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@cbde214](#) on November 21, 2022.

Authors

- **Jiewen Luo**
  · ⬡ [Noomi-Luo](#)
  Department of CEE, University of Illinois at Urbana&Champaign

- **Rourou Ji**
  · ⬡ [JadeJi](#)
  Department of CEE, University of Illinois at Urbana&Champaign

- **Bernardo Burbano**
  · ⬡ [BernieJBA](#)
  Department of CEE, University of Illinois at Urbana&Champaign

CEE 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:

1. Time - The first column is the observation time of 2015

2. Station - The second column is the station name, there are 25 observation stations, those stations are showing at the table [1](#).

Table 1: A table contain all stations in Taiwan.

		station		
Banqiao	Cailiao	Datong	Dayuan	Guanyin
Guting	Keelung	Longtan	Pingzhen	Sanchong
Shilin	Songshan	Tamsui	Taoyuan	Tucheng
Wanhua	Wanli	Xindian	Xinzhuang	Xizhi
Yangming	Yonghe	Zhongli	Zhongshan	Linkou

3. Items - From the third column to the last one

4. item - `unit` - description

- SO_2 - `ppb` - Sulfur dioxide
- CO - `ppm` - Carbon monoxide
- O_3 - `ppb` - ozone
- PM_{10} - `µg/m^3` - Particulate matter
- $PM_{2.5}$ - `µg/m^3` - Particulate matter
- NO_x - `ppb` - Nitrogen oxides
- NO - `ppb` - Nitric oxide
