Project Final:
Seasonal Analysis and Forecasting of Air Quality in Maricopa County

Joseph Angel
College of Engineering and
Applied Science
University of Colorado, Boulder
Boulder, CO, United States
Joseph.Angel@colorado.edu

ABSTRACT

This study analyzes air quality trends in Maricopa County, Arizona, using over 30 years of historical Air Quality Index (AQI) data from 1994 to 2024. This study decomposes and forecasts air quality patterns using time series decomposition and predictive modeling techniques. Seasonal decomposition was applied to identify long-term trends, seasonal patterns, and residual noise in the AQI data. The analysis revealed a clear annual cycle in air quality, with higher AQI values during certain seasons, alongside a notable upward trend in the late 2010s, followed by a more recent decline.

We further employed a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to forecast future AQI values. The model captured seasonal and non-seasonal patterns by fitting historical data, allowing for accurate trend projection. The SARIMA model, with parameters (5, 1, 0) for non-seasonal components and (1, 1, 1, 12) for seasonal components, provided a robust forecast, particularly emphasizing the seasonal variations and potential outliers due to exceptional events, such as high pollution periods in 2019 and 2020.

This study's findings are critical for policymakers, environmentalists, and public health officials, offering insights into long-term air quality trends and providing a framework for predicting future air quality conditions. The predictive capabilities of the model can aid in strategic planning for pollution control and public health interventions, especially in areas prone to seasonal air quality variations.

INTRODUCTION

Air quality has become an increasingly important issue, particularly in urban areas where pollution levels can greatly impact public health and the environment. In the United States, the (AQI) is the most well-known and widely used metric for assessing the quality of the air on a given day, providing a standardized measure of the concentration of pollutants such as ozone, particulate matter (PM), carbon monoxide, sulfur dioxide, and nitrogen dioxide. Poor air quality is known to cause adverse health outcomes, ranging from respiratory diseases to cardiovascular problems. Therefore, understanding the patterns and trends in air quality over time is crucial for effective environmental management and public health planning.

Maricopa County, which encompasses the Phoenix metropolitan area in Arizona, is one of the fastest-growing regions in the United States. As the population has surged, so too have concerns over air pollution, driven by increased vehicle emissions, industrial activities, and climate-related factors such as high temperatures and dust storms. Monitoring air quality trends in this region is particularly important given the county's desert climate and geography, which can exacerbate certain types of pollution, especially particulate matter.

This paper aims to analyze historical AQI data for Maricopa County from 1994 to 2024 and make predictions about future AQI values and trends. The paper provides a clearer understanding of long-term shifts and seasonal variations in air quality through seasonal decomposition. Additionally, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model is employed to forecast future air quality trends for the next five years. The insights gained from this study can inform both policy decisions and public awareness initiatives aimed at improving air quality and mitigating its health impacts.

This study contributes to the existing research on air quality monitoring and forecasting by applying advanced statistical techniques to a large dataset of historical pollution data.

The use of SARIMA modeling allows for both seasonal and non-seasonal pattern detection, making it a robust tool for predicting future trends in air quality. The findings identify trends in Maricopa County's air quality and provide a framework that can be applied to other regions facing similar environmental challenges.

RELATED WORK

Air quality data and trends are of high interest to both the government and health organizations. The EPA and Lung.org publish their own findings and insights from the AQI data. The EPA primarily focuses on identifying concentrations of specific pollutants.[2] Lung.org has their own yearly "State of the Air Report" that is insightful informational with insights on interesting facts and trends regarding air quality.[3]

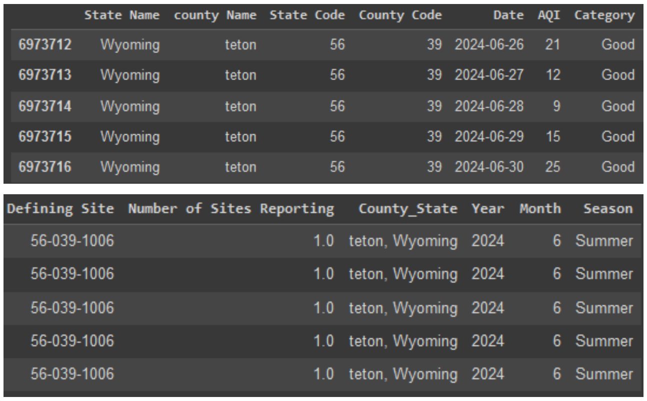
Additionally, previous studies have explored air quality and its health impacts, employing various statistical and anomaly detection methods. For instance, "Long-Term Exposure to Air Pollution and Effects on Respiratory Health" investigates the adverse effects of prolonged exposure to air pollutants, including PM10, on respiratory health across multiple locations. This study differs by focusing on health outcomes rather than forecasting, however, it uses similar long term AQI data.

