A Statistical Approach to Estimating Burned Area of Forest Fires

Matt Parker

School of Statistics, University of Minnesota

STAT 4893W: Consultation and Communication for Statisticians

Professor Kuzmak

December 18, 2023

Abstract

Using data provided by A U.S. Forest Service Incident Management Team (USFS IMT) in Colorado, our statistical research explored the relationship between burned area of forest fires and 12 spatial, temporal, and meteorological variables. Our objectives were to create an optimal model for predicting burned area, and simulating burned area over a variety of conditions. For model optimizations, we explored the predictive power of a multiple linear regression model and a random forest model. For model simulation, we analyzed three regression trees to explore the predicted burned area over various classifications of our predictor variables. Ultimately, we were not able to create an optimal model that provides statistically significant outcomes. With that said, we created interpretable visuals in the form of our regression trees, while improving upon our initial MLR model with random forest when comparing test MSE values.

Introduction

In the midst of major changes to our environment caused by climate change, it's important to be aware of the various effects this ongoing issue will pose to our communities. One of the most concerning effects of climate change is forest fires, which we saw hit record levels throughout California's wildfire season in 2020 (Kerlin, 2022). Various environmental factors can lead to forest fires, which can be extremely difficult to extinguish given quick spread and lack of detection ("Why Is It So Hard For Firefighters To Put Out Wildfires?", 2023). With that said, being able to predict the damage caused by forest fires can lead to harm reduction on a variety of counts, not limited to forest preservation and the safety of our communities. Though descriptive data can provide a glimpse into how large forest fires will become, we can dive

deeper, using statistical methods to determine the relationship between burned area and various spatial, temporal, and meteorological variables.

A U.S. Forest Service Incident Management Team (USFS IMT) in Colorado has tasked us with two research objectives in relation to forest fire management support: creating an optimal statistical model using a data set providing 517 forest fire incidents from Portugal, and simulating the burned area of a forest fire over various factors. Several statistical methods will be employed to determine the importance of each variable present in our dataset, using the most important predictors to simulate the burned area of a forest fire. In doing so, the USFS IMT will be able to detect fires more effectively, and utilizes their resources more efficiently.

To address our research objective, our statistical analysis will address two key questions: Which model provides the best predictive performance, and which variables are most effective in predicting burned area of forest fires?

Methods and Materials

The statistical analysis will employ data retrieved from an observational study of 517 forest fires in Montesinho natural park in Portugal. Each sample provides the burned area of the region under observation, which we denote as our response variable, and 12 predictors variables. The statistical techniques may not include all 12 predictors, depending on their respective significance to our response.

The 12 predictors are split into three categories: spatial, temporal, and meteorological variables. Spatial and temporal variables include X and Y coordinates on a $1-9$ scale, as well as the month and day of the week in which the fire occurred. Four of the meteorological variables can be derived from the Forest Weather Index (FWI), which include Fine Fuel Moisture Code

(FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI) (Cortez, Morais, 2007, p. 3). The last four predictors include basic meteorological variables comprised of temperature in degrees celsius, relative humidity in percentage, wind speed in km/h, and rain in mm/m^2.

Before we dive into our statistical techniques, we begin by analyzing the predictor and response variables and their respective outcomes present in the data set. It's especially vital to analyze the results of the response, as issues with skewness may lead to unfavorable results in our statistical models, especially in multiple linear regression where a few underlying assumptions must be met (Saxena, 2020). Transformations are an essential tool that we will explore to address this issue, as they can improve fit and correct violations of statistical assumptions made when fitting our model, which includes ensuring variance is homogenous across the spread of the distribution (Segall, 2023).

We'll begin our analysis by conducting multiple linear regression (MLR) and stepwise selection. The goal of MLR is to explain the relationship between two or more predictors and the response variable, while stepwise selection is a form of regression that utilizes statistical algorithms to determine which predictors remain in the model. MLR and stepwise selection are especially advantageous in that they are easy to interpret and implement. That said, stepwise selection runs into several statistical problems, including biased estimates as a result of overfitting the dataset and increased Type I error (Harrell, 2015). In addition, stepwise selection does not consider the context of the study when selecting variables, making the final model difficult to interpret from a scientific perspective. Ultimately, MLR and stepwise selection will provide a baseline for our analysis, as we can seek to further improve upon our regression model by utilizing Mean Square Error (MSE) and AIC as a means of model comparison.

