ISyE 6402 Final Project
Time Series Modelling and Forecasting of Air Pollutants

Nevin Thomas[1]

[1]MS in Operations Research, GT ID: 903387577, nevinthomas@gatech.edu

Abstract. *Air pollution has been a serious issue around the world for the past few decades. Identifying the rise and fall in the concentration of individual pollutants and forecasting their future concentrations is of primary importance in determining the quality of air in a particular area. Although, there have been multiple studies in the past aimed at developing time series models to forecast the concentration of individual species, there are very few which addresses the inter-dependencies among pollutant species and environmental factors. Therefore we propose a multivariate time series analysis of the four most common air pollutants: NO_2, O_3, SO_2 and CO in the Dekalb county of Georgia State in USA from 10-01-2010 to 01-31-2016. We start out by doing a parametric trend and seasonality estimation and building ARMA models separately on the residuals for each species. We then move on to build VAR and VARMA models to capture the relations between different pollutants. We also add temperature as an exogenous variable to capture the meteorolgical effects on pollutant concentrations. The residual analysis of these models point us to violation of the normal assumption and constant variance assumption. Therefore, we develop a multivariate GARCH model to capture the volatility in the data. We use the MAPE on test set as the performance metric for the forecasts produced by different models. While all the models present us with good quality forecasts, multivariate GARCH stands out as the best performer. We also observe that adding temperature as an exogenous variable does not really contribute significantly to predictive power of models.*

1 Introduction

The aftermath of industrial revolution has resulted in a massive increase in the demand for energy sources. This increase in demand has been catered largely by conventional sources of energy like and coal and fossil fuels. While exploiting these natural resources have opened the gateway to large-scale economic development, environmental pollution across the world was a consequence. Conventional power plants and industries which generate energy by burning fossil fuels has contributed a great deal towards the accumulation of greenhouse gases in the atmosphere. This has been a cause of concern for countries around the world and especially US as the leading industrial nation. Many regulations have been put in place and migration to green energy sources have been proposed in the past to address this deterioration in the air quality.

However, studying the behavior of air pollutants and their effect on air quality would help us gain valuable insights into dealing with the problem. Looking at the air pollution data from previous years might also help us forecast the concentration of pollutants in future and might help us in recommending strategies which can improve air quality. Very often, the concentration of individual air pollutants in an area is related to each other. Hence, we propose a statistical time series analysis of the concentration of NO_2, O_3, SO_2 and CO in the Dekalb County of Georgia, United stated from 10-01-2010 to 01-31-2016.

Photochemical smog has become a common phenomenon in many part of the world. Chemical reactions between pollutant like hydrocarbons and nitrogen oxides (eg. NO_2) results in the production of oxides such as ozone (O_3) which can sustain in the atmosphere for over a month. Ozone is quite a potent pollutant which can cause eye and lung irritation in humans and damage to animals and vegetation [1].

The chemistry of transformation of primary pollutants into the secondary ones is very complicated. The time dependent response and production of secondary pollutants in a particular area has a strong dependence on hydrocarbon and nitrogen oxide composition there. Many past studies have demonstrated the time dependent rise and fall of the NO_2 and O_3 concentration. The time lags of these concentration fluctuations are further affected by meteorological factors such as temperature, wind speed and wind direction [1].

While there are numerous past studies which model pollutant species individually using stationary autoregressive moving average models, they fail to capture the complete picture. The inter-correlations among the pollutants cannot be investigated using such models. In a practical scenario, pollutants may be introduced continuously into the environment in an irregular pattern. It would be a very cumbersome task to track all the source inventory and the time dependent concentration of primary and secondary pollutants along with meteorological variables. Therefore, we focus on investigating the interdependence between the NO_2, O_3, SO_2 and CO. We also include temperature as an exogenous variable in our studies to account for the meteorological effects. We implement multiple statistical time series models and compare their ability to produce forecasts of good quality.

2 Data set

The data set we use for this study consist of four time series corresponding to the daily average concentrations of NO_2, O_3, SO_2 and CO. It has 1949 observations from 10-01-2010 to 01-31-2016 as collected from a checkpoint in the Dekalb county in the Georgia state of the United States of America. We obtained the data from the pre-generated data files repository on the Unites State Environmental Protec-

tion agency website.