"Anomalous Detection in Air Quality Measurements" emphasizes detecting anomalies in air quality datasets through computational algorithms. While their work focuses on historical anomaly detection, this project combines forecasting techniques with anomaly analysis, aiming to predict future air quality trends and highlight irregularities over a five-year horizon. Therefore, this study's work builds on the predictive and detection approaches found in these studies while introducing forecasting based on seasonal decomposition, and providing a predictive analysis of air quality.

PROPOSED WORK

The dataset utilized in this study comprises daily (AQI) reports collected from 771 U.S. counties between 1994 and 2024. The data includes essential variables such as date, county name, state, and AQI values. The data initially started in separate .csv files and needed to be combined accordingly. Conflicts appeared when combining data across files, like for the 'Date' variable where some data was in the MM/DD/YYYY format while others were in the YYYY-MM-DD format. However, these were combined and converted to a uniform datetime format.

The AQI data was resampled to obtain monthly averages and used a month-start ('MS') frequency, to analyze seasonal patterns. This approach provided a clearer view of AQI trends over time and allowed the capture of inherent seasonal variation. Furthermore, a seasonal classifier was created by extracting the month from the date, enabling the analysis of AQI patterns across different seasons.



	State Name	county Name	State Code	County Code	Date	AQI	Category
6973712	Wyoming	teton	56	39	2024-06-26	21	Good
6973713	Wyoming	teton	56	39	2024-06-27	12	Good
6973714	Wyoming	teton	56	39	2024-06-28	9	Good
6973715	Wyoming	teton	56	39	2024-06-29	15	Good
6973716	Wyoming	teton	56	39	2024-06-30	25	Good

Defining Site	Number of Sites Reporting	County_State	Year	Month	Season
56-039-1006	1.0	teton, Wyoming	2024	6	Summer
56-039-1006	1.0	teton, Wyoming	2024	6	Summer
56-039-1006	1.0	teton, Wyoming	2024	6	Summer
56-039-1006	1.0	teton, Wyoming	2024	6	Summer
56-039-1006	1.0	teton, Wyoming	2024	6	Summer

Figure 1: Head of the combined data files.

Now that the data was ready for analysis, the data was grouped by season and calculated the average AQI values for each season across all years in the dataset (1994-2024). The data was for all counties with AQI data over the period (774 counties total). The groupby function was used to aggregate the data based on the season classifier, which was extracted from the date field. Plotting the seasonal averages using a bar chart allows us to visually compare AQI levels across Winter, Spring, Summer, and Fall.

The results showed noticeable seasonal patterns, with the highest AQI values in the Summer. Specifically, the Summer months have an average AQI around 22% higher than any other season. This pattern likely corresponds to increased emissions Summer caused by industrial activities and wildfires. The analysis highlights the need for seasonal-specific policies or interventions to manage air quality more effectively during these periods.

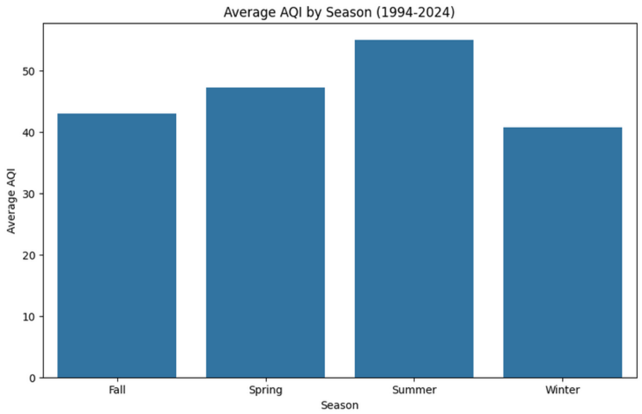


Figure 2: Average AQI by season (1994-2024).

