Development and Implementation of Machine Learning Tools for Ozone Formation in the LADCO Region

Hantao Wang

hwang819@email.unc.edu

University of North Carolina at Chapel Hill

2022 summer internship report

prepared for Lake Michigan Air Directors Consortium (LADCO)

Abstract

Recently, the United States Environmental Protection Agency (EPA) issues periodic reports that describe meteorologically adjusted ozone trends with ground observations in the US. However, the influence of meteorological conditions on ozone formation is still less clear. The main objective of this internship is to apply the generalized additive model (GAM) analysis to different areas in the LADCO region and to adjust the annual ozone trends for meteorology in these areas. During this internship, we developed and extended a machine learning tool to analyze the nonlinear effects of meteorological variables in the LADCO region (Sheboygan and SWFP). We then hope to use these observations and results to better understand the influence of meteorological conditions on ozone formation in the LADCO region. We evaluated effects of meteorological variables on ozone formation by the partial response from the model, with respect to the daily maximum 8 h average (MDA8) observations from 2000 to 2019. Here, a generalized linear model was used to analyze the correlation between MDA8 ozone observations and meteorological conditions, as well as the temporal data. Quantile regression was employed to examine the reliability of meteorological data at various ozone concentrations. Our results highlight that 90th and 98th percentile meteorology adjusted ozone trends are flatter at SWFP but has decreased a lot at Sheboygan. In general, interannual variability is much less at the 50th percentile level in both regions. This result suggests that high ozone concentrations are continuing to decrease at Sheboygan much more than at SWFP, and the effect of meteorology is stronger on peak MDA8 ozone conditions.

Development and Implementation of Machine Learning Tools for Ozone Formation in the

LADCO Region

1. Background

1.1 Overview of Ozone in the LADCO Region

Ground-level ozone is a secondary pollutant formed by photochemical reactions of precursor pollutants and is not directly emitted from sources. Anthropogenic sources of ozone are formed through photochemical reactions of precursor pollutants, which include volatile organic compounds (VOCs), and nitrogen oxides (NO_X), and carbon monoxide (CO).¹ Evidence from atmospheric chemistry and exposure assessment studies indicates that short-term (i.e., hours, days, weeks) or long-term (i.e., months to years) exposure to ozone above EPA standards (70 ppb) can have health effects ranging in severity from mild subclinical effects to mortality.^{2,3} Ambient ozone is estimated to have caused about 365,000 deaths globally or 0.65% of all global deaths in 2019. 4

Ground-level Ozone observations at monitoring sites in the LADCO region have consistently violated the National Ambient Air Quality Standards (NAAQS) over the past 40 years. From Figures 1 and 2, the monitoring sites with the high ozone concentrations are typically located downwind of major source areas, suggesting that long-range transport of emissions may be responsible for ozone formation. As of May 2022, 11 areas in the LADCO region were designated as nonattainment for the 2015 Ozone NAAQS (70ppb). The designation of nonattainment areas requires states to reduce ozone concentrations to meet NAAQS. Therefore, it is important to understand the factors that drive ozone formation in the LADCO region.

1.2 Adjusted ozone trends by Meteorology Effects

As can be seen in Figure 1, ozone concentrations are higher around Lake Michigan than in other areas, especially along the lake shore in Wisconsin and western Michigan. Ozone precursors emitted from other areas are blown onto the lake by onshore winds, and they react in the shallow marine boundary layer to form ozone that accumulate around Lake Michigan. Similar lake drivers influence ozone formation in the LADCO region, which in turn affects daily average ozone concentrations. Therefore, in order to develop reasonable plans to reduce ozone concentrations in these areas, the effects of meteorological variations on ozone need to be adjusted. The adjustment of long-term trends in ozone concentrations to the effects of meteorological changes has been the subject of scientific research since the early 1990s.⁵ The EPA issues periodic ozone trend reports that reconcile the effects of fluctuating meteorological conditions, and the results of these periodic reports are typically based on a relatively limited number of meteorological parameters.⁶ However, these reports may not explain specific regions more broadly, as Ozone formation in different locations may be driven by different meteorological conditions. In recent years, researchers have used the latest statistical and numerical methods to improve estimates of adjusted ozone trends in U.S., such as quantile regression and machine learning algorithms, and have also applied variable selection to identify important meteorological factors in different geographical areas. 7

$$
Adj(O_{3i}) = \exp\left(\mu + \sum_{j=1}^{n} f_j(x_j)\right)
$$

Above is the function to calculate adjusted mean of ozone by meteorology in this internship. Adj (O_{3i}) is the adjusted mean of MDA8 ozone on day "i", and exp is the natural exponential function. μ is the intercept of the model, which represents the long-term average of 8-hour peak ozone. x_j is meteorological variable "j", and $f_j(x_j)$ is the function generated by the generalized linear model in section 2.2, which parameterizes the relationship between meteorological variable "j" and 8-hour peak ozone on day "i".

