Utilization of Multiple Regression for Climate Predictions in Delhi, India (2013-2017)

Mitchell Gelband

February 25, 2025

Introduction

The analysis was performed on the Delhi data set from Kaggle, aiming to make future climate predictions. The data set consists of two CSV files—one for training and one for testing, the data contains information such as mean temperature (°C), mean pressure (mbar), humidity (g/m^3), and wind speed (km/h). In this study, time series data was analyzed using Fourier Transforms and correlative analysis which confirmed annual and semi-annual periodicity and multivariate relationships. Machine Learning was employed utilizing Multiple Linear regression, which predicted humidity, pressure, temperature, and wind speed with varying degrees of accuracy.

Methodology

Data Importation/Cleaning

The data was imported via the Pandas library in python. The data was cleaned using a Simple Moving Average (SMA) algorithm only for visualization (not for training). For some variables, outliers larger than 10 standard deviations from the mean were classified/assumed due to measurement error and set to the mean.

Fourier Analysis

The Fourier transform is a powerful analytic tool that transforms domain data into frequency domain data to determine the modes in which a signal is comprised of.

$$X(k) = \sum_{n=0}^{N-1} x_n \cdot e^{-i\frac{2\pi kn}{N}} \tag{1}$$

$$f_k = \frac{k}{\Delta t} \tag{2}$$

Multiple Regression

MLMs were used in this study to make predictions on future climate data. The method used was MR, the below equation is the general form.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \epsilon \tag{3}$$

Y is the predicted value for the output, while X_i are the variable inputs that go into predicting that output. β_i are the model coefficients, which can be estimated via the Residual Sum of Squares (RSS).

Results

Data Analysis

Applying the SMA yielded a much clearer relationship via a time-series of the variables. Figure 1 visually illustrates the quantitative relationship between the variables.

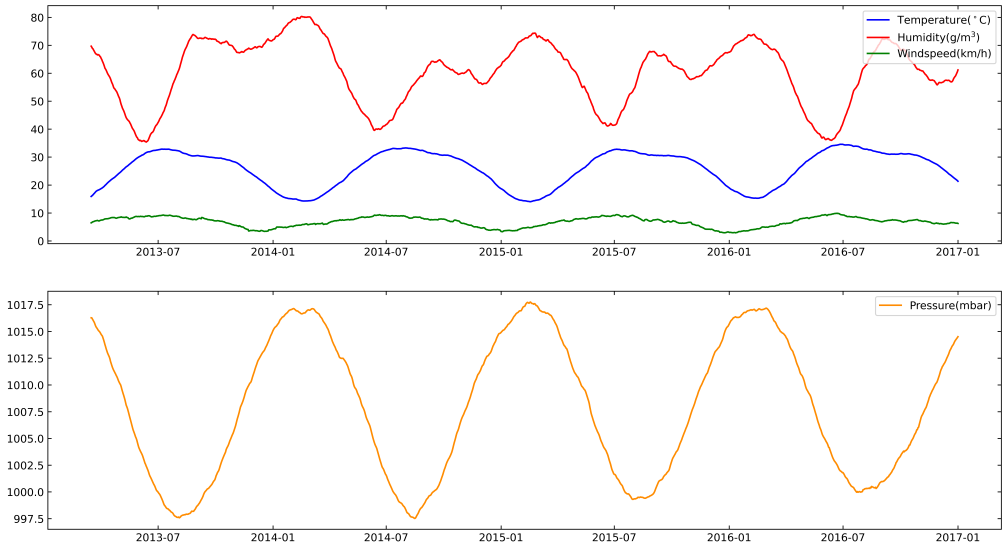

Figure 1: *Top*: Time-series data of Temperature((°C) Humidity(g/m^3), and Wind speed (km/h) *Bottom*: Time-series data of Pressure(mbar)

Here we see the variables seems to follow a periodic nature, from first glance it seems that Temperature and Wind speed are inversely correlated with Humidity and Pressure. A correlation matrix measures this:

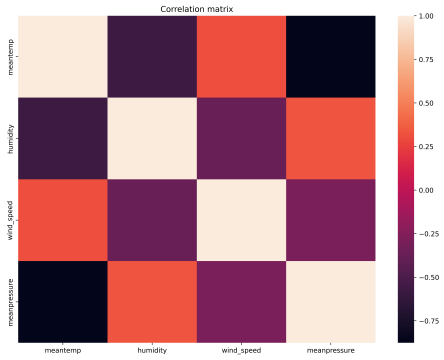

Figure 2: Correlation matrix

The correlation matrix supports this observation to a significant extent. Temperature and wind speed exhibit a positive correlation of 0.31, while humidity and pressure show a correlation of 0.33. In contrast, when comparing these variables across categories, negative correlations emerge where values such as -0.57 between temperature and humidity, and as high as -0.87 between temperature and pressure are observed.

After analyzing the relationship between temperature, pressure, wind speed, and humidity the periodicity of these variables were investigated. By apply a discrete Fourier transform the frequency domain revealed annual and semi-annual periodicity. The below figure illustrates this.

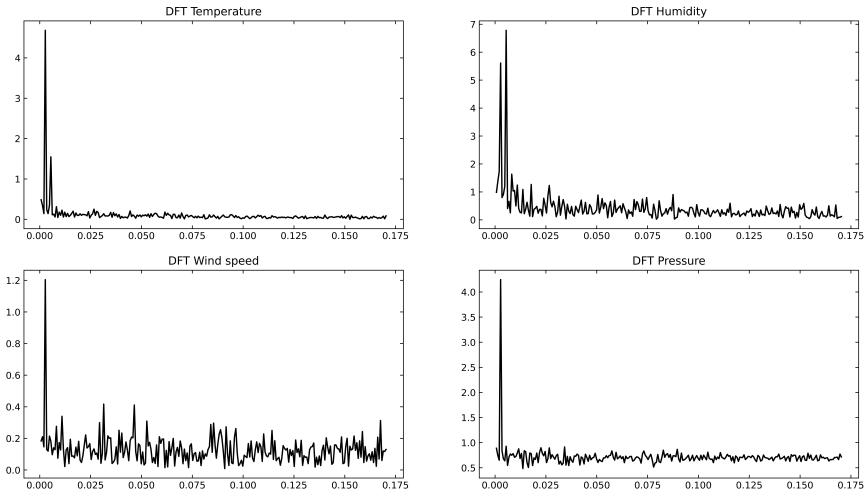

Figure 3: *Top Left*: Fourier Transform of the temperate time series. *Top Right*: Fourier Transform of the humidity time series. *Bottom Left*:Fourier Transform of the Wind speed time series. *Bottom Right*: Fourier Transform of the pressure time series.

Figure 3 shows the frequency domain with respect to the amplitude, the spikes represent the sinusoidal functions which have the largest amplitude and at what frequency. A prominent frequency on all cases is $T \approx 365.5$, while the Fourier transforms of Temperature and Humidity reveal a another spike (relative to noise), at $T \approx 182.75$, indicating annual periodicity across all variables and semi annual periodicity across humidity and temperature.

Model Predictions & Error

Utilizing multiple regression, from the sklearn library in python, the model made predictions on temperature, wind speed, pressure, and humidity. The below table provides information on the coefficients of the model, error, variable, F-statistic, and $F_{critical}$ value for each variable the model was tested on.

Model coefficients & accuracy

Category	Variable	β_i	RSME	R^2	F	F_{critical}
Temperature	Intercept	2.50				
	X_1	-0.14	3.00	0.79	126.32	2.68
	X_2	-0.07				
	X_3	-0.77				
Humidity	Intercept	1849.33				
	X_1	-2.64	14.47	0.42	26.48	2.68
	X_2	-0.88				
	X_3	-1.70				
Pressure	Intercept	1042.65				
	X_1	-1.01	3.29	0.67	74.12	2.68
	X_2	-0.12				
	X_3	-0.14				
Wind Speed	Intercept	1042.65				
	X_1	-1.02	3.81	-0.14	4.46	2.68
	X_2	-0.12				
	X_3	-0.14				

