UCI Air Quality Analysis in Rome (March 2004 - April 2005)

Final Report

Data 2010: Tools and Techniques for Data Science
Professor: Ruwani Rasanjali Herath Mudiyanselage

Felix Vo (7924848)
Duc Do (8005501)
Thuan Khang Dinh (8003655)
Parth Ashvinbhai Pansara (7937067)

04 April 2025

I. Introduction:

The UCI Air Quality dataset provides a comprehensive record of pollution levels in Rome, Italy, capturing hourly measurements of key pollutants—including Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Total Nitrogen Oxides (NO_x), Benzene (C_6H_6), and Non-Methane Hydrocarbons ($NMHC$)—alongside sensor responses and environmental variables such as temperature and humidity. This dataset offers a unique opportunity to analyze sensor performance, assess air quality trends, and develop predictive models for pollutant concentrations.

This analysis is structured into four key components: sensor calibration and air quality assessment, which evaluates the correlation between sensor readings and true pollutant concentrations while detecting cross-sensitivities and analyzing seasonal AQI variations; regression-based CO prediction, which aims to develop models for estimating CO levels based on sensor and environmental data; time series and hybrid modeling, which explores temporal trends and advanced modeling techniques for pollutant forecasting; and machine learning comparison, which assesses different algorithms to determine the most effective approaches for air quality prediction and monitoring.

II. Preprocessing

The air quality data was initially cleaned by removing empty rows and columns, and by converting character columns—where commas were used as decimal points—into proper numeric format. The Date column was also converted from string format to a proper date type.

CO.GT.	NMHC.GT.	C6H6.GT.	NOx.GT.	NO2.GT.	T	RH	AH
1683	8443	366	1639	1642	366	366	366

PT08.S1.CO.	PT08.S2.NMHC.	PT08.S3.NOx.	PT08.S4.NO2.	PT08.S5.O3.
366	366	366	366	366

Number of missing data	n
0	827
1	6138
2	467
3	364
4	1195
9	26
10	291
11	6
12	12
13	31

To handle missing data, we first identified gaps using the is.na() and summary() functions. Since -200 was used as a placeholder for missing values, it was temporarily replaced with 0 to allow for five-number summary analysis and help determine which variables were most affected. During this process, we observed that the median value of NMHC_GT is 0, indicating that this variable may contain a large number of zero values, which could impact interpretation and imputation strategies.

CO.GT.	PT08.S1.CO.	NMHC.GT.	C6H6.GT.	PT08.S2.NMHC.	NOx.GT.	PT08.S3.NOx.
Min. :-200.00	Min. :-200	Min. :-200.0	Min. :-200.000	Min. :-200.0	Min. :-200.0	Min. :-200
1st Qu.: 0.60	1st Qu.: 921	1st Qu.:-200.0	1st Qu.: 4.000	1st Qu.: 711.0	1st Qu.: 50.0	1st Qu.: 637
Median : 1.50	Median :1053	Median :-200.0	Median : 7.900	Median : 895.0	Median : 141.0	Median : 794
Mean : -34.21	Mean :1049	Mean :-159.1	Mean : 1.866	Mean : 894.6	Mean : 168.6	Mean : 795
3rd Qu.: 2.60	3rd Qu.:1221	3rd Qu.:-200.0	3rd Qu.: 13.600	3rd Qu.:1105.0	3rd Qu.: 284.0	3rd Qu.: 960
Max. : 11.90	Max. :2040	Max. :1189.0	Max. : 63.700	Max. :2214.0	Max. :1479.0	Max. :2683

NO2.GT.	PT08.S4.NO2.	PT08.S5.O3.	T	RH	AH
Min. :-200.00	Min. :-200	Min. :-200.0	Min. :-200.000	Min. :-200.00	Min. :-200.0000
1st Qu.: 53.00	1st Qu.:1185	1st Qu.: 700.0	1st Qu.: 10.900	1st Qu.: 34.10	1st Qu.: 0.6923
Median : 96.00	Median :1446	Median : 942.0	Median : 17.200	Median : 48.60	Median : 0.9768
Mean : 58.15	Mean :1391	Mean : 975.1	Mean : 9.778	Mean : 39.49	Mean : -6.8376
3rd Qu.: 133.00	3rd Qu.:1662	3rd Qu.:1255.0	3rd Qu.: 24.100	3rd Qu.: 61.90	3rd Qu.: 1.2962
Max. : 340.00	Max. :2775	Max. :2523.0	Max. : 44.600	Max. : 88.70	Max. : 2.2310

After examining the data's summary statistics and missingness patterns, we found that some variables exhibit Missing Not At Random (MNAR) behavior, while others are Missing At Random (MAR). Imputation was then applied based on variable type: for time-series columns such as T, RH, and AH, we used multiple imputation methods including forward fill, backward fill, and interpolation. For correlated sensor variables, we applied KNN imputation, as these features included both continuous and categorical data. The improvements from cleaning and imputing were evident upon reviewing the updated five-number summaries.

CO.GT.	PT08.S1.CO.	C6H6.GT.	PT08.S2.NMHC.	NOx.GT.	PT08.S3.NOx.
Min. : 0.10	Min. : 647	Min. : 0.000	Min. : 383.0	Min. : 2.0	Min. : 322.0
1st Qu.: 1.00	1st Qu.: 941	1st Qu.: 4.000	1st Qu.: 741.0	1st Qu.: 96.0	1st Qu.: 642.0
Median : 1.70	Median :1072	Median : 7.900	Median : 920.0	Median : 169.0	Median : 796.0
Mean : 2.06	Mean :1116	Mean : 9.689	Mean : 956.8	Mean : 235.7	Mean : 822.8
3rd Qu.: 2.70	3rd Qu.:1253	3rd Qu.:13.600	3rd Qu.:1136.0	3rd Qu.: 309.0	3rd Qu.: 960.0
Max. :11.90	Max. :2040	Max. :63.700	Max. :2214.0	Max. :1479.0	Max. :2683.0

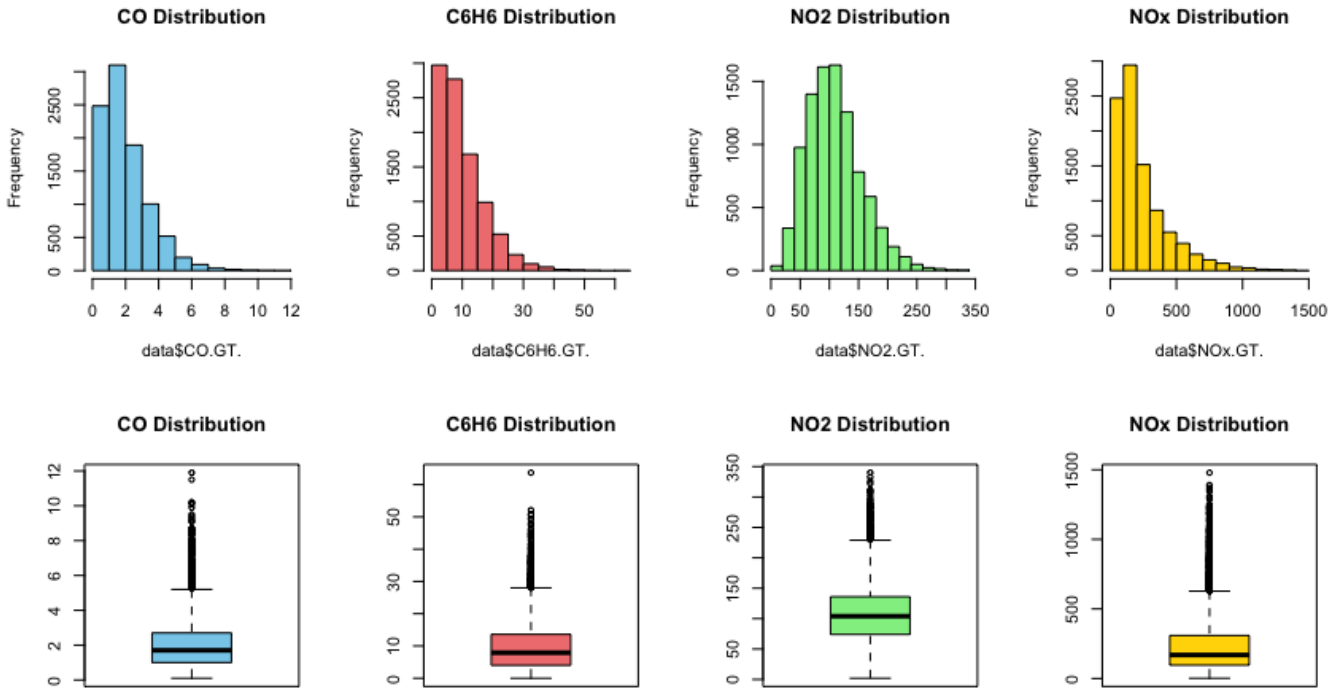
NO2.GT.	PT08.S4.NO2.	PT08.S5.O3.	T	RH	AH
Min. : 2.0	Min. : 551	Min. : 221	Min. :-1.90	Min. : 9.20	Min. :0.1847
1st Qu.: 74.0	1st Qu.:1242	1st Qu.: 742	1st Qu.:11.90	1st Qu.:35.40	1st Qu.:0.7262
Median :104.0	Median :1479	Median : 979	Median :17.60	Median :48.90	Median :0.9875
Mean :109.2	Mean :1480	Mean :1044	Mean :18.32	Mean :48.82	Mean :1.0174
3rd Qu.:136.0	3rd Qu.:1702	3rd Qu.:1307	3rd Qu.:24.30	3rd Qu.:61.90	3rd Qu.:1.3067
Max. :340.0	Max. :2775	Max. :2523	Max. :44.60	Max. :88.70	Max. :2.2310

III. Correlation Analysis and Air Quality Index

1. Data Characteristics and Correlation Analysis

1.1 Exploratory Data Analysis (EDA)

The Histograms reveal right-skewed distributions for all pollutants. This non-normality motivates the use of non-parametric correlation measures. Additionally, boximages highlight the presence of significant outliers across all pollutants. The consistent pattern of median concentrations being lower than the means confirms the positive skewness of the data.

