

STAT-S 650 - TIME SERIES ANALYSIS - Spring 2023

RELATIVE HUMIDITY FORECASTING

SATYA PRIYANKA PONDURU

sponduru@iu.edu

Data Science at Indiana University

1. INTRODUCTION

In our everyday lives, the weather affects so much more than just whether we need an umbrella or a coat. It impacts how our crops grow, the quality of the air we breathe indoors, and even our overall health. A big part of understanding the weather is predicting something called relative humidity, which tells us how much moisture is in the air compared to how much it could hold at a given temperature.

But figuring out what the humidity will be, especially at specific times like 3 PM, is no easy task. The atmosphere is a complex system with lots of moving parts, like temperature, wind, and other factors that all interact in intricate ways. Traditional weather forecasts often struggle to get it right, leaving us with uncertainties and sometimes inaccurate predictions.

That is where our project comes in. We are using ARIMA and ARIMAX models to help us forecast humidity levels at 3 PM. These models crunch a lot of data about past humidity, temperature, wind speed, and more to give us a better idea of what to expect in the future.

What is cool about our approach is that we are not just looking at the obvious stuff. We are also considering things like how humid it is in the morning and what the temperature is like. By considering these extra factors, we can make our predictions even more accurate and consider how humidity changes throughout the day.

Our project is all about recognizing how important it is to get these forecasts right. They are crucial for farmers who need to know the best conditions for their crops, and for folks managing energy use, where humidity can affect how well heating and cooling systems work. By giving decision-makers timely and accurate forecasts, we are helping them make smarter choices, use resources more efficiently, and avoid any problems that might pop up because of humidity.

With our high-tech methods and careful analysis of data, we are aiming to make a real difference in how we understand and deal with humidity in our environment.

2. DATASET DESCRIPTION

The dataset comprises observations pertaining to various meteorological variables recorded at 9 AM, potentially instrumental in weather forecasting and environmental analysis.

air_pressure_9am: This variable denotes the air pressure measured at 9 AM. The minimum recorded pressure is 908.0, while the maximum is 929.3 millibars. There are three missing values in this variable.

air_temp_9am: Represents the air temperature measured at 9 AM. The range of temperatures spans from 36.75 to 98.91 degrees Fahrenheit, with five missing values.

avg_wind_direction_9am: Indicates the average wind direction observed at 9 AM. The direction varies between 15.50 and 343.40 degrees. Four missing values are present.

avg_wind_speed_9am: Reflects the average wind speed measured at 9 AM. The speed ranges from 0.6935 to 23.5550 miles per hour. There are three missing values.

max_wind_direction_9am: Denotes the maximum wind direction recorded at 9 AM. The direction varies from 1.186 to 29.841 degrees. Three missing values exist.

max_wind_speed_9am: Represents the maximum wind speed observed at 9 AM. The speed ranges from 28.90 to 312.20 miles per hour, with four missing values.

rain_accumulation_9am: This variable indicates the amount of rain accumulation by 9 AM. The range of values spans from 0.0000 to 24.0200 inches, with six missing values.

rain_duration_9am: Reflects the duration of rain recorded by 9 AM. The duration ranges from 0.0 to 17704.0 minutes, with three missing values.

relative_humidity_9am: Represents the relative humidity measured at 9 AM. The values vary from 6.09% to 92.62%, with no missing values.

relative_humidity_3pm: Denotes the relative humidity recorded at 3 PM. The range of values spans from 5.30% to 92.25%, with no missing values.

Overall, the dataset offers a comprehensive snapshot of meteorological conditions at 9 AM, providing valuable insights for weather analysis and forecasting applications.

3. OBJECTIVES


1. **Temporal Dependency of Relative Humidity:** The degree of temporal dependency in 3 PM relative humidity values are explored, particularly in relation to morning relative humidity and other meteorological variables.
2. **Impact of Lagged Relative Humidity:** Including lagged relative humidity measurements at 9 AM significantly improves the accuracy of 3 PM relative humidity forecasts.
3. **Influence of Morning Weather Conditions:** Morning meteorological variables (e.g., air temperature, wind speed, rainfall) have a significant impact on afternoon relative humidity levels.
4. **Effectiveness of ARIMA vs. ARIMAX Models:** ARIMAX models outperform simple ARIMA models in forecasting 3 PM relative humidity, as evidenced by lower AIC, BIC, and improved forecast accuracy.
5. **Stationarity and Autocorrelation:** Differencing the data improves stationarity and reduces autocorrelation, leading to more accurate forecasting models.
6. **Impact of Other Meteorological Variables:** The contribution of variables like wind speed or air pressure to the variability in afternoon humidity levels is assessed beyond the effect of morning humidity alone.

4. METHODOLOGY & RESULTS

4.1 Importing Data and Necessary Libraries:

To begin the analysis for our weather forecasting project, we first need to import the necessary data from the CSV file into R. Additionally, we will load the essential packages like tidyr, tseries, ggplot2, forecast, fable, reshape2 and lmtest to facilitate the analysis process. These packages provide the functionality required to perform various statistical analyses and visualize the data effectively.

```
##{r}
data <- read.csv("daily_weather.csv")
summary(data)
head(data)
```



Description: df [6 x 11]

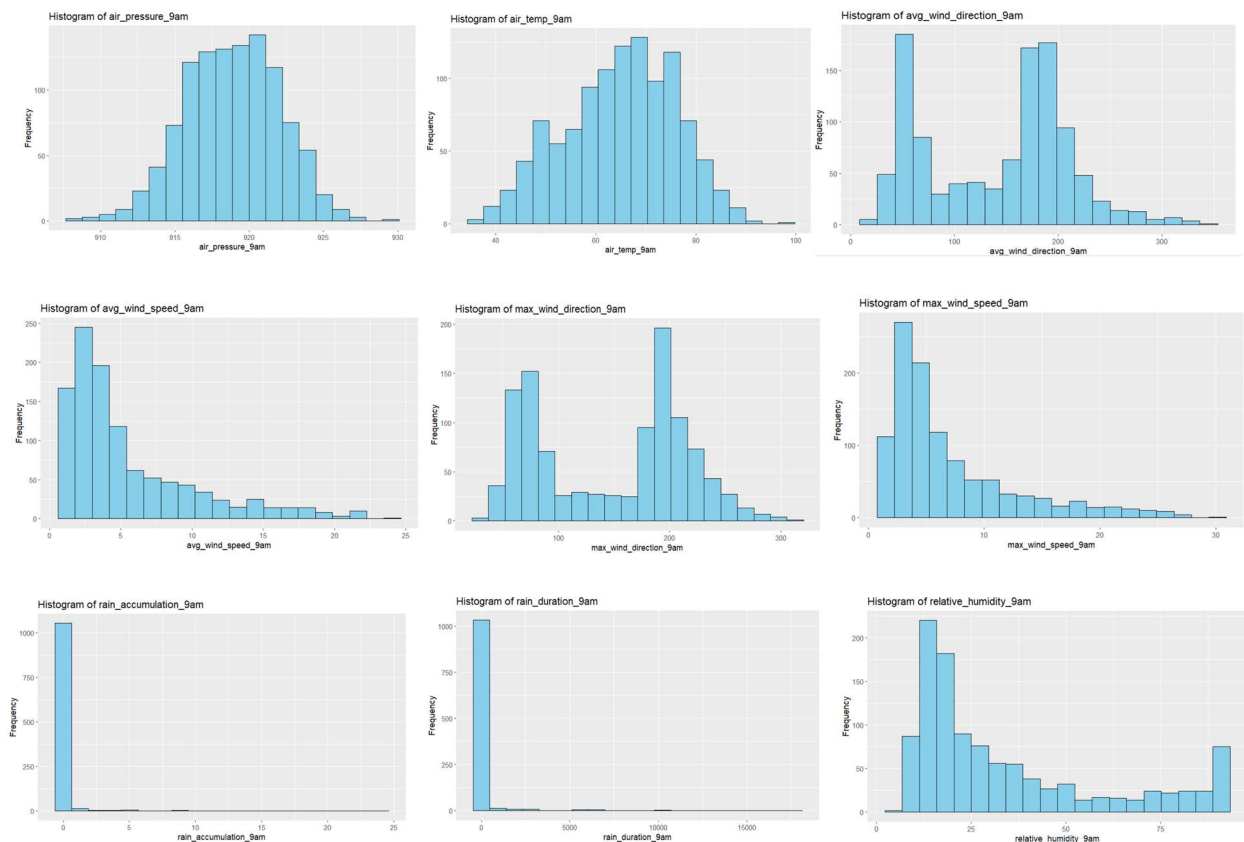
	number <int>	air_pressure_9am <dbl>	air_temp_9am <dbl>	avg_wind_direction_9am <dbl>	avg_wind_speed_9am <dbl>	max_wind_direction_9am <dbl>
1	0	918.0600	74.82200	271.1000	2.080354	295.4000
2	1	917.3477	71.40384	101.9352	2.443009	140.4715
3	2	923.0400	60.63800	51.0000	17.067852	63.7000
4	3	920.5028	70.13889	198.8321	4.337363	211.2033
5	4	921.1600	44.29400	277.8000	1.856660	136.5000
6	5	915.3000	78.40400	182.8000	9.932014	189.0000

