PGR207 - Predictive Analytics Examination Report

Candidate nr. 12
School of Economics, Innovation, and Technology
Kristiania University College, Oslo, Norway

Abstract

This report investigates the application of Vector Autoregression (VAR) and Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) models for forecasting air quality data. Using a dataset comprising hourly pollutant and environmental readings, we focus on three variables: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Relative Humidity (RH). The objective is twofold: to evaluate whether these variables are better modeled using a unified multivariate approach or as separate univariate time series, and to demonstrate a comprehensive understanding of time series modeling. The findings show that variable preferences vary greatly, with RH favoring a unified multivariate approach, while NO2 demonstrates a preference for separate univariate modeling. A detailed comparison of model performance is provided using statistical metrics, residual analysis, and diagnostic tests. This study highlights the importance of selecting appropriate modeling strategies based on variable characteristics and interrelations.

CONTENTS

I Introduction 3

II Methodology 3
- II-A Air Quality Dataset . . . 3
  - II-A1 CO . . . 3
  - II-A2 NO2 . . . 3
  - II-A3 RH . . . 3
- II-B Statsmodels . . . 3
- II-C VAR Model . . . 3
  - II-C1 Mathematical Formulation . 3
  - II-C2 Example . . . 4
  - II-C3 Key Assumptions . . . 4
- II-D ARIMA Model . . . 4
  - II-D1 Mathematical Formulation . 4
  - II-D2 Key Assumptions . . . 4
- II-E SARIMAX Model . . . 4
  - II-E1 Mathematical Formulation . 4
  - II-E2 Key Assumptions . . . 4
- II-F Evaluation Metrics . . . 5
  - II-F1 ADF Test . . . 5
  - II-F2 AIC . . . 5
  - II-F3 BIC . . . 5
  - II-F4 Residuals . . . 5
  - II-F5 Rolling Statistics . . . 5
  - II-F6 RMSE . . . 5
  - II-F7 MAE . . . 5
  - II-F8 Ljung-Box Test . . . 5
  - II-F9 Shapiro-Wilk Test . . . 5

III Experiments and Results 5
- III-A Data Processing . . . 5
- III-B Data Exploration . . . 6
  - III-B1 Granger-Causality Analysis . 6
  - III-B2 Correlation Matrix . . . 6
  - III-B3 Augmented Dickey-Fuller (ADF) Test . . . 6
  - III-B4 Box-Plot . . . 6
- III-C Code Structure . . . 6
- III-D VAR Code Structure . . . 7
  - III-D1 Data Splitting and Scaling . 7
  - III-D2 Model Implementation . . . 7
  - III-D3 Forecast Implementation . . 7
- III-E SARIMAX Code Structure . . . 7
  - III-E1 Data Splitting and Scaling . 7
  - III-E2 Model Implementation . . . 7
  - III-E3 Forecast Implementation . . 8
- III-F Model Results Comparison . . . 9
  - III-F1 Evaluation . . . 9

IV Discussion 10

V Conclusion 11

VI Areas of Improvement 11
- VI-1 Time Management . . . 11
- VI-2 Lag Order Selection . . . 11
- VI-3 SARIMAX RH . . . 11

VII Acknowledgment 11

VIII References 12

I. Introduction

Time series analysis plays a critical role in understanding and forecasting temporal dependencies in multivariate datasets. This report explores the application of Vector Autoregression (VAR) and Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) models for forecasting air quality data. The experiments are conducted using a dataset containing hourly readings of pollutants and environmental factors from a polluted Italian city.

VAR models are well-suited for analyzing the relationships among multiple variables simultaneously. In contrast, the SARIMAX model is an ARIMA variant that allows for seasonal adjustments and the inclusion of external variables, offering a complementary approach to forecasting univariate time series.

Of the 15 variables within the dataset, this study focuses on three: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Relative Humidity (RH). The objectives of this report are twofold: First, to evaluate whether the selected variables—CO, NO2, and RH—are better modeled using a unified multivariate approach or as separate univariate time series. Second, to serve as a demonstration of my understanding and application of time series modeling concepts, fulfilling the requirements of this examination task.

II. Methodology

In Methodology each of the different technologies and tools relevant to the report is introduced, including an overview of the dataset experimented on and the different methods used in said experiments.

A. Air Quality Dataset

The dataset consists of 9358 hourly readings from five metal oxide chemical sensors monitoring air quality in a polluted Italian city at road level, collected from March 2004 to February 2005. It includes ground truth data represented as (GT) for CO, non-methane hydrocarbons, benzene, NOx, and NO2 from a certified analyzer. The data includes cross-sensitivities, sensor drifts, and missing values marked as -200. It is intended for research purposes only (De Vito et al., 2008). In this report, we will only concern ourselves with the CO, NO2, and RH readings from this dataset.

1) CO: Carbon Monoxide (CO) is a colorless, odorless gas primarily emitted by motor vehicles and industrial processes. It serves as an indicator of urban air pollution and is a standard variable in assessing air quality (Organization, 1999).

In the dataset, CO is reported as ground truth (GT) values obtained from certified analyzers, making it a reliable measure despite known sensor drifts and cross-sensitivities. These drifts are corrected through interpolation, ensuring data quality for analysis. Peaks in CO levels are expected during high-traffic periods, making it a suitable variable for time-series forecasting.

2) NO2: Nitrogen Dioxide (NO2) is a toxic gas formed during the combustion of fossil fuels, often serving as a precursor to smog and acid rain. It is an important marker for urban air pollution and is strongly correlated with traffic density (Organization, 2021).

The dataset provides ground truth NO2 values, which show daily and seasonal variations tied to traffic and weather patterns. Similar to CO, NO2 readings are subject to sensor drift, but preprocessing ensures the data is consistent for modeling.

3) RH: Relative Humidity (RH) measures the percentage of water vapor in the air relative to the maximum amount it can hold at a given temperature. While RH is not a direct pollutant at water level, it plays a significant role in influencing sensor readings for gases like CO and NO2, as shown by Spinelle et al. (2020).

In this dataset, RH provides contextual information to understand and adjust for environmental effects on sensor readings. Its variability over time adds complexity to the modeling process but enhances the robustness of predictions.

B. Statsmodels

Statsmodels is a Python library for statistical modeling and analysis, focusing on tools for estimating, testing, and interpreting machine and statistical models. It supports tasks like regression, time series analysis, generalized linear models, and hypothesis testing. Its strength lies in providing detailed statistical summaries and diagnostics, making it ideal for understanding model behavior Statsmodels Developers (2024a).