1.3 Generalized Additive Model (GAM)

Machine learning has been widely used to analyze ozone formation due to the ability to fit the nonlinear dependence of ozone on predictor variables using indeterminate curves. A generalized additive model (GAM) was developed by Camalier et al. in 2007 and used by EPA to determine ozone trends adjusted for weather variability. The model was cited by EPA (2018) as the weight of evidence (WoE) for ozone attainment arguments analysis. The EPA GAM used a natural curve, and by analyzing the model parameter to distinguish the observed sensitivity of ozone to different weather variables. Dr. Charles L. Blanchard developed and extended this EPA GAM in 2020 to describe the relative impact of weather, emissions on MDA8 ozone in the southern Lake Michigan region (under contract to LADCO and WDNR). We further extended the GAM based on Dr. Blanchard's work to analyze the relative impact of weather fluctuations on MDA8 ozone at two sites (Sheboygan, SWFP) in the Lake Michigan region and used quantile regression to determine the adjusted value for weather variability on different ozone concentrations.

2. Methods

2.1 Input data

In this internship, we applied the GAM analysis in the LADCO region (Sheboygan and SWFP). We collected input datasets for these 2 sites. Each input dataset has the same file format, independent variables, and dependent variables, so that only the input datasets need to be changed when applying the model to other regions. The dependent variable is the MDA8 Ozone at the monitoring site, and there are 65 independent variables in 7305 days, which are ozone precursor concentrations, meteorological conditions, and time variables.

Ozone precursor concentrations were used as independent variables for the data inputs. Annual average emissions of VOCs and NOx were included to account for long-term emission trends. In addition, multisite daily average concentrations of CO , NOx , and $SO₂$ (primarily at sites in Chicago and Milwaukee) were used to represent daily changes in precursors upwind of the city. Carbon monoxide data were used to represent mobile source VOC_s emissions.

Meteorological conditions were used as input independent variables. Upper air measurements were taken from NOAA and these included daily maximum temperature (Tmax), precipitation, daily average wind speed, surface-level barometric pressure (BP), and direction of the fastest gust in 2 minutes. Daily average relative humidity (RH) was from surface measurements at air quality monitoring sites. Solar radiation measurements were from NREL's solar radiation model (https://www.nrel.gov/docs/fy12osti/54824.pdf). Surface water temperatures for Lake Michigan were obtained from NOAA [\(https://coastwatch.glerl.noaa.gov/statistic/\)](https://coastwatch.glerl.noaa.gov/statistic/).

Temporal variables were also used as input data. The start time of each MDA8 Ozone was included as a predictor variable. Other temporal variables include year, month, day of week, and day of year.

2.2 Generalized linear model

$$
L(O_{3i}) = \mu + \sum_{j=1}^{n} f_j(x_j)_i + \sum_{k=1}^{m} g_k(y_k)_i + \sum_{p=1}^{l} h_p(z_p)_i + e_l
$$

According to Dr. Charles L. Blanchard, $L(O_{3i})$ is the logarithm of the 8-hour peak ozone on day "i". The variable μ is the intercept of the model, which represents the long-term average of 8hour peak ozone. The variable x_i is meteorological variable "j", the term $f_i(x_i)$ parameterizes the relationship between meteorological variable "j" and 8-hour peak ozone on day "i". The variable y_k is the logarithm of the concentration of emission "k", $g_k(y_k)$ parameters the relationship between ambient concentrations of ozone precursor "k" and 8-hour peak ozone on day "i". The term z_p represents the temporal variable "p", including "day of the week" and "year", $h_p(z_p)$ i parameters the relationship between time variable "p" and 8-hour peak ozone on day "i". The last variable e_i is the difference between observed and predicted 8-hour peak ozone on day "i". Each function "f", "g", "h" is generated by the generalized linear model. Each term parameterizes the response of the daily 8-hour peak ozone as a deviation from the long-term average. Since it is the observed long-term average ozone, its value is independent of the choice of parameters in the model.