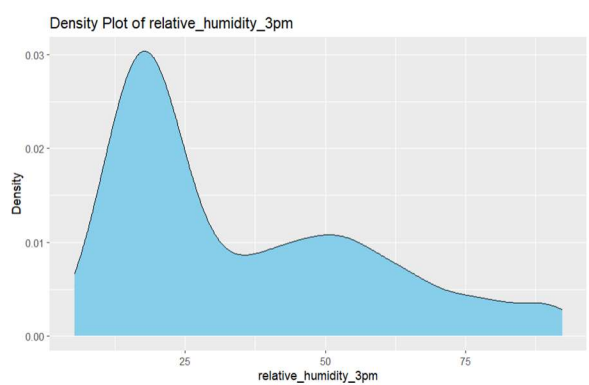
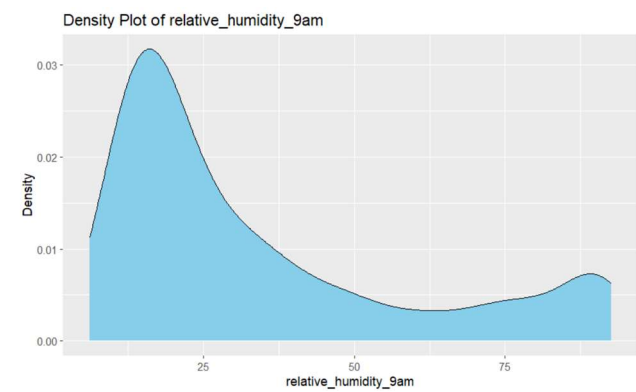
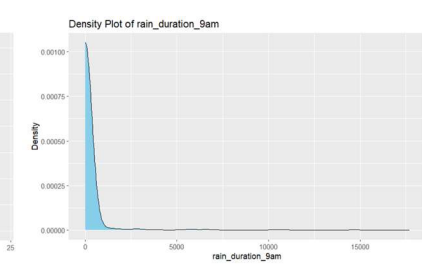
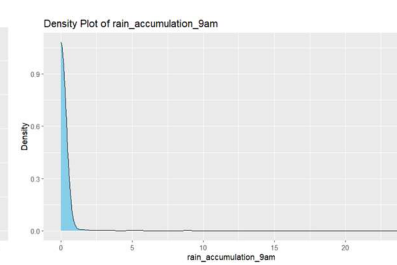
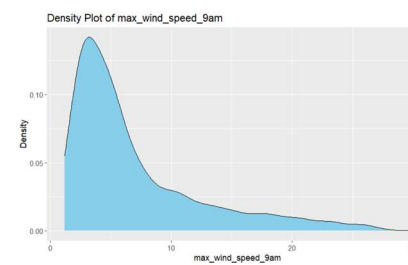
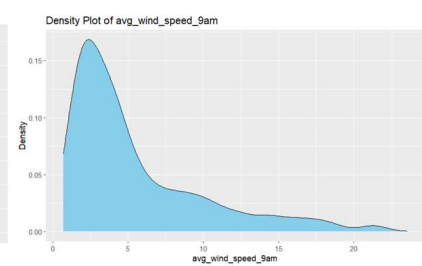
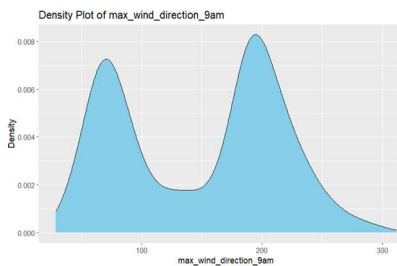
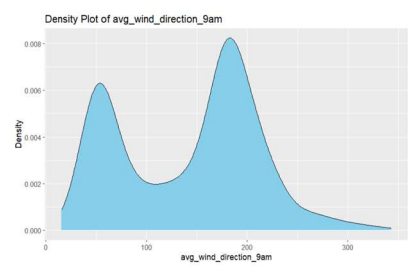
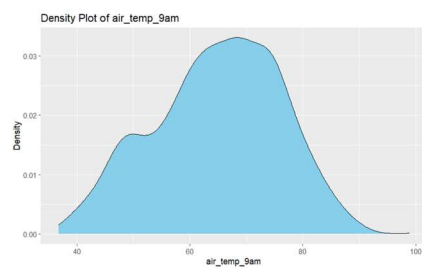
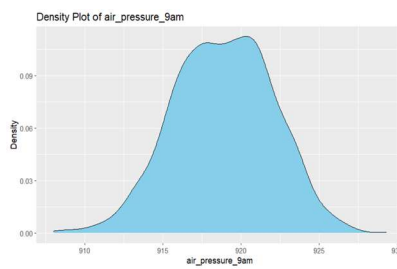
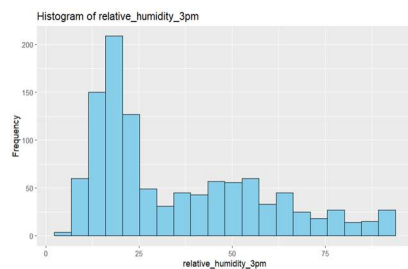
6 rows | 1-7 of 11 columns

4.2 Data Preprocessing:

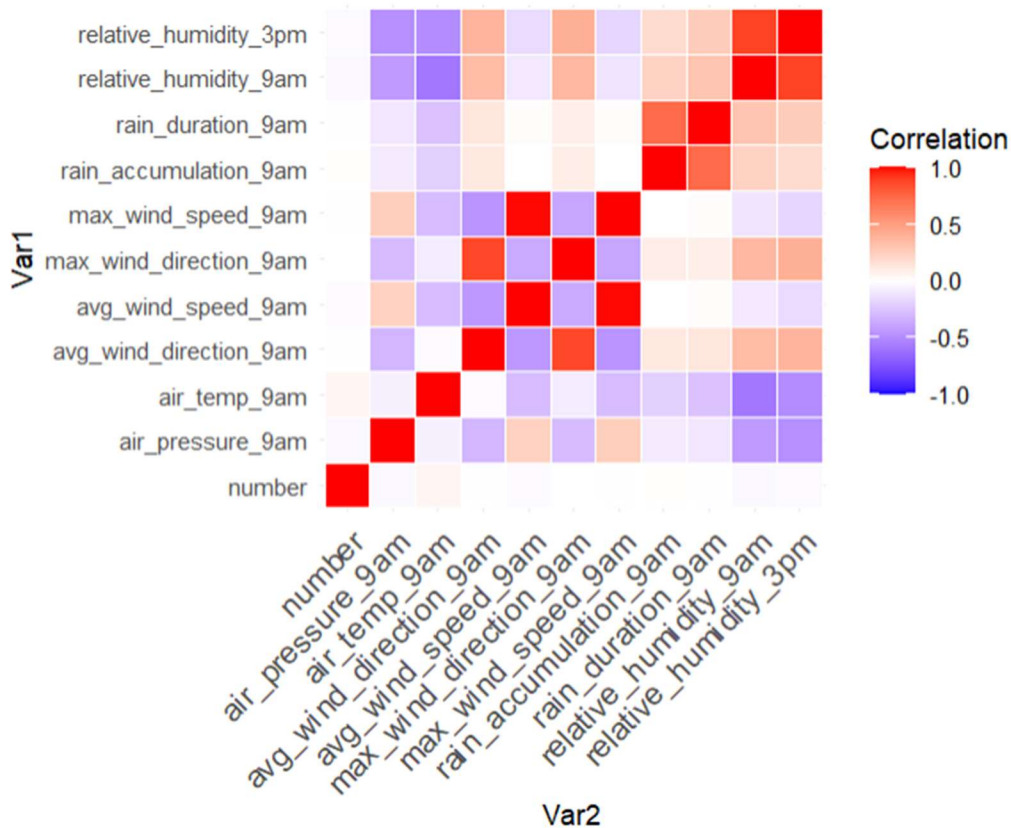
4.2.1 Handling Missing Values: Before proceeding with any analysis, it is essential to manage missing values present in the dataset. In this case, the dataset contains missing values for several variables, such as `air_pressure_9am`, `air_temp_9am`, `avg_wind_direction_9am`, `avg_wind_speed_9am`, `max_wind_direction_9am`, `max_wind_speed_9am`, `rain_accumulation_9am`, and `rain_duration_9am`. One common approach to dealing with missing values is to remove the corresponding observations or rows from the dataset using the `na.omit()` function.

4.2.2 Visualizing Variable Distributions: To gain insights into the distribution of each variable, we can plot histograms and density plots. These visualizations help identify the shape of the distribution, potential outliers, and any skewness or modality present in the data. The `ggplot2` package in R provides powerful tools for creating these plots in a concise and visually appealing manner.





4.2.3 Correlation Analysis: Understanding the correlation between variables is crucial for identifying potential predictors and multicollinearity issues. We can compute the correlation matrix using the `cor()` function and visualize it as a heatmap using `ggplot2`.



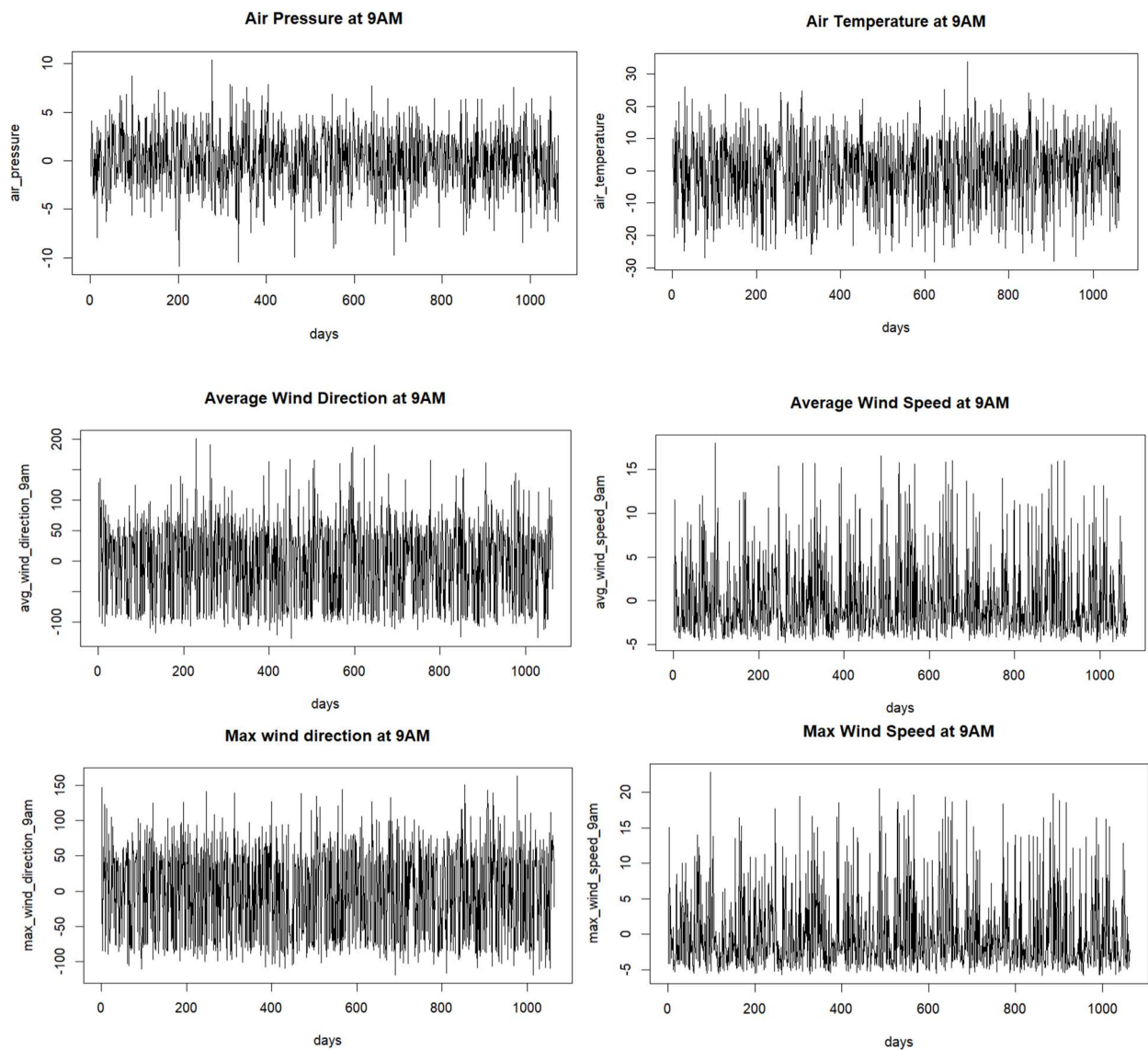
- Relative humidity at 3pm has the highest positive correlation with relative humidity measured earlier at 9am. This suggests using 9am relative humidity as an important predictor variable for forecasting 3pm relative humidity levels.
- There are moderate positive correlations between relative humidity at 3pm and variables like rain duration, rain accumulation, and wind speeds measured at 9am. Including these rainfall and wind variables could potentially improve the 3pm relative humidity forecast.
- Relative humidity at 3pm exhibits a moderate negative correlation with air temperature at 9am. Incorporating morning temperature readings may help account for the inverse relationship between temperature and afternoon humidity levels.
- Wind direction variables (avg/max at 9am) do not show strong correlations with 3pm relative humidity, indicating they may not be as relevant for this specific forecasting task.
- Air pressure at 9am also has a weak correlation with the target 3pm humidity, suggesting limited predictive value from this variable alone.

