George Washington University

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Time series Analysis and Modeling
DATS 6313: Fall 2023
Final Term Project Report
Air Pollution Forecasting

Submited By: Anushka Vuppala
Submited To: Prof. Reza Jafari

Table of Contents

Table of Figures and Tables

Abstract
This time series project revolves around the prediction of air pollution levels using an extensive Air Quality dataset. The dataset spans a five-year period, encompassing hourly reports on both weather conditions and pollution levels at the US embassy in Beijing, China. Our primary objective is to leverage this rich dataset to formulate a forecasting problem. Specifically, we aim to predict pollution levels in the subsequent hour based on historical weather conditions and pollution data from prior hours. The project focuses on developing and evaluating forecasting models that make use of the hourly features available to us for predicting pollution levels, thereby contributing to a better understanding of air quality dynamics in this urban setting.

Introduction:
In this study, we employ various air quality index metrics spanning a 5-year period, collected at hourly intervals, to estimate air pollution levels in Beijing, China. Our methodology commences with data cleaning, followed by a crucial assessment of dataset stationarity, a fundamental consideration in any time series analysis. In instances where the data is found to be non-stationary, we will proceed with detrending and deseasonalizing procedures to render it stationary. Once stationarity is achieved, we proceed with feature selection, eliminating variables that do not contribute significantly to the predictive model. Subsequently, we employ a range of forecasting models including linear regression, ARIMA, SARIMA, and LSTM models to identify the most suitable predictive model.

This dataset is from the Kaggle Source: https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate

Description of Dataset:
The dataset downloaded from Kaggle has no missing value for any column. There are a total of 43,800 rows in the dataframe. A brief description of the all the variables are tabulated below:

Feature	Description
pollution	PM2.5 concentration in the air (target variable)
dew	Dew point - temperature at which the air is saturated with moisture (in °C)
temp	Current temperature (in °C)
press	Atmospheric pressure (in millibar)
wnd_dir	Wind direction – categorical variable with 4 values: NE, NW, SE, CV (calm and variable)
wnd_speed	Cumulated wind speed (in mph)
Snow	Snow depth precipitation (in cm)
rain	Precipitation measurement (in mm)

Table 1: Dataset Description

An initial time series plot was generated to depict the relationship between the air pollution level (considered as the dependent variable) and time. It is discernible, upon visual inspection, that a distinct seasonality is evident on an annual basis. Notably, the zenith of pollution levels is consistently observed toward the conclusion of each year. This observation aligns logically with the heightened use of automobiles, as opposed to bicycles, during the winter season, contributing

to an escalation in pollution levels. Furthermore, an uptick in the utilization of indoor heating systems in buildings and homes is identified as an additional factor influencing the observed pattern.

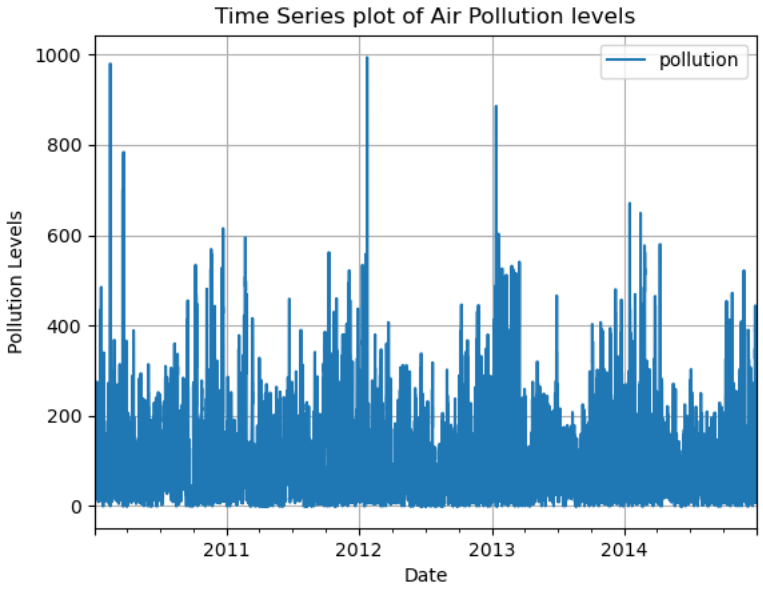


Figure 1: Time Series Plot of Air Pollution Levels

An autocorrelation function (ACF) plot was generated with a lag parameter set to 100. Upon examination, a gradual diminishing trend is evident, suggesting a decline in autocorrelation as the lag increases. Additionally, distinctive seasonal patterns are noticeable in the dataset, manifested as recurring bumps in the ACF plot. This phenomenon serves to further affirm the presence of seasonality within the data. The observed seasonal bumps in the ACF plot reinforce the notion that certain time intervals exhibit recurring patterns, highlighting the significance of seasonality in the underlying time series.

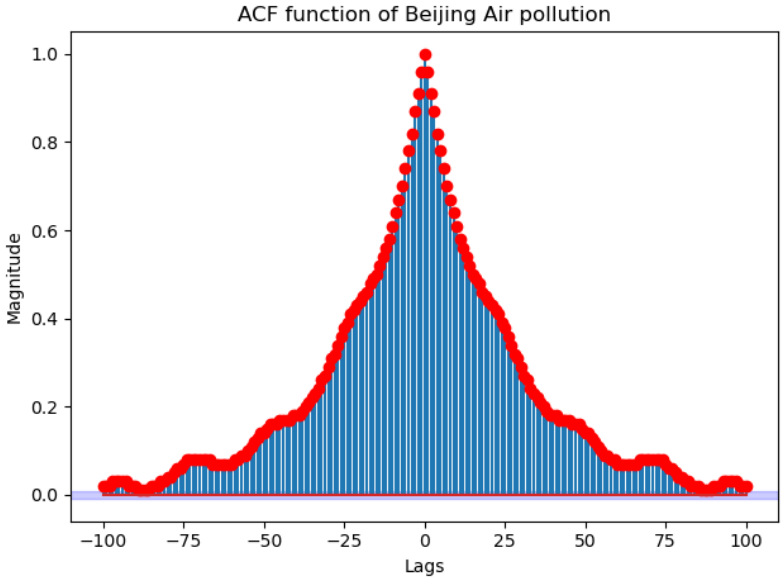


Figure 2: ACF plot of Air Pollution Levels

A comprehensive correlation matrix was also generated, offering valuable insights into the relationships among various meteorological variables. Noteworthy findings include

- a negative correlation between dew and pressure, suggesting that as the dew point increases, atmospheric pressure tends to decrease.
- a negative correlation was evident between temperature and pressure, indicating that rising temperatures coincide with a decrease in atmospheric pressure.
- a positive correlation was observed between temperature and dew, implying that higher temperatures correspond to an elevated dew point.

These nuanced correlations highlight the intricate interplay between meteorological factors and provide an understanding of their interconnected dynamics within the dataset.

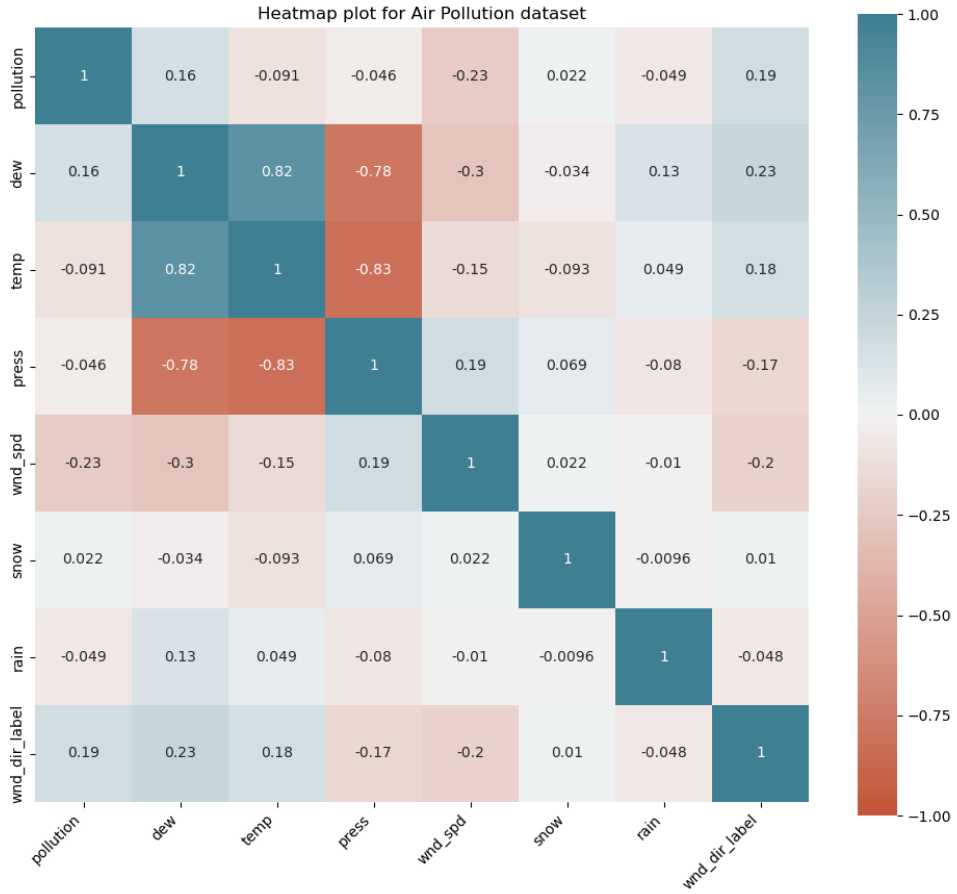
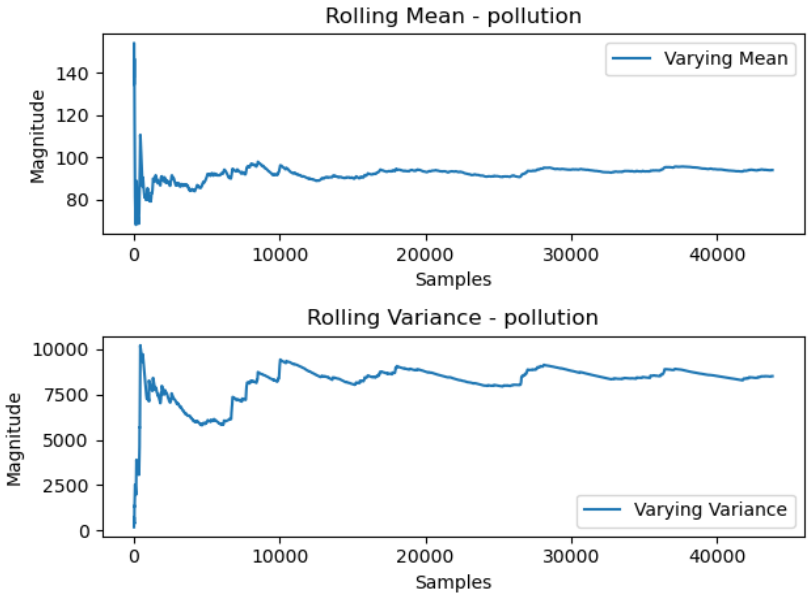


Figure 3: Correlation Matrix of Air Pollution Forecasting Features

Stationarity

The examination of the rolling mean and variance plots for the pollution variable, considered as the target variable, reveals apparent non-stationarity, evident in the interruptions observed in the rolling variance plot.

