California Polytechnic State University, Pomona

College of Science

STA 5250

Time Series Analysis

Time Series Analysis of Air Quality in Urban Environments (Naples, Italy) Employing Various Modeling Approaches

Group Members :
Ryan Flynn
Kevin Tsoi
Taylor Paerels

Contents

1 INTRODUCTION

Air quality provides valuable insight into the macroscopic health of a geographic region, all else considered equal. To drive investigation into the topic at large, consider the following analysis on the air quality data set of ground truth values of the carbon monoxide (CO) measured in the urban environment of Naples, Italy.

2 DATA PRE-PROCESSING

2.1 PRELIMINARY DATA ANALYSIS

From March 10, 2004 to April 4, 2005, hourly recordings were taken concerning the air quality in Naples. Among those were hydrocarbon levels, benzene levels, and NOx levels. The focus of this analysis was specific to the carbon monoxide levels though. Due to the nature of having a date and time associated with each value, it was necessary to combine these vectors into an all encompassing change-in-time vector. The table of values could then be simplified to include only this new time vector and the vector of the average carbon monoxide levels.

2.2 DATA SET LIMITATIONS

This dataset consisted of 9357 observations, however, 1683 values, or 18% of the total, were unavailable for the carbon monoxide data in question. The set used -200 in place of each piece of absent data. These were changed to “NA” in order to eliminate the use of numerical values before further analysis could be conducted. Properly approximated values were required to show the changes which occurred over the time interval. Several attempts were made to fill in these gaps, with varying degrees of success.

3 DATA IMPUTATION

3.1 LINEAR SMOOTHING

Linear interpolation is a method of smoothing which uses linear functions to approximate values existing between two known points. It is useful in filling in the gaps in a data set as is the case here. As the name suggests, the points filling in these spaces reside on a linear segment connecting the known values on either side. It works best when the holes in the function are small, however certain sections of our data are quite large. Because of this, the plot of our set is not as cohesive as we would have liked.

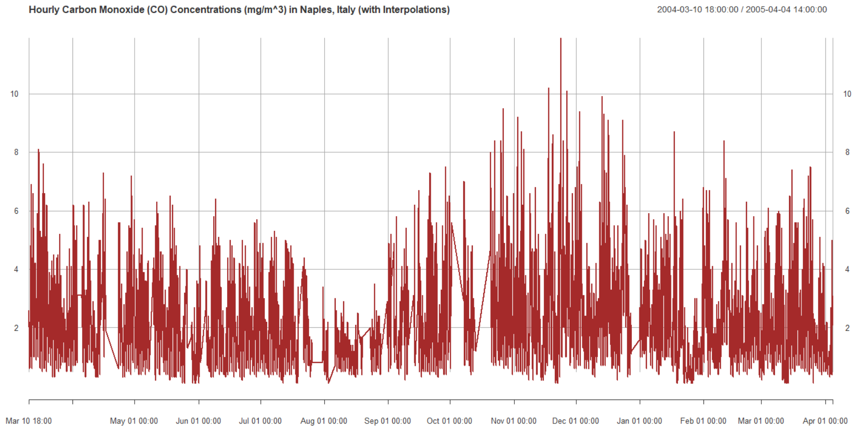


Figure 1: The original untransformed Air Quality Time Series

4 TIME SERIES MODELING

4.1 DATA ANALYSIS

Analyzing the time series plot of the data, we noticed that there were numerous spikes in variance which could be due to the linear smoothing imputations. Taking a natural logarithm and then a difference made the plot look stationary. There were occasional spikes near the end of the plot, but we concluded that it could be due to randomness in white noise.

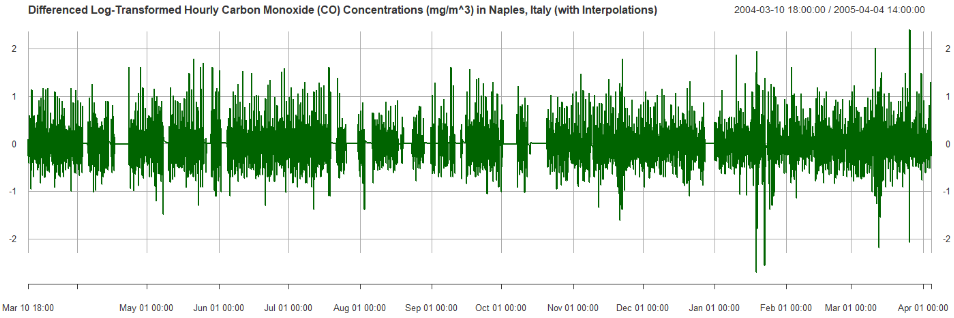


Figure 2: Application of $\nabla \log x_t$

Looking the acf and pacf plots, there is an obvious seasonal trend in both the acf and pacf plots. The trend reoccurs every 24th time difference which correlates to our data being hourly.

