Proposal: PM2.5 Prediction with Machine Learning*

[*]For the fulfillment project proposal of AT82.01 Computer Programming for Data Science and Artificial Intelligence course by Dr. Chantri Polprasert

1st Kyi Thin Nu
Department of Data Sciences and Artificial Intelligence
Asian Institute of Technology
Pathum Thani, Thailand
st124087@ait.asia

2nd Thongtong Eamsaard
Department of Industrial System Engineering
Asian Institute of Technology
Pathum Thani, Thailand
st123300@ait.asia

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—Particulate Matters, PM2.5, Machine Learning**

I. Introduction

A. Introduction to PM2.5

PM2.5, or Particulate Matter 2.5, is a critical measure of air quality that refers to tiny airborne particles or droplets with a diameter of 2.5 micrometers or smaller. These minuscule particles can originate from a variety of sources, including industrial emissions, vehicle exhaust, construction activities, natural dust, and even chemical reactions in the atmosphere. PM2.5 is significant because it has a substantial impact on both human health and the environment. Understanding PM2.5 levels is essential for assessing air quality, making informed policy decisions, and implementing measures to safeguard public well-being.

B. Global Perspective View on PM2.5

A global perspective on PM2.5 levels is vital to comprehend the scale and variations in air quality across different regions. By monitoring PM2.5 on a global scale, we can identify trends, sources of pollution, and areas where air quality may be particularly hazardous. This global view often involves the use of satellites and international air quality monitoring networks. It helps nations collaborate in addressing transboundary air pollution and sharing information to mitigate the impact of airborne particles on a global scale.

C. Local Perspective View for Thailand

On a local level, such as within a country like Thailand, monitoring PM2.5 is crucial for assessing the immediate air quality conditions that people are exposed to. Local monitoring networks, government agencies, and environmental organizations collect data on PM2.5 levels to provide real-time information to citizens. This local perspective helps individuals make informed decisions about outdoor activities, and it assists policymakers in implementing measures to improve air quality and protect public health.

Fig. 1: PM 2.5 Impact on environment and humans

Fig. 2: PM 2.5 Impact in Thailand

D. Why Do We Do PM2.5 Projects?

1) Protecting Public Health: PM2.5 particles are so small that they can penetrate deep into the respiratory system,

posing significant health risks. PM2.5 projects aim to reduce exposure to these particles and thereby protect the health of communities. High PM2.5 levels have been linked to respiratory diseases, cardiovascular issues, and even premature death.

2) Environmental Impact: PM2.5 particles can also harm the environment. They can contribute to smog formation, damage ecosystems, and affect water quality. PM2.5 projects seek to reduce these environmental impacts.
3) Policy and Regulation: Monitoring PM2.5 is essential for setting air quality standards and regulations. Governments and regulatory agencies use PM2.5 data to implement measures to limit emissions from various sources.
4) Awareness and Education: PM2.5 projects help raise public awareness about air quality issues. They encourage people to take action to reduce their own contributions to air pollution and to advocate for cleaner air.

In summary, PM2.5 projects serve the vital purpose of safeguarding both human health and the environment. They provide essential information for informed decision-making, regulation, and action at both the local and global levels.

II. Problem Statement

To predict the PM2.5 values based on given weather conditions and trends of the PM2.5.

III. Related Works

This project goal is to train the model to able to predict the PM2.5. Here the focus is on how other concerned in PM2.5 at different area and weather, some well-known techniques to make the model be able to predict desired target variables. Thus, These can make the model predicted the PM2.5 with high accuracy, precision, and recall.

Vahid Mehdipour [4] and his team from Tehran compared different methods for modeling PM2.5 in the capital city of Iran, Tehran. They proposed decision trees (DT), Bayesian Network (BN), and support vector machine (SVM). Using the data for over three periods, they concluded that PM10, NO_2, SO_2, and O_3 are critical factors for PM2.5 with the best model is SVM.