To extend beyond classical regression, we will simulate the burned area of a forest fire over a variety of conditions using regression trees. Regression trees are the most simple statistical approach among tree-based methods, producing interpretable results in the form of decision trees. These trees split our data set into distinct regions, where predictions for the observations are made using the mean of the response variable for the observations within each region (James, et. al., 2023, p. 330). To construct these regions, denoted by $R_1,...,R_J$, we select predictor variables Xj and cutpoints *s* that minimize the residual sum of squares (RSS) when dividing the data into regions ($\{X|Xj < s\}$ and $\{X|Xj \geq s\}$) (James, et. al., 2023, p. 330). For any given observation, the observed value of X_j determines the region Rj it belongs to based on whether it is greater than or less than the value of the cutpoint *s*. Since regression trees are built using a top-down approach, observations will be assigned to regions R_1 or R_2 (James, et. al., 2023, p. 330). We can repeat this process by fitting X_j and *s* in the resulting regions, creating more regions. In producing large or unpruned trees, fitting continues until no more than five observations are contained within each region (James, et. al., 2023, p. 331). Once all regions are defined and the terminal nodes are reached, we make predictions on the observations. In our case, the regression tree should provide a visually appealing prediction of the burned area of forest fires, given the node results of several spatial, temporal, and meteorological variables. Despite their interpretability, regression trees suffer from high variance, as they are highly tuned to the training set used in constructing the tree (Boehmke, 2018).

Sticking with tree-based methods, we will proceed with our initial objective, focusing on model optimization using the random forest algorithm. Random forest is a special case of the bootstrap aggregating tree-based method, otherwise known as Bagging. In order to reduce variance of a statistical learning method, we need to split our data into many training sets, create a statistical model for each set and take the average of each prediction produced by the models. Bagging conducts these steps over B regression trees, leading to variance reduction of the trees. Random forest follows the same bootstrap steps involved in bagging, but at each split, instead of considering all p predictor variables, a random sample of m predictors (m ≈ p/3) is chosen (James, et. al., 2023, p. 343). Though it may sound disadvantageous to limit the number of predictors we consider at each split in our decision trees, it helps reduce the problem of highly correlated trees commonly found in Bagging. This is evident when working with a data set that contains particularly strong predictors: the collection of trees used in our bootstrap will contain these strong predictors at the top split, producing a strong similarity amongst the bagged trees, resulting in high correlation (James et al., 2023, p. 344). In addition to reducing the number

of predictors tried at each split, we will tune the model by testing various values for *m*, which may assist in discovering the optimal model. Though random forest won't provide the easily interpretable results of regression trees, we are able to visualize the variable importance of all variables present in predicting burned area of forest fires by the mean decrease accuracy measure, which measures the decrease in MSE due to splits over a given predator, and the increase in node purity, which is calculated using the reduction of sum of squared errors when a predictor is chosen at any node (Hoare, 2022).

