Air Pollutant Prediction Using Precipitation

Patrick Chang
Henry M. Gunn High School, 780 Arastradero Rd, Palo Alto, CA 94306

Abstract
Elevated air pollutants impact human health in various ways. Ozone, Nitrogen Dioxide, Sulfur Dioxide, and PM_{10} can aggravate symptoms of pre-existing lung diseases. Carbon Monoxide is potentially lethal. Younger children exposed to lead are subject to lowered IQ and mental issues. Hence, an efficient air pollutant prediction system is key to public health. It is believed that air pollutant levels are affected by the amount of precipitation. For example, levels of pollutants will decrease when precipitation increases. By studying the relationship between precipitation and air pollutants, scientists may save time and money by only measuring precipitation to predict air pollutants. In this paper, we studied whether we can infer the levels of ozone, Carbon Monoxide, Nitrogen Dioxide, Sulfur Dioxide, PM_{10} particulates, and lead from precipitation. Regression analysis tools including simple linear regression and nonlinear regression methods such as neural networks are used in this study. First, we studied whether the air pollutant levels could be predicted by precipitation alone. Second, the wind factor was also included for improving air pollutant prediction.

Key Words: air pollutant, prediction, linear regression, neural networks

1. Introduction

Air pollutants have harmful effects on the whole human body. Ozone, Nitrogen Dioxide, Sulfur Dioxide, and PM_{10} can cause various symptoms such as chest tightness, coughing, difficulty breathing, and the aggravation of previously existing symptoms from other lung diseases. Pure Carbon Monoxide is odorless and will cause nausea, vomiting, dizziness, higher risk of heart disease, and headaches, and is potentially lethal at higher concentrations, as hundreds of people die in the United States from accidental Carbon Monoxide poisoning. Lead is able to be distributed around the body fairly easily, and therefore harms many different systems along with an increased risk of heart disease. Younger children exposed to lead will have a lowered IQ, behavioral issues, and trouble learning. Hence, efficient air pollutant prediction is key to public health.

Air pollutants considered under this study include ozone, carbon monoxide, nitrogen dioxide, sulfur dioxide, PM_{10} particulates, and lead. With the increase of the amount of precipitation, levels of all air pollutants listed will decrease. Nonetheless, there are many other factors that affect the levels of the various pollutants, such as humidity, air pressure, precipitation, wind speed, temperature, and the varying amounts of pollutants released by human activities or nature (See Figure 1). For example, PM pollution peaks between 50% to 55% humidity and about 15° Celsius. PM pollution also increases once air pressure reaches or past 900 hectopascals, but decreases as wind speed increases to 3 meters per second, and continually decreases until there is 1500 mm of annual precipitation [1]. The relations between these factors and pollutant levels are in general non-linear.

The topic of this study was inspired by an experience of a substantial increase in $PM_{2.5}$ in the vicinity of the Bay Area due to a fire in far northern California, which caused discomfort for many. Nonetheless, the levels of air pollutants decreased significantly right after a rain fall. The intention of this study is to find which natural and predictable factor affects pollutant levels most significantly and find an accurate way to use multiple factors to predict pollutant levels. In particular, we examine the feasibility of predicting the levels of air pollutants via average daily precipitation in inches and wind speed in knots.

