Air Quality Analysis and Prediction in Tamil Nadu

Phase-5

INTRODUCTION : Technological advancements lead to the emissions of air pollutants over the decades. Major concerns in industrial cities which experience air pollution, can be harmful not only for the environment but also for human health. Due to this urban resident are more likely to live in less polluted neighborhoods to avoid the health impact of air pollution. Atmospheric pollution can be classified into three types based on the sources mobile, stationery and area sources. Mobile sources are due to the motor vehicles, airplanes, locomotives and other engines and equipment that are able to move to different locations. Stationary sources include foundries, fossil fuel burning, food processing plants, power plants, refineries and other industrial sources. Area sources is caused by certain local actions. Air pollution can be caused due to the pollutants which are emitted directly from a source or which are not directly emitted as such. It can result in the degradation of ambient air quality in the industrial cities. Also daily exposure of people to air pollution results in diseases like asthma, wheezing, and bronchitis.

DATASET :

The data is obtained from
https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014

Original dataset with columns and rows

```
air
```

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
0	38	01-02-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	12-10-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-14	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

2879 rows × 11 columns

COLUMNS USED :

From Tamil Nadu_Air quality analytsishe following columns are used :

. stn code

. Sampling Date

. State

. City/Town/Village

. Location of agency

. Type of location

. SO2

. NO2

. RSPM/PM10

. PM2.5

LIBRARIES USED:

The Python 3 environment comes with many helpful analytics libraries installed and several helpful packages to load.

The essential libraries used in this project are :

- Importing OS (for kaggle inputs)

- Numpy and Pandas libraries
- Matplotlib
- Seaborn

TRAIN AND TEST:

Training the dataset by describe(), isnull().sum(), drop(), show(), and by using k-means algorithm we train the data

Testing the data by importing sklearn.cluster from k-means with ensuring the plot range and axis labels producing the k value, scattering the data by kmeans.cluster_centers and producing 3D plot.

Data Collection:

The samples are collected from NAMP stations are analysed for the Respirable Suspended Particulate matter (RSPM) and gaseous pollutants such as Sulphur dioxide(SO2) and Nitrogen dioxides(NO2)

Data analysis:

ANOVA (one way), Tukey HSD, and Pearson correlation coefficient (*r*) were computed using self-coded software on Microsoft Excel 2019 to statistically analyze the collected data.

ALGORITHMS USED:

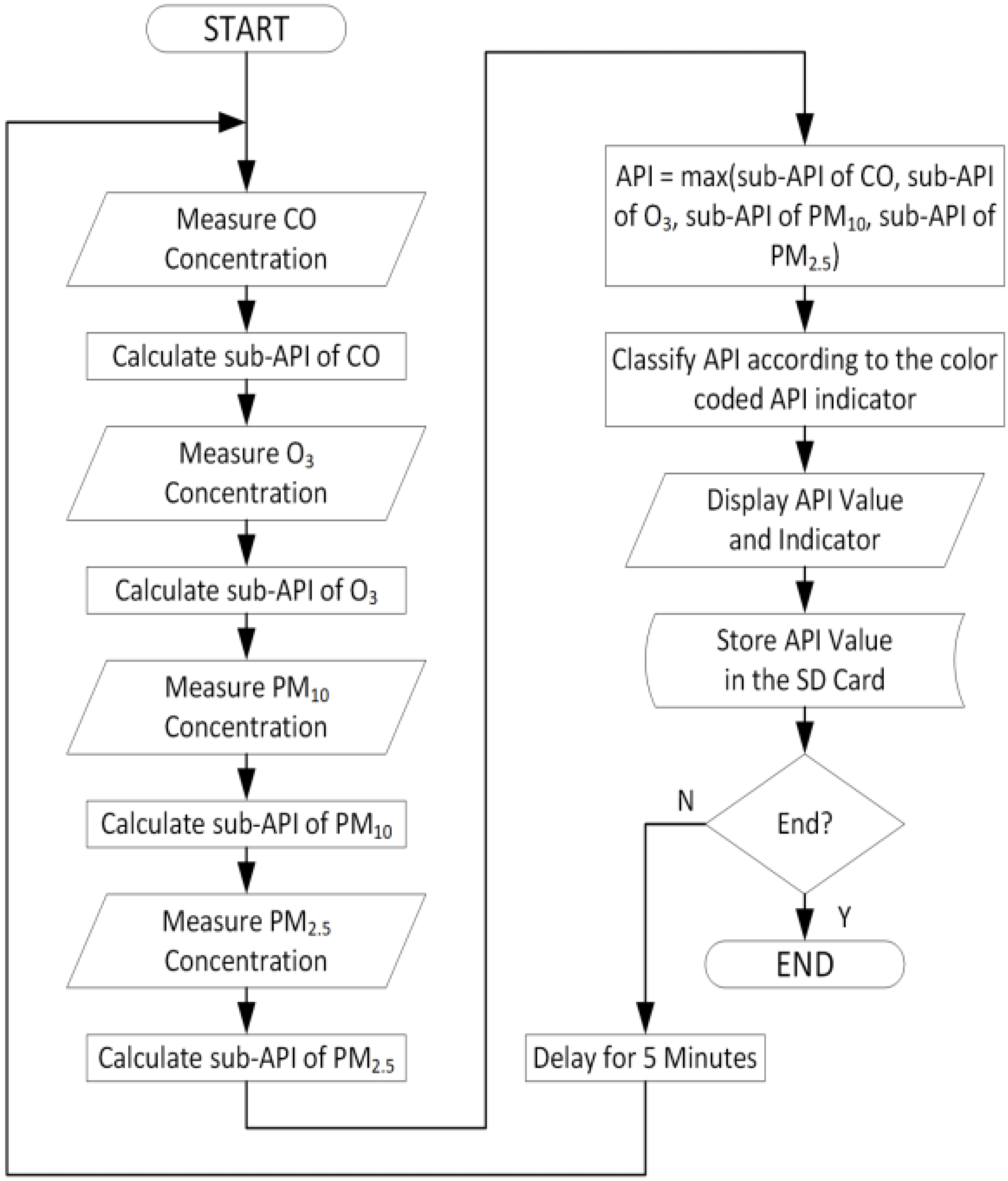
Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

Visualization: Visualize the customer segments using techniques like scatter plots, bar charts, and heatmaps. Interpretation: Analyze and interpret the characteristics of each customer segment to derive actionable insights for marketing strategies.

