“Predictive Analytics of Air Quality Dataset in Urban Italy 2004 - 2005”

1) Introduction

This project explores the application of predictive analytics to a real-world air quality dataset, comprising hourly measurements of air pollutants, collected between March 2004 and February 2005 from a densely populated urban area in Italy at road level.[5] This report aims to analyze this time series data, assess their statistical properties, and develop a forecasting model to predict future values using time-series models Vector Autoregressive (VAR) techniques. The study investigates variable correlations, ensures stationarity, and compares the performance of VAR with alternative methods like ARIMA. It leverages predictive analytics techniques and data science tools for data manipulation, analysis, and visualization in a practical setting. The report is structured into sections covering methodology, results, and evaluation for each task, culminating in a conclusion that highlights key findings and their practical implications.

2) Methodology

I) Tools Overview:

The first presentation in the lectures on **“Chapter 1 & 2 - Overview of Predictive Analytics and Setting up the problem”** [6] suggests tools specifically for predictive analytics and time series purposes. These tools enabled efficient data manipulation, analysis, and visualization to achieve the objectives of the project. The tools include:

Python	Used as the primary programming language for data analysis and modeling.
Pandas	For data manipulation and preprocessing.
NumPy	For numerical computations.
Matplotlib and Seaborn:	Generate visualizations like histograms, scatter plots, and more.
Stats models	For data visualization and exploratory data analysis.
Scikit-learn	For time series modeling and statistical tests like KPSS and ADF.
Jupyter Notebook	For organizing and executing code in a step-by-step format.

II) Understanding the Dataset and Context (Task 1)

a) Data overview

The Air quality dataset contains 9358 instances of hour averaged readings from five metal Oxide chemical sensors collected by different sensor devices. It was collected between March 2004 and February 2005 from a densely populated urban area in Italy at road level and provides useful information about historical air quality trends. The dataset included 15 features as follows [5]:

- **Target pollutants:** Carbon Monoxide (CO), Non-Methane Hydrocarbons (NMHC - GT), Nitrogen Oxide (NOx), a Nitrogen Dioxide (NO2), (Ozone) (O3) and Benzene (C6H6)
- **Environmental data:** (T) The temperature, (RH) Relative Humidity, and (AH) Absolute Humidity.

- **Sensor response**: (PT08.S to PT08.S5) for specific pollutants.

This data includes missing values, represented (-200) and exhibits the characteristics of a multivariate time series, with the original dimension of 9358 rows and 15 columns, providing insight into air quality trends.

b) Data Preparation

1. **Data loading:** The dataset was downloaded through "AirQualityUCI.xlsx" file using pandas data frame with only 3 features were selected, carbon monoxide CO (GT), nitrogen dioxide NO2(GT), and relative humidity (RH) along with date and time features.
2. **Feature engineering:** Pandas was used for feature engineering to prepare a dataset for time series analysis. Because the time series analysis requires proper numeric data types, CO(GT), NO2(GT), and RH were retained as numeric (float) types, while the data date and time column were combined into a single column called **"datetime"**, formatted to enable time series indexing. The final dataset was 9357 rows and 4 columns (Datetime, CO(GT), NO2(GT) and RH).

c) Exploratory data analysis (EDA)

In this section, visual plots were created using **Matplotlib** and **Seaborn**, focused on three features CO(GT), NO2(GT) and RH to observe their trends and patterns. The plot shows the original trends from February 2004 - April 2005 with hourly measurements of air pollutants recorded overtime. Observed from the plot that the data had missing values (denoted as -200), indicating data gaps, spikes, dip and seasonal variations. **Correlation analysis, Mean and Median** were conducted to identify dependencies between the variables using Pandas and NumPy, providing insights into how these variables relate to one another.

III) Utility value (Task 2)

This step assessed the dataset's suitability for predictive analysis by focusing on **data preprocessing**, **stationarity checks**, and **determining whether to use univariate or multivariate models (VAR)**. The methodology as follows:

1. **Data Preprocessing and cleaning action:**
   Like other analysis techniques, the initial exploration was conducted to explore the dataset and identify any part that needs to be cleaned and adjusted for analysis purposes.

   a) Data Type Verification:
   Ensured that selected features **CO(GT)**, **NO2(GT)**, and **RH** were in numeric (float) format and the Datetime column was properly indexed with an hourly frequency for time series analysis. Since all pollutant levels and RH are already float 64, no additional changes are required for those

```
DatetimeIndex: 9357 entries, 2004-03-10 18:00:00
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   CO(GT)   9357 non-null   float64
 1   NO2(GT)  9357 non-null   float64
 2   RH       9357 non-null   float64
dtypes: float64(3)
memory usage: 292.4 KB
```

b) Handling Missing Values

After plotting the time series data, missing values (placeholders: -200) were identified with the following percentage of missing values: CO(GT): 17.99%, NO2(GT): 17.55% and RH: 3.91%. As those missing values are important for the analysis, they were addressed using interpolation, which estimates the value of missing values based on the surrounding trends and patterns to preserve the data's continuity and accuracy [1].

```
Number of NaN values after interpolation:
CO(GT)     0
NO2(GT)    0
RH         0
dtype: int64

Number of NaN values in each column:
CO(GT)     1683
NO2(GT)    1642
RH          366
dtype: int64
```

c) Duplicate Check:

This step confirms that there were no duplicate values in the dataset, -200 were properly handled during the data preprocessing process.

d) Post cleaning and verification:

Visual analysis using histogram and density plots after interpretation confirmed that the missing values are properly removed. However, retaining a high skewness and outlier in **CO(GT)** and **NO2(GT)**, while **RH** data was symmetric with no major skewness, requires further attention to solve the issue.