- NO_2 - `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 O_3 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 (NO_x) and Volatile Organic Compounds (VOC), which are also known as Hydrocarbons. Therefore it could be hypothesized that using measurements of NO_x as an independent variable a model could be developed to predict O_3 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 O_3 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 O_3 concentrations.

Exploratory Data Analysis:

In order to explore the relation between the dependent variable and independent variables several scatter plots were created between meteorological variables, pollutant concentrations and ozone concentrations. Additionally, a heatmap was generated to investigate the correlation values between ozone concentration and independent variables. The most correlated variables are RH(relative humidity) and UVB(Ultraviolet index). RH is negatively correlated with ozone,value of -0.51, while UVB is positively correlated, value of 0.51. Another relevant observation was that concentrations of nitrogen-related chemicals are highly correlated between each other, which is close to the truth.
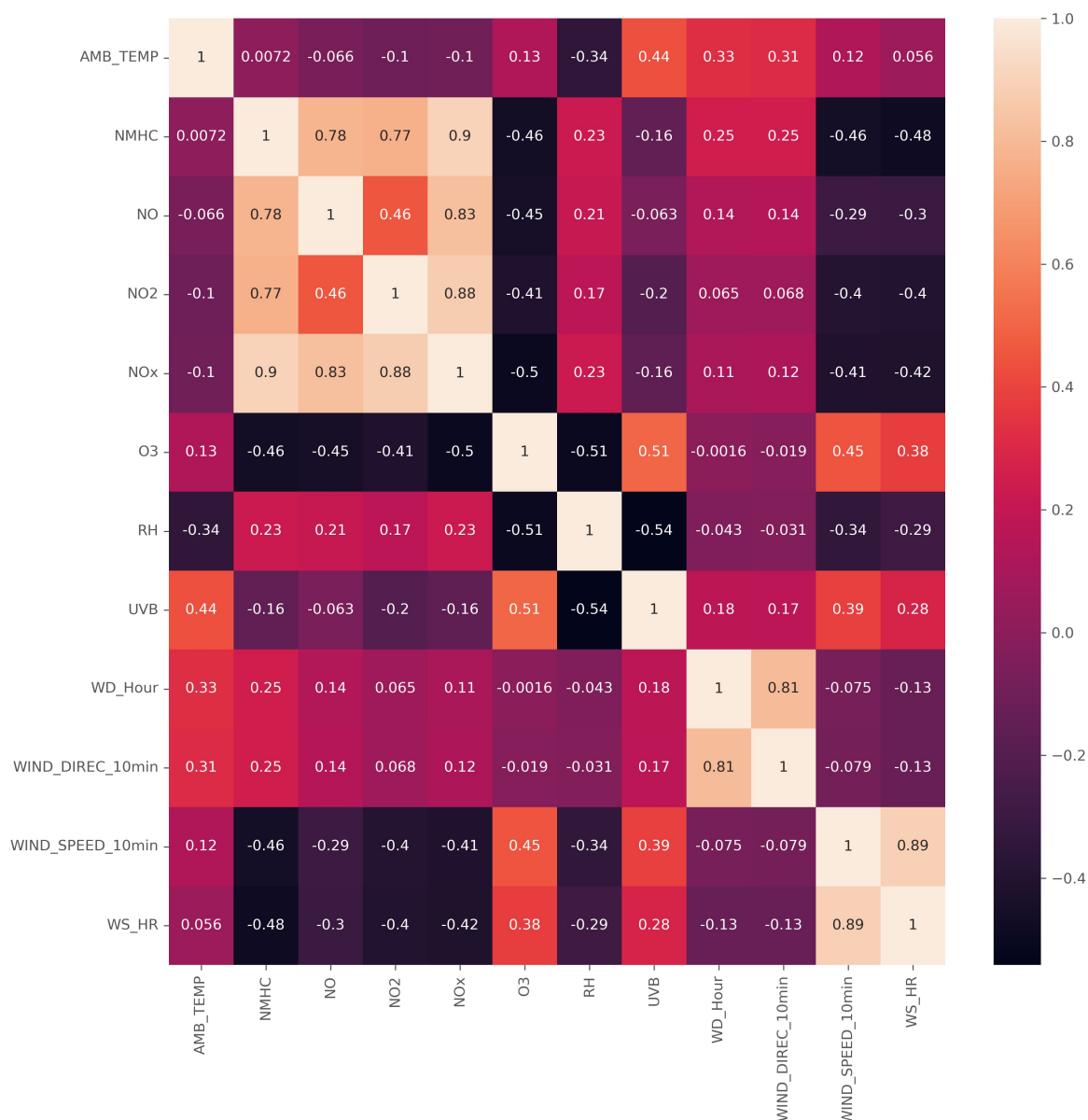
Figure 1: Heatmap for the variables we considered in this analysis