Notably, this observation persists despite the Augmented Dickey-Fuller (ADF) test indicating a p-value very close to zero and a test statistic falling below all critical values showing stationarity.

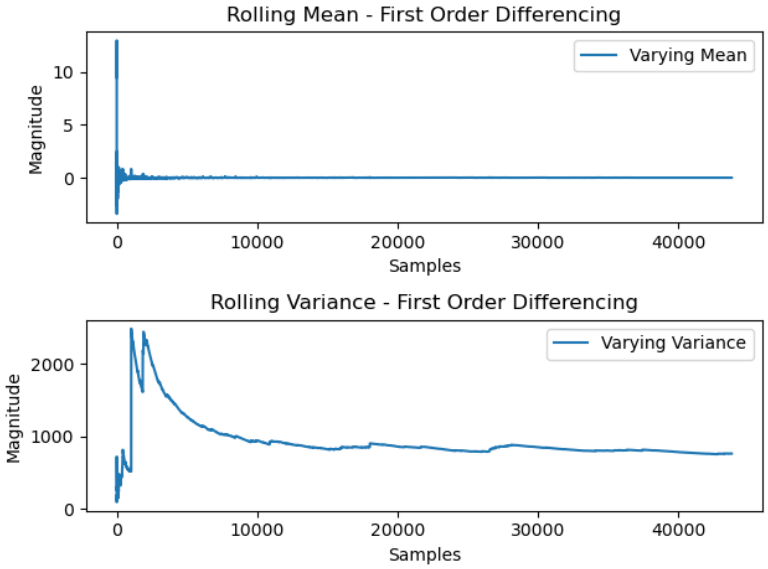



```
ADF Test for Beijing Air Pollution dataset:
ADF Statistic: -21.004109
p-value: 0.000000
Critical Values:
    1%: -3.430
    5%: -2.862
    10%: -2.567

KPSS Test for Beijing Air Pollution dataset:
Results of KPSS Test:
Test Statistic            0.078133
p-value                   0.100000
Lags Used               114.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

Figure 3: Non Stationarity of Dataset

Consequently, to address the non-stationarity, a first-order differencing of the pollution variable was performed. Subsequent to this differencing, the ADF test statistic further diminishes, providing empirical confirmation of the attained stationarity. Furthermore, even after the first-order differencing, the rolling variance exhibits a more stabilized and constant profile, thereby corroborating stationarity visually. These analytical steps collectively contribute to a refined understanding of the temporal dynamics of the pollution variable and validate the achieved stationarity through both statistical and visual assessments. Thus going forward, the first order differencing value of pollution would be used in model predictions.




```
ADF Test for First Order transformation:
ADF Statistic: -36.858231
p-value: 0.000000
Critical Values:
    1%: -3.430
    5%: -2.862
    10%: -2.567

KPSS Test for First Order transformation:
Results of KPSS Test:
Test Statistic            0.001725
p-value                   0.100000
Lags Used               119.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

Figure 4: Stationarized Dataset Results

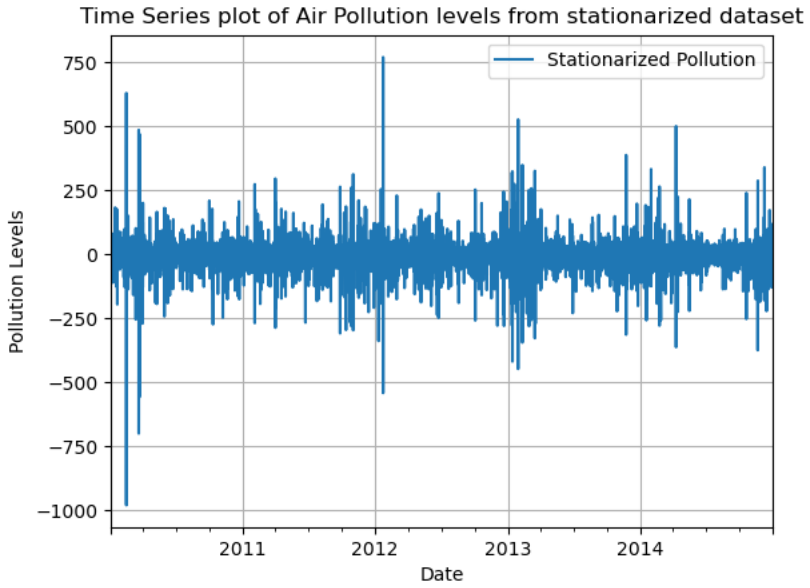


Figure 5: Time Series plot of Stationarized Dataset

Time Series Decomposition

Utilizing Seasonal and Trend decomposition using LOESS (STL), the time series data was effectively decomposed into its constituent elements, namely trend, seasonality, and residual. The choice of a period equal to 24 was informed by the recognition of a daily seasonality pattern, a phenomenon prominently evident in the autocorrelation function (ACF) plot. The ACF plot had previously indicated recurring patterns at a daily frequency, aligning with the 24-hour period, making it a judicious selection for decomposition.

Upon applying STL with a period of 24, the resulting decomposition showcased distinct components for trend, seasonality, and residual. Visual representation of these components was conducted through comprehensive plotting, both collectively and selectively for the initial 100 values, facilitating a more detailed inspection of their individual characteristics.

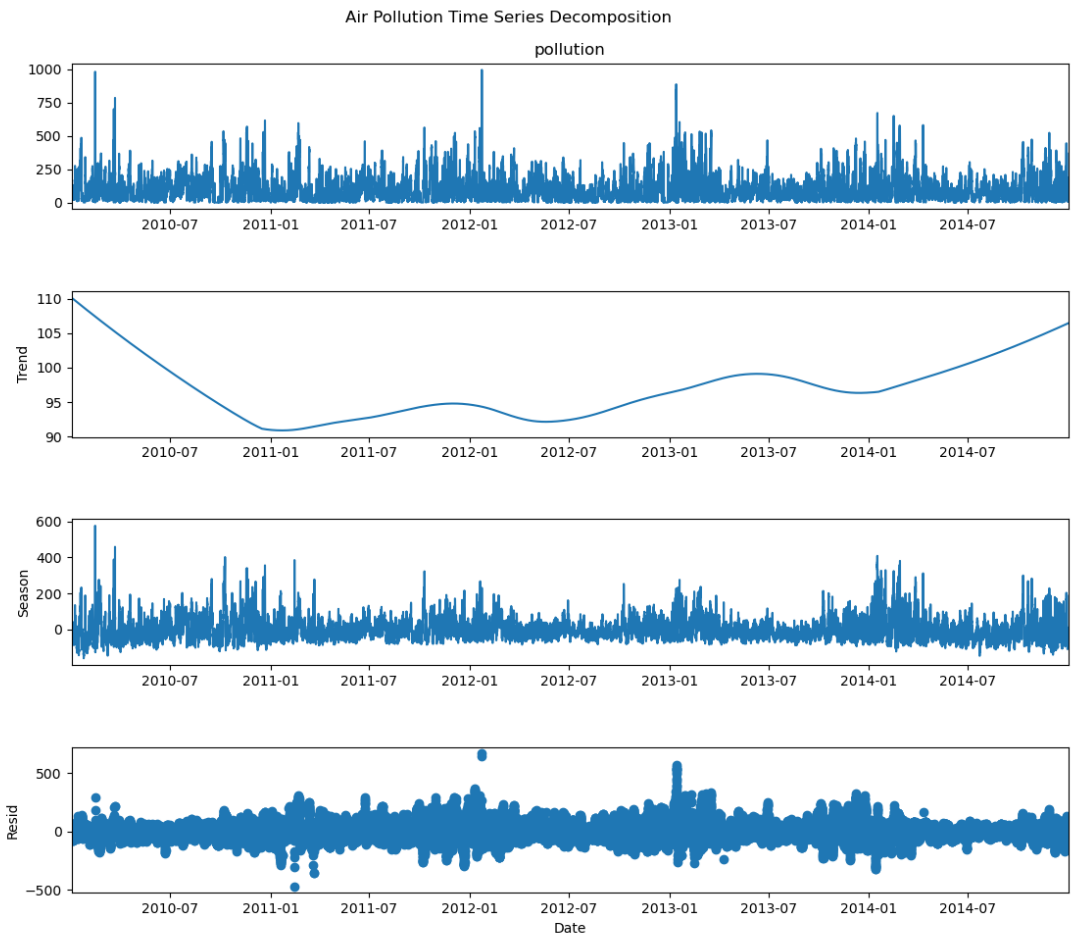


Figure 6: Time Series Decomposition into Trend, Seasonal, Residual

We also plot all 3 splits onto 1 graph to compare the scales of all 3:

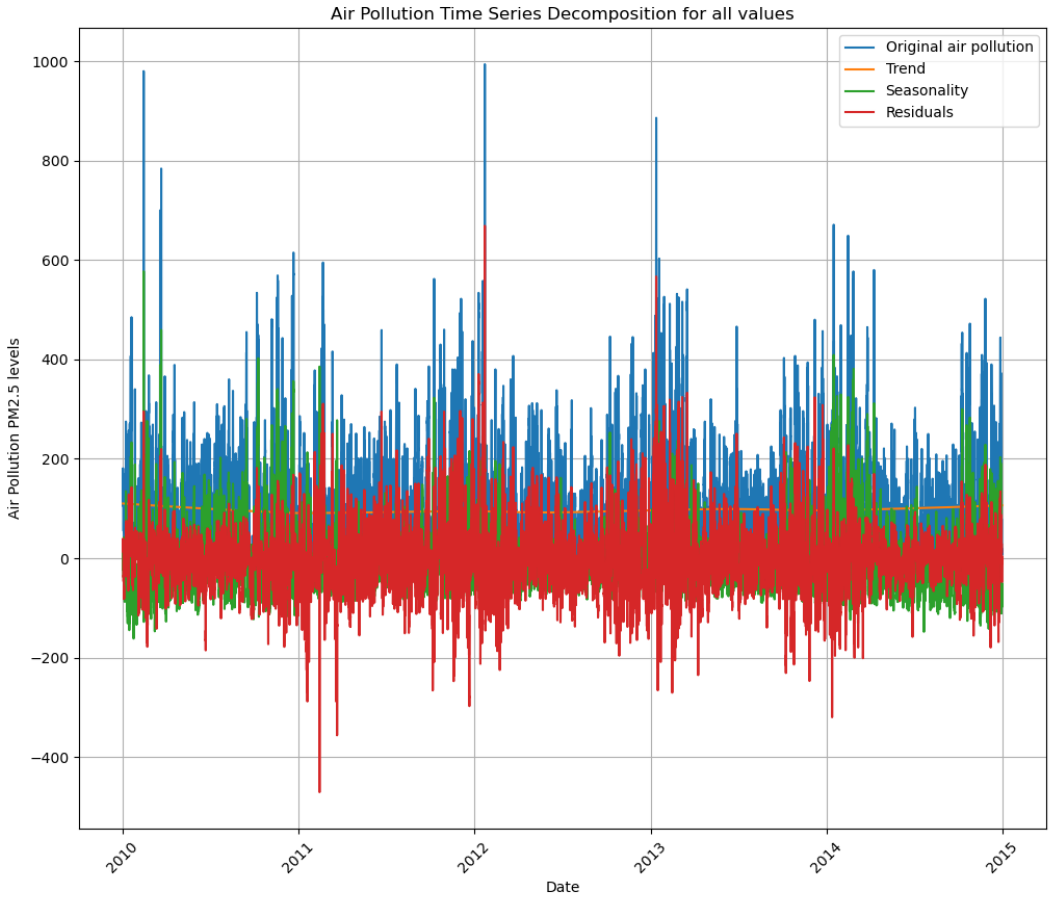


Figure 11: Time Series Decomposition of Trend, Seasonal, Residual in one plot

Since the above plot reveals less information, for closer inspection, we plot the these 4 splits onto a single plot but only the first 50 values.

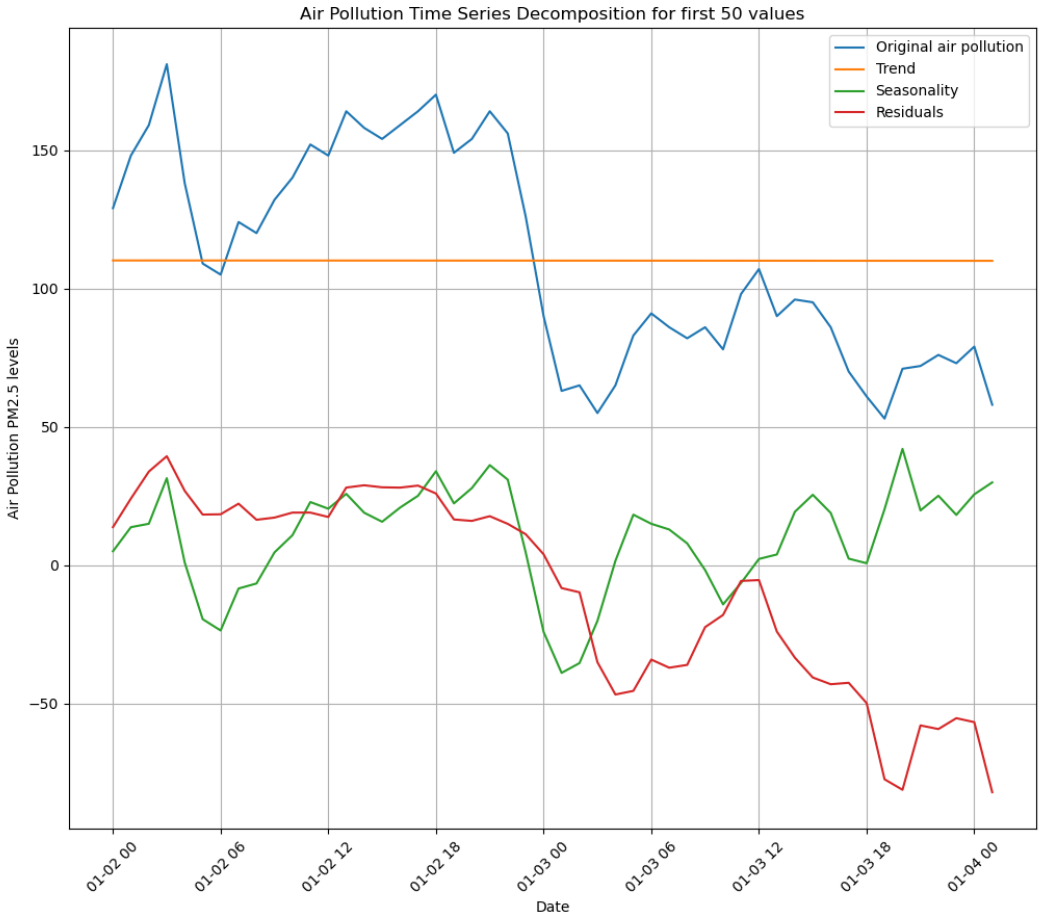


Figure 12: Time Series Decomposition: a closer look with first 50 values

```
The strength of trend for the air pollution levels is = 0.776%
The strength of Seasonality for the air pollution levels is = 50.151%
```

The time series analysis of air pollution levels reveals a modest upward trend of 0.776%, suggesting a gradual increase over the observed period. However, the more prominent feature is the substantial seasonality with a strength of 50.151%. This indicates significant recurring patterns in air pollution, likely influenced by cyclic factors such as weather changes, industrial activities, or other periodic events. Understanding the nature of this seasonality, whether it follows a daily, monthly, or yearly cycle, is crucial for effective pollution management and intervention strategies. Further exploration of the data, considering its statistical significance and the length of the time series, would provide a more comprehensive understanding of the dynamics influencing air quality over time.

Feature Selection

```
Singular values = [98433.31460546 38656.43968767 34889.13559701 33298.81122023
 27616.05390007  7489.03247924  4897.21251033]
Condition number is 4.48
```

Singular values: These are the singular values obtained during the Singular Value Decomposition (SVD) of the design matrix. The singular values give information about the spread or scale of the variables in your dataset. Larger singular values indicate greater variability. The singular values in descending order convey information about the relative importance of each component in the decomposition. The larger singular values correspond to more significant components, contributing more to the overall structure of the data. Conversely, smaller singular values represent less influential components.
In the above specific case, the singular values are presented in descending order, indicating their significance from left to right. The first singular value (98433.31460546) is the largest, followed by the second (38656.43968767), and so on. This pattern suggests that the first few components (associated with larger singular values) are more dominant in capturing the variability and structure of the data.

Condition number: The condition number is a measure of how well-conditioned the design matrix is. A high condition number (relative to the scale of your singular values) can indicate potential multicollinearity issues, making the matrix ill-conditioned. A lower condition number is generally better. In our case, the condition number is 4.48, which is relatively low and suggests that the design matrix is well-conditioned.

Backward Stepwise Regression:
In addition to our initial regression analysis, we also conducted a Backward Stepwise Regression and Variance Inflation Factor (VIF) removal to further refine our model. In the Backward Stepwise Regression, the remaining columns considered were 'dew,' 'temp,' 'press,' 'wnd_spd,' 'snow,' 'rain,' and 'wnd_dir_label.' Notably, no columns were eliminated during this step, indicating that all the initially included variables retained their significance in explaining the variation in the target variable.

Backward Stepwise Regression

OLS Regression Results

Dep. Variable:	y	R-squared:	0.232
Model:	OLS	Adj. R-squared:	0.232
Method:	Least Squares	F-statistic:	1515.
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	0.00
Time:	16:08:51	Log-Likelihood:	-2.0349e+05
No. Observations:	35040	AIC:	4.070e+05
Df Residuals:	35032	BIC:	4.071e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	93.3594	0.430	217.000	0.000	92.516	94.203
dew	57.3654	0.867	66.141	0.000	55.665	59.065
temp	-72.5054	0.925	-78.417	0.000	-74.318	-70.693
press	-18.8625	0.797	-23.676	0.000	-20.424	-17.301
wnd_spd	-10.1865	0.464	-21.973	0.000	-11.095	-9.278
snow	-1.1356	0.434	-2.616	0.009	-1.986	-0.285
rain	-9.0733	0.438	-20.709	0.000	-9.932	-8.215
wnd_dir_label	11.2003	0.450	24.892	0.000	10.318	12.082

Omnibus:	12451.849	Durbin-Watson:	0.159
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65367.662
Skew:	1.636	Prob(JB):	0.00
Kurtosis:	8.836	Cond. No.	4.48

Results from Backward Stepwise Regression:

```
Remaining Columns:  Index(['dew', 'temp', 'press', 'wnd_spd', 'snow', 'rain', 'wnd_dir_label'], dtype='object')
Columns eliminated:  Index([], dtype='object')
```

VIF Test:

```
VIF Test
VIF data:
          feature       VIF
1            temp  4.618718
0             dew  4.064124
2           press  3.429035
3         wnd_spd  1.161093
6   wnd_dir_label  1.093777
5            rain  1.037051
4            snow  1.017999
Columns retained:  Index(['dew', 'temp', 'press', 'wnd_spd', 'snow', 'rain', 'wnd_dir_label'], dtype='object')
Columns eliminated []
```

Following this, the VIF removal process was applied, and the columns 'dew,' 'temp,' 'press,' 'wnd_spd,' 'snow,' 'rain,' and 'wnd_dir_label' were retained, with no columns being eliminated. This lack of elimination suggests that, based on the VIF criteria used, there were no significant issues of multicollinearity among the variables.