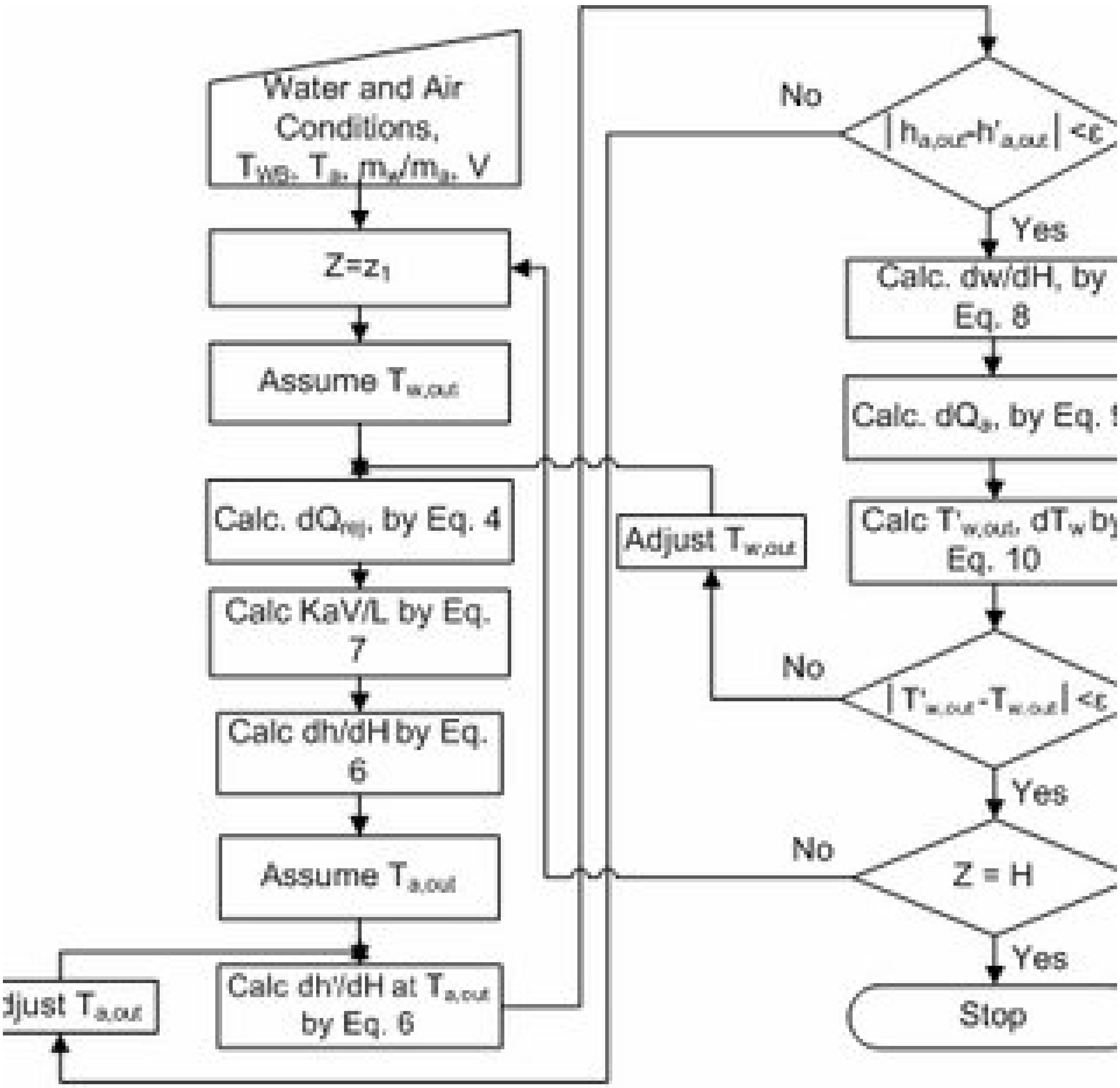
DESIGN AND DATAFLOW:

1.Physical data flow diagram:

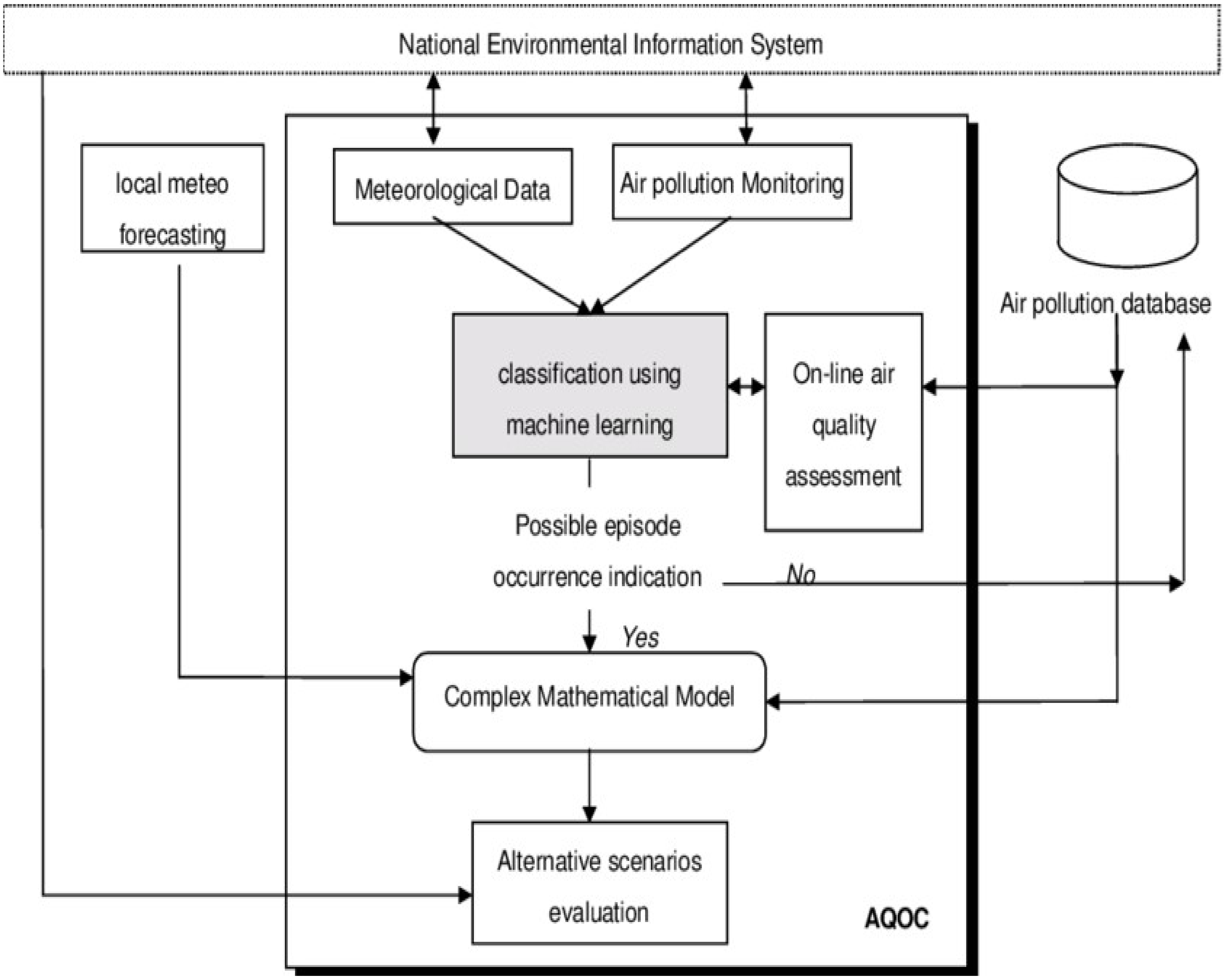
AIR QUALITY ANALYSIS AND PREDICTION IN TAMIL NADU


