

NORTHWEST NAZARENE UNIVERSITY

Statistical Analysis of Forest Burn Extent Data

THESIS


Submitted to the Department of Mathematics and Computer Science
in partial fulfillment of the requirements
for the degree of
BACHELOR OF SCIENCE

Tyler J. Shea
2022


THESIS
Submitted to the Department of Mathematics and Computer Science
in partial fulfillment of the requirements
for the degree of
BACHELOR OF SCIENCE

Tyler J. Shea
2022


Statistical Analysis of Forest Burn Extent Data

Author: 

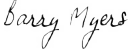
Tyler J. Shea

Approved: 

Dale A. Hamilton, Ph.D., Associate Professor, Department of
Mathematics and Computer Science, Faculty Advisor

Approved: 

Stephen Riley, Ph.D., Associate Professor, Department of Theology,
Second Reader

Approved: 

Barry L. Myers, Ph.D., Chair,
Department of Mathematics & Computer Science

ABSTRACT

Statistical Analysis of Forest Burn Extent Data

SHEA, TYLER (Department of Mathematics and Computer Science),

HAMILTON, DR. DALE (Department of Mathematics and Computer Science)

Various machine learning algorithms have been shown to be effective methods of mapping forest fire burn extent and tree mortality. The algorithms use drone imagery to classify pixels as burned or unburned. Recent efforts used a mask region-based convolutional neural network (MR-CNN) and support vector machine (SVM) to label pixels in a post-fire forest as being within the fire's extent. These algorithms reclassified the pixels using the Unburned Tree Noise and Sub-Crown Burn Reclassifications. The objective was that these reclassifications would produce more accurate results than the previously computed Surface Burn Classification. The purpose of this project was to use analysis methods to determine statistical significance in the results, and decide whether the reclassifications gave significantly better results than the original classification. The primary tools used in the analysis were the one- and two-tailed paired Student's t-tests. These tests were conducted on the sensitivity results given by the algorithms, because sensitivity was considered the metric of most importance due to precedence being placed on minimizing the false negative percentage. Results calculated from the t-tests demonstrated that the new reclassifications produced a statistically significant increase in sensitivity over relying solely on the Surface Burn Classification for burn extent mapping.

Acknowledgements

First of all, I would like to thank my family for their support and encouragement throughout my summer of research. I also want to thank Dr. Hamilton for allowing me to join his FireMAP team on late notice and be a co-author on his published paper, as well as Dr. Myers for his support along the way. Finally, I would like to thank all the students who worked on FireMAP, specifically Kamden Brothers, Cole McCall, and Bryn Gautier for all their hard work on this project and the corresponding paper.

Table of Contents

ABSTRACT	iii
Acknowledgements	iv
Table of Contents	v
List of Figures and Tables	vi
Overview	1
Background	1
Objective	4
Methods	4
Results	9
Future Work	10
Conclusion	11
References	12
Appendix A: Code	13
a. Two-tailed paired Student's t-test Code.....	13
b. One-tailed paired Student's t-test Code	13

List of Figures and Tables

Figure 1. Sample orthomosaic imagery from the Hoodoo fire (Hamilton et al., 2021)	2
Figure 2. Surface Burn Classification sample imagery (Hamilton et al., 2021)	3
Figure 3. Bar graph modeling the percent increase in sensitivity	5
Figure 4. Line graph modeling the percent increase in sensitivity.....	5
Figure 5. Two-tailed Student's t-test graph	8
Figure 6. One-tailed Student's t-test graph.....	8
Figure 7. Results of the two-tailed paired Student's t-test.....	10
Figure 8. Results of the one-tailed paired Student's t-test.....	10
Table 1. Results from SVM and 5600 threshold	4
Table 2. Percent increase in sensitivity for each fire.....	5

Overview

This project was a study of the improvement of forest fire burn extent mapping through the use of two machine learning algorithms: mask region-based convolutional neural networks (MR-CNN) and support vector machines (SVM). These two models were used to classify the pixels from hyperspatial imagery of a forest as burned or unburned, forming what was called the Unburned Tree Noise and Sub-Crown Burn Reclassifications. Various metrics, namely sensitivity and accuracy, were used to compare the results from the reclassifications to the results from the previously used Surface Burn Classification, which classified pixels using only the SVM with no additional tools. The primary goal of this specific effort was to contribute to the published paper *Mapping Forest Burn Extent from Hyperspatial Imagery Using Machine Learning* (Hamilton et al., 2021) by investigating if there was a statistically significant improvement in the results of the mapping through the new reclassifications compared to the original classification.

