Forest Fire Data Analysis

Joshua Shook

June 12th, 2020

1. Abstract

Fire prediction has gotten to become quite complicated over the years with remote sensing and satellite imaging being able to be utilized in order to help firefighters contain wildfires. However, being able to predict fires based on simple data should be an important goal to have. So, given a simple forest fire dataset of data accumulated about the area of a fire and the corresponding weather situation on the day of the fire, I attempted to create a multiple linear regression model in order to see if these simple measurements could be used to help firefighters reliably predict the size of the fire without the utilization of massive amounts of technology.

2. Introduction

Forest fires, as all Californians know, can wreak havoc on the surrounding population and wildlife. If we could predict the likelihood and size of the fire given variables that can be easily measured, such as temperature and humidity, it could help prepare firefighters for dealing with such natural disasters before they get out of control. In order to identify certain periods in which firefighters should be on high alert and attempt to create a model that reliably predicts the future size of fires based on independent variables that can easily be measured, I will be using the UCI Machine Learning Forest Fire Data Set. It possesses 12 independent variables and the area of each fire in every case. Although this data set is based on recordings from the northeast region of Portugal, the methods used within this paper can be applied to a data set of any region. Making sure that the data only comes from one specific region is important because otherwise the consistency in the data would be lost, and the accuracy of the model would be affected negatively.

Since the early 1900s scientists began using controlled wildfires in order to understand the complexity of wildfire spread. Computers lead to a huge advancement in fire predicting technology in the 1980s.[1] Even later, the use of remote sensing and satellite imaging data has revolutionized the ability to predict wildfires as well.[2] Even though such technology is extremely useful, it is possible that not every fire department, especially in poorer countries, has access to these technological advancements. An example of a model that does not use advanced technological methods but still is quite complex in itself is a model that was developed in Canada called the Fire Weather Index. It is used to predict the fire danger of a certain area using six premeasured components, all combined into one Fire Weather Index number. All that is needed to conduct these measurements are the measurements of air temperature, precipitation levels, relative humidity, and wind speed. These measurements are then used to compile the components that use the moisture of the fuel present within the wildlife area and the weather conditions within the area to predict the behavior of a fire.[3] Four of these components are present within the wildfire dataset. This is a well-established model for predicting fires in Canada, but we should still be seeking for new mathematical and statistical methods for predicting these

[1] https://apfmag.mdmpublishing.com/wildfire-prediction-systems-for-the-future/
[2] http://www.borealforest.org/world/innova/fire_prediction.htm
[3] https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi

natural disasters. Even though we have many potentially useful methods when it comes to fire science, we still are unable to reliably predict and prevent wildfires from exploding in size, as proven by the extent of the California fires that occur year after year.

Modern fire predicting models are quite detailed and advanced, so I do not expect to match these models' level of technological and scientific progress. However, being able to create a model with minimal amount of details using regression analysis could very well be useful if advanced methods such as remote sensing and access to PhD-earning statistics students are lacking in availability. After all, some insight is better than none.

3. Questions of Interest

I have two main questions of interest that I'd like to tackle in this paper. Firstly, the frequency of fire is important to know. For my first Question of Interest I would like to examine the days and months in which forest fires within this data set occur the most. This will give insight on when the ill-fated fire season occurs, which is a relatively simple question to answer. My second Question of Interest will be quite a bit more complex. I would like to find a reliable regression model to predict the size of a fire given some easily measurable input variables. I am given 12 variables to help me with this analysis, and I may have to remove or transform any of these variables to achieve the desired result. This determination of removal or transformation will also be a part of my analysis.

4. Data and Methods

Data

The data I will be using for this paper will be, as previously stated, collected from a dataset found on University of California’s machine learning site. [4]

All the variables in my dataset are relevant to my paper except the X and Y variables which measure the coordinate location of the fire. The variables month, day (of the week), and temp (degrees Celsius) are all self-explanatory. The FFMC, DMC, DC, and ISI variables are all measurements that are used to predict the FWI, or Fire Weather Index. Each of these variables contain separate easy to obtain measurements that could be potentially extremely useful when it comes to predicting wildfires. FFMC stands for Fine Fuel Moisture Code and it measures the fuel moisture levels of litter fuels that are scattered along the ground, examples being grass, leaves, and needles. The DMC index, or Duff Moisture Code, is the measurement of the moisture levels of decomposed organic matter present beneath the litter. The third FWI measurement is the DC index, which measures the dryness deep into the soil. Finally, the ISI index measurement

[4] [Cortez and Morais, 2007] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: [https://archive.ics.uci.edu/ml/datasets/Forest+Fires]

is used to determine the potential for spread of a fire in a certain region.[5] It takes into account the surface windspeeds in an area and the moisture levels of fine dead fuels, which are grass, leaves, and needles of a certain size. The remaining independent variables are RH, which is the relative humidity in percent, wind, which is wind speed in kilometers per hour, and rain, which is the measurement of rainfall in mm per meter squared. Finally, we have the dependent variable, area, which is the burned area of forest measure in hectares.

