JOURNAL OF
EMERGING INVESTIGATORS

Article

COVID-19 and air pollution in New York City

William Ning[1], Michael Ning[2]

[1] Hunter College High School, New York, New York
[2] Fordham University, New York, New York

SUMMARY

The COVID-19 pandemic may have improved air quality, as the restriction on travel and the slowdown of social and economic activities may have helped to reduce the emission of greenhouse gases. In this analysis, a seasonal Autoregressive Integrated Moving Average (ARIMA) model is proposed to forecast the level of $PM_{2.5}$ particulate pollutants in New York City, using differencing at a lag equal to trailing 12 months to remove seasonal effects. Comparing the actual data with the model's prediction during the period from January 2020 to January 2021, no statistically significant difference was found when taking long-term trends into consideration. Overall, COVID-19 did not show a convincing temporary positive impact on air quality in New York City.

INTRODUCTION

Air pollution refers to the presence of substances, or pollutants, in the air or atmosphere that have harmful effects on humans, other animals, the environment/climate, or materials (1). These pollutants can come from many sources, but the vast majority of air pollutants are emitted as byproducts during energy production and use. While air pollutants do exist from natural causes (notably, wildfires), high levels of air pollutants create serious health concerns. According to the United States Environmental Protection Agency (EPA), exposure to air pollutants may grant "an increased chance of getting cancer" as well as "damage to the immune system, as well as neurological, reproductive (e.g., reduced fertility), developmental, respiratory and other health problems (2)." As such, it is important to monitor and try to reduce the levels of these particles. The particles that pose the greatest risk to health are known as $PM_{2.5}$, which refers to a category of particulate matter that is 2.5 microns or smaller in size (3), because they can get deep into the lungs and respiratory system as well as into the bloodstream. $PM_{2.5}$ is one of the primary pollutants monitored in New York City with the longest record available to the public in the World Air Quality Index (WAQI) project database (4). The WAQI database also provides information on many other pollutants for New York City, such as carbon monoxide, nitrogen dioxide, ozone, and sulphur dioxide. However, only data on these pollutants from after 2016 is available, so $PM_{2.5}$ is the most suitable parameter for developing a statistical model over a longer time series (5).

The WAQI database lists the Air Quality Index (AQI) daily averages, with $PM_{2.5}$ data in New York City available from 2014 to 2021. AQI scores are calculated by averaging the $PM_{2.5}$ data from midnight to midnight and converting into a different 0–500 scale, with 0 being the best and 500 representing the worst quality air (6). Any score from 0-100 is acceptable and the air pollution poses little to no risk to human life, but scores above 100 denote unhealthy air that poses health risks. Overall, air quality in New York City has improved in the past few years due to actions taken to reduce these emissions (7) (**Figure 1**). The $PM_{2.5}$ levels in New York City have remained less than 50 since 2017. More progress, however, did not appear until 2020, when the world was impacted by the deadly COVID-19 pandemic.

In New York City, $PM_{2.5}$ levels typically peak around January when consumption of energy rises and then gradually recede until May. As the weather warms up, emissions rise again as vehicle ridership increases. The second peak levels usually occur in July, after which air quality steadily improves until Labor Day when a new cycle begins (**Figure 2**). These seasonal air pollution trends are associated with the so-called summer and winter seasons, which are also known as the driving and heating seasons. During the first half of 2020, the $PM_{2.5}$ levels in New York City experienced a significant drop (**Figure 1**). The observation was even more evident when the data series was plotted against the individual seasons in which the air quality was measured.

Is the large drop in $PM_{2.5}$ in the first six months of 2020 in New York City a result of social distancing mandates and lockdowns? A number of research papers confirm improvements in air pollution in other parts of the United States, China, and South Korea during the COVID-19 pandemic (8, 9, 10). However, investigations specific to air quality in New York have so far shown no statistically significant improvement in air pollution (11, 12). It is true that the COVID-19 lockdown reduced transportation activities, which resulted in less

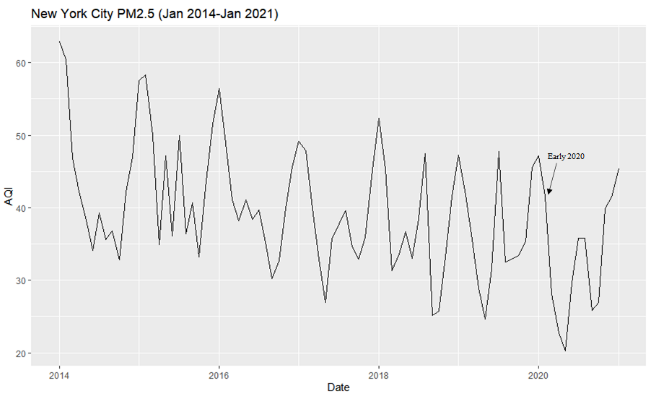

Figure 1: Declining Trend of $PM_{2.5}$ Pollution in NYC. Source: World Air Quality Index project.