Next, how seasonal patterns have changed over the years was investigated. The average AQI values were analyzed for each season over time by grouping the data based on Year and Season. Doing so observed long-term variance and shifts in seasonal air quality patterns from 1994 to 2024. Figure 3 shows how the AQI values have changed seasonally across different years and visualizes these trends. The results highlighted significant variations and spikes in certain years, particularly during summer, such as 2003, 2012, and 2023, which may correlate with specific environmental events or policy changes.

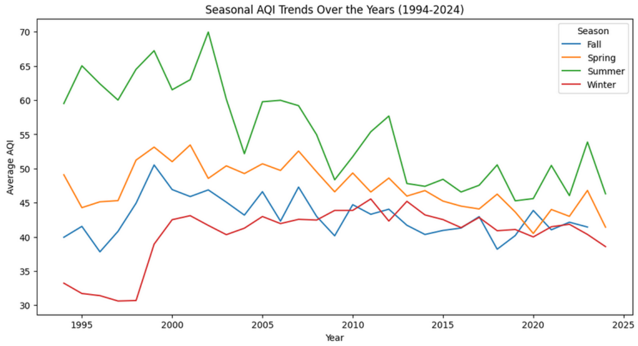


Figure 3: Average AQI Over the Years by Season (1994-2024).

Finally, the historical data is used to implement predictions for future pollution levels. Time series forecasting techniques were employed using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. This process consisted of the following key steps: First, the data was further preprocessed to fit the requirements of the next steps. The AQI values were then resampled to a monthly frequency, which allowed the capture of longer-term seasonal patterns. Next, seasonal decomposition was performed to separate the time series into four components: observed values, trend, seasonality, and residuals. These graphs provide insights into the underlying patterns of the AQI data over the observed period and provide an understanding of the data necessary to create an effective model.

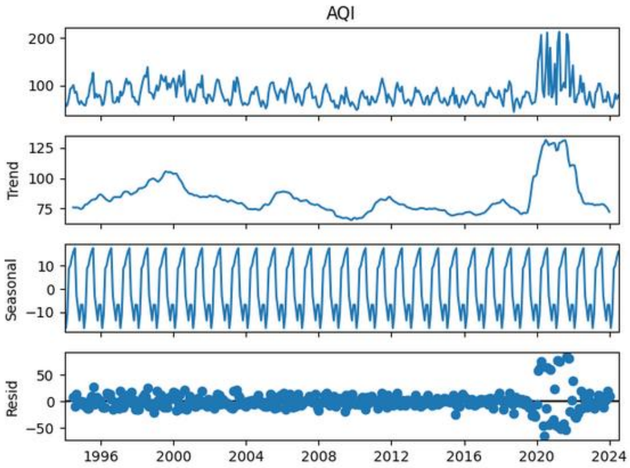


Figure 4: Seasonal Decomposition Plot.

To predict future AQI values, the SARIMA model was chosen because it effectively accounts for both trend and seasonality within the data. The model was configured with non-seasonal parameters `(5, 1, 0)` and seasonal parameters `(1, 1, 1, 12)`, reflecting a monthly seasonality with a 12-month cycle. The SARIMA model was then trained on the historical AQI data to generate forecasts for the next 60 months (5 years). This allowed the anticipation of potential high-pollution periods and observe how AQI levels might fluctuate in the future.

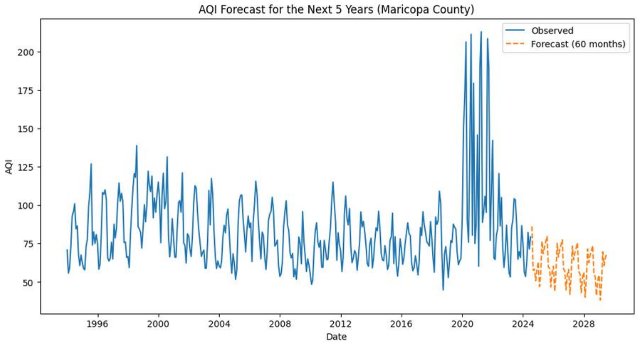


Figure 5: SARIMA Forecast Plot. Predictions shown in orange.