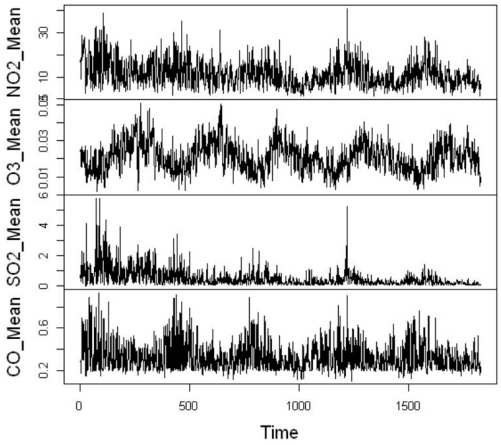

Figure 1: Time series of the four species under analysis from 10-01-2010 to 01-31-2016

The variation in the daily avaerage concentration of the four species during the above mentioned duration is depicted in Figure 1. As is evident from the time series plot, NO_2, O_3 and CO exhibit a more or less yearly seasonal trend. While we cannot make any such statement about the SO_2 series, its concentration seems to be decreasing over the years, which is indeed a good sign. The yearly seasonality can be more clearly distinguished using the month wise box-plots attached in the Appendix Figure 6. We also explore if there exists any day of the week seasonality in Appendix Figure 7.

All the time-series models are trained on the data till the end of 2015. The performance is evaluated by calculating the mean absolute percentage error (MAPE) of the forecasts on the test set which is the data for the first month of 2016.

3 Methodology

A time series can usually be decomposed as follows

$$Y_t = m_t + s_t + X_t \tag{1}$$

where m_t is the trend component, s_t is the seasonal component with known periodicity and X_t is the stationary component with a probability distribution that does not change with time.

3.1 Trend and Seasonality Estimation

Very often, m_t and s_t are first estimated and subtracted from Y_t to have left the stationary process X_t to be modelled using time series modeling approaches. There are many approaches toward estimating trend and seasonality separately from a series and removing it to get the stationary residuals. In this study we adopt a parametric regression approach to estimate both together. We use the variables t and t^2 in the regression to capture the nonlinear trend. The seasonality is captured by fitting a mean for each seasonality groups - month and day of the week. We used a regression model with intercept, 11 dummy variables for month and 6 dummy variable for the day of the week.

3.2 ARIMA and ARIMAX

An autoregressive integrated moving average (ARIMA) model is defined as follows. If X_t is process for which,

$$\begin{aligned} (1-B)^d X_t &= Y_t \\ \phi(B) Y_t = \theta(B) X_t,\ Z_t &\sim WN(0, \sigma^2) \end{aligned} \tag{2}$$

where d is a non-negative integer, B is the backward operator and the functions $\phi(x)$ and $\theta(x)$ are defined as,

$$\begin{aligned} \phi(x) &= 1 - \phi_1(x) - ... - \phi_p x^p \\ \theta(x) &= 1 - \theta_1(x) - ... - \theta_q x^q \end{aligned} \tag{3}$$

where p and q are non-negative integers, then X_t is said to be an ARIMA(p, d, q) process. ARMA models are a special case of ARIMA model when $d = 0$.

When an exogenous variable is added as a regressor variable to an ARIMA framework, then it is called an ARIMAX model.

3.3 VAR and VARX

The vector autoregression (VAR) model is one of the most flexible and easy to use models for the analysis of multivariate time series. The basic p-lag vector autoregressive (VAR(p)) model has the following form:

$$Y_t = c + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + ... + \pi_p Y_{t-p} + \epsilon_t \tag{4}$$

For example, the bivariate case can be written as follows,

$$\begin{aligned} \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} &= \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} \\ &+ \begin{pmatrix} \pi_{11}^2 & \pi_{12}^2 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2+} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} \end{aligned} \tag{5}$$

where $Cov(\epsilon_{1t}, \epsilon_{2s}) = \sigma_{12}$ for $t = s$ and 0 otherwise.

When an exogenous variable is added as a regressor variable to the VAR framework, then it is called a VARX model. If the moving average part of the ARMA model is also included in the multivariate format, then it is called a VARMA model.

