Predicting Post-Wildfire ReGreen Rates:
An application of multi-factor regression modeling

by

Jessica Ogden Eselius

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

December 2017

Copyright © 2017 by Jessica O. Eselius

To my husband who has supported me through this entire endeavor, to my boys who have waited patiently, and to my Lord and Savior Jesus Christ who gave me the strength to press on

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to my advisor, Professor Kemp, for the direction, encouragement, and editing I needed to get through this process. I thank LTC Devin Eselius, my husband, who has helped me with my many questions and issues I had with RStudio and coding.

List of Abbreviations

A horizon	Alpha horizon, the top most mineral layer
adNBR	adjusted dNBR
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AVIRIS	Airborne Visible and Infrared Imaging Spectrometer
CAL FIRE	California Department of Forestry and Fire Protection
CDF	California Department of Forestry and Fire Protection
DEM	Digital Elevation Model
dNBR	difference in NBR
EROS	Earth Resource Observation and Science Center
ESPA	EROS Science Processing Architecture
Esri	Environmental Systems Research Institute
EVI	Enhanced Vegetation Index
FRAP	Fire and Resource Assessment Program
GIS	Geographic information system
GISci	Geographic information science
IFI	Integrated Forest Index
lm	Linear Regression Function
MAE	Mean Absolute Error
MIR	Mid-Infrared
MOD13Q1	MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid
MODIS	Moderate Resolution Imaging Spectroradiometer
NA	not available/missing values

NAD	North American Datum
NASA	National Aeronautics and Space Administration
NBR	Normalized Burn Ratio
NDII	Normalized Difference Infrared Index
NDVI	Normalized Difference Vegetation Index
NIR	Near Infrared
NPS	National Park Service
NRCS	Natural Resources Conservation Service
O horizon	Organic horizon
OLI	Operational Land Imager
pRI	pixel-based Regeneration Index
RdNBR	Relative dNBR
RMSE	Relative Mean Squared Error
S-G	Savitzky-Golay
SLC	Scan Line Corrector
SNF	Stanislaus National Forest
SSI	Spatial Sciences Institute
SSURGO	Soil Survey Geographic Database
SWIR	Short Wave Infrared
TES	Terrestrial Ecosystem
TRIS	Thermal Infrared Sensor
USC	University of Southern California
USDA	United States Department of Agriculture

USFS	United States Forest Service
USGS	United States Geological Survey
UTM	Universal Transverse Mercator

Abstract

Recovery from wildfires is related to a series of interacting factors. This study was conducted to reproduce and attempt to improve upon the work of Casady et al. (2010) by building a regression decision tree model for predicting post-fire recovery based on interacting environmental factors using two spatial resolutions. Mimicking the efforts of Casady et al. in evaluating post-fire vegetation regeneration rate, their term has been renamed throughout this study as ReGreen Rate, since this is a more accurate representation of how the imagery can be interpreted. This present study used a combination of ArcGIS and R to prepare data from 30 m and 240 m spatial resolutions and analyze model attributes' impact on recovery rates. This study answers two questions. First, does the use of higher spatial resolution data create a more accurate regression tree model predicting the post-fire ReGreen Rate? Second, do different indices of fire severity show a different result in model accuracy? The resulting models all demonstrated a strong correlation between fire severity and rate of vegetation recovery, where greater fire severity lead to faster recovery. As for the first question, 30 m spatial resolution data did provide a marginally more accurate predictive model. However, the model built from the 240 m spatial resolution data was nearly as accurate as the model developed from the 30 m spatial resolution data when applied to the 30 m data. Second, different indices of fire severity did not provide statistically different accuracy in the resulting model. Further research into modeling various forest recovery rates could be useful in constructing generalizable models based on 240 m data to produce a good prediction of recovery for application in forest management, enabling targeted areas for post-fire replanting and optimizing resources allocation.

Casady, Grant M., Willem J. D. van Leeuwen, and Stuart E. Marsh. 2010. "Evaluating Post-wildfire Vegetation Regeneration as a Response to Multiple Environmental Determinants." *Environmental Modeling & Assessment* 15: 295-307.

Chapter 1 Introduction

In the United States, forest fire management has evolved since the early days of forestry, moving from a 1910 U.S. Forest Service (USFS) 100% fire suppression policy to a recognition of the importance of fire on the ecosystem (USFS 2015). The U.S. National Parks Service (NPS) currently notes that various plant and animal species need cyclical fires to thrive in the western forest environment and that without such fire events, dry vegetative matter builds up resulting in more destructive wildfires (NPS 2016). It is in this context that this present study seeks to understand the role of fire severity as a contributing factor to recovery rates after a major forest fire. A better understanding of how multiple environmental factors impact post-fire recovery is essential in guiding responses to future fires in forests containing significant amounts of fuel due to drought, plant pathogens and insect damage (Virginia et al. 2001).

This present study uses a multi-factor predictive modeling method to identify the influences of different environmental factors affecting post-fire ReGreen Rate in California's Stanislaus National Forest. Predicting the influences of various environmental factors on natural ReGreen Rate can help inform post-fire management practices and help optimize recovery efforts by understanding what combination of factors contribute to faster recovery and where additional efforts may be needed to support recovery in areas with little vegetation regrowth.

This present study is based on the previous work of Grant M. Casady, Willem J. D. van Leeuwen, and Stuart E. Marsh who set out to assess post-wildfire plant recovery as a response to various environmental factors using a predictive model (Casady et al. 2010). The research group used a time series of 250 m pixel size enhanced vegetation index (EVI) data to calculate an indication of the rate of recovery, which they called the *post-fire vegetation regeneration rate*. The calculated vegetation regeneration rate was then used with a set of environmental factors to

develop a regression tree model of the 2005 Rodeo-Chediski fire in Arizona. Casady et al. found the regression tree model based on a time series of vegetation data was a useful tool for identifying the dominant factors involved in post-fire recovery by setting the vegetation regeneration rate as the response variable to the set of environmental factors.

The Casady et al. study had three objectives: 1) to observe post-fire forest recovery using a time-series vegetation index derived from satellite imagery; 2) to assess the correlation amongst the post-fire vegetation recovery and a set of environmental factors that may cause variations in post-fire regeneration; and, 3) to estimate the strength of the chosen environmental factors to define post-fire vegetation types appropriately. Casady et al. found that the post-fire response depended on elevation (evaluated as a proxy for water availability), fire severity, pre-burn vegetation type, and post-burn forest management activities. Additionally, Casady et al. postulated that higher spatial resolution data would lead to a more accurate model.

Building on the previous work of Cassady et al., the main objective of this present study is to determine if increasing the resolution of data used in developing the regression tree model can significantly improve the predictive accuracy of the model. Therefore, this present study compares results obtained following a similar workflow conducted for a different fire using EVI data from both 240 m Moderate Resolution Imaging Spectroradiometer (MODIS) imagery and 30 m Landsat imagery.

The EVI value is not able to determine which individual species of vegetation recovered, and species identification requires a combination of remotely sensed data with substantial field sampling for validation. The Casady et al. term *vegetation regeneration rate* is potentially misleading because *regeneration* implies post-fire recovery of the same plant species as were present before the fire. However, the term was used to describe simply the rate of change in the

annual sum of normalized EVI values. For ease of comprehension, hereafter Casady et al.'s vegetation regeneration rate is referred to as the ReGreen Rate. Thus, like Casady et al.'s proposed measure, the ReGreen Rate is a numeric representation of the rate of change in the annual sum of normalized EVI values over each of the three post-fire years at each pixel location.

1.1. Research Questions of This Present Study

In using the work of Casady et al. as a guide, this current study looked at four categories of environmental factors in building the regression tree model with ReGreen Rate as the response variable: topographic factors, soil types, pre-fire vegetation types, and fire severity. This present study explored two key questions. The first question is: does the use of higher spatial resolution data create a more accurate regression tree model predicting the post-fire ReGreen Rate? A secondary question asks: do different indices of fire severity show a different result in model accuracy? The alternate fire severity index was developed by Miller and Thode (2007) and is known to more accurately reflect field sample data (Lydersen et al. 2014).

1.2. The Rim Fire

The Rim Fire is, to date (2017), the largest fire in California's Sierra Nevada Mountains. It was ignited by a lost hunter in a canyon in the Stanislaus National Forest (SNF) in Tuolumne County on August 17, 2013 (Gabbert 2015). The fire burned fast, and in just 12 days blackened nearly 236,000 acres or approximately 92% of the fire's final acreage. The fire burned in areas of challenging topography, as depicted in Figure 1, where ground crews could not combat the fire in the steep terrain thus requiring aerial support for fire suppression. By September 1, aerial crews dropped over three million gallons of fire retardant and water on the fire (NASA 2013). It was over two months before fire crews could contain the Rim Fire, and over a year before it

could be declared fully extinguished, though logs could still be seen smoldering in November of 2014. In the end, a total of 257,314 acres burned, at the cost of over $127 million dollars, with 112 structures lost (CDF 2013; NPS 2013; Potter 2014; USFS 2015).

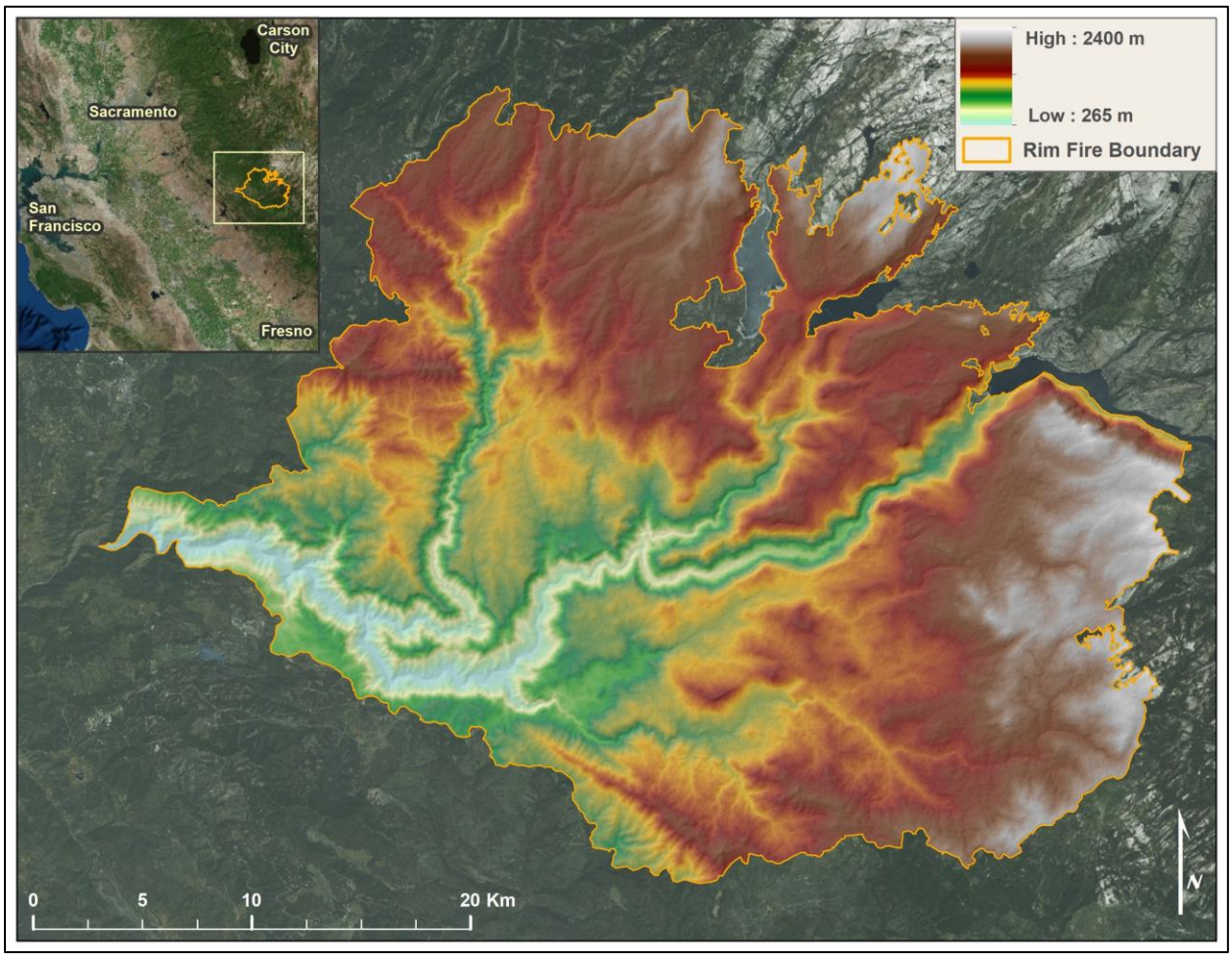

Figure 1 Map showing the fire boundary and elevation across the burned area

The Rim Fire was different from past wildfires: aside from its size, there was the unprecedented severity of the damage to burned areas. USFS survey teams estimated that 28 square miles of soil structure were scorched down to 1-2 inches in depth, destroying fine roots and organic matter, and what remained was a loose surface material which could be easily blown or washed away (Flores et al. 2013). USFS estimated a further 152 square miles had a minimum of 90% tree and vegetation mortality (USFS 2015).

1.3. Fire Management in the Sierra Nevada Mountains

The USFS describes the southern Sierra Nevada Mountains where the Rim Fire burned as a Mediterranean climate with cold, wet winters and hot, dry summers. The historic forest evolved with low severity fires that would burn through every 5-20 years replenishing the soil with essential nutrients. The ancient forest was heterogeneous, having a variety of species, with differing ages and sizes (Parsons and DeBenedetti 1979; Scholl and Taylor 2010; Lydersen et al. 2013; Lydersen, North and Collins 2014).

In the early 20th century the natural fire process was halted with the implementation of the 100% fire suppression policy of the USFS (USFS 2015). As a result, the forests became denser, with shade tolerant trees, shorter than the apex canopy of mixed conifer and Ponderosa pine. This lack of fire had also made the forests less diverse, with many stands at the same age, height, and species making them more susceptible to infection and fire.

In the 1960's, forest scientists began showing that forests were dependent on fire, and the past fire suppression policy was relaxed allowing fires that started naturally to burn if the fire did not threaten private property (USFS 2015). Further, in the 1970's the NPS began a policy whereby, in addition to allowing natural fires to burn like the USFS, they also set prescribed burns to mimic the natural process of cyclical fires (Rothman 2005). However, the forests had already changed and ground fuels, like ponderosa needles, windfall branches, and other dry vegetative matter had built up in the absence of fires to burn away the duff. The forest of the Rim Fire was not at all like the ancient forest.

1.3.1. The Ecological Role of Forest Fires

Frequent and low-intensity wildfires are part of the natural cycle of ponderosa mixed conifer habitat. This regular disturbance aids in the renewal of nutrients as low-intensity fires

burn vegetation. The action of the wildfire benefits the habitat at a few levels. At the local scale, there are changes in the soil structure and mineral composition, making nutrients more available for plant root uptake. Wildfire also plays a factor in plant species distribution and plant competition (Reilly et al. 2006; Lhermitte et al. 2011). At the ecosystem level, low intensity and frequent wildfires change the number and density of different organisms that make up a forest ecosystem, changing the appearance of the forest, and how the forest ecosystem interrelates (Eva and Lambin 2000; Viedma 2008; Lhermitte et al. 2011). Lastly, at the global scale, forest fires change the makeup of forests (Hoelzemann et al. 2004; Lhermitte et al. 2011). Low intensity and frequent wildfire can help to maintain a climax forest, but if a stand-clearing, highly intense wildfire occurs, it can take hundreds of years for the climax forest to regenerate or it might never return. (Nepstad et al. 1999; Lhermitte et al. 2011). With that in mind, the impact of the unusually high intensity of the Rim Fire on the ReGreen Rate is of concern to better inform post-fire reconstruction efforts there and in future California wildfires.

1.4. Structure of This Document

The remaining chapters of this document lay out the research process and conclusions. Chapter 2 contains background on various measures of forest health and fire severity, a review of related works and a description of regression tree modeling as a method of decision tree analysis. Chapter 3 provides a description of the satellite imagery-based data and calculations used to determine ReGreen Rates, as well as all the processing steps required to prepare data for the regression factors of topography, soil types, pre-fire vegetation types, and fire severity. Chapter 3 also describes the construction of the regression tree model. Chapter 4 discusses the results of the modeling effort by comparing the accuracy of the model built using 30 m spatial resolution data and the model developed using 240 m spatial resolution data, as well as the accuracy of models

based on different fire severity indices. Chapter 5 concludes with a discussion of opportunities for further studies to assist in predicting ReGreen Rates.

Chapter 2 Background

The literature identifies four steps that are required to examine post-fire recovery. First, a process for assessing fire severity is needed. Secondly, it is necessary to determine different factors, beyond fire severity, that account for the post-fire vegetation response such as topographic and biological factors as well as pre- and post-fire forest management factors such as burn history or reseeding. Thirdly, the rate of recovery must be determined, which requires a method that establishes a comparable rate of recovery throughout the fire-affected area. Finally, the fourth step is the construction of a predictive model. This chapter examines the important concepts behind each of these steps as they are described in the literature and concludes with a description of how the four steps were applied in the Casady et al. study and addresses an alternate method of determining the rate of recovery.

2.1. Satellite Imagery for Forest Analysis

Sensors on NASA's Landsat 8 and Terra constellation are used to calculate many index values in the earth sciences. At the time of the Casady et al. study the Landsat 7 imagery had (and still has) striping across all its images due to a breakdown of the scan line corrector (SLC) in 2003. The most recent addition to the Landsat family of constellations, Landsat 8, was set into orbit in February 2013, only months before the 2013 Rim Fire. In addition to the elimination of the striping problem, the Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS), hereafter referred to collectively as Landsat, has capabilities not previously available in earlier Landsat systems. The Landsat constellation provides 16 day temporal resolution imagery but does not offer any post-capture processing to eliminate cloud obscuration.

Due to Landsat 7’s problems, Casady et al. used data from the Terra constellation equipped with the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, hereafter

referred to as MODIS, which provides processed imagery every 16 days. MODIS processing involves aggregating 16 daily images to mask and fill obscuring pixels to produce one image that represents the 16 day period. MODIS imagery comes in a sinusoidal projection with trapezoidal pixel cells.

Both Landsat and MODIS provide multispectral bands that are used to calculate various index values. Casady et al.'s method uses the Enhanced Vegetation Index (EVI), which is useful in characterizing vegetation density or health based on relative chlorophyll reflectance and absorbance (Weier and Herring 2000). MODIS-based EVI has a pixel resolution of 10 arc seconds (~240 m) whereas Landsat has a ~30 m pixel size.

The Normalized Burn Ratio (NBR), also used in this study, uses bands in the near infrared (NIR) and mid-wave infrared (MIR) that are sensitive to the changes of living green plant matter, moisture content, and soil environments which may occur after fire (Miller and Thode 2007). The Landsat TIRS captures the MIR equivalent in the shortwave infrared (SWIR2) range as shown in Table 1 which summarizes the different bands used in this present study. Highlighted in gray are the Landsat bands.

Table 1 Satellite sensor bands for MODIS, OLI, and TIRS used in the creation of EVI and NBR

Satellite	Sensor	Band	Wavelength (μ)	Resolution (m)	Index Use	Description
Landsat	OLI	2	0.45 - 0.51	30	EVI	blue
Terra	MODIS	3	0.459 - 0.479	~240	EVI	blue
Terra	MODIS	1	0.620 - 0.670	~240	EVI	red
Landsat	OLI	4	0.64 - 0.67	30	EVI	red
Terra	MODIS	2	0.841 - 0.876	~240	NBR, EVI	NIR
Landsat	OLI	5	0.85 - 0.88	30	NBR, EVI	NIR
Terra	MODIS	7	2.105 - 2.155	~500	NBR	MIR
Landsat	TIRS	7	2.11 - 2.29	30	NBR	SWIR2

2.2. Comparing NDVI and EVI

Casady et al. compared regeneration rates calculated from MODIS-based EVI and Normalized Difference Vegetation Index (NDVI) for a forest dominated by ponderosa pine (*Pinus ponderosa*) with small representation of mixed conifer forest at higher elevations (over 2000 m) and deciduous oak and juniper trees in the lower elevations (around 1800 m). Casady et al. concluded that EVI values provided "consistently better results" (p. 296) than NDVI and only presented the results of their use of EVI as a basis for model construction. Given the similarity of vegetation and elevation mixtures to this present study, and to remain consistent with the foundational study, EVI was selected as the preferred index when calculating the relative vegetation level over time for the Rim Fire region.

NDVI, shown in Equation 1, is a common standard used in quantifying the amount of vegetation in a pixel and is derived from the difference between the NIR and the red bands, which are sensitive to differences in chlorophyll. A limiting factor of this method is that it does not consider the differences of vegetation types or density of vegetation that reflect in the relevant spectral bands red and NIR (Table 1 shows sensor band ranges).

$$\text{NDVI} = \frac{(\text{NIR}-\text{red})}{(\text{NIR}+\text{red})} \quad (1)$$

The enhanced vegetation index (EVI) is like NDVI but includes compensating coefficients on the red and blue bands to remove the influence of the aerosols from the denominator as shown in Equation 2.

$$EVI = G \times \frac{(NIR-red)}{(NIR+C_1 \times red - C_2 \times blue + L)} \quad (2)$$

Where L=1 is a canopy background adjustment that addresses NIR and red radiant transfer through a canopy (Huete et al. 2002, p. 198)), $C_1 = 6$ and $C_2 = 7.5$ are coefficients used to compensate for aerosol influences on the red band, and G = 2.5 is a gain factor. The USGS

product catalog describes the coefficients as reducing background noise, atmospheric noise, and saturation in most cases (Vermote et al. 2016).

Analysis of the differences between NDVI and EVI for a pre-fire year and the first two years post-fire by Chen et al. (2011) found a high correlation with ground-based samples. The correlations were weak beyond the second-year post-fire. These findings indicate a temporal limitation in the correlation between remotely sensed and field samples for both the NDVI and EVI indices beyond three years after a fire incident.

2.3. Measuring Fire Severity

Severity is a qualitative descriptor of the degree of distress resulting from the intensity, heat, and duration of the fire (Díaz-Delgado et al. 2003; Key and Benson 2006). A fire has a spectrum of impacts across the biologic, atmospheric and social dimensions. Depending on which perspective is taken, the "severity" of fire is measured and classified differently (Jain and Graham 2003). For example, when assessing atmospheric severity, one could look at the CO_2, particulates and toxic gasses released by the fire to gauge the impact on air quality or climate change. Even when considering only the perspective of assessing the biologic severity, there are a multitude of ways that can be used to evaluate the significance of the fire's impact beyond the direct results of the fire, such as soil erosion, stand-replacement mortality, nutrient cycling, or vegetation recovery (Kokalya et al. 2007).

Some studies categorize fire severity into levels using extensive field studies or aerial collection systems with high spatial resolution (e.g. 2.4 m pixel size) (Díaz-Delgado et al. 2003; Kokalya et al. 2007). Studies such as those by Kokalya et al. (2007) and Robinchauda et al. (2007) indicate that models based on higher spatial and temporal resolution provide more accurate descriptions of the environment which in turn influence the understanding of post-fire

effects. While a more spatially precise instrument can provide higher resolution severity maps that better match field data, the more precise data collection tools are expensive and do not typically provide access to the timeline of both pre- and post-fire conditions available through satellite imagery.

Casady et al. used the Normalized Burn Ratio to derive three different indices to characterize fire severity: the differenced Normalized Burn Ratio (dNBR), the relative differenced Normalized Burn Ratio (RdNBR), and the adjusted Normalized Burn Ratio (adNBR). This present study examines model accuracy when using RdNBR versus adNBR as index values for fire severity. Discussion of each of the three fire severity index methods are in the following sections.

2.3.1. Normalized Burn Ratio (NBR)

The NBR is calculated as a ratio of near infrared (NIR) and shortwave infrared bands (SWIR). According to Miller and Thode (2007), in a post-fire examination, the MIR band is particularly adept at differentiating dead wood from soil, ash and charred wood. The Landsat SWIR2 band is the functional equivalent of the MODIS MIR band.

The equation for NBR is:

$$\mathrm{NBR} = \frac{(\mathrm{NIR}-\mathrm{MIR})}{(\mathrm{NIR}+\mathrm{MIR})} \times 1000 \quad (3)$$

For the MODIS sensor, MIR is band 7, captured in the range 2.105 - 2.155μm, and NIR is band 2, captured in the range 0.841-0.876 μm. For Landsat 8, the SWIR is band 7 (range 2.11 - 2.29 μm), and NIR is band 5 (range 0.85 - 0.88 μm) (NASA n.d.). Note, the convention is to multiply the NBR value by 1000 to turn the index into an integer value.

2.3.2. Differenced Normalized Burn Ratio (dNBR)

While the NBR by itself is not a burn index, the difference between the pre- and post-fire NBR called the differenced NBR or dNBR (sometimes called the ΔNBR), is an indication of the amount of vegetation destroyed in the fire. It is calculated as follows:

$$dNBR = NBR_{\text{pre-fire}} - NBR_{\text{post-fire}} \quad (4)$$

Miller and Thode (2007) found that the dNBR produces classification errors. As shown in Figure 2, the high severity fire in "A" has a lower dNBR than the moderate severity fire in "C" meaning that a classification threshold set based on the dNBR value of "A" would over represent fire severity by including areas of moderate severity in the classification of high severity. This present study does not use classification thresholds, but rather the continuous value. However, the fact that the dNBR can misconstrue the level of fire severity in categorical methods indicates it is not reliable as a method of determining fire severity by itself. Miller and Thode and Casady et al. proposed to use instead the relative differenced Normalized Burn Ratio (RdNBR) and adjusted difference Normalized Burn Ratio (adNBR), respectively.

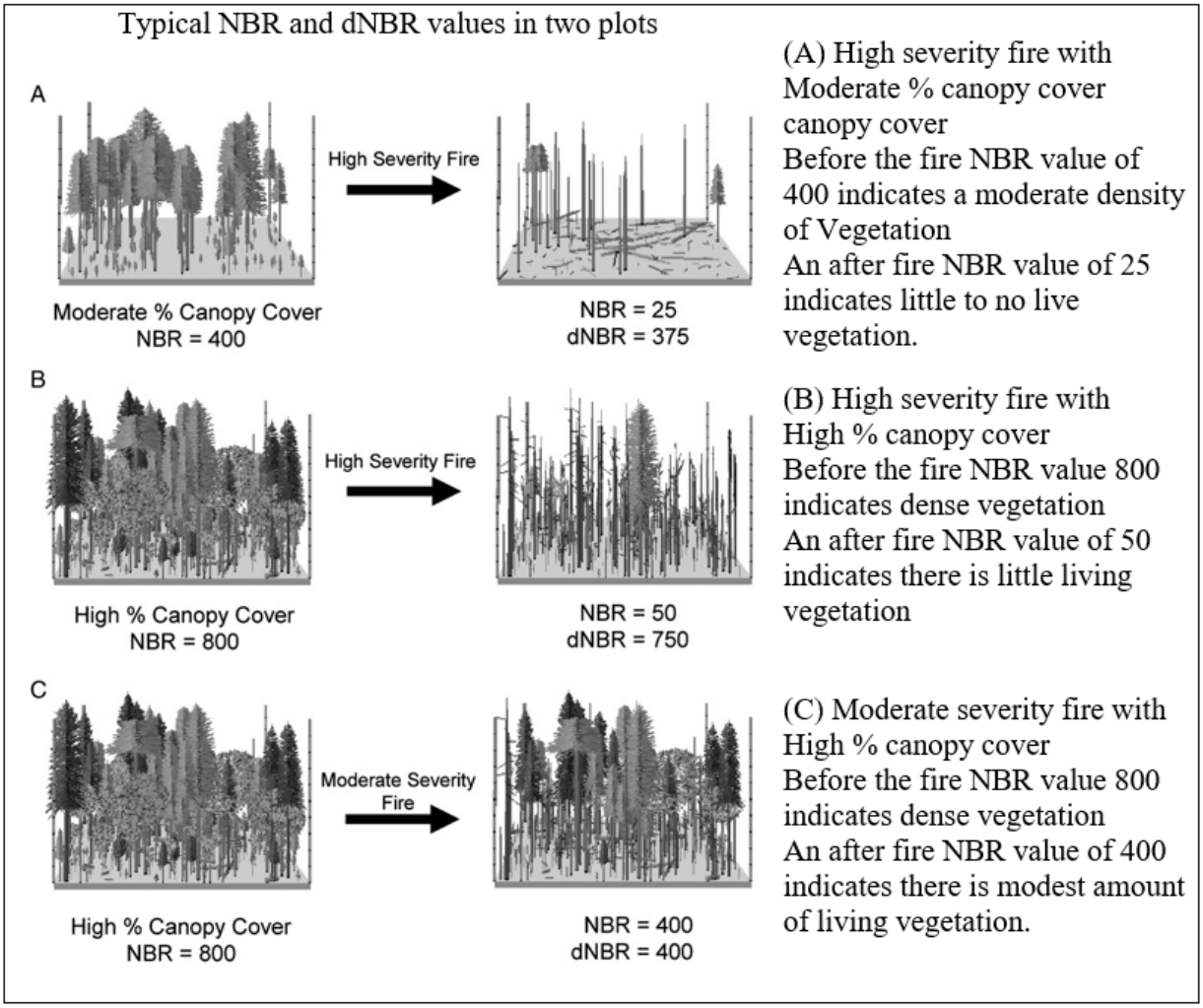

Figure 2 Classification errors inherent to dNBR (reproduced from Miller and Thode 2007)

2.3.3. Relative Differenced Normalized Burn Ratio (RdNBR)

As discussed in the previous section, Miller and Thode (2007) found that the dNBR produces classification errors based on the impact of the classification thresholds applied to different pre-fire conditions and may over-represent fire severity by including areas of moderate severity in the classification of high severity. Miller and Thode improved the dNBR by eliminating the correlation to the pre-fire NBR by dividing by the square root of the raw pre-fire NBR value. This method is the relative differenced Normalized Burn Ratio (RdNBR), which

indicates the amount of vegetation killed relative to the amount of pre-fire vegetation as shown in Equation 5:

$$\text{RdNBR} = \frac{\text{PreFireNBR} - \text{PostFire NBR}}{\sqrt{\frac{\text{PreFireNBR}}{1000}}} \quad (5)$$

The terms PreFireNBR and PostFireNBR are the NBR values just before the fire and just after the fire. Recalling that the convention is to multiply raw NBR values by 1000 to convert them to integer values, Miller and Thode divided by the square root of the raw (i.e., no scaling factor applied) PreFireNBR

Several studies of the Rim Fire found that the RdNBR correlated well with their field-collected data on fire severity and with the ratio of change in the average amount of an area occupied by tree stems and canopy cover (Lydersen et al. 2013; Potter 2014; Harris and Taylor 2015).