e) Addressing skewness and outliers:

The initial analysis of the dataset identified the high skewness and presence of outliers in CO(GT) and NO2(GT). In contrast, the histogram for RH showed that the data was already symmetric. To prepare the data for VAR and ARIMA models, it was essential to handle skewness as it can negatively impact the dataset, and ensure the data was stationary and systematic, as these models require those conditions for accuracy predictions. From several trials, the Box-cox transformation was applied, it was chosen because of its ability to reduce skewness while maintaining positive values.[1] The step taken were as followed:

1) **Applying Box-Cox transformation:**
    a) Box-cox method automatically determines the optimal transformation parameter to minimize skewness for each variable.
    b) Plot the transformed distributions to confirm the skewness has been reduced.
    c) Skewness values were calculated before and after Box-Cox to quantify the improvements.
2) **Density and correlation plots:**
    a) Visualized the transformed data using density plots to assess the impact on the distributions and confirm improvements.
    b) Correlations between the variables were checked to ensure the transform did not alter the relationship.

    However, the **Relative Humidity (RH)** column retained its original form because applying the **Box-Cox transformation** to RH would unnecessarily introduce skewness. Since the RH data was already symmetric, further transformation was not required. The dataset was finalized with the following columns for modeling : CO(GT_boxcox) NO2(GT_boxcox) and RH.

2. **Stationary check of the time series**
   To ensure the dataset was suitable for time-series modeling, the stationarity of the series was verified. It is important to to ensure stationarity, where the mean and variance are constant overtime and the covariance between yt and y+k depends only on k. [3] The following step were taken:

   a) **Data Preparation:**
      The dataset was preprocessed to handle missing values, outliers, and skewness. Confirmed that Datetime indexes are proper format for time series analysis, ensuring that the time index is relative and not absolute.

   b) **Stationarity testing with A unit root test**
      - Augmented Dickey-Fuller (ADF) is the measure that helps in identifying whether a time series is stationary or non-stationary. A p-value close to 1 indicates that there is likely a unit root, meaning that the series is not stationary. A p-value closer to 0 means that we can likely reject the assumption that there is a suggestion for the presence of a stationary.[4]
      - If non stationary, transformations as differencing or lag scaling were considered

   c) **Autocorrelation analysis:**
      We perform the autocorrelation using ACF and PACF plots analysis, the following technique are employed:
      - **Autocorrelation (ACF):** Measures the correlation between a time series and its lagged values, identifying how correlated the values are with each other. Time series that show no correlation are called white noise.
      - **Partial autocorrelations (PACF):** gives the partial correlation of a stationary time series with its own lagged values, regressing the values of the time series at all shorter lags. The PACF plays for AR terms The PACF plot helps in determining the value of p. [4]

3. **Univariate vs. Var model:**
   This step investigates whether the three time series (CO(GT), NO(GT) and RH) should be treated as a univariate time series or a multivariate Vector Autoregressive (VAR) for modeling. Utilize the Granger causality test to determine if one variable can predict another by analyzing lag relationships. If there is casualty, a VAR model may suit most for these prediction analysis. The methodology includes:

- **Stationery requirements:** Both Univariate models (e.g., ARIMA) and multivariate VAR models required stationary data. First, we ensure that the data was processed and stationary before running Granger causality test.

- **Find the best lag selection:** Before the granger test, to avoid trial and error in lag selection, information criteria **Akaike Information Criterion (AIC)**, **Bayesian Information Criterion (BIC)**, **Hannan-Quinn Information Criterion (HQIC)**, and

Final Prediction Error (FPE) were used to identify the optimal lag. Based on this criteria the lag of 10 was selected for the casualty test.

- **Granger Causality Testing:** The test checks if lagged values of one variable (e.g., CO(GT) improve the prediction of another variable (e.g., NO2(GT)), perform the Granger test for all pairs of time series. Used the **grangercausalitytests() function from Statsmodels with the determined maxlag (5).**

- **The output** includes the statistical output for each lag like F statistic and associated p value so if the p-value for each lag is significant (e.g., p-value < 0.05), the lagged variable causes the other variable.

The confirmation of casual relationships through Granger causality testing, the analysis validated the selection of the most suitable model for forecasting air quality dynamics.

IV) Analysis, modeling and prediction (Task 3)

The objective of this step was to perform a short term prediction using the three time series variables CO(GT), NO2(GT), RH. The following methodology was implemented:

1. **Split data into Train and Test sets:**
    - Display all the column names to ensure the dataset included the necessary features : Index CO(GT) , NO2(GT), RH , CO(GT_boxcox), NO2(GT_boxcox) , dtype=object.
    - Split the dataset into:
        - **Train set:** comprising 90% of the data.
        - **Tesst set:** comprising the last 24 points for validating the prediction.
    - **Verified split shape:**
        - Train Set Shape: (9333, 3)
        - Test Set Shape: (24, 3)
2. **VAR Model and Lag Selection Order (p):**
    - **Optimal Lag selection:**
        - The optimal lag order (p) using information criteria, (AIC, BIC, HQIC, FPE).
        - Lag order of **15** was selected based on these criteria.
3. **VAR Model Fitting:**
        - The VAR model was trained using the training set and summarized the results (Coefficients, t-statistics, p-values).
        - Evaluated residuals using:
            - **Durbin-Watson statistic:** Check for uncorrelated residuals.
            - **Ljung-Box Test:** Verified no autocorrelation in residuals.
4. **Forecasting and Evaluate with Test Data:**
   To evaluate the accuracy and reliability of the predictions, forecasted data and the actual data were compared. The following steps were implemented:

    a. **Preparation for comparison:**

        - Used the (df_original as an original dataset for comparison, confirming that it was free of missing value and without Box-Cox transformation.
        - Then, applied inverse Box-Cox transformation to forecasted values to match the original data scale.

