Submitted to the Annals of Applied Statistics

STAT 444 FINAL PROJECT

BY KAARTHICK RAMESHKUMAR

1. Introduction.

Ozone is a gas made of three oxygen atoms bonded together, which naturally exists in the upper atmosphere and can also be created at the surface by reactions between other pollutants. Although it is often celebrated by the public as a gas which protects the planet from harmful radiation, its presence at the surface level can cause serious health problems. Its immediate effects upon inhalation include causing shortness of breath, wheezing, asthma attacks, and an increased risk of respiratory illness. In the long term, ozone can cause premature death for high risk individuals (seniors, individuals with pre-existing lung conditions, etc.) It also reacts with other pollutants and can increase the body's response to them as well [2].

The concentration of surface level ozone has been found to be determined by a variety of meteorological factors. Ozone is created at the ground level by reactions between nitrous oxides and volatile organic compounds (VOCs). Temperature has a significant impact on the rates of these reactions, and also increases the amount of VOCs which are emitted by biogenic sources such as plants [3]. Furthermore, these reactions are initiated by solar radition, so it also plays an important role in the amount of surface level ozone [5]. Wind is another important meteorological factor which can disperse reactants and ozone at higher speeds, or cause them to travel from other locations and become concentrated at lower speeds [1].

Since ozone is detrimental to the health of certain individuals upon exposure and negatively impacts the general public, understanding how certain meteorological factors impact its surface level concentration is of interest. If the mechanisms behind the creation of ozone can be better understood, public health officials may be able to create guidelines for when the public should avoid going outdoors. To this end, the research question this paper proposes is to what extent do solar radiation, wind, and temperature serve as reliable indicators of ozone levels? To answer this question, a dataset involving surface level ozone concentrations and the meteorological factors in question will be used to fit a variety of predictive models which will be evaluated based on their predictive power. The goal of this paper is to find an adequate model which can be used to predict ozone concentrations given solar radiation, wind, and temperature.

2. Data.

The dataset that will be used to create the desired model is the ozone dataset from the Elements of Statistical Learning Textbook [4]. This dataset includes observations of ozone concentration (ppb), daily high temperature (degrees Fahrenheit), solar radiation in the 4000-7700 Angstrom frequency range (Langleys), and average wind speed from 0700 to 1000 hours (miles per hour) in New York City from May to September 1973. The temperature and wind speed measurements were taken from LaGuardia Airport, whereas the solar radiation measurement was taken from Central Park and the ozone measurement was taken at Roosevelt Island. The ozone concentration data was collected by the New York State Department of Conservation and the meteorological data was collected by the National Weather Service. A preliminary analysis of the data was done and can be found in 1.

The plots in 1 correspond to each of the covariates plotted against ozone, with a simple lin-

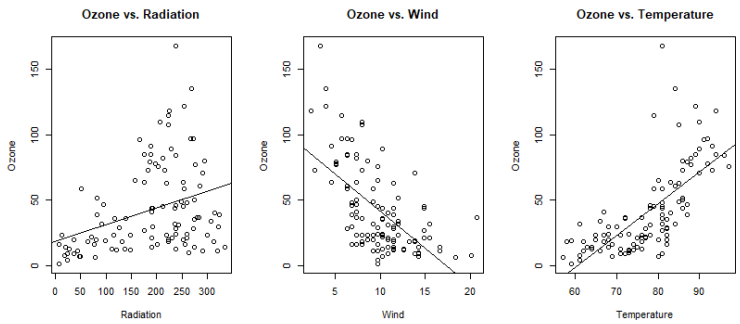

Fig 1: Preliminary Plots

ear model fit to each pairing. From these plots, some basic conclusions may be made. First, none of the covariates appear to exhibit any strong linear relationship with the response. It appears that as temperature and radiation increase, ozone concentration also increases. This is to be expected because of how these two covariates impact the chemical reactions which create ozone. It appears that lower average wind speeds are also correlated with higher ozone concentrations. This suggests that New York City produces a lot of pollution which is not dispersed and so they react to produce ozone. This may also suggest that reactants from other locations gather in New York City due to lower average wind speeds.

After this preliminary analysis was performed, various models were fit to this data. Of note in the fitting of these models is the potential interactions between the meteorological covariates. Intuitively, one would assume that temperature, wind, and solar radiation all have some sort of impact on one another, and so when fitting a model which hopes to describe ozone concentrations using these factors are covariates, their interaction must be kept in mind.

