Forest Fires Initial Spread Index Prediction: A Training and Validation Study

C2 Team 4* Zhiquan Shen, Katherine Albrecht, Jiachen Chen, Jiahe Zhang

October 25, 2022

Abstract

This study uses a dataset from Montesinho Natural Park to develop a linear regresssion model for the prediction of Initial Spread Index (ISI) in order to improve forest fire management. The proposed model was generated using a training dataset and tested on a validation dataset to evaluate predictive accuracy. Our proposed model shows relatively good performance for moderate ISI values in the training dataset with acceptable generalization to the validation dataset, but tends to underestimate ISI at extremely high observed values. Further improvement is still needed to ensure predictive accuracy under the most dangerous conditions.

1 Introduction

Forest fires are serious environmental problems, which threaten both forest preservation and human life. For example, Portugal, a country seriously affected by forest fire, lost over 2.7 million hectare of forest area due to fire from 1980 to 2005. In 2003 and 2005, with dramatic fire seasons, 4.6 percent and 3.1 percent of the territory were affected and caused 39 total human deaths [Cortez and Morais, 2007]. Therefore, fire detection has become extremely crucial and challenging. Since fire occurrences are correlated to weather conditions such as temperature, humidity, and wind, collecting meteorological data seems necessary for better fire detection.

The Canadian forest Fire Weather Index (FWI) was developed to help rate fire danger. One of the six components included in FWI is Initial Spread Index (ISI), which relates to fire velocity spread. In this project, we used the given forest fire data from the Montesinho Natural Park to build a model that predicts ISI, with the aim of helping to detect scenarios with high risk of rapid fire spread. Ideally, these efforts could be used to guide more efficient approaches to forest fire management.

*Introduction & Background: Shen, Albrecht; Modeling: Albrecht, Chen; Analysis, diagnostic & validation: Albrecht, Chen, Zhang; Prediction & discussion: Chen, Albrecht

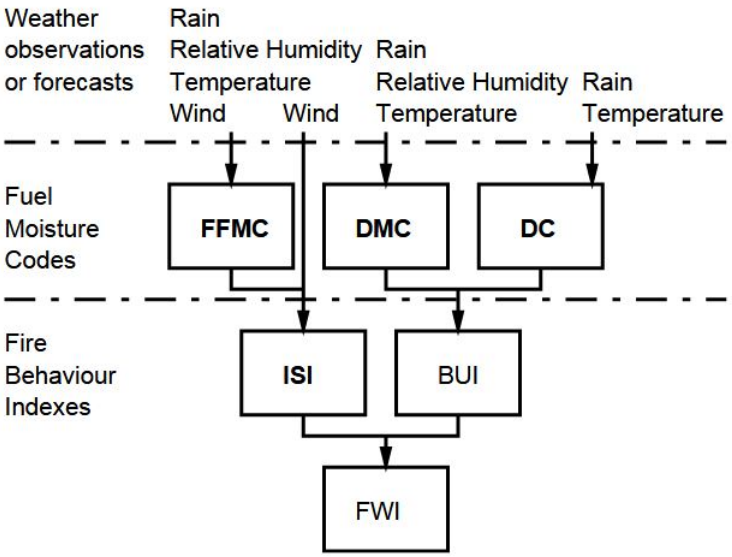

Figure 1: FWI flow chart

2 Background

Historically, meteorological data, including temperature, relative humidity, rain, and wind, have been used for fire prevention and management. In the 1970s, the Canadian forest Fire Weather Index (FWI) was designed to describe general fire intensity, which required only simple calculations from several meteorological observations. The forest Fire Weather Index (FWI) includes six components: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), and FWI. Our outcome variable ISI represents the fire velocity spread.

The forest fire data we are analyzing is derived from two databases. The first is collected by the inspector responsible for monitoring forest fires in Montesinho Natural Park. This includes details about each fire occurrence. The second consists of meteorological data collected by Bragança Polytechnic Institute. The data includes 517 observations collected from January 2000 to December 2003, each corresponding to a fire occurrence in the park. The variables recorded include the date (month and day of the week), spatial location, weather observations (temperature, wind, rain, and relative humidity), FWI components, and area burned. Figure 1 [Cortez and Morais, 2007] is a flow chart demonstrating the development of the FWI, which indicates that we should be able to predict ISI based on the included weather variables.

3 Modeling and Analysis

We aimed to build a linear model predicting ISI from the information included in the forest fires dataset. Usually, the estimate resulting from the model fit on the original training dataset will be biased. Hence, we utilized the validation set approach [James et al., 2013] by splitting the full dataset randomly into two halves. The first half constitutes the training set and the other half the validation set. We built our model based on the training set, then tested its performance by fitting it on the validation set.