C. VAR Model

The Vector Autoregression (VAR) model was introduced as an alternative to large-scale structural economics models by Sims (1980). VAR models are particularly powerful as a framework for analyzing multi-variable time series as they can capture the evolution and the inter-dependencies between multiple time-series.

1) Mathematical Formulation: A VAR model of order p, denoted as VAR(p), models a vector of k time series variables, $\mathbf{y}_t$, as a linear function of their own lagged values and the lagged values of other variables in the system:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t$$

Here:

- $\mathbf{y}_t$ is a $k \times 1$ vector of k endogenous variables at time t:

$$\mathbf{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{kt} \end{bmatrix}$$

- $\mathbf{c}$ is a $k \times 1$ vector of intercept terms.
- $\mathbf{A}_i$ is a $k \times k$ matrix of coefficients for lag i, capturing how each variable affects the others.
- $\mathbf{u}_t$ is a $k \times 1$ vector of error terms (innovations) with:

$$\mathbf{u}_t \sim \mathcal{N}(0, \Sigma_u)$$

where Σ_u is the covariance matrix of the residuals.

2) Example: Consider a VAR(1) model with two variables ($k = 2$):

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

This expands to the following system of equations Hassani (2023):

$$y_{1t} = c_1 + a_{11} y_{1,t-1} + a_{12} y_{2,t-1} + u_{1t}$$

$$y_{2t} = c_2 + a_{21} y_{1,t-1} + a_{22} y_{2,t-1} + u_{2t}$$

3) Key Assumptions: The VAR model makes the following assumptions, as explained by Sims (1980):

1) The time series $\mathbf{y}_t$ is stationary or made stationary (e.g., through differencing).
2) The errors $\mathbf{u}_t$ are white noise, with zero mean, constant variance, and no autocorrelation.
3) The lag order p is finite and appropriately chosen.

D. ARIMA Model

The Autoregressive Integrated Moving Average (ARIMA) model was introduced as part of the Box-Jenkins methodology for time series forecasting by Box and Jenkins (1970). The ARIMA model was built to analyze and predict univariate time series by capturing both temporal dependencies and trends through a combination of autoregressive, differencing, and moving average components.

1) Mathematical Formulation: An ARIMA(p, d, q) model integrates three components:

- Autoregression (AR): Incorporates lagged values of the variable.
- Differencing (I): Ensures stationarity by differencing the series d times.
- Moving Average (MA): Models the error terms as a linear combination of past errors.

The ARIMA(p, d, q) equation looks like this:

$$\Phi(B)(1 - B)^d y_t = \Theta(B)\epsilon_t$$

Here:

- y_t: The time series data at time t.
- $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$: Autoregressive polynomial of order p.
- $\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$: Moving average polynomial of order q.
- B: Backshift operator, where $By_t = y_{t-1}$.
- d: The differencing order to ensure stationarity.
- ϵ_t: White noise (innovations) with:

$$\epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

2) Key Assumptions: As outlined by Box and Jenkins (1970), ARIMA models rely on the following assumptions:

1) Stationarity: The time series is stationary or made stationary through differencing (d).
2) Linearity: The relationship between variables is linear.
3) White Noise Residuals: The residuals (ϵ_t) are uncorrelated and follow a normal distribution.
4) Finite Lags: The orders p, d, q are finite and determined based on the data.

E. SARIMAX Model

The Seasonal Autoregressive Integrated Moving Average + Exogenous Variable (SARIMAX) model is a combination of two ARIMA model types; the SARIMA model introduced by Box and Jenkins (1970), which takes into account seasonal patterns, and the ARIMAX model, which integrates external variables to enhance accuracy (Majka, 2024). SARIMAX extends the ARIMA framework by accounting for both seasonal patterns and external variables. This advancement allows the SARIMAX model to encapsulate complex behaviors within time-series data Shah et al. (2024).

1) Mathematical Formulation: A SARIMAX(p, d, q, P, D, Q, s) model incorporates both seasonal and non-seasonal components, along with exogenous variables. The general equation is:

$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t = \Theta_q(B)\Theta_Q(B^s)x_t + \mathbf{X}_t\boldsymbol{\beta} + \epsilon_t$$

Here:

- y_t: The time series at time t.
- B: The backshift operator, $By_t = y_{t-1}$.
- p, d, q: Non-seasonal autoregressive (p), differencing (d), and moving average (q) orders.
- P, D, Q, s: Seasonal autoregressive (P), seasonal differencing (D), and seasonal moving average (Q) orders with seasonal period s.
- $\Phi_p(B)$: Non-seasonal AR polynomial of order p, $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$.
- $\Phi_P(B^s)$: Seasonal AR polynomial of order P, $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$.
- $\Theta_q(B)$: Non-seasonal MA polynomial of order q, $\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$.
- $\Theta_Q(B^s)$: Seasonal MA polynomial of order Q, $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs}$.
- $\mathbf{X}_t\boldsymbol{\beta}$: Exogenous variables ($\mathbf{X}_t$) and their coefficients ($\boldsymbol{\beta}$).
- ϵ_t: White noise (innovations), $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

2) Key Assumptions: As outlined by Box and Jenkins (1970) and further extended in (Statsmodels Developers, 2024b), the SARIMAX models inherit the assumptions of the standard ARIMA model, including stationarity, linear relationships, and white noise residuals. Additionally, SARIMAX relies on the following specific assumptions:

1) Seasonality: Seasonal patterns are regular and periodic with a defined seasonal period (s).

2) Exogenous Predictors: External variables ($\mathbf{X}_t$) are independent of the residual errors (ϵ_t).

F. Evaluation Metrics

1) ADF Test: Augmented Dickey-Fuller (ADF) Test is a statistical test to check if properties like mean and variance are stationary over time. It was first introduced by Dickey and Fuller (1979). Its Null Hypothesis (H_0) states that the time series has a unit root, meaning it is non-stationary. The Alternative Hypothesis (H_a) suggests that the time series does not have a unit root and is stationary.

The test works by fitting a regression model to the time series data and calculating a test statistic. This statistic is then compared against critical values to decide whether to reject H_0. If the test statistic is smaller than the critical value, H_0 is rejected, indicating the time series is stationary.

2) AIC: Akaike Information Criterion (AIC) is a statistical measure introduced by Akaike (1974) used to evaluate the goodness-of-fit of a model while accounting for its complexity. AIC balances the trade-off between model fit and complexity by penalizing models with more parameters. It is calculated as: AIC $= 2k - 2\ln(L)$, where k is the number of model parameters and L is the maximum likelihood of the model. Lower AIC values indicate better models.

