Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters

**Abstract**

Outdoor air pollution costs millions of premature deaths annually, mostly due to anthropogenic fine particulate matter (or PM2.5). Quito, the capital city of Ecuador, is no exception in exceeding the healthy levels of pollution. In addition to the impact of urbanization, motorization, and rapid population growth, particulate pollution is modulated by meteorological factors and geophysical characteristics, which complicate the implementation of the most advanced models of weather forecast. Thus, this paper proposes a machine learning approach based on six years of meteorological and pollution data analyses to predict the concentrations of PM2.5 from wind (speed and direction) and precipitation levels. The results of the classification model show a high reliability in the classification of low (<10 µg/m3) versus high (>25 µg/m3) and low (<10 µg/m3) versus moderate (10–25 µg/m3) concentrations of PM2.5. A regression analysis suggests a better prediction of PM2.5 when the climatic conditions are getting more extreme (strong winds or high levels of precipitation). The high correlation between estimated and real data for a time series analysis during the wet season confirms this finding. The study demonstrates that the use of statistical models based on machine learning is relevant to predict PM2.5 concentrations from meteorological data.

**Introduction**

* 98% of cities in low/middle income countries do not meet WHO guidelines
* 3.3 million premature deaths worldwide
* Depth of mixing layer depends on solar radiation and temperature
  + Shallower mixing depth, less dilution of emissions
* Wind speeds, relative humidity, and precipitation all have an effect
* Three current methods: statistical, chemical transport, and machine learning
  + Chemical transport requires updated source list that is difficult to produce, topography complicates chemical transport
  + ANN is most accurate
* Uses three variables (wind speed, wind direction, and precipitation)
* Uses Boosted Trees and Linear Support Vector Machines

**Methods**

* Quito is under direct sunlight (equator) and is on a long plateau near a volcano, due to complex terrain wind speed is almost unpredictable
* Jun 2007-Jul 2013, weekend values removed
* To obtain general trends, data used to generate convolutional based spatial representations
  + Strong winds result in low PM2.5 concentrations, generally come from a similar direction
* Used Boosted Trees (BT) and Linear Support Vector Machines (L-SVM)
  + BT combines simple rules to create classification algorithm, where each misclassified data point gains weight

**Results**

* Two classes are used (above 15 ug/m3 and below), yields high difference in performance between two sites
* ROC curve used to evaluate binary classifier
  + Plots true negatives vs. true positives
* Three classes yield that values of 10–25 μg/m3 and >25 μg/m3 are not influenced by meteorological parameters
* Made a neural network as well
* RMSE graph suggests that prediction of PM2.5 is more reliable for extreme than for moderate climactic conditions
  + Meteorological inputs do not accurately describe PM2.5 levels if they are more than 20 μg/m3

**Conclusions**

* Regression analysis shows that prediction is possible for levels less than 20 ug/m3
  + Accuracy improved in conditions of strong winds and high precip
* Future research includes hybrid between machine learning and chemical transport models
* Festivals, wildfires, accidents could explain PM2.5 levels exceeding 20 ug/m3

**Citation**

Deters, J. K., Zalakeviciute, R., Gonzales, M., & Rybarczyk, Y. (2017) Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *Journal of Electrical and Computer Engineering*. https://doi.org/10.1155/2017/5106045

**Purpose**

I will be using this research on my background research paper, as well as to supplement the background information section of my final research paper. Also, this gives me important information on how to do binary classifications, if I decide to do so by setting healthy/unhealthy limits for different pollution parameters (PM10, PM2.5, etc.)