2.3.4. Adjusted Differenced Normalized Burn Ratio (adNBR)

Casady et al. attempted to use the RdNBR but found that it resulted in an unbounded solution. This result may have been due to using the indefinite or very large RdNBR values stemming from dividing by a near zero value (i.e. when $\sqrt{PreFireNBR/1000}$ is very small). Instead of using RdNBR, Casady et al. used an "adjusted dNBR" as a modeling factor. They arrived at this result by generating a least squares line to a plot of the pre-fire NBR against the dNBR (with an R^2 value of 0.53) and using the residual differences from that line on each pixel to capture the positive (more severe) and negative (less severe) departures from that best fit line. Their explanation was that this method allowed use of a factor which accounted for deviations from an expected value of dNBR based on the pre-fire NBR as opposed to directly using a

relative severity factor. Their examination found that the adNBR provided a better model than one based on using the dNBR as a factor.

2.4. Environmental Factors Affecting Forest Recovery

While fire severity is an important factor in post-fire recovery, research has demonstrated that there are many other contributing factors influencing recovery and that there are phases of the recovery process (DeBano et al. 1998; Amiro et al. 2000; Pollet and Omi 2002; Goetz et al. 2006). The most important common factor is the availability of water. There are several factors considered to be analogous to water availability, such as flow accumulation, elevation, and aspect. Many studies have looked at how different topographic and biological factors individually affect recovery, but according to work published by Shalizit (2009), only a few studies have examined the impact of combining multiple factors to predict a response variable.

Casady et al. used flow accumulation, elevation, and aspect as potential analogs for the availability of water. For aspect, the sine and cosine of the aspect served to indicate East-West and North-South with the idea that a northerly facing slope would receive less sun and therefore less evaporation and retain more moisture. Casady et al. used the USFS's Terrestrial Ecosystems Survey (TES) data to capture soil type and vegetation type in the form of "Map Units." Map Units combine areas of similar vegetation, topography, and geology into survey areas to uniquely classify regions based on a set of similar parameters. Different forests use different defining parameters which limit the utility of map units as predictors of fire recovery to a given survey area.

In constructing of their regression tree models, Casady et al. used the map units to calculate their factors used. The predictive factors created from the Map Units are, four vegetation types (tree cover, shrub cover, sprouter [sic] cover, and herbaceous cover) and six soil

characteristics (percent clay, percent rock fragment, percent soil organic matter, depth of the organic horizon (O horizon), and the top-most mineral layer (A horizon), and total soil depth). These are based on the map unit's relative percentage of the factor.

2.5. Regression Trees

The term decision tree analysis covers two methods: classification decision trees which use categorical response variables, and regression decision trees which use continuous response variables. Since this present study used a continuous response variable, only regression decision trees are relevant here.

Regression tree analysis uses multiple factors as inputs to predict an outcome of a continuous variable based on the descriptive power of the input elements (Loh 2016). A typical example of regression tree modeling is finding the selling price of a house based on a mixture of factors such as square footage, attached garages, the rating of a local school, and availability of public transportation. Regression tree modeling uses localized decisions as branches about a single factor (e.g., square footage is greater than 1500, yes or no) to lead to a prediction of the value of an output variable (leaves). The recursive process works to find the best fit by minimizing the variance within the two sides of the split, also known as the branches or nodes. Recursive, or looping, attempts at building different route structures (which order of factors and which way to go – yes, no) produces a regression tree with good (though not necessarily perfectly accurate) predictions based on a set of training data.

Overfitting, making the regression tree fit the data too precisely, leads to a model that has little predictive power outside the training data. Thus, "trimming" of the regression tree, which depends on setting definitions for stopping the recursive process, is necessary to ensure that the model retains its usefulness as a model. The final "leaves" of the tree consist of elements with a

minimized “within-leaf” variance and are summarized as the average response value for all training data that followed the same decision path (Loh 2016). When using regression trees, measuring model accuracy is a matter of determining the difference between the observed and the predicted values.

2.6. Modeling Post-Fire Recovery

The research conducted in 2008 and published in 2010 by Casady et al. found that few studies of post-fire regeneration examined more than one variable in the analysis of the forest’s post-fire response. The Casady et al. team used regression tree analysis to predict the EVI-based rate of recovery based on a set of predictive parameters. While they note that there are elements that could be improved, they concluded that their fundamental method provides a good predictive estimation of post-fire recovery. A similar method for determining recovery is used in the later work by Lhermitte et al. (2011) who also used a remotely sensed vegetation index to capture an indication of the amount of green vegetation present. These two studies are next considered individually in more detail.

2.6.1. The Example Study

Casady et al. (2010) used EVI data covering the period three years before and five years after the fire. Using a time series of bi-monthly snapshots that characterized the amount of green vegetation present in each 250 m pixel, they divided each pixel’s EVI value by the corresponding pixel average of the bi-monthly values for the three pre-fire years. This method created normalized EVI values at each pixel which were then summed at each pixel on an annual basis for each of the five post-fire years. The summed values represented the amount of green in each of those years at each pixel. For the first year after the fire, the summed value was low. Subsequent years had larger annual sums of the normalized EVI values indicating recovery.

The annual sum of normalized EVI values for each of the post-fire years produced a set of five points that were then used to generate a least-squares fit line. The slope of that line, which they called the post-fire EVI slope, was then used as the indicator of the rate of regeneration at each pixel.

Then, with the post-fire EVI slope value held as a response variable at each pixel, Casady et al. constructed three regression trees to identify and model the importance of several factors (pre-fire NBR, adNBR, dNBR, elevation, sine and cosine of aspect, flow accumulation, map units and derived soil and vegetation factors). For digital elevation model (DEM)-derived data used in the regression modeling, high spatial resolution information was averaged to provide the required 250 m pixel level attributes. Thus, the elevation values in 625 contiguous 1/3 arc second pixels were averaged to produce a value for each 250 m MODIS pixel.

The first Casady et al. model used all the environmental and fire severity (excluding the adNBR) factors, including map units and the derived soil and vegetation values. The second used all the same factors but replaced the fire severity dNBR with adNBR. The first two models of the Casady et al. study served the same purpose as this present study's examination of adNBR versus RdNBR; evaluating different methods of defining fire severity. The third model looked at all the factors except dNBR and the map units. This process was done to evaluate the impact of any particular soil or vegetation factor as a predictive variable.

Casady et al. used the R^2 value between the predicted and observed values to determine accuracy of the three regression trees and found that the second model using adNBR and including map units provided a marginally better model ($R^2 = 0.181$) than when using the traditional dNBR ($R^2 = 0.179$) and even better than model three ($R^2 = 0.148$) where map units were removed so that individual soil and vegetation factors were more important in the model.

It is worth pointing out some issues of concern with Casady et al.'s preparation of the data. First, the pre-fire EVI values were averaged across the three years. As discussed in Lhermitte et al. (2011), this technique does not account for normal variations in vegetation due to seasonal changes. The second point of concern is the use of smoothing for the EVI values. Casady et al. applied a Savitzky-Golay (S-G) smoothing technique on their bi-monthly EVI values. This approach was based on the work of Jonsson and Eklundh (2002) who accounted for sensor and environmental variability by smoothing daily measurements to arrive at a seasonally smoothed set of NDVI values. In describing this S-G smoothing process in Chapter 14 of *Numerical Recipes in FORTRAN; The Art of Scientific Computing* (cited over 115K times), Press et al. offered the following critique:

> We must comment editorially that the smoothing of data lies in a murky area, beyond the fringe of some better posed, and therefore more highly recommended, techniques that are discussed elsewhere in this book. If you fit data to a parametric model, for example, it is almost always better to use raw data than to use data that has been pre-processed by a smoothing procedure (Press et al. 1993, p. 644).

While two of the Casady et al. models included clay, rock fragment, and shrub factors, their selected optimal model did not contain any of these factors as a determinant. The chosen model did find that map units were the most significant factor followed by elevation (interpreted as an analog for water). Since map units represent an amalgamation of environmental components, they noted that those environmental factors are therefore important, though the use of the map unit alone precludes any insight into which components are determinant.

2.6.2. An Alternative Method for Describing Post-Fire Recovery

An alternate method for determining the rate of recovery of a forest is provided by Lhermitte et al. (2011) who developed a method of using NBR to determine the rate of vegetation regrowth after a fire. Lhermitte et al. looked specifically at the first year after the fire

to examine the intra-annual vegetation regrowth rate using an index they created called the pixel-based Regeneration Index (pRI). When calculating the pRI using the NBR, each pixel was normalized by dividing each post-fire NBR time series value (VI_{burn}) by the NBR value from one-year prior ($\overline{VI}_{control}$) as shown in Equation 6:

$$\text{pRI} = \frac{VI_{burn}}{\overline{VI}_{control}} \tag{6}$$

The resulting NBR-based pRI values represent an index value which has been normalized by a value that approximates the vegetation health as if the fire had not occurred. However, this method of creating a normalized index has the disadvantage that it depends on a complex sliding average and is heavily influenced by any drought conditions that may have preceded the fire. This method may serve as a basis for further study in modeling post-fire recovery because it addresses seasonal variability across the recovery period.

2.7. Summary

The research discussed in this chapter provides strong support for several decisions made in the design of this present study. The literature provides a solid foundation for the merit of using remotely sensed fire severity and vegetation index values as these have been shown to provide good correlation with field observations. Research has also shown that such index values remain valid over the three-year post-fire period. Environmental factors that indicate access to water and the types of pre-fire vegetation and soil were shown to have a strong correlation to post-fire recovery. Regression tree analysis has been shown to be a useful means to model post-fire recovery, and it allows for an easily interpretable predictive model of a response variable. Both Casady et al. and Lhermitte et al. used a technique of normalizing an index value, and Casady et al. created a linear regression line through an annual sum of the normalized EVI value at each of the post-fire years to model the rate of recovery. The next chapter outlines how this

study generally follows the method described by Casady et al. with departures as needed to account for data and environmental differences.

Chapter 3 Data and Methods

This chapter provides an overview of the satellite imagery used, the data acquisition process, the data processing workflow, and model construction. Much of the effort in this study was spent on acquisition of the data and data processing. The data used and its acquisition is described in the following section.

Later sections in this chapter describe the processing that was carried out on each data source to create the data used in the model. Figure 3 summarizes the data processing and model building stages. Processing using ArcGIS entailed reprojection, clipping to the study area, rasterization and registering all raster layers to a common template. With the attributes in a common template, missing or obscured pixels were processed in R to filter and smooth the anomalous data. Finally, R was used in the construction of the models and model comparisons.

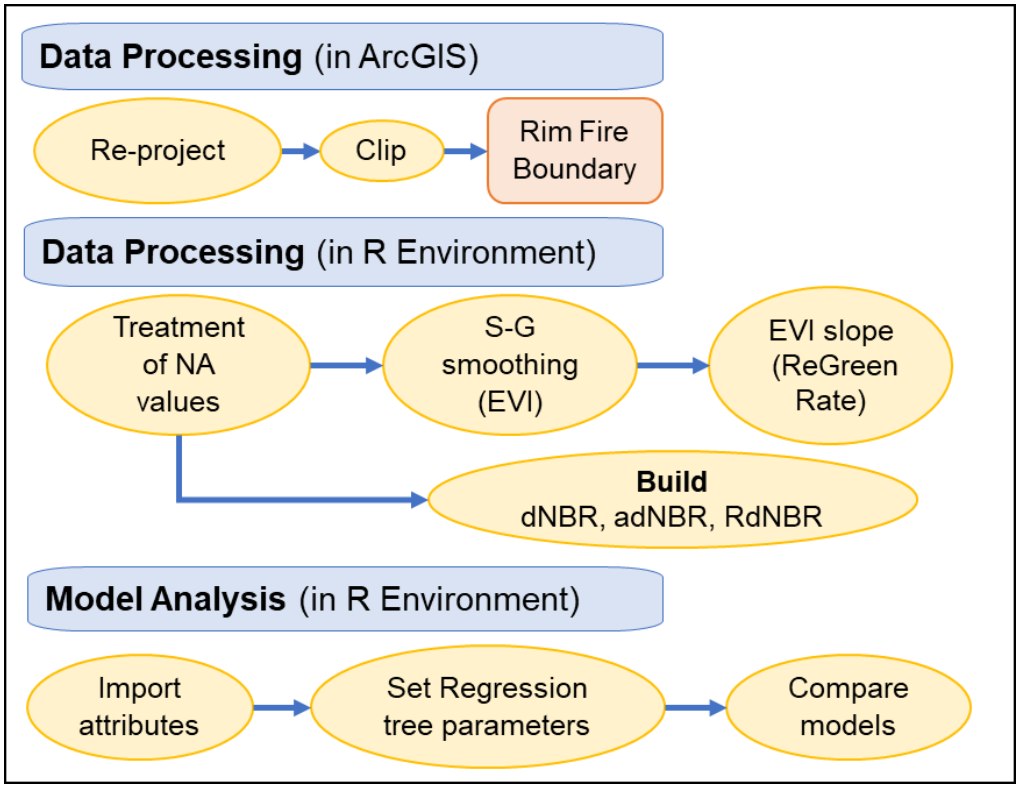

Figure 3 Overview of data processing steps

With respect to the time period used for this present study, while the Casady et al. study used five years post-fire to establish the recovery period, the accuracy of EVI as an indicator of post-fire vegetation response has a diminishing validity beyond a two-year span, and forest recovery is nonlinear in its progression (Chen et al. 2011). For this reason, the three-year period after the 2013 Rim Fire was assessed as sufficient to observe the initial phase of forest ReGreen Rate.

3.1. Data Acquisition

The first step in the model building process was to gather the needed data. As shown in Table 2, USGS provided DEM and EVI data while CALFIRE and Esri provided the remaining data. As with Casady et al., this present study focused on aspects of the environment that capture topography, hydrology, vegetation, and soil characteristics. However, this present study differs in that no Terrestrial Ecosystems Survey Map Unit data were available for the study area, and so vegetation and soil information were derived from CALFIRE and the Soil Survey Geographic Database (SSURGO).

Table 2 summarizes the data and sources used in this present study. Most of the data came directly from the USGS Earth Explorer website (https://earthexplorer.usgs.gov), Esri, or was ordered for downloading from the Earth Resource Observation and Science Center (EROS) Science Processing Architecture (ESPA https://espa.cr.usgs.gov).

Table 2 Environmental factors and sources

<table>
<tr><th>Initial Data</th><th colspan="2">Derived Data</th><th>Data Source</th><th>Organization</th><th>Type</th></tr>
<tr><td>rim fire 9_24</td><td colspan="2">Fire Boundary</td><td>ArcGIS online Story Maps</td><td rowspan="2">Esri</td><td rowspan="2">Vector</td></tr>
<tr><td>SSURGO</td><td colspan="2">soil type</td><td>SSURGO Downloader 2014 for ArcGIS</td></tr>
<tr><td>FVEG15_1</td><td colspan="2">vegetation type</td><td>FRAP</td><td>CALFIRE</td><td rowspan="4">Raster</td></tr>
<tr><td>30 m EVI</td><td colspan="2">30 m ReGreen Rate</td><td rowspan="2">Landsat 8 OLI/TIRS</td><td rowspan="3">USGS ESPA</td></tr>
<tr><td>30 m Pre-Burn NBR
30 m Post-Burn NBR</td><td colspan="2">30 m adNBR</td></tr>
<tr><td>250 m 16 Day EVI</td><td colspan="2">240 m ReGreen Rate</td><td>MODIS MOD13Q1 V6</td></tr>
<tr><td>250 m MIR Pre-Burn
250 m NIR Pre-Burn
250 m MIR Post-Burn
250 m NIR Post-Burn</td><td rowspan="2">NBR</td><td>240 m RdNBR</td><td></td><td></td><td></td></tr>
<tr><td>240 m adNBR</td><td></td><td></td><td></td></tr>
<tr><td>DEM</td><td colspan="2">Elevation
Cosine Aspect
Sine Aspect
Slope
Flow Accumulation</td><td>Aster Global DEM</td><td>USGS Earth Explorer</td><td></td></tr>
</table>

3.1.1. Fire Boundary

For this present study, the extent of the fire was obtained through Esri's portal, ArcGIS online. Posted in September of 2014 by Esri, the vector layer "rim fire_9_24" captured the Rim Fire extent on September 24th, 2013 when it was fully contained.

3.1.2. EVI Data

Figure 4 outlines the basic steps required to acquire the EVI data. The USGS Earth Explorer site provides multiple products for both MODIS 240 m imagery and Landsat 30 m imagery.

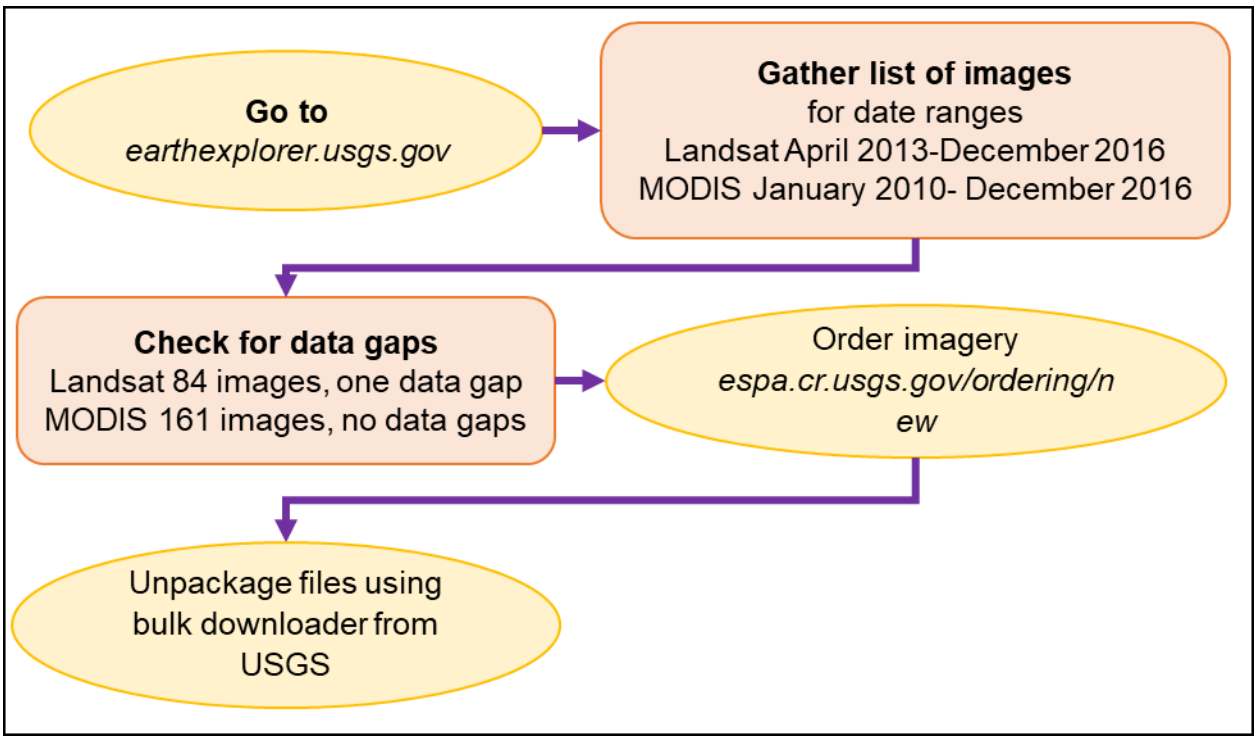

Figure 4 Basic acquisition steps for EVI

The 240 m EVI imagery used in this present study comes from the NASA Land Processes Distributed Active Archive Center Terra Satellite MOD13Q1 product. It is supplied in a global sinusoidal projection. Images for January 2010 to December 2016 were acquired. The MOD13Q1 product comes with four bands (red, blue, NIR, MIR) and the EVI product. The band information from images just before and just after the fire were used in calculating the NBR which is described in Section 3.2.2.

The Landsat 8 30 m imagery comes from the combined OLI and TRIS instruments, and available products include both EVI and NBR. Images for the period April 2013 to December 2016 were acquired. As mentioned before, the striping error on Landsat 7 prevented the possibility of normalizing the Landsat-based EVI values using a multi-year average.

MODIS imagery is collected every 1 to 2 days, and over a 16-day period the images are analyzed and processed together (on the "processed date") to create an average image. Having

the MODIS imagery averaged over a 16-day period produces a collection of virtually cloudless images. On the other hand, Landsat data is taken once every 16 days and distributed as single frames for each date of acquisition. This difference was significant when it came to processing the images for calculating the ReGreen Rate because the cloudy days in the Landsat images required more interpolation and smoothing than the MODIS images.

Once all the data was downloaded, and dates reconciled, one Landsat image was missing. The image was missing because the Landsat thermal sensor that collects the Surface Reflectance data was unavailable from 30 January 2015 to 19 February 2015. The missing image was from February 10th, 2015. To address this gap, ArcGIS 10.4 Raster Calculator was used to construct an image in which each pixel's values were the average of those on the images taken before and after the data gap. There were three additional Landsat images that were completely or nearly completely covered in clouds which required similar filling and smoothing. Though constructing data may not be preferred or precise, this method did produce an approximation of the missing data needed in the subsequent processing.

Ultimately 150 MODIS and 77 Landsat images were needed. The 150 MODIS images provided three years prior to the fire, as Casady et al. used, and three years after. The average of the three years prior was used to normalize the three post-fire MODIS-based EVI values. Landsat 8 was not available for the full three years before the fire, and so only 77 images were available, with only 4 months or 8 images before the fire available to establish an average for use in normalization of post-fire EVI values.

3.1.3. NBR Data

While Landsat pre-processing provided a specific NBR product, MODIS did not. However, MODIS products come with the required band information, specifically the MIR and

NIR bands, to manually calculate NBR values from one pre-fire (July 2013) and one post-fire (November 2013) image. Figure 5 shows the basic steps in acquiring the needed layers for NBR.

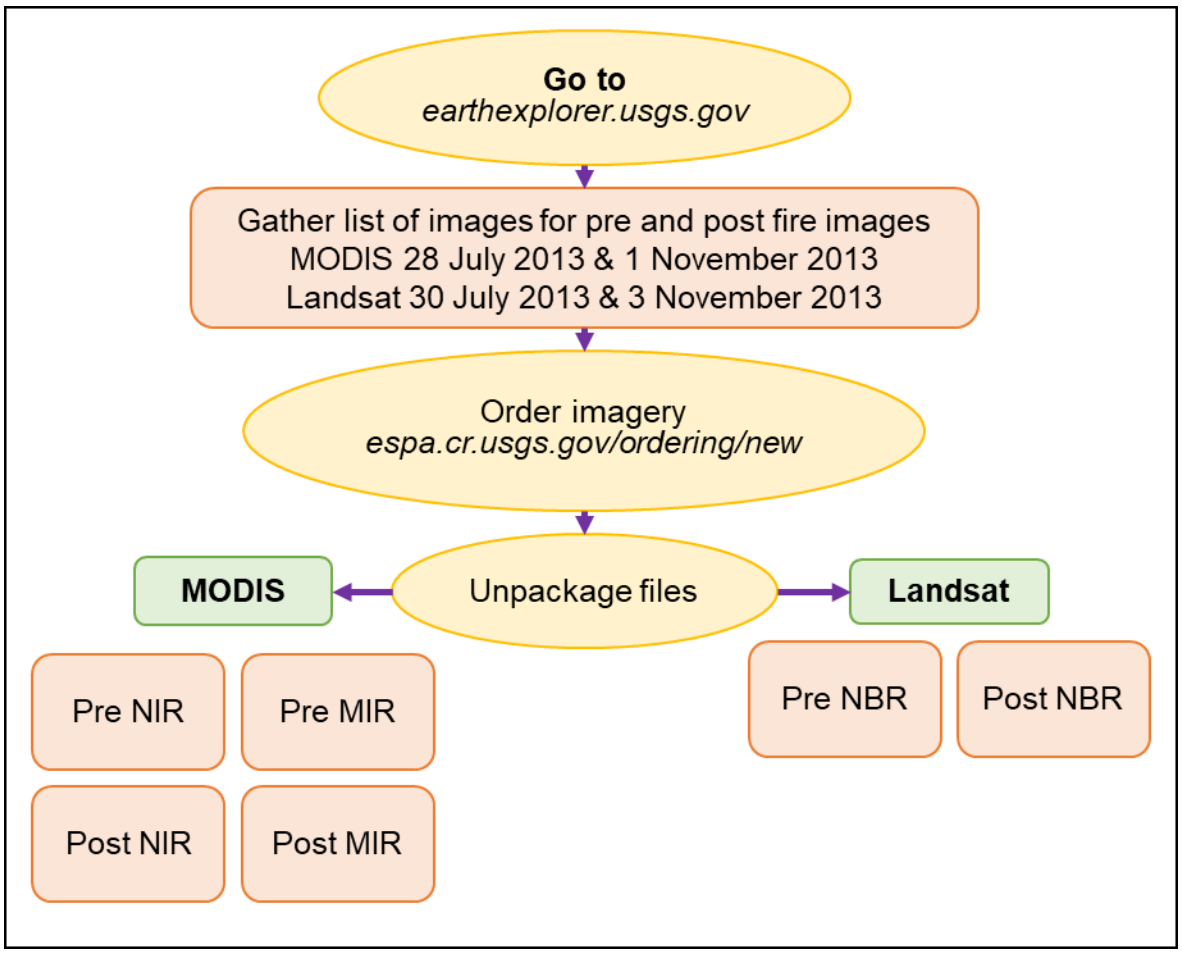

Figure 5 Basic acquisition steps for NBR data

3.1.4. DEM Data

The fire extent required four tiles of 30 m DEM data. Elevation was measured in meters. Using the mosaic tool in ArcGIS, the files were combined into a single raster that was then reprojected into NAD 83 UTM zone 10N and clipped to the fire boundary. The data came from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model Version 2 (GDEM V2) provided by NASA through the Land Processes Distributed Active Archive Center (LP DAAC) section of the USGS Earth Explorer website.

