CS 299: Data Rich ML for Air Quality Interpolation

Karan Gandhi, Under the supervision of Prof. Nipun Batra

February 2024

1 Problem Statement

Air pollution is one of the major environmental challenges these days. It adversely impacts public health and kills nearly seven million people worldwide. Since air quality sensors are expensive, sparse deployment and interpolation is a natural solution to predict air quality. Hence in this project, the main aim is to identify the major sources of air pollution from the PollutionMapper Model [1], monitor the concentrations levels of the components of $PM_{2.5}$ (which is a key indicator of Air Quality) in these identified areas and use this information, along with meteorological data to interpolate air quality in surrounding regions using machine learning methods.

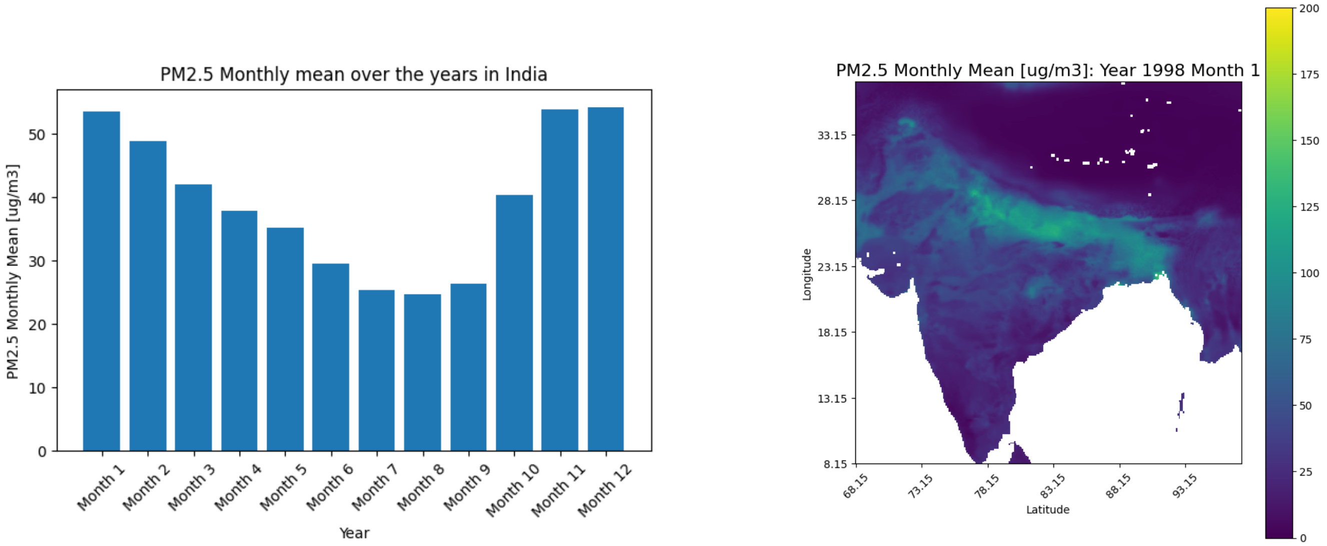
2 Current Progress

2.1 $\mathbf{PM_{2.5}}$ Data analysis

First, I started out with the $PM_{2.5}$ Data analysis of India. $PM_{2.5}$ is short for fine particulate matter, which is 2.5 microns or less in diameter. The concentration of $PM_{2.5}$ is a key indicator of air quality. I created an animation of the $PM_{2.5}$ data over India after taking the monthly snapshots from WUSTL. From here, we make the following observations:

1. The $PM_{2.5}$ levels in India follows a periodic nature.
2. The region near the Indo-Gangetic plain experiences the maximum concentration of $PM_{2.5}$. This could be because of various factors like meteorology, the planetary boundary layer height, etc.

The periodic nature can be confirmed by plotting the monthly mean of $PM_{2.5}$ over the years. It reaches its peak value around January and reaches a minimum around August (As seen in Figure 1 (a)).