START
Measure CO Concentration
Calculate sub-API of CO
Measure O_3 Concentration
Calculate sub-API of O_3
Measure PM_{10} Concentration
Calculate sub-API of PM_{10}
Measure $PM_{2.5}$ Concentration
Calculate sub-API of $PM_{2.5}$
API = max(sub-API of CO, sub-API of O_3, sub-API of PM_{10}, sub-API of $PM_{2.5}$)
Classify API according to the color coded API indicator
Display API Value and Indicator
Store API Value in the SD Card
End?
N
Y
END
Delay for 5 Minutes

2.Logical data flow diagram:



3. Data flow diagram



Information About Dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Stn Code                        2879 non-null   int64
 1   Sampling Date                   2879 non-null   object
 2   State                           2879 non-null   object
 3   City/Town/Village/Area          2879 non-null   object
 4   Location of Monitoring Station  2879 non-null   object
```

```
 5   Agency                       2879 non-null   object
 6   Type of Location             2879 non-null   object
 7   SO2                          2868 non-null   float64
 8   NO2                          2866 non-null   float64
 9   RSPM/PM10                    2875 non-null   float64
 10  PM 2.5                       0 non-null      float64
dtypes: float64(4), int64(1), object(6)
memory usage: 247.5+ KB
```

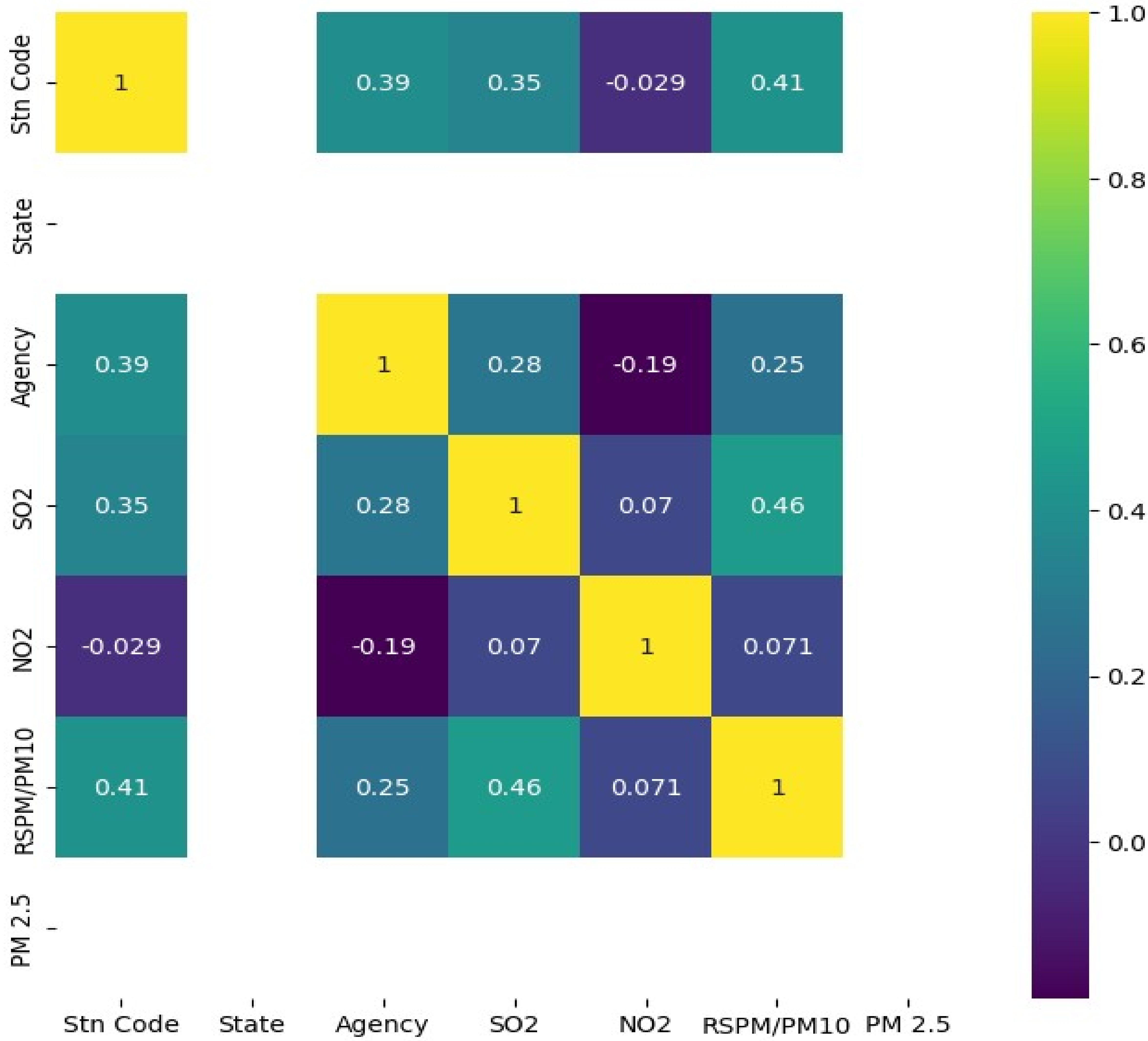
Checking Missing Values:

checking missing values