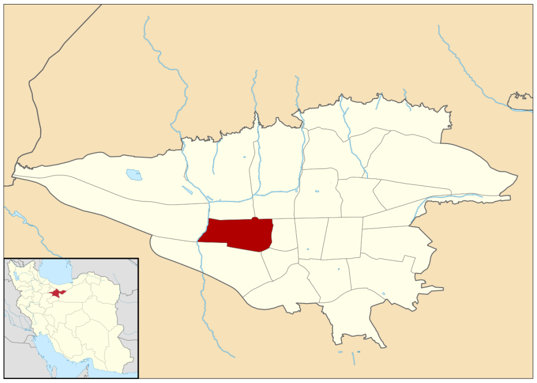


Fig. 3: Area of interest, Tehran, Iran [4]

Delhi, another mega-city in India, also faced an enormous of air pollution because of rapid development for a while. Nidhi Sharma and her colleagues [6] forecast pollution load in an atmosphere using time-series regression forecasting. In the results, predicted trends are shown after 2017.

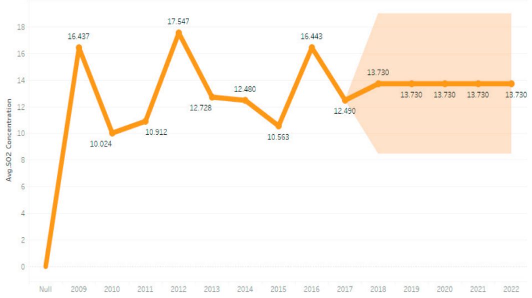


Fig. 4: trend of SO_2 in $\mu g/m^3$ [6]

Another interesting paper used Taiwan Air Quility Monitoring (TAQMN) data set. Doreswamy and his team [7] did the forecasting using also machine learning regression models. The data used are from 2012 to 2017. Models were evaluated by Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) as shown in eq. 1 - 4. They used Fourier arrangement and spline multinomial to fill the missing values in data set. The model they used are random forest regressor (RFR), gradient boosting regressor GBR), k neighbors regressor (KNR), MLP regressor (MLPR), and decision tree regressor CART. To select the best model, they used cross-validation and determined that gradient boosting regressor model is better in forecasting ait pollution in TAQMN data.

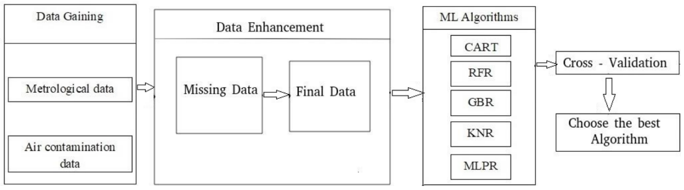


Fig. 5: The proposed prediction pipeline model for air pollution [7]

$$MAE = \frac{\sum_{i=1}^{m} |x_i - \hat{x_i}|}{m} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \hat{x_i})^2}{m}} \tag{2}$$

$$MSE = \frac{1}{m} \sum_{j=1}^{m} (x_i - \hat{x_i})^2 \tag{3}$$

$$R^2 = [\frac{1}{M} \frac{\sum_{j=1}^{M} [(Y_j - \overline{Y})(X_j - \overline{X})]}{\sigma_y \sigma_x}]^2 \tag{4}$$

where,

m and M are the number of observations
$\hat{x_i}$ is the predicted value
x_i is the actual value
σ_x is the standard deviation of the observation X
σ_y is the standard deviation of the observation Y
X_j is the observed values
$\overline{X}$ is the mean of the observed values
Y_j is the calculated values
$\overline{Y}$ is the mean of the calculated values.

IV. Dataset

A. Description

The data set contains a record of PM2.5 per hour recorded in 2019 from 5 stations distributed in Bangkok. In total, 5 stations' data set for this project are given by Dr. Chantri via the Pollution Control Department of Thailand. The station are numbered as follows (5 from all 66 stations established in 2019)

- Station 03: Bang Khun Thian, Bangkok
- Station 50: Pathum Wan, Bangkok
- Station 52: Thonburi, Bangkok
- Station 53: Chok Chai, Bangkok
- Station 54: Din Daeng, Bangkok

The data set is in EXCEL spread sheet format.

B. Features

The dataset contains following features:

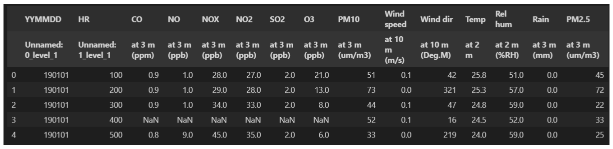
- Date and time of record
- Various air quality parameters (CO, NO, NO2, NOX, O3, PM10, PM2.5)
- Meteorological data (wind speed, wind direction, temperature)

V. Methodology

To begin with, this project use scikit-learn [5] packages as it is simple and efficient tools for predictive data analysis. Next, we received data set, and explored data inside, did pre-processing, modeling, evaluation, and deployed into simple website to demonstrate our powerful PM2.5 prediction.

