Forensics on Fires: Analysis on atmospheric conditions on forest fires.

Denny Lehman, Tymon Silva, Savita Chari

December 8, 2021

1.0 Introduction

Forest fires in the Pacific Northwest have been the most severe in centuries, causing destruction of property and animal habitat, air pollution, and death. For our group, this is personal - each one of us has been affected by recent fires in some way. It is therefore of great concern to reduce the damage caused by forest fires. One way to combat fires is through human intervention like fire fighting. However, fire fighting can cost huge amounts of money and resources. For out of control wildfires, firefighters may deploy expensive measures like using aircraft or creating physical barriers to limit the spread. Deploying the wrong resources, like sending aircraft to a small fire, is a costly mistake. If we knew the size of the fire based on data, we could respond appropriately for big and small fires alike. In this project, we attempt to answer what conditions cause larger fires so that we may recommend the correct human response and better prepare our communities.

In this research paper, we argue that our best model is sufficient for explaining the size of forest fires. Before building our model, we reason about the causes of forest fires based on the combustion reaction. From this ground work, we build our causal model of forest fires that will be the foundation of our study. After carefully considering available variables, proxy variables, and omitted variables, we selected rainfall, wind, and two variables representing the type of forest fuel for our casual model on explaining fire size. In the data section, we explain how our data was collected and discuss its source. We use exploratory data analysis (EDA) to explore our variables and transform them when necessary. Our modeling section outlines the effectiveness of our model and hypothesis. Finally, results and next steps can be found in the conclusion section.

2.0 Research Design

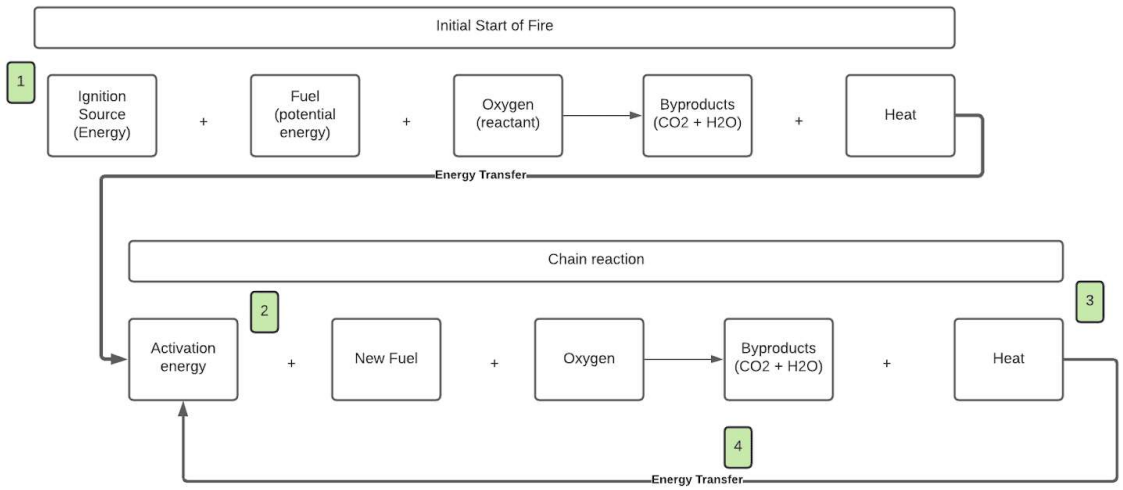
2.1 Research Question

Our main goal in this research project is to examine if a statistical significant relationship exists between environmental conditions and the size of forest fires. Chiefly, we are motivated by the power of atmospheric variables and forest floor conditions to explain the area burned. The outcome of our research will be identifying which, if any, of our derived variables best explains forest fire severity.

2.2 Underlying Causal model

In this section, we reason about environmental factors as causes of forest fires.

To better understand forest fires, we must first understand fire - specifically the combustion reaction.



1. Fires need a source of ignition
   In the initial start of a combustion reaction, an ignition source in the presence of fuel and oxygen starts a fire. The result of this reaction is carbon dioxide, water and heat. The energy used to start the reaction (green 1 callout) is called the ignition source and is our first concept of interest. All fires need some energy as input to start the reaction, otherwise the wood from this researcher's desk and the oxygen in the air would be enough to cause an office fire. A spark from a campfire or strike of lightning is often the ignition source for a forest fire. Without the ignition source, a forest fire will not start and therefore have no area burned which makes it critical to our model.

Summary - fires require an ignition source.

2. Fires need low activation energy to start chain reactions
   Activation energy (green 2 callout) is the amount of energy required to start a combustion reaction. Activation energy and ignition source are similar in that they both are energy at the start of combustion however the difference between them is important. Ignition source describes a small amount of energy used to begin the initial fire. This could be a spark causing small leaves to burn. Activation energy is a general term for the energy required to combust any fuel type, not just the initial source. This term could be used for the energy needed to combust a piece of paper (low) or a massive log (high). We define activation energy as the energy required to burn a fuel source AFTER the initial ignition.

   The source of the activation energy is fascinating. When something burns, it gives off heat which can be used as activation energy to combust other fuel. This is an example of the products of one reaction becoming the reactants of another. This energy cycle of product heat being used as a source for the next exothermic reaction creates a chain reaction. The chain reaction phenomena is how fires spread from one piece of fuel to another. Therefore, the lower the activation energy required, the more likely the fire will grow and spread. In opposition, when activation energy is high, combustion can halt and the fire stops. Examples of fuel with high activation energy are water logged wood. This fuel requires more activation energy to burn than dry wood.