3.1 Variable Selection

Because the dependent variable in this analysis was ISI, the FWI components shown in Figure 1 were not appropriate predictors for our model. The area variable was also not appropriate for inclusion, as a fundamental causal assumption of the FWI system is that the area burned in a fire is a consequence of the conditions reflected in the FWI component scores. Weather (relative humidity, temperature, wind, and rain) and time (month) variables were left remaining as potentially relevant predictors.

Based on a correlation plot using our full training dataset, we proceeded with the exploratory data analysis to observe the distribution and correlation of variables in the training set. In our initial model, we included the predictors wind, temp, and a categorical variable called summer, which was defined to be 1 for observations occurring in August and September and 0 otherwise. We identified one extreme outlier with a studentized residual of 13.93. Upon examination, this observation had an ISI of 56.1, yet the next highest ISI in the dataset was 22.7. Since ISI values greater than 15 are considered extreme [Alberta Ministry of Agriculture and Development, 2020], we believed that this observation was erroneous and would result in biased parameter estimates if included.

Our subsequent analyses considered the training dataset excluding this outlier. Figure 2 contains the scatterplots and correlation matrix of variables. Rain has an extremely skewed distribution with very few non-zero observations and no correlation with ISI, and thus was excluded from our model. Month is significantly correlated with ISI, but the relationship appears to be nonlinear, with the highest ISI values in August and September, most likely due to the cumulative effect of hot and dry weather over the course of the summer. The other variables are approximately close to normally distributed. The correlation coefficient between relative humidity (RH) and temperature (temp) is higher than between any other pair of variables. Since temperature shows a stronger correlation with ISI than RH, we chose to include temperature and exclude RH in order to avoid multicollinearity. Based on initial exploration of pairwise correlations formed from the full training set, we fitted models using the predictors wind, temp, and a categorical variable (summer) defined as described previously.

We used an iterative process to fit our model. We initially tried fitting two separate models: one for data from August and September and another for all other months. These showed poor performance in terms of non-constant variance and non-normality of standardized residuals, likely due to the small number of observations. Because these models were unlikely to generalize well to

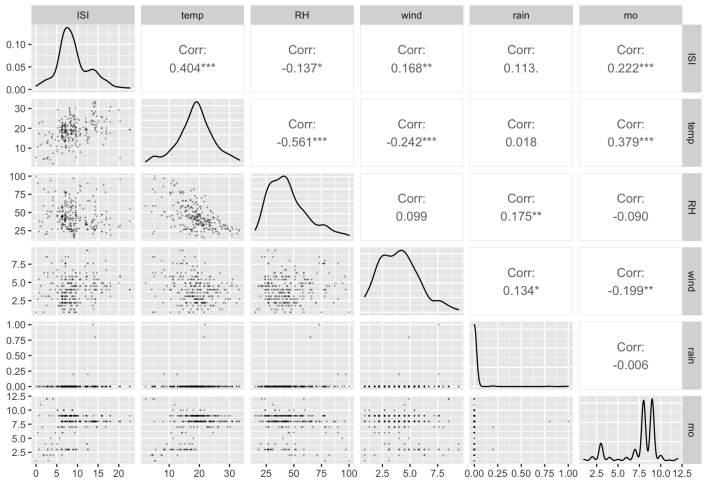

Figure 2: Scatterplots and correlations

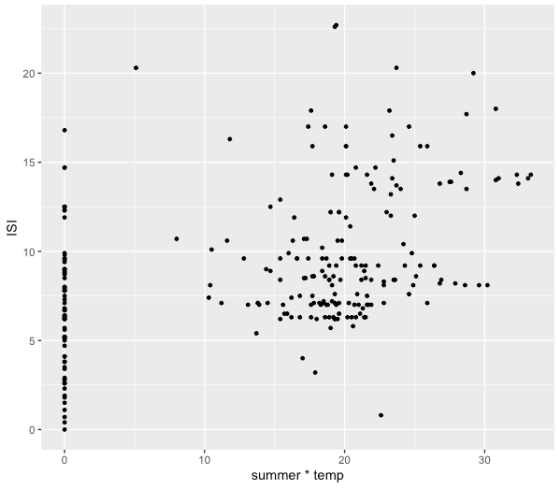

Figure 3: ISI by different season*temperature

the validation set, we chose to use a model based on the full training data excluding the previously-mentioned outlier.