3.4 GARCH

The assumption of constant conditional is violated in many practical time series datasets. Auto Regressive Conditional Heteroskedasticity (ARCH) models are of ten used to model the time varying volatility. A generalized version which can overcome the limitations of the ARCH model is the GARCH. The simplest specification GARCH(1,1) can be defined as,

$$\begin{aligned} Y_t &= \mu + Z_t, \quad Z_t | F_{t-1} \sim \mathcal{N}(0, \sigma_t^2) \\ \sigma_t^2 &= \gamma_0 + \gamma_1 Z_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned} \tag{6}$$

More generically, there are GARCH(m, n) specifications where σ_t^2 is given by,

$$\sigma_t^2 = \gamma_0 + \gamma_1 Z_{t-1}^2 + ... + \gamma_m Z_{t-m}^2 + \beta_1 \sigma_{t-1}^2 + ... + \beta_n \sigma_{t-n}^2 \quad (7)$$

GARCH can be viewed as an ARMA form of heteroskedasticity. In this study we use a multivariate version with a robust version of VAR based on the multivariate Least Trimmed Squares Estimator described in Croux *et. al.* [2] as the mean model. We implemented this using the `rmgarch` package [3] in R software. We can extend this model to a GARCHX model by adding external regressors to the basic framework.

3.5 Performance Metric: MAPE

The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula:

$$M = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad (8)$$

where A_t is the actual value and F_t is the forecast value. We use MAPE on the test set to evaluate the performance of the different models that we implement in this study

4 Analysis and Results

Before starting with the modelling, we plot the Autocorrelation and Cross-correlation among different species which is displayed in Figure 2. The sinusoidal nature of autocorrelation (all except SO_2) and cross-correlation plots affirms the presence of annual seasonality in the data. This implies that the series is non-stationary. The partial autocorrelation plots are attached in Appendix Figure 8.

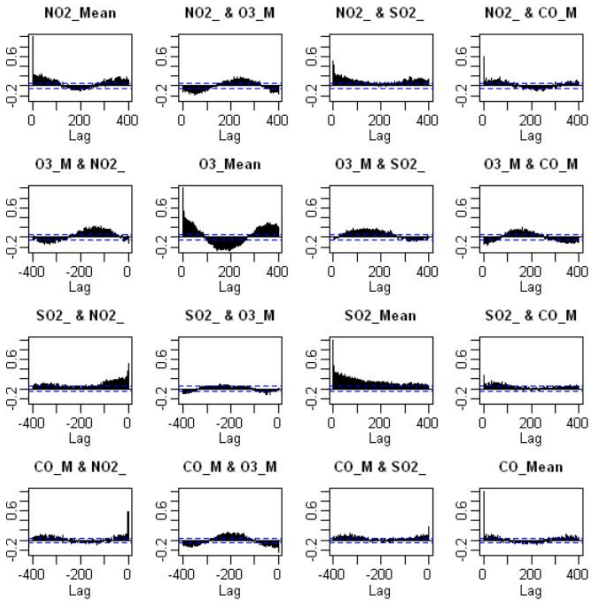

Figure 2: ACF of pollutant concentration time series

Therefore, we start out by estimating the trend and seasonality in the series corresponding to each of the specimen. We estimated the trend and seasonality using a simple linear regression model with the following independent variables: time points, square of time points, day of the week and month. We included the weekday variable because it improved the performance of the regression model as indicated by anova. The fitted trend-seasonality model is plotted separately for each of the specimen in Figure 3. We see that this regression model did a pretty goof job of capturing the variance in our data.

