Model Application

to Air Pollution Data

of SCCX

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Kuei-yu Chien

2008
The thesis of Kuei-yu Chien is approved.

H Cantor

H

Jan de Leeuw, Committee Chair

University of California, Los Angeles

2008
To my family
# Table of Contents

1 Introduction .................................................. 1

2 Model .......................................................... 4
   2.1 ARIMA Model ............................................. 4
       2.1.1 Background ....................................... 4
       2.1.2 Statistical Model ................................. 5
   2.2 Seasonal ARIMA Model .................................. 6
       2.2.1 Background ....................................... 6
       2.2.2 Statistical Model ................................. 7

3 Model Identification for ozone data ....................... 9
   3.1 Ozone Data ............................................. 9
   3.2 Autocorrelation Function (ACF) ....................... 13
   3.3 Partial Autocorrelation Function (PACF) ............. 15

4 Estimations and Diagnostic Checking ..................... 22
   4.1 Parameter Estimation .................................. 22
   4.2 Diagnostic Checking ................................... 23
       4.2.1 Normality ....................................... 24
       4.2.2 Randomness ..................................... 26

5 Forecasting .................................................. 28
   5.1 Seasonal $ARIMA(1,0,0) \times (1,1,0)_7$ Model .......... 28
6 Model Comparison ................................................. 32
  6.1 Non-Seasonal ARIMA Model ............................. 32
  6.2 Mixed ARMA Model .......................................... 33
    6.2.1 Model Identification ................................. 34
  6.3 Concluding Remarks ......................................... 36

7 Conclusions ..................................................... 38
  7.1 Result ......................................................... 38
  7.2 Further Studies .............................................. 39

References .......................................................... 41
**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Average Hourly Maximum ozone data-9,000 data points</td>
<td>10</td>
</tr>
<tr>
<td>3.2</td>
<td>Seasonality</td>
<td>11</td>
</tr>
<tr>
<td>3.3</td>
<td>Power Transformation—the optimal $\lambda$</td>
<td>16</td>
</tr>
<tr>
<td>3.4</td>
<td>$T(\text{ozone})$, transformed with $\lambda = -0.5$</td>
<td>17</td>
</tr>
<tr>
<td>3.5</td>
<td>Seasonality in $T(\text{Ozone})$</td>
<td>17</td>
</tr>
<tr>
<td>3.6</td>
<td>$T(\text{Ozone})$ after 1st order seasonal differencing</td>
<td>18</td>
</tr>
<tr>
<td>3.7</td>
<td>Normal Q-Q plot of $T(\text{Ozone})$</td>
<td>19</td>
</tr>
<tr>
<td>3.8</td>
<td>Sample ACF—Tail off trend</td>
<td>20</td>
</tr>
<tr>
<td>3.9</td>
<td>Sample ACF of the First Differences</td>
<td>20</td>
</tr>
<tr>
<td>3.10</td>
<td>Sample Partial ACF—decaying pattern</td>
<td>21</td>
</tr>
<tr>
<td>4.1</td>
<td>Residuals from the $SARIMA(0, 1, 1) \times (0, 1, 1)_7$ Model</td>
<td>23</td>
</tr>
<tr>
<td>4.2</td>
<td>Normality test</td>
<td>24</td>
</tr>
<tr>
<td>4.3</td>
<td>Residual ACF</td>
<td>27</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparison: Actual Value and Predicted Value</td>
<td>30</td>
</tr>
<tr>
<td>5.2</td>
<td>Forecast of Seasonal ARIMA Model, leading=200</td>
<td>31</td>
</tr>
<tr>
<td>6.1</td>
<td>Forecast of ARIMA Model, leading=200</td>
<td>37</td>
</tr>
</tbody>
</table>
# List of Tables

3.1 KPSS Test ........................................................................... 12

4.1 parameter estimation ....................................................... 22

4.2 17 Innovative Outliers .................................................... 25

4.3 Parameters Estimation after Modification ......................... 26

6.1 Parameters Estimation for Non-Seasonal Model ............... 33

6.2 The ACF and PACF for ARMA Models ......................... 34

6.3 Extended ACF for Difference of OZMAX1HR ............... 35

6.4 Parameters Estimation for $ARMA(2,2)$ Model ............... 35
Research on the variation in Ozone concentration contains the analysis of its changing in the time. We will present a detailed discussion about how to apply the seasonal ARIMA (SARIMA) model, also known as seasonal Box-Jenkins approach, to the daily average maximum ozone data extracted from the California Air Resource Board, and based on the fitted model, we will forecast the level of ozone in the years to come. The SARIMA model is parsimonious enough as well as sufficient to describe non-stationary data such as the underlying ozone data. We firstly reduce a stochastic series to a stationary series and apply the MA and AR model to make further illustration of the data; considering the seasonal pattern occurring in the original data, we may as well use the seasonal ARIMA model to adjust the seasonality and improve the results of our forecast. We will examine our model by three methods: the model checking process; comparison with other possible models such as the non-seasonal ARIMA model and the mixed ARMA model, and applying the in-sample prediction concept and then comparing with
the actual values. After conducting the three method, we can reach our conclusion that the SARIMA model provides a good description of our data.
CHAPTER 1