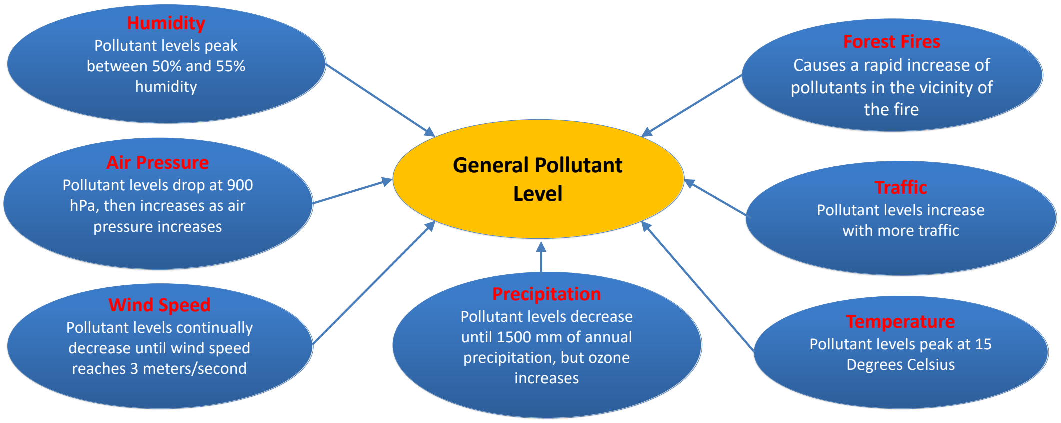

Humidity
Pollutant levels peak between 50% and 55% humidity
Air Pressure
Pollutant levels drop at 900 hPa, then increases as air pressure increases
Wind Speed
Pollutant levels continually decrease until wind speed reaches 3 meters/second
General Pollutant Level
Precipitation
Pollutant levels decrease until 1500 mm of annual precipitation, but ozone increases
Forest Fires
Causes a rapid increase of pollutants in the vicinity of the fire
Traffic
Pollutant levels increase with more traffic
Temperature
Pollutant levels peak at 15 Degrees Celsius

Figure 1: Air Pollutant Factors

2. Methods

2.1 Data

Daily total precipitation data [2] of Portland, Oregon and Seattle, Washington along with the pollutant data [3] from 2013 to 2016 were used for linear regression. Daily precipitation, average wind speed, and pollutant data of San Jose, California also from 2013 to 2016 were used for multiple linear regression and neural networks. Precipitation is measured in millimeters (mm), wind speed is measured in meters per second (m/s), carbon monoxide is measured in parts per million (ppm), lead is measured in micrograms per meters cubed ($\mu g/m^3$), nitrogen dioxide is measured in parts per billion (ppb), ozone is measured in parts per million, PM_{10} is measured in micrograms per meters cubed ($\mu g/m^3$), and sulphur dioxide is measured in parts per billion.

2.2 Linear Regression

Linear regression was used initially to find how correlated precipitation was to the pollutant levels of all six pollutants on the next day, to provide insight on whether precipitation was an effective factor for prediction. Linear regression assumes that precipitation is linearly correlated to the pollutant levels. The pollutant levels can be predicted by the following equation: $pollutant = w \cdot precipitation$, where w is a scalar number.

2.3 Multiple Linear Regression

Since pollutant levels are not just impacted by precipitation alone, multiple linear regression was used with the assumption that the pollutant levels are linearly correlated to multiple independent factors. In my study, both wind and precipitation are assumed to be linearly correlated to pollutant levels. Based on the multiple linear regression, the pollutant levels are predicted by the following equation:

$$pollutant = w_1 \cdot precipitation + \ w_2 \cdot wind,$$

where w_1 and w_2 are both scalar numbers.

Multiple linear regression was also used to compare the correlation between precipitation and pollutant levels and the correlation between wind and pollutant levels, after discovering that precipitation may not be the best factor for prediction of pollutant levels.

2.4 Neural Networks

Since the correlation between precipitation, wind, and pollutant levels are not linear [4], neural networks were used to apply non-linear fits to predict pollutant levels of the next day more accurately. Figure 2 models the pollutants using precipitation and wind speed. Here the multilayer perceptron neural network is used to model the pollutant levels using inputs of precipitation and wind speed. The network consists of two hidden layers. There are 10 nodes in the first hidden layer and 5 nodes in the second hidden layer. The output of the neural network is the predicted pollutant level at the next day. The RELU activation function is used in this study.