```
air.isnull().sum()
```

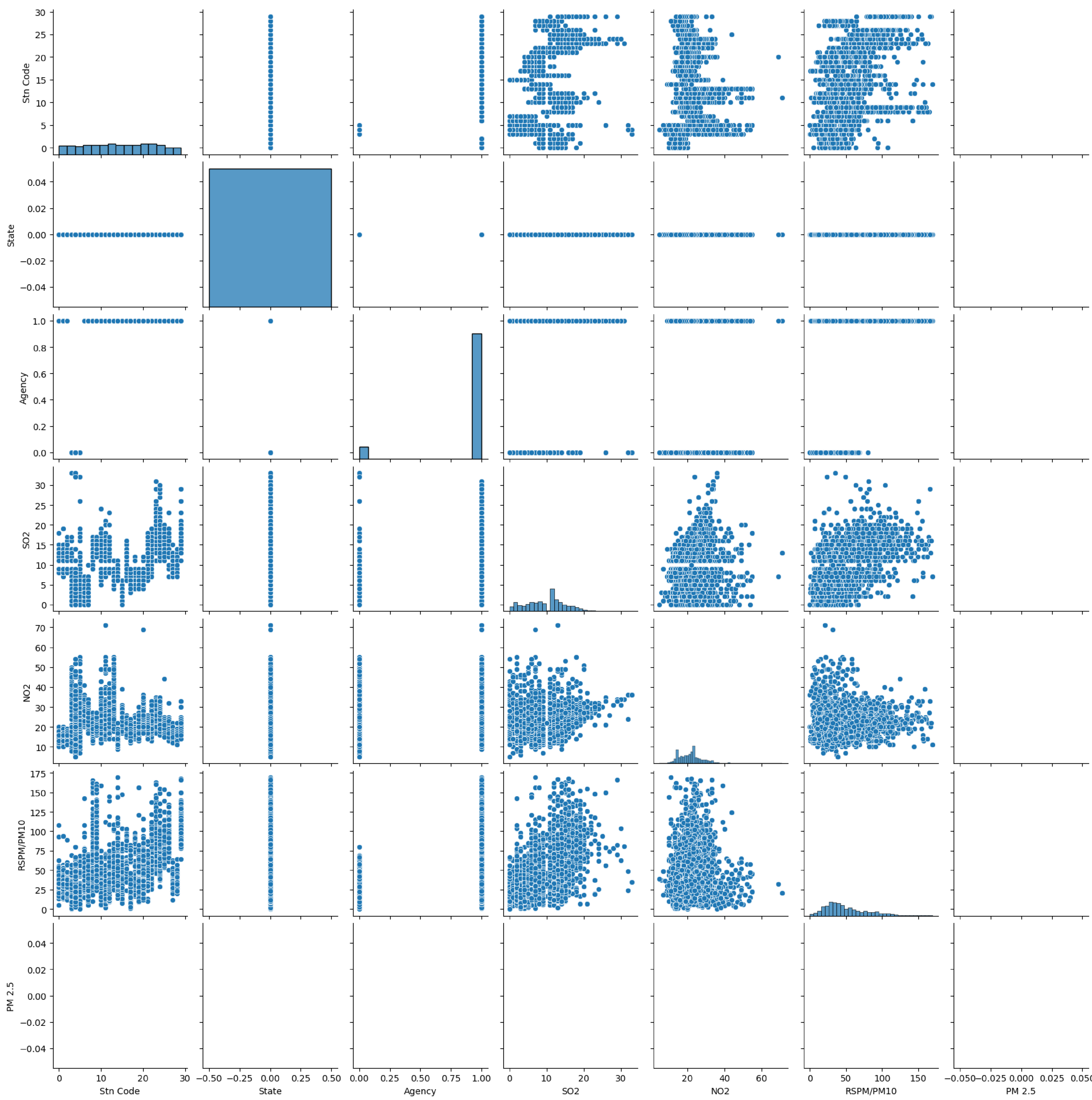
```
Stn Code                              0
Sampling Date                         0
State                                 0
City/Town/Village/Area                0
Location of Monitoring Station        0
Agency                                0
Type of Location                      0
SO2                                  11
NO2                                  13
RSPM/PM10                             4
PM 2.5                             2879
dtype: int64
```

Model Analysis:

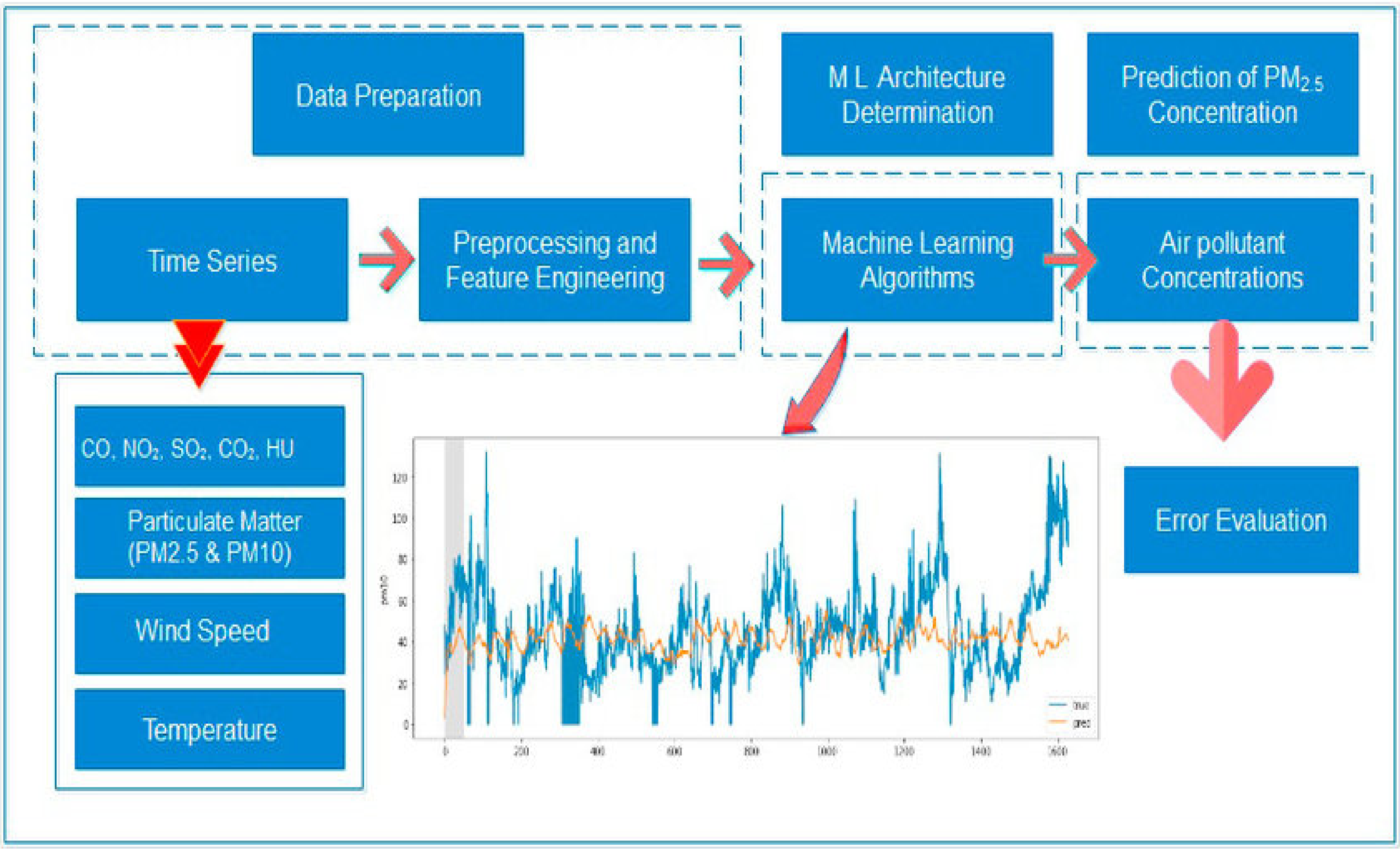


Stn Code
State
Agency
SO2
NO2
RSPM/PM10
PM 2.5
1
0.39
0.35
-0.029
0.41
0.28
-0.19
0.25
0.07
0.46
0.071
1.0
0.8
0.6
0.4
0.2
0.0

Model comparision:



Data preprocessing:



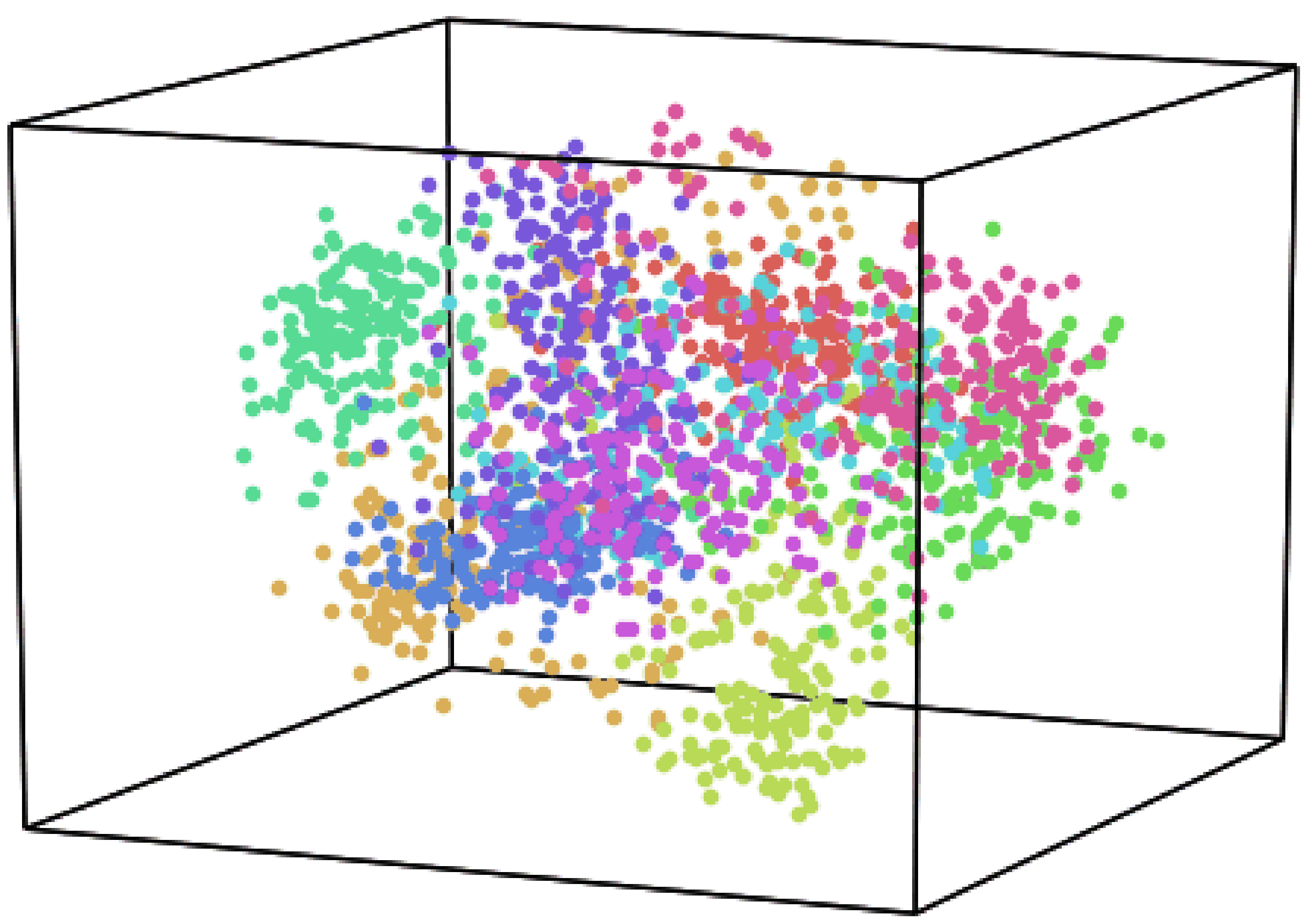
In data preprocessing,

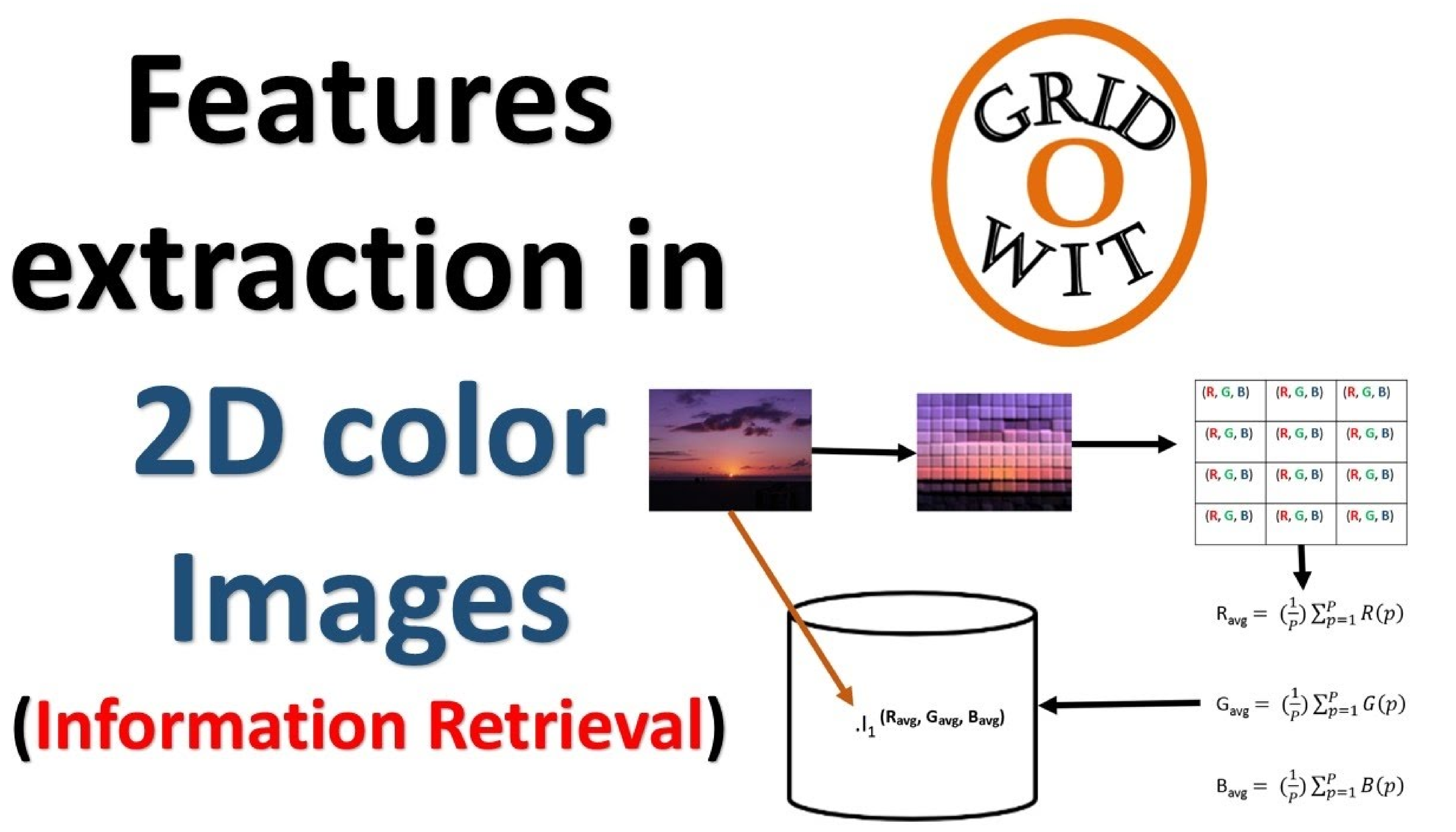
they cleaned the original dataset and extracted the New Delhi, Bangalore, Kolkata, and Hyderabad city data. Because these are major cities in India, it is important to analyze the pollution levels in different urban cities in India as they are the major contributors to the pollution.