There were no missing values within the dataset. There was one outlier data value I removed because it would result in a misleading model. Also, I performed a transformation on one of the variables, specifically the area variable, due to skewness. Additionally, I had to convert month and day variables from categorical to numerical variables through dummy coding because they were deemed important to the regression model. I will go into more detail in the exploratory analysis part of the paper.

In terms of principle of measurements, the data present within my dataset is definitely relevant to the question I am asking. The variables present within the dataset are all potentially useful in determining whether a fire will blow up in size. The data is also quite precise. I would not want to include data that would be far too precise for the average forest ranger or meteorologist to obtain. The values present within the Forest Fire Dataset are all precise enough to be useful and not too precise as too be useless. The measurements within the data set in no way affect the landscape in which a fire occurs, and so worry over system distortion is unfounded. And finally, although the FIW measurements do require a little effort, but they can all be calculated using simple relative humidity, wind speed, rainfall, and temperature measurements. Another potential cost could be that if people were to use a fire prediction model and it produced an area likelihood that was vastly underestimated, it could have disastrous consequences if the people that were in charge of containing wildfires put too much trust in the model.

Ethically, not much is concerning with this dataset. It doesn’t breach anyone’s privacy or put anyone in danger. The only concern I could possess would be whether the data is correct or not, which could be potentially disastrous as previously stated.

Methods

Answering the first question of which days and months are the most important to watch for forest fires is an easy yet important one, since the model doesn’t take into account the frequency of the fires, only the size. This was just a matter of creating two simple histograms in order to count the occurrences of each fire in the dataset. In order to answer the second Question of Interest I will need to create a regression model, but first I have to preprocess the data by getting rid of any columns that do not help with the data analysis, removing any outliers, making sure there is no multicollinearity, and one hot encoding. After this preprocessing I will create the multivariate regression model and test its usefulness by calculating the average squared error as well as the r squared value.

[5] https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

Exploratory Analysis

I would first like to attempt to answer the first Question of Interest which is viewing the impact the day of the week and the month of the year have on the frequency of fire. Even though this is a simple question, it is important because my regression model, as previously stated doesn't take into account the frequency of the fires, only the size. So, it is still very important for the fire department to know when the most fires occur even if the size of the fires might not expand to drastic proportions. Here are two histograms showing the fire frequency count for each day of the week and each month of the year (Figure 1 and Figure 2, respectively).

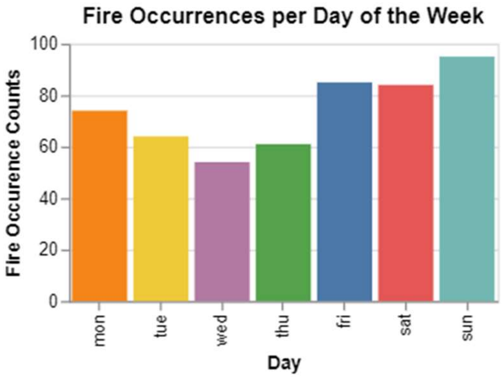


Figure 1, Occurrences of fires based on the day of the week

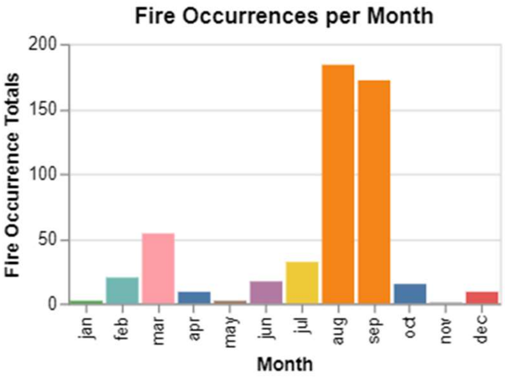


Figure 2, Occurrences of fires based on the month of the year

As evident from Figure 1, there is some difference in the occurrences of fires on the weekend (from Friday to Sunday) as opposed to the other days of the week. Fire occurrences seem to be much more common on these days. With Figure 2, the stark contrast in fire occurrences is quite impressive. August and September completely dominate the fire count totals. This concludes the exploratory analysis for the first Question of Interest and leads us to the more complex problem of creating a reliable model for predicting fire size with the data at hand.