Furthermore, the fraction of available measurements, meaning the number of data points available divided by the number of hours in a year, was computed for all stations and all measured variables. This computation helped visualize the stations that missed the least data points as well as the variables whose values are recorded the most consistently through different stations. The station with the highest fraction of available measurements was Banquiao, as seen in the Figure 2. For this reason the remaining portion of this EDA was devoted to this station.

| Station	Dependent Variable	Independent variable: Nitrogen based				Independent variable: Metereological variables							Average Fraction
	O3	NMHC	NO	NO2	NOx	Temperature	RH	UVB	W Direction 10 min	W Speed 10min	W Speed hourly	Wind Direction hourly	
Banqiao	0.991	0.984	0.966	0.966	0.966	0.991	0.991	0.991	0.991	0.991	0.991	0.991	0.984
Cailiao	0.978	0.000	0.958	0.958	0.958	0.977	0.978	0.000	0.977	0.977	0.977	0.977	0.794
Dayuan	0.988	0.000	0.975	0.975	0.975	0.980	0.955	0.000	0.852	0.910	0.922	0.922	0.770
Guanyin	0.980	0.000	0.969	0.969	0.969	0.973	0.980	0.000	0.977	0.980	0.979	0.977	0.798
Guting	0.988	0.966	0.983	0.983	0.983	0.985	0.983	0.000	0.985	0.985	0.985	0.985	0.901
Keelung	0.987	0.970	0.977	0.977	0.977	0.977	0.980	0.000	0.986	0.986	0.986	0.986	0.891
Linkou	0.973	0.000	0.966	0.966	0.966	0.973	0.973	0.000	0.972	0.972	0.964	0.972	0.793
Longtan	0.982	0.000	0.961	0.961	0.961	0.982	0.982	0.000	0.982	0.982	0.981	0.981	0.798
Pingzhen	0.988	0.000	0.968	0.968	0.968	0.988	0.988	0.000	0.983	0.988	0.988	0.988	0.802
Shilin	0.983	0.000	0.957	0.957	0.957	0.983	0.983	0.000	0.983	0.983	0.980	0.983	0.797
Songshan	0.985	0.967	0.976	0.976	0.976	0.985	0.985	0.000	0.984	0.984	0.984	0.984	0.891
Tamsui	0.961	0.000	0.956	0.956	0.956	0.000	0.000	0.961	0.000	0.000	0.000	0.000	0.348
Taoyuan	0.983	0.958	0.976	0.976	0.976	0.982	0.982	0.976	0.982	0.982	0.982	0.982	0.978
Tucheng	0.981	0.974	0.965	0.965	0.965	0.980	0.981	0.000	0.980	0.980	0.980	0.980	0.886
Wanhua	0.986	0.000	0.979	0.979	0.979	0.986	0.986	0.000	0.986	0.986	0.986	0.986	0.805
Wanli	0.984	0.000	0.976	0.976	0.976	0.983	0.976	0.000	0.976	0.982	0.983	0.975	0.800
Xindian	0.985	0.000	0.979	0.979	0.979	0.985	0.982	0.000	0.985	0.985	0.985	0.985	0.804
Xinzhuang	0.978	0.000	0.970	0.970	0.970	0.978	0.978	0.000	0.978	0.978	0.978	0.978	0.798
Xizhi	0.963	0.000	0.954	0.954	0.954	0.962	0.962	0.000	0.962	0.962	0.958	0.958	0.784
Yangming	0.979	0.000	0.964	0.964	0.964	0.979	0.979	0.000	0.000	0.000	0.000	0.000	0.441
Yonghe	0.987	0.972	0.978	0.978	0.978	0.984	0.987	0.000	0.987	0.987	0.987	0.987	0.893
Zhongli	0.988	0.973	0.967	0.967	0.967	0.986	0.986	0.000	0.986	0.986	0.985	0.985	0.890
Zhongshan	0.986	0.976	0.976	0.976	0.976	0.986	0.986	0.000	0.986	0.986	0.986	0.986	0.893

Figure 2: Data availability of air quality stations

Other relevant statistics from the Banquio air quality station are shown as Table 2.

Table 2: Descriptive statistics of features that compose in Banqiao Station

|	count	mean	std	min	25%	50%	75%	max
AMB_TEMP	8682	1.18E-15	1.00	-2.47	-0.89	0.15	0.68	2.25
NMHC	8619	-4.74E-15	1.00	-1.32	-0.62	-0.30	0.29	16.26
NO	8462	3.20E-15	1.00	-0.75	-0.50	-0.32	0.05	23.55
NO2	8462	1.56E-15	1.00	-1.94	-0.78	-0.10	0.57	5.48
NOx	8462	1.62E-15	1.00	-1.54	-0.69	-0.20	0.42	14.67
O3	8685	-8.78E-16	1.00	-1.35	-0.80	-0.07	0.54	6.06
RH	8684	-1.56E-15	1.00	-4.85	-0.75	0.17	0.76	2.35
UVB	8684	3.53E-15	1.00	-0.56	-0.56	-0.56	0.12	4.81
WD_Hour	8680	2.64E-16	1.00	-1.58	-0.78	-0.59	1.01	2.32
WIND_DIREC_10min	8682	6.96E-16	1.00	-1.56	-0.78	-0.58	1.01	2.31
WIND_SPEED_10min	8682	-6.21E-15	1.00	-1.43	-0.72	-0.18	0.53	7.92
WS_HR	8680	-2.83E-15	1.00	-1.67	-0.80	-0.12	0.56	7.86