A. Data Acquisition

To obtain the data set, Prof. Chantri received the data set for us. These data sets came from the Pollution Control Department of Thailand (PCD) [2]. Generally, the data are recorded daily for public use. However, we can ask PCD for more details hourly records. Original data looks as shown in Fig. 6.



	YYMMDD	HR	CO	NO	NOX	NO2	SO2	O3	PM10	Wind speed	Wind dir	Temp	Rel hum	Rain	PM2.5
	Unnamed: 0_level_1	Unnamed: 1_level_1	at 3 m (ppm)	at 3 m (ppb)	at 3 m (ppb)	at 3 m (ppb)	at 3 m (ppb)	at 3 m (ppb)	at 3 m (um/m3)	at 10 m (m/s)	at 10 m (Deg.M)	at 2 m	at 2 m (%RH)	at 3 m (mm)	at 3 m (um/m3)
0	190101	100	0.9	1.0	28.0	27.0	2.0	21.0	51	0.1	42	25.8	51.0	0.0	45
1	190101	200	0.9	1.0	29.0	28.0	2.0	13.0	73	0.0	321	25.3	57.0	0.0	72
2	190101	300	0.9	1.0	34.0	33.0	2.0	8.0	44	0.1	47	24.8	59.0	0.0	22
3	190101	400	NaN	NaN	NaN	NaN	NaN	NaN	52	0.1	16	24.5	52.0	0.0	33
4	190101	500	0.8	9.0	45.0	35.0	2.0	6.0	33	0.0	219	24.0	59.0	0.0	25

Fig. 6: Original sensor data from Thonburi station, Bangkok

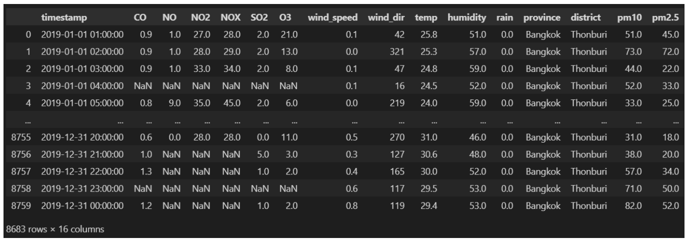
B. Exploratory Data Analysis (EDA)

For this task, we just exploring the data to see if there are potential problems in the data set (outliers, mislabeled data, unwanted correlations between variables/samples, etc) but no actual work done in here. The steps includes changing column names as shown in Fig. 7, datetime formatting, removing duplicates, add city and province columns, and identifying missing values and deciding how to handle them, whether by imputing missing values, removing rows with missing values, or using other strategies.

```
# rename columns
df_52.rename(columns = {'YYMMDD':'yymmdd',
                        'HR':'hr',
                        'CO':'CO',
                        ' NO ':'NO',
                        ' NOX ':'NOX',
                        ' NO2':'NO2',
                        ' SO2 ': 'SO2',
                        ' Wind speed': 'wind_speed',
                        ' Wind dir': 'wind_dir',
                        'PM10':'pm10',
                        'PM2.5':'pm2.5',
                        ' Temp':'temp',
                        ' Rel hum':'humidity',
                        ' Rain':'rain'
                         }, inplace = True)
```

Fig. 7: Renaming all column names

Then, after we did some cleansing and format the data, we could show some insights in the data as shown in Fig. 8.