3) BIC: Bayesian Information Criterion (BIC) is a similar measure to AIC introduced by Schwarz (1978) to evaluate model fit while penalizing complexity. However, it applies a stricter penalty than AIC, especially as the sample size increases. It is calculated as: BIC $= k\ln(n) - 2\ln(L)$, where n is the sample size, k is the number of parameters, and L is the likelihood. Like AIC, lower BIC values indicate better models.

4) Residuals: Residuals are the differences between observed values and those predicted by a statistical model In regression analysis, residuals are calculated as: Residual $=$ Observed Value $-$ Predicted Value. Examining residuals can reveal various model deficiencies such as Non-linearity, Outliers, or Heteroscedasticity. Alternately it can produce white noise, this is the desired outcome as that means the data does not contain any of the aforementioned deficiencies. Residual analysis is crucial in assessing model adequacy in both ARIMA models, as emphasized by Box and Jenkins (1970), and VAR models, as highlighted by Sims (1980).

5) Rolling Statistics: Rolling statistics involve calculating moving averages or variances over a specified window of time, providing a dynamic view of changes in the data. These metrics are useful in time series analysis for assessing stationarity and identifying trends. In ARIMA modeling, as described by Box and Jenkins (1970), rolling statistics complement formal stationarity tests like ADF by offering visual confirmation of constant mean and variance over time. Similarly, in Vector Autoregression (VAR) models, Sims (1980) emphasized the importance of understanding trends in multi-variable time series, where rolling statistics can play a key diagnostic role.

6) RMSE: Root Mean Squared Error (RMSE) is a standard metric for evaluating model accuracy, representing the square root of the average squared differences between observed and predicted values. It provides a measure in the same units as the original data and emphasizes larger errors due to the squaring process. RMSE is widely used in time series forecasting, including ARIMA models as highlighted by Box and Jenkins (1970), and VAR models, as emphasized by Sims (1980).

7) MAE: Mean Absolute Error (MAE) measures the average absolute difference between observed and predicted values. Unlike RMSE, MAE treats all errors equally, making it less sensitive to outliers. Similar to RMSE, it has been extensively applied to evaluate time series models like ARIMA (Box and Jenkins, 1970) and VAR (Sims, 1980).

8) Ljung-Box Test: The Ljung-Box test is a statistical test used to detect autocorrelation in time series analysis. As explained by Ljung and Box (1978), it evaluates the null hypothesis (H_0) that the residuals in are independently distributed, or in other words, do the residuals look like white noise. A high p-value ($p > 0.05$) indicates no significant autocorrelation. Conversely, a low p-value ($p \leq 0.05$) suggests autocorrelation. Box and Jenkins (1970) outlines that avoiding residual autocorrelation is crucial in identifying the appropriate ARIMA model. Similarly, Sims (1980) states that VAR model residuals should not exhibit autocorrelation.

9) Shapiro-Wilk Test: The Shapiro-Wilk test is a statistical test used to assess the normality of a dataset. As outlined by Shapiro and Wilk (1965) it evaluates the null hypothesis (H_0) that the data follows a normal distribution. A high p-value ($p > 0.05$) suggests that the data is consistent with a normal distribution, while a low p-value ($p \leq 0.05$) indicates a deviation from normality.

III. Experiments and Results

Two models were chosen for this assignment: A VAR model as required by the given task, and SARIMAX. The SARIMAX model was selected over the standard ARIMA model due to its innate ability to handle seasonal patterns, in addition to incorporating exogenous variable, aiming to build the best univariate time series model possible. During the implementation of the VAR model it was discovered that the data exhibited both seasonal data signs of inter-column correlation, making SARIMAX suitable for experimentation. Both models were build using statsmodels. A SARIMAX model should be well suited for the task of testing if the data suit univariate time series better or not due to incorporating exogenous variables.

This section will first outline the data processing and data exploration process and the key findings. This will be followed by an explanation of the code structure and model implementations. Lastly, I will present the results.

A. Data Processing

The dataset documentation (De Vito et al., 2008) specifies that missing values are marked as '-200', These were specified as na_values during the data import.

To clean the dataset:

- The two trailing empty columns were removed due to being empty.
- The `Date` and `Time` columns were combined into a single `DateTime` column, which was then converted into a proper datetime format and set as the index for the dataframe.
- Then the columns—`CO(GT)`, `NO2(GT)`, and `RH`—were extracted into its own dataframe as specified by the given task.

Next, all rows containing missing values were dropped. This decision was made because they made up a significant percentage of the total data, filling such a large number of `NaN` values would risk introducing significant bias into the model. Finally, each row was converted to float data type. These steps were taken for both models.

B. Data Exploration

This section showcases the data exploration performed to understand the relationships and properties of each time series(variable) before modeling. The data exploration steps were the same for both models.

The process began by examining the data structure using `df.info` and printing the three time series to understand their characteristics. Subsequently, the three series were plotted together to explore their relationships. The dataset consists of 6,941 data points spanning 13 months. A significant difference in unit values among the series is observed, with `NO2(GT)` ranging from 50 to 300 , `RH` from 25 to 80 , and `CO(GT)` remaining below 10 . This is illustrated in Figure 1.

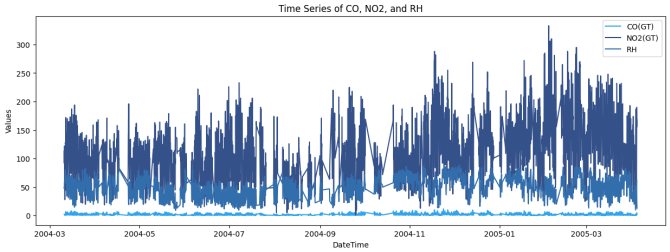

Fig. 1. Graph showing the full dataset after processing.

1) Granger-Causality Analysis: Granger-causality tests were conducted to evaluate whether one time series variable could be used to predict another. The results varied but show significant relationships at earlier lags:

- `CO(GT)` Granger-causes `RH` with significant p-values fro lag 1 and 2.
- `CO(GT)` Granger-causes `RH` at lags 3, 4, and 5, with decreasing p-values as the lag increases.
- `NO2(GT)` Granger-causes `RH` to some degree up to lag 4.
- `RH` Granger-causes both `CO(GT)` and `NO2(GT)` at lag 1 to some degree.

These findings suggest some temporal dependencies among the variables.