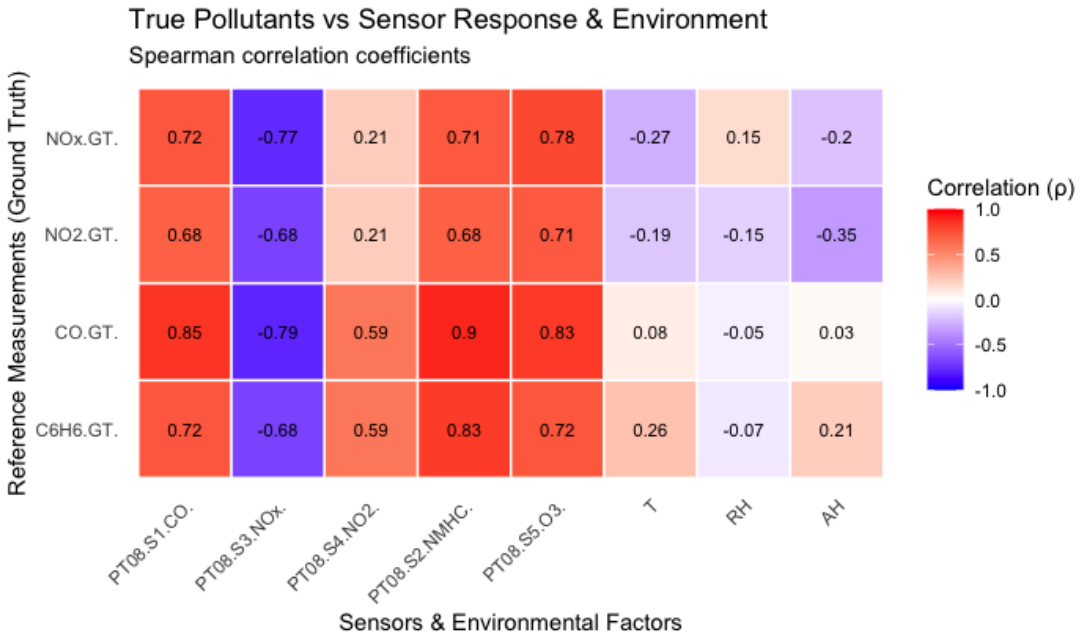


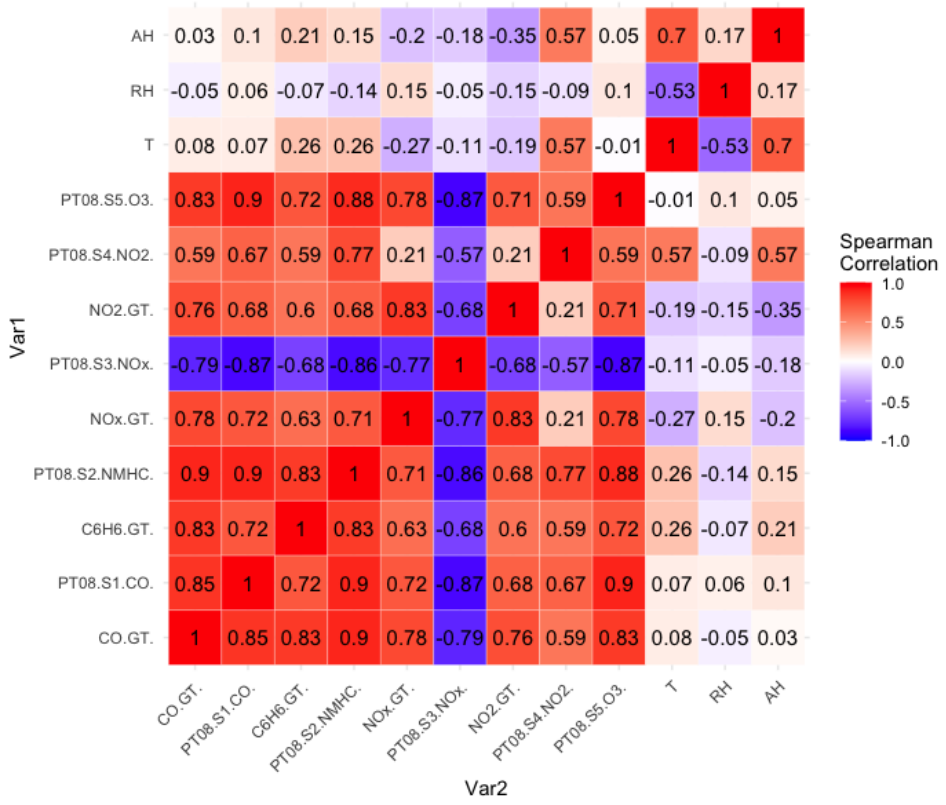
1.2 Correlation Matrix Interpretation

Methodology

Given the non-normal distributions and the presence of extreme values, we utilize Spearman's rank correlation (ρ), a non-parametric measure that captures monotonic relationships without assuming linearity.

Sensor-Pollutant Relationships:





The Correlation Heatmap reveals: The relationships between pollutants and sensors show a strong positive correlation for `CO` and `Sensor CO` ($\rho = 0.85$) and a strong negative correlation for NO_x and `Sensor NOx` (ρ = -0.77), while `NO2` and `Sensor NO2` have a weak correlation ($\rho = 0.21$). Pollutants exhibit weak or no correlation with environmental factors ($\rho \in$ (-0.35, 0.28)). However, cross-sensitivity detection suggests potential sensor interference, as `CO` correlates highly with `Sensor NOx, NMHC, and O3`, NO_x and `NO2` show moderate correlation with `Sensor CO, NMHC, and O3`, and C_6H_6 has high correlation with `Sensor NMHC` and moderate correlation with other sensors.

1.3 Robust Correlation Validation by Permutation Test

Methodology

To verify sensor-reference pollutant correlations beyond traditional correlation metrics, we conducted permutation tests (n = 1000 iterations). This non-parametric approach: First we compute the observed Spearman's ρ, then generating a null distribution by shuffling reference values, using those to calculate empirical p-values

Result

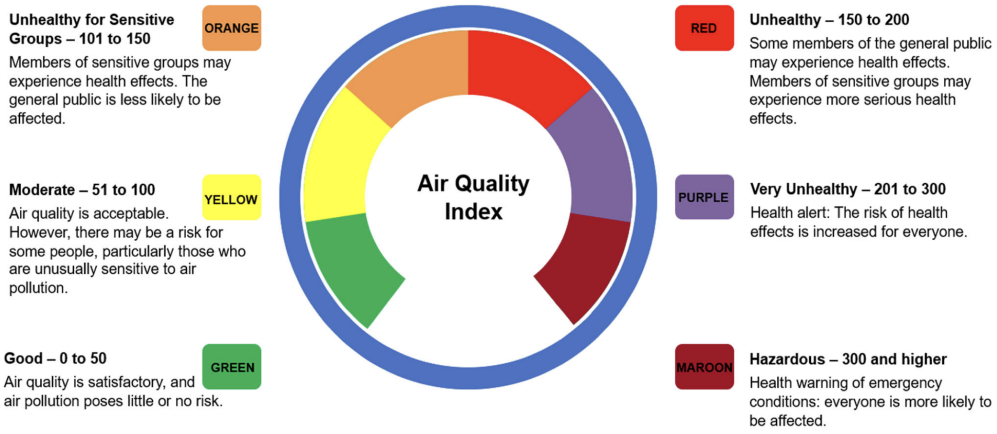
The permutation test for CO, NO_2 and NO_x against their respective sensors yielded empirical **p-value < 0.001**, we have sufficient evidence to conclude that there are strong and statistically significant correlation between sensor and corresponding concentration. As a **non-parametric test**, this validation method strengthens the confidence in our findings **without relying on normality assumptions**.

