Data Acquisition Performance Assessment:

Executive Summary

John Foster

Department of Information Technology, Western Governors University

D214: Data Analytics Graduate Capstone

Professor Daniel Smith

July 17, 2023

Research Question and Hypothesis

The research question that I have chosen to explore is as follows: “With what degree of accuracy can a time series model forecast daily AQI values for Seattle, Washington?”

Air Quality Index (AQI), is a metric developed by the US Environmental Protection Agency to warn the American public of dangerous levels of air pollution in a given geographic area (American Lung Association, 2022). Various airborne pollutants contribute to this index, and prolonged outdoor exposure to elevated values are strongly associated with adverse health effects. The construction of a time series model that can accurately forecast these daily values over time for a given region would have multiple practical applications for organizational decision making, including long term planning for regional medical infrastructure development, adjustments to projected taxations on carbon emissions, and even calculations for health insurance premiums.

In order to adequately explore this research question, we have constructed a null hypothesis and an alternative hypothesis that can be objectively tested against the results of our analysis. These hypotheses are as follows:

Null Hypothesis

A time series model cannot be created using the training dataset that generates predictions with a Mean Absolute Percentage Error (MAPE) below 10%.

Alternate Hypothesis

A time series model can be created using the training dataset that generates predictions with a Mean Absolute Percentage Error (MAPE) below 10%.

Analysis Summary

We loaded the source csv files into dataframes using Pandas, then filtered the data by the CBSA Code corresponding to Seattle, Washington (‘42660’), which reduced the total size of the dataset when the five dataframes were concatenated. At that time, we dropped every column from our dataframe excluding the ‘AQI’ and ‘Date’ column, the latter of which was used as the index of our time series. We finally split our dataset into training and testing sets, with the training set encompassing the first 80% of the data and the testing set covering the final 20%. This split was chosen because it fit neatly into our dataset’s total time period of five years.

After we successfully prepared our dataset, we performed exploratory data analysis to identify obvious trends and seasonality in the raw dataset. We then utilized a custom grid search to iteratively train ARIMA/SARIMA models with a range of base and seasonal orders in a large parameter grid. These models were fitted to the training set and evaluated for their AIC score so that their goodness-of-fit could be compared. The model with the lowest AIC score was selected for forecasting on the testing data so that its MAPE could be calculated. We then compared the model’s MAPE to that of a linear regression to better contextualize our results.

Findings

In the end, the MAPE of our selected model’s forecasting on the testing dataset was just over 30%. According to Lewis' interpretation of MAPE results on time series models, this

indicates that the forecasting accuracy of our model is classified as "reasonable," but does not approach the 10% threshold we established to reject our null hypothesis (1982). To put these results in better perspective, when a simple linear regression was performed on the training dataset and extended to the testing set, the resulting forecast on the test data produced a MAPE of 22%.

Limitations

Over the course of this analysis, it became increasingly apparent that a time series model was unsuited to generate actionable insights related to our identified practical applications when we applied this forecasting technique to raw AQI data. This is largely due to the inherent noisiness of raw AQI data, and this noisiness constituted the most significant limitation of our analysis.

Another significant limitation of this analysis was the chosen python implementation of the base ARIMA/SARIMA model. We used the Statsmodels implementation of time series models and experienced a wide variety of issues with model fitting related to seasonality on our chosen dataset. Consequently, we were precluded from utilizing any existing libraries for our grid search due to the workarounds required to successfully fit our model. Instead, we developed a very large and complicated custom grid with our necessary training and storage configurations. The full grid search took nearly 24 hours of time to complete, and our final model required 4 hours to fit to the training data, necessitating the serialization of our trained model to make development and demonstration reasonably efficient. Upon reflection, this author concludes that Meta's "Prophet" time series toolset would likely be more suited to future analyses of this scope.

Proposed Actions

Based on the final model forecast's MAPE of 30% compared to the linear regression's MAPE of 22%, we can strongly recommend that the methodology of this time series analysis be disregarded as a viable approach to understanding the relative air quality of a given region for the purposes of public health care-related organizational decision making. The complexity and computational demand of this methodology combined with its demonstrated inferiority to a simple linear regression clearly indicate that a time series model is not suited to inform the identified types of business decisions.

Nevertheless, we can identify two future avenues of research to more adequately explore the use cases for a statistical regional understanding of air quality using this dataset.

The first method would seek to address the inherent noisiness of our raw data in order to generate better forecasting and regression models. If we were to apply smoothing techniques such as moving averages or exponential smoothing to our raw data, we would filter a great deal of the noise out of the time series, which would substantially simplify the construction of future generative models, producing results that more closely resemble seasonal trendlines than inaccurate spikes of trailing data. By then calculating the integral of the resulting time series plot (evaluating the total area under the forecast line on the plot), analysts would have a normalized quantitative metric by which to compare the air quality of different geographical regions over an arbitrary number of trailing years in subsequent analyses.

The second approach would be to disregard the quantitative measure of AQI altogether and instead evaluate a region's overall air quality based on the percentage of the total dataset represented by each class in the "Category" column of the source csv files. This column corresponds to the specific risk thresholds of AQI identified by the EPA, and would enable an analyst to drastically simplify how the air quality of a given region is conceptualized. These percentages would also then comprise a moving average of the data for an arbitrary trailing number of years, ultimately enabling analysts to evaluate our identified business needs.

Both methods would much more effectively distill this data to quantitative metrics that could easily be compared across regions and statistically evaluated against the geographical incidence of various medical conditions in future analyses.

Expected Benefits

Because our analysis did not yield a model that can be confidently applied to our identified practical and organizational needs, the benefits of this analysis are instead oriented to our insights into the limitations of these techniques on this type of data. By ruling out time series forecasting as an effective tool to analyze raw AQI data, we can refine how we interact with these datasets and tailor our approach based on the lessons learned.

First, we can propose that using raw AQI data as a time series is an inappropriate technique toward the end of gaining a high level understanding of this data for a given geographical region. Instead, our analysis indicates that the exploration of noise reduction techniques or the transformation of this data using summary statistics would likely yield much better results in generating representative regional metrics for AQI in subsequent research.

We can also infer that the Statsmodels implementation of SARIMAX model fitting is not particularly suited to the task of time series forecasting on large datasets with large seasonal periodicity, and that future analyses performed on daily time series with clear annual seasonality should utilize more efficient alternatives to these toolsets, particularly if a grid search is employed in the model identification process. This will hopefully enable those conducting similar research in the future to avoid the time spent developing around the identified limitations of these toolsets in the course of their own analyses.

Sources

Lewis, C.D. (1982). *Industrial and Business Forecasting Methods*. Butterworths Publishing. London.