2.3 Variables selection

Dr. Charles L. Blanchard found that many candidate input variables are highly correlated (e.g., $r2 \sim 0.8$). Highly correlated predictor variables usually have unstable coefficients, i.e., removing one of the two correlated variables changes the coefficient of the other. Therefore Dr. Charles L. Blanchard retained only one predictor variable per set of correlated predictor variables in his previous work. We performed further variable selection on this basis. Because the sensitivity of ozone to meteorological variables varies in each region, a method is explored here to determine which variables are most closely related to the observed MDA8 ozone concentrations. As can be seen in Table 1&2, we determined the importance of each meteorological variable in the model using the Akaike Information Criterion (AIC) and analysis of variance (ANOVA). The AIC is an estimated measure of the quality of the model after each variable is added and represents a tradeoff between the complexity of the model and its goodness of fit. When the AIC is smaller, it means that the variable is better to be added to the model. ANOVA is a hypothesis test based on

the parameters of a variable with the null hypothesis that the parameters of that variable are equal to 0. The F-values in Table 1&2 are the results of ANOVA, and the larger the value the less likely it is that the parameters of the variable are 0. Collectively, when the AIC value of a weather variable is smaller, the larger the F-value, the higher the importance of this variable in the model. By this method, our model automatically outputs the results of the analysis of each independent variables and selects the meteorological variables that have the greatest impact on ozone in each area. Then use them to calculate the adjusted ozone mean trends. Rather than using all meteorological variables together, this approach takes into account the variability of each region and reduces overfitting by reducing the complexity of the model.⁸

2.4 Quantile regression

In this internship, we added a quantile regression of meteorological variables on ozone observations for 50%, 90% and 98% quantile. Our model uses a nonlinear optimization to support different magnitudes and the ability to change the weather conditions in the regression when needed by selecting variables. Based on the results of the quantile regression, we used the function in section 1.2 to calculate the adjusted ozone values for different concentrations of ozone conditions.

3. Results

3.1 Model Performance

In previous work, Dr. Charles L. Blanchard used a variety of methods and metrics to assess the quality of the model fit. Table 4. GAM performance summary in Dr. Charles L. Blanchard's 2020 report summarizes model performance for each of the 20 sites in terms of fit, correct model prediction above or below the MDA8 O3 threshold of 70 ppbv, false alarm rate, probability of detection, and composite success index.

In this internship, the GAM in section 2.2 were fitted using data from 2000-2019 for all available days (7305 days) at each site which is same as Dr. Charles L. Blanchard's work. The logtransformed MDA8 ozone was fitted to the model. Figure.3 shows the regression between fitting value and MDA8 ozone observations, a high R-square (0.725) obtained, which means that GAM works properly.

3.2 Partial response of meteorological variables

In this internship, we analyzed the relationship between different meteorological variables and ozone concentration, which can be seen in Figures 4 to 10. Because the relationship between meteorological variables and ozone concentration is similar in Sheboygan and SWFP areas, we use Sheboygan as an example to illustrate it for convenience. The horizontal coordinates represent the values of different meteorological variables. The vertical coordinate represents the effect of the variable on ozone and is the ratio of the predicted ozone concentration for that condition to the average MDA8 ozone concentration for that year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

From Figure 4, it can be seen that as the height of 850 millibar increases, the effect on ozone changes from negative to positive as the barometric pressure decreases, increasing all the time, and the greatest suppression of ozone production occurs when the height of 850 millibar equals about 1400meters. This is consistent with the results from Figure 10, where surface-level barometric pressure (SondeBP) has a negative effect on ozone, and the negative effect increases with the increase of air pressure. As can be seen in Figure 5, relative humidity has a negative effect on ozone, and the negative effect increases with increasing humidity. From Fig. 6, it can

be seen that the effect on ozone keeps increasing with increasing solar radiation, but the limit is 1.05. From Fig. 7, it can be seen that ozone increases in a specific range of wind directions, while it keeps decreasing in other wind directions. This is related to the direction of the lake wind at the site. From Fig. 8, it can be seen that as the wind speed increases, its negative effect on ozone decreases, while the positive effect keeps increasing. As can be seen in Figure 9, the effect on ozone changes from positive to negative as the temperature increases and then increases again, with the greatest suppression of ozone production occurring at about 17 degrees.

3.3 Adjustment of ozone trends for meteorology

Our script allows the selection of different meteorological variables to calculate the adjusted ozone mean. Therefore, according to the method of selecting variables in section 2.3, we selected different meteorological variables to calculate adjusted ozone mean for Sheboygan and SWFP. In the Sheboygan area, we selected Tmax, SR_max, MeanRH, WS.local, and WD.local. In the SWFP area, we selected MeanRH, Tmax, SR_max, Ht850mb, and WD.local.