2. Air Quality Index (AQI) Analysis

2.1 Understanding AQI

What is AQI?

The Air Quality Index (AQI) is a standardized metric used to communicate the health risks of air pollution. It translates complex pollutant concentration data into a single, easy-to-understand value on a 0 - 500 scale, with higer values indicating worse air quality.



Custom AQI Methodology

In this study, we employ a **Custom AQI** methodology where CO and NO_2 follow standard **EPA AQI** breakpoints, benzene (C_6H_6) is assessed using **WHO exposure limits**, and NO_x is evaluated using the same breakpoints ranges as NO_2. This approach ensures alignment with established regulatory standards while incorporating WHO-recommended guidelines for benzene exposure.

2.2 AQI Calculation Methodology

The AQI is computed as:

$$AQI = \max\left(\frac{I_{HI} - I_{LO}}{BP_{HI} - BP_{LO}}(C - BP_{LO}) + I_{LO}\right) \text{ across all pollutants}$$

where:

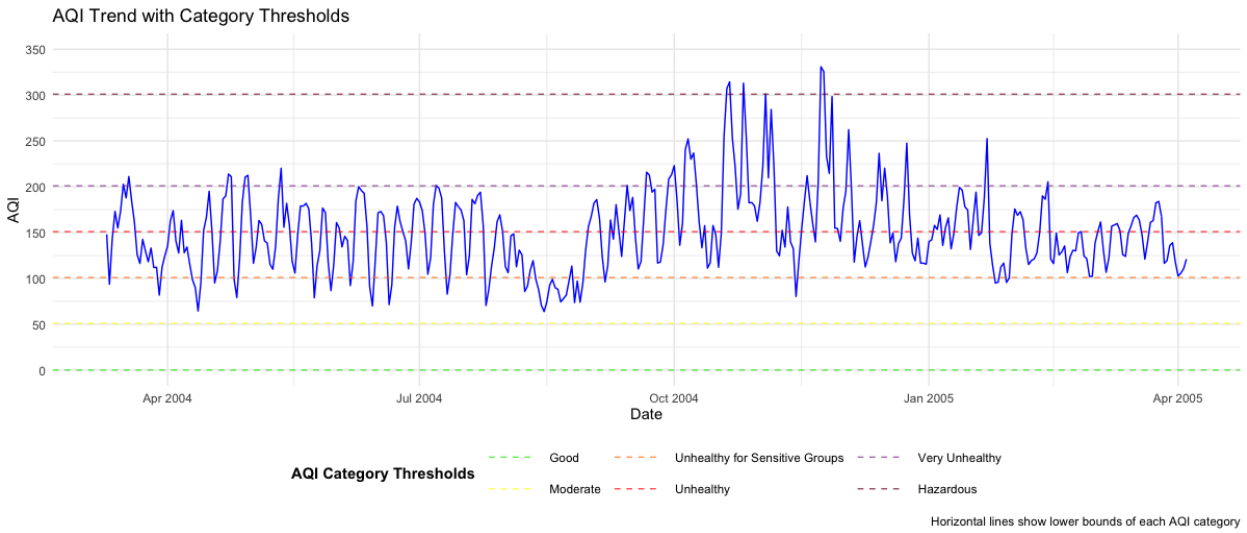
- $I_{HI/LO}$ = AQI category thresholds
- $BP_{HI/LO}$ = Breakpoints concentrations
- C = Observed pollutant concentration

Breakpoints reference table:

Category	$AQI Range$	CO	NO_2	NO_x	C_6H_6
Good	0-50	0-4.4	0-53	0-53	0-3
Moderate	51-100	4.4-9.4	53-100	53-100	3-7

Category	$AQIRange$	CO	NO_2	NO_x	C_6H_6
Unhealthy for Sensitive Groups	101-150	9.4-12.4	100-360	100-360	7-10
Unhealthy	151-200	12.4-15.4	360-649	360-649	10-15
Very Unhealthy	201-300	15.4-30.4	649-1249	649-1249	15-20
Hazardous	301-500	30.4-50.4	1249-2049	1249-2049	20-30

2.3 Temporal AQI Patterns



The AQI trend shows significant fluctuations, frequently exceeding the `Unhealthy for Sensitive Groups` and `Unhealthy` thresholds, with occasional peaks reaching `Very Unhealthy` levels. The highest peaks, reaching `Hazardous` range, are observed during late 2004. A sharp rise in AQI during Winter suggests increased pollution, likely due to heightened combustion activities, while Summer peaks align with high benzene (C_6H_6) contributions. Early 2005 shows a slight decline in AQI, though persistent fluctuations suggest evolving emission patterns and seasonal influences on air quality.

2.4 Pollutant-Specific Contributions

Season	Primary Pollutant	Contribution
Spring(2004)	C_6H_6	`87.95`%
Summer	C_6H_6	`90.22`%
Fall	C_6H_6	`76.92`%
Winter	NO_x	`67.78`%
Spring(2005)	NO_x	`54.29`%

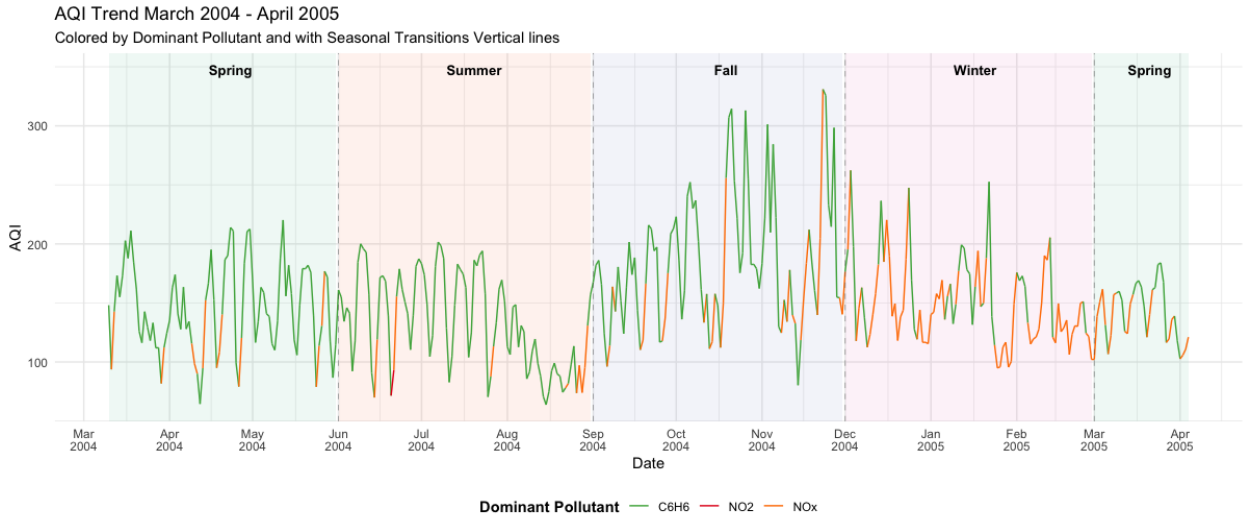
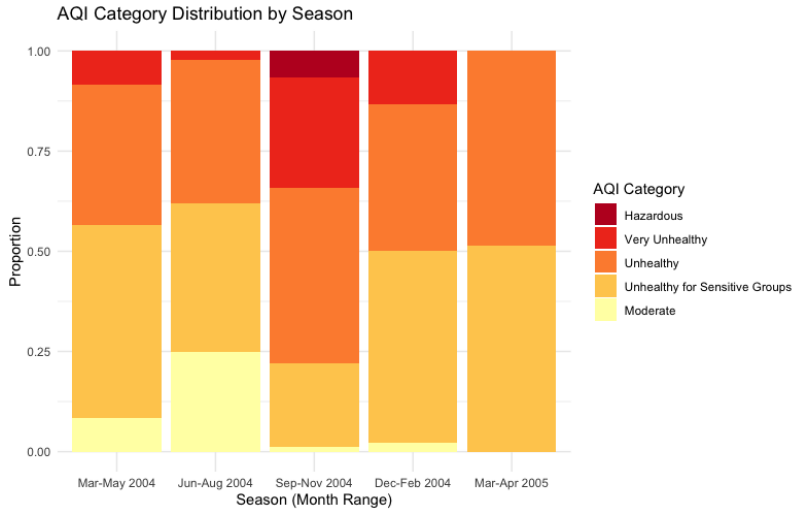


Table summarizes the seasonal contributions of individual pollutans to AQI, revealing that benzene (C_6H_6) is the dominant pollutant in Spring, Summer and Fall, accounting for over 75% of AQI excrescences, with a peak contributes of 90.22% during Summer. In contrast, Winter AQI is primarily influenced by NO_x , which contributes 67.78%, likely due to increased combustion activities. The transition period in early Spring 2005 marks a shift, where NO_x and C_6H_6 exhibit nearly equal contributions, indicating a changing pollution patterns over time.

