Predicting NO2 concentrations

Authors

- **Tessa Clarizio**
  · ⬡ [tessac2](#)

- **Marcel Briguglio**
  · ⬡ [marcelbriguglio](#)

- **Sudheer Salana**
  · ⬡ [sudheersalana1](#)

- **Xiaokai Yang**
  · ⬡ [xiaoky97](#)

- **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 NO_2 is set at a 53 ppb annual average [1]. NO_2 can cause respiratory irritation and can aggravate respiratory diseases such as asthma (US EPA, n.d., B). NO_2 can also react with other chemicals in the atmosphere to form both particulate matter (PM) and tropospheric ozone (O_3) (US EPA, n.d., B). PM and O_3 are also criteria air pollutants and are harmful to human health. NOO_2 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 NO_2 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 NO_2 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 PM2.5

Chen et al (2018) explored the use of random forest (RF) models to predict $PM_{2.5}$ concentrations spatially in China and compared them to multiple linear regression (MLR) and generalized additive models. 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 root mean squared error (RMSE) of 28.1 μg/m^3 on a daily scale (R^2 = 83%), improving to 10.7 μg/m^3 (R2 = 86%) and 6.9μg/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) 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) 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 P$_{2.5}$ data collected from 63 stations between 2001 to 2014 as well as 3km resolution aerosol optical depth (AOD) data from Moderate Resolution Imaging Spectroradiometer (MODIS). This study found that the cubist model had the highest accuracy (RMSE=2.64 µg/m^3 and R^2=0.48) and the the MLR had the lowest accuracy (MSE = 3.24 µg/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) SVM with a radial basis kernel, 4) 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 (^{2} = 0.84), respectively.

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 NO$_2$, SO$_2$, CO, PM$_{10}$ and O$_3$ over a period of 7 years in Malaysia. The best RMSE and R^2 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 PM$_{2.5}$, PM$_{10}$, SO$_2$, CO, NO$_2$ and O$_3$. 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 NO$_x$ in an Italian city. Parameters included historical hourly averaged AQI concentrations for PM$_{2.5}$, O$_3$, SO$_2$, PM$_{10}$ and NO$_2$ in Beijing (five years), and hourly averaged responses for CO, non-methane hydrocarbons, benzene, NO$_x$ and NO$_2$ in the Italian city (1 year). The SVM performed better in predicting AQI in Beijing with a RMSE 7.666 (R2=0.9776), but the RF model performed better in predicting NO$_x$ concentrations in the Italian city (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, as essentially RF are ensemble decision trees. The parameters included in the model are 5 years of data on: daily air quality measurements (SO_2, NO_2, suspended particulate matter and respirable suspended particulate matter) meteorology (air temperature, relative humidity, wind speed, evaporation, and daily sunshine period). The DTF and DTB models outperformed the SVM models. DTB performed the best, with a RMSE of 4.38 (R^2 = 0.92).

2.1.3 Comparison of PM2.5 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. For example, meteorological data tended to be included in $PM_{2.5}$ studies, but not in studies examining API/AQI. Additionally, different types models tended to perform best depending on the target prediction.

A common limitation of all of the 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. 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, 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.

The models in each of these studies is summarized in Table 2.1 below:

Table 2.1 Models used in literature

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 $PM_{2.5}$ than AQI, and the ones that were used in AQI studies were also used in predicting $PM_{2.5}$. The best-predicting models in each study are shown in Table 2.2, alongside their RMSE and R^2 values.

Table 2.2 Best models in each study

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 $PM_{2.5}$ or NO_x 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 R^2 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 (NO_2 concentrations), then the models used by $PM_{2.5}$ 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

Artifical neural networks are based on the design philosophy of the neural connections in our brain. They consist of a group of nodes interconnected to each other through edges. The edges transmit the signals (which in a machine learning model would be a real number) from one neuron to another just like a synapse functions in a brain.

A typical neural network has many layers. Each layer is composed of a set of neurons and each layer transforms and processes data in a different way based on the hyperparameters of the model.

There are different kinds of neural networks. The most commonly used of which are: 1. Feed Forward Neural Network 2. Convolutional Neural Network (CNN) 3. Recurrent Neural Network (RNN) 4. Long-short term memory neural network (LSTM) 5. Gated Recurrent Unit (GRU)

There are several other advanced ones such as General Adversarial Networks,Auto-encoders and Deep Belief Neural networks.

