**Predictive and Decision-Making Report for Beijing Weather**

University Canada West

BUSI 650 (Section- 20)- Business Analytics

Abhilasha Rajan (2307174)

Aditya Prashant Arte (2305505)

Amandeep Kaur (2222418)

Anitha Juturu (2242148)

Fariha Nasrin (2302615)

Mohammed Sameer khan (2227080)

Date: March 17, 2024

**Table of Contents**

[**Introduction 3**](#_q2t35g699yft)

[**Methodology 4**](#_uye2un5b1yjg)

[**Analysis 5**](#_b1w3boi4shsk)

[Data Division 5](#_h09jcz7mx3rt)

[Descriptive Analysis 9](#_gdqfqtjtkk7r)

[Data Description: 9](#_wjvpgsq7aglz)

[Regression analysis 10](#_35a77cbfn83q)

[Regression 10](#_hp0ghib0c2bu)

[Piecewise linear regression 11](#_7yp6z1iox15i)

[Time Series Forecasting 13](#_z8g0gk2nmahu)

[**Conclusion 20**](#_t7fdxq2hpl4k)

[**References 21**](#_ugu8r5l12r0x)

[**AIgarism 22**](#_k30a1f8afu3a)

# 

# 

# 

# 

# 

# 

# **Introduction**

In this paper, we will discuss the weather conditions and pollution levels in Beijing for five years from the dataset of the US embassy based in Beijing, China. The primary purpose of this report is to analyze Beijing's environmental situation using the US embassy dataset. Analyzing weather and pollution levels is crucial for people who travel to Beijing for business or other essential purposes. Weather conditions can impact both companies and customers. Accurate weather and pollution forecasting help customers plan their travels, outdoor activities, and commute for work or other lifestyle-related things. It also allows the health department to provide health advisory services for the region.

On the other hand, businesses can plan according to the situation to conduct their operations by assessing environmental hazards. Predictive models help businesses mitigate risk and optimize resource allocation. This study will comprehensively approach weather conditions and pollution levels through descriptive analysis, regression analysis on pollution prediction through data division, and time series forecasting for crucial weather parameters. This paper will represent a critical analysis, which will help customers and business infrastructures to get insights about specific environmental conditions and help them to plan accordingly.

# **Methodology**

To conduct this study, we adopted a comprehensive approach consisting of descriptive analysis, regression analysis, and time series forecasting techniques. Utilizing the given dataset for analysis requires data division. Data should be divided into training and testing, which include hourly observations of pollution levels and weather parameters recorded by the US embassy in Beijing. Descriptive analysis helps the study by connecting summary statistics and visualization and identifying trends and patterns in the data. On the other hand, regression analysis plays an essential role in predicting pollution levels using multiple weather variables since it helps us to get linear, non-linear, regular, and piecewise regression models. Moreover, time series forecasting models forecast future values using historical data weather parameters and predict the trends and fluctuations in temperature, wind speed, and dew point.

