Modeling Particulate Matter 2.5 as a Time Series

Author*: Ethan Biegeleisen
Advisor†: Dr. Angel R. Pineda
Department of Mathematics, Manhattan College, Bronx, NY

May 4, 2020

Abstract

This project models Manhattan's daily recorded levels of particulate matter 2.5 from 2013 to 2018 as a time series using RStudio. The time series of observed data is decomposed into trend data, seasonal data, and the remaining error when the trend data and seasonal data are removed from the observed data. In addition, the Holt-Winters method is utilized to apply a best fit to the observed data. The Holt-Winters fit is then used to predict Manhattan's 2019 daily particulate matter 2.5 levels, which are compared with Manhattan's observed 2019 data for particulate matter 2.5 levels.

Introduction

Particulate matter 2.5 (PM 2.5) is an air pollutant that can impact respiratory function and worsen medical conditions like heart disease and asthma [3]. The pollutant is typically emitted from vehicle exhausts and burned fuel [3]. PM 2.5 can also cause short-term irritation in locations like the eyes and throat, and it can lead to coughing, sneezing, or shortness of breath [3]. For this project, Manhattan's daily recorded levels of PM 2.5 are modeled as a time series that is examined through decomposition and the Holt-Winters method. The Holt-Winters method is also used to predict PM 2.5 levels at a later time period. Modeling PM 2.5 can allow for estimates of near-future PM 2.5 levels which can help warn people in advance if an upcoming period is expected to have especially high levels.

Background

The observed data is from the Environmental Protection Agency's (EPA) Air Quality Index (AQI) Data Values Report [4]. The EPA's website has PM 2.5 levels for parts of the United States available from 1999 onward, but Manhattan only has daily PM 2.5 data available from 2013 and later [4]. The data sets are searched for by year and location, with location being searchable by general city area or by county [4]. The annual data sets can be saved as PDF or CSV files, with the later being chosen for this project.

*ebiegeleisen01@manhattan.edu
†angel.pineda@manhattan.edu

Date	PM2.5 AQI	AQI Category	Site Name	Site ID	Source
1/1/2018	51	Moderate	Intermediate School 143	36-061-01	AQS
1/2/2018	56	Moderate	Intermediate School 143	36-061-01	AQS
1/3/2018	61	Moderate	Intermediate School 143	36-061-01	AQS
1/4/2018	48	Good	PS 19	36-061-01	AQS
1/5/2018	48	Good	Intermediate School 143	36-061-01	AQS
1/6/2018	46	Good	Intermediate School 143	36-061-01	AQS
1/7/2018	52	Moderate	Intermediate School 143	36-061-01	AQS
1/8/2018	67	Moderate	Intermediate School 143	36-061-01	AQS
1/9/2018	61	Moderate	Intermediate School 143	36-061-01	AQS
1/10/2018	55	Moderate	PS 19	36-061-01	AQS
1/11/2018	74	Moderate	IS 45	36-061-00	AQS
1/12/2018	41	Good	PS 19	36-061-01	AQS
1/13/2018	44	Good	Intermediate School 143	36-061-01	AQS
1/14/2018	43	Good	Intermediate School 143	36-061-01	AQS
1/15/2018	58	Moderate	PS 19	36-061-01	AQS

Figure 1: Daily PM 2.5 levels in Manhattan from January 1, 2018 to January 15, 2018 [4].

Figure 1 shows the first 15 days of PM 2.5 data from Manhattan in 2018 opened in Microsoft Excel. The major categories are date observed, PM 2.5 AQI, AQI category, and the site where the data was observed. The AQI category is determined by the daily PM 2.5 level [1]. The AQI category is Good if the level is 0-50, Moderate if the level is 51-100, Unhealthy for Sensitive Groups if the level is 101-150, and Unhealthy if the level is 151-200 [1]. The pollutant level can go above 200 into more severe AQI categories, but the observed data is not close to the levels required for these higher categories. It should be noted that even in a single county, PM 2.5 levels are recorded at different sites each day. This appears to be one limitation with the data set, but it is the primary reason for using county data instead of general city data in this project. Although county observations are still recorded at different sites each day, the sites should be geographically closer together than the sites recording observations for data in a general city area. Ideally, this should lead to a bit more consistency in observations across a singular county than observations across a city-wide location.

