Air Pollution PM2.5 Data Analysis In Los Angeles Long Beach With Seasonal ARIMA Model

Weiqiang Wang
School of computer science, Beijing Institute of Technology,
Beijing, China, 10081
Department of Statistics, University of California-Los Angeles, Los Angeles
Weiqiang.wang1983@gmail.com

Ying Guo
School of Management and Economics, Beijing Institute of Technology,
Beijing, China, 10081
violet7376@gmail.com

***Abstract*—An autoregressive integrated moving average (ARIMA) is one of the popular linear models in time series forecasting during the past three decades. Recently many environmental and socioeconomic time series data can be adequately modeled using the seasonal ARIMA model, also known as seasonal Box-Jenskins approach, and based on the fitted model. this paper presented a general expression of seasonal ARIMA models with periodicity and provide parameter estimation , diagnostic checking procedures to model, and predict PM2.5 data extracted from the California Air Resource Board using seasonal ARIMA models, we show experimental results with Los Angeles long beach PM 2.5 data sets indicate that the seasonal ARIMA model can be an effective way to forecast air pollution.**

Keywords- Seasonal ARIMA, Air Pollution, pm2.5

I. INTRODUCTION

Particulate matter is the term used for a mixture of solid particles and liquid droplets found in the air. PM2.5 refers to particulate matter that is 2.5 micrometers or smaller in size. The sources of PM2.5 include fuel combustion from automobiles, power plants, wood burning, industrial processes, and diesel powered vehicles such as buses and trucks[1].

As we know, the Los Angeles metropolitan area has the worst PM2.5 pollution in the United State. The main air quality observation institution in California is the California Air Resources Board (ARB)[2], established in 1967 and aimed at maintaining healthy air quality. They conduct research into the causes of poor air quality and provide. Therefore, in this paper, we will use the database from ARB.

Air pollution is an effected by many environmental factors, and the factors have a complicated correlation, it's hard to explain the correlation with a structure casual model. At this time it is an effective method to establish the time series dynamic model in accordance with its own law, and air pollution often have long-term trends, seasonal, cyclical, short-term fluctuations and irregular changes. The predictive of analysis of such data always need to consider these characteristics. If we used a simple time-series model, such as the model of ARMA, often does not fully reflect the characteristics of the environmental damage itself.

Seasonal ARIMA is a good analytical method which is based on a time series and interrelated dynamic data. We have selected Los Angeles Long Beach PM2.5 datasets extracted from the California Air Resource Board and used the Seasonal ARIMA model to construct an analysis.

II. DATA

The PM-2.5 data was produced by the California Air Resources Board. The mission of the California Air Resources Board is to promote and protect public health, welfare and ecological resources through the effective and efficient reduction of air pollutants while recognizing and considering the effects on the economy of the state. LA long beach is No.174 station, which is located at 17 Latitude: 33 deg 47 min 50 N, Longitude: 118 deg 05 min 38 W. The monitor gives daily measurements of PM-2.5 (fine particulate matter). A monthly average is calculated from January 1999 to January 2006.The data is organized in a data frame, with value as rows and days as columns and the data is 12×8, with no missing data, so we have 96 observations for each variable.

Let's first look at the standard. For PM-2.5 California only has an annual standard of 12μg/m3 (i.e. microgram per cubic meter). There are no hourly or daily standards. There is a federal 24 hour standard of 35μg/m3 and a federal annual standard of 15μg/m3.If we look at the average of our 2 observations , we find 18.608 μg/m3, this is over the federal standard as well as the annual state standard.[2]

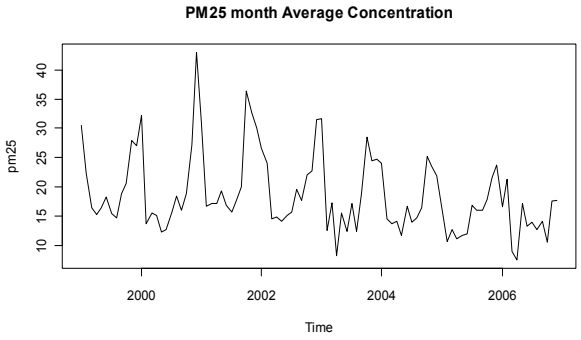

Figure 1. LA long beach PM2.5 month average

978-0-7695-3819-8/09 $26.00 © 2009 IEEE
DOI 10.1109/ICEET.2009.468

Figure 1. shows the monthly PM2.5 concentration levels from January 1999 through December 2006. There is obviously a strong downward trend. In this display, PM2.5 concentrations in Long Beach tend to be highest in the late fall and winter months. This is due mainly to meteorological conditions which occur more frequently in late fall and winter[4]. Observation suggests that the variation series is not constant over time and that there is a trend as well as a seasonal pattern in the data. This PM2.5 time series have to be determined as non stationary.