For the second question of interest, I first needed to see the distribution of the area variables to make sure no transformations needed to be made on the data. To do this, I created a histogram of binned area values. As you can see from figure 3 below, the areas are extremely skewed towards an area of 0.0 hectares. This apparent skewness could just be due to the presence of some outliers, so get a better idea of the skewness I removed the zero values and created an upper bound of 20 hectares, and even then the data was skewed.

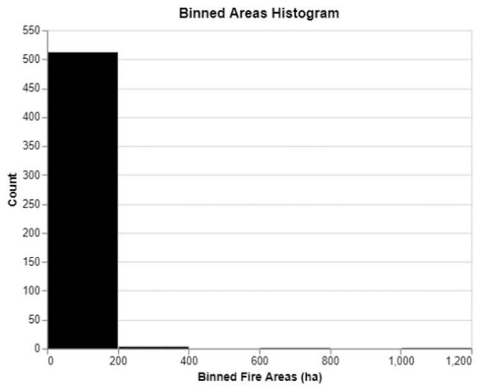


Figure 3, Raw Distribution of Areas

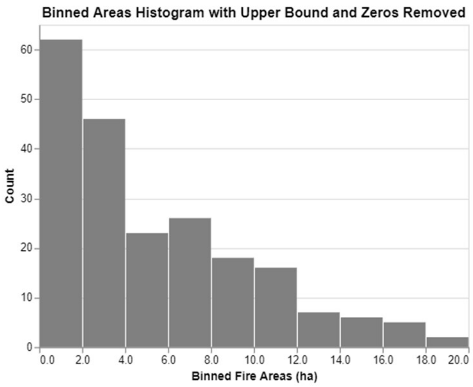


Figure 4, Distribution of Areas with an upper bound of 20 hectares and all zero removed.

I would like to perform a log transformation on the areas burned column in order to make the data follow an easier to predict distribution. This involves talking the natural log of the areas column after adding a value of 1 to each value in the column to avoid undefined values. The results of such a transformation are seen below in Figure 5.

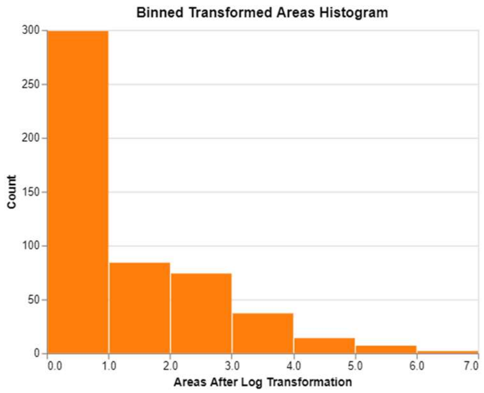


Figure 5, Distribution of Binned Log Transformed Areas.

It is very clear that the log transformation helped reduce the skewness in the data set and could improve the regression analysis prediction capability. Before I perform any regression, I'd first like to see the impact the day of the week and the month of the year have on the frequency of fire. This will give me an indication of whether I should include these categorial variables in the dataset. If there isn't really any difference between the days of the week or month when it comes to fire size it is best to simplify the model by leaving those variables out. Figures 6 and 7 below are the average fire sizes for each day of the week and each month of the year, respectively.