# 

# 

# **Analysis**

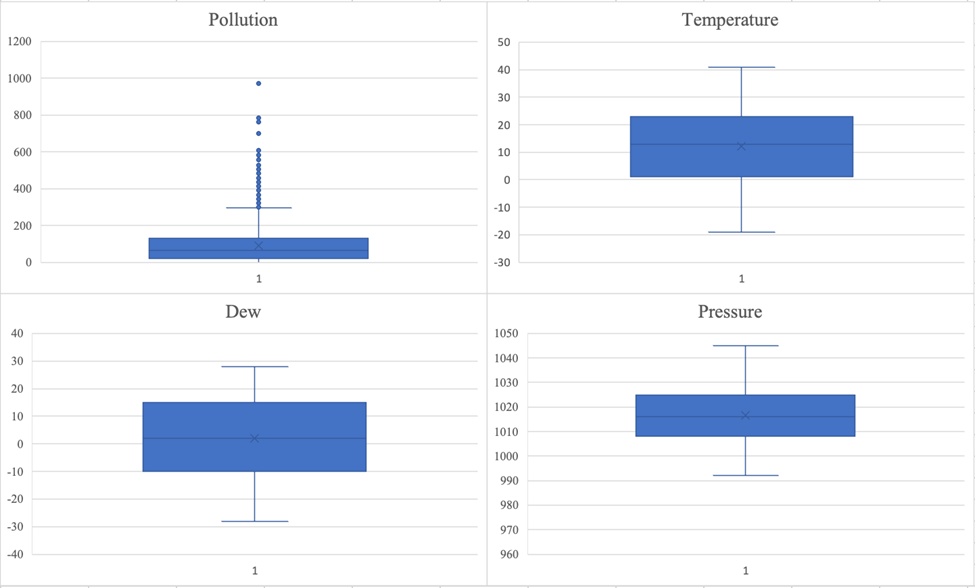
## **Data Division**

Seasonality plays a crucial role in time series predictions when working with data concerned with climate and environment. These produce a periodic and recurring pattern influenced by pollution, dew, rain, snow, and wind in the given data. Data division into training and testing data sets aims to formulate a model, train it on past data, and test the model on future data. (Puma-Villanueva et al., 2006, 4740-4747) This ensures the validity of the data and provides accurate insights based on the observations. Precision can be achieved by cleaning the data and eliminating outliers.

The authors identified outliers visually using a box-and-whisker plot for each variable: pollution, dew, temperature, pressure, wind direction, wind speed, snow, and rain.

**Figure 1**

Variable-wise Box-and-Whisker Plot *I*

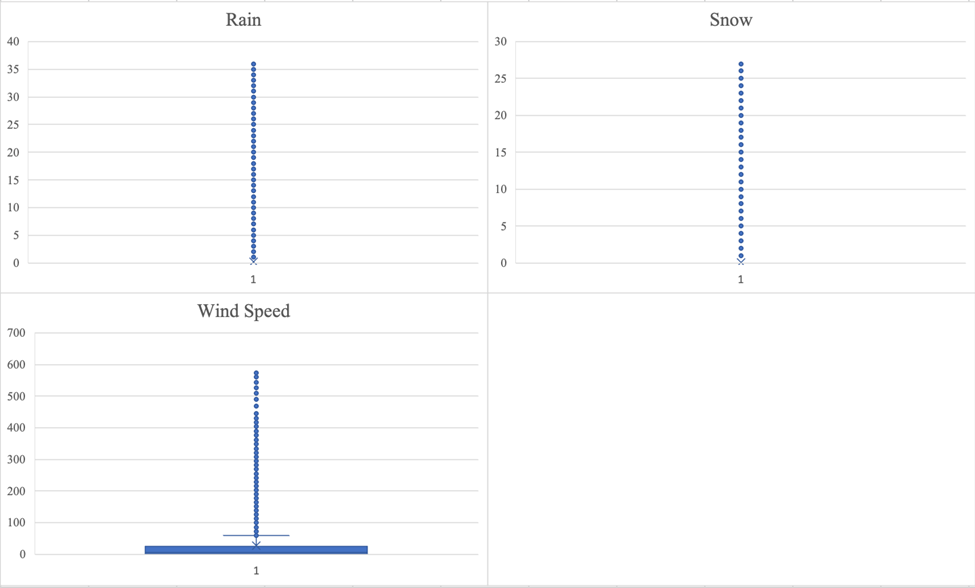


*Note:* Created by the authors

# 

**Figure 2**

Variable-wise Box-and-Whisker Plot *II*



*Note:* Created by the authors

The anomalies in pollution and windspeed were identified. These values were eliminated by calculating the variable's quartiles and upper and lower limits. The values are shown in Table 1.

**Table 1**

Quartile Values for Pollution and WindSpeed

| Pollution | | Wnd\_Speed | |
| --- | --- | --- | --- |
| QTL1 | 22 | QTL1 | 1.79 |
| QTL3 | 132 | QTL3 | 25.02 |
| IQR | 110 | IQR | 23.23 |
| Lower Limit | -143 | Lower Limit | -33.055 |
| Upper Limit | 297 | Upper Limit | 59.865 |

*Note:* Created by the authors

## 

## **Descriptive Analysis**

### 

### Data Description:

· As per Weather terms, the mode value represents the most recurring occurrence of the particular weather, whether its temperature, wind, or rainfall.