Results

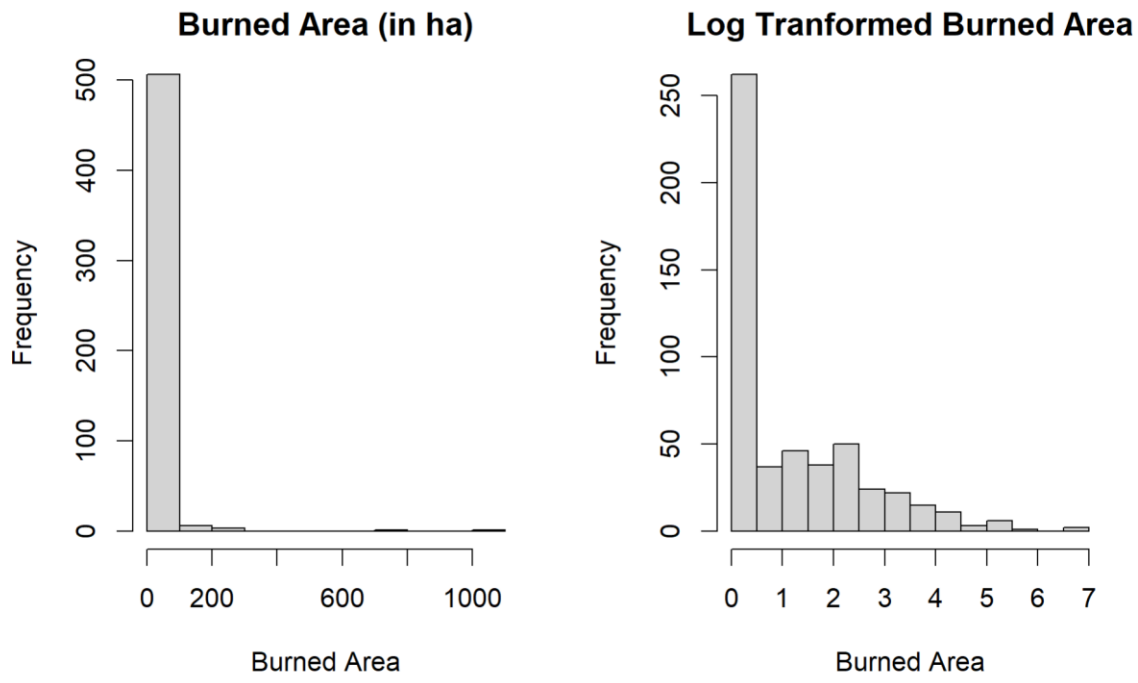
We'll begin with a table of descriptive statistics of the predictor and response variables, providing a glimpse into how we may interpret and transform some of these variables in our analysis. We decided to remove coordinates X and Y as predictors, given that the effects of coordinates would differ from Portugal to Colorado.

	Minimum	**Median**	**Mean**	**Max**
Month	1.00	7.00	6.76	12.00
Day	1.00	4.00	3.74	7.00
FFMC	18.70	91.60	90.64	96.20
DMC	1.10	108.30	110.87	291.30
DC	7.90	664.20	547.94	860.60

ISI	0	8.40	9.02	56.10
Temperature	2.20	19.30	18.89	33.30
RH	15.00	42.00	44.29	100.00
Wind	0.40	4.00	4.02	9.40
Rain	0	0	0.02	6.40
Burned Area	0	0.52	12.85	1090.84

Amongst the 11 variables present in the dataset, DC and response variable Burned Area stand out with particularly large differences between their minimum and maximum values. Burned Area has an extremely small median at 0.52, and what we can likely deem an outlier with a maximum of 1090.84. We can thus conclude that the data is highly skewed, and would benefit from a log transformation that handles both outliers and the large number of zeros (247 of them) in the dataset.

Below are the results of our pre- and post-transformed response variable in histogram form.