We added an interaction term between the summer and temperature variables to allow for seasonal differences in the relationship between temperature and ISI. Based on a scatterplot of summer*temp vs. ISI (shown in Figure 3), we observed a positive relationship between temperature and ISI during the summer months which appeared to be quadratic. Our final model included wind, linear and quadratic temperature, season, and interaction between season and linear and quadratic temperature.

3.2 Results

Using the method of least squares, our model can be expressed as:

$$ISI = \beta_0 + \beta_1 * summer + \beta_2 * temp + \beta_3 * wind + \beta_4 * I(temp^2) + \beta_5 * I(summer * temp) + \beta_6 * I(summer * temp^2)$$

We fit the model in R. The results are displayed in Table 1.

The R^2 is 0.3364, which implies that our model explains slightly less than 34% of the observed variance in ISI. While this estimate tends to be inflated in multiple linear regression, the adjusted R^2 value is quite similar at 0.3204, suggesting that there is not a major problem with irrelevant variables inflating R^2. With an F-statistic of 21.12 on 6 numerator and 250 denominator degrees of freedom, we reject the global null hypothesis and conclude that the model performs better than an intercept-only model.

Predictor	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.515308	1.946176	-1.292	0.197399
summer	14.403554	3.507131	4.107	5.44e-05 ***
temp	0.623960	0.245916	2.537	0.011780 *
wind	0.670241	0.121738	5.506	9.11e-08 ***
I(temp^2)	-0.011853	0.007718	-1.536	0.125843
I(summer*temp)	-1.311434	0.377229	-3.476	0.000599 ***
I(summer*temp^2)	0.033897	0.010240	3.310	0.001069 **
Residual standard error: 3.338 on 250 degrees of freedom				
Multiple R-squared: 0.3364, Adjusted R-squared: 0.3204				
F-statistic: 21.12 on 6 and 250 DF, p-value: $<$ 2.2e-16				

Table 1: R output for the summary of the proposed linear model

At a significance level of 0.05, the variables for temperature, wind, summer (defined as month = August or September), and interaction variables summer*temp and summer*temp^2 are significant predictors of the outcome ISI.

In terms of estimated effects of the predictors on mean ISI, for a one km/hr increase in wind, we would expect mean ISI to increase by approximately 0.67 (95% CI: [0.43, 0.91]). Interpretation of the effect of season and temperature is more complex, due to the presence of a significant interaction. At a fixed temperature, we would expect the mean ISI in August or September compared to any other month to be higher by approximately 14.4 (95% CI: [7.50, 21.31]). Outside the months of August and September, we would expect a linear increase in ISI with temperature, with an increase in mean ISI of approximately 0.62 (95% CI: [0.14, 1.11]) per degree Celsius increase in temperature. In contrast, the significance of the interaction terms indicates that during August and September, we would expect the trend in mean ISI with temperature to follow the quadratic equation $ISI = -0.687474 * temp + 0.033897 * temp^2$.

Overall, our estimated model can be represented by the equation:

$$\begin{aligned} ISI = &- 2.515308 + 0.670241 * wind + 0.623960 * temp + 14.403554 * summer \\ &- 1.311434 * I(summer * temp) + 0.033897 * I(summer * temp^2). \end{aligned}$$

3.3 Diagnostic analysis and evaluation

We used a Q-Q plot of standardized residuals to check the normality assumption. According to this plot in Figure 4, the predicted values conform tightly to the line in the lower end. There is a pronounced upward deviation starting at 1 standard deviation above the mean, indicating high values of ISI are over-represented in the data relative to theoretical expectations for a normal distribution. Thus we conclude that there is some violation of the normality assumption in this model.

We used the standardized residuals plot to check the assumption of equal variance. There does not seem to be any extreme violation of the equal variance assumption. However, there is a slight

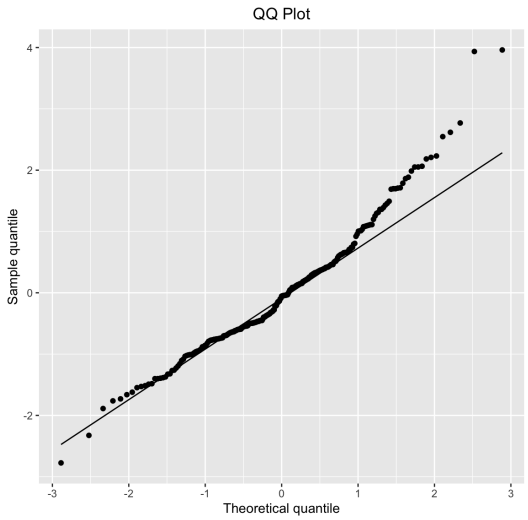

Figure 4: Q-Q plot

spread of variance as shown in Figure 5, with standardized residuals for fitted values of ISI between 8 and 12 showing higher variance than standardized residuals at lower fitted values. Thirteen standardized residuals (5%) have an absolute value greater than 2, with 11 of these reflecting underestimated ISI. Similarly, there is also a trend shown in Figure 6 that observations with higher actual ISI have larger standardized residuals. These patterns are consistent with the observed over-representation of high values seen in the Q-Q plot.