Figure 11 shows the annual ozone adjusted values for the Sheboygan area by reducing the effect of fluctuations in meteorological conditions. the observed values for Sheboygan have been fluctuating and it is difficult to see whether the ozone in this area is in an increasing or decreasing trend. By comparing the observed and adjusted values it can be seen that the annual average of ozone concentration becomes smoothed after eliminating the changes in meteorological variables and is in a slow increasing trend from 2000 to 2019. By adjusting for fluctuations in meteorological conditions, we are able to better detect the trend of ozone concentration in the region.

Figure 12 shows the annual ozone adjusted values for the SWFP region by reducing the effect of fluctuations in meteorological conditions. By comparing the observed and adjusted values, we

can find that the annual average of ozone concentration is still in constant fluctuation from 2000 to 2019, although it becomes smooth after eliminating the variation of meteorological variables. This result may due to the meteorological variables do not change much in SWFP.

3.4 Adjustment value of ozone for different percentile

In this internship, we calculate the adjusted values of meteorological variables for different ozone concentrations by the results of quantile regression. Although the quantile regression is a poorer fit than GAM, this method allows us to analyze the effects of fluctuations in meteorological conditions under peak MDA8 ozone conditions. By the method in section 2.3, we select the more important meteorological variables, and since the fit is weaker than the GAM, more meteorological variables are included in the calculation of the ozone adjustment values. The variation of ozone concentration excluding the effect of meteorological changes in Sheboygan area is shown in Figure 13. After adjusting for meteorological variables, the fluctuations in ozone concentrations become smoother. Unlike the adjusted mean trend, we can see that the meteorologically adjusted ozone concentrations are clearly in a decreasing trend under the peak MDA8 ozone condition, which indicates that the high ozone concentrations in the Sheboygan area are decreasing year by year.

The variation of ozone concentration excluding the effect of meteorological changes in the SWFP area is shown in Figure 14. It can be seen that the observed values of the peak MDA8 ozone in the SWFP area fluctuate particularly widely, and the fluctuation of ozone concentration becomes smoother after the adjustment of meteorological variables. Unlike the adjusted mean trend, the difference between the adjusted and observed ozone values under the peak MDA8 ozone condition is large, which indicates that the sensitivity of ozone concentrations to changes

in meteorological conditions is higher than usual under high ozone concentration conditions in the SWFP region.

Overall, interannual variability at the 50th percentile level is much smaller, while ozone at the 90th and 98th percentile is flatter at SWFP, but declines considerably at Sheboygan. This suggests that the high ozone concentrations at Sheboygan continue to decline much more than at SWFP (which is consistent with other LADCO analyses).

4. Discussion

The main objective of this internship is to understand the trend of average ozone concentration after adjusting for meteorological conditions. To understand the trends of ozone in different LADCO regions by reducing the fluctuating disturbances of meteorological conditions. Trends in MDA8 peak ozone concentrations are adjusted by implementing quantile regression methods. These trends can help air quality modelers understand trends in peak ozone levels for a given year excluding fluctuations in meteorological conditions. We observed much larger reductions in meteorologically adjusted ozone at the 90th and 98th percentiles for the Sheboygan site than for the SWFP, indicating that high peak ozone concentrations decreased more in the Sheboygan area than in the SWFP.

A major advantage of the GAM used in this internship over other nonlinear statistical methods is that it can handle linear and nonlinear relationships between response variables and independent indicators⁹. We used AIC and ANOVA to select the meteorology variables. This internship has some improvements compared to previous studies that Dr. Charles L. Blanchard's GAM to model ozone concentrations. We performed quantile regressions of MDA8 ozone concentrations with meteorological variables from 2000 to 2019 and analyzed how the peak MDA8 ozone were

influenced by meteorological variations for the 90%, 98% quantile. The methods in this internship can be extend to model any air pollutant at other regions.

There are still many limits in this internship that can be improved in future study. First, the fit of quantile regression is weaker than GAM this may affect the accuracy of the prediction. We can improve the prediction accuracy and reduce the bias by combining GAM and quantile regression. Second, we did not add automated selecting variables to the GAM and still need to rely on our own evaluation to select variables. Because of the limitations of evaluating a variable only by AIC or ANOVA alone, we need to consider the weights of these two parameters and combine past studies to evaluate meteorological variables. This can be automated by machine learning tools to evaluate the importance of variables and then censoring in future study. Third, GAM analyzes the area where one monitoring station is located at a time and does not take into account the geographic distribution of ozone near each monitoring site. This is a limitation of the GAM model itself, which does not take spatial information into account. This aspect can be improved in the future by combining geostatistics methods.