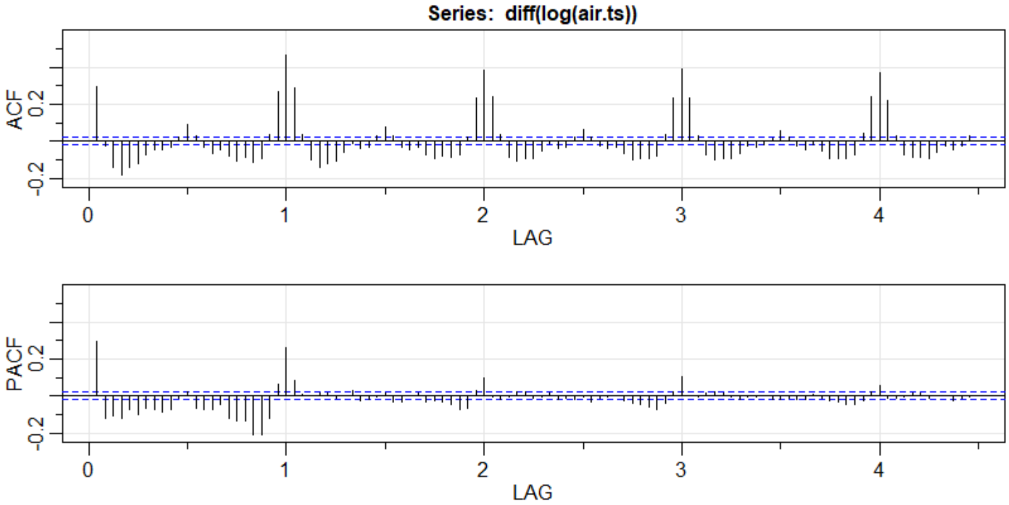


Figure 3: ACF, PACF plots of $\nabla \log x_t$

4.2 MODEL GENERATION

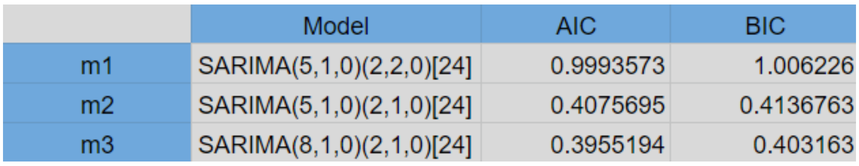
Looking the acf, and pacf, it strikes out to us to do some sort of SARIMA model to account for seasonality. Taking one or two seasonal differences for seasonal stationarity, we found 3 potential models we will consider:

1. ARIMA(5,1,0) x $(2,2,0)_{24}$
2. ARIMA(5,1,0) x $(2,1,0)_{24}$
3. ARIMA(8,1,0) x $(2,1,0)_{24}$

All three models were found using the same function `auto.arima`. For instance, model 1 was an experiment to see if adding another seasonal difference would provide a better model. Model 3 was the best model and model 2 was the 2nd best model from the auto.arima function on the basis of aic in an exhaustive search. Unfortunately, we could only perform a step-wise search on the 2nd seasonal difference due to how computationally intensive exhaustive searches are for seasonal models.

4.3 MODEL SELECTION

The performance metric we used to compare these models were AIC, and BIC. Below shows the table for comparison.



	Model	AIC	BIC
m1	SARIMA(5,1,0)(2,2,0)[24]	0.9993573	1.006226
m2	SARIMA(5,1,0)(2,1,0)[24]	0.4075695	0.4136763
m3	SARIMA(8,1,0)(2,1,0)[24]	0.3955194	0.403163

Figure 4: Information Criteria for SARIMA models

We found that the best model is Model 3: ARIMA(8,1,0) x $(2,1,0)_{24}$ on the basis of AIC and BIC. For Model 1, we do not have a direct comparison, but we want to see if it will satisfy the model assumptions.

4.4 MODEL DIAGNOSTICS

Looking at the diagnostics of Model 3, we see that many of our assumptions fail. For instance, the Q-Q plot shows that our standardized residuals from our data is not normally distributed. In addition, the ACF of our residuals show some peaks above the confidence band which suggests that we don't just have white noise or that our residuals is dependent upon each other. Finally, the p-values for the Ljung-Box statistic reveal that the p-values for the test is not significant. This means that our model may inadequate.

