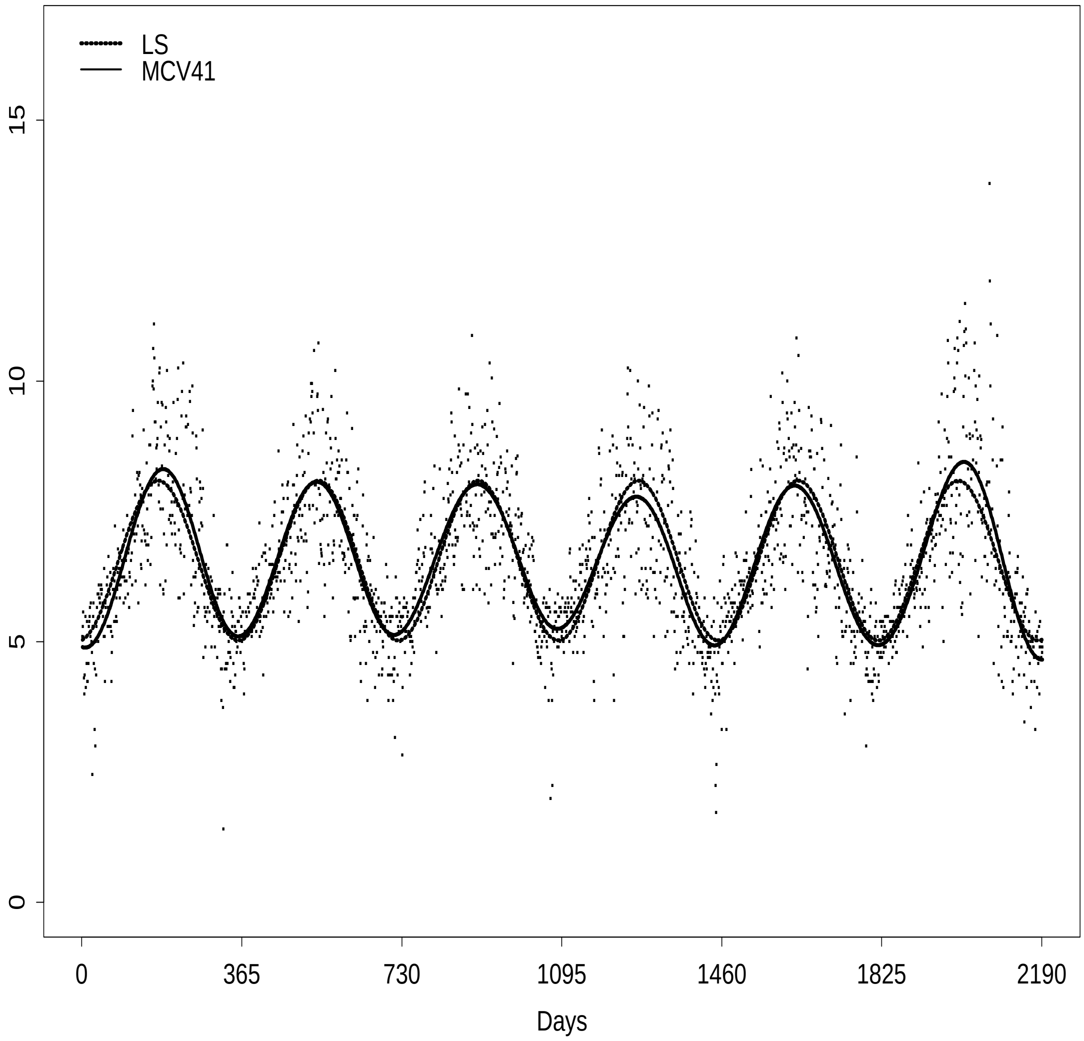

Figure 5.26: *Fits for Square Root Transformed Daily Levels of O_3 at Site #7.*

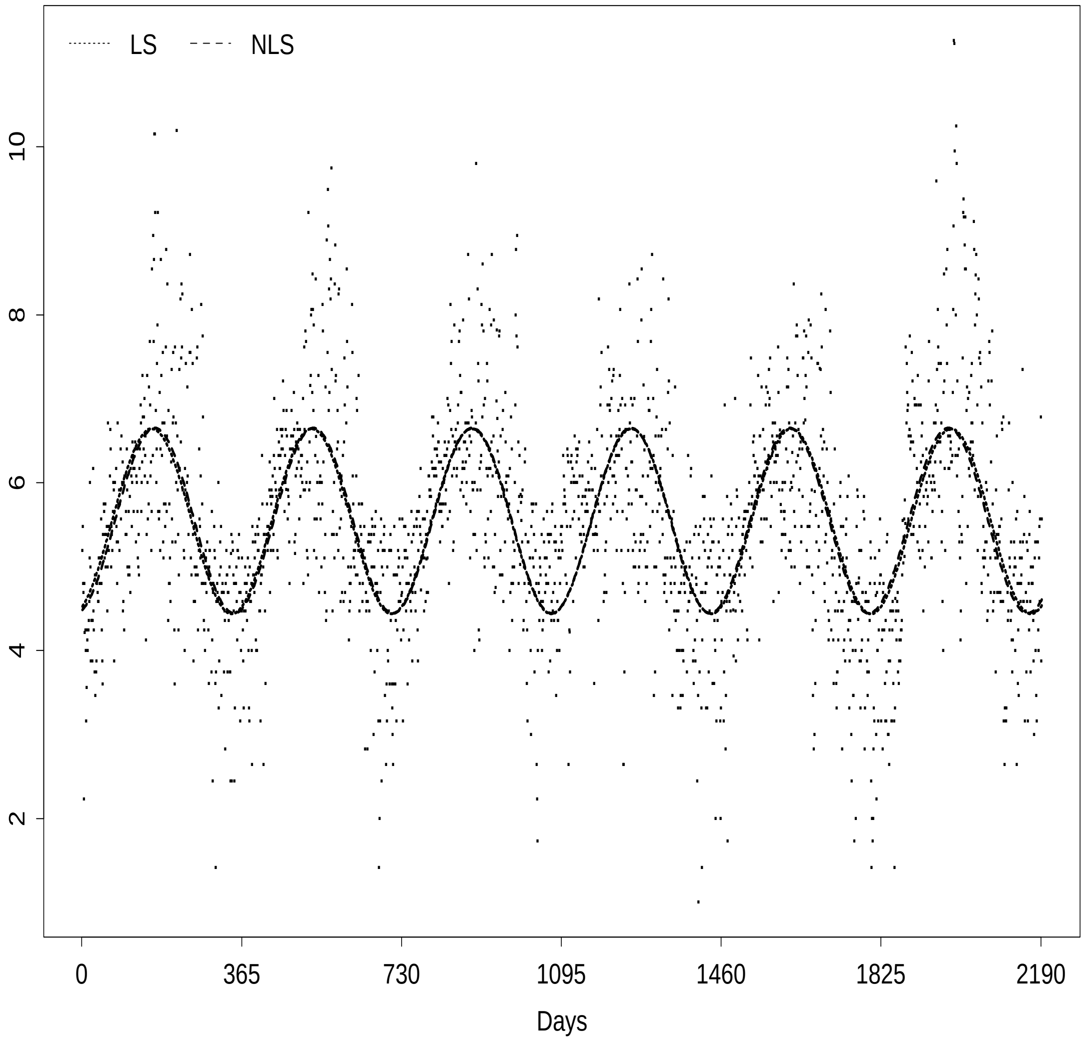