2) Correlation Matrix: The correlation matrix revealed the following key relationships:

- `CO(GT)` and `NO2(GT)` have a strong positive correlation ($r = 0.67$).
- `RH` has a weak positive correlation with `CO(GT)` ($r = 0.06$) and a weak negative correlation with `NO2(GT)` ($r = -0.08$).

These correlations indicate shared patterns between CO and NO2 readings, while relative humidity shows only a weak or negligible correlation with both.

3) Augmented Dickey-Fuller (ADF) Test: The Augmented Dickey-Fuller (ADF) test was applied to assess the stationarity of each time series variable. Stationarity is a crucial assumption for many time series models, such as VAR and ARIMA, as it ensures that the statistical properties of the series, like mean and variance, remain constant over time. The test evaluates the null hypothesis (H_0) that a unit root is present, indicating non-stationarity. If the null hypothesis is rejected, the series is deemed stationary.

a) Results and Interpretation:

- `CO(GT)`:
  - ADF Statistic = -9.89
  - Critical Values: 1% = -3.43, 5% = -2.86, 10% = -2.57
  - p-value < 0.01
  - The ADF statistic is less than the critical value at all significance levels and the p-value is very low, leading to the rejection of H_0.
- `NO2(GT)`:
  - ADF Statistic = -7.22
  - Critical Values: 1% = -3.43, 5% = -2.86, 10% = -2.57
  - p-value < 0.01
  - Similarly, the ADF statistic for `NO2(GT)` is below all critical values, and the p-value indicates high confidence in rejecting H_0.
- `RH`:
  - ADF Statistic = -7.40
  - Critical Values: 1% = -3.43, 5% = -2.86, 10% = -2.57
  - p-value < 0.01
  - The results for `RH` follow the same pattern. The test statistic is well below the critical thresholds, and the null hypothesis is rejected.

b) Summary: H_0 were rejected for all three variables—`CO(GT)`, `NO2(GT)`, and `RH`, indicating that all series are stationary.

4) Box-Plot: Box plots for `CO(GT)`, `NO2(GT)`, and `RH` revealed the presence of a multitude of outliers for `CO(GT)` and `NO2(GT)`. These outliers were retained as they predominantly occur in the second half of the dataset and are likely indicative of an emerging pattern or yearly cycles rather than errors. Removing them could potentially exclude meaningful insights about the underlying trends.

C. Code Structure

This section will be split into one for each of the model.

D. VAR Code Structure

The implementation of the VAR model is organized into three stages: data splitting and scaling, model implementation, and forecasting.

1) Data Splitting and Scaling: The dataset was split into training and testing subsets, with 85% of the data used for training and the remaining 15% for testing. Each variable in the dataset was scaled independently using the `StandardScaler` from the `sklearn` library. This ensured that each variable had a mean of 0 and a standard deviation of 1, preventing variables with larger ranges from dominating the model.

The split was then visualized to help understand the change. This graph can be seen in Figure 2

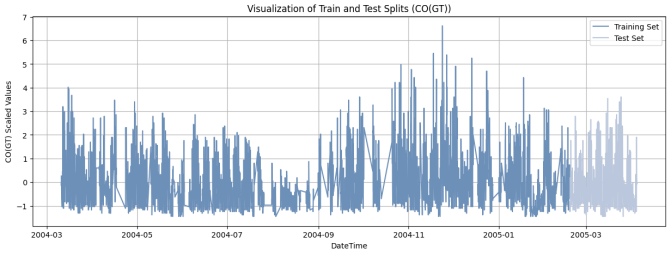

Fig. 2. Graph showing the Scaled Train/Test split on CO(GT)

2) Model Implementation: The VAR model was fitted on the scaled training dataset:

- **Lag Order Selection**: The optimal lag order 24 was determined based on AIC and BIC. The following logic was used to refine the choice programmatically:

```
optimal_lag = int(lag_order_results.bic + (
    abs(lag_order_results.aic -
    lag_order_results.bic) / 2))
```

  Since AIC suggested a higher lag than BIC, the formula adds half the absolute difference between AIC and BIC to the optimal BIC value. This results in a lag order that is as close as possible to the midpoint between the two criteria.
- **Model Fitting**: A single VAR model was then fitted to all three of the variables, the results of which can be viewed in Table I.
- **Model Stability Check**: The stability of the fitted VAR model was verified using `.is_stable()` to ensure that the predictions would not diverge over time.

TABLE I
SUMMARY OF VAR MODELING RESULTS

```
==================================
Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:              Sun, 08, Dec, 2024
Time:                      22:31:14
----------------------------------
No. of Equations:           3.00000
Nobs:                      5875.00
Log likelihood:           -5088.31
AIC:                       -6.70689
BIC:                       -6.45794
HQIC:                      -6.62034
FPE:                    0.00122247
Det(Omega_mle):         0.00117801
----------------------------------
Results for equation CO(GT)
==================================
```

3) Forecast Implementation: Using the trained VAR model:

- Forecasts were generated for the test set, using the last values from the training set as input. The forecast was performed for as many steps as there are in the test set.
- **Inverse Transformation**: To interpret the forecasted values in their original scale, the inverse of the scaling transformation was applied to both the forecasted and actual test data.
- The forecasted values for `CO(GT)`, `NO2(GT)`, and `RH` were compared with the actual test set values to evaluate the model's performance.

E. SARIMAX Code Structure

The implementation of the SARIMAX model is organized into three stages: data splitting and scaling, model implementation, and forecasting.

1) Data Splitting and Scaling: Similar to the VAR model, the dataset was split into training and testing subsets, with 85% of the data allocated for training and the remaining 15% for testing. Each variable in the dataset was scaled independently using the `StandardScaler` from the `sklearn` library. This scaling ensured that each variable had a mean of 0 and a standard deviation of 1, preventing variables with larger ranges from dominating the model.

2) Model Implementation: The SARIMAX model were implemented as separate models for each variable using the other two variables as exogenous variables. Selection of hyperparameters was done through an extensive grid search process. The steps involved are as follows:

- **Defining Target and Exogenous Variables**: For each target variable (`CO(GT)`, `NO2(GT)`, and `RH`), the remaining variables were used as exogenous predictors. This setup allows the model to leverage the interdependencies between variables for more accurate forecasting.
- **Grid Search for Hyperparameter Optimization**: A grid search was conducted to identify the optimal SARIMAX parameters for each target variable. The parameters explored included:

- Non-seasonal orders: $p \in \{1,2,3\}$, $d \in \{0,1\}$, $q \in \{1,2,3\}$
- Seasonal orders: $P \in \{1,2,3\}$, $D \in \{0,1\}$, $Q \in \{1,2,3\}$, $s = 24$ (seasonal period)

Separate searches were conducted to test values 4 and 5 as well. however, 3 was determined to be the highest desired value.