Remarkably, in both the Backward Stepwise Regression and VIF removal, no columns were eliminated in the current iteration. These results imply that, according to the specified criteria for variable selection and multicollinearity assessment, all the original columns are considered important and do not exhibit multicollinearity issues in this particular stage of the analysis. This convergence of results reinforces the stability and relevance of the selected variables within the model, substantiating their collective importance in capturing the underlying patterns in the dataset.

Linear Regression

In the context of regression analysis, the assessment of variable significance and overall model goodness-of-fit relies on the examination of p-values associated with individual coefficients and the F-test. A p-value less than 0.05 for each coefficient and a F-test p-value of 0.00 generally indicate the statistical significance of the model as a whole.

Specifically, individual coefficients with p-values below 0.05 signify their statistical significance in predicting the response variable.

The extremely low p-value (effectively 0.00) for the F-test emphasizes that at least one predictor variable is related to the response variable, providing valuable information beyond random chance. Despite not being precisely zero due to numerical precision limitations, the F-test p-value suggests a robust association between the predictors and the response variable.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.232
Model:	OLS	Adj. R-squared:	0.232
Method:	Least Squares	F-statistic:	1515.
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	0.00
Time:	16:08:51	Log-Likelihood:	-2.0349e+05
No. Observations:	35040	AIC:	4.070e+05
Df Residuals:	35032	BIC:	4.071e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	93.3594	0.430	217.000	0.000	92.516	94.203
dew	57.3654	0.867	66.141	0.000	55.665	59.065
temp	-72.5054	0.925	-78.417	0.000	-74.318	-70.693
press	-18.8625	0.797	-23.676	0.000	-20.424	-17.301
wnd_spd	-10.1865	0.464	-21.973	0.000	-11.095	-9.278
snow	-1.1356	0.434	-2.616	0.009	-1.986	-0.285
rain	-9.0733	0.438	-20.709	0.000	-9.932	-8.215
wnd_dir_label	11.2003	0.450	24.892	0.000	10.318	12.082

Omnibus:	12451.849	Durbin-Watson:	0.159
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65367.662
Skew:	1.636	Prob(JB):	0.00
Kurtosis:	8.836	Cond. No.	4.48

However, the low adjusted r-squared (=0.232) suggests that not all training data information was captured by the model. This can also be proved from the analysis performed in the Time Series decomposition where we concluded that the strength of trend is very low (almost nearing zero). And since linear regression models captures the linear trend (which is absent in this dataset), the adjusted r-squared is very low. The examination of the Autocorrelation Function (ACF) of Prediction Errors is also essential to identify any remaining structure or patterns in errors. Ideally, a flat ACF signifies uncorrelated and random errors. However, we see that autocorrelation is observed. And there are spikes in the plot signifying that seasonality is not taken into account by the model when training.

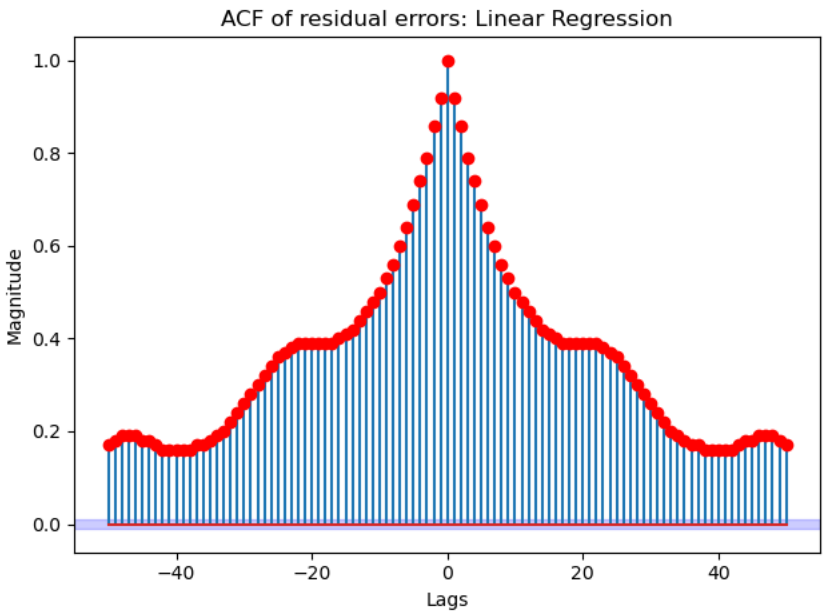


Figure 13: ACF of Linear Regression residual errors

Moving beyond significance, the Mean Squared Error (MSE) values were computed, revealing a MSE of 6484.26 on the training set and 6832.14 on the test set. A higher MSE on the test set compared to the training set may indicate potential overfitting, emphasizing the importance of balancing model complexity for training data performance and generalization to new, unseen data.

Turning attention to the Ljung-Box test, the calculated Q-value of 104960.57 exceeds the critical value of 66.21, leading to the rejection of the null hypothesis of no autocorrelation in residuals. This result indicates significant autocorrelation in the residuals, suggesting that the model might not fully capture temporal dependencies or patterns in the data.

Further analysis of the residuals reveals a mean of 3.27 and a variance of 6821.44. The non-zero mean suggests potential bias in predictions, while the variance indicates the dispersion of residuals around the mean. These insights collectively provide valuable information about the center and spread of residuals, guiding considerations for model refinement and improvement. In summary, while the model exhibits statistical significance, indications of residual autocorrelation suggest avenues for enhancing predictive performance through adjustments in model structure or feature incorporation.

```
MSE test value =  6832.13630619873
MSE train value =  6484.261740346981
RMSE test value =  82.65673781488555
Q-value:
Q-value =  104960.56800000001
Chi Critical value =  66.20623628399322
The residual is NOT white
Mean of residuals =  3.2706335616425526
Variance of residuals =  6821.439262304188
```

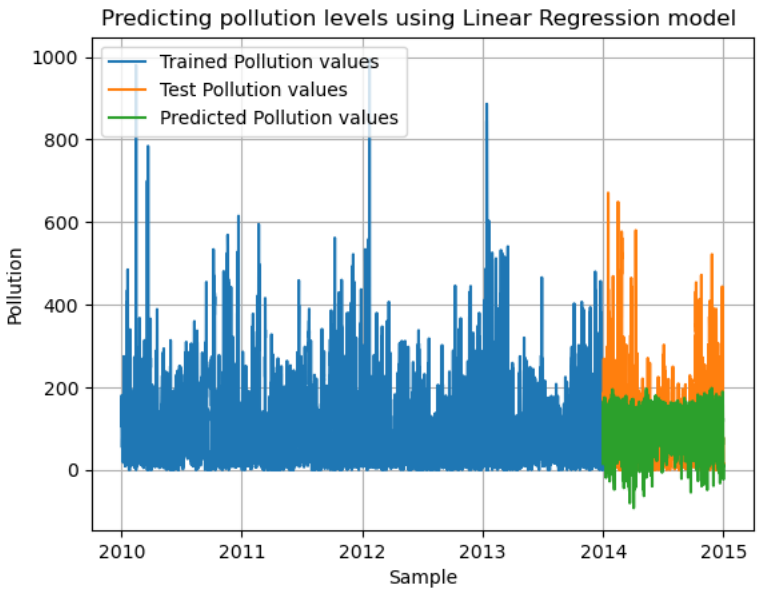


Figure 14: Test vs predicted values from linear regression model

Base Models

1) Average: The average time series model predicts future values based on the historical average of the observed data, assuming a constant mean over time.

   Residual Error (MSE) 8448.914
   Forecast Error (MSE) 8765.707
   Variance of prediction error: 8447.022
   Variance of forecast error: 8755.013
   Q-value from in-built function = 293831.367

