****

**AIR POLLUTION: The Chokehold on China**

*A time series forecasting analysis*

**OPIM 5671: Data Mining and Business Intelligence**

**Spring 2024, Section 713**

**Group 9**

**Madhavi Muppani**

**Stephanie Sullivan**

**Divyam Jalota**

**Steeven Raj Burugula**

**Arun Bezawada**

**TABLE OF CONTENTS**

1. EXECUTIVE SUMMARY

2. OPPORTUNITY FOR EXPLORATION

3. DATA UTILIZATION

3.1 DATA PREPARATION

3.2 SEASONALITY

3.3 TREND

3.4 STABILITY

3.5 CROSS-CORRELATION

4. MODEL DEVELOPMENT

4.1 SIMPLE EXPONENTIAL SMOOTHING

4.1.1 ADDITIVE SEASONAL EXPONENTIAL SMOOTHING

4.2 ARIMA (0,0,0)(0,1,0)s

4.3 PREWHITENING

4.4 ARIMAX (0,0,0)

4.5 ARIMAX (1,0,0)

4.6 ARIMAX (1,0,1)

4,7 ARIMAX (2,0,1)

5. MODEL EVALUATION

5.1 SIMPLE EXPONENTIAL SMOOTHING

5.1.1 ADDITIVE SEASONAL EXPONENTIAL SMOOTHING

5.2 ARIMA (0,0,0)(0,1,0)s

5.3 ARIMAX (0,0,0)

5.4 ARIMAX (1,0,0)

5.5 ARIMAX (1,0,1)

5.6 ARIMAX (2,0,1)

6. CONCLUSION

7. LESSONS LEARNED

8. REFERENCES

**1.** **EXECUTIVE SUMMARY**

In response to the pressing issue of air pollution in China, our team performed a comprehensive analysis to assist Chinese policymakers in understanding and combating the problem. Utilizing a dataset from Kaggle comprised of five years of hourly weather measurements, we aimed to forecast the concentration of air pollution in China for 2015. Despite challenges in analyzing hourly seasonality and trends, we prepared the data by switching to a monthly interval for clearer insights. Our analysis involved exploring various models, including Simple Exponential Smoothing, ARIMA, and ARIMAX, with the ARIMAX(1,0,0) model emerging as the most promising.

The evaluation of these models relied on metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) to identify the most effective approach. Pre-whitening techniques played a pivotal role in enhancing the quality of the time series data, resulting in more precise and reliable forecasts for pollution.The incorporation of pressure, wind speed, snow, and rain as predictor variables demonstrated significance in forecasting pollution, providing valuable insights into the dynamics of air pollution.

Our best model, ARIMAX(1,0,0) suggests increasing anti-pollution campaigns leading up to Q1 and October 2015, accompanied by educational efforts to promote green living practices. Our team highly recommends collaborating with policymakers for future analyses to monitor the impact of these actions on air pollution levels in China.

**2.** **OPPORTUNITY FOR EXPLORATION**

According to the [World Health Organization](https://www.who.int/china/health-topics/air-pollution#:~:text=Air%20pollution%20is%20responsible%20for%20about%202%20million,million%20deaths%20in%20the%20same%20period%20in%20China.), air pollution is responsible for “about 2 million deaths per year in China.” Chinese policymakers are eager to understand air pollution trends further to plan for more effective and efficient anti-pollution campaigns. As young professionals eager to make a difference in the world, we began exploring the opportunity to equip policymakers with time series forecast analysis to inform appropriate marketing and political messages for the upcoming year. We sought credible data to further understand correlation between air pollution and weather patterns to further understand potential seasonality and trends worth investigating. After a team-wide effort to find a reliable dataset, we felt comfortable performing our analysis to determine a response to “What is the expected concentration of air pollution in China in 2015?”

**3.** **DATA UTILIZATION**

The [dataset](https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate/data) leveraged for this analysis came from Kaggle and consists of 43.8K rows of hourly weather measurements over 5 years. The 12 attributes for each row include:

| **Column Name** | **Variable Description** |
| --- | --- |
| Year | Calendar year between 2010-2014 |
| Month | Calendar month between January 2010 and December 2014 |
| Day | Calendar day of each month between January and December |
| Hour | Hour of day (24 hours per day) |
| PollutionConcentration\* | Measured in PM2.5 – a measurement of air pollution that calculates the concentration of air particles with a diameter of 2.5 micrometers or less ([US Environmental Protection Agency](https://www.epa.gov/air-trends/particulate-matter-pm25-trends)) |
| Dewpoint | Measured in degrees Celsius |
| Temperature | Measured in degrees Celsius |
| Pressure | Measured in barometers |
| WindDirection | Combined wind direction |
| WindSpeed | Cumulated wind speed |
| SnowHrs | Cumulated number of hours snowed |
| RainHrs | Cumulated number of hours rained |

\***Dependent variable**

**3.1 DATA PREPARATION**

By first running Time Series Exploration, we quickly discovered the difficulty in analyzing hourly seasonality and trends. In addition, the lag indicators of dew point, temperature, pressure, wind direction, wind speed, snow hours, and rain hours presented little correlation with upstream and downstream air pollution levels. Because of this, our team determined it necessary to utilize a monthly interval for a clearer understanding of seasonality and trend direction. We utilized accumulation functions to transform hourly data to daily data, then to monthly data.

The dataset used for this analysis was sourced from Kaggle.

**Data Cleaning and transformation:**

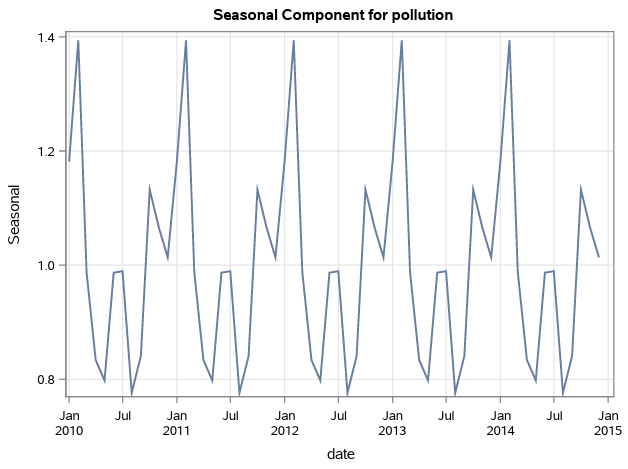
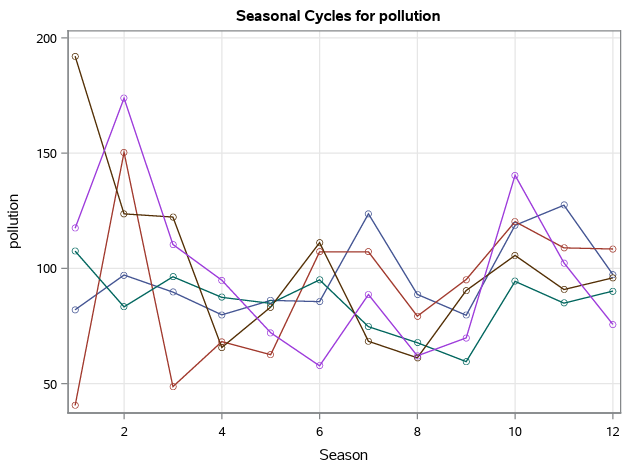
Our dataset exhibits missing values in the column, during the years 2010 to 2014. To address this issue, we need to perform imputation techniques to handle these gaps in the data. Imputation methods such as interpolation or regression analysis can help us estimate and fill in these missing values, ensuring the completeness and integrity of our time series dataset for further analysis and insights.

**Handling Missing Values:**

In our dataset there are missing values. So, to handle the missing values we have used the median value of accumulated time series to imputation in the Data Preparation step.

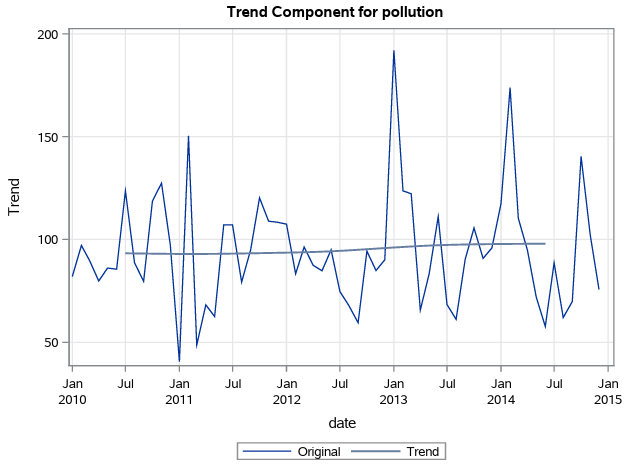
**3.2 SEASONALITY**

There is no seasonality in air pollution in China.