III. ARIMA MODEL

In the late 1960s,Box and Jenkins advocated ARIMA methodology for time series based on finite-parameter models.

An autoregressive integrated moving average (ARIMA) model is fitted to time series data either to better understand the data or to predict future points in the series[5]. The model is generally denoted to as an ARIMA(p,d,q) model where p is the number of autoregressive terms, d is the number of non seasonal differences, and q is the number of lagged forecast errors in the prediction equation[6].

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \quad (1)$$

In equation (1),where L is the lag operator, the ϕ_i are the parameters of the autoregressive part of the model, the θi are the parameters of the moving average part and the ε_t are error terms. The error terms ε_t are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

IV. SEASONAL ARIMA MODEL

The seasonal part of an ARIMA model has the same structure as the non-seasonal one: it may have an AR factor, an MA factor, and an order of differencing. The seasonal autoregressive integrated moving average model of Box and Jenkins(1970) [3]is given by

$$\left[\nabla^d \nabla_s^D Y_t-\mu\right]=\frac{\theta(B)\Theta(B^S)}{\phi(B)\Phi(B^S)}e_t \quad (2)$$

and is denoted as an ARIMA(p,d,q)×(P,D,Q)S

where P is number of seasonal autoregressive (SAR) terms, D is number of seasonal differences, Q is number of seasonal moving average (SMA) terms. 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, P, Q is numbers of seasonal, autoregressive terms and seasonal moving average terms, respectively. Further,

$$\phi(B)=1-\phi_1 B-\phi_2 B^2-\ldots-\phi_P B^P \quad (3)$$

$$\theta(B)=1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q \quad (4)$$

$$\Phi(B)=1-\Phi_1 B-\Phi_2 B_s^2-\ldots-\phi_p B_s^P \quad (5)$$

$$\Theta(B)=1-\Theta_1 B-\theta_2 B_s^2-\ldots-\theta_Q B_s^Q \quad (6)$$

and

$$\omega_t=\nabla_s^D\nabla^d Y_t \quad (7)$$

The last equation illustrates the multiplicative seasonal behavior indicating that seasonal and consecutive differencing may be required to induce stationary. 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.

V. ESTIMATIONS AND DIAGNOSTIC CHECKING

The strong seasonal autocorrelation relationships are shown in Figure1. Evidence shows that there is substantial other correlation that needs to be modeled. Clearly we need at least one order of differentiation.

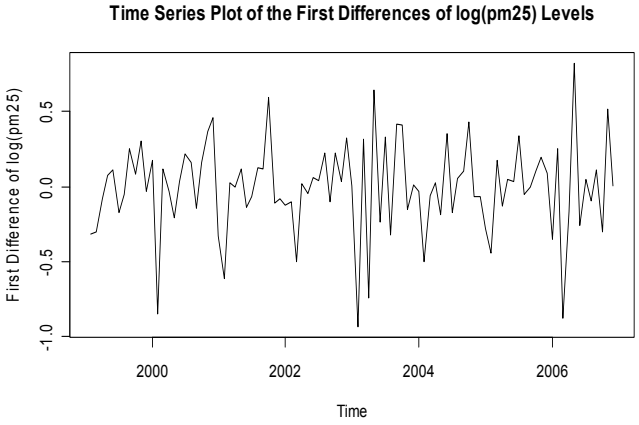

Figure 2. First Difference of Log(PM2.5) Levels

Figure 2 shows the time series plot of the log(PM2.5) levels after we take a first difference. The downward trend has now disappeared, but the strong seasonality is still present as evidenced by the behavior shown in Figure 2. In addition, the inconstant variation still seems to be a big problem. Maybe seasonal differencing will bring us to a series that may be modeled parsimoniously.[7]

The numbers of AR and MA can be identified in a more systematic way by ACF and PACF plot.