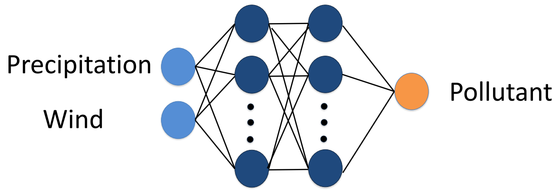


Figure 2: Neural Network Design

3. Results

3.1 Linear Regression

Figure 3 shows the linear regression models [5] of the pollutants over precipitation using data collected in Seattle, Washington. From top left to top right are carbon monoxide, lead, and nitrogen dioxide respectively. From bottom left to bottom right are ozone, PM_{10}, and sulphur dioxide respectively. Carbon monoxide and precipitation have very little negative linear correlation, with an R^2 of 0.000842. Lead and nitrogen dioxide are slightly more negatively correlated with an R^2 of 0.011713 and 0.011198 respectively. Contrary to the hypothesis, ozone is positively correlated to precipitation with an R^2 of 0.001129. PM10 and sulphur dioxide are both negatively correlated with R^2 values of 0.032429 and 0.02232 respectively. Small R^2 values indicate insignificant correlations between the air pollutants and precipitation.

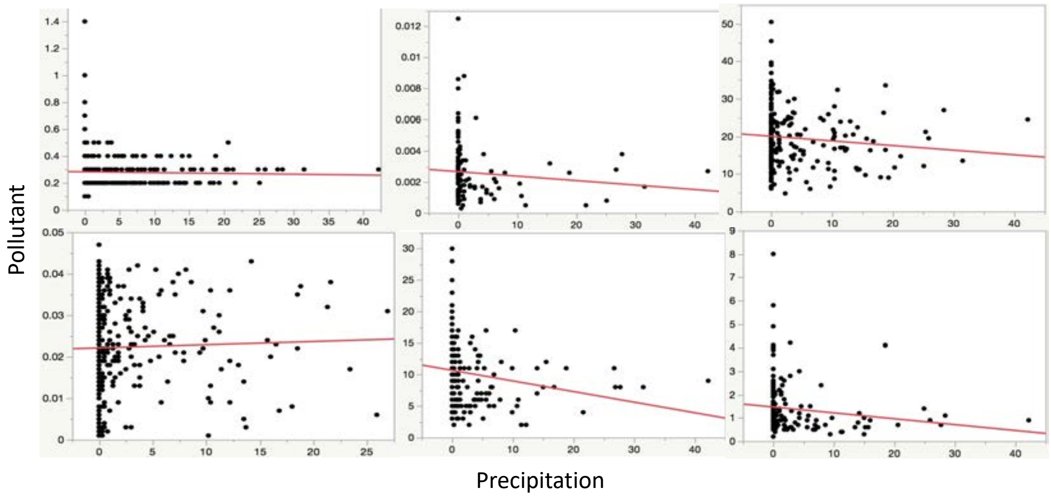
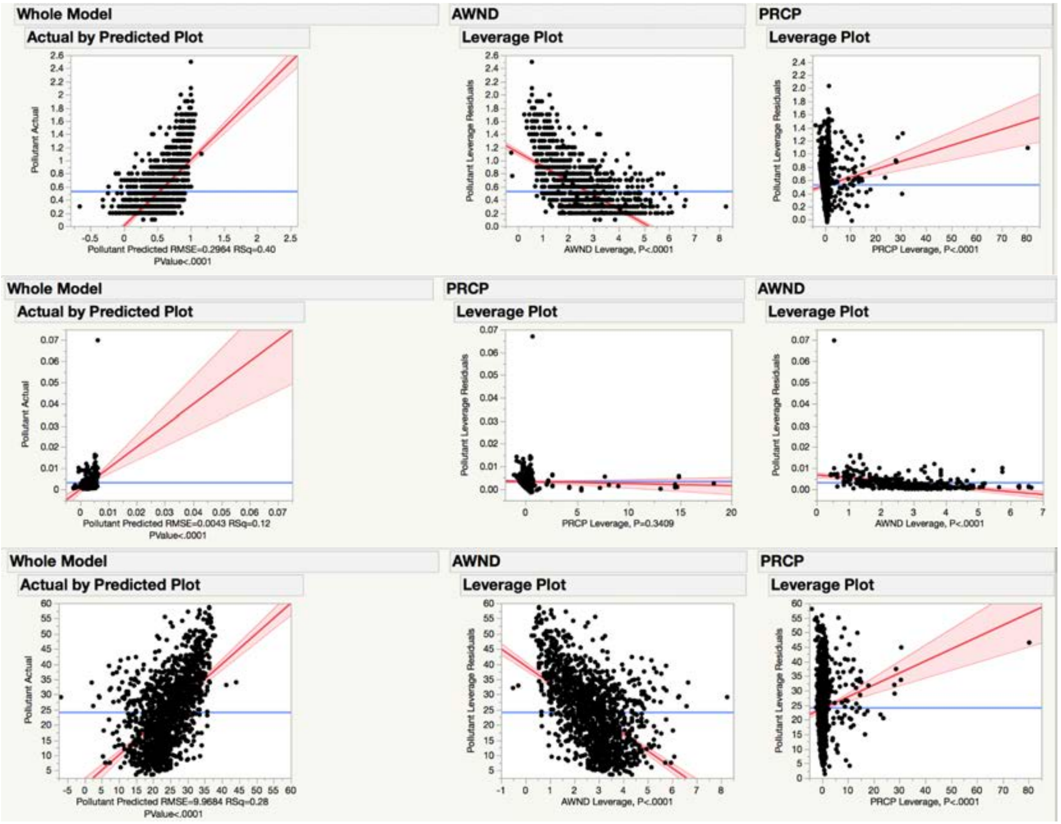