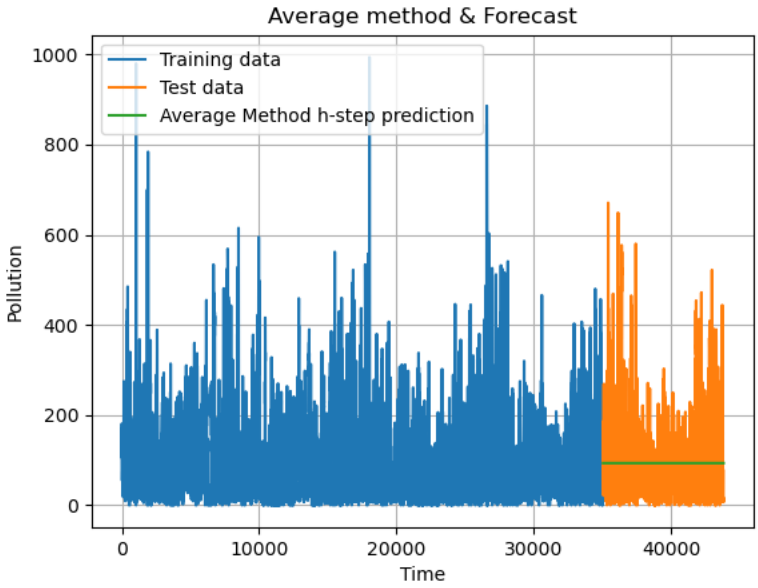


Figure 15 – Average Base model test vs predicted values

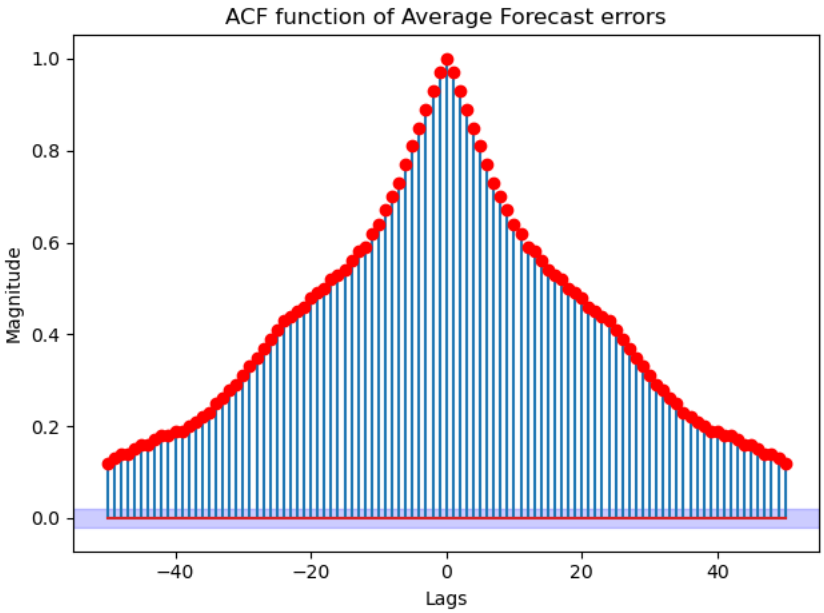


Figure 16: ACF of residual errors from Average Base Model

From the above prediction plot, we see that the average model is a flat prediction as it draws a horizontal line as the prediction which is very different from the test values.

The ACF plot of the errors is also clearly not white which shows correlation which was not captured by the model.

2) Naïve: The naive time series model predicts future values by simply using the most recent observed value as the forecast, without considering trends or seasonality.

Prediction Error (MSE) 799.753
Forecast Error (MSE) 14176.393
Variance of prediction error: 799.753
Variance of forecast error: 8755.013
Q-value from in-built function = 82079.942

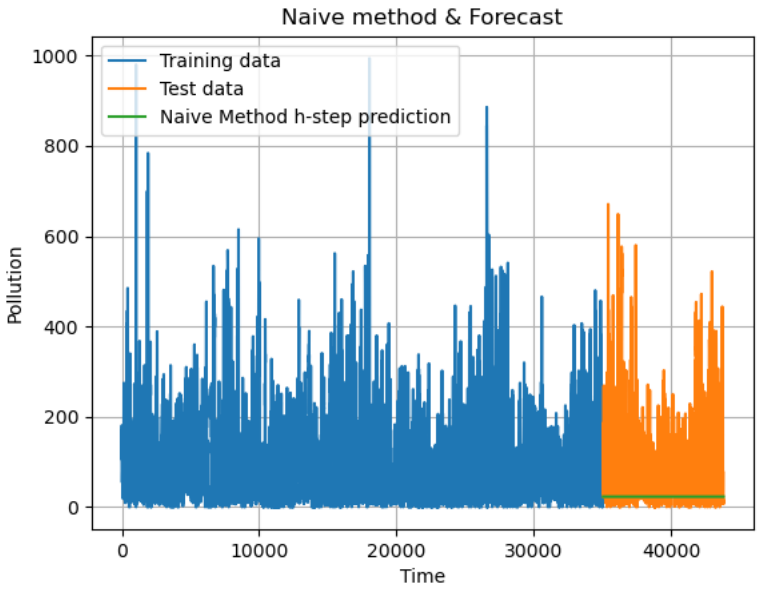


Figure 17: Naïve Base Model Test vs Predicted values

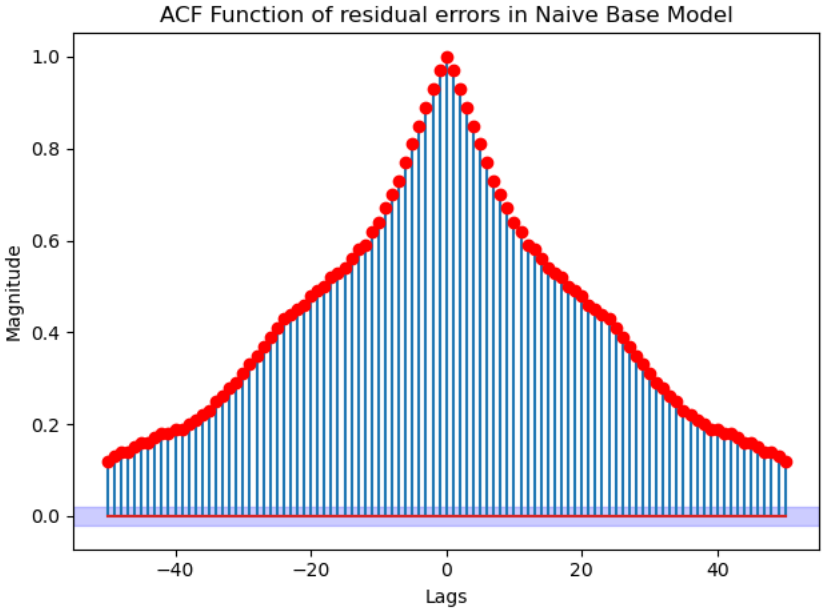


Figure 18: ACF function of residual errors in Naïve Based Model

From the above prediction plot, we see that the naive model is also flat prediction as it draws a horizontal line as the prediction which is very different from the test values. The ACF plot of the errors is also clearly not white which shows correlation which was not captured by the model.

3) Drift: The drift time series model incorporates a linear trend by extrapolating the historical slope, allowing for a gradual increase or decrease in future values.

Prediction Error (MSE) 799.995
Forecast Error (MSE) 16201.983
Variance of prediction error: 799.995
Variance of forecast error: 8653.516
Q-value from in-built function = 81276.541

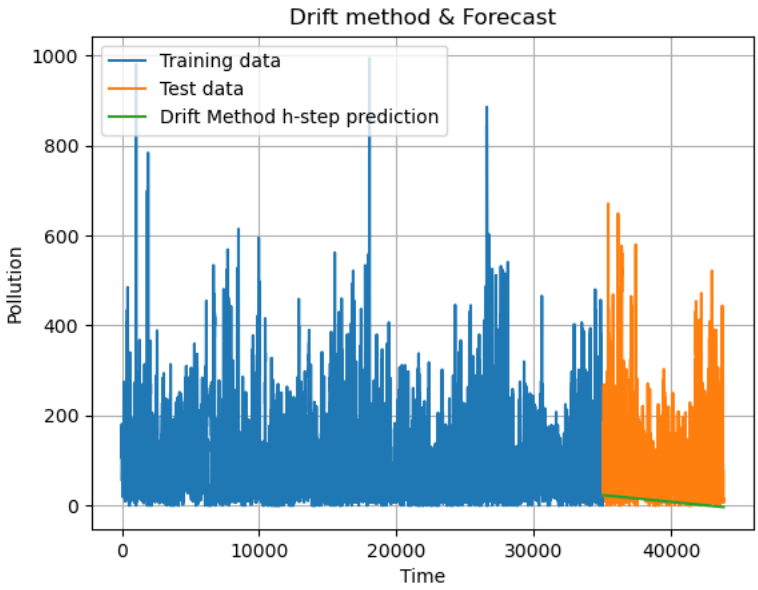


Figure 19: Drift Model test vs predicted values

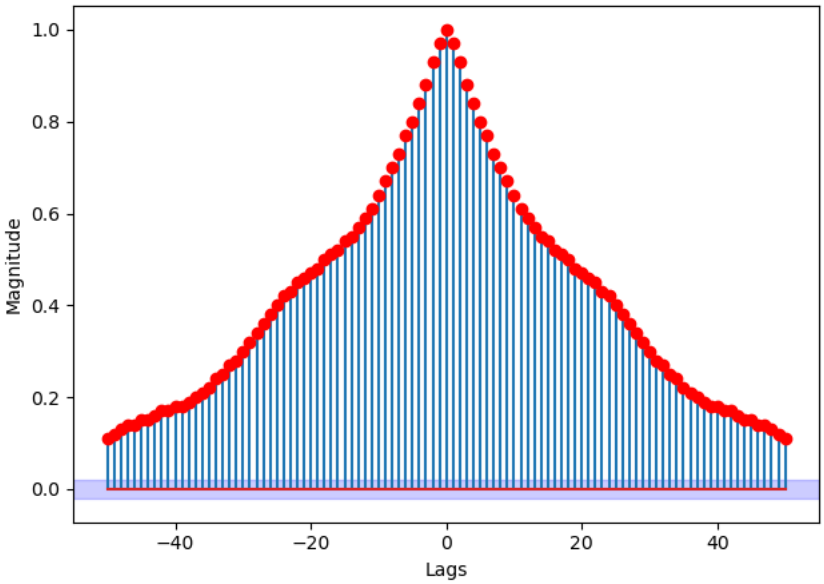


Figure 20: ACF function of residual errors from Drift Base Model

From the above prediction plot, we see that the drift model also does not predict the test values accurately. It shows a line that converges to the time axis.

The ACF plot of the errors is also clearly not white which shows correlation which was not captured by the model.

4) Simple and Exponential Forecasting: The simple exponential smoothing model forecasts future values by assigning exponentially decreasing weights to past observations, giving more importance to recent data while smoothing out noise.