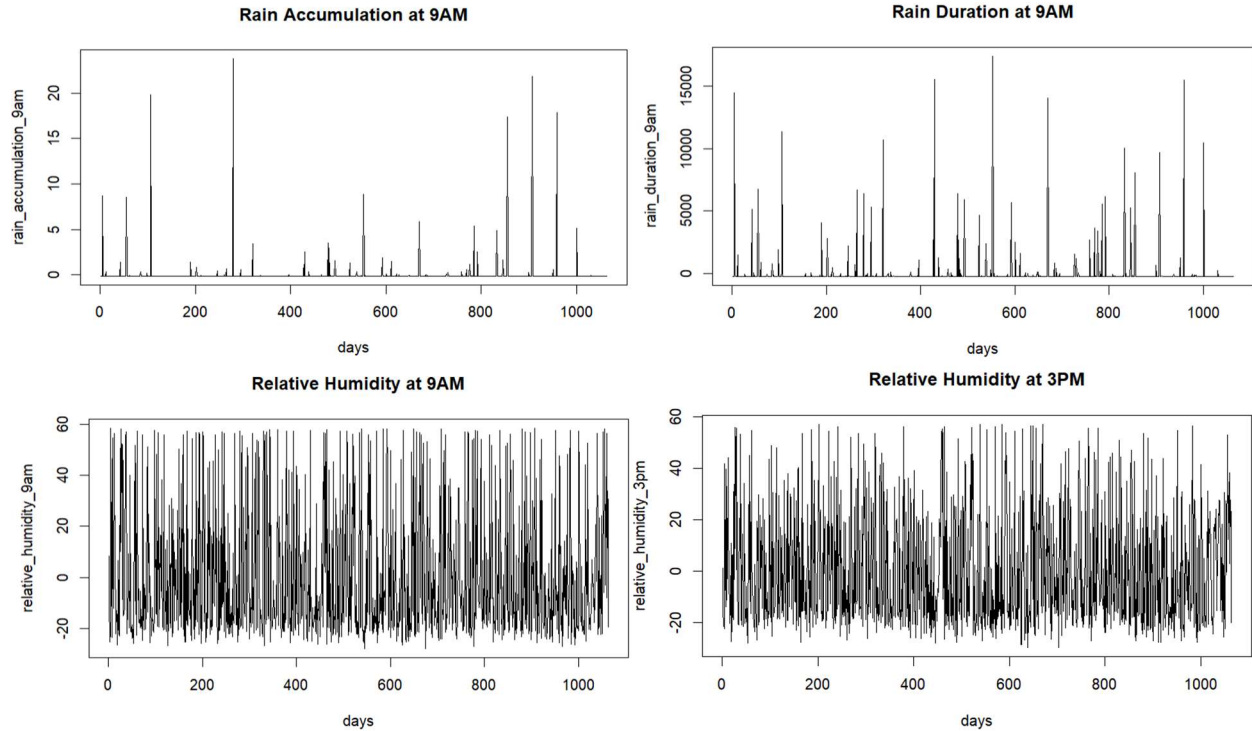
Overall, the matrix highlights that incorporating same-day morning readings of humidity, rainfall, wind speed, and temperature could be beneficial for building an effective model to predict afternoon relative humidity values. Wind direction and pressure appear less correlated and potentially less important predictors for this specific forecasting problem.

4.3 Time Series Analysis:

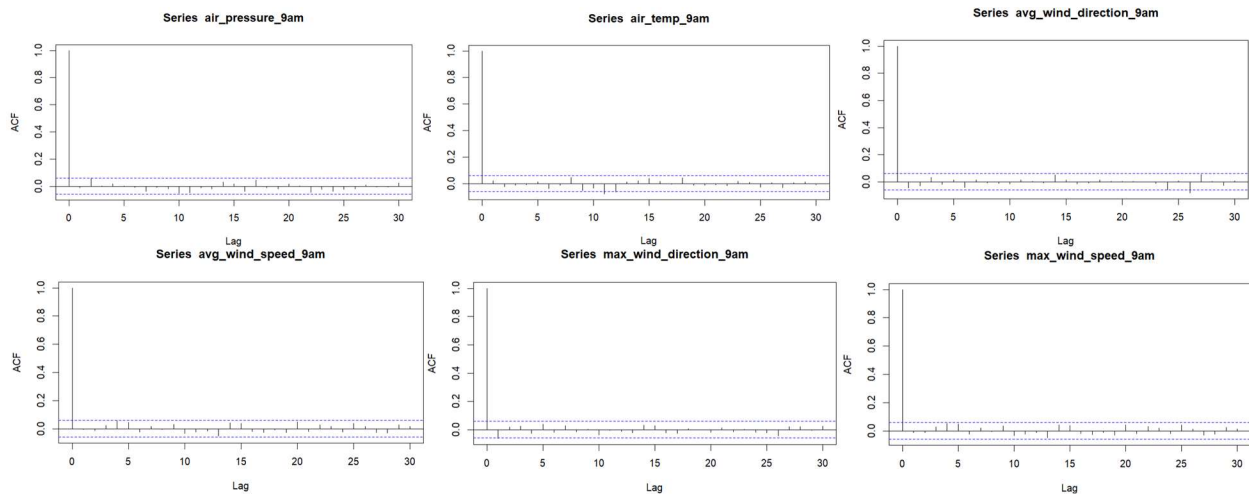
4.3.1 Data Standardization and Time Series Conversion:

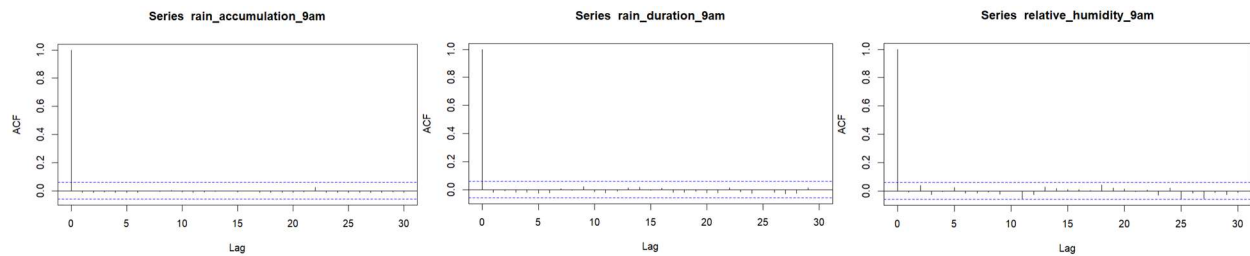
- **Standardizing the Data:** Each variable in the dataset is standardized by subtracting its mean value from every data point. This process ensures that each variable has a mean of zero, making them comparable across different scales. Standardization is essential for many statistical analyses, including time series modeling, as it removes the influence of scale differences between variables.
- **Converting to Time Series Object:** After standardization, the variables are converted into time series objects using the `ts` function in R. Time series objects are suitable for analyzing data collected over time, allowing for the detection of temporal patterns and trends. The **start** and **end** arguments specify the beginning and end points of the time series, respectively.



