Improving Ozone Forecasting in the Northeast

CareyAnne Howlett
Northeastern University Khoury College of Computer Science
Boston, MA
DS5500: Capstone Course
EPA Region 1 Pathways Intern

Abstract

Ground-level ozone is known to be harmful to human health in large concentrations. The EPA issues an ozone forecast for the US on a daily basis and is require to alert the public when ozone concentrations are predicted to be high. In Region 1 (the Northeast), there has been some difficulty correctly forecasting days with high ozone levels in southwest Connecticut due to precursor gases being transported from metropolis areas, such as New York City, upwind. The purpose of this research is to use supervised machine learning models to assist in the prediction of high ozone levels in southwest Connecticut. The four different types of models tested were binary classification, multi-classification, regression, and time series. The best models found in each of these categories were SVM, KNN, Linear Regression, and ARIMA respectively.

Introduction

During the summer months, the main pollutant of concern in the Northeast is typically ground-level ozone. In the presence of sunlight, nitrogen oxides (NOx) and volatile organic compounds (VOCs) combine to form ground-level ozone [US 16c]. Ozone is a highly reactive molecule. When the concentrations of ozone are sufficiently high, it is linked to an array of health effects including inflamed airways, difficulty breathing, and increased frequency of asthma attacks [US 16a]. The Environmental Protection Agency (EPA) is required to alert the public when ozone levels are predicted to be unhealthy. This is why it's so important to forecast these events correctly.

Background

In 1970, the Clean Air Act was written into law by President Nixon. It requires, still to this day, for the EPA to set National Ambient Air Quality Standards (NAAQS) for the six principle pollutants that are harmful to public health and the environment [US 14]. The six pollutants are carbon monoxide, lead, nitrogen dioxide, ozone, particulate matter, and sulfur dioxide. For this project, the focus will be on ozone. According to the NAAQS set by the EPA, ozone is considered to be in exceedance when it is found in concentrations greater than 70 parts per billion (ppb) over an 8-hour period [US 15]. Regions around the US are given a status of "attainment" or "nonattainment" for each of the NAAQS pollutants. Using the 70 ppb threshold, a region is considered to be in nonattainment when the average of the annual fourth maximum 8-hour ozone concentration of the previous 3 years is over 70 ppb [US 16b].

AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Figure 1: This table depicts the ozone thresholds the EPA uses to determine the risk for public health.

The EPA issues a forecast for ozone on a daily bases. When concentrations reach a certain level, they are required to alert the public. The EPA uses a threshold categorical system to alert the public of different concentrations of ozone known as the air quality index (AQI) (See Figure 1). However, there have been instances when the EPA has incorrectly forecasted an ozone exceedance day resulting in a high ozone day when the public was not notified. This can be very dangerous, especially for those that already have existing health conditions. This is why improving ozone forecasting in the Northeast is very important.

Related Work

In 2018, the Long Island Sound Tropospheric Ozone Study (LISTOS) identified a large amount of nitrogen oxide (NOx) emissions, an important ozone precursor, in the New York–New Jersey–Connecticut ozone nonattainment area, which is a dense urban area with a population of approximately 20 million people [LaR]. The LISTOS study implicated this plume of NOx as a key reason why ozone monitors along Connecticut's coastline record the highest ozone levels east of the Mississippi River. In addition, several meteorological variables, such as wind (speed and direction),

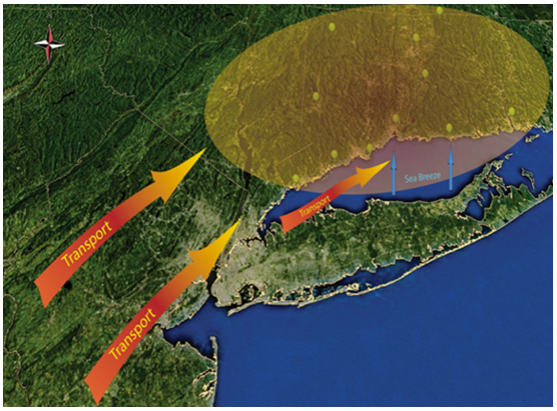


Figure 2: This image shows how poor air quality is transported from Washington, DC and New York City to southwest Connecticut.