    b. **Data Combination:**
        - Combine the actual test data and forecasted data into a single dataframe for easier comparison.
        - Compare the transformed forecast values with the original test data, plotted actual values from (df_original) vs. forecasted values for all features in a single chart to visualize performance.

    c. **Short term prediction:**
        - Predicted the next 24 steps(hours) that match the test set length.
        - Convert the forecasted value to a dataframe for comparison with the test set.
        - Forecasted an additional the next 2 days (48 hours, beyond the test set) to evaluate long term trend and model reliability.

    d. **Visualization and Comparison:**
        - Plotted actual vs. forecasted values for all variables, **CO(GT)**, **NO2(GT)**, and **RH** were displayed in a single chart to visualize performance.

5. **Evaluation metrics overview:**

It is important to use an evaluation metric to evaluate the model performance in the statistical tasks, providing the quantitative evidence to support the results, especially when comparing different models. In this study, we introduce seven distinct metrics to comprehensive evaluate the VAR model, including [4] :

- **MAE (Mean Absolute Error)**: Measures the average distance between Predicted and original values. Lower values mean better performance.
- **RMSE (Root Mean Squared Error)**: Penalizes larger errors more heavily. Lower RMSE values means more accurate predictions, particularly with the model with sensitive outliers.
- **MAPE (Mean Absolute Percentage Error)**: Expresses accuracy as a percentage of actual values and the model's predictions, with lower values indicates better performance.
- **ME (Mean Error)**: Indicates overall error in the predictions, identifies whether the model consistently over-predicts or under–predicts.
- **MPE (Mean Percentage Error)**: Shows the average of mean error from actual values.
- **Correlation:** a statistical measure, values close to 1 0r -1 show strong relationships, while near 0 mean weak or no relationship.
- **Min max Error:** calculate a minimum and maximum error of each measurement for the final results, indicating the deviations and insights into the model's worst-case performance.

These metrics were used to provide a comprehensive evaluation framework, which helps identify a model's strengths and weaknesses across multiple evaluation of

accuracy, bias and variability. The combination of these metrics aids in comparing VAR model's performance with alternatives experiments like ARIMA.

V) Bonus Task: Fitting the ARIMA report:

In this task, the Arima model applied for each variable individually CO(GT, NO2(GT), and RH - to to evaluate their performance as univariate forecasting models. The goal is to compare ARIMA's performance with the multivariate VAR and evaluate its suitability for air quality time series dataset. ARIMA, well known as a linear modeling technique, helps predict data as a nonlinear component and requires that the data be made stationary before fitting a linear equation to the data.[2] In order to evaluate forecasting accuracy and suitability for univariate times series prediction and comparing it with the VAR model . The methodology followed these steps:

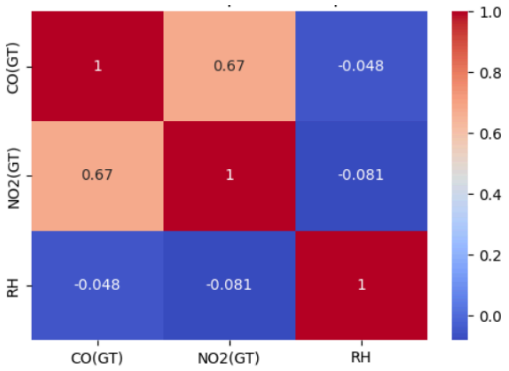
1. **Data Splitting:**
   The data was split similarly to the VAR model, (9333 instances) and Test set with the last 24 points for validation.
2. **Arima parameter selection**:
   Unlike the VAR model that uses information criteria (AIC/BIC) to define lag orders. ARIMA requires identifying the optimal parameters (p,d,q) p (autoregressive terms), d (differencing order), and q (moving average terms) for each time series. To achieve the most accurate results, the auto_arima were employed to find the best parameters, avoiding manual inspection on ADF and PACF plots. Optimal parameter are **CO(GT_boxcox)**: (5, 1, 5), **NO2(GT_boxcox)**: (2, 1, 2) and RH (3,1,2)
3. **ARIMA model fitting and forecasting:**
   - We fit the ARIMA model independently for each variable using the selected parameters.
   - Forecasted the next 24 steps (matching the test set length). The steps include creating the data frame, Applying inverse Box-Cox transformations for CO(GT) and NO2(GT) to revert predictions to the original scale, while RH retains the original form.
4. **Evaluation metrics:**
   Similarly with Evaluation metrics overview, standard metrics such as MAE, RMSE, MAPE, ME, MPE Correlation, Min max Error were used for the ARIMA model to evaluate its performance. Supporting the quantitative evidence when comparing VAR's model performance with ARIMA.
5. **Comparison with VAR model:**
   - Actual and forecasted values were plotted for each variable.
   - Create a table of evaluation metrics to compare the strength and weakness of ARIMA versus VAR.