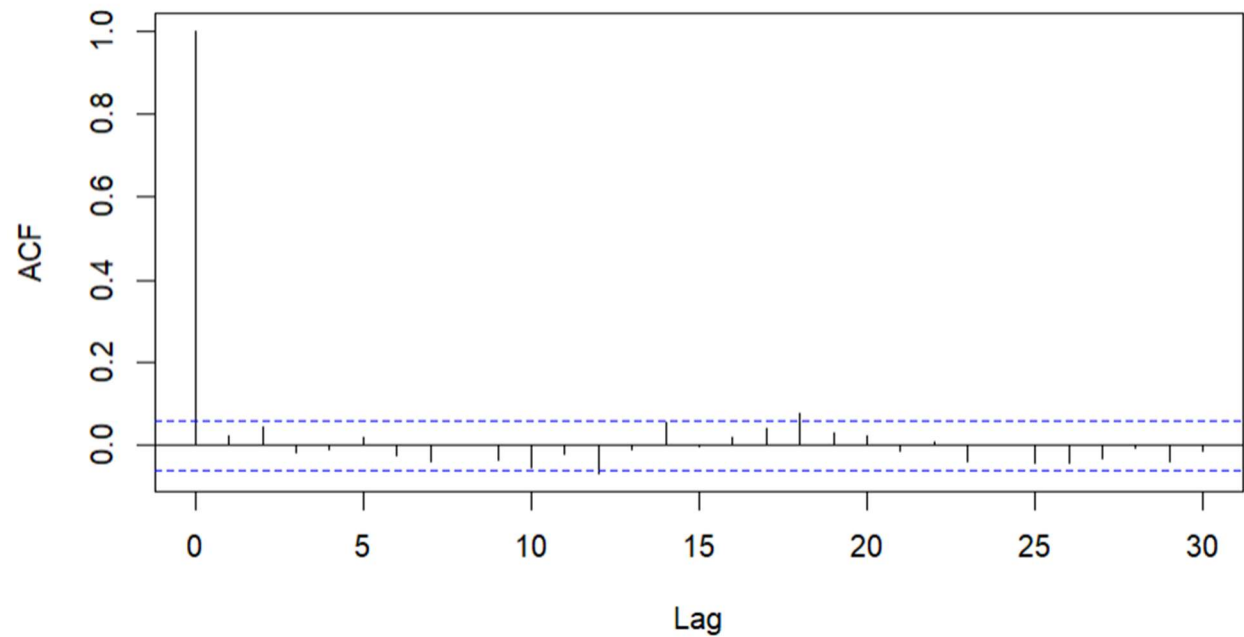


4.3.2 Autocorrelation Analysis: Autocorrelation Function (ACF) plots provide insights into the temporal dependencies within time series data. The ACF plots generated before differencing allow us to assess the degree of correlation between each variable and its lagged values. Peaks and patterns in these plots indicate serial correlation, aiding in model selection. We will conduct ACF analysis after differencing to further refine our understanding of the data's temporal structure.

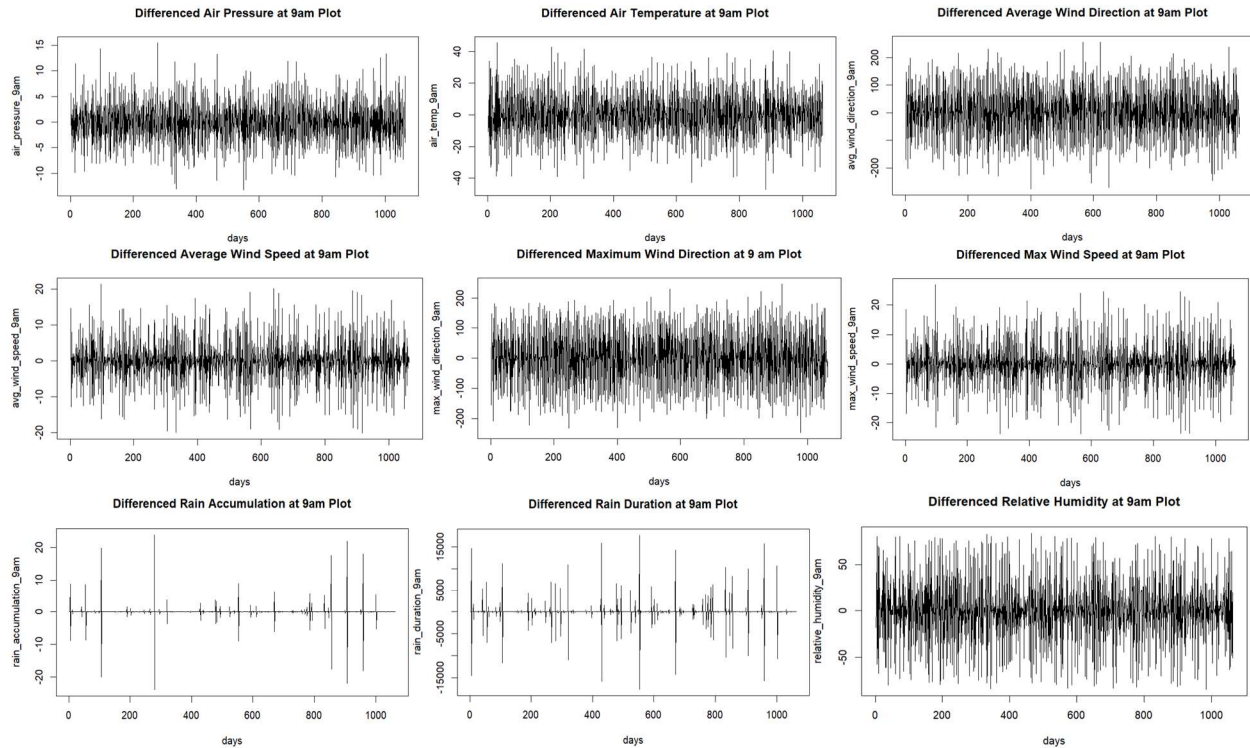




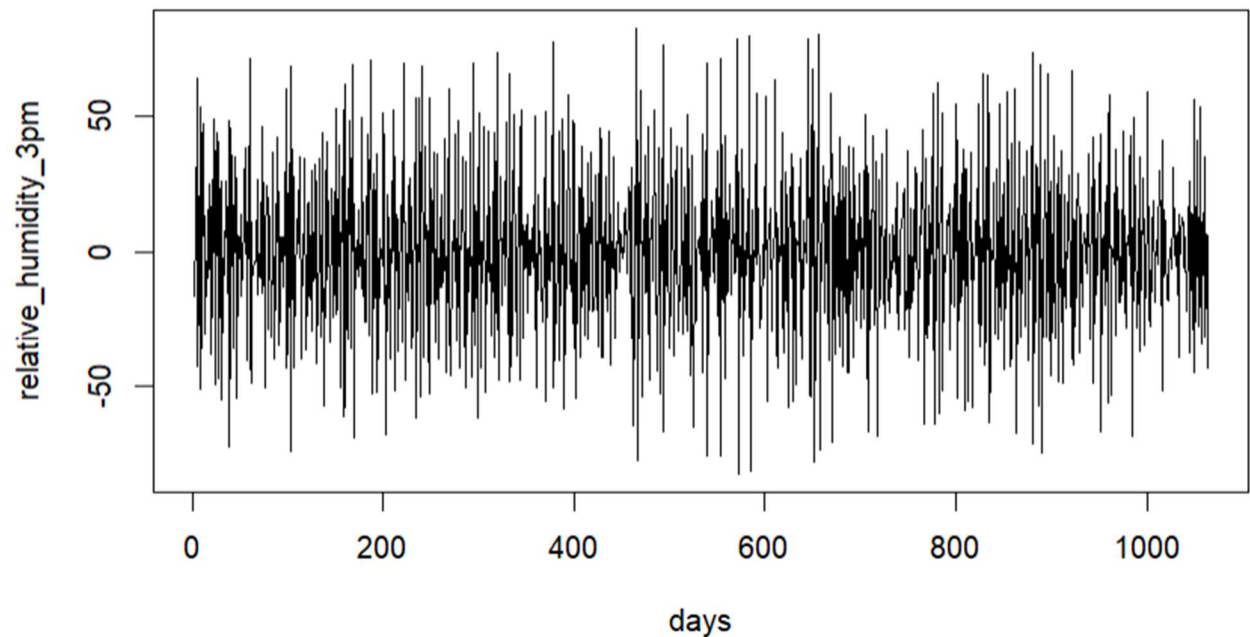
Series relative_humidity_3pm



4.3.2 Differencing for Stationarity: Differencing the variables is a crucial step in time series analysis to achieve stationarity, where the mean and variance of the data remain constant over time. In this section, we compute the first differences of the variables by taking the lag-1 differences between consecutive observations. This process helps stabilize the series and remove trends or seasonality.



Differenced Relative Humidity at 3pm Plot



4.3.3 ADF Test: The Augmented Dickey-Fuller (ADF) Test is a statistical test used to determine whether a time series is stationary or non-stationary. Here ADF Test results indicate whether differenced time series variables have achieved stationarity after applying differencing techniques.

The key components of the ADF Test results are as follows:

Dickey-Fuller Statistic: This statistic measures the significance of the test. A more negative value suggests stronger evidence against the presence of a unit root (i.e., non-stationarity).

Lag Order: This indicates the number of lagged differences included in the test. It helps capture any autocorrelation in the differenced series.

p-value: The p-value determines the statistical significance of the test. A small p-value (typically less than 0.05) suggests strong evidence against the null hypothesis of non-stationarity, indicating that the series is likely stationary.

In the results provided, the warning message "p-value smaller than printed p-value" suggests that the actual p-value is smaller than the default printed value, indicating high statistical significance.

