Predicting NO2 concentrations

This manuscript (permalink) was automatically generated from tessac2/498_NO2_pred@f0597cc on December 6, 2020.

Authors

- **Tessa Clarizio**
  · tessac2

- **Jane Roe**
  · janeroe

- **Jane Roe**
  · janeroe

- **Jane Roe**
  · janeroe

- **Jane Roe**
  · janeroe

- **Jane Roe**
  · janeroe

Abstract

test commit on abstract

1. Introduction

NO_2 is defined by the U.S. Environmental Protection Agency (EPA) as a criteria air pollutant, meaning it poses a risk to human and environmental health. The primary National Ambient Air Quality Standard (NAAQS) for NO2 is set at a 53 ppb annual average [1]. NO2 can cause respiratory irritation and can aggravate respiratory diseases such as asthma (US EPA, n.d., B). NO2 can also react with other chemicals in the atmosphere to form both particulate matter (PM) and tropospheric ozone (US EPA, n.d., B). PM and ozone are also criteria air pollutants and are harmful to human health. NO2 also contributes to the formation of acid rain, smog, and nutrient pollution in coastal waters (US EPA, n.d., B). The primary source of NO2 emissions is fossil fuel combustion, particularly from traffic and power plants (US EPA, n.d., B).

Therefore, understanding and predicting the spatial variability of NO2 emissions is of great importance to public health. However, prediction of air quality can be complicated due to the number of factors that affect local air quality, ranging from meteorology to land use. Machine learning models are a useful tool to interpret and find relationships in complex data.

[introduce Bechle study...] Bechle et al (2015) explores the impact of.. [Grace please add here]

This report proposes a machine learning model to predict NO2 concentrations spatially. First, a literature review was undertaken to understand what machine learning models have typically performed well in predicting air quality. Next, an exploratory data analysis (EDA) was performed on the Bechle et al (2015) dataset. Finally, multiple linear regression, neural network and random forest models were built and results were compared to see which method had the lowest mean-squared error (MSE).

2. Methods

2.1 Literature Review

There are a number of studies examining how machine learning models can be used to predict air quality. Seven studies were examined as part of this literature review, and can be broadly categorized into 2 areas: predicting PM2.5 and predicting the Air Quality Index (AQI)/ Air Pollution Index (API). One exception is that one of the studies examining AQI also predicted NOx concentrations.

2.1.1 PM2.5

Chen et al (2018) explored the use of random forest models to predict PM2.5 concentrations spatially in China and compared them to multiple linear regression and generalized additive models. Random forest models are non-parametric learning algorithms, and have been shown to have high accuracy. While the study began with a large number of predictors, these were narrowed down to ground-based measurements, satellite retrieved AOD data, urban cover data and meteorological data.The random forests model had the greatest predictive power of all the models considered, with a RMSE of

28.1 µg/m3 on a daily scale (R2 = 83%), improving to 10.7 µg/m3 (R2 = 86%) and 6.9µg/m3 (R2=86%) on monthly and annual time-scales, respectively.

Xu et al (2018) likewise considered a number of machine learning models for PM2.5 prediction in British Columbia, Canada. 8 models were examined in this study: 1) multiple linear regression (MLR), 2) Bayesian Regularized Neural Networks (BRNN), 3) Support Vector Machines with Radial Basis Function Kernel (SVM), 4) Least Absolute Shrinkage and Selection Operator (LASSO), 5) Multivariate Adaptive Regression Splines (MARS), 6) Random forest (RF), 7) eXtreme Gradient Boosting (XGBoost), and 8) Cubist.The predictors included humidity, temperature, albedo, normalized difference vegetation index (NDVI), height of the planetary boundary layer (HPBL), wind speed, distance to the ocean, elevation, and calendar month beside the ground level monthly averaged PM2.5 data collected from 63 stations between 2001 to 2014 as well as 3km resolution AOD data from MODIS. This study found that the cubist model had the highest accuracy (RMSE =2.64 microg/m3 and R2=0.48) and the the MLR had the lowest accuracy (MSE = 3.24 µg/m3 and R2=0.22). The predictors with the most influence were monthly AOD and elevation.

