|  | AIR QUALITY REQUEST FOR PROPOSAL  Blue Team 10  October 22, 2020 |
| --- | --- |
|  |  |
|  |  |

**Table of Contents**

[**OVERVIEW**](#_heading=h.gjdgxs) **2**

[**Methodology**](#_heading=h.30j0zll) **2**

[*Data Used*](#_heading=h.1fob9te) *2*

[*ARIMA*](#_heading=h.3znysh7) *2*

[UCM](#_heading=h.f07zuwowyy75) 4

[ARIMAX](#_heading=h.xfk0i6hnxvs5) 5

[Ensemble Model](#_heading=h.k8m2m7xlh90e) 5

[**ANALYSIS**](#_heading=h.atqjzsd82h88) **5**

[ARIMA](#_heading=h.l8jwygcbr2c7) 5

[ARIMAX](#_heading=h.53jrycixvybn) 5

[UCM](#_heading=h.6nrrctb7jv4i) 6

[Ensemble Model and Overall Comparison](#_heading=h.p5h7ckqklhnq) 6

[**Results & Recommendations**](#_heading=h.1t3h5sf) **7**

[Results](#_heading=h.y53fwnqsgjpq) 7

[Recommendations](#_heading=h.er142rlxlxnx) 8

[**Conclusion**](#_heading=h.17dp8vu) **8**

**List of Figures**

[**Figure 1. Time plot 3**](#_heading=h.2et92p0)

[**Figure 2. Time plot of predicted versus actual 4**](#_heading=h.4d34og8)

**LIST OF TABLES**

[**Table 1. Measures of Accuracy on Test Data 4**](#_heading=h.2s8eyo1)

Air Quality Request for Proposal

# **OVERVIEW**

A local medical facility has requested our assistance to investigate the air quality at the Millbrook School station, located in Wake County, North Carolina. The client is interested in a model that can accurately predict daily 8-hour max ozone concentration for the Millbrook School station in Wake County, NC for two weeks.

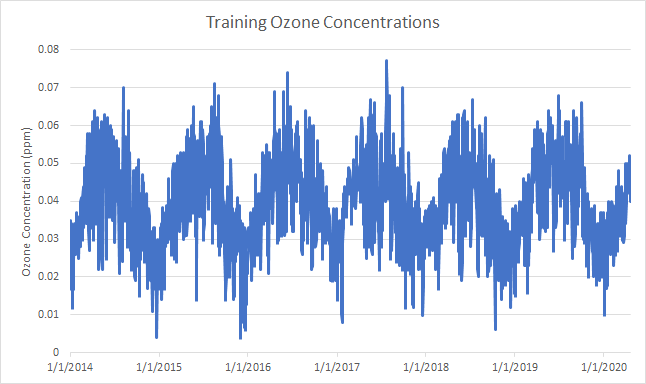
Within the dataset, our team explored ARIMA, ARIMAX, Unobserved Component Model (UCM) and ensemble modeling techniques to obtain the required forecast. Mean absolute percentage error (MAPE) to judge the model’s performance. Our team discovered that the UCM model provided the most accurate forecast on the test data with a recorded MAPE of 28.99% when forecasting the last 14 days of the dataset (May 18, 2020 - May 31, 2020); however, we recommend the client monitor ozone concentrations going forward to determine if an intervention variable provides meaningful improvements to the model as the COVID-19 landscape changes.

# **Methodology**

### *Data Used*

The data examined contains daily observations of ozone levels at the Millbrook School station located in Wake County, North Carolina. The dataset contains 2342 daily observations of max ozone levels in Wake County, ranging from January 1st, 2014 to May 31st, 2020. For modeling and analysis purposes, our team removed any leap days from the data, resulting in a total of 2341 observations.  The data was initially split into training (2299 days, Jan 1st 2014 - April 19th 2020), validation (28 days, April 20th-May 17 2020) and test (14 days, May 18-May 31st 2020) sets. After exploring the data, the training and validation data sets were merged to report the mean absolute error (MAE) and MAPE on the test data set. The ARIMA, ARIMAX, UCM and ensemble models all utilized this data split for model building and analysis. For validation and test data, weather data supplied was used due to the complexity of forecasting weather information; however, covariate pollutant levels were forecasted as well to avoid overestimating the ability of the models to forecast on future data.