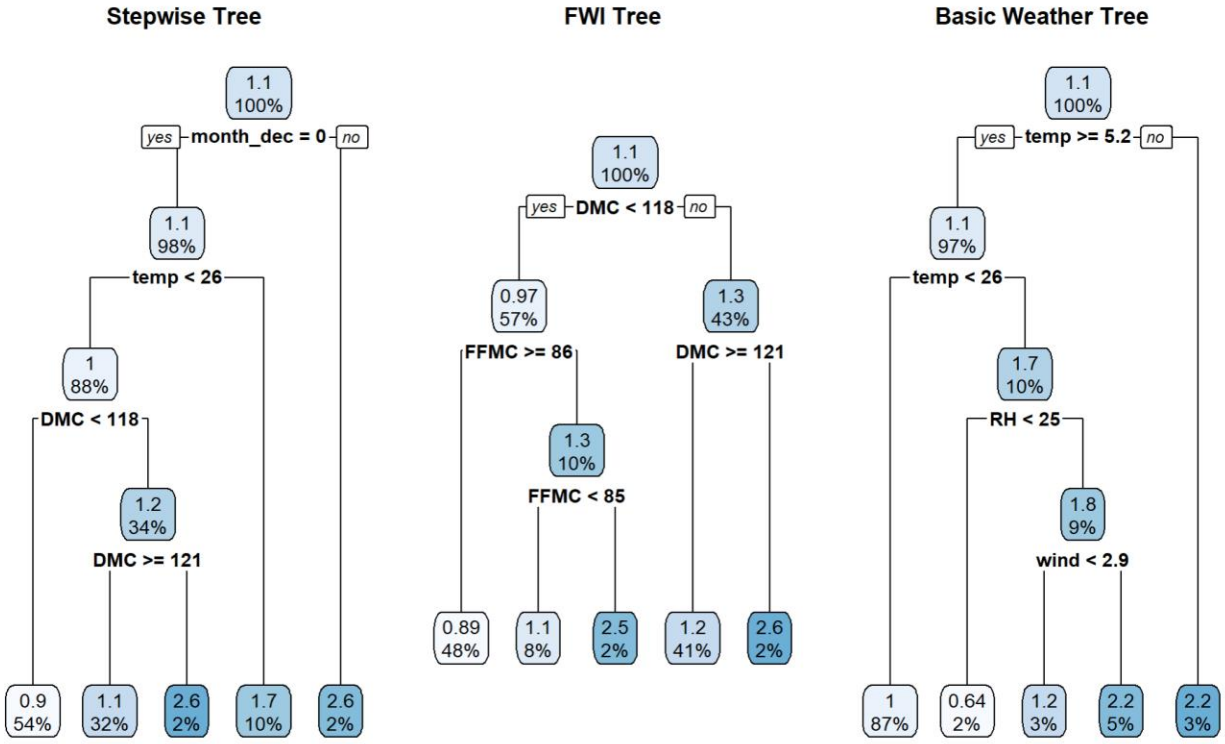


The log transformation helps marginally with the extreme right skewness exhibited in the pre transformed burned area plot, so we can only expect marginal improvements in the model performance present in our three methods.

We'll begin our analysis by exploring MLR and stepwise selection. First, we convert both the month and day variables to a one-hot coding structure. When dealing with categorical variables with multiple categories, we can assign a zero or one to each category depending on its presence in a given observation. This format is often easy to read for the machine learning algorithms, and thus why it's pursued. We then construct an MLR model and apply backward stepwise selection to extract the model with the lowest AIC amongst all available models. In doing so, the final model includes 5 subcategories of variable Month (August, December, July, June, March) as well as DMC, DC, Temperature, and Wind. The model produces an RMSE value of 1.84 and an AIC value of 335.97, a moderate decrease over the full model AIC of

361.12. Though the machine algorithm didn't choose variables in the context of the study, the model does include a mix of temporal and meteorological variables, and thus may be worth analyzing in decision tree models

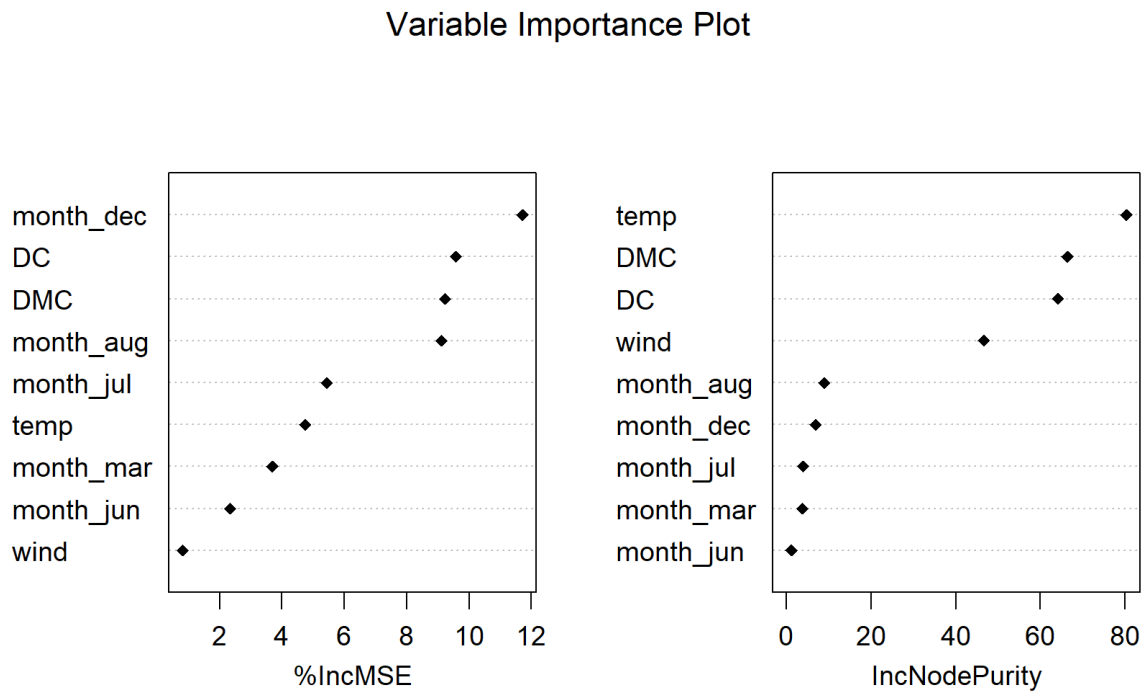
Next are the results of the burned area simulation of three different variable combinations using regression trees. The first regression tree takes all variables present in the final stepwise regression model, the second will take Fire Weather Index variables, and the third will include four basic weather variables. Below are the results of the three regression trees.