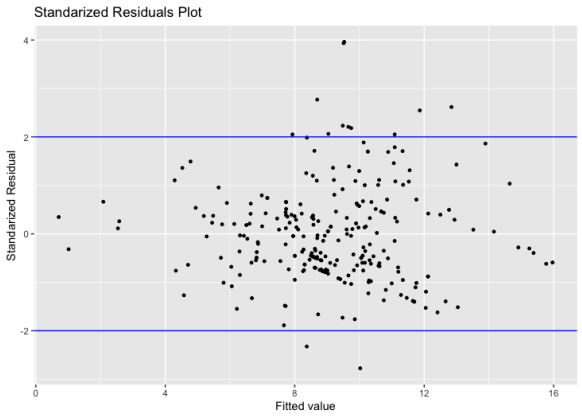

Figure 5: Standardized residuals versus fitted values

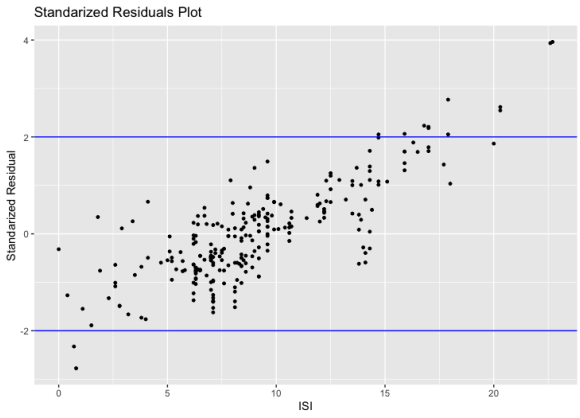

Figure 6: Standardized residuals versus ISI

Based on the deviation from normality in the Q-Q plot, we are not fully confident with the p-values observed in our model. With respect to the validity of the hypothesis tests, we believe that there is a true statistically significant association between wind, temperature, season and ISI

based on the scatterplot patterns. This is also consistent with commonsense expectations of the effect of weather and seasonal conditions on propensity for fires to spread. However, based on the diagnostics, the magnitude of the effect of these factors may be exaggerated.

4 Prediction

Based on the regression model in Section 3 and the covariates in the validation data, we generated the predicted ISI for the validation dataset.

The MSE of ISI for validation data was 12.94, compared to 10.84 for training data, which is not an extreme difference. The relative mean squared error for the validation data was 0.1385, indicating our model performs much better at predicting ISI in the validation set than a model using only the mean ISI value.

We conclude that the model obtained from the training data has acceptable generalizability to the validation data. Figure 7 shows the plot of the actual ISI versus the predicted ISI on the validation set. The range of predicted values is considerably narrower than the range of observed values, resulting in considerable horizontal dispersion around the 45 degree line. ISI tends to be systematically underestimated at higher values and overestimated at lower values, with reasonable predictive accuracy at moderate values. This pattern is also observed in Figure 8, which presents the comparisons between predicted and actual ISI by index. Based on this figure and the zoomed in plot shown in Figure 9, our model generally succeeds at capturing trends in ISI. However, it underestimates the magnitude of increases in ISI at the extreme values that are most critically important for fire management.