3. Seasonal Variation Analysis



The AQI category distribution varies by season, showing a shift from moderate air quality in Spring (Mar–May 2004) to worsening conditions in Summer (Jun–Aug 2004) and Fall (Sep–Nov 2004), where "Unhealthy" and "Very Unhealthy" categories dominate, with a small proportion reaching "Hazardous" levels. Winter (Dec–Feb 2004) continues to have high pollution, with a majority in the "Unhealthy for Sensitive Groups" and "Unhealthy" categories, while early Spring (Mar–Apr 2005) shows slight improvement with a higher proportion of moderate air quality. This seasonal pattern suggests increased pollution in Fall and Winter, likely due to combustion-related emissions, while air quality improves slightly in transitional periods.

IV. Linear Regression-Based Prediction of CO(GT)

1. Regression Modelling for CO Prediction

Baseline Model In the Baseline Model, the RMSE was **1.416849**. This model simply predicts the mean CO(GT) for all instances without using any predictors. It serves as a basic reference point for evaluating other models.

Single Variable Model (CO Sensor) When using a Single Variable Model with only the CO Sensor, the RMSE dropped to **0.7662274**. This shows a noticeable improvement compared to the baseline, indicating that the CO Sensor captures important information about CO(GT).

Full Model (All Predictors) The Full Model, which uses all available environmental and sensor variables, achieved an even lower RMSE of **0.4108955**. This substantial reduction highlights that including multiple predictors further enhances model performance.

Model	RMSE
Baseline Model	1.416849
Single Variable (CO Sensor)	0.7662274
Full Model (All Predictors)	0.4108955

Overall, the results demonstrate that while the CO Sensor alone offers significant predictive power, combining it with other environmental variables leads to even more accurate predictions of CO(GT).

2. Impact of Including SensorCO in Linear Regression Models

To assess the importance of sensor data, particularly the CO sensor (`PT08.S1.CO.`), we compared the Linear Regression models: one using only environmental variables (Temperature, Humidity, and Hour), and another that included the CO sensor reading as an additional predictor.

Without CO Sensor The Linear Regression model using only Temperature, Humidity, and Hour resulted in an RMSE of **1.304**. These value indicate limited predictive power, suggesting that environmental factors alone are not sufficient to accurately estimate CO(GT) levels in a linear framework.

With CO Sensor After including SensorCO, the model's performance improved significantly, achieving an RMSE of **0.748**. This highlights the importance of sensor data, as it captures key variations in CO(GT) that environmental variables miss.sor input captures meaningful variance in the CO(GT) values, significantly boosting model accuracy.

Model	RMSE
Without CO Sensor	1.304
With CO Sensor	0.748

This shows that the sensor input captures meaningful variance in the CO(GT) values, significantly boosting model accuracy.

3. Introducing Random Forest Models: With and Without SensorCO

Random Forest is a powerful ensemble learning method that builds multiple decision trees and averages their outputs, allowing it to capture complex nonlinear relationships that linear models often miss. Unlike linear regression, which assumes a straight-line relationship between predictors and the target, Random Forest can adapt to patterns and interactions among variables, making it especially effective for modeling air quality data.

Without SensorCO When applied to our dataset, the Random Forest model using only Temperature, Humidity, and Hour achieved an RMSE of **1.087** showing modest improvements over linear regression.

With SensorCO However, after adding the SensorCO variable, the model's performance improved significantly, achieving an RMSE of **0.595**.

Model	RMSE
Random forest Without CO Sensor	1.087
Random forest with CO Sensor	0.595

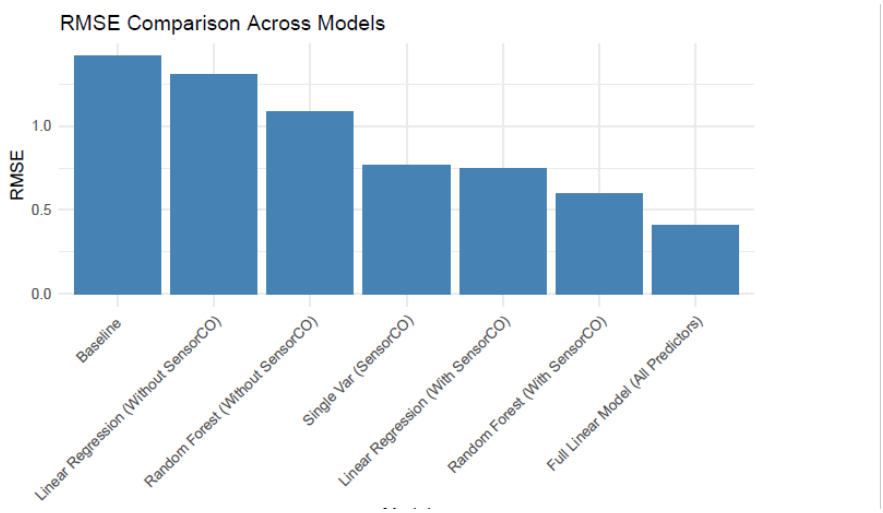
This confirms that Random Forest not only handles nonlinearities better but also maximizes the predictive power of key sensor inputs like SensorCO.

4. Evaluation and Model Comparison

To determine the best approach for predicting CO(GT), we compared several models based on their use of environmental variables, sensor inputs, and modeling techniques (Linear Regression vs. Random Forest).

Model	RMSE
Baseline Model	1.416849
Single Variable (CO Sensor)	0.7662274
Full Linear Model (All Predictors)	0.4108955
Linear Regression (Without SensorCO)	1.304
Linear Regression (With SensorCO)	0.748
Random Forest (Without SensorCO)	1.087
Random Forest (With SensorCO)	0.595

From the results, the Random Forest model with SensorCO clearly outperforms all other models across all evaluation metrics. It captures the non-linear relationships between the predictors and the target variable better than linear models, and the inclusion of SensorCO dramatically improves prediction accuracy.

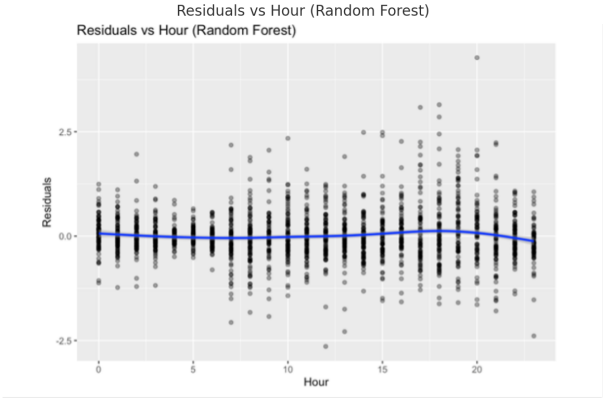


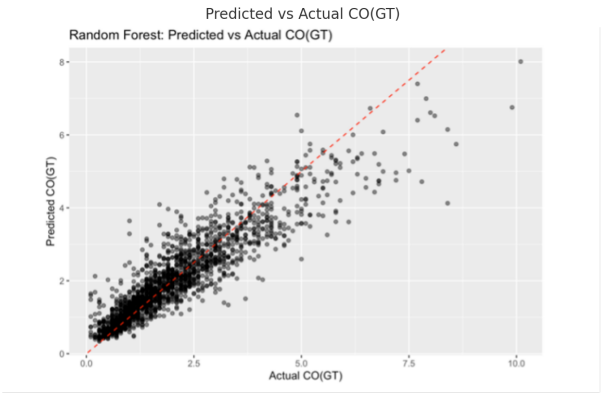
5. Diagnostic Analysis

Residual analysis was conducted to evaluate whether the models exhibited any systematic biases or time-based error patterns.

For the selected-variable model, residuals were plotted against the hour of the day, revealing no significant time-dependent trends or heteroscedasticity.

Predicted versus actual plots for the Random Forest model showed a tight clustering around the 45-degree line, especially for low and mid-range CO values. This confirmed that the model was accurately predicting most values and only slightly underestimating extreme values, likely due to fewer observations in the higher range.