The grid search evaluated each combination based on the AIC values, selecting the combination with the lowest AIC as the optimal set of parameters.

- **Storing Best Parameters**: The best SARIMAX parameters for each target variable were stored for model fitting and forecasting. The summary of the best parameters is presented in Table II.
- **Model Fitting**: Using the parameters found during the grid search, a SARIMAX model was fitted to each target variable with the corresponding exogenous variables. The fitting process included enforcing stationarity and invertibility to ensure model stability. The results of these can be found in Tables III, IV, and V.

TABLE II
OPTIMAL SARIMAX PARAMETERS FOR EACH TARGET VARIABLE

Variable	p	d	q	P	D	Q
CO(GT)	2	0	3	3	0	1
NO2(GT)	2	0	3	3	0	1
RH	3	1	3	1	0	1

TABLE III
SUMMARY OF SARIMAX MODELING RESULTS FOR CO(GT)

```
==================================
Summary of SARIMAX CO(GT) Results
==================================
Dep. Variable:                  CO(GT)
No. Observations:                5899
Model:             SARIMAX(2, 0, 3)x(3, 0, [1], 24)
Log Likelihood:           -2714.955
Date:              Sun, 08 Dec 2024
Time:                     22:05:11
AIC:                      5453.909
BIC:                      5534.100
HQIC:                     5481.781
Covariance Type:               opg
==================================
```

TABLE IV
SUMMARY OF SARIMAX MODELING RESULTS FOR NO2(GT)

```
==================================
Summary of SARIMAX NO2(GT) Results
==================================
Dep. Variable:                 NO2(GT)
No. Observations:               5899
Model:             SARIMAX(2, 0, 3)x(3, 0, [1], 24)
Log Likelihood:           -1580.973
Date:              Sun, 08 Dec 2024
Time:                     22:07:44
AIC:                      3185.947
BIC:                      3266.137
HQIC:                     3213.818
Covariance Type:               opg
==================================
```

TABLE V
SUMMARY OF SARIMAX MODELING RESULTS FOR RH

```
==================================
Summary of SARIMAX RH Results
==================================
Dep. Variable:                     RH
No. Observations:                5899
Model:             SARIMAX(3, 1, 3)x(1, 0, [1], 24)
Log Likelihood:            -86.286
Date:              Sun, 08 Dec 2024
Time:                     22:08:04
AIC:                      194.572
BIC:                      268.078
HQIC:                     220.120
Covariance Type:               opg
==================================
```

3) Forecast Implementation: With the SARIMAX models trained, forecasting was performed on the test set as follows:

- **Generating Forecasts**: For each target variable, forecasts were generated for the entire test period using the fitted SARIMAX model. The forecasts incorporated the exogenous variables to enhance prediction accuracy.
- **Inverse Transformation**: To interpret the forecasted values in their original scale, the inverse of the scaling transformation was applied to both the forecasted and actual test data. This step should ensure that the results are comparable to the original data.
- **Performance Evaluation**: The forecasted values were compared against the actual test set values using evaluation metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).
- **Visualization**: The actual and forecasted values were plotted to visually assess the model's forecasting capability. These plots are referenced as Figure 3, Figure 4, and Figure 5 for `CO(GT)`, `NO2(GT)`, and `RH` respectively.

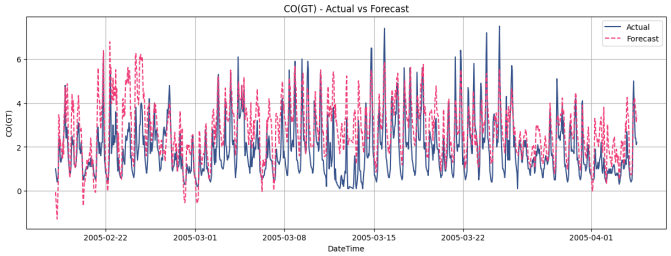

Fig. 3. SARIMAX Forecast vs Actual for `CO(GT)`

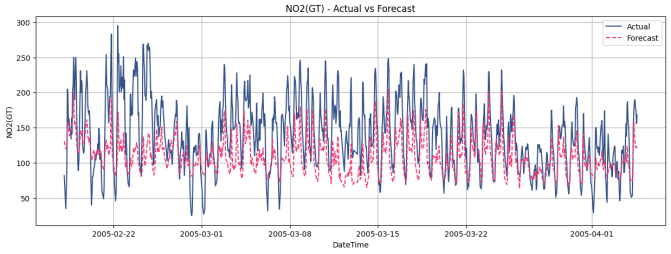

Fig. 4. SARIMAX Forecast vs Actual for `NO2(GT)`

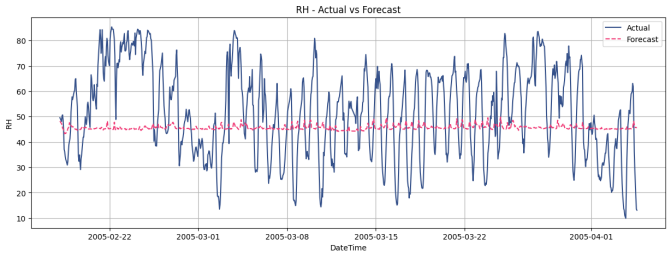

Fig. 5. SARIMAX Forecast vs Actual for `RH`

F. Model Results Comparison

The forecasting results for both models can be found in Figure 6 for the VAR model, and Figure 7 for the SARIMAX model. Both models generally follow the trends for the short forecast. The SARIMAX model exhibits much more erratic behavior, with pronounced spikes and variability from point to point. Despite this, these spikes often align with the general trends and, in some instances, match the actual data more closely. This suggests that SARIMAX captures some finer-grained fluctuations in the data that the VAR model smooths over. In contrast, the VAR model produces smoother forecasts that adhere closely to the overall trends but may fail to capture sudden changes or short-term variability.

1) Evaluation: The performance of the VAR and SARIMAX models was evaluated using three key approaches: root mean squared error (RMSE) and mean absolute error (MAE), residual analysis via residual plots, and Ljung-Box tests for autocorrelation in residuals. These methods collectively assess the predictive accuracy, residual behavior, and goodness-of-fit of the models.

