Ozone Concentration Prediction

Lijing Zhai, Minyu Ma, Zephren Collison, Zongrun Li

Introduction

Ozone is an air pollutant which can cause harmful respiratory and other health effects on humans [1]. Predicting ozone concentrations in different conditions is useful for ozone control and public health management. Ozone concentrations depend on both weather conditions (wind, temperature, etc.) and emissions; especially VOC (volatile organic carbon) and NO_x (NO_2 + NO). These interactions can be empirically modeled using machine learning models trained by observation data to predict ozone levels over time [2]. Furthermore, multiple models can be utilized to predict ozone concentration between weather observation stations or even to forecast ozone levels under different weather and emission conditions.

Problem Definition

We will focus on the South Coast Air Basin (SoCAB) region in California from 1980 to 2020. Since myriad variables potentially affect ozone concentration, we will identify which variables are more important for predicting. Also, relative influence of these identified variables on ozone levels will be studied using different machine learning models. Based on empirical model established by machine learning methods, spatial distribution of ozone in SoCAB region and ozone isopleth (ozone concentration under different VOC and NO_x emission levels) can also be studied.

Data Collection

Our project focus on SoCAB region in California. To train ozone empirical model, we include several features which have potential relationships with ozone concentration.

Emissions

Emissions of VOC and NO_x for each year are estimated by previous research [2] (Figure 1). Obviously, emissions of VOC and NO_x decrease in these years while the decrease rate is lower in recent years especially for VOC.

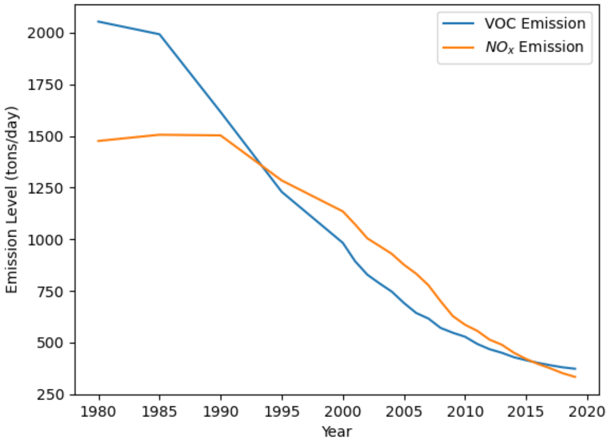


Figure 1. Emissions of VOC and NO_x in SOCAB

Ozone

We use hourly ozone concentration data from EPA website. Based on SoCAB boundaries, 68 monitors are included in the project. The spatial distribution of monitors is shown below. Since previous research shows that the yearly fourth high 8-hour maximum ozone (ozone design value) has relative significant effect to human health, we calculate ozone design value (select maximum value of all 8-hour consecutive ozone for each day and then find the fourth high for whole year) for each year each site and use them as labels in our machine learning model. Spatial averaged ozone design values are shown below. Ozone concentration is decreasing during these years while it fluctuates in recent years.

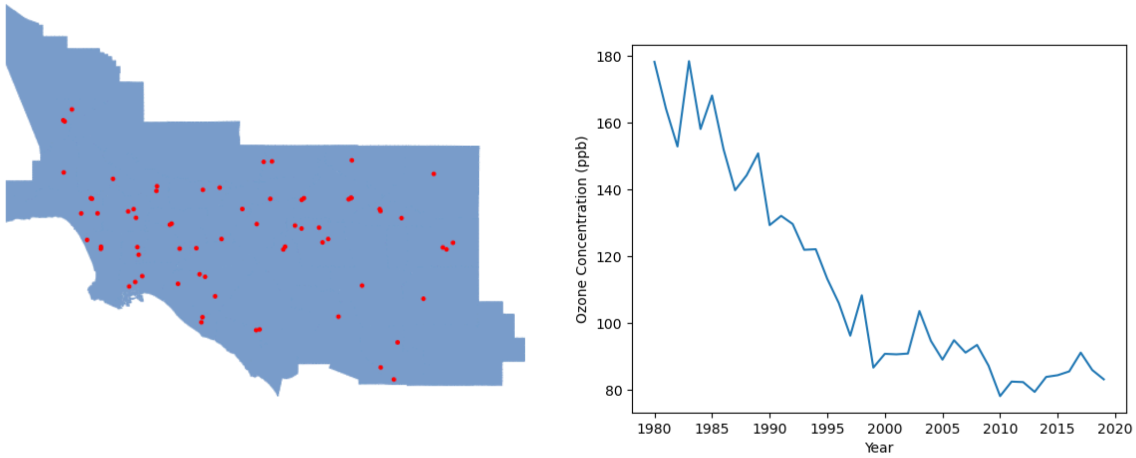


Figure 2. Distribution of selected monitors in SoCAB (left figure). Spatial averaged ozone concentration in SoCAB from 1980 to 2019 (right figure).