3.1.5. Soil Data

Soil information was obtained from the Esri Hydro Reference Overlay. The Esri data set is a polygon layer using a map unit system to code for 157 discreet information fields. Soil data is also available through the United States Department of Agriculture (USDA) Natural Resources Conservation Service (NRCS), Web Soil Survey site. There, soil attributes are incorporated into detailed taxonomic descriptions of Map Units, identified by a code. However, it is an arduous process to interpret what the numeric codes mean from the USDA information, so the Esri-provided information was used. The Esri information is derived from SSURGO, a compilation of soils information collected over the last century by the NRCS, and comes as vector polygons with 12 distinct greater group soil classification types, in an Albers Equal Area projection. Thus, the information is from the same source ultimately, though the Esri processing made the information more accessible for this present study. The study area required three contiguous data layers from the Esri site to compile the complete soil layer within the study area. How the soil data was decomposed for use in the model is described in Section 3.2.5.

3.1.6. Vegetation Data

Vegetation types are available through the California Department of Forestry and Fire Protection (CAL FIRE) (California Department of Forestry and Fire Protection 2017) database for the Fire and Resource Assessment Program (FRAP) Vegetation (FVEG15_1). Data comes as a Statewide Geodatabase in the California Teale Albers NAD83 projection and is current as of 2015. CAL FIRE maintains the FRAP to assess the amount and extent of the forest and rangelands in California to support the analysis of conditions and enable assessments of different management and policy guidelines (CAL FIRE 2012). The site provides free access to 18 sets of GIS data for a variety of applications ranging from county boundaries and facility locations to

fire threats and vegetation data consisting of various levels of taxonomic specificity (this study used 8 general landcover types from the vegetation data).

3.2. Data Processing

All data used in this present study were reprojected, using the Project Raster tool, into the projected coordinate system UTM NAD 1983 Zone 10N. Data derived from Landsat had a final pixel size of 30 m x 30 m, and the MODIS data had to be registered to the Landsat pixels by setting the spatial resolution to a whole multiplier of 30 to enable the Landsat pixels to reside completely within a MODIS pixel. This study used 8 as a multiplier which resulted in a 240 m MODIS pixel size. One data layer, Soils, needed to be rasterized after projection. During the reprojection process, raster layers were resampled to a common raster template and clipped to the fire boundary.

Each interpolation and smoothing step was examined for the impact on a selection of spatial points (pixels) to gauge the result of the processing step on the data set and confirm its validity. The procedures outlined in Section 3.2.1 are not only an effort to minimize distortion of the original data but also to ensure a comprehensive data set for use in the model construction.

3.2.1. Pre-Processing for ReGreen Rate Calculation

As explained in Chapter 1, while Casady et al.'s method addressed "post-fire vegetation regeneration rate," their term has been renamed throughout this study as ReGreen Rate since this is a more accurate representation of what can be determined from the imagery. The ReGreen Rate is the slope of a regression fit line through the three annual sums of normalized EVI values for the three post-fire years. These kinds of calculations are not possible in ArcGIS and necessitated the use of a mathematical coding environment such as R.

The R environment is an open source platform where users make contributions in the form of packages that provide tailored functionality. While base R has functions for calculating common statistical results such as regression fit lines, it needs special packages for functions such as reading raster files or performing smoothing operations. All R code used in this present study was self-taught and may not represent the most efficient means of accomplishing the desired result. Figure 6 summarizes the processing steps undertaken in the ArcGIS and R environments, and a full rendering of all R code used in this present study is captured in Appendix A.

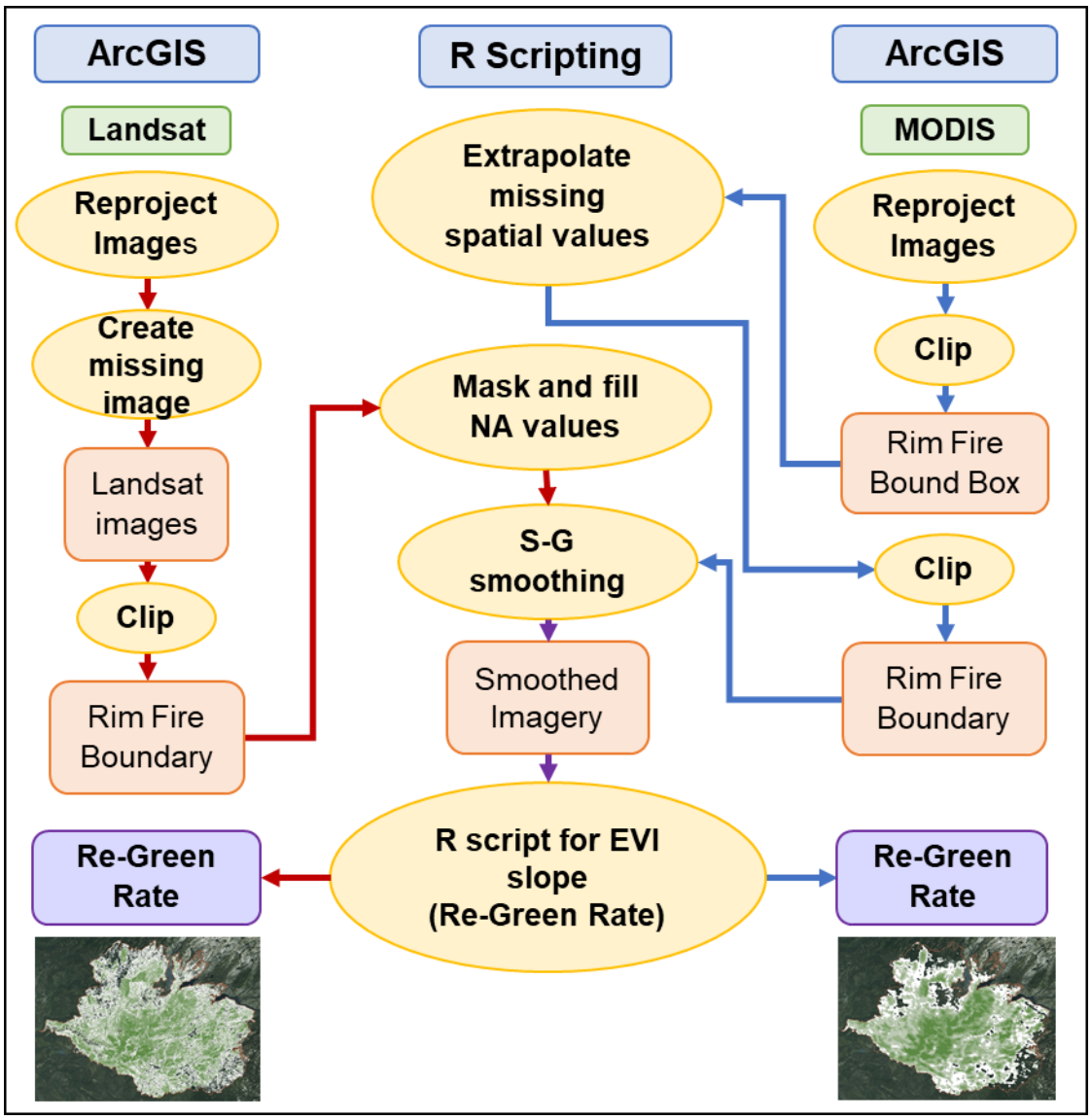

Figure 6 Data processing workflow for EVI conditioning and ReGreen Rate

Calculating the ReGreen Rate required a significant amount of data preparation. Anomalous values that required smoothing came from a variety of sources. The fire burned an area in Tuolumne County that had 65% of average rainfall for the 2012-2013 rain season. Given the drought conditions and the corresponding impact on EVI values, subsequent wet months created seasonal water bodies that influenced the data set and created outliers in individual or small groups of pixels. Additionally, other non-fire related conditions, such as clear-cut areas,

bare rock, clouds or shadows based on the angle of image capture also introduced anomalous values.

Additionally, there were values outside the range of typical cloudless images. Inspection of clear days from a representative set of images before and after the fire found that more than 99% of valid EVI values range between 0 and 6500. This range was therefore selected as the filtering range for both MODIS and Landsat. Pixels that were out of those bounds were given a value of NA. The resulting gaps and remaining errors in the data required processing to form a data set that is usable by the regression decision tree model as discussed in Section 3.3.

To calculate the ReGreen Rate, the anomalous values needed to be first filled and smoothed; then the data was normalized and summed into annual post-fire values. Processing steps for MODIS and Landsat imagery were generally similar but differed in the treatment of anomalous data. MODIS already had most missing or obscured pixels processed as part of the EVI product development as discussed in Section 3.1, but there were still a few missing pixels in each image which in aggregate created a lot of missing values when looking at all 150 images across the time series.

Landsat had many areas and several complete images obscured by clouds. Missing and anomalous EVI values result in over- or under-estimating the annual sum of normalized EVI values. Since the annual sum of normalized EVI values is used to calculate the regression line slope; the ReGreen Rate would be flawed. Therefore, special care was taken to ensure a full data set, with minimal changes to the data integrity. The following subsections describe the steps for processing each imagery set in greater detail.

3.2.1.1. MODIS Imagery EVI Processing

MODIS imagery was reprojected from sinusoidal to UTM NAD 1983 Zone 10N. The resultant pixels were resampled using the bilinear analysis method to form 240 m square pixels. The pixel size of 240 m was selected to ensure a whole number of 30 m pixels lay wholly within the reprojected MODIS-based 240 m imagery. A spatial resolution of 240 m was selected rather than 270 m or 210 m because 240 m is closest to the 250 m pixel size of the MODIS data at the equator. Having the MODIS and Landsat images line up precisely enabled analysis across a common reference frame. The data was then read into the R environment as a GeoTIFF for subsequent processing.

It is worth noting a few key structures used by the R environment to handle raster data. For some operations, a raster image can be converted into a single *vector*. The images are vectorized by starting at the pixel at the top left of the image extent and ending at the bottom right of that image as shown in Figure 7. A *raster stack* is a collection of images layered together, and when extracted into a matrix each column is a vectorized image.

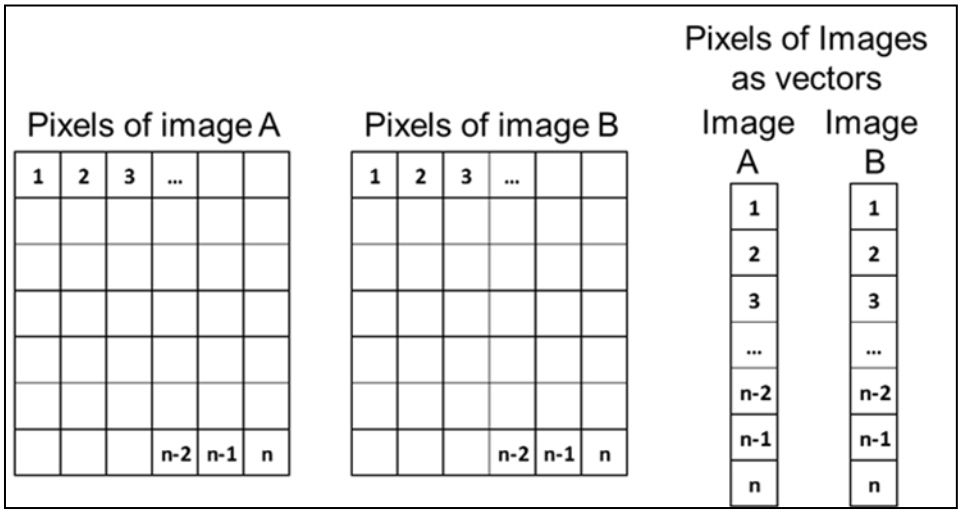

Figure 7 Vectorizing an image and forming a raster stack

A raster template was constructed from one of the MODIS images. The template contained all the structure (extent, rows, columns, the number of cells, projection, etc.) from the MODIS image and was filled with NA values. This template provided a shell that was used in later processing steps where it was necessary to convert images to vectors, process the vectors and then return the values to the raster again.

It was determined that the oblique MODIS image capture angle from the southeast created areas of NA values in the reprojected image. Figure 8 shows where these NA values occur in all 150 reprojected MODIS images. It can be seen that the preponderance of NA values occurs along the southeast faces of the canyon walls. Additionally, using the same range of valid EVI values as the Landsat images (greater than 0 and less than 6500) to mask the images, 6648 pixels were identified as less than zero and 32 pixels were identified as higher than 6500. Thus, the total percentage of missing pixels was 9.4% from the image capture and an additional 0.24% from out of range values within the fire boundary across the time series.

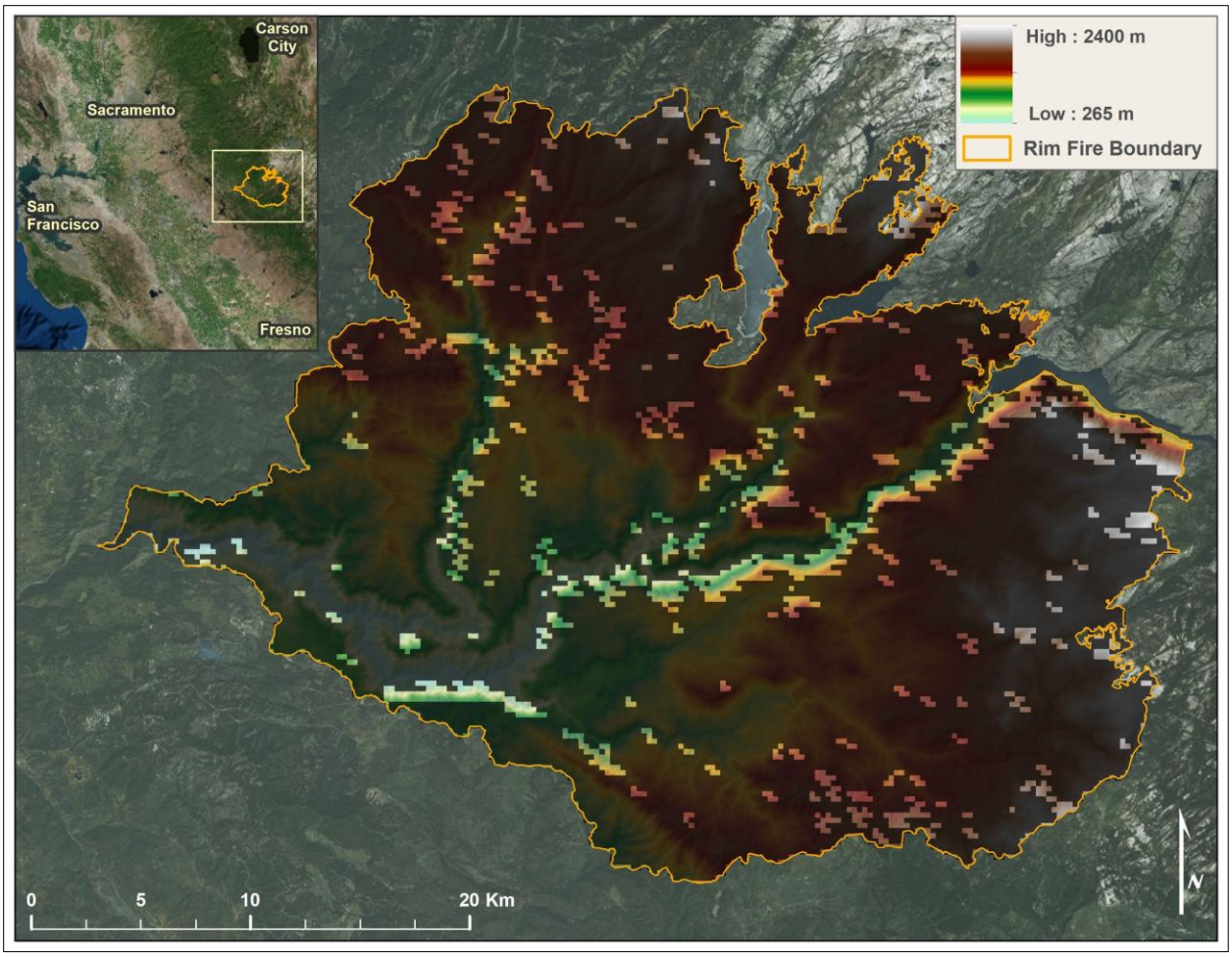

Figure 8 Image of all NA pixels in the MODIS time series

With the missing and erroneous data identified, the R function na.fill (from the *zoo* package) was used to fill missing values across the spatial extent. The na.fill function creates an interpolation from the last valid numbers in a vector on either side of a missing value by continuing the trend from those valid numbers. This step could have been accomplished in ArcGIS, but doing so over 150 images would have been very labor intensive. This na.fill technique allows for horizontal averaging to fill in the missing values, which was deemed sufficient given the large number of pixels and images involved. Code Chunk 1 specifies how this was applied to the vectorized raster stack.

```
for(i in 1:number.of.images) {
temp.vector.column <- as.vector(values.of.stack[,i])
if(!all(is.na(temp.vector.column)))
{
temp.vector.column <- na.fill(temp.vector.column, "extend")
values.of.stack[,i] <- temp.vector.column
}
next
}
```

Code 1. R code to fill NA pixels in the MODIS time series

In the code above, the `values.of.stack` variable is a matrix formed from all 150 images in the time series with each vectorized image as a column in the matrix. When the code extracts `values.of.stack[,i]` (i to the RIGHT of the comma in the column position and selecting all rows in that column) it is selecting each of the images and applying the na.fill function. An effect of this process is that the NA values outside the fire boundary are sequentially filled with an interpolated value using the last pixel in the vector with a value and the next pixel inside the fire boundary with a value. To remove these spurious values, the mask function from the R *raster* package was used to clip the data back to the fire edge.

To apply the mask, the filled values of the stack were transformed back into a raster stack, the mask function used, then the values were extracted back into the `values.of.stack` variable for the subsequent step. Code Chunk 2 shows the R code for applying the mask.

```
EVI_Mask[] <- values.of.stack

#applying the MODIS mask returns the values of the stack to
#the fire boundary
EVI_Mask <- mask(EVI_Mask,MODIS_Mask)

#and now the NA values have been spatially filled in the
#areas outside the fire boundary
values.of.stack <- getValues(EVI_Mask)
```

Code 2. Clipping the filled values to the fire boundary

Next, the images were smoothed across the time domain using the Savitzky-Golay (SG) smoothing filter (sgolayfilt function in R from the *signal* package) using the code shown in Code Chunk 3.

```
for(i in 1:NROW(values.of.stack)) {
 temp.vector.row <- as.vector(values.of.stack[i,])
 smoothed.row <-
sgolayfilt(na.pass(temp.vector.row),p=7,n=9,m=0)
 values.of.stack[i,] <- smoothed.row
 }
```

Code 3. R code of the temporal smoothing using Savitzky-Golay

Now the `values.of.stack[i,]` (i to the LEFT of the comma in the row position and selecting all columns in that row) are extracted to create a vector out of each row of the matrix. The ReGreen Rate is across the time domain and is calculated for each pixel and is therefore independent of the spatial domain. Smoothing across the time domain ensures that the EVI values used in calculating the ReGreen Rate have anomalous features smoothed out. The S-G method works by creating a localized polynomial (typically at least 4th order) function around a given point in a time series by looking at sets of points to the left and right in the vector (i.e. before and after) and adjusting the given point to fit that curve. The S-G typically involves left and right points numbered in the dozens. Given the paucity of annual data points (23) and the lack of information regarding how many points Casady et al. used on the left and right of the sliding average in the smoothing process, this author was reluctant to use this method. However, given the high number of pixels involved in the study and the existence of clouds in the data and other anomalies, smoothing was determined to be necessary.

Code Chunk 3 takes single pixels and applies the SG smoothing temporally across the time series. Different polynomials from 3 through 11 were experimented with, and a 7^{th} order

polynomial (p=7 in the code) was found to be most effective at reducing extreme maximum and minimum values while retaining the same basic statistics for mean and first and third quartiles. The intent behind the S-G smoothing is to maintain the same basic structure of the time series EVI values while reducing the effect of extraneous values caused by clouds.

The next step was to average individual pixel values across all the pre-fire images and use that pixel average to normalize all the pixel values. The result of this final step is illustrated in Figure 9 which compares representative pre-fire EVI values (left side) with the filled, smoothed and normalized EVI (right side). The colors follow Low pre-fire EVI values= Red/Orange, Mid-pre-fire EVI values= Green/Cyan, High pre-fire EVI values = Blue/Purple. The fire is easy to identify by the sudden drop in values just before 2014 and is marked by the text "Rim Fire." The gradually increasing recovery for the mid- and high-pre-fire EVI values can also be seen.

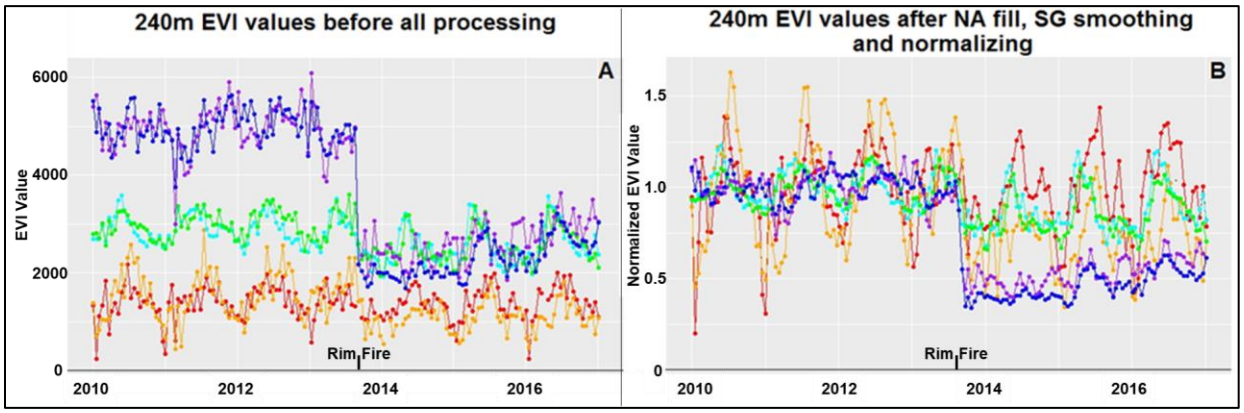

Figure 9 Comparison of individual MODIS EVI pixel values across the time series before processing and after the process fill, smooth, and normalization

3.2.1.2. Landsat Imagery EVI Processing

In general, the processing of the Landsat images proceeded in a similar manner to that described above for the MODIS images, using similar R code. Unlike MODIS, the Landsat images did not have any missing values due to capture angle or high reflectance. However, the

Landsat images had a high frequency of clouds that required additional processing steps described in this section.

In addition to the missing image from 10 February 2015, three additional images were completely covered in clouds with meaningless EVI values, and other clouds created sharp negative spikes in the time series of particular pixels. An examination of histograms of EVI values for pre- and post-fire clear days as compared to cloud-obscured images, shown in Figure 10, revealed why it was important to set bounds on which EVI values represent reflectance values from the surface and those from clouds. The histograms of pre-fire cloudless days show a range of 0-6500 with the horizontal axis representing EVI values in an image and the vertical axis representing the count of pixels in each bin. This examination of histogram values from clear and obscured images provided confidence that an EVI value range of 0-6500 would include the valid pixels in any given image. Figure 11 identifies that the range of 0-6500 remains valid in clear images throughout the years after the fire. Therefore, values outside this range were set to NA and filled as described in the treatment of MODIS EVI values in Section 3.2.1.1.

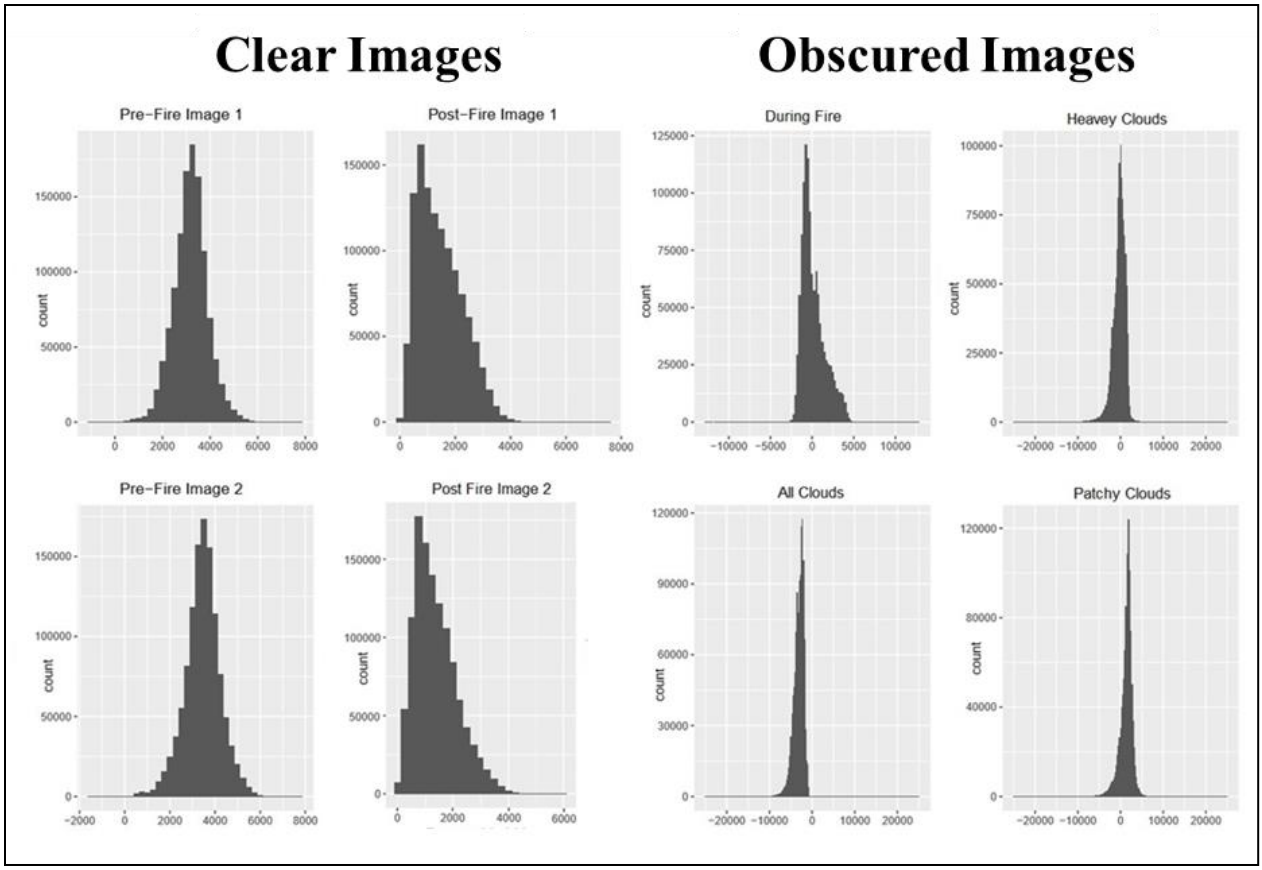

Figure 10 Comparing EVI values of clear vs. obscured images

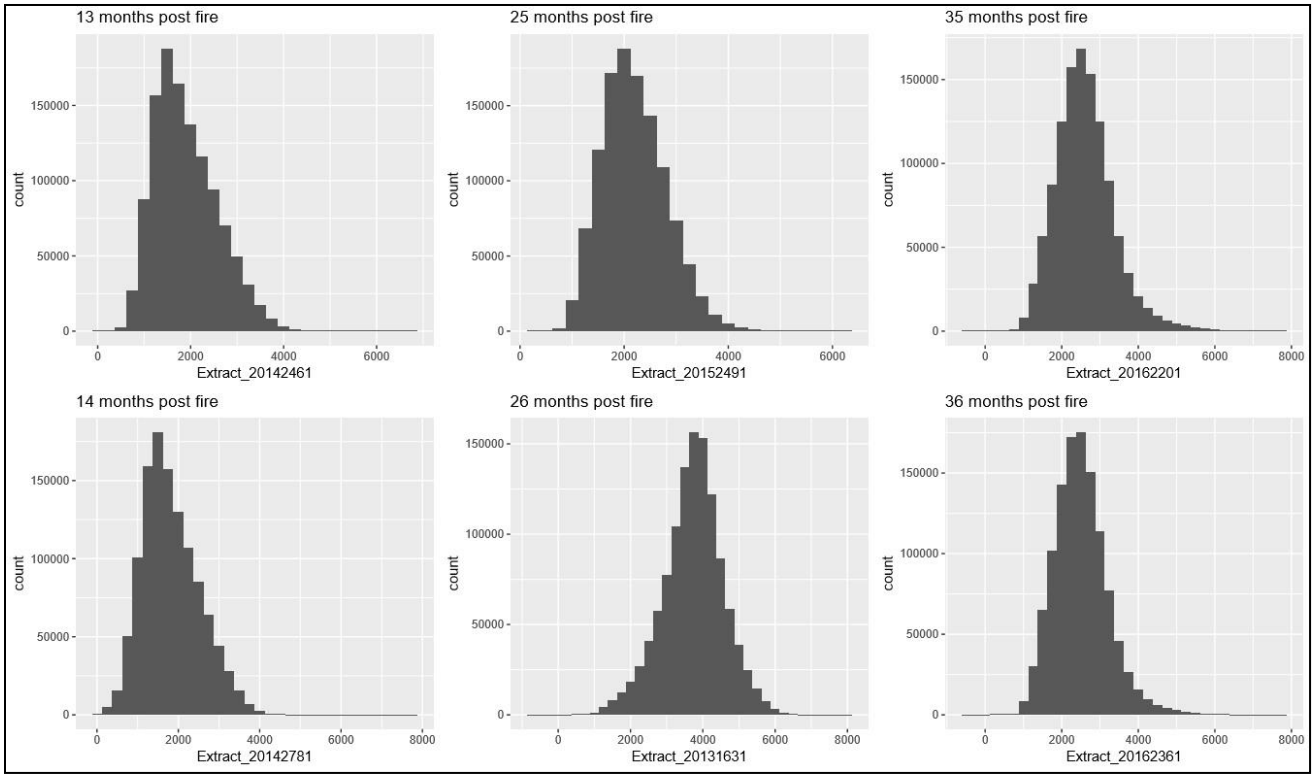

Figure 11 Comparing clear Post-fire images

The Landsat raster stack consisted of 89,017,390 EVI values; of which 62,575 had a greater than 6500 EVI value and 4,211,269 had a less than 0 EVI value, indicating that 4.8% of pixels needed adjustment across the 77 images. The clouds in the Landsat images added significantly to the coding challenges and to the veracity of the information. With almost 5% of the data corrupted (albeit less than the missing values from MODIS), each step was treated very deliberately to preserve the data integrity as much as possible. The cloud pixels could not be discarded either. Holes (NA values) in the time series of any given pixel would have introduced undercounting over the time series when creating the sum across the normalized EVI anniversary years and thereby affected the slope, that is the ReGreen Rate. No pixel in the temporal extent was unaffected by clouds, and so all images required the same treatment.

