Assessing The Measures Of Air Quality And Their Relation To Each Other

Sahithi Ancha

ABSTRACT

The presence of several different particulates in the atmosphere can drastically alter the quality of the air around us. Factors such as sunshine, rain, higher temperatures, wind speed, turbulence, can all affect the pollutant concentrations[1]. Knowing how the increase or decrease in certain particulate matter is going to affect the quality of air can help keep air pollution low while also taking precautionary measures to keep it good. For this purpose, I chose to answer some basic questions from the Air Quality dataset from the UCI Machine Learning repository to perform some basic analysis on the data. In the end, I chose the ARIMAX model to fit the data and pick the best model to forecast values. It's based on this model summary that our questions from the hypothesis are answered.

INTRODUCTION

Finding a pattern or relation between data points over a period of time can play a huge role in helping us forecast data and make predictions and changes based on that. It can be used in analysing the stock market, sales, yields, population, etc. Being able to take appropriate corrective measures if needed based on these forecasts is an important part of data-driven decision making. The scope of what we can use time series analysis for is quite large and hence we limit ourselves to answering some questions about air quality using some basic approaches.

The quality of air can drastically affect our way of lives and also have long lasting effects on the environment. Pollutants and certain gas concentrations in the air can affect the quality of the water and even destroy the ecosystem in the worst case scenario. Multipollutant exposure may imply great risk for the at-risk populations. In situations like these, it is important to understand how these factors (or variables) that determine the quality of air are correlated and the effect they might have on each other [2].

The "Air Quality" dataset that we picked has several variables that consist of gas concentrations, humidity and temperature values collected over a period of time. Through this project, I attempt to answer three questions specifically. The first one is - out of the input variables picked, which ones are the best predictors of our target variable Relative Humidity. The second is that - out of

the inputs for gas concentrations, which one predicts Relative Humidity the best. And finally, the last hypothesis is about the relation between the variables Temperature and Absolute Humidity.

METHODS

Air pollution is one of the most critical problems the world is tackling right now. There is a constant need for us to monitor the quality of the air and keep it from getting worse. The dataset that we are using is from the UCI Machine Learning repository [3], which consists of air quality values. The data has been collected as part of a research to study and expand monitoring of air pollution [4][5]. This data was recorded over a span of one year from March 2004 to February 2005 and has 9358 instances.

The dataset contains the responses of a gas multisensor device deployed on the field in a significantly polluted area at road level within an Italian city. In addition to this, it also contains hourly average responses from an array of five metal oxide chemical sensors which were recorded along with gas concentration references from a certified analyzer. The five metal oxides were namely CO, Non Methane Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2). It has 15 attributes and missing values have been depicted by the value -200.

The measurements obtained are as follows -

0 Date (DD/MM/YYYY)
1 Time (HH.MM.SS)
2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7 True hourly averaged NOx concentration in ppb (reference analyzer)
8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12 Temperature in Â°C
13 Relative Humidity (%)
14 AH Absolute Humidity

The first step after obtaining the data would be to check its structure and remove any missing values that are present. In this case the null values have been denoted by -200. These null values have been replaced by the mean of all the values in the column it is a part of. However, the column NMHC had 8443 such values and thus the entire column had to be dropped. I have also combined the Date and Time columns into one and put them in the format "ymd-hms". Finally, I applied a log on all the remaining columns to convert the dataset to additive. The NaNs produced have again been aggregated.

The input variables being used were chosen at random and are as follows -

- CO
- PT08.S3 (tungsten oxide)
- NO2
- Temperature
- Absolute Humidity

The target variable -

- Relative Humidity

Next the variables being used are centered so that the scale shifts to a new '0' point and plot it afterwards as shown in Figure 1.

