Wake County Air Quality Analysis

CATHY TRAN
GRANT CLARK
EVAN LI
PRICE BURNETT
SUFYAN SHAHIN

September 16, 2019

Table of Contents

Wake County Air Quality Analysis

Executive Summary

With 79% accuracy based on validation data, Team Orange 2 identified a Holt-Winters Multiplicative Exponential Smoothing Model (ESM) as the best method to model and forecast $PM_{2.5}$ levels in Wake County, North Carolina. The state reports acceptable levels of these particulates within the county, but the model shows some potential areas for concern.

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. Air quality data collected between 2014 and 2018 at the Millbrook School station show a gradual increase in these tiny particulates since 2017. The benchmark set by the EPA is based on a 3-year average, and there is time for the county to address the issue in order to determine what, if anything, needs to be done.

The North Carolina Department of Environmental Quality (NCDEQ) reports Wake County's 3-year average from 2016 to 2018 to be 7.7 $\mu g/m^3$. This may mislead lawmakers and other stakeholders into a false sense of security around this issue. The average $PM_{2.5}$ for Wake County in 2018 was 9.5 $\mu g/m^3$, and the report of the 3-year rolling average is likely to make the increase in levels seem more gradual than they actually are.

Results

Holt-Winters Multiplicative ESM was used to forecast the average monthly $PM_{2.5}$ from July 01, 2018 to December 31, 2018. The MAPE was 0.21. In other words, the forecast model maintained 79% accuracy when compared against the validation data set. Figure 1 plots the predicted values (red line) versus the actual values (black line) of $PM_{2.5}$ in $\mu g/m^3$ on the validation data. The x-axis is the year and the y-axis is the mean $PM_{2.5}$ concentration in $\mu g/m^3$. From August to November, the forecast appears to mirror the actual data. It is only in the first and last month that we see a difference, which accounts for part of the 21% loss in accuracy.

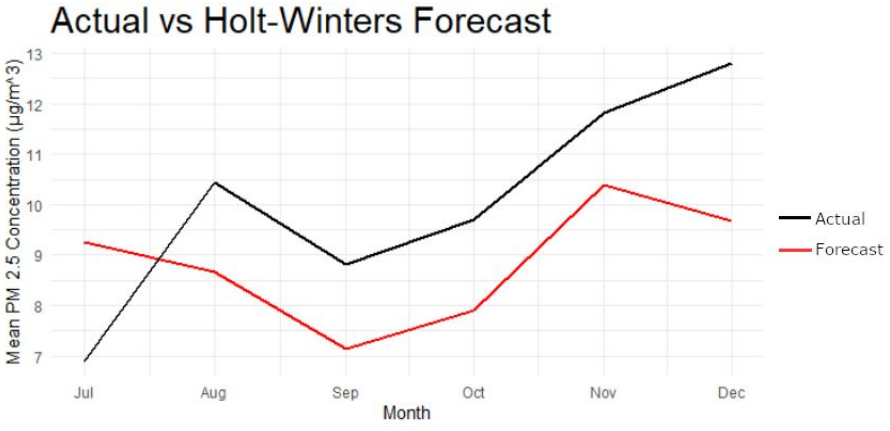

Figure 1. Holt-Winters Multiplicative ESM: Predicted $PM_{2.5}$ Values vs Actual $PM_{2.5}$ Values from July 01, 2018 to December 31, 2018

The Holt-Winters Multiplicative ESM model was selected based on our analysis of the raw data. Figure 2 depicts the mean $PM_{2.5}$ concentration from 2014 to 2018. The x-axis represents the year and the y-axis is the average $PM_{2.5}$ concentration from 2014 to 2018 in $\mu g/m^3$. Across those 4 years, the average $PM_{2.5}$ was 9.85 $\mu g/m^3$. The maximum average $PM_{2.5}$ was 16.08 $\mu g/m^3$ and the minimum average $PM_{2.5}$ was 5.50 $\mu g/m^3$. The most significant improvement in air quality was at the end of 2016. The mean $PM_{2.5}$ prior to 2017 was 10.64 $\mu g/m^3$ whereas the mean $PM_{2.5}$ after 2017 was 8.66 $\mu g/m^3$. This suggests that the air quality has gotten better after 2017 but we still need to be wary because the mean $PM_{2.5}$ after 2017 was increasing gradually.