Enebish et al (2020) considered 6 different machine learning models for PM2.5 prediction in Mongolia: 1) RF, 2) gradient boosting, 3) support vector machine (SVM) with a radial basis kernel, 4) multivariate adaptive regression splines (MARS), 5) generalized linear model with elastic net penalties (a type of MLR), and 6) generalized additive model. These models were run for annual data, cold season and warm season. Parameters considered were air pollution monitoring data, meteorology, land use and population. Across all time periods, the RF had the best R2 and RMSE values. Over the entire period using the hold-out test set, RF had a RMSE of 12.92 (R2 = 0.96), and the cold season and warm season had RMSE of 21.23 (R2 = 0.92) and 7.44 (R2 = 0.84), respectively.

A common limitation of all three studies is the volume of missing data. In Chen et al (2018), the model had only two years of ground-based measurements to train the model on (2014-2016), and then predicted PM2.5 concentrations for a ten year period (2005 to 2014). Xu et al, 2018 also discussed the challenge of missing data, averaging hourly and daily measurements where available to monthly concentrations to use in model development. Finally Enebish et al, 2020 discussed there being few air quality monitoring stations and insufficient data to well represent the high seasonal variability of PM2.5 concentrations.

Additionally, all studies considered meteorology when constructing the machine learning model. The dataset in our study does not include meteorology, potentially leaving out an important predictive factor.

2.1.2 AQI/API

Azid et al (2014) used a multilayered perceptron feed-forward artificial neural network model to predict API, using daily measurements of NO2, SO2, CO, PM10 and O3 over a period of 7 years in Malaysia. The best RMSE and R2 occurred when the hidden nodes were set to 6, and were 0.618 and 10.017, respectively.

Gu et al (2020) focuses on predicting the AQI in Shenzen, China. The dataset consists of 365 sets of daily pollution data over one calendar year (2018), and the purpose was to develop a model to predict AQI. Pollution measurements included PM2.5, PM10, SO2, CO, NO2 and O3. Unlike other studies, Gu et al 2020 did not take into account meteorological factors and related urban characteristics. Two SVM models were developed: smart adaptive particle swarm optimization and particle swarm optimization, SAPSO-SVM and PSO-SVM, respectively. Additionally, a back propagation (BP) neural network model was developed. SAPSO-SVM had a test set classification accuracy of 91.62%, and PSO-SVM 88.56%. For the BP-neural network model, ten iterations of the algorithm best fit the test data set, where the percent error ranged from 18.41% for PM2.5 to 30.29% for SO2. While Gu et al stated that both models were a good fit for the data, by using different statistical comparisons to explain model fit, it is not clear which of the two models has a better predictive ability, although it appears to be SAPSO-SVM. The paper listed a number of limitations associated with its neural network, particularly the limited data points.

Liu et al (2019) developed SVM and RF models to predict hourly AQI in Beijing, China and hourly NOx in an Italian city. Parameters included historical hourly averaged AQI concentrations for PM2.5, O3, SO3, PM10 and NO2 in Beijing (five years), and hourly averaged responses for CO, non-methane hydrocarbons, benzene, NOx and NO2 in the Italian city (1 year). The SVM performed better in predicting AQI with a RMSE 7.666 (R2=0.9776), but the RF model performed better in predicting NOx concentrations (RMSE = 83.67, R2 = 0.8401).

Singh et al (2013) used ensemble learning methods to predict air quality index in Lucknow, India. They trained four different models: single decision tree (SDT), decision tree forest (DTF), decision treeboost (DTB) and SVM. While decision trees can be different from random forest, it appears in Singh's methodology that the DTF and DTB involve randomization with replacement from the training dataset to create separate models, which are then used to predict the entire data from the subsets. This is consistent with RF models–essentially RF are ensemble decision trees. The parameters included in the model are 5 years of data on: daily air quality measurements (SO2, NO2, SPM and RSPM) meteorology (air temperature,T(C), relative humidity, RH (%), wind speed, WS(km h-1), evaporation (mm), and daily sunshine period, SS (h)). The DTF and DTB models outperformed the SVM models. DTB performed the best, with a RMSE of 4.38 (R2 = 0.92).

Similar to the papers examining PM2.5 concentrations some of these papers also discussed the limitations due to limited data. For example, Gu et al (2020) only used pollution data from one year within one region in China, and Liu et al (2019) used pollution data from one Italian city over one year.

