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@ef4048d] on December 2, 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

Introduction:

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.

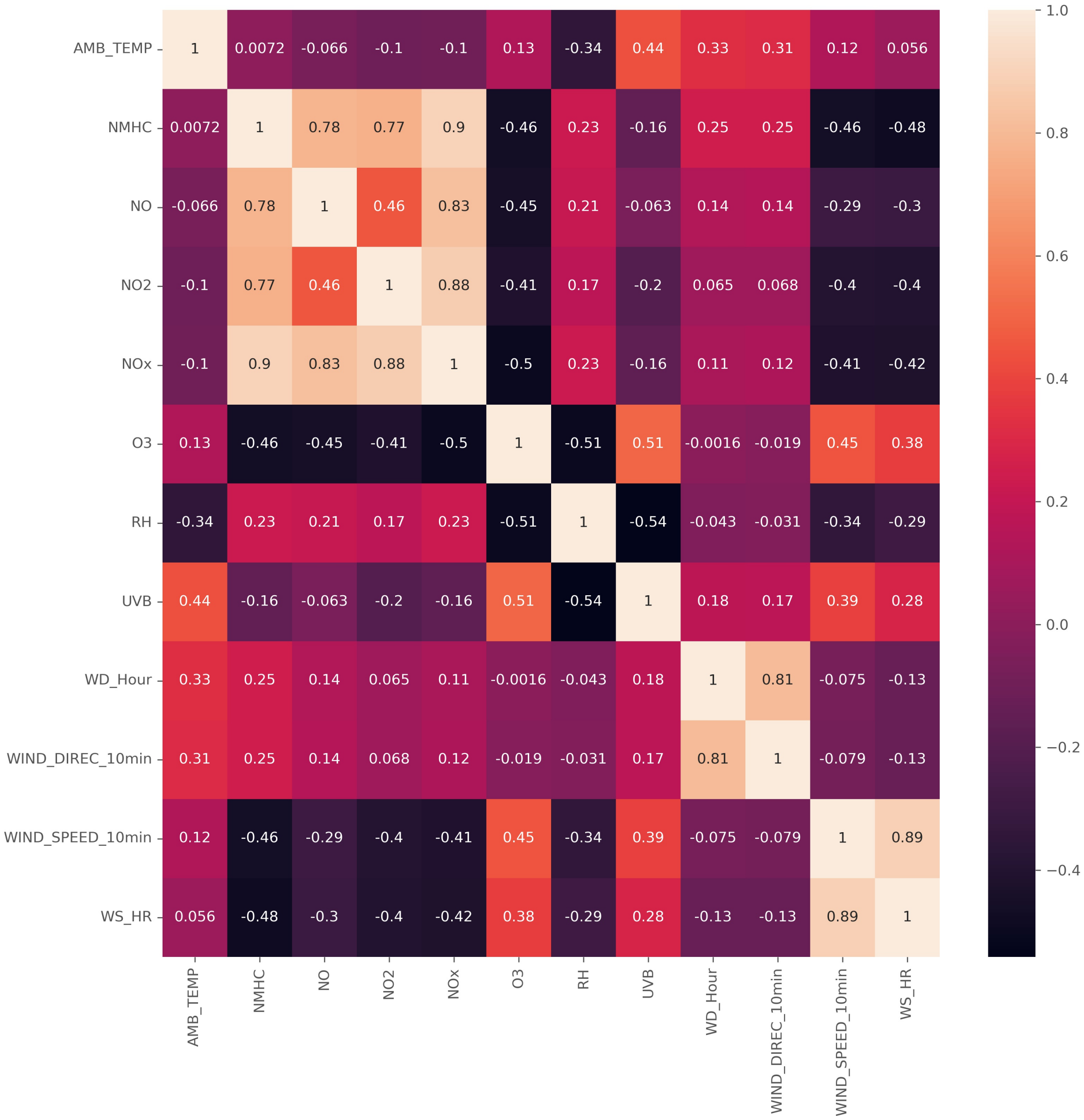

Figure 1: Correlation matrix for hourly values

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 3. For this reason the remaining portion of this EDA was devoted to this station.