3. Methods.

3.1. *Cross Validation.* Since the goal of this paper is find the best model to predict ozone concentration given wind, solar radiation, and temperature, the models that are created will be evaluated primarily based on their generalized cross validation (GCV) score under square loss. The models which did not have linear predictors will be evaluated based on leave-one-out cross validation instead (LOO-CV). Cross validation was also used as a method of evaluation instead of using a train-test split because the dataset only has 111 observations. If a smaller training set was used, model bias would increase significantly due to the lack of data. Thus, all the data was used to train the models and cross-validation was used to evaluate their predictive power. Prediction error and R^2 values were also examined, but were not used to compare the different models that were fit.

3.2. *Multiple Linear Regression.* The first model that was fit was a multiple linear regression. The plots corresponding to the residuals and multicollienarlity of this model are shown below in 2.

From the first plot in 2, we see that the residuals of the multiple linear model are slightly right-skewed. Furthermore, the quantile plot of the residuals shows some variation from the theoretical quantiles for the smaller and larger residuals. As a result, the linear model may not be entirely adequate, although the evidence is not very strong. Furthermore, the multiple

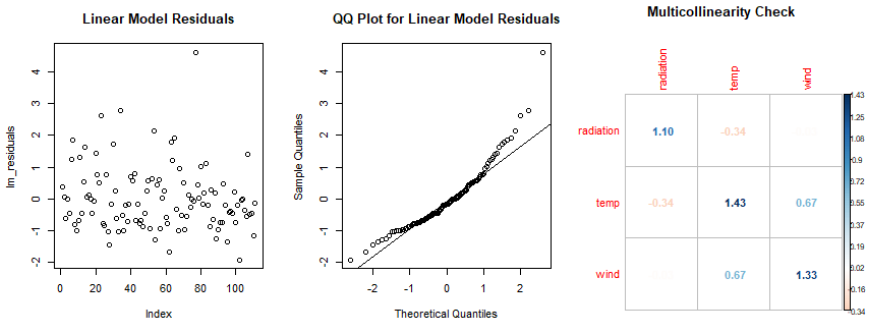

Fig 2: Linear Model Plots

linear model does not account for any interaction between covariates which is ideal for this dataset. The covariates do not show any problems with respect to multicollinearity, as shown in the third plot. Due to this fact and since the number of covariates is already quite low, the ridge and lasso models which were fit will not be discussed. There was no need to perform any variable selection or penalization.

3.3. *Additive Model.* The next model which was fit was an additive model using the default settings of the `gam` package in R. Models with all possible combinations of variates were fit, and the best one in terms of GCV was the model including smooth terms for all of them. All the smooth terms in this additive model were statistically significant, however it did not perform well in terms of its GCV. Furthermore, similar to the linear model this additive model did not account for any interaction between the covariates.

3.4. *Additive Model with Interactions.* In order to account for the interaction between the covariates that is likely present in the environment, some additive models with interaction terms were fit. The first model that was fit included smooth terms for all the covariates and for all pairs of interactions between them. However, many of the smooth terms in this model were not significant.

In order to get a simpler model with significant terms, the covariate or interaction term with the largest p-value was iteratively removed and the model was refit. This was done until all the remaining smooth terms had a p-value of less than 0.1. The final model which was reached included a smooth term for wind and the interaction terms involving temperature and wind, and temperature and solar radiation. In mathematical notation, this additive model can be written as:

$$f(\mathbf{x}) = f_1(x_{wind}) + f_2(x_{wind}, x_{temp}) + f_3(x_{radiation}, x_{temp}) + \epsilon$$

The estimated degrees of freedom were 5.22, 18.30, and 9.96 for each function respectively. The smooth function f_2 was the most statistically significant of the three with a p-value of 1.15×10^{-6}. The other two smooth functions had p-values of 0.075 and 0.082 respectively. This additive model indeed accounts for the interaction between the covariates. Furthermore, it suggests that the interactions are quite complex, since the estimated degrees of freedom are quite high.

3.5. *Decision Tree and Random Forest.* The last class of models which were fit were tree based models. These models were fit because they could capture the complex interactions between the covariates. A decision tree was fit with no limits on its depth. The `rpart`

package in R was used to do this. The resultant model had a depth of 3 and it was found that temperature was the most significant variable by far and accounted for 60% of the model's improvements.

A random forest model with 10 trees was also fit to the data using the `randomForest` package in R. However, the data used did not satisfy the model assumptions for a random forest. This is because a random forest model needs a large amount of data in order to fit a large number of diverse trees. The dataset used in this paper only has 111 observations, and so this assumption was not met. In order to compensate for this fact, the random forest was fit with only 10 trees, in hopes that the resultant trees would still be diverse enough.

Since these two models do not have linear predictors, a GCV score cannot be computed for them. As a result, LOO-CV was performed instead, and this score was used to compare these models with the others.

4. Results. The following table includes the GCV or LOO-CV scores of the models which were fit to the data, along with their prediction error under square loss, and the adjusted R^2 values for the linear predictor models. The additive model with interactions refers to the model that was presented in 3.4 through the iterative removal of smooth terms.