Looking at the cloud-obscured pixels over the time series, five pixels were selected as representative of different types of behavior. Figure 12 shows two that were selected as having

the most frequent occurrence of EVI values less than 0 or greater than 6500 (blue and green lines), another two that were selected for having the least frequent number of EVI values out of range (red and orange lines), and the black line shows the pixel with the most consistent value over the time series. The deep downward spikes at image 19 and close to 65 are days with complete or almost complete cloud cover. However, a winter storm in early 2015 (near the 40th image sequence number) created a set of images that consisted of high, but mostly within range EVI values that were in no way a reflection of the ground EVI conditions.

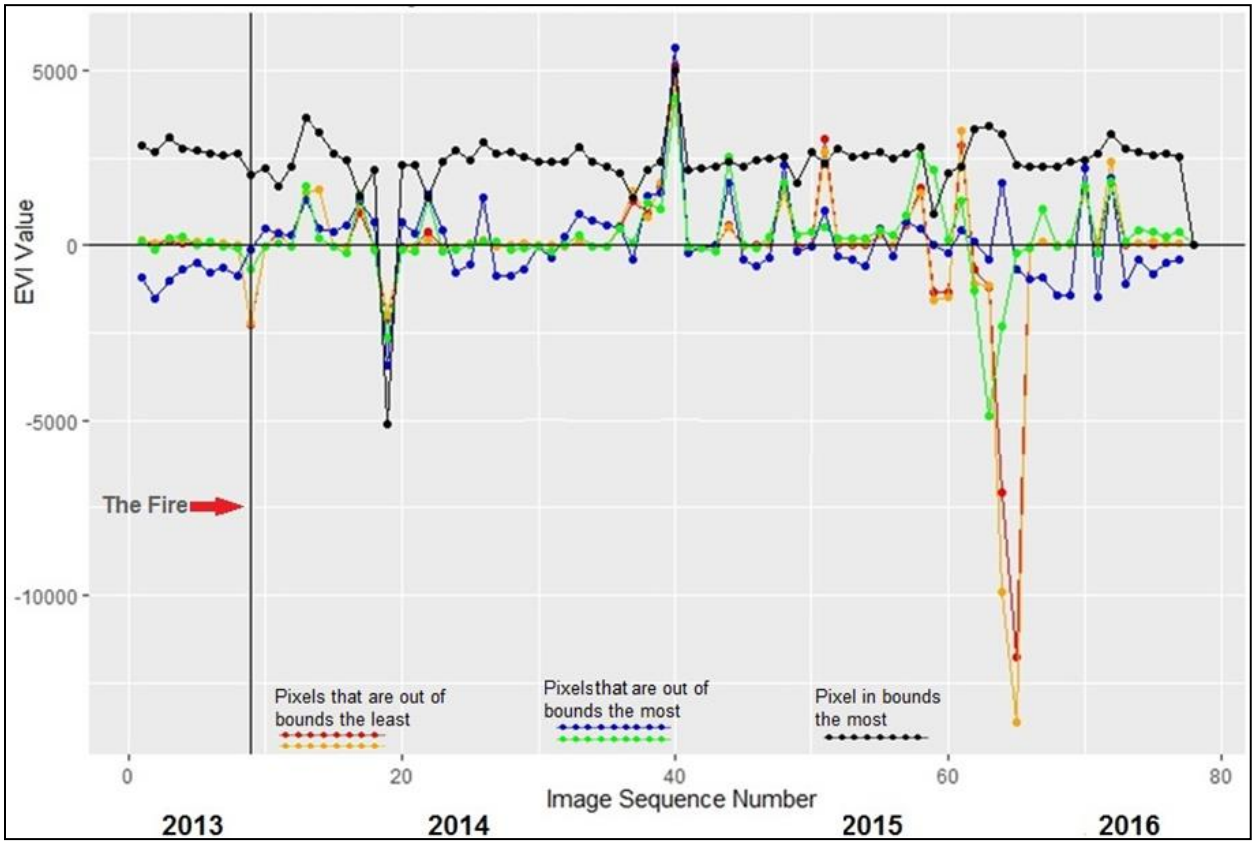

Figure 12 Examination of pairs of pixels that are most and least out of bounds and a single pixel that is the most in bounds

To assess the impact of the filling and smoothing steps, Figure 13 shows a similar set of pixels tracked through each step in the processing. On the top graph, holes appear where the 2015 winter cloud images were removed. Once these winter 2015 cloud images were eliminated

from the time series, the search and selection of the most out-of-range and least out-of-range pixels produced a different set of pixels than in the figure above. However, they do continue to represent the behavioral cases. The third image down shows that the S-G smoothing process re-introduced negative EVI values. Those values were set to NA and again filled using the na.fill function.

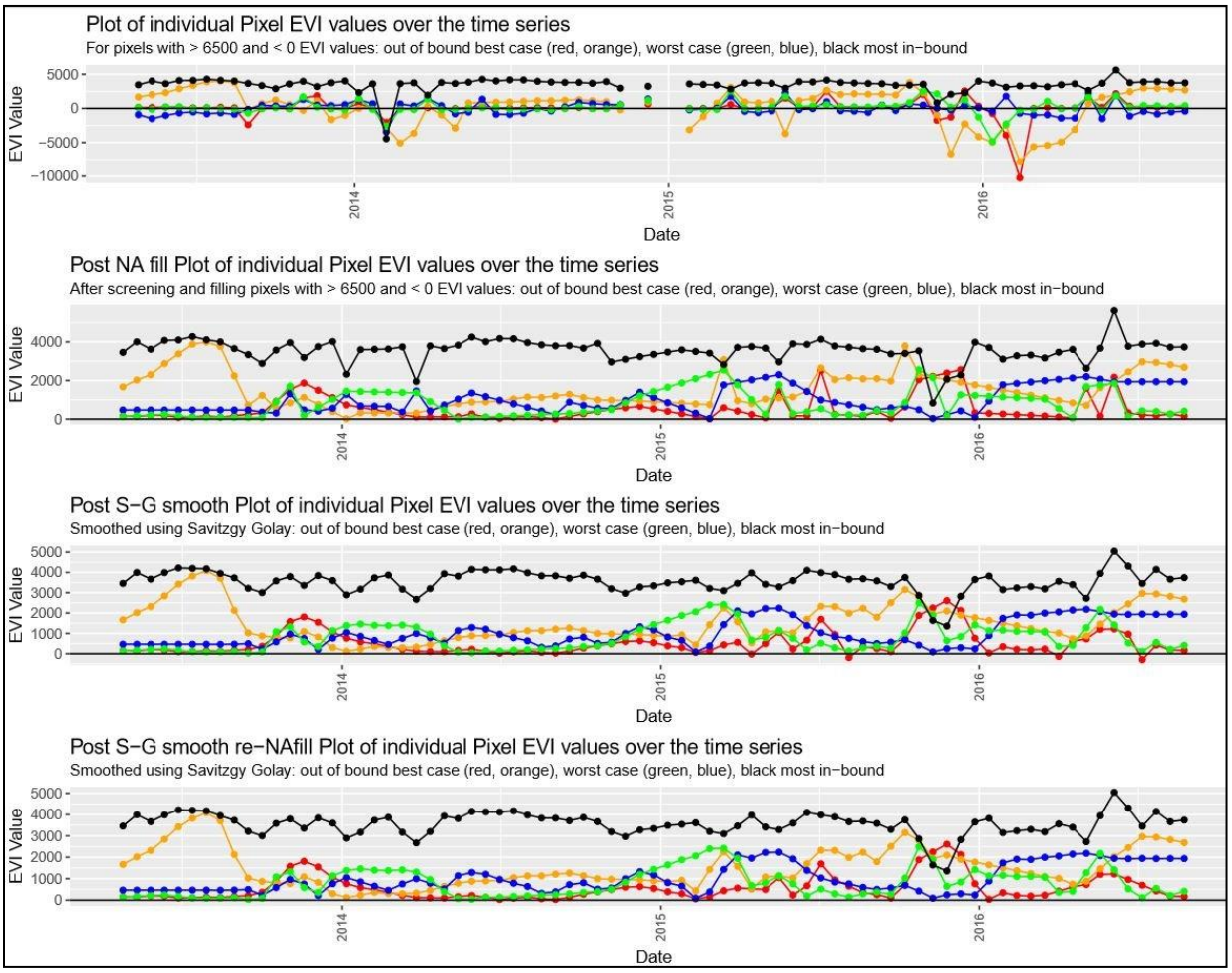

Figure 13 Tracking five pixels through the processing steps

3.2.2. ReGreen Rate Calculations

Figure 14 illustrates how the ReGreen Rate is calculated. The solid gray line is the processed and normalized EVI values of a single pixel across the time series and uses the left vertical axis for scale. The black triangle values are the annual sum of normalized EVI values

collected from each post-fire year; they use the right vertical axis scale. The dashed line is the linear best fit for the annual sum across the three post-fire years. The figure also shows the resulting equation describing that best fit line as well as the ReGreen Rate derived from the slope and the R^2 for the line.

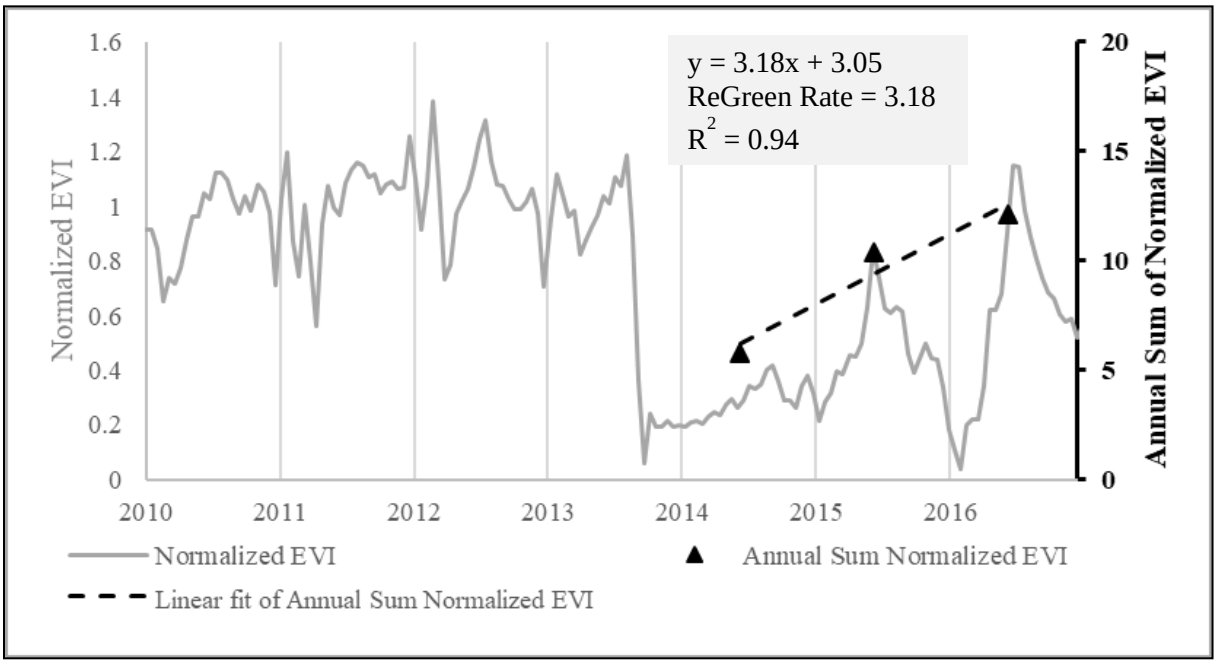

Figure 14 Illustration of the ReGreen Rate as a regression slope

Following the processing of MODIS and Landsat data, calculation of the ReGreen Rate is shown in Code Chunk 4. A sum of the normalized EVI values from each post-fire annum is used to create a matrix called Postfire.reveg that contains the data points needed in calculating the ReGreen Rate slope. The three columns in the matrix are the yearly cumulative EVI values for each fire annum, and the rows are the pixel values within an image extent (both inside the fire boundary with values and outside the fire boundary where the values are NA). Using the second vector called the years.gone.by, the linear regression function (lm) from the base R package was used to determine the slope at each pixel for a best-fit line through the three post-fire annual

cumulative EVI points. In keeping with the Casady et al. example, slope values less than zero were assigned as NA. The tempslope variable contains the final ReGreen Rate pixel values.

```
years.gone.by <- as.vector(c(1,2,3))
#where 1 is the first year post-fire, 2 is the second and 3
#is the third.

Postfire.reveg[is.na(Postfire.reveg)] <- 0
#should set NA values to 0 for the lm function to work
#properly

tempslope <- vector(length=NROW(Postfire.reveg))
#is the template into which each slope value is placed

for(i in 1:NROW(Postfire.reveg)){
 temprow <- as.vector(Postfire.reveg[i,])
 tempfit <- lm(temprow ~ years.gone.by)
#regression analysis of the relationship between year and
#the cumulative annual normalized EVI values for each pixel

 tempslope[i] <- tempfit$coefficients[2]
#this returns the slope of the regression line to the
#template
 }

tempslope[tempslope%in%(0)] <- NA
#A check revealed the no calculated slope values equaled
#zero so reapply the NA values to outside the fire
#boundary.
```

Code 4. Calculation of ReGreen Rate

Calculating slope values in Code Chunk 4 takes about 20 seconds with the MODIS images. In generating the slope coefficient, the lm function also generates information about the residuals, that is the difference between the observed points and the expected values based on the linear regression. Figure 15 depicts the distribution of those residual values from the ReGreen Rate calculation for the MODIS imagery. While annual cumulative normalized EVI values typically range from 5 to 20, most residual values were between -1 and 0 indicating an overall good fit between the regression line and the observed cumulative annual normalized EVI values.

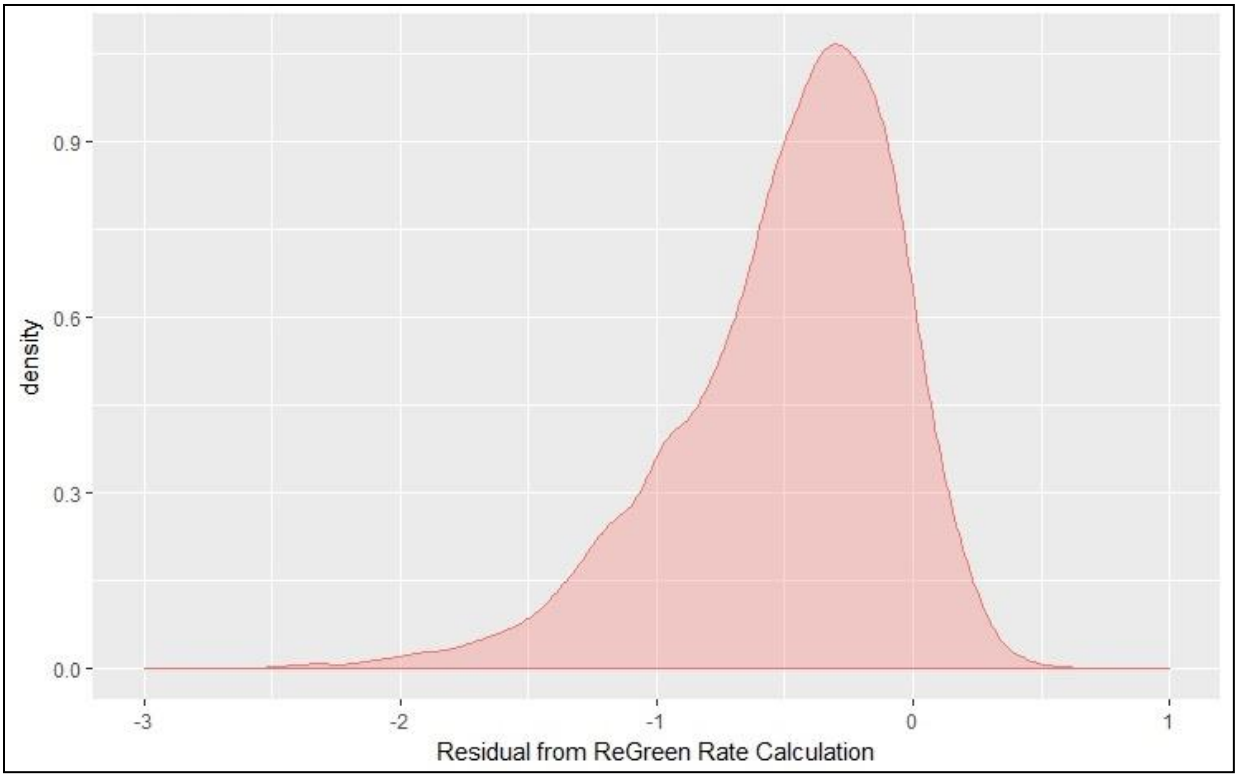

Figure 15 Density distribution of residuals from ReGreen Rate calculations for 240m data

Figure 16 shows a histogram of the ReGreen Rate of Landsat-based ReGreen Rate on a scale from 0-10. As per Casady et al. the less-than-zero values were replaced with NA values for both the Landsat and MODIS derived ReGreen Rates. Looking at only the positive ReGreen Rates supported looking at the factors associated with fire recovery rather than degradation. Elimination of negative ReGreen Rate values removed ~12.5% of the study area.

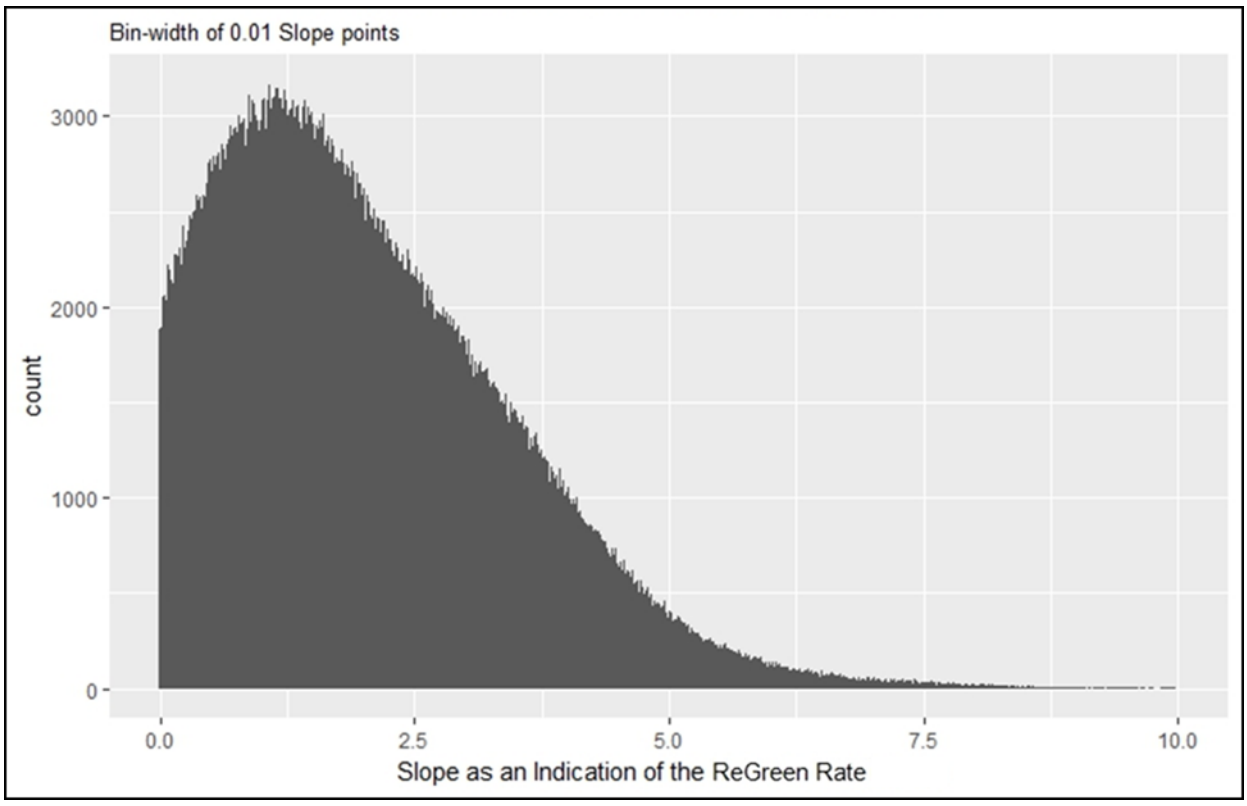

Figure 16 Histogram of Landsat ReGreen Rate values in the 0-10 range

Most of the remaining ReGreen Rate values fell within the range of 0 to 10. The pixels depicted in Figure 17 are from Landsat-based ReGreen Rates values greater than 10. A visual inspection shows that these values fell on water features such as ponds or river beds, and heavily shadowed areas on rocky outcroppings. Given the conditions at these pixels, and the errors they would introduce, pixels with values greater than 10, representing an additional 0.015% of available 30 m pixels, were also removed from the ReGreen Rates used for model development.

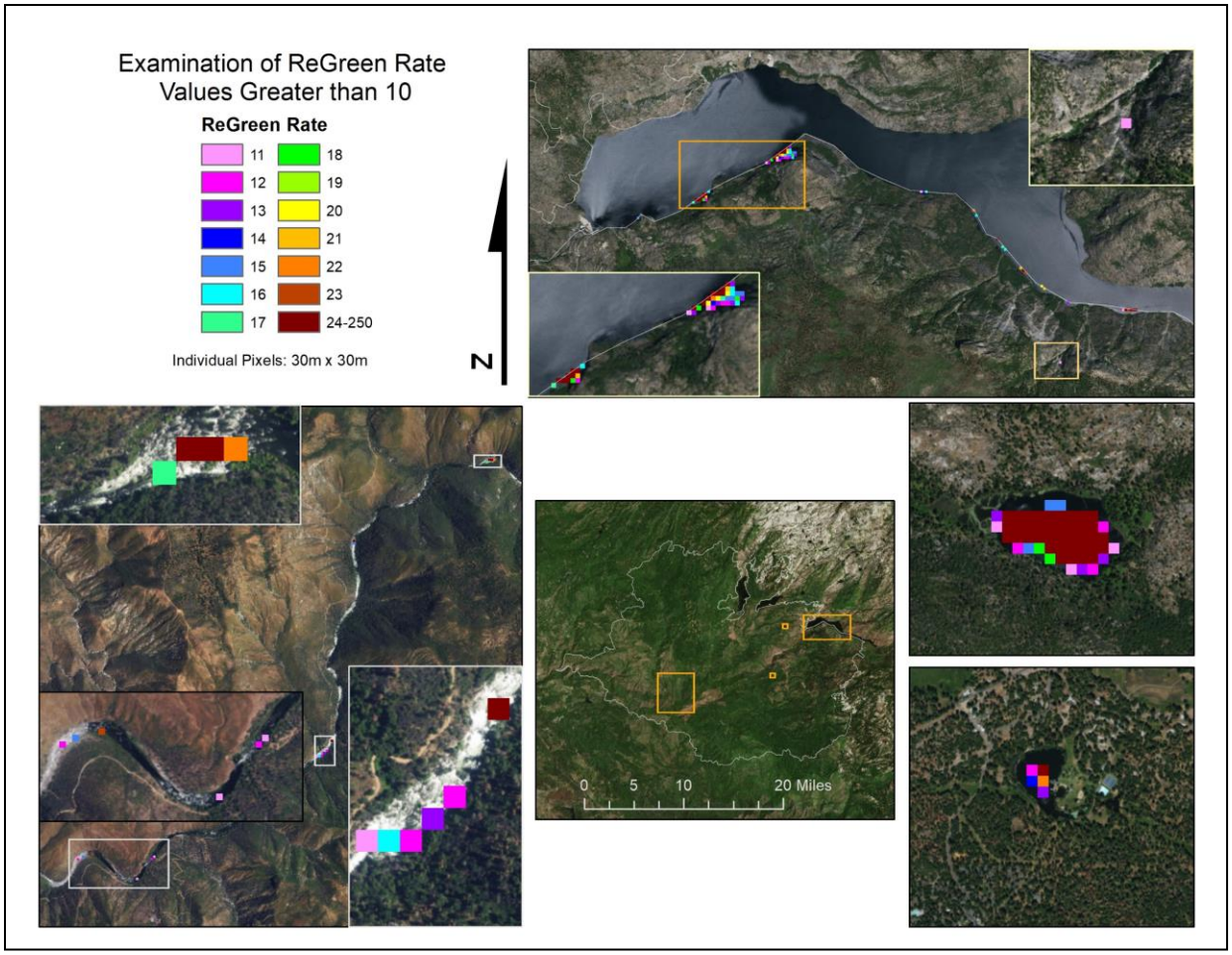

Figure 17 Examination of ReGreen Rate values greater than 10

Given that the Landsat images consist of 64 times as many pixels in the extent than the MODIS images, processing steps took substantially longer. Processing steps with MODIS images that took on the order of tens of seconds took tens of minutes for Landsat images. The lengthy computational time could likely be mitigated with more efficient coding and use of parallel processing, but optimizing the processing procedure is outside the scope of this present study.

As the final step, the tempslope vector was used to fill the values of the raster template, and the result was saved as a GeoTIFF raster. This produced a raster of ReGreen Rates that could be examined in ArcGIS and used in the construction of the regression tree model.

