Comparing Ensemble and Linear Methods for Predicting PM2.5 Levels from AOD Data

Yuval Levental
RIT Center for Imaging Science
54 Lomb Memorial Dr, Rochester, NY 14623
yhl3051@rit.edu

Abstract

PM2.5 is defined as particulate matter which is smaller than 2.5 μm. A major challenge involving pollution detection is to measure the average PM2.5 concentration over major metropolitan areas. One useful method for predicting PM2.5 over a relatively wide range are AOD measurements. AOD is Aerosol Optical Depth, which is a measurement of how much light is reflected from satellite sensors. Terra and Aqua MODIS satellites collect AOD data on a daily basis from all around the world. The Terra satellite crosses the equator at 10:30 AM in every timezone, and the Aqua satellite crosses the equator at 1:30 PM in every timezone.

After the data was cleaned, the PM2.5 values were set as the y value, and the rest of the values were set as X values except for AOD1 since there is limited AOD1 data available. Various scikit-learn machine learning algorithms then compared the X and y values to determine the overall algorithm scores, where 1.0 is classified as a perfect score. The strongest linear correlations were found for higher values of AOD and PM2.5, as they are more noticeable. However, the highest scikit-learn scores were for the ensemble learning algorithms when compared to the linear algorithms. This is because ensemble algorithms use several weak estimators to find the best score, drawing on a far wider range of data.

1. Introduction

NASA's International Space Apps Challenge provides open source data to individuals and teams interested in solving various problems involving Earth and space phenomena. This specific project is from NASA's Space Apps 2019 Surface-to-Air (Quality) Mission. The main objective is to predict air quality that displays the most accurate data based on parameters such as location and time [1].

The sample data provided by NASA comes from Addis Ababa, Ethiopia, Delhi, India, and Los Angeles, CA, USA. However, there was not much pollution data provided from the Delhi data set, so only Addis Ababa and Los Angeles data was used.

The specific objective for this project is to predict PM2.5 from AOD data. It is known that predicting PM2.5 concentrations from ground data is difficult, as there is high variability over a relatively small location. Additionally, interactions between PM2.5 particles and other kinds of air pollution need to be taken into account.

AOD measurements cover all atmospheric levels, from space to ground level. PM2.5 particles only exist close to the ground. There needs to be some way to take into account higher levels of aerosols above ground level, which could give false correlation. Therefore, the best method to use wide-ranging AOD and PM2.5 data from different locations, taking into account the years, months, and days. This is because different kinds of correlations can be obtained, which could lead to a stronger overall correlation.

The AOD data comes from NASA's Terra and Aqua satellites. The orbits are sun-synchronous, and the orbital period for both satellites is 99 minutes. The Terra satellite, which was launched in 1999, crosses the equator at 10:30 AM in every timezone. The Aqua satellite, which was launched in 2002, crosses the equator at 1:30 PM in every timezone.

The PM2.5 data was taken from the OpenAQ Project. The OpenAQ project provides air quality parameters for various kinds of pollutants from different monitoring stations in different countries. The data points are from 2016-2019.

All of the programming was done in Python 3, using the pandas and scikit-learn libraries [7]. Pandas was used to remove missing data as needed (which had negative values), to sort and group columns, and to concatenate different files together as needed. Scikit-learn provides a function to split the data into training and testing sets, where the y-value is the PM2.5 concentration, and the X-value is most of the other values. Several kinds of models were fitted and scored on the data [2].

The AOD data was structured very differently from the

PM2.5 data, so most of the time was spent on data preparation. On average, data scientists spend 80% of their time properly organizing data [3]. By contrast, the scikit-learn libraries were very easy to import and use for all data points in this project.

2. Related Works

Much of this work is inspired by a data science project conducted by Robert Ritz, titled "Predicting Air Pollution in Ulaanbaatar, Mongolia" [4]. One of the reasons that Ritz chose PM2.5 is that these particles have the ability to enter a person's bloodstream, creating a higher risk of danger. Additionally, PM2.5 data has been collected for a longer time period. The metric he chose to predict PM2.5 levels was the Root Mean Squared Error (RMSE). This is a measure of the average distance predictions have from the ground truth.