Meteorology and Topology Features

Since different locations and meteorology conditions also affect ozone formation and deposition, we include meteorology and topology features into our model. For topology data, we extract elevation for each site from LANDFIRE website which provides high resolution topology data. Also, we include latitude and longitude of each sites which can be used to train the model and analyze spatial distribution of ozone in model application. Julian day and hour for ozone design value are selected to describe sun flux which may affect photochemistry of ozone.

For the meteorological data, we mainly focus on four features, which are Winds, Temperature, Barometric Pressure and RH. We download all the data from EPA website. To make our features more convincing and scientific, we decide to use the hourly data and calculate the average of maximum 8 hours for all of the meteorological features. However, there still exists missing data since the EPA uses different site ids to monitor each of the features. To make our data consistent, we use the site which has the minimum distance to the ozone monitors and capture the data from the closest meteorology site for the missing data and directly replace it.

Methods

The ozone concentration and weather data are provided by the EPA. Since there are several missing features for the data, we dropped the data which has more than two missing features and then k-Nearest Neighbors imputation method is conducted to estimate missing features. A weather simulation model based upon satellite data is a potential data source when we apply our model to study ozone spatial distribution since the simulation model could provide fields of meteorology parameters. Emissions of VOC and NO_x are estimated by previous work [2]. Also some geographic information such as elevation, latitude, longitude, and time will be included since these affect solar flux; which is an important driving force for chemical reactions in the atmosphere. In this project, we will provide two routines to train the empirical models. First we will use principal component analysis to reduce the dimensions of our data and then use polynomial regression on the reduced data to predict ozone concentrations. A backward elimination method is conducted to reduce polynomial terms for the regression analysis. The other method we will use is a self-organizing map to reduce dimensions and then an artificial neural network to predict ozone levels [3]. We can potentially compare these two model training routines by using all features directly. Random forest is another potential algorithm since it has better performance for discrete relationships than the supervised algorithms we listed above [4]. To avoid overfitting and estimate the uncertainty of our model, data withholding will be conducted.

Results and Discussion

We extract 1171 data points from data sources while 206 of them are missing at least two features. These data are most concentrated in early year due to limited monitors in California. We drop these data and conduct 3 nearest neighborhood imputation method to estimate the missing values for the left data.

The covariance matrix between each feature is shown below. NO_x and VOC emissions have the highest correlation values with ozone among selected features. It is understandable since VOC and NO_x emissions are two dominant chemical species which affect ozone concentration [6].

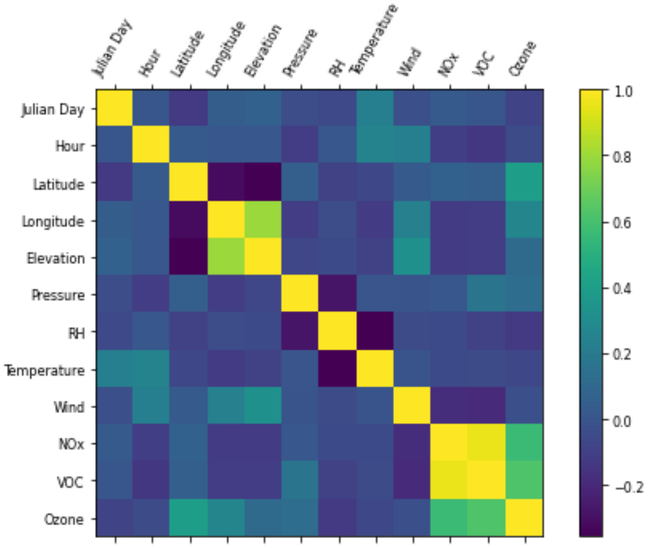


Figure 3. Correlations between features.

Polynomial Regression

Principal Components Analysis

Firstly, we use principal components analysis to reduce the dimension of our data. The ratio of explained variance for each component is shown in figure 4. The highest three features are explained 99.19% variance. So, we extract these three components and to find the dominant features in the principal component.

The principle components are the linear combination of original features. From figure 5, NO_x, VOC, pressure, elevation have the highest absolute coefficients. It means that these features can reproduce principal components without too much loss. So, we select these features for training polynomial regression model.

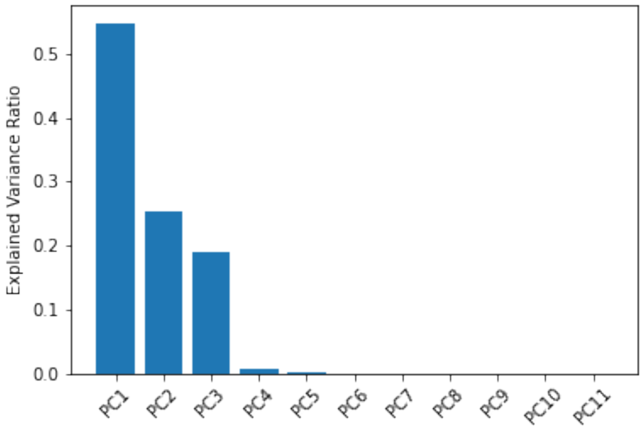


Figure 4. Ratio of explained variance for different components.

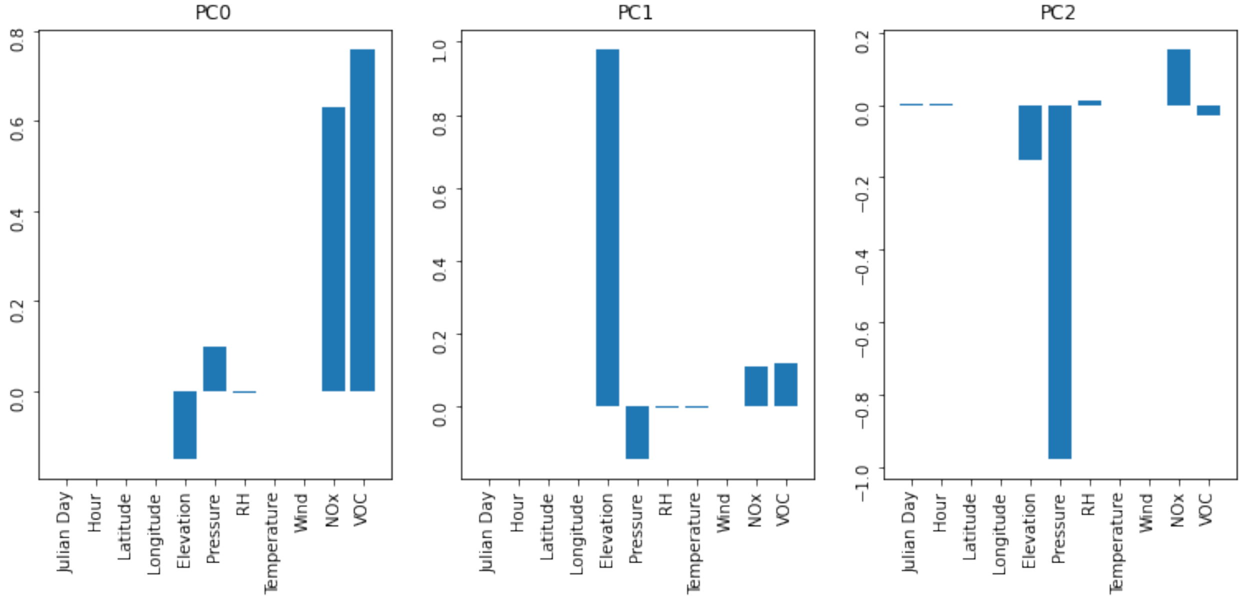
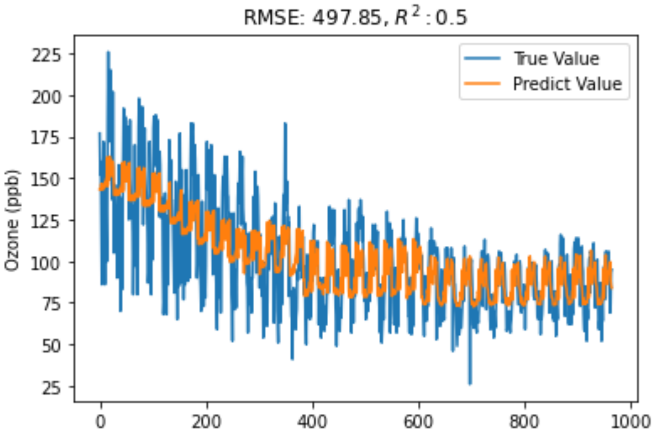


Figure 5. Ratio of explained variance for different components.

Polynomial Regression (ongoing)

Second order polynomial regression is conducted to predict ozone under selected feature in PCA step. There are 14 terms which order is lower than 2. Backward elimination method is conducted to eliminate these terms. Nine of them are remained after elimination and polynomial regression is applied by using remained terms and labels (ozone concentration). Regression performance is shown in figure 5. The trend and peaks are captured in regression model while the bias is relatively high in early year (high ozone concentration) and low in recent year (low ozone concentration) which is favorable since stakeholders are more interested to predict ozone concentration for the future which is expected as low ozone concentration.



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.

[6] Steinfeld, J. I. (1998). Atmospheric chemistry and physics: from air pollution to climate change. *Environment: Science and Policy for Sustainable Development*, *40*(7), 26-26.