Introduction

Naturally, the atmosphere contains greenhouse gases which include water vapor ($H_2O$), carbon dioxide ($CO_2$), methane ($CH_4$), nitrous oxide ($N_2O$), and ozone ($O_3$). The greenhouse gases cause the temperature on the surface to change smoothly and make the environment suitable for life. However, for the past over 200 years, the increasing amounts of burning fossil fuels (such as coal and oil), urbanization, and the sharp decline in tropical forests[1] have caused greenhouse gases to increase significantly in our atmosphere. The over production of greenhouse gases mainly resulting from human activities has broken the balanced composition of our atmosphere. These extra gases keep the heat and humidity on the surface of our earth thus making the earth warmer than it otherwise would be.

The most obvious impact on the earth from greenhouse gases is well-known as climate change. Climate change refers to any significant change in measures of climate such as temperature, wind, or precipitation lasting for decades or longer[2]. Among all the greenhouse gases, ozone ($O_3$) is the main factor that actively interacts with climate[3]. Interactions between ozone and climate naturally occur in the stratosphere as well as at the Earth’s surface (troposphere); chemicals participating in ozone formation include two groups of compounds: nitrogen oxides ($NO_x$) and volatile organic compounds (VOCs). The compounds also enter into consideration when detecting climate change. Besides, high ground-level ozone is harmful to both human health and plants. It is a key ingredient of urban smog,
and repeated exposure to ozone pollution may cause permanent damage to the lungs. Even low level of Ozone, inhaling it can trigger health problems including coughing, congestion, nausea, throat irritation, and chest pains [4]. Therefore, early information and prediction of high concentrations of ozone is necessary to early detection of unusual change in the climate earlier and early warning allows for more effective public warning and response to the threat to health posed by pollution.

In southern California, the Los Angeles metropolitan area has the worst ozone pollution in the United State [5]. The main air quality observation institution in California is the California Air Resources Board (ARB), established in 1967 and aimed at maintaining healthy air quality and conducting research into the causes of poor air quality and solutions to the problem of air pollution. Therefore, in this paper, we will use the database from ARB. The placement of its fifteen air quality monitoring basins covers the whole of California, and the data from each basin consists of the concentrations of toxic gases in the air. In order to detect the concentration of ozone, we thus focus on the observed value of ozone in the basin called SCCX (Southern Central Coast), which covers two severely polluted California counties: Kern county (Rank 3) and Ventura county (Rank 17) [6], and one moderately polluted county, Santa Barbara. The observations ranging from 1980 January 1st to 2005 December 31st, contain 9,497 uninterrupted measurements. For better fitting the data, we apply a general time series model called the Auto Regressive Integrated Moving Average model (ARIMA) model to the dataset.

In addition, in order to test our prediction, we apply the concept of in-sample and post-sample prediction in this paper [7]. In other words, firstly we separate the data into two parts, say, a training set and a testing set, and then we choose the first part of our data points (training set) from the 9497 values. We would
like to keep the training set as large as possible so that our error estimates are robust. Because we have a long series before the intervention, we choose the first 9000 observations in the Ozone data to identify the model, and fit the ARIMA model to the training set. After, we make a prediction for the remaining 497 data points (testing set) and compare our prediction with the actual values in the testing set to inspect the accuracy. If the actual values lie in the 95% confidence interval of our prediction, we conclude that our model is accurate and make further predictions based on the actual values.

In the following chapter, we describe the ARIMA model and discuss the seasonality in the seasonal ARIMA model. In chapter 3, we show the model specification process by examining the plot of the original series and autocorrelation functions. In chapter 4, parameter estimations and model diagnosis of are illustrated. In chapter 5, we forecast the ozone level and take a close look at our results; within-sample forecasting is used to examine our prediction. In chapter 6, we examine other possible models and prove that the seasonal as well as the non-stationary process should be taken into account when we analyze the underlying series. In the last chapter, we draw conclusions and discuss further possible research in the field.
CHAPTER 2