The Ulaanbaatar project used a regression model over a classification model. This is because the results from a regression model are easier to plot, and that knowing the severity of PM2.5 concentration ranges is more important than the specific numbers. Most of the calculations were done in Python, and all of the code is open source, available online.

The relationship between PM2.5 and other factors is nonlinear, so many studies use generalized additive models (GAMs) to model the relationship between AOD and PM2.5. One large study discovered that using a GAM gave a R^2 correlation of 0.79, and that the linear R^2 was only 0.46. There were two GAM stages, and the authors assumed that the model was normally distributed. They noted that the two-stage GAM model that was created could not account for the changing PM2.5 spatial patterns with time. However, using a GAM would be too complicated for this project, so the main focus was using the best aspects of the given data [9].

Additionally, the study noted that PM2.5 is easier to predict when AOD is available. The conclusion also claimed that cloud cover and high surface reflectance are the major reasons why AOD data cannot be collected, and therefore is missing.

3. AOD and PM2.5 Datasets Introduction

All of the data was from Addis Ababa, Delhi, and Los Angeles, but only the stations from Addis Ababa and Los Angeles were used. The station data from Addis Ababa that was used were from "central" and "school", and the stations from Los Angeles that were used were "Anaheim", "Clarita", "Glendora", "North-Main", "Reseda", and "South Long Beach".

One file that was provided was the "AOD_Data.zip" file. The AOD data was produced from the NASA MODIS instruments on the Terra (morning) and Aqua (afternoon) satellites using NASA's Deep Blue and Dark Target algorithm. Files ending with MOD04 are from the morning satellite, and those ending with MYD04 are from the afternoon satellite. Each AOD file has a number of rows corresponding to best quality measurement days spanning from February 2000 for MODIS-Terra and from July 2002 for MODIS-Aqua through April 2019.

There was no header included; however, the readme file stated that the header was "YYYY, MM, DD, Latitude, Longitude, AOD1, AOD3, STD3". AOD1 is the AOD for the 10km x 10km grid point closest to the ground site with a wavelength of 550 nm. AOD3 is the average AOD for the 3×3 grid point centered on the closest grid point, and STD3 is the standard deviation for this grid.

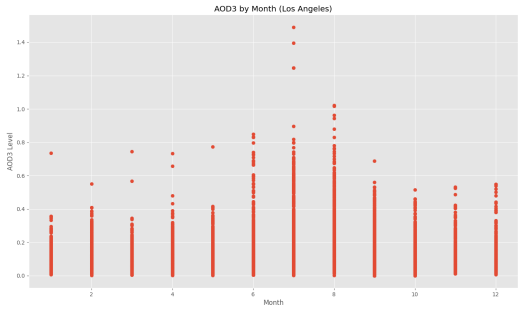

Figure 1. AOD3 levels by month in Los Angeles. As can be seen, AOD3 levels are substantially higher in summer months.

The .zip file "Reference.zip" contains measurements of various ground pollutants at each of the locations for various time periods between 2016 and 2019. These values are from the OpenAQ project. The header values are "date, parameter, location, value, unit, city, attribution, averaging-period, coordinates, country, sourcename, sourcetype, mobile". The dates both in UTC and Local zones, and the measurements are taken hourly, stored under the "value" column. The parameter value lists the type of pollution, including PM2.5. The PM2.5 values need to be separated from the other values. The coordinates contain both latitude and longitude, which need to be separated and rounded to match their respective AOD values.

Data values in these sets are equal to -1 or -999 where there is missing data. These values are placeholder values.

