Wake County Air Quality Analysis

CATHY TRAN
GRANT CLARK
EVAN LI
PRICE BURNETT
SUFYAN SHAHIN

September 25, 2019

Table of Contents

Wake County Air Quality Analysis

Executive Summary

A Moving Average (MA) term of 1 was fitted to the model to fully capture autocorrelations and leave only the white noise. Using the Augmented Dickey-Fuller test (ADF), Orange Team 2 determined the data to be stationary around the trend. Air quality data was collected exhibits a gradually decreasing trend between 2014 and 2018 at the Millbrook School station.

In 2012, the EPA set the acceptable level for yearly averages of $PM_{2.5}$ at 12 $\mu g/m^3$. Wake County continues to remain below this standard, but recent trends and forecasts may suggest a need to examine why these levels are on the rise again after years of improvement.

Results

After identifying that a trend exists within the data, the team used the ADF test to identify if the data was stationary around the trend or if a random walk was present. The ADF test showed sufficient evidence that the data was stationary around this trend line. After modeling the trend, the team proceeded to add an MA term to the model. This MA term, in addition to accounting for the trend, was able to fully model the data resulting in only white noise.

Recommendations

Considering the noticeable spike in $PM_{2.5}$ in 2016, and the seeming change in variance after that spike, it is recommended that similar modeling be conducted in the surrounding counties (Figure A, Appendix). If similar trends, spikes, and changes in variance can be observed nearby, then it may suggest that the changes in pattern are more regional. If these abnormalities are not observed in the surrounding counties, then it may suggest a more local issue. Once identified, this may provide insight as to how resources could be utilized in order to better understand why these differences occurred.

Methodology & Analysis

Data Used

The given dataset contains information about the air quality between 2014 and 2018 at the Millbrook School station at Wake County, North Carolina. Specifically, there are 18 variables and 1473 observations in the dataset.

Due to the 353 missing dates and values within the original time series, Team Orange 2 aggregated the data by monthly average and split the last 6 months as the test dataset for model testing and selection. The training dataset contains 54 observations from January 2014 to June 2018.

Testing Stationarity

Team Orange 2 performed the ADF test to check for stationarity around the trend going back up to two lags. The significant p-values (0.0007, 0.0010, and 0.0171) led us to conclude that there is stationarity about the trend.

White Noise

Since the stationarity about the trend assumption was met, a linear regression line was fitted to account for the trend component (Figure A, Appendix). After fitting a linear regression, there was no obvious trend in the residuals plot so we concluded that the residuals plot is stationary (Figure 1). Our team moved on to check for white noise. The x-axis in the Ljung Box Test P-values refers to the number of lags and the y-axis is the white noise probabilities (Figure 2). Since none of the lags were above the 0.05 probabilities, this indicates that there is no white noise and the autocorrelation structure needs to be modeled.