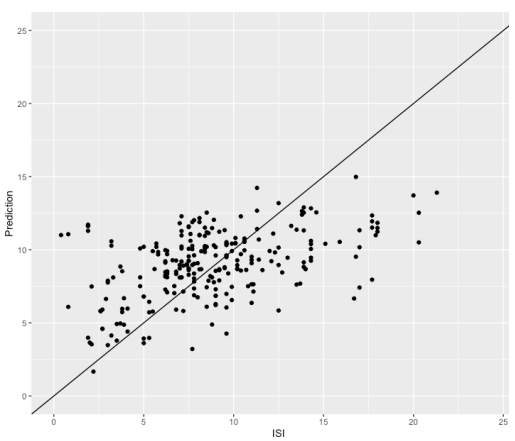

Figure 7: Model fit for ISI on the validation dataset

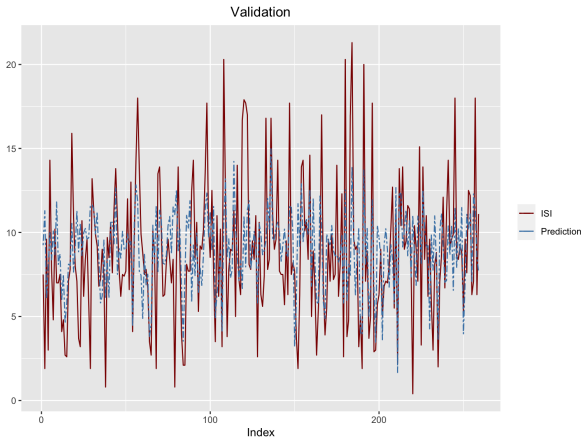

Figure 8: Predicted vs. actual ISI in the validation dataset

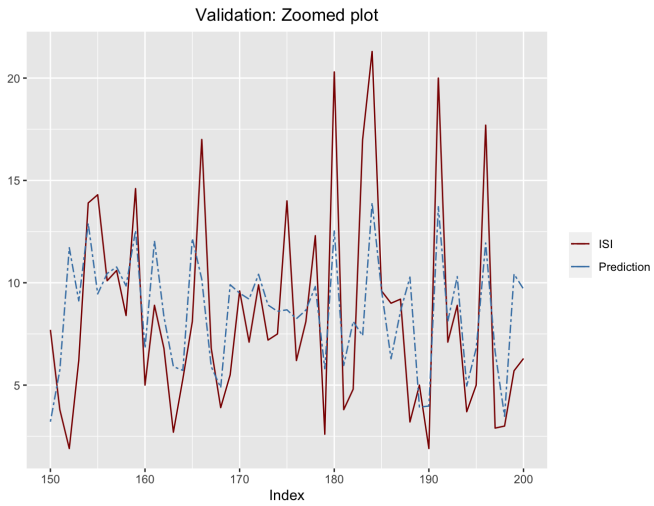

Figure 9: Predicted vs. actual ISI (zoomed in)

5 Discussion

The regression model for ISI prediction generated from the training data has relatively good performance when ISI is moderate, but shows limitations in predicting high ISI in both the training and validation datasets. The goal of the study is to sensitively detect scenarios with extremely high ISI that might lead to rapid forest fire spread. Unfortunately, when we apply our model to predict the ISI of the validation data, our model underestimates the most critical observations with extreme values of ISI. In order for our model to contribute efficiently to the control of fires and resources planning, we must improve its performance in identifying days with higher ISI.

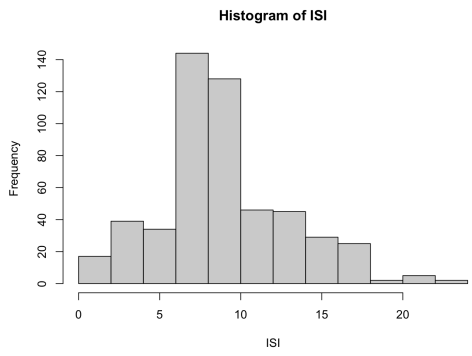

Figure 10: Histogram of ISI

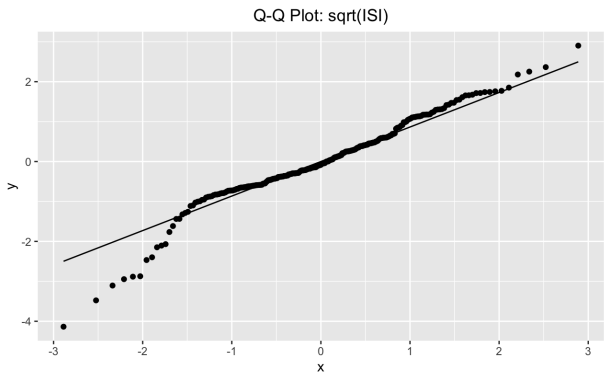

Figure 11: Q-Q plot of model using sqrt(ISI)

Certain features of this dataset present challenges to model-building. One difficulty arises from