Most government pollution sensors are high-quality, but they are also expensive. As an example, one provider of government sensors is ThermoFisher scientific. One method that is used by their sensors is gravimetric sensing, which is performed by drawing ambient air through a filter, weighing the filter, and then calculating concentrations of particulate matter. The frequency of oscillation is dependent upon the physical characteristics of the tapered tube and the mass on its free end. Another method that is used is

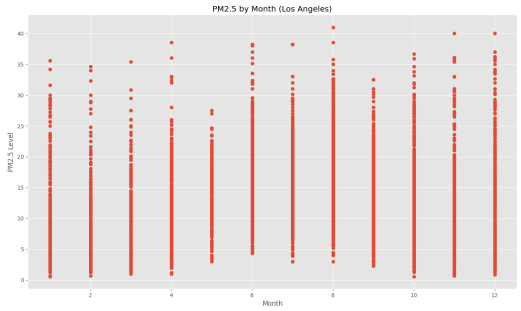

Figure 2. PM2.5 levels by month in Los Angeles. PM2.5 levels are somewhat higher in summer months, and lowest during the spring time. This is similar to the data at https://www.epa.gov/outdoor-air-quality-data

beta attenuation, which absorbs energy from beta particles as they pass through particulate matter that is collected on a filter media.[5]

4. Data Preparation Methods

After the headers were added to the AOD files, they were manually transferred to folders which contained the name of the air quality locations. All of these folders were placed in another folder titled 'City'. A for loop was created, which was run for each air quality station in a region.

To properly align the AOD files with the PM2.5 files, the YYYY, MM, and DD columns were converted into the DateTime format. The latitude and longitude columns were also rounded to three decimal places. Using a separate script, a column named "Time" was added, which was set equal to 1 for the MOD files and was set equal to 2 for the MYD files. After all of the grouping was done, the results were saved to a file titled 'test.csv'.

All the PM2.5 values were placed into a folder titled 'Citypm25'. If the local time value under the date column was less than 12, the "Time" column was set to 1. If the local time value under the date column was greater than 12, the "Time" column was set to 2. Additionally, the entire date was converted to the DateTime format.

Only the rows where the "parameter" column was equal to "pm25" were kept. Additionally, PM2.5 values that were below 1% and above 99% of all values were filtered out. To match the coordinate values of the AOD files, they were also rounded to three decimal places. The rows were grouped by the "Date" and "Time" columns so there would not be repeating date and time values. After all of the grouping was done, the results were saved to a file titled 'test2.csv'. The files 'test.csv' and 'test2.csv' were then grouped together based on "Date, Time, Longitude, Latitude", and negative PM2.5 values were filtered out.

After filtering and reorganization is complete for each data station, the year, month, day, and weekday are retrieved from the Date column using various methods. All the data was concatenated together and saved as 'test5.csv'.

5. Machine Learning Methods

scikit-learn was used for all the machine learning algorithms. The output column, y, was set equal to the PM2.5 values. The input columns X excluded the date column (which was redundant), and also excluded AOD1 and STD3, which had too many missing data values. Additionally, the data had to be split into training and testing sets. The training set comprised 70% of the data, as it needed to be significantly larger for the best possible chance of succeeding. Following from this, the testing set comprised 30% of the data [2].

The model types that were used were linear, neural, ensemble, and xgboost. The specific models that were used were Linear, Neural Network, Lasso, ElasticNet, Decision forest, Extra Trees, Boosted decision tree, and XGBoost.

The linear model uses the Ordinary Least Squares equation, which is the simplest algortihm that is utilized. The model minimizes the residual sum of squares between the testing dataset and the targets predicted by the linear approximation using the following equation:

$$\min_{w} ||Xw - y||_2^2 \tag{1}$$

Lasso and ElasticNet are variants of the linear model with added regularization terms. For the Lasso model, an additional objective is to take the minimum of the ℓ_1-norm of the coefficient vector. One possible advantage is that solutions tend to have fewer non-zero coefficients, requiring less features. The equation for Lasso is written as the following:

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha||w||_1 \tag{2}$$

By contrast, the ensemble methods combine the predictions of several estimators with a given learning algorithm in order to improve the outcomes of a single estimator. Most of the ensemble methods involve forests of randomized trees. Decision Trees (DTs) are a supervised learning method based on decision rules inferred from the data features. In a DT, each feature is partitioned into the following subsets:

$$\begin{aligned} Q_{left}(\theta) &= (x, y)|x_j <= t_m \\ Q_{right}(\theta) &= Q \setminus Q_{left}(\theta) \end{aligned} \tag{3}$$