3) Results and Evaluation (Task 4)

The analysis of the air quality dataset, focusing on CO, NO2, and RH, has provided several insights into predictive modeling and time series analysis. The results and analysis from this study are summarized below:

I) Task 1: Understanding the Dataset:

The data represent a multivariate time series for analyzing interactions and dependencies between each variable. The visualization revealed seasonal variations particularly in NO(GT) with a higher value during colder months. Strong correlations were observed between CO(GT) and NO2(GT) while RH showed limited correlation but likely influenced pollutants dispersion **(Fig. 1)**. Additionally, the data also represent frequent missing values (-200). The findings summary are illustrate in the the table of observations as followed:



Variable	Missing values	Trend analysis	Distribution	Dependencies	Statistic
Co	17.986534%	No consistent patterns, Frequently missing values (-200), spike and dips likely due to sensor issue.	Large skewed with a peak near 0 and high outliers.	Found correlations with NO2(GT) with share emission source.	Mean: -34.2075 Median: 1.5000
NO2	17.548360%	Higher value in the colder months during winter, likely because seasonal patterns or emission related spikes.	broader spread with a peak near 100.	Similarly, correlations with NO2(GT) share emission sources.	Mean: 58.1359 Median: 96.0000
RH	3.911510%	Stable, occasional shape drop likely due to sensor error or environmental changes.	Looked normal but slightly negatively outliers.	Might not directly generate pollutants but can lead to pollutant dispersion and transformation.	Mean: 39.4836 Median: 48.5500

II) Task 2: Utility Value

a) Data preprocessing:

1. Handling missing values:

Missing values (17.99% for CO (GT), 17.55% for NO2 (GT) and RH (3.911510%)) were filled using interpolation techniques. This approach is more feasible to use when missing values are not scattered too much to avoid dropping rows and losing a significant portion of the dataset. The results as shown in **Fig 2**, the frequency distribution for **CO(GT)**, **NO2(GT)**, and **RH** before and after preprocessing, -200 were properly replaced.

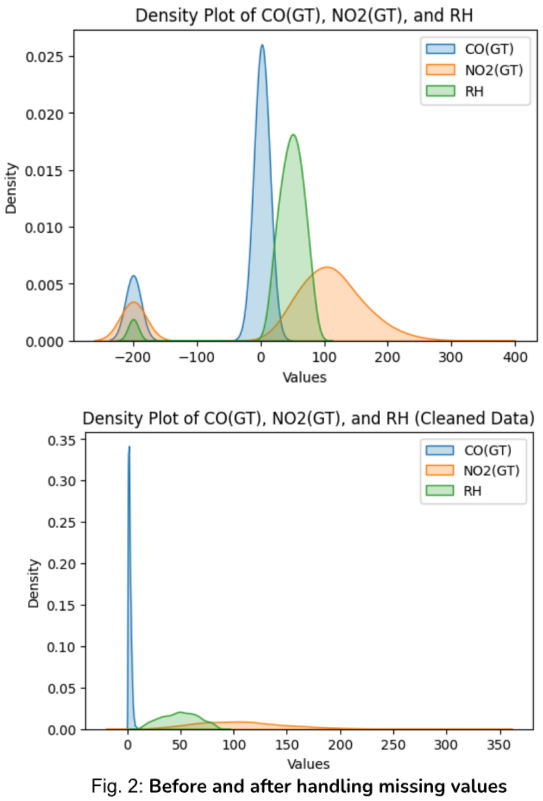


Fig. 2: **Before and after handling missing values**

2. Skewness and outlier:

CO(GT) and NO2(GT) show high skewness in their original distributions with **CO(GT)** original Skewness: 1.2923 (After Box-Cox: -0.0176) and **NO2(GT)** original Skewness: 0.7072 (After Box-Cox: 0.0076). The results prove that the Box-Cox

transformation effectively reduced skewness in **CO(GT)** and **NO2(GT)** while retaining the statistical properties of the data. While RH original skewness is -0.043 close to 0 suggest no transformation is required and distribution is close to normal, so retain its origin as it was already symmetic. The visual chek uses histogram plots **(Fig 3)** to confirm improved distributions, highlighting successful reduction in skewness for **CO(GT)** and **NO2(GT).**

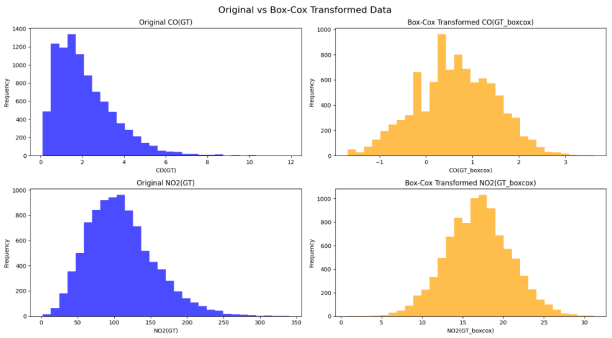
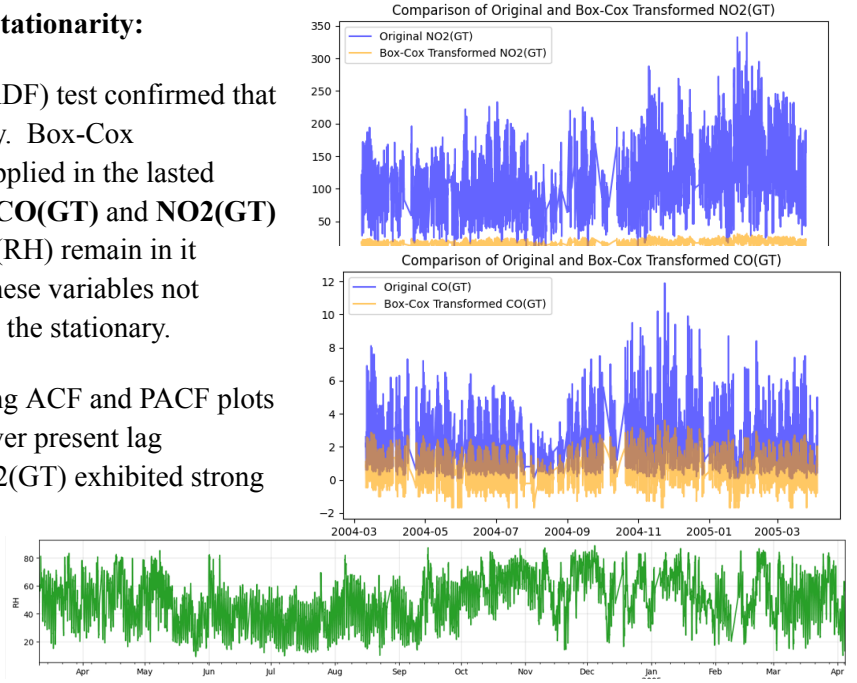


