Classification of Safe States Suitable for Solar and Wind Renewable Energy Development

Eric Zhang (zhange6@rpi.edu)

Dr. Munasinghe
Data Analytics
December 2023

Abstract and Introduction

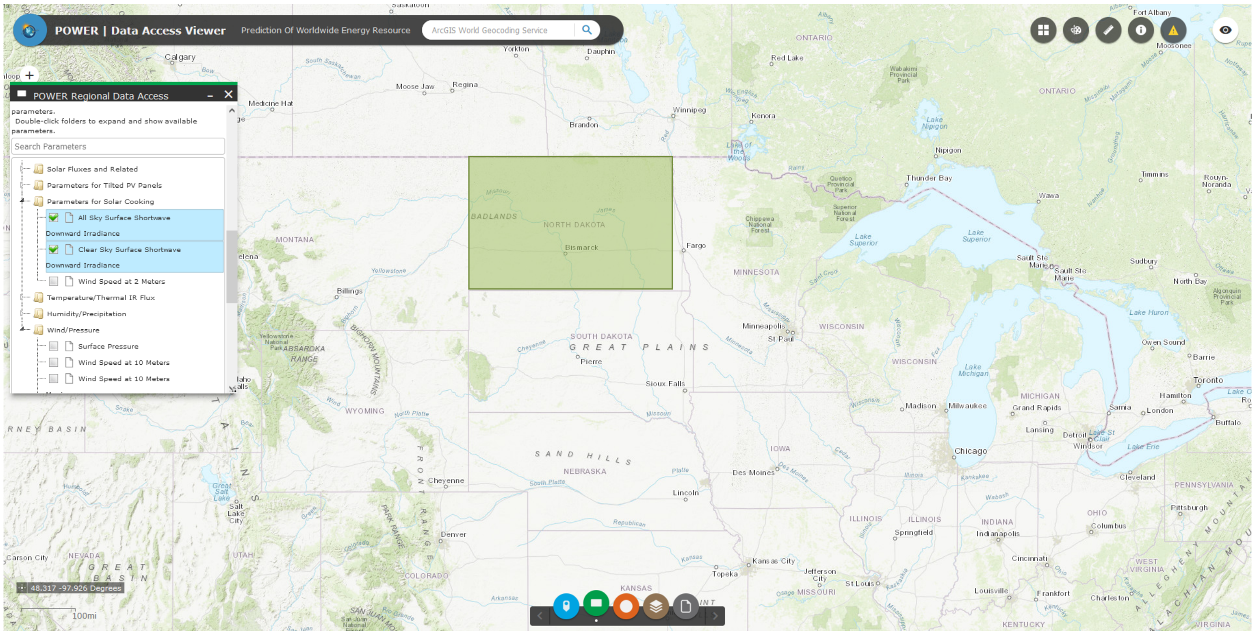
My motivation for this project was the continuation and expansion of the X-Informatics project I did last semester. The original was based on examining the environment, demographics, and hydropower plant development in the southeastern United States. I wanted to expand the scope a bit and change the area of focus. Instead of focusing solely on the Southeast, I wanted to try and classify states into being suitable or not for a different type of renewable energy, wind energy. Alongside that, I thought having access to climate data would allow me to make a small model on the side regarding potential solar energy and solar power generation. My initial idea was that given a risk score from FEMA and with a measure of average wind speeds, I would be able to classify a state as suitable for wind power or unsuited for wind power. For the solar part, I initially thought I could use past solar presence to predict the amount of solar energy available to be generated in that state. With only around 13% of the energy the United States uses being renewable and only 6% being wind and solar[1], it appeared to be an area worth investigating into where more wind plants and solar plants could be developed. Additionally, by using the FEMA risk data, I thought it would be useful in determining which states should be considering developing renewable resources while considering their safety.

Data Description and Exploratory Data Analytics

The first dataset was collected and computed by the Federal Emergency Management Agency, or FEMA. The National Risk Index data is collected yearly and is limited to either county level resolution or state level resolution. In addition, only the county level resolution has the risk scores in the dataset[2]. When approaching this problem, I initially thought I would be able to find climate data that could be matched on a county level and would thus be able to use the risk scores. However, I was unable to find climate data that matched the resolution that I was looking for and had to use the state level data instead. As a result, the Expected Annual Loss score, or EAL score, was used instead. The data itself contains an EAL score for each state and territory of the United States, which I filtered down into the continental United States. It also contains an EAL score as well as other EAL measurements for each natural disaster or risk, such as avalanches, flooding, strong winds, tornadoes, and lightning. The risk scores for the disasters were not available as well.

For the other half of my dataset, I tried to find climate data for each state. Unfortunately, most weather stations that had climate data did not have wind data. This may be due to a lower priority in assessing wind as a factor in everyday life. However, I was able to use NASA Langley's data access map to create and retrieve rectangular sections of climate data from the United States[3]. As seen in Figure 2.1 below, I needed to create sections of data myself to download.

Although I needed to include parts of states outside of the state I was trying to select in order to get the most data, I was able to obtain the features and measurements that I wanted. The main feature was wind speed at 50 meters above the ground, along with solar irradance under all kinds of weather. I also downloaded clear sky irradiance and a wind speed range, but did not find those to be that useful after processing them.