Background

This project was a continuation of the Fire Monitoring and Assessment Platform (FireMAP) research at Northwest Nazarene University in Nampa, Idaho. The goal of FireMAP is to gather data on wildfire burn severity and extent by using small, unmanned aerial vehicles (sUAV) to capture hyperspatial imagery of the affected area. Throughout the evolution of this project, various machine learning models have been used to analyze the sUAV imagery and classify each pixel as burned or unburned.

The imagery data was collected from four different locations in southwestern Idaho that had recently been hit by wildfire. This included three fires, the Cottonwood,

Hoodoo, and Corner fires, that were located in the Boise National Forest, as well as one fire, the Mesa fire, that was located in the Payette National Forest (Hamilton et al., 2021). The data captured was “hyperspatial (5cm) resolution imagery” taken from a sUAS flying at an altitude of 120 meters above ground level (Hamilton et al., 2021). This imagery created post-fire orthomosaics of the burned areas that were then used for analysis via machine learning algorithms.

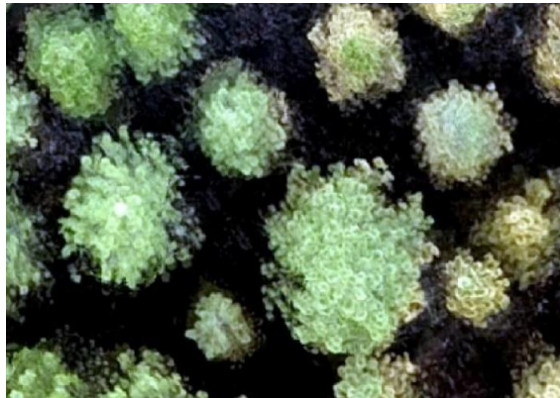


Figure 1. Sample orthomosaic imagery from the Hoodoo fire (Hamilton et al., 2021)

The first machine learning model that was used to analyze the imagery was the Surface Burn Classification, which used the SVM by itself without any other additional tools. The effectiveness of this classification was measured through three primary metrics: accuracy, specificity, and sensitivity. Accuracy measures the total percentage of correctly labeled pixels, specificity deals with the percentage of correctly labeled negative pixels, and sensitivity computes the percentage of correctly labeled positive pixels. The formal equations for accuracy, specificity, and sensitivity are given in Equations (1)-(3) respectively.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (1)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

The Surface Burn Classification yielded averages of 77.6% accuracy, 95.3% specificity, and 59.4% sensitivity for the four fires. The noticeably low sensitivity percentage was caused by the classifier incorrectly labeling pixels as unburned in cases where an unburned tree crown was surrounded by burned surface vegetation. This mislabeling created a high number of false negatives, thus affecting the sensitivity metric. An example of orthomosaic imagery classified using the Surface Burn Classification is shown in Figure 2. The green shapes represent unburned tree crowns and the black area represents burned surface vegetation.

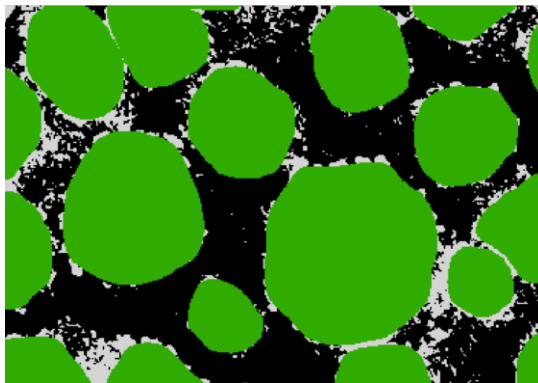


Figure 2. Surface Burn Classification sample imagery (Hamilton et al., 2021)

In hopes of increasing the metrics, two new algorithms were combined to form the Unburned Tree Noise and Sub-Crown Burn Reclassifications. These algorithms combined the SVM and MR-CNN, and used a specific pixel cluster threshold of 5600 pixels when removing noise pixels (Hamilton et al., 2021). They gave results of 86.7% accuracy, 94.6% specificity, and 77.7% sensitivity on average.