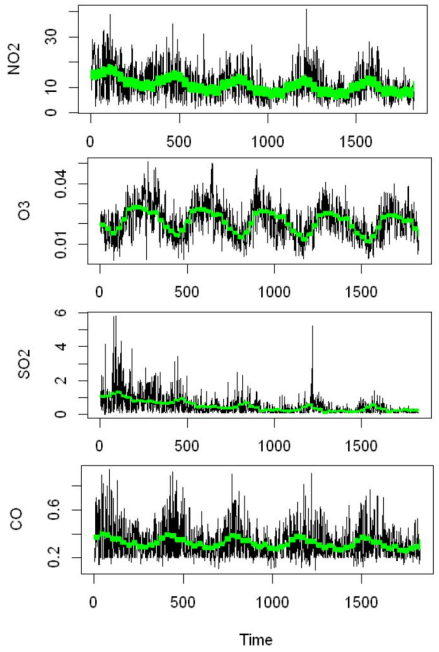

Figure 3: Fitted parametric trend and seasonality estimation

Now, we obtain a stationary series by subtracting the fitted values of the trend-seasonality model from the original time series. All the further modelling effort is done on these residuals which is plotted in Appendix Figure 9. The ACF plots for these residual series are attached in Figure 4, which look very close to that of white noise. The partial autocorrelation plots for the residuals are attached in Appendix Figure 10. We conduct an augmented Dickey-Fuller test on these residuals and the resulting p-values support the hypothesis that they are stationary.

We start out the modelling effort by fitting separate ARIMA models on the time series corresponding to the concentration of each of the species. Since, the residuals are already stationary, $d = 0$ for all the species resulting in ARMA models. The (p, q) order selection for each of the four models are done separately using the AIC criterion, details of which can be found in the jupyter notebook '*pollution_arma.ipynb*'. After developing the ARMA models, we compute the Box-Pierce and Ljung-Box test statistics for examining the null hypothesis of independent residuals. The p-values indicate that we can accept the null hypothesis for NO_2 and O_3 residuals and not for SO_2 and CO. Further, the normal Q-Q plots of NO_2 and O_3 exhibit

a very similar behavior to normal distribution, while SO_2 and CO displayed significant deviation. However, the Box tests on the squared residuals resulted in p-values much lesser than 0.05 for all the four series, indicating the presence of heteroskedasticity. We extend this ARMA model to ARMAX by adding temperature as a exogenous varibale. However, the model performance in terms of MAPE on the test set improved only for CO.

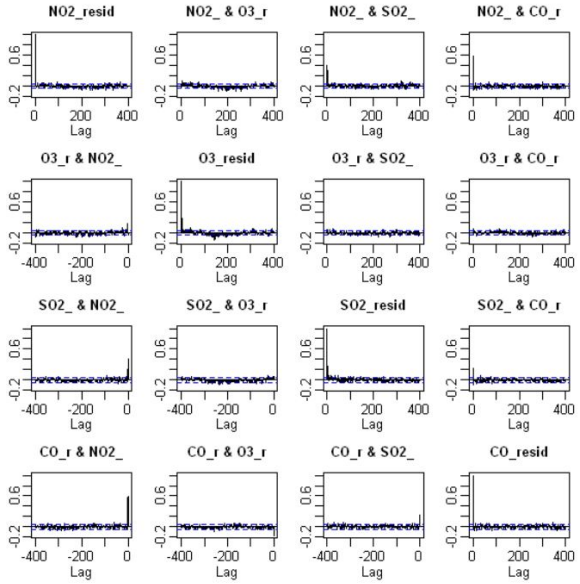

Figure 4: ACF of residuals obtained after trend-seasonality estimation

Now, we build a VAR model on the residuals of the four series together using the `MTS` package in R software. We normalized all the four series to a 0-1 scale before feeding it to the VAR function. We selected an order of $p = 2$ from the VAR model based on AIC criteria. Details of the implementation can be found in the jupyter notebook '*pollution-MVmodels*'. The MAPE produced on the test set is summarized in Table 1. We observe that the VAR model on the residuals do not really perform as well as the individual ARMA models. The inclusion of temperature as an exogenous variables worsens the performance further. We further extend the model to a VARMA(2,2) framework which performs better the VAR(2) model, although it still could not beat the individual ARMA models. The residual analysis of the VAR model fails the constant variance and normality assumptions as expected. The results of the serial test for checking uncorrelated errors assumption seems to be on the boundary and we could not make a solid conclusion about the correlation among residuals. The roots analysis results supports the stability of the VAR model.