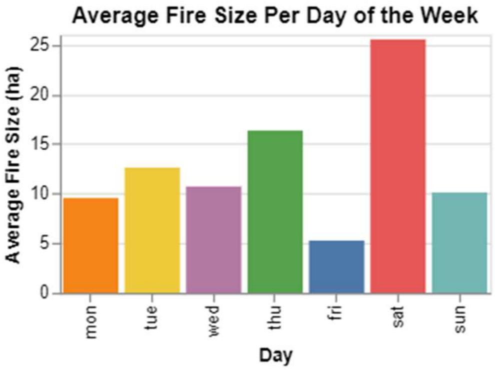


Figure 6, Average Fire Size per Day of the Week in hectares

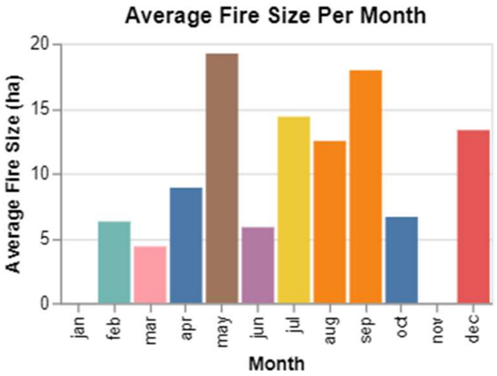


Figure 7, Average Fire Size per Month of the Year in hectares

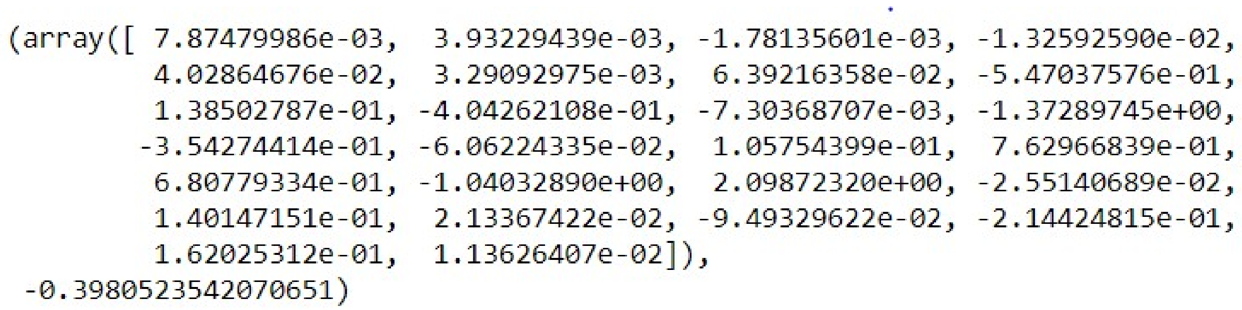
It does seem that there is some variance when it comes to the size of a fire based on the day of the week it occurs in and the month of the year it occurs in as well. So, I will be including these in my regression model. The only caution I would have from this result would be the fact that fires during May seem to be extremely large. However, the count of May fires is extremely low, as displayed by Figure 2. So, the extremely high average fire size in May could be the result of a potential outlier. This is in fact the case. There are only 2 fires that occur in the month of March, and both fires are drastically different in size from each other. I think it would be misleading to believe that may has such a high average fire area when only two fires occur in such a time frame. Because of this fact, I will remove the high area variable relative to May of 38.48 hectares from the dataset.

Next, I will use the pairs function to find any linear relationships between the numerical variables and the log transformed area variable. I have decided to remove the X and Y coordinate variables from further consideration in the regression analysis because linear regression using coordinate variables does not make intuitive sense. The pairs function will also give us any indication of multicollinearity within the data. Multicollinearity within this data set was a concern of mine, since the DMC, DC, ISI, and FFMC variables all use temperature, relative humidity, wind speed, and precipitation levels within their calculations.

From the scatterplot matrix in the Appendix section of the report (Figure 8), it appears that there are no obvious linear relationships between the different independent variables and the logarithm transformation of the area. This is probably because it is not only one aspect of the environment that causes a fire to reach a large size, but rather the combination of them. Because rain seems to be normally at the value of zero it is probably okay if we remove it from the data set. Counting the number of rainfall values that are 0 confirms this supposition, and so I will remove rainfall as a feature in the dataset. Additionally, it seems that multicollinearity is not a concern because there does not seem to be any strikingly strong relationships between the independent variables within the dataset according to the scatter plot matrix.

The next problem I will have to deal with is dealing with the categorical variable's month and day. I cannot just numerically rank them from 1 to 12 and 1 to 7 respectively because that implies that 2 * Monday = Tuesday, for example, an absurd notion. This is why I will have to use one hot encoding, which will create a column for each month of the year and day of the week. If a fire occurs on a certain day of the week W or month of the year M, the column for day=W and month=M will be 1. Otherwise, it will possess the value of 0.