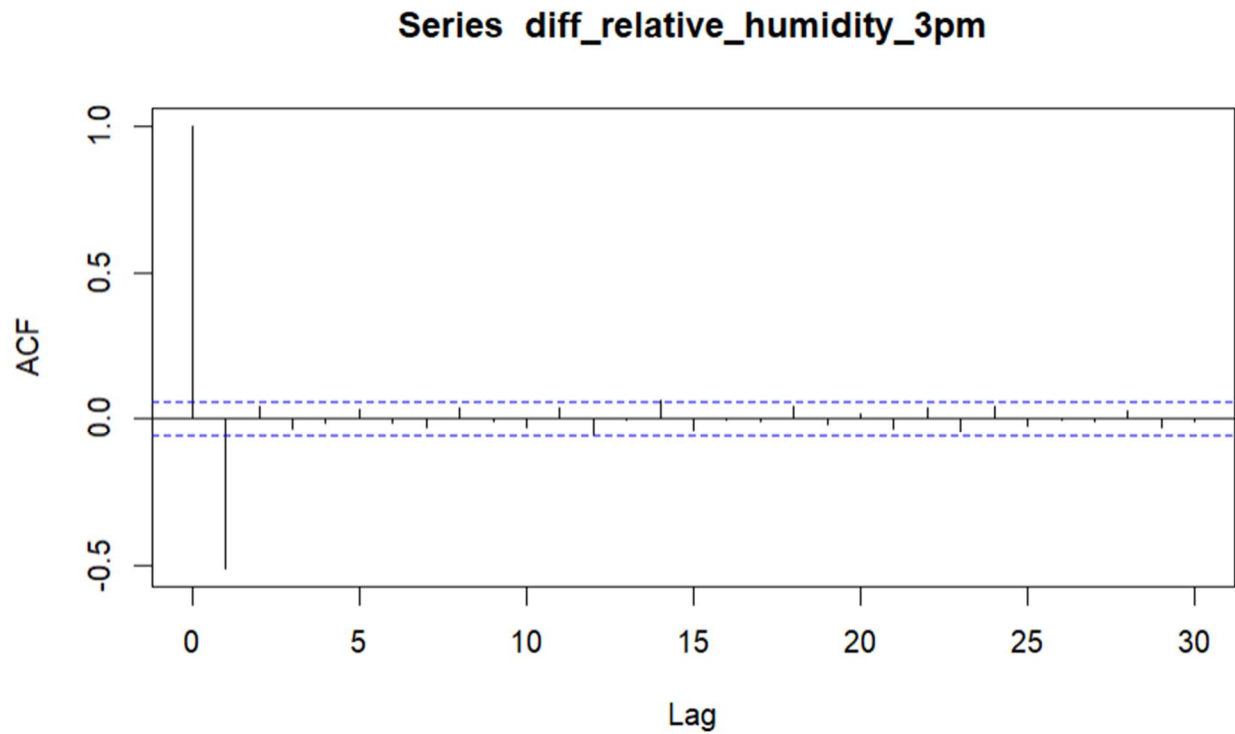
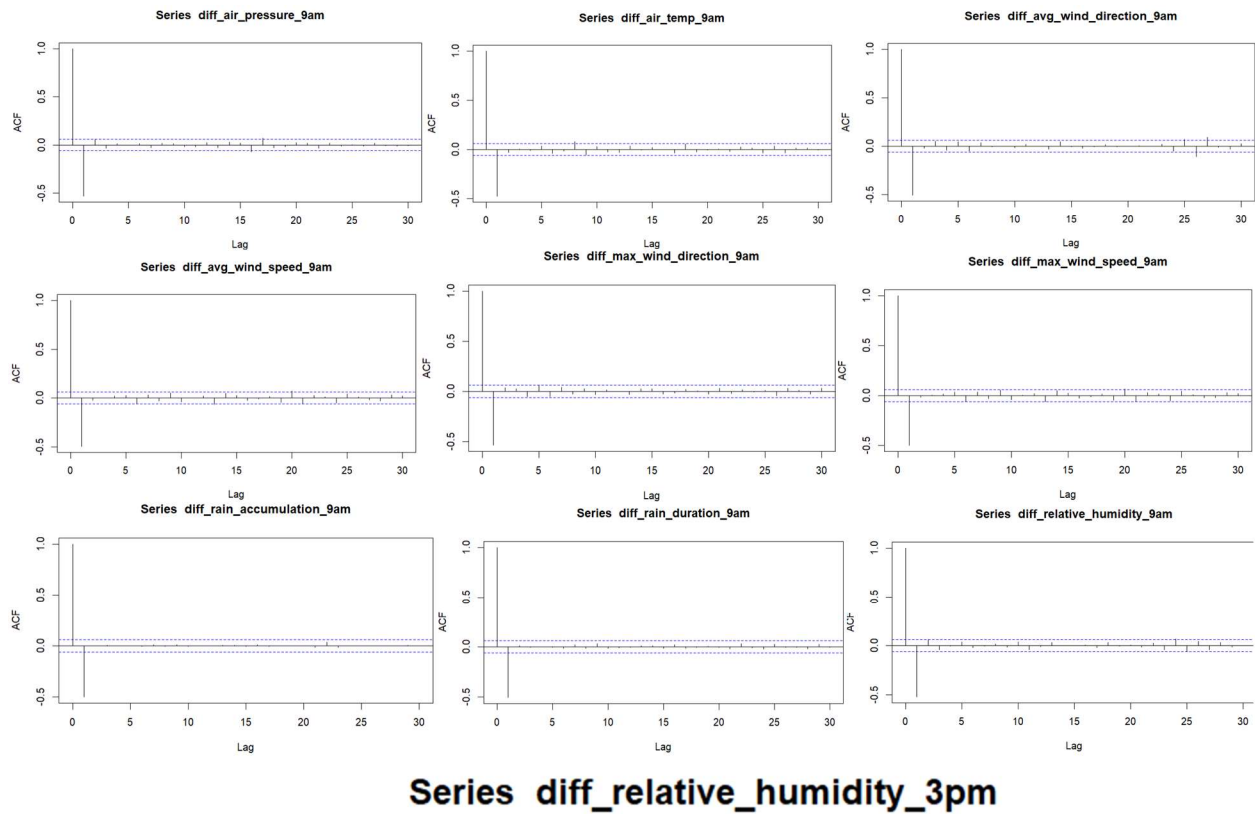
Therefore, for each variable tested, the null hypothesis of non-stationarity is rejected, and the alternative hypothesis of stationarity is accepted. In other words, the variables have achieved stationarity after differencing, making them suitable for further time series analysis and modeling.

<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_air_pressure_9am Dickey-Fuller = -16.148, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>	<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_max_wind_speed_9am Dickey-Fuller = -15.698, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>
<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_air_temp_9am Dickey-Fuller = -15.332, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>	<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_rain_accumulation_9am Dickey-Fuller = -16.832, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>
<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_avg_wind_direction_9am Dickey-Fuller = -17.135, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>	<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_rain_duration_9am Dickey-Fuller = -16.862, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>
<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_avg_wind_speed_9am Dickey-Fuller = -15.752, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>	<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_relative_humidity_9am Dickey-Fuller = -16.118, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>
<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_max_wind_direction_9am Dickey-Fuller = -16.261, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>	<p>Warning: p-value smaller than printed p-value Augmented Dickey-Fuller Test</p> <p>data: diff_relative_humidity_3pm Dickey-Fuller = -15.81, Lag order = 10, p-value = 0.01 alternative hypothesis: stationary</p>

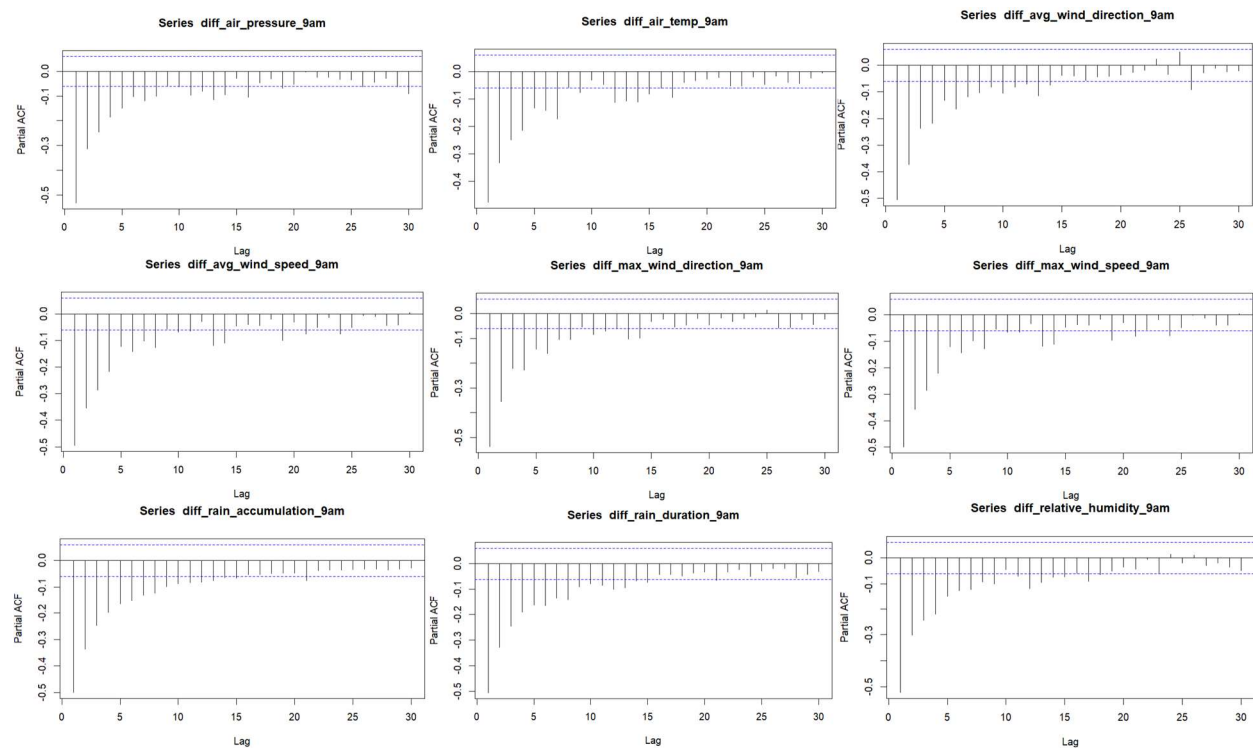
4.3.4 Autocorrelation Function (ACF) Plots: Autocorrelation Function (ACF) plots are utilized to visualize the autocorrelation structure of the differenced time series. ACF plots display the correlation between each observation and its lagged values at different time lags. These plots help identify the presence of significant autocorrelation at specific lags, which can inform the selection of appropriate lag values for time series models.