Figure 3. Linear Regression of Pollutants over Precipitation

3.2 Multiple Linear Regression

Figure 4 shows the multiple linear regression models [5] of precipitation and wind speed to pollutants using data collected in Santa Clara, California. From top to bottom are carbon monoxide, lead, nitrogen dioxide, ozone, PM_{10}, and sulphur dioxide respectively. Table 1 shows the R^2 values of the multiple linear regression models. Figure 5 shows the comparison between the predicted pollutant levels of the next day and the actual pollutant levels of the next day.



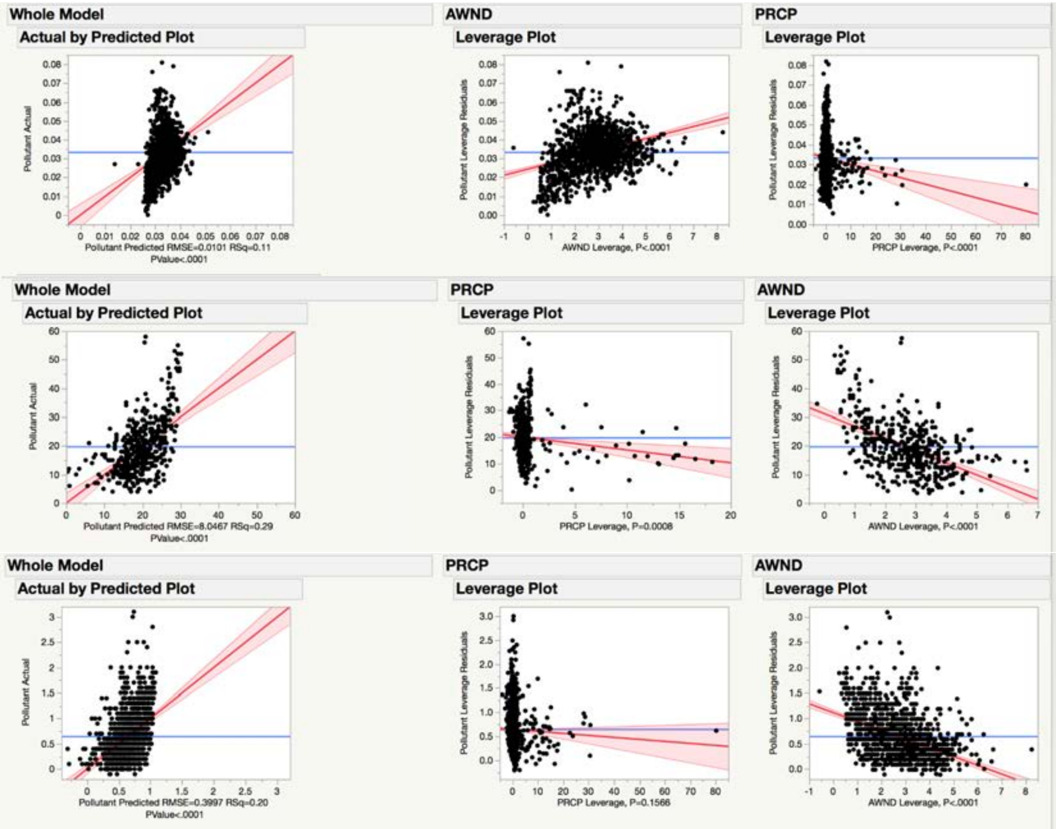


Figure 4. Multiple Linear Regression of Pollutants

Table 1. R^2 values of Multiple Linear Regression Models

	CO (ppm)	Lead (µg/m^3)	NO_2 (ppb)	Ozone (ppm)	PM10 (µg/m^3)	SO_2 (ppb)
R Square	0.396	0.122	0.277	0.11	0.29	0.202