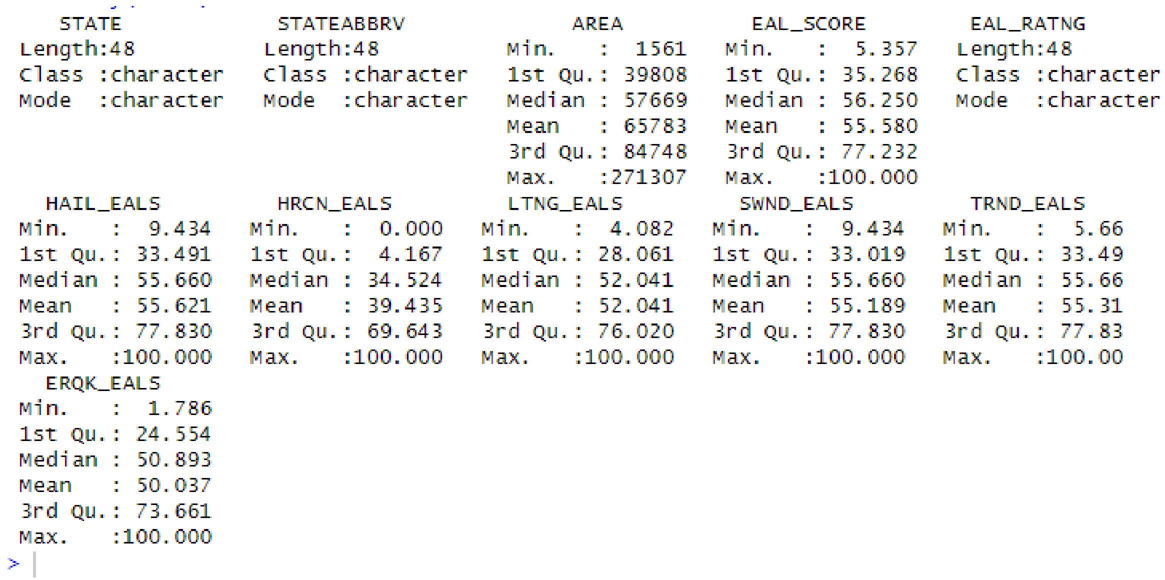
parameters.
Double-click folders to expand and show available parameters.

- Solar Fluxes and Related
- Parameters for Tilted PV Panels
- Parameters for Solar Cooking
  - ☑ All Sky Surface Shortwave Downward Irradiance
  - ☑ Clear Sky Surface Shortwave Downward Irradiance
  - □ Wind Speed at 2 Meters
- Temperature/Thermal IR Flux
- Humidity/Precipitation
- Wind/Pressure
  - □ Surface Pressure
  - □ Wind Speed at 10 Meters
  - □ Wind Speed at 10 Meters

Figure 2.1 - Demonstration of data retrieval from NASA Langley data access map.

Analysis

I was required to transform and preprocess both datasets a lot. First, for the FEMA Risk Index data, I sorted it by ascending EAL scores in order to find the top ten least riskiest states or top ten least lost due to natural disasters while taking the area of the states. In addition, there were a number of natural disaster risks that were not applicable to wind and solar power, so I extracted the risks that made sense while filtering out the non-continental United States as well. A summary of the dataset after processing is shown in Figure 3.1 below.



```
    STATE            STATEABBRV              AREA           EAL_SCORE         EAL_RATNG        
 Length:48          Length:48          Min.   :  1561   Min.   :  5.357   Length:48         
 Class :character   Class :character   1st Qu.: 39808   1st Qu.: 35.268   Class :character  
 Mode  :character   Mode  :character   Median : 57669   Median : 56.250   Mode  :character  
                                       Mean   : 65783   Mean   : 55.580                     
                                       3rd Qu.: 84748   3rd Qu.: 77.232                     
                                       Max.   :271307   Max.   :100.000                     
   HAIL_EALS         HRCN_EALS         LTNG_EALS         SWND_EALS         TRND_EALS     
 Min.   :  9.434   Min.   :  0.000   Min.   :  4.082   Min.   :  9.434   Min.   :  5.66  
 1st Qu.: 33.491   1st Qu.:  4.167   1st Qu.: 28.061   1st Qu.: 33.019   1st Qu.: 33.49  
 Median : 55.660   Median : 34.524   Median : 52.041   Median : 55.660   Median : 55.66  
 Mean   : 55.621   Mean   : 39.435   Mean   : 52.041   Mean   : 55.189   Mean   : 55.31  
 3rd Qu.: 77.830   3rd Qu.: 69.643   3rd Qu.: 76.020   3rd Qu.: 77.830   3rd Qu.: 77.83  
 Max.   :100.000   Max.   :100.000   Max.   :100.000   Max.   :100.000   Max.   :100.00  
   ERQK_EALS      
 Min.   :  1.786  
 1st Qu.: 24.554  
 Median : 50.893  
 Mean   : 50.037  
 3rd Qu.: 73.661  
 Max.   :100.000  
> 
```

Figure 3.1 - Summary statistics of the dataset after extracting relevant features

One important aspect of the dataset is the fact that it has already been somewhat standardized. Due to the nature of the data, one state or natural disaster is guaranteed to have a score of 100 as they are most affected or are the most dangerous.

Another transformation I did was to take out the small states. While they may be less costly than others, I believed that they did not have proper infrastructure or space available to build wind and solar plants, so I decided to take them out of consideration.

After transformation, I took out the top ten least costly states to build my models on. This is what led me to select the following states, as they were the least riskiest - Wyoming, Maine, Montana, Idaho, New Mexico, North Dakota, South Dakota, Arizona, Wisconsin, and Nebraska. Some clustering can already be seen here, as most of the northern midwest states are shown to be the least costly with a few in the west/southwest as well. Maine is the only outlier as far as state selection is concerned, being in the northeast.

