# Predicting air pollution level in a Delhi

Mcmillon Walton

2017chb1047@iitrpr.ac.in

## 1. INTRODUCTION

The regulation of air pollutant levels (AQI) is rapidly becoming one of the most important tasks for the government of India, especially in Delhi. According to the global Environment Performance Index (EPI) 2018, India is ranked at 177 with an EPI of 30.57, and it is disheartening to hear that Delhi, the national capital of the country, is being tagged as one of the most heavily polluted capital cities in the world. It is the world's worst city in terms of air pollution, with with an unhealthy air quality index for the majority of the year.

Among the pollution that is caused by the particulate matters, Fine particulate matter (PM2.5) is a significant one because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated.

Air pollution is not only caused by the concentration of the particulate matters but also gases. These gases include carbon, Sulphur and nitrogen oxides. While some of these gases occur naturally, like carbon dioxide in the expulsion of air from the lungs, the serious polluters come from the burning of fossil fuels: coal, oil and natural gas. In this study however we will focus more on the particulate matters.

The air quality index (AQI) is a simple color coded, unitless index that is an effective way to communicate air pollution concentration to the general public . Simple equations are used to convert the concentration of the particulate matters into AQI (These equation vary).

For example in **India**

$$\text{AQI} = \text{Max}\ (I_p)\ (\text{where; } p = 1,2,...,n\text{; denotes n pollutants})$$

I is the sub index of the pollutant and it varies for each pollutant, like for Sulphur dioxide

$$I_{SO2} = 84 * X^{0.431}$$

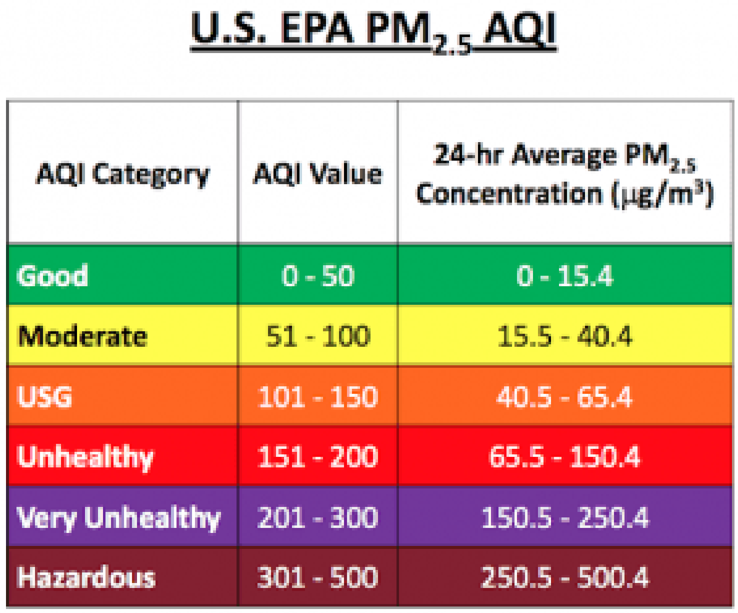
X is the observed pollutant concentration.

In **US**

$$\text{AQI} = \frac{(PM_{obs} - PM_{min}) \times (AQI_{max} - AQI_{min})}{(PM_{max} - PM_{min})} + \text{AQI}_{min}$$

$PM_{obs}$ = observed 24-hour average concentration in μg/m3

$PM_{max}$ = maximum concentration of AQI color category that contains $PM_{obs}$
$PM_{min}$ = minimum concentration of AQI color category that contains $PM_{obs}$
$AQI_{max}$ = maximum AQI value for color category that corresponds to $PM_{obs}$
$AQI_{min}$ = minimum AQI value for color category that corresponds to $PM_{obs}$



## U.S. EPA $PM_{2.5}$ AQI

| AQI Category | AQI Value | 24-hr Average $PM_{2.5}$ Concentration ($\mu g/m^3$) |
|---|---|---|
| Good | 0 - 50 | 0 - 15.4 |
| Moderate | 51 - 100 | 15.5 - 40.4 |
| USG | 101 - 150 | 40.5 - 65.4 |
| Unhealthy | 151 - 200 | 65.5 - 150.4 |
| Very Unhealthy | 201 - 300 | 150.5 - 250.4 |
| Hazardous | 301 - 500 | 250.5 - 500.4 |

Even healthy people can experience health impacts from polluted air including respiratory irritation or breathing difficulties during exercise or outdoor activities. Your actual risk of adverse effects depends on your current health status, the pollutant type and concentration, and the length of your exposure to the polluted air.