   Summary - the wetter the log, the harder it is to catch on fire

3. Fires need to produce lots of heat to keep chain reactions going
   The heat that is released from a combustion reaction (green 3 callout) is important to spreading fire. As mentioned before, if the heat produced is greater than the activation energy of nearby fuel, it will cause the new fuel to burn. With enough heat, even soggy logs will combust and keep a fire going. For this reason, fuel that gives off a lot of heat (high energy density fuels) will cause forest fires to grow more so than reactions with little heat output.

   Summary - the bigger the log the more energy it outputs

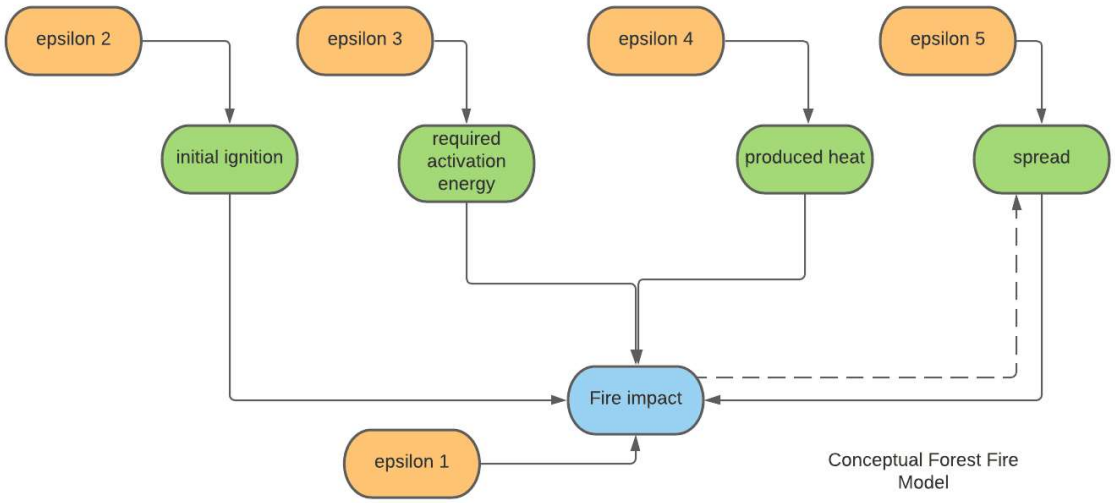
4. Heat needs to be transferred to new fuel
   Fires give off heat, which is used to ignite new fuel and keep the chain reaction going. But that heat needs to be transferred to the next stick. If the fuel is far away or if there is no movement of the heat, the energy will simply rise into the air and the chain reaction will stop. Fire must have a way to spread to new fuel and that requires a transfer of heat energy from combustion to fuel source (green 4 callout). The density of the forest affects the proximity of fuel and wind could play a role in transferring heat to new fuels.

   Summary - Wind transferring heat causes larger fires

In conclusion, forest fires are all about combustion. We theorize that the factors that cause forest fires impact are:

- Initial ignition (spark!)
- Low required activation energy (dryness)
- Product heat (energy density)
- Spread to new fuel (transportation of heat)

Our simple conceptual model is as follows

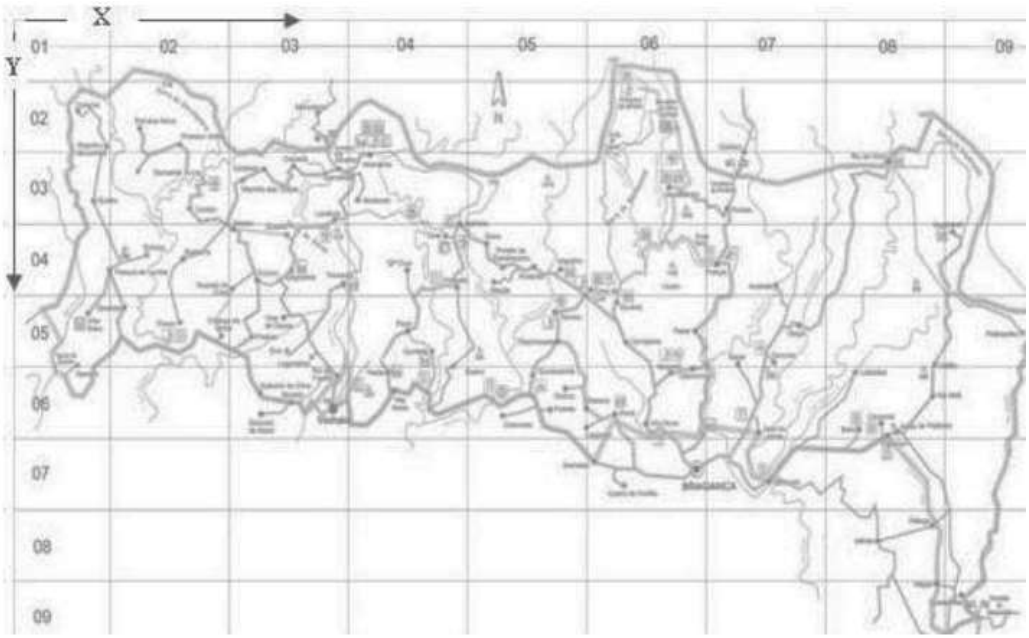


3.0 Data

In this section, we will introduce the variables that will represent the concepts introduced in the research design. From these operationalized variables, we will redefine our causal model and investigate omitted variables.

3.1 The Dataset

The data for our analysis comes from observations on forest fires that occurred in Montesinho Natural Park, which is located in northeastern Portugal and spans an area roughly 300 square miles. This data was collected from January 2000 to December 2003. It contains a total of 517 observations with thirteen columns, four of which are categorical and the remaining numeric. The forest fire data was collected from two sources; (1) a database that contained fire occurrences collected by the park inspector responsible for identifying when and where fires occurred, and (2) a database from a meteorological station located in the center of the park that recorded weather observations within a 30-minute period. Every time a forest fire occurred, key data features were registered into these databases, such as the time, date, spatial location, some of the components of the Canadian Forest Fire Weather Index (FWI)[1] system (e.g. FFMC, DMC, DC, and ISI), wind speeds, temperature, relative humidity, and rain.