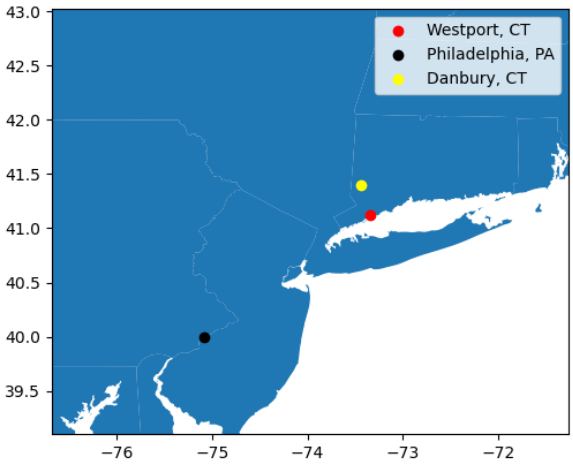


Figure 3: The three points plotted on the map indicate the location of the stations in which the data was used for training the models.

temperature, and humidity, are also factors in determining when these high ozone events set up and ultimately impact Connecticut.

Much of the ozone measured in Connecticut originates outside of Connecticut. During the summer when the sky is clear and the winds are from the southwest, ozone precursors from the megalopolis extending from Washington, DC to New York are transformed into ozone, which impacts Connecticut (See Figure 2). For several reasons, the Long Island Sound enhances ozone formation, with the highest monitors right along the Connecticut coast. A continuation of the work done in the LISTOS study is important to better understand and characterize these events and to, ultimately, support the air quality forecasting community in the Northeast to better forecast when they are likely to happen. In this study, the focus will be to forecast ozone exceedances in Westport, Connecticut that occurred during the summers of 2019, 2020, and 2021.

Project Description

Data

The data used for this study was taken from the EPA's Air Quality System (AQS) pre-generated data files [EPA]. The parameters listed below were packaged in their own respective CSV files containing information on the date and time of the measurement of the given parameter, latitude, longitude, etc. for all the given sites in the US on an hourly time scale for a given year. However, only three stations were used for this study (see Figure 3).

- Surface Temperature
- Wind Speed and Direction
- NO_2 Concentration
- Ozone Concentration
- Barometric Pressure
- Dew Point and Relative Humidity

Preprocessing Before training the models, preprocessing of the data was necessary in order to get the data in a tidy format. To reduce the size of the dataset, a subset was taken that contained data for the Northeast. Since some of the parameters have more than one measurement (for example, the wind data contained measurements for the speed and direction), these needed to be extracted from the files and merged correctly on the date, time, and location of that given parameter. Once this was complete, the same "tidying" techniques could be used on the other files downloaded for the same parameters for different year (2019, 2020, and 2021). When all the files are in the same tidy format for a given parameter, they then can be concatenated together to make a complete time series. The complete time series of each parameter was then written to a temporary CSV to be combined with all the other parameters after all of them were in a tidy format. Once all the temporary files were merged, a final dataset was created containing all the measurements of the parameters in a tidy format. More specifically, the data contained observations for all the parameters listed above for stations in the Northeast on an hourly time scale from 2019 through 2021.

Once the data were in a tidy format, two columns were added to define the classes for the binary and multi-classification models for ease. More information about these classes and their thresholds will be discussed in the Methods section. After these two columns were added, the final dataset was written to a CSV file to be used to train the models in phases 1 and 2.

Exploratory Data Analysis After achieving a reduced and tidy format, exploratory data analysis was done to further inspect the data. It was found that the barometric pressure, relative humidity, and dew point measurements were nearly all missing. In the interest of keeping a complete dataset, these parameters were removed.