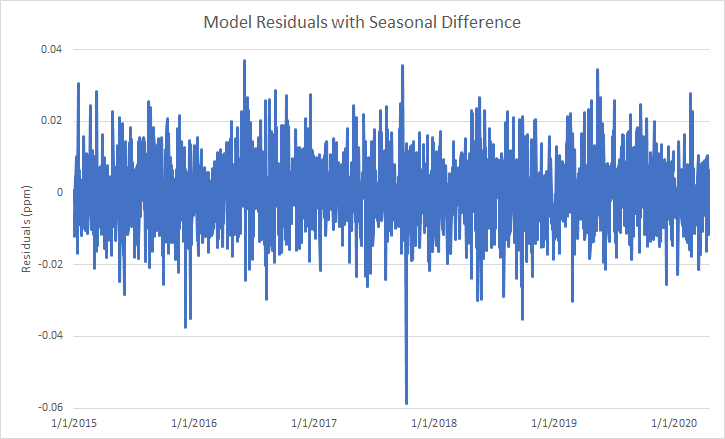
### *ARIMA*

The model building process began with our team splitting the data into training, validation, and test sets. After plotting and exploring the data, our team concluded that there is clearly seasonality in our data but not a strong trend. This is evident in Figure 1 below. The data oscillates about a fixed mean on 12 month cycles rising to a peak in summer months and falling to a trough in the winter. To address the seasonality present in the data, several methods were tested to determine the best technique. 

Because the data are daily observations, the team determined that it would not be appropriate to model the seasonality using dummy variables. With the long nature of the seasonal component, trigonometric variables were created into order to fit the observed seasonal component present. 6 cosine and 6 sine functions were inserted as input variables in order to best approximate this seasonal trend. Upon incorporating this into the ARIMA model, the team saw some improvement in accounting for seasonality, but a clear seasonal trend still existed.

The second technique used to account for seasonality in the data was to incorporate a seasonal difference of 365 days. With leap days removed from the dataset, this was intended to account for the year to year pattern evident in the time series. Upon review of the output of this ARIMA model, we decided that the difference appropriately accounted for the seasonality and further modeling could continue.

No clear trend component was visible in the data, so we did not investigate incorporating a trend component. Using the Augmented Dickey-Fuller test, the dataset was now shown to oscillate around a mean of zero and was confirmed to be stationary about that mean. This can be seen in Figure 2 below.



Now that the model accounted for both seasonality and was stationary about a zero mean, the team moved on to fitting autoregressive (AR) and moving average (MA) components to the model. Based on the autocorrelation and partial correlation plots, we iterated through several ARIMA models to account for any autocorrelation in the model, with the goal to be left with white noise. Once this was completed, MAPE was calculated on this model in order to obtain a goodness of fit test. From there, the model was tested and scored on the validation dataset. Again MAPE was calculated, but this time only on the newly forecasted 28 observations in order to obtain an accuracy measurement for the model prediction. This MAPE was recorded to compare against additional models.

### *UCM*

In addition to the ARIMA techniques used to model ozone levels, an Unobserved Components Model (UCM) was used. While the model building process for UCM’s is not as well developed as for ARIMA and ARIMAX forecasting, the ability of the UCM to incorporate deterministic and stochastic effects gives it desirable properties for our modeling.

Given previous information from the ARIMA model building process, similar effects were tested for incorporation.As expected, the slope component was shown to have no statistically significant impact on ozone levels. A stochastic level component was included while seasonality with trigonometric variables was not found to be significant. Given prior knowledge on the seasonality of the data, a seasonal autoregressive term was used instead. This proved significant. Inspection of the autocorrelation function showed no indication of further need for autoregressive or moving average terms. Similar to the ARIMAX approach, additional pollutants and weather variables which have proven effects on ozone concentration were considered for addition to the model. CO, NO2, and SO2 levels as well as precipitation, fastest two minute wind speed, average wind speed, and average temperature were found to be statistically significant and improve the white noise plot of residuals. These covariates along with the seasonal autoregressive term and stochastic level term made up the candidate UCM model.

As well as forecasting ozone levels, UCMs were used to forecast the other pollutant levels as well for use in the validation and test data. The three pollutants had similar models including a stochastic level component, a seasonal autoregressive component, and various regular autoregressive components as identified in their autocorrelation functions.

