Wake County Air Quality Analysis

CATHY TRAN
SAVANNAH HAMPTON
ZACH LEWIS
DAVID ANDEXLER
KYRSTEN RUDOCK

October 13, 2019

Table of Contents

Wake County Air Quality Analysis

Executive Summary

Team Orange 9 partnered with the Environmental Protection Agency (EPA) to investigate the Air Quality at the Millbrook School station located in Wake County, North Carolina. The partnership aimed to predict the average monthly Particulate Matter 2.5 ($PM_{2.5}$) levels to inform oversight and regulatory initiatives. Based on previous models, the EPA believed that seasonality played an important role in estimating $PM_{2.5}$ and was interested in using a seasonal ARIMA model.

After exploratory analysis and decomposition, Team Orange 9 generated a series of seasonal ARIMA models. The final model was a model with design variables to account for seasonality, a quadratic function to account for trend, one autoregressive term (AR), and one seasonal moving average term (MA). This model achieved white noise and had a MAPE measurement of 17 percent.

Moving forward, Team Orange 9 recommends that the EPA collect air quality data for 2019 to help generate a more accurate model for the future. Using the final model in conjunction with additional data, the EPA will be able to more accurately predict the average monthly $PM_{2.5}$ for the Millbrook School station, potentially influencing regulatory and environmental initiatives in the area.

Results And Recommendations

Team Orange 9 generated nine seasonal and trending ARIMA models. The top three models were selected because they best modeled the correlation structure after achieving white noise. The final model consisted of an ARIMA(1, 0, 0) $(0, 0, 1)_{12}$ model with design variables to account for seasonality and a quadratic function to account for trend. This model was selected because it was the most parsimonious and it had the best relative predictive capability with a MAPE of 17 percent.

Statistics for all tested models are provided in the Appendix. The final three models and their associated accuracy statistics are listed in Table A1. All models that were built on the training data are listed in Table A2. Figure 1 below shows the forecasted values produced from the final model compared to the observed values for the same time period.

The EPA should deploy the seasonal ARIMA model to inform regulatory action surrounding air quality near the Millbrook School station. Further opportunities for additional analysis involve deploying the final ARIMA model in conjunction with additional data collection to further refine the results.