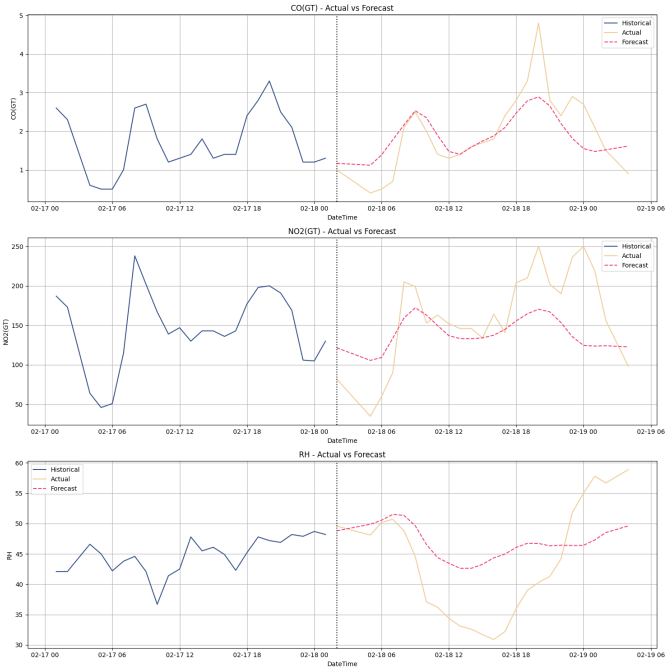

Fig. 6. 24 hours forecast by the VAR model.

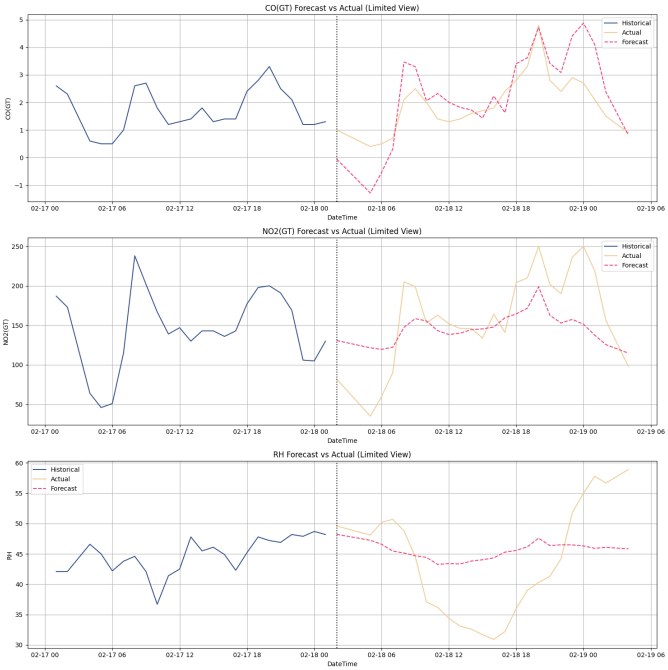

Fig. 7. 24 hours forecast by the SARIMAX model.

a) RMSE and MAE: The RMSE and MAE values for each model and variable are presented in Table VI bellow. These metrics quantify the predictive accuracy of the models, with lower values indicating better performance.

TABLE VI
RMSE AND MAE FOR VAR AND SARIMAX MODELS

Model	Variable	RMSE	MAE
VAR	CO(GT)	1.3137	1.0592
VAR	NO2(GT)	54.1367	43.0682
VAR	RH	17.8101	14.9613
SARIMAX	CO(GT)	1.2144	1.0202
SARIMAX	NO2(GT)	45.8254	36.9386
SARIMAX	RH	18.5715	15.4967

For both `CO(GT)` and `NO2(GT)`, the SARIMAX model outperformed the VAR model with marginally lower RMSE and MAE values, suggesting better predictive accuracy. However, for `RH`, the VAR model demonstrated slightly better performance.

b) Residual Analysis: Residual plots for both models were examined to evaluate the distribution and behavior of forecast errors. Both models display a slight leftward skew in the residuals for `CO(GT)` and `NO2(GT)`, with the skew more pronounced in the VAR model. Residuals for `RH` appeared more symmetric in comparison. This skewness indicates that forecast errors are not perfectly centered around zero. These plots can be found at Figure 8, 9, 10 for the VAR model, and at Figure 11, 12, 13 for the SARIMAX model.

c) Ljung-Box Tests: The Ljung-Box test was conducted to assess the presence of autocorrelation in the residuals. The results are summarized in Table VII. For the VAR model, the p-values for all three `CO(GT)`, `NO2(GT)` and `RH` suggest no autocorrelation. For SARIMAX, the p-values for `CO(GT)` and `NO2(GT)` indicate no significant autocorrelation. However, the p-value for `RH` is bellow 0.05 which means H_0 is rejected, suggesting residual autocorrelation and potential overfitting or misfit for this variable. This observation potentially aligns with the forecast observed in Figure 5 which appear to fail to forecast to some degree.

TABLE VII
LJUNG-BOX TEST RESULTS FOR VAR AND SARIMAX MODELS

Model	Variable	Ljung-Box Statistic (p-value)
VAR	CO(GT)	15.0274 (0.9690)
VAR	NO2(GT)	26.1028 (0.5129)
VAR	RH	17.4645 (0.9190)
SARIMAX	CO(GT)	0.0008 (0.9771)
SARIMAX	NO2(GT)	0.1018 (0.7497)
SARIMAX	RH	12.5760 (0.0004)

d) Shapiro-Wilk Test for Normality: The Shapiro-Wilk test was applied to evaluate whether residuals follow a normal distribution. For both models and all variables, the p-values were approaching zero, clearly rejecting the null hypothesis of normality. This highlights deviations from the normality assumption.

IV. DISCUSSION

The VAR model was built with the goal of testing how well the data performs as a single Vector Autoregressive Model, as stipulated by the task. The SARIMAX model was chosen to investigate whether treating the data as three univariate time series would yield better performance. Determining which model performed better overall proved challenging; however, specific areas clearly highlight the strengths of each model. A notable example is the RH forecasts. Although the RMSE and MAE values do not suggest a significant difference between the models, Figures 5 and 16 illustrate that both models produced relatively flat forecasts over the entire test set. The VAR model forecast begins at the start of the test set, relying on the last observed values from the training set as inputs. In contrast, the SARIMAX model generates a rolling forecast throughout the test set, continuously updating predictions based on its internal structure and the provided exogenous variables. This methodological difference is evident when comparing the CO(GT) and NO2(GT) forecasts, as shown in Figures 14 and 3 for CO forecasts, and Figures 15 and 4 for NO2(GT), these plots should not be similar. The discrepancy in the RH forecasts becomes even clearer when considering the Ljung-Box test results, where the RH variable rejected H_0 for the SARIMAX model, indicating residual autocorrelation and suggesting that the SARIMAX model may be unfit for this variable. Additionally, in the short-term predictions, the VAR model appears to follow the actual trends (yellow plot) more closely.