The forecast captured repeating seasonal patterns and a slight negative trend in the AQI levels. Both of these patterns would be expected considering the previous graphs and we've shown which typically show a slight downslope over time. Additionally, when aggregating the slopes for AQI over time for each county, the mean was approximately -.00077.

How accurate these predictions are would remain to be seen, however, from a policy planning perspective, this provides valuable insights into the seasonal and yearly trends in air quality in Maricopa County.

EVALUATION

The performance of the SARIMA model was evaluated using the test set, which consisted of the last 12 months of AQI data from Maricopa County. The model's predictive accuracy was assessed using two key error metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results of these metrics were as follows: MAE: 10.56 and RMSE: 13.38.

Given that the typical AQI values for Maricopa County range between 50 and 150, these error values suggest that the SARIMA model performs reasonably well. The MAE value indicates that, on average, the model's predictions deviate from the actual AQI values by approximately 10 points. This is relatively low considering the overall scale of AQI values, which suggests that the model can capture the primary trends and seasonal variations effectively.

The RMSE, which penalizes larger errors more heavily, is only slightly higher than the MAE, indicating that the model is not making significantly large errors for most predictions. This is an encouraging sign of the model's robustness and consistency in its predictions.

DISCUSSION

The findings of this study provide valuable insights into the seasonal patterns and long-term trends of air quality in Maricopa County, Arizona, over 30 years. By using seasonal decomposition techniques and the SARIMA model, AQI tends could be identified and forecasted, highlighting potential high-pollution periods and overall air quality fluctuations. The results show clear seasonal variations, with consistently higher AQI values observed during the summer months. These results correspond with increased industrial activities, higher temperatures, and potential wildfire events during this season.

The trend analysis revealed that Maricopa County experienced short periods of significant increases in AQI levels. These include the late 1990s and 2020-2022, potentially due to specific environmental conditions or regulatory changes. However, a general downward trend in AQI was observed in recent years, which could reflect the effectiveness of local environmental policies and interventions aimed at improving air quality.

The SARIMA model effectively captured both seasonal and long-term trends in AQI levels. Its forecast for the next five years indicates that while there may be fluctuations in AQI levels, the general trend should remain relatively stable or continue its slight downward trajectory. The trend suggests that current pollution control measures and regulations are producing healthier air quality, although continuous monitoring and policy adjustments are still necessary, particularly during high-risk periods. Additionally, from the evaluation, we can be see the model is fairly good at predicting AQI values, however, there may be room for improvement and other models that may produce even more accurate predictions.

Overall, time series forecasting methods such as SARIMA provide a strong foundation for understanding and predicting air quality trends. However, it is essential to note that the model's predictions are based solely on historical data and do not account for future changes in environmental regulations, population growth, or other unforeseen factors. Therefore, while the model offers valuable insights, conclusions should be made cautiously and used in the context of surrounding environmental and policy considerations.

CONCLUSION

This study examined air quality trends in Maricopa County, Arizona, using over 30 years of historical AQI data to understand seasonal patterns and predict future air quality conditions. Through time series decomposition and implementing a SARIMA model, both the cyclical nature of AQI values and the underlying long-term trends were captured.

The analysis revealed that AQI levels vary significantly by season and that higher values were consistently observed during summer months. Furthermore, while the AQI has shown occasional spikes in recent decades, there appears to be a general downward trend, suggesting that regulatory efforts and public awareness campaigns may be positively impacting air quality in the region.

Our SARIMA model provided a reliable forecast of future AQI levels, with predictions indicating relatively stable air quality over the next five years. Although the model performed well based on standard error metrics such as MAE and RMSE, there are still potential limitations. For example, external variables like meteorological factors or sudden policy changes could influence future pollution levels and would not be captured by the model.