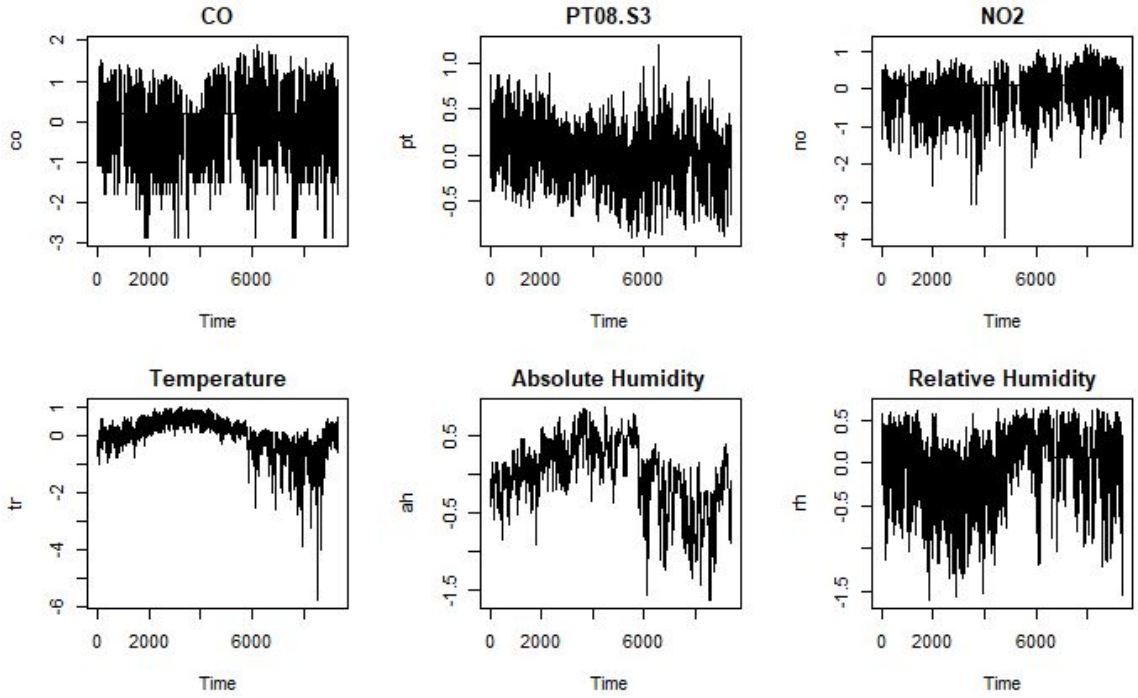


Figure 1: Time Series plots for each input variable.

From the plots, it is evident that both the mean and variance are not constant. The data is seasonal. For this purpose, we next difference our variables and again plot them. The results have been depicted in Figure 2. However, it is evident that further analysis needs to be carried out before trying to fit and model data. For this, we use the autocorrelation, cross correlation and spectral analysis plots.

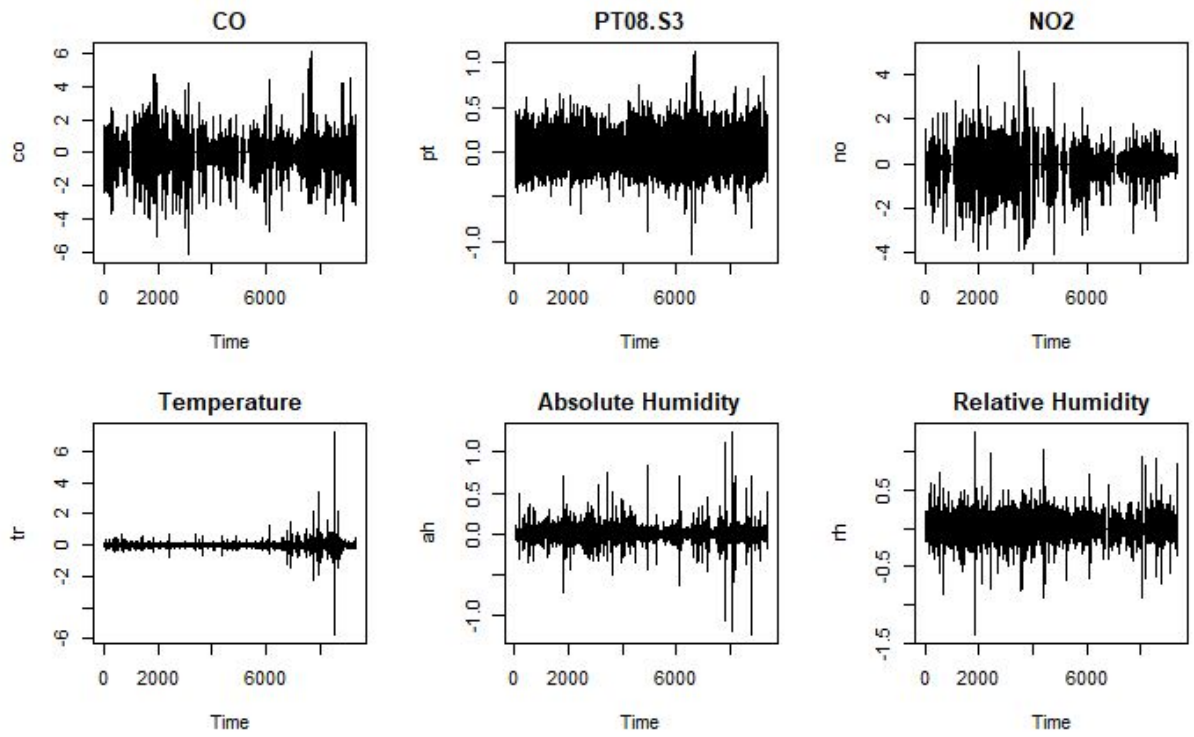
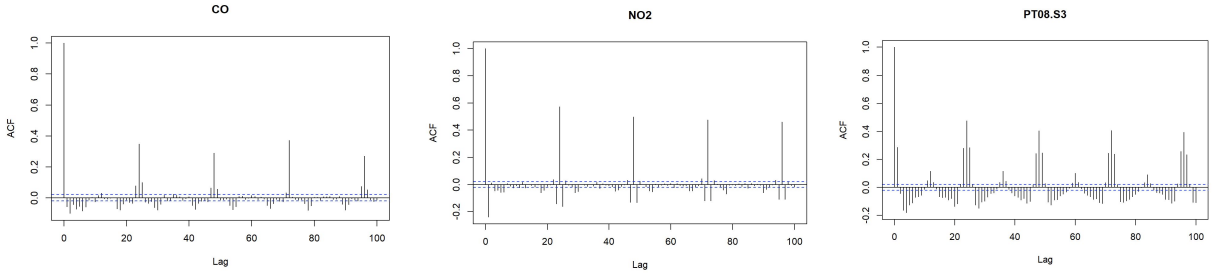