The ACF plots no longer exhibit significant spikes at lags beyond the first few observations. This is a significant improvement compared to the ACF plots before differencing, which likely indicated non-stationarity. In a stationary time series, the ACF should ideally decay to zero relatively quickly, indicating that the current value is not significantly dependent on past values at

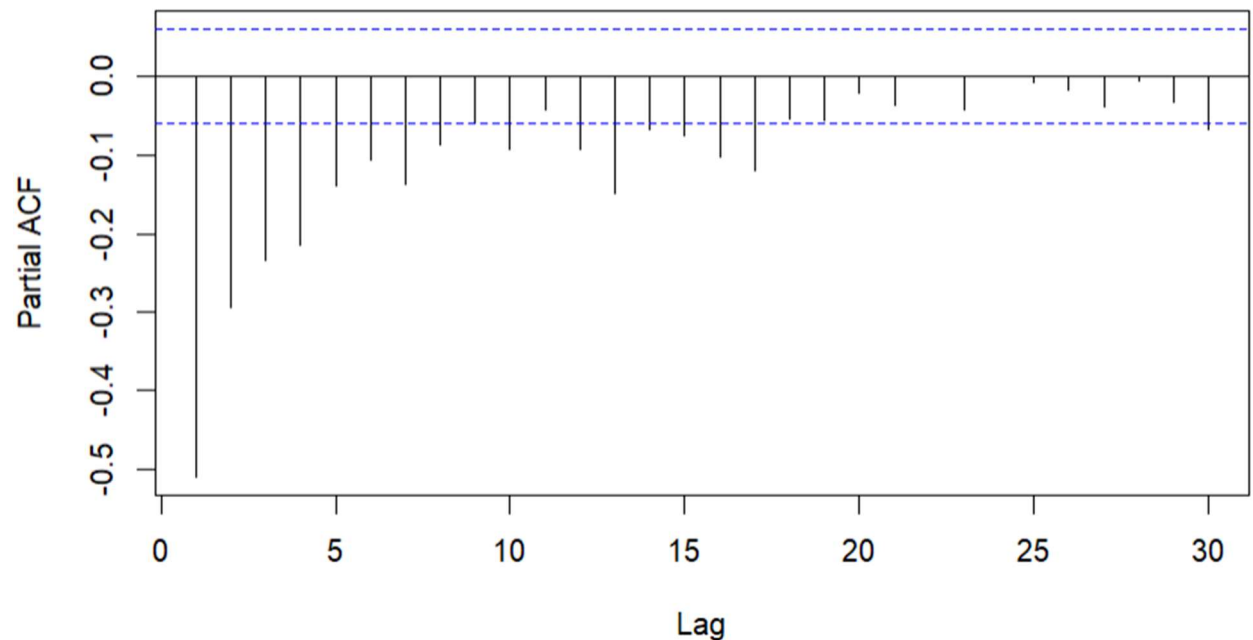
those lags. Overall, the ACF plot of the differenced relative humidity at 3 pm data suggests that differencing was an effective approach to achieve stationarity.



4.3.5 Partial Autocorrelation Function (PACF) Plots: Partial Autocorrelation Function (PACF) plots complement ACF plots by showing the correlation between observations at different lags after removing the effects of intervening observations. PACF plots help identify the direct relationships between observations at specific lags, aiding in the determination of the order of autoregressive terms in time series models.



Series diff_relative_humidity_3pm

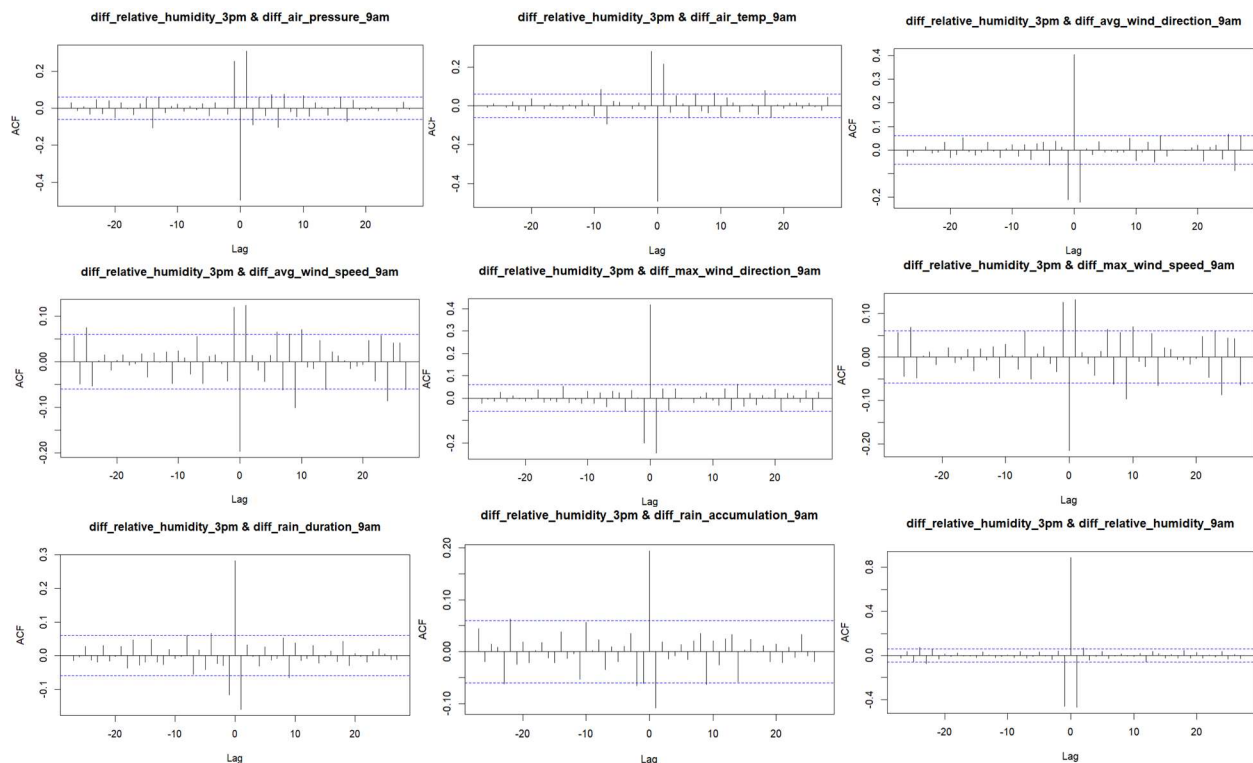


The PACF plot for the differenced relative humidity at 3 pm time series indicates:

1. Rapid decrease in PACF values after the first few lags, suggesting diminishing autocorrelation as lag increases.
2. Significant partial autocorrelation at lag 1, implying a correlation between the current value and its value one period ago.
3. Insignificant PACF values at higher lags, indicating no significant autocorrelation beyond lag 1.
4. Overall pattern resembling an autoregressive process of order 1 (AR(1)), where the current value primarily depends on the previous value.

This information guides the selection of an appropriate ARIMA model for forecasting, potentially incorporating an AR term of order 1 to capture the identified autocorrelation pattern.

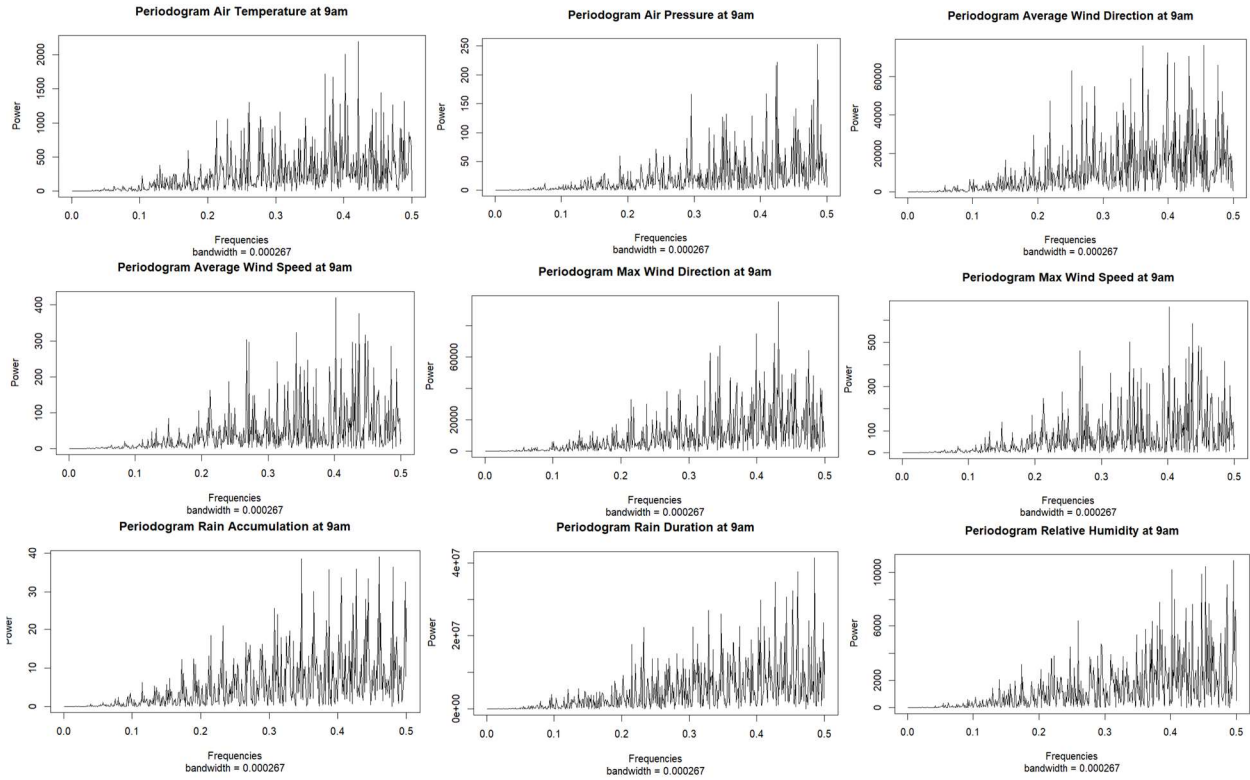
4.3.5 Cross-Correlation Function (CCF) Plots: Cross-Correlation Function (CCF) plots examine the relationship between two differenced time series by plotting their cross-correlation coefficients at different lags. These plots are particularly useful for analyzing the relationship between variables and identifying potential lead-lag relationships or dependencies between them.



