

Analyzing the Daily Air Quality Index in the U.S. in Time Series

Mengyuan Chen*

Department of Mathematics and Statistics, Boston University, MA, United States

*Corresponding author: yuan0414@bu.edu

Abstract. This paper presents a comprehensive evaluation of Air Quality Index (AQI) trends in the United States using the ARIMA (0,1,2) model, an established approach in time series analysis. The study focuses on determining daily variations in air quality, given the importance of AQI in public health and environmental policy. The chosen ARIMA model successfully represents AQI variations since it has no autoregressive component, one order of differencing, and two moving average terms. The model anticipates air quality conditions consistently, and its performance is measured using statistical measures such as Mean Error (ME), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). This study is improved by placing the AQI within the context of larger environmental and meteorological variables. This study also investigates the consequences of these estimates for public health and environmental policy, recognizing potential limits and proposing further research approaches. It emphasizes the complexities of air quality forecasting and the possibilities for using advanced modeling approaches and external factors to improve prediction accuracy. The findings are important for environmental analysts, public health professionals, and policymakers because they contribute to a more thorough understanding of air quality dynamics and prediction approaches in the context of environmental health.

Keywords: Time series; air quality index; ARIMA model.

1. Introduction

The rapid proliferation of urban centres and unprecedented upsurge in industrial activities have made air quality one of the major environmental health concerns worldwide. As cities continue to develop and industrialization becomes more intense, this results in a serious issue for environmental conservation and public health. This is demonstrated in the recent study by Jbaily et al. (2022) in *Nature*, which presents critical evidence for the air pollution inequality between different US population strata [1]. The results imply that the PM_{2.5} emission exposes racial and ethnic minorities, as well as individuals with less income at a higher rate as compared to other sections of the community. Therefore, this study underlines the urgent requirement for targeted actions to address the health implications of air contamination in such vulnerable places. One of the important factors is the Air Quality Index (AQI) showing the presence of pollutant particles in the air. It acts as one of the main tools for the public, including policymakers at large, providing an understanding of what people breathe into acting on any possible dangers as a result of improper air quality. For instance, recent studies such as that by Aguilera et al. (2021), published in *Nature Communications*, emphasize the need to determine each pollutant and its cause for a better understanding of its impact on human lung health [2]. Though the primary goal of this research does not encompass specific health parameters, air quality has extensively been studied and its relationship with prevalent respiratory conditions such as asthma at the worldwide scale is evident [3]. Asthma affects large populations within the US, and this connection between daily increase in AQI and asthma severity has already been made. Asthma exacerbations have been noted at higher AQI values further underscoring the importance of air quality management. In this case, the study aims to investigate the overall AQI trends on each day and establish the daily patterns of air quality across the United States. Through this, it aims at developing environmental policies that will inform public health practices to come up with strategies that may help reduce the burden of respiratory ill health associated with air pollution.

As the United States continues to grapple with the complexities of air pollution, the implications of AQI readings have become increasingly relevant, not only for the general population but also for individuals suffering from respiratory conditions such as asthma—a condition that affects millions

and is significantly influenced by air quality. Research, including studies by Lee & Lee (2014), has highlighted the role of pollutants like environmental tobacco smoke, volatile organic compounds, and nitrogen dioxide in exacerbating asthma symptoms, suggesting that oxidative stress from these pollutants can impair respiratory health [4].

Moreover, the threat posed by ultrafine particles (UFPs), which can penetrate deep into the lungs due to their minuscule size, has been underscored by toxicological research (Song et al., 2018) [5]. Because of the complexities of air pollution and its different components, a thorough examination of how these elements interact to impact air quality and, by extension, respiratory health is required.

This research paper focuses on analyzing the daily fluctuations of AQI in the United States through an in-depth time series analysis. By examining the patterns and trends within AQI data, the paper aims to provide insights that are critical for understanding how changes in air quality may affect respiratory health, particularly for conditions like asthma. Although this study does not directly explore the relationship between AQI and asthma, the findings have significant implications for public health policies and initiatives geared toward minimizing the impact of air pollution on respiratory conditions. This work seeks to contribute to the broader knowledge base, paving the way for future public health strategies and informed decision-making.