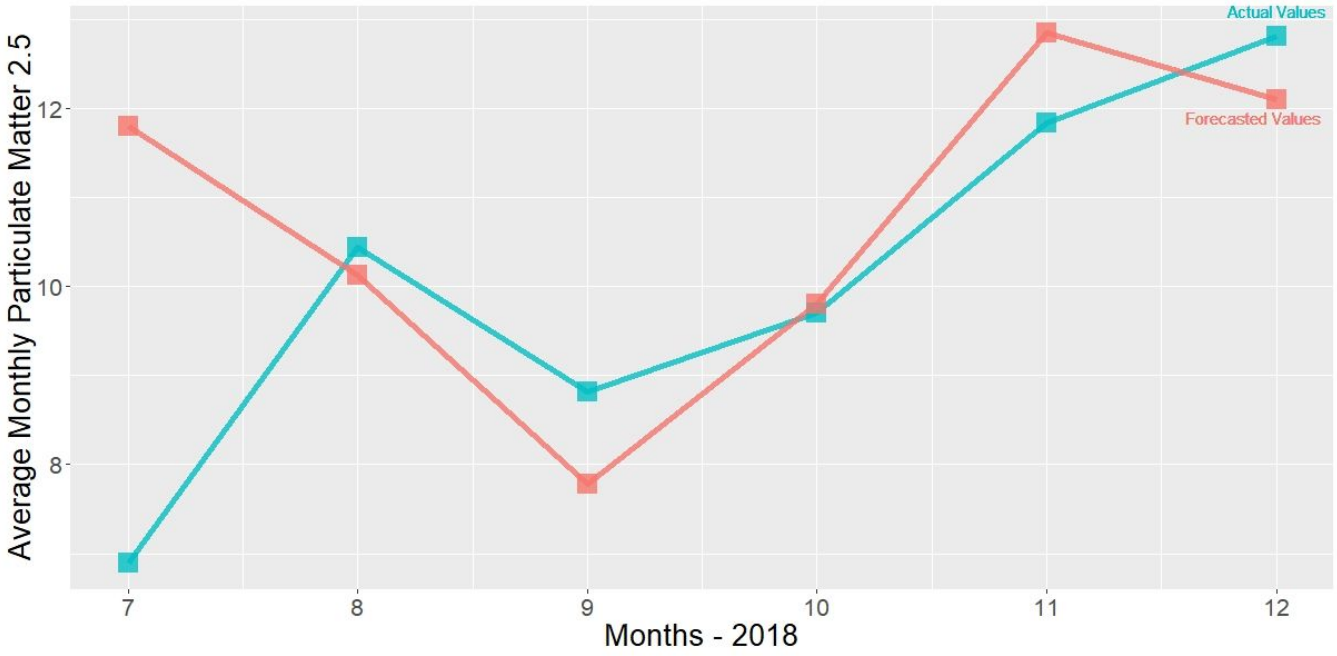

Figure 1. Forecasted values obtained from the final model compared to actual values observed

Methodology And Analysis

Data Used

The given dataset contains information about the air quality between 2014 to 2018 at the Millbrook School station at Wake County, NC. Specifically, there are 18 variables and 1,473 observations. This data describes the daily $PM_{2.5}$ measurements from January 1, 2014 to December 31, 2018.

Due to the 353 missing dates and values within the original time series, Team Orange 9 aggregated the data by monthly average and segmented off the last 6 observations to create the validation dataset for model testing and selection. The training dataset for modeling and prediction contains 54 observations from January 2014 to June 2018.

Decomposition Process

Team Orange 9 selected STL decomposition to decompose this time series to analyze the Trend/Cycle and Seasonality components, demonstrated in Figure 2.

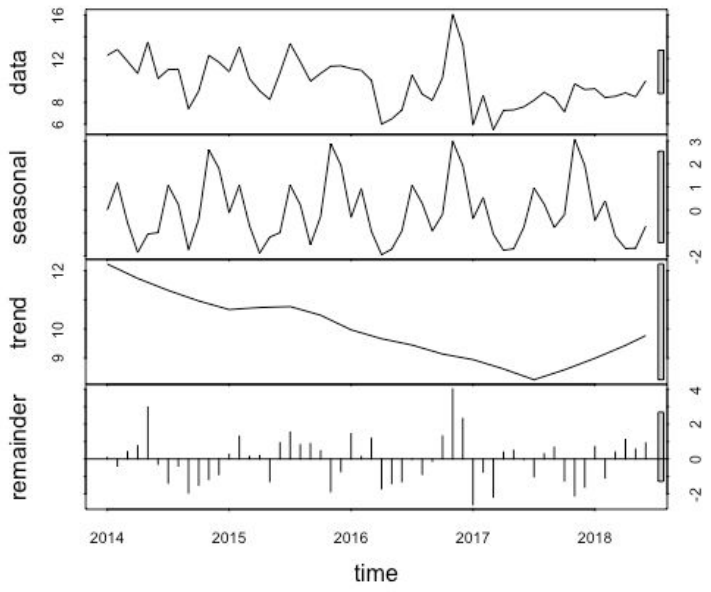

Figure 2. Raw Data and the Three Components of the Training Data Set Using STL Decomposition

The decomposition plot demonstrates the existence of a seasonal pattern and a general trend. The trend plot presents a downward trend from 2014 to mid-2017 and an upward trend since mid-2017, possibly indicating the presence of a non-linear trend.

Functional Seasonality & Trend

Before moving forward with modeling, the functional form of the seasonality and trend components had to be incorporated into the model. Team Orange 9 confirmed seasonality using a Seasonal Augmented Dickey-Fuller Test. Design variables accounted for the functional form of seasonality on the training data. After modeling seasonality, the team observed a clear trend pattern in the residuals. The team subsequently tested for stationarity of the time series using an Augmented Dickey-Fuller Test. This test indicated the presence of a deterministic trend. The team fit both a quadratic trend and a linear trend to assess which one optimally accounted for the trend observed in the residuals.