Figure 2: Data availability of air quality stations

Figure 2: Data availability of air quality stations

Other relevant statistics from the Banquio air quality station are shown in the following Table.

index	AMB_TEMP	NMHC	NO	NO2	NOx	O3	RH	UVB	WD_Hour	WIND_DIREC_10min	WIND_SPEED_10min	WS_HR
std	5.72	0.19	8.74	10.38	16.35	19.41	11.94	2.24	92.38	92.94	1.12	1.03
min	10.00	0.00	-0.40	1.90	3.00	0.20	13.00	0.00	0.20	0.10	0.50	0.00
mean	24.11	0.25	6.16	22.05	28.20	26.44	70.95	1.24	145.83	145.47	2.10	1.72
max	37.00	3.27	212.00	79.00	268.00	144.00	99.00	12.00	360.00	360.00	11.00	9.80
count	8682.00	8619.00	8462.00	8462.00	8462.00	8685.00	8684.00	8684.00	8680.00	8682.00	8682.00	8680.00
75%	28.00	0.30	6.60	28.00	35.00	37.00	80.00	1.50	239.00	239.00	2.70	2.30
50%	25.00	0.19	3.40	21.00	25.00	25.00	73.00	0.00	91.00	92.00	1.90	1.60
25%	19.00	0.13	1.80	14.00	17.00	11.00	62.00	0.00	74.00	73.00	1.30	0.90

Figure 3: Statistics of air quality and relevant meteorological variables from Banqiao station

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 4.

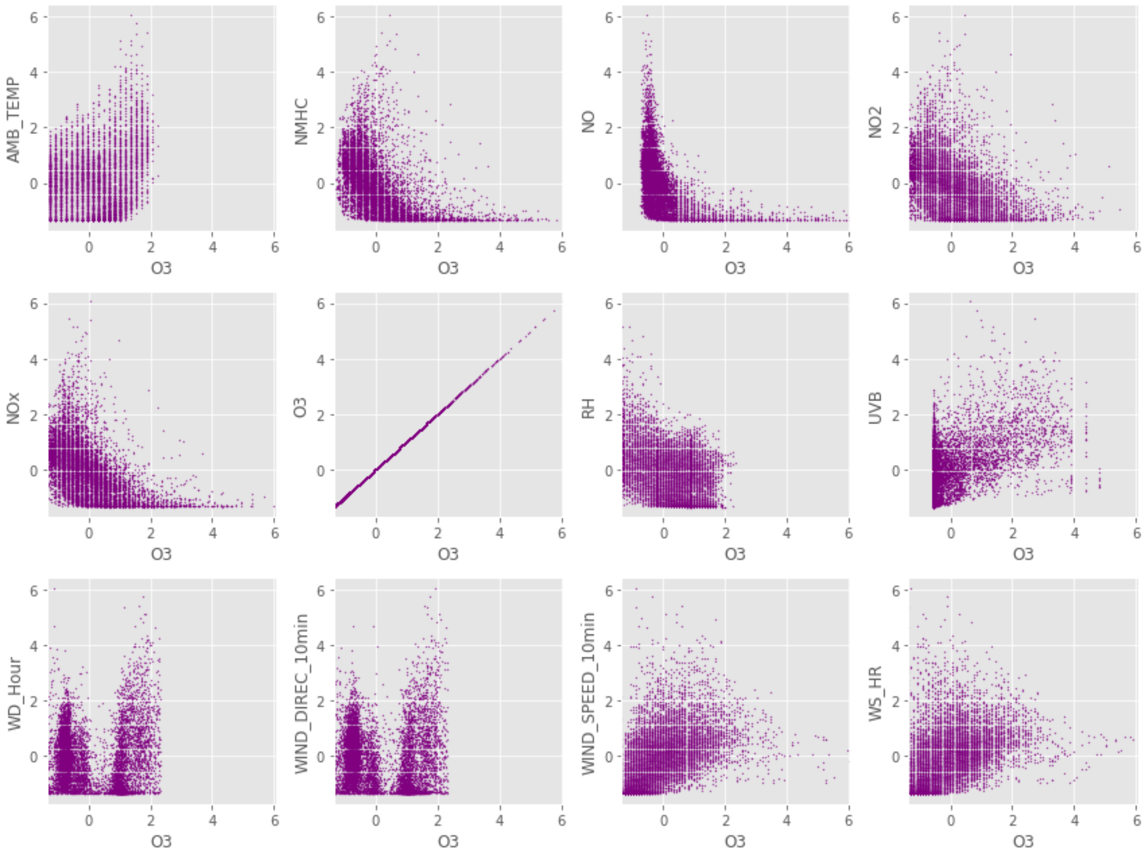

Figure 4: Scatter plots of hourly measurements of variables

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 5.