Objective

The primary objective of this project was to determine if the results from the Unburned Tree Noise and Sub-Crown Burn Reclassifications indicated an increase in effectiveness that was statistically significant when compared to the original Surface Burn Classification.

Methods

The first decision to be made was to determine which metric should be used to test for statistical significance. The main purpose of the Unburned Tree Noise and Sub-Crown Burn Reclassifications was to lower the number of false negatives. This is due to the fact that in the original classifications, the average sensitivity amongst the four fires was only 59.4%. As previously mentioned, sensitivity is directly related to the false negative percentage. Therefore, it was decided that sensitivity should be the primary metric for analysis rather than accuracy or specificity, as it would give the best representation of the effectiveness of the new reclassifications.

Fire	Threshold	Accuracy	Specificity	Sensitivity
Hoodoo	SVM	81.85%	91.41%	66.68%
Hoodoo	5600	99.45%	99.35%	99.54%
Cottonwood	SVM	84.59%	96.15%	71.28%
Cottonwood	5600	92.65%	95.83%	88.93%
Mesa	SVM	78.95%	89.48%	70.34%
Mesa	5600	84.58%	87.63%	82.09%
Corner	SVM	65.01%	95.99%	29.19%
Corner	5600	69.99%	95.77%	40.19%

Table 1. Results from SVM and 5600 threshold

In order to calculate the most accurate measure of improvement given by the Unburned Tree Noise and Sub-Crown Burn Reclassifications, it was important to compare each fire individually, rather than just comparing the averages. For example, as seen in Table 1, the sensitivity from the original classification of the Cottonwood fire was 71.28%, while the Corner fire was only 29.19%. Therefore, it would not be constructive to just look at the average sensitivity, but rather it was necessary to look at the percentage increase of sensitivity in each individual fire, seen in Figure 3, to get a better picture of the effectiveness of the reclassifications.

Sensitivity % Increase	
Hoodoo	32.86%
Cottonwood	17.65%
Mesa	11.75%
Corner	11.00%

Table 2. Percent increase in sensitivity for each fire

It was also important to model the data in order to visualize what the increase looks like. These models were created in Python’s Matplotlib library using the data given in Tables 1 and 2. Two graphical representations of the percent increase in sensitivity are depicted in Figures 3 and 4.