Figure 5.27: *Fits for Square Root Transformed Daily Levels of O_3 at Site #17.*

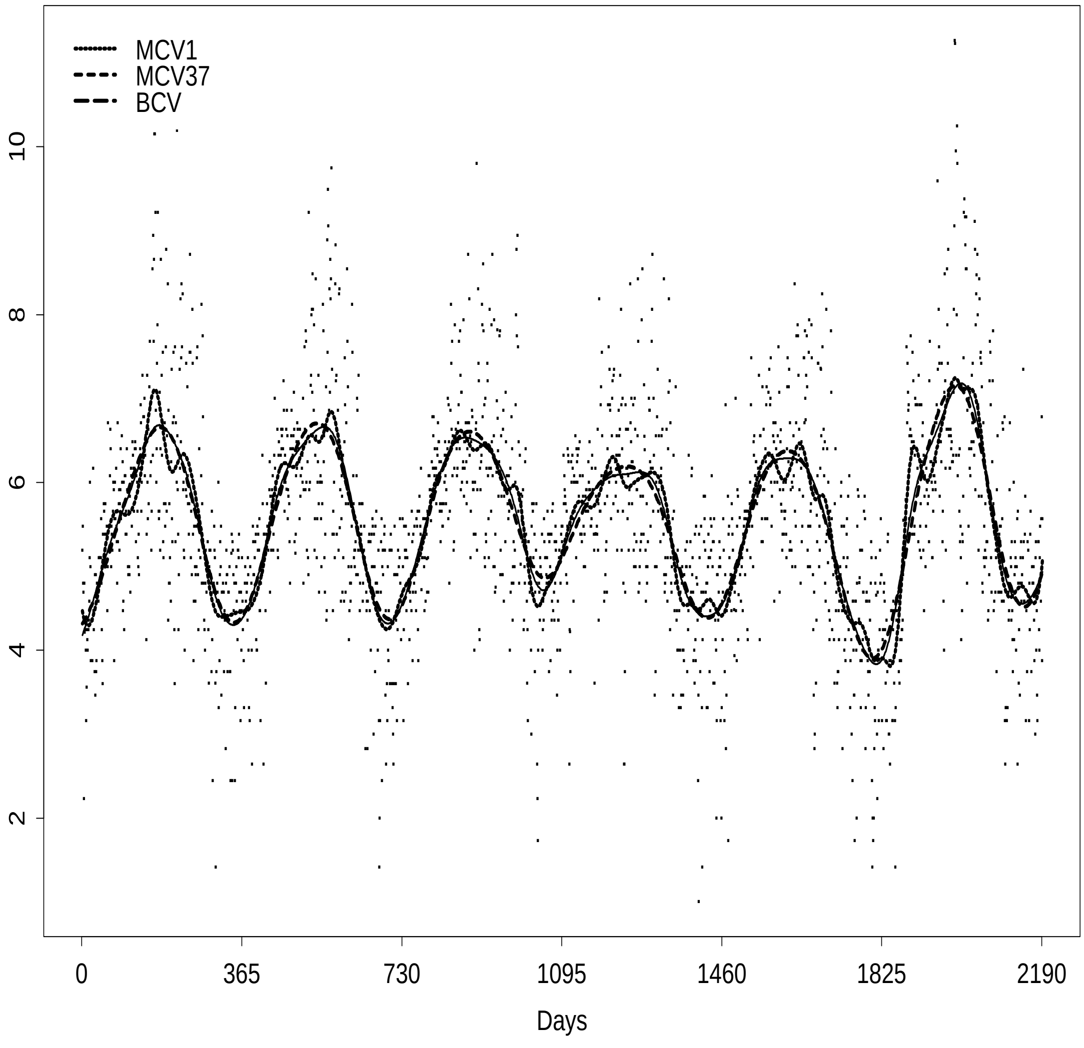

Figure 5.28: *Fits for Square Root