A. Parameter Estimation

Applying the SARIMA model to our pm2.5 data, the following table shows:

TABLE I. PARAMETER ESTIMATION

coefficient	θ Θ
	-0.8645 -0.6894

s.e	0.0790 - - - 0.1268
AIC	495.3

Thus our SARIMA(0,1,1) ×(0, 1, 1)12 model is:

$$Yt=Y_{t-1}+Y_{t-12}-Y_{t-13}+\varepsilon_t+0.8645\,\varepsilon_{t-1}+0.6894$$

$$\varepsilon_{t-12}+(0.8645)(0.6894)\ \ \varepsilon_{t-13}$$

B. Diagnostic Checking

Diagnostic checking is necessary to ensure the best forecasting model has been built.[8] The coefficient estimates are highly significant and the estimated value is small. However, in order to check the estimated ARIMA(0, 1, 1) × (0, 1, 1)12 model, there are some other things to check. Let's look at the time series plot of the residuals first.

Firstly, look at the time series plot of the residuals. Figure 5.2 gives the standardized residuals.

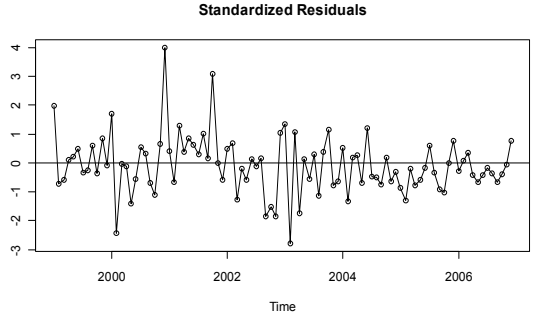

Figure 3. Residuals from the ARIMA(0,1,1)×(0,1,1)12 Model

We can see from Figure 3, the standardized residuals form the ARIMA(0,1,1)×(0,1,1)12 Model. Other than some strange behaviors in the middle of the series, such as 2000 winter, this plot does not suggest any major irregularities with the model. The residuals even balance out around zero, it seems constant. And neither additive outliers nor innovative outliers have been detected. To take a further look, we have plotted the sample ACF of the residuals.

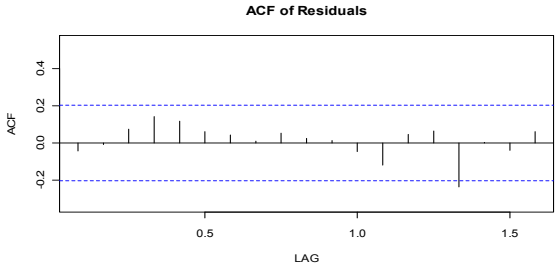

Figure 4. ACF of Residuals Plot

The sample autocorrelation function (ACF) plot with the vertical axis is almost close to 0.Taking the first differences produces a very clear pattern in the sample ACF[9], so successive differencing has no need to be carried out on this data.

Secondly, we use Histogram and Q-Q plot approaches for testing normality and identifying outliers.

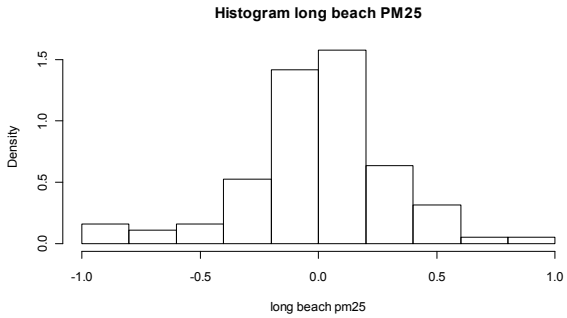

Figure 5. Histogram of the residuals

By plotting the histogram of the residual we can see the center seems a little bit off from zero, and the shape of histogram appears "good shaped". However, the Shapiro-Wilk test of normality has a test statistic of W = 0.9643, leading to a p-value of 0.01088, and normality hypothesis cannot be rejected.

TABLE II. SHAPIRO-WILK NORMALITY TEST

Data: Residuals:
W=0▷9643 p-value=0.01088

To take a further look, another way to test the normality is Q-Q plot. we would like to determine if outliers exist. We can take a look at the normal Q-Q plot formed by residuals, from the Q-Q plot below, we can see there are not many outliers, almost all the points are laid on the line.

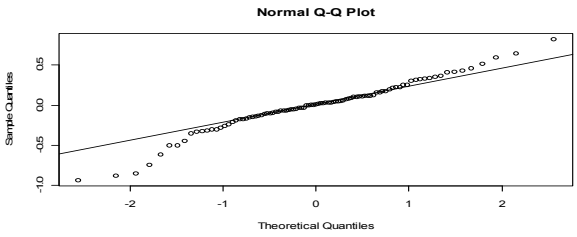

Figure 6. QQ-plot of the residuals

Above at all, through the diagnostic checking including standardized residuals, histogram of the residual, QQ-plot of the residuals, the experiment results indicate that model expresses very well. We could conclude PM2.5 can be represent very well by ARIMA(0,1,1)×(0,1,1)12.