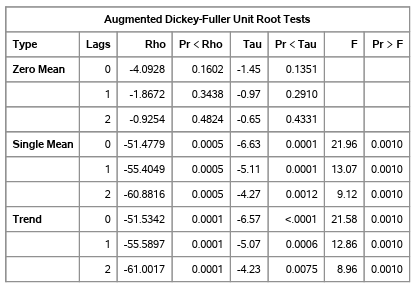
**3.3 TREND**

There is no trend in air pollution in China.

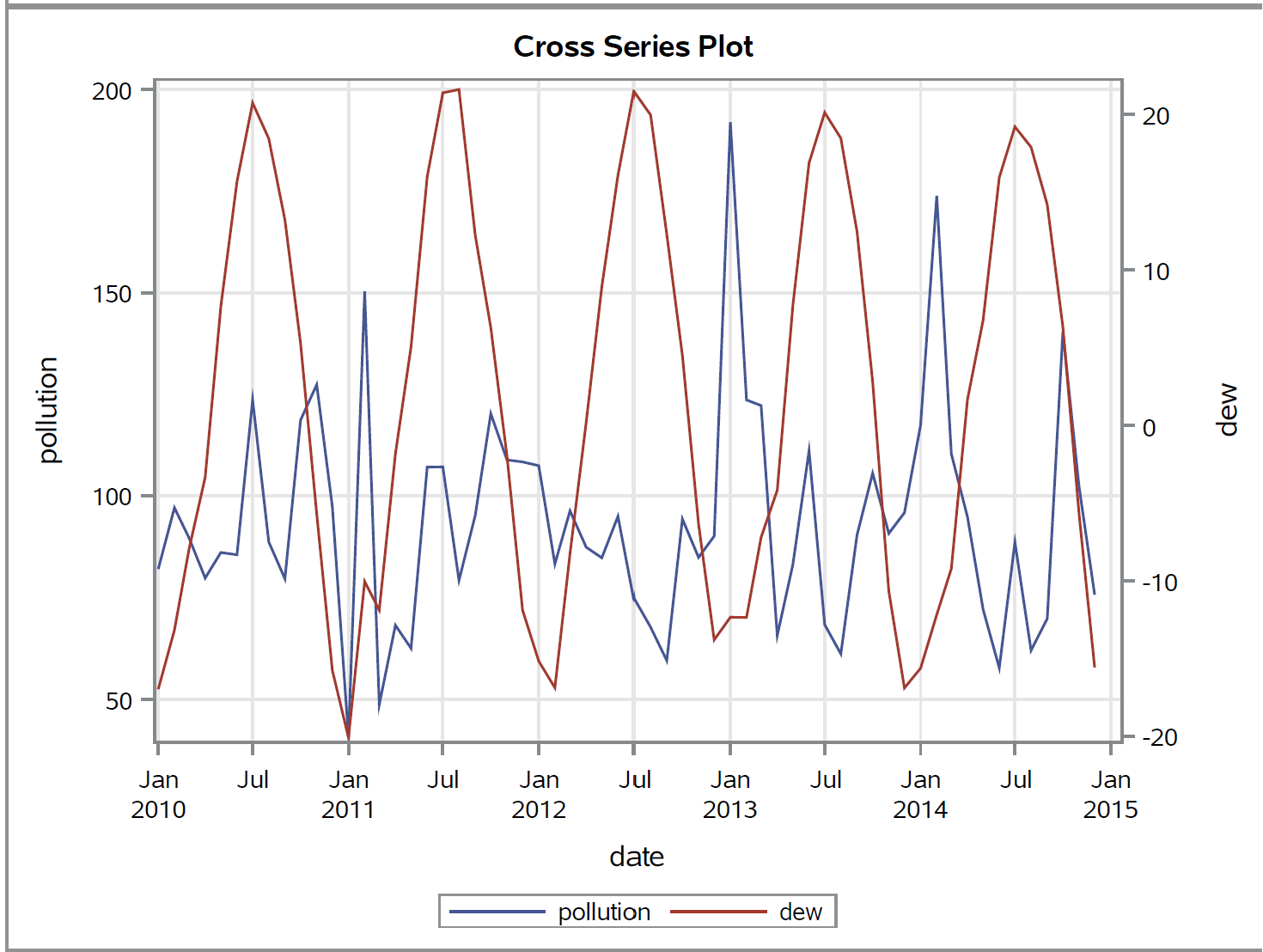


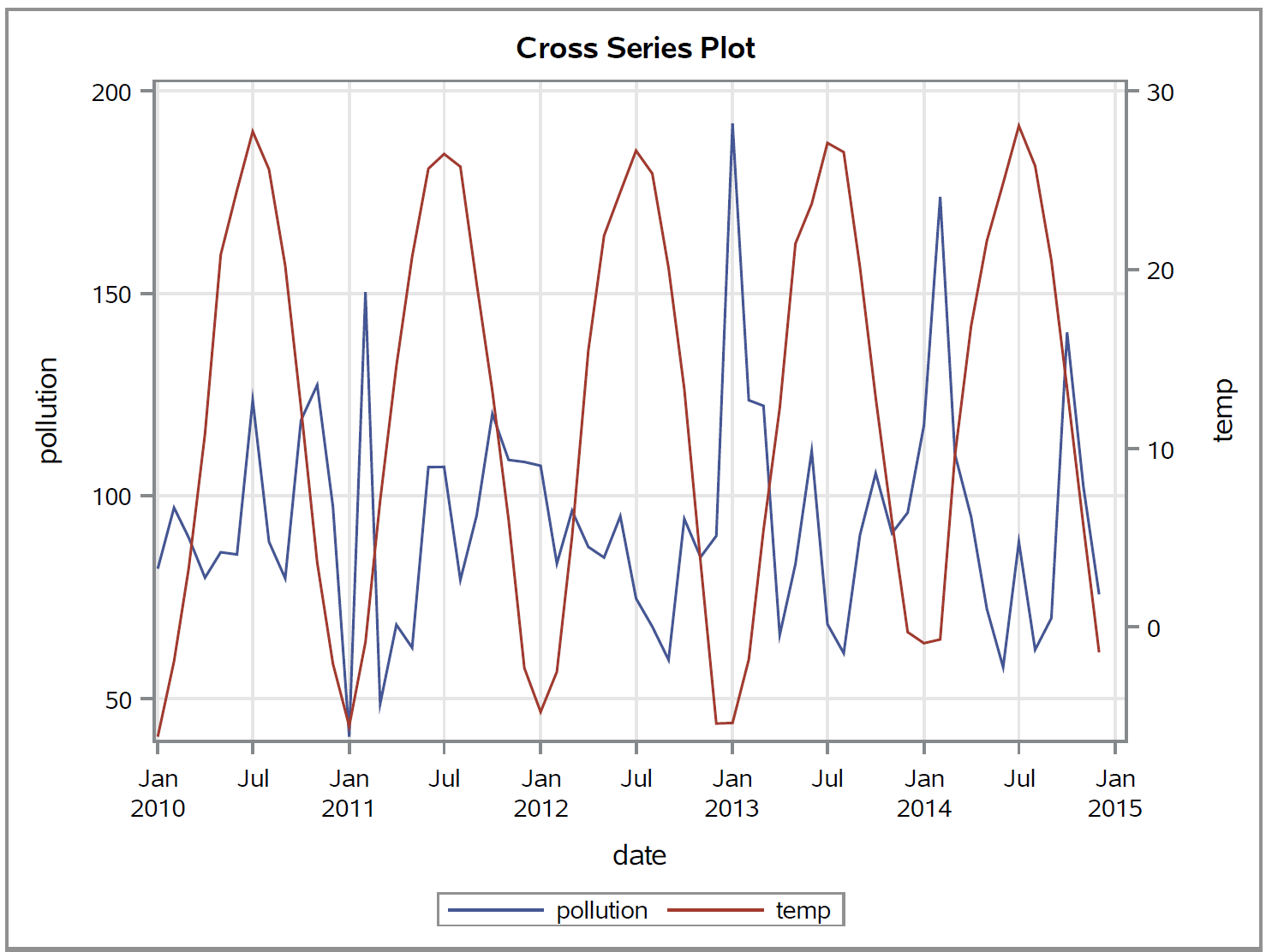
**3.4 STABILITY**

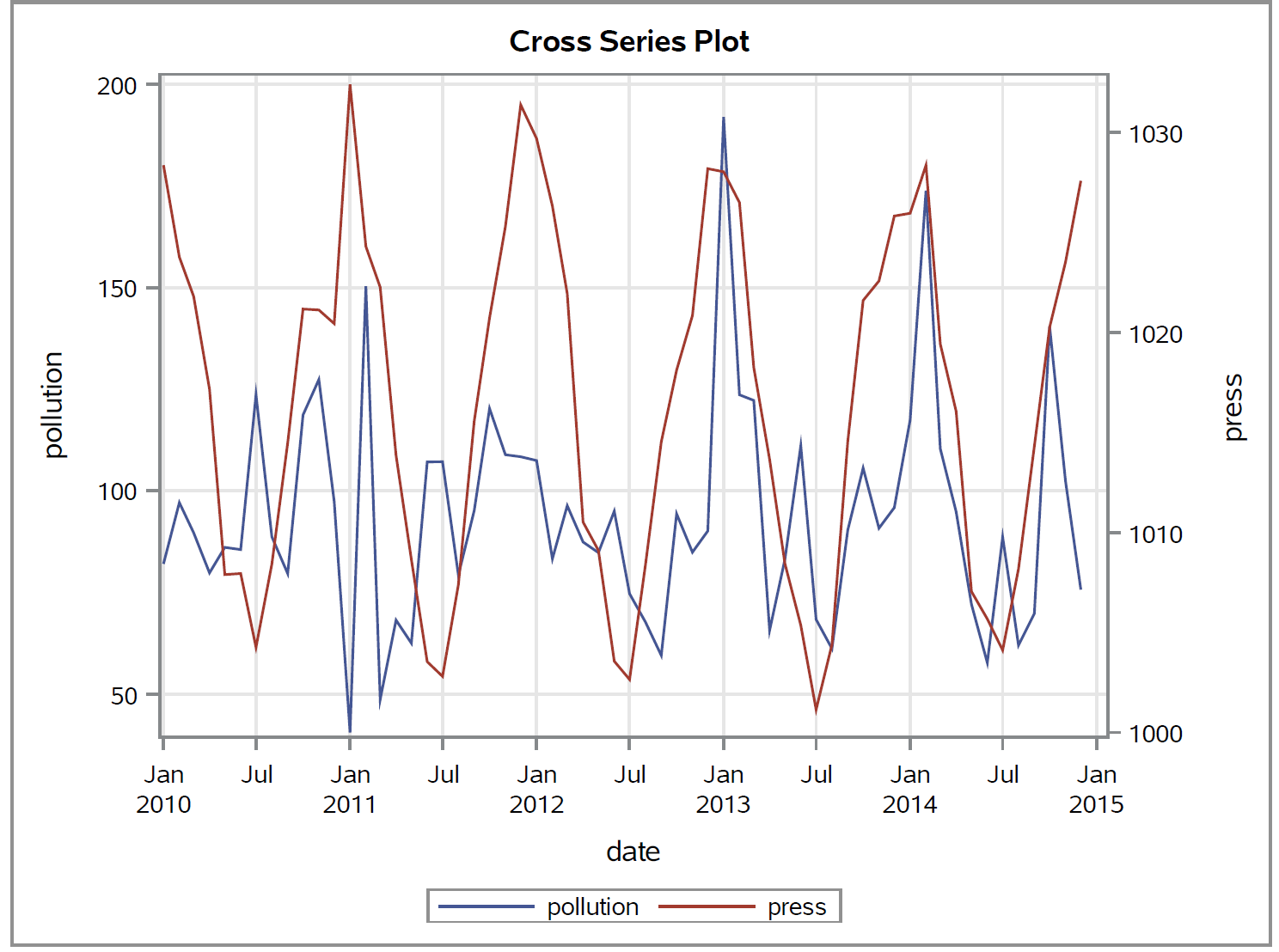
The Augmented Dickey-Fuller unit root test can be used to determine whether or not the data is stationary. P-values are less than 0.05 from the ADF test, indicating strong evidence to reject the null hypothesis of a unit root, confirming stationarity in the time series data after differencing. Here, we see this data is a stationary time series.