High air pollution levels can cause immediate health problems including:

- Aggravated cardiovascular and respiratory illness
- Added stress to heart and lungs, which must work harder to supply the body with oxygen
- Damaged cells in the respiratory system

Long-term exposure to polluted air can have permanent health effects such as:

- Accelerated aging of the lungs
- Loss of lung capacity and decreased lung function
- Development of diseases such as asthma, bronchitis, emphysema, and possibly cancer
- Shortened life span

Those most susceptible to severe health problems from air pollution are:

- Individuals with heart disease, coronary artery disease or congestive heart failure
- Individuals with lung diseases such as asthma, emphysema or chronic obstructive pulmonary disease (COPD)
- Pregnant women
- Outdoor workers

- Older adults and the elderly
- Children under age 14
- Athletes who exercise vigorously outdoors

People in these groups may experience health impacts at lower air pollution exposure levels, or their health effects may be of greater intensity.

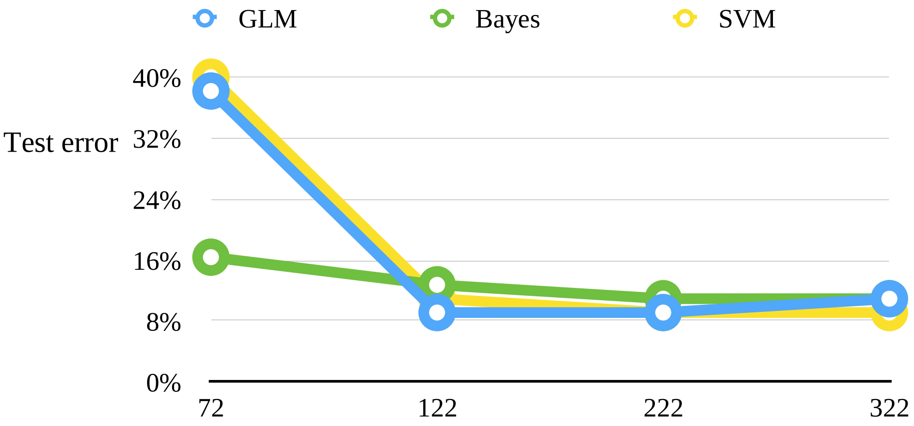
This study is mainly conducted as the relationships between the concentration of these particles with the meteorological and traffic factors are poorly understood. To shed some light on these connections, some of these advanced techniques have been introduced into air quality research. This study utilized the selected technique, Support Vector Machine (SVM), to predict ambient air pollutant levels based on mostly weather and sometimes traffic variables.
This project attempted to apply some machine learning techniques to predict PM2.5 levels on locations where stations are not installed based on a dataset consisting of daily weather and traffic parameters in Delhi. With this study we could determine the least polluted way to reach from point A to B, source apportionment could be determined and then policies could be made to reduce pollution more efficiently.

## 2. Background

A similar study was conducted in Beijing(China) by Dan Wei, He tried out 3 methods

- Logistic regression
- Naïve Bayes Classification
- Support vector machines (SVM)

Figure 1. Test error curve for three models



Their data size is 322. For them they found out that SVM was the most precise of them all. However he also states that other than the meteorological and traffic factors, industrial factors should also be considered.

In the city of Lowa (USA) a study was conducted by Dixian Zhu, Changjie Cai, Tian bao Yang and Xun Zhou. They came up with 11 models.

Their general formulation was , they let ($\mathbf{x}_i$; $y_i$) denote the *i*th training data, where $y_i \in \mathrm{R}^{24\times1}$ denotes the concentration of a certain air pollutant on a day, and $\mathbf{x}_i = (\mathbf{u}_i; \mathbf{v}_i)$ denotes the observed data on the previous day that include two components, where a semicolon “;” represents the column layout. The first component $\mathbf{u}_i = (\mathbf{u}_{i,1}; \ldots; \mathbf{u}_{i,D}) \in \mathrm{R}^{24\cdot D\times1}$ includes all meteorological data over 24 h for the previous day, where $\mathbf{u}_{i,j} \in \mathrm{R}^{24\times1}$ denotes the *j*th meteorological feature of the 24 h and $D$ is the number of meteorological features; the second component $\mathbf{v}_i \in \mathrm{R}^{24\times1}$ includes the hourly concentration of the same air pollutant on the previous day. The general formulation can be expressed as

$$\min_{W} \frac{1}{n}\sum_{i=1}^{n} \mathrm{k}f(W, \mathbf{x}_i) - y_i\|_2^2 + \varphi(W) \qquad (1)$$

where $W$ denotes the parameters of the model, $f(W, \mathbf{x}_i)$ denotes the prediction of the air pollutant concentration, and $\phi(\cdot)$ denotes a regularization function of the model parameters $W$.