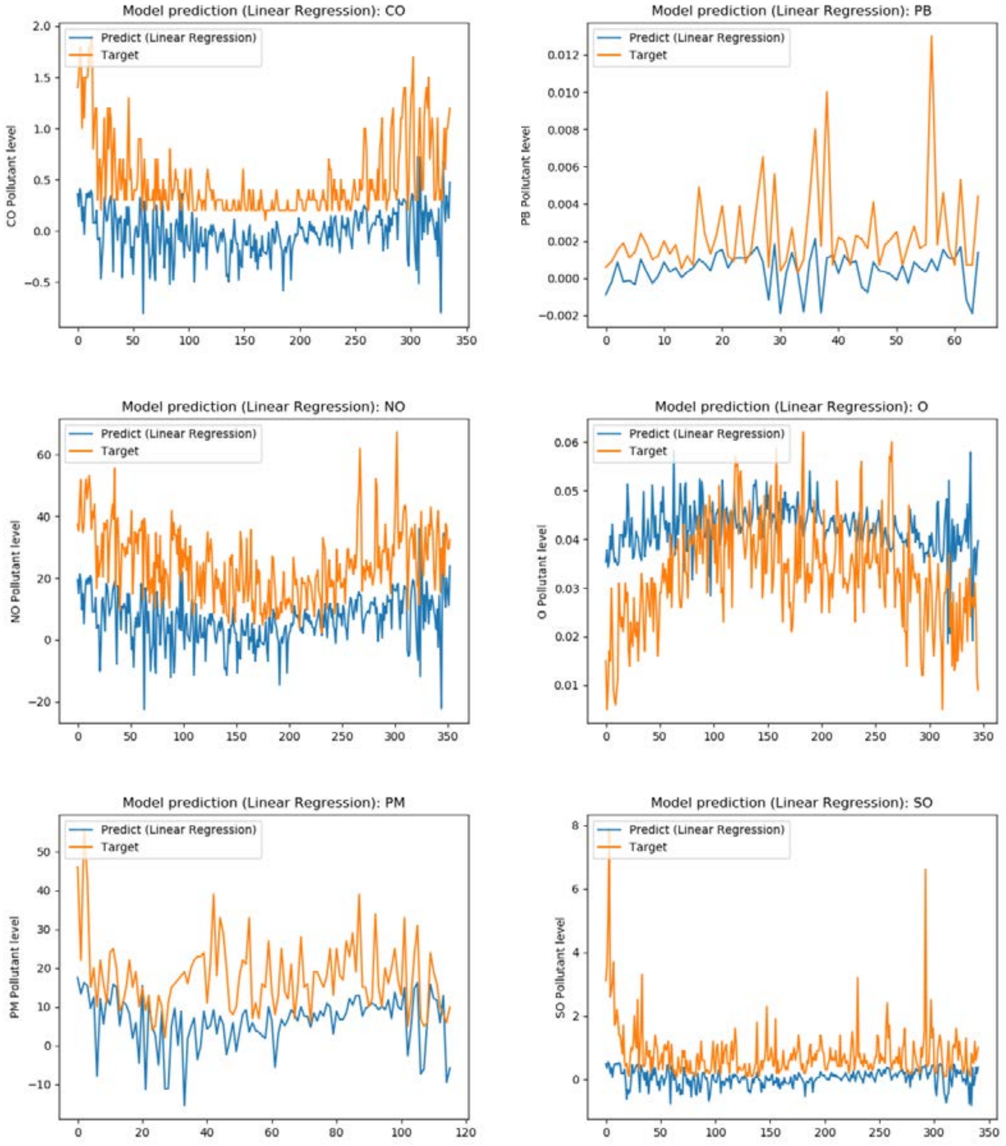


Figure 5. Prediction Performance of Testing Data (Multiple Linear Regression)

3.3 Neural Networks

Figure 6 shows the predicted values of the pollutants using precipitation and wind speed with neural networks compared to the actual values of the pollutants of the next day [6]**.** Table 2 compares the root mean square error between the predicted values and the actual values using neural networks and multiple linear regression.