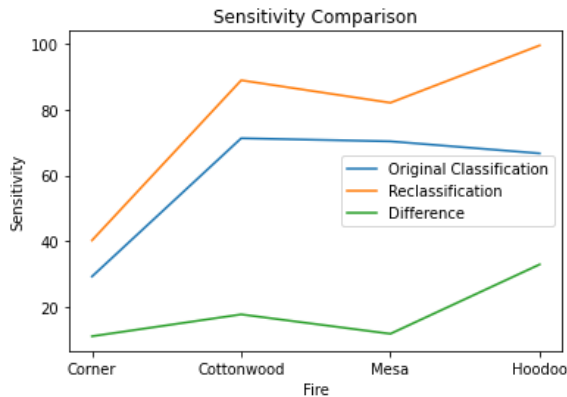


Figure 4. Line graph modeling the percent increase in sensitivity

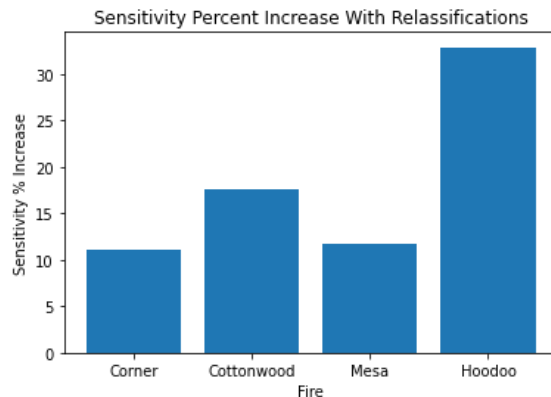


Figure 3. Bar graph modeling the percent increase in sensitivity

After the data was extracted and modeled, the next step was to determine how to test the data for statistical significance. Statistical significance is defined as the “claim that a result from data generated by testing or experimentation is not likely to occur randomly or by chance, but is instead likely to be attributable to a specific cause” (Wilkerson, 2008). In other words, statistical analysis is the process of testing to see if the data output is significantly different than the mean or the standard. This is important because it validates the results and proves that they did not happen by random chance. In this specific case, the goal of significance testing is to determine if the improvements gained in the sensitivity data by the new reclassifications are great enough to be declared significant, or are just attributable to random chance.

There are many different statistical hypothesis tests that could have been used to determine significance. However, after careful research and consideration, it was decided that the paired Student’s t-test was the most applicable test for this project. There are two types of Student’s t-tests: the standard t-test and the paired t-test. A standard Student’s t-test compares two independent data samples, while the paired Student’s t-test compares the means of two related data samples (Brownlee, 2018a). As mentioned earlier, for this project it was necessary to compare the sensitivity results from each fire individually. Therefore, the standard t-test cannot be used because the data samples are paired together. The paired t-test corrects for this fact to form a modified version of the standard t-test, making it applicable to this project.

Many other statistical analysis tests were also researched and considered before the Student’s t-test was ultimately chosen. One of the potential options was the Analysis of Variance Test (ANOVA). This test acts similarly to the Student’s t-test because it

calculates if two or more data samples are significantly different from each other (Brownlee, 2018a). The ANOVA test is generally used when there are more than two data samples as a more convenient way to compare more data at once. The Repeated Measures ANOVA test acts like the paired Student's t-test because it can deal with data samples that are related or dependent (Brownlee, 2018a). This test was strongly considered as the potential primary test in this project, but one major drawback was that Python's SciPy library does not currently contain a Repeated Measures ANOVA package. Because this research only involves the comparison of two data samples, the Repeated Measures test was not necessary, as the paired Student's t-test would produce the same results.

Another analysis test that was looked into was McNemar's Test. This test is primarily used to analyze complex machine learning models, specifically deep learning models (Brownlee, 2018c). However, it was not seriously considered for this project because its main function is to compare two machine learning models on the same data set, which was not included in the scope of this research (Brownlee, 2018c). The final test that was briefly examined was Pearson's Chi-Squared Test, which observes whether a data sample follows an expected distribution (Brownlee, 2018b). This test was not seriously considered either because there is no expected result for the sensitivity data. Overall, despite there being viable alternatives, the paired Student's t-test was clearly the correct statistical test for this project.

The paired Student's t-test begins with an assumption, or null hypothesis (H_0), that there is no difference between the data sets (Brownlee, 2018a). In order to reject the null hypothesis, there has to be enough evidence to suggest that the data sets are different

(Lutes, 2020). The null hypothesis can only be rejected if the test satisfies the pre-determined significance level. For this project, the significance level, also called the p-value, was set at 0.05, meaning that to reject the null hypothesis, the t-test would have to show at least a 95% certainty that the two data sets are statistically different (Lutes, 2020). If the test outputs a p-value of less than 0.05, then the null hypothesis can be rejected, but if the p-value is greater than 0.05, the null hypothesis cannot be rejected.

Two different paired Student's t-tests were used in the analysis of sensitivity. The first, called the two-tailed paired Student's t-test, was used initially to determine if the means of the Surface Burn Classification and the Unburned Tree Noise and Sub-Crown Burn Reclassifications were considered equal or not. Then, a one-tailed paired Student's t-test was used to test if the mean results from the reclassifications were actually greater than the original classifications. The two-tailed test is more general because it considers both sides of the distribution curve, while the one-tailed test is specific to just one side. Graphical representations of the basic structures of the two-tailed and one-tailed paired Student's t-tests are shown in Figures 5 and 6, respectively. The red shaded sections of the graph refer to the rejection regions that each test allows.