Figure 2: Input variables after differencing.

We use the autocorrelation function to check how the current value relates to past values in the time series. The autocorrelation plots (Figure 3) here indicate that there's a significant amount of fluctuation present in the data. In other words, there's strong autocorrelation at every 24th lag as well as at the variables right before and after it.



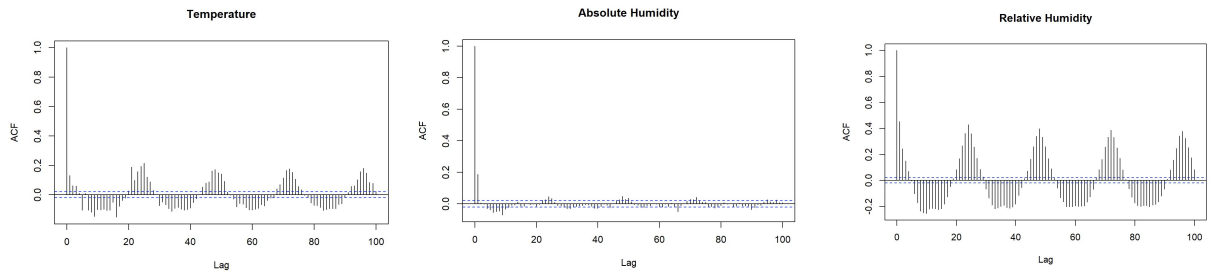


Figure 3: ACF plots for the variables

When we try to see how each of our input variables measures or relates to the target variable, in other words, we are plotting the cross correlation between them. Here, we observe strong correlations in a cyclical pattern.

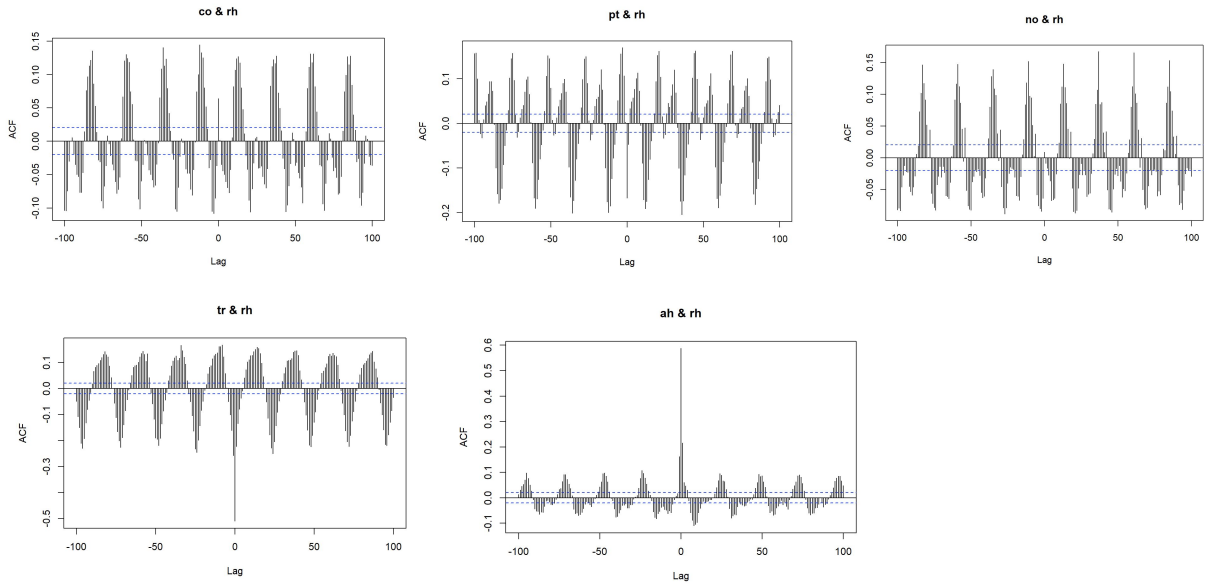
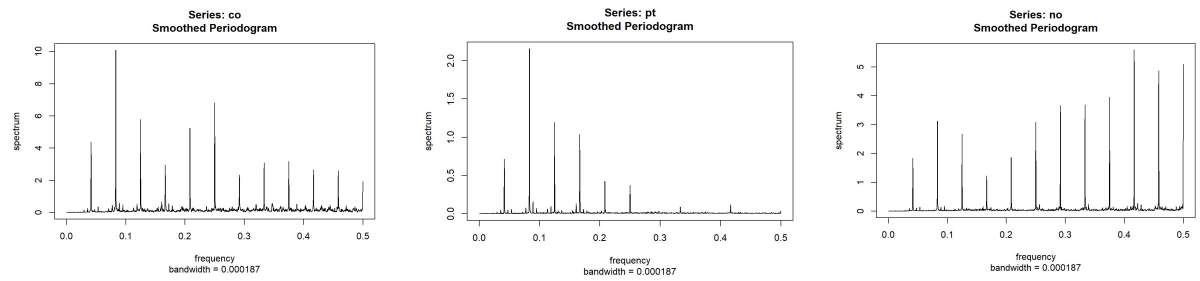