	timestamp	CO	NO	NO2	NOX	SO2	O3	wind_speed	wind_dir	temp	humidity	rain	province	district	pm10	pm2.5
0	2019-01-01 01:00:00	0.9	1.0	27.0	28.0	2.0	21.0	0.1	42	25.8	51.0	0.0	Bangkok	Thonburi	51.0	45.0
1	2019-01-01 02:00:00	0.9	1.0	28.0	29.0	2.0	13.0	0.0	321	25.3	57.0	0.0	Bangkok	Thonburi	73.0	72.0
2	2019-01-01 03:00:00	0.9	1.0	33.0	34.0	2.0	8.0	0.1	47	24.8	59.0	0.0	Bangkok	Thonburi	44.0	22.0
3	2019-01-01 04:00:00	NaN	NaN	NaN	NaN	NaN	NaN	0.1	16	24.5	52.0	0.0	Bangkok	Thonburi	52.0	33.0
4	2019-01-01 05:00:00	0.8	9.0	35.0	45.0	2.0	6.0	0.0	219	24.0	59.0	0.0	Bangkok	Thonburi	33.0	25.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8755	2019-12-31 20:00:00	0.6	0.0	28.0	28.0	0.0	11.0	0.5	270	31.0	46.0	0.0	Bangkok	Thonburi	31.0	18.0
8756	2019-12-31 21:00:00	1.0	NaN	NaN	NaN	5.0	3.0	0.3	127	30.6	48.0	0.0	Bangkok	Thonburi	38.0	20.0
8757	2019-12-31 22:00:00	1.3	NaN	NaN	NaN	1.0	2.0	0.4	165	30.0	52.0	0.0	Bangkok	Thonburi	57.0	34.0
8758	2019-12-31 23:00:00	NaN	NaN	NaN	NaN	NaN	NaN	0.6	117	29.5	53.0	0.0	Bangkok	Thonburi	71.0	50.0
8759	2019-12-31 00:00:00	1.2	NaN	NaN	NaN	1.0	2.0	0.8	119	29.4	53.0	0.0	Bangkok	Thonburi	82.0	52.0

8683 rows × 16 columns

Fig. 8: data after do some cleansing before put into EDA

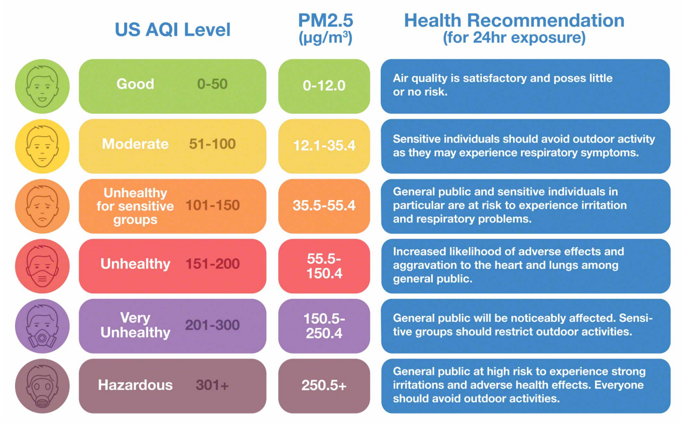
At first, we did some plotting to see the trend of PM2.5 daily in Thonburi station (52t). Additionally, we also did some histogram plot to see the distribution of PM2.5. Then, we did some scatter plot to see more insight of that predictor. Next, we did the correlation heat map plot to see potential for predictor to use in Fig. 11a. We also computed the power score in Fig. 11b, which calculate how strong each predictor can predict other (target) variables. We created the box plot to see the distribution, outliers, mean, and minimum and maximum range for each numerical variables as shown in Fig. 12.

C. Pre-processing

After exploring what our data set looks like, we do pre-processing - the steps required to go from raw data to a format suitable to input to your ML model [1].

Replace missing values - Since all columns are received from sensor data. Assuming that all data were calibrated correctly, we filled with forward fill, and some for backward fill.

Feature Engineering - We do add the Air Quality Index (AQI) index based on Thailand Air Pollution Department criterion as shown in Fig. 9. We added city and province columns for later if data set from other stations are acquired.



US AQI Level		PM2.5 (µg/m³)	Health Recommendation (for 24hr exposure)
Good	0-50	0-12.0	Air quality is satisfactory and poses little or no risk.
Moderate	51-100	12.1-35.4	Sensitive individuals should avoid outdoor activity as they may experience respiratory symptoms.
Unhealthy for sensitive groups	101-150	35.5-55.4	General public and sensitive individuals in particular are at risk to experience irritation and respiratory problems.
Unhealthy	151-200	55.5-150.4	Increased likelihood of adverse effects and aggravation to the heart and lungs among general public.
Very Unhealthy	201-300	150.5-250.4	General public will be noticeably affected. Sensitive groups should restrict outdoor activities.
Hazardous	301+	250.5+	General public at high risk to experience strong irritations and adverse health effects. Everyone should avoid outdoor activities.