1.1 Potential Improvements. There are a few potential improvements that could be made to this project. To further improve the strength of the model, adding exogenous variables could improve prediction accuracy. For example, things like weather conditions, and industrial activities on that day are known to influence the AQI value read on a given day. The gathering and combining of that data into our model can therefore give it more information to base its predictions on and potentially increase the accuracy at the evaluation stage. Additionally, this would inform policy makers of what variables are having the greatest impact on AQI rather than simply showing how seasons impact the air quality.

Another potential improvement would be to implement our model on all counties we have data for and create a geospatial map showing how different seasons impact the AQI values within different regions. The data and model we have now is obviously applicable to Maricopa County, however, other regions have different geography and therefore may be affected by seasons differently. Therefore, at this stage it would not be right to take this data and apply it to other counties without a geospatial approach.

Finally, all the predictions made have been done with the SARIMA model. While this model is highly applicable to our study, there may be more effective machine learning models and predicting the future AQI values while still capturing the seasonal variance that this project sought to identify. Other models can be implemented and evaluated accordingly and after being compared to the current model, we can properly determine if the SARIMA model should still be the model used to make our seasonal AQI predictions.

1.2 Future Work. Given the basis that the project has laid out, there is lots of room for future work to build from it. What path that future work would take would just depend on how this data is going to be used, and what areas of interest will be focused on.

For example, one possibility for future work would be to link our AQI data with health data to determine how health outcomes are impacted by the AQI values. In the context of our seasonal model, and the fact that Summer consistently has the largest AQI values, it could follow that people that are outdoors the most during the Summer are more likely to suffer from conditions like asthma. Therefore, linking these values could inform individuals of the risks of being outdoors during periods of high AQI values and help us further understand what healthy AQI values and what the safest and most dangerous periods to go outdoors would be.

Another potential area for future work to build off this project would be energy policy. Air quality is closely linked to energy consumption patterns, particularly those involving fossil fuel combustion. Future research could explore the relationship between AQI values and energy consumption data to identify how different sources of energy contribute to air pollution across various seasons. For instance, this research could examine whether periods of higher AQI values correspond with increased use of coal or natural gas in power generation. Additionally, if policies are passed into law that restrict or subsidize the usage of certain energies, we can use AQI values as a metric to determine the effectiveness of these policies. By understanding these correlations, policymakers can make more informed decisions about energy policies and regulations that target emissions reductions during peak pollution periods.

Furthermore, the project could be expanded to analyze the impact of renewable energy adoption on AQI values. With many regions transitioning to greener energy sources such as wind and solar, it would be valuable to assess how these changes are affecting air quality trends over time. Integrating renewable energy usage data with AQI trends could reveal the effectiveness of clean energy initiatives and provide a case for accelerating renewable energy adoption in regions with historically poor air quality. Such studies could guide the development of policies that optimize the mix of energy sources to achieve better air quality outcomes, particularly in areas where seasonal pollution levels pose a recurring challenge.

Ultimately, AQI values are of a large interest to many parties and areas of research. As good air quality becomes an increasing concern among advocacy and research groups, the seasonal trends found in this project can be taken and applied to many different areas.

In conclusion, this study offers a comprehensive view of historical and predicted air quality trends in Maricopa County. The insights can inform future efforts to mitigate air pollution and promote healthier living environments. Continued research and model refinement by integrating additional variables will be essential to further enhance our understanding and forecasting capabilities for air quality in Maricopa County and beyond.

REFERENCES

[1] U.S. Environmental Protection Agency. (2024). *Download Daily Data*. Retrieved from https://www.epa.gov/outdoor-air-quality-data/download-daily-data.

[2] American Lung Association. (2024). *Key Findings: State of the Air Report*. Retrieved from https://www.lung.org/research/sota/key-findings.

[3] U.S. Environmental Protection Agency. (2024). *Air Quality - National Summary*. Retrieved from https://www.epa.gov/air-trends/air-quality-national-summary.

[4] Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W. W., Katsouyanni, K., Brunekreef, B., & Hoek, G. (2014). Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 European cohorts. *Epidemiology, 25*(3), 368-378. https://doi.org/10.1097/EDE.0000000000000076

[5] Paraschiv, M., & Sitaru, A. (2024). Anomalous detection in air quality measurements: A machine learning approach for smart cities. *Journal of Environmental Informatics*, 45(2). https://doi.org /10.1016/j.jenv.2024.08.005