A great number of technologies and instruments both for sampling and determination of the concentration levels.

Extraction techniques:

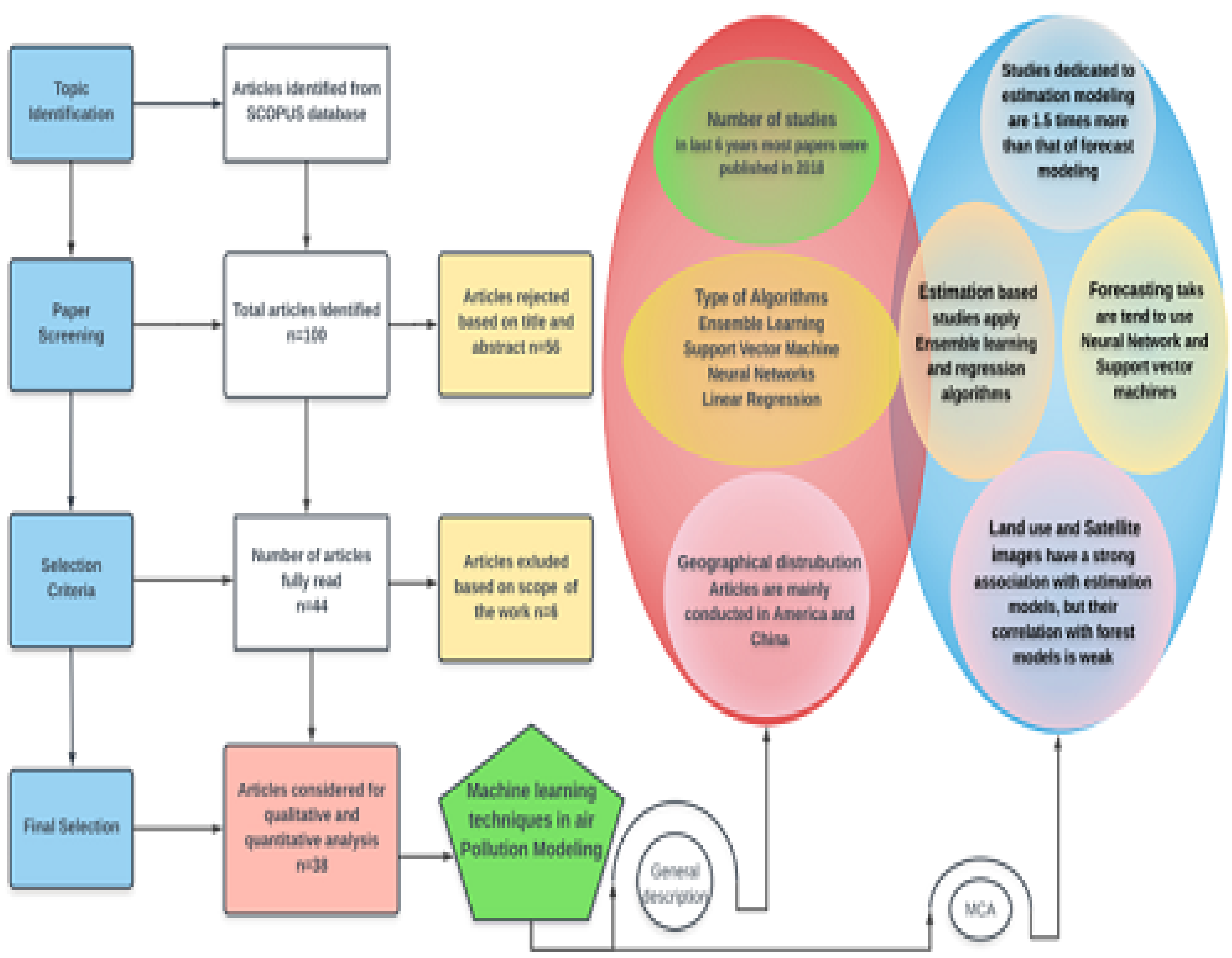
Some other pollutants such as chlorine, ammonia and hydrogen cyanide can be determined by **Infrared spectroscopy**. The organic pollutant collected and concentrated from air can be determined by freeze out techniques. Gas chromatography is a great method to

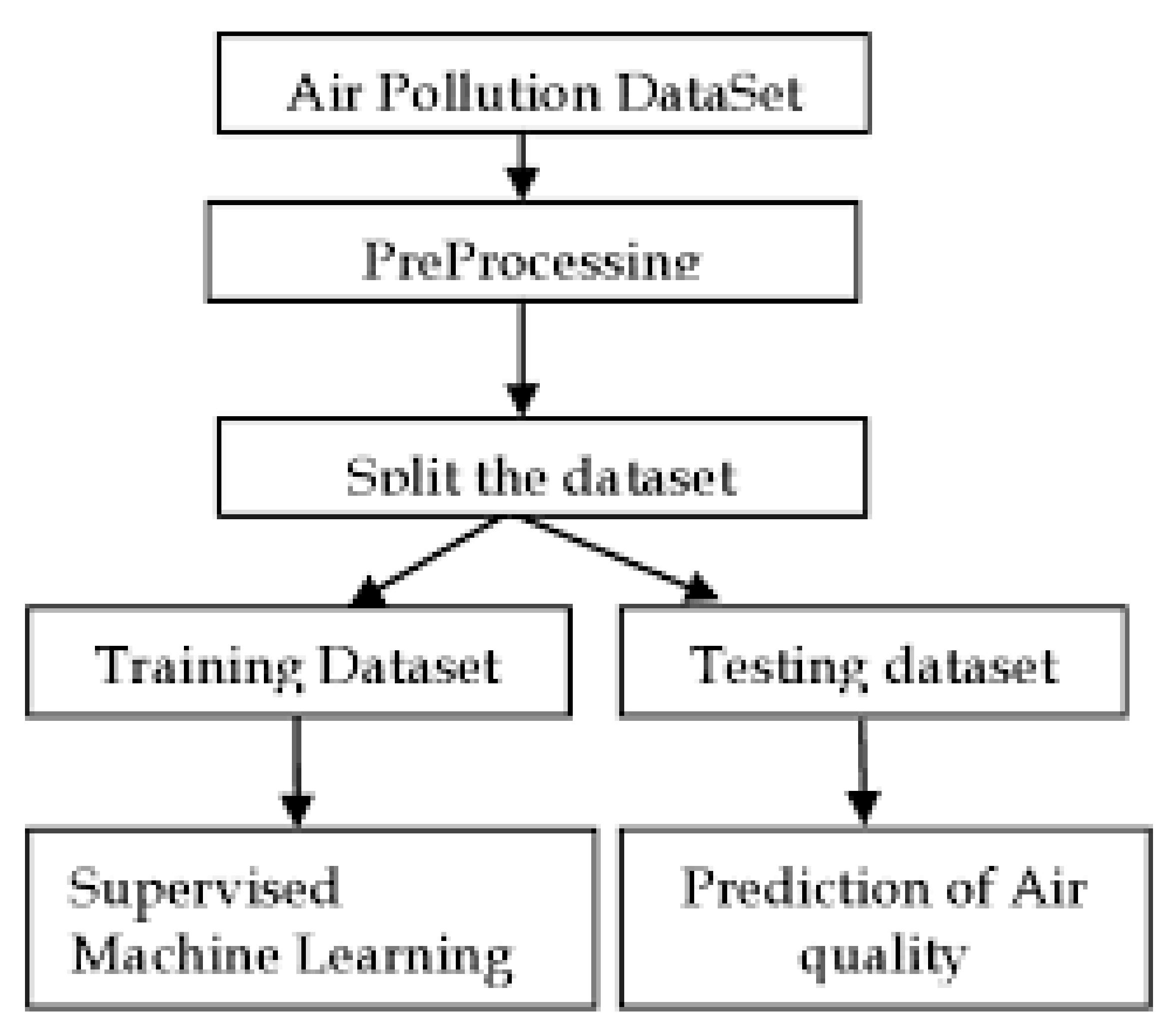




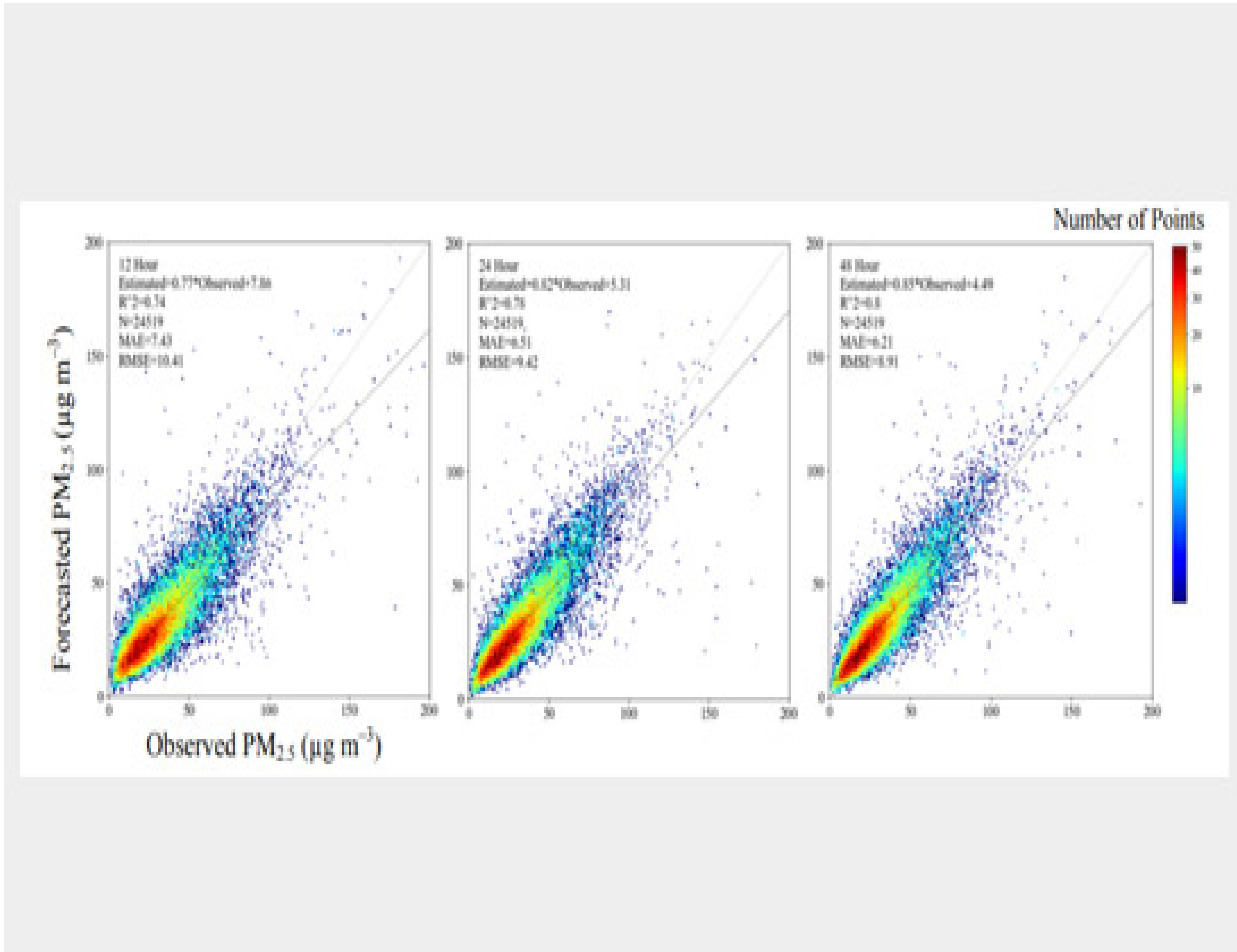
Algorithms used:

Linear regression was used as a machine learning algorithm to predict air quality for the next day using sensor data from three specific locations in the Capital City of India-Delhi and the National Capital Region (NCR). The model's performance was assessed using four performance measures: MAE, MSE, RMSE, and MAPE.





Model Training:

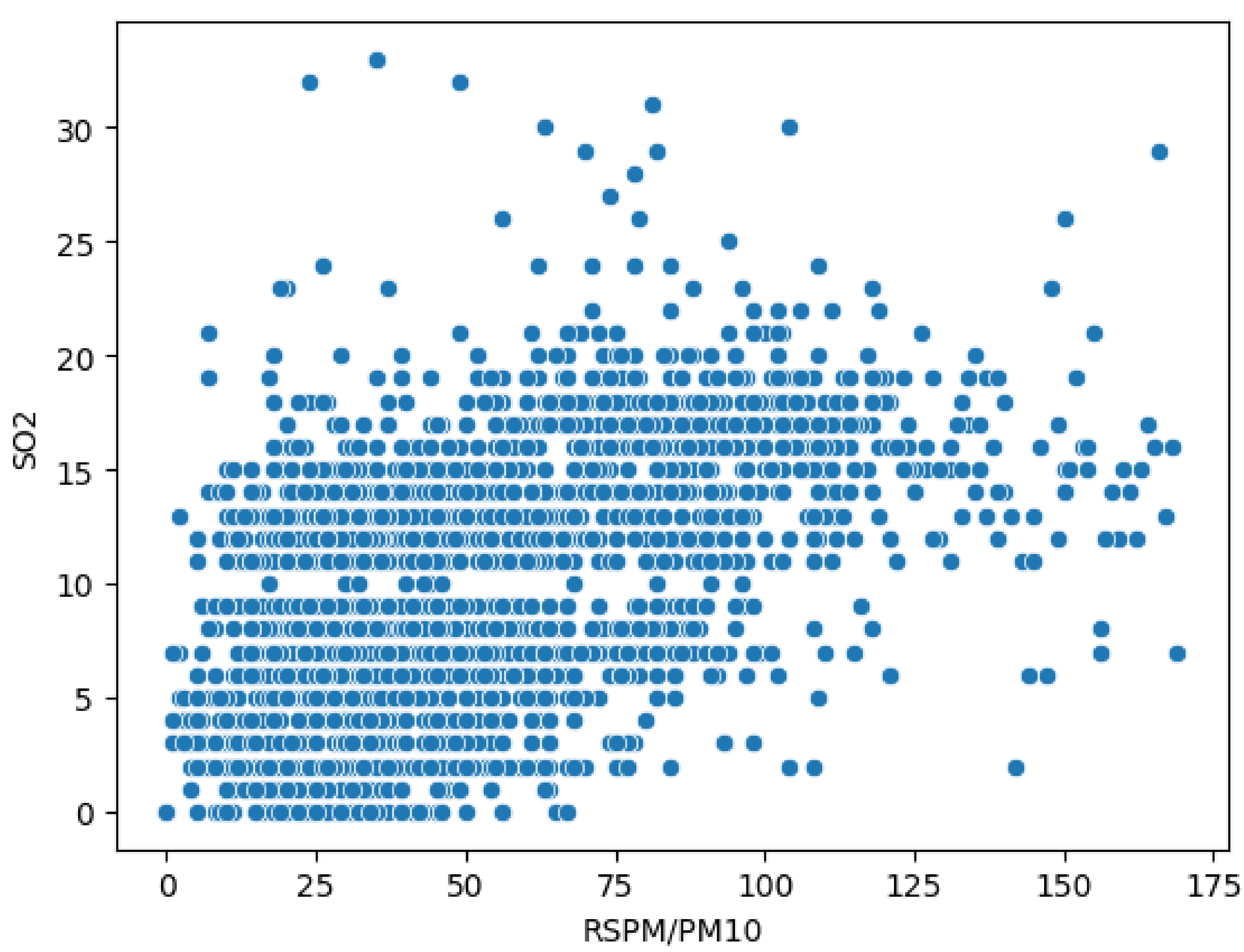


The air quality forecasts are generated using high-resolution meteorological forecast models coupled with a sophisticated air-mass trajectory analysis (HY-SPLIT) and, in the case of ozone, complex, photochemical grid models.