I then extracted another dataset with only the numerical features and plotted a pairs plot of the features, shown in Figure 3.2. In this plot I noticed that the hurricane risk I originally selected was irrelevant, so I dropped that risk as well.

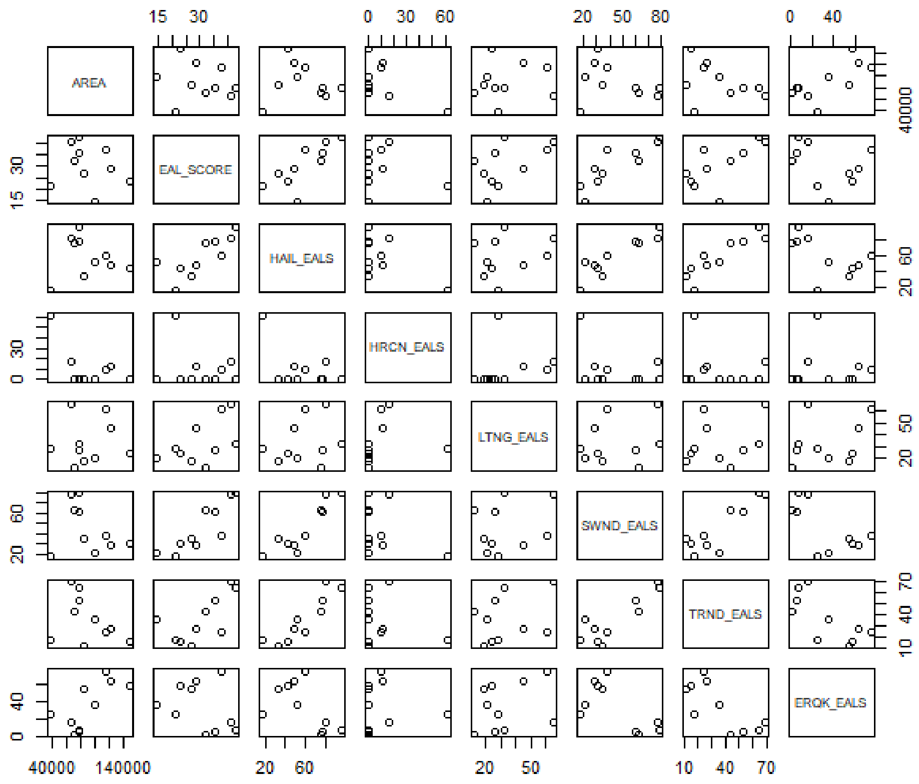
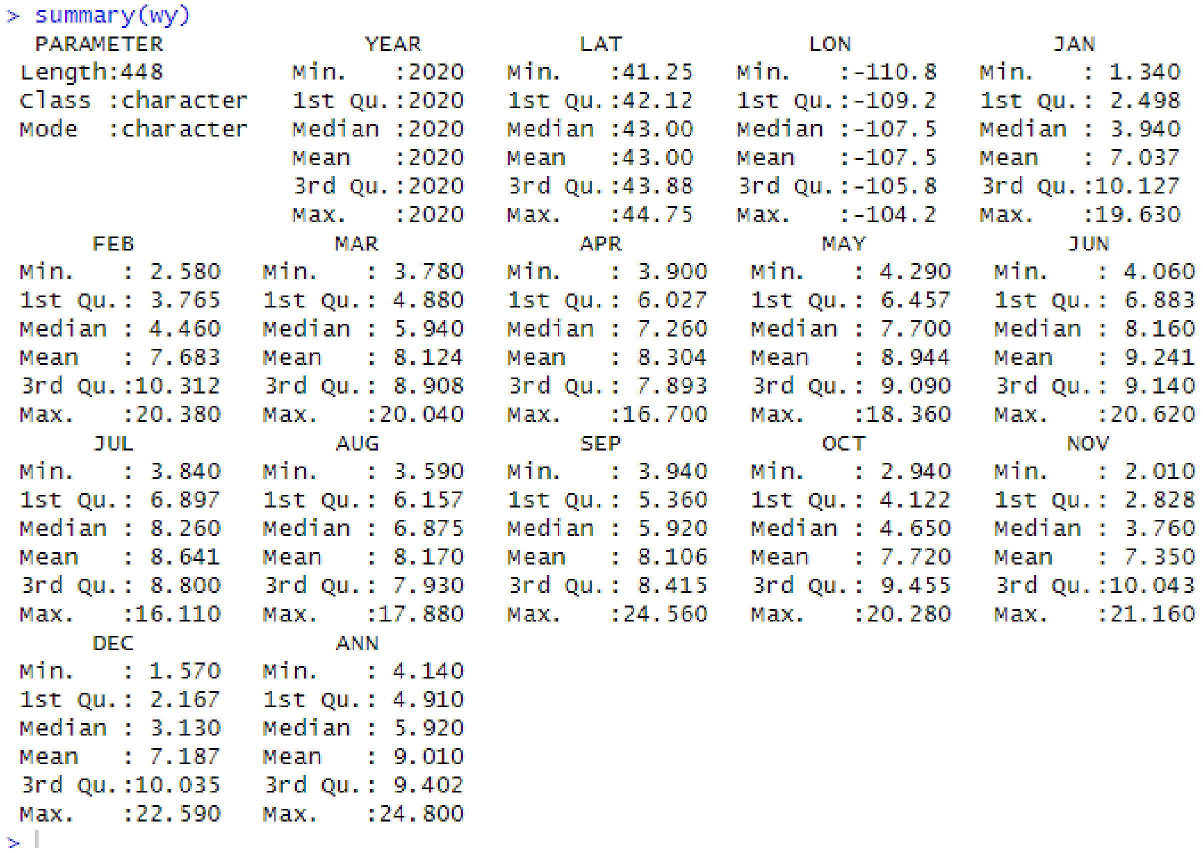