Fig. 9: US AQI levels, equivalent PM2.5 standards by $\mu g/m^3$, and health recommendations for each level. [3]

Split train-test - we splited original into test set ratio of 0.2 with fixed random state. After that, we then imputed and scaled the data in train set.

Impute and Scale - For imputing, forward and backward fill are the most suitable since the data came from sensors directly. In the data, most columns are numerical. First, we'll check the distribution of the data. If it was normal distribution, we'll fill it with StandardScaler(). If it was not, we'll fill it with MinMaxScaler(). Thus, we'll do some scaling to make the model trained better.

D. Modeling

The model we selected came from most in regression models.

1) Linear Regression will be used for basic prediction of PM2.5. It is the baseline for our goals.
2) Gradient Boosting Regression
3) ϵ-support Vector Regression (SVR)
4) K Neighbor Regression
5) Decision Tree Regression
6) Random Forest Regression
7) AdaBoost Regression
8) ARIMA (Auto Regressive Integrated Moving Average) is a time series forecasting model in Python. Since the problem involves forecast with time-series, this model was considered.

To select the best model for this purpose, we do k-folds cross validation with k = 5.

After we got the best model, we did grid search for finding the best hyper-parameters for that model.

E. Training

Training will be done by the scikit-learn .fit method.

F. Evaluation

After trained the model, we'll evaluated the model using the cross entropy loss or log loss $L(w)$ as shown in eq. 5. This evaluation process were nicely provided by scikit-learn for easy implementation.

$$y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \tag{5}$$

G. Deployment

Deployment will use Flask and some basic html to show the results. Also we'll deploy the code on Github.

VI. Preliminary Results

After we do the EDA in V-B and pre-processing in V-C, we do some visualization and got some interesting results.

At first, We do uni-variate analysis, plotting to see the trend of average PM2.5 daily in Thonburi station in Fig. 10a. It can be seen that PM2.5 is quite high in the first two months. Then it drops dramatically at each its lowest on September. Then, it starts rising again up until the end of December. Additionally, we also did some histogram plot to see the distribution of PM2.5 in Fig. 10b. It normal range is from between 15 to 35 most of the times. Then, we did some scatter plot to see more insight between PM2.5 and PM10 of that predictor in Fig. 10. As predicted, it shows that PM2.5 strongly related to PM10. Furthermore, if these two are high, CO value will also be high too.

We did multivariate analysis, we did the correlation heat map plot to see potential for predictor to use in Fig. 11a. We also computed the power score in Fig. 11b, which calculate how strong each predictor can predict other (target) variables. We created the violin plot to see outliers, mean, and minimum and maximum range for each numerical variables as shown in Fig. 12.

References

[1] Leah Luyen, “EDA, Data Preprocessing, Feature Engineering: We are different!”. Medium.com. https://medium.com/@ndleah/eda-data-preprocessing-feature-engineering-we-are-different-d2a5fa09f527 (access October 23, 2023).

[2] Pollution Control Department of Thailand. “Particulate Matter Historical Data”. http://air4thai.pcd.go.th/webV2/history/ (access October 16, 2023).

[3] https://www.iqair.com/th-en/newsroom/thailand-2021-burning-season

[4] Mehdipour, V., Stevenson, D.S., Memarianfard, M. et al. “Comparing different methods for statistical modeling of particulate matter in Tehran, Iran”. Air Quality Atmosphere Health 11, pp.1155—1165 (2018). https://doi.org/10.1007/s11869-018-0615-z

