







Predicting NO2 concentrations

This manuscript ([permalink](#)) was automatically generated from [sfiala2/498_NO2_predict@d9e9962](#) 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

NO₂ 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 PM_{2.5} and predicting the Air Quality Index (AQI)/ Air Pollution Index (API). One exception is that one of the studies examining AQI also predicted NO_x concentrations.

2.1.1 PM_{2.5}

Chen et al (2018) explored the use of random forest models to predict PM_{2.5} 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 $\mu\text{g}/\text{m}^3$ on a daily scale ($R^2 = 83\%$), improving to 10.7 $\mu\text{g}/\text{m}^3$ ($R^2 = 86\%$) and 6.9 $\mu\text{g}/\text{m}^3$ ($R^2=86\%$) on monthly and annual time-scales, respectively.

Xu et al (2018) likewise considered a number of machine learning models for PM_{2.5} 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 PM_{2.5} 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 $R^2=0.48$) and the the MLR had the lowest accuracy (MSE = 3.24 $\mu\text{g}/\text{m}^3$ and $R^2=0.22$). The predictors with the most influence were monthly AOD and elevation.

Enebish et al (2020) considered 6 different machine learning models for PM_{2.5} 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 R^2 and RMSE values. Over the entire period using the hold-out test set, RF had a RMSE of 12.92 ($R^2 = 0.96$), and the cold season and warm season had RMSE of 21.23 ($R^2 = 0.92$) and 7.44 ($R^2 = 0.84$), 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 PM_{2.5} 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 PM_{2.5} 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

2.1.3 Comparison of PM_{2.5} and AQI/API studies

The main difference between the PM_{2.5} and the AQI studies is that studies examining PM_{2.5} 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 PM_{2.5} 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:

| PM_{2.5} | Both PM_{2.5} 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) | | ## 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>