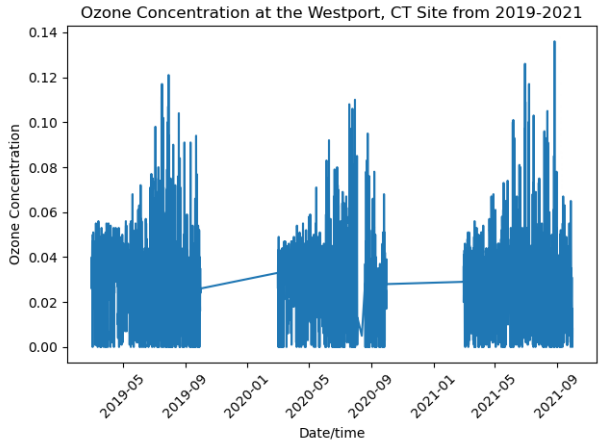


Figure 4: Depicted are the three years of data used to train the models. It is clear that the ozone monitor at the Westport, Connecticut site was only turned on during the "ozone season" which is defined to be between March 1st and September 30th. It is also clear that there is some missing data found in August of 2020.

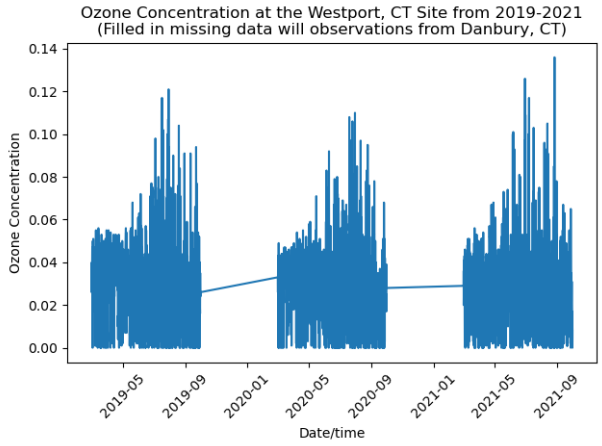


Figure 5: Depicting the three years of data used to train the models. It is clear that the ozone monitor at the Westport, Connecticut site is only turned on during the "ozone season" which is defined to be between March 1st and September 30th. It is also clear that there is some missing data found in August of 2020.

Under further inspection, it was found that the ozone monitor at the Westport, Connecticut site (the target variable) was only turned on for the ozone season which is define to be between March 1st and September 30th which is seen clearly in Figure 4. Though this was not something that would affect the study, it is certainly a factor to keep in mind for the time series portion of the study. In Figure 4 it is also clear that there is some missing data found in August of 2020. In the effort of keeping the dataset as complete as possible, these missing datapoints were filled in by a neighbouring station in Danbury, Connecticut. Figure 5 depicts the complete ozone concentration for Westport, Connecticut which will be the target variable.

The project has been split into 2 phases. Phase 1 focused on binary and multi-classification models. Instead of predicting the ozone concentration itself, these models predicted the class of the ozone concentration. Phase 2 focused on regression and time series models, meaning the models predicted the ozone concentrations directly. From these four categories, the models were compared against each other to determine which preformed the best.

Phase 1

In an effort to keep the models simple for phase 1, the problem was turned into a classification problem rather than regression. In phase 1, an array of models are trained to predict the ozone classification category (the target variable) using only the data at the Westport, Connecticut site. For binary classification, the ozone concentrations greater than 50 ppb were classified as class 1 and concentrations less than or equal to 50 ppb were classified as class 0. The threshold of 50 ppb was chosen because concentrations over the 50 ppb threshold is when ozone levels start to become a concern.