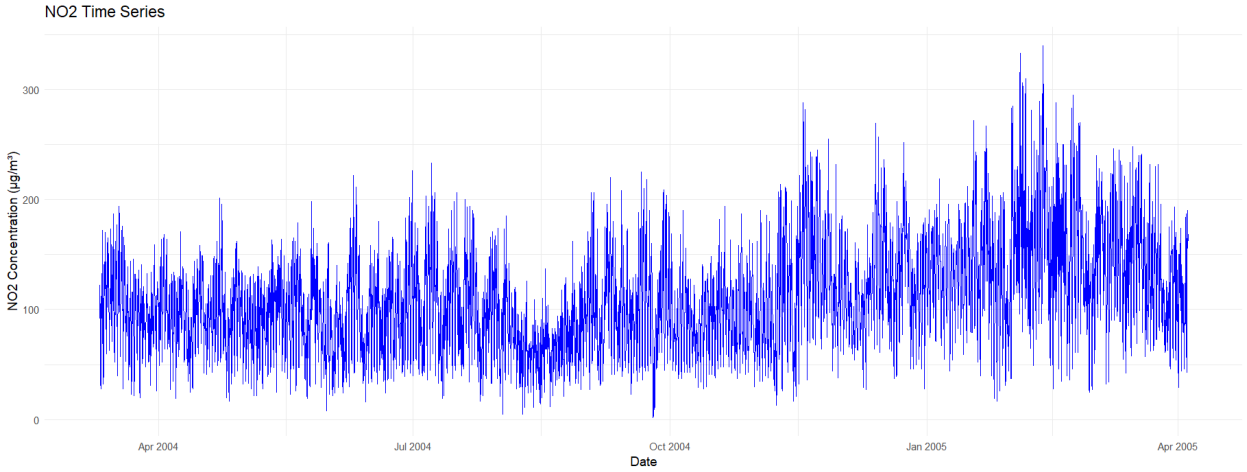
6. Conclusion

The full model lowers RMSE by including more predictors like NOx.GT. and C6H6.GT., but at the cost of overfitting and reduced interpretability. Using only Temperature, Humidity, Hour, and SensorCO offers strong performance ($R^2 = 0.83$, RMSE = 0.60) with simpler inputs. Random Forest models further enhance accuracy by capturing nonlinear relationships, making the selected-variable model both effective and practical.

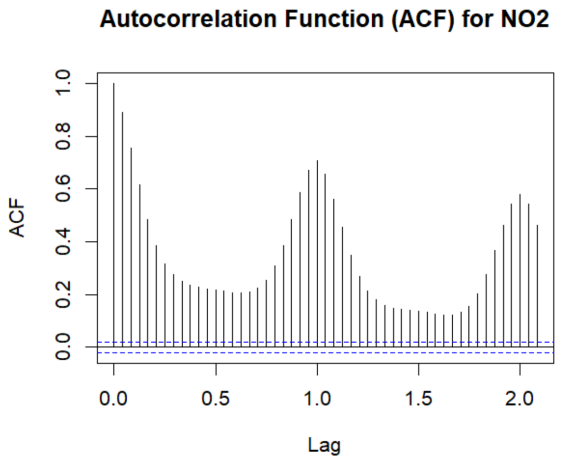
V. Hybrid Time Series Forecasting of NO_2 Using SARIMAX and SAPRC Simulation Proxies

1. Data Preprocessing Time Series Data and Stationarity Test

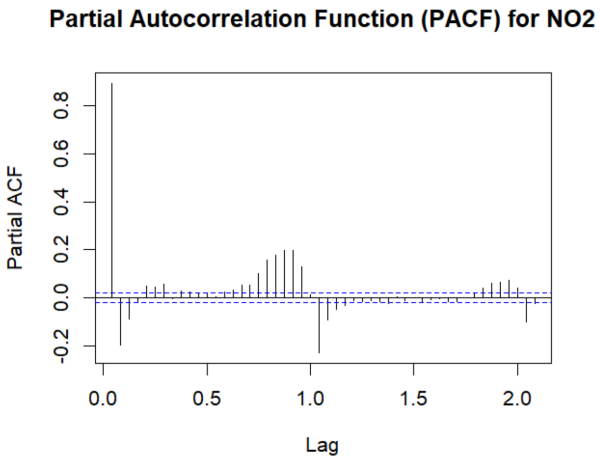
The Augmented Dickey-Fuller (ADF) test was applied to assess stationarity in the NO_2 series. Results p-value approximately 0.01, with a significance level of 0.05, the p-value is below the threshold, leading to the rejection of the null hypothesis. Consequently, there is evidence to suggest that NO_2 series in the data is stationary, hence, no transformation needed for NO_2. Although the ACF shows strong seasonality, the ADF test confirms stationarity (p = 0.01). The PACF suggests an AR(1) structure with possible seasonal components. A seasonal autoregressive (AR) model could still be appropriate.



ACF for NO2

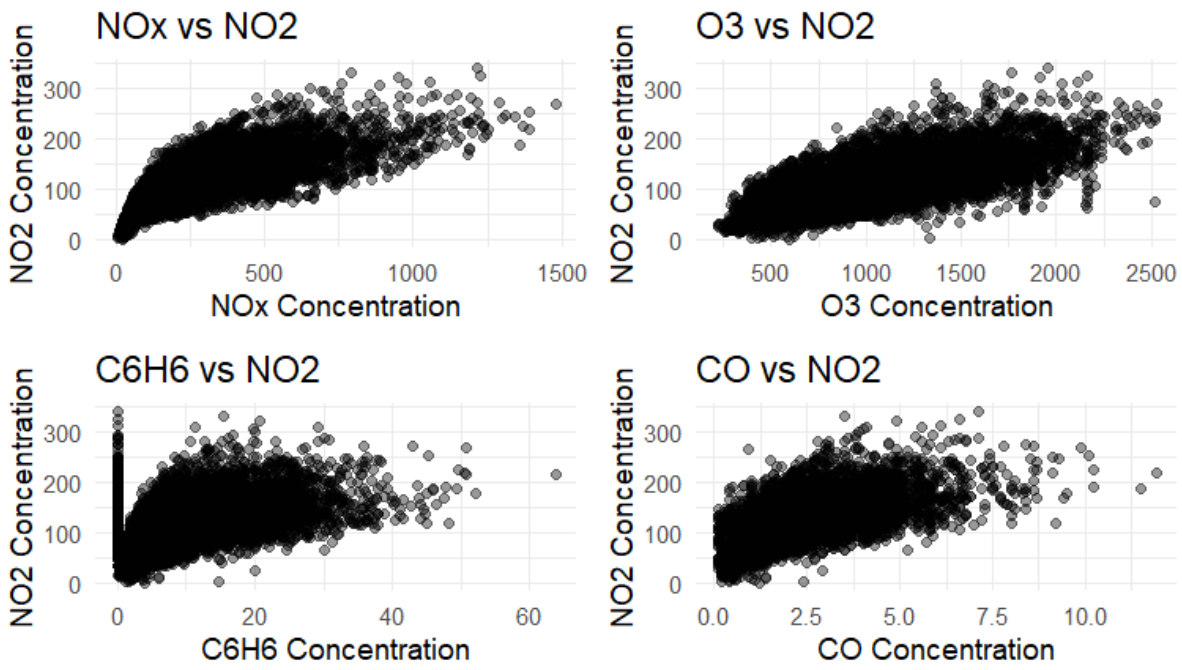


PACF for NO2



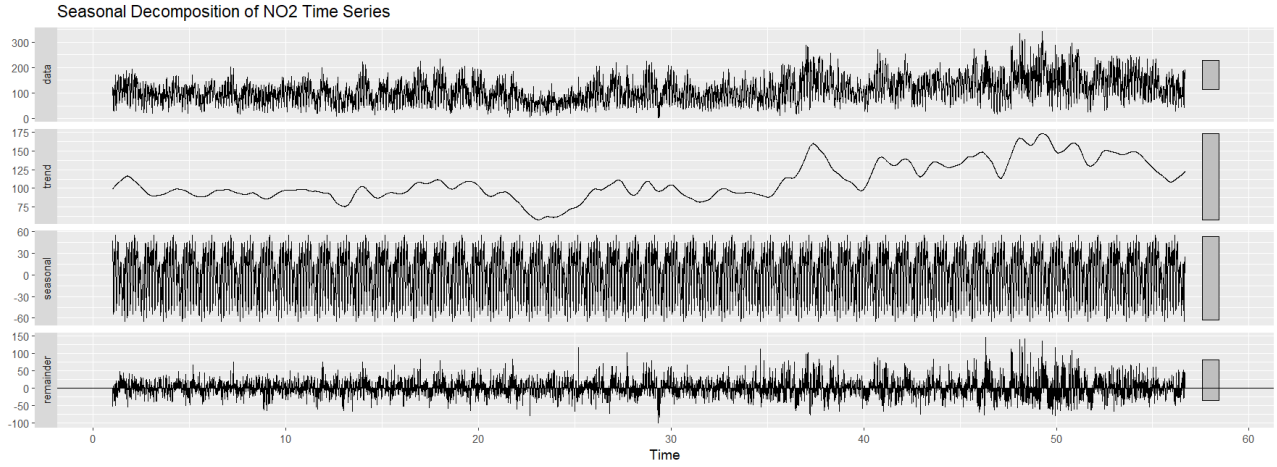
2. Exploratory Data Analysis (EDA)