Next, we introduce two levels of model regularization. The first level is to explicitly control the number of model parameters. The second level is to explicitly impose a certain regularization on the model parameter. For the first level, we consider three models that are described below:

- **Baseline Model**. The first model is a baseline model that has been considered in existing studies and has the fewest number of parameters. In particular, the prediction of the air pollutant concentration is given by

$$f_k(W, \mathbf{x}_i) = \sum_{j=1}^{D} \mathbf{e}_k^{\top} \mathbf{u}_{i,j} \cdot w_j + \mathbf{e}^{>}_k \mathbf{v}_i \cdot w_{D+1} + w_0, \qquad k = 1, \ldots, 24$$

where $\mathbf{e}_k \in \mathrm{R}^{24\times1}$ is a basis vector with 1 at only the *k*th position and 0 at other positions; $w_0, w_1, \ldots, w_D, w_{D+1} \in \mathrm{R}$ are the model parameters, where $w_0$ is the bias term. We denote this model by $W = (w_0, w_1, \ldots, w_{D+1})^{>}$. It is notable that this model predicts the hourly concentration on the basis of the same hourly historical data of the previous day and that it has $D + 2$ parameters. This simple model assumes that all 24 h share the same model parameter.

**Heavy Model**. The second model takes all the data of the previous day into account when predicting the concentration of every hour of the second day.

**Light Model**. The third model is between the baseline model and the heavy model. It considers the 24 h pattern of the air pollutants in the previous day and the same hourly meteorological data of the previous day to predict the concentration at a particular hour.

With these 3 base models and with 4 regularization of model parameters which are:-

1. **Frobenius norm regularization**
2. $l_{2,1}$ **norm regularization**
3. **Nuclear norm regularization**

4. **Consecutive close (CC) regularization**.

They came up with 11 such models

- Baseline: the baseline model with standard Frobenius norm regularization.
- Heavy–F: the heavy model with standard Frobenius norm regularization.
- Light–F: the heavy model with standard Frobenius norm regularization.
- Heavy–$\ell_{2,1}$: the heavy model with $\ell_{2,1}$-norm regularization.
- Heavy–nuclear: the heavy model with nuclear-norm regularization.
- Heavy–CCL2: the heavy model with CC regularization using the $\ell_{2}$-norm.
- Heavy–CCL1: the heavy model with CC regularization using the $\ell_{1}$-norm.
- Light–$\ell_{2,1}$: the light model with $\ell_{2,1}$-norm regularization.
- Light–nuclear: the light model with nuclear-norm regularization.
- Light–CCL2: the light model with CC regularization using the $\ell_{2}$-norm.
- Light–CCL1: the light model with CC regularization using the $\ell_{1}$-norm.

The authors of the study still our working on this project.

We have training data set which has 26 observation points . Each point gives meteorological and pollutant level in the atmosphere at specific areas of Delhi.
The data comes from the website of Delhi Pollution Control Committee. We don't have that many data points but they area we have to cover is also not that much, Since we are using the same method (SVM) that is being used in Beijing we can expect a precision not more than 72%.

A variety of meteorological, traffic and industrial parameters affect the air pollution level. After taking consideration of the importance of the data, this project should use the following five features:

**$X_1$ - Temperature**

Temperature affect air quality because of temperate inversion: the warm air above cooler air acts like a lid, suppressing vertical mixing and trapping the cooler air at the surface. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground.

**$X_2$ - Wind velocity**

Wind speed plays a big role in diluting pollutants. Generally, strong winds disperse pollutants, whereas light winds generally result in stagnant conditions allowing pollutants to build up over an area. The direction of the wind also play a key role.

**$X_3$ - Relative Humidity**

Humidity could affect the diffusion of contaminant. (Rain also affects the concentration of the particulate matter).

**$X_4$ - Traffic index**

The large number of cars on the road cause high level of air pollution and traffic jam may increase the pollutants concentration from vehicles. The definition of traffic index is a index reflecting the smooth status of traffic. The index range is from 0 to 10. 0 represents smooth and 10 represents sever traffic jam.

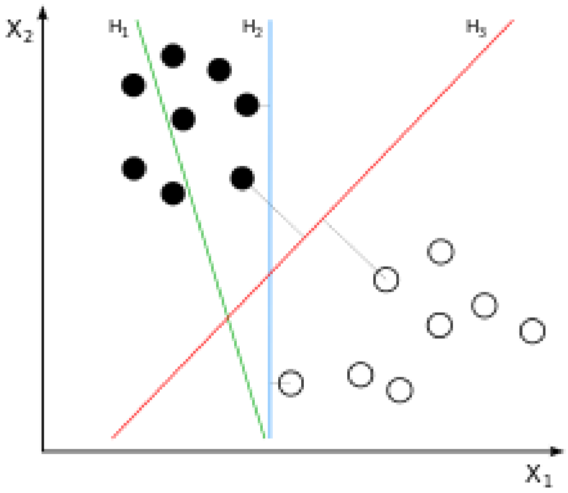
**$X_5$ - Air quality of previous day**

The air pollution level is influenced by the condition of the previous day to some extent. If the air pollution level of the previous day is high, the pollutants may stay and affect the following day.