Model

2.1 ARIMA Model

2.1.1 Background

Before 1960, the concept of prediction had prevailed in a variety of fields such as management, economics, engineering and the social sciences, and the models used include the Moving Average (MA) model, the Auto Regression (AR) model, the Exponential Smoothing model, etc.. The models assume that the underlying process is stationary, which means that the mean, the variance, and the autocovariance of the process are invariant in the time [8]. In other words, the mean and variance are constant, and the autocovariances depends only on the time lag. The general model used in these methods is:

\[ z_t = f(x_t, \beta) + \varepsilon_t \]  \hspace{1cm} (2.1)

where \( x_t \) represents the predictor variables or functions of time and \( \beta \) refers to the parameters. The error term \( \varepsilon_t \) are assumed to be uncorrelated errors, however the assumption may not be practical in real world. Many observed time series are serially correlated because the data is collected in a sequential span. Models that can capture the correlation structure have to be built for better prediction. The Auto Regressive Integrated Moving Average (ARIMA) model was first presented by Box and Jenkins in 1960. The model, integrating the AR and MA
models, is a highly generalized method for forecasting. In addition, it can be used in detecting a variety of different correlation structures. Theoretically an observed time series can be thought of as a random sample from a stochastic process, and by the ARIMA model, the non-stationary situation if occurring can be eliminated by transformations such as differencing or power transformation. Therefore, the AR, the MA, and the ARMA models are all special cases of the ARIMA model.

### 2.1.2 Statistical Model

An ARIMA model is classified as an ARIMA(p,d,q) model:

\[
(1 - \sum_{i=1}^{p} \phi_i B^i)[\nabla^d Y_t - \mu] = (1 - \sum_{j=1}^{q} \theta_j B^j) e_t
\]

(2.2)

Where \( B \) denotes the backward shift operator such that \( B Y_t = Y_{t-1} \), \( p \) is the order of auto regressive terms, \( d \) is the order of differences, and \( q \) is the number of moving average order. \( Y_t \) is the time series value at time equal to \( t \), \( \mu \) is the mean after \( d \) differencing. The parameters for the auto regressive part are \( \phi_1, \phi_2 \cdots \phi_p \), which can be one or more polynomials of order \( p \); the parameters for the moving average part are \( \theta_1, \theta_2 \cdots \theta_q \), which can be one or more polynomials of order \( q \), and \( e_t \) denotes a purely random process, and follows NID(0,\( \sigma^2 \)).

ARIMA (p,d,q) can be reduced to many special cases. Take ARIMA(1,0,0). It can be reduced to the AR(1) model, but the AR model should only be applied to a stationary time series. Also, ARIMA(1,0,1) can be reduced to ARMA(1,1), and the ARMA model does not consider the differencing and can only be used in a stationary sequence. In summary, the ARIMA model is suitable when one wants to forecast continuous data and exogenous variables [7]. The forecasting process in ARIMA has four steps: capture the trend of the original data, decide the number of differencing (d), do necessary transformation to make the sequence
stationary, and fit the stationary series with the ARMA(p,q) model.

2.2 Seasonal ARIMA Model

2.2.1 Background

The ARIMA model uses historical data to forecast trends in the future, however it cannot forecast periodic fluctuations in temperature, variables affected by temperature, and more generally variable data like retail sales figures. The most commonly used equations that describe the seasonal behaviors are [9]:

\[ X_t = m_t + S_t + \varepsilon_t \]  
\[ X_t = m_tS_t + \varepsilon_t \]  
\[ X_t = m_tS_t\varepsilon_t \]

where \( S_t \) is the seasonal effect at time \( t \), \( m_t \) is the deseasonalized mean level at time \( t \), and \( \varepsilon_t \) is the random error.

The above equations describe the seasonality. Equation (2.3) is the additive case, while equation (2.4) and equation (2.5) both involve multiplicative seasonality. The time plot should be examined to choose an appropriate model.

In addition, when dealing with series containing a seasonal fluctuations, the method used also depends on whether we want to measure the seasonal effect or eliminate seasonality. If the series shows a slight trend, it is usually sufficient to estimate the seasonal effect by averaging the values within each cyclic time span and subtract the corresponding yearly average in the additive case. While in the multiplicative case, we can divide the values within each cyclic time span by the yearly average. If the series shows an obvious trend, it may require a more complicated approach to deseasonalize the data [8].
The main feature of seasonal data is that there are high correlations among observations from a certain time span (day, week, month or quarter). Thus the difference between the ARIMA model and the Seasonal ARIMA (SARIMA) model is that the later, besides consecutive differencing, includes seasonal differencing. Based on the ARIMA model, we can fit a seasonal model to data using the SARIMA method. In other words, we consider the seasonal character of data, do seasonal differencing if needed, and then apply the seasonal ARIMA model to the sequence. So as the consecutive differencing shared by both ARIMA and SARIMA captures the long-term trend, the seasonal differencing in SARIMA removes the seasonal pattern.