· The Kurtosis value is positive for pollution, wnd\_spd, snow, and rain, which means frequent heavy rainfall occurrences occur in the area.

· The Kurtosis value is negative for dew, temp, and press, which implies that the temperature tends to remain relatively stable without extreme fluctuations.

· The skewness value for pollution, press, wnd\_spd, snow, and rain is positive, indicating that the distribution of a meteorological variable is skewed towards higher values. For example, a positive humidity skewness indicates that humidity levels usually are higher, with rare increases.

· The skewness value for dew and temp is negative, indicating that the distribution of a meteorological variable is skewed towards lower values. For example, a negative humidity skewness indicates humidity levels usually are lower, with occasional bursts of severe wind.

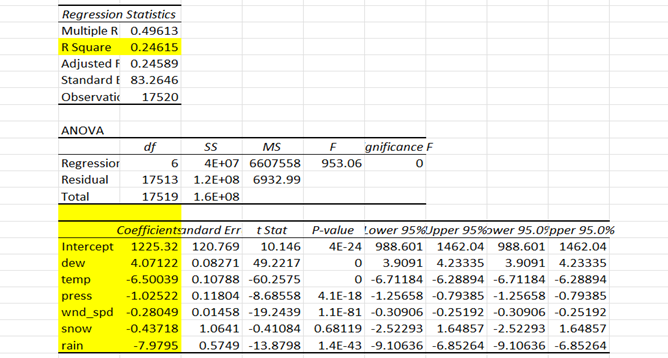
## **Regression analysis**

### Regression

A collection of statistical techniques called regression analysis is used to estimate the relationships between a dependent variable and one or more independent variables. The given data shows the differences in R squared without wind direction and with wind direction of train and test data. The value of R squared without wind direction for train and test data is similar, around 0.25, and the intercept is also positive.

**Figure 3**

Regression analysis of train and test data



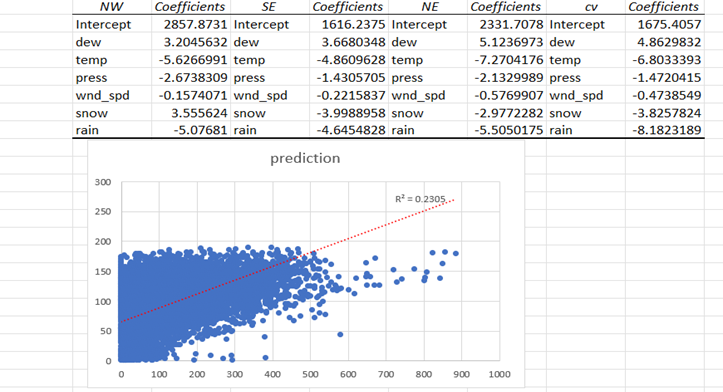
*Note:* Created by the authors

### Piecewise linear regression

The fundamental tenet of piecewise linear regression is that the regression function should be modeled in "pieces" if the data exhibit distinct linear trends across various data regions. The given data was divided based on wind direction to find the relation between actual and predicted value, and a piecewise linear regression model was performed. Firstly, divide the data based on wind direction and perform regression analysis on each wind direction. After that, combine the coefficient value of each wind direction and calculate the predicted value. In the given test data, there is a positive relation between actual and expected value, but there is a negative relation in train data.

**Figure 4**

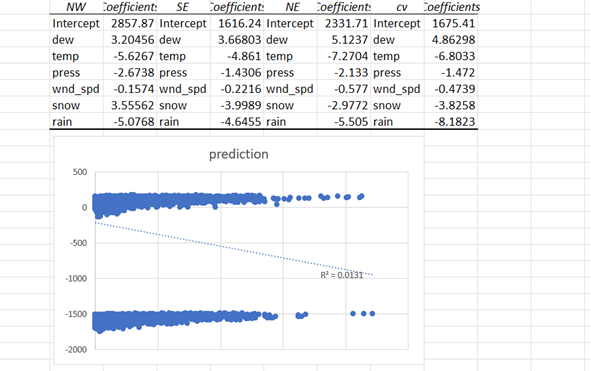
Regression analysis of test data with wind direction



*Note:* Created by the authors

**Figure 5**

Regression analysis of train data with wind direction



*Note:* Created by the authors

## **Time Series Forecasting**

Time series forecasting is a method of analysis in which predictions are made based only on time or historical data. In time series forecasting, the patterns data follows over time are also considered during analysis. Using time series forecasting enables users to predict events that recur with time.