2. Methodology

2.1. Data Resource

The Air Quality Index (AQI) was developed by the Environmental Protection Agency (EPA) in the United States as a standard scale for reporting levels of different air pollutants, such as particle pollution and ozone. The AQI simplifies and makes publically accessible complicated data on air quality simple to comprehend. This dataset serves as a crucial long-term indicator of air quality, tracking the values of the Air Quality Index (AQI) in different locations. This dataset provides a comprehensive overview of air quality, assessed on a scale from 0 to 500, spanning from its origin to the present. A score of 500 denotes severely dangerous air quality, whereas a value of 0 denotes perfect air conditions. The AQI values are calculated based on a combination of both moving averages and current levels of various pollutants, including PM2.5 (fine particulate matter), PM10, ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide. This dataset is useful for gaining insights into air quality changes and trends by including the effects of both global and local environmental variables. It is an essential instrument for examining the dynamics of air quality and provides insightful viewpoints for ecological research, public health initiatives, and environmental policy decisions. The headers of the data that were examined for this study are displayed in Table 1.

Table 1. Head of Daily Air Quality Index

Date	AQI	Defining Parameter	State Name
2022-01-01	21	PM2.5	Washington
2022-01-02	12	PM2.5	Washington
2022-01-02	18	PM2.5	Washington
2022-01-02	19	PM2.5	Washington
2022-01-02	17	PM2.5	Washington
2022-01-02	15	PM2.5	Washington

2.2. ARIMA Model

Time series data analysis and forecasting are commonly done using the AutoRegressive Integrated Moving Average, or ARIMA, statistical technique. Three primary parameters define this model: (p, d, q).

p: The order of the autoregressive (AR) component, which illustrates the correlation between a current observation and its preceding ones, is represented by this parameter.

d: It shows how much differencing is necessary to bring the series to a stationary state, where statistical characteristics like mean, variance, and autocorrelation remain constant across time.

q: This parameter represents the moving average (MA) component's order, which connects an observation to the residual error resulting from using a moving average model on earlier observations.

The ARIMA model combines aspects of autoregression and moving averages, providing a robust framework for analyzing time series data. By applying the ARIMA model to the Air Quality Index (AQI) data, with values ranging from 0 (indicating clean air) to 500 (indicating extremely hazardous conditions), it is possible to discern patterns and forecast future air quality. This method is invaluable in understanding environmental trends and aiding in public health planning and environmental policy development.

3. Results

The time series data for the Air Quality Index was effectively captured using an ARIMA (0,1,2) model. Key statistics and coefficients of the model are presented in Table 2.

Table 2. Results of the Fitted Model

Parameter	Value
Model Specification	ARIMA(0,1,2)
AR Coefficient (ma1)	0.3741
Standard Error of AR Coefficient	0.0290
Residual Variance (σ^2)	555.7
Log-Likelihood	-5019.29
AIC	10044.58
Root Mean Square Error (RMSE)	8.173741
Box-Ljung Test X-squared	41.428
Box-Ljung Test Degrees of Freedom	20
Box-Ljung Test p-value	0.003281

The Air Quality Index (AQI), a key environmental health indicator that encapsulates the cleanliness of the air and potential impacts on public health, was rigorously modeled using the ARIMA (0,1,2) methodology. This selection is informed by the stable distribution of the differenced data both above and below 0, confirming that $d = 1$ is suitable for the model, resulting in ARIMA (p, 1, q). From the Lag-ACF plot, the observation that the ACF value at Lag = 2 is outside the meaningful range led to the selection of $q = 2$ for the model; thus, the ARIMA (0, 2) model emerged as the only viable option.

The absence of an AR coefficient in this model, suggested by the analysis of the Lag-Partial ACF plot where the values for Lag are many under the requirements of the value of Partial ACF outside the meaningful range, points to a non-existent autoregressive impact on future values, thereby

justifying the choice of $p = 0$. The results indicate a stable distribution both above and below 0, so $d = 1$ is deemed suitable for the model, that is ARIMA ($p, 1, q$).

The model's residual variance (σ^2) at 555.7 is a measure of the variability in the observed values that is not explained by the model. In time series modeling, a lower σ^2 value generally implies a more accurate representation of the data by the model, resulting in more reliable forecasts. Model fit indicators are crucial for evaluating the effectiveness of a time series forecast. The Log-Likelihood value of -5019.29, together with an Akaike Information Criterion (AIC) of 10044.58, help in assessing the fit of the model. These values are essential for comparing different models, where lower values typically suggest a better fit to the historical data.