(a) $PM_{2.5}$ Monthly Mean over the years in India

(b) $PM_{2.5}$ Snapshot of India on Jan, 1998

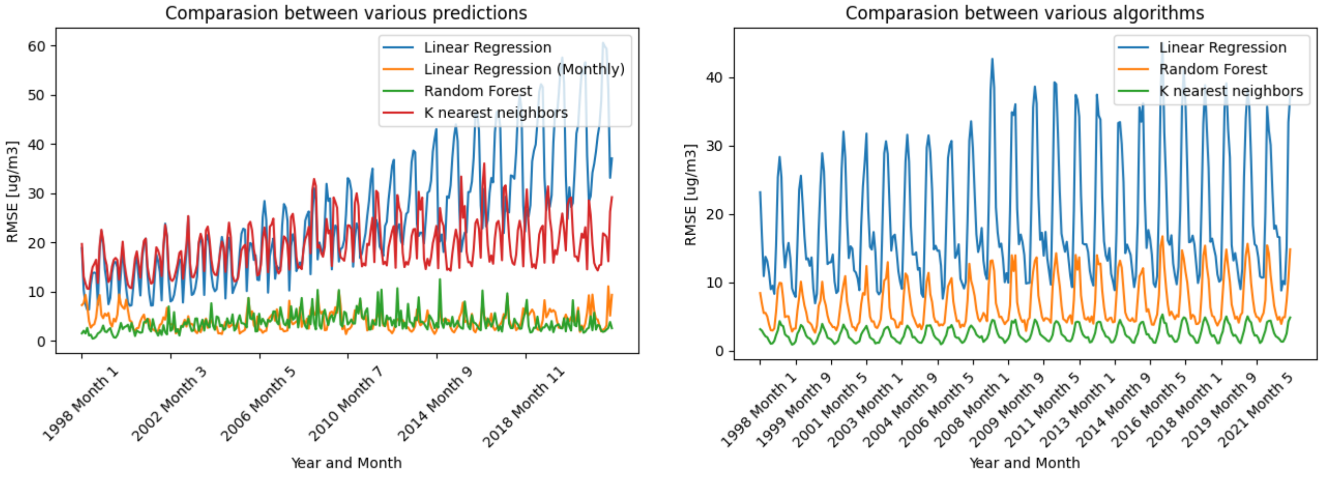
Figure 1: $PM_{2.5}$ Data analysis

2.2 $\mathbf{PM_{2.5}}$ Forecasting

After doing $PM_{2.5}$ data analysis, I try to forecast $PM_{2.5}$. For this, I train and compare 4 models: Linear Regression, Linear Regression (Monthly)[1], Random Forest (with n-estimators = 20), K Nearest Neighbours (with n-neighbors = 2). The input feature of the model is the timestamp, and each model predicts the concentration of the $PM_{2.5}$. I divide the same dataset from WUSTL into 40% train and 60% test data. You can see the final comparison between the different models in Figure 2 (a).

2.3 Spatial Prediction of $\mathbf{PM_{2.5}}$ Concentration

Here, I try to predict the spatial variation of $PM_{2.5}$. I again train and compare 3 models: Linear Regression, Random Forest (with n-estimators = 20), and K Nearest Neighbours (With n-neighbors = 8). The input features are the latitude and longitude of each grid, and the model predicts the concentration of the $PM_{2.5}$ in that grid. I again train the models on the same $PM_{2.5}$ dataset from WUSTL after dividing it into 40% train and 60% test data. You can see the final results in Figure 2 (b).



(a) $PM_{2.5}$ Forecasting Results for different models

(b) Spatial Prediction of $PM_{2.5}$ for different models

Figure 2: Predicting $PM_{2.5}$ concentrations

2.4 Other Findings

Apart from this, I also have also explored the following:

1. **Correlation of $\mathbf{PM_{2.5}}$ with $\mathbf{NO_2}$ Data:** Since one of the components of $PM_{2.5}$ is PM_{Ni} (Nitrates), which is a secondary pollutant that is formed when NO_2 gets oxidised with O_3. Hence, there might be some correlation with $PM_{2.5}$; hence, I plotted $PM_{2.5}$ with NO_2 for the California region (Los Angeles, Orange and Ventura) to find some correlation.

2. **Correlation of $\mathbf{PM_{2.5}}$ with the planetary boundary layer height:** During the winter season, at night, the air near the earth's surface gets cooler and is trapped by a layer of warmer air above it. This forms an envelope over the earth's surface, trapping all the particulate matter in the bottom layer. The height of this first envelope is known as planetary boundary layer height. Hence, after finding the dataset, I plotted the average boundary layer height and the $PM_{2.5}$ concentration for the California region from 2021 to 2022, and found that the average $PM_{2.5}$ does increase when the boundary layer height is low in the winters.

References

[1] Dhruv Agarwal, Srinivasan Iyengar, and Pankaj Kumar. "PollutionMapper: Identifying Global Air Pollution Sources". In: *ACM Journal on Computing and Sustainable Societies* 2.1 (2024), pp. 1–23.

[1] I create a Linear Regression model for each month separately