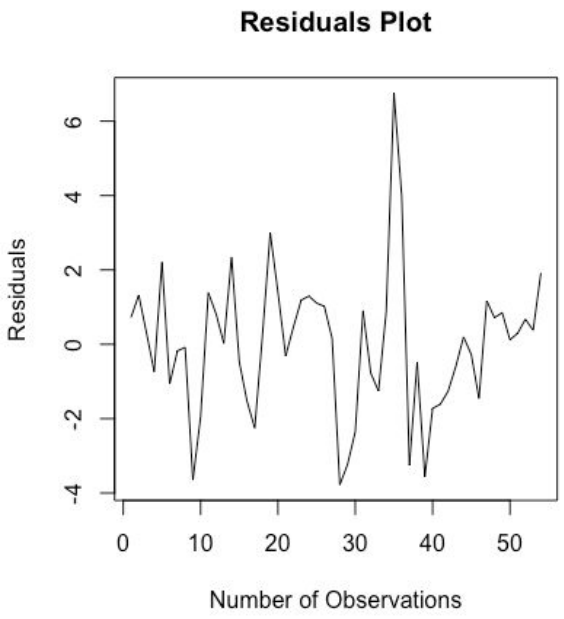

Figure 1. Residuals Plot

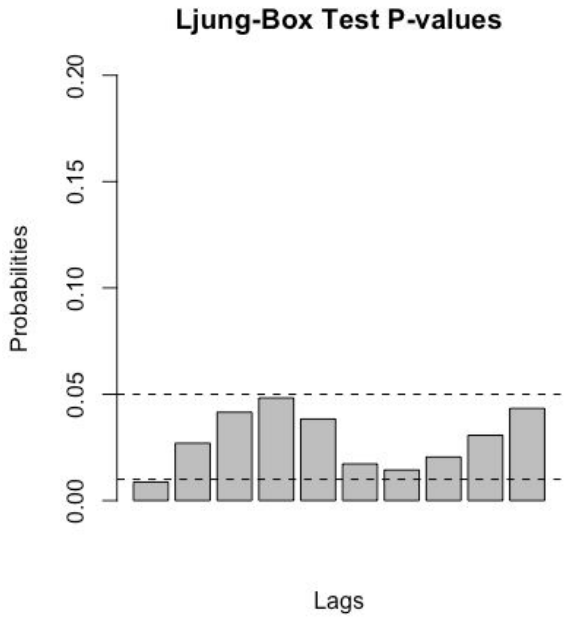

Figure 2. Ljung Box Test

Moving Averages and Autoregressive Terms

Team Orange 2 modeled the autocorrelation structure by figuring out how many MA and Autoregressive (AR) terms are needed to fully capture all of the signals. For both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plot of the residuals before fitting AR and MA terms, the x-axis represents the number of lags and the y-axis shows the amount of autocorrelation (Figure 3 and Figure 4). The big spike in lag 1 of the ACF plot suggests that we should fit an MA term of 1 whereas the big spike in lag 1 of the PACF plot suggests that we should fit an AR term of 1.

After modeling AR(1), lag 10 still fell outside the confidence limit (Figure 5). This suggested remaining autocorrelation. In contrast, after modeling MA(1), all of the lags in the PACF plot were within the confidence limit (Figure 6). This led us to conclude that an MA(1) is the most appropriate model that captures all of the autocorrelation. What was left was the unexplained variation in the error terms. The x-axis in Figure 7 is the number of lags and the y-axis is the white noise probabilities. All lags, except the first one, exhibit white noise.

