Preliminary Data Analysis: Modeling Solar and Wind Across the United States

Jack Murray

4/13/2022

## Project Question

* The project seeks to examine how wind speed is related across 31 locations in Northeastern United stats and Southern Canada. Then, identify trends in wind speed between each location and if the wind power generated at some locations or regions could reliably supplement the electrical needs at other locations. In order to to create a supply of electricity at any location regardless of the local wind speed and power production.
* This question is very critical and interesting for the future of the the full scale implementation of renewable energy. Since wind and solar energy must be as dependable and accessible as fossils fuels to eventually replace them. Because, non-renewable energy sources can produce electric regardless of the local weather conditions.

## Difficulty

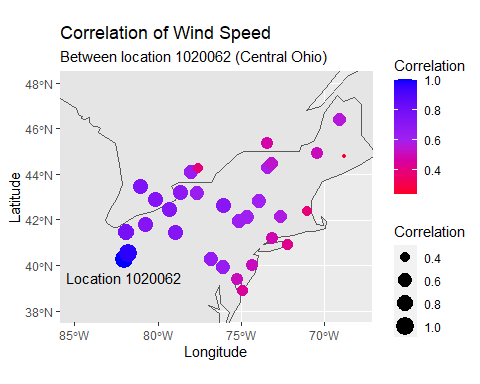
* The main difficultly with this project was retrieving and preprocessing the data used to fit the wind models. I used the NSRDB (National Solar Radiation Database) to gather the data on 31 location for the past 21 years. However the downloaded data is very messy and divided on individual files by the year and location. Therefor, Each file had to be cleaned and stored in a more usable format, then combine into one large data set.
* Moreover, the model I created to simulate wind speed is very complicated and can be slow to fit each location. Which makes updating the model and algorithm difficult.

## Convo

* A consistent and reliable power supply is critical for everyday life. The majority of American’s power comes from non-renewable sources partly due to the dependent availability of production. If clean renewable energy is to become a viable and sustainable alternative, it must offer the same reliability as current non-renewable sources. Power companies will be interested in the results of this project to optimize the placement of renewable energy plants to ensure a stable power supply for their customers.
* The weather fluctuates constantly throughout the year and varies significantly at each location. Moreover, locations can not be selected solely based on the amount of wind and solar energy it receives throughout the year. Since these locations may be far removed from the population centers and if these high wind and solar localisations do experience a rare windless cloudy day, the local grid could not create any power. Therefore, the typical cyclical weather patterns of solar radiation and wind speed at each location needs to be analyzed, regardless of the total amount available. Identifying locations that can supplement lactation during cloudy windless will be crucial to creating a sustainable renewable energy grid.
* By gathering daily average solar radiation and wind speed data across the Northeast, a simulation will be created to model the wind speed and solar radiation throughout at each selected location. Then, using the simulated data, the correlation and trends between the different regions can be visualized with scatter and correlation plots. More advance methods, like cluster analysis, will identify similar locations to evaluate which what regions have complementary solar and wind energy. Finally, power companies can examine the simulated data to verify locations that can supply other locations reliably during low wind or solar days.
* If the initial results are sufficiently insightful and some relationship can be established between the Northeastern locations wind and solar production, then the model will be improved further. More locations will be modeled and the larger network of simulated data will create more accurate outputs. Then, collection of simulated data and models will help create a model that will optimization the location of future renewable energy power plans.

## Example of Trends

* To identify possible trends the models and simulated data will explore, the correlation was calculated between the wind speeds of 1 sample consolation and all other Northeastern locations wind speed. The central Ohio location,denoted as location 1020062 on the map below, is highly correlated with other inland locations. This could indicate that if the wind speed is slow in Ohio, then the locations in Pennsylvania and New York would also have low wind speeds. Meaning different locations are needs to supplement the power shortage in Ohio. Conversely, the locations on the East Coast and in Canada are less correlated. These locations could possibly be used to support the electrical grind in Ohio.
* The models of wind speed at each of these locations will further aid in detecting these tends between locations. Ad will identify any other trends between regions.

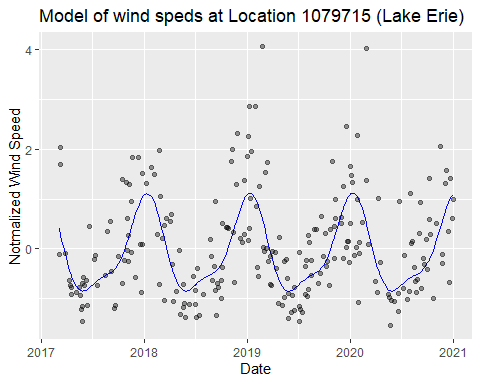


## Model Fitting

## Model Fitting

* The basic model is built from 6 sinusoidal terms with periods of 1 year, 1/2 year, and 1/4 year. This reflects the seasonal change of wind speed during the summer and winter months. To construct each model, the parameters are fitted with thousands of different combinations of coefficients. Then the errors between the model predictions and the observed average wind speed was calculated for each day.
* Then, which ever combination of parameters result in the lowest amount of error was selected to be the model used at that location. This process was repeated for all 31 location.
* In order to improve the time to fit all 31 models, only 15% of the data was selected at random to be in the training set.

