EPA Analysis Results

Blue Team 5:
Katlyn Limer
Iqra Munawar
Sam Scarpino
Ben Sullivan
Michael Zabawa

September 25, 2019

Executive Summary

The Environmental Protection Agency (EPA) has requested proposals regarding measures of Particulate Matter 2.5 (PM 2.5) at the Millbrook School station, in Wake County, North Carolina. Understanding and predicting levels of PM 2.5 is important for assessing the health and safety of the public that reside in the area. Previous analysis showed that the provided data lack significant seasonal and trend components. Exponential smoothing models can be used to predict future levels of PM 2.5, however more complex models could capture the variance and signal shape of the data. During the current analysis, the analyst team found the data are stationary and suggests modeling using AR(1), MA(1), and mixed ARMA models.

Recommendations

Based on the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs, the analyst team recommends using the current, unaltered, monthly data to model the signal using autoregressive (AR) and moving average (MA) terms. The graphs suggest possible candidate models include, but are not limited to, AR(1), MA(1), and mixed ARMA models.

Results

The team determined that the data are stationary based on a single-mean Augmented Dickey-Fuller (ADF) test. All 3 lags that were tested produced significant p-values, indicating stationarity along a single mean. Moving forward potential candidate models could include AR(1), MA(1), or ARMA(1,1).

Analysis

Data

The data contained 1,437 daily observations of PM 2.5 at the Millbrook School station in Wake County, North Carolina. The data began in January 2014 and ended in December 2018. These daily measures were averaged into monthly values resulting in 60 observations. Following this aggregation, the data was split into training and testing sets. The training data consisted of the first 54 observations and the testing data contained the final six months of the series.

Methodology

Analysis included:

- Augmented Dickey-Fuller (ADF) Test to determine if the data are a random walk
- White Noise Analysis to ensure there is further variance to be explained
- Autocorrelation Function & Partial Autocorrelation Function examination to inform further modeling

Signal Evaluation

Using the aggregated data, the analysts performed an ADF test to determine if the raw series was stationary or a random walk. The results from the ADF test are listed in Table 1. The unconditional mean of the series is 9.82. By inspection, the team concluded the absence of a trend. Thus, the series has a single mean. The p-values listed in Table 1 confirm that the signal is stationary about the single mean.

Table 1 Results from Augmented Dickey Fuller Test

Type	*Lags*	*Rho*	*Pr < Rho*	*Tau*	*Pr < Tau*
Single Mean	0	-30.8063	0.0005	-4.46	0.0007
	1	-36.0688	0.0005	-4.15	0.0017
	2	-30.8046	0.0005	-3.38	0.0158

Table 2 shows the p-values associated with the Ljung-Box test for the first 20 lags. All of the lags are significant at an alpha level of 0.05. Using these results, the team concluded that the data are not white noise.

Table 2 Ljung-box test

Lag	*p-value*	*Lag*	*p-value*
1	0.0002	11	0.0120
2	0.0005	12	0.0024
3	0.0013	13	0.0016
4	0.0030	14	0.0028
5	0.0068	15	0.0043
6	0.0136	16	0.0063
7	0.0246	17	0.0096
8	0.0355	18	0.0127
9	0.0206	19	0.0170
10	0.0257	20	0.0243

Figure 1 shows the ACF and PACF graphs of the data. The graphs indicate that an AR(1) term may be appropriate, as indicated by the single spike at the first lag in the PACF and that the ACF tails off. The graphs also suggest other potential candidate models could be MA(1) or ARMA(1,1).

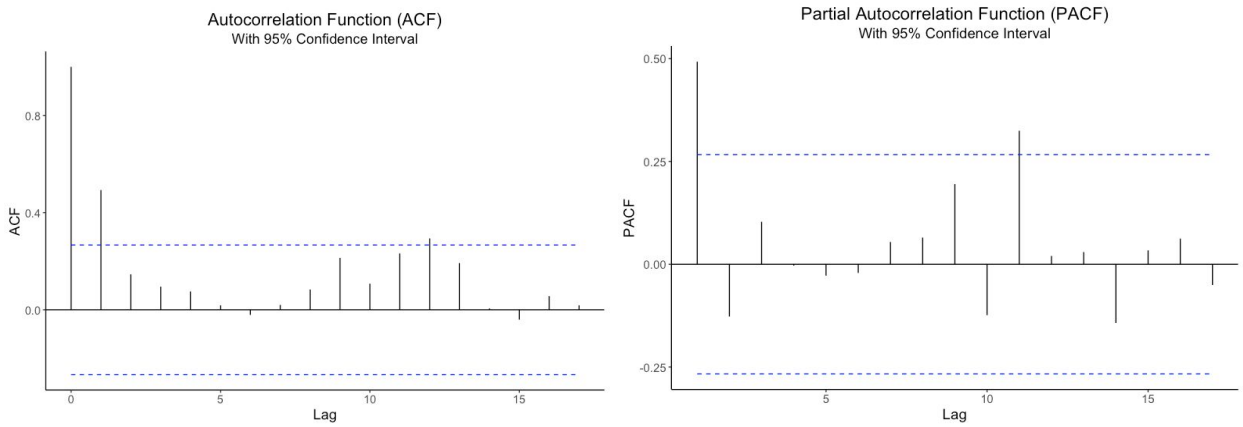

Figure 1 Autocorrelation Function and Partial Autocorrelation Function for Monthly Mean PM2.5

Conclusion

Previous analysis showed that the data lack significant seasonal and trend components within PM 2.5 levels in Wake County, North Carolina, from 2014-2018. The current analysis found the data are stationary and can move forward to further modeling. The analytics team recommends using the current, unaltered, monthly data to model the signal using autoregressive (AR) and moving average (MA) terms. Possible candidate models include, but are not limited to, AR(1), MA(1), and mixed ARMA models.