The regression trees visualize how specific classifications of predictors lead to specific results of the response. In the first tree, a fire occurring in the month of december holds particular importance, concluding fires in december cause larger burned area versus non december fires. Though both the first and third trees include the effects of temperature on burned

area, the third tree shows a particularly notable relationship where a temperature below 5.2 degrees celsius exhibits a relatively large burned area versus the other terminal nodes. As mentioned previously, the high variance associated with regression trees makes the models less statistically relevant, though providing visuals with interpretable outcomes is notable, especially for those outside of the realms of statistical practice.

Lastly, we visualize the results of our random forest model. The random forest model was created using the variables present in our final stepwise model. In creating this model, a bootstrap sample of 500 unpruned regression trees was included. Additionally, three different values for the tuning parameter 'mtry' were tested, which is the number of variables tried at each split. The tested values were 2, 3, and 4, with 2 variables at each split producing the highest variance explained score at -3.86% and the lowest RMSE score at 1.383. Below are the variable importance charts of our random forest model.



The month of December, DMC, and Temperature seem to be the most important variables in our dataset according to our random forest model. In regards to variable Month, it is worth noting the differences in climate between Portugal and Colorado. Portugal's December climate, with lows averaging 48°F and highs averaging 59°F (Walker, 2021), is more comparable to Colorado's September to November climate ("Denver Colorado weather & temperature info", 2023). Though our visuals give us a sense of which variables have the largest and smallest impact, our variance explained outcome plays a vital role in the interpretability of these charts. Given this value falls into the negatives, we can't make statistically significant conclusions on variable importance. That said, we can deduce that our RF model provides an improvement over the MLR model, with a test MSE of 1.383 versus 1.832.

Discussion

To conclude, we utilized MLR and RF to analyze the relationship between spatial, temporal, and meteorological variables and the burned area of a forest fire, and utilized regression trees to predict burned area given the classification of specific predictor variables.

Our first research objective tasked us with finding an optimal model to predict forest fires over our predictor variables. We developed an MLR model using backward stepwise selection to choose the most appropriate predictors based on predictive power. Though we produced a model that lowered our AIC value, it's difficult to describe why each variable was included in our model outside the realms of the machine algorithm process, which focuses solely on variable elimination that will cause the largest drop in AIC. We know that temporal, FWI, and basic weather variables were included in our model, but how can we describe the inclusion of these variables in the context of the study? It is especially difficult considering our lack of knowledge

on the science behind forest fires, but a basic interpretation may have been plausible if the inclusion of variables was more centric to their respective subgroups. For example, if our final model included all of the FWI variables, we may have concluded that FWI variables in general play a significant role in predicting forest fires, given that the model would produce a low RMSE or AIC value. Ultimately this was not the case, and we thus use MLR solely as a means of comparing models using the AIC.