Figure 4: CCF plots for the variables.

Next we perform a spectral analysis to calculate a periodogram and plot it against frequency. Similar to how we observed in the earlier analysis of the acf and ccf plots, the periodogram also has large spikes.



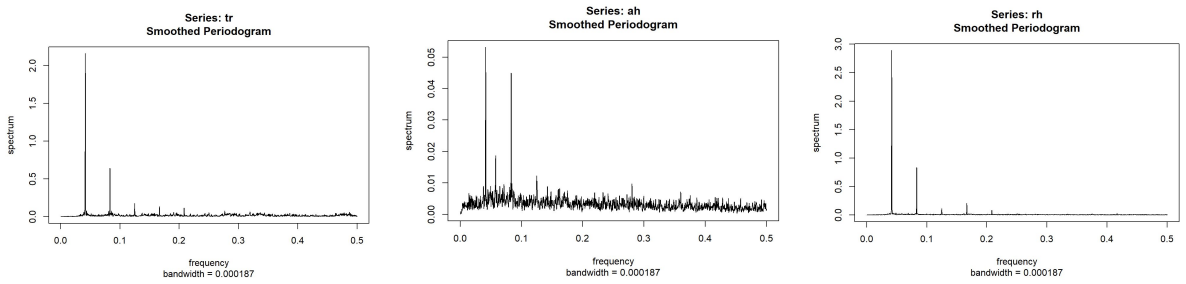


Figure 5: Spectral analysis plots

Trying to change in the window sizes resulted in the spikes in the plot not being as obvious.

It is necessary to detrend this data before proceeding with fitting the models and interpreting them. To detrend the data, we difference each of the variables again with a lag of 24. This is because when the acf plots were observed, there was a peak in the data for every 24th value (24, 48,...). After differentiating, we also plot the acf, ccf and spectrum to see how differentiating has changed the data. Differencing makes the model more stable and easier to manipulate. After this, ACF, CCF and Spectral analysis are performed again to see how differencing has changed the data.

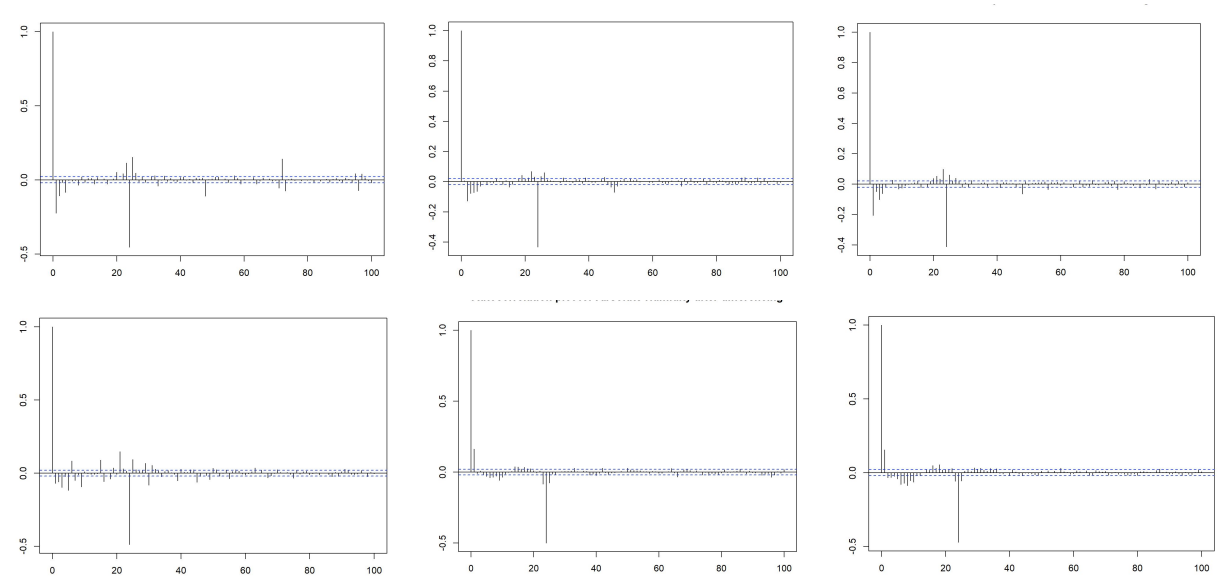


Figure 6: ACF plots after differentiating the variables.