However, unlike the PM2.5 studies, many of the API/AQI studies were also interested in classification, and determining which pollutant contributes the most to API/AQI, rather than examining the role meteorology or other factors have on predicting the ambient concentrations of a certain pollutant. This means these studies were more likely to pick areas with suitable air pollutant concentration data to detect patterns between a particular pollutant and AQI.

Another difference from the PM2.5 studies is that only one AQI study included meteorology in the parameters of the model (Singh et al, 2019). This is because these studies were looking for patterns between ambient concentrations of certain pollutants and API/AQI, rather than necessarily predicting air quality in relation to other environmental factors.

2.1.3 Comparison of PM2.5 and AQI/API studies

The main difference between the PM2.5 and the AQI studies is that studies examining PM2.5 tended to only examine one pollutant, whereas AQI studies consisted of measuring and modeling a number of different pollutants. Therefore, some AQI models were more interested in classification than predicting a specific pollutant spatially or temporally. As a result, different parameters tended to be included in the model depending on if it was predicting PM2.5 or AQI. Additionally, different models tended to perform best depending on the target prediction.

The models in each of these studies is summarized in Table 2.1 below:
Table 2.1 |PM2.5|Both PM2.5 and AQI| |—-|——————| |**MLR** (Xu et al, 2018; Enebish et al, 2020; Chen et al, 2018) | **RF** (Chen et al, 2018; Xu et al, 2018; Singh et al, 2013; Liu et al, 2019; Enebish et al, 2020)| | **LASSO** (Xu et al, 2018) | **Neural Network** (Azid et al, 2014; Xu et al, 2018, Gu et al, 2020) | | **MARS** (Xu et al, 2018; Enebish et al, 2020) | **SVM** (Xu et al, 2018; Gu et al, 2020; Liu et al, 2019; Enebish

et al, 2020; Singh et al, 2013) | | **Gradient Boosting** (Xu et al, 2018; Enebish et al, 2020) | | | **Cubist** (Xu et al, 2018) | | | **Generalized additive model** (Enebish et al, 2020; Chen et al, 2018)| | | **Mixed effects models** (Chen et al, 2018) | |

As we can see above more models were used to predict PM2.5 than AQI, and the ones that were used in AQI studies were also used in predicting PM2.5. The best-predicting models in each study are shown in Table 2.2, alongside their RMSE and R2 values.

Study	Target Prediction	Best Model	RMSE	R2
Chen et al (2018)	Annual average PM2.5	Random Forest	6.9	0.86
Xu et al (2018)	Monthly average PM2.5	Cubist	2.6	0.48
Enebish et al (2020)	Annual average PM2.5	Random Forest	12.9	0.96
Azid et al (2014)	Daily AQI	Neural network	10.0	0.62
Gu et al (2020)	Daily AQI	SVM	n.a.	n.a
Singh et al (2013)	Daily AQI	Random Forest	4.4	0.92
Liu et al (2019)	Hourly AQI	SVM	7.7	0.98
Liu et al (2019)	Houly NOx	Random Forest	83.7	0.84

Table 2.2 demonstrates that RF models tend to provide the most accurate prediction when considering a single pollutant, with 3/4 studies looking at PM2.5 or NOx having RF as the best predicting model. When examining AQI, SVM models tend to work best, with 2/4 studies finding SVM provides the best prediction.

The RMSE and R2 values vary significantly for each study. This can be attributed to the different geographic areas considered, varying spatial resolutions, amount of uncertainty in the data sources, prediction type and different parameters included in the model.

Because the objective of this study is to predict a single variable (NO2 concentrations), then the models used by PM2.5 studies are the most relevant. Therefore, in our model analysis, we will use MLR, neural networks and RF.

2.2 Exploratory Data Analysis

2.3 Model

2.3.1 Multiple Linear Regression

2.3.2 Neural Networks

2.3.3 Random Forest

3. Results

4. Discussion

References

1. **Primary National Ambient Air Quality Standards (NAAQS) for Nitrogen Dioxide**
   OAR US EPA
   US EPA (2016-07-01) https://www.epa.gov/no2-pollution/primary-national-ambient-air-quality-standards-naaqs-nitrogen-dioxide