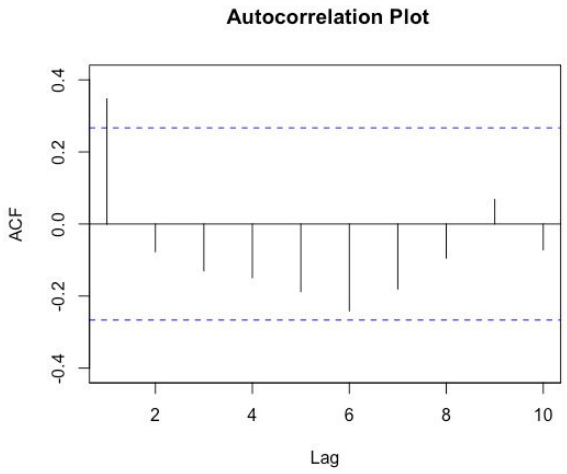

Figure 3. ACF Before MA and AR Terms

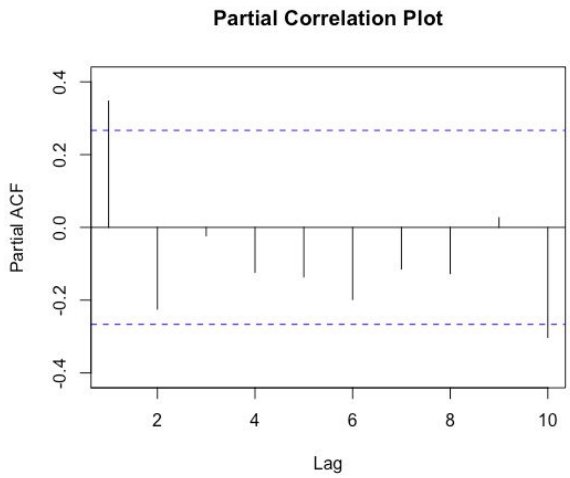

Figure 4. PACF Before MA and AR Terms

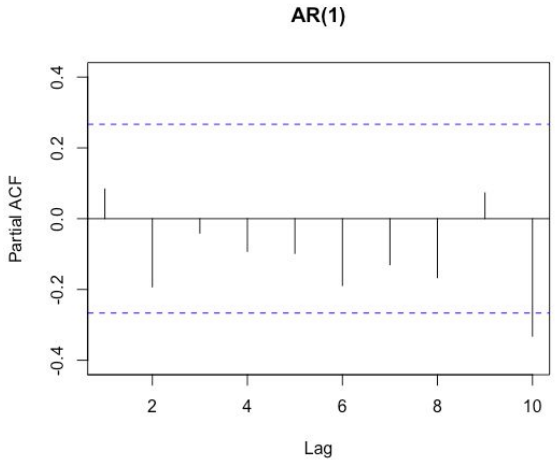

Figure 5. AR (1) PACF Plot of Residuals

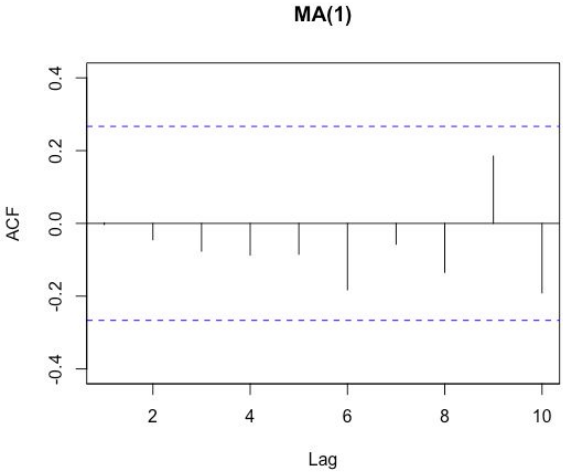

Figure 6. MA (1) ACF Plot of Residuals

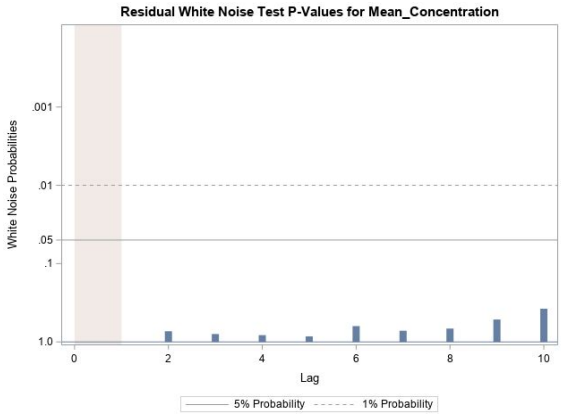

Figure 7. MA(1) Exhibits White Noise

Conclusion

By utilizing an ADF test, Orange Team 2 identified the data as being stationary around a trend line. After adding a moving average term to the model, the model was left with only white noise. The findings from the white noise table appear to be conclusive, so our future plans involve making sure the model continues to accurately address the variability in $PM_{2.5}$ values. We recommend gathering the data from the first half of 2019 and using it as a test dataset.

APPENDIX

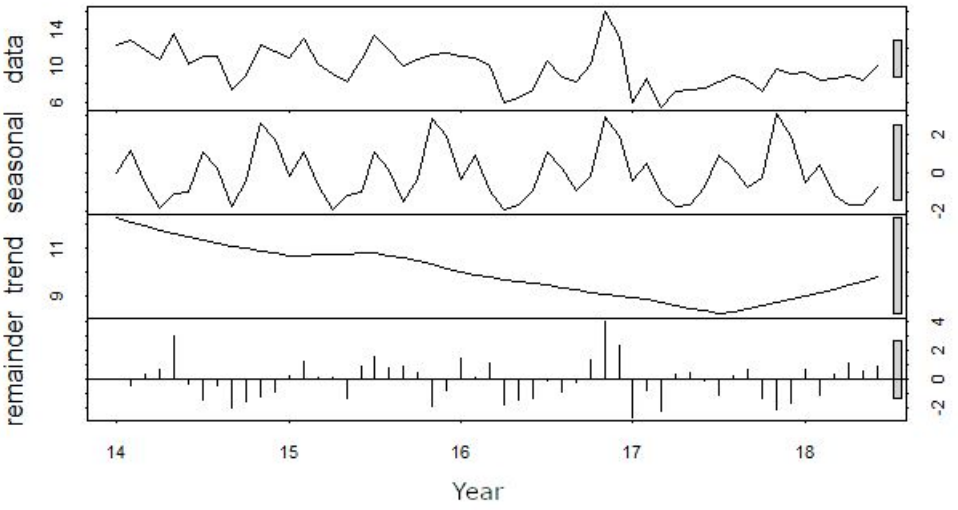

Figure A. STL Decomposition Plot