Methods

The Holt-Winters method is built on top of exponential smoothing, which is meant to predict x_{n+k} when observations are available from x_1 to x_n [2]. The exponential model is

$$x_t = \mu_t + \omega_t \tag{1}$$

where μ_t is the non-stationary mean at time t and ω_t are independent random deviations with mean 0 and standard deviation σ. a_t is the exponentially weighted moving average at time t such that

$$a_t = \alpha x_t + (1-\alpha)a_{t-1}, \qquad 0 < \alpha < 1 \tag{2}$$

where α is the smoothing parameter. The closer α is to 1, the closer a_t is to x_t [2].

The Holt-Winters method focuses on the level (the seasonally adjusted mean), the slope (the change in level from one time period to the next), and the seasonal effect [2]. The method's key equations over period p are

$$a_t = \alpha(x_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{3}$$

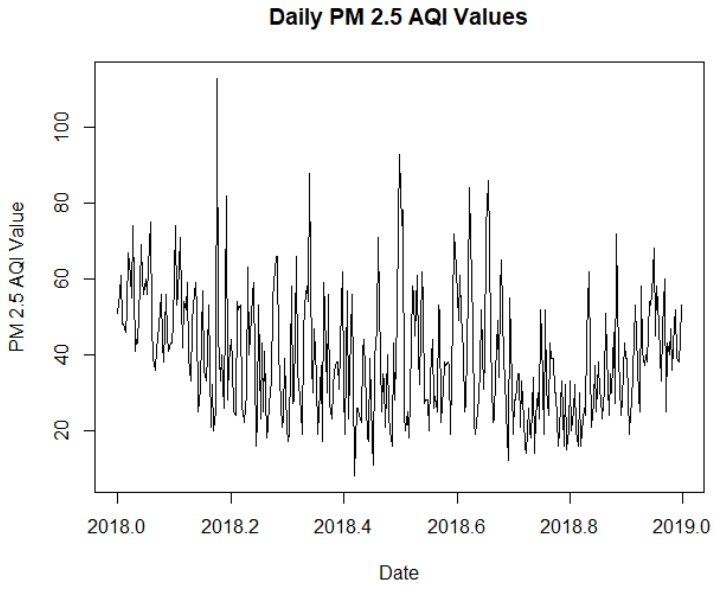

Figure 2: Daily PM 2.5 levels in Manhattan from January 1, 2018 to December 31, 2018.

$$b_t = \beta(a_t - a_{t-1}) = (1-\beta)b_{t-1} \tag{4}$$

$$s_t = \gamma(x_t - a_t) + (1-\gamma)s_{t-p} \tag{5}$$

where a_t, b_t, and s_t are the estimated level, slope, and seasonal effects at time t and α, β, and γ are the smoothing parameters. The forecasting model based on these estimates for x_{n+k} is

$$\hat{x}_{n+k|n} = a_n + kb_n + s_{n+k-p}, \qquad k \leq p \tag{6}$$

where $a_n + kb_n$ is the expected level at time $n+k$ and s_{n+k-p} is the exponentially weighted estimate of the seasonal effect at time $n = k - p$ [2].

Results

Manhattan's PM 2.5 data sets for 2013-2018 were saved as a single CSV file and the data sets for 2013-2019 were saved into a second CSV file. The first step in RStudio was importing the data from 2013-2018. Next, all of the entries in the PM 2.5 AQI level column were converted into a time series starting on January 2013 with a frequency of 365.25 due to the leap year in the data set. Figure 2 displays the 2018 data plotted as its own time series. After checking the time series and plotting it to make sure the data was converted properly, the decompose function was utilized to extract the trend, seasonal data, and error from the observed data. Both additive and multiplicative models of decomposition were attempted, but additive was decided on due to both models yielding similar results. Figure 3 displays the decomposition of the 2013-2018 time series.