While differentaiting hasn't particularly improved the ACF plots for the input variable CO, For the rest of the variables, while there's a spike in the beginning, the relation between the current and past variables eventually dies out and stays within the confidence interval.

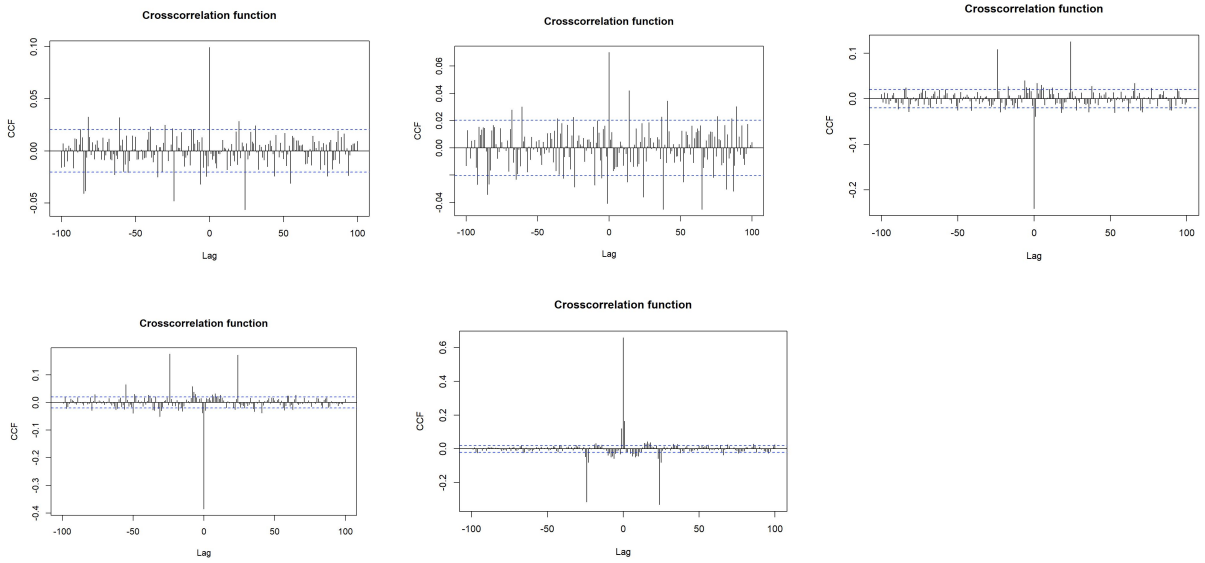


Figure 7: CCF plots depicting the relation between the input variables and the target variable.

The CCF plots too have considerably improved when compared to Figure 4. While there are a few spikes indicating strong correlations, the correlation stays between the confidence interval (blue-dotted lines) most of the time.