3.2.3. NBR Calculations

Figure 18 captures the processing steps used in creating the adNBR and RdNBR fire severity index values. The RdNBR was only produced from MODIS data to enable comparison between the regression trees built using the RdNBR and adNBR at the 240 m spatial resolution. The MIR bands for the pre- and post-fire images were originally captured at a 500 m spatial resolution, but when resampled at the 240 m resolution to match the EVI MODIS image, the values aligned with the 240 m NIR pixels. Unfortunately, the resampling process introduced 8 NA values in the MIR pre-fire image and the NIR post-fire image already had 2 NA values. After using the same method as was used with the EVI values to fill the NA values, the pre- and post-fire NBR values were calculated using the R code shown in Code Chunk 5:

```
NBR_pre <- (values.PreNIR –
 PreMIR.nafill)/(values.PreNIR +
 PreMIR.nafill)*1000
#note that the multiple of 1000 is by convention for
#calculating NBR to put them into integer form.
```

Code 5. Calculating the pre-fire NBR

All NBR layers were masked using the MODIS_Mask to ensure the overfill areas were clipped out of the layers.

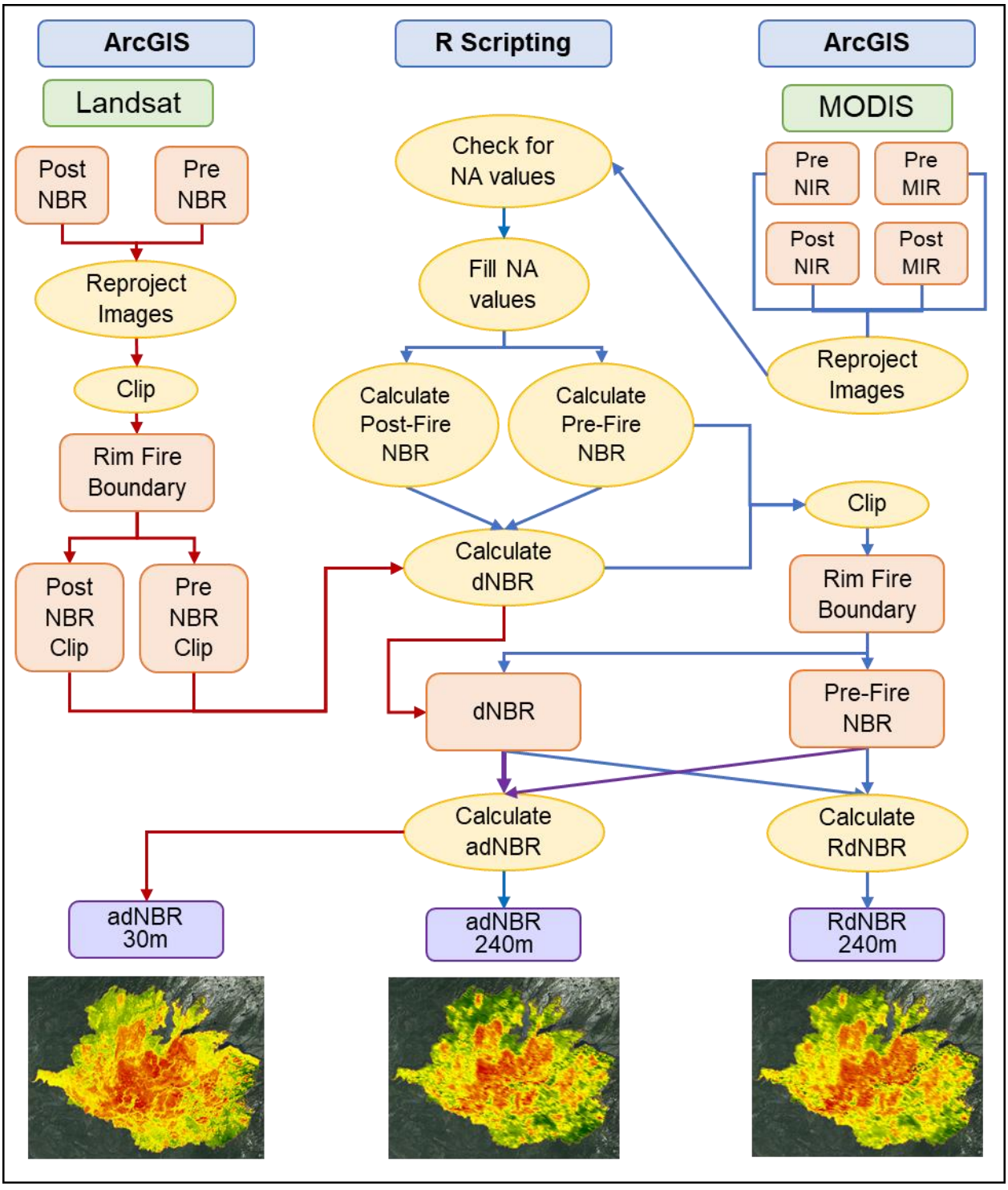

Figure 18 Process flow diagram for producing adNBR and RdNBR fire severity index values

3.2.3.1. dNBR Calculations

The pre-fire NBR values were subtracted from the post-fire NBR value to create the dNBR. The effect is that a negative post-fire NBR value (indicating deep burn) will be added to the pre-fire NBR value creating a larger value and small NBR values (indicating a significant burn) will be subtracted from the pre-fire NBR resulting in a large dNBR. The dNBR was calculated to support the calculation of the adNBR and not used as a factor in the regression tree model.

3.2.3.2. adNBR Calculations

The adjusted difference normalized burn ratio (adNBR) as described by Cassady et al. uses a linear best fit line from a plot of pre-fire NBR and the dNBR to create a predicted dNBR value for each pre-fire NBR value. The adNBR is the difference between the observed dNBR and the predicted dNBR. Note, it is important to remove the pixels from outside the fire boundary, so they do not influence the relationship between the dNBR and the pre-fire NBR. Figures 19 and 20 show the relationship and coefficients used in creating the predicted dNBR values for MODIS and Landsat pixels with the resulting linear regression equation included. Positive adNBR values reflect dNBR values that were greater than the linear model would expect for a given pre-fire NBR value. The converse goes for negative adNBR values.

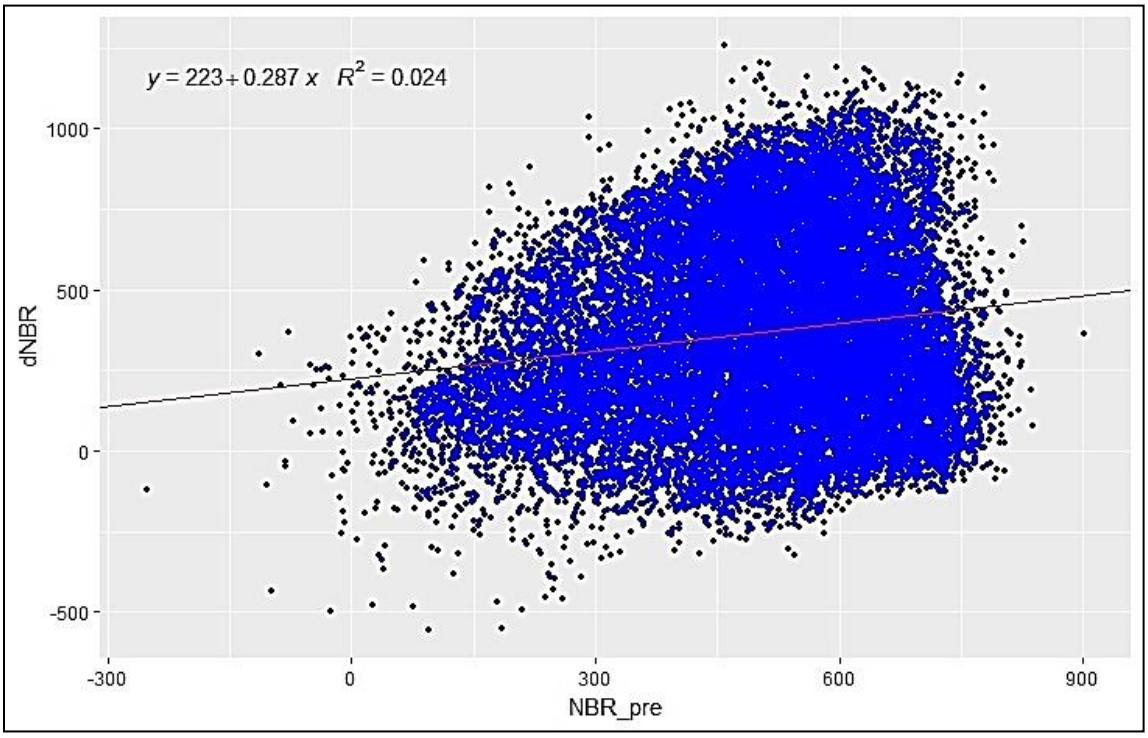

Figure 19 Visualizing the adNBR for MODIS

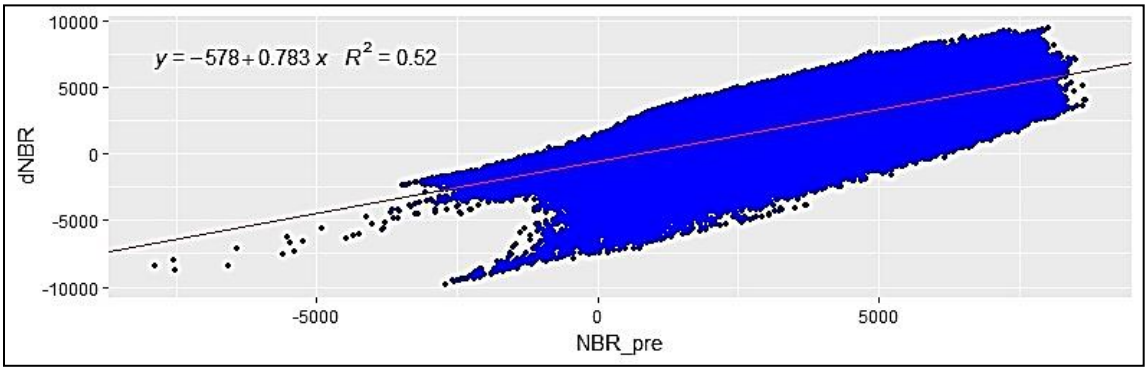

Figure 20 Visualizing the adNBR for Landsat

The coefficients of the linear regression equations shown in Figure 19 and Figure 20 were used to calculate the adNBR by the code described in Code Chunk 6.

```
pre.dNBR.relation <- lm(dNBR ~ NBR_pre,
na.action=na.exclude)
adNBRCoeff <- coefficients(pre.dNBR.relation,1:2)
#use these to populate the expected dNBR function used in
#calculating adNBR

exptdNBR <- adNBRCoeff[1]+adNBRCoeff[2]*NBR_pre

#Construct the adNBR vector
adNBR <- as.vector(dNBR-exptdNBR)
```

Code 6. Finding the linear regression line between dNBR and pre-fire NBR to determine the adNBR

3.2.3.3. RdNBR Calculations

Casady et al. were unable to use the RdNBR as a fire severity parameter due to the model failing to converge. After examining the distribution of RdNBR values using the raw pre-fire NBR values, it is evident that the pre-fire NBR values between -1 and +1 created 3 extremely high and indefinite values. These extremes caused problems with the model construction. Code Chunk 7 shows how those near zero values were identified and set to 1 before calculating the RdNBR.

An inspection of the distribution of RdNBR values revealed that over 99% of pre-fire RdNBR values reside between -500 and 1500. Code Chunk 7 also sets those values greater than 1500 or less than -500 to NA then smoothed using the na.fill function. These outliers can create errors in the model and are removed to facilitate building the regression decision tree model.

```
near.zero <- which(NBR_pre<1 & NBR_pre>-1,arr.ind = TRUE)
#this identifies values that greatly increase the RdNBR
value when calculating RdNBR as the RdNBR calculation
divides by the pre-fire NBR values.

NBR_pre[NBR_pre%in%(NBR_pre[near.zero])] <- 1

RdNBR <- dNBR/sqrt(abs(NBR_pre/1000))
out.of.range <- which(RdNBR>1500 | RdNBR< (-500))
# identifies pixels that are more than 1500 and less than -
# 500.

ModRdNBR[ModRdNBR%in%(ModRdNBR[out.of.range])] <- NA
# sets those out of range values to NA
```

Code 7. Processing NBR values for use in determining the RdNBR

3.2.4. DEM Processing

The DEM data originally had a 30 m spatial resolution and was projected to match the Landsat EVI images such that all derived products (elevation, flow accumulation, and aspect) retained their alignment with the Landsat EVI layer. However, all derivative products required resampling to 240 m spatial resolution to align with the MODIS EVI layer: elevation used bilinear resampling; aspect and flow accumulation used majority resampling. Bi-linear resampling returns an interpolation within the range of the original data using a weighted distance average for the four nearest input cell centers. Majority resampling finds the 4x4 input cells closest to the center of the output cell and takes the majority as the value. Bi-linear resampling is well suited for continuous data such as elevation. Aspect and flow accumulation are more like discrete information for which majority resampling is well suited.

3.2.4.1. Elevation

All digital elevation model processing, shown in Figure 21, was completed in ArcGIS. The first step was to mosaic together the four DEM tiles into a single layer, project it into NAD 83 UTM Zone 10N, align it with the Landsat EVI, and clip it to the fire boundary. The raster

resulting from this step became the raster template used to ensure the other layers were all aligned.

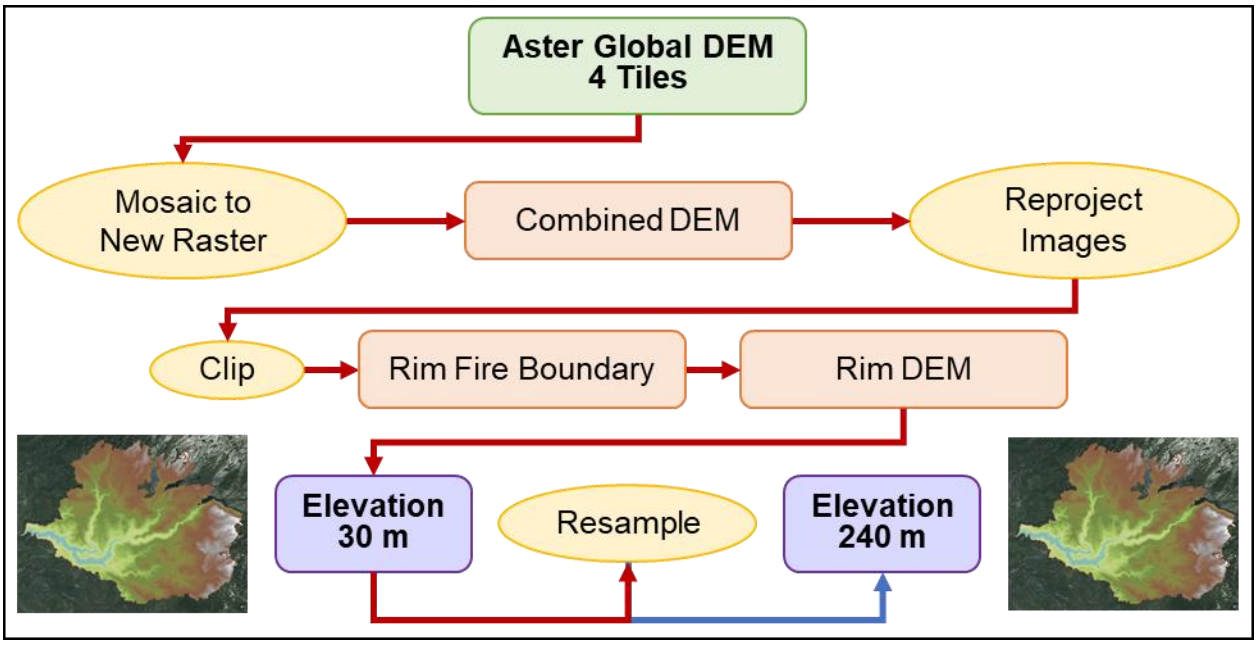

Figure 21 Elevation processing steps

3.2.4.2. Flow Accumulation

Figure 22 shows the steps involved in calculating flow accumulation at each pixel. First, the flow direction tool is used with the elevation layer to determine the flow direction for each pixel. The flow accumulation tool then uses flow direction to calculate the cumulative number of pixels that flow into a pixel. The 240 m pixel flow accumulation range was 0-17,721 pixels, and the 30 m pixel flow accumulation range was 0-124,734 pixels.

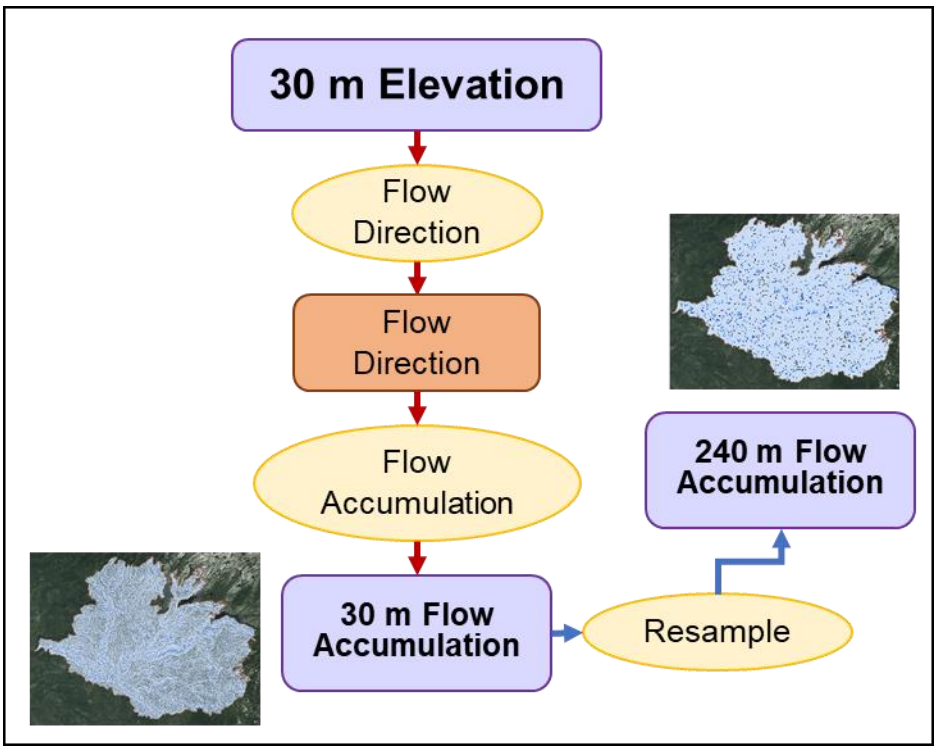

Figure 22 Flow accumulation processing steps

3.2.4.3. Aspect

Aspect was calculated from the elevation layer using the aspect tool as shown in Figure 23. Aspect is given in degrees, requiring a conversion to radians to calculate the Cosine (South = -1, North = 1) and Sine (West = -1, East = 1) of the aspect as needed in the model.

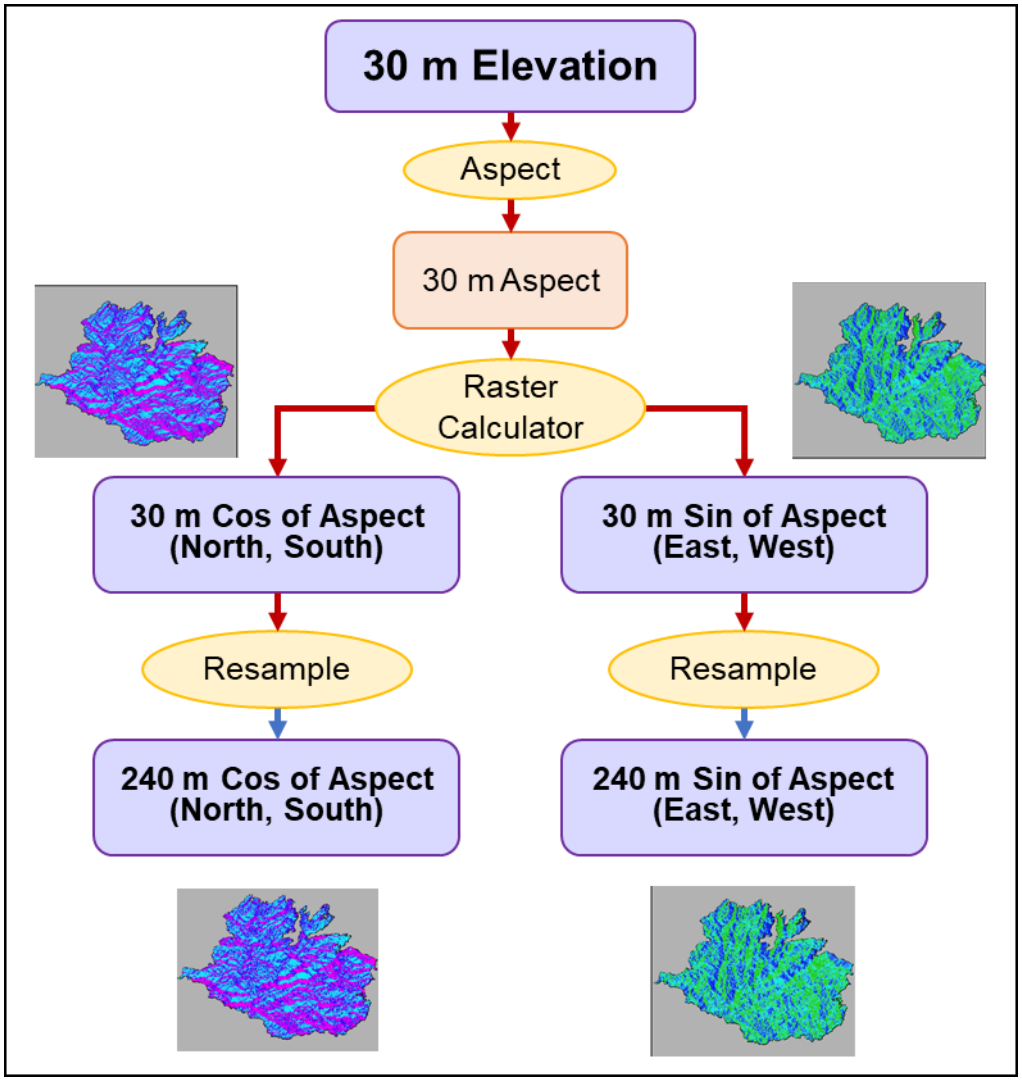

Figure 23 Aspect processing steps

3.2.5. Soil Data Processing

ArcGIS was used to merge and rasterize the data from the three contiguous soil vector datasets to create a single 30 m raster layer based on the taxonomic soil types as shown in Figure 24. The 30 m raster layer was resampled using majority aggregation into 240 m pixels.

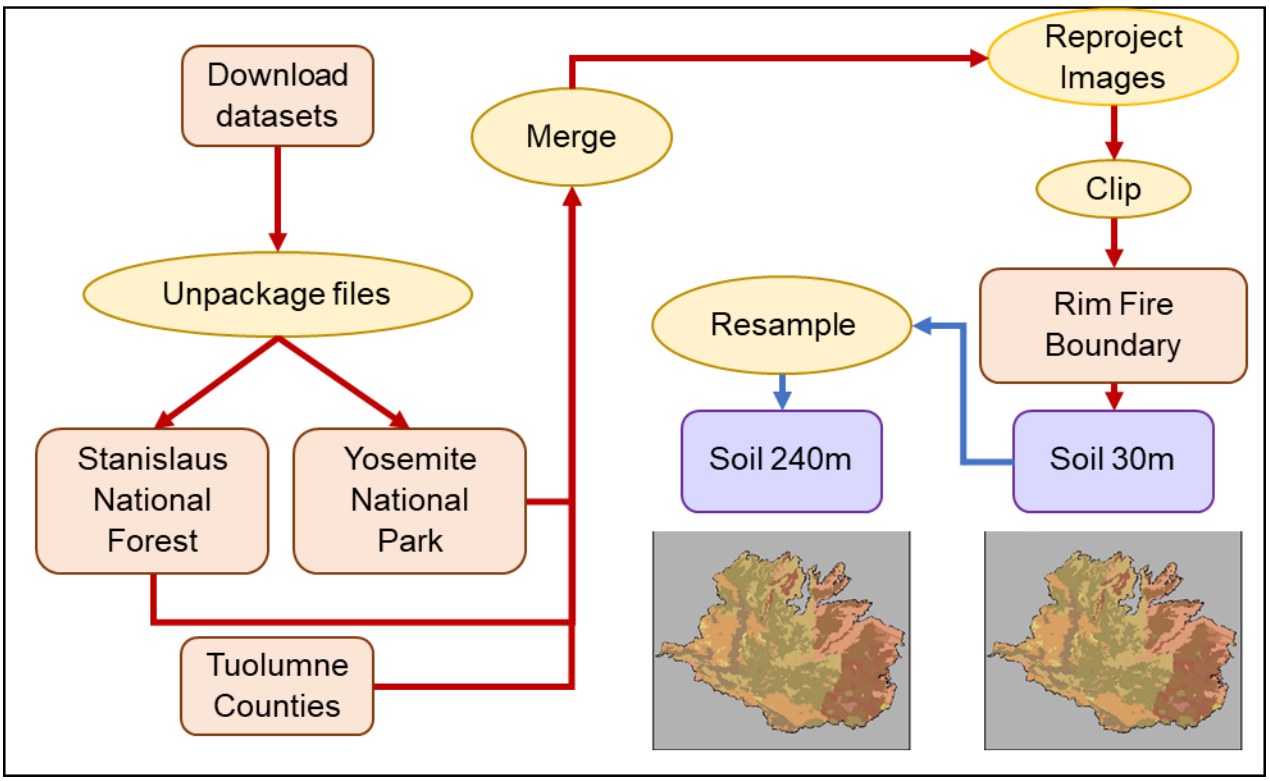

Figure 24 Soil data processing steps

The initial use of the most granular classification (164 different types in the taxonomic sub-group) produced an illegible regression decision tree as large groupings of very different soil types formed nodes on the tree. For easier interpretation and better utility, this present study used a broader taxonomic level (great group) to capture 12 soil types. An additional soil type attribute that may be useful in subsequent research would be to use the geomorphic description (e.g., alluvial flats, mountains, depression, mountain slopes, etc.) as a predictive attribute in modeling.

3.2.6. Vegetation Data Processing

A summary of the basic processing steps for vegetation is shown in Figure 25. CAL FIRE vegetation data was already in a raster format, and so was clipped to the fire boundary, then reprojected from California Teale Albers NAD83 projection to UTM NAD 1983 Zone 10N and registered to the Landsat EVI layers to ensure consistency with all the other layers. The 30 m

raster was resampled using a majority aggregation, which is suitable for categoric or discrete data, to create the 240 m raster needed for the MODIS model.

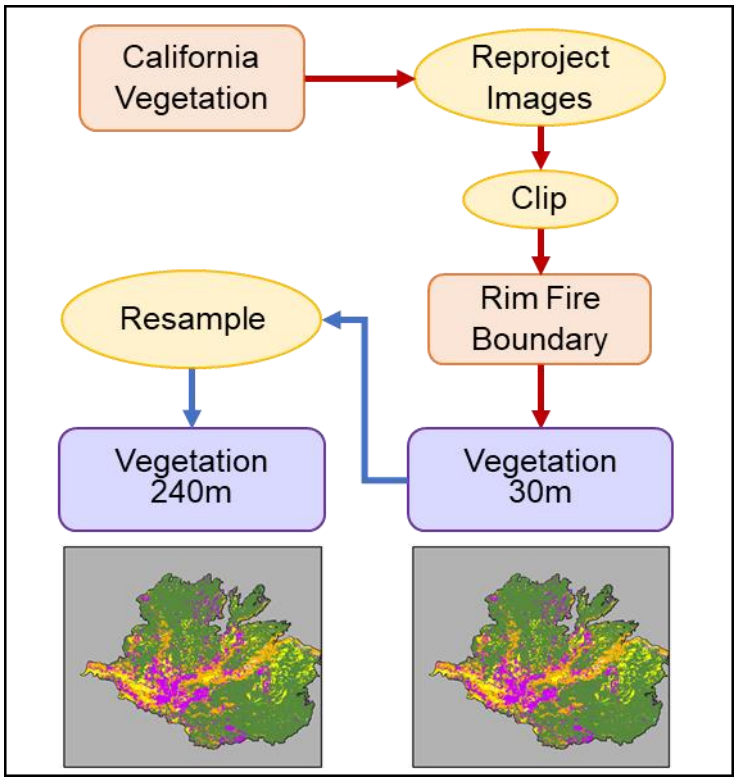

Figure 25 Vegetation data processing steps

3.3. Model Construction

The R environment allows the use of a wide range of packages to provide broad and complex functionality. Installing and accessing the large number of tools needed for advanced analyses such as decision tree modeling can be overwhelming. Fortunately, the R package *Rattle* was developed by Dr. Graham Williams as a tool for aspiring and bonified data scientists through his company Togaware (Williams 2009). *Rattle* provides an access point to 40+ R packages that are useful in data manipulation and processing through a single graphical user interface (GUI). Through a set of tabs across the top of the GUI, a user can load the data, build a model, and evaluate the model. It allows access to basic machine learning tools such as

regression decision trees as well as plotting and evaluation tools. This GUI provided an easy way to access all the tools needed to build the model and explore the results.