The variables the models were trained on were NO_2 surface concentrations and meteorological observations, such as wind speed, wind direction and surface temperature, to predict the ozone classification category at the Westport, Connecticut site. The supervised machine learning models that were trained were Logistic Regression, Decision Trees, SVM, and KNN. These models were chosen because they are known to work well with classification problems. To evaluate the performance of each model, the F1 scores were compared.

Initially the models were trained on a dataset that spanned from 2020 through 2021 and used under and over sampling to balance the data. After reviewing the results, it was clear that the models were overfitting with both under and oversampling. To combat this, an additional year of data was added to the dataset and Synthetic Minority Oversampling Technique (SMOTE) was used to balance the data. Since the overfitting problem was fixed by adding an additional year of data and using SMOTE, the dataset used going forward spanned from 2019 through 2021.

As mentioned before, the EPA uses a categorical system to alert the public for different concentration levels of ozone (see Figure 1). These thresholds were implemented to create a multi-classification problem to predict the AQI level indicated in Figure 1. The F1 scores were used to evaluate the multi-classification model performance. The models

mentioned above were also used for multi-classification. All models in phase 1 were trained and tested with an 80% and 20% split of the data respectively.

Since it is thought that most of the ozone that impacts the Connecticut coast originates outside the state, additional data were added to the dataset in an effort to improve the models performance. Data from the Philadelphia, Pennsylvania site, including wind speed, wind direction, surface temperature, NO_2, and ozone concentration, was added as features to the dataset.

All the binary classification and multi-classification model performances were compared for both datasets.

Phase 2

In phase 2, regression and time series models were tested. Here, the target variable was the ozone concentration at the Westport, Connecticut site. The regression models that were trained using 80% of the data and tested on the remaining 20% were Linear Regression, Random Forests, Decision Trees, and Gradient Boosting.

The two time series models tested were Auto Regressive Integrated Moving Average (ARIMA) and Holt Winter's Exponential Smoothing (HWES). Before running these models, it is necessary to check if the data was stationary. To do this, the Augmented Dickey-Fuller test was done.

With the nature of the ARIMA model, only the target value itself could be used to train the model. The python function `auto_arima()` was used to get the optimal hyperparameters for the model. Different iterations were run to find the optimal lag (or `m` value) as this parameter was put in manually.

The HWES model was also run with just the Westport, Connecticut ozone concentration due to time constrains of the project.

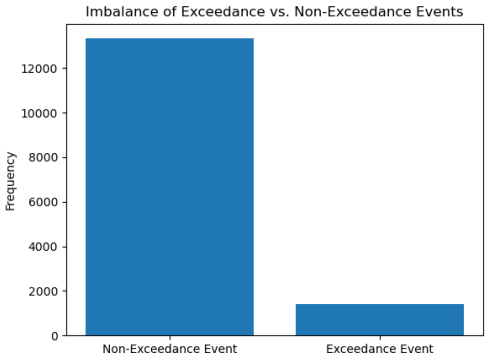
The RMSE was used to evaluate model performance of all the model tested in phase 2. They were compared to determine which gave the best results.

Empirical Results

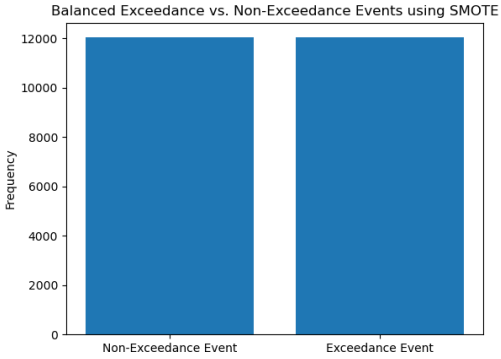
Phase 1

After training the binary classification models on the Westport, Connecticut data, the F1 score was calculated to assess the model performance. The scores are listed in Table 1. Under and over sampling was used to balance the dataset. However, from the results shown in Table 1, it is clear that the models were overfitting. In an effort to combat this, SMOTE was implemented as well as adding an additional years worth of data to the dataset. Shown in Figure 6, SMOTE was used to balance the respective classes of each binary and multi-classification set. By doing this, it seemed to improve the overfitting problem (see Table 2). Going forward, the data spanning from 2019 through 2021 was used as well as SMOTE to deal with imbalanced data.