After further analysis, the team determined that the quadratic regression better removed the trend in the residuals compared to the linear regression. The team did not observe a trend after fitting the quadratic regression, whereas an observable trend pattern remained after fitting the linear regression. Specifically, using quadratic regression, the residuals centered around a zero mean.

Team Orange 9 also utilized the Fourier method with four sine and four cosine functions to capture the seasonal form and linear regression to capture the trend component. This method had higher AIC and BIC than the models with design variables and quadratic regression and was eliminated from further consideration.

Model Generation

Based on the residual ACF and PACF plots, there were significant spikes at lag 1 and lag 12. The Ljung-Box test indicated that there was a correlation structure to be modeled. Nine different models were built to account for the remaining dependency structure and subsequently evaluated based on their goodness-of-fit measures after applying seasonal AR and MA terms. After fitting the necessary correlation components, the Ljung-Box Test confirmed that significant autocorrelation no longer existed in the model and that the time series achieved white noise.

Model Selection

Out of the nine models that were built, three models were chosen to assess their predictive capacity on the validation data based on three reasons: all lags were within the confidence limits in the ACF and PACF plots, the residual plots appeared normally distributed, and AIC/BIC were minimized. A final model was selected using MAPE as the accuracy statistic. The model with one AR term and one seasonal MA term achieved the lowest MAPE of 17% compared to the other two models that were tested on the validation data. These accuracy results are listed in the Appendix.

Conclusion

Team Orange 9 provided the EPA with a seasonal ARIMA model that utilized design variables, a quadratic trend, one autoregressive term, and one seasonal moving average term. This model outperformed the top three models with a MAPE of 17 percent. By deploying this model and collecting more data, the EPA can more accurately predict future air quality levels at the Millbrook School station, informing regulatory initiatives regarding air quality in the area.

APPENDIX

Table A1. Accuracy Statistics for Top Three Seasonal ARIMA Model Sorted By Ascending MAPE

Seasonality Functional Modeling Method	Trend	Model	Accuracy	
			MAE	MAPE
Design Variables	Quadratic	ARIMA(1,0,0)(0,0,1)$_{12}$	1.35	0.17
Design Variables	Quadratic	ARIMA(0,0,1)(1,0,0)$_{12}$	1.45	0.18
Design Variables	Quadratic	ARIMA(1,0,1)(1,0,1)$_{12}$	1.54	0.19

Table A2. Models Built Using Training Data Sorted by Ascending AIC

Seasonality Functional Modeling Method	Trend	Model	Goodness Fit of Model on Training Dataset		
			AIC	AICc	BIC
Design Variables	Quadratic	ARIMA(1,0,0)(0,0,1)$_{12}$	193.63	210.63	227.44
Design Variables	Quadratic	ARIMA(1,0,1)(1,0,1)$_{12}$	194.30	216.65	232.09
Design Variables	Quadratic	ARIMA(0,0,1)(1,0,0)$_{12}$	194.32	211.32	228.13
Design Variables	Quadratic	ARIMA(0,0,0)(1,0,1)$_{12}$	202.61	219.61	236.42
Design Variables	Quadratic	ARIMA(1,0,0)(1,0,0)$_{12}$	208.28	225.28	242.09
Design Variables	Quadratic	ARIMA(0,0,1)(0,0,1)$_{12}$	208.49	225.49	242.30
Design Variables	Linear	ARIMA(1,0,0)(1,0,0)$_{12}$	209.57	224.27	241.40
Design Variables	Quadratic	ARIMA(1,0,1)(0,0,0)$_{12}$	219.82	236.82	253.63
Fourier (k=4)	Linear	ARIMA(1,0,2)(1,0,0)$_{12}$	238.86	243.97	258.75