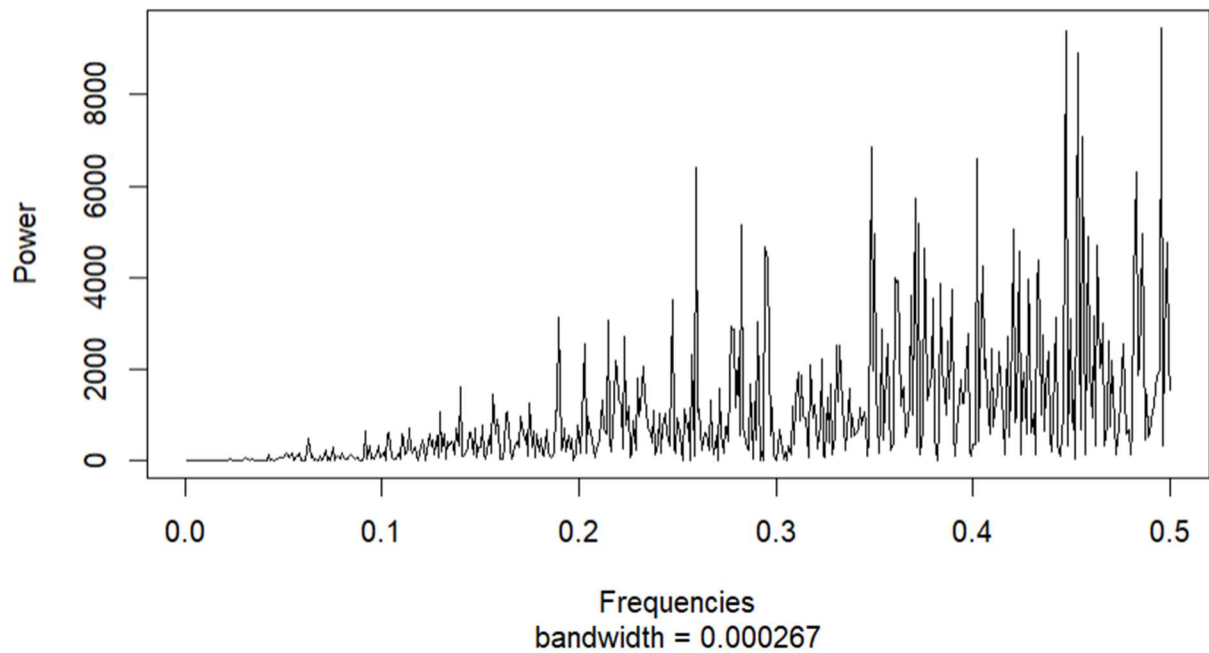
- CCF Plot of diff_relative_humidity_3pm & diff_air_pressure_9am: This cross-correlation function plot shows no significant correlation between the difference in relative humidity at 3 pm and the difference in air pressure at 9 am for any time lag between -20 and 20 days. The CCF values remain within the approximate 95% confidence interval, indicating no statistically significant relationship between the two variables at any lag.

- CCF Plot of `diff_relative_humidity_3pm` & `diff_air_temp_9am`: The CCF values are mostly insignificant and within the confidence bounds (dashed lines), indicating no strong cross-correlation between changes in relative humidity at 3 pm and changes in air temperature at 9 am, regardless of the lag. This suggests that morning temperature changes may not be a good predictor of afternoon humidity changes.
- CCF Plot of `diff_relative_humidity_3pm` & `diff_avg_wind_speed_9am`: The CCF shows a significant positive peak at lag 0, indicating that increases (decreases) in average wind speed at 9 am tend to coincide with increases (decreases) in relative humidity at 3 pm on the same day. However, there are no significant correlations at other lags, suggesting that the relationship is limited to contemporaneous changes.
- CCF Plot of `diff_relative_humidity_3pm` & `diff_max_wind_direction_9am`: This CCF plot shows the correlation between the relative humidity at 3 pm and the maximum wind direction at 9 am. The plot suggests no significant correlation between these two variables at any lag value.
- CCF Plot of `diff_relative_humidity_3pm` & `diff_max_wind_speed_9am`: This CCF plot shows the correlation between the relative humidity at 3 pm and the maximum wind speed at 9 am. The plot suggests a weak negative correlation at lag 0, indicating that higher maximum wind speeds at 9 am are associated with slightly lower relative humidity at 3 pm on the same day.
- CCF Plot of `diff_relative_humidity_3pm` & `diff_rain_duration_9am`: The CCF exhibits a significant negative correlation at lag 0, implying that longer rain durations at 9 am are associated with decreases in relative humidity at 3 pm on the same day. This counterintuitive relationship may be due to other confounding factors or non-linear effects. There are no significant correlations at other lags.
- CCF Plot of `diff_relative_humidity_3pm` & `diff_rain_accumulation_9am`: This CCF plot shows the correlation between the relative humidity at 3 pm and the rain accumulation at 9 am. Similar to the previous plot, there is no significant correlation observed between these variables.
- CCF Plot of `diff_relative_humidity_3pm` & `diff_relative_humidity_9am`: The plot suggests a strong positive correlation at lag 0, indicating that high (low) relative humidity at 9 am is likely to be associated with high (low) relative humidity at 3 pm on the same day.

4.3.6 Spectrum for Periodograms: Periodograms, visualized through spectrum plots, provide insights into the frequency domain characteristics of the differenced time series. Spectrum plots display the power spectrum of the data, revealing dominant frequencies present in the series. Analyzing periodograms helps identify periodic patterns or cycles in the data, which can inform the selection of appropriate frequency components in time series models.



Periodogram Relative Humidity at 3pm



4.4 ARIMA and ARIMAX Modeling:

4.4.1 Creation of Lagged Variables: In this subsection, lagged variables are generated based on insights obtained from the autocorrelation analysis. The code snippet provided below illustrates the process of creating lagged inputs and outputs for various meteorological parameters. These variables are essential for capturing temporal dependencies and are instrumental in forecasting models such as ARIMA.

4.4.2 Simple ARIMA Models: We started our analysis by exploring various simple ARIMA models with different orders of autoregressive (AR) and moving average (MA) components. We tested combinations of (p, q) within the range of 1 to 5 and fitted ARIMA models to the differenced relative humidity data at 3 pm. The purpose was to identify the best-fitting model based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

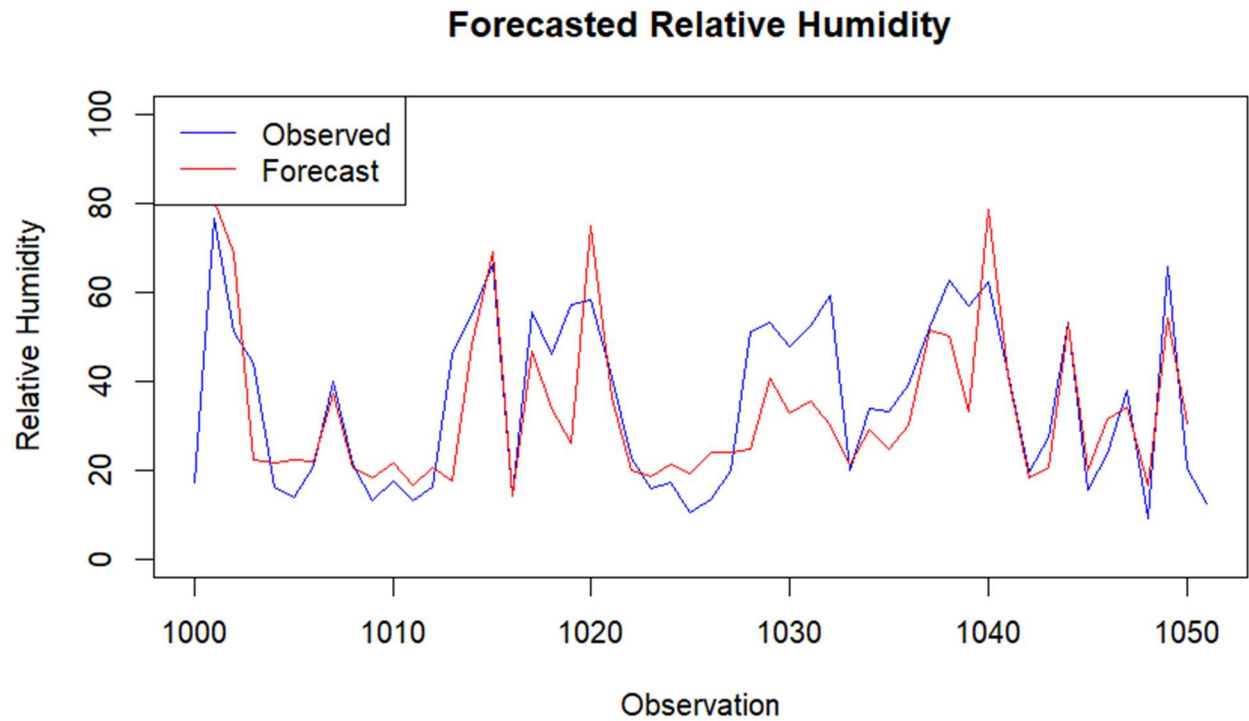
```
AIC: 9074.406 BIC: 9094.037 p: 1 q: 1
AIC: 9076.126 BIC: 9100.665 p: 1 q: 2
AIC: 9077.978 BIC: 9107.425 p: 1 q: 3
AIC: 9079.405 BIC: 9113.759 p: 1 q: 4
AIC: 9079.756 BIC: 9119.018 p: 1 q: 5
AIC: 9075.475 BIC: 9100.014 p: 2 q: 1
AIC: 9077.663 BIC: 9107.109 p: 2 q: 2
AIC: 9073.938 BIC: 9108.293 p: 2 q: 3
AIC: 9068.856 BIC: 9108.118 p: 2 q: 4
AIC: 9069.232 BIC: 9113.402 p: 2 q: 5
AIC: 9076.505 BIC: 9105.952 p: 3 q: 1
AIC: 9078.539 BIC: 9112.893 p: 3 q: 2
AIC: 9074.797 BIC: 9114.059 p: 3 q: 3
AIC: 9080.866 BIC: 9125.036 p: 3 q: 4
AIC: 9080.726 BIC: 9129.803 p: 3 q: 5
AIC: 9078.304 BIC: 9112.658 p: 4 q: 1
AIC: 9079.957 BIC: 9119.219 p: 4 q: 2
AIC: 9076.99 BIC: 9121.159 p: 4 q: 3
AIC: 9078.994 BIC: 9128.071 p: 4 q: 4
AIC: 9078.601 BIC: 9132.587 p: 4 q: 5
AIC: 9079.83 BIC: 9119.092 p: 5 q: 1
AIC: 9081.592 BIC: 9125.761 p: 5 q: 2
AIC: 9078.971 BIC: 9128.049 p: 5 q: 3
AIC: 9080.99 BIC: 9134.975 p: 5 q: 4
AIC: 9081.844 BIC: 9140.737 p: 5 q: 5
```

4.4.3 ARIMA Models with Exogenous Variables: In addition to simple ARIMA models, we considered ARIMA models with exogenous variables. Two models were fitted: one with lagged inputs (**data_arima1**) and another with only the differenced relative humidity at 3 pm. The performance of these models was evaluated based on their AIC and BIC values.