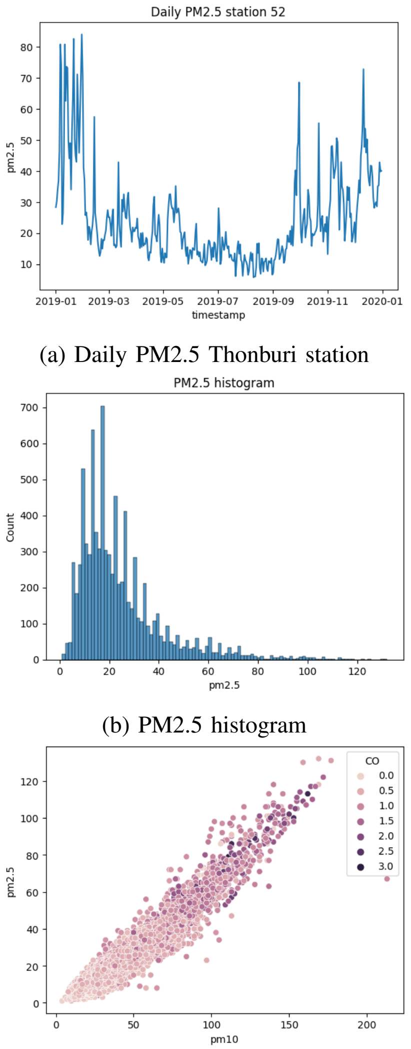
[5] https://scikit-learn.org/stable/

[6] Nidhi Sharma, Shweta Taneja, Vaishali Sagar, and Arshita Bhatt. “Forecasting air pollution load in Delhi using data analysis tools”. Procedia Computer Science. vol. 132. 2018. pp.1077–1085. https://www.sciencedirect.com/science/article/pii/S1877050918307555.

[7] Doreswamy, Harishkumar K S, Yogesh KM and Ibrahim Gad. “Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models”. Procedia Computer Science. vol. 171. 2020. pp. 2057–2066. https://www.sciencedirect.com/science/article/pii/S1877050920312060

[8] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[9] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.



(a) Daily PM2.5 Thonburi station

(b) PM2.5 histogram

(c) Scatter plot of PM2.5 relatives to PM10 and colored by CO

Fig. 10: Some PM2.5 Insights

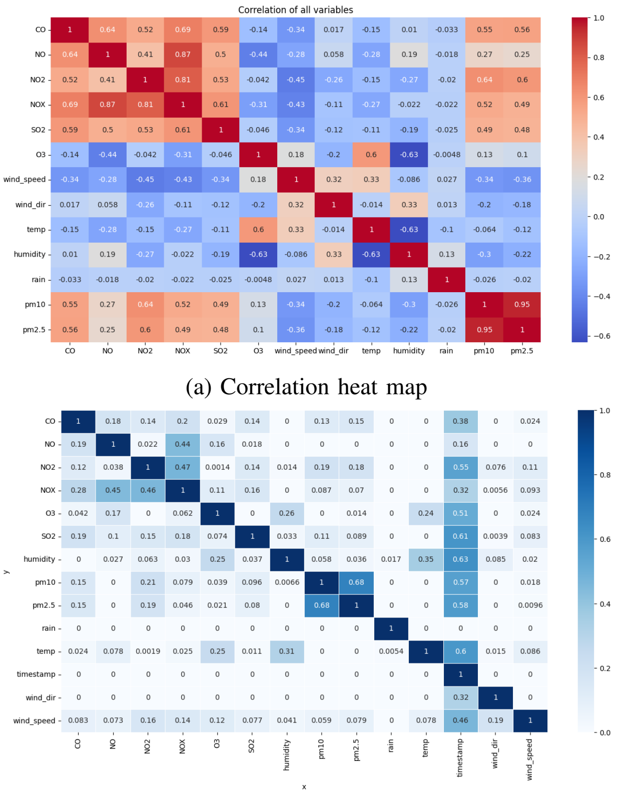
[10] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[11] K. Elissa, "Title of paper if known," unpublished.

[12] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[13] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[14] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.



(a) Correlation heat map

(b) Power score

Fig. 11: Correlation and Power score between variables

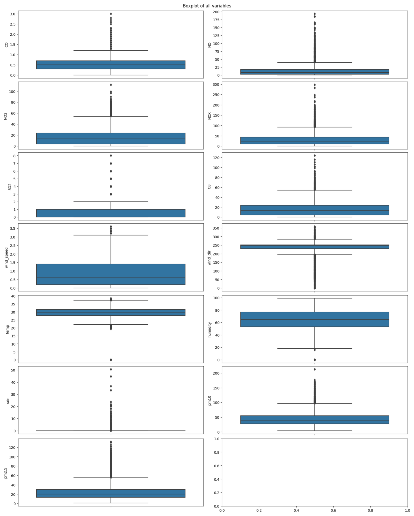


Fig. 12: Box plot of all variables