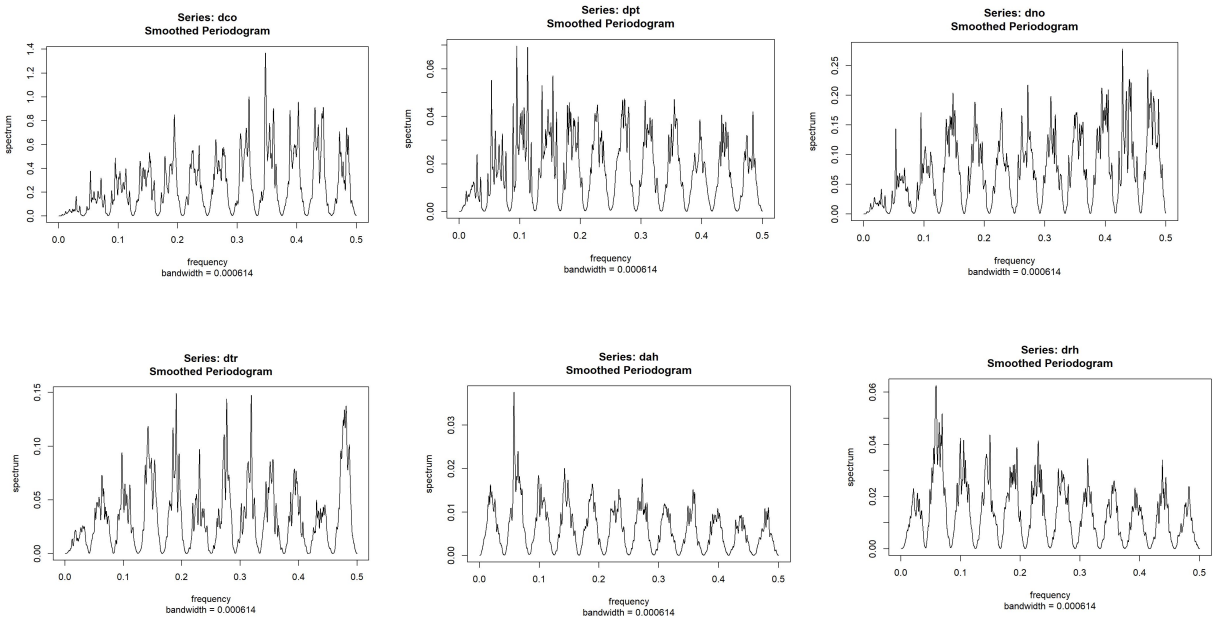


Figure 8: Spectral plots for the differenced variables.

The spectral plots however, shows large systematic cycles sticking out.

To find the relation between our variable Temperature and Absolute Humidity, we use the ccf plot and depict the results in Figure 9.

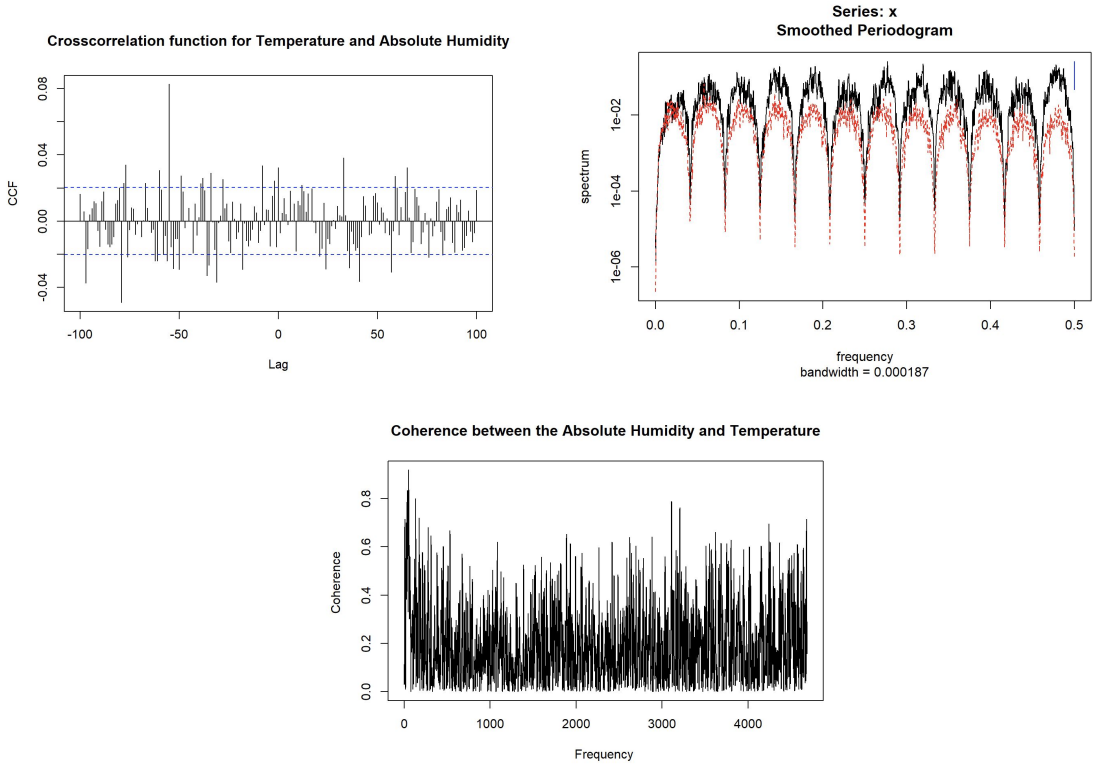


Figure 9: Ccf and Spectral plots depicting the relation between Temperature and Absolute Humidity.

Based on the ACF, CCF and spectral plots above, the data is going to be fit using an ARIMA model. A lag is applied to each of the differentiated variables and lag is again applied to the lag 1 differentiated terms. This is followed by formulating two models, X1 and X2. X1 contains lag 1 and lag 2 terms and X2 contains only lag 1 terms. The first model we depict has xreg=X1 with all the external variables lag 1 and lag 2 whereas the second model only has lag 1 variables with xreg=X2. Next the results from these models are summarised and compared. Trying to vary the AR and MA terms did not yield much of a difference, hence I kept them as 1.