Next, the same models were run with the multi-classification dataset using the AQI indices in Figure 1. The results from these models can be found in Table 3. Since it is thought that most of the ozone impacting the Westport, Connecticut originates from outside the state [LaR], additional



(a) The data on it's own was imbalanced as there were more non-exceedance events than there were exceedance events.



(b) By using SMOTE, the imbalance of the data was neutralized.

Figure 6: SMOTE was implemented to balanced the binary classes that define exceedance versus non-exceedance events.

data was added to the dataset in the effort to improve the model performances. In Table 4, the results indicate a slight increase in performance. Because of this, the additional data was used to train other models going forward.

Phase 2

As stated before, the regression models were run with the 2019 through 2021 data including observations from Westport, Connecticut and Philadelphia, Pennsylvania. From the results shown in Table 5 it is clear that Gradient Boosting gives the best result without severely overfitting.

For the time series models, only the ozone concentration at the Westport, Connecticut site could be used due to the nature of these models. As mentioned before, it was necessary to check if the data was stationary before running the time series models. The Augmented Dickey-Fuller test was run and gave a resulting p-value less than 0.05 meaning the dataset was stationary. The `auto_arima()` funtion was run to determine the right hyperparameters needed to optimize the ARIMA model. However, the lag, or the number of previous observations used to make a prediction, needed

Binary Classification: 2020-2021 Under and Over sampling						
	Imbalanced		Oversampling		Undersampling	
	Train	Test	Train	Test	Train	Test
Logistic	0.4464	0.4245	0.8183	0.7914	0.8155	0.8205
Decision Trees	1	0.4756	1	0.7687	1	0.9733
SVM	0	0	0.8447	0.7988	0.8275	0.8374
KNN	0.6648	0.5247	0.8861	0.8025	0.9533	0.9394

Table 1: Depicted above are the F1 score results from running the binary classification models with just the Westport, Connecticut data from 2020 through 2021. Under and over sampling was used to balance the dataset to see which one gave better results.

Binary Classification: 2019-2021, Westport				
	Imbalanced		Balanced	
	Train	Test	Train	Test
Logistic Regression	0.4066	0.4066	0.5948	0.6016
Decision Trees	1	0.5253	1	0.6529
SVM	0	0	0.6522	0.6359
KNN	0.6778	0.5682	0.8102	0.6979

Table 2: Depicted above are the F1 score results from running the binary classification models with just the Westport, Connecticut data from 2019 through 2021. SMOTE was used to balance the dataset.

Multi-Classification: 2019-2021, Westport				
	Imbalanced		Balanced	
	Train	Test	Train	Test
Logistic Regression	0.9130	0.9123	0.7155	0.7157
Decision Trees	1	0.9062	1	0.9437
SVM	0.9047	0.8938	0.8423	0.8389
KNN	0.9445	0.9193	0.9561	0.9346

Table 3: Depicted above are the F1 score results from running the multi-classification models with just the Westport, Connecticut data from 2019 through 2021. SMOTE was used to balance the dataset.

Multi-Classification: 2019-2021, Westport and Philadelphia				
	Imbalanced		Balanced	
	Train	Test	Train	Test
Logistic Regression	0.9073	0.9008	0.7175	0.7161
Decision Trees	1	0.9282	1	0.9760
SVM	0.9080	0.9021	0.8606	0.8538
KNN	0.9473	0.9289	0.9812	0.9685

Table 4: Depicted above are the F1 score results from running the multi-classification models with Westport, Connecticut and Philadelphia, Pennsylvania data from 2019 through 2021. SMOTE was used to balance the dataset.