Using the tabs along the top of the tool as separate work processes, *Rattle* provides three elemental stages: data ingest, growing the regression decision tree, and trimming the regression decision tree. Once the model is constructed and trimmed, it can be applied to a different data frame (R's matrix data format) if the data structure (i.e. identical column names) remains the same as the data frame used in building the model.

For this present study, three models were constructed: a 30 m model using adNBR for fire severity, a 240 m model using adNBR for fire severity, and a 240 m model using RdNBR for fire severity. The models were constructed using the same method. The 30 m and 240 m models using adNBR for fire severity were used to answer the first study question: does the use of higher spatial resolution data create a more accurate regression tree model predicting the post-fire ReGreen Rate? The two 240 m models with the different fire severity methods were used to answer the second study question: do different indices of fire severity show a different result in model accuracy?

3.3.1. Data Ingest for Model Construction

The data ingest portion requires a single data frame from the R environment as input. All the ArcGIS data layers for elevation, aspect, flow accumulation, soil, vegetation, and the R environment-calculated values for ReGreen Rate and fire severity (adNBR and RdNBR) are individual raster layers. Reading those raster layers into the R environment and stacking them together into a raster stack then extracting the values into a matrix creates a single data frame. The columns of the resulting data frame are vectorizations of each data layer and the rows are the same pixel across each layer (see Chapter 3.2.1.1, Figure 5 for illustration).

With the data frame loaded into the *Rattle* tool, the first step in model construction was to establish the training and validating data sets (95% of data for training, 5% for validating) using random selection to construct the two data sets (note the default seed 42 was used – this value is possibly a reference to the Hitchhiker's Guide to the Galaxy (Adams 1979) as 42 is "the answer to life the universe and everything"). Using the *Rattle* tool, the ReGreen Rate was set as the target variable and the other factors (sin aspect, cos aspect, elevation, adNBR, RdNBR, flow accumulation, soil type, and vegetation type) as inputs.

3.3.2. Growing the Regression Decision Tree

The *Rattle* GUI tool uses the recursive partitioning for classification, regression and survival trees package (*rpart*) to construct the decision tree. The growth of the decision tree is controlled using a complexity parameter (CP). The CP sets the limit for incremental improvement (decrease) in the relative error. The relative error is the error rate computed on the training data at each number of splits (nsplit). When the user defined CP is met, the model stops splitting the data and growing the tree. For example, taking the initial relative error as 1; after the first split, the relative error is 0.55. Thus the CP for the first split is 0.45. After a second split, the relative error is 0.54, and the CP for the second split is 0.01. The tree continues to grow until the CP meets the user-defined threshold. As well as keeping track of the overall relative error from each split, *rpart* also uses an internal 10-fold cross validation error while building the tree.

3.3.3. Trimming the Regression Decision Tree

Two common methods for pruning the decision tree are: 1) prune at the point with a minimum cross-validation error (x-error) which indicates that further construction of the model decreases its generalizability; and, 2) prune at the point where the sum of the relative error and the standard deviation of the cross-validation error (std x-error) is less than the x-error (Simpson

2017). This is to help prevent overfitting and to improve the generalized applicability of the model. Using the second method for pruning, a plot of the cross-validation error at each of the nodal splits identifies the diminishing value of adding additional nodes (Figure 26) where the X indicates where the sum of the rel error and the std x-error is less than the x-error which corresponds with a CP = 0.005.

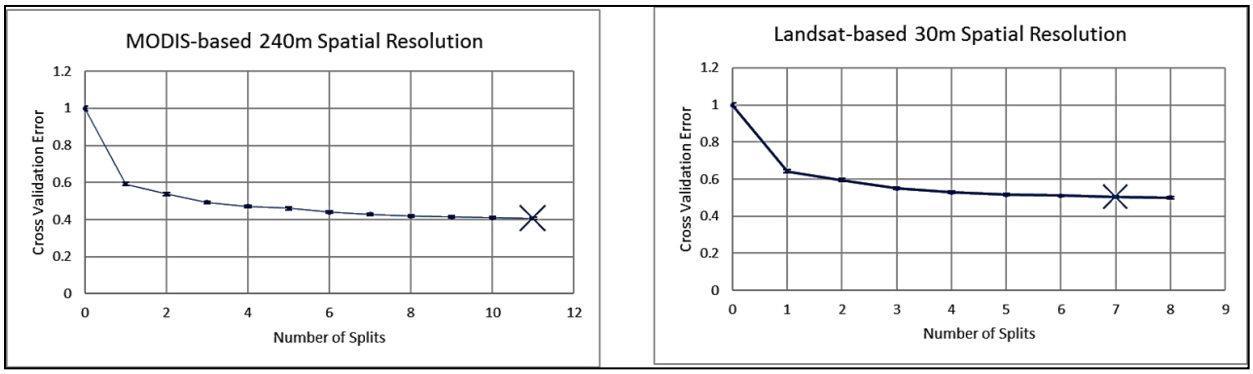

Figure 26 Comparison of 240 m and 30 m models' xerror to number of splits

Additionally, several measures can be used to provide an initial assessment of the model accuracy. The root node error is calculated from the available data rows, divided by the number of rows used in constructing the model. This value is then used to compute a measure of predictive performance for cross-validation error *rate*, which is equal to the x-error times the root node error. Tables 2 and 3 show a summary of model construction parameters including the CP, the pruning point (at CP = 0.005) and various error results from growing the 240 m and 30 m regression trees at each of the splits. Based on these results, the MODIS-based 240 m model shows better promise as a predictive model as indicated by a lower cross-validated error rate, and lower relative error (rel error) and x-error at the pruning point.

Table 2 Summary of model construction parameters in 240 m predictive model

Root node error: 29486/14005 = 2.1054						**cross validation error rate**
CP	**nsplit**	**rel error**	**xerror**	**xstd**	**Prune**	
0.4086	0	1.000	1.000	0.012		2.11
0.0536	1	0.591	0.592	0.008		1.25
0.0451	2	0.538	0.538	0.008		1.13
0.0218	3	0.493	0.494	0.007		1.04
0.0168	4	0.471	0.472	0.007		0.99
0.0158	5	0.454	0.460	0.007		0.97
0.0118	6	0.438	0.441	0.006		0.93
0.0082	7	0.427	0.429	0.006		0.90
0.0073	8	0.418	0.419	0.006		0.88
0.0067	9	0.411	0.415	0.006		0.87
0.0053	10	0.404	0.410	0.006		0.86
0.0050	11	0.399	0.407	0.006	X	0.86

Table 3 Summary of model construction parameters in 30 m predictive model

Root node error: 1911347/961469 = 1.9879						**cross validation error rate**
CP	**nsplit**	**rel error**	**xerror**	**xstd**	**Prune**	
0.3593	0	1.000	1.000	0.002		1.99
0.0456	1	0.641	0.641	0.001		1.27
0.0447	2	0.595	0.596	0.001		1.18
0.0226	3	0.550	0.551	0.001		1.09
0.0124	4	0.528	0.528	0.001		1.05
0.0061	5	0.515	0.516	0.001		1.03
0.0060	6	0.509	0.510	0.001		1.01
*0.00501	7	0.503	0.505	0.001	X	1.00
0.0050	8	0.498	0.501	0.001	X	1.00

*note that a CP 0.00501 exceeds the allowable precision of the input for selecting a CP value. Therefore, the CP value of 0.005 was used

3.3.4. Testing Models with Different Data

To test the generalizability of each model, the models were used to predict the vegetation ReGreen Rate of the other data set. That is, a model constructed from the 240 m data (referred to as the MODIS-based model) was used to predict the vegetation ReGreen Rate of the 30 m data.

Similarly, the 30 m model (referred to as the Landsat-based model) was utilized with the 240 m data to predict the ReGreen Rate.

The *Rattle* package contains two built-in methods for evaluating the predictions of the models: Predicted versus Observed (PrvOb), and Score. As an evaluation of the predictive accuracy, the PrvOb output is a plot of the predicted values and the corresponding observed value with a pseudo-R^2 which Graham Williams describes as akin to the R-squared value of linear regression. The pseudo-R^2 is calculated as the square of the correlation (cor function in R) between the predicted and observed values. Like the R-squared of linear regression, Pseudo-R^2 values closer to 1 have greater consistency between predicted and observed values.

As described by R help, the cor function is based on the Pearson correlation coefficient which is a measure of the linear correlation between two variables x and y using the following equation where n is the number of sample points and $\bar{x}$ and $\bar{y}$ are the sample means.

$$r(x,\mathrm{y}) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\sqrt{(\sum y_i^2 - n\bar{y}^2)}} \quad (7)$$

The second method is to output a score which shows the observed values with the corresponding predicted values as a comma separated value table. This data set allows for a statistical analysis of the difference between observed and predicted values. The results of the model accuracy for Predicted versus Observed (PrvOb) and Score are discussed in the next chapter.

3.4. Summary

This chapter looked at where the data came from and how it was processed into a regression tree model. The differences in image capture and preprocessing for MODIS and Landsat resulted in different processing challenges. Since USGS processes sets of 16 raw

MODIS images into an average single image, cloud free images were obtained, however, the look angle and the topography of the study area resulted in missing data values throughout the study area. Landsat images are collected once every 16 days which led to many cloud-obscured images. At each step, care was given to retain as much data as possible without distorting values too much when filling or smoothing anomalous values. The processed images were used to create normalized EVI values for the three post-fire years. The sum of each of those normalized post-fire years made a single value for each year which together was used to build a regression line for each pixel. The slope of the regression line, the ReGreen Rate, for each pixel was interpreted as an indication of the rate of recovery after the fire.

Using ArcGIS to build the other environmental factors derived from DEM (elevation, aspect, flow accumulation), the soil taxonomic great groups, and the vegetation classes and the R environment to calculate the fire severity (adNBR, RdNBR) values enabled the construction of regression tree models. The regression tree models used the environmental factors to produce predicted values for the ReGreen Rate at each pixel for the 240 m and 30 m spatial resolutions. The results of the regression decision trees and an analysis of the difference between the predicted and the observed ReGreen Rate values are discussed in Chapter 4.

Chapter 4 Results

This chapter compares the accuracy for each of the models and examines the regression decision trees which show the relative importance of each factor used in predicting the ReGreen Rate. Additionally, the two study questions are answered.

4.1. Regression Decision Tree Results

Figures 27 and 28 show the resulting regression decision trees for the 240 m model (MODIS-based) and the 30 m model (Landsat-based). Note, these alternative names, MODIS and Landsat, are used to help differentiate the models from their use with different spatial resolution data. The models showed a strong correlation between fire severity and ReGreen Rate with greater fire severity resulting in a larger ReGreen Rate. As can be seen in the decision trees, fire severity defined the first two splits for both spatial resolutions. Those areas with the lowest ReGreen Rates correspond with the lowest fire severity. This present study interpreted the results as meaning low fire severity areas did not need to recover very much over the three years. Therefore, the rate was relatively flat. In contrast, those areas with high fire severity experienced the greatest loss in green vegetation, therefore over the study period those areas had the greatest potential for recovery and had high ReGreen Rates.

As with the Casady et al. findings, elevation was also significant as a predictor, however in this present study lower elevations (measured in meters) were predictive of higher ReGreen Rates. This difference is likely because of the different climates of Arizona and California and the fact that barren rock dominates the high elevations in this present study. Vegetation as a predictor was common to both MODIS and Landsat-based models with shrub lands experiencing a faster recovery than those dominated by conifer, hardwoods or herbaceous cover. Since the focus of this present study is on the comparative accuracy of the models, detailed examination of

the relationships between soil, vegetation, and elevation on post-fire recovery is recommended to the forestry community for future study.

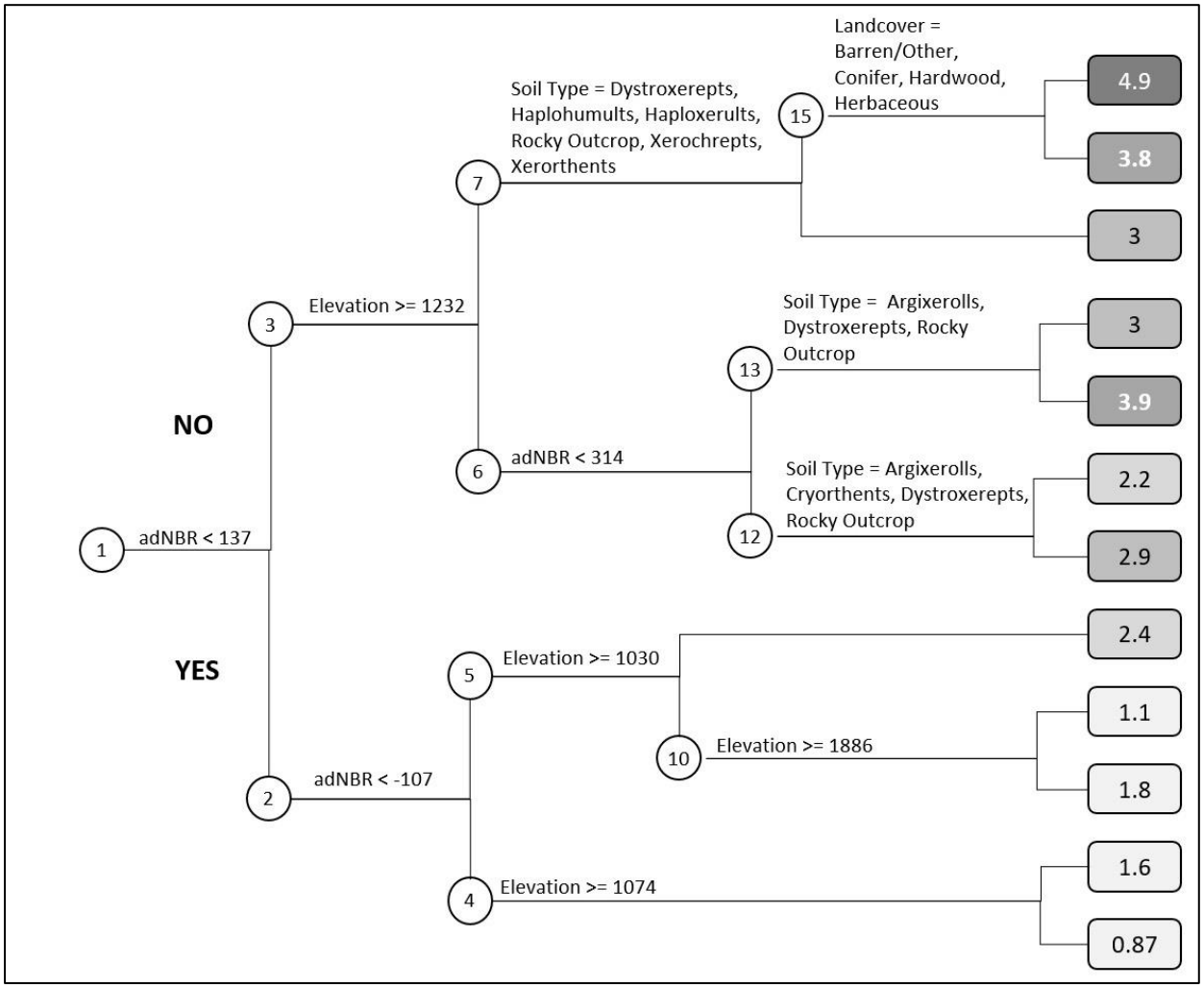

Figure 27 MODIS 240 m Decision Tree Factors to Determining ReGreen Rate

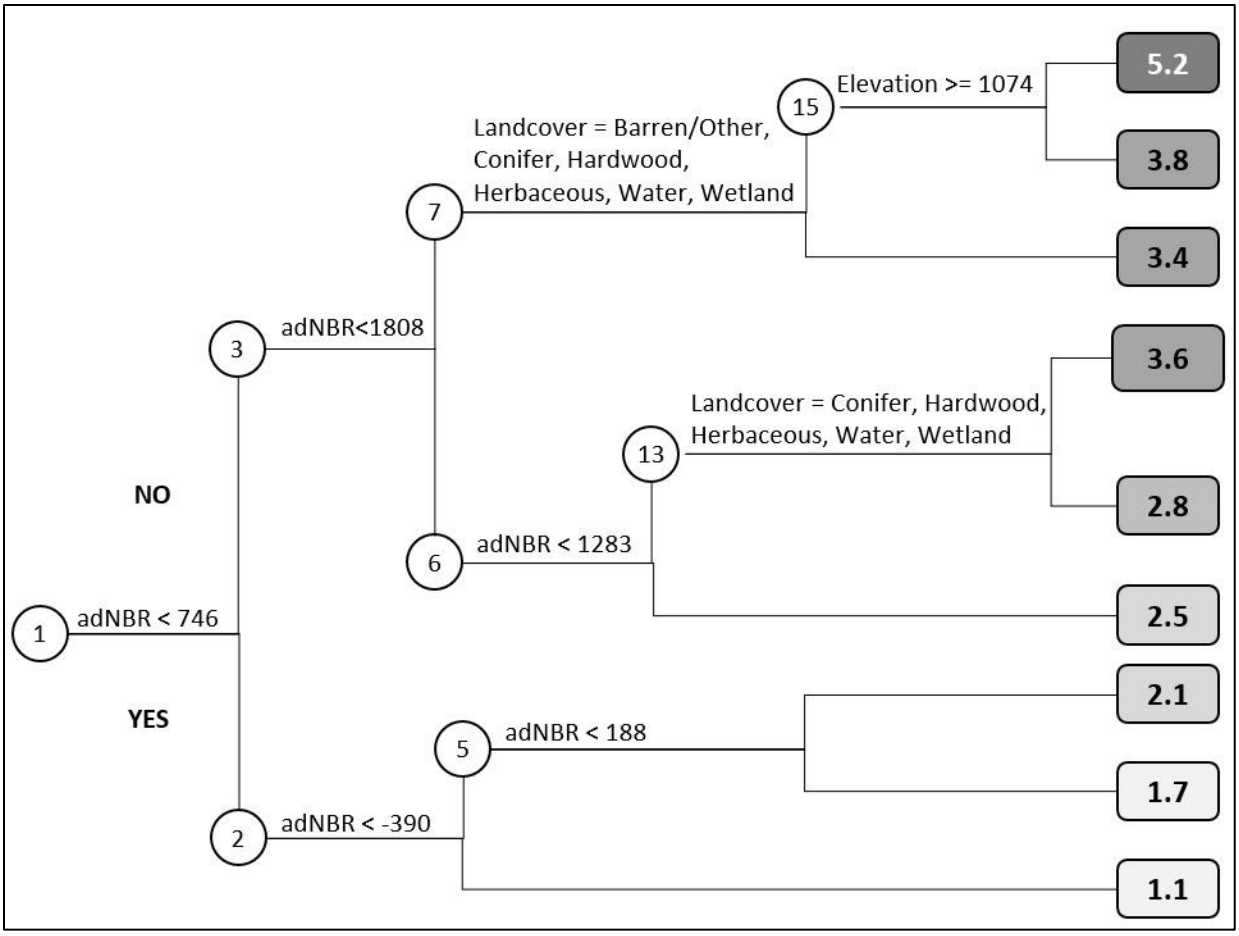

Figure 28 Landsat-based 30 m resolution Decision Tree Factors to Determining ReGreen Rate

Maps of those environmental factors that were used and not used by the models are depicted in Figures 29 and 30. The maps enable a visualization of the decision tree factors in the study area. One can see the regions of high fire severity generally correspond with mid- to low-elevation areas and the shrub lands which are found predominantly in the lower elevation in the confluence of drainage from the highlands.

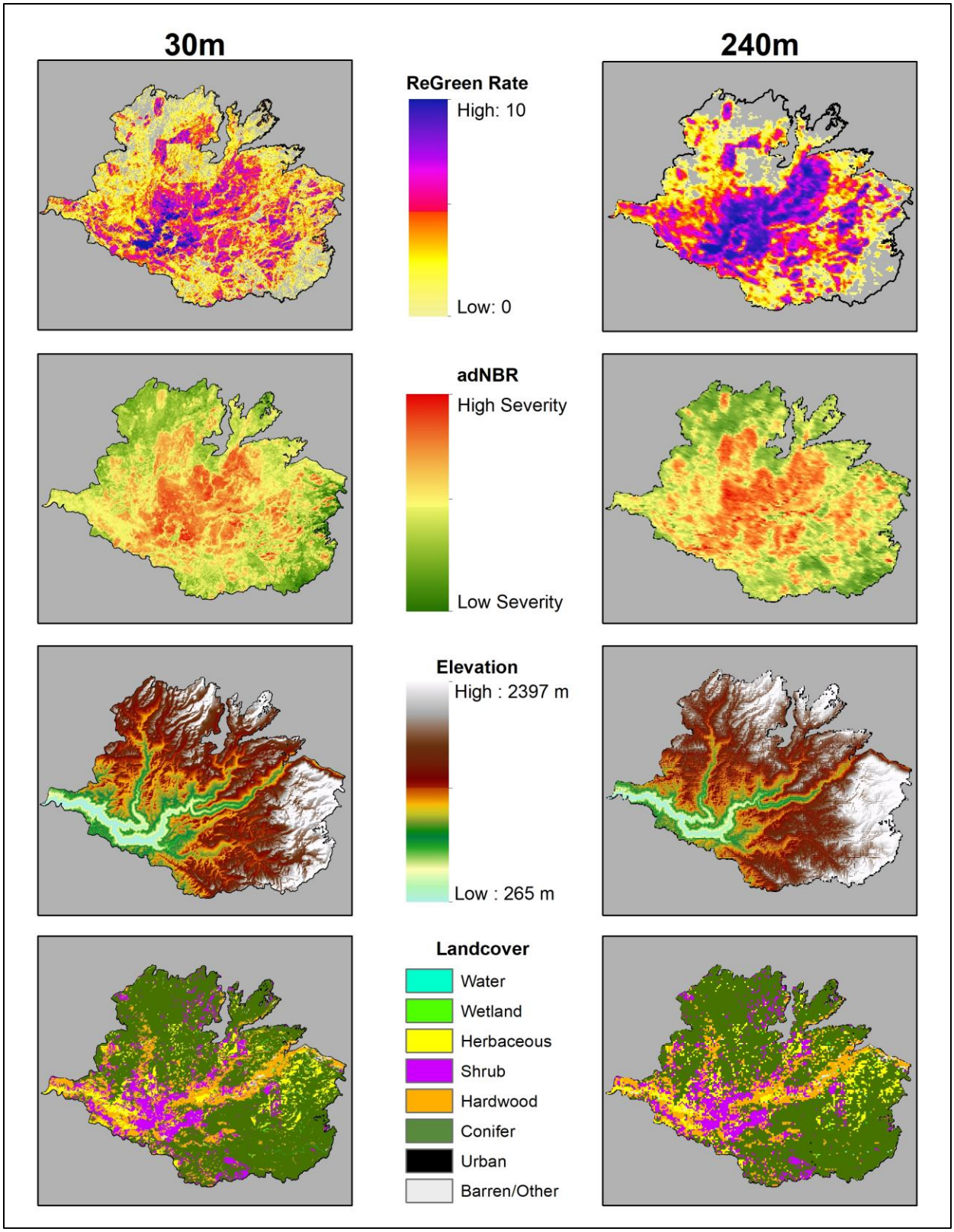

Figure 29 Dominant attributes used by both models (soil was used only by the MODIS model)

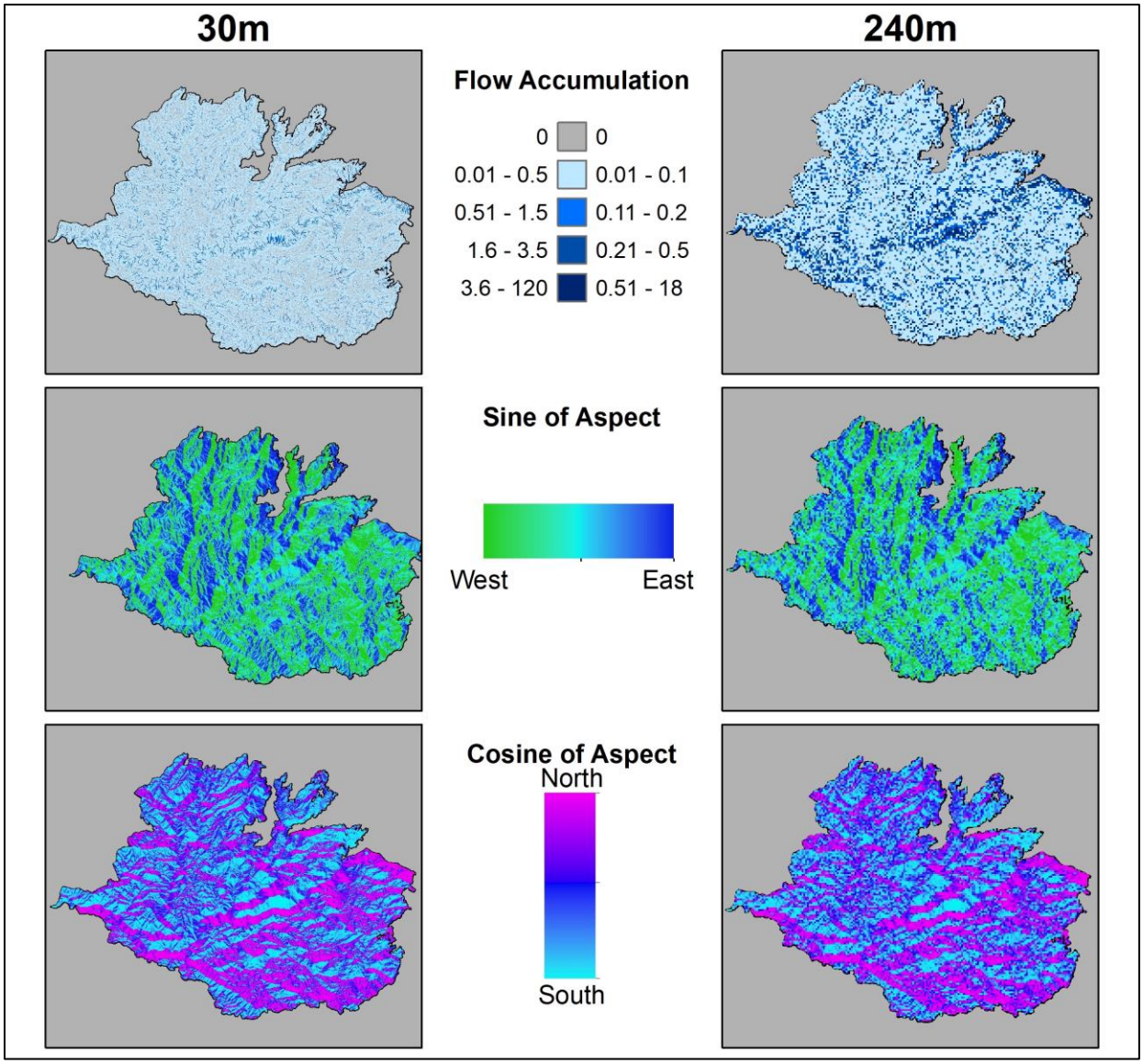

Figure 30 Attributes not used in either model

4.2. Predicted versus Observed Results

An assessment of the predicted versus observed values by use of the square of the Pearson correlation coefficient called the pseudo-R^2 is summarized and shown in Table 4. These values indicate how well the model matched the observed values with predicted values. The grayed in areas highlight the data resolution used in building the MODIS and Landsat-based models, the white areas are from the cross-resolution testing. Both models performed better with

the data resolution used to construct the model than with the cross-resolution testing data. The MODIS-based model more accurately predicted the ReGreen Rate with new data from the other resolution than did the Landsat-based model, indicating the MODIS-based model is a better-generalized model. However, the Landsat-based model produced a more accurate model using 30 m spatial resolution data.