References

[1] Crutzen, P. J. (1973). "A discussion of the chemistry of some minor constituents in the stratosphere and troposphere." Pure and Applied Geophysics 106-108: 1385-1399.

[2] Que, L. G., et al. (2011). "Pulmonary function, bronchial reactivity, and epithelial permeability are response phenotypes to ozone and develop differentially in healthy humans." J Appl Physiol (1985) 111(3): 679-687.

[3] Devlin, R. B., et al. (2012). "Controlled exposure of healthy young volunteers to ozone causes cardiovascular effects." Circulation 126(1): 104-111.

[4] GBD 2019 Risk Factor Collaborators (2020) Global burden of 87 risk factors in 204countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet, 396: 1223-49, doi: 10.1016/S0140-6736(20)30752-2

[5] Lou Thompson, M., et al. (2001). "A review of statistical methods for the meteorological adjustment of tropospheric ozone." Atmospheric Environment 35(3): 617-630.

[6] Camalier, L., et al. (2007). "The effects of meteorology on ozone in urban areas and their use in assessing ozone trends." Atmospheric Environment 41(33): 7127-7137.

[7] Porter, W. C., et al. (2015). "Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression." Atmos. Chem. Phys. 15(18): 10349-10366.

[8] Wells, B., et al. (2021). "Improved estimation of trends in U.S. ozone concentrations adjusted for interannual variability in meteorological conditions." Atmospheric Environment 248: 118234. [9] Gao, Z., et al. (2022). "Separating emissions and meteorological impacts on peak ozone concentrations in Southern California using generalized additive modeling." Environmental Pollution 307: 119503.

Tables

Table 1

Variables selection at Sheboygan

Note: Pr(F) is the p value of ANOVA test, which is significant if less than 0.05.

Table 2

Variables selection at SWFP

Note: Pr(F) is the p value of ANOVA test, which is significant if less than 0.05.

Figure 1. 2019-2021 ozone design values for the entire LADCO region. Nonattainment and maintenance areas for the 2008 and 2015 ozone NAAQS are shown for comparison. Where the two nonattainment areas overlap, the area appears purple.

Figure 2. 2019-2021 ozone design values for the nonattainment areas (labeled) in the LADCO region. Nonattainment and maintenance areas for the 2008 and 2015 ozone NAAQS are shown for comparison. Where the two nonattainment areas overlap, the area appears purple. The nonattainment status of areas is given as of mid-May 2022.

Figure 3. The regression between the observed and fitted values of ozone in the Sheboygan and SWFP regions shows a high R-square (0.725), implying that the model works well.

Figure 4. The partial response between the Ht850mb (height of 850 millibar) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect

on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Figure 5. The partial response between the MeanRH (mean relative humidity) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Figure 6. The partial response between the SR max (solar radiation maximum) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Figure 7. The partial response between the WD.local (local wind direction) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Figure 8. The partial response between the WS.local (local wind speed) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Figure 9. The partial response between the tmax (maximum temperature) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Figure 10. The partial response between the MeanSondeBP (mean surface-level barometric pressure) and deviation ratio of long-term mean ozone in the Sheboygan region. The vertical coordinate represents the effect on ozone and is the ratio of the ozone concentration at that condition to the average MDA8 ozone concentrations for the year, with greater than 1 indicating an increase and less than 1 indicating a decrease.

Sheboygan 8-hr Ozone Trends

Figure 11. Adjusted ozone mean trends by meteorology and observed MD8A ozone for each year from 2000-2019 at Sheboygan. The adjusted value of ozone is to eliminate the effect of fluctuations in meteorological variables.

Figure 12. Adjusted ozone mean trends by meteorology and observed MD8A ozone for each year from 2000-2019 at SWFP. The adjusted value of ozone is to eliminate the effect of fluctuations in meteorological variables.

Figure 13. Adjusted ozone value trends by meteorology and observed MD8A ozone at different quantile for each year from 2000-2019 at Sheboygan. The effect of adjusting by meteorological variables can be seen by quantile regression for peak MDA8 ozone concentrations (0.9, 0.98).

Figure 14. Adjusted ozone value trends by meteorology and observed MD8A ozone at different quantile for each year from 2000-2019 at SWFP. The effect of adjusting by meteorological variables can be seen by quantile regression for peak MDA8 ozone concentrations (0.9, 0.98).