The above described behavior raises a question about the power of the lagged concentration values of one species in predicting the concentration of another. We did Granger causality tests to address this and the results indicate that each of the four species taken separately Granger causes the other three. One possible explanation for this under-performance of VAR model compared to the simple ARMA model might be the the time dependent variance of the time series involved. Another possible reason might be the fact that we are working on the daily averages of concentration and temperature. Since the chemical reactions between different pollutants occur at a much smaller time scale, these averages might not actually be able to capture accurately the interdependencies among different species and the temperature.

Method	MAPE			
	NO_2	O_3	SO_2	CO
ARMA	0.453	0.391	1.811	0.401
ARMAX	0.485	0.418	2.083	0.383
VAR	0.478	0.397	2.094	0.403
VARX	0.504	0.400	2.312	0.412
VARMA	0.470	0.394	2.074	0.401
mGARCH	0.455	0.395	1.593	0.382
mGARCHX	0.467	0.393	1.595	0.374

Table 1: Table of MAPE for all the implemented methods

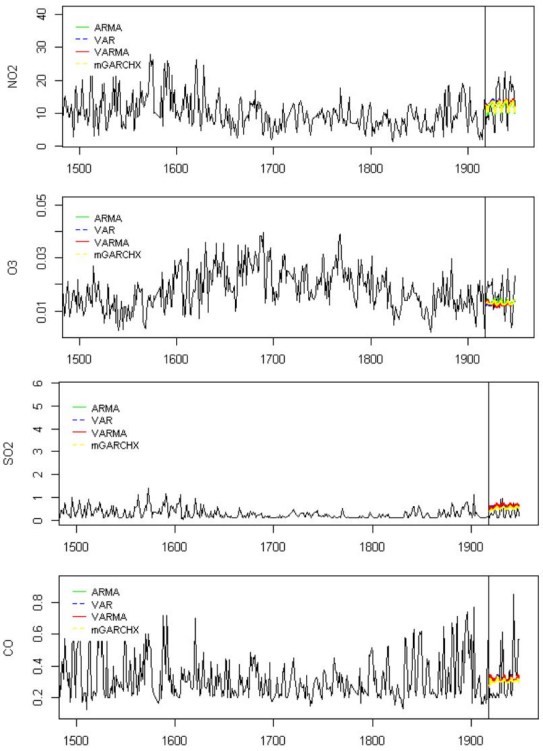

Figure 5: Forecasts produced by different models. The data before the time point corresponding to the vertical line is used to train the models and the data after is used for the evaluation of models.

Now to address the heteroskedasticity in the series, we develop a multivariate GARCH model on the pollutant time series. We chose a VAR(4)-GARCH(2,2) model and a VARX(5)-GARCH(2,2) model towards this purpose. We observe that the MAPE on the test set does improve when

the forecasts are done using the GARCH models. This is so because the GARCH models have the capability to account for the volatility in the series. The MAPE of SO_2 and CO forecast by the GARCH are way better than the other models, while the forecasts for NO_2 and O_3 are as good as the best among the rest of the models. The significant improvement in the accuracy of SO_2 and CO can be attributed to the fact that these were the two species whose concentration deviated significantly from the normal assumption.

Finally we list the MAPE values produced by all the implemented models on the test set in Table 1. We see that the MAPE values for SO_2 forecasts are particularly large compared to other species. However, this is due to the near zero values of the original SO_2 concentration series which comes in the denominator while calculating MAPE. We also plot the forecasts produced by the different models in Figure 5. We observe that, there are not really many significant differences in the species concentration forecasts. In fact, most of them overlap each other at many a time points.

5 Conclusion

A statistical time series analysis is the best way to analyze the rise and fall in the concentration of air pollutants in a particular area at a certain point in time. Time series models can produce reasonably accurate forecasts of the pollutant concentrations which can be crucial in estimating the air quality and identifying potential root causes of the pollution. Multivariate time series models are more suited for this purpose than univariate models because of inter-relations between the concentration of multiple pollutants. Adding meteorological variables as exogenous regressors might complement the predictive power of time series models. However, adding temperature as exogenous variable did not really improve the performance of the models in our study. This might be due to the fact that we are dealing daily average values of pollutant concentrations and temperature. Since most of the (photo-)chemical reactions involving pollutants happen at much smaller time scales, we might have to analyze higher frequancy data to exploit these relations improve the forecasts.