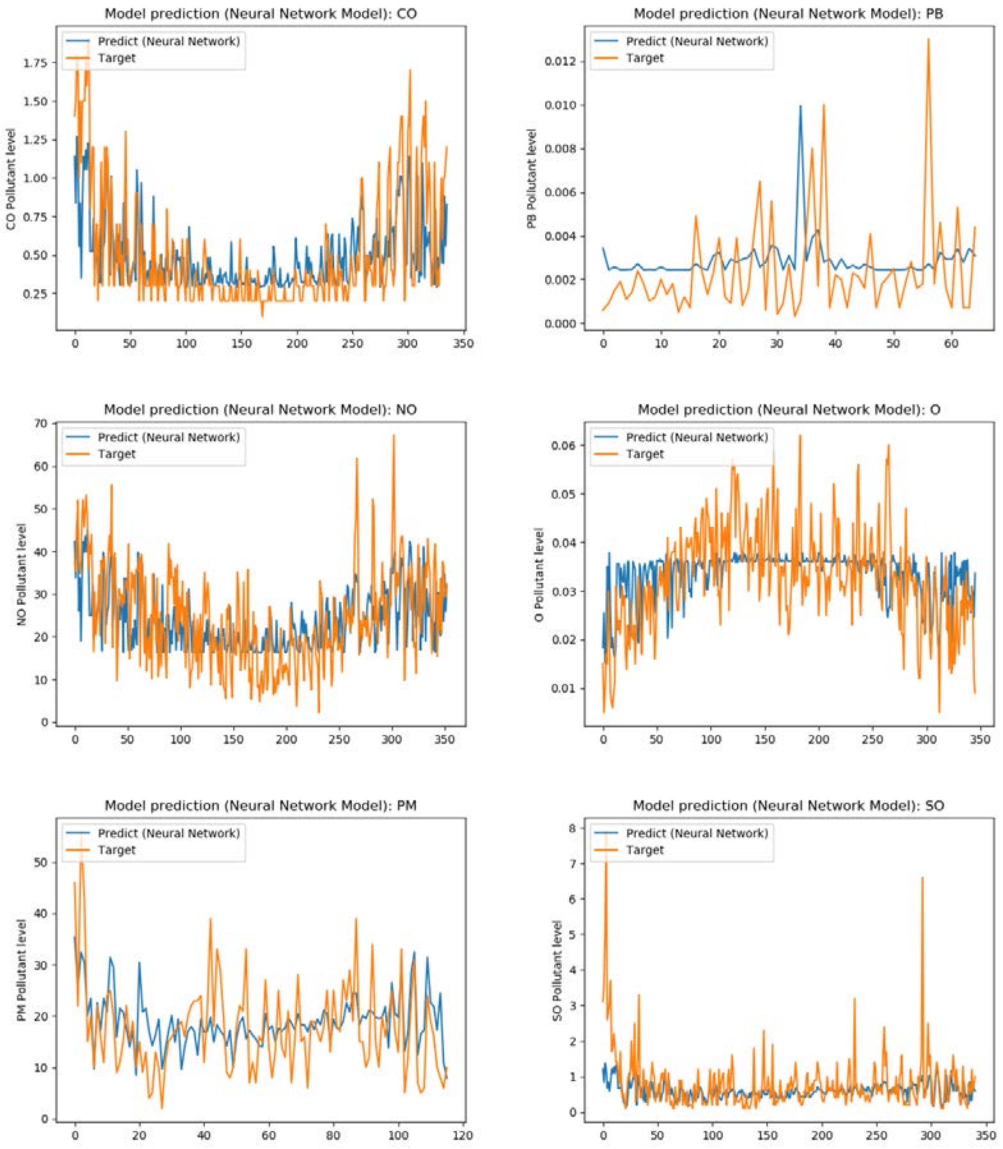


Figure 6. Prediction Performance of Testing Data (Neural Network)

Table 2: Root Mean Square Error Comparison

	CO (ppm)	Lead (μg/m³)	NO_2 (ppb)	Ozone (ppm)	PM10 (μg/m³)	SO_2 (ppb)
Neural Networks RMSE	0.257	0.0025	9.34	0.0095	7.53	0.75
Linear Regression RMSE	0.607	0.0028	21.13	0.014	13.96	1.05

4. Conclusion

It is not simple to predict air pollutant levels of the next day as there are many unpredictable factors, such as wildfires or human activities. Using precipitation to predict pollutant levels proves to be insufficient as precipitation is not very correlated to pollutant levels, disproving the original hypothesis, and rain can be very infrequent in certain areas. There are also more factors that can be measured, which may also have a major impact on pollutant levels. Wind speed has more of an impact to pollutant levels than precipitation, so it is more suitable for pollutant prediction. In this study, neural networks usually predict pollutant levels more effectively than bivariate linear regression, as proven by the smaller root mean square error values. Neural networks allow for the use of non-linear fits, leading to more accurate predictions with more flexibility to various situations. However, even using precipitation and wind is not enough as both bivariate linear regression and neural networks could not predict more extreme spikes in pollutant levels. Future research will extend this study to include other factors for predicting pollutant levels.

5. Acknowledgements

This study owes its inspiration and early editing to the STEAMS organization.

6. References

[1] Xu, W. Y., Zhao, C. S., Ran, L., Deng, Z. Z., Liu, P. F., Ma, N., ... & Yu, J. (2011). Characteristics of pollutants and their correlation to meteorological conditions at a suburban site in the North China Plain. Atmospheric Chemistry and Physics, 11(9), 4353-4369.

[2] National Centers for Environmental Information, https://www.ncdc.noaa.gov/cdo-web/search. (Last access on September 8, 2019)

[3] U.S. Environmental Protection Agency, https://www.epa.gov/outdoor-air-quality-data/download-daily-data. (Last access on September 8, 2019)

[4] Aldrin, Magne, and Ingrid Hobæk Haff. "Generalised additive modelling of air pollution, traffic volume and meteorology." Atmospheric Environment 39.11 (2005): 2145-2155.

[5] JMP®, SAS Institute Inc., Cary, NC, 1989-2019.

[6] Chollet, F. (2015) keras, https://github.com/fchollet/keras.