Transformed Daily Levels of O_3 at Site #17.*

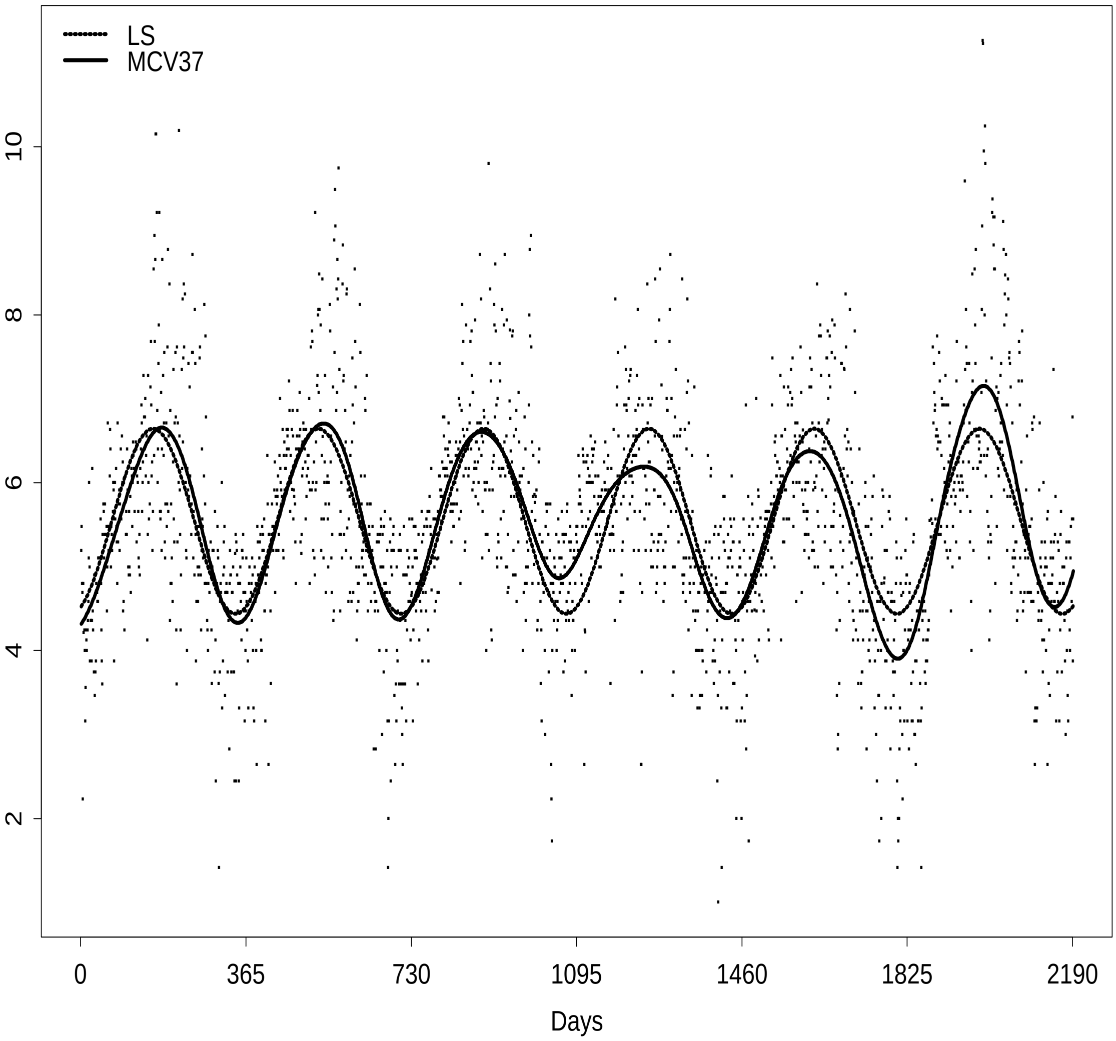

Figure 5.29: *Fits for Square Root Transformed Daily Levels of O_3 at Site #17.*