Model 2 was our second best performing model on the basis of both aic and bic. Looking at the diagnostics, we notice that our assumptions again fail. It doesn't seem like there was any improvement for our assumptions (Q-Q plot, ACF plot, Ljung-Box). Likewise, Model 1 still fails all of our assumptions even though we addded another seasonal difference.

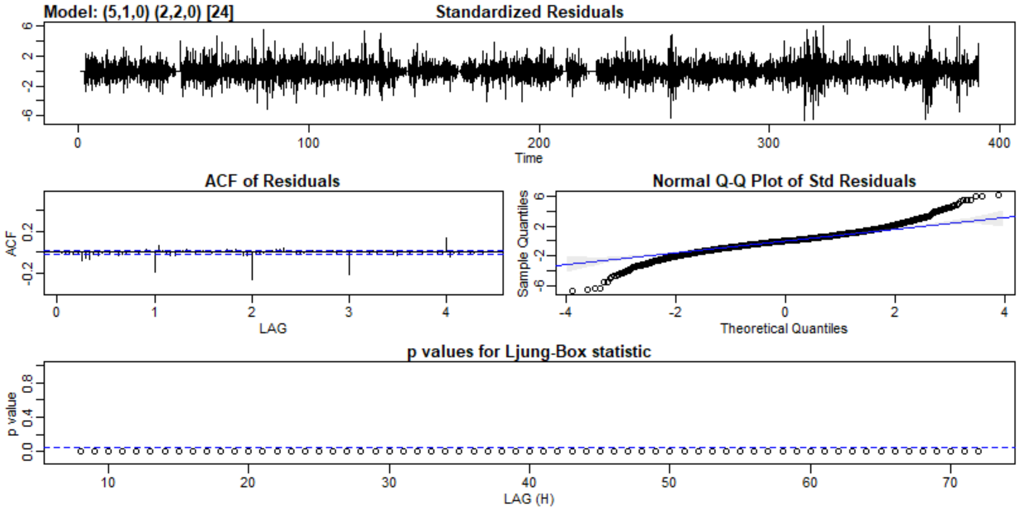


Figure 5: Diagnostics for SARIMA(5,1,0)(2,1,0)[24]

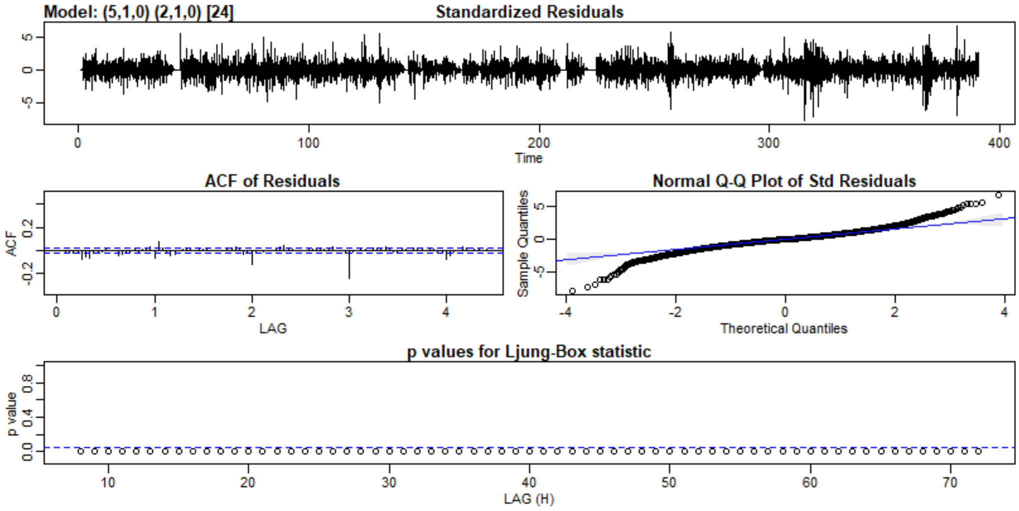


Figure 6: Diagnostics for SARIMA(5,1,0)(2,1,0)[24]