Residual Error (MSE) 1039.952
Forecast Error (MSE) 14363.531
Variance of prediction error: 1039.952
Variance of forecast error: 8755.013
Q-value from in-built function = 82079.942

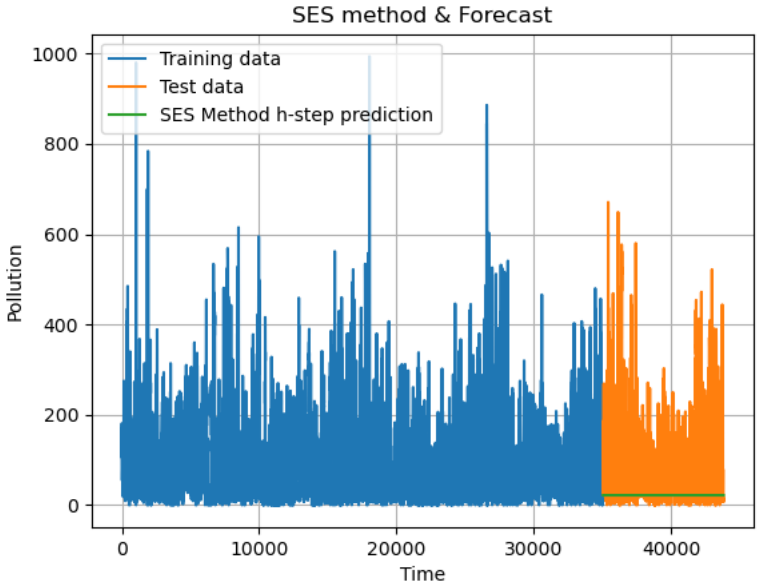


Figure 21: SES Model test vs predicted values

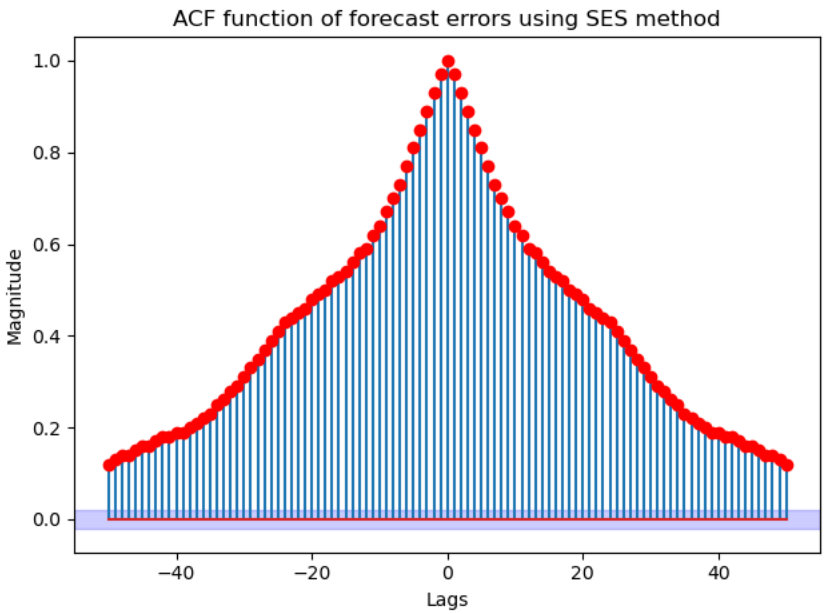


Figure 22: ACF function of residual errors using SES model

From the above prediction plot, we see that the SES model also does not predict the test values accurately. It is a flat prediction line which is parallel to the time axis line. This plot was calculated with alfa value = 0.5.

The ACF plot of the errors is also clearly not white which shows correlation which was not captured by the model.

Below is a plot which shows a comparison between 4 possible values of alfa = 0, 0.25, 0.75, 0.99.

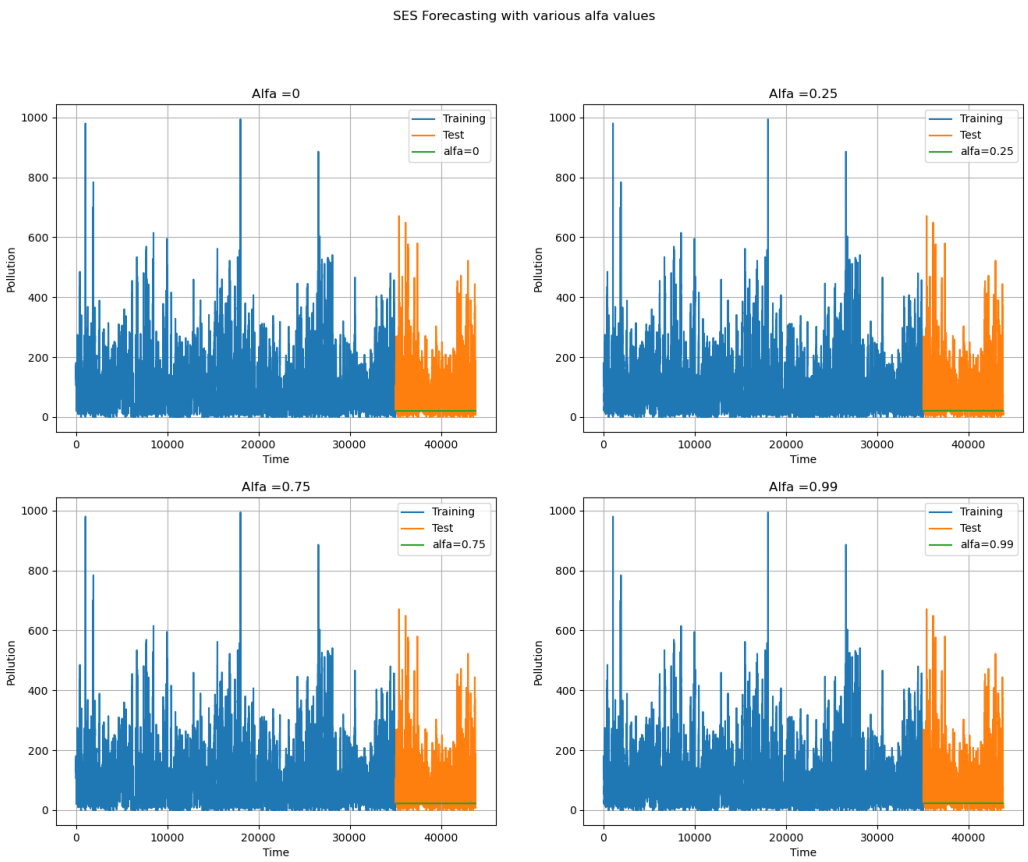


Figure 23: SES forecasting with various alfa values

Holt Winter Method

The Holt-Winters method is a popular time series forecasting technique that extends the exponential smoothing approach to handle seasonality and trends in the data. It comprises three components: level, trend, and seasonality.

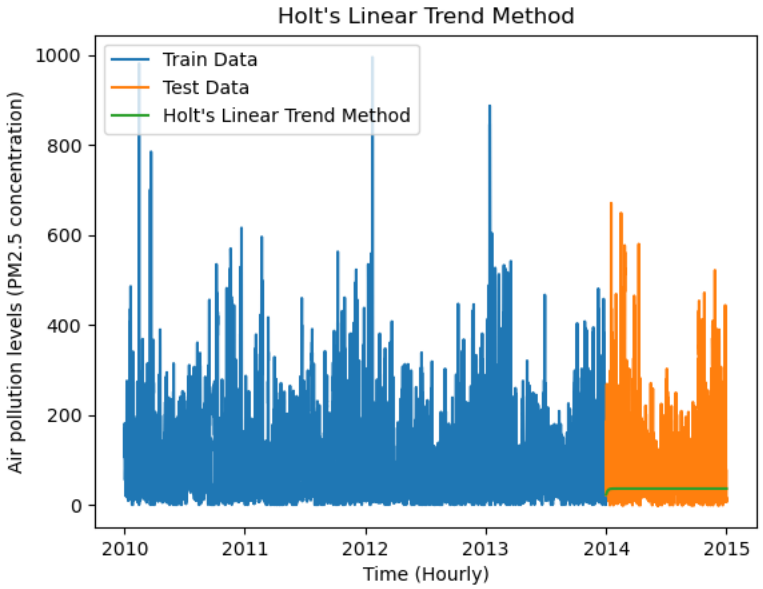


Figure 24: Holt Winter Linear Trend Method

When we plot the Holt Winter trend method, we see that the values predicted are a parallel line drawn to the time axis and does not correctly predict the y_test values.

We then continue to plot the Holt Winter Multiplicative method which captures seasonality as well. From the prediction, we see that the predicted values are not close to that of the true values. We also calculate the metrics that will help us in comparing this model with the others.

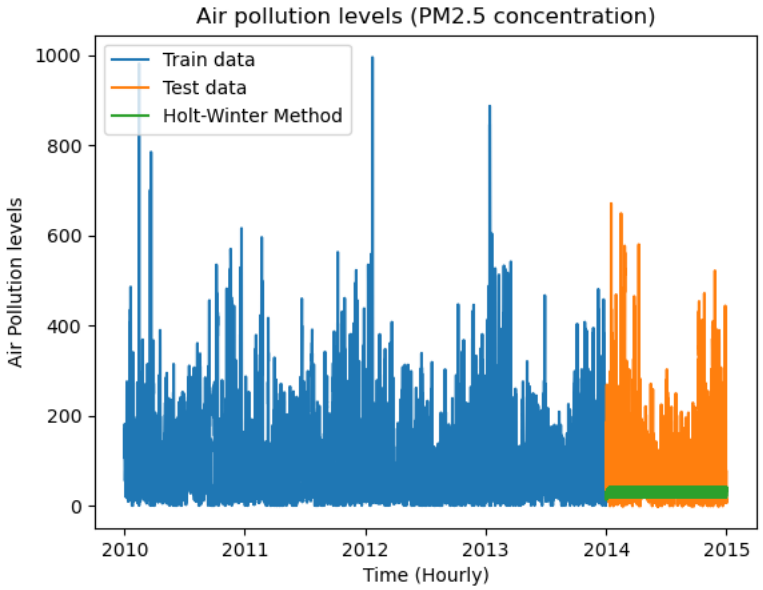


Figure 25: Holt Winter Multiplicative Method

```
Mean square error of forecast errors for holt-winter method is  13235.992695882525
Q =  81410.91732428284
Mean of forecast errors: 66.67205658674024
Variance of forecast errors: 8790.829566377031
```

From the above, we can see that the mean of forecasts error is very high and not close to zero at all. The Q-value is also high.