Figure 3.2 - Pairs plot of numerical dataset

After this plot, I read in all of the data I gathered for each state area. A summary is shown in Figure 3.3. This data contained data readings in increments of latitude and longitude. In order to make it usable, I had to average out each parameter's annual feature and add them to the numerical dataset under the AVG_WIND_SPEED, AVG_ALL_IRR, and AVG_CLR_IRR features. While this leads to a much weaker strength in the data as a particularly windy location could be offset by one that gets little to no wind, there was no other way for me to get climate data on a level that would work well with the resolutions I had. This is a large source of uncertainty and a great potential source of error. If there was a way in the future to easily get data on a county level, these datasets would work much nicer together.



```
> summary(wy)
  PARAMETER              YEAR            LAT              LON               JAN        
 Length:448         Min.   :2020   Min.   :41.25   Min.   :-110.8   Min.   : 1.340  
 Class :character   1st Qu.:2020   1st Qu.:42.12   1st Qu.:-109.2   1st Qu.: 2.498  
 Mode  :character   Median :2020   Median :43.00   Median :-107.5   Median : 3.940  
                    Mean   :2020   Mean   :43.00   Mean   :-107.5   Mean   : 7.037  
                    3rd Qu.:2020   3rd Qu.:43.88   3rd Qu.:-105.8   3rd Qu.:10.127  
                    Max.   :2020   Max.   :44.75   Max.   :-104.2   Max.   :19.630  
      FEB              MAR              APR              MAY              JUN        
 Min.   : 2.580   Min.   : 3.780   Min.   : 3.900   Min.   : 4.290   Min.   : 4.060  
 1st Qu.: 3.765   1st Qu.: 4.880   1st Qu.: 6.027   1st Qu.: 6.457   1st Qu.: 6.883  
 Median : 4.460   Median : 5.940   Median : 7.260   Median : 7.700   Median : 8.160  
 Mean   : 7.683   Mean   : 8.124   Mean   : 8.304   Mean   : 8.944   Mean   : 9.241  
 3rd Qu.:10.312   3rd Qu.: 8.908   3rd Qu.: 7.893   3rd Qu.: 9.090   3rd Qu.: 9.140  
 Max.   :20.380   Max.   :20.040   Max.   :16.700   Max.   :18.360   Max.   :20.620  
      JUL              AUG              SEP              OCT              NOV        
 Min.   : 3.840   Min.   : 3.590   Min.   : 3.940   Min.   : 2.940   Min.   : 2.010  
 1st Qu.: 6.897   1st Qu.: 6.157   1st Qu.: 5.360   1st Qu.: 4.122   1st Qu.: 2.828  
 Median : 8.260   Median : 6.875   Median : 5.920   Median : 4.650   Median : 3.760  
 Mean   : 8.641   Mean   : 8.170   Mean   : 8.106   Mean   : 7.720   Mean   : 7.350  
 3rd Qu.: 8.800   3rd Qu.: 7.930   3rd Qu.: 8.415   3rd Qu.: 9.455   3rd Qu.:10.043  
 Max.   :16.110   Max.   :17.880   Max.   :24.560   Max.   :20.280   Max.   :21.160  
      DEC              ANN        
 Min.   : 1.570   Min.   : 4.140  
 1st Qu.: 2.167   1st Qu.: 4.910  
 Median : 3.130   Median : 5.920  
 Mean   : 7.187   Mean   : 9.010  
 3rd Qu.:10.035   3rd Qu.: 9.402  
 Max.   :22.590   Max.   :24.800  
> 
```