2.2.2 Statistical Model

A general multiplicative seasonal model is classified as $\text{ARIMA}(p, d, q) \times (P, D, Q)_s$:

$$\nabla^d \nabla_s^D Y_t - \mu = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)} e_t$$

(2.6)

We can decide $s$(Seasonal Span) by the time span for each cycle: for quarterly data, we apply $s = 4$; for daily data, we set $s = 7$; for monthly data, we set $s = 12$; and for hourly data, we set $s = 24$. $Y_t$ is the time series value at time equal to $t$, $d$ is the number of the consecutive differencing, and $D$ is the number of the seasonal differencing. $\mu$ is the mean after differencing. $e_t$, the error term, follows NID(0, $\sigma^2$). $P, Q$ is numbers of seasonal, auto-regressive terms and seasonal moving average terms, respectively. Further,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

(2.7)

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

(2.8)

$$\Phi(B) = 1 - \Phi_1 B_s - \Phi_2 B_s^2 - \cdots - \phi_p B_s^p$$

(2.9)

$$\Theta(B) = 1 - \Theta_1 B_s - \theta_2 B_s^2 - \cdots - \theta_Q B_s^Q$$

(2.10)
and

$$\omega_t = \nabla^p s Y_t$$  \hspace{1cm} (2.11)$$

The last equation illustrates the multiplicative seasonal behavior indicating that seasonal and consecutive differencing may be required to induce stationarity [10]. Seasonal ARIMA should be used discreetly. If we applied the seasonal model to non-seasonal data, the forecast would show a cycle that may be far from the truth. We should make sure the data contains seasonality before applying the model.
CHAPTER 3

Model Identification for ozone data

The first step in fitting a time series to certain model is to check the stationarity, assessing values of $d$ and $D$ that reduce the series to stationarity and eliminate most of the seasonality. Then we look at the Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (PACF) and compare them with the theoretical ACF and PACF in order to determine $p$, $P$, $q$ and $Q$.

3.1 Ozone Data

The data that we use was extracted from the California Air Resource Board (ARB). The basin we choose (SCCX) covers Santa Barbara, Ventura, and San Luis Obispo, which is located in the southern central coastal region of California. Within the basin, there are twenty-five monitoring sites. The variable OZMAX1HR stands for the average hourly maximum concentration of ozone (ppm, parts per million); the observed time span is daily from 1980 January 1 to 2005 December 31 with no missing values. Thus there are 9,497 values. Figure 3.1 shows the OZMAX1HR along the time:
The data plot illustrates the overall trend and also shows decreasing dispersion in the seasonal variation of the ozone level; the ozone level was higher in 1980 and constantly decreased till the end of 2005. Also, within each year, the concentration of ozone started at a lower level, increased in the middle of the year, and then decreased toward the end of the year (see figure 3.2). Observation suggests that the variation of the series is not constant over time and that there is a trend as well as a seasonal pattern in the data.
Figure 3.2: Seasonality

The first step in the ARIMA forecasting procedure is to induce stationarity. The data inspection has shown that the variability in the data decreases with time, so it would be necessary to apply the power transformation to the original dataset to stabilize the variance. The power transformations, introduced by Box and Cox (1964) [11], could be employed here. For a given value of the parameter $\lambda$, the transformation is defined by:

$$Y^{(\lambda)} = \begin{cases} 
\frac{y^{\lambda-1}}{\lambda} & \lambda \neq 0 \\
\log Y & \lambda = 0
\end{cases}$$  

(3.1)

In order to find an optimal $\lambda$, figure 3.3 gives the 95% interval of the value of $\lambda$.

Figure 3.3 suggests that $\lambda$ equal to -0.5 as the closest value to the 95% confidence interval of $\lambda$. After transformation with $\lambda = -0.5$, our original data becomes relatively stable. The downward trend does not exist significantly. We can examine
the trend by plotting the data after power transformation (see Figure 3.4).

For the next step, we analyze the seasonality within the transformed data. Apparently, from Figure 3.5, even after a transformation, say $T(\text{ozone})$, our data contains a seasonal periodic component; the concentration of ozone rose in the middle of the year (around the 200th day in each year), and went down at the end of each year. Several ozone studies have proven that the concentration of ozone and temperature have effects on each other [12], so we can reasonably take seasonality into consideration.

Since a distinct seasonal pattern exists, we do seasonal differencing, which is a technique to remove seasonal trend within the data. The result is given in figure 3.6.