As described in previous sections, the dataset consists of hourly observations of ozone (dependent variable) and several pollutant concentrations and meteorological measurements (independent variables). The first step was to plot ozone against all of the independent variables to visualize if the data collapsed into any identifiable pattern, thus to later on use such a pattern to identify potential models. The measurements in an hourly time scale did not show any discernible pattern between the dependent and independent variables. As seen in figure 3.
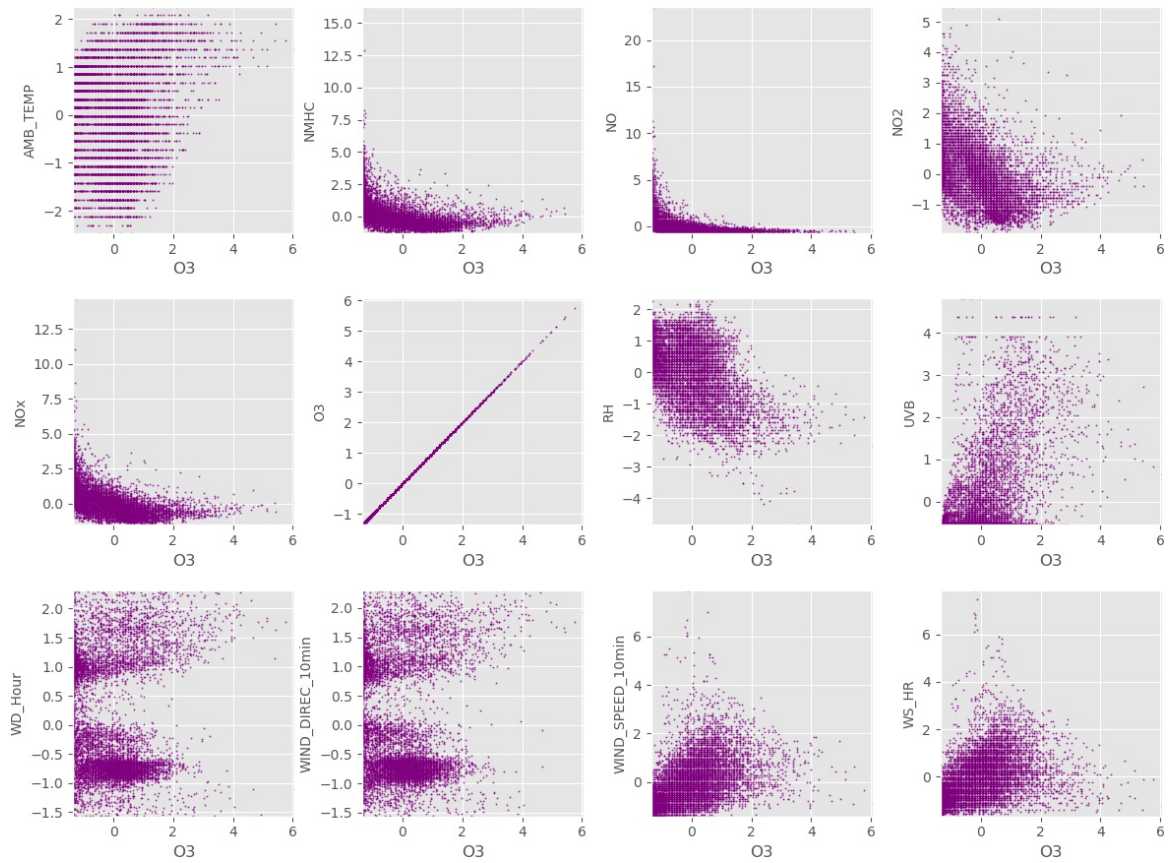
Figure 3: Scatter plots of hourly measurements of variables

Plotting the raw data, i.e. the available measurements without any processing or transformation, did not yield any insights that could help elucidate the relation between the variables. Therefore, the data was normalized. Notwithstanding, normalization did not translate into plots where patterns could be identified. Thus, the data was processed again following two consecutive steps. First the values were averaged over a day and over a month producing a dataset of daily and monthly measurement. Second, such values were standardized by dividing them by the corresponding daily and monthly averages.