Map of Montesinho Natural Park in Portugal[2]

The two databases were integrated into a single dataset, providing us the 517 records used in our research.

[1] https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi
[2] http://www3.dsi.uminho.pt/pcortez/fires.pdf

3.2 Operationalization

We will be using the following covariates from this dataset to achieve our modeling goals:

Concept	Proxy Variable	Units	Range	Description	Source
Fire impact	Area Burned	Hectares (ha)	0.00 to 1090.84	the burned area of the forest	Portugal Dataset
Initial ignition	Fine Fuel Moisture Code (FFMC)	Unitless index	18.7 to 96.2	Represents fuel moisture of forest litter fuels	Portugal Dataset
Required activation energy	Rainfall	millimeters per square meter (mm/m^2)	0.0 to 6.4	Current rainfall in the environment	Portugal Dataset
Produced heat	Build Up Index (BUI)	Unitless index	2.2 to 315.6	Represents the total amount of combustible fuel	Calculated Variable
Spread	Wind speed	Kilometer per hour (km/h)	0.40 - 9.40	Current wind speed	Portugal Dataset

Area Burned

The outcome of our model is the size and impact of forest fires. To operationalize this concept, we use the area burned by the fire in hectares. This operationalization comes with some flaws, namely that the area burned may not directly correlate with the damage of the fire. Brush fires may do little damage to the environment but cover huge land areas while intense wildfires may do more damage over less area. Additionally, we must consider how we operationalize large and small fires. We have many 0 valued records for the area burned in our data. This does not represent the absence of a forest fire, merely one that is smaller than the measurement tolerance of our data gathering methods. We believe it is not prudent to remove the 0 value records from our data set because they represent small fires which are of interest to our study. We wish to understand the causes of small fires as a comparison to large ones, which is why we selected area burned as our outcome variable.

FFMC

To operationalize our concept of initial ignition, we use the Fine Fuel Moisture Code (“FFMC”). The FFMC represents the flammability of the forest floor by taking into account dryness of the

kindling on the ground. The higher the index, the more likely the fuel is to catch fire. This is NOT related to current rainfall because FFMC calculates moisture content with a 16 hour time lag[3].

Rain

The lower the activation energy for a combustion reaction, the less energy is required for fuel to catch on fire. We operationalize this concept with current rainfall. Wet fuel requires more activation energy to start a combustion process than dry fuel. Rain introduces water to the fuel making it hard for forest fires to continue burning. Therefore, we use rain as a proxy variable to the activation energy of the combustion reaction.

BUI

The more heat produced from the current combustion reaction, the more energy there is to start another combustion reaction. This concept will be operationalized with the Build Up Index ("BUI"), which is a metric that we have calculated based on two variables in the dataset, DC and DMC.[4] BUI describes the quality of the subsurface fuel in a forest, namely the dryness and amount of medium and large sized logs on the forest floor. If these are present in a forest fire, the fires will be more intense, generating more heat. The more heat, the more likely the fire will overcome wet logs and live trees, which require high activation energy to start burning. Looking at the live tree example, if only kindling is present, little heat will be produced from combustion which will not be enough activation energy to ignite the live trees. However, when more hearty, energy dense fuel is present, enough heat will be produced to ignite live trees. This will impact the area burned because small kindling fires will not generate enough heat to ignite larger, wetter, fuel and spread to larger areas.

Note: a damp log and a dry log will have the same amount of produced heat, which is why, conceptually, rain is not a confounding variable with produced heat. We understand that in practice, our operationalized variables contain dryness and rain. These are related as examined in the later section on bias.

Wind

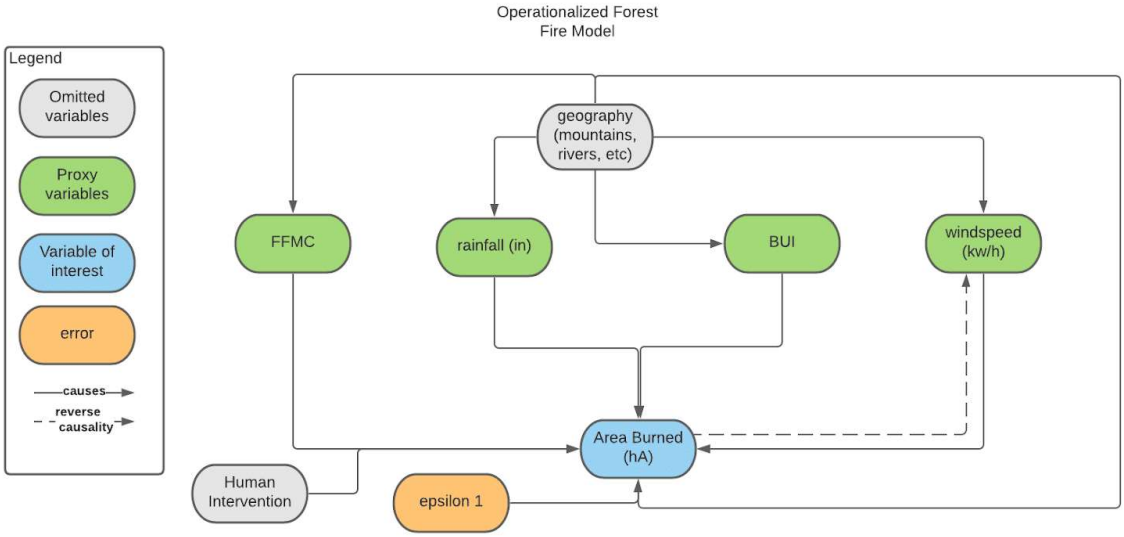
In order for fire to grow, a flame must spread to a new fuel source. We might represent the spreadability of fire as a factor of current wind speed. Wind will move embers of a fire to new fuel sources, thus increasing the area burned. Additionally, wind will also provide oxygen to the blaze. We expect wind to have reverse causality with area burned because of research that suggests fires can cause strong winds[5]. See the later section on bias for more details.