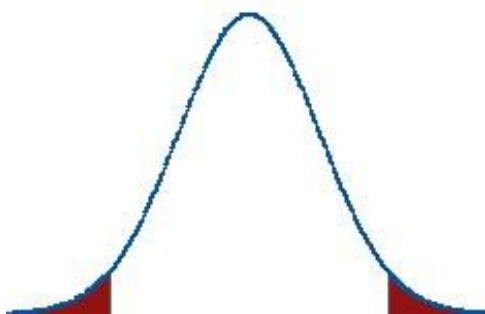


Figure 5. Two-tailed Student's t-test graph

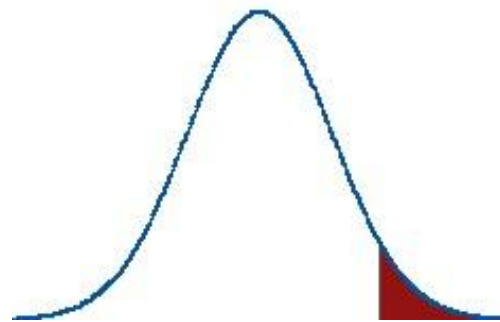


Figure 6. One-tailed Student's t-test graph

These two versions of the paired Student's t-test were run using two different platforms. The two-tailed t-test was calculated using the `scipy.stats` package within Python's SciPy library. This package contains a test called `ttest_rel`, which automatically runs the two-tailed paired Student's t-test when given two data sets, and outputs the resulting p-value (Brownlee, 2018a). This method worked effectively for the two-tailed test; however, Python does not currently have a package for a one-tailed paired t-test. Instead, the t-test package in R was used, with the "paired" parameter set to "true," and the "alternative" parameter set to "less" to indicate that it would only test if the first data set was significantly less than the second data set (*T.Test Function - RDocumentation*). Therefore, the packages provided by Python and R were the only two packages that were used in this project.

Results

The null hypothesis (H_0) for the two-tailed paired Student's t-test was that the Unburned Tree Noise and Sub-Crown Burn Reclassifications created no significant difference in sensitivity compared to the original Surface Burn Classification produced by the support vector machine. Conversely, the alternative hypothesis (H_1) was that there is a significant difference in the sensitivity data as a result of the reclassifications. Running the two-tailed paired t-test, a p-value of 0.036 was computed. This was under the predetermined significance level of 0.05, so H_0 was rejected in favor of H_1 . Therefore, it could be said with greater than 95% certainty that the two sensitivity data sets are significantly different from each other. The results given by the `ttest_rel` Python test are shown in Figure 7.

```
pvalue: 0.03646365885547794
```

```
The two data sets are from different distributions
```

Figure 7. Results of the two-tailed paired Student's t-test

Given this information, the one-tailed paired Student's t-test was then used, with a null hypothesis (H_0) that there was a decrease in sensitivity after the reclassifications, and an alternate hypothesis (H_1) that there was an increase in sensitivity. After running the test, the resulting p-value was 0.018, which was again under the significance level of 0.05. As a result, H_0 was rejected, and it was determined, with greater than 95% certainty, that the reclassifications produced a statistically significant increase in the sensitivity results over the initial classifications. The results given by the R t-test package are shown in Figure 8.

```
Paired t-test
```

```
data: sensitivitySVM and sensitivity5600  
t = -3.6116, df = 3, p-value = 0.01823
```

Figure 8. Results of the one-tailed paired Student's t-test

Future Work

Future work on this project is dependent on the ability to collect more post-fire data. With only four fires, there was limited data available for analysis. Although four fires gave enough data to establish statistical significance for sensitivity, it was not enough to find significance in the accuracy. When more fires are flown and subsequent post-fire data is gathered, these tests could easily be rerun to validate both accuracy and sensitivity results. More data points would make it feasible to analyze a wider variety of metrics, which would further enhance the strength of the overall research project.