The resulting daily and monthly standardized averages were plotted against time. Plotting the daily averaged variables shown plots where the fluctuations of the values happened in a relatively short time and thus such fluctuations obscured any pattern that could be observed in the data, as seen in the next figure. Conversely, when the monthly standardized averages were plotted against time it was visible that the pollutants concentrations shown similar time patterns as seen in figures 4.
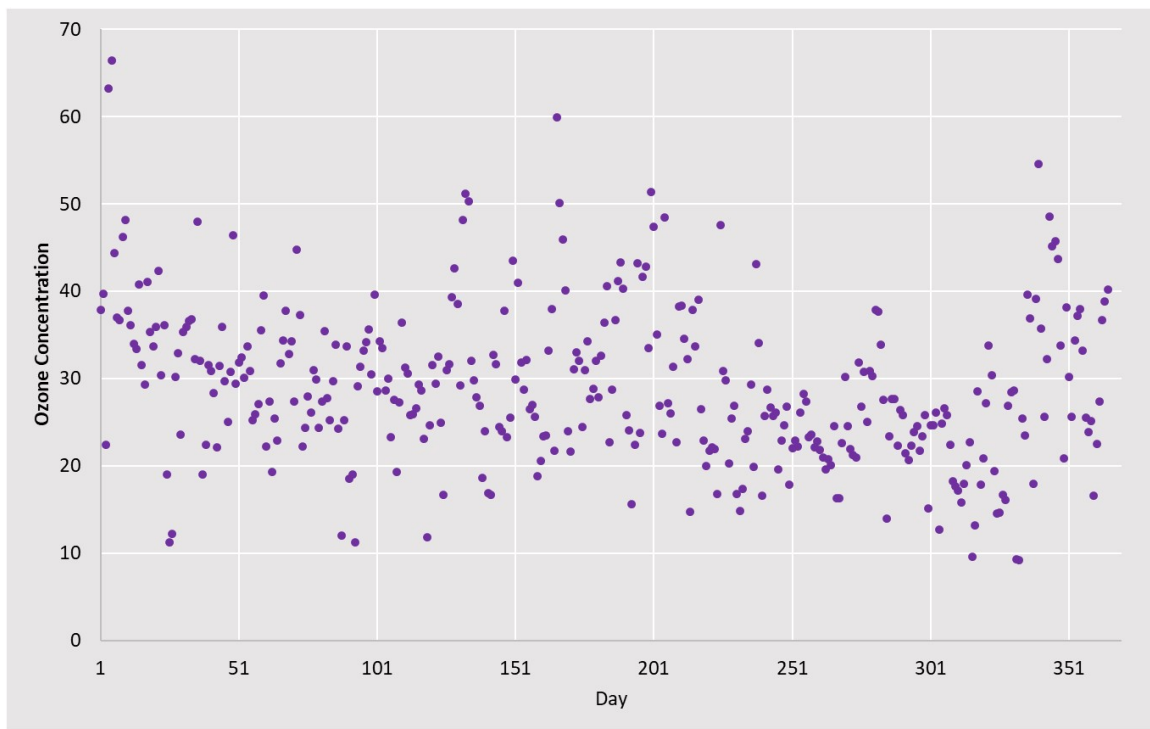
Figure 4: Daily concentrations of ozone
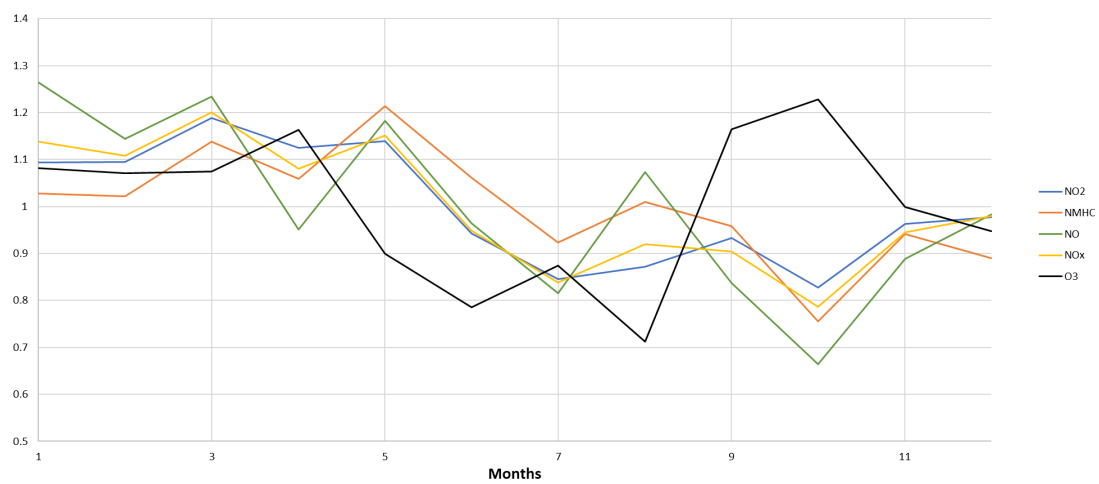
Figure 5: Standardized pollutants and ozone monthly concentration changes
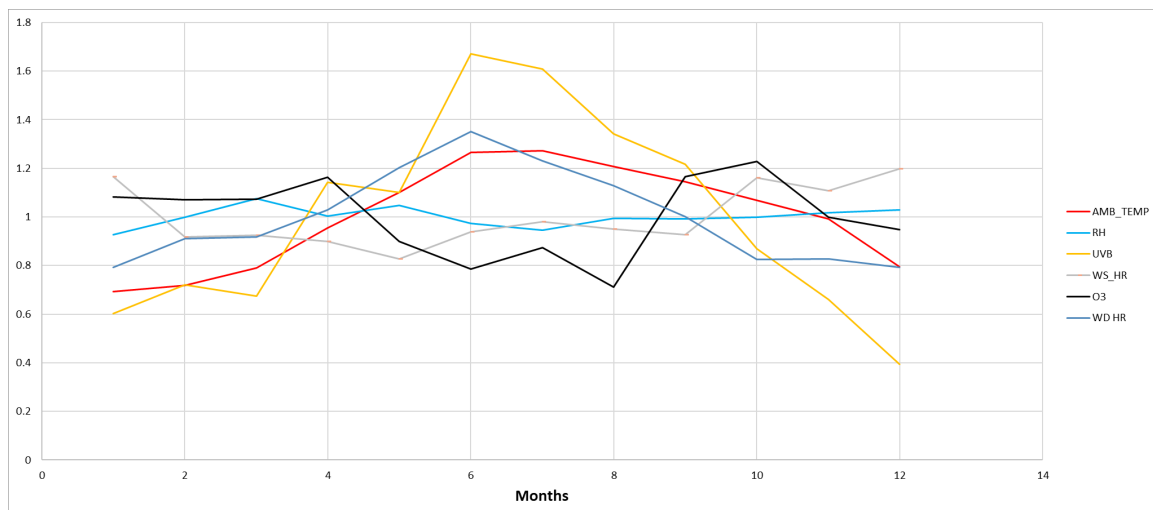
Figure 6: Standardized meteorological measurements and standardized ozone monthly concentration changes