The Holt-Winters method was utilized on the time series through the HoltWinters function in RStudio, which automatically determines the values of α, β, and γ. However, the function originally

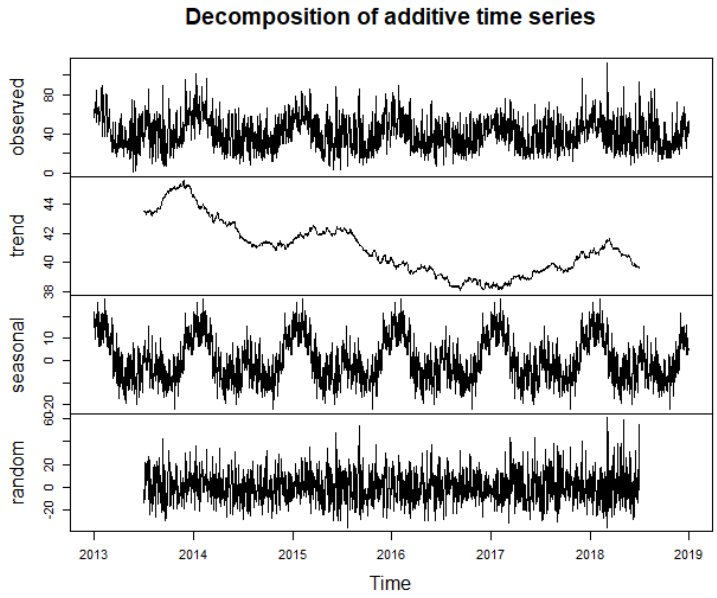

Figure 3: Decomposition of daily PM 2.5 levels in Manhattan from January 1, 2013 to December 31, 2018.

gave an error message saying it required two or more periods to work when the time series had a frequency of 365.25. This seems to imply that when the frequency is 365.25 days, a single period is considered four years long instead of one year long. If this speculation is correct, then eight full years of daily observations would be necessary for the function to run correctly with a time series that has a frequency of 365.25. Due to the data set containing less than 8 years of daily observations, the observed data from February 29, 2016 was excluded from the time series and the frequency was changed to 365. This allowed the HoltWinters function to properly run, and the loss of one observation in a time series with over 2000 entries does not appear to be a significant omission. The values for α, β, and γ that were determined by the HoltWinters function were 0.024, 0, and 0.

Figure 4 shows the fitted Holt-Winters values plotted alongside the original time series. The predict function in RStudio was used on the Holt-Winters values to provide an estimate of Manhattan's daily PM 2.5 levels in 2019. Figure 5 plots the elements of the Holt-Winters fit and Figure 6 compares the daily 2019 estimates with the actual 2019 observations.

Figure 7 plots the observed PM 2.5 levels from 2013-2019 with the noise removed by adding together the trend and seasonal variation from the decomposition of the 2013-2019 time series. Figure 7 also plots the predicted 2019 PM 2.5 levels based on the Holt-Winters fit for the 2013-2018 time series. In addition, Figure 8 plots the observed PM 2.5 levels from 2013-2018 along with a forecast for daily PM 2.5 levels in 2019-2022 that is based on the same Holt-Winters fit used for the 2019 prediction.