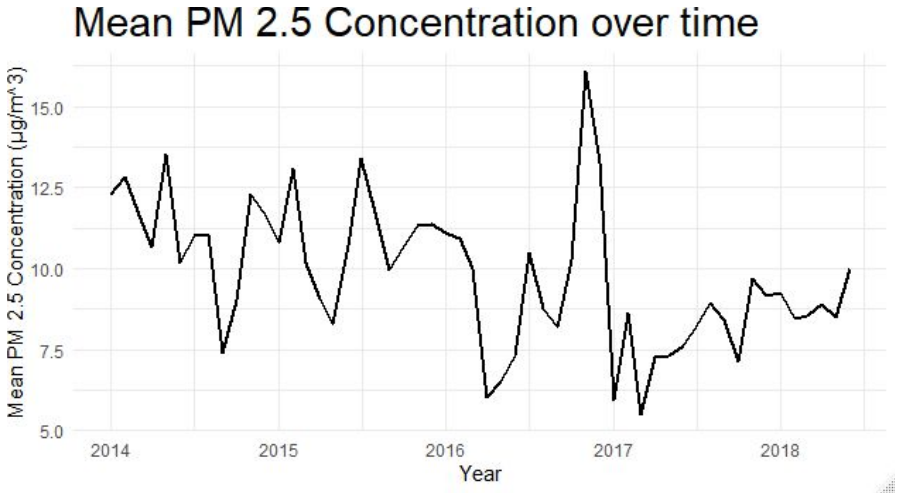

Figure 2. Mean $PM_{2.5}$ Concentration From 2014 to 2018 on the Training Data Set

Recommendations

We recommend evaluating the irregularities within the data, and completing research to understand their cause. These irregularities are illustrated within Figure 2 where there is high variability in the $PM_{2.5}$ values between 2014 and the end of 2016. However, in 2017 and the early part of 2018, the mean $PM_{2.5}$ concentration is much more stable from month to month, suggesting unequal variance across the data. This suggests that a multiplicative model would be useful to capture changing variance over time, but does not help to explain the reason for this change in pattern.

In addition, we recommend considering alternate techniques for data aggregation. While the monthly means are informative for evaluating $PM_{2.5}$, other metrics such as monthly minimums, maximums, or medians may also be useful. For example, the NCDEQ measures air quality by averaging the 98th percentile value of daily $PM_{2.5}$ over three year periods. It may be useful to explore models that evaluate the highest levels of $PM_{2.5}$ as the goal is to keep these levels down.

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. This model 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 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 for modeling and prediction contains 54 observations from January 2014 to June 2018.

Decomposition Process

During this process, the STL decomposition technique was chosen to decompose this time series in order to analyze the Trend/Cycle and Seasonality components for model selection (Figure 3). Team Orange 2 chose STL over Classical decomposition method since STL uses local regressions to estimate trend and seasonality that allows changes within the decomposition components. In this way, a more accurate description and understanding of the data could be gained from the decomposition plot and further assist Team Orange 2 selecting models.