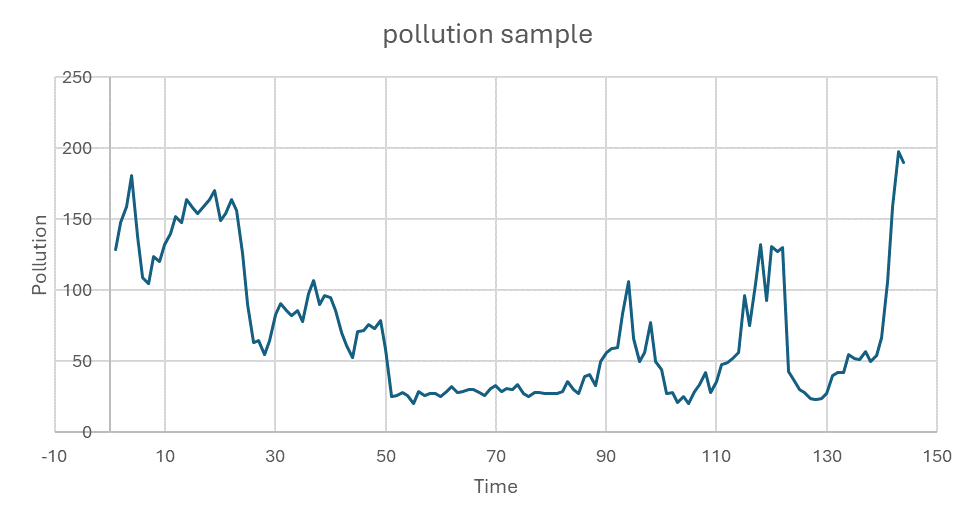
There are many time series forecasting methods; in this paper, the authors have only used Moving Average, Exponential Smoothing, Non-Linear Regression, and Auto Regression. These methods were implemented after the authors observed the patterns in the available dataset.

As the data available to the authors had many data points, they had to check patterns for time for the whole data range and check small samples to get a clear idea of the type of pattern the data followed.

The authors were asked to make predictions for the level of pollution, dew point, temperature, and wind speed. The authors started the analysis by checking for patterns in data. To check the patterns in data, they plotted the parameters individually for time. The pattern observed for pollution and wind speed is a horizontal pattern, and the pattern observed for dew and temperature is a trend pattern (Refer to figures 1 to 8).