Order Estimation using GPAC Table

We first draw a GPAC table with the training data and estimate various potential combinations of na and nb.

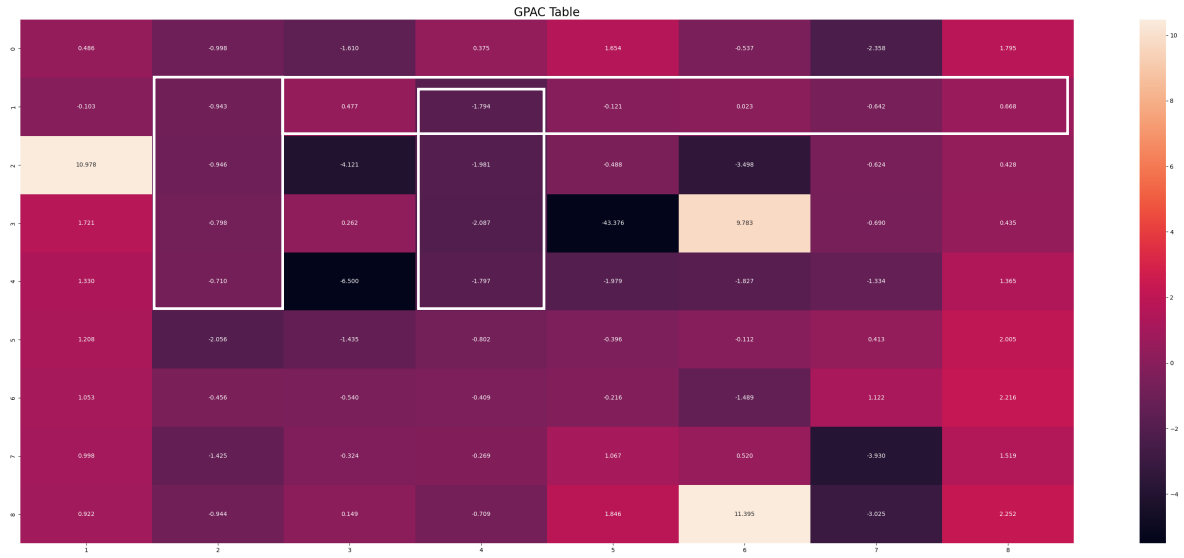
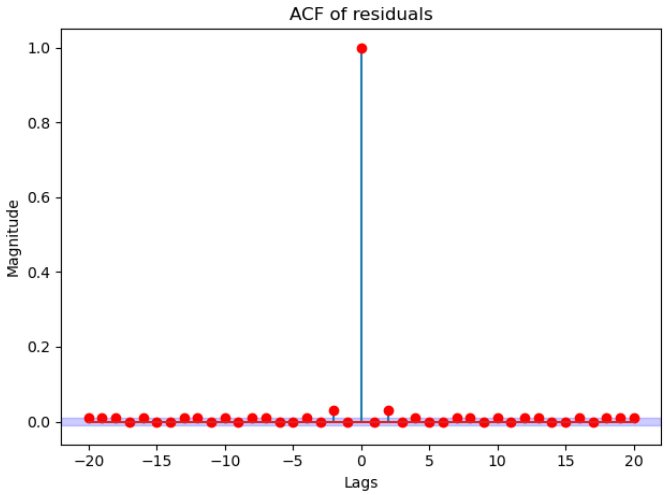


Figure 26: GPAC table

Here, we don't see a clear cut na and nb defined. However, there are 2 possible combinations seen from the above. We shall try them out one at a time and then we shall compare their residual ACF and their Q-value to check for white noise.

There are 2 possible combinations of (na,nb). One is that of (2,1) and the other is (4,1). Let's try **2,1** now. The following are the screenshots of the ACF of residual error as well as the Q-value to compare with the chi-squared value.

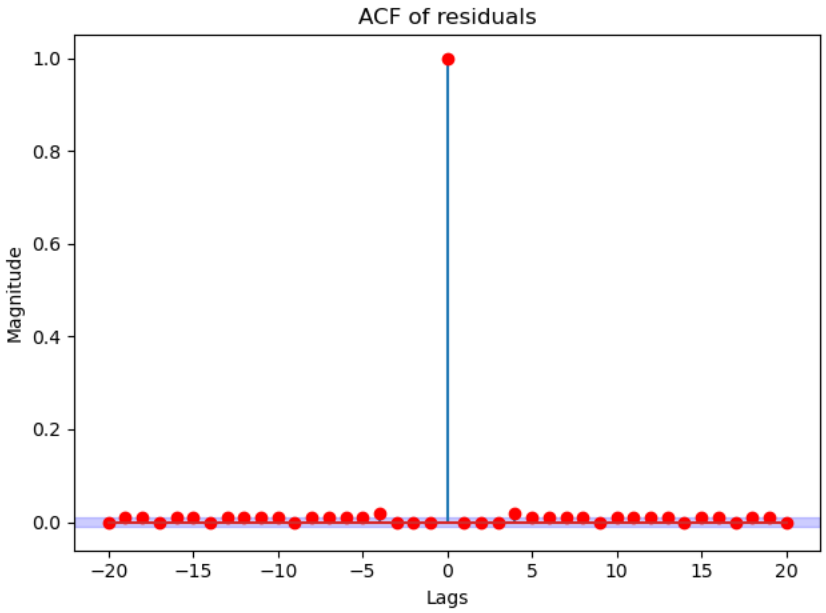



```
Q =  72.97749605699136
Chi Critical =  33.40866360500461
The residual is NOT white
```

Figure 27: Whiteness using the (2,1) order

Clearly from the above, we can see that the ACF is not a white noise with the patterns lying acf values falling outside the insignificant band. Additionally we can confirm, the Q value is very high in comparison to the chi squared value but in comparison to other models we executed above, it is very less which looks positive.

Now, we shall try the second possible combination of **(4,1)**. Below is the screenshot for ACF of residuals as well as the Q-value to compare with the




```
Q =  56.25223873746438
Chi Critical =  30.57791416689249
The residual is NOT white
```

Figure 28: Whiteness check with (4,1) model

This looks much better in terms of Q than the model obtained from (2,1). Additionally, the ACF of residual errors also looks much more like white noise than that of (2,1). However, it is not perfectly white and would indicate much more room for adjustment. Hence, we calculate the residual error's ACF and then create a GPAC with those values which can then be added to the na,nb combination to get a new order. So, we get a GPAC like the following:

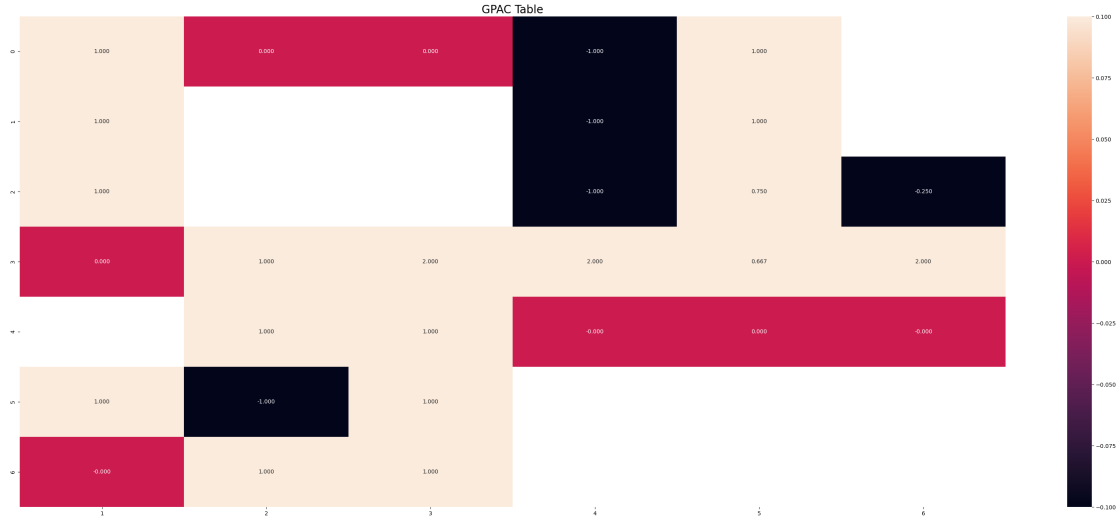


Figure 29: GPAC of autocorrelation of residual error

Unfortunately, this GPAC also does not result in anything conclusive and no specific patterns seen. Hence, there are no further additions that can be performed. Additionally, we would be limiting ourselves to the order (4,1) despite not having a perfect white noise.

We also execute the SARIMA model using the statsmodel library. But before doing so, we also add the seasonal order in the parameters of the ARIMA model. Ideally the seasonal order is 8760 (annual seasonality of 365 multiplied by 24 since the data is performed on an hourly basis). But since, we do not have enough computation to perform a seasonal order of 8760, we will have a smaller seasonal order of 24 indicating a daily frequency. Now, we run the SARIMA model and get the following results:

SARIMAX Results

Dep. Variable:	pollution_new	No. Observations:	35040
Model:	ARIMA(4, 0, 1)x(0, 1, [], 24)	Log Likelihood	-177960.104
Date:	Fri, 01 Dec 2023	AIC	355932.209
Time:	13:20:09	BIC	355982.990
Sample:	01-02-2010 - 12-31-2013	HQIC	355948.384
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9254	0.001	685.818	0.000	0.923	0.928
ar.L2	0.0283	0.001	19.251	0.000	0.025	0.031
ar.L3	-0.0227	0.002	-10.785	0.000	-0.027	-0.019
ar.L4	-0.0106	0.003	-4.044	0.000	-0.016	-0.005
ma.L1	-1.0000	0.004	-279.288	0.000	-1.007	-0.993
sigma2	1519.9427	5.785	262.727	0.000	1508.604	1531.282

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	6038325.92
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	0.85	Skew:	-0.17
Prob(H) (two-sided):	0.00	Kurtosis:	67.33

From here, we can see that all the p-values are below 0.05 and hence they are significant. Another test we see is that of confidence interval and we see that all intervals are significant with no zero value in between. We also get the coefficents from SARIMA model from here.

