Forecasting and time variability analysis of Ozone concentrations using nitrate oxide and meteorological variables as predictors

This manuscript (permalink) was automatically generated from uiceds/cee-492-term-project-fall-2022-hydrograds@24e49c2 on September 22, 2022.

Authors

- **Rourou, Bernardo, Jiewen**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of CEE, University of Illinois at Urbana&Champaign · Funded by Grant XXXXXXXX

- **Rourou Ji**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

- **Bernardo Burbano**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

CE 492 Final Project Selection

1. Dataset description:

The dataset used in this project is a CSV file about the air quality in northern Taiwan collected in 2015 (https://www.kaggle.com/datasets/nelsonchu/air-quality-in-northern-taiwan), which include air quality data and meteorological monitoring data for research and analysis, originally from Environmental Protection Administration, Executive Yuan, R.O.C. (Taiwan). There are 25 observation stations in total. Columns in this CSV file are the following.

Time - The first column is the observation time of 2015

Station - The second column is the station name, there are 25 observation stations [Banqiao, Cailiao, Datong, Dayuan, Guanyin, Guting, Keelung, Linkou, Longtan, Pingzhen, Sanchong, Shilin, Songshan, Tamsui, Taoyuan, Tucheng, Wanhua, Wanli, Xindian, Xinzhuang, Xizhi, Yangming, Yonghe, Zhongli, Zhongshan].

Items - From the third column to the last one

item - unit - description

SO2 - ppb - Sulfur dioxide

CO - ppm - Carbon monoxide

O3 - ppb - ozone

PM10 - μg/m3 - Particulate matter

PM2.5 - μg/m3 - Particulate matter

NOx - ppb - Nitrogen oxides

NO - ppb - Nitric oxide

NO2 - ppb - Nitrogen dioxide

THC - ppm - Total Hydrocarbons

NMHC - ppm - Non-Methane Hydrocarbon

CH4 - ppm - Methane

UVB - UVI - Ultraviolet index

AMB_TEMP - Celsius - Ambient air temperature

RAINFALL - mm

RH - % - Relative humidity

WIND_SPEED - m/sec - The average of the last ten minutes per hour

WIND_DIREC - degrees - The average of the last ten minutes per hour

WS_HR - m/sec - The average of an hour

WD_HR - degrees - The average of an hour

PH_RAIN - PH - Acid rain

RAIN_COND - μS/cm - Conductivity of acid rain

2. Proposal:

The purpose of this project is to predict O3 concentrations using measurements of concentration of other pollutants and available meteorological measurements. Ozone might be formed when heat and sunlight cause chemical reactions between oxides of nitrogen (NOx) and Volatile Organic Compounds (VOC), which are also known as Hydrocarbons. Therefore it could be hypothesized that using measurements of NOx as an independent variable a model could be developed to predict O3 concentrations. Additionally, meteorological variables such as air temperature, relative humidity(RH) and ultraviolet index (UVB - UVI) could be included as independent variables to assess their influence on temporal variability of ozone. As an additional step wind-related variables such as mean wind velocity and direction will be included to study their effect on temporal variability of ozone.

After the air quality data has been processed the strongest O3 predictors will be determined using PCA. PCA could be used to identify the main axes of variance within the dataset and explore underlying correlations that exist in a set of variables. Variables that are highly correlated cluster together. Using PCA 2D figures per each pair of variables are not needed, instead all the variables could be visualized simultaneously. Differences on PC1 are more important than differences on PC2. After plotting PCA plots, a heatmap could also be plotted to check the results. As additional criteria to identify the strongest predictors a LSTM network (long short-term memory network) can be used since the data used is time dependent. The network should contain several LSTM layers and fully-connected layers. The output should contain the pollution concentration and will point out the weights assigned to each correlated criterion, the values of such weights should also indicate what the strongest predictors are. Once the strongest predictors have been identified, genetic programming will be used to develop the models to predict O3 concentrations.

Dataset description: The dataset used in this project is a CSV file about the air quality in Northern Taiwan collected in 2015 (https://www.kaggle.com/datasets/nelsonchu/air-quality-in-northern-taiwan), which includes air quality data and meteorological monitoring data for research and analysis, the originator is Environmental Protection Administration, Executive Yuan, R.O.C. (Taiwan). There are 25 observation stations in total. Columns in this CSV file are:

Proposal: The dataset used in this project is a CSV file about the air quality in Northern Taiwan collected in 2015 (https://www.kaggle.com/datasets/nelsonchu/air-quality-in-northern-taiwan), which includes air quality data and meteorological monitoring data for research and analysis, the originator is Environmental Protection Administration, Executive Yuan, R.O.C. (Taiwan). There are 25 observation stations in total. Columns in this CSV file are the following. Time - The first column is the observation time of 2015 Station - The second column is the station name, there are 25 observation stations: [Banqiao, Cailiao, Datong, Dayuan, Guanyin, Guting, Keelung, Linkou, Longtan, Pingzhen, Sanchong, Shilin, Songshan, Tamsui, Taoyuan, Tucheng, Wanhua, Wanli, Xindian, Xinzhuang, Xizhi, Yangming, Yonghe, Zhongli, Zhongshan]. Items - From the third column to the last one item - unit - description SO2 - ppb - Sulfur dioxide CO - ppm - Carbon monoxide O3 - ppb - ozone PM10 - µg/m3 - Particulate matter PM2.5 - µg/m3 - Particulate matter NOx - ppb - Nitrogen oxides NO - ppb - Nitric oxide NO2 - ppb - Nitrogen dioxide THC - ppm - Total Hydrocarbons NMHC - ppm - Non-Methane Hydrocarbon CH4 - ppm - Methane UVB - UVI - Ultraviolet index AMB_TEMP - Celsius - Ambient air temperature RAINFALL - mm RH - % - Relative humidity WIND_SPEED - m/sec - The average of the last ten minutes per hour WIND_DIREC - degrees - The average of the last ten minutes per hour WS_HR - m/sec - The average of an hour WD_HR - degrees - The average of an hour PH_RAIN - PH - Acid rain RAIN_COND - µS/cm - Conductivity of acid rain

Proposal:

The purpose of this project is to predict O3 concentrations using measurements of concentration of other pollutants and available meteorological measurements. Ozone might be formed when heat and sunlight cause chemical reactions between oxides of nitrogen (NOx) and Volatile Organic Compounds (VOC), which are also known as Hydrocarbons. Therefore it could be hypothesized that using measurements of NOx as an independent variable a model could be developed to predict O3 concentrations. Additionally, meteorological variables such as air temperature, relative humidity(RH) and ultraviolet index (UVB - UVI) could be included as independent variables to assess their influence on temporal variability of ozone. As an additional step wind-related variables such as mean wind velocity and direction will be included to study their effect on temporal variability of ozone.

After the air quality data has been processed the strongest O3 predictors will be determined using PCA. PCA could be used to identify the main axes of variance within the dataset and explore underlying correlations that exist in a set of variables. Variables that are highly correlated cluster together. Using PCA 2D figures per each pair of variables are not needed, instead all the variables could be visualized simultaneously. Differences on PC1 are more important than differences on PC2. After plotting PCA plots, a heatmap could also be plotted to check the results. As additional criteria to identify the strongest predictors a LSTM network (long short-term memory network) can be used since the data used is time dependent. The network should contain several LSTM layers and fully-connected layers. The output should contain the pollution concentration and will point out the weights assigned to each correlated criterion, the values of such weights should also indicate what the strongest predictors are. Once the strongest predictors have been identified, genetic programming will be used to develop the models to predict O3 concentrations.

References