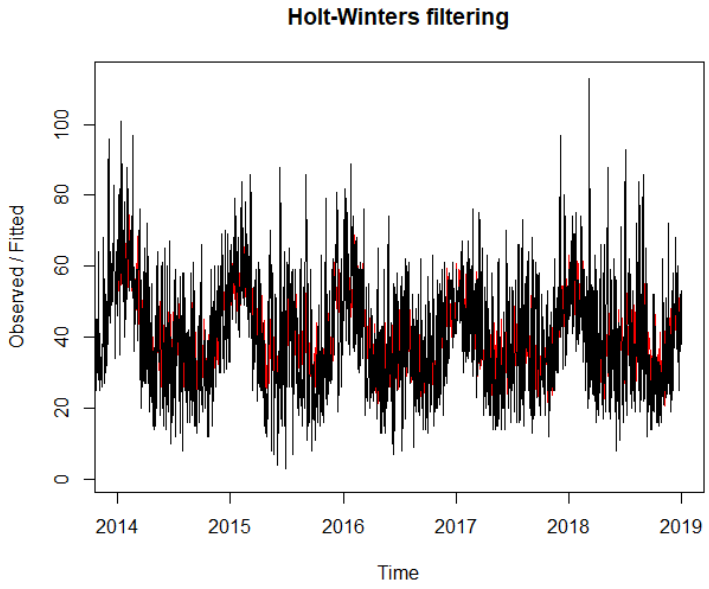

Figure 4: The black line consists of the observed daily PM 2.5 levels in Manhattan from January 1, 2013 to December 31, 2018. The red line consists of the Holt-Winters fit for the daily PM 2.5 levels.

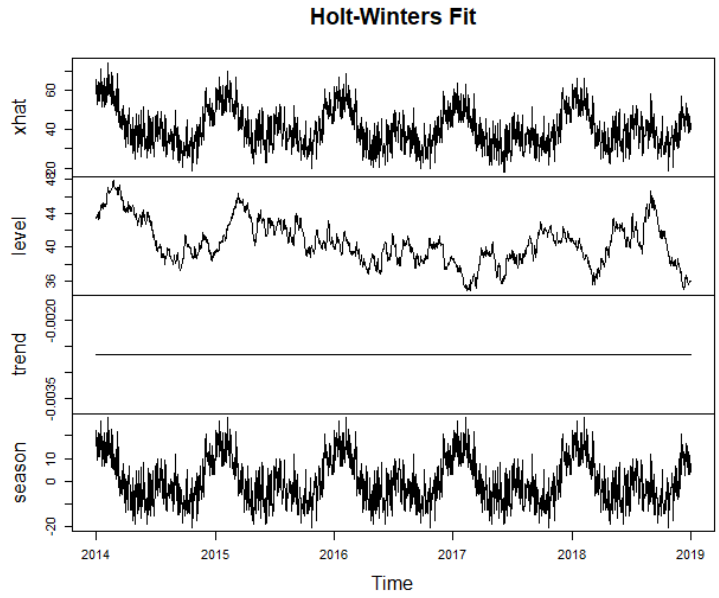

Figure 5: Plotting the components of the Holt-Winters fit. $\hat{x}$ is the Holt-Winters approximation of the observed data, the level is the seasonally adjusted mean, the trend/slope is the change in level from one time period to the next, and the season refers to the seasonal effect. [2]

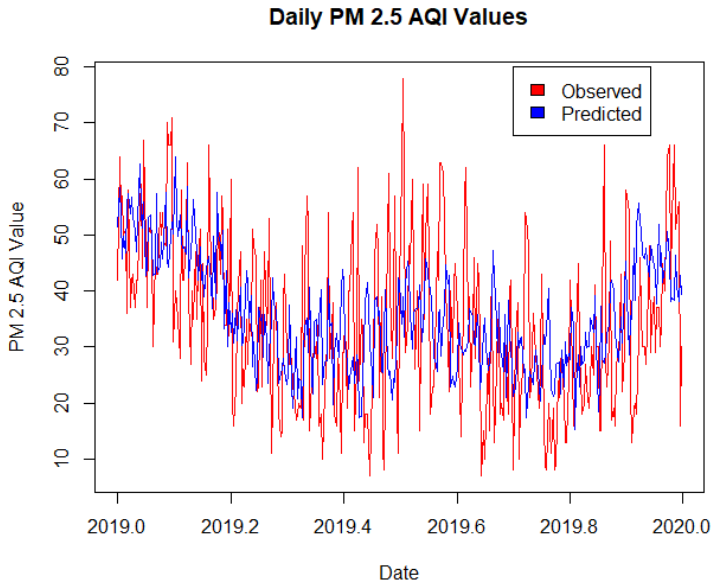

Figure 6: The red line consists of the observed daily PM 2.5 levels in Manhattan from January 1, 2019 to December 31, 2019. The blue line consists of the predicted daily PM 2.5 levels across the same period of time.