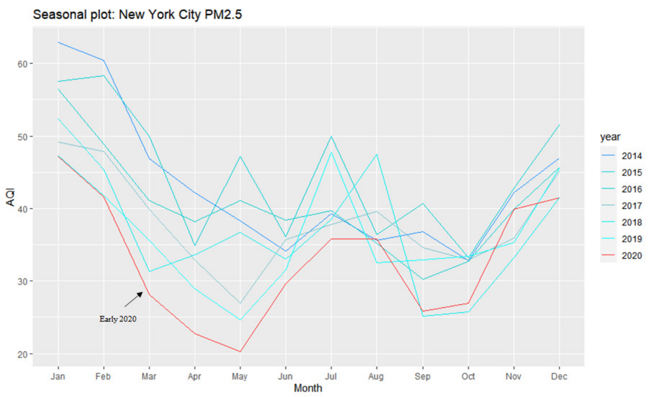

Figure 2: NYC Seasonal Variability of PM$_{2.5}$. Source: World Air Quality Index project.

energy consumption and lower oil demand from vehicles (13), but soon after the government eased the lockdown in May and June 2020, New York drivers resumed their normal driving tendencies (14). As of August 2020, traffic congestion has returned to normal, reaching pre-COVID levels in New York City. The increase in traffic appears even more striking considering that only 15% of workers had returned to their Manhattan offices as of October 2020 (15). Understanding the effects of the COVID-19 pandemic on air quality and PM$_{2.5}$ levels in New York is important because inadequate attention to the results may jeopardize public health and impact policy decisions.

To validate the causality and relationship between the COVID-19 pandemic and the reduction of PM$_{2.5}$, a model quantifying the normal range of PM$_{2.5}$ is needed as a benchmark against 2020 data. Previous works on the topic focus on using a linear regression model (11, 12). However, a preliminary analysis shows that linear regression fits the data set poorly, as a linear model can only capture long-term trends, which results in significant discrepancies and a very wide band of prediction intervals, or ranges surrounding the forecasts at different levels of confidence.

The linear model is insufficient because the historical levels of PM$_{2.5}$ in New York City follow two patterns: the first is a long-term declining trend over time and the second is the seasonal fluctuations around this long-term trend. The seasonal patterns are reflected in the shape of the annual circles, with May and September close to the center, and January and July farthest (**Figure 3**). This seasonal plot uses polar coordinates, where the time axis is circular rather than horizontal, and the distance to the center describes PM$_{2.5}$ levels. The swirl line shrank inward towards the center inter and intra each year, indicating the two movements were happening concurrently.

Given the complex nature of PM$_{2.5}$ movements, a seasonal Autoregressive Integrated Moving Average (ARIMA) model was proposed. ARIMA is a class of model that explains a given time series based on its own past values— that is, its lags and lagged forecast errors—so that the equation can be used to forecast future values. ARIMA is a powerful tool: in its general form, any time series that exhibits patterns and is not a random white noise can be modeled as ARIMA(p, d, q) with three parameters (p, d, q), where p is the order of the autoregressive term, d is the number of differencing required to make the time series stationary, and q is the order of the moving average term.

A seasonal ARIMA model ARIMA(p,d,q)x(P,D,Q)$_m$ is formed by including additional seasonal terms (P,D,Q)$_m$ where m is the number of observations per year and P,D,Q are similar to non-seasonal components of the model (p,d,q) but involve backshifts of the seasonal period. A seasonal ARIMA model provides a simple method for making skillful time series forecasts with seasonal components (3).

This analysis focuses on ARIMA(0,1,1)x(0,1,1) $_{12}$, the most commonly used seasonal ARIMA model with monthly data. Specifically, it uses exponential smoothing to track both the long-term trend and the seasonal pattern of the data. The long-term trend component is a simple exponential smoothing model, which is a technique for smoothing time series data using an exponential window function without constant:

$$\hat{y}_t = y_{t-1} - \Theta_1 \epsilon_{t-1}$$

where is the coefficient for the first order moving average, and is the model discrepancy at period . Over time the exponential function assigns exponentially decreasing weights to the older observations. As approaches 1, the exponential window function becomes a very long-term moving average (16).