Figure 3.3 - Summary of one state's weather data, Wyoming shown

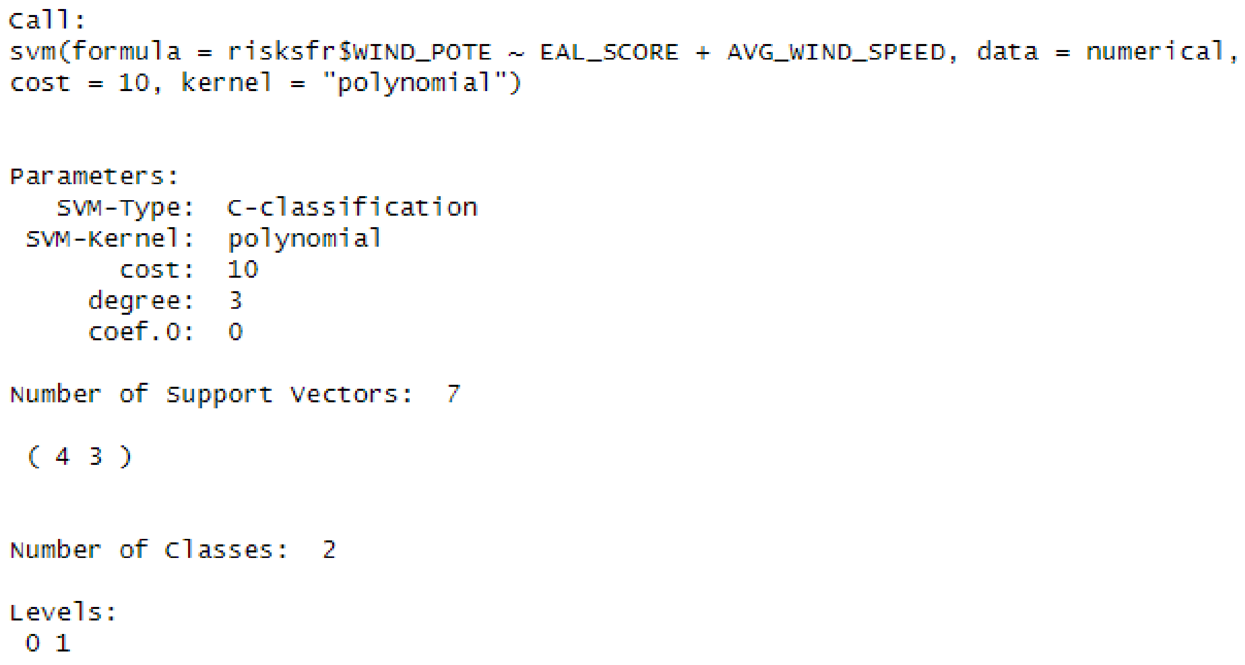
In addition, I had to create a target variable WIND_POTE for wind power plant potential. As a binary variable, I assigned 1 to states that had potential to develop wind power plants and a 0 to those who did not, basing it on the EAL_SCORE and AVG_WIND_SPEED.

Model Development and Application of Models

I generated 4 models for this project. I made a support vector machine to try and classify states, a logistic regression model, a KNN classifier, and a linear regression plot.

Support Vector Machine

A support vector machine helps classify points by providing a boundary, then a margin separating the two classes where some error is allowed. This model was used to try and classify states based on the EAL_SCORE and AVG_WIND_SPEED with WIND_POTE being the target variable. A polynomial kernel was used after testing different kernels, as it appeared to fit the data the best while providing tangible results.



```
Call:
svm(formula = risksfr$WIND_POTE ~ EAL_SCORE + AVG_WIND_SPEED, data = numerical,
cost = 10, kernel = "polynomial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  10
     degree:  3
     coef.0:  0

Number of Support Vectors:  7

 ( 4 3 )


Number of Classes:  2

Levels:
 0 1
```

Figure 4.1 - Support Vector Machine call

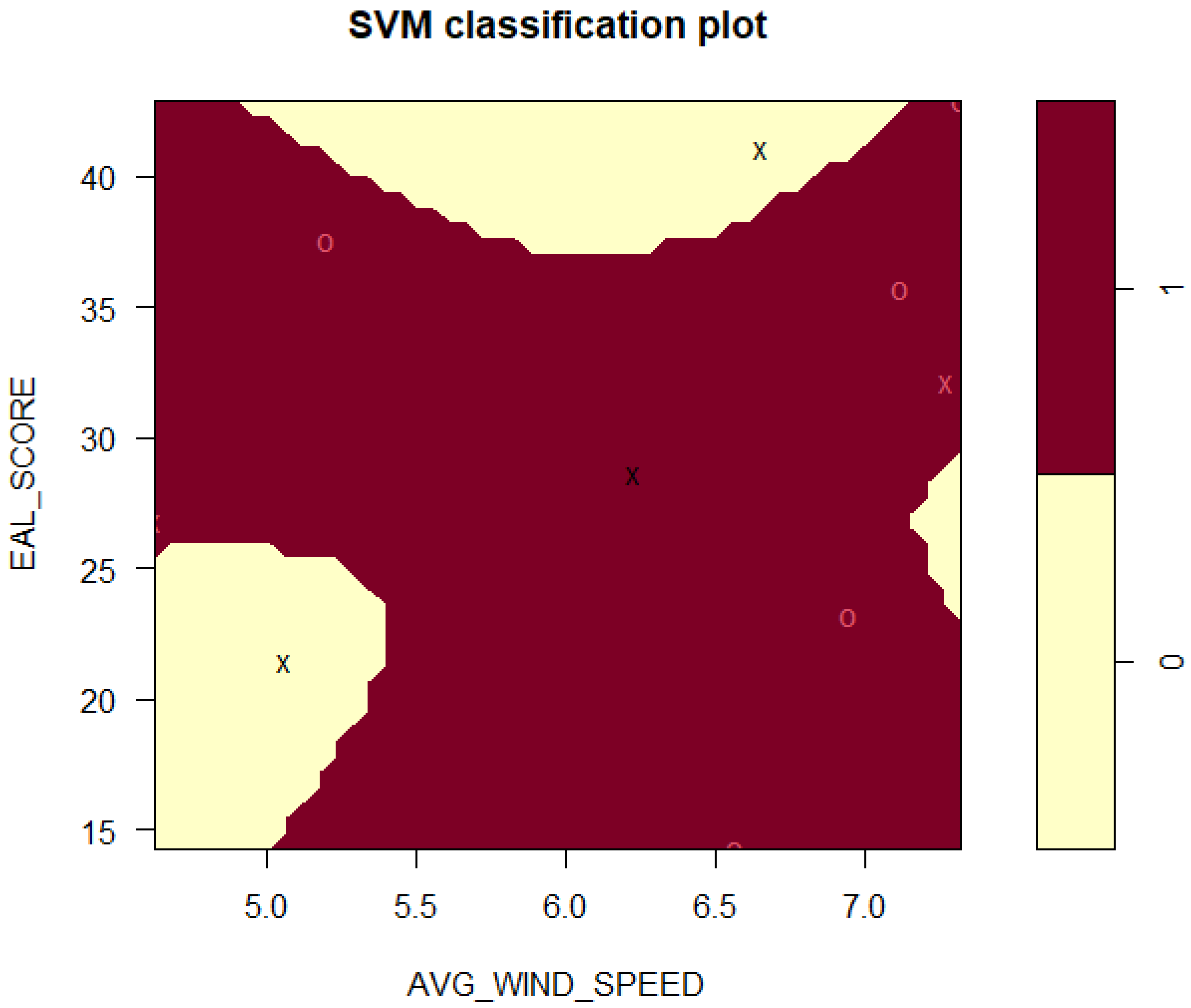
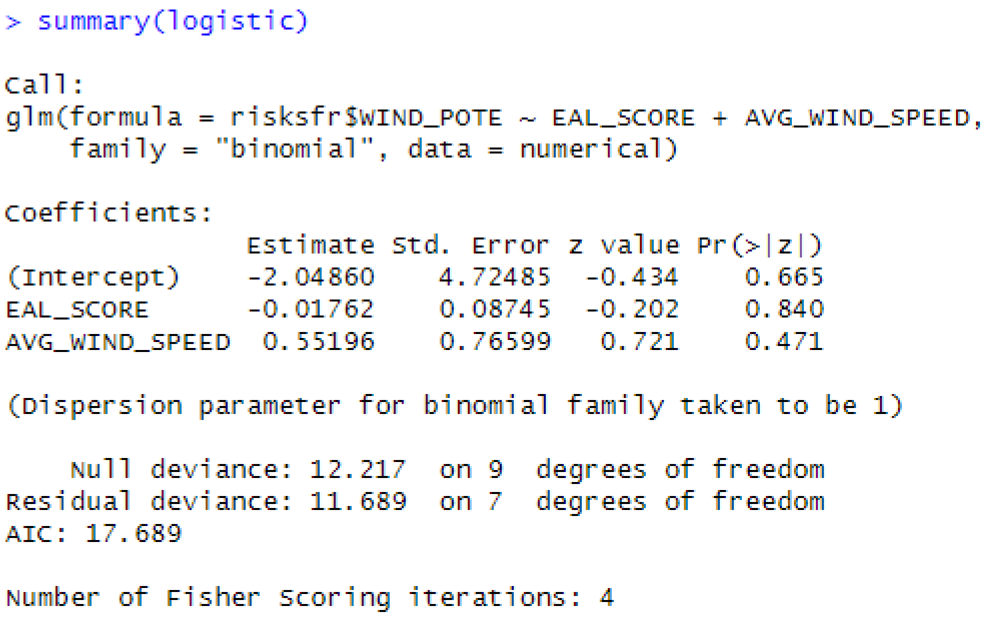