After I perform the one hot encoding approach to interpreting categorical variables for linear regression, it is time to see if what I have done so far is enough to create a workable regression model. The multivariate regression model I have produces the coefficients listed below in Figure 9, with an y intercept value of -0.3980523542070651.



```
(array([ 7.87479986e-03,  3.93229439e-03, -1.78135601e-03, -1.32592590e-02,
         4.02864676e-02,  3.29092975e-03,  6.39216358e-02, -5.47037576e-01,
         1.38502787e-01, -4.04262108e-01, -7.30368707e-03, -1.37289745e+00,
        -3.54274414e-01, -6.06224335e-02,  1.05754399e-01,  7.62966839e-01,
         6.80779334e-01, -1.04032890e+00,  2.09872320e+00, -2.55140689e-02,
         1.40147151e-01,  2.13367422e-02, -9.49329622e-02, -2.14424815e-01,
         1.62025312e-01,  1.13626407e-02]),
 -0.3980523542070651)
```

Figure 9, Coefficients of the regression model that are produced

Now even though this model was created successfully, it still might be a poor one. I will create a function in order to measure the error of the regression, specifically the average squared loss. The error loss for this regression was found to be 1.8073716479554158. This is seemingly quite a low value, considering the mean value of area is around 12 hectares. However, the log transformation severely reduced the size of the are variables, so the error is not as favorable as it appears at first glance. In order to investigate further I will need to see the r squared value of my model.

My model obtained an r squared score of 0.06973189656038548, a dismal result. Overall, the model seems to be unreliable in terms of fire prediction.

5. <u>Analysis, Results and Interpretation</u>

My First question of interest was easy to answer. The highest count values for the number of fires occur in August and September. These must be the fire seasons within this area of Portugal. Additionally, the most popular times of fire occurrences were in the days of Friday, Saturday, and Sunday. This could be because these are the days in which people tend to travel to national forests and so there subsequently is a higher risk of a fire happening.

The second question of interest was much tougher to answer, and my model overall was too simple to answer the question of interest in a successful way. I assumed that I would be able to produce a multiple linear regression model using the dataset provided, and unfortunately my model was very subpar, with a relatively high error and low r squared score value. Taking a look at the residuals diagram for the regression gives us some idea of what is occurring (Figure 10).

We would ideally want to see a horizontal line of points at zero which would indicate a perfect prediction. This error surprisingly forms quite a strong linear relationship between the logarithm of the error and the difference between the predicted and actual values. The model has high error and is this most likely too simple. The model systematically predicts too high of area for larger fire sizes. I explained earlier how the methods for fire prediction have gotten to be quite advanced, and I naively believed a model as simple as a multiple linear regression model could be used to predict such a complex thing as fire size. Obviously this was incorrect. Fire prediction science is complex for a reason and going back to simple statistical tools such as simple linear regression is unrealistic.

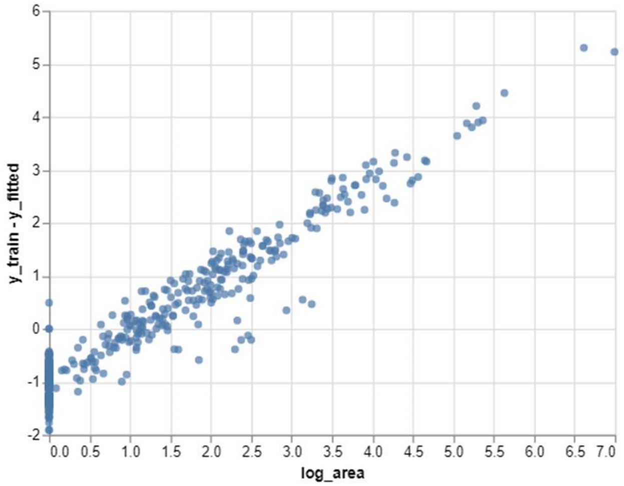


Figure 10, Residual vs Fit Plot

6. Conclusion and Future Work

The main message one should gain from my paper is that it is difficult to interpret the area of a fire given just basic weather information. Even though I was given information such as FWI and other weather metrics I still failed to obtain even a remotely close to reliable model that could predict fire area. It is safe to say that I completely and utterly failed in my initial goal, but this was definitely a learning experience because it helped me realize that predicting fire area is a very difficult process. After all, if we currently had reliable and working model of which fires to watch out for, we would not be experiencing massive wildfires every year in California. It should be noted that my regression model only took one specific approach to attempting to model the area burned. It is possible that another transformation method could be used, certain other variables could have been removed, more outliers deleted. But the fact that the r squared was so low implies that there isn't much room to improve at this point when it comes to multiple linear regression. My results are not very general due to the fact that they produce a model that only "works" for the Portugal northwestern forest, but the techniques used in this regression analysis could technically be applied to any dataset that possesses the same measurements. Being able to predict fire area size is not something that can be done just with simple regression techniques, and so I will give this venture a rest until I can apply more advanced machine learning methods to the data.