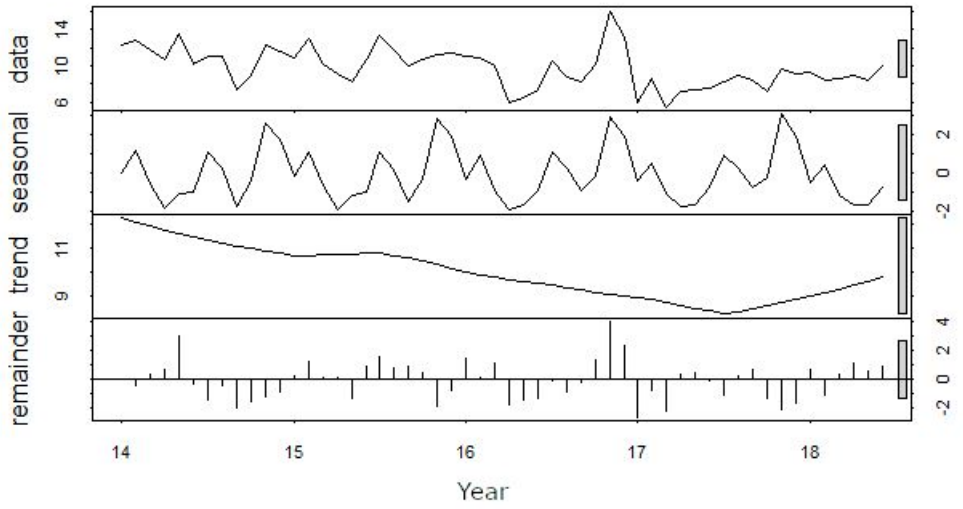

Figure 3. Actual Data and the Three Components of This Time Series From 2014 To 2018 of the Training Data Set Using STL Decomposition (From Top To Bottom: Actual Data; Seasonality; Trend/Cycle; Residuals)

This decomposition plot informs us that there exists a seasonality pattern and a general trend which facilitates the model selection process. In this decomposition plot, the horizontal axis represents the timeline from Jan. 2014 to Jun. 2018. The graph consists of 4 components: the top one is the plot of actual data over time; the second is the decomposed trend piece over time; the third describes the seasonality pattern of this timeseries (12 months as a season); and the bottom plot is the residuals left after removing the trend and seasonality.

The Seasonality plot denotes the existence of a yearly seasonal pattern, in which $PM_{2.5}$ values decreased and the air quality was better during Spring and Summer, yet $PM_{2.5}$ values increased back to a higher level during Winter. The Trend plot presents a noticeable trend which will be described in further detail in the next section.

Trend vs. Seasonality Adjusted Component

The Trend and Seasonality and their relations to the actual data were further investigated and compared. Figure 4 overlays the trend component (red line) on the original data of $PM_{2.5}$ values (black line) in $\mu g/m^3$ from the training dataset.

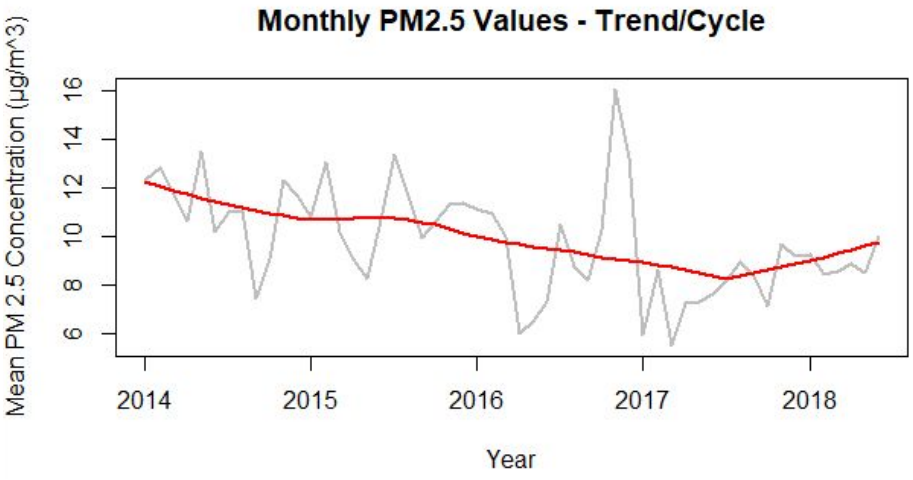

Figure 4. Trend Component Overlaid on the Original Data of the Monthly $PM_{2.5}$ Values on the Training Data Set

The trend component did successfully capture the gradual decreasing and increasing trend of the actual data. There is a steady decrease in monthly $PM_{2.5}$ values between 2014 and the mid-2017, where the $PM_{2.5}$ values dropped from 12 $\mu g/m^3$ to about 8 $\mu g/m^3$. After that, there is an increasing trend from mid-2017 to mid-2018 where the monthly $PM_{2.5}$ values returned to around 10 $\mu g/m^3$. However, the $PM_{2.5}$ values of solely the trend component did not effectively capture some extreme variations in the data, specifically the big spike towards the end of 2017.

In addition to the Trend/Cycle component, Figure 5 overlays the trend component as well as residuals on the original data. The black line is the original data of $PM_{2.5}$ values and the red line is the seasonally adjusted data. Along with the error terms, the seasonally adjusted $PM_{2.5}$ values manage to describe not only the general trend but also the drastic changes within the data that the trend component left out previously in Figure 4.