Conclusion

The results of the statistical analysis tests clearly proved that the Unburned Tree Noise and Sub-Crown Burn Reclassifications gave higher sensitivity metrics compared to relying solely on the Surface Burn Classification. This means that the reclassifications were significantly more effective at identifying positive burn pixels within the post-fire orthomosaic, with fewer false negatives.

These findings were crucial to the overall FireMAP research project, as they established statistical significance in the results of the new reclassifications. Without this validation, the Unburned Tree Noise and Sub-Crown Burn reclassifications could not be proven to be effective. The results of this research were published in an article entitled *Mapping Forest Burn Extent from Hyperspatial Imagery Using Machine Learning* in the *Remote Sensing* Journal, which has an Impact Factor of 4.8. The Impact Factor measures the frequency with which the average article in a journal is cited in a particular year. A rating of 4.8 signifies that *Remote Sensing* is a high-quality journal.

Overall, this project led to gaining valuable experience in the applications of data analysis and statistical analysis. Learning how to implement analysis tools and strategies into a real-life research project was challenging, fun, and rewarding. In the end, the statistical analysis methodologies used for this project were the same as the processes used in three associated papers by Dr. Dale Hamilton. Independently coming to the same conclusions about the most effective statistical analysis methods was more positive validation for this research. In total, this project provided many learning opportunities, as well as a variety of experiences that can be drawn upon in the future.

References

- Brownlee, J. (2018a, May 17). How to Calculate Parametric Statistical Hypothesis Tests in Python. *Machine Learning Mastery*.
<https://machinelearningmastery.com/parametric-statistical-significance-tests-in-python/>
- Brownlee, J. (2018b, June 14). A Gentle Introduction to the Chi-Squared Test for Machine Learning. *Machine Learning Mastery*.
<https://machinelearningmastery.com/chi-squared-test-for-machine-learning/>
- Brownlee, J. (2018c, July 24). How to Calculate McNemar's Test to Compare Two Machine Learning Classifiers. *Machine Learning Mastery*.
<https://machinelearningmastery.com/mcnemars-test-for-machine-learning/>
- Hamilton, D.; Brothers, K.; McCall, C.; Gautier, B.; Shea, T. (2021). Mapping Forest Burn Extent from Hyperspatial Imagery Using Machine Learning. *Remote Sens.* 13, 3843. <https://doi.org/10.3390/rs13193843>
- Lutes, J. (2020, October 4). *Statistical Significance with the help of Python*. Medium.
<https://towardsdatascience.com/statistical-significance-with-the-help-of-python-1fbb318ce216>
- t.test function—RDocumentation*. (n.d.). Retrieved February 2, 2022, from
<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test>
- Wilkerson, S. (2008). *Application of the Paired t-test*. 5, 6.

Appendix A: Code

a. Two-tailed paired Student's t-test Code

```
main.py

# TWO-TAILED PAIRED T-TEST CODE TO BE USED FOR TESTING
THE STATISTICAL SIGNIFICANCE OF SENSITIVITY DATA

from scipy.stats import ttest_rel

sensitivitySVM = [29.19, 71.28, 70.34, 66.68]
# sensitivitySVM is an array of the four sensitivity metrics
# for the four fires from using just the Surface Burn Classification SVM

sensitivity5600 = [40.19, 88.93, 82.09, 99.54]
# sensitivity5600 is an array of the four sensitivity metrics for the four fires
# from using the Unburned Tree Noise and Sub-Crown Burn Reclassification
#with threshold of 5600 pixels

stat, pvalue = ttest_rel(sensitivitySVM, sensitivity5600)
print("\nt statistic:", stat, "\npvalue:", pvalue)

if pvalue > 0.05:
    print("The two data sets are from the same distribution")
else:
    print("\nThe two data sets are from different distributions")
```

b. One-tailed paired Student's t-test Code

```
SensitivityAnalysis.Rproj

sensitivitySVM <- c(29.19, 71.28, 70.34, 66.68)
sensitivity5600 <- c(40.19, 88.93, 82.09, 99.54)

t.test(sensitivitySVM, sensitivity5600, paired = TRUE, alternative = "less")
#less means x is less than y
#This means that the 5600 threshold is significantly better than the original SVM
```