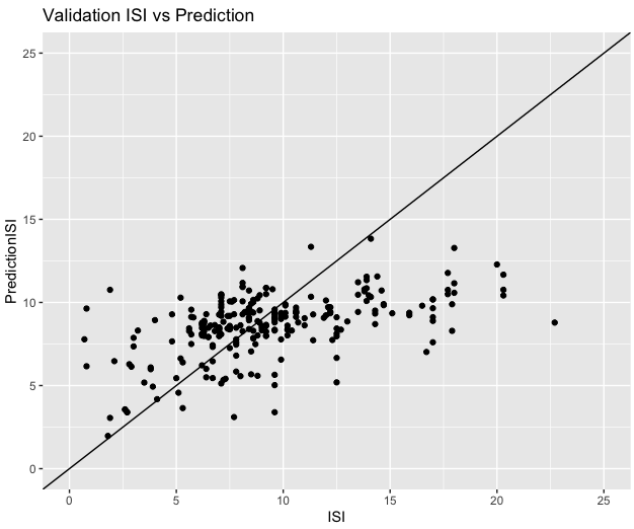

Figure 12: Model fit for ISI transforming sqrt(ISI) back

the distribution of ISI. Figure 10 shows a histogram of ISI from the entire forest fires dataset excluding the previously-mentioned outlier. We can see that ISI is not normally distributed. Using new training and validation datasets, we explored the possibility of fitting a model with the square root of ISI as the dependent variable. While the Q-Q plot (Figure 11) did appear better for high ISI values (at the cost of underestimating lower observations), when transforming the predictions back, this approach did not perform better than our original model at predicting high ISI observations in the training dataset (Figure 12). Therefore, for simplicity and ease of interpretation, we prefer our original model.

Another potentially even more serious limitation is the fact that observations in this dataset are both temporally and spatially correlated, violating the assumption of independence. If we examine all observations where ISI is at least 20 (Figure 13), it appears based on similarity of FWI components that many of these observations take place on the same day. While ISI is only calculated once per day, instantaneous temperature and windspeed measurements can vary widely over the course of a day. For example, on the day with ISI equal to 20.3, temperature ranged from 5.1 to 23.7 degrees Celsius. The high variability in wind and temperature for the same ISI observation is expected to weaken the association between these variables and ISI, ultimately biasing the model toward moderate predicted ISI values.

It may be theoretically possible to achieve better predictive performance with a time series model. Unfortunately, the only temporal data included in this dataset are month and day of the week. We explored the possibility of inferring a more precise temporal structure by 1) assuming observations with similar FWI component values in each month occurred in the same year and 2) ordering data within the same month based on the day of week and similarity of FWI component values. Unfortunately, due to the sparse representation of data in winter and spring months combined with high seasonal variability in FWI component values, we were unable to identify a clear

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
12	7	5	sep	sat	92.8	73.2	713.0	22.6	19.3	38	4.0	0	0.00
25	7	4	aug	sat	93.5	139.4	594.2	20.3	23.7	32	5.8	0	0.00
136	3	5	aug	sat	93.5	139.4	594.2	20.3	17.6	52	5.8	0	0.00
207	2	2	aug	sat	93.5	139.4	594.2	20.3	22.9	31	7.2	0	15.45
212	7	4	aug	sat	93.5	139.4	594.2	20.3	5.1	96	5.8	0	26.00
267	6	5	aug	tue	94.3	131.7	607.1	22.7	19.4	55	4.0	0	0.17
486	2	4	aug	mon	95.0	135.5	596.3	21.3	30.6	28	3.6	0	2.07
504	2	4	aug	wed	94.5	139.4	689.1	20.0	29.2	30	4.9	0	1.95
505	4	3	aug	wed	94.5	139.4	689.1	20.0	28.9	29	4.9	0	49.59

Figure 13: ISI $\geq$ 20

sequence for the dataset as a whole. This prevented us from fitting an autoregressive model.

Given these limitations, a more productive approach may be to consider each unique combination of the month, day of week, FFMC, DMC, DC, and ISI variable values as a distinct day, then calculate daily summary statistics for weather variables (e.g. maximum temperature, mean windspeed) to be used as predictors of ISI. This model would still violate the assumption of independence, but would address the problems arising from same-day weather variability. We did not apply this approach because it entails a major transformation of the dataset, but we believe it warrants further exploration.