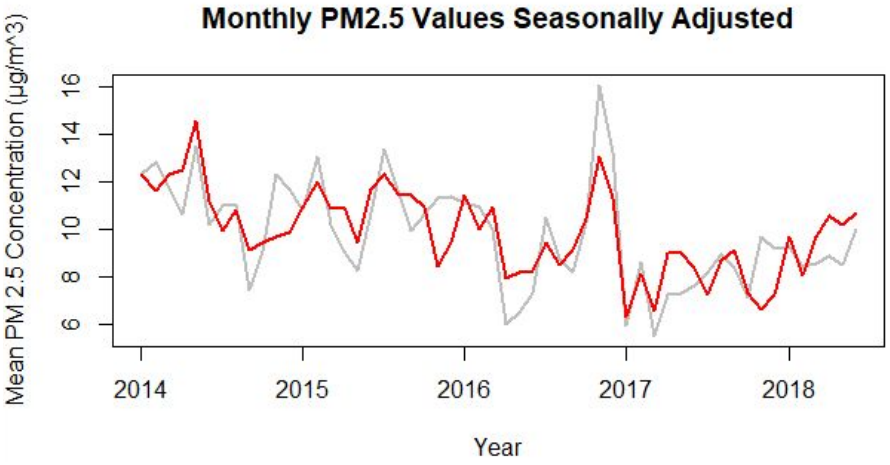

Figure 5. Seasonally Adjusted Monthly $PM_{2.5}$ Values Overlaid on the Original Data in $\mu g/m^3$ on the Training Data Set

This finding informs us that the Trend component is crucial in describing the time series data. Based on this, the Seasonal component may not be an impactful factor in forecasting the monthly average $PM_{2.5}$ values; however, due to the seasonal nature of the data, the team decided to utilize a model that accounts for seasonality.

Model Selection

Considering the decomposition analysis, accuracy statistics, and the general context of topic, Team Orange 2 chose the Holt-Winters Multiplicative ESM method to forecast $PM_{2.5}$ levels in Wake County, North Carolina**.** Although the Single Exponential Smoothing Model has the smallest MAPE and MAE values among all five models (as shown in the Table of accuracy statistics in the Appendix), we identified the Holt-Winters Multiplicative ESM as the best model since it consists of Level, Trend, and Seasonality component.

The Decomposition analysis of the Trend and Seasonally Adjusted plots shows a necessity to include Trend as an important factor in describing the data and forecasting the monthly average$PM_{2.5}$ values.

Although seasonality does not appear to be an impactful factor based on the seasonally adjusted values, it is still present in this time series as noted in Figure 3. In a broader context of air quality, the seasonality is present year to year. Therefore, Team Orange 2 believes it necessary to include seasonality in the model.

The multiplicative Holt-Winters ESM was chosen over additive Holt-Winters ESM based on the plot of actual data (Figure 2) as well as their performances on the test dataset in accuracy statistics. As mentioned in previous sections, the high variability of the $PM_{2.5}$ values between 2014 and the end of 2016 was gradually minimized and leveled off in 2017 and the early part of 2018. This diminishing variations informed us to choose multiplicative ESM model in order to capture the changing variations.

Conclusion

During this analysis Team Orange 2 identified, with 79% accuracy, a Holt-Winters Multiplicative ESM as the best method to model and forecast $PM_{2.5}$ levels in Wake County, NC. Moving forward, the team would like to check for random walks and stationarity in the data set. After these have been accounted for, the team would explore other modeling techniques such as ARIMA, if applicable, to describe any autocorrelation within the data.

APPENDIX

Sources:

"PM2.5 Average Values." *North Carolina Department of Environmental Quality*. https://deq.nc.gov/about/divisions/air-quality/air-quality-monitoring/historical-data-summaries/design-value-0

Table 1. Accuracy Statistics for Each Model Used in Testing

Model	MAE	MAPE
Single Exponential Smoothing Model	1.888392	0.1810493
Linear Exponential Smoothing Model	2.200177	0.2074771
Linear Damping Exponential Smoothing Model	2.008073	0.1906637
Holt-Winters ESM - **Additive**	2.247700	0.2285497
Holt-Winters ESM - **Multiplicative**	2.029253	0.2090463