While looking at short term forecasts, the SARIMAX model appears to perform slightly better in predicting CO levels. The SARIMAX forecast is more erratic but follows the general trends well; however, this erratic nature also results in overshooting forecasts more frequently compared to the VAR model. The VAR CO forecast captures the general trends accurately but fails to capture the extremes of the data. These observations are supported by the RMSE and MAE scores, which are marginally better for the SARIMAX forecasts. This suggests that the SARIMAX model is slightly better at capturing systematic patterns and dependencies in the CO data compared to the VAR model.

The NO2 forecasts, however, present a more clear-cut comparison. The SARIMAX model significantly benefited from its more erratic behavior, which, while not as pronounced in the plotted forecasts, is evident in the evaluation metrics. The SARIMAX model achieved a 6-point lower MAE and nearly a 9-point lower RMSE compared to the VAR model. This performance is further supported by the Ljung-Box test results, where SARIMAX scored 0.25 points higher in p-value than the VAR model. Additionally, the residual distribution of the NO2 model for SARIMAX is more aligned around 0 and denser compared to the VAR model. Residuals being more aligned around 0 indicate that the SARIMAX model better captures the central tendency of the data, leading to forecasts that, on average, are closer to the observed values. The denser residual distribution implies that the SARIMAX

model produced fewer extreme errors, as supported by the forecasts shown in Figures 7 and 6. However, the SARIMAX model exhibits considerably higher skewness than the VAR model, despite its higher density. This suggests that while most errors are small, a few relatively larger errors are pulling the distribution away from symmetry. This could suggest overfitting in some cases or sensitivity to certain data points, like the cluster of outliers at the second half of the year within the dataset.

V. Conclusion

Based on the observations, the RH forecast demonstrates a clear prefrence for modeling using the VAR approach, as it effectively captures the interdependencies across the entire dataset rather than treating the variable as an isolated univariate time series. In contrast, the NO2 forecast significantly benefited from being modeled as a standalone variable, highlighting the effectiveness of the SARIMAX model for this case. The CO variable, however, shows no clear preference between the two approaches, suggesting that it can be modeled effectively either as part of an interconnected system or independently. These findings show a clear preferential difference for each separate variable in the dataset, and underscores the importance of understanding the nature of each variable when selecting modeling strategies for multivariate time series data.

VI. Areas of Improvement

1) Time Management: Late in the task, I recalled the KPSS test as an alternative method for assessing stationarity. While the Augmented Dickey-Fuller (ADF) test is generally effective, it can occasionally overlook subtle quadratic trends. Upon conducting the KPSS test, it marginally rejected H_0, indicating a slight non-stationarity in the data. However, applying logarithmic transformations or differencing to address this issue negatively impacted the overall forecasting performance. Despite these limitations, the RMSE and MAE metrics remained within acceptable ranges. Due to time constraints, I decided to proceed without fully addressing these factors.

2) Lag Order Selection: During the development of the VAR model, I used an unconventional method for determining the optimal lag order by combining the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC):

```
optimal_lag = int(lag_order_results.bic + (abs(
    lag_order_results.aic - lag_order_results.bic) /
     2))
```

While writing the report, I realized this approach is not standard practice in time series analysis. Typically, lag order selection is based on the criterion that independently minimizes either AIC or BIC, rather than a hybrid combination of the two. In hindsight, it would have been more rigorous to experiment with both criteria separately and determine which one yielded the best-performing model for this dataset.

3) SARIMAX RH: The Ljung-Box test issues identified with the RH variable in the SARIMAX model may have been caused by an implementation error on my part. Unfortunately, I did not have sufficient time to investigate this issue thoroughly after discovering it.

VII. Acknowledgment

I would like to thank the creators of ChatGPT and Claude for providing valuable tools that assisted with code debugging as well as spelling and thought clarification during the development of this report.

Appendices

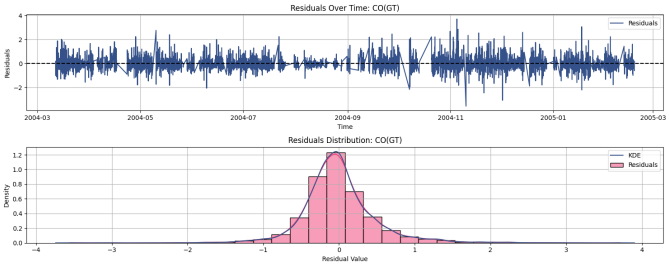

Fig. 8. Graph depicting residual spread and density for the VAR forecast on the CO(GT) variable.

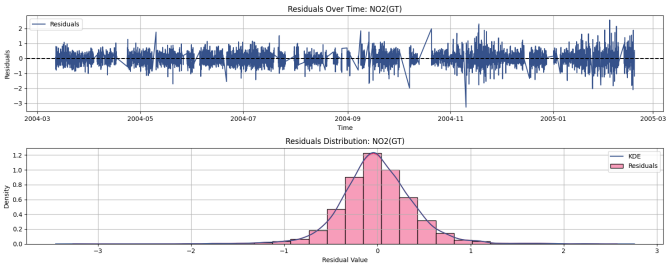

Fig. 9. Graph depicting residual spread and density for the VAR forecast on the NO2(GT) variable.

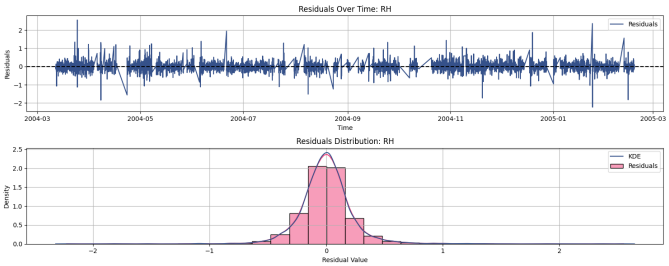

Fig. 10. Graph depicting residual spread and density for the VAR forecast on the RH variable.