Appendix

```
# Create training and validation datasets
 fires <- fires[sample(nrow(fires)),]
 firestrain <- fires[1:258,]
 firesval <- fires[259:517,]
 write.csv(firestrain, "firestrain.csv")
 write.csv(firesval, "firesval.csv")

# Create scatterplot
attach(firestrain)
data <- data.frame(ISI, temp, RH, wind, rain, mo)
ggpairs(data, lower = list(continuous = wrap("points",
        alpha = 0.3, size=0.1)))

# Initial model (all data)
lm1 <- lm(ISI ~ temp+wind+as.factor(month %in% c("aug","sep")))
summary(lm1)
which(abs(studres(lm1)) > 3)
studres(lm1)[20]
hist(ISI)
# remove outlier
firetrain_use=firestrain[-20,]



# define categorical variable
firetrain_use$summer=as.numeric(firetrain_use$month %in%
                                  c("aug","sep"))

# Scatterplot of summer*temp interaction
ggplot() +
geom_point(data=firetrain_use,
           aes(x=summer*(temp), y=ISI, color = "MLS"), size = 1)

# Multiple linear regression model
m.mls <- lm(ISI~wind+summer*temp+I(summer*temp^2),
            data=firetrain_use)
summary(m.mls)

Call:
lm(formula = ISI ~ summer * temp + wind + summer * I(temp^2),
    data = firetrain_use)
```

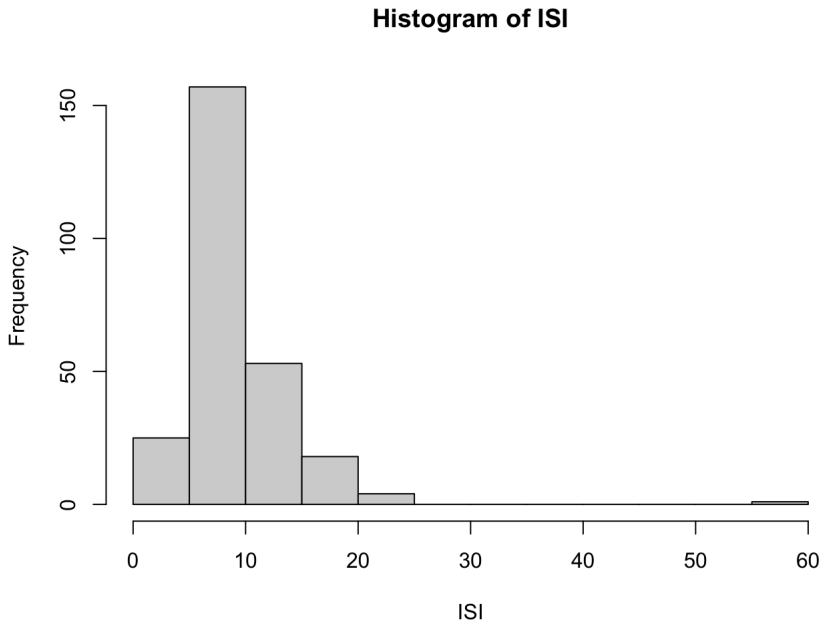

Figure 14: Histogram of ISI with outlier

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.2232 -2.1439 -0.1784  1.5286 13.1714

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.515308   1.946176  -1.292 0.197399
summer           14.403554   3.507131   4.107 5.44e-05 ***
temp              0.623960   0.245916   2.537 0.011780 *
wind              0.670241   0.121738   5.506 9.11e-08 ***
I(temp^2)        -0.011853   0.007718  -1.536 0.125843
summer:temp      -1.311434   0.377229  -3.476 0.000599 ***
summer:I(temp^2)  0.033897   0.010240   3.310 0.001069 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


# Standard Residuals vs ISI
StanResMLS <- rstandard(m.mls)
```

```
dataMLS <- data.frame(ISI=firetrain_use$ISI,StanResMLS)


ggplot() +
  geom_point(data=dataMLS, aes(x=ISI, y=StanResMLS), size = 1) +
  geom_hline(yintercept=2,color='blue') + geom_hline(yintercept=-2, color='blue') +
  scale_color_manual(name = element_blank(), labels = c("MLS"), values = c("black")) +
  labs(y = "Standarized Residual") + ggtitle("Standarized Residuals Plot")

# Standardized Residuals vs Fitted
Fitted = fitted(m.mls)
dataMLSFitted <- data.frame(Fitted,StanResMLS)

ggplot() +
  geom_point(data=dataMLSFitted,
             aes(x=Fitted, y=StanResMLS, color = "MLS"), size = 1) +
  geom_hline(yintercept=2,color='blue') +
  geom_hline(yintercept=-2, color='blue') +
  scale_color_manual(name = element_blank(), labels = c("MLS"),
                     values = c("black")) +
  labs(y = "Standardized Residual") + labs(x = "Fitted value") +
  ggtitle("Standardized Residuals Plot")

# Test of Normality for Standardized Residuals of MLS
p <- ggplot(data.frame(StanResMLS), aes(sample = StanResMLS)) +
     ggtitle("QQ Plot")
p + stat_qq() + stat_qq_line()