Fig. 3: **Before and after Box-Cox Transformation**

b) Check and address stationarity:

The Augmented Dicky-Fuller (ADF) test confirmed that all three data series are stationary. Box-Cox transformation previously was applied in the lasted process to reduce skewness for **CO(GT)** and **NO2(GT)** and confirmed stationary, while (RH) remain in it symmetic and stationary form, these variables not requiring differencing to achieve the stationary.

Post transformation analysis using ACF and PACF plots also confirmed stationary, however present lag dependencies. CO(GT) and NO2(GT) exhibited strong autocorrelations, likely due to underlying patterns such as seasonal or repetitive behaviors, supporting their suitability for time-series modeling. On the contrary, RH showed weaker autocorrelation, confirming its stability in the original form. Thus, all variables were confirmed stationary and ready for time series analysis. The summarize of the finding listed in table below (Fig 4):



Variables	ADF Test (Original)	ADF Test (Box-Cox)	Stationary Status
CO(GT)	p-value < 0.05	p-value < 0.05	Stationary
NO2(GT)	p-value < 0.05	p-value < 0.05	Stationary
RH	p-value < 0.05	Not transformed	Stationary

Fig. 4: VAR - Comparison of Box-Cox Transformation

C) Investigating Univariate vs. VAR Models

The results from conducting the Granger causality test to analyze predictive relationships among CO(GT_Boxcox), NO2(GT_boxcox) and RH, show significant interdependencies at lags ≥ 3. These findings indicate that the historical values of one variable provide predictive information about the others, confirming their mutual dependence.

These interdependencies strongly highlight the suitability of these variables for environmental monitoring and prediction systems. The significant relationships between pollutants (**CO** and **NO2**) and the environmental variable (**RH**) underscore the impact of environmental factors on air quality dynamics. Based on these results, **Vector Autoregressive (VAR)** models are well-suited for capturing

multivariate dependencies and its ability to model such interrelationships make it highly recommended for effective forecasting in air quality dataset.

III) Analysis, modeling and prediction (Task 3)

a) Model Results

VAR models were fitted to predict the relationship of these three variables: **CO(GT_boxcox)**, **NO2(GT_boxcox)**, and **RH**. Based on the lag selection test, 15 was selected as the max lag for the VAR model. During the fitting process, the VAR model effectively captured the relationships between these variables. Table for these variables were created to analyzes how much the residual error of one variable is related to another, highlights insights into the dependencies and accuracy of the model:

- **Interdependence:** CO(GT_boxcox) and NO2(GT_boxcox), show a strong mutual influence, aligning with its nature that share pollution sources (e.g., vehicles or industrial emissions). While RH has weak correlations with the other two, meaning it might behave more independently but less directly influence on pollutant levels but shows seasonal patterns emphasizing its roles in depression modeling.

Variables	CO(GT_boxcox)	NO2(GT_boxcox)	RH
CO(GT)	1.000	0.691	0.118
NO2(GT)	0.691	1.000	0.063
RH	0.118	0.063	1.000

- **The low AIC, BIC and AIC score** verify that the VAR model is good for this dataset, making its robust choice for forecasting to related variables.

Thus, the correlation matrix of residuals confirms the independence relationship between better pollutants (**CO** and **NO2**) and the independent nature of **RH**. This finding supports the suitability of selecting the VAR model for forecasting environmental and pollution interactions.

b) Residual Diagnostics

- **Durbin-Watson statistic:**
  The Durbin - Watson test confirms that residuals for **CO(GT_boxcox)** (2.0), **NO2(GT_boxcox)** (2.01), and **RH** (2.01) exhibit no autocorrelation. This support supports the validity of the VAR model and indicates that it sufficiently captures the dynamic of the data.
- **Ljung-Box Test:**
  The Ljung Box test for residuals indicates that the model well captured the dynamics of **CO(GT_boxcox)** and **NO2(GT_boxcox)**, with p-values (0.736 and 0.693) above the 0.05 threshold, confirming autocorrelation. While RH, a borderline p-value (0.052) suggests mild residual autocorrelation, indicating the additional technique might be needed for further exploration.

The verification from Durbin-Watson statistics and the Ljung-Box test confirmed that the residuals are well-behaved and minimal autocorrelation for most variables, this serve as another validation, supporting the suitability of the VAR model as an appropriate choice for modeling and forecasting the air quality dataset effectively.

c) Forecast Evaluation with Test Data :

Below is the performance comparison of actual and forecasted values and graphs for monoxide (CO), nitrogen dioxide (NO2), and relative humidity (RH) for the last 24 hours (24 points from the test set) from **2005-04-01** 15:00:00 **to 2005-04-06** 14:00:00 The graph shows actual value (solid lines) versus forecasted line (dashed lines). These variables were observed to understand the air quality trends and predicting pollution levels over time using the VAR model.

1. **CO(GT) (Carbon Monoxide):**

   Performance: The trend (Green line) shows that predicted value aligns with actual trends, but the model struggles with sudden spikes. This highlights that the model is good at predicting normal levels but struggles with sudden changes (like unexpected sudden surge).