The second component is a seasonal adjustment:

$$y_{t-12} - y_{t-13} - \varnothing_1 \epsilon_{t-12} - \Theta_1 \varnothing_1 \epsilon_{t-13}$$

where $0 \leq \Theta_1 \leq 1$ is the coefficient for the first order seasonal moving average. Since the data frequency is the monthly average, the seasonal adjustment introduces a differencing at a lag equal to trailing 12 months to eliminate seasonal effects. Combining the two components, the forecasting equation is constructed as follows:

$$\hat{y}_t = y_{t-12} + y_{t-1} - y_{t-13} - \Theta_1 \epsilon_{t-1} - \varnothing_1 \epsilon_{t-12} - \Theta_1 \varnothing_1 \epsilon_{t-13}$$

Note that there are only two coefficients Θ_1 and $\varnothing_1$ to be estimated in the equation.

Our seasonal ARIMA(0,1,1)x(0,1,1)$_{12}$ model is built with the historical time series data. The model outputs include the forecasts of future PM$_{2.5}$ levels in New York and prediction intervals. More specifically, the 2020 observations are compared with the prediction interval with a confidence level of 95%.

We employed a hypothesis testing process, where the null

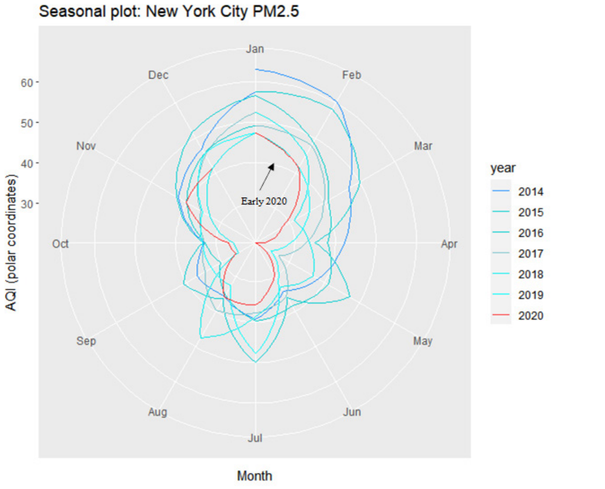

Figure 3: NYC Seasonal Variability and Long-term Trend of PM$_{2.5}$. Source: World Air Quality Index project.

hypothesis was defined as:

H_0: The large drop in $PM_{2.5}$ in the first six months of 2020 in New York City was explained completely by long-term trends and changes in New York air quality.

If the confidence interval contains the value claimed by the null hypothesis H_0, then the sample result is close enough to the predicted value, and we therefore do not reject (17). In our case, if many actual data points fall outside of the 95% confidence interval, the model confirms COVID-19 pandemic reduced $PM_{2.5}$ levels in New York.

RESULTS

To accurately model the air quality of New York City taking into account long-term seasonal trends, we used an ARIMA model. Such a model has the benefit of being more recursive as it is based on its own past values, so any variations or irregularities are considering appropriately. In the ARIMA model, there are only two coefficients that need to be estimated and computed given previous data: namely, Θ_1 and $\emptyset_1$.

The estimated coefficients of the fitted ARIMA(0,1,1) x(0,1,1) $_{12}$ model are Θ_1:0.9285 and $\emptyset_1$:0.6046. With the two coefficients, the model fits the data with a Root Mean Square Error (RMSE) 4.63. RMSE is defined as the standard deviation of the model residuals, which are the prediction errors obtained by subtracting the observed values from the predicted values (3). Our results are consistent with other models that use a linear regression; however, models employing linear regression yield a RMSE of 7.64, which is 65% higher than our model predicts. As the model with lower RMSE is considered superior (18), our model fits the historical data much better.

The model forecasts, prediction intervals and actual observations are summarized in **Table 1**. Prediction intervals are an estimate of an interval in which an observation is likely to fall into with a certain probability. Our model therefore predicts that 80% of the time, the actual level will lie between the low 80% and high 80% markers, and similarly for the 95% prediction interval. Notably, while there are several months where the model forecast and actual level differ by almost 7, the total trends match up well and there is no statistically-significant difference between the model's outputs and the actual data. Model outputs from the R toolbox will come with expected $PM_{2.5}$ levels, as well as prediction intervals. If the actual observations fall outside the 95% prediction interval, the reduction of pollution level in the past year is considered statistically significant.