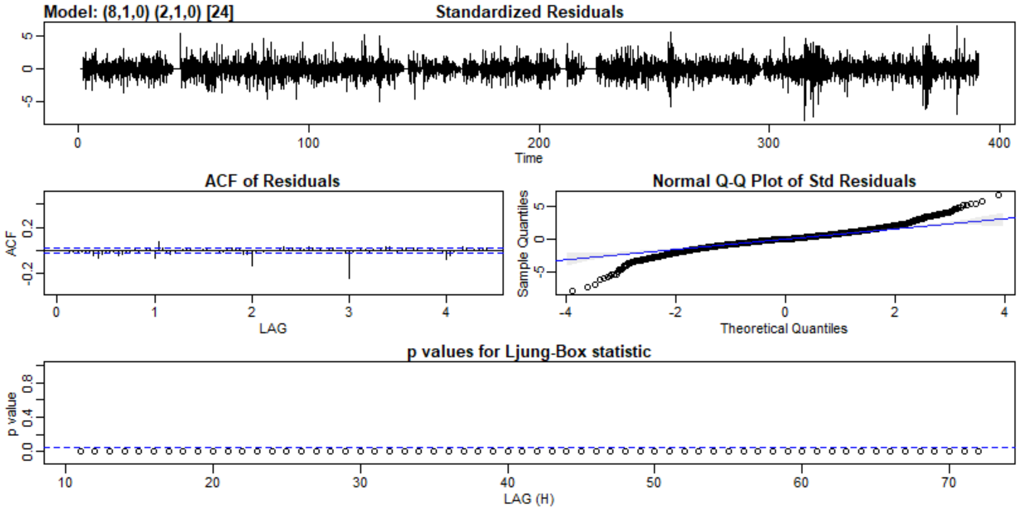


Figure 7: Diagnostics for SARIMA(8,1,0)(2,1,0)[24]

5 FORECASTING

Although we could not find an adequate model that satisfies our model assumptions, we have decided to use model 3 for forecasting since it had the best aic and bic. Below shows the 10-day forecast results. Obviously, it does not fit the data very well. The forecast looks like it believes that the data has a spiking upward trend.

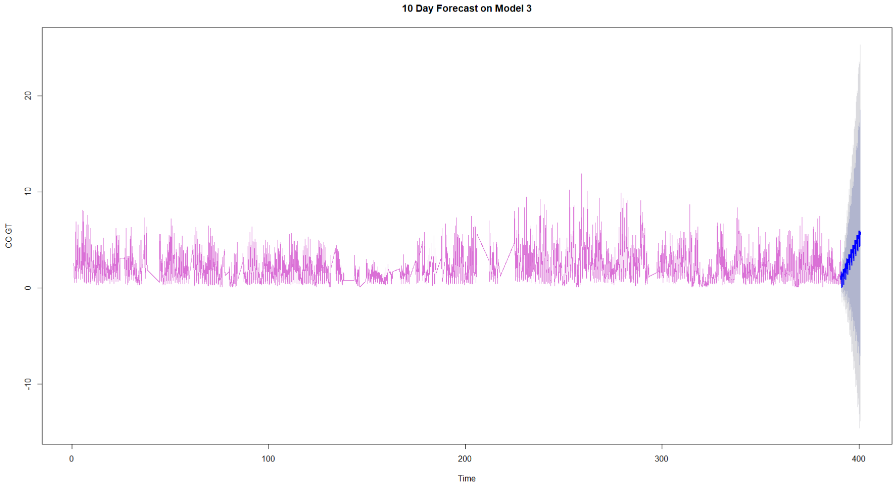


Figure 8: Forecast of 10 days beyond the final date (4/4/05 - 4/14/05

6 CONCLUSION

The air quality in Naples, Italy was studied extensively from March 2004 to April 2005. Among the many gases and particulates observed was carbon monoxide. Over 9000 hours worth of data were recorded during this time period, but only 82% of the hours yielded information. We attempted to fill in these holes and fit a model to the now completed set. Unfortunately, problems arose in with the former which is turn led to problems with the latter.

When attempts to use maximum likelihood estimation failed, methods of linear smoothing were utilized to bridge the gaps between the known values. Linear functions mapping out areas of a point or two worked well, however, spaces in several instances excluded several days worth of readings. With no other options, these line segments were kept.

To remove spiking variances, differencing and taking the natural logarithm of the data forced stationary on the set. As expected, the acf and pacf plots showed seasonal trends at 24 value intervals corresponding to each new day. With this in mind, three models were chosen. The diagnostics from each model did not help matters as they all failed the Ljung-Box statistic and had poor Residual ACF and Q-Q plots. As a result, our 10-day forecast did not perform well as expected.