   Assumption: The original data is impacted by 17% of missing values and high skewness, an outlier in CO(GT), while this study proves that the Box-Cox transformation improved distribution, but skewness and high number of missing values still affect predictions accuracy.

datetime	CO(GT)	NO2(GT)	RH	CO(GT)_forecast	NO2(GT)_forecast	RH_forecast
2005-04-03 15:00:00	1.10	100.4	11.075000	1.159746	95.731716	13.536232
2005-04-03 16:00:00	1.30	132.0	10.375000	1.383062	107.186749	14.905964
2005-04-03 17:00:00	1.40	156.3	9.875000	1.645477	118.173037	17.138432
2005-04-03 18:00:00	1.20	137.9	21.699999	1.793479	125.871936	20.427926
2005-04-03 19:00:00	2.70	181.3	33.050000	1.825683	128.355443	23.879668
2005-04-03 20:00:00	2.50	186.8	40.724999	1.672275	122.740017	27.770087
2005-04-03 21:00:00	1.50	157.8	46.550000	1.405112	113.244339	31.541857
2005-04-03 22:00:00	1.60	153.2	48.975000	1.161356	101.438906	35.828535
2005-04-03 23:00:00	1.20	127.6	52.500000	0.953460	90.330567	39.393017
2005-04-04 00:00:00	0.90	93.0	51.450000	0.847060	82.939982	41.837736
2005-04-04 01:00:00	0.60	58.3	51.150001	0.776991	77.880071	43.491378
2005-04-04 02:00:00	0.50	54.6	56.300000	0.731903	74.838864	44.443062
2005-04-04 03:00:00	0.40	51.3	58.899999	0.749000	74.784805	44.917057
2005-04-04 04:00:00	0.45	42.5	55.975000	0.799757	76.246664	44.871038
2005-04-04 05:00:00	0.50	53.1	59.875000	0.870260	79.096707	44.191401
2005-04-04 06:00:00	1.10	93.0	63.150000	0.948792	83.100630	42.740502
2005-04-04 07:00:00	4.00	154.6	61.924999	1.041718	88.271433	40.642243
2005-04-04 08:00:00	5.00	173.6	48.875000	1.146574	93.976327	38.050957
2005-04-04 09:00:00	3.90	186.5	36.275001	1.252979	99.549852	35.279847
2005-04-04 10:00:00	3.10	189.8	29.250000	1.356274	104.839985	32.608010
2005-04-04 11:00:00	2.40	179.2	23.725000	1.444864	109.229330	30.320642
2005-04-04 12:00:00	2.40	174.7	18.350000	1.522792	112.749483	28.604778
2005-04-04 13:00:00	2.10	155.7	13.550000	1.581999	115.024208	27.633772
2005-04-04 14:00:00	2.20	167.7	13.125000	1.623445	116.242662	27.481716

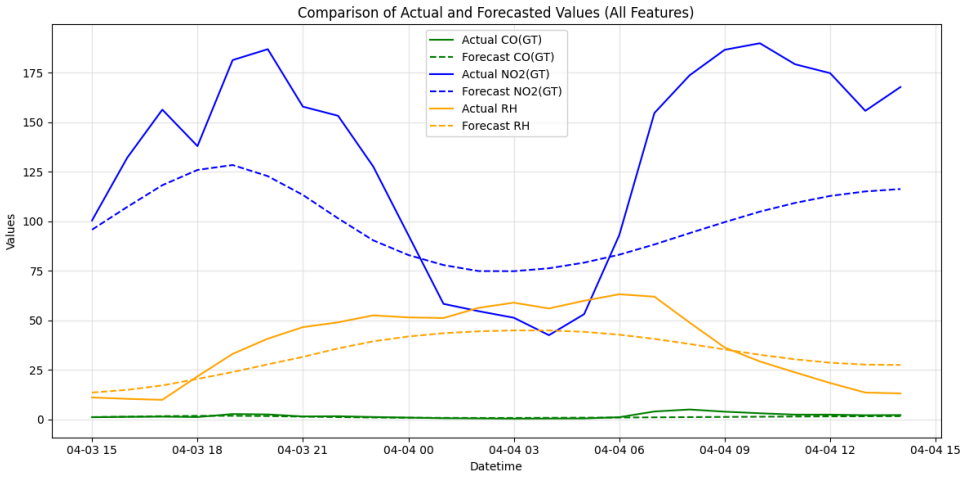


Fig. 5: VAR - Comparison of actual and forecasted values

2. **NO2(GT) (Nitrogen Dioxide):**

   Performance: The forecast (Blue line) follows general trends but struggles to predict the peaks accurately, highlighting the poor ability in handling extreme values, similarly with CO2.

   Assumption: The trend at the peak during winter time, indicates that NO2(GT) is influenced by external conditions (e.g., weather data).

3. **RH (Relative humidity):** The prediction line (Orange line) aligns closely with actual values, showing the model effectiveness for stable variables likely because it has less effect from external factors.