Table 4 Summary of Pseudo-R^2 values for Models and Data Resolution

Data resolution	MODIS-based Model	Landsat-based Model
30 m data	0.4105	**0.4982**
240 m data	**0.588**	0.3858

The pseudo-R^2 served as an indication of model accuracy and identified that the Landsat model produced a more accurate result than the MODIS model when both used the 30 m data. To understand the magnitude of difference in model accuracy on the scale of the ReGreen Rate range, Section 4.3 examines the root mean square difference between the observed and predicted values.

4.3. Result of Study Question One – Comparing 240 m and 30 m Derived Models

Distinct from the pseudo-R^2 value, the score output from *Rattle* was used to extract the data table of each pixel's observed and predicted ReGreen Rate values of the MODIS-based and Landsat-based models. Examination of the difference between the internal predictive power and cross resolution predictive power showed that the MODIS-based model more accurately predicts the ReGreen Rate on a cross resolution data test, that is when the model uses data from the spatial resolution *not* utilized in the model construction. However, the present study's question asks if higher resolution data produces a more accurate model. The Landsat-based model did more accurately predict the ReGreen Rate of 30 m spatial resolution data.

To understand the difference in accuracy at the scale of the ReGreen Rate for the MODIS and Landsat-based models, "accuracy" here is now defined as the mean difference between the predicted value and the observed value of each pixel. A greater mean difference indicates a less accurate model and a smaller mean difference indicates a more precise model.

A root means squared error (RMSE) was used to quantify the difference between each models' accuracy where values closer to zero indicates a more accurate model. Table 5 summarizes the error of the models for internal validation and cross-resolution testing. It is important to note that ReGreen Rate values range from 0 to 10, and the errors are relative to that scale. Values in parentheses are the percent of the range.

Table 5 Summary of Model Errors

Model and Data	RMSE
Landsat Model when using 240 m data (cross-test)	1.35 (13.6%)
Landsat Model when using 30 m data (internal validate)	0.99 (9.9%)
MODIS Model when using 30 m data (cross-test)	1.16 (11.8%)
MODIS Model when using 240 m data (internal validate)	0.93 (9.3%)

A visual inspection of the model results confirms there is a minimal difference in model accuracy. Figure 31 contrasts the calculated ReGreen Rate based on EVI values with those generated by the models. For a given model, the spatial resolution used to develop the model is labeled as the internal validation (C and B), and as a cross-resolution test (A and D) when the model used data from the spatial resolution not involved in building the model. The embedded table shows the difference in RMSE percent error for the models. An example reading of the table for B>A is to say the "accuracy of B is greater than A by 1.70%", that is B (Landsat model using 30 m data) is 1.70% more accurate than the A (MODIS model using 30 m data).

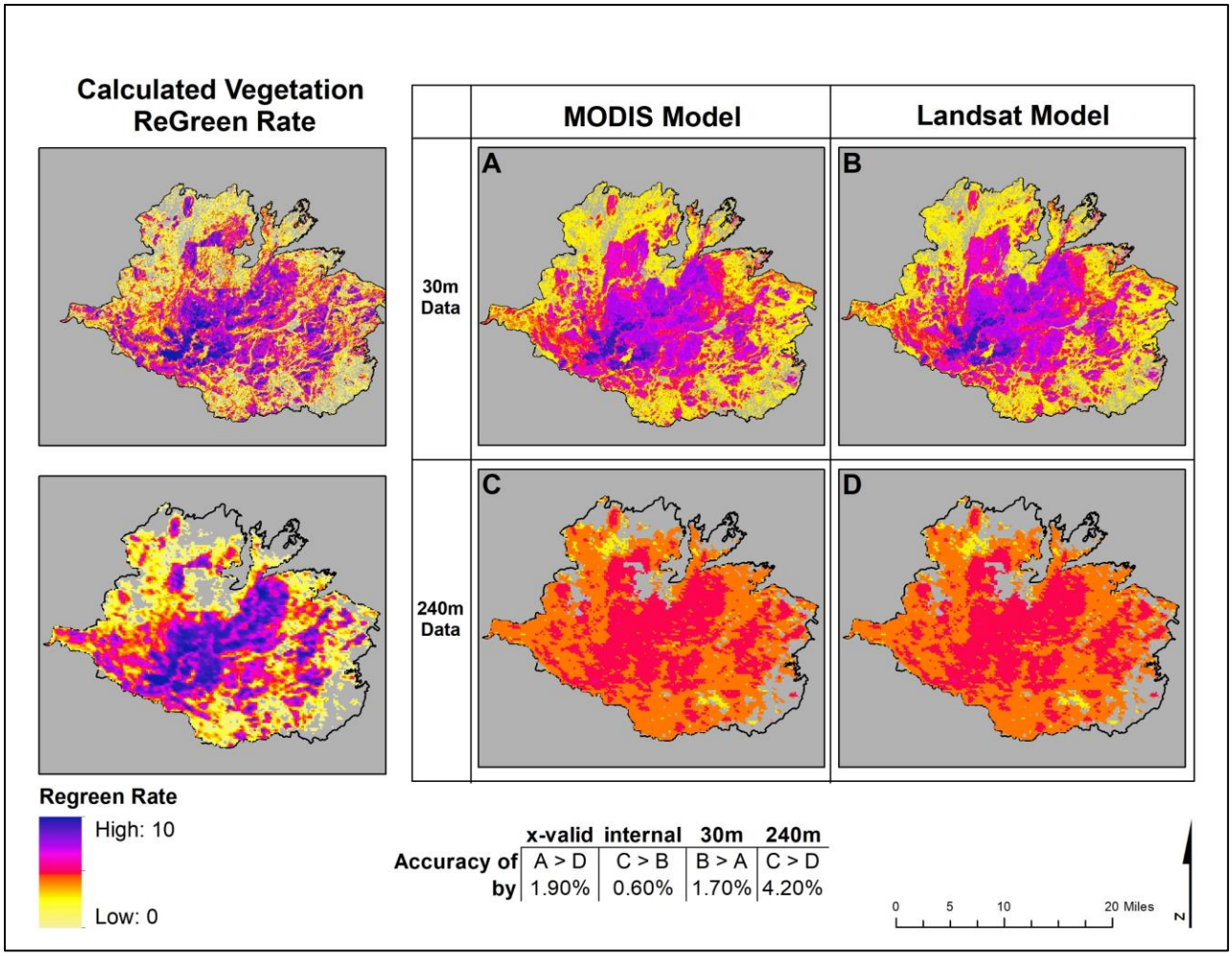

Figure 31 Maps of calculated and modeled ReGreen Rate

As identified by the embedded table in Figure 31, the difference in relative percent error between the models is about 1.7% for the 30 m resolution data. While the difference in internal resolution accuracy is less than 1%, the MODIS is almost 2% more accurate in cross-resolution validation.

4.4. Result of Study Question Two – Use of Different Indices for Fire Severity

As discussed in Chapter 2, RdNBR is known to more accurately describe fire severity when compared to field observations than dNBR. The adNBR is an interpretation of fire severity based on the difference between the observed and a linear regression-based predicted value from

pre-fire NBR and dNBR values. Two regression tree models were constructed using the 240 m data with these different indices as the fire severity predictive variable using CP = 0.005 as a stopping point, with 5% retained for testing.

The score output from *Rattle* produced a comma separated value (CSV) file consisting of observed and predicted ReGreen Rate values. The plots in Figure 32 depict the density distribution for the difference between observed and predicted values for the test data from the 240 m data set for both fire severity index models. A simple visual inspection of Figure 32 indicates no significant increase in model accuracy using the RdNBR rather than adNBR as a measure of fire severity.

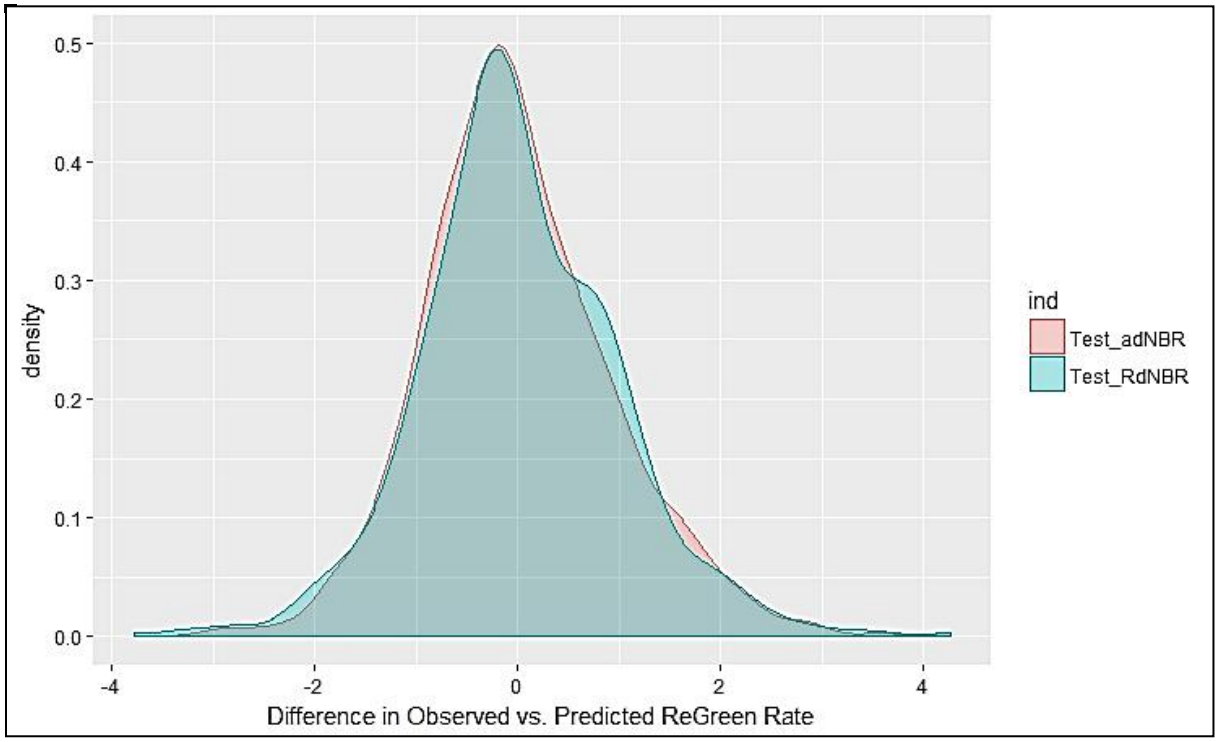

Figure 32 Difference between Observed and Predicted ReGreen Rates for MODIS-based models on 5% test set

A two-tailed t-test was used to establish if there is a statistically significant difference between the residuals of the observed and predicted ReGreen Rates with the null hypothesis: there is no difference between the average residuals for ReGreen Rates using adNBR or RdNBR

as predictive variables. A two-tailed t-test confirms the visual inspection, where t = 0.068 is inside the 95% confidence interval of -0.111 to 0.119. Therefore, the null hypothesis cannot be rejected. There is not sufficient evidence to show a statistical difference in the predictive power of using adNBR versus RdNBR on 240 m spatial resolution imagery.

4.5. Summary

Analysis of study question one indicated that higher spatial resolution information did result in a marginally more accurate model than when using the MODIS-based model with 30 m data. However, there is a bias associated with this result since both models performed markedly better when evaluated against the data resolution for internal validation. The MODIS-based model is a better-generalized model when using data from the other resolution. Considering to the complex processing for Landsat EVI data and the marginal improvement in model accuracy, this study finds that a robust model based on a 240 m spatial resolution EVI data can be constructed with the advantage of minimal correction for atmospheric or image capture errors and reduced computational time. Additionally, different methods for determining fire severity did not produce statistically different accuracy in the resulting model.

Chapter 5 Conclusion

Both Casady et al. and this present study used a temporally smoothed time series of EVI data to formulate rates of post-fire recovery after a wildfire. This present study constructed three models to answer two questions. First, does higher spatial resolution data produce a more accurate predictive model and second does use of the RdNBR index instead of the adNBR index for defining fire severity produce a more precise predictive model? This present study found that 30 m spatial resolution data can produce a marginally more accurate model than a model constructed from 240 m when both models use the 30 m data to predict the ReGreen Rate. Additionally, no significant difference in model accuracy was observed when different fire severity indices were used to construct the regression tree model.

Regression tree models can be applied to a post-fire environment where enough data is available to create the needed model attributes Earlier studies have shown the significance of individual properties influencing the recovery of vegetation after a fire. Casady et al.'s method provides for the combination of attributes and consequently leads to greater insight into the dominant factors affecting post-fire recovery.

The most significant conclusion is that the MODIS model performed almost as well (within 2%) using 30 m data as did the Landsat model. Constructing the ReGreen Rate is much simpler and faster with the 240 m data. With the ReGreen Rate and environmental factors scaled from 30 m to 240 m, one could construct a MODIS model for use with 30 m data arriving at a higher spatial resolution understanding of local influences on post-fire recovery without the confounding aspects of obscured pixels when using Landsat EVI data to build a model.

Landsat is better suited for analyzing smaller areas. As this present study found, the 18,093 pixels in the MODIS EVI images were sufficient to construct a robust model. If the study

looked at an area covered by 18,093 Landsat 30 m pixels (approximately a 4 km x 4 km square), there would only be 283 MODIS 240 m pixels which would not construct a robust model. This suggests there is a point of diminishing returns depending on the size of the study area. For small areas, it is likely more appropriate to use 30 m pixels, whereas, in an area such as the 235,841 acre Rim Fire, the MODIS pixels are better suited for building a generalized model. Additionally, as discussed above, the issue of cloud cover in the Landsat images increases the complexity of processing and invites errors without an appreciable increase in model accuracy over using the MODIS-based model with 30 m spatial resolution data.

That the ReGreen Rates were highest in those areas with high fire severity exposes a limit of this type of analysis. Those regions with the greatest decrease in EVI values due to the fire had the fastest recovery over the three-year period likely due to having the largest potential for recovery. However, the model does not indicate what grew back in the area as compared to the pre-fire conditions. The difference between pre-fire EVI and the EVI value after several years of recovery could be used as an additional factor to get a measure of complete recovery.

5.1. Opportunities for Future Research

Examination of the study area identified areas that had been clear-cut. The clear-cut harvesting practice has a noticeable impact on the recovery after a forest fire. Figure 33 compares the calculated (observed) ReGreen Rate and what the Landsat-based model predicted for this area. The distinct border at the edge of the clear-cut area shows the observed recovery as much lower than in the vicinity. Based on the environmental parameters used, the model predicted a similar recovery inside and outside of the clear-cut area. Future studies could look at incorporating forest management practices as parameters in the regression decision trees to understand their relative impact on post-fire recovery.

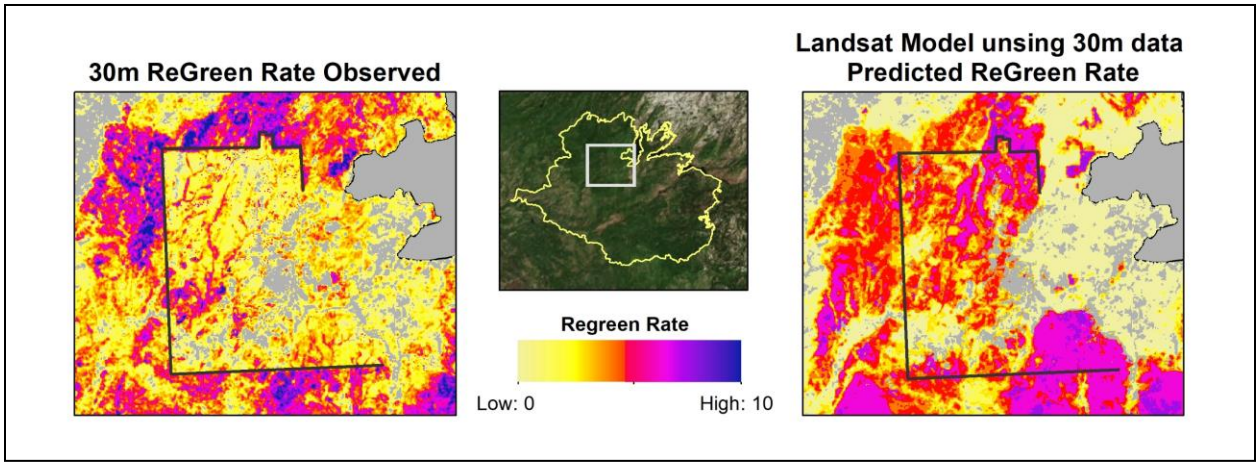

Figure 33 Difference in Observed and Predicted ReGreen Rates in a Clearcut Area

Further research could also look at using the *Rattle* tool for the rapid production of insightful decision trees. Using the *Rattle* tool, each factor could be examined for its relationship to other factors and complex relationships such as flow accumulation and soil type as predictors for vegetation type can be considered. Research could also examine the relationship of fire history and clear-cut practices to fire severity. It may be interesting to see if a more accurate regression tree model is possible using a ReGreen Rate that is based on seasonally adjusted annual EVI values based on the work of Lhermitte et al. as described in Section 2.6.2.

This present study focused on the post-fire ReGreen Rate which is different from saying this current study concentrated on vegetation regeneration. Using remotely sensed data is useful in providing insight into the health of vegetation, but this current study could not provide insight into which vegetation species is recovering. Additional work could focus on field validation of recovery rates and specific species recovery. Finally, examination of the regions with negative recovery despite favorable environmental factors could provide insight into the driving factors influencing negative post-fire recovery.

5.2. Summary

This present study helped support the assumption of Casady et al. that higher spatial resolution data does produce a more accurate model. More significantly, this current study showed that researchers armed with a basic understanding of the R environment could take advantage of the increasing computational power of traditional computers and the growing array of tools that demystify the art of data science to produce robust predictive models. A broader application and understanding of predictive modeling in natural resource management can lead to greater insights about the interconnected aspects of ecosystems. Such insights are crucial to the efficient allocation of resources to preserve and protect global natural resources.

References

Adams, Douglas. 1979. *Hitchhiker's Guide to the Galaxy.* London: Pan Books.

Amiro, B. D., Chen, J. M., & Liu, J. 2000. "Net primary productivity following forest fire for Canadian ecoregions." *Canadian Journal of Forest Research/Revue Canadienne De Recherche Forestiere* 30 (6): 939–947.

CAL FIRE. 2012. *FRAP HOME.* Accessed February 25, 2017. http://frap.fire.ca.gov/index.

California Department of Forestry and Fire Protection. 2017. *FRAP Vegetation (FVEG15_1).* 03 15. http://frap.fire.ca.gov/data/statewide/FGDC_metadata/fveg15_1.xml.

Casady, Grant M., Willem J. D. van Leeuwen, and Stuart E. Marsh. 2010. "Evaluating Post-wildfire Vegetation Regeneration as a Response to Multiple Environmental Determinants." *Environmental Modeling & Assessment* 15: 295-307. doi:10.1007/s10666-009-9210-x.

CDF. 2013. *Cal Fire- Incident Information- Rim Fire.* Accessed October 7, 2016. http://cdfdata.fire.ca.gov/incidents/incidents_details_info?incident_id=905.

Chen, Xuexia, James E. Vogelmann, Matthew Rollins, Donald Ohlen, Carl H. Key, Limin Yang, Chengquan Huang, and Hua Shi. 2011. "Detecting post-fire burn severity and vegetation recovery using multitemporal remote sensing spectral indices and field-collected composite burn index data in a ponderosa pine forest." *International Journal of Remote Sensing* 32 (23): 7905-7927.

DeBano, L. F., D. G. Neary, and P. F. Folliott. 1998. *Fire's effect on ecosystems.* New York: Wiley.

Díaz-Delgado, R., F. Lloret, and X. Pons. 2003. "Influence of fire severity on plant regeneration by means of remote sensing imagery." *International Journal of Remote Sensing* 24 (8): 1751-1763.

Eva, Hugh, and Eric F. Lambin. 2000. "Fires and Land-Cover Change in the Tropics: A Remote Sensing Analysis at the Landscape Scale." *Journal of Biogeography* 27 (3): 765-776.

Flores, Mary, Curtis Kvamme, Brad Rust, Kellen Takenaka, and David Young. 2013. *BAER Assessment Soils Report – Rim Fire.* Yosemite National Park and Wilderness: USFS.

Gabbert, Bill. 2015. *Deaths of two witnesses result in dropped charges against person accused of starting Rim Fire.* Accessed October 2016, 7. http://wildfiretoday.com/tag/rim-fire/.

Goetz, S. J., Fiske, G. J., & Bunn, A. G. 2006. "Using satellite time-series data sets to analyze fire disturbance and forest recovery across Canada." *Remote Sensing of Environment* 101 (3): 352–365.

Harris, Lucas, and Alan H. Taylor. 2015. "Topography, Fuels, and Fire Exclusion Drive Fire Severity of the Rim Fire in an Old-Growth Mixed-Conifer Forest, Yosemite National Conifer Forest, Yosemite National." *Ecosystems* 18: 1192–1208. doi:10.1007/s10021-015-9890-9.

Hoelzemann, Judith J., Martin G. Schultz, Guy P. Brasseur, and Claire Granier. 2004. "Global Wildland Fire Emission Model (GWEM): Evaluating the use of global area burnt satellite data." *Journal of Geophysical Research* 109: 1-18. doi:10.1029/2003JD003666.

Huete, A., K Didana, T. Miuraa, E.P. Rodrigueza, X. Gaoa, and L.G. Ferreirab. 2002. "Overview of the radiometric and biophysical performance of the MODIS vegetation indices." *Remote Sensing of Environment* 83 (1-2): 195-213.

Jain, Theresa B., and Russell T. Graham. 2003. "Fire severity classification: Uses and abuses." Orlando, FL, U.S.A.: Second International Wildland Fire Ecology and Fire Management Congress and Fifth Symposium on Fire and Forest Meteorology.

Jonsson, P., and L. Eklundh. 2002. "Seasonality Extraction by Function Fitting to Time-Series of Satellite Sensor Data." *IEEE Transactions on Geoscience and Remote Sensing* 40 (8): 1824-1832.

Key, Carl H., and Nathan C. Benson. 2006. *Landscape Assessment Sampling and Analysis Methods.* United States Forest Service.

Kokalya, Raymond F., Barnaby W. Rockwella, Sandra L. Haireb, and Trude V.V. King. 2007. "Characterization of post-fire surface cover, soils, and burn severity at the Cerro Grande Fire, New Mexico, using hyperspectral and multispectral remote sensing." *Remote Sensing of Environment* 106 (3): 305–325.

Lhermitte, S., J. Verbesselt, W.W. Verstraeten, S. Veraverbeke, and P. Coppin. 2011. "Assessing intra-annual vegetation regrowth after fire using the pixel based regeneration index." *ISPRS Journal of Photogrammetry and Remote Sensing* 66: 17-27.

Loh, Wei-Yin. 2016. "Classification and Regression Tree Methods." In *Encyclopedia of Statistics in Quality and Reliability*, by Fabrizio Ruggeri, Ron S. Kenett and Frederick Faltin (eds.), 315-323. Wiley.

Lydersen, Jamie M., Malcolm P. North, and Brandon M. Collins. 2014 . "Severity of an uncharacteristically large wildfire, the Rim Fire, in forests with relatively restored frequent fire regimes." *Forest Ecology and Management* 328: 326–334.

Lydersen, Jamie M., Malcolm P. North, Eric E. Knappc, and Brandon M. Collins. 2013. "Quantifying spatial patterns of tree groups and gaps in mixed-conifer forests: Reference

conditions and long-term changes following fire suppression and logging." *Forest Ecology and Management* 304: 370–382.

Miller, Jay D., and Andrea E. Thode. 2007. "Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR)." *Remote Sensing of Environment* 109: 66–80.

NASA. n.d. *Moderate Resolution Imaging Spectroradiometer (MODIS) Specification Sheet.* Accessed February 15, 2017. https://modis.gsfc.nasa.gov/about/specifications.php.

—. 2013. *Progression of California's Rim Fire.* Accessed October 7, 2016. http://earthobservatory.nasa.gov/IOTD/view.php?id=81971.

National Parks Service. 2016. *National Parks Service Fire and Aviation Management - Wildland Fires.* Accessed November 14, 2016. https://www.nps.gov/fire/wildland-fire/about.cfm.

National Parks Service. 2013. *Yosemite NP Rim Fire, Burned Area Emergency Response Plan.* National Park Service, 4,6,8,9,19-21,29-32.

Nepstad, Daniel C., Adalberto Verissimo, Ane Alencar, Carlos Nobre, Eirivelthon Lima, Paul Lefebvre, Peter Schlesinge, et al. 1999. "Large-scale impoverishment of Amazonian forests by logging and fire." *NATURE* 398: 505-508.

Parsons, David J., and Steven H. DeBenedetti. 1979. "Impact of fire suppression on a mixed-conifer forest." *Forest Ecology and Management* 2: 21-33.

Pollet, J., & Omi, P. N. 2002. "Effect of thinning and prescribed burning on crown fire severity in ponderosa pine forests." *International Journal of Wildland Fire* 11 (1): 1-10.

Potter, Christopher. 2014. "Geographic Analysis of Burn Severity for the 2013 California Rim Fire." *Natural Resources* 5: 597-606. doi:http://dx.doi.org/10.4236/nr.2014.511052.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1993. *Numerical Recipes in FORTRAN; The Art of Scientific Computing.* 2. New York, NY: Cambridge University Press.

Reilly, M.J., M.C. Wimberly, and C.L. Newell. 2006. "Wildfire effects on plant species richness at multiple spatial scales in forest communities of the southern Appalachians." *Journal of Ecology* 94 (1): 118–130.

Robichauda, Peter R., Sarah A Lewisa, Denise Y.M. Laesb, Andrew T. Hudaka, Raymond F. Kokalyc, and Joseph A. Zamudiod. 2007. "Postfire soil burn severity mapping with hyperspectral image unmixing." *Remote Sensing of Environment* 108 (4): 467–480.

Rothman, Hal K. 2005. *A Test of Adversity and Strength: Wildland Fire in the National Park System.* Washington, D.C: U.S. Dept. of the Interior, National Park Service.

Scholl, Andrew E., and Alan H. Taylor. 2010. "Fire regimes, forest change, and self-organization in an old-growth mixed-conifer forest, Yosemite National Park, USA ." *Ecological Applications* 20 (2): 362–380.

Shalizit, Cosma. 2009. *Classification and Regression Trees.* Accessed October 2016. www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf.

Simpson, Gavin, interview by Jessica Eselius. 2017. *What is the difference between rel error and x error in a rpart decision tree?* (March 22). https://stackoverflow.com/questions/29197213/what-is-the-difference-between-rel-error-and-x-error-in-a-rpart-decision-tree.

USFS. 2016. *Rapid Assessment of Vegetation Condition after Wildfire (RAVG).* Accessed October 8, 2016. http://www.fs.fed.us/postfirevegcondition/whatis.shtml.

—. 2015. *Stanislaus National Forest: All Firefighters go Home to their Families After their Shift.* Accessed October 5, 2016. http://www.fs.usda.gov/detail/stanislaus/home/?cid=stelprd3824723.

—. 2015. "U.S. Forest Service Fire Suppression." March17. Accessed October 7, 2016. http://www.foresthistory.org/ASPNET/Policy/Fire/Suppression/Suppression.aspx.

USGS. 2016. *Frequently Asked Questions about the Landsat Missions.* Accessed October 8, 2016. http://landsat.usgs.gov/band_designations_landsat_satellites.php.

Vermote, E., C. Justice, M. Claverie, and B. Franch. 2016. "Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product." *Remote Sensing of the Environment*, April: 46-56.

Viedma, Olga. 2008. "The influence of topography and fire in controlling landscape composition and structure in Sierra de Gredos (Central Spain)." *Landscape Ecology* 23 (6): 657–672.

Virginia, Dale H., Joyce A. Linda, Mcnulty Steve, Neilson P. Ronald, Ayres P. Matthew, Flannigan D. Michael, Hanson J. Paul, et al. 2001. "Climate Change and Forest Disturbances." *BioSciences* 51 (9): 723-734.

Weier, John, and David Herring. 2000. *Measuring Vegetation (NDVI & EVI).* Accessed October 9, 2016. http://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php.