---

[3] https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system
[4] https://wikifire.wsl.ch/tiki-index8720.html?page=Buildup+index
[5]https://www.accuweather.com/en/weather-news/how-destructive-wildfires-create-their-own-weather/346337

3.3 Operationalized Model

Below is the operationalized casual model. The variables are measurable proxies for our concepts. These variables will allow us to derive conclusions about our research question.



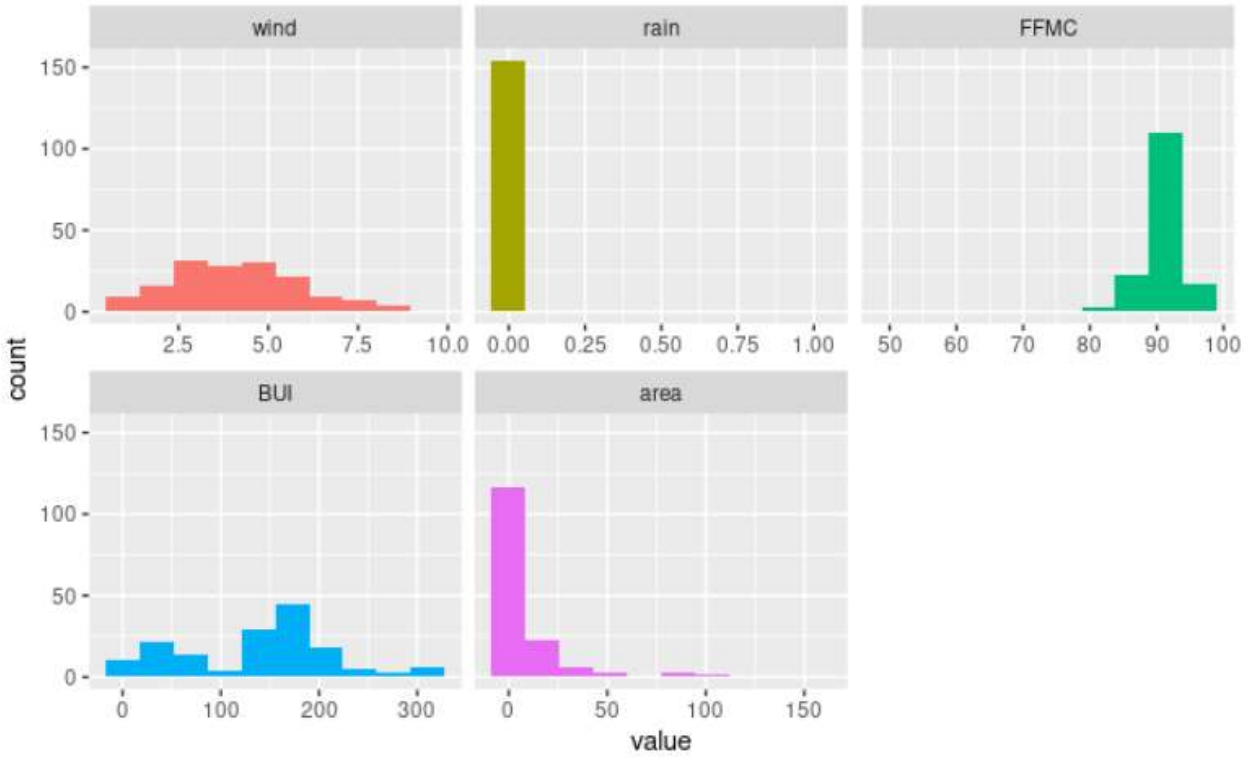
4.0 Exploratory Data Analysis (EDA)

In this section, we will explore the data and introduce important information about the variables. Then, we will explore their distributions and consider transformations to meet modeling requirements.

4.1 Variable Distributions

In order to perform our EDA, we randomly selected 30% of the records in our dataset and reserved them for our data exploration. The remaining 70% was set aside for our statistical analysis and model building.

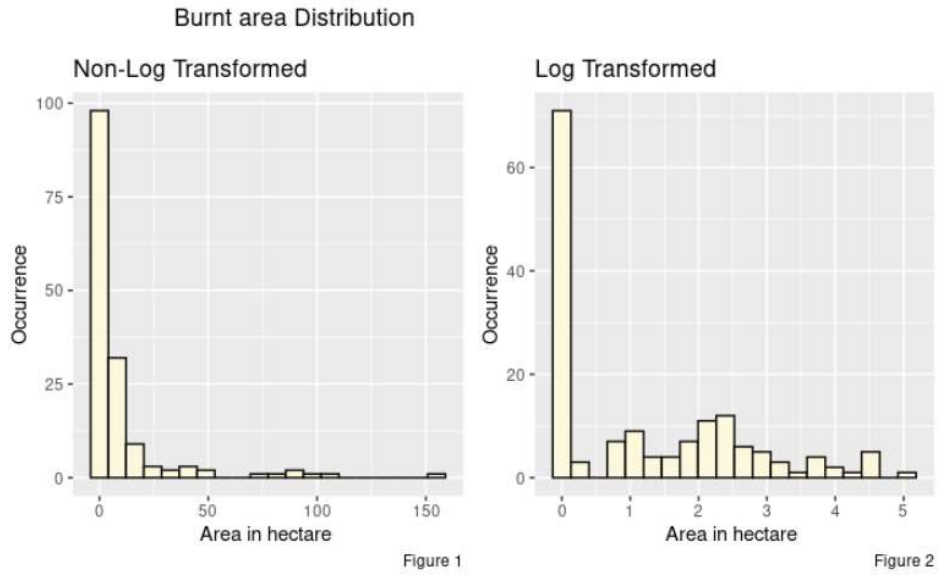
We begin with an initial review of our data by looking at each variable's distribution. The histograms below illustrate the distributions of our variables of interest:



Visualization of all variable distributions

At a glance, all of our variables, besides wind, show varying degrees of skewness. For this reason, many variables show potential to be improved by transformations. It is important to note that BUI was not part of the original dataset. Instead, BUI was calculated using two other variables that were provided in the dataset, namely the Duff Moisture Code (DMC) and the Drought Code (DC) (see reference 5 in Reference section for calculation details).

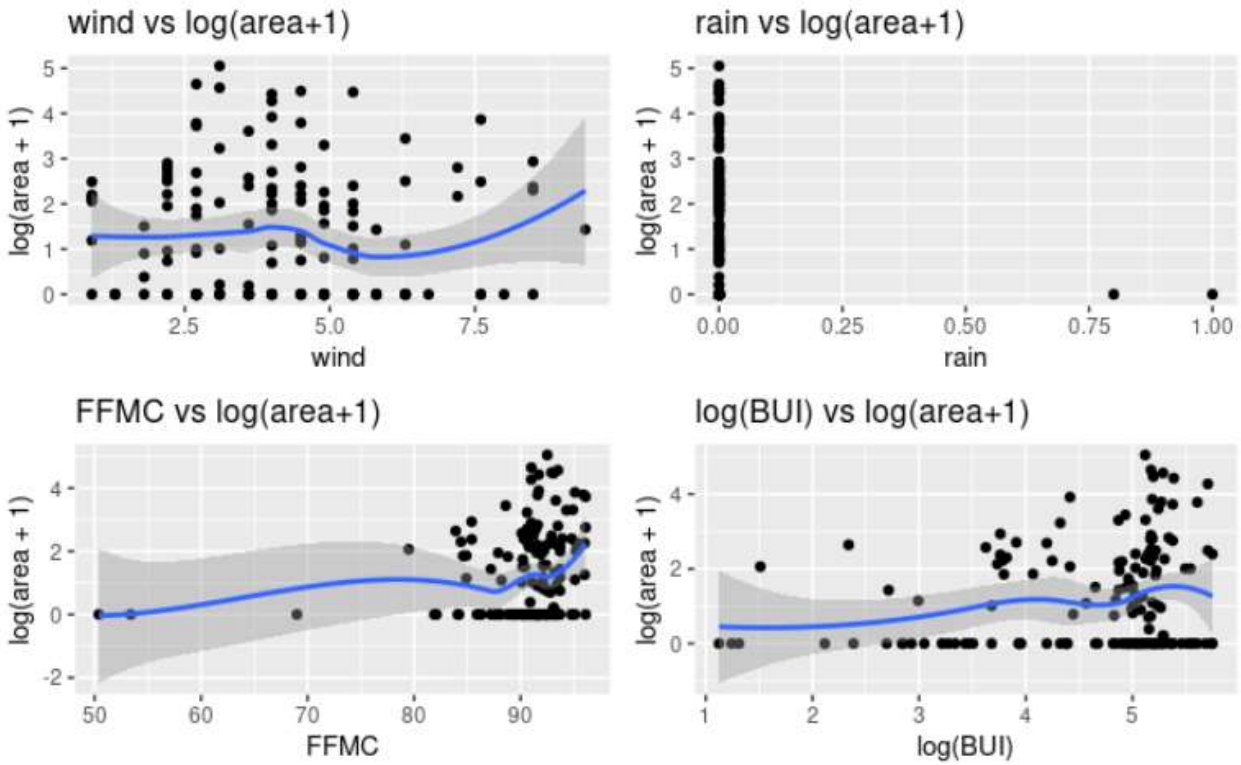
The outcome variable, area, was highly skewed to the right. Most of the area burned had a value of 0 hectares. One data cleaning option was to remove the 0 values for burn area if they represented no fire. However, as mentioned in section 3.2, these zero valued burns were actual fires. The data gathering method had very poor tolerance for measuring small fires and most were recorded as 0.0. To study the difference in causes of small and large fires, we chose to keep the 0.0 records and focus on finding transformations of this variable. After performing many transformations on area and observing the effect each had on the linear relationship with our predictor variables, it was decided a log transformation of area would improve the linear relationships with each of the predictor variables. The histograms below show a slight improvement (towards a normal distribution) of the outcome variable by comparing the distribution of the area variable as is with the distribution of log(area +1).



Comparison of area distributions

4.2 Linear Relationships

Similar investigations and analyses were performed with the other variables to see if any other transformations could help improve the linear relationships they had with the log transformed area variable. Ultimately, we decided that only the linear relationship of BUI visibly benefited from a log transformation. The following grid of scatter plots describes the linear relationships for each of our predictors with all appropriate transformations applied.