### *ARIMAX*

In order to account for seasonality using ARIMAX, the first technique to model the seasonal trend was to incorporate different independent variables into the model with the target and assess whether or not any one could account for the seasonality. Upon comparison of all these variables, we determined that the dependent variables did not sufficiently account for that season trend. After using dependent variables to account for seasonality was ruled out, we proceeded with the same process noted above in the ARIMA model building in order to determine that a seasonal difference best accounted for the seasonal variation. The team did consider the residuals output by every subsequent ARIMAX model in order to make sure that introducing any independent variables models did not introduce or increase seasonal components.

After accounting for seasonality present in the data, independent variables were considered in terms of model fit in order to account for as much variation in the data as possible. As each independent variable was included or removed, the Autocorrelation and Partial Autocorrelation plots were considered to detect where appropriate moving average and autocorrelation terms should be added. Different combinations were considered until three candidate models were selected.

After selection of the three candidate models, a forecast was run on the validation data for each. From this forecast, the MAPE was calculated for each. With the MAPEs available, they were weighed with the parsimony of the model as a whole to decide which was the best ARIMAX to move forward with.

### *Ensemble Model*

A final candidate for forecasting ozone levels was an ensemble of the previous modelling techniques. A simple mean ensemble was proposed where forecasts would be made as the average forecast of the candidate ARIMA, ARIMAX, and UCM models discussed above. This final candidate model was then compared to the component models on the validation data.

# **ANALYSIS**

### *ARIMA*

The ARIMA model had a MAPE on the validation data of 20.79%. This was the worst model compared to the ARIMAX, UCM, and ensemble models. The forecasts on the validation data were less accurate than most of the other models. However, one benefit of the ARIMA model as compared to the others is the simplicity. The ARIMA is based only on previous values of ozone concentration, seasonal differencing, and a few autoregressive and moving average terms. The other models require forecasting additional prediction variables.

### *ARIMAX*

In order to construct the ARIMAX, several variables were considered to be modeled to help predict the target going forward. Prior research suggested that some specific variables may have some relationship with ozone levels. Temperature, wind speed and NO2  levels are some of those variables. Based on this information, we tried several of the variables related to each of those in different combinations on the training data. With every new combination of independent variables, the resulting data was checked for seasonality, stationarity and to see if only white noise was left unmodeled. After several iterations, three specific models were found to be the best fit. The resulting models can be seen below in Table 1, along with their calculated MAPE for the training dataset.

| Candidate Model | MAPE | Independent Variables |
| --- | --- | --- |
| 1 | 4.01% | 5 |
| 2 | 4.04% | 5 |
| 3 | 4.44% | 3 |

We decided that these three candidate models were good enough to be tested on the validation data to get a better gauge of their accuracy. After forecasting the independent variables into the validation dataset, each of these three models was used to determine the ozone levels for the new timeframe. The resulting MAPEs can be found below in Table 2.

| Candidate Model | MAPE | Independent Variables |
| --- | --- | --- |
| 1 | 20.67% | 5 |
| 2 | 20.51% | 5 |
| 3 | 23.23% | 3 |

The ARIMAX model performed better on the validation data than the ARIMA model but not as well as the UCM and ensemble models. The ARIMAX model had a MAPE of 20.51% on the validation data. One drawback of the ARIMAX model is the need to forecast the X variables. The accuracy of the predictions of ozone concentration depends on the accuracy of the prediction of the X variables. This introduces additional uncertainty around the ozone forecast because of the added uncertainty of the X forecasts.

### *UCM*

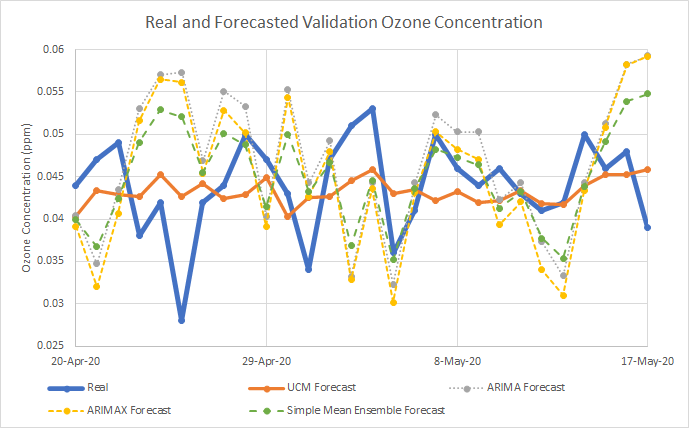
The UCM model had the best MAPE of all models tested on the validation data with a value of 10.57%. The MAPE was by far the most accurate model tested on the validation data, however, it suffers from some of the same drawbacks as ARIMAX because the X variables need to be forecasted.