RESULTS

The summary for Model 1 is as follows -

```
Call:
arima(x = drh, order = c(1, 0, 1), xreg = X1)

Coefficients:
          ar1     ma1  intercept   dco_2    dco_1    dpt_2   dpt_1    dno_2
      -0.3209  0.3419      0e+00  0.0024  -0.0034  -0.0618  0.0136  -0.0026
s.e.   0.1768  0.1753      8e-04  0.0018   0.0018   0.0069  0.0070   0.0032
        dno_1    dtr_2    dtr_1   dah_2   dah_1
      -0.0066  -0.2555  -0.0372  0.9621  0.0252
s.e.   0.0033   0.0042   0.0042  0.0101  0.0101

sigma^2 estimated as 0.005412:  log likelihood = 10991.12,  aic = -21954.25

Training set error measures:
                      ME       RMSE        MAE MPE MAPE      MASE
Training set 5.701527e-08 0.07356293 0.04283063 NaN  Inf 0.4113394
                     ACF1
Training set -0.006533106
```

Summary for Model 2 is as follows -

```
Call:
arima(x = drh, order = c(1, 0, 1), xreg = X2)

Coefficients:
          ar1     ma1  intercept   dco_1    dpt_1    dno_1    dtr_1   dah_1
      -0.2097  0.2318      0e+00  0.0024  -0.0623  -0.0014  -0.2528  0.9644
s.e.   0.1855  0.1842      8e-04  0.0017   0.0069   0.0031   0.0042  0.0101

sigma^2 estimated as 0.005471:  log likelihood = 10942.13,  aic = -21866.27

Training set error measures:
                      ME       RMSE        MAE MPE MAPE      MASE
Training set 1.491759e-06 0.07396388 0.04298673 NaN  Inf 0.4128386
                     ACF1
Training set -0.002683856
```

Summary for Model 3 -

```
Call:
arima(x = drh, order = c(1, 0, 1))

Coefficients:
          ar1     ma1  intercept
      -0.1311  0.2955    -0.0001
s.e.   0.0552  0.0530     0.0014

sigma^2 estimated as 0.01334:  log likelihood = 6825.89,  aic = -13643.79

Training set error measures:
                      ME      RMSE        MAE MPE MAPE      MASE
Training set 6.169638e-06 0.1155198 0.07844289 NaN  Inf 0.7533546
                     ACF1
Training set -0.0008605847
```

The AIC values for each model are as follows -

Model 1: -21954.25

Model 2: -21866.27

Model 3: -13643.79

The BIC values for each model are as follows -
Model 1: -21854.43
Model 2: -21802.1
Model 3: -13615.26

On comparing both the values, it can be ascertained that Model 1 is the best one to fit the data since it has the least AIC and BIC values. Model 2 is close behind in AIC and BIC values but for the rest of the analysis, we go with Model 1. On plotting the residuals for model one and analysing them to see if the model is white noise, we obtain the following results. The ACF plot is given in Figure 10.

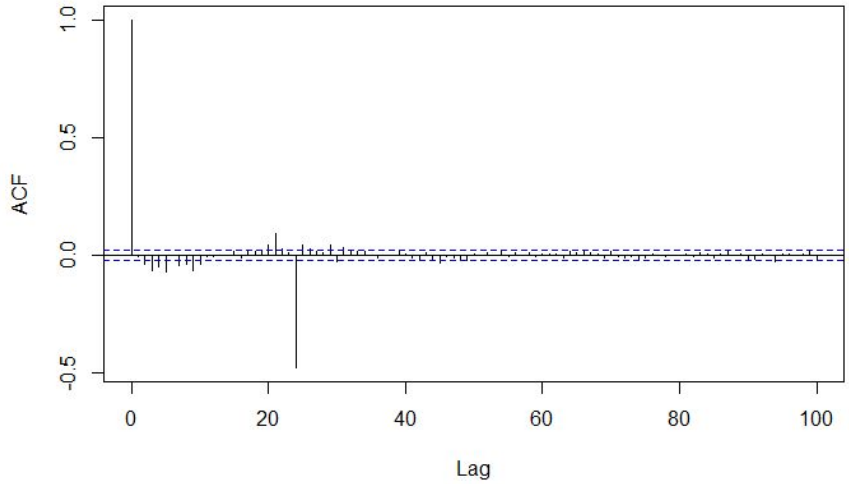


Figure 10: ACF plot for residuals of model 1.