Visualization of scatter plots describing linear relationships

In the scatter plots above, most variables show reasonable relationships with our outcome variable. However, the rain variable is of particular concern. As mentioned earlier, the data comes from Portugal and it experiences very little rainfall, especially during the hot summer months. Additionally, rain was poorly measured in the data gathering stage. Researchers did not have accurate tools to measure low rainfall which therefore reduced any small amounts of rain to 0 inches. In our full dataset, the rainfall ranges between 0.0 and 6.4 with 509 out of 517 records having 0 value. This distribution is so extreme that all transformations attempted on this variable failed to produce a meaningful relationship with our outcome variable. Furthermore, we expect the robust standard errors to be relatively large, since this variable would be providing very little information on what happens to the area burned when it rains. With that said, we also recognize the importance of rain as a causal variable on fire and fire spread. In order to defend keeping rain as a predictor in our model, we decided to convert it to an indicator variable, augmenting the variable such that it indicates whether rain is present or not. As a binary variable, we might have a more explainable model with less biased coefficients. The binary rain variable in our model, if statistically significant, would now be interpreted as having a percentage change impact on the amount of area burned conditional on rain being present during the time the forest fire occured.

5.0 Modeling

In this section, we will outline our modeling approach and highlight 3 models with which we will test our hypothesis that our predictor variables explain forest fire burn area. We use linear regression as our model and interpret the coefficients of our predictor variables. Our approach is borrowed from RDADA which recommends starting with a simple model and then adding complexity. In the complex model, we will determine which variables were most impactful. Finally, our last model will include the most important variables from our analysis of the first two models.

Here are the three models we ran:

1. Model 1: Natural logarithm of burn area on natural logarithm of BUI

$$log(area + 1) = \beta_0 + \beta_1 log(BUI) + \varepsilon$$

2. Model 2: Model 1 plus the remaining variables of interest

$$log(area + 1) = \beta_0 + \beta_1 log(BUI) + \beta_2 wind + \beta_3 rain\_binary + \beta_4 FFMC + \varepsilon$$

3. Model 3: Natural logarithm of burn area on wind

$$log(area + 1) = \beta_0 + \beta_1 wind + \varepsilon$$

6.0 Results

In this section, we test the 3 different models that explain forest fire burn areas and compare each one. We will select the most appropriate model that supports our goal.

Of note is that our outcome variable has been transformed from area burned to log(area burned +1). The way we reason about our new outcome variable is that the coefficients of predictors will represent a percentage change in area burned (in hectares) as opposed to a unit change in area burned. Put differently, our coefficients now represent the percent increase or decrease of area burned.

6.1 Model 1: BUI only

Our simple model will contain only one variable. We found that BUI represented a variable that satisfies both causal theory and showed linearity during EDA. In theory, BUI is critical to explain areas burned by forest fires. When log transformed, it appears to have some linearity with our variable of interest. For these reasons, we start with a simple model representing log(BUI) explaining the log(area burned + 1). Below is the summary of the coeff test on the model.

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.946883   0.394191  2.4021  0.01681 *
log(BUI)    0.024429   0.082624  0.2957  0.76766
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the coeff test show the estimations, the standard error, t-test statistic, and the p value of the coefficients. The null hypothesis of the t-test is that log(BUI) has no effect on area burned. With a p-value > 0.05, we fail to reject the null hypothesis. We cannot conclude that log(BUI) causes forest fires, which is a serious blow to our theory. Notable in this model is that the intercept term is statistically significant. The intercept term represents when the BUI term is 1 (log(BUI) = 0) and explains the expectation of the outcome variable unconditional on BUI. The result of this model is that log(BUI) fails to reject the null hypothesis and does not explain forest fires well.

6.2 Model 2: All variables of interest

We chose this next model because it contains all of our variables that we theorize will explain forest fires. This model contains log(BUI), wind, rain as a binary variable, and FFMC. With the additional variables, this model is more complex than the first. As a reminder, our hypothesis is that these variables, which come from our theory on forest fires, will explain area burn. Below are the results of model 2.

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9629663  1.0575243  0.9106   0.3631
log(BUI)     0.0954846  0.1002442  0.9525   0.3415
wind         0.0961630  0.0402457  2.3894   0.0174 *
rain_binary -0.6367068  0.4003307 -1.5905   0.1126
FFMC        -0.0079667  0.0139260 -0.5721   0.5676
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coeftest shows the estimations of the coefficients and each p-value. The null hypothesis for the t-test is that the coefficient has no effect in explaining area burned. All variable coefficients except the one for wind had p values > 0.05 which fail to reject the null hypothesis. The wind variable had a p value < 0.05. We reject the null hypothesis for wind in favor of the alternative. We interpret this to mean that wind likely causes forest fires to spread. The coefficient for wind explains that for every increase in unit of wind, the area burned increases by ~9.6%. Overall, we reject the null hypotheses for all variables besides wind in model 2.

6.3 Model 3: “Best” combination of variables

In our third model, we choose the best estimators for our outcome variable based on the first two models. Unfortunately, our first model gave us little information about the causality of forest fires. Our second model did better, showing that wind likely explains the area burned. For our third model, we will simply test if wind alone makes for a better model than the others. Below are the results of our coeff test.

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.735300   0.174045  4.2248 3.036e-05 ***
wind        0.082652   0.039619  2.0862   0.03767 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results show that both the intercept and the wind are statistically significant to our model. With p-values < 0.05 we reject the null hypothesis for the wind coefficient. Our standard error dropped compared to model 2 which means we have a tighter confidence interval on the true effect we are estimating. Our coefficient for wind was 0.0826, representing that for a unit change in wind, area burned changes by ~8.3%.

6.4 Comparison of models

Finally, to compare the three models, we use the f-test. Our f statistic will look at the change in variance as the models change the number of variables. Our null hypothesis is that the addition of variables will have no significant effect on the explained variance of the model. Put differently, the null hypothesis is that adding variables does not improve the explanation of forest fire area

burn. Below are the results of our model comparison.