One point of concern that I noticed is that some coefficients are very close to zero and would deem as insignificant. To confirm this, let's do the LM algorithm to perform parameter estimation.

Parameter Estimation

The following are the results from the LM algorithm.

```
The AR coefficient 1 is: 0.007
The AR coefficient 2 is: 0.003
The AR coefficient 3 is: 0.013
The AR coefficient 4 is: 0.016
The MA coefficient 1 is: -0.009
```

When we compare these results with that from the SARIMA model, we see that the coefficients of: AR1, AR2 and MA1 do not match accordingly.

We also run the confidence interval check just to check if there might be a zero pole cancellation. Here are the results:

```
Confidence Intervals with lower and upper bound:
-231604.212  <  0.007  <  231604.225
-0.229  <  0.003  <  0.234
0.002  <  0.013  <  0.024
0.005  <  0.016  <  0.026
-231604.227  <  -0.009  <  231604.209
Roots of numerator =  [0.009]
Roots of Denominator =  [ 0.248+0.277j  0.248-0.277j -0.252+0.225j -0.252-0.225j]
```

From here, we can clearly see that the confidence intervals for coefficients of AR1, AR2 and MA1 have zero in between them. Additionally, the confidence interval of AR1 and MA1 do not look realistic at all. Hence, we conclude that these 3 coefficients (AR1, AR2, MA1) are not significant. After removing them from the equation, our final model equation is in the form below:

$$\mathbf{y(t) + 0.013y(t-3) + 0.016y(t-4) = e(t)}$$

Below we define the prediction steps:

1-step prediction: $\hat{y}_t\mathbf{(1) = -0.013y_{t-2} - 0.016y_{t-3}}$

2-step prediction: $\hat{y}_t\mathbf{(2) = -0.013y_{t-1} - 0.016y_{t-2}}$

3-step prediction: $\hat{y}_t\mathbf{(3) = -0.013y_{t} - 0.016y_{t-1}}$

4-step prediction: $\hat{y}_t\mathbf{(4) = -0.013y_{t+1|t} - 0.016y_{t}}$

General h-step Equation: $\hat{y}_t(\tau) = \mathbf{-0.013y_{(\tau-3)} - 0.016y_{(\tau-6)}}$

We also this equation for predicting the y_train values. Below are the first 500 values when comparing the true train value and the SARIMA predicted values:

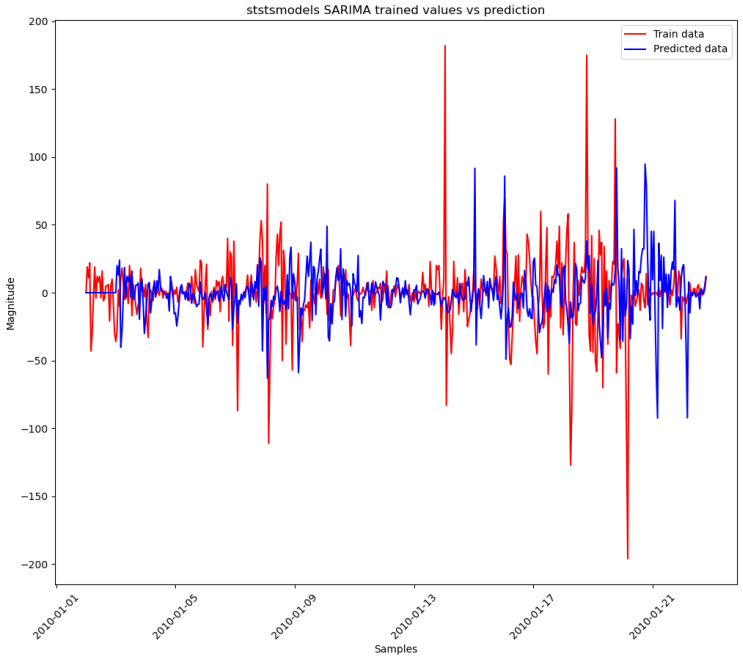


Figure 30: SARIMA train vs predicted values for first 500 samples

We see that the model has done a good prediction for the trained values.

By using the same for predicting the test values, we get the following metrics to compare this model with the rest:

```
Mean of residuals =  0.0034408958163306236
Variance of residuals =  712.0197884547474
MSE test value of SARIMA model =  712.0198002945107
MSE train value of SARIMA model =  1519.4305768803672
RMSE test value of SARIMA model =  26.683699149377897
```

Lastly, we reverse transform and compare the test values with the values that we predicted:

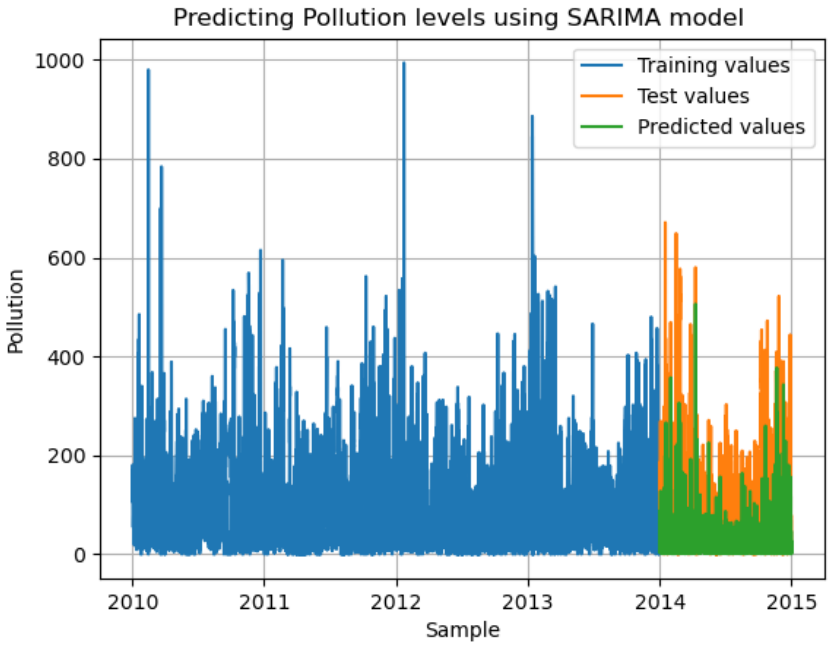


Figure 31: Predicting pollution levels using SARIMA model

This shows that the model has not completely predicted the test values correctly and that there is room for improvement. Another suggestion would be to perform a seasonal order of 365*24 = 8,760 using high computation resources on the SARIMA model which can then be used in order determination.

Final Model Selection

In the course of this project, we systematically evaluate all models and make a selection based on their comparative performance.

MODEL	MSE train	MSE test	Q-value	Variance of forecast Error
Linear Regression	6484.26	6832.136	104960.56	6821
Average Base Model	8765.707	8448.914	293831.367	293831.367
Naïve Base Model	799.753	14176.393	82079.942	8755.013
Drift Base Model	799.995	16201.983	81276.541	8653.516
SES Base Model with α=0.5	1039.952	14363.531	82560.162	8755.013
Holt Winter Method	12592.941	13235.992	81410.917	8790.829
SARIMA	1519.43	481.203	56.25	712.019

Table 2: Time Series Model Comparisons

In evaluating the performance metrics of the various models, it is evident that the SARIMA (Seasonal Autoregressive Integrated Moving Average) model stands out as a compelling choice for forecasting in this context.

The Mean Squared Error (MSE) on the test set, a crucial indicator of a model's ability to generalize to new and unseen data, is notably minimal for the SARIMA model, recorded at 56.25. This suggests that the SARIMA model outperforms other contenders in providing accurate predictions on data it has not been trained on.

Furthermore, the Q-value, a measure of forecast accuracy that considers the ratio of the mean squared forecast error to the variance of the forecast error, is impressively low for SARIMA at 712.019. A lower Q-value is indicative of a model that makes more accurate predictions, and in this regard, SARIMA excels.

Considering the training data, the SARIMA model demonstrates proficiency with an MSE of 1519.43. While the training MSE is higher than the test MSE, the model still maintains a competitive performance, indicating that it has successfully learned and adapted to the underlying patterns within the data.

In summary, the SARIMA model not only exhibits superior performance in terms of MSE on both training and test sets but also showcases a low Q-value, emphasizing its efficacy in generating accurate forecasts. Therefore, based on the comprehensive analysis of the provided metrics, the SARIMA model emerges as the most suitable and reliable choice for this forecasting task.

Summary and Conclusion

it is noteworthy to acknowledge that, given more extensive computing resources, a refinement in the modeling approach could be considered. Specifically, increasing the seasonality from a daily pattern to an annual one could provide a more nuanced understanding of the underlying behavior in the dataset. This adjustment could potentially uncover and capture long-term trends, contributing to a more comprehensive and accurate forecasting model.

It is essential to highlight, however, that the determination of the order for the SARIMA model posed a challenge. The GPAC (Generalized Partial Autocorrelation) table, a crucial tool in identifying the optimal parameters for the model, did not yield a clear order. The absence of a distinct pattern in the GPAC table complicated the process of selecting the appropriate order for the model components.

Despite this challenge, it is hypothesized that introducing an annual seasonality component with a period of 8760 hours might have facilitated a more definitive interpretation of the GPAC results. The extended seasonality period could potentially reveal clearer autocorrelation patterns, aiding in the identification of an optimal order for the SARIMA model and, consequently, leading to improved forecasting accuracy.

In summary, while constrained by computational resources and faced with the ambiguity in the GPAC table, the consideration of an extended seasonality, if feasible, could enhance the model's ability to capture annual behavioral patterns and provide a more refined forecast for the dataset.