Discussion

With seasonal effects and noise removed from the observed 2013-2018 data, the general trend in PM 2.5 levels appears to be slightly decreasing over time. The Holt-Winters fit for the data appears to match the original observations fairly well, but the individual components of the Holt-Winters fit raise a few questions. The seasonally adjusted mean in Figure 5 looks normal, but the slope is a straight horizontal line. This seems to be connected with β getting assigned the value of 0 in the HoltWinters function. In addition, the seasonal effect in Figure 5 appears to be mostly identical to the Holt-Winters approximation of the observed data. Figure 5's results are curious when Figure 4 shows that the Holt-Winters approximation still appears to accurately match the observed data to a substantial degree. The seasonal effect in Figure 5 also appears to match the seasonal effect in Figure 3.

Figure 6 shows the general pattern in the 2019 estimates appears to match the 2019 observations quite well in terms of when the PM 2.5 levels increase, decrease, or remain constant. However, the estimated levels contain a smaller spread than the observed levels where the estimated levels rarely go as high or as low as the true observations. The reason for this disparity is that the Holt-Winters approximation and predictions based on it cannot account for noise. A more accurate way to test the effectiveness of the prediction model is to compare it with the combined trend and seasonal effect from the decomposition of the 2013-2019 data. Figure 7 shows that when the noise is removed from the observed data, the 2019 observations match the 2019 predictions quite well. However, adding the decomposed elements of the 2013-2019 time series excludes data from the first 6 months and the last 6 months so the 2019 predictions are only getting plotted against the the combined

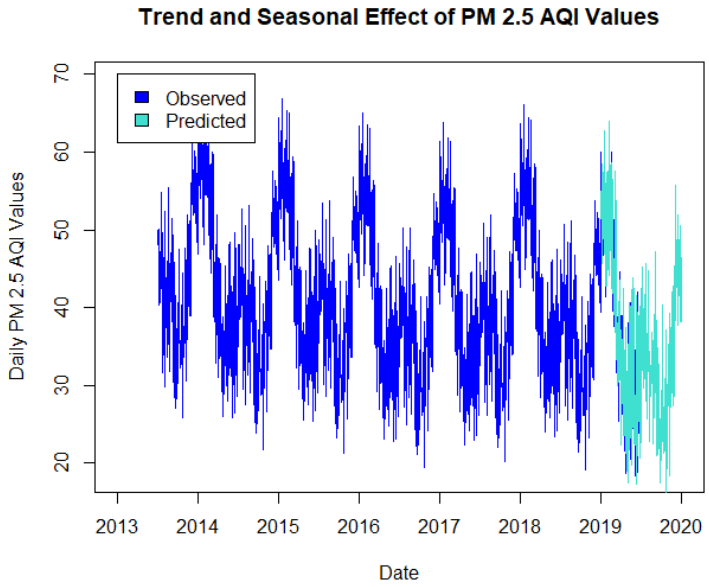

Figure 7: The blue line consists of the observed trend + the seasonal effect of the observed daily PM 2.5 levels in Manhattan from 2013-2019. The turquoise line consists of the predicted daily PM 2.5 levels throughout 2019.