Regression: 2019-2021, Westport and Philadelphia		
	Train	Test
Linear Regression	0.006799	0.006918
Random Forests	0.002054	0.005531
Decision Trees	7.09e-19	0.007947
Gradient Boosting	0.005473	0.006068

Table 5: Depicted above are the RMSE results from running the regression models with Westport, Connecticut and Philadelphia, Pennsylvania data from 2019 through 2021.

Time Series: 2019-2021, Westport Ozone Concentration	
Auto Regressive Integrated Moving Average	0.007358
Holt Winter's Exponential Smoothing	0.007807

Table 6: Depicted above are the RMSE results from running the time series models with Westport, Connecticut ozone concentration data from 2019 through 2021.

to be specified. Different trials were done to find the optimal lag. For instance, a lag of `m = 72` (previous 3 days) was tested and gave an RMSE of 0.01502. After setting `m = 24`, the ARIMA result in Table 6 was the best result found. After optimization, the final ARIMA parameters used were `ARIMA(1, 1, 0)(2, 0, 1)[24]`. The HWES model used similar parameters and gave similar results as the ARIMA model.

Conclusion and Future Work

The best results found for the binary and multi-classification models were SVM and KNN respectively with out severely overfitting. It was clear from the results that Decision Trees was not a good model for this data because it overfit the data significantly in all instances.

For the regression and time series models, the Linear Regression and ARIMA models were the best respectively with out overfitting. Gradient Boosting gave the best results, however it did overfit. If time had allowed, some hyperparameter tuning could have made this model one of the best in this category.

Some ways this study could be improved is to refine the input data used to train the models. More specifically, other stations around the Northeast should be used to train the models to check if better performance metrics can be achieved. Also, filling in the missing data that was initially removed could also help the model performance. If more time allowed, Long Short-Term Memory (LSTM) would have been included in the study to see how its model performance compared to the others tested. Additional future work would be to create a GUI that the EPA can use to help predict the ozone concentrations for the following day. Ideally, the data would be fed into the GUI, and the models would predict the ozone concentration for the following day.

With the models created in this study, the EPA could use them as a tool to help predict ozone concentrations during their daily forecasts. This would aid in the prediction of ozone exceedance days in the Northeast, ultimately, keep-

ing the public safe.

References

[US 14] OAR US EPA. *NAAQS Table*. en. Other Policies and Guidance. Apr. 2014. URL: https://www.epa.gov/criteria-air-pollutants/naaqs-table (visited on 06/25/2022).

[US 15] OAR US EPA. *Timeline of Ozone National Ambient Air Quality Standards (NAAQS)*. en. Data and Tools. Dec. 2015. URL: https://www.epa.gov/ground-level-ozone-pollution/timeline-ozone-national-ambient-air-quality-standards-naaqs (visited on 06/25/2022).

[US 16a] OAR US EPA. *Health Effects of Ozone in the General Population*. en. Data and Tools. Mar. 2016. URL: https://www.epa.gov/ozone-pollution-and-your-patients-health/health-effects-ozone-general-population (visited on 06/25/2022).

[US 16b] OAR US EPA. *Nonattainment Areas for Criteria Pollutants (Green Book)*. en. Collections and Lists. Apr. 2016. URL: https://www.epa.gov/green-book (visited on 06/25/2022).

[US 16c] OAR US EPA. *What is Ozone?* en. Data and Tools. Mar. 2016. URL: https://www.epa.gov/ozone-pollution-and-your-patients-health/what-ozone (visited on 06/25/2022).

[EPA] EPA. *AirData website File Download page*. en. Data & Tools. URL: https://aqs.epa.gov/aqsweb/airdata/download_files.html (visited on 08/14/2022).

[LaR] Ali Aknan : NASA LaRC. *NASA Airborne Science Data for Atmospheric Composition*. URL: https://www-air.larc.nasa.gov/missions/listos/index.html (visited on 06/25/2022).