In regards to pollutant concentrations, O3 peaked in the months when concentration of the nitrogen based pollutants and non-methane hydrocarbons dropped. This is especially the case for NO concentrations (green line). This pattern of corresponding decreasing pollutant concentrations and increasing ozone could suggest that the pollutant concentrations are negatively correlated with ozone concentrations. This is also consistent with figure 5 (correlation plot)

In regards to the meteorological variables, UVB (ultraviolet index) and air temperature peak in the same months. Both temperature and UVB experience an increase in their values from the beginning of the year peaking in June. After June, both values experience a steady decrease. No matching pattern can be observed in terms of the relation of the latter two variables and ozone concentrations.

The increase or decrease of the values of wind direction shown in figure 6 correspond to a relative shift in direction of the wind compared to the yearly wind direction. These shifts in the direction of the wind can be used later on the forecasting of O3 concentration. Wind direction could help elucidate if O3 concentration from upwind neighboring locations could affect O3 values in the location of interest, Banquiao.

Predictive Modeling

Selection of neural network model

Different configurations of neural networks where used: fully connected layers, here called NN for simplicity, convolutional neural networks (CNN), and long short-term memory (LSTM).

Long short-term memory and convolutional neural networks

Convolutional neural networks (CNN) uses filters to extend the depth of the input volume. Compared with regular neural network, it indicates the relationship between and better capture edge features of the network. One drawback of CNN is that it's gradients can explode or vanish which may restrict neural network performance. Long short-term memory use two path for long (cell state) and short memories (hidden state) to avoid exploding/vanishing gradient problem. LSTM has three gates that determined the output: forget gate determines which percentage of long-term memory is remembered using a Sigmoid function; input gate calculates the potential memory using a Tanh function as well as the percentage of potential memory to remember using the same approach with the forget gate; and output gate multiply a Tanh function with long term memory result to obtain the output.

| Normalize	Activation function	Number of neurons	Number of Dense Layers	Epochs	η	RMSE
No	Relu	18	3	1.E+05	1.E-04	10.4
No	Tanh	18	3	1.E+06	1.E-10	24
Yes	Relu	18	3	1.E+05	1.E-04	8.5
Yes	Tanh	18	3	1.E+06	1.E-10	33
Yes	Relu	18	4	1.E+06	1.E-04	7.5
Yes	Relu	36-18-12	4	1.E+06	1.E-04	7.5
Yes	Relu	18	5	1.E+06	1.E-04	7
Yes	Relu	18	8	1.E+06	1.E-04	6.7

Figure 7: RMSE values for different hyperparameters and NN configurations

Two convolutional neural network (CNN) were also tested. A first CNN had 8, fully connected layer with 18 neurons, Relu as activation function and a convolution window of size 5. The number of epochs used was 100 with η = 0.01. RMSE of CNN was 8.63 and mean average error (MAE) was 5.85, for reference the best preforming NN, hereafter called NN 8-L, had a RSME of 6.7 and a MAE of 4.64 for the training dataset. The second CNN had a one convolution layer of 64 neurons, two fully connected layers, Relu as activation function and a convolution window of size 2. The number of epochs used was 1000 with a η = 0.0001. RMSE was 6.26.

The LSTM had one 32 neuron layer, a second 64 neuron layer followed by two dense layers of 8 and 1 neuron, activation function used was Relu. Also LSTM training was don with the measurements from the previous 5 hours to predict ozone concentration of the 6th hour. Also, not only the most correlated predictors were used but all of the predictors except for wind speed and wind direction. RMSE for LSTM on the training data was 1.76, clearly LSTM outperformed the other neural network configurations.

The following plots shows predicted O3 values with LSTM and observed O3 values
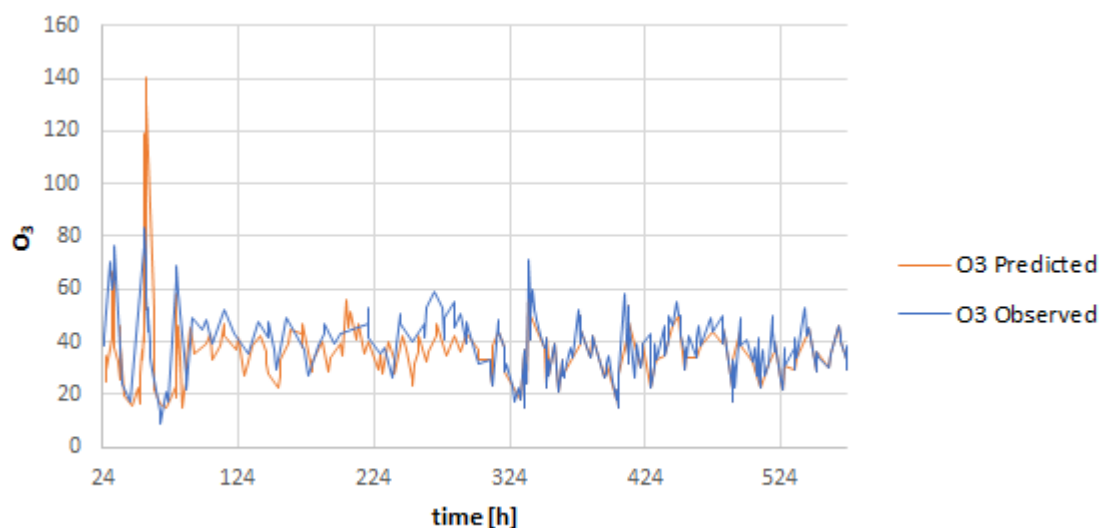
Figure 8: LSTM O3 predicte and observed values