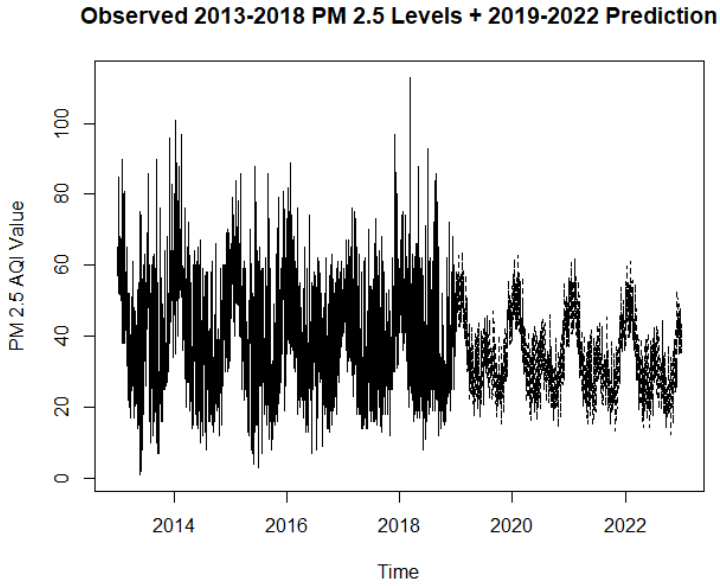

Figure 8: The line from 2013-2018 displays Manhattan's daily observed PM 2.5 levels, while the dotted line consists of predicted PM 2.5 levels for 2019-2022.

trend and seasonal effects from the first half of the 2019 observations. Even with this limitation, though, the forecast based on the Holt-Winters approximation appears highly accurate at predicting the combined trend and seasonal effect of PM 2.5 levels in the near future. Figure 8 shows a more long term prediction for PM 2.5 levels from 2019-2022 compared to observed PM 2.5 levels from 2013-2018, but the accuracy of these estimates cannot be determined at the time of this writing.

The current prediction model seems to decently capture the upcoming trend for the year following the original data set, but the model appears to have more trouble forecasting exact PM 2.5 levels on specific dates due to the forecasting model not being able to account for noise. Figure 3 shows that noise makes up a significant component of the original time series, so precise estimates on specific dates might be more difficult for this data set than it would be for a time series that contained significantly less noise in its decomposition. However, the prediction model might also be sufficient for determining if specific upcoming weeks or months could see a general increase or decrease in PM 2.5 levels. For example, Figure 6 shows that the forecast of a general decrease in PM 2.5 levels from the start of 2019 to roughly the middle of 2019 is accurate.

One potential way to build on modeling this data set in the future would be including another observed air pollutant in Manhattan across the same time period. Manhattan's PM 2.5 levels could also be compared with the PM 2.5 levels of another country during the same period.

Conclusions

Modeling Manhattan's PM 2.5 emission levels from 2013-2018 as a time series allowed it to be decomposed and fitted with the Holt-Winters method in RStudio. The decomposed time series displayed a decreasing trend along with a standard appearance for the seasonal trend and random noise. The Holt-Winters fit of the time series matched the original data set fairly well, and the fit was used to predict 2019 PM 2.5 levels. The pattern in the predicted pollution levels matched the pattern of Manhattan's observed 2019 PM 2.5 levels, but predictions on a daily level appear to be limited by the amount of noise in the time series for the original data set. However, the Holt-Winters method could be used to predict future patterns more reliably overall. In addition, daily predictions appeared to be more accurate when they were compared with a plot of the observed time series that excluded noise.

References

[1] AirNow (June 2019). Air Quality Index (AQI) Basics. Retrived from: `https://www.airnow.gov/index.cfm?action=aqibasics.aqi`

[2] Cowpertwait, P. S. P., Metcalfe, A. V. (2009), *Introductory Time Series with R*. Springer Science + Business Media.

[3] Department of Health (February 2018). Fine Particles (PM 2.5) Questions and Answers. Retrieved from: `https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm`

[4] Environmental Protection Agency (2020). Air Quality Index Daily Values Report. Retrieved from: `https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report`