	Actual Level	Model Forecast	Low 80%	High 80%	Low 95%	High 95%
Jan-20	47.17	48.00	41.32	54.68	37.78	58.22
Feb-20	41.57	43.13	36.43	49.83	32.89	53.37
Mar-20	28.10	34.11	27.40	40.83	23.85	44.38
Apr-20	22.73	29.50	22.77	36.23	19.20	39.79
May-20	20.30	28.49	21.74	35.23	18.17	38.81
Jun-20	29.60	30.48	23.71	37.24	20.13	40.82
Jul-20	35.80	39.86	33.08	46.64	29.49	50.23
Aug-20	35.77	34.74	27.95	41.54	24.35	45.14
Sep-20	25.87	28.63	21.82	35.45	18.21	39.05
Oct-20	26.97	28.26	21.43	35.09	17.81	38.71
Nov-20	39.87	33.05	26.20	39.90	22.58	43.52
Dec-20	41.47	41.86	35.00	48.73	31.36	52.36
Jan-21	45.37	46.64	39.09	54.19	35.09	58.19

Table 1: Forecasts of Jan 2020-Jan 2021 using Prior Data vs. Actual Observations.

DISCUSSION

The present work developed a theoretical model to validate the causality between the COVID-19 pandemic and the reduction of $PM_{2.5}$. A seasonal ARIMA model was proposed to forecast the level of $PM_{2.5}$ in New York City, using differencing at a lag equal to trailing 12 months to remove seasonal effects. Comparing the actual data with the model prediction during the period from January 2020 to January 2021, no statistically-significant difference was found when taking long-term trends into consideration. In other words, COVID-19 did not show convincing temporary positive impact on air quality in New York City.

Comparing the model forecasts and the actual observations, with the exception of April and May 2020, forecasts for all other periods were within the 80% prediction confidence interval. April was a close miss, but both April and May were comfortably sitting in the 95% prediction confidence interval. Thus, the hypothesis test failed to reject the null hypothesis at a 95% level of confidence. We therefore concluded that there was no evidence that $PM_{2.5}$ levels in the first six months of 2020 in New York City were significantly different from previous years.

Additionally, a model is a good fit only when residuals are white noise, meaning they are randomly centered around zero with constant variance and uncorrelated over time (19). Our results suggest that our model residuals have a near-constant variance over time (**Figure 4**, top chart). Plotting the autocorrection function (ACF) at different lag for the residuals shows that there is no spike at any particular frequency (**Figure 4**, lower left chart). Residuals are normally distributed around zero, and the model outputs suggest that such residuals are white noise that can be ignored (**Figure 4**, lower right chart).

One pitfall of this study is that the distribution of the model residuals has a sizeable right tail indicating the presence of significant prediction errors, which reduces confidence in our model output. A next investigative step would be to analyze the cause of this right tail and improve the model accordingly.

MATERIALS AND METHODS

The statistical toolbox used for the analysis is the Forecast Package in R. R is an open-source, object-oriented scripting language with a large set of available add-on packages (20). Additional details about the package may be found in its

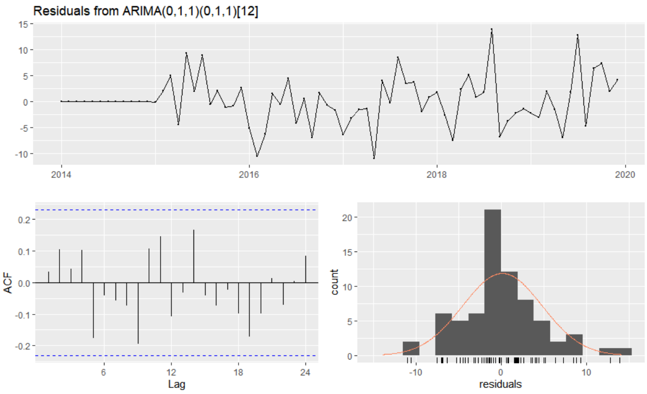

Figure 4: Residuals from the Fitted ARIMA(0,1,1)x(0,1,1) $_{12}$ Model.

manual (3).

The model inputs are the historical monthly $PM_{2.5}$ levels in New York City. The raw data comes from daily observations taken from the database of the World Air Quality Index (WAQI) project (4). From this data, the monthly averages were calculated, and the data was processed and made readable in R.

The seasonal ARIMA(p,d,q)x(P,D,Q) m model is configured to have (p,d,q) = (0,1,1), (P,D,Q) = (0,1,1), and m = 12 in the R setting. It is fitted to the historical time series data between 2014 and 2019 and predicts future points in January 2020 and January 2021.