But due to lack of availability of data in the websites that are being scraped we have currently only used wind velocity, humidity and temperature. We could however get an idea of the traffic index if we get to know how google implements the traffic density on roads (GPS for shortest time). Also if there is a computer to save everyday data from the website then the 5 th data could also be generated.

## 3. METHODOLOGY

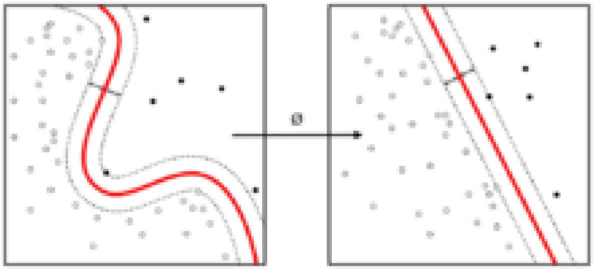
Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes (poor or good), and the goal is to decide in which class the new data point will be in. A simple and the most basic way of classifying two classes would be to draw a hyperplane between the



$H_1$ does not separate the classes.
$H_2$ does, but only with a small margin.
$H_3$ separates them with the maximal margin.

$H_3$ is the ideal hyperplane here to separate the 2 classes.

Of course sometimes the classes can’t be separated as easily by drawing a hyperplane so each point first is undergone a transformation through a function known as kernel first to transform the original dimensional space to a higher dimensional space then the hyperplane can be drawn.



Kernel Machine

The library sklearn of python takes other parameters into consideration as given above in feature selection and is much more complex when predicting the pollutant level at that point.

The meteorological and pollutant dataset is provided by dpcc live website. If more stations are installed in the future then the data predicting will be more accurate.

1) Support Vector Machine

The classifier that we have used here is more complex than the basic linear svc (discussed above) , It will not only include the location of the stations but also the meteorological dataset for parameters to predict the pollution level of a point . A python library sklearn is used to the calculations for us. It uses the some of the datasets that we have to train a model and use the rest as test datasets to see the performance / accuracy.

2) Data Scraping

This was done by using the VBA Macro feature available in Excel. Macro has been designed in such a way that the data in the sheets be updated every 1 minute by checking the data in the live website, (also when the excel sheet is opened the macro runs automatically)

3) Data Sharing

To share the data between the excel sheet to the ML code in python another library known as xlrd is used which reads all the contents of the table and gives the input to the ML Code

The models are all from Python library – sklearn , xlrd .

## 4. RESULTS

For now the data for location of the stations was unavailable and so a proper test was not done. However when we tried it out on a rough basis by implementing the algorithm on a cartesian plane and determining the AQI (color code),the accuracy was 100% whereas when quantitative analysis is done accuracy is around 70%. In the future  when mobile vans with the air quality monitoring systems and when more stations are installed it is expected that the accuracy will improve provided the data from these sources are available along with their location updates.

## 5. FUTURE WORKS

Since AQMS cant be built everywhere in Delhi. We have planned to have mobile vans built with the sensors to monitor the pollutant concentrations. Of course these vans will update data of not only pollutant concentrations but also the meteorological and its location on some time interval basis .We would also add some industrial parameters. If we could also get data on traffic from google then the traffic index parameter could also be included. Data of previous years would also come in handy to predict the pollution with more accuracy . After accomplishing the goal at hand we also want to predict the pollution in the future. For example pollution of 5th September 2019 could be determined  based on the pollution in 5th September 2018 and the days before 5th September 2019.

## REFERENCES

[1] Pandey, Gaurav, Bin Zhang, and Le Jian. "Predicting submicron air pollution indicators: a machine learning approach." Environmental Science: Processes & Impacts 15.5 (2013): 996-1005.

[2] Athanasiadis, Ioannis N., et al. "Applying machine learning techniques on air quality data for real-time decision support." First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland. 2003.

[3] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. "Classification techniques for air

quality forecasting." Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[4] M. Caselli & L. Trizio & G. de Gennaro & P. Ielpo. "A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model." Water Air Soil Pollut (2009) 201:365–377.

[5] S. Bordignon, C. Gaetan and F. Lisi, "Nonlinear models for ground- level ozone forecasting." Statistical Methods and Applications, 11, 227-246, (2002).

[6] http://www.sparetheair.com/health.cfm