The 'scores' of the methods are found and compared. A score is the coefficient of determination $R\hat{2}$ of the prediction. $R\hat{2}$ is defined as (1 - u/v), where u is the residual sum of squares ((y_true - y_pred) ** 2).sum() and v is the total sum of squares ((y_true - y_true.mean()) ** 2).sum(). The best possible score is 1.0 and it can be negative (because

the model can be arbitrarily worse). Additionally, the root-mean-square-errors are found. This is found by comparing the default test values to the predicted values, and then taking the square root of the results.

6. Results and Discussion

For Addis Ababa and Los Angeles, known values of AOD1 were plotted against the average values of PM2.5. Addis Ababa's data has an r-value of 0.52, but Los Angeles' data only has an r-value of 0.26. One reason is that most air quality sensors are better at detecting greater PM2.5 values, which typically exist at higher AOD levels. Additionally, PM2.5 values are much greater for the city of Addis Ababa, since Ethiopia is less technologically developed compared to the United States.

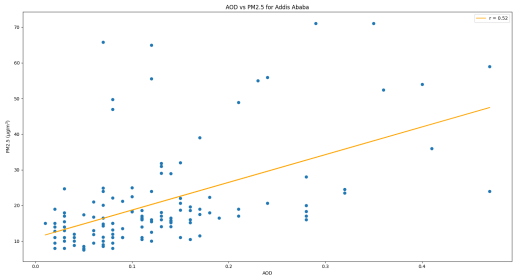

Figure 3. AOD vs PM2.5 for Addis Ababa. The correlation is better for higher AOD values.

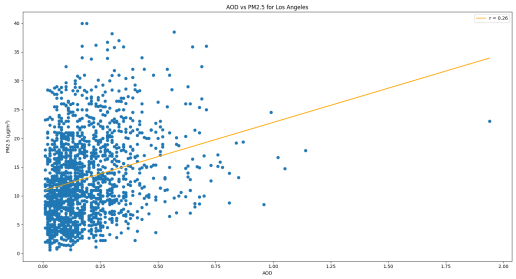

Figure 4. AOD vs PM2.5 for Los Angeles. The correlation is better for higher AOD values, though the values are more spread out than Addis Ababa.

For both Addis Ababa and Los Angeles, the Linear regression scores were relatively low. This is because AOD and PM2.5 data can be very inconsistent, as these measurements are dependent on many factors. The Linear regression score for Addis Ababa was somewhat higher because there is less data in general, implying fewer outliers.

The ensemble methods yielded much better results. They work by combining several weak estimators to produce a powerful ensemble. There are many columns with different kinds of data, which are dependent on location, time, and AOD3 values. These variables can be combined to find many different patterns, which can be synthesized into much better predictions.

Addis Ababa		
Algorithm	Score	RMSE
Linear regression	0.38	10.09
Neural network	0.008	12.77
Lasso regression	0.045	12.52
ElasticNet regression	0.05	12.49
Decision forest	0.52	8.86
Extra Trees	0.51	8.97
Boosted decision tree	0.49	9.14
XGBoost	0.47	9.33

Table 1. Algorithm scores and errors for Addis Ababa

Los Angeles		
Algorithm	Score	RMSE
Linear regression	0.09	6.35
Neural network	0.02	6.57
Lasso regression	0.0117	6.61
ElasticNet regression	0.012	6.61
Decision forest	0.48	4.79
Extra Trees	0.476	4.81
Boosted decision tree	0.46	4.90
XGBoost	0.5	4.71

Table 2. Algorithm scores and errors for Los Angeles

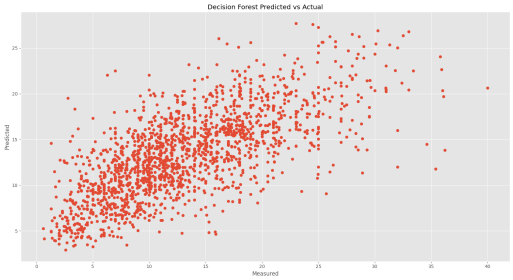

Figure 5. Measured vs predicted PM2.5 values for the Decision Forest model for Los Angeles. This is one of the better machine learning models.