**Figure 6**

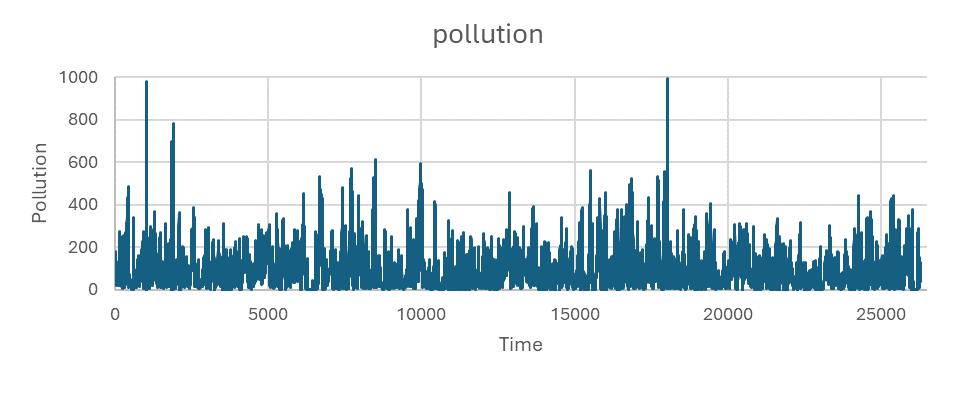
Pollution over time sample plot

****

*Note:* Created by the authors

**Figure 7**

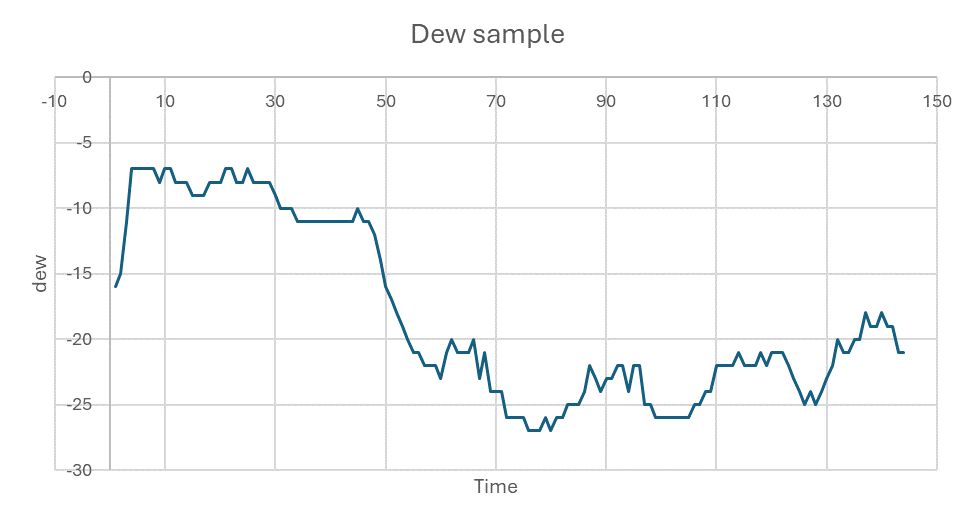
Pollution over time plot



*Note:* Created by the authors

**Figure 8**

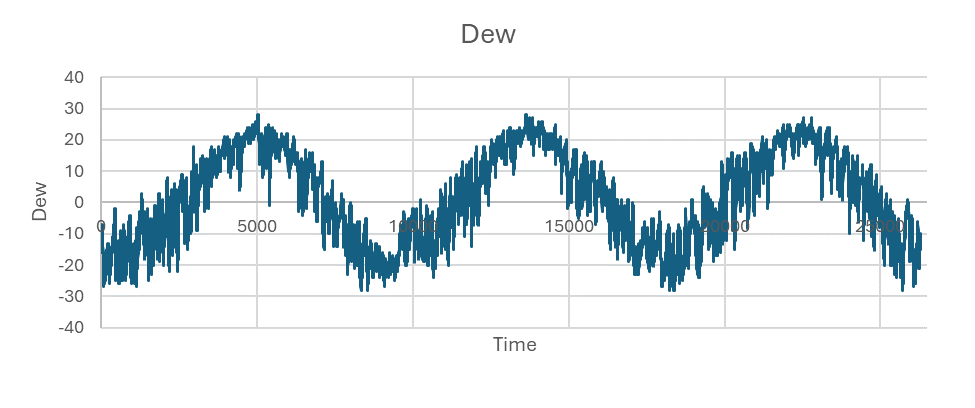
Dew point over time sample plot



*Note:* Created by the authors

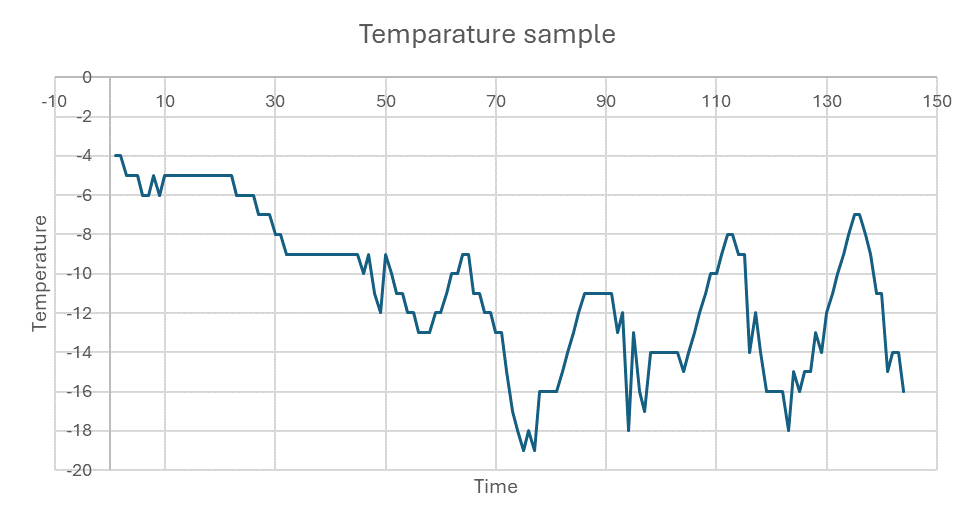
**Figure 9**

Dew point over time plot



*Note:* Created by the authors

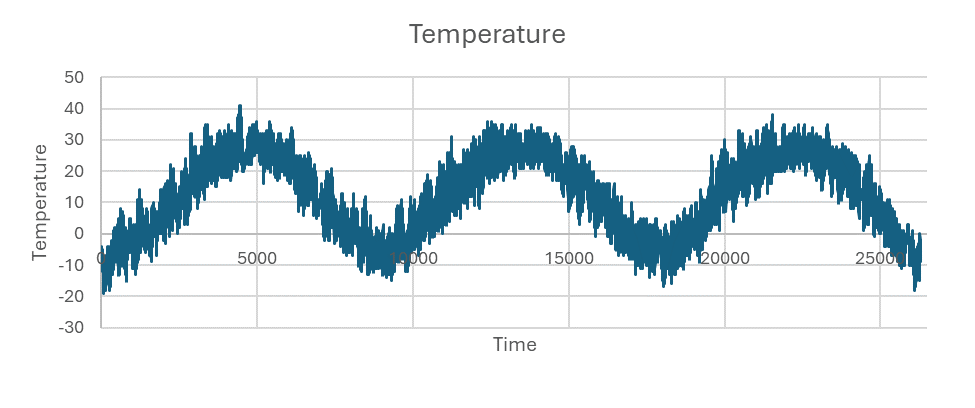
**Figure 10**

Temperature over time sample plot

*Note:* Created by the authors

**Figure 11**

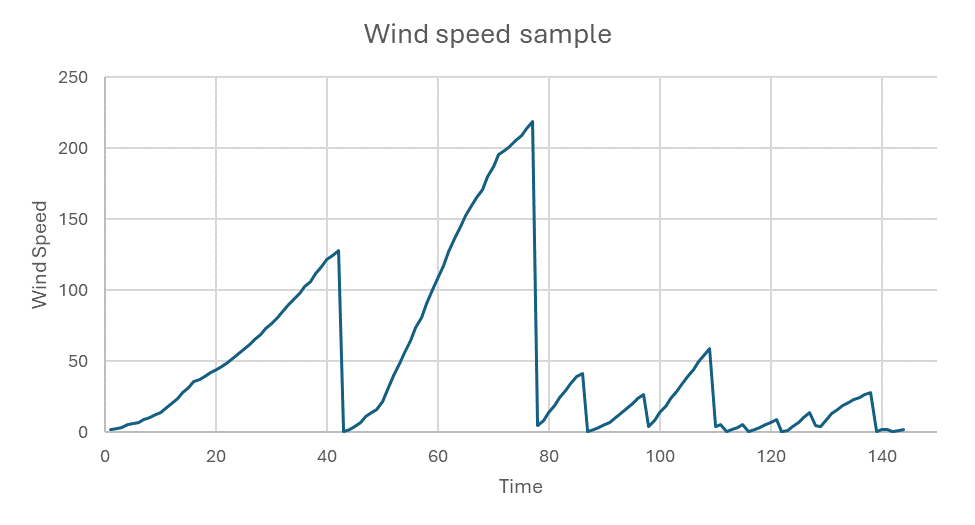
Temperature over time plot



*Note:* Created by the authors

**Figure 12**

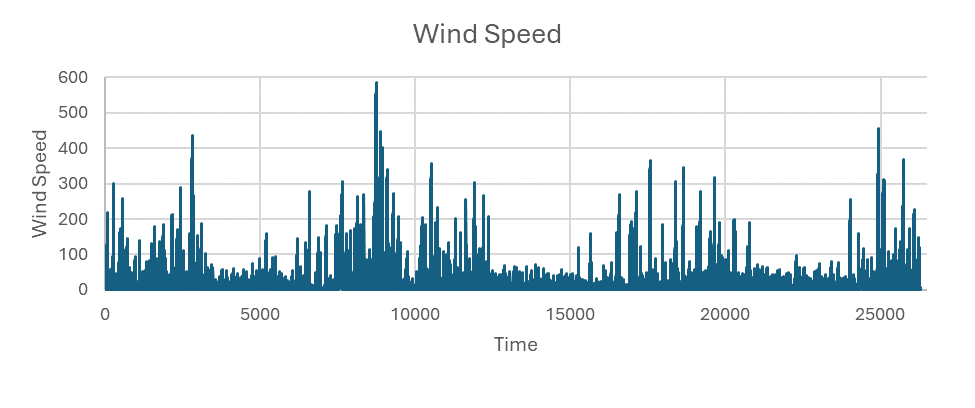
Wind speed over time sample plot



*Note:* Created by the authors

**Figure 13**

Wind speed over time plot



*Note:* Created by the authors

The authors used the Moving Average Method on the test data. As the records are for each hour of the day and the authors are predicting events for the next hour, they decided to use the past six-hour (six records) data to predict. After completing this analysis for pollution, wind speed, dew point, and temperature, the R square values the authors observed were – 0.7956, 0.7623, 0.9678, and 0.9132, respectively. This means that the moving average model predicts future events with an accuracy of 79.56%, 76.23%, 96.78%, and 91.32% for pollution, wind speed, dew point, and temperature, respectively.