Figure 4.2 - Support Vector Machine plot

I chose a 10 as the cost as a lower cost ended up classifying the entire plot as having potential with no boundary to be seen. A polynomial kernel was also chosen as the data was not linearly separable at all, and seemed to fit the data the best. With these results, it appears that a higher average wind speed is more weighted than the EAL_SCORE in determining a state's potential for wind energy.

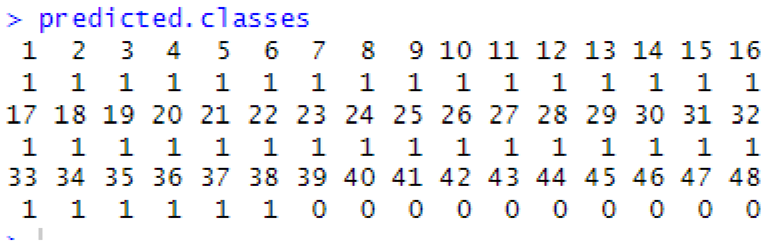
Logistic Regression

The next model I chose was a logistic regression model to try and classify the states again. A summary of the results can be found in Figure 4.3 below.



```
> summary(logistic)

Call:
glm(formula = risksfr$WIND_POTE ~ EAL_SCORE + AVG_WIND_SPEED,
    family = "binomial", data = numerical)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.04860    4.72485  -0.434    0.665
EAL_SCORE      -0.01762    0.08745  -0.202    0.840
AVG_WIND_SPEED  0.55196    0.76599   0.721    0.471

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12.217  on 9  degrees of freedom
Residual deviance: 11.689  on 7  degrees of freedom
AIC: 17.689

Number of Fisher Scoring iterations: 4
```

Figure 4.3 - Logistic Regression summary



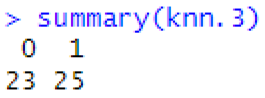
```
> predicted.classes
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1 
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1 
33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 
 1  1  1  1  1  1  0  0  0  0  0  0  0  0  0  0 
```

Figure 4.4 - Logistic Regression Predictions

With the logistic regression model, I was able to test it on the rest of the states to try and classify them based on their EAL_SCORE and AVG_WIND_SPEED. To help with the issue of having no wind data for the other states, I had to fill in the data with the mean for the AVG_WIND_SPEED feature in the dataset.