	Dependent variable:		
	log(area + 1)		
	(1)	(2)	(3)
log(BUI)	0.024	0.095	
	(0.087)	(0.105)	
wind		0.096*	0.083*
		(0.042)	(0.041)
rain_binary		-0.637	
		(0.581)	
FFMC		-0.008	
		(0.016)	
Constant	0.947*	0.963	0.735***
	(0.416)	(1.267)	(0.178)
Observations	361	361	361
R2	0.0002	0.017	0.011
Adjusted R2	-0.003	0.006	0.008
Residual Std. Error	1.402 (df = 359)	1.396 (df = 356)	1.394 (df = 359)
F Statistic	0.080 (df = 1; 359)	1.509 (df = 4; 356)	4.083* (df = 1; 359)
Note:			*p<0.05; **p<0.01; ***p<0.001

Model 3 had an f-statistic of 4.083. When converted to a p-value, we find that it is statistically significant. With a p-value < 0.05, we reject the null hypothesis for the f-test that this model does not explain more variance than the other models. Therefore, we select model 3 as our best explanation of forest fire area burn. To summarize, both our coefficients for wind and the intercept term are statistically significant. The wind variable represents the spreadability of the fire and the coefficient represents the conditional expectation of log(area burned +1) given the km/h of wind. The intercept term represents the unconditional expectation of area burned, which is a constant.

After a statistically significant f-test, we select model 3 as our best explanatory model for forest fire area burn. Our exploratory data analysis was damning to our expectation that we had data to support our hypothesis, so we were surprised that we found a statistically significant model. Our result is that wind is the only variable in our theory that might explain forest fire area burn.

7.0 Limitations

7.1 Statistical Limitations

Before data cleanup, our data set contains 500+ records of forest fires. Given the size of our dataset is larger than 100 data points, we will use large sample assumptions for our linear model. There are two requirements for large sample models. In this section, we will introduce both.

IID

The first assumption for a large sample linear model is that the sampled data is independent and identically distributed ("IID"). Some common concerns surrounding this assumption is if the data is in some way conditional on other data points, such as from time series data sets or datasets that suffer from clustering. For IID assumptions to hold, researchers avoid using time series data. The seasonality of the data means that current observations are not independent of past observations. Our data was collected over the course of many years which imply time series data. The fires from a year prior could absolutely affect the forest conditions in the future and in turn impact the size of a fire in the following years. If a large fire burned in March, April will have less fuel to burn, hence our data will be conditional on past events.

We also have cause for concern for IID due to the geography. The data comes from a single forest, where the location of the fires impacts other fires. Fires and atmospheric variables alike could conditionally change other data points due to the location proximity. We must be cautious of geographic clustering effects that could damage the IID assumption.

However, studying forest fires has few alternatives. This data comes from a natural experiment. There is no real alternative to recording forest fires, which are temporal and geographic in nature. Fires have start and end times and occur in specific locations. There is no way of sampling fires all at once or have all fires be geographically separated. One alternative would be to generate simulated data, which was outside of the scope of this project. For these reasons, we proceeded with our study even though the IID assumption had flaws.

Unique BLP Exists

The last assumption for a linear model under the large sample assumption is that the Best Linear Predictor ("BLP") exists. We expected our dataset to be rich with information on weather and forest floor conditions measured as numeric values. These would make for interesting models. However, against our expectations, our data was filled with single valued records which added little to our dataset or understanding. Variables like area burned and rainfall were mostly 0's missing even the slightest noise. This type of data is not helpful in building linear models because it has heavy tails - even with transformation. In the case of rain, we tried transforming the data into a binary class (rain or no rain), but the overwhelming number of 0 values caused a class imbalance where no rain dominated. Without using methods for correcting for class

imbalance, we were left with a poor variable for modeling. The very high robust standard error for rain_binary in model 2 supports our concern. Still, we believe that a unique BLP exists and this assumption for large-sample models is satisfied, albeit that belief is fragile. The rain variable in particular, is nearly a vertical line when plotted with area burned. If it weren't for those handful of observations where rainfall was present, defending the unique BLP assumption would be impossible for this relationship.

To conclude our assumption, we require no multicollinearity among the variables. We only had one relationship that was cause for concern because it could have strong collinearity. That relationship is between rain and FFMC and BUI. FFMC and BUI are indexes that use dampness of the forest floor in their calculation. While rain could impact the dampness of the forest floor that changes FFMC and BUI, the indexes are calculated on a time lag. This means that only prior period rain will cause changes to BUI or FFMC, not our variable of current rainfall. With that relationship explained, no other variables were a theoretical cause for multicollinearity concern and we meet this requirement. In future work, we would recommend challenging this assumption with a VIF test in R. For this study, we place this test outside of our scope.

7.2 Structural Limitations

In this section, we will describe the most interesting structural limitations of our model. Considering the complexity of weather and geography, we will not use this area to list all the areas of bias but only the most interesting to our audience.

Reverse Causality

In our model, wind causes forest fires to grow. However, forest fire spread also causes wind to increase. The positive feedback loop of wind causing fires and fires causing wind means that our coefficient for wind is biased and the sign of that bias is positive. Therefore, we may be overestimating the effect of wind on area burned by fire.

Outcome Variable on the Right Hand Side (RHS) - Confounding Variables

No confounding variables were identified in our causal pathway diagram. Although this seems unlikely to reflect reality, we can't find good reasoning to believe that any of our concepts/ predictor variables would have a causal effect with one another. However, the omitted variables Human Intervention and Geography (discussed in the Omitted Variable Bias section below) would likely by confounding variables, since Geography would have a causal relationship with our atmospheric variables wind and rain and those atmospheric variables in turn may have a causal effect on Human Intervention. If those variables were included in our model, our interpretation of the beta coefficients would be problematic.