We will first discuss, briefly, the CNN, RNN, LSTM and GRU before moving on to the Feed formal neural network which is the neural network model used in this project. #### 2.3.2.1 CNN Largely used for classifying image and audio data, the convolutional neural network, similar to neural networks has an input and output layer with hidden layers in between them. Further, a CNN has three layers called convolutional layer, pooling layer and a fully-connected layer which is similar to the regular neural network. The convolutional layer takes the input from the input layer and convolutes the data and sends it downstream for further processing. In simpler terms, in CNNs an input image is taken and a filter is applied on it repeatedly acorss the image to create a feature map which can be used to identify and classify the input image.

2.3.2.2 RNN

These neural networks are used to predict the temporal trend of the data.In these kind of models, the 'memory' of the previous inputs of is stored and used for further processing of future inputs and outputs.This relationship between input and output data could be both unidirectional (moving in the forward direction) or bidirectional where future data could also be used to improve the current inputs and outputs.These neural networks could use a variety of relationships between their inputs and outputs and could be described as one to one, one to many, many to many and many to one relationships. Sigmoid, ReLu, Tanh are the common activations used in RNNs.

2.3.2.3 LSTM

LSTMs are a type of RNN which solve the common problems that RNNs face, that is their inability to efficiently handle short-term memory over a lengthy series of steps. LSTMs solve this problem using a memory cell and gating units and consists of a set of gates called input, output and forget.The input gate is responsible for monitoring and deciding the kind and quantity of data that is allowed to enter the cell, the memory gate is responsible for deciding the proportion of data that should be 'forgotten' and which information is useless and to be discarded. Finally, the output gate is responsible for deciding the amount of data that is passed as output from the cell.

2.3.3 Random Forest

Trees and tree algorithm is among the most widely applied machine learning algorithms. It includes random forest, gradient boosting decision trees, XGBoost, etc. The fundament of trees algorithm is decision tree, which can be divided into classification tree and regression tree. In this project, the label is the observed NO_2 concentration, thus, regression tree algorithm can be applied to make the prediction.

Random forest is a combination of tree predictors based on model aggregation ideas. It is realized by creating an ensemble of trees by generating random vectors that govern the growth of trees and letting them vote for the most popular label [2].

Random forest have the advantages of low overfit, low noise affect. However, for random forest regression, a relatively large number of features are required to reduce the test set error[2].

Variable selection and feature engineering are very important in data preprocess of random forest. The two main objectives are to find variables highly related to the response variable, and to find a small number of variables sufficient to a good prediction of the response variable[3].

Based on previous exploratory data analysis, the raw dataset contains valuable land-use variables in series including impervious surfaces, population, major road length, residential road length, and total road length within different buffers. It also include elevation, distance to coast, latitude, longitude,

linear combinations of satellite data and WRF-Chem output. The variables within their series are highly correlated, so a few representative ones should be selected out based on the second criteria mentioned before. Here we select one for each kind of land-use variables. It is interesting that high correlations have been found between some of these land-use variables, like road lengths. Thus, we only selected the road length having the highest relationship with the variable to be predicted. Latitude and longitude are binned and one-hot coded. Around 20% of the raw data were selected out for prediction validation.

Random forest model in this project is created with scikit-learn RandomForestRegressor library. GridSearchCV was imported to conduct hyperparameter optimization. The model generally get a MSE around 3.2 in the validation data. However, on the test data it did not show a good performance (MES = 4.2). The first reason might be that random forest algorithm perform worse on regression problems than on classification problems since it cannot give a continuous prediction. It is also weak on predicting values with a magnitudes beyond the train dataset, meaning that it cannot confidently handle noise or outliers. It prefers high dimensional, large-scale data, however, in this project the data is in a relatively low scale.

In conclusion, random forest is possibly not a very appropriate algorithm to make prediction in regression problems from dataset in a low-scale and a low-dimension. # 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

2. **Random Forest**
   Leo Breiman
   Machine Learning (2001)
   DOI: doi:10.1023/a:1010933404324

3. **Variable selection using random forests**
   Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot
   Pattern Recognition Letters (2010-10) https://doi.org/czr7p4
   DOI: 10.1016/j.patrec.2010.03.014