In order to test the stationarity of our data so far, we may conduct the KPSS test [13], the hypothesis test is:

$$
\begin{align*}
H_0 : & \text{ The data is a stationary series } \\
H_1 : & \text{ The data is not a stationary series }
\end{align*}
$$

The test result is shown in the following table:

<table>
<thead>
<tr>
<th>KPSS Level</th>
<th>0.0065</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncation Lag of Parameter</td>
<td>21</td>
</tr>
<tr>
<td>p-value</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 3.1: KPSS Test

The $p$-value of 0.1 is greater than 0.05, so we do not reject $H_0$, the stationarity
holds for our final series.
Different from the original data, the transformed, differenced data is adequate for consideration as a stationary sequence, therefore, the processes above suggests that we use the seasonal ARIMA model. For daily data, we choose the time span $s$ equals to 7 and the order of seasonal differencing $D$ equals to.

Figure 3.7 illustrates the normality of our data. As can be judged by the plot, the normality holds. Normality holds partly because of the large samples size ($n=9,000$), and the differences that we have applied to the data also reinforce the normality.

### 3.2 Autocorrelation Function (ACF)

After having made the necessary transformations of the data to assure stationarity and after also having insured constant variance, we then examine the sample autocorrelation function. The autocorrelation function measures the correlation between observations of differences in distance[9]. In other words, unlike the correlation coefficient, which measures the relation between two different variables, autocorrelation is a correlation coefficient that describes the correlation between two values of the same variable at times $t$ and $t+k$. The autocorrelation function has the following applications: 1) model identification in a time series model, and 2) detecting non-randomness in data. If a time series is completely random, such autocorrelations should be near zero for any and all time-lag separations [14]. If a time series shows some non-random behavior, then one or more of the autocorrelations will be significantly non-zero. Also, if the autocorrelation function fails to die out quickly as the lag $k$, increases, we may need further differencing [10].
In our data, \( N = 9,000 \) observations are provided, say \( Y_1, Y_2, Y_3, \ldots, Y_{9000} \), on a discrete time series. Given measurement, the lag \( k \), the autocorrelation function is defined as:

\[
r_k = \frac{\sum_{i=1}^{n-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\]  

(3.2)

In practice, non-randomness may be present if the values plotted in the autocorrelation coefficients do not diminish at large lags. As can be seen in Figure 3.8, our data, even after power transformation and seasonal differencing, is not sufficient to be a random series:

Figure 3.8, the sample autocorrelation function (ACF) plot with the vertical axis equal to \( r_k \), and the horizontal axis equal to the lag, \( k \), shows a smoothly decreasing trend. Apparently, even until the lag equals to 35 the autocorrelation coefficient is not close to 0, so successive differencing has to be carried out on this data.

Taking the first differences produces a very clear pattern in the sample ACF. The autocorrelation function cuts off after lag equals to 1, therefore, so far we can describe our data as a seasonal ARIMA(0,1,1) model with \( d \) equals to 1, for the first differences and \( D \) equals to 1, for the seasonal differencing. The result can be seen in figure 3.9.
3.3 Partial Autocorrelation Function (PACF)

Unlike the autocorrelations, partial autocorrelations cannot be estimated using a straightforward formula [10]. The partial autocorrelation coefficients however complement the autocorrelation coefficients. In general, supposeing that we have $Y_t$ as a normally distributed time series, the partial autocorrelation function at lag equals to $k$ can be defined as:

$$\phi_{kk} = \text{Corr}(Y_t, Y_{t-k}|Y_{t-1}, Y_{t-2}, \cdots, Y_{t-k-1})$$  \hspace{1cm} (3.3)

Because our tentative model so far is seasonal ARIMA(0,1,1), we consider the MA(1) model and examine the partial autocorrelation plot in figure 3.10.

The partial autocorrelation decays to zero without cut off, which, according to the theoretical ACF and PACF plots for the MA(1) model, tentatively identifies our model as seasonal ARIMA(0,1,1) $\times$ (0,1,1)$_7$. 

Figure 3.3: Power Transformation—the optimal $\lambda$
Figure 3.4: $T(ozone)$, transformed with $\lambda = -0.5$

Figure 3.5: Seasonality in $T(Ozone)$
Figure 3.6: T(Ozone) after 1st order seasonal differencing
Figure 3.7: Normal Q-Q plot of $T$(Ozone)
Figure 3.8: Sample ACF–Tail off trend

Figure 3.9: Sample ACF of the First Differences
Figure 3.10: Sample Partial ACF—decaying pattern
CHAPTER 4

Estimations and Diagnostic Checking

4.1 Parameter Estimation

Applying the SARIMA model to our ozone data, the following table shows the result:

<table>
<thead>
<tr>
<th>coefficient</th>
<th>$\theta$</th>
<th>$\Theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.2581</td>
<td>-0.9965</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.0144</td>
<td>0.0013</td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>19903.98</td>
</tr>
</tbody>
</table>

Table 4.1: parameter estimation

Thus our SARIMA$(0, 1, 1) \times (0, 1, 1)_7$ model is:

$$Y_t = Y_{t-1} + Y_{t-7} - Y_{t-8} + \varepsilon_t + 0.2581\varepsilon_{t-1} + 0.9965\varepsilon_{t-7} + (0.2581)(0.9965)\varepsilon_{t-8} \quad (4.1)$$

We apply the method of Maximum Likelihood Estimates to estimate our parameters. The advantage of the method is that all the values provided in the data are used, thus achieving a better estimation than methods using only a subset of the data. Also, many large-sample results like our data are known under very general conditions [15].

4.2 Diagnostic Checking

In order to ensure that the best forecasting model has been built, diagnostic checking must be performed. Model diagnosis is concerned with testing if the model can describe the series properly. As with most statistical models, diagnosis usually begins by examining the residuals, continues by checking if outliers exist, and finally looks at the index of goodness of fit (we will apply AIC in the part). Firstly, we look at the time series plot of the residuals. Figure 4.1 gives the standardized residuals. Other than some strange behavior in the middle of the series, the plot doesn’t suggest any major pattern in the residuals.

![Residuals from ARIMA(0,1,1)x(0,1,1)](image)

Figure 4.1: Residuals from the $SARIMA(0, 1, 1) \times (0, 1, 1)_7$ Model
4.2.1 Normality

To look further, we would like to investigate if outliers exist. We can firstly take a look at the normal Q-Q plot formed by residuals:

![Normal Q-Q Plot](image)

Figure 4.2: Normality test

Comparing with the straight line (the normal distribution), the normality ap-
pears in doubt when the data points have extreme values, so outliers might exist where the deviation occurs. Calculation shows that there are seventeen outliers (Table 4.2). However, normality holds at the data points that are not extreme, so roughly we can conclude that normality holds for the overall series.

<table>
<thead>
<tr>
<th>t</th>
<th>299</th>
<th>399</th>
<th>378</th>
<th>406</th>
<th>415</th>
<th>650</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$-4.757793$</td>
<td>$-5.796749$</td>
<td>$-5.299856$</td>
<td>$-5.134212$</td>
<td>$5.080049$</td>
<td>$-4.893164$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t</th>
<th>720</th>
<th>730</th>
<th>782</th>
<th>1149</th>
<th>1326</th>
<th>1445</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$-6.063102$</td>
<td>$-4.657533$</td>
<td>$4.594734$</td>
<td>$5.044263$</td>
<td>$-5.226107$</td>
<td>$5.006307$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t</th>
<th>1716</th>
<th>2073</th>
<th>3355</th>
<th>3582</th>
<th>4576</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$-4.548122$</td>
<td>$-5.688542$</td>
<td>$4.73371$</td>
<td>$-4.628942$</td>
<td>$-7.1803$</td>
</tr>
</tbody>
</table>

Table 4.2: 17 Innovative Outliers

The above table shows significant outliers under the Bonferroni’s method [16]. The $t$-th observation is deemed an innovative outlier if $\lambda_1$ is larger than the upper $0.025/n \times 100$ percentile of the standard normal distribution. Because the outliers will inflate the maximum likelihood estimate of standard deviation, an adjustment is needed.

Taking the outliers into account, we then update our ARIMA model. The result is shown in Table 4.3:
Comparing this result with the earlier results shown in Table 4.1, the estimates of \( \theta \) and \( \Theta \) have only slightly changed, yet the AIC is better in the modified model. So our final model become:

\[
Y_t = Y_{t-1} + Y_{t-7} - Y_{t-8} + \varepsilon_t + 0.2453\varepsilon_{t-1} + 0.9962\varepsilon_{t-7} + (0.2453)(0.9962)\varepsilon_{t-8}
\] (4.2)

### 4.2.2 Randomness

For a good model, we expect the residuals to be random [10]. To illustrate the randomness, we can now exhibit the residuals through the autocorrelation functions.

Figure 4.3 illustrates that the autocorrelation function of residuals shows significant correlations before lag 7. We can also observe that although a seasonal pattern exists, the correlations become smaller as the lag number becomes larger. Since we have a large dataset containing 9,000 data points, we may conclude that the model seems to have captured the essence of the randomness.
Figure 4.3: Residual ACF
CHAPTER 5

Forecasting

5.1 Seasonal $ARIMA(1, 0, 0) \times (1, 1, 0)_7$ Model

After model specification, we determined that the model should be Box-Cox transformed, seasonal ARIMA model with the auto regression order equals to 1, the seasonal auto regression order equals to 1, and the seasonal differencing order equals to 1.