VI. FORECASTING

To examine our prediction, we compare the actual values with the predicted values. The result appears in Figure 7.

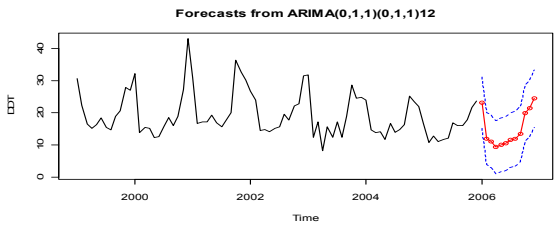

Figure 7. Comparison: Actual Value and Predicted Value

In Figure 7 the first plot is LA long beach PM2.5 actual value from January 1999 to December 2006, like figure 1, there has a obviously strong downward trend in the plot, and the second plot is the same from January 1999 to December 2005, but with the prediction 2006. Figure 6.1 shows the forecasts and 95% forecast limits for a lead time of one year for the ARIMA(0, 1, 1) × $(0, 1, 1)_{12}$ model.[10] The last one year of observed data is also shown. The forecasts mimic the stochastic periodicity in the data quite well, and the forecast limits give a good feeling for the precision of the forecasts.

VII. 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.

A. non-seasonal ARIMA Model

If we apply a non-seasonal ARIMA model to the Long Beach data with the same power transformation, λ equals to -0.5, the parameters for ARIMA(0,1,1) are:

TABLE III. Parameters Estimation for Non-Seasonal Model

coefficient	θ
	-0.0504
s.e.	0.1093
AIC	600.16

The parameter θ changed slightly without considering the seasonality. However, the AIC becomes worse going from 495.3 to 600.16.

B. Mixed ARMA Model

The mixed model is also called ARMA (Autoregressive Moving Average process).

In general, if Yt is a mixed ARMA process of orders p and q, we abbreviate the name to ARMA(p,q), and it can be defined as:

$$Y_t = \mu_t + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j} \quad (8)$$

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

TABLE IV. Parameters Estimation for ARMA(2,2) Model

coefficient	ϕ_1	ϕ_2	θ_1	θ_2
	1.47	-0.72	-0.84	0.19
s.e.	0.15	0.14	0.23	0.19
AIC	586.53			

So the ARMA(2,2) model is:

$Yt=1.47Yt\text{-}1+\text{-}0.72Yt\text{-}2+\varepsilon_t$ -(-0.84) ε_{t-1} -(-0.19) ε_{t-2} (9)

AIC changes significantly, which indicates that the ARMA model may be an inappropriate one for the Long beach data. Both results show increasing ranges of AIC, and using the ARMA model provides a poor description of the data.

C. 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.

VIII. Conclusions

The reason why we don't use ARMA model is that the model can only be applied to a stationary data. A seasonal ARIMA(0, 1, 1)×(0.1.1)12 model has been fitted. A forecasting plot has been drawn. The forecast of PM2.5 illustrates the pattern as well as the seasonality of the data. The model will be helpful to predict the air pollution PM2.5. So we can see seasonal ARIMA model has board applicability in the field.

References

[1] http://www.epa.gov/

[2] http://www.arb.ca.gov.

[3] G.E.P. Box and G.M.Jenkins. Time Series Analysis: Forecasting and Control.Holden-Day, revised edition, 1976.

[4] Bernhard Pfaff. Analysis of Integrated and cointegrated Time Series with R. Springer, September 2005.

[5] W.N Venables and B.D.Ripley. Modern Applied Statistics with S, Springer, January 2002.

[6] Robert F. Nau. Introduction to arima: nonseasonal models, 2005. http://www.duke.edu/rnau/411arim.htm.

[7] Maindonald, J., and Braun, J., Data Analysis and Graphics Using R, An Example-Based Approach, Cambridge University Press, Second Edition,2007.

[8] Shumway, R.H., and Stoffer, D.S., Time Series Analysis and Its Applications with R Examples, Second Edition, Springer, 2006.

[9] H. Akaike, Time Series Analysis and Control Through Parametric Models, Applied Time Series Analysis, New York 1978.

[10] Bovas & Johannes. Statistical Methods for Forecasting.John Inc, 2005.