Math6450 Assignment3: Time Series AQI Analysis

Ian Tai Ahn

October 21, 2025

Exploratory Data Analysis (EDA): The dataset was loaded and filtered to keep only the `State Name` and `County Name` columns. Rows were then further filtered to include only entries where the state was Utah and the county was Weber. The dataset contained 365 records for 2022, 365 for 2023, and 366 for 2024 confirming that there was one valid data entry for every single day, with no missing or NA values.

The daily AQI shows considerable day-to-day fluctuations, but overall, each year follows a similar pattern, suggesting the presence of seasonality.

Aggregating the data into monthly averages creates a smoother line, providing a clearer, big-picture view of long-term changes. However, this level of smoothing can hide short-term AQI fluctuations. The monthly averages clearly highlight seasonal trends, particularly the consistent AQI increase during July.

Weekly aggregation captures both short-term variations and broader seasonal trends. It provides a balance between the noisy daily data but not overly smoothed like monthly data.

Reason for Aggregation: Initial exploration showed that while there was enough data for daily time series modeling, the distribution was right-skewed and the daily trends appeared jagged. When comparing aggregated views, the monthly data looked overly smooth, while weekly data retained meaningful variability and visible seasonal patterns.

To capture as many AQI dynamics as possible, both weekly and monthly aggregations were chosen for analysis and forecasting. Aggregating daily AQI values helps normalize the variance, which was initially right-skewed. Monthly AQI distributions still showed some skew, though less pronounced. To further stabilize the variance and achieve stationarity, a log transformation was applied.

Transformations: After applying the log transformation, both the weekly and monthly AQI distributions appeared more normally distributed and exhibited more consistent variance when plotted.

Seasonal Patterns: Seasonality was first evident in the daily AQI plot and became even clearer in the aggregated monthly averages. To ensure stationarity for modeling, seasonal differencing was applied.

A more negative ADF statistic and a smaller p-value both indicate stronger evidence for stationarity. In this case, both metrics confirmed that the differenced data was stationary. Seasonal decomposition further supported this, showing oscillating patterns characteristic of seasonality in air quality data.

Taking the first seasonal difference improved the ADF results with more negative statistics and smaller p-values confirming that the series had become stationary. Comparing ADF results across weekly and monthly datasets (both raw and log-transformed) showed consistently low p-values, reinforcing that the transformations effectively achieved stationarity.

Autocorrelation: In the weekly ACF plot, correlation reverses around lag 8, likely reflecting how weather and environmental conditions shift seasonally. The PACF plot starts high but drops sharply after lag 1 and then tails off without a clear pattern, indicating that only the most recent week has a direct influence on current AQI values.

The monthly ACF and PACF plots show similar behavior. The seasonal patterns seen in these plots reflect the influence of changing seasons on AQI. The steep drop-off after lag 1 in the PACF plot again suggests limited direct influence from earlier months once the most recent month is accounted for.

Correlation Coefficients: The weekly correlation results reinforce the seasonal interpretation. A positive correlation at lag 1 suggests persistence within the same season, while negative correlations at lags 12 and 26 correspond to comparisons across different seasons (e.g., winter versus summer). For instance, data from January compared to 26 weeks earlier (around July) would naturally show an inverse relationship in AQI levels.

The monthly data supports this interpretation: a negative correlation at lag 6 indicates that winter months tend to have lower AQI values than summer months. Overall, both correlation coefficients and ACF plots confirm the strong seasonal influence on AQI.

Data Splitting: An 80/20 train-test split was used for modeling.

Model Selection: Various combinations of (p, d, q, P, D, Q) parameters were tested for the SARIMAX model, guided by ACF/PACF plots and stationarity results. Differencing and seasonal differencing consistently improved model performance, especially for the log-transformed AQI series. Setting both p/P and q/Q to 1 provided better forecasting accuracy, helping the model capture short-term autocorrelation and seasonal patterns observed in the data.

Model Parameters and Diagnostics: The best-performing model for the monthly AQI data was identified as a **SARIMA(1, 1, 1) × (1, 1, 1, 12)** model.

In full notation:

$$\text{SARIMA}(p = 1, d = 1, q = 1) \times (P = 1, D = 1, Q = 1, s = 12)$$

The general form of the model is:

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(1-B)(1-B^{12})Y_t = (1+\theta_1 B)(1+\Theta_1 B^{12})\varepsilon_t$$

The model performance, evaluated across several criteria, is summarized below:

- **Lowest RMSE:** Monthly AQI (RMSE = 5.9618)
- **Lowest MAE:** Monthly AQI (MAE = 5.0084)
- **Lowest MAPE:** Monthly AQI (MAPE = 9.48%)
- **Lowest AIC:** Monthly AQI (AIC = -1.26)

Conclusion: This study was valuable for understanding seasonality and variance in time series data. The dataset was well-structured with no missing values, which allowed for focused analysis on trends, stationarity, and autocorrelation. The exploration of ACF/PACF patterns and differencing choices provided deeper insight into how seasonality impacts AQI.

The best-performing model was the base monthly SARIMA model using monthly average AQI. However, even this model struggled to fully capture the late-year spike observed in the data. Future improvements might involve experimenting with exogenous variables (e.g., temperature, wildfire activity, or emissions) or alternative model architectures to better capture those sharp changes.