Model	Prediction Error	CV Score	Adjusted R^2
Linear	432.11	465.02	0.60
Additive	276.96	292.56	0.72
Additive with Interaction	126.59	133.72	0.83
Decision Tree	333.78	326.86	N/A
Random Forest (n = 10)	74.19	342.72	N/A

Table 1: Model Performance Results

From table 1, we can see that the additive model with interaction terms and the random forest have the lowest prediction errors by a significant margin. This can likely be attributed to them capturing the complex interactions between the meteorological factors at hand. In terms of the CV score, the additive model with interaction terms performed significantly better than all the other models, including the random forest. This is likely because the random forest overfit to the data, leading to low prediction error, but a high LOO-CV score. The additive model with interaction also had the highest adjusted R^2 value of the linear predictor models. Thus, from the metrics outlined in the previous section, it can be concluded that the additive model with interaction was the best performing model. It captured the complex interactions between the covariates and also is expected to perform well on new data as a result of its low GCV score.

From 3 we see that the residuals for the additive model with interaction are less skewed than that of the original linear model. They are slightly closer on average to the center. However, they still exhibit the deviance from theoretical quantiles at the extremal values, albeit slightly less than the linear model.

The estimated degrees of freedom for each smooth term in the additive model with interaction also provides some insight. In particular, since the estimated degrees of freedom of the temperature and wind interaction term was so high (18.30), one may conclude that temperature and wind together interact in a complex manner to determine ozone concentration.

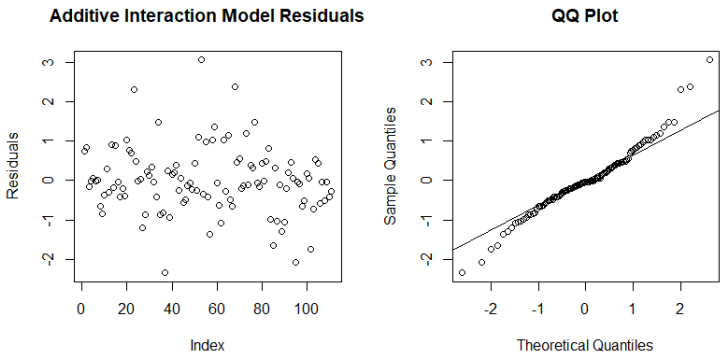

Fig 3: Residual Plots for Additive Model With Interaction

Furthermore, temperature appears to be one of the more important covariates in the dataset. It was the variable of most importance in the decision tree and (more importantly) its interactions with wind and radiation appear to be of great significance. This suggests that higher temperatures are necessary for the chemical reactions which form ozone, and the other variates depend on this being true in order to come to affect ozone concentration.

5. Conclusion.

In conclusion, an additive model involving wind and interaction terms with wind and temperature, and radiation and temperature performed the best out of the models which were fit. It had the lowest GCV score, meaning it likely has adequate predictive power to predict ozone concentrations given meteorological conditions. Thus, the goal of this paper has been achieved, and this model could potentially be used by public health officials to predict future ozone concentrations and warn high risk inidivuals.

One of the large limitations of this study was the lack of a large dataset. With more data, more complex models such as random forests could have performed better. These models should perform well with this type of data, since the meteorological factors likely interact in complex ways to determine ozone concentrations. However, without a larger dataset, these models cannot be adequately fit. In the future, the same models should be fit to a larger dataset. Another limitation of this study was that the data only pertained to New York City. Thus, the additive model likely cannot be used to predict ozone concentrations in other locations. Although a general model which works for all geographic locations likely cannot be obtained due to the immensely complicated nature of such a problem, other locations should be similarly analyzed and the differences in the models obtained should be further examined.

Finally, with respect to the goal of being able to use the model from this paper to allow public health officials to give recommendations, some improvement could be done with respect to the model's interpretability. If possible, this model should be visualized in a manner that the public can understand, so that it may be used for their benefit.

REFERENCES

[1] United States Environmental Protection Agency.
Ozone concentration.

[2] American Lung Association.
Ozone, 2024.

[3] Jane Coates, Kathleen A. Mar, Narendra Ojha, and Tim M. Butler.
The influence of temperature on ozone production under varying nox conditions – a modelling study.
Atmospheric Chemistry and Physics, 2016.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning.
Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[5] Deying Wang, Jizhi Wang, Yuanqin Yang, Wenxing Jia, Xiaofei Jiang, and Yaqiang Wang.
Impact of meteorological conditions on tropospheric ozone and associated with parameterization methods for quantitative assessment and monitoring.
Frontiers in Environmental Science, 2022.