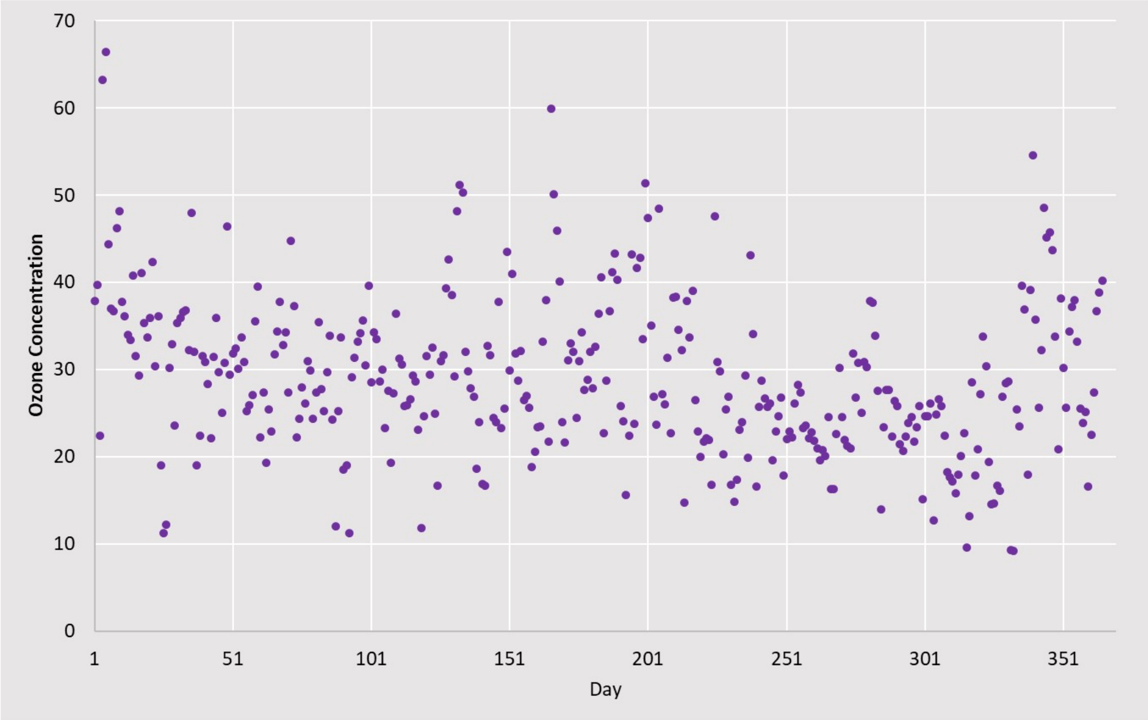

Figure 5: Daily concentrations of ozone

Figure 6: Standardized pollutants and ozone monthly concentration changes

Figure 6: Standardized pollutants and ozone monthly concentration changes

Figure 7: Standardized meteorological measurements and standardized ozone monthly concentration changes

Figure 7: 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 1 (correlation plot). As shown in the figure correlation values for the nitrogen species are negative and vary from -0.41 to -0.5.

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. In the case of UVB a correlation of 0.51 can be observed in 1.