7. Appendix

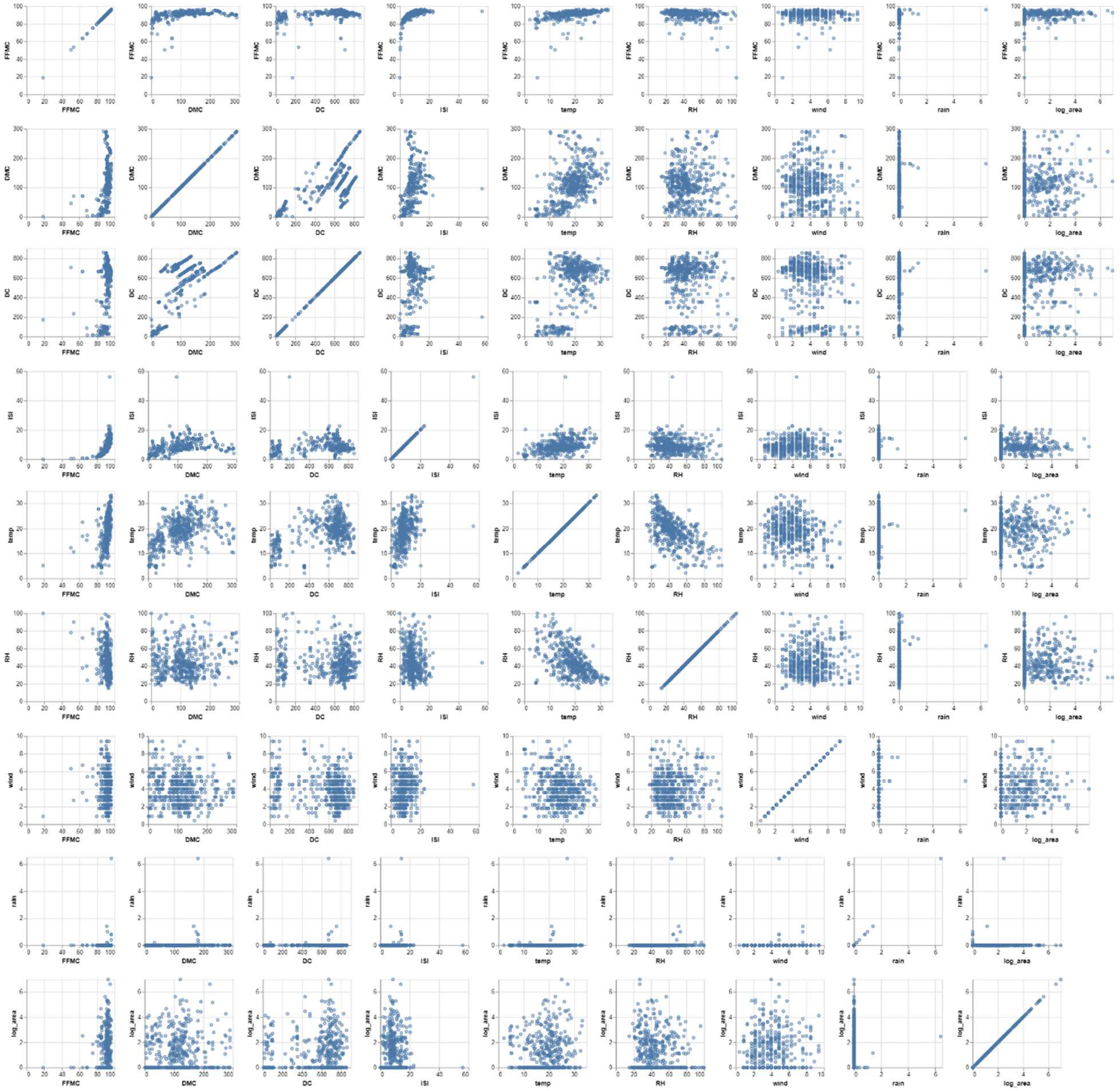


Figure 8, Scatter Plot Matrix