To examine our prediction, we compare the actual values with the predicted values. The result appears in Figure 5.1.

The actual values lie in the 95% confidence interval of our prediction, and the predicted interval describes the pattern of the actual data. Thus we can safely use the model to predict the ozone as we wish.

Based on the new value in hand, we can update our forecasts. In Figure 5.2, we forecast 200 days ahead for the concentration of the ozone. The 95% forecast and the 80% interval are presented, and the horizontal line at the estimate of the process mean is shown. We can observe the forecasts approach the mean exponentially as the lead time increases and the prediction limits increase in width. Owing to the non-constant variation of our original data, the forecast limits allow more space for a variance larger than the mean value and less space for a variance smaller then the mean value. This provides an acceptable description of the
original data. Furthermore, the width of forecast limits increase with time, which is reasonable because we have less information on which to base our prediction. In summary, the forecast performed by our fitted Seasonal ARIMA model is adequate at this stage.
Figure 5.1: Comparison: Actual Value and Predicted Value
Forecasts from ARIMA(0,1,1)(0,1,1)[7]

Figure 5.2: Forecast of Seasonal ARIMA Model, leading=200
CHAPTER 6

Model Comparison

In order to support the results so far, we can compare the seasonal ARIMA model to different models such as the non-seasonal ARIMA model and the mixed model (ARMA). However, the order of \((p, d, q)\) remains unchanged for the non-seasonal ARIMA model to demonstrate the pure effects caused by not considering seasonality for seasonal data; for the ARMA model, we briefly introduce the process of model identification and apply the result of the identification process to the underlying data. And we will focus on the difference in criteria measuring the goodness of fit between each model, and the forecast outcome from each model will be examined closely.

6.1 Non-Seasonal ARIMA Model

If we apply a non-seasonal ARIMA model to the OZMAX1HR data with the same power transformation, \(\lambda\) equals to -0.5, the parameters for \(ARIMA(0, 1, 1)\) are:
The parameter $\theta$ changed slightly without considering the seasonality. However, the AIC becomes worse going from 19,903.98 to 20,741.24.

Then we take a look at the forecast. The prediction of the non-seasonal ARIMA model is shown in Figure 6.1.

Comparing the two plots in Figure 6.1, we can observe that the overall range of forecasts looks the same, but the horizontal line at the estimate of the process mean shows the cyclic variation only in the forecast of the seasonal model, not in the non-seasonal model. The horizontal line indicates that seasonal model is probably a more accurate forecasting method for our OZMAX1HR data.

### 6.2 Mixed ARMA Model

The mixed model is also called ARMA (Autoregressive Moving Average process). In general, if $Y_t$ is a mixed ARMA process of orders $p$ and $q$, we abbreviate the name to $ARMA(p,q)$, and it can be defined as:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}$$ (6.1)
6.2.1 Model Identification

For a mixed ARMA model, its theoretical ACF and PACF have infinitely many nonzero values, making it difficult to identify mixed models from the sample ACF and PACF. [16] Many graphic tools have been proposed to identify the ARMA orders. However, one of the methods we briefly describe here is the Extended autocorrelation (EACF) method. Firstly, if the AR part of the ARMA model is known, eliminate or filter out the autoregression part from the underlying time series thus making our series a pure MA process. By doing this, we can observe the cut-off in the series’ ACF. Secondly, once we set the order of MA, the AR coefficients may be estimated by a finite sequence of regressions. The following table summarizes the behavior of ACF and PACF for specifying the mixed ARMA model. [7]

<table>
<thead>
<tr>
<th></th>
<th>AR(p)</th>
<th>MA(q)</th>
<th>ARMA(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Tails off</td>
<td>Cuts off after lag q</td>
<td>Tails off</td>
</tr>
<tr>
<td>PACF</td>
<td>Cuts off after lag p</td>
<td>Tails off</td>
<td>Tails off</td>
</tr>
</tbody>
</table>

Table 6.2: The ACF and PACF for ARMA Models

The EACF of the OZMAX1HR data is shown in the following table with the differencing applied to the data because the data could be considered stationary after differencing:
Table 6.3: Extended ACF for Difference of OZMAX1HR

In this case, we can try the ARMA model with \( p = 2 \) and \( q = 2 \). The parameter estimation is shown in the following table:

\[
\begin{array}{c|ccccc}
\text{coefficient} & \phi_1 & \phi_2 & \theta_1 & \theta_2 \\
\hline
\text{s.e.} & -0.1461 & 0.3927 & -0.1805 & -0.7086 \\
\text{AIC} & 0.0649 & 0.0426 & 0.0618 & 0.0586 \\
\end{array}
\]