We also observe that the multivariate GARCH models significantly outperforms the VAR and VARMA models. This demonstrates the importance of checking the validity of the assumptions we make while developing models. Residual analysis of ARMA and VAR models indicate significant deviance from the normal assumption and constant variance assumption we make. The GARCH model, by the virtue of its ability to capture volatility in the time series data performs better.

5.1 Code

The R codes developed for doing the analysis can be found here.

References

[1] Kuang-Jung Hsu. Time series analysis of the interdependence among air pollutants. *Atmospheric Environment. Part B. Urban Atmosphere*, 26(4):491–503, 1992.

[2] Christophe Croux and Kristel Joossens. Robust estimation of the vector autoregressive model by a least trimmed squares procedure. In *COMPSTAT 2008*, pages 489–501. Springer, 2008.

[3] Alexios Ghalanos. *rmgarch: Multivariate GARCH models.*, 2019. R package version 1.3-6.

APPENDIX

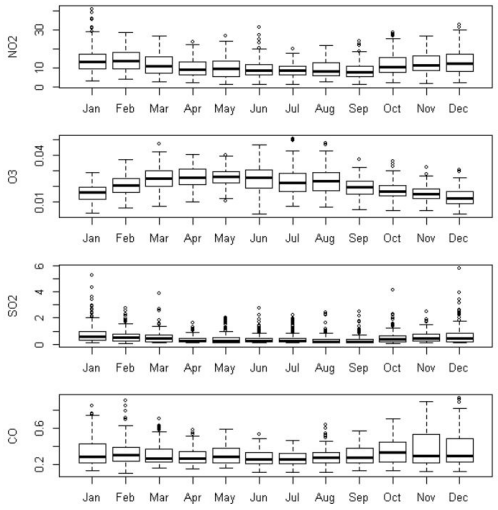

Figure 6: Box-plots demonstrating the day of the monthly seasonality

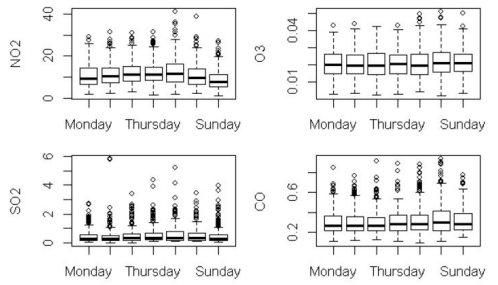

Figure 7: Box-plots demonstrating the day of the week seasonality

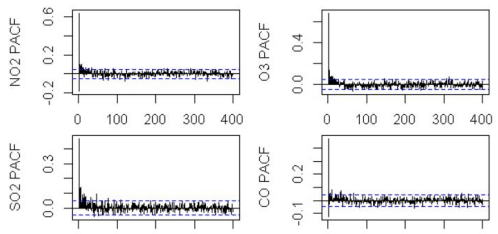

Figure 8: Residual Series

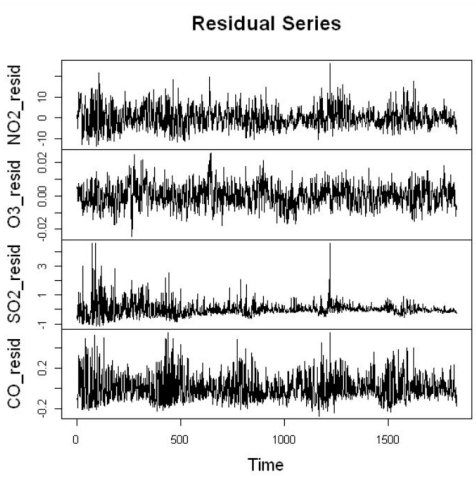

Figure 9: Residual Series

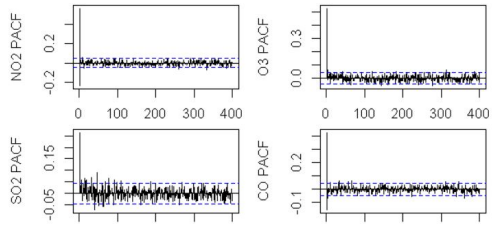

Figure 10: Residual Series