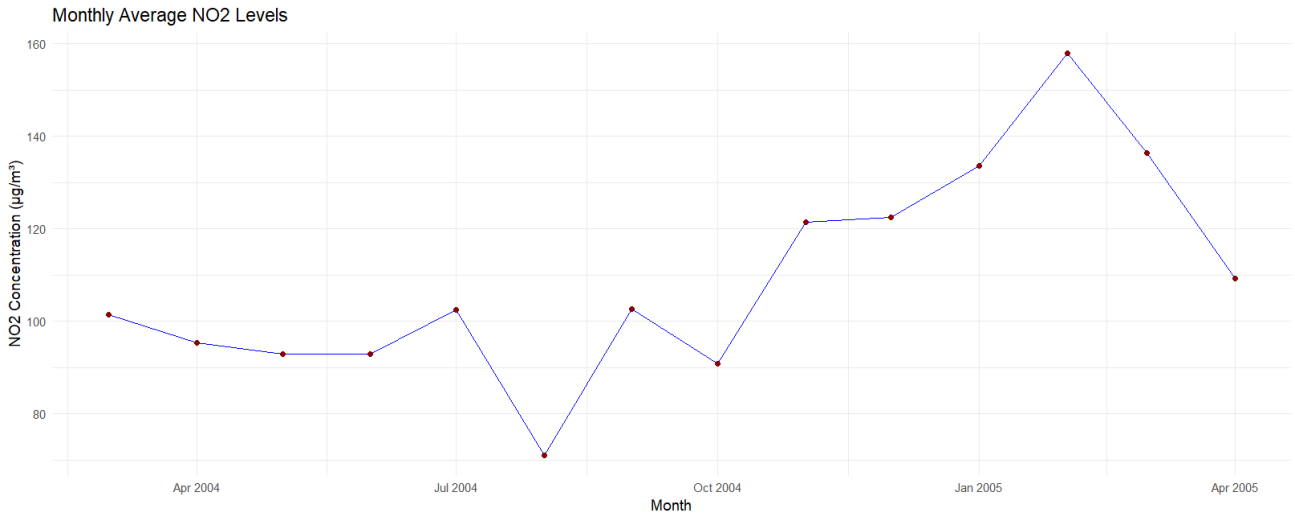
Key variables such as NO_x, CO, C_6H_6, and O_3 were examined in scatter plots to evaluate their relationship with NO_2. These plots revealed strong positive correlations, particularly between NO_2 and NO_x.



STL Decomposition of the NO_2 Time Series Using Seasonal-Trend Decomposition based on Loess (STL), the time series was broken down into trend, seasonal, and residual components. The trend captures gradual shifts in pollution levels over longer periods, while the seasonal pattern reflects consistent weekly cycles—possibly linked to differences in weekday versus weekend activity. The residual component isolates short-term fluctuations that are not explained by the trend or seasonality.

Higher concentrations of NO_2 were observed in the winter months (January–February 2005), with dips during summer, likely due to heating emissions and atmospheric conditions.



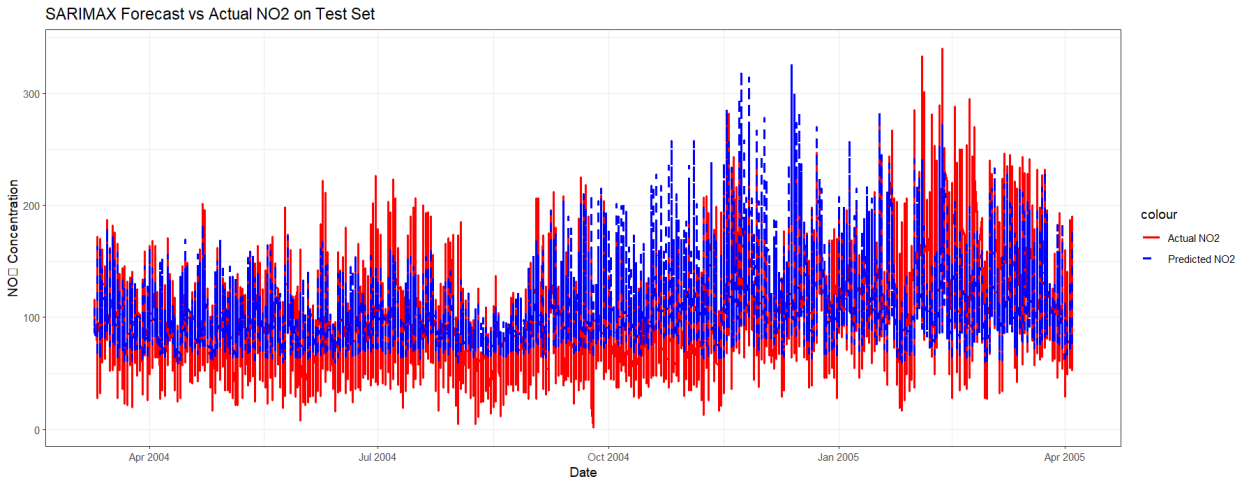


3. Time Series Analysis and Forecasting: SARIMAX

SARIMAX—Seasonal Autoregressive Integrated Moving Average with Exogenous Variables—is an advanced stochastic forecasting technique that extends the traditional ARIMA model. SARIMAX combines seasonal patterns (S), autoregressive (AR), Integrated Component (I), and moving average (MA) components to model time series data. It also incorporates exogenous variables (X) to account for external influences.

3.1 SARIMAX Baseline Modeling

The baseline SARIMAX model was developed using CO, NO_x, O_3 as exogenous variables. This achieved a Root Mean Square Error (RMSE) of **28.87**, indicating decent performance in capturing the general trend. However, the model underestimated peak values and exhibited lag during abrupt changes, showing limitations in short-term prediction.



3.2 Hybrid Modeling with SAPRC Proxy Variables

3.2.1 Motivation for Using SAPRC Proxies By simulating SAPRC chemical reactions, we can include intermediate species like PAN and HONO and general SAPRC proxy, which play crucial roles in NO_2 behavior but aren't directly measured.

Peroxyacetyl Nitrate (PAN) decomposition releases NO_2 under warm conditions:

$$\text{PAN} \xrightarrow{k_4} \text{CH}_3\text{C(O)OO}\cdot + \text{NO}_2$$

Nitrous Acid (HONO) photolysis contributes to NO production:

$$\text{HONO} + h\nu \rightarrow \text{NO} + \cdot\text{OH} => \text{NO} + \text{O}_3 \rightarrow \text{NO}_2 + \text{O}_2$$

3.2.2 SAPRC Proxy Equations Peroxyacetyl Nitrate (PAN) Approximation (Polynomial Regression):

$$PAN_{proxy} = a \times NOx + b \times NOx^2 + c \times C6H6 + d \times C6H6^2 + e \times O_3 + f \times T$$

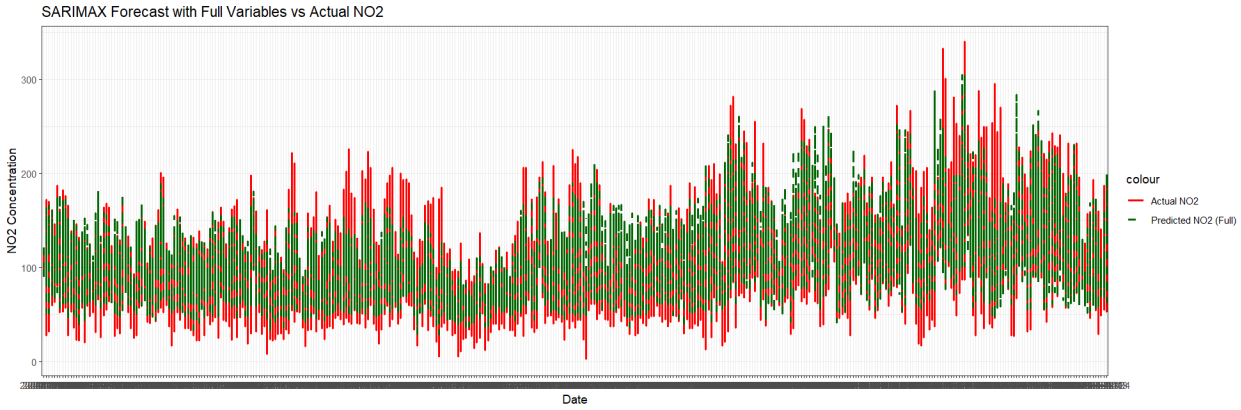
Nitrous Acid (HONO) Approximation (Non-Linear Regression):

$$HONO_{proxy} = f \times NOx + g \times RH + h \times (NOx \times Rh) + i \times T + j \times (NO_x \times T)$$

SAPRC Proxy Approximation (Linear Regression):