Table 6.4: Parameters Estimation for ARMA\((2, 2)\) Model

So the ARMA\((2, 2)\) model is:

\[
Y_t = -0.1461Y_{t-1} + 0.3927Y_{t-2} + e_t - (-0.1805)e_{t-1} - (-0.7086)e_{t-2} \quad (6.2)
\]

The AIC changes significantly, which indicates that the ARMA model may be an inappropriate one for the OZMAX1HR. The forecasts using ARMA\((2, 2)\) model
is shown in Figure 6.2. By observing the differenced result, the flat trend of the forecasting value is reasonable because of the differencing; however, comparing to Figure 6.1, both plots show increasing ranges of forecasts, and using the ARMA model provides a poor description of the OZMAX1HR data.

6.3 Concluding Remarks

In this chapter, we reaffirm that the seasonal integrated model is the best choice for our data. The forecast from the SARIMA model provides a similar trend as the original data, while the forecast from the ARIMA model can only describe the forecasting range and fails to describe the seasonal variation in each cycle. Further, the prediction from the mixed ARMA model performs even worse. So we can conclude that when dealing with seasonal or non-stationary data, the integrated model as well as the seasonal model should be considered. Otherwise we would either make an inaccurate forecast or choose the incorrect model fitting process.
Figure 6.1: Forecast of ARIMA Model, leading=200
CHAPTER 7

Conclusions

7.1 Result

The seasonal ARIMA model provides an acceptable description of the ozone daily data. We made reasonable assumptions at the first stage of model identification and delved into the model identification process to examine and verify the adequacy of the model, and the results were qualified in the model diagnosis stages. The forecast of OZMAX1HR (daily, average of maximum hours) illustrates the pattern as well as the seasonality of the data. According to the seasonal ARIMA model, the volume of OZMAX1HR is associated closely with the temperature, we can observe a systematic change that the concentration of ozone rises during the summer months of each year. However, the overall trend goes down due to the policies and actions of concerned authorities [17]. The seasonal ARIMA model may not be able to predict the effect of the environmental protection policy, so the downward trend will not be presents in our forecast. Again, the reason why we are not using the AR (Autoregressive), the MA (Moving average) or a mixed model such as ARMA (Autoregressive moving average) is that the models can only be applied to a stationary data. Applying the models to our OZMAX1HR data may be misleading and insufficient.
7.2 Further Studies

Plausibly, combinations of forecasts from different methods are generally better than forecasts from individual methods, and fitting a good model to historical data does not necessarily minimize the error in the post-sample forecasts [9]. So one purpose of our further study is to fit more flexible but parsimonious models to the non-stationary, seasonal data that can be seen frequently in practice. Furthermore, forecasting the future values of an observed time series is apparently one of the most important goals of doing time series analysis. In this paper, we applied the forecasting method called the Seasonal ARIMA (Box-Jenkins) procedure, which is an objective method for prediction. The significant advantage of the ARIMA model is that the forecast can be developed in a very short time [10]. As can be proved in this paper, more time is spent in examining and validating the data than building the models. However, the shortcoming of ARIMA models is the limited explanatory capability due to the essence of univariate models. Especially when the expectation is that the underlying data is affected by other variables, the univariate ARIMA model may not serve as the best choice. Thus, for better forecasting, we can induce more complicated methods such as 1) the Vector ARIMA model, which extends the ARIMA model to a multivariate process that can depict the relationship between variables and make the forecasts better, and 2) the Non-automatic approach, which requires some subjective inputs from the forecaster (for example, the opinion of experts considered as inputs in the process of forecasting). These two methods may lead to very different results from the Box-Jenskin method used in this paper.

Another possible area of research is the air pollution particles that affect our everyday life. In California, especially in the Los Angeles metropolitan area, the air pollution problem is an unavoidable issue. Besides the greenhouse gases,
there are tons of toxic gases emitted into the atmosphere every day. These gases include carbon monoxide (by-product of combustion), nitrous oxide and sulfur dioxide (a byproduct of the generation of electricity). Research on these gases can also contribute to more perceptive solutions to the problem of air pollution or policies to control damage to the environment as well as to our health. Furthermore, we can also explore the data of the 15 Basins established by the ARB (California Air Resource Board). For instance, if we focus on the climate change issue, we may choose the basins located along the coast; if we focus on the health issue, we may select the urban area for special study. The variety of model applications to the air data can provide different solutions to a variety of problems, however, we should always remember that the more precise our forecast, a consequence of the careful and correct application of statistical models to our data, the more confidence that we can place in the decision making process.
REFERENCES