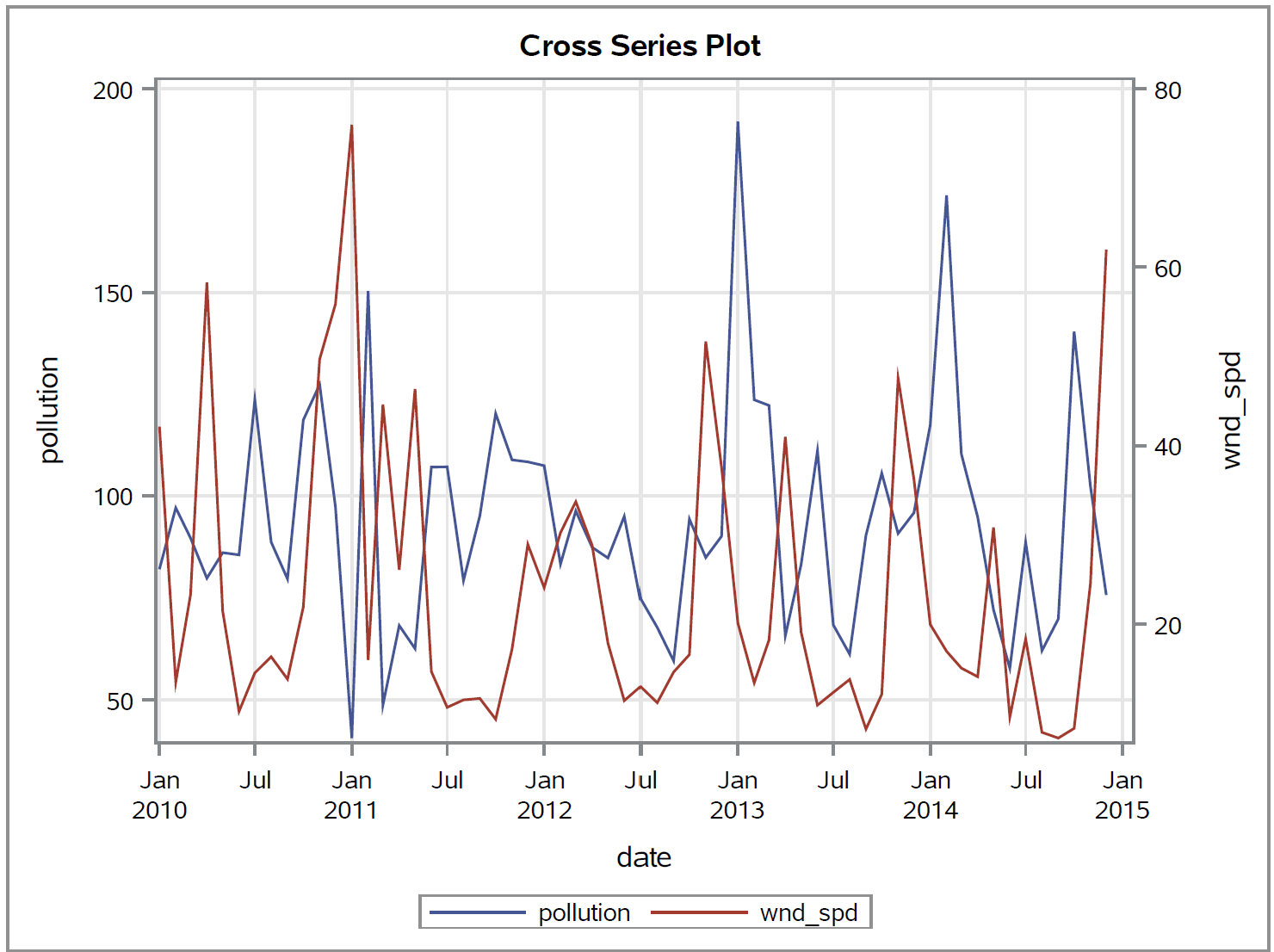
****

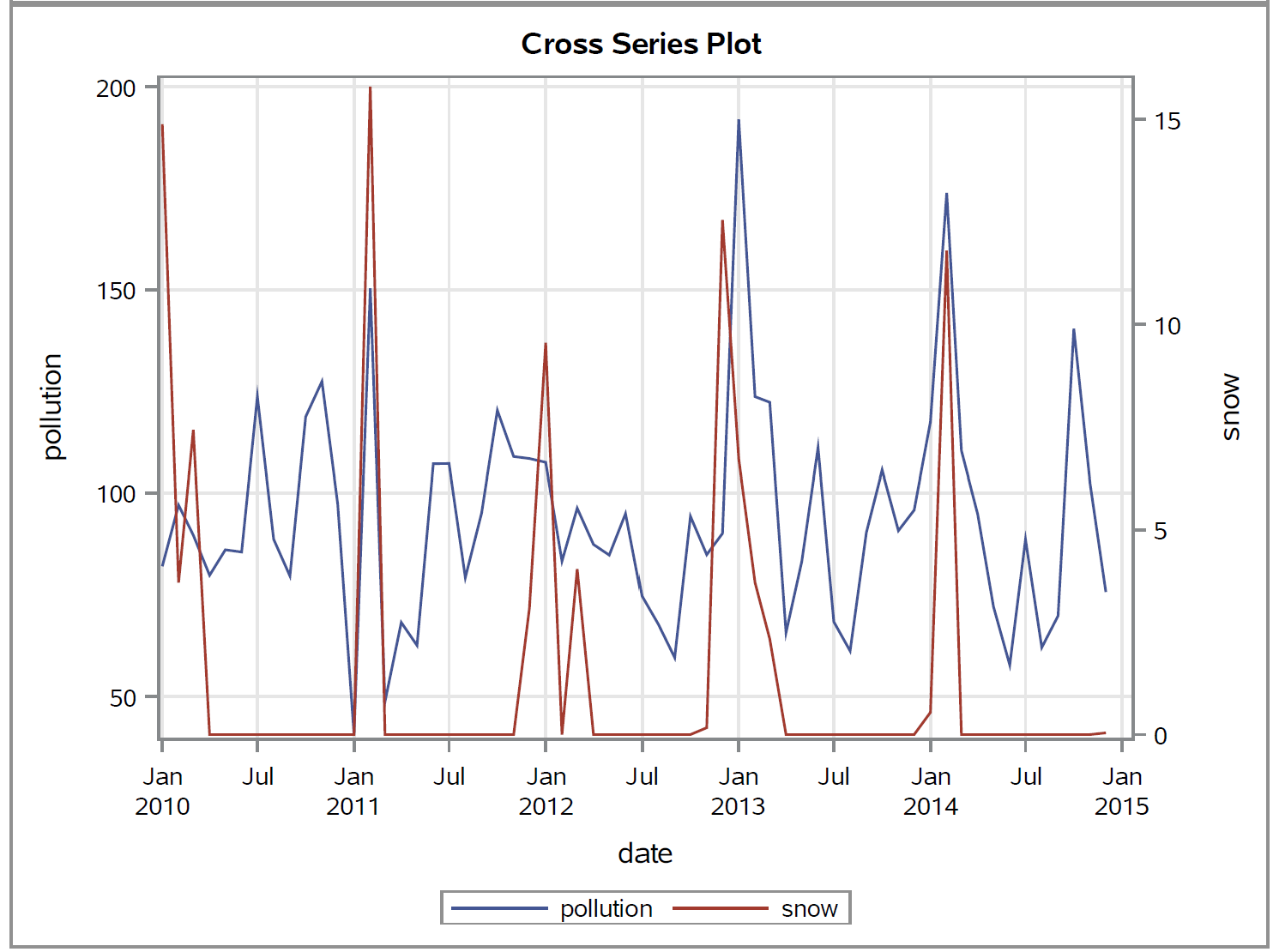
**3.5 CROSS-CORRELATION**

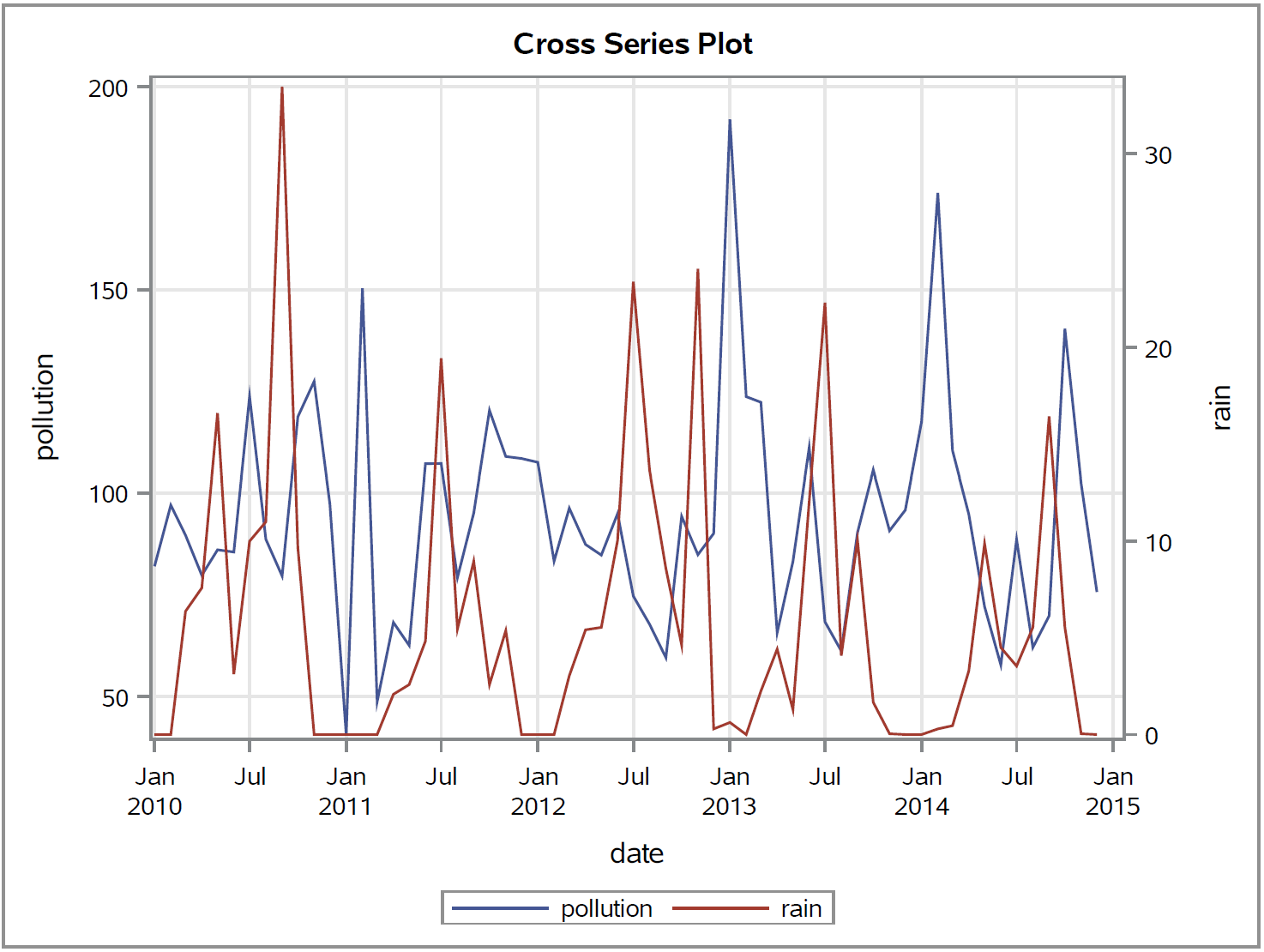
****

****

****

****

****

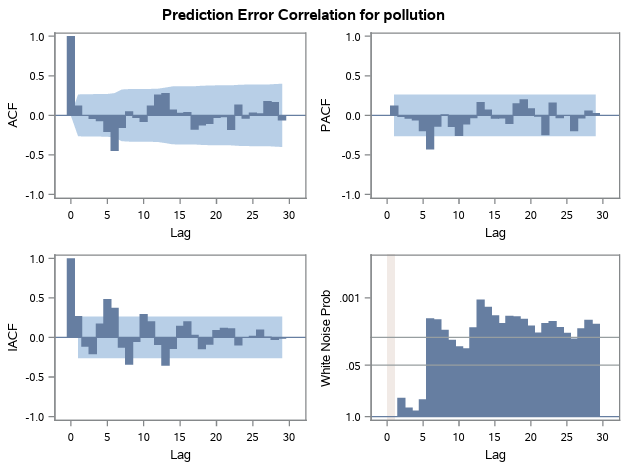
****

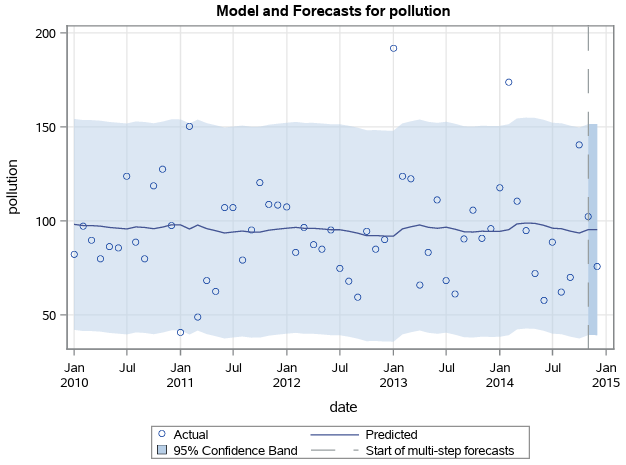
**4.** **MODEL DEVELOPMENT**

With seasonality, instability, and no trend, our team began building Seasonal Exponential Smoothing and ARIMA models.