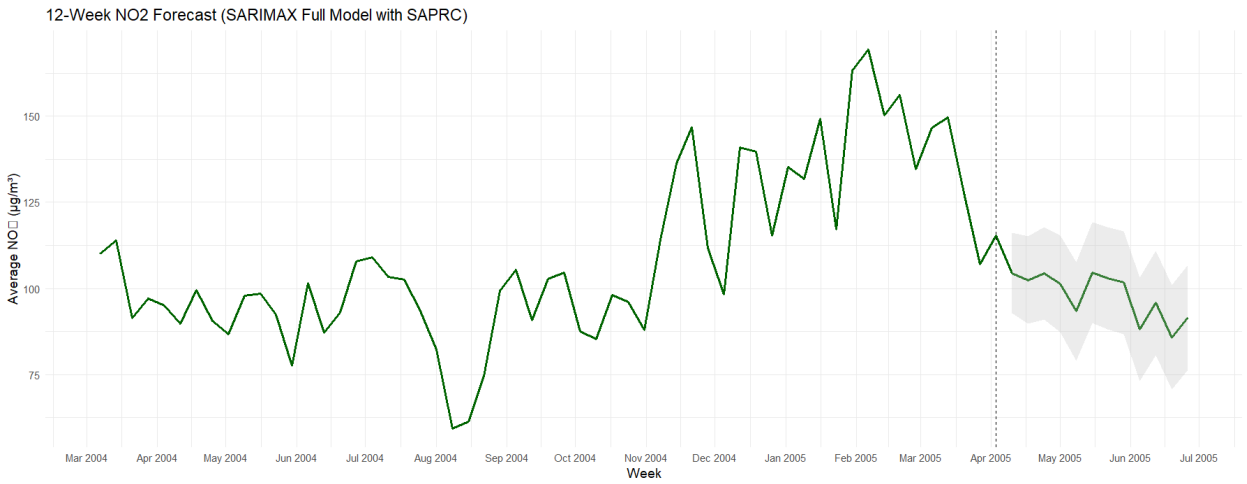
$$SAPRC_{proxy} = k \times NOx + p \times O_3 + q \times T + r \times RH$$

3.2.3 Hybrid SARIMAX with SAPRC: A comprehensive SARIMAX model was developed by incorporating the SAPRC-based proxy variables (PAN, HONO, and overall SAPRC chemical activity) as additional predictors. This hybrid approach significantly improved the model's forecasting accuracy, reducing RMSE to **21.73** compared to the earlier baseline.



3.3 12-Week Forecast of Weekly NO_2 Using SARIMAX with SAPRC

The 12-week SARIMAX forecast shows a gradual decline in NO_2 levels from April to July 2005, consistent with seasonal patterns seen in previous years. The model captures the typical drop in concentrations during warmer months, with uncertainty increasing over time. This suggests improved air quality is expected as the region transitions into summer. The widening 80% confidence interval reflects increasing uncertainty and variability in NO_2 predictions over time.



4 Conclusion

The forecasting confirmed strong seasonality in NO_2 levels and strong correlations with pollutants like NO_x. While the basic SARIMAX model captured trends (RMSE = 28.87), adding SAPRC proxies like PAN and HONO improved accuracy (RMSE = 21.73).The proxy variables better captured the chemical interactions affecting NO_2 concentrations, making predictions more accurate, especially during peak emissions and seasonal shifts.

VI. Machine Learning for Predicting Pollutant Concentrations and AQI Categories from Sensor and Environmental Data

1. Regression for Pollutant Concentrations Using KNN and Random Forest

We extract the **month** column as an integer feature. Before training, all covariates (excluding target variables) undergo **Min-Max normalization** to eliminate unit bias:

$$X_{normalized} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

The dataset is split into **80% for training** and **20% for testing**.

To ensure fair comparison across targets, we define the **normalized RMSE**, which expresses RMSE as a percentage of the test set mean:

$$\text{RMSE}^{\text{target}}_{\text{normalized}} = \frac{\text{RMSE}^{\text{target}}}{\text{mean}(\text{test}^{\text{target}})}$$

a. KNN Regression Results

Through experiments, `k=5` yielded the best performance.

Using Only Sensor Data

Target	RMSE	Normalized RMSE (%)
CO	0.5779	28.07%
C6H6	3.6012	37.14%
NOx	75.7171	32.31%
NO2	21.0551	**19.53%**

Using Sensor Data, Environmental Data, and Time (Month)

Target	RMSE	Normalized RMSE (%)
CO	0.4979	24.19%
C6H6	1.8391	18.97%
NOx	59.5803	25.42%
NO2	17.6079	**16.33%**

Including **environmental and temporal features** significantly improves prediction accuracy across all pollutants. Among them, regression performed best on **NO2**.

b. Random Forest regression

We use default parameter for `randomForest()` function.

Using Only Sensor Data

Target	RMSE	Normalized RMSE (%)
CO	0.4806	23.35%
C6H6	2.9008	29.92%
NOx	67.9865	29.01%
NO2	20.0721	**18.62%**

Using Sensor Data, Environmental Data, and Time (Month)

Target	RMSE	Normalized RMSE (%)
CO	0.4385	21.30%
C6H6	1.5624	16.11%
NOx	55.1452	23.53%
NO2	16.7216	**15.51%**

Likewise, adding environmental and temporal features greatly enhances prediction accuracy and the model performs best for **NO2**. Compared to KNN, Random Forest reduces normalized RMSE by an average of **4.0%** for sensor covariates and **2.1%** for all covariates, though it requires more training time.

c. Make use of 3 SAPRC Proxy variables to predict NO2(GT)

Since Random Forest generally outperforms KNN, we will use it for this part.

To improve prediction accuracy for **NO2**, we incorporate domain knowledge by adding three new variables from *Part V* (**PAN_proxy, HONO_proxy, SAPRC_proxy**) to the covariates.

Target	RMSE	Normalized RMSE (%)
NO2	13.9966	**12.98%**

Compared to the best previous performance on **NO2** (**15.51%**), incorporating domain knowledge reduces the normalized RMSE by **2.5%**, a notable improvement.

2. Pattern recognition visualization using PCA and K-Means clustering

Can we define a meaningful classification problem from this dataset? To explore this, we use visualization techniques to detect patterns, applying PCA for dimensionality reduction.

a. By Month

Month is the only categorical variable in the dataset, so we perform PCA using all sensor and environmental variables.

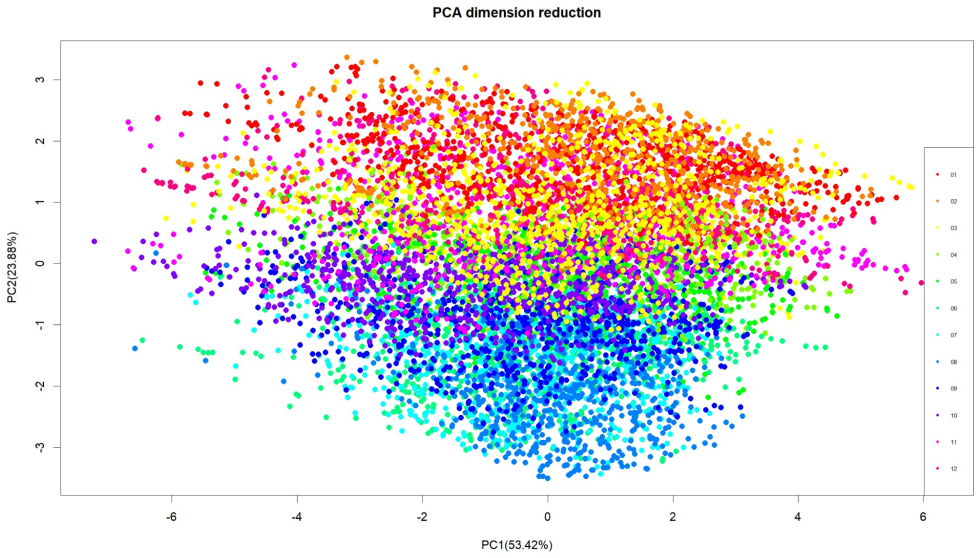


Figure 1: PCA by Month

A clear pattern emerges: months are clustered vertically, with colder months generally appearing in the upper half of the plot and hotter months in the lower half.

However, classifying months based on air quality data is not particularly meaningful. A more relevant classification task is **AQI group classification**. As observed in *Part III*, AQI is linked to seasonal variations, suggesting that a month-based pattern likely corresponds to patterns in pollutant levels.

b. By AQI Category

The target variable for classification is `AQI_Category` from *Part III*, computed using four pollutant concentrations. We perform PCA using all sensor, environmental variables, and the month variable, with points colored by `AQI_Category`.

Figure 2: PCA by AQI Category

As expected, a clear pattern emerges, indicating that this target is classifiable: pollutant levels increase horizontally from right to left.

c. Can K-Means Clustering detect any pattern?

Although the average silhouette method suggests that `k=2` yields the most meaningful clusters, we set `k=6` to examine whether K-Means can capture a similar pattern to `AQI_Category`.