Order (p, q): 1 , 1	AIC: 7540.58 BIC: 7565.119
Order (p, q): 1 , 2	AIC: 7538.819 BIC: 7568.265
Order (p, q): 1 , 3	AIC: 7540.743 BIC: 7575.097
Order (p, q): 1 , 4	AIC: 7542.324 BIC: 7581.586
Order (p, q): 1 , 5	AIC: 7544.323 BIC: 7588.493
Order (p, q): 2 , 1	AIC: 7538.693 BIC: 7568.139
Order (p, q): 2 , 2	AIC: 7540.688 BIC: 7575.043
Order (p, q): 2 , 3	AIC: 7542.816 BIC: 7582.078
Order (p, q): 2 , 4	AIC: 7544.353 BIC: 7588.523
Order (p, q): 2 , 5	AIC: 7545.47 BIC: 7594.547
Order (p, q): 3 , 1	AIC: 7540.688 BIC: 7575.042
Order (p, q): 3 , 2	AIC: 7542.652 BIC: 7581.914
Order (p, q): 3 , 3	AIC: 7544.452 BIC: 7588.622
Order (p, q): 3 , 4	AIC: 7546.506 BIC: 7595.583
Order (p, q): 3 , 5	AIC: 7542.65 BIC: 7596.635
Order (p, q): 4 , 1	AIC: 7545.723 BIC: 7584.985
Order (p, q): 4 , 2	AIC: 7544.233 BIC: 7588.403
Order (p, q): 4 , 3	AIC: 7540.643 BIC: 7589.721
Order (p, q): 4 , 4	AIC: 7546.476 BIC: 7600.461
Order (p, q): 4 , 5	AIC: 7531.617 BIC: 7590.51
Order (p, q): 5 , 1	AIC: 7547.714 BIC: 7591.884
Order (p, q): 5 , 2	AIC: 7549.715 BIC: 7598.792
Order (p, q): 5 , 3	AIC: 7542.624 BIC: 7596.61
Order (p, q): 5 , 4	AIC: 7531.978 BIC: 7590.872
Order (p, q): 5 , 5	AIC: 7528.755 BIC: 7592.556

4.5 Forecasting and Evaluation:

4.5.1 Forecasting Relative Humidity: We proceeded to forecast relative humidity at 3 pm using the ARIMA model with the best performance. This model incorporated exogenous variables, specifically relative humidity at 9 am. The forecast was generated for observations 1001 to 1050. The forecasted values were compared against the observed data through visualization.



4.5.2 Model Residuals Analysis: To ensure the adequacy of the ARIMA model, we analyzed its residuals. This involved plotting the residuals, conducting the Box-Ljung test for residual autocorrelation, and examining the autocorrelation function (ACF) and spectral analysis of the residuals. These diagnostic checks helped verify whether the model adequately captured the underlying patterns in the data.

Box-Ljung test

```
data: model_residuals  
X-squared = 0.00066847, df = 1, p-value = 0.9794
```

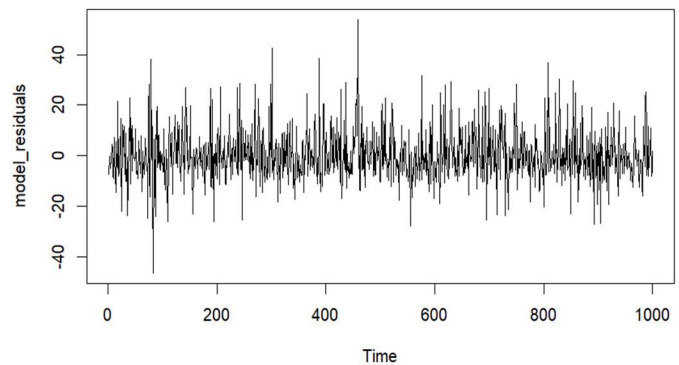
The residual plot demonstrates random scatter around zero, indicating unbiased predictions with no systematic errors.

While the mean of the residuals aligns closely with zero, additional testing for constant variance (homoscedasticity) would further validate the model's performance and ensure consistency in residual variance.

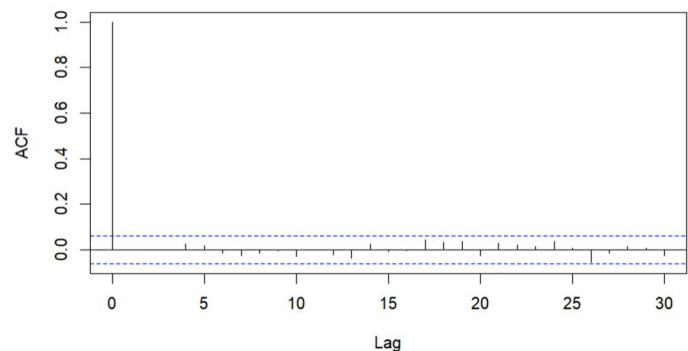
Examining the ACF of the residuals aids in identifying any residual patterns unaccounted for by the model. The absence of notable spikes in the ACF plot suggests that the model adequately captures all temporal dependencies present in the data.

The absence of notable peaks in the periodogram indicates that there are no discernible patterns or trends present in the residual time series. This suggests that the residuals follow a random walk pattern.

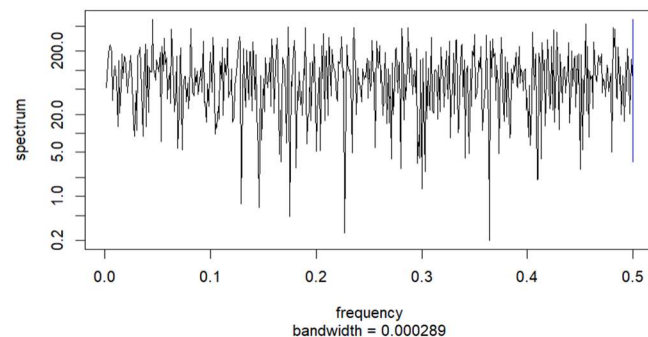
Model Residuals



ACF of Residuals



Spectral Analysis of Residuals



5. INTERPRETATION OF RESULTS

Finally, we evaluated the accuracy of the forecasts using various metrics. This step allowed us to quantify the performance of the ARIMA model in predicting relative humidity at 3 pm. This comprehensive analysis enabled us to understand the performance of different ARIMA models, select the most appropriate one for forecasting, and evaluate the accuracy of the forecasts.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.01256763	10.23568	7.572173	-12.32641	27.80614	0.3050435	-0.0008163751

The Box-Ljung test for residual autocorrelation yields a p-value of 0.9794, indicating that there is no significant evidence of autocorrelation in the residuals. This suggests that the model adequately captures the temporal dependencies in the data. Additionally, the accuracy metrics for the training set indicate a mean error (ME) of 0.0126, a root mean squared error (RMSE) of 10.24, and a mean absolute error (MAE) of 7.57. These metrics provide insight into the overall performance and predictive accuracy of the model.

5. CONCLUSION

In conclusion, our study focused on leveraging ARIMA and ARIMAX models to forecast relative humidity at 3 PM, considering various meteorological variables. Through thorough data preprocessing, time series analysis, and model fitting, we successfully developed and evaluated several ARIMA models. Our analysis revealed the importance of incorporating lagged variables, particularly morning relative humidity, for accurate forecasting. The diagnostic checks on model residuals confirmed the adequacy of our chosen model in capturing underlying data patterns. Moreover, the accuracy evaluation metrics demonstrated the model's capability in providing reliable forecasts. Overall, our study underscores the significance of advanced modeling techniques in improving weather forecasting accuracy, thereby facilitating informed decision-making in various sectors reliant on weather information.

6. LIMITATIONS AND FUTURE SCOPE

6.1 Limitations:

1. Data Limitations: The study relies on a specific dataset, which may not capture all the relevant meteorological variables or cover a sufficiently diverse range of weather conditions. The performance of the models may vary when applied to data from different geographical regions or time periods.
2. Stationarity Assumptions: The analysis assumes that the time series data can be made stationary through differencing. However, in cases where the underlying processes exhibit non-linear or complex trends, the stationarity assumption may not hold, potentially affecting the model's accuracy.

3. Exogenous Variable Selections: The selection of exogenous variables in the ARIMAX models was guided by correlation analysis and domain knowledge. However, there may be other relevant variables or interaction effects that were not considered, potentially limiting the model's predictive power.

4. Model Complexity: While the ARIMA and ARIMAX models capture linear dependencies, they may not adequately represent non-linear or complex relationships present in the data, leading to potential biases or inaccuracies in the forecasts.

6.2 Future Work:

1. Incorporation of Additional Data Sources: Future studies could explore integrating additional data sources, such as satellite imagery, radar data, or numerical weather prediction models, to enhance the predictive power of the forecasting models.

2. Non-linear and Machine Learning Models: Investigating the application of non-linear time series models or advanced machine learning techniques, such as artificial neural networks or support vector machines, could potentially capture more complex patterns and improve forecast accuracy.

3. Real-time Forecasting and Updates: Developing a real-time forecasting system that can continuously update and refine the models as new data becomes available could enhance the practical utility of the research for weather forecasting applications.

By addressing these limitations and exploring the proposed future research directions, our understanding of relative humidity forecasting could be further advanced, leading to more accurate and reliable weather predictions for various sectors and applications.

7. REFERENCES

- [1] https://jbusemey.pages.iu.edu/time/time_series.htm
- [2] Cowpertwait, P. S. P. & Metcalfe, A V. (2009) Introductory Time Series with R.
- [3] Springer. Shumway, Robert H., & Stoffer, David S. (2017) Time series analysis and its applications. Springer.
- [4] <https://www.kaggle.com/datasets/apratik46/daily-weather-dataset>
- [5] <https://github.com/PacktPublishing/Time-Series-Analysis-with-Python-Cookbook/tree/main/datasets>