**4.1 SIMPLE EXPONENTIAL SMOOTHING**

Although we did not see a trend amongst the data, we want to rule out models that would otherwise be catered toward data with a trend and no seasonality.

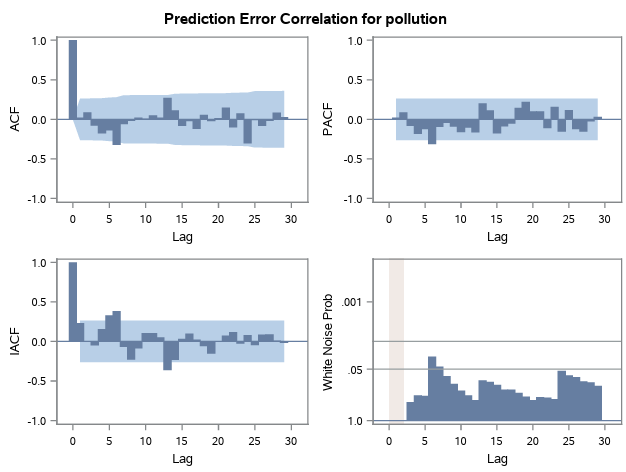


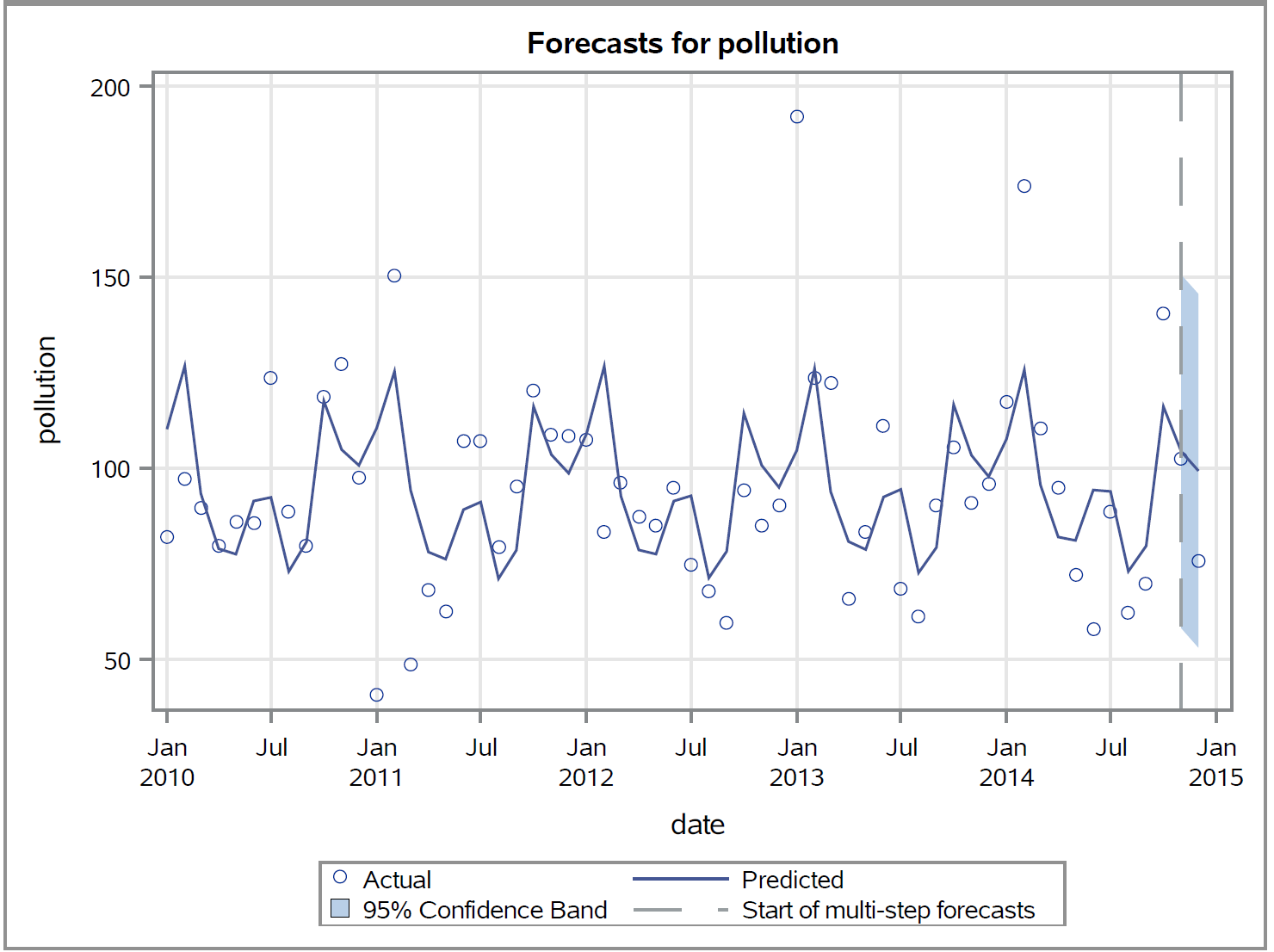


As expected, this model will not be a good forecasting model for our dataset, as demonstrated by the White Noise chart. Because of this, we continued exploring other models.

**4.1.1 ADDITIVE SEASONAL EXPONENTIAL**

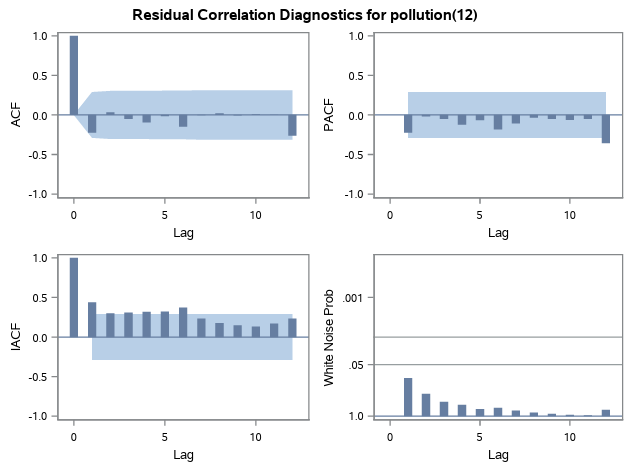
We see a lag(1) with no other significant lags. There seems to be some variation not caused by White Noise at around lag(6), although this may be due to happenstance. We are comfortable with the output of this quality model.

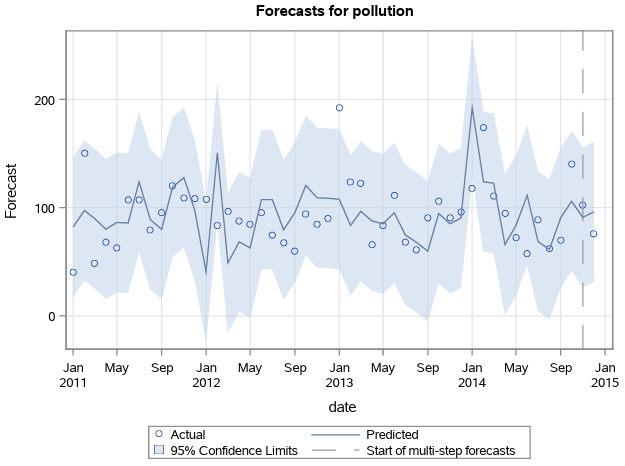
**

**

**4.2 ARIMA (0,0,0)(0,1,0)s**

Our team ran several permutations of p and q for ARIMA and used these insights to adjust our model and discovered the following residual graphs:

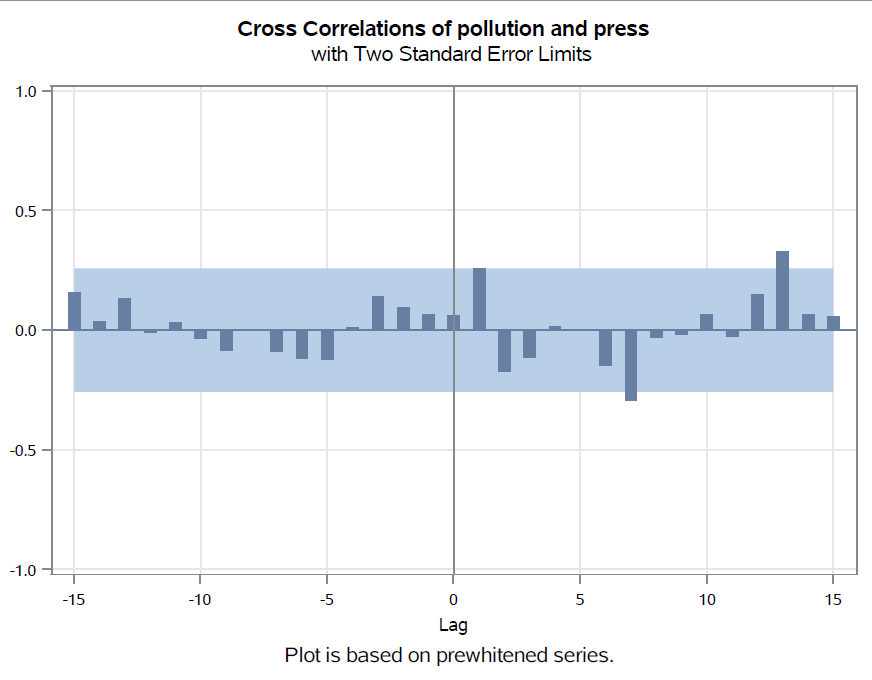
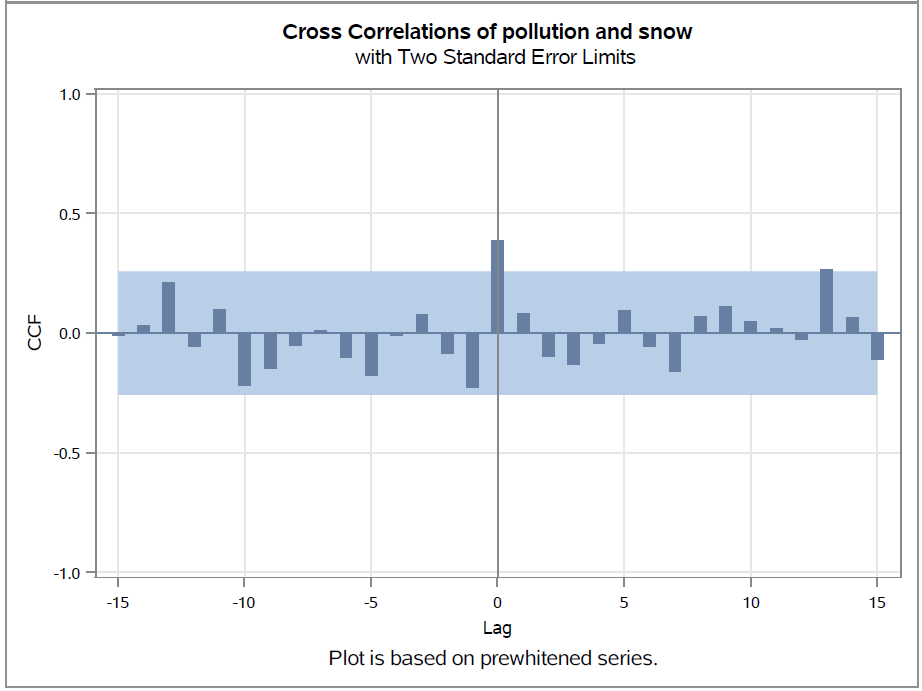
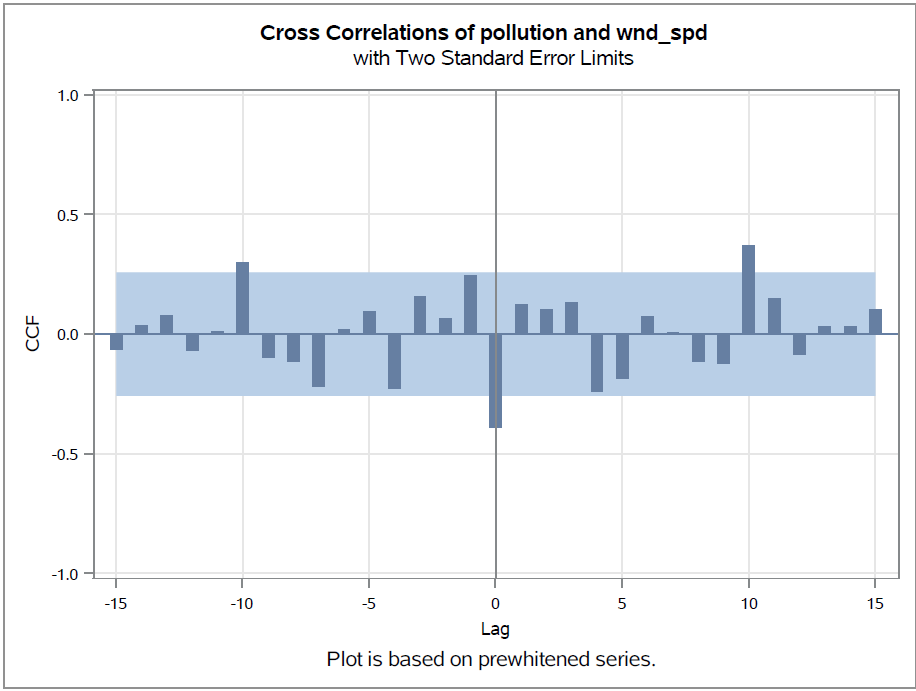
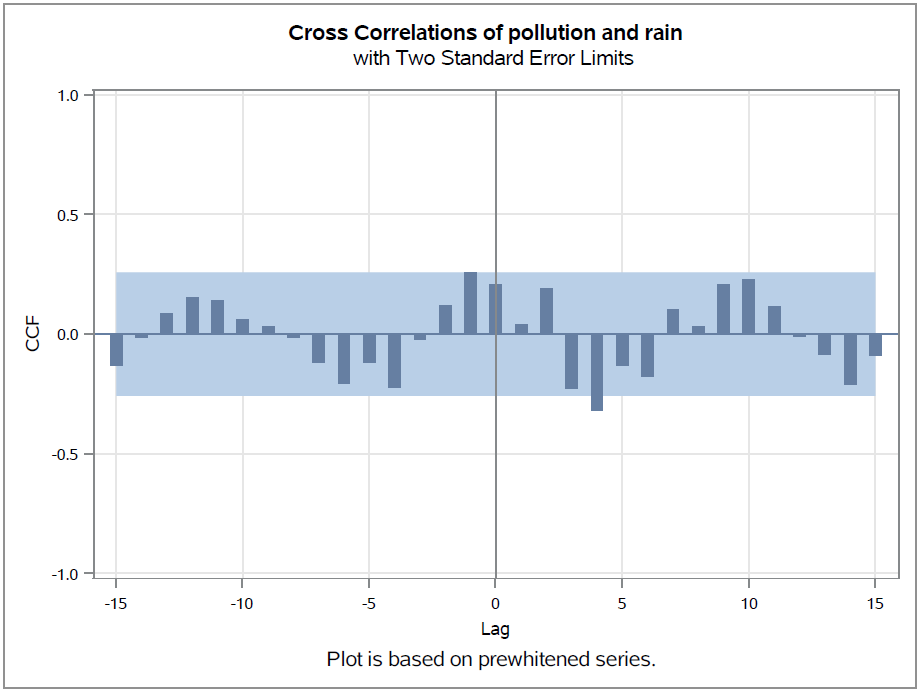




Overall, we see a lag(1) with no other significant lags. Additionally, seeing this model pass our White Noise Test is pleasing.