Omitted Variable Bias

Human Intervention

Human Intervention (HI) is an omitted variable that is also responsible for putting out fires. Human firefighters are an example of how humans can reduce the area burned regardless of the other variables. In our true causal model, HI is included as a covariate. However, since this data was not available to us, and we had no way of operationalizing it, it was an omitted variable. And, because it is an omitted variable, we can reason about the bias it introduces to our coefficients. We argue the sign of the beta coefficient for HI is negative - the more human intervention we have the smaller the burn area. We can further speculate the sign of the bias to be positive for both Wind and Rain predictor variables. In this case, the direction of the bias for wind is away from zero whereas the direction of the bias for rain is towards zero. The direction of bias for wind is of particular concern, since wind was our only statistically significant predictor. This calls into question the interpretability and reliability of this coefficient. The impact of the bias from omitted HI on FFMC and BUI is unknown since we can't reason about the relationship between HI and these two variables.

Geography

Geography is a huge omitted variable that could affect bias in many different ways. Rivers surrounding a forest could limit the area burned by being a natural barrier to additional fuel. Mountains could limit area burned in the same way - limiting the fire's spread to other areas. However, mountains may also cause meteorological effects, like limiting rain cloud formation and causing the rainshadow effect[6] (ie. drought)[7]. Geography is quite complex and likely impacts most, if not all, of our variables in multiple ways. It is because of this complexity that makes it difficult to reason about the sign and the direction of the introduced bias by omitting this variable. We believe the concept of Geography might be too general, and would have to be unpacked such that there is a distinction between, for example, a mountain that has a positive causal relationship with atmospheric variables and a mountain that has a negative causal relationship with atmospheric variables. Currently, our concept of Geography does not make that distinction.

8.0 Conclusion

In this study, we attempt to test whether environmental conditions cause greater forest fire size. To do this, we built a foundational theory that rain, wind, and both quality and quantity of forest debris cause increased burn area of forest fires. To represent these concepts, we used data gathered on environmental conditions during forest fires that took place in Portugal. After making appropriate transformations, we modeled our data using linear regression to seek insights into what causes large fires. We found an increase in wind speed to be statistically

[6] https://www.thinglink.com/scene/779748923773288450
[7] https://www.nps.gov/moja/learn/education/classrooms/upload/MDD-Unit-II-Deserts.pdf

significant in explaining a percent increase in burn area. All other coefficients were not statistically significant.

With proper modeling assumptions met, our study would be conclusive. However, we found some cause for concern with our assumptions, namely the IID and BLP assumptions for large model assumptions. We recommend using our model with reservation.

8.1 Discussion

The outcome of this study would help firefighters, home owners, policy makers, and other stakeholders to understand and identify when they are in danger of a major blaze. Providing these stakeholders the variables that will help them anticipate the size of a forest fire, before it reaches its potential, enables them to better prepare communities and to choose more appropriate responses to mitigate and minimize damage without overreaction.

Our conclusion is that wind is the best explanation of forest fire burn area. Using data from a natural experiment that has modeling assumption concerns, we tested our theory on the cause of forest fire spread with linear regression models. After checking simple and complex models, we arrived at the conclusion that unit increases in wind speed (km/h) cause roughly 8.3% increases in the area burned by forest fires. Our data showed concern for meeting IID and best linear predictor assumptions. So, the results of this study should be taken at face value.

This study is a case where gathering more data may not help improve the model. The data gathering process used poor measurement methods which in turn produced poor quality data. Put differently, our model would not improve with more 0 valued area burn or 0 valued rainfall data points. Instead of gathering more data, we believe researchers should focus on collecting higher quality data, perhaps achieved with better measurement tools. We recommend a review of the data gathering process before future studies can be performed.

This dataset was challenging to use for linear regression. With so many records with 0 area burned, there may be an opportunity to make this a classification problem, potentially making this dataset a good exercise to perform logistic regression. We recommend future researchers consider classifying forest fires into small and large fires and attempt to model these classes, rather than a numerical outcome variable. The purpose of preparing stakeholders for large fires will still be met. We also recommend using corrections for class imbalances. Large fires are much more important to predict than small ones, so correcting for type 2 error rate on large fires is a must.

All in all, we hope that our findings that wind causes increased forest fire spread can be used to help firefighters select the best response to forest fires.

9.0 Reference

1. “Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System.” Canadian Wildland Fire Information System / Système Canadien d’information Sur Les Feux de Végétation, https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi. Accessed 8 Dec. 2021.
2. Cortez, Paulo, and Anibal Morais. “A Data Mining Approach to Predict Forest Fires Using Meteorological Data.” Department of Information Systems/R&D Algoritmi Centre, University of Minho. Accessed 8 Dec. 2021.
3. “Deserts.” National Park Service Classroom, National Park Service, https://www.nps.gov/moja/learn/education/classrooms/upload/MDD-Unit-II-Deserts.pdf. Accessed 8 Dec. 2021.
4. “Fire Weather Index (FWI) System | NWCG.” NWCG | NWCG Is an Operational Group Designed to Coordinate Programs of the Participating Wildfire Management Agencies., https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system. Accessed 8 Dec. 2021.
5. “Fire Weather Indices Wiki | Buildup Index.” Wikifire, https://wikifire.wsl.ch/tiki-index8720.html?page=Buildup+index. Accessed 8 Dec. 2021.
6. Schmidt, Amanda. “How Destructive Wildfires Create Their Own Weather.” Accuweather, https://www.accuweather.com/en/weather-news/how-destructive-wildfires-create-their-own-weather/346337. Accessed 8 Dec. 2021.
7. Thinglink. “Rain Shadow on the Gobi Desert.” ThingLink: Create Unique Experiences with Interactive Images, Videos & 360° Media, https://www.thinglink.com/scene/779748923773288450. Accessed 8 Dec. 2021.
8. UCI Machine Learning Repository: Forest Fires Data Set. http://archive.ics.uci.edu/ml/datasets/Forest+Fires. Accessed 8 Dec. 2021.