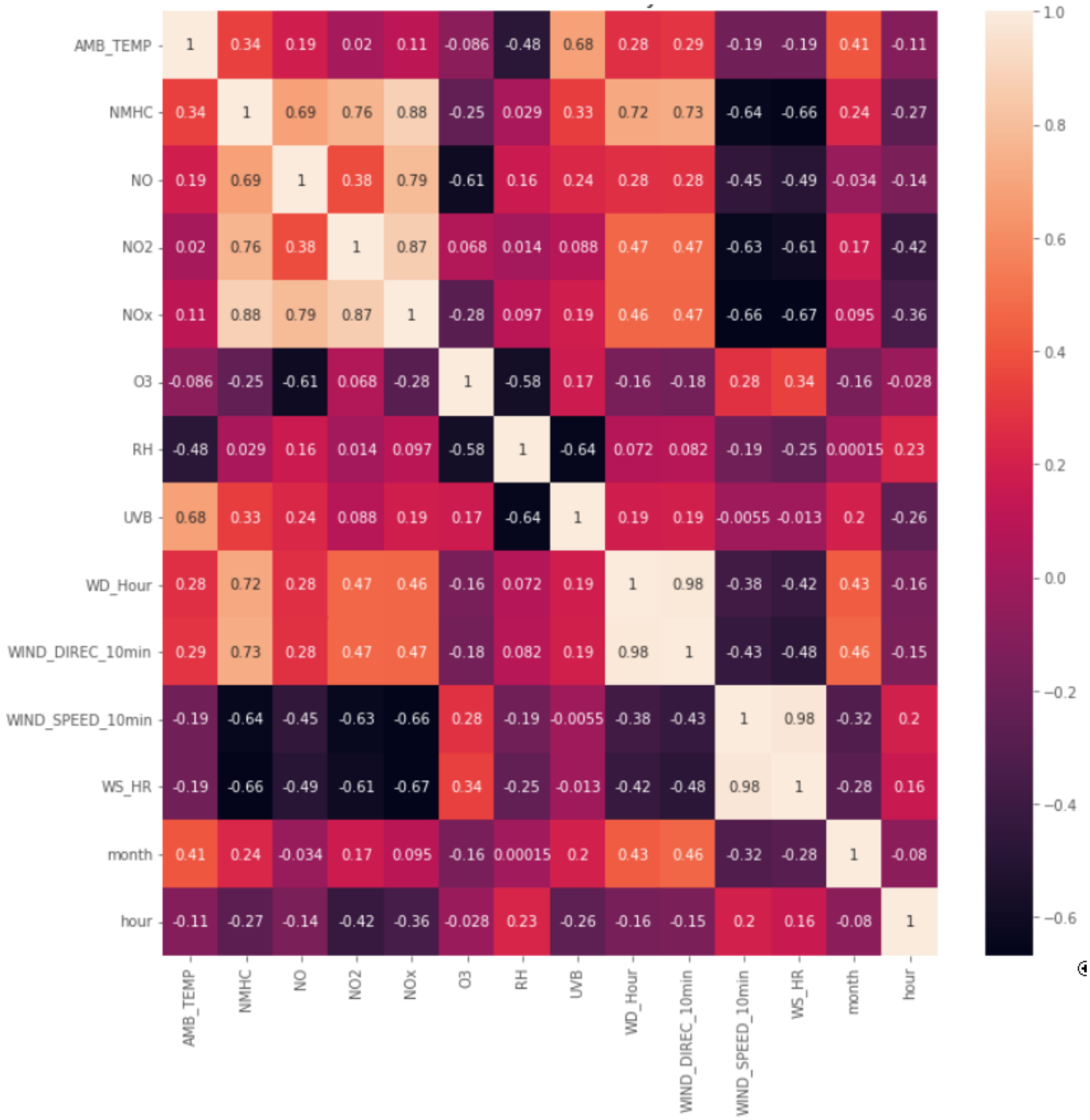

Figure 8: Correlation matrix for daily values

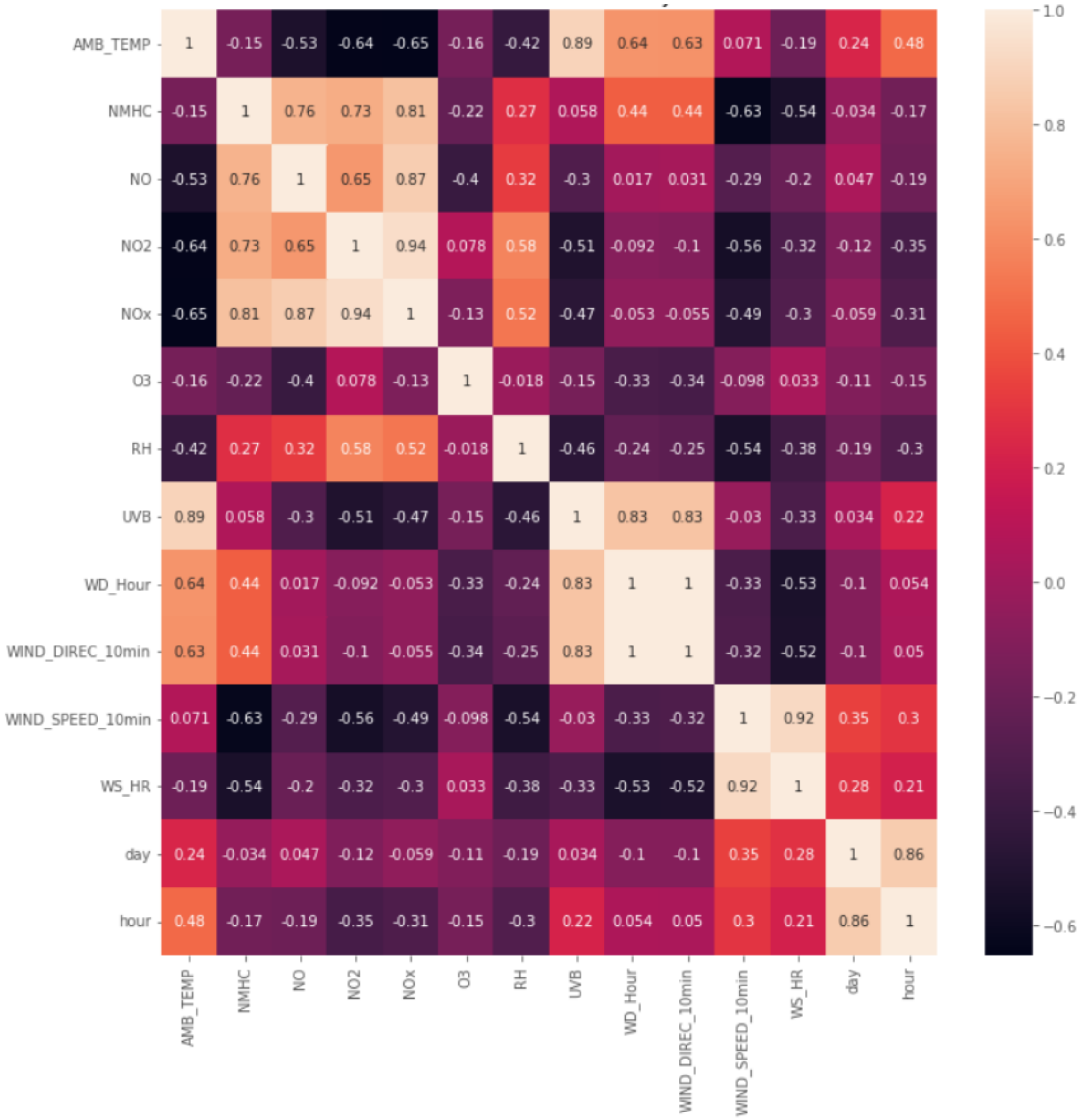

Figure 9: Correlation matrix for monthly values

Predictive Modeling

Two additional correlation matrices were produced. One for monthly average values and another for daily average values. As observed in the figures 1, 8, 9, correlation values of ozone with respect to hourly and monthly values of nitrogen-containing compounds, NMHC and temperature are generally lower compared to correlations values of hourly measurements. Thus, the hourly measurements will be used for predictive modeling.

The low values of the coefficient of determination (r2) between O3 and each of the other variables suggest non-linear relations. Considering the non-linearity of the relation between the predictand O3 and the potential predictors a neural network seems suitable to develop the predictive model. Three different types of neural networks will be tested along with different values of hyperparameters.

Predictive model

Different configurations of neural networks were used, fully connected layers, hereafter called NN for simplicity and convolutional neural networks (CNN). Both neural networks were used to predict hourly concentrations of O_3 (predictand) using as predictors the most correlated variables found in the EDA.

The variables use as predictors were hourly meassurements of: relative humidity (RH), ultraviolet radiation (UVB rays), NMHC, NOx, NO and NO2.