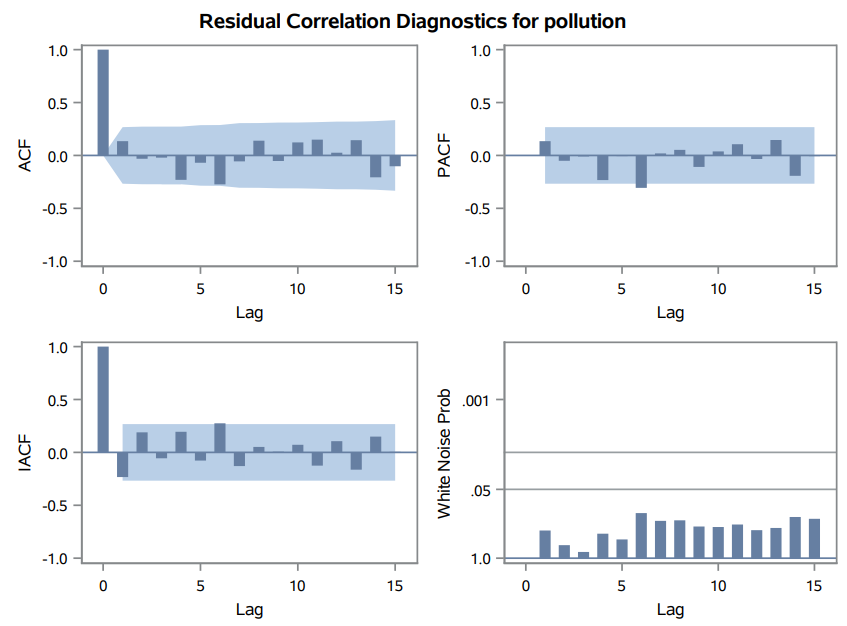
**4.3 PREWHITENING**

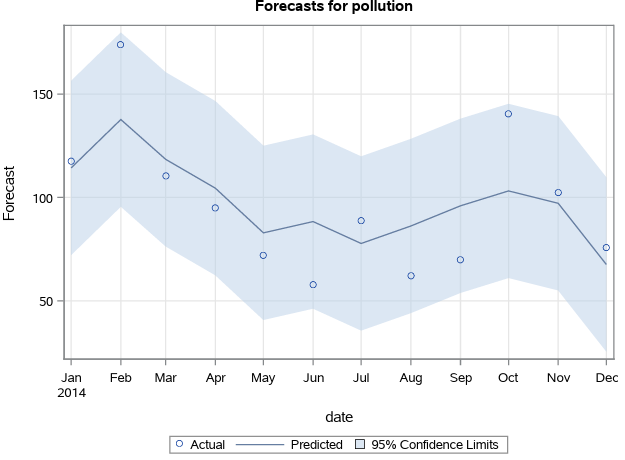
Each independent variable is a time series of its own, so we performed pre-whitening for all independent variables.

This generated stability for our data to be applied to an ARIMAX Model.

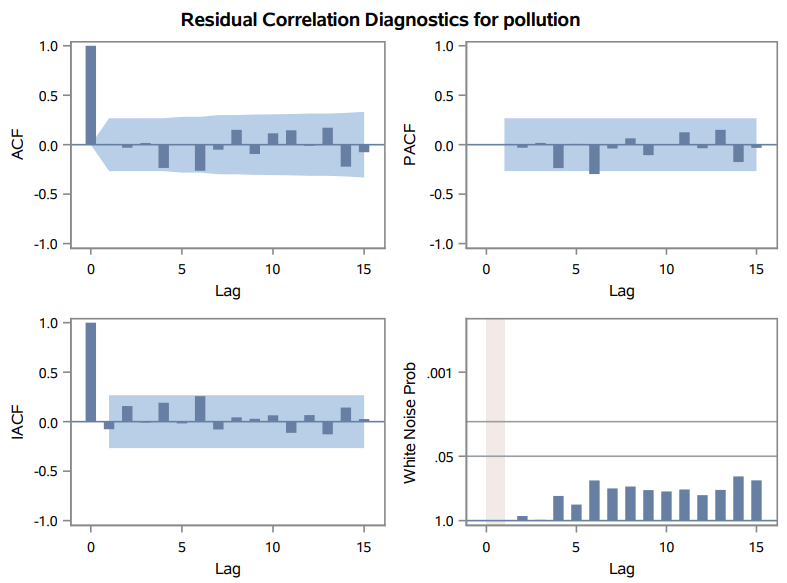
**4.4 ARIMAX(0,0,0)**

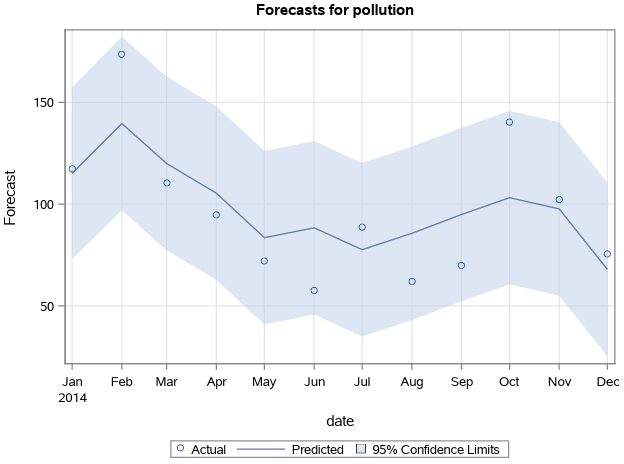
****

****

From the above results, we can see the residuals are less auto-correlated and passed the White Noise Test.

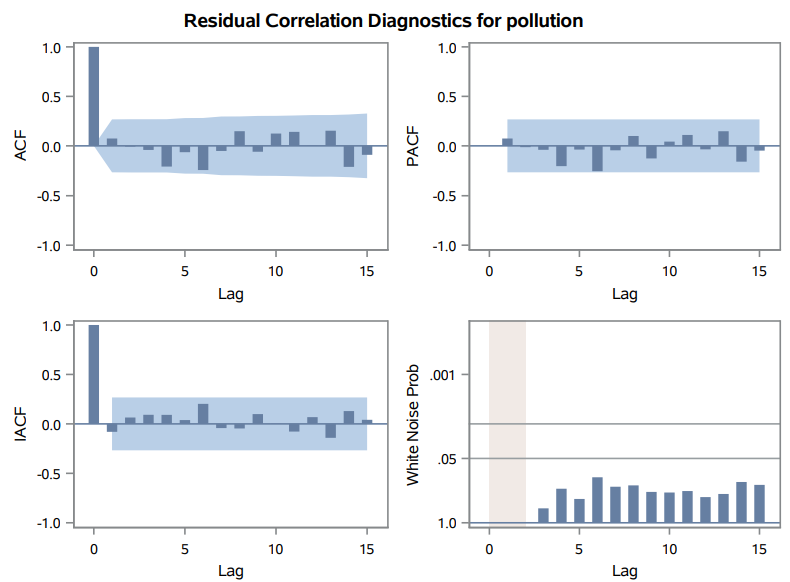
**4.5 ARIMAX(1,0,0)**

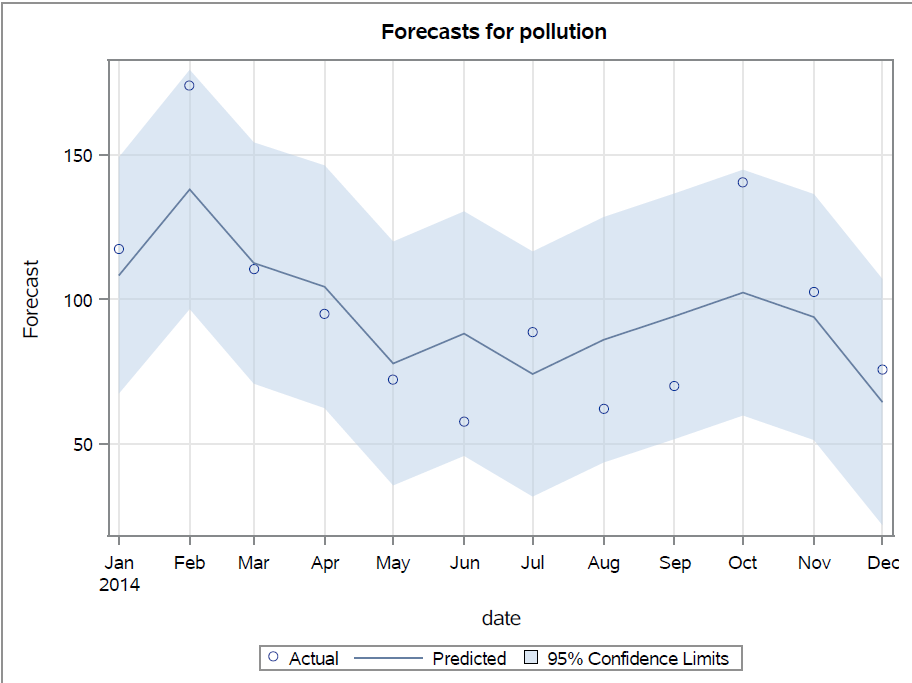
****

****

The above model residuals don't show much correlation and the White Noise is passed.