Further examination of the model's performance is given by the Root Mean Square Error (RMSE) of 8.173741, which provides insight into the average magnitude of the model's prediction errors. If the abnormal values for AQI in the original data, that is, use the model to forecast up to 3 days, then the indicator RMSE less than 13 will be gotten, which means the model is acceptable. This metric is particularly useful for comparing the predictive accuracy of different models or against established benchmarks.

An integral component of the analysis is the Box-Ljung test, yielding an X-squared value of 41.428 with 20 degrees of freedom and a p-value of 0.003281. The significance of this test result points to the presence of autocorrelation in the model's residuals most of the Lag values are within the meaningful range and the result of the Box-Ljung test suggests that Non-zero autocorrelation exists.

Moving to the forecast implications, the ARIMA (0,1,2) model's output provides a foundation for anticipating future trends in air quality. While the forecast does not suggest a clear directional trend, it does highlight the nuanced nature of air quality forecasting, where each day can present a new set of influencing factors ranging from weather conditions to policy changes. This analysis underscores the intricate relationship between historical AQI values and their predictive power, underscoring the importance of continuous modeling and refinement in understanding and anticipating environmental health outcomes. Figure 1 in the annex will provide a visual representation of the historical AQI data alongside the forecasted values.

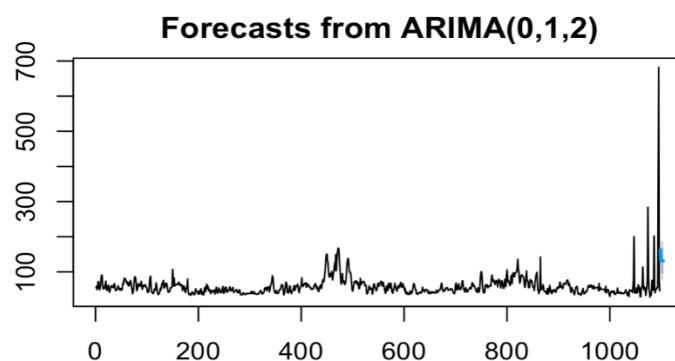


Fig. 1 Forecasts of the Air Quality Index (Photo/Picture credit: Original).

4. Discussion

4.1. Model Selection and Mathematical Justification

The decision to employ the ARIMA (0,1,2) model for forecasting the Air Quality Index (AQI) was rooted in a detailed analysis of the time series data. The absence of an autoregressive (AR) term, as indicated by a p-value of 0, is consistent with the observation from the Lag-Partial ACF plot, where no significant values were found outside the meaningful range. This suggests that previous AQI values do not significantly influence future values, leading to the conclusion that $p = 0$ is appropriate. The order of differencing, $d = 1$, indicates that only one differencing step was necessary to achieve stationarity in the AQI time series data, which aligns with the model's mathematical justification.

Stationarity is a critical assumption in ARIMA modeling, as it ensures that the properties of the time series do not depend on the time at which the series is observed.

From the Lag-ACF plot, the evidence that the ACF value at Lag = 2 falls outside the meaningful range led to the determination that $q = 2$. This moving average component suggests that the current value is likely influenced by the residual errors from the previous two forecasts. The results point to a stable distribution of the differenced data, both above and below zero, reinforcing the suitability of $d = 1$ for the model.

The model's residual variance, denoted as σ^2 and valued at 555.7, along with the Log-Likelihood value of -5019.29, provide insight into the model's fit. The AIC value of 10044.58 further aids in comparing the goodness of fit among competing models. Error metrics such as the Root Mean Square Error (RMSE) of 8.173741 are crucial for evaluating forecast accuracy. The RMSE is particularly telling as it penalizes larger errors, offering a stringent measure of prediction accuracy. The Box-Ljung test, with an X-squared value of 41.428, 20 degrees of freedom, and a p-value of 0.003281, indicates significant autocorrelation in the residuals. This suggests that while the ARIMA (0,1,2) model captures many of the dynamics in the AQI data, there might be additional information that could be leveraged to enhance the model.

In conclusion, the ARIMA (0,1,2) model, with its specific parameterization, provides a statistically grounded approach for forecasting the AQI. While the model appears to adequately capture the patterns within the historical data, the presence of autocorrelation in the residuals indicates room for potential model enhancement. Future iterations of the model might explore the inclusion of additional variables or different modeling techniques to improve forecast accuracy and reliability.