In the case of the NNs, different numbers of layers, neurons and activation functions were tested ??. As seen in the figure, the lowest root mean square error achieved was 6.7 PPM using an 18 neuron 8 layer NN architecture with 1E6 epochs and a learning rate (η) of 1E-4.

Two CNN were also tested. A CNN with 8 fully connected layers of 18 neurons with Relu activation function and a second CNN with one convolution layer of 64 neurons and two fully connected layers with Relu activation function. The two CNNs were trained with 1E3 epochs and η = 1E-4. The corresponding RMSEs were 8.63 PPB and 6.26 PPB. All the error values were calculated using the testing data i.e. the 30% held out fraction of available observations.

For reference, the mean value of ozone concentration in the dataset was 25.0 PPB. Considering the lowest RMSE measured in all the different NN and CNN model was 6.26 PPB, the level of error of the predictive model might be deemed unsatisfactory. This motivated to test another neural network configuration: long-short term memory neural network. The latter is often used to model sequential data, thus instead of a predictive model a forecast model was developed.

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 10: RMSE values for different hyperparameters and NN configurations

Forecast model

Long short-term memory neural networks were use to forecast hourly concentrations of O_3 (predictand) using hourly measurements recorded on the previous 5 hours of different variables. The variables used were: relative humidity (RH), ultraviolet radiation (UVB rays), NMHC, NOx, NO and NO2 and ambient temperature.

Long short-term memory neural networks (LSTMs)

In order to explain the utility of LSTMs a drawback of CNNs have to be discussed. Convolutional neural networks (CNN) use filters to extend the depth of the input volume. One drawback of CNN is that its 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 the exploding/vanishing gradient problem. LSTM has three gates that determined the output: forget gate to determine the percentage of long-term memory that is remembered via a Sigmoid function; input gate to calculate both the potential memory using a Tanh function and the percentage of potential memory that is remembered; and a third gate, called the output gate, to multiply a Tanh function with the long term memory results to obtain the output.

The LSTM tested had one 32 neuron layer, a second 64 neuron layer followed by two dense layers of 8 and 1 neuron, and Relu as activation function. LSTM training was done with hourly measurements from the previous 5 hours to forecast ozone concentration of the 6th hour. RMSE for LSTM on the testing data was 1.76.

The following plots shows predicted O3 values with LSTM and observed O3 values

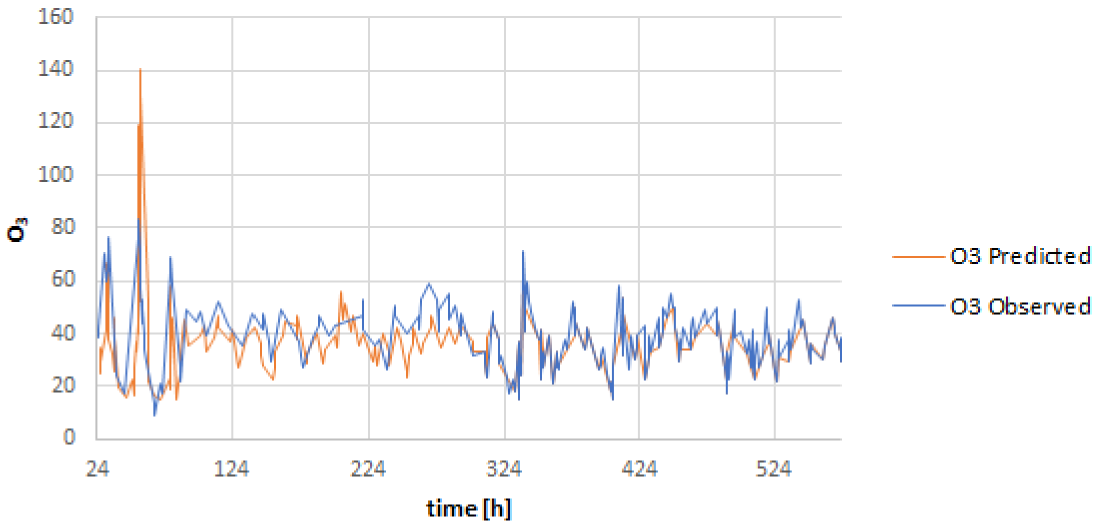

Figure 11: LSTM O3 predicted and observed values

Additional avenues of improvement

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”

Discussion

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”