Ozone Concentration Prediction

Lijing Zhai, Minyu Ma, Zephren Collison, Zongrun Li

Introduction

Ozone is considered as an air pollutant which causes respiratory and other health effects on humans [1]. Predicting ozone pattern under different conditions is useful for ozone controlling and public health. Since ozone concentrations depend on both weather conditions and emissions especially VOC (volatile organic carbon) and NO_x (NO_2 + NO), empirical model such as machine learning models trained by observation data can be conducted to predict ozone levels [2]. And these models can be utilized to predict ozone concentration in the places missing observations or forecast ozone levels under different weather, emission conditions.

Problem Definition

We will focus on South Coast Air Basin (SoCAB) in California during 1980 to 2020. Since myriad variables potentially affect ozone concentration, we will decide which variables are more important for predicting ozone. Also, influences of different variables to ozone levels will be studied by using different machine learning models.

Methods

The ozone concentration data and weather data (including temperature, relative humidity, wind speed) are provided by EPA. The weather simulation model based on satellite data is potential data source if too much weather data is missing. Emissions of VOC and NO_x are estimated by previous work [2]. Also, some geographic information such as elevation, latitude, longitude and time data will be included since these will affect solar flux which is an important driving force for chemical in atmosphere. In this project, we will provide two routines to train the empirical models. One is that we use principal component analysis to reduce dimensions of data and then use polynomial regression to predict ozone by using selected features. The other is that we use self-organizing map to reduce dimensions and then use artificial neural network to predict ozone [3]. Potentially, training models by using all features directly can be used to compare with these two routines. Random forest is another potential algorithm since it has better performance for discrete relationships than supervised algorithms we used above [4]. To avoid overfitting and estimate the uncertainty of our model, data withholding will be conducted.

Potential Results and Discussion

Performances of models established by different methods will be compared with each other. If the ozone concentration can be estimated by our empirical models, there are several potential applications for the models. One is that spatial distribution of ozone will be predicted by applying the model for all places in SoCAB. The ozone levels for every place in the whole region can be used for estimating ozone exposure which is the product of ozone concentration and population. The ozone exposure is helpful for medical research related to ozone [5]. Another is that we can simulate ozone levels under different emission conditions. The simulation results can be used for environment engineer to decide emission control policies.

Reference

[1] Balmes, J. R. (1993). The role of ozone exposure in the epidemiology of asthma. *Environmental health perspectives*, *101*(suppl 4), 219-224.

[2] Qian, Y., Henneman, L. R., Mulholland, J. A., & Russell, A. G. (2019). Empirical development of ozone isopleths: Applications to Los Angeles. *Environmental Science & Technology Letters*, *6*(5), 294-299.
[3] Andras, P. (2013). Function approximation using combined unsupervised and supervised learning. *IEEE transactions on neural networks and learning systems*, *25*(3), 495-505.
[4] Friedman, J. H. (1994). An overview of predictive learning and function approximation. *From statistics to neural networks*, 1-61.
[5] Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., ... & Thun, M. (2009). Long-term ozone exposure and mortality. *New England Journal of Medicine*, *360*(11), 1085-1095.