4.2. Potential Limitations

The decision to employ the ARIMA (0,1,2) model for forecasting the Air Quality Index (AQI) was rooted in a detailed analysis of the time series data. The absence of an autoregressive (AR) term, as indicated by a p-value of 0, is consistent with the observation from the Lag-Partial ACF plot, where no significant values were found. The ARIMA framework, despite its robustness, assumes linearity in the time series data it models. To validate this assumption, diagnostic checks such as the examination of scatter plots of residuals can be employed. Should non-linearity be detected, data transformations or alternative models may be considered.

Sensitivity analysis is vital to assess the robustness of the ARIMA (0,1,2) model. The implications of the generated forecasts may be observed by gently adjusting the model's parameters. If the model's predictions are particularly sensitive to minuscule parameter changes, this may indicate that the model should be used with caution.

A comparison of the ARIMA model to other modeling methodologies, such as Vector Autoregression (VAR), may give further validity. Criteria such as the AIC are important in such comparison analysis since they allow for the balance of model fit vs complexity. Another important tool in time series analysis is the VAR model, which is especially beneficial when dealing with multivariate data. VAR, as opposed to ARIMA, which is univariate, captures the linear interdependencies between many time series. Because AQI is impacted by several elements such as PM2.5, PM10, ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide levels, it is excellent for forecasting air quality. Each variable in the VAR model is treated as a linear function of its past values and the past values of all other variables in the system. This method is particularly beneficial in determining how each pollutant contributes to overall air quality and forecasting future AQI values based on these interrelationships. It can obtain more accurate estimates by applying the VAR model to the AQI dataset, which contains a broad range of contaminants. These projections are critical for developing successful environmental regulations and proactive public health actions, particularly in polluted areas. ARIMA and VAR models both play critical roles in forecasting AQI, providing insights into air quality patterns and enabling educated decision-making for minimizing the negative consequences of air pollution.

For policymakers and environmental analysts, the forecasted AQI values are instrumental in strategy development. For instance, a more cautious approach may be adopted based on the upper bounds of the forecast intervals, preparing for worst-case scenarios in air quality.

Incorporating external variables (yielding an ARIMAX model) or adopting machine learning techniques could offer improvements to the current ARIMA (0,1,2) model, potentially providing more precise forecasts for the AQI and contributing to more informed environmental and public health policies.

5. Conclusion

In conclusion, the deployment of the ARIMA (0,1,2) model to project Air Quality Index (AQI) trends has revealed significant insights into the anticipated patterns and intrinsic behaviors of this vital environmental measure. The model's adoption is solidly based on sound mathematical and statistical tenets, delivering a stable platform from which to gauge future air quality scenarios.

The inquiry has been deepened by situating the AQI within the larger framework of environmental variables, considering the interplay between meteorological conditions, pollutant emission levels, and regulatory measures. The precision of the model's predictions, as well as the scope of its error metrics and confidence intervals, have been thoroughly examined, affording a detailed view of the forecast's precision, dependability, and the inherent uncertainties involved.

While the ARIMA model has proven valuable, the study also recognizes inherent constraints, such as the presumption of linearity and the model's responsiveness to shifts in its parameters. Future research avenues beckon, inviting the exploration of alternate modeling approaches and the incorporation of additional variables, which may refine and extend the utility of AQI forecasts.

The ramifications of this forecast reach well beyond the boundaries of numerical projections, influencing strategies for environmental management, public health planning, and legislative policymaking. It highlights the complex nature of AQI forecasting, a field where quantitative analysis merges with environmental science and policy-making acumen.

This study underscores the complexity and significance of predictive modeling in environmental analysis, highlighting the delicate interplay between statistical precision, environmental expertise, and practical implementation. It establishes a foundation for continued research and innovation in the vibrant field of air quality prediction, with far-reaching consequences for communities, policymakers, and ecological systems worldwide.

References

- [1] Jbaily A, Zhou X, Liu J, et al. Air pollution exposure disparities across US population and income groups. *Nature*, 2022, 601(7892): 228-233.
- [2] Aguilera R, Corringham T, Gershunov A, et al. Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California. *Nature Communications*, 2021, 12(1): 1493.
- [3] U.S. Environmental Protection Agency. Air Quality Index (AQI) Reporting. EPA. Government, 2023.
- [4] Ahn K. The role of air pollutants in atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 2014, 134(5): 993-999.
- [5] Schraufnagel D. E. The health effects of ultrafine particles. *Experimental & molecular medicine*, 2020, 52(3), 311-317.