   Assumption: RH is a stable variable. The original data shows a normal distribution, with no significant skewness or outlier, and only 3% missing values. Due to its well-structured nature,

no transformation techniques were required for RH, which preserve the variable's predictability and contributing to accurate modeling and forecasting

d) Evaluation metric summary:

The performance of a VAR model for predicting the air quality metric shows similar results from plotting the prediction, and was accessed using standard predictive analysis metrics. These metrics go beyond the performance as it indicates the model strength weakness and potential improvement areas:

Metrics	CO(GT_boxcox)	NO2(GT_boxcox)	RH
MAE	0.7011	4.5944	12.4402
RMSE	0.8594	4.9409	15.0444
MAPE	91.7998%	25.5188%	28.6848%
ME	0.3440	2.5318	11.6514
MPE	87.9795%	7.6276%	21.3671%
Correlation	0.1080	-0.1500	0.9475
Min max Error	0.2459	0.3588	0.2335

- **CO(GT)**: Predictions align for stable levels but struggle with sudden spikes (MAPE = 91.8%, Correlation = 0.108).
- **NO2(GT)**: Moderate prediction accuracy (MAPE = 25.5%, Correlation = -0.150)
- **RH**: Strong performance with high correlation (0.95) and low error (MAPE = 28.7%).

e) Recommendations and Conclusion

- **CO(GT) :** The findings from this study suggest further improvement, applying a log transformation or different imputation techniques could reduce skewness. Additionally, as CO(GT) highly depends on external factors (e.g., traffic and weather), incorporating more factors into the model would improve prediction reliability.
- **NO2(GT):** Including additional data and changing the model for seasonal patterns like winter heating emission could improve peak prediction accuracy.
- **RH:** The correlation values and the line plots verify that RH behaves independently, suggesting that the ARIMA model could be more effective.

f) Conclusion:

Both forecasted plotted and evaluation metrics provide similar observations and real world implication. CO(GT) and NO2(GT) are harder to predict due to their dependence on unpredictable external factors like traffic or weather, which result in spikes and variations in the predictions. While RH on the other hand is easy to predict, it shows the highest accuracy prediction because it follows stable weather patterns, making it less sensitive for external disturbance. The application using the VAR model is shown to be effective for independent variables, while the model captures general trends, its accuracy can be enhanced by incorporating additional data.

g) Additional implementation, Forecasting the next 2 days (48 hours - beyond the test set)

The graph represents the forecasted and actual values for **CO(GT)**, **NO2(GT)**, and **RH** over a 48-hour extended forecast period using the VAR model panning from 2005-04-01 15:00:00 to 2005-04-06 14:00:00 (a total duration of 4 days). The forecasts align with previous trends:

- **CO(GT)**: Captures general trends but struggles with spikes, likely due to skewness, outliers, and missing data.

- **NO2(GT)**: Predicts general patterns but underestimates peaks, possibly influenced by seasonal factors.
- **RH**: Shows accurate predictions likely due to its stability and minimal external influences.

Overall, the VAR model performs well for future forecasting for **RH** but struggles with pollutants (**CO** and **NO2**) due to external dependencies.

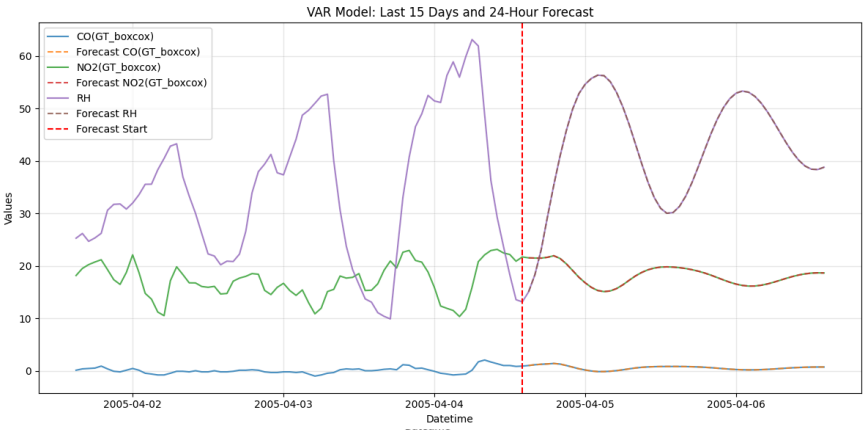


Fig. 6: VAR - Comparison of actual and forecasted values

IV) Bous task Analysis, modeling and prediction (Task 3)

Individual ARIMA models were fitted for each independent variable based on their identified parameters, the forecasted values for CO(GT), NO2(GT), and RH were plotted alongside actual values to evaluate its performance:

a) ARIMA Model Fitting

Variable	Model	AIC	BIC	Log Likelihood	Ljung-Box (p-value)	Jarque-Bera (p-value)	Kurtosis
CO(GT)	ARIMA(5, 1, 5)	6870.519	6949.072	-3424.259	0.36	0.00	5.36
NO2(GT)	ARIMA(2, 1, 2)	33688.384	33716.090	-16835.192	0.98	0.00	5.27
RH	ARIMA(3, 1, 2)	51643.452	51686.300	-25815.726	0.13	0.00	11.38

- **CO(GT):** The finding shows a good fit with a low AIC and BIC with no autocorrelation in residuals (Ljung-Box test passed). However, residuals skewness and heavy tails.
- **NO2 :** The model effectively captures short-term patterns in NO2(GT_boxcox), with no autocorrelation in residuals (Ljung-Box test passed). While the findings show the heavy tailed residuals.

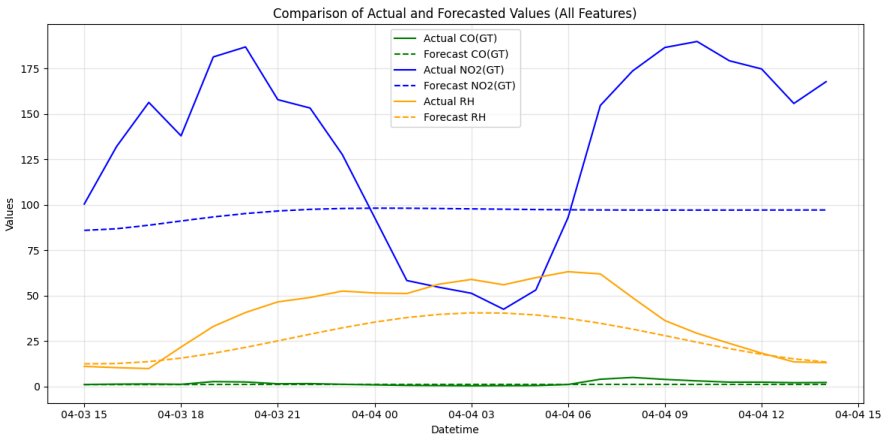


Fig. 7: ARIMA - Comparison of actual and forecasted values

- **RH:** The results show strong fit with significant AR and MA terms, residuals show no autocorrelation. The limitation is a high residual and deviations from normality (Jarque-Bera test failed).