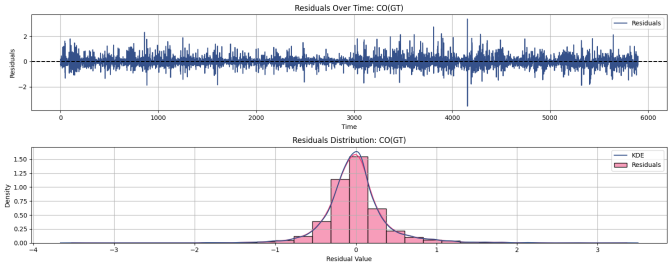

Fig. 11. Graph depicting residual spread and density for the SARIMAX forecast on the CO(GT) variable.

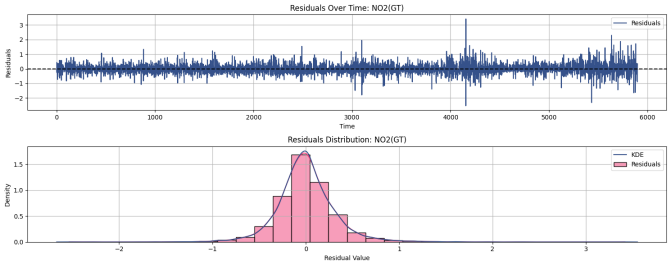

Fig. 12. Graph depicting residual spread and density for the SARIMAX forecast on the NO2(GT) variable.

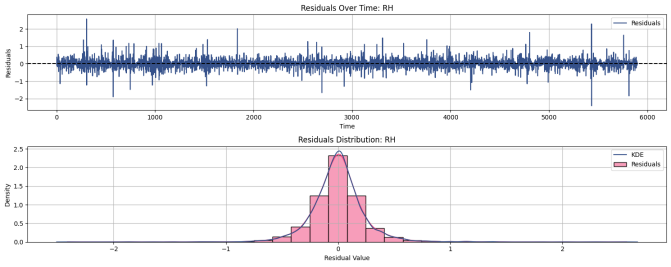

Fig. 13. Graph depicting residual spread and density for the SARIMAX forecast on the RH variable.

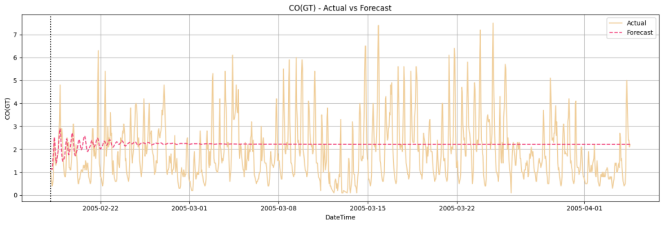

Fig. 14. Graph showing the VAR model CO forecast displayed with the entire Test dataset.

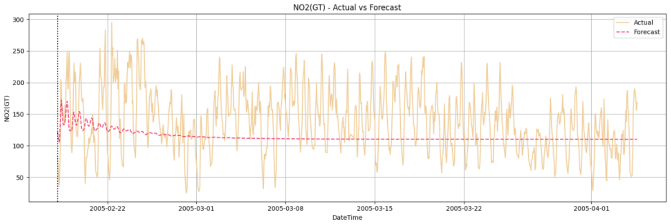

Fig. 15. Graph showing the VAR model NO2 forecast displayed with the entire Test dataset.

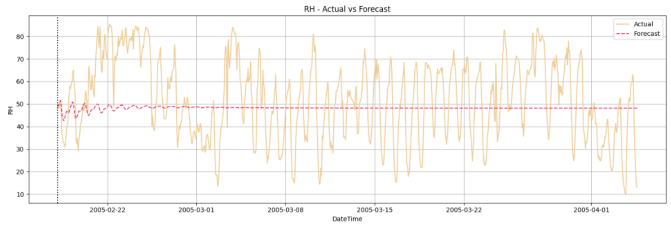

Fig. 16. Graph showing the VAR model RH forecast displayed with the entire Test dataset.

VIII. References

References

S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "Air quality dataset," 2008, accessed: 2024-12-06. [Online]. Available: https://archive.ics.uci.edu/dataset/360/air+quality

W. H. Organization, "Environmental health criteria 213: Carbon monoxide," 1999, accessed: 2024-12-06. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/42180/WHO_EHC_213.pdf

——, "Who global air quality guidelines: Particulate matter (pm2.5 and pm10), ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide," 2021, accessed: 2024-12-06, Page 111. [Online]. Available: https://iris.who.int/bitstream/handle/10665/345329/9789240034228-eng.pdf

L. Spinelle, M. Gerboles, G. Kok, V. Puygrenier, T. Sauerwald, and F. Bonavitacola, "Influence of humidity and temperature on the performance of low-cost gas sensors for air quality monitoring," *Sensors*, vol. 20, no. 18, p. 5175, 2020, accessed: 2024-12-06. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5175

Statsmodels Developers, *Statsmodels Documentation*, 2024, accessed: 2024-12-07. [Online]. Available: https://www.statsmodels.org/stable/index.html

C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980. [Online]. Available: https://doi.org/10.2307/1912017

V. Hassani, "Time series analysis: Vector auto-regressive (var) models," 2023, lecture slides, accessible through Canvas. Accessed on December 7, 2024. [Online]. Available: https://kristiania.instructure.com/courses/12539/files/1448319?wrap=1

G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day, 1970.

M. Majka, "Arimax: Time series forecasting with external variables," *ResearchGate*, 2024, accessed on December 7, 2024. [Online]. Available: https://www.researchgate.net/publication/384196976_ARIMAX_Time_Series_Forecasting_with_External_Variables

V. Shah, N. Patel, D. Shah, D. Swain, M. Mohanty, B. Acharya, V. C. Gerogiannis, and A. Kanavos, "Forecasting maximum temperature trends with sarimax: A case study from ahmedabad, india," *Sustainability*, vol. 16, no. 16, p. 7183, 2024. [Online]. Available: https://www.mdpi.com/2071-1050/16/16/7183

Statsmodels Developers, *statsmodels.tsa.statespace.sarimax.SARIMAX: Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors*, 2024, accessed on December 7, 2024. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979, accessed: 2024-12-06. [Online]. Available: https://www.jstor.org/stable/2286348

H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974, accessed: 2024-12-06. [Online]. Available: https://doi.org/10.1109/TAC.1974.1100705

G. E. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978, accessed: 2024-12-06. [Online]. Available: https://doi.org/10.1214/aos/1176344136

G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978. [Online]. Available: https://doi.org/10.1093/biomet/65.2.297

S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965. [Online]. Available: https://doi.org/10.2307/2333709