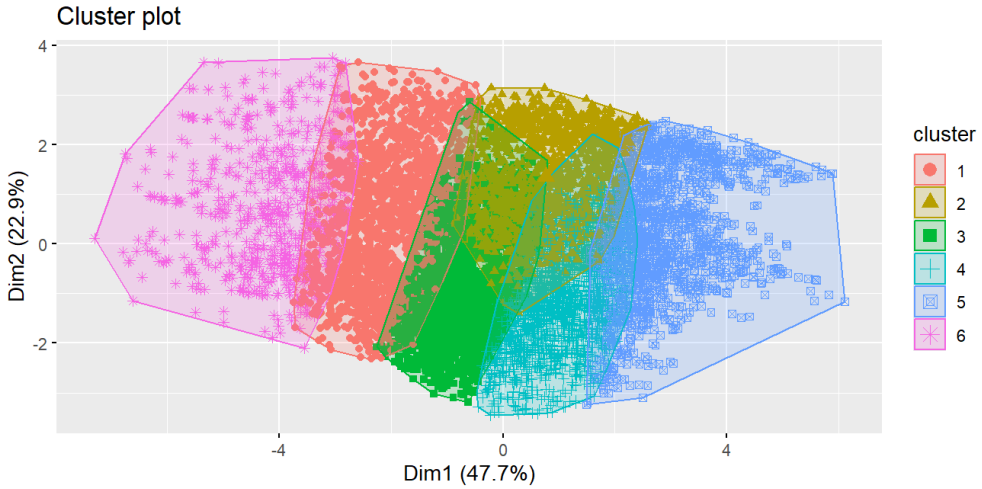


Figure 3: K-Means clustering

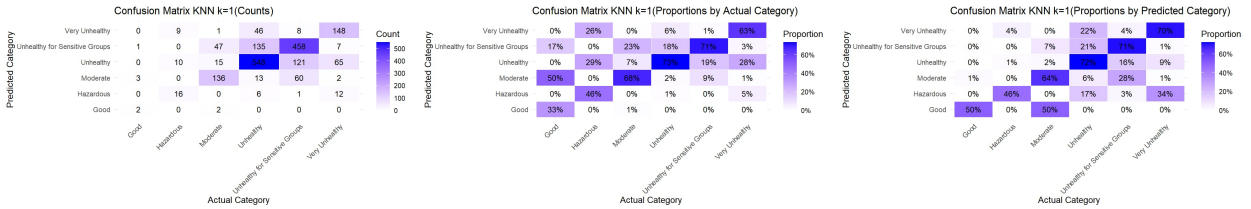
The result confirms this hypothesis: K-Means forms six clusters that roughly align with the horizontal trends observed in `AQI_Category` (*part b*). This finding is useful, as it suggests that pollutant levels can be categorized with reasonable accuracy using K-Means clustering alone, without the need for explicit AQI calculations.

3. AQI Category classification

It has been showed that `AQI Category` (pollutant levels) can be classified using month, sensor and environmental data. We will use `KNN` and `Random Forest`.

a. KNN

Through experiments, `k=1` yielded the best performance with **accuracy = 69.87%**. The details result are as follows:

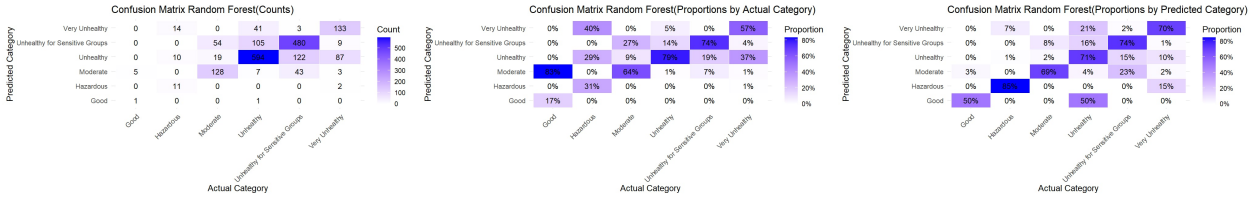


The performance is best for `Unhealthy` and `Unhealthy for Sensitive Groups`, which makes sense since the dataset is imbalanced, and these two groups represent the majority of the data. `Good` performs the worst as there are very few data points belonging to this group.

Some insights: **73%** of Unhealthy group are correctly predicted as Unhealthy.

b. Random Forest

The performance is slightly better than KNN with **accuracy = 71.96%**.



What we observed for KNN holds true for Random Forest as well.

Some insights: **79%** of the `Unhealthy` group are correctly predicted as `Unhealthy`. **85%** of data predicted as `Hazardous` actually belong to that group.

c. Regression to Classify

Another approach is to use regression to compute pollutant concentrations, then use those to calculate AQI and derive the AQI Category.

We use the best regression result from part *1.b* and achieve **accuracy = 57.59%**, which is much worse than just classification.

4. Conclusion

In summary, Random Forest outperforms KNN in predicting pollutant concentrations and AQI categories. Incorporating environmental and temporal features, as well as domain knowledge such as SAPRC proxy variables, significantly improves model accuracy.

While K-Means clustering provides a practical alternative for AQI categorization without explicit AQI calculations, regression to classify pollutants proves less effective than direct classification methods. These findings underscore the importance of selecting appropriate models and features for air quality prediction tasks.

VII. Final Conclusion

This project conducted a comprehensive analysis of air quality in Rome based on the UCI Air Quality dataset from March 2004 to April 2005. Through detailed preprocessing, correlation analysis, modeling, and machine learning, we extracted significant insights into pollutant behavior, sensor performance, and predictive modeling strategies.

Our study found that:

- **Sensor readings** generally correlated strongly with their corresponding pollutants, particularly for CO and NOx, validated through robust permutation tests.
- **Seasonal AQI analysis** revealed clear pollution patterns, with benzene dominating AQI during Spring–Fall and NOx during Winter. Seasonal combustion activities significantly influenced winter pollution levels.
- **Linear regression and Random Forest models** demonstrated that including sensor data, especially SensorCO, greatly improved prediction accuracy for CO concentrations. Random Forest with SensorCO achieved the best performance with an RMSE of 0.595.
- **Time series forecasting using SARIMAX** captured seasonal patterns in NO_2 levels. Incorporating SAPRC-based proxy variables (PAN, HONO, SAPRC proxies) further improved forecast accuracy by accounting for unmeasured chemical interactions, reducing RMSE from 28.87 to 21.73.
- **Machine learning techniques (KNN and Random Forest)** effectively predicted pollutant concentrations and AQI categories. Adding environmental and temporal features, and leveraging domain-specific proxy variables, further enhanced model accuracy. Random Forest outperformed KNN, achieving up to 71.96% accuracy in AQI classification.
- **Clustering techniques like PCA and K-Means** indicated that pollutant patterns aligned well with AQI categories, suggesting unsupervised learning as a viable method for AQI categorization.

In conclusion, effective air quality analysis and forecasting require a combination of sensor calibration, domain knowledge, advanced statistical modeling, and machine learning. Our results show that integrating chemical proxy variables and choosing appropriate modeling techniques can significantly enhance the accuracy and interpretability of air quality predictions, supporting more informed environmental decision-making.

Code Appendix

Follow this Github link for code appendix. ***Press link here***

References

- Technical Assistance Document for the Reporting of Daily Air Quality – the AQI, AirNow.gov. https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quailty.pdf
- Philadelphia Air Quality Report 2023 (Final Draft), City of Philadelphia. https://www.phila.gov/media/20241213133616/Item-12_Philadelphia-Air-Quality-Report-2023_Final_Draft.pdf
- AQI Breakpoints – Code Tables, EPA Air Quality System (AQS). https://aqs.epa.gov/aqsweb/documents/codetables/aqi_breakpoints.html
- AQI Basics – Understanding the Air Quality Index, AirNow.gov. https://www.airnow.gov/aqi/aqi-basics/?utm_source=chatgpt.com
- Benzene – WHO Guidelines for Indoor Air Quality: Selected Pollutants, World Health Organization. https://www.ncbi.nlm.nih.gov/books/NBK138708/
- Seasonality Detection in Time Series Data, GeeksforGeeks. https://www.geeksforgeeks.org/seasonality-detection-in-time-series-data/?ref=ml_lbp
- Time Series Analysis: Methods, Applications, and Best Practices, Built In. https://builtin.com/data-science/time-series-model
- A Complete Tutorial on Time Series Modeling in Python, Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/
- Seasonal Decomposition of Time Series by LOESS (STL), GeeksforGeeks. https://www.geeksforgeeks.org/seasonal-decomposition-of-time-series-by-loess-stl/
- Relative humidity dependence of HONO chemistry in urban areas, Stutz et al., *Journal of Geophysical Research* (2004). https://doi.org/10.1029/2003JD004135
- Mechanism of Peroxyacetyl Nitrate Formation, Hanst, *Journal of the Air Pollution Control Association* (1971). https://doi.org/10.1080/00022470.1971.10469527
- Peroxyacetyl Nitrates: Ozone in the Cooling PAN, Qureshi, *Atmospheric Chemistry* (Spring 2010). *(No direct link available.)*
- A Detailed Mechanism for the Gas-Phase Atmospheric Reactions of Organic Compounds, Carter, *Atmospheric Environment* (1990).
- Gas-phase Tropospheric Chemistry of Organic Compounds: A Review, Atkinson, *Atmospheric Environment* (1990).
- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics. https://doi.org/10.1007/b98835
- Lecture notes from MATH 2740: Math for Data Science, Professor Julien Arino, University of Manitoba, Fall 2024.