### *Ensemble Model and Overall Comparison*

The ensemble model had a MAPE of 15.60%, coming in as the second best of all models tested on the validation data. The MAPE values for all of the models can be found in Table 3 below. ARIMA is the simplest but least accurate of all models and the ensemble was the most complex. UCM provides balance between model complexity and highest accuracy so the UCM model was passed to the test data.

| Candidate Model | MAPE | Independent Variables |
| --- | --- | --- |
| ARIMA | 20.79% | - |
| ARIMAX | 20.51**%** | 5 |
| UCM | 10.57% | 7 |
| Ensemble | 15.60% | 7 |

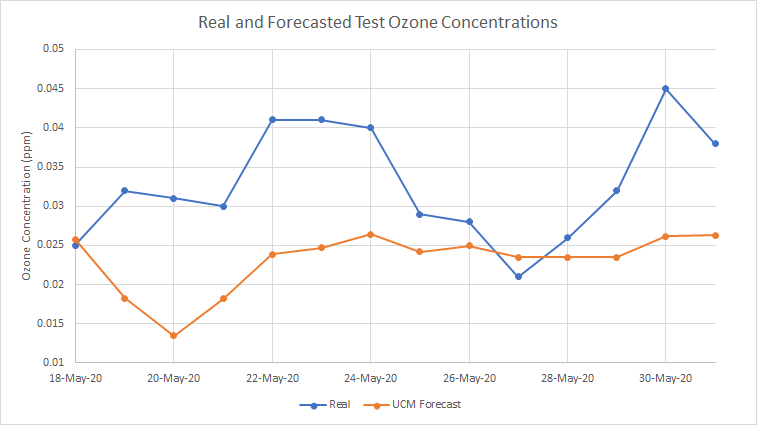
To explore the forecasted values of each model graphically, Figure **X** shows the forecasted values for each model as well as the actual ozone values for the validation data (April 20th-May 17 2020).



# **Results & Recommendations**

### *Results*

Out of all the models our team created, the UCM model performed the best on the test data set. Figure 4 depicts a time plot of the actual values of ozone concentrations as well as the predicted values of ozone concentrations for the test set. The MAPE on the test data for the UCM model was 28.99%.



In context, a naive model which forecasts ozone levels as the mean value of all previous readings will forecast on the test dataset with a MAPE of 29.69%. While the UCM model does produce some improvement on the naive model, the high variability of daily ozone measurements as shown previously in Figure 1 results in a high signal-to-noise ratio (SNR). This high SNR limits the improvement possible by forecasting on daily data as models that perform well on training data will tend to overfit the signal due to the presence of noise. Thus, accurate forecasting of ozone concentrations could be improved by rolling up the data into longer term averages. Previously, monthly averages were explored with greater success, but for a shorter time frame, weekly averages could also be explored.

### *Recommendations*

Based on the success of the UCM model and the variability in the daily data, we recommend that the client do the following:

* If a daily forecast is required, utilize the UCM model
* Utilize the weekly or monthly average to forecast for a more accurate, though less granular, projection compared to daily data
* Consider using an intervention variable if, in the future, the COVID-19 landscape significantly impacts ozone concentration

# **Conclusion**

Our team discovered that the UCM model provided the most accurate forecast on the test data with a recorded MAPE of 28.99% when forecasting the last 14 days of the dataset (May 18, 2020 - May 31, 2020); however, we recommend the client monitor ozone concentrations going forward to determine if an intervention variable provides meaningful improvements to the model as the COVID-19 landscape changes. Due to the large amount of variation in the daily observations, building a model on the weekly or monthly average level may provide more accurate, though less granular, forecasts.