## Model Visual: Observed Wind Speed

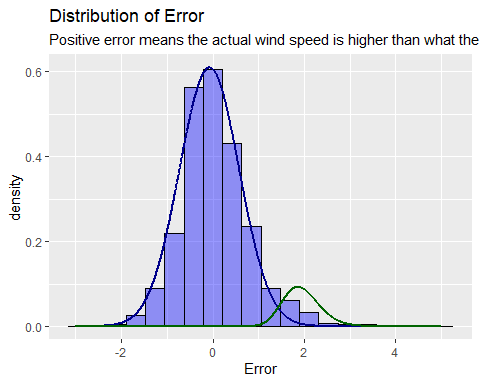


## Model Visual: Observered Wind Speed

* The model captures the sinusoidal change in wind speed each year. The winters have higher wind speeds than the summer and peak quickly before dropping again rapidly in the spring. Then in the summers, the wind speeds tend to flatten out before slightly increasing in to the fall. This was captured by making the coefficient for the year long cosine ans sine terms the largest. Then blunting the troughs with positive 1/2 year terms, and tuning the slopes with small coefficients for the 1/4 terms.
* Some issues with the models are the high positive errors. Since the model predicts few very high average wind speeds, it structured to captures the winter peaks. also some locations with low wind speeds have more error. Since low peaks and trough make the sinusoidal pattern is less pronounced and harder to fit.

## Model Error

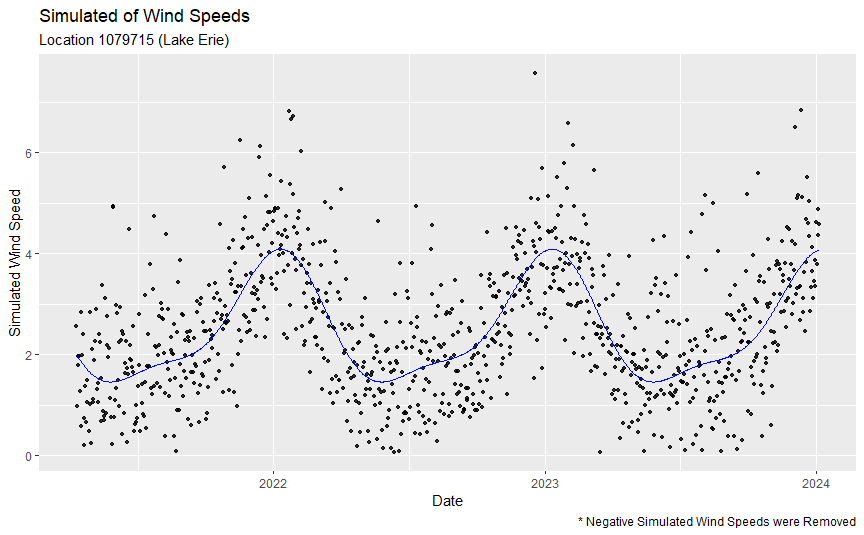
* After selecting the best model, the error was plotted on the histogram below and a probability distribution was fitted to capture the error. The error needs a probability distribution so it can randomly be added to the models predictions. In order to replicate the observed randomness in the simulated data. Therefore,the error distribution must fit the model well.



## Model Error

* In order to fit the error properly, a normal distribution is applied for the majority of the of the error. Since the data is normalized and the model fits the observed data well, most of the errors are centered around 0 with a standard deviation closer to 1. A gamma distribution is also added to capture the positive skewing seen in most models error distributions. The high positive errors resulted form less common high wind days in winter the models does not account for.
* Moreover, to improve the fit of the error distributions a cut off point was selected to dived the errors in the bell shaped part of the distribution form of the errors from the high positively skewed portion. Then, the gamma and normal distributions were evaluated based on the log-likelihood of their fit and which ever cut of point roughly maximized both log likelihoods was selected to molded the error.

## Model Visual: Simualted Wind Speeds



* After the models was fitted, the next 1000 days of wind speeds were simulated. Each day randomly deviates from the model prediction based the error distribution fitted for the models error. The majority of days add a random error value to model the normal distribution to replicated the randomness. The gamma distribution is also used to model the radon error by calculating the percentage of errors that are above the cutoff point, then adding an error value using the fitted gamma distribution that percentage of the time. this captures the high wind speeds observed in winter.
* Although the models appropriately models the error, the summer wind speeds are much lower and vary more then expected.

## Next steps

* For the project, I plan to add the previous day’s wind speed to the model in order to improve the quality of the predicts wind speed and capture how year typically have slightly different wind speeds. Then I will create the simulated data for all 31 location using the improved models. With the simulated data i hope to identify insights to on how the different locations are correlated and use cluster analysis to examine which locations are similar.
* Following this semester, the solar radiation for these location will also be modeled and simulated solar radiation. More location will also be added in the future to identify more differences between geographical regions. Wis will eventual be used to optimize placement of wind/solar farms.