The second model we implemented to optimize predictive performance was the random forest model. Our model produced variable importance charts that visualized the importance of variables including Temperature, DMC, and the month of December. That said, our model ultimately held very little significance from a statistical perspective given that its variance explained score was in the negatives. What this means is that we were better off predicting any given sample as the average of burned area over the entirety of the dataset, which indicates very poor performance. We can conduct further analysis with our random forest model by exploring the tuning parameters involved in the model construction, including number of trees, maximum depth of trees, node splitting, and node purity. That said, it's likely that the random forest model won't provide a statistically significant interpretation of our data, and exploring different methods is likely to be our best bet to find the optimal model for predicting forest fires.

References

Cortez, P., & Morais, A.D. (2007). *A data mining approach to predict forest fires using meteorological data.*

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed.). New York, NY: Springer*.

Hoare, J. (2022, September 13). *How is variable importance calculated for a random forest?*. Displayr.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). *An introduction to statistical learning: With applications in R*. Springer.

Kerlin, K. E. (2023, April 17). California's 2020 wildfire season. UC Davis.

Saxena, A. (2020, October 26). How transformation can remove skewness and increase accuracy of linear regression model. Medium.

Segall, J., & Bursa, F. (2023, August 29). Response transformation: When and how?. Quantics Biostatistics.

Walker, K. (2021, February 18). Portugal in December: Travel tips, weather, and more. kimkim.

Data Tricks. (2021, December 10). One-hot encoding in R: Three simple methods.

Denver colorado weather & temperature info. Visit Denver. (n.d.).

Finnstats. (2021, April 13). *Random Forest in R: R-bloggers*. R.

Regression trees. Regression Trees · UC Business Analytics R Programming Guide. (n.d.).

Why is it so hard for firefighters to put out wildfires?. Capstone. (2023, April 5).

Appendix

Response Variable Transformation

```
forestfires <- read_csv("forestfires.csv")
par(mfrow=c(1,2))
hist(forestfires$area, main = "Burned Area", col = "red", xlab = "Burned Area")
forestfires$area <- log(forestfires$area + 1)
hist(forestfires$area, main = "Log Tranformed Burned Area", col = "orange", xlab = "Burned
Area")
```

Descriptive Statistics

```
summary(forestfires)
```

Linear Regression

```
forestfires$month <- as.factor(forestfires$month)
forestfires$day <- as.factor(forestfires$day)
new.forest <- one_hot(as.data.table(forestfires))

m1 <- lm(area ~ . - X - Y, data = new.forest)
step(m1, direction = "backward")
m2 <- lm(area ~ month_aug + month_dec + month_jul + month_jun +
   month_mar + DMC + DC + temp + wind, data = new.forest)
mean(m1$residuals^2)
mean(m2$residuals^2)
```

Regression Trees

```
set.seed(123)

rt.1 <- rpart(formula = area ~ month_aug + month_dec + month_jul + month_jun +
   month_mar + DMC + DC + temp + wind, data = new.forest, method = "anova", control =
list(minsplit = 55, maxdepth = 13))
rt.2 <- rpart(formula = area ~ FFMC + DMC + DC + ISI, data = new.forest, method = "anova")
rt.3 <- rpart(formula = area ~ temp + RH + wind + rain, data = new.forest, method = "anova")

par(mfrow=c(1,3))
rpart.plot(rt.1, main = "Stepwise Tree")
rpart.plot(rt.2, main = "FWI Tree")
rpart.plot(rt.3, main = "Basic Weather Tree")
```

Random Forest

```
set.seed(123)

fire <- sample(2, nrow(new.forest), replace = TRUE, prob = c(0.7, 0.3))
train <- new.forest[fire==1,]
test <- new.forest[fire==2,]

rf <- randomForest(area ~ month_aug + month_dec + month_jul + month_jun +
    month_mar + DMC + DC + temp + wind, data = train, type = "regression", keep.forest =
FALSE, importance = TRUE, mtry = 2)
print(rf)
plot(rf)
sqrt(rf$mse[which.min(rf$mse)])
varImpPlot(rf, main = "Variable Importance Plot", pch = 18)
```