Williams, Graham j. 2009. "Rattle: A Data Mining GUI for R." *the R Journal* 1/2: 45-55. Accessed June 2, 2017. https://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.

Appendix A R Code

```
library(zoo)
library(signal)
library(Rattle)

#########################
#SETUP FOR MODIS BUILD
#########################
MODIS_Mask <-
raster('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\EVI\\M
ODIS\\Mask\\MODIS_Mask.tif')
MODIS_Mask[!is.na(MODIS_Mask)] <- 1 #32550 total pixels per
image
pixel.count <- na.omit(as.vector(MODIS_Mask)) #18093 value
pixels per image x 161 images = 2912973 pixels in study
area

#change directories to where the images reside
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\EVI\\MO
DIS')

#read in the data from the folder
rlist=list.files(getwd(), pattern="tif$", full.names=FALSE)
for(i in rlist) { assign(unlist(strsplit(i, "[.]"))[1],
raster(i)) }
#clean up
rm(i)
number.of.images <- length(rlist)

EVI <- stack()# creates place holder stack

#loop to build the RasterStack
ptimestack <- system.time({
 for(i in 1:NROW(rlist)){
  tempraster <- raster(rlist[i])
  EVI <- raster::stack(EVI,tempraster)
 }
})
ptimestack #16.87sec,14.79 sec

rm(i)
rm(rlist)

EVI_Mask <- mask(EVI, MODIS_Mask) #apply the mask from
ArcGIS to ensure the pixels of the image match the standard
template

values.of.stack <- getValues(EVI_Mask)
```

```
number.pixels <- na.omit(values.of.stack) #2638951 pixels
in stack or 90.59% of all pixel values inside the fire
boundary.
saveRDS(values.of.stack,"setup_VoS")
grt6500 <- na.omit(values.of.stack[values.of.stack>6500])
#32 pixels
values.of.stack[values.of.stack>6500] <- NA
less0 <- na.omit(values.of.stack[values.of.stack<0]) #6648
pixels
values.of.stack[values.of.stack<0] <- NA

#Replace located NA with na.fill. This looks down each
column (which is a vectorized image) of the values. This
fills the NA values across the same image, i.e. spatially.
pre <- which(is.na(values.of.stack),arr.ind = FALSE) #find
the placement of the NA values
str(pre) #2343806 NA values

for(i in 1:number.of.images) {
 temp.vector.column <- as.vector(values.of.stack[,i])
 if(!all(is.na(temp.vector.column))){#do if the column has
values
  temp.vector.column <- na.fill(temp.vector.column,
"extend")
  values.of.stack[,i] <- temp.vector.column
 }
 next
}

EVI_Mask[] <- values.of.stack
plot(EVI_Mask) #extent-based pixels have been filled as
well
EVI_Mask <- mask(EVI_Mask,MODIS_Mask) #but applying the
mask returns the values of the stack to the fire boundary
plot(EVI_Mask) #and now the NA values have been spatially
filled
values.of.stack <- getValues(EVI_Mask) # update the
values.of.stack variable with the new NA Filled values.

saveRDS(values.of.stack,"NAfillEVIValues")

#Conduct S-G Smoothing over the time series
ptimesmooth <- system.time ({
 for(i in 1:NROW(values.of.stack)) {
  temp.vector.row <- as.vector(values.of.stack[i,])
  smoothed.row <-
sgolayfilt(na.pass(temp.vector.row),p=7,n=9,m=0)
  values.of.stack[i,] <- smoothed.row
 }
})
saveRDS(values.of.stack,"NAf_Smooth_EVIValues")
values.of.stack <- readRDS("NAf_Smooth_EVIValues")
```

```
ptimesmooth #19 sec

smooth.dataframe <- as.data.frame(values.of.stack)
#Generating the Normalized EVI values

nPre <- 84 #number of MODIS images prior to the fire. This
has to be set based on knowing about the data set.

avg.prefire <- rowMeans(smooth.dataframe[,1:nPre]) #uses
the first "nPre" columns of the data frame.
summary(avg.prefire)

normalized.smooth.dataframe <-
as.matrix(smooth.dataframe/avg.prefire) #this normalizes
the entire dataframe based on the prefire average EVI
values from the images
saveRDS(normalized.smooth.dataframe,"Normalized_smooth_fill
ed_EVI")
summary(normalized.smooth.dataframe)

#determine the post-fire cumulative annual EVI values.
PostFire.image.per.year <- 23 #total number of images in
each of the 12 month years after the fire

#get the sum of the post year normalized EVI values in each
pixel
interval1 <- c((nPre+1):(nPre+PostFire.image.per.year))
interval2 <-
c((max(interval1)+1):(max(interval1)+PostFire.image.per.yea
r))
interval3 <-
c((max(interval2)+1):(max(interval2)+PostFire.image.per.yea
r))

year1 <- normalized.smooth.dataframe[,interval1]
year2 <- normalized.smooth.dataframe[,interval2]
year3 <- normalized.smooth.dataframe[,interval3]

year1.sum <- as.vector(rowSums(year1)) #need to be vectors
for the rbind
year2.sum <- as.vector(rowSums(year2))
year3.sum <- as.vector(rowSums(year3))
summary(year1.sum)
summary(year2.sum)
summary(year3.sum)

#rebuild a dataframe that has all three years of summed EVI
values and calculate the ReGreen Rates
Postfire.reveg <- rbind(year1.sum,year2.sum,year3.sum)
Postfire.reveg <- t(as.data.frame(Postfire.reveg)) #need
the transpose of the data frame in order to calculate the
```

```
linear regression slopes using the number of years vector
(called reveg.slp)

years.gone.by <- as.vector(c(1,2,3))

Postfire.reveg[is.na(Postfire.reveg)] <- 0 #have to set NA
values to 0 for the lm function to work properly

tempslope <- vector(length = NROW(Postfire.reveg)) #is the
template into which each slope value is placed
ptimeslope <- system.time({
 for(i in 1:NROW(Postfire.reveg)){
  temprow <- as.vector(Postfire.reveg[i,])
  tempfit <- lm(temprow ~ years.gone.by, na.action =
na.omit) #regression analysis of the relationship between
year and the cumulative annual normalized EVI values
  tempslope[i] <- tempfit$coefficients[2] #this returns the
slope of the regression line to the template
 }
})
saveRDS(tempslope,"tempslope")
ptimeslope # 20sec for MODIS. note that processing for
LANDSAT is approximately 27 minutes.
tempslope[tempslope%in%(0)] <- NA #reapply the NA values to
those areas that were converted to 0 in order to calculate
the slopes
rate.of.reveg <- as.data.frame(tempslope) #format as a
dataframe for easier inspection
saveRDS(rate.of.reveg,"Rate of Revege Dataframe")

summary(tempslope)

Re.Green.Rate <- MODIS_Mask #from Step 1. The MODIS Mask
has all the correct meta data for creating a GeoTIFF from
the ReGreen.rate vector data.
Re.Green.Rate[] <- tempslope #fill the shell with the slope
data
plot(Re.Green.Rate)
names(Re.Green.Rate) <- "ReGreen Rate" #rename the Raster
to match the data it contains.
writeRaster(Re.Green.Rate,
filename="ReGreen_with_neg_vals", format="GTiff",
overwrite=TRUE)#putting the resulting
#however, this plot contains negative values
tempslope[tempslope<(0)] <- NA #remove negative slopes as
per Cassidy et. al. It would be interesting to explore the
causal or correlated attributes associated with the
negative slopes in a future study.
Re.Green.Rate.clipped <- MODIS_Mask
Re.Green.Rate.clipped[] <- tempslope #fill the shell with
the slope data
plot(Re.Green.Rate.clipped)
```

```
names(Re.Green.Rate) <- "ReGreen Rate" #rename the Raster
to match the data it contains.
writeRaster(Re.Green.Rate.clipped, filename="ReGreen",
format="GTiff", overwrite=TRUE)#putting the resulting
raster back into a geotiff in the working directory

############################################################
#########################

#Gather the descriptive attributes (i.e., fire severity,
slope, elevation, aspect, vegetation type, soil type, and
flow accumulation)

############################################################
#########################

###############
#STEP 1
#Calculate NBR, adNBR, and RdNBR  Must correct for the NA
values
###############
rm(list=ls())

#read in the data from the MODIS Mask folder
MODIS_Mask <-
raster('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\EVI\\M
ODIS\\Mask\\MODIS_Mask.tif')
MODIS_Mask[!is.na(MODIS_Mask)] <- 1

#change directories to where the NBR images reside
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\NBR\\MO
DIS\\R-NBR\\Loop_build')
rlist=list.files(getwd(), pattern="tif$", full.names=FALSE)
for(i in rlist) { assign(unlist(strsplit(i, "[.]"))[1],
raster(i)) }
#clean up
rm(i)
number.of.images <- length(rlist)

shell <- PreNIR #use the PreNIR as a shell
shell[] <- NA #remove all the values for the shell
plot(PreNIR)
#Check for NA values in Rasters
summary(PreMIR) # 8 NA
summary(PreNIR) # 0 NA

summary(PostMIR) # 2 NA
summary(PostNIR) # 0 NA

#fill the NA values of the vectorized Raster using na.fill.
note the values must be a vector or matrix
values.PreMIR <- as.vector(getValues(PreMIR))
```

```
PreMIR.nafill <- na.fill(values.PreMIR,"extend")
summary(PreMIR.nafill)

values.PreNIR <- as.vector(getValues(PreNIR)) #there are no
NA values to fill

values.PostMIR <- as.vector(getValues(PostMIR))
PostMIR.nafill <- na.fill(values.PostMIR,"extend")

values.PostNIR <- as.vector(getValues(PostNIR)) #there are
no NA values to fill

NBR_pre <- (values.PreNIR - PreMIR.nafill)/(values.PreNIR +
PreMIR.nafill)*1000 #note that the multiple of 1000 is by
convention for calculating NBR to put them in integer form.
summary(NBR_pre)

NBR_post <- (values.PostNIR -
PostMIR.nafill)/(values.PostNIR + PostMIR.nafill)*1000 #
summary(NBR_post)

dNBR <- (NBR_pre-NBR_post) #

#need to trim out the non-fire affected areas to ensure
that the resultant dNBR only accounts for those fire-
affected pixels in subsequent calculation of adNBR

dNBR_raster <- shell
names(dNBR_raster) <- "dNBR"
dNBR_raster[] <- dNBR
#writeRaster(dNBR_raster, filename="dNBR", format="GTiff",
overwrite=TRUE)

NBR_pre_raster <- shell
names(NBR_pre_raster) <- "NBR_pre"
NBR_pre_raster[] <- NBR_pre
#writeRaster(NBR_pre_raster, filename="NBR_pre",
format="GTiff", overwrite=TRUE)

NBR_post_raster <- shell
names(NBR_post_raster) <- "NBR_post"
NBR_post_raster[] <- NBR_post
#writeRaster(NBR_post_raster, filename="NBR_post",
format="GTiff", overwrite=TRUE)

#use the same MODIS Mask as was used in the caculation of
the Reveg Rate
NBR_stack <- raster::stack(dNBR_raster,NBR_pre_raster,
NBR_post_raster)
NBR_Masked <- mask(NBR_stack, MODIS_Mask)
saveRDS(NBR_Masked,"NBR_Masked")
```

```
values.of.NBR <- getValues(NBR_Masked)
#extract the individual layers
dNBR <- as.vector(values.of.NBR[,1])
NBR_pre <- as.vector(values.of.NBR[,2])
NBR_post <- as.vector(values.of.NBR[,3])

pre.dNBR.relation <- lm(dNBR ~ NBR_pre,
na.action=na.exclude)
summary(pre.dNBR.relation) #R-squared: 0.02414
coefficients(pre.dNBR.relation,1:2)
#plot the dNBR and pre NBR with the linear regression line
https://stackoverflow.com/questions/7549694/adding-regression-
line-equation-and-r2-on-graph
NBR.df <- as.data.frame(values.of.NBR)
library(ggpmisc)
my.formula <- dNBR~NBR_pre
bf1 <- ggplot(NBR.df,aes(x=NBR_pre,y=dNBR)) +
 geom_smooth(method = "lm", se=FALSE, color="black",
formula = my.formula) +
 stat_poly_eq(formula = my.formula,
        aes(label = paste(..eq.label.., ..rr.label.., sep =
"~~~")),
        parse = TRUE) +
 geom_point(colour = 'blue', size = 1)+
 geom_abline(slope=adNBRCoeff[2],intercept=
adNBRCoeff[1],color="red")+
 ggtitle("MODIS dNBR vs. NBR pre-fire")
bf1
####adNBR
adNBRCoeff <- coefficients(pre.dNBR.relation,1:2)
adNBRCoeff #use these to populate the expected dNBR
function used in calculating adNBR

exptdNBR <- adNBRCoeff[1]+adNBRCoeff[2]*NBR_pre

#Construct the adNBR
adNBR <- as.vector(dNBR-exptdNBR)


adNBR_raster <- MODIS_Mask #build a shell out of the MODIS
mask
adNBR_raster[] <- NA #ensure all values removed prior to
applying the adNBR values to the shell
names(adNBR_raster) <- "adNBR"
adNBR_raster[] <- adNBR
writeRaster(adNBR_raster, filename="adNBR", format="GTiff",
overwrite=TRUE)

#remove extreme RdNBR values beyond 2000 (ensure no INF
values as well)
near.zero <- which(NBR_pre<1 & NBR_pre>-1,arr.ind = TRUE)
#this identifies values that will greatly increase the
```

```
RdNBR value when calculating RdNBR as the RdNBR calculation
divides by the pre-fire NBR values.
ModNBR_pre <- NBR_pre
ModNBR_pre[ModNBR_pre%in%(ModNBR_pre[near.zero])] <- 1
#removes 3 values
ModRdNBR <- dNBR/sqrt(abs(ModNBR_pre/1000))
out.of.range <- which(ModRdNBR>1500 | ModRdNBR< (-500)) #
identifies 153 pixels that are more than 1500 and less than
-500.
ModRdNBR[ModRdNBR%in%(ModRdNBR[out.of.range])] <- NA
#removes 153 values or 0.47% of the data. These outliers
can create errors in the model and are removed to
facilitate building the Decision Tree Regression model.

#fill the removed values by NA.fill and correct overfill
using a mask

ModRdNBR.nafill <- na.fill(ModRdNBR,"extend")
tempRdNBR_raster <- MODIS_Mask
names(tempRdNBR_raster) <- "RdNBR"
tempRdNBR_raster[] <- ModRdNBR.nafill

plot(tempRdNBR_raster)
RdNBR_raster <- mask(tempRdNBR_raster,MODIS_Mask)
plot(RdNBR_raster)
writeRaster(RdNBR_raster, filename="RdNBR", format="GTiff",
overwrite=TRUE)

###############
#STEP 2
# BUILD THE RASTER STACK FROM ALL THE RASTER LAYERS
ASSOCIATED WITH THE DESCRIPTIVE ATTRIBUTES
###############
rm(list=ls())
#read in the data from the MODIS Mask folder
MODIS_Mask <-
raster('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\EVI\\M
ODIS\\Mask\\MODIS_Mask.tif')
MODIS_Mask[!is.na(MODIS_Mask)] <- 1

#read in the data from the attributes folder
https://gis.stackexchange.com/questions/136231/importing-
several-geotiff-files-into-r
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\Attribu
tes\\MODIS')
rlist=list.files(getwd(), pattern="tif$", full.names=FALSE)
for(i in rlist) { assign(unlist(strsplit(i, "[.]"))[1],
raster(i)) }
#clean up
rm(i)

# CONVERT RASTER STACK TO A DATA.FRAME
```

```
Attributes <- stack()# creates place holder stack

#loop to build the RasterStack

#Stack the attribute layers
 for(i in 1:NROW(rlist)){
  tempraster <- raster(rlist[i]) #note the tempraster is
important later
  Attributes <- raster::stack(Attributes,tempraster)
 }

Masked_attributes <- mask(Attributes,MODIS_Mask) #ensure
all layers have the same extent and boundary

Attributes.values <- getValues(Masked_attributes)

MD_Attribute.dataframe <- data.frame(Attributes.values)
MD_Attribute.dataframe$order <-
seq(len=nrow(MD_Attribute.dataframe)) #ensures the data
comes out of the model in the correct sequential order for
re-populating a raster frame.
head(MD_Attribute.dataframe,10)
#Merge veg ref table to change veg numbers to common name.
Note that the reference tables are built in ArcGIS as an
export of the attributes table of the soil and vegetation
layers.
Veg.lookup<-
as.data.frame(read.csv(file="Rim_veg_Table.csv",
header=TRUE))
head(Veg.lookup)
Veg.lookup$Rim_Veg <- as.factor(Veg.lookup$ï..Value)
merged_MD.Attribute <-
merge.data.frame(MD_Attribute.dataframe,Veg.lookup,
all.x=TRUE, sort = FALSE, by="Rim_Veg")
head(merged_MD.Attribute)

soil.lookup<-
as.data.frame(read.csv(file="Rim_Soil_Table.csv",
header=TRUE))
head(soil.lookup)
soil.lookup$Rim_SoilGrp <- as.factor(soil.lookup$Value)
merged_MD.Attribute <-
merge.data.frame(merged_MD.Attribute,soil.lookup,
all.x=TRUE, sort = FALSE, by="Rim_SoilGrp")
head(merged_MD.Attribute)
colnames(merged_MD.Attribute)[which(names(merged_MD.Attribu
te) == "WHR10NAME")] <- "Land_Cover"
colnames(merged_MD.Attribute)[which(names(merged_MD.Attribu
te) == "taxgrtgrou")] <- "Soil_Type"
head(merged_MD.Attribute)
```

```
excess <- -1*c(1,2,11,12,14) #the excess data columns to be
removed
MD_Attributes <-
merged_MD.Attribute[sort.list(merged_MD.Attribute$order),ex
cess]
head(MD_Attributes)
saveRDS(MD_Attributes,"MODIS_Attributes")
#MD_Attributes <- readRDS("MODIS_Attributes")
MD_Attributes_adNBR <- MD_Attributes[,-5]#removes the RdNBR
layer for the comparison between 30 m and 240 m spatial
resolution
head(MD_Attributes_adNBR)
saveRDS(MD_Attributes_adNBR,"MODIS_Attributes_adNBR")

# bring in the saved Modis Attribute table
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\Attribu
tes\\MODIS')
MD_Attributes_adNBR <- readRDS("MODIS_Attributes_adNBR")
############################
#STEP 3
# Grow Rgression Decision Tree
############################

Rattle(dataset="MD_Attributes_adNBR", useGtkBuilder=TRUE)
#Rattle version 4.1.0
#using a 95% Model 5% test split of the data (so that the
sample is approximately 1/20 of population) using a seed of
42, and a complexity parameter of 0.005 to build all
models.


########################
#SETUP LANDSAT BUILD
########################
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\EVI\\La
ndsat\\LANDSAT_DATA\\Extracts')

#read in the data from the folder
https://gis.stackexchange.com/questions/136231/importing-
several-geotiff-files-into-r
rlist=list.files(getwd(), pattern="tif$", full.names=FALSE)
for(i in rlist) { assign(unlist(strsplit(i, "[.]"))[1],
raster(i)) }
#clean up
rm(i)
#rm(ReGreen)
number.of.images <- length(rlist)
EVI <- stack()# creates place holder stack

#loop to build the RasterStack
ptimestack <- system.time({
 for(i in 1:NROW(rlist)){
```

```
  tempraster <- raster(rlist[i]) #note the tempraster is
important later
  EVI <- raster::stack(EVI,tempraster)
 }
})
ptimestack #4.75 sec
#plot(EVI)
tempraster
tempraster[] <- NA
names(tempraster) <- "tempraster"
saveRDS(tempraster,"tempraster")
rm(i)
rm(rlist)

values.of.stack <- getValues(EVI) #convert stack to matrix
summary(values.of.stack)
values.of.stack.df <- data.frame(values.of.stack)

#removeing bad/ poor/ problem/ cloudy images from analysis,
by filling them with NA values
values.of.stack.df$Extract_20143581[] <- NA
values.of.stack.df$Extract_20143261[] <- NA
values.of.stack.df$Extract_20150091[] <- NA
summary(values.of.stack.df)

values.of.stack <- as.matrix(values.of.stack.df)
save(values.of.stack,number.of.images,tempraster,
file="setup_Landsat")
saveRDS(values.of.stack.df,"setup_VoSdf")
saveRDS(number.of.images, "Nbr_Images")
saveRDS(tempraster,"tempraster")
rm(list=ls())
gc()
# values.of.stack <- readRDS("setup_VoS")
load("setup_Landsat")
c1 <- values.of.stack[,1]
c1.noNA <- na.omit(c1) #1156070 non-NA values x 74 images
(77 less the three cloud images)=

##############################
#Find the out of range pixels and process with: NA fill -
SG Smooth - NA fill
##############################

grt6500 <- na.omit(values.of.stack[values.of.stack>6500])
#62575 pixels
values.of.stack[values.of.stack>6500] <- NA
less0 <- na.omit(values.of.stack[values.of.stack<0])
#4211269 pixels
values.of.stack[values.of.stack<0] <- NA

#find the NA values
```

```
pre <- which(is.na(values.of.stack),arr.ind = TRUE) #find
the placement of the NA values
pre.row <- unique(pre[,1])

#Replace located NA with smooth fNA.fill.
ptimeNAfill <- system.time({for(i in pre.row) {
 temp.vector.row <- as.vector(values.of.stack[i,])
 if(!all(is.na(temp.vector.row))){#do if the row has values
  temp.vector.row <- na.fill(temp.vector.row, "extend",
maxgap = 3)
  values.of.stack[i,] <- temp.vector.row
 }
 next #otherwise skip to the next pre.row item
}
})
saveRDS(values.of.stack,"NAf_")
ptimeNAfill #about 630 sec

# apply S-G smoothing
ptimesmooth <- system.time ({
 for(i in 1:NROW(values.of.stack)) {
  temp.vector.row <- as.vector(values.of.stack[i,])
  if(!all(is.na(temp.vector.row))){
   smoothed.row <-
sgolayfilt(na.pass(temp.vector.row),p=7,n=9,m=0)
   values.of.stack[i,] <- smoothed.row
  }
  next
 }
})
saveRDS(values.of.stack,"NAf_Smooth")
ptimesmooth #740 sec ~ 12.3 min for LS : 30.75sec for MODIS

# based on earlier observations, valid EVI values fall
bewteen 0 and 6500
values.of.stack[values.of.stack<0] <- NA

#find the NA values
pre <- which(is.na(values.of.stack),arr.ind = TRUE) #find
the placement of the NA values
pre.row <- unique(pre[,1])

#Replace located NA with smooth NA.fill.
ptimeNAfill <- system.time({for(i in pre.row) {
 temp.vector.row <- as.vector(values.of.stack[i,])
 if(!all(is.na(temp.vector.row))){#do if the row has values
  temp.vector.row <- na.fill(temp.vector.row, "extend",
maxgap = 3)
  values.of.stack[i,] <- temp.vector.row
 }
 next #otherwise skip to the next pre.row item
}
```

```
})
saveRDS(values.of.stack,"NAf_Smooth_NAf")
ptimeNAfill
#values.of.stack <- readRDS('NAf_Smooth_NAf')
smooth.dataframe <- as.data.frame(values.of.stack)

##########################################
#Normalize the EVI values
##########################################
nPre <- 8 #number of LandSat or MODIS images prior to the
fire. This has to be set based on knowing about the data
set.

avg.prefire <- abs(rowMeans(smooth.dataframe[,1:nPre]))
#uses the first "nPre" columns of the data frame, i.e. the
images prior to the fire.
#apply the average prefire EVI values in order to normalize
the
normalized.smooth.dataframe <- smooth.dataframe/avg.prefire
#this normalizes the entire dataframe based on the prefire
average EVI values from the images. note that the data
frame can be divided by a vector but not a different sized
dataframe.
saveRDS(normalized.smooth.dataframe,"norm_smooth_dataframe"
)

PostFire.image.per.year <- 23 #total number of images in
each of the 12 month years after the fire for the sensor
(Landsat or MODIS)

#get the sum of the post year normalized EVI values in each
pixel
interval1 <- c((nPre+1):(nPre+PostFire.image.per.year))
interval2 <-
c((max(interval1)+1):(max(interval1)+PostFire.image.per.yea
r))
interval3 <-
c((max(interval2)+1):(max(interval2)+PostFire.image.per.yea
r))

year1 <- normalized.smooth.dataframe[,interval1]
year2 <- normalized.smooth.dataframe[,interval2]
year3 <- normalized.smooth.dataframe[,interval3]

year1.sum <- as.vector(rowSums(year1))#need to be vectors
for the cbind
year2.sum <- as.vector(rowSums(year2))
year3.sum <- as.vector(rowSums(year3))

#rebuild a dataframe that has all three years of summed EVI
values and calculate the ReGreen Rates
Postfire.reveg <- cbind(year1.sum,year2.sum,year3.sum)
```

```
Postfire.reveg <- as.matrix(Postfire.reveg)

years.gone.by <- as.vector(c(1,2,3))

Postfire.reveg[is.na(Postfire.reveg)] <- 0 #have to set NA
values to 0 for the lm function to work properly

tempslope <- vector(length = NROW(Postfire.reveg)) #is the
template into which each slope is placed
ptimeslope <- system.time({
 for(i in 1:NROW(Postfire.reveg)){
  temprow <- as.vector(Postfire.reveg[i,])
  tempfit <- lm(temprow ~ years.gone.by, na.action =
na.omit) #regression analysis of the relationship between
year and the cumulative annual normalized EVI values
  tempslope[i] <- tempfit$coefficients[2] #this returns the
slope of the regression line to the template
 }
})
saveRDS(tempslope,"tempslope")
ptimeslope # 1389 sec or ~27min for LS : 23sec for MODIS

tempslope[tempslope%in%(0)] <- NA #reapply the NA values to
those areas that were converted to 0 to calculate the
slopes

shell.raster <- readRDS("tempraster") #from Step 1 in
populating the global environment, this was a place holder
to build the RasterStack. It also has all the correct meta
data for creating a GeoTIFF with the ReGreen.rate vector
data.
shell.raster[] <- tempslope

Re.Green.Rate <- shell.raster
names(Re.Green.Rate) <- "ReGreen Rate"
plot(Re.Green.Rate)
writeRaster(Re.Green.Rate, filename="ReGreen",
format="GTiff", overwrite=TRUE)#putting the resulting
raster back into a geotiff in the working directory

tempslope[tempslope>10] <- NA #remove high slopes as per
visual inspection
tempslope[tempslope<(0)] <- NA #remove negative slopes as
per Cassidy et. al.

Re.Green.Rate.clipped <- readRDS("tempraster") #from Step 1
in populating the global environment, this was a place
holder to build the RasterStack. It also has all the
correct meta data for creating a GeoTIFF with the
ReGreen.rate vector data.
Re.Green.Rate.clipped[] <- tempslope
names(Re.Green.Rate.clipped) <- "Clipped ReGreen Rate"
```

```
plot(Re.Green.Rate.clipped)
writeRaster(Re.Green.Rate.clipped,
filename="ReGreen_clipped", format="GTiff",
overwrite=TRUE)#putting the resulting raster back into a
geotiff in the working directory

############################################################
#########################

# Gather Attributes for REGRESSION MODEL CONSTRUCTION

#
############################################################
#########################
#
# ###############
#STEP 1
#Calculate NBR, adNBR
###############
rm(list=ls())
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\NBR\\La
ndsat\\R-NBR')
#read in the data from the folder
https://gis.stackexchange.com/questions/136231/importing-
several-geotiff-files-into-r
rlist=list.files(getwd(), pattern="tif$", full.names=FALSE)
for(i in rlist) { assign(unlist(strsplit(i, "[.]"))[1],
raster(i)) }
#clean up
rm(i)
rm(rlist)
shell <- LS_Pre_Fire_NBR
shell[] <- NA

#Check for NA values in Raster
NBR_pre <- as.vector(getValues(LS_Pre_Fire_NBR))
NBR_post <- as.vector(getValues(LS_Post_Fire_NBR))

dNBR_raster <- (LS_Pre_Fire_NBR - LS_Post_Fire_NBR)

dNBR <- as.vector(getValues(dNBR_raster))

# inspect the NBR values
dNBR.df <- data.frame(dNBR,NBR_pre,NBR_post)

#calculate the best fit line between