b) Performance analysis

- **CO(GT):**
  Performance: Prediction (Green Line) aligns with actual trend during stable period but poorly performing to capture the spike, shown that ARIMA struggles with underpredicable surge in CO.
  Assumption: CO highly dependent by external factors like traffic and industrial activities which assume causing these spikes.
- **NO2(GT):**
  Performance: Poor performance which does not follow the trends and consistently underestimate high peaks.
  Assumption: Assume that the high underestimation of peak due to the impact of missing external variables, seasonal influence like winter emission.

- **RH:**
  Performance: This study verifies that RH might suit ARIMA as its independent variables, which the trend shows the line well closely with actual values, emphasizing that ARIMA suits for stable variables with low variability and independence.
  Assumption: RH has less effects from external disturbances, making it inherently easier to predict compared to pollutants like CO and NO2.

c) Evaluation metric summary:

Metrics	CO(GT_boxcox)	NO2(GT_boxcox)	RH
MAE	0.9996 (Low error)	55.9876 (High error)	12.4402 (Moderate error)
RMSE	1.3903	61.4026	15.0444
MAPE	59.73%	47.12%	28.68%
ME	0.697	36.1396	11.6514
MPE	-1.37%	8.71%	21.37%
Correlation	0.2016 (Weak positive)	-0.1184 (Weak negative)	0.9475 (Strong positive)
Min max Error	0.8350	0.6296	0.5095

- **CO(GT):** the metrics shows **moderate accuracy** with low MAE and RMSE values indicate low error for predicting general trends but struggles with variability **(MAPE = 59.73%)** and **weak correlation** (0.2016): Indicates poor predictive performance when capturing sudden changes with **MinMax Error** (0.8350).
- **NO2(GT)**: High errors (**MAE = 55.98, RMSE = 61.40**) and negative correlation (-0.1184) reflect poor accuracy. **MAPE** (47.12%) shows frequent mismatches between predicted and actual values.
- **RH**: While the metric shows best accuracy with **low MAPE (28.68%)**, high correlation (0.9475), and minimal **MinMax Error** (0.5095), indicating strong performance.

d) Conclusion and recommendation:

The finding proves that ARIMA models, suited for linear and dependent variables, are not appropriate for predicting **CO2 and NO2** due to their interdependent and nonlinear relationships. ARIMA assumes that present data are a linear function of past data points and errors, making it unsuitable for variables influenced by external factors like traffic and weather[2]. While ARIMA performs well for stable variables like **RH**, it fails to generalize effectively to volatile pollutants. To improve the accuracy for real world air quality monitoring, adding additional data to the model and leveraging a machine learning approach are recommended.

e) Comparison of VAR and ARIMA for Air quality Dataset:

- **Performance :**

  VAR : VAR outperforms ARIMA across all metrics (MAE, RMSE, MAPE) with significantly lower errors (MAE, RMSE, MAPE) for the air quality dataset as it captures strong interdependencies between variables like CO and NO2.

  ARIMA: slightly outperforms VAR in MAPE (28.68% vs. 33.89%**)** but is not recommended for the overall air quality dataset due to poor handling of interdependent pollutants.

Metrics	CO	NO2	RH
VAR (MAE)	0.5232 (Better)	3.3858 (Better)	10.4572 (Better)
ARIMA (MAE)	0.7011	4.5944	12.4402
VAR (RMSE)	0.7148 (Better)	3.7800 (Better)	11.7460 (Better)
ARIMA (RMSE)	0.8594	4.9409	15.0444
VAR (MAPE)	72.93% (Better)	18.20% (Better)	33.89%
ARIMA (MAPE)	91.80%	25.52%	28.68% (Better)

- **Best use case:**

  VAR: The model is well suited for CO(GT) and NO2(GT) due to their interdependence from shared pollution sources.

  ARIMA: Suitable for RH, which is independent and stable, requiring no modeling of interdependencies.

4) Conclusion

The analysis of the air quality dataset for CO, NO2 and RH demonstrates the importance of selecting the appropriate preprocessing techniques and the models can significantly impact forecasting accuracy. VAR proved its strong potential in predicting interdependencies between variables like CO and NO2, making it the most suitable choice for this study in the air quality dataset. ARIMA, on the other hand, performed better for independent variables like RH. Overall VAR outperformed ARIMAR in this study. For further improvement, incorporating additional data that directly influence these variables and refinding different preprocessing techniques could improve model accuracy and forecasting performance.

5) References

[1] ABBOTT, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst Dean Abbott*. John Wiley and Sons.

[2] Babu, C. N., & Reddy, B. E. (2014). A moving-average filter-based hybrid Arima–Ann Model for forecasting time series data. *Applied Soft Computing*, *23*, 27–38. https://doi.org/10.1016/j.asoc.2014.05.028

[3] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice*. OTexts.

[4] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-56706-x

[5] Vito, S. (2008). Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F.

[6] V. Hassani, "Chapter 1 & 2 - Overview of Predictive Analytics and Setting up the Problem," unpublished.