Model evaluation techniques:

Air quality modeling refers to the use of mathematics and computer programs to estimate concentrations of pollutants in the air. Air quality modeling is a United States Environmental Protection Agency (U.S. EPA) approved method for evaluating air quality impacts from air emission sources such as factories and roads.

Scatter Plots for Air quality Analysis;



CODE:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn import metrics
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.tree import DecisionTreeRegressor
import xgboost as xgb
from sklearn.cluster import KMeans

air=pd.read_csv('/content/Air quality-analysis-2014.csv')
air
air.describe()
air.info()
air.isnull().sum()
air_fillna = air
air_fillna.fillna(air_fillna.mean(), inplace=True)
#count the number of NaN values in each column
print(air_fillna.isnull().sum())

air_fillna
le=LabelEncoder()
air['State']=le.fit_transform(air['State'])
air
le=LabelEncoder()
air['Stn Code']=le.fit_transform(air['Stn Code'])
```

```
air
le=LabelEncoder()
air['SO2']=le.fit_transform(air['SO2'])
air
le=LabelEncoder()
air['Agency']=le.fit_transform(air['Agency'])
air
le=LabelEncoder()
air['RSPM/PM10']=le.fit_transform(air['RSPM/PM10'])
air
air['Sampling Date'] =air['Sampling Date'].str.replace('-','')
air
air
air.columns
corr =air.corr()
plt.figure(figsize=(8,8))
sns.heatmap(corr,cmap='viridis',annot=True)
sns.pairplot(air)
sns.regplot( y="Agency",x="Type of Location", data=air)
sns.scatterplot( y="SO2",x="RSPM/PM10", data=air)
sns.displot(air, x="State",hue="SO2", common_norm=False)
sns.scatterplot(air,x='Type of Location',y="State")
sns.displot(air, x="State",kde=True)
sns.displot(air, x="City/Town/Village/Area",kde=True)
sns.regplot( y="State",x="Stn Code", data=air)
x=air[['Stn Code','Sampling Date', 'State', 'City/Town/Village/Area','Location of
Monitoring Station','Agency','Type of Location','SO2', 'NO2', 'RSPM/PM10',
 'PM2.5']]
air
y=air[['RSPM/PM10']]
```

```
y
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=42)

x_train
x_test
y_train
y_test
LR=LinearRegression()
dataset=pd.read_csv('/content/Air quality-analysis-2014.csv')
data = dataset.sample(frac=0.9,random_state=786).reset_index(drop=True)
data_unseen = dataset.drop(data.index).reset_index(drop=True)

print('Data for Modeling: ' + str(data.shape))
print('Unseen Data For Predictions: ' + str(data_unseen.shape))
dataset_fillna = dataset
dataset_fillna.fillna(dataset_fillna.mean(),inplace=True)
#count the number of NaN values in each column
print(dataset_fillna.isnull().sum())

le=LabelEncoder()
dataset['State']=le.fit_transform(dataset['State'])
dataset
=LabelEncoder()
dataset[le'Stn Code']=le.fit_transform(dataset['Stn Code'])
dataset
le=LabelEncoder()
dataset['Agency']=le.fit_transform(dataset['Agency'])
dataset
le=LabelEncoder()
dataset['Type of Location']=le.fit_transform(dataset['Type of Location'])
dataset
```

```
dataset['Sampling Date'] = dataset['Sampling Date'].str.replace('-', ' ')
dataset
```

CONCLUSION:

The process of evaluating the dataset and the objective of project air quality analysis and visualization techniques dispalying the air quality levels in tamilnadu has been done.

Performed the instructions on how to replicate the analysis of the project and performed some calculations on the dataset and created some visualizations in this project using python.