KNN Classifier

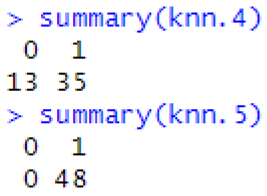
For the third model, I chose a K Nearest Neighbors classifier. I used the entire dataset and initialized the training with k = 3 neighbors. Based on the numerical dataset and taking in the modified dataset with the other states, I was able to predict a classification for the other states using this classifier.



```
> summary(knn.3)
 0  1 
23 25 
```

Figure 4.5 - KNN Classification, k = 3

I thought this was a bit too low of a number of states with potential for wind energy, so I changed k to k = 4 and k = 5 and made more models.



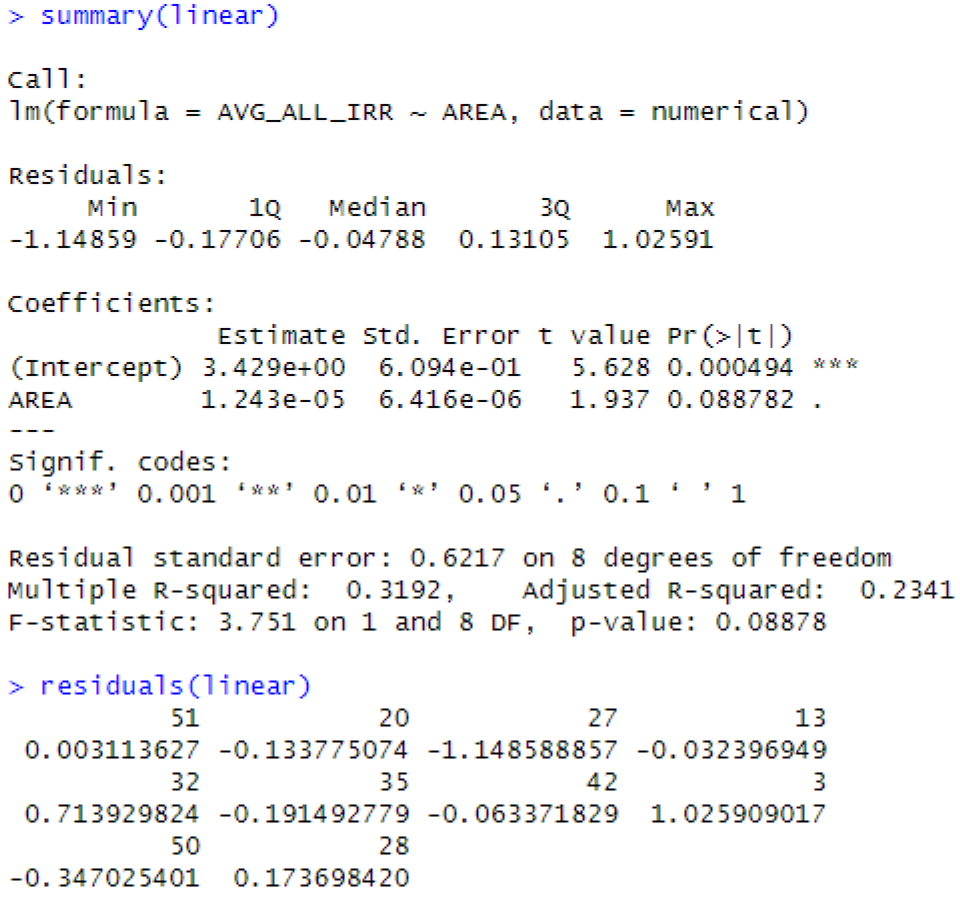
```
> summary(knn.4)
 0  1 
13 35 
> summary(knn.5)
 0  1 
 0 48 
```

Figure 4.6 - KNN Classification, k = 4 and k = 5

With k = 4, the predictions were around what I expected them to be with around .25 not being good candidates for wind energy. With k = 5, the classification skewed much more to having all the states be good candidates for wind energy plants.

Linear Regression

For my final model, I chose to use a linear regression model to try and predict the average solar irradiance in all sky conditions given an area. To do this, I used the area measurements found in the FEMA data and the solar irradiance data I put together from NASA Langley. I also had the average solar irradiance given a clear sky, but I thought that it would be too unrealistic to use. The results of the linear regression model as well as the plot with the best fit line are shown in Figures 4.7 and 4.8 below.



```
> summary(linear)

Call:
lm(formula = AVG_ALL_IRR ~ AREA, data = numerical)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.14859 -0.17706 -0.04788  0.13105  1.02591 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.429e+00  6.094e-01   5.628 0.000494 ***
AREA        1.243e-05  6.416e-06   1.937 0.088782 .  
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6217 on 8 degrees of freedom
Multiple R-squared:  0.3192,	Adjusted R-squared:  0.2341 
F-statistic: 3.751 on 1 and 8 DF,  p-value: 0.08878

> residuals(linear)
          51           20           27           13 
 0.003113627 -0.133775074 -1.148588857 -0.032396949 
          32           35           42            3 
 0.713929824 -0.191492779 -0.063371829  1.025909017 
          50           28 
-0.347025401  0.173698420 
```