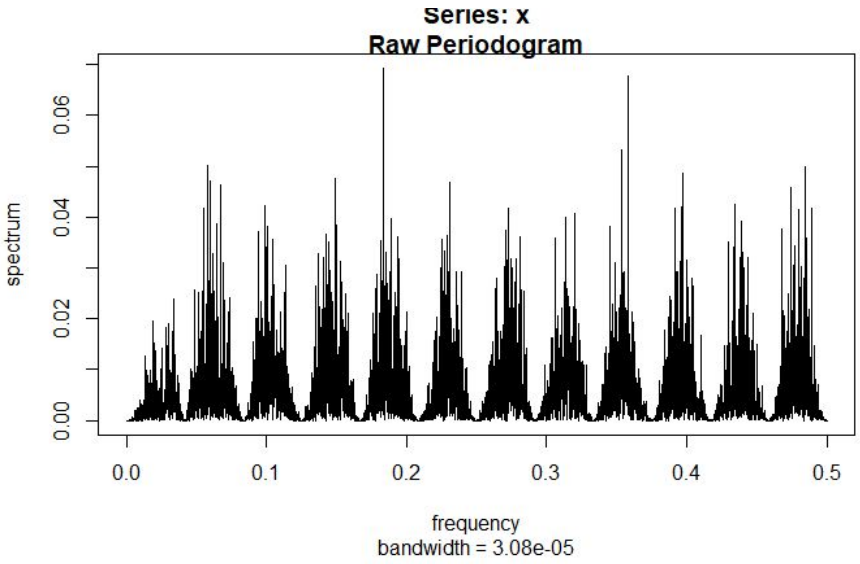


Figure 11: Spectral plot for the residuals.

From the Box-Pierce test, the p-value obtained is 0.5302, which is greater than 0.05 and we reject the alternative hypothesis. We thus cannot reject our null hypothesis "The residuals are stationary". We can thus say that our residuals are white noise. Next we predict/forecast future values based on the selected model 1, using the last 50 values of our dataset and compare the same. The forecasted values are depicted by a dotted line. This is given in Figure 12.

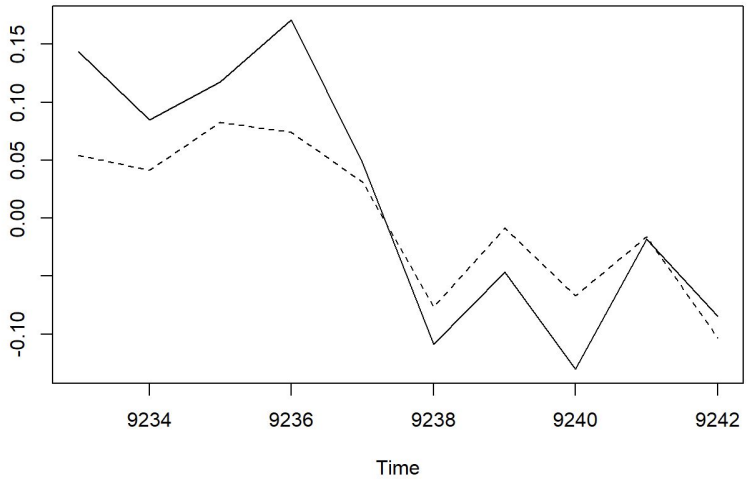


Figure 12: Data vs Forecasted values.

DISCUSSION

On comparing the coefficients for model 1 for each of the parameters in it, we observe that dtr_2 is the coefficient with the least value. This means that Temperature is the best predictor for Relative Humidity. This also complies with research that affirms that temperature is a good predictor of humidity. "Warm air can hold more water vapor than cool air, relative humidity falls when the temperature rises if no moisture is added to the air" [6]. Out of all the gas concentration variables taken as input, NO2 is the best predictor for Relative Humidity and performs better than CO as well as PTS8.S3. Also, from Figure 9, we can deduce that Temperature and Absolute Humidity have a close relationship and follow a similar pattern. The Model 1 selected is also a good fit for our dataset. To make the analysis more robust, we can include all the variables in the dataset to build a more comprehensive understanding.

REFERENCES

[1]US Department of Commerce, and Noaa. “Clearing the Air on Weather and Air Quality,” April 6, 2017. https://www.weather.gov/wrn/summer-article-clearing-the-air.

[2] “Research on Health and Environmental Effects of Air Quality.” EPA. Environmental ProtectionAgency,May15,2019.https://www.epa.gov/air-research/research-health-and-environmental-effects-air-quality.

[3] Dataset:- http://archive.ics.uci.edu/ml/datasets/Air+Quality

[4]Saverio De Vito, Marco Piga, Luca Martinotto, Girolamo Di Francia, CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors and Actuators B: Chemical, Volume 143, Issue 1, 4 December 2009, Pages 182-191, ISSN 0925-4005

[5]S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005

[6] Skilling, Tom. “The Relationship between Relative Humidity, Temperature and Dew Point.” chicagotribune.com,August24,2018.
https://www.chicagotribune.com/news/ct-xpm-2009-11-15-0910190209-story.html.

[7]Frost, Jim, Aritra Rawat, Michael J Buono, Indira, Curt Miller, Seman Kedir Ousman, Bunga Aisyah, et al. “Identifying the Most Important Independent Variables in Regression Models.” Statistics By Jim, June 13, 2019. https://statisticsbyjim.com/regression/identifying-important-independent-variables/.