Next, the authors used the Exponential Smoothing method. The authors used the training data to get the optimum value of alpha for each of the parameters to be predicted. After adjusting the alpha values, they were observed as – 0.98 for pollution, 0.78 for wind speed, 0.98 for dew, and 1 for temperature. Then, these alpha values were used in the test data to get the predicted values. After completing this analysis for pollution, wind speed, dew point, and temperature, the R square values the authors observed were – 0.9256, 0.8978, 0.9893, and 0.9832, respectively. This means that the exponential smoothing model predicts future events with an accuracy of 92.56%, 89.78%, 98.93%, and 98.32% for pollution, wind speed, dew point, and temperature, respectively.

Lastly, the authors used the auto-regression method. In auto-regression, they split the data day-wise. As the data had 24 records for one day, the authors split the data after every 24 records to get predictions. After using regression on this data, the authors got R square values for each of the four parameters. Except for temperature, all the other R square values were undesired. The R square value for temperature was found to be 0.9261, which meant that for temperature, auto-regression predicted future events with an accuracy of 92.61%.

# **Conclusion**

To understand the project's results, readers must realize how weather factors, such as temperature, pressure, wind direction and speed, and pollution levels, interact.

The descriptive analysis used in this study showed the distribution properties of important meteorological and pollutant factors, including skewness and kurtosis. In this way, the main trends in the data could be looked at in more depth. This study found places that are sensitive to sudden changes in the weather and confirmed that there are times when it rains a lot and when the temperature stays the same.  
 Lastly, this study explains how the different weather factors, like temperature, pressure, wind direction and speed, and pollution levels, follow specific patterns over time and how developing different regression and time series forecasting models for predicting future trends will prove helpful for public use.

# 

# **References**

Madhuri V. M., Samyama G. G. H., & Kamalapurkar S. (2020). Air pollution prediction using machine learning supervised learning approach. *International Journal of Scientific & Technology Research*, 9(4),118–123.

Gopalakrishnan V. (2021, January 25). Hyperlocal air quality prediction using machine learning. *Towards Data Science*. <https://towardsdatascience.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71>

Fatimetou Zahra Mohamad Mahmoud, “The application of Predictive Analytics: Benefits, Challenges & how it can be improved,” International Journal of Scientific and Research Publications, Volume 7, Issue 5, May 2017, ISSN 2250-3153.

Puma-Villanueva, W. J., dos Santos, E. P., Von Zuben, F. J., & IEEE. (2006, July 20). Data partition and variable selection for time series prediction using wrappers. *Data partition and variable selection for time series prediction using wrappers*, *1*(1), 4740-4747.

# **AIgarism**