Figure 4.7 - Linear Regression model summary and residuals

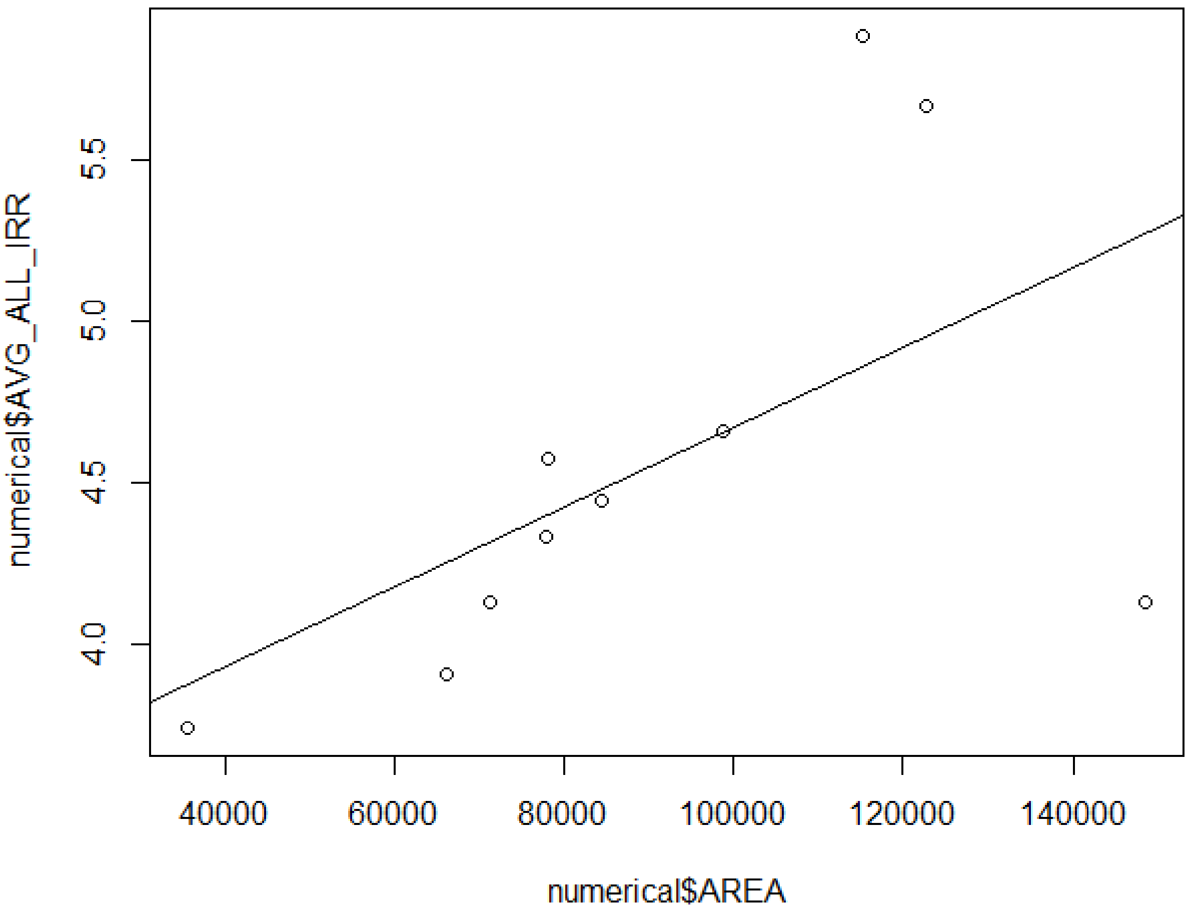


Figure 4.8 - Best fit line plotted over scatter plot

With the residuals, the numbers above the residuals correspond to a state as the row names were kept from the original FEMA data. A fairly regular linear relationship is seen, although it varies wildly as area increases, which may be due to an increase in potential cloud coverage. While the p-value is very small, it is still above 0.05, which means that there is a linear relationship between the two variables. To use this for predicting solar energy generation, a general idea of more energy generation as area increases appears to be supported.

Some of the original models I considered like a K-Means cluster analysis were not possible due to the size of the dataset and the characteristics of the data.

Conclusions and Discussion

Before data analysis, the hypothesis was that a state could be assessed to have potential for wind power generation based on their Expected Annual Loss score and average wind speeds. In addition, a relationship between area and average solar irradiance in all sky conditions was hypothesized.

After creating the models and analyzing them, I think that there is a way to classify states based on Expected Annual Loss and average wind speed, but also think that there is a better variable to do it with. I was able to get some parts of a boundary out of the support vector machine, as well as use the logistic regression to classify the states along with the KNN classifier. However, the data that I used was heavily modified from its original state and may have ended up not representing the states fairly in the end. In addition, the lack of a target variable meant that I had to come up with my own thresholds for classifications and add the classifications in, introducing a large amount of potential error depending on how I classified the states.

As for the linear model, I found it to be much less related than I thought. The average solar irradiance was barely dependent on area, which makes sense when taking into account other factors like cloud coverage and average amount of sunshine. A more robust multivariate regression could be done on all of those variables if available to obtain a better relationship for determining solar irradiance, and therefore predicting energy generation.

In future applications, if wind and climate data was available on a county by county basis there could be a lot more models built off of this dataset. In addition, knowing which states have potential or not for wind power would be very helpful to have as a target to train the other variables on. I did not realize that there would be no target when trying to use the existing classification as a variable, and this led to a large amount of uncertainty and potential problems being introduced into this project. As for climate data, I had to carve out rectangular approximations by hand as NASA Langley was the only source capable of handling the data requests. If climate data was available more readily like mentioned above, these models would be much easier to both create and refine, and potentially draw new conclusions from.

GitHub URL:

https://github.com/Wyvail/DataAnalyticsFall2023_Term_Project_Eric_Zhang

References:

https://css.umich.edu/publications/factsheets/energy/us-renewable-energy-factsheet

https://hazards.fema.gov/nri/data-resources#csvDownload

https://power.larc.nasa.gov/data-access-viewer/