The historical $PM_{2.5}$ data set and the R code are available for download at https://github.com/williamzning/William-Research.

REFERENCES

1. Mackenzie, Jillian, and Jeff Turrentine. “Air Pollution: Everything You Need to Know.” NRDC, 22 June 2021, www.nrdc.org/stories/air-pollution-everything-you-need-know.
2. “Health and Environmental Effects of Hazardous Air Pollutants.” EPA, Environmental Protection Agency, www.epa.gov/haps/health-and-environmental-effects-hazardous-air-pollutants.
3. “Particulates.” *Wikipedia*, Wikimedia Foundation, 16 Sept. 2021, https://en.wikipedia.org/wiki/Particulates.
4. The World Air Quality Index project. *Contacting the World Air Quality Index team*. aqicn.org. aqicn.org/contact/.
5. Hyndman, Rob, and George Athanasopoulos. *Forecasting: Principles and Practice*. 3rd ed., Otexts, 2021.
6. AirNow.gov, U.S. EPA. (n.d.). *AQI Basics*. AQI Basics | AirNow.gov. www.airnow.gov/aqi/aqi-basics/.
7. New York City Department of Health and Mental Hygiene, *Tracking Air Quality across New York City* 2008-2015, Epi Data Brief, April 2017, No. 88, www1.nyc.gov/assets/doh/downloads/pdf/epi/databrief88.pdf
8. Berman, Jesse D., and Keita Ebisu. “Changes in U.S. Air Pollution during the COVID-19 Pandemic.” *Science of The Total Environment*, Elsevier, 1 June 2020, https://www.sciencedirect.com/science/article/abs/pii/S0048969720333842.
9. He, Guojun, *et al*. “The Short-Term Impacts of Covid-19 Lockdown on Urban Air Pollution in China.” *Nature News*, Nature Publishing Group, 7 July 2020, https://www.nature.com/articles/s41893-020-0581-y.
10. Seo, Ji Hoon, *et al*. “Changes in Air Quality during the Covid-19 Pandemic and Associated Health Benefits in Korea.” *MDPI*, Multidisciplinary Digital Publishing Institute, 5 Dec. 2020, https://www.mdpi.com/2076-3417/10/23/8720.
11. Zangari, S., *et al*. “Air Quality Changes in New York City during the COVID-19 Pandemic.” *The Science of the Total Environment*, U.S. National Library of Medicine, 25 June 2020, https://pubmed.ncbi.nlm.nih.gov/32640401/.
12. Smith, D. *Data Study: Did the COVID-19 Pandemic Reduce PM2.5 Levels in New York?* Kaiterra. learn.kaiterra.com/en/air-academy/data-study-covid-19-pandemic-pm2.5-nyc.
13. “How COVID-19 Is Affecting New York City Traffic.” The Daily Gazette. dailygazette.com/how-covid-19-is-affecting-new-york-city-traffic/.
14. Caspani, Maria, and Nathan Layne. “Traffic Jams Signal Return to Normal in New York but COVID-19 Cases Jump Elsewhere.” *Reuters*, Thomson Reuters, 22 June 2020, https://www.reuters.com/article/us-health-coronavirus-usa/traffic-jams-signal-return-to-normal-in-new-york-but-covid-19-cases-jump-elsewhere-idUSKBN23T25C.
15. Dan Rivoli, *Traffic Roars Back, Reaching Pre-Pandemic Levels*. Spectrum News NY1. www.ny1.com/nyc/all-boroughs/news/2020/10/01/traffic-roars-back--reaching-pre-pandemic-levels.
16. Duke University. Introduction to ARIMA models. https://people.duke.edu/~rnau/411arim.htm#ses.
17. “Using Confidence Intervals to Test Hypotheses”. University of Wisconsin. mat117.wisconsin.edu/3-using-confidence-intervals-to-test-hypotheses/.
18. Teh, Powers. “An Introduction to Applied Machine Learning with Multiple Linear Regression and Python.” *Medium*, Medium, 25 July 2018, https://medium.com/@powersteh/an-introduction-to-applied-machine-learning-with-multiple-linear-regression-and-python-925c1d97a02b.
19. “White Noise.” *Wikipedia*, Wikimedia Foundation, 7 Sept. 2021, https://en.wikipedia.org/wiki/White_noise.
20. “The R Project for Statistical Computing”. R. www.r-project.org/.

Received: March 19, 2021
Accepted: June 29, 2021
Published: October 1, 2021