# Validation
# Residuals for training data
ResMLS <- resid(m.mls)

output<-predict(m.mls, se.fit = TRUE,
                newdata=data.frame(temp=firesvalidate$temp,
                                   wind=firesvalidate$wind,
                                   summer=firesvalidate$summer))
ResMLSValidation <- firesvalidate$ISI - output$fit


# MSE for training data
mean((ResMLS)^2)
```

```
# MSE for validation data
mean((ResMLSValidation)^2)

# Relative Mean Square Error for validation data
mean((ResMLSValidation)^2)/mean((firesvalidate$ISI)^2)

# Model fit vs. ISI plot
ggplot(data = test, aes(x = ISI, y = Prediction)) + geom_point() +
  geom_abline(intercept = 0, slope = 1)

# Plot ISI vs Prediction for Validation Data Set
ggplot(data = test, aes(x = Index)) +
  geom_line(aes(y = ISI, color = "ISI")) +
  geom_line(aes(y = Prediction, color="Prediction"), linetype="twodash") +
  scale_color_manual(name = element_blank(), labels = c("ISI","Prediction"),
                     values = c("darkred", "steelblue")) + labs(y = "") +
  ggtitle("Validation")

# Plot Zoomed ISI vs Prediction for Validation Data Set
zoomset = test[150:200,]
ggplot(data = zoomset, aes(x = Index)) +
  geom_line(aes(y = ISI, color = "ISI")) +
  geom_line(aes(y = Prediction, color="Prediction"), linetype="twodash") +
  scale_color_manual(name = element_blank(), labels = c("ISI","Prediction"),
                     values = c("darkred", "steelblue")) + labs(y = "") +
  ggtitle("Validation: Zoomed plot")

# Square root model
mlr2 <- lm(data=firestrain2, sqrt(ISI) ~ temp*summer+I(temp^2)*summer+wind)
summary(mlr2)

Call:
lm(formula = sqrt(ISI) ~ temp * summer + I(temp^2) * summer +
   wind, data = firestrain2)

Residuals:
      Min       1Q     Median    3Q     Max
   -2.48835 -0.35900 -0.04806 0.35377 1.79142

Coefficients:
                 Estimate   Std. Error   t value Pr(>|t|)
```

```
(Intercept)      0.818011   0.345729     2.366   0.018739 *
temp             0.122574   0.044864     2.732   0.006740 **
summer           2.658478   0.754217     3.525   0.000504 ***
I(temp^2)       -0.001667   0.001433    -1.163   0.245807
wind             0.063633   0.022446     2.835   0.004956 **
temp:summer     -0.222727   0.078210    -2.848   0.004767 **
summer:I(temp^2) 0.004793   0.002057     2.331   0.020559 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6196 on 251 degrees of freedom
Multiple R-squared: 0.3221, Adjusted R-squared: 0.3059
F-statistic: 19.88 on 6 and 251 DF, p-value: < 2.2e-16

stdres2 <- stdres(mlr2)
fit2 <- fitted(mlr2)
MLS2data <- data.frame(stdres2, fit2)

# Test of Normality using Q-Q plot
p <- ggplot(data.frame(MLS2data), aes(sample = stdres2))
p + stat_qq() + stat_qq_line()+ ggtitle("Q-Q Plot: sqrt(ISI)")

# Prediction for validation data
output<-predict(mlr2, se.fit = TRUE, newdata=data.frame(temp=firesvalidate$temp,
wind=firesvalidate$wind,summer=firesvalidate$summer))
ResMLSQValidation <- firesvalidate$sqrtISI - output$fit

# Create data frame with validation observation and prediction
test = data.frame(firesvalidate$sqrtISI,firesvalidate$ISI,(output$fit)^2,output$fit,
1:length(output$fit));
colnames(test)[1] = "sqrtISI"
colnames(test)[2] = "ISI"
colnames(test)[3] = "PredictionISI"
colnames(test)[4] = "PredictionsqrtISI"
colnames(test)[5] = "Index"

# Plot ISI vs Prediction for Validation Data Set
ggplot(data = test, aes(x = ISI, y = PredictionISI)) + geom_point() + geom_abline
(intercept = 0, slope = 1) + xlim(0, 25)+ ylim(0,25) + ggtitle("Validation ISI vs
Prediction")
```

References

[Alberta Ministry of Agriculture and Development, 2020] Alberta Ministry of Agriculture, F. and Development, R. E. (2020). Understanding fire weather.

[Cortez and Morais, 2007] Cortez, P. and Morais, A. d. J. R. (2007). A data mining approach to predict forest fires using meteorological data.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.