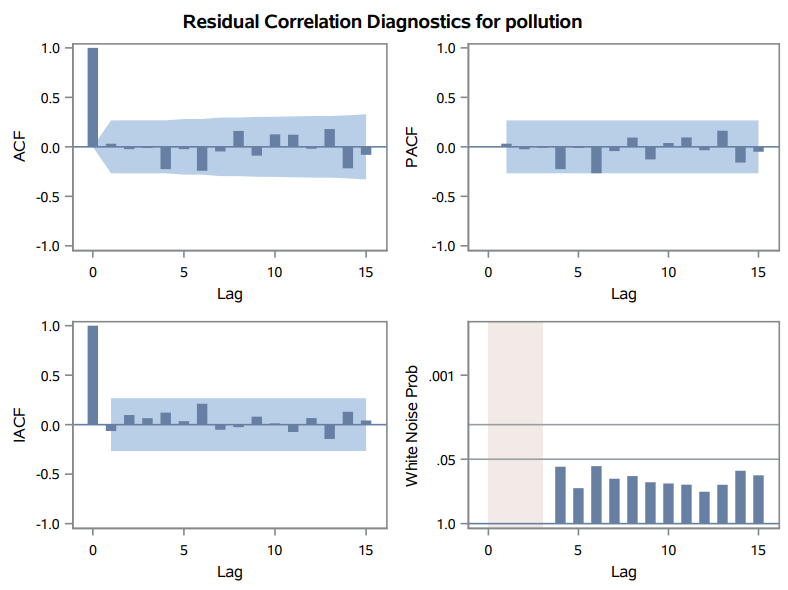
**4.6 ARIMAX(1,0,1)**

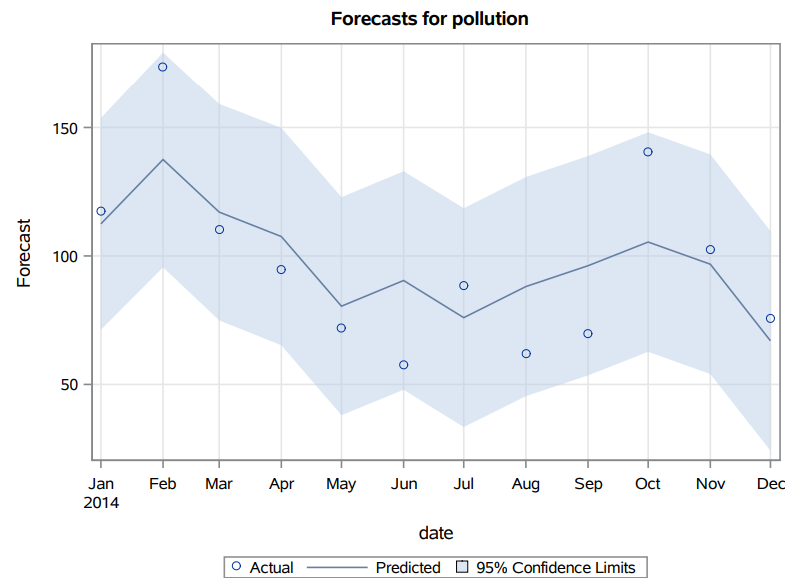
****

****

This model also passed White Noise and the residuals don’t seem much correlated.

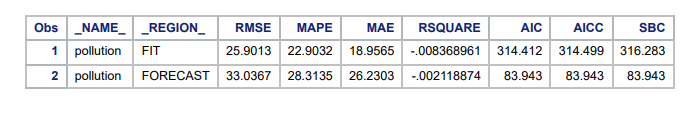
**4.7 ARIMAX(2,0,1)**

****

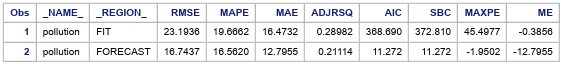
****

**5.** **MODEL EVALUATION**

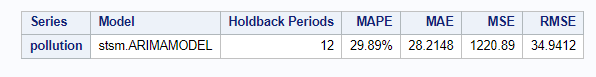
**5.1 SIMPLE EXPONENTIAL SMOOTHING**

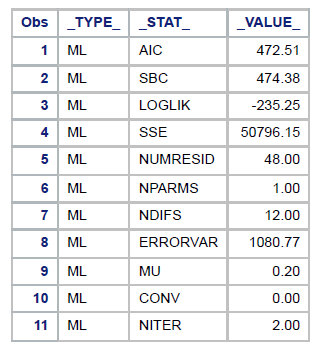
****

**5.1.1 ADDITIVE SEASONAL EXPONENTIAL SMOOTHING**

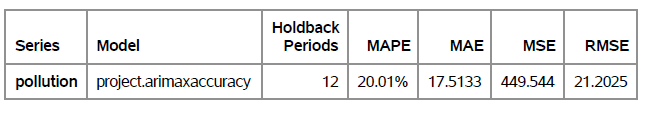
****

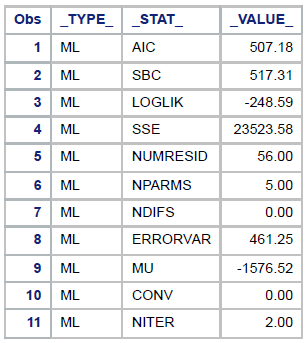
**5.2 ARIMA (0,0,0)(0,1,0)s**

****

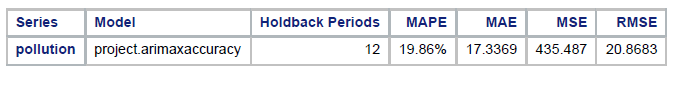
****

**5.3 ARIMAX (0,0,0)**

****

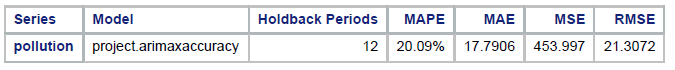
****

**5.4 ARIMAX (1,0,0):**

****

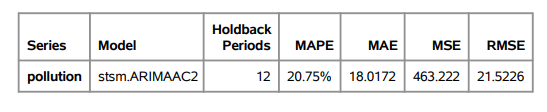
****

**5.5 ARIMAX (1,0,1):**

****

****

**5.6 ARIMAX (2,0,1)**

****

****

**Model Comparison:** As outlined above, ARIMAX(1,0,0) seems to be our most promising of the models evaluated. This is further validated by comparing accuracy measures across the models:

| **Model** | **AIC** | **SBC** | **MAPE** | **ACCURACY** |
| --- | --- | --- | --- | --- |
| **ARIMA(0,0,0)(0,1,0)S** | **472.51** | **474.38** | **29.89%** | **70.11%** |
| **ARIMAX(0,0,0)** | **507.18** | **517.31** | **20.01%** | **79.99%** |
| **ARIMAX(1,0,0)** | **508.05** | **520.20** | **19.86%** | **80.14%** |
| **ARIMAX(1,0,1)** | **507.27** | **521.44** | **20.09%** | **79.91%** |
| **ARIMAX(2,0,1)** | **508.97** | **525.18** | **20.75%** | **79.25%** |

**6.** **CONCLUSION**

While nature is often out of human control, there are likely factors of human behavior that can increase or mitigate the risk of air pollution. After forecasting Chinese air pollution using the ARIMAX(1,0,0) model, our team will work closely with policymakers to recommend increasing anti-pollution campaigns leading up to Q1 of 2015 and October 2015. This should help the population become more mindful of behavioral changes they can make to help decrease these trends.

Our team recommends focusing on educational opportunities in the months leading up to these anticipated spikes in air pollution to spread awareness of green living. In Q1 and October, this education may shift to helping the population understand how to monitor their own usage and respond appropriately to maintain a greener lifestyle. Policymakers may opt to partner with our team in the future to perform a similar analysis and understand whether these actions have positively influenced the air pollution concentration in China.

**7. LESSONS LEARNED**

While our team was first analyzing the data, we found a correlation amongst pollution and the other variables. However, while trying to fit models, it became frustrating that nearly every model had significant lags as far out as lag(6). After speaking with the Professor, we realized our initial exploration had grouped data by month, whereas our models were built on the raw hourly data. We realized a much stronger model output after cleaning and accounting for our data appropriately.

**8.** **REFERENCES**

Air pollution. (n.d.). Www.who.int. <https://www.who.int/china/health-topics/air-pollution#:~:text=Air%20pollution%20is%20responsible%20for%20about%202%20million>

Air Pollution Forecasting - LSTM Multivariate. (n.d.). Www.kaggle.com. Retrieved February 18, 2024 <https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate/data>

US EPA,OAR. (2016, July 19). Particulate Matter (PM2.5) Trends | US EPA. US EPA. <https://www.epa.gov/air-trends/particulate-matter-pm25-trends>

<https://github.com/divyamjalota25/Air-Pollution-Forecasting-for-China.git>