One additional set of data that could help produce better correlations is MERRA2 Meterological data, which is provided by NASA. Each file contains hourly measurements of water vapor, temperature, wind speed, and trace gas concentrations. The data comes from satellite observations and atmospheric model forecasts, combined together in a "reanalysis" to provide the best estimates. This data is often utilized to improve correlations between AOD and PM2.5, or for AOD and other particulate matter data.

Additionally, as was noted before, there are many missing AOD1 and PM2.5 data points, which were indicated with negative values that are equal to or less than -1. One possible way to significantly improve overall correlations between AOD and PM2.5 is to improve the accuracy and overall availability of measurements. More air quality stations could be created with higher-quality sensors, or there could be a volunteer project where many people measure air quality from their own homes, which would be similar to the Folding@home project.

In fact, the EPA recently launched the "Air Sensor Toolbox", a project to improve the quality and availability of air sensors. One major current initiative is to research the long-term capabilities of air sensors. The Fire and Smoke Map at AirNow is now accepting data from home air quality sensors on a limited basis. The U.S. Government now loans sensors from companies such as PurpleAir and Airbeam to various individuals and groups [6].

Evidence for effectiveness of this strategy comes from results presented in a paper from China titled "Towards End-to-End License Plate Detection and Recognition: A Large Dataset and Baseline". The Chinese City Parking Dataset (CCPD) contains 250k license plate (LP) images, which is much larger than previous LP datasets. The algortihm that the authors created resulted in 98.5% accuracy over a wide variety of images, and was generally far more accurate in every category compared to standard LP detection algorithms [8].

7. Conclusion

The main purpose of this project was to learn how to predict PM2.5 levels based on AOD and other parameters. Additionally, learning how to perform data preparation with Python and pandas was an important aspect of this project, as the data had to be efficiently sorted and organized. All of the data was taken from NASA's Space Apps Challenge. The machine learning models were from the scikit-learn library.

The most effective algorithms are ensemble methods, which are based on decision trees. A diverse set of classifiers is created from these methods. They are more effective because they use more patterns compared to linear learning methods. What matters most is the number of data points and the number of types of data that correlate with PM2.5 levels.

Data preparation was the most difficult aspect, since different datasets had different headers and used different notations. PM2.5 data was more stable over time at higher values. At lower values, the numbers varied greatly from day to day. Hopefully, when data collection methods dramatically improve, AOD data will be a far more useful prediction method in the future since AOD readings are based off a very wide area from satellites.

8. Acknowledgements

Many thanks to Dr. John Kerekes of the RIT Center for Imaging Science for suggesting this project and providing guidance on data preparation and analysis. Additional thanks go to data scientist Robert Ritz for helping with machine learning algorithms, which are from an air quality project he worked on in Mongolia.

References

[1] NASA. Surface-to-air (quality) mission, 2019.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Emre Rençberoğlu. Fundamental techniques of feature engineering for machine learning, Apr 2019.

[4] Robert Ritz. Predicting air pollution in ulaanbaatar, mongolia - part i, introduction, Jun 2018.

[5] ThermoFisher. Air quality analysis information - us.

[6] ORD US EPA. Air sensor toolbox, Apr 2016.

[7] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[8] Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, Huan Huang, Changchun Ying, and Liusheng Huang. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, page 261–277. Springer International Publishing, 2018.

[9] Liu Yang, Paciorek Christopher J., and Koutrakis Petros. Estimating regional spatial and temporal variability of pm2.5 concentrations using satellite data, meteorology, and land use information. *Environmental Health Perspectives*, 117(6):886–892, Jun 2009.