Table 1: Model coefficients, accuracy, and F-statistics

Temperature and pressure have relatively high R^2 and RMSE values. Humidity scores lower in accuracy obtaining an R^2 value of 0.42, along with an RSME of 14.47. Lastly, wind speed has a relatively low RSME along with a bad R^2 value, not even predicting the mean of the data. F statistic tests were conducted to further reject the null hypothesis. All models scored above $F_{critical}$ ($F > F_{critical}$). Temperature scored the highest F-value with a value of 126.32, giving a p-value well below the 0.05 mark. This is likely due the high correlation it had with pressure in the correlation matrix and it's high R^2 value. Wind

speed had the lowest F-value of 4.46 however, still well below the $p \leq 0.05$ value and likely had the lowest due to it's low correlation with any of the predictors coupled with it's low R^2 value. Given that these values are above the critical value, this model serves as a more accurate predictor than the mean model, or the null hypothesis (H_0).

Model Predictions

While the accuracy of the model results vary, each model whether it was predicted, temperature, pressure, wind speed or humidity generally trended in that of the actual data. Another interesting piece is that the relative errors seem to have some correlation for example, in wind speed and humidity the relative errors are very similar in temporal structure. The plot below depicts this:

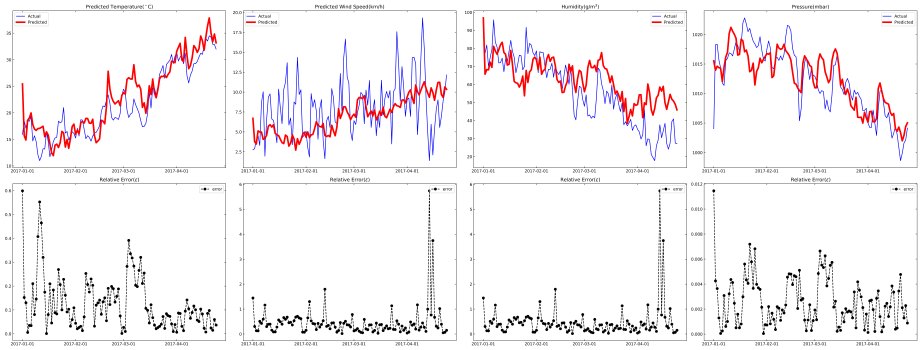

Figure 4: *Far Left*: Time-series data of predicted (red) and actual (blue) temperature (°C), along with its relative error below.
Left: Time-series data of predicted (red) and actual (blue) wind speed (km/h), along with its relative error below.
Right: Time-series data of predicted (red) and actual (blue) humidity (g/m^3), along with its relative error below.
Far Right: Time-series data of predicted (red) and actual (blue) pressure (mbar), along with its relative error below.

Discussion

This study found multiple analytic and predictive results. When calculating a correlation matrix between temperature, pressure, humidity, and wind speed, it was found that variables like temperature wind speed were weakly positive correlatives comparatively to pressure, humidity, and temperature that were found to be strongly negative correlatives. Applying Fourier analysis to each variable yielded annual periodicity across all variables and semi-annual periodicity prevalence was observed in humidity and temperature. After applying data analysis ML methods were utilized to make predictive results on all variables.

Multi-Linear regression was applied to these variables that resulted in differing accuracies depending on the variable predicted. Applying Multi-Linear regression to temperature produced the highest accuracy with an RSME of 3 and a R^2 value of 0.8. Wind speed proved to be on the other side of this spectrum, with an R^2 value of -0.14, not even predicting the mean. All variables scored above the $F_{critical}$ value of 2.68, allowing a reject of the null hypothesis (H_0). Graphically however, all variables followed the general trend of the data.

References

[1] International Statistical Literacy Project. (2025). *International Statistical Literacy Project*. Retrieved from `https://iase-web.org/islp/`

[2] Rao, S. V. (n.d.). *Daily Climate Time Series Data*. Kaggle. Retrieved from `https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data`