ADVANCED DATA ANALYSIS AND MACHINE LEARNING

CLIMATE A2 - INVESTIGATE THE FORECASTING POWER FOR PREDICTING HUMIDITY DATA USING MULTIVARIATE DATA

LUT University

2025

Group: Amin Hassanzadehmoghaddam, Antonio Oliva, Nico Niemelä

https://github.com/AntoniooOliva/Climate-A2

Content

1 Visualization of the data and initial exploratory analysis

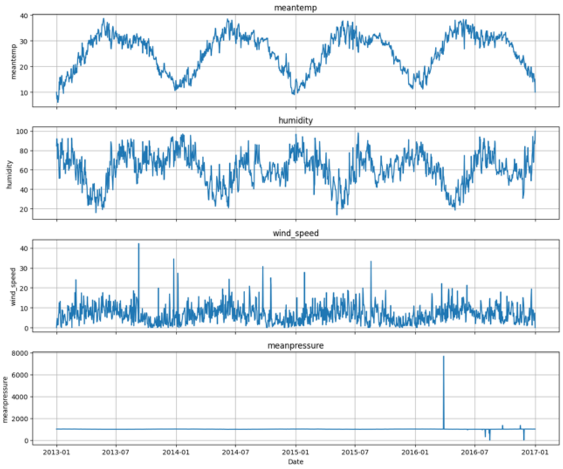

Figure 1. Data visualization

The dataset contains 1,462 daily observations of weather parameters including mean temperature, humidity, wind speed, and mean pressure from beginning of 2013 to the beginning of 2017.

- Mean Temperature:

  The average temperature is around 25.5, with a minimum of 6 and a maximum of 38.7. The standard deviation (7.35) indicates moderate variability and clear seasonal fluctuations typical of annual temperature cycles. The cycles peak in summer and dipping in winter consistently each year.

- Humidity:

  Humidity averages 60.8 and its range is from 13.4 to 100. This large range suggests both dry and humid periods throughout the year. The plot shows that humidity follows an inverse seasonal pattern, peaking during cooler or monsoon periods.

- Wind Speed:

  The average wind speed is 6.8, with a widespread up to 42.2. This indicates occasional strong wind events or storms, and most days remain relatively calm. This parameter is highly variable but doesn’t show strong seasonality.

- Mean Pressure:

  The average pressure is 1011, but the unusually large standard deviation (180.2) and maximum value (7679) indicate the presence of outliers.

There are no missing values across any columns, ensuring data completeness for analysis.

2 Time-series decomposition analysis of long-term trend, seasonality, and residuals

For doing the decomposition analysis, because the data consists of multiple years range, 365 days are considered as period. The trend component starts around 68 humidity in early 2013 and declines to about 58 by end of 2016, indicating a steady long-term decrease of about 15% in average humidity over four years.

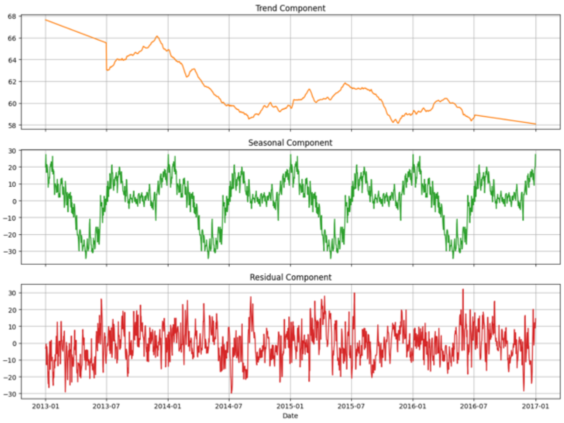

Figure 2. Long term decomposition on Humidity variable

The seasonal component oscillates between approximately 25 and –30, showing a clear annual cycle. Humidity rises during the monsoon season (June–September) and drops during dry winter months (November–February).

The residual component fluctuates within ±30 and reflects short-term daily variations due to weather anomalies, local storms, or measurement noise.

3 An autocorrelation analysis of the dataset

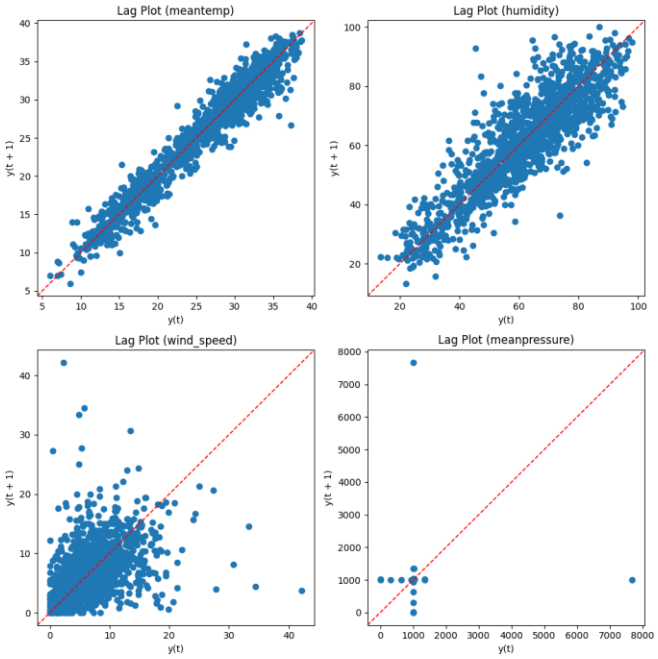

Figure 3. Lag plots for each feature

Mean temperature

	t	t+1	t+2	t+3
t	1	0.974	0.956	0.942
t+1	0.974	1	0.974	0.956
t+2	0.956	0.974	1	0.974
t+3	0.942	0.956	0.974	1

Humidity

	t	t+1	t+2	t+3
t	1	0.878	0.783	0.718
t+1	0.878	1	0.878	0.783
t+2	0.783	0.878	1	0.878
t+3	0.718	0.783	0.878	1

Wind Speed

	t	t+1	t+2	t+3
t	1	0.436	0.223	0.171
t+1	0.436	1	0.436	0.223
t+2	0.223	0.436	1	0.436
t+3	0.171	0.223	0.436	1

Mean pressure

	t	t+1	t+2	t+3
t	1	0.003	0.011	0.002
t+1	0.003	1	0.003	0.011
t+2	0.011	0.003	1	0.003
t+3	0.002	0.011	0.003	1

Table 1. Correlation matrixes for each feature

As the lag plots in Figure 3 and the correlation matrixes in Table 1 show, there is a strong autocorrelation between the values y(t) and y(t+1) in mean temperature (~ 0.974), and humidity (~ 0.878), with t representing days.

For wind speed, as expected, the correlation is not so strong, meaning the measurements of a particular day do not influence much the measurement of the following day (~ 0.437).

The mean pressure behaves differently. The lag plot reveals a strong outlier (7679) that makes the correlation difficult to analyse from the plot. There are also outliers closer to the normal behaviour of the data, and all outliers in the datasets cause a near 0 correlation between the measurements.

Just to obtain a more meaningful analysis, outliers were removed and the calculations were repeated.

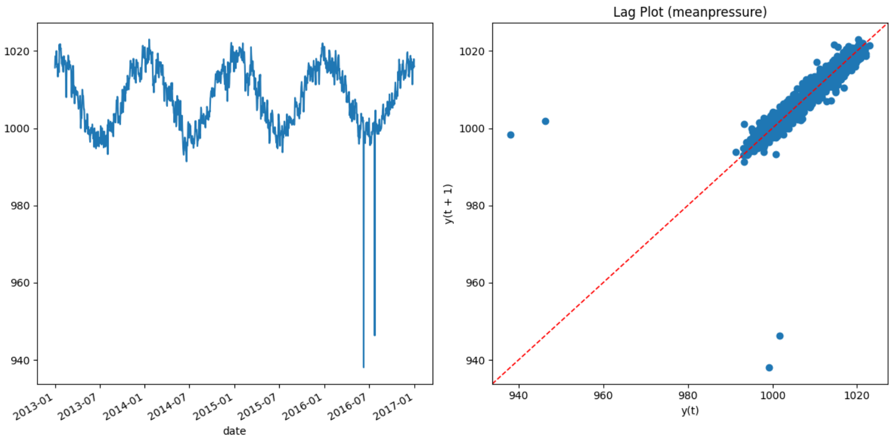

Figure 4. New lag plot for "mean pressure"

Mean pressure				
	t	*t+1*	*t+2*	*t+3*
t	*1*	*0.902*	*0.873*	*0.859*
t+1	*0.902*	*1*	*0.902*	*0.873*
t+2	*0.873*	*0.902*	*1*	*0.902*
t+3	*0.859*	*0.873*	*0.902*	*1*

Table 2. New correlation matrix for "mean pressure"

After removing most outliers, the analysis is clearer: in the mean pressure feature column, there is a strong correlation between the values (~ 0.902), behaving like a sinusoidal wave, within the range 995-1200. It is possible to conclude that these strong outliers can complicate further analyses, and a proper way to handle these values must be found.

4 Plan for partitioning the time-series data for model formation

To prevent data leakage and preserve temporal integrity, the time-series data must be partitioned so that the training data occurs before the test data in time. One method that can be used is regular Day Forward Chaining where the data is split into many different training, validation and test sets. (Chess, S. 2020) When using this method, we successively consider each day of the data set as the test set and assign all previous data into the training set. Concept of the plan is shown in table 3.

Fold	Training period	Validation period	Testing period
1	Day 1	Day 2	Day 3
2	Day 1 ➔ Day 2	Day 3	Day 4
3	Day 1 ➔ Day 3	Day 4	Day 5
4	Day 1 ➔ Day 4	Day 5	Day 6
m	Day 1 ➔ Day m	Day m+1	Day m+2

Table 3. Conceptualization of the data partitioning plan.

It should be noted that the original data partitioning plan might need to be revisited if it is computationally too expensive.

5 References

Chess, S. (2020) 'Cross-validation in time series', Medium, 6 October. Available at: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4 (Accessed: 8 November 2025).