Additional avenues of improvement

<<<<<<< HEAD The current modeling effort might indicate that ozone concentrations might not be predicted with lower error values using the available dataset. Thus repurposing of the modeling effort towards a more error tolerant goal might be an alternative to yield further utility from the available dataset. For instance, the available dataset could be used for a binary classification model to predict if ozone levels are above or below 70 ppb. The aforementioned value is the threshold of the primary (public health) and secondary (public welfare) 8-hour ozone standards defined by the "2015 Revision to 2008 Ozone National Ambient Air Quality Standards (NAAQS) Related Documents" =======

The current modeling effort might indicate that ozone concentrations might not be predicted with lower error values using the available dataset. Thus repurposing of the modeling effort towards a more error tolerant goal might be an alternative to yield utility from the available dataset. For instance, the available dataset could be used for a binary classification model to predict if ozone levels are above or below 70 ppb. The aforementioned value is the threshold of the primary (public health) and secondary (public welfare) 8-hour ozone standards defined by the "2015 Revision to 2008 Ozone National Ambient Air Quality Standards (NAAQS) Related Documents"

Long Short-term Memory Modeling(LSTM)

Convolutional neural networks (CNN) use filters to extend the depth of the input volume. Compared with regular neural network, it indicate the relationship between each cells and better capture edge (changed) feature of the network. One draw back of CNN is that it's gradients can explode or vanish which may restrict neural network performance. Long short-term memory use two path for long (Cell state) and short memories (hidden state) to avoid exploding/vanishing gradient problem. It have three gates that determined the output value together: forget gate determine which percentage of long-term memory is remembered using a Sigmoid function; input gate calculate the potential memory using a Tanh function as well as the percentage of potential memory to remember using the same approach with the forget gate; and output gate multiply a Tanh function with long term memory result to obtain the output. long short-term memory network which has two hidden LSTM layers to increase the depth of the model. It will insure the model become more accuracy since the upper LSTM layer could serve as a sequence observations over time.

While predict the model, we compare two methods with the same size of layers, CNN and Stacked LSTM. Our CNN model is formed by a convolutional layer with 64 filters, (2,2) window, and a relu activation function, as well as two dense layers. Similarly, long short-term memory model input data into two LSTM layers and pass 64 features into two dense layers with the same size as CNN. Notice that we took multivariate time series to achieve O3 concentration forecast, thus, we took the previous 5-hour data of 13 features which we think is relative to the O3 concentration to predict the concentration at the 6 th hour. The 13 features are as table 2 except O3, wind speed and wind direction data for ten minutes. Since last two features are highly correlation and be well captured by average wind speed and wind direction data in one hour. We Processed the time data and expand it into four variables which change periodic monthly and annual (cos function and sin function) to capture the affected by the time of day, week and month. The predict result is shown as below (draft). The red line represent the actual value of O3 concentration and the blue line shown the predict value. The root mean squared error of CNN is 6.2564 and 1.7582 for LSTM model after 1000 training steps with 0.0001 learning rate.

>>>>>>>

82dadc6c9aaa6d07240a5d460a3e495d21b2b921

References

Reference: https://machinelearningmastery.com/return-sequences-and-return-states-for-lstms-in-keras/

https://machinelearningmastery.com/stacked-long-short-term-memory-networks/

https://en.wikipedia.org/wiki/Convolutional_neural_network#Convolutional_layer

https://www.youtube.com/watch?v=YCzL96nL7j0 "Long Short-Term Memory (LSTM), Clearly Explained"

https://www.youtube.com/watch?v=kGdbPnMCdOg "Multivariate Time Series Forecasting Using LSTM, GRU & 1d CNNs"

https://github.com/Dana2021/CEE498DS-Project1

https://blog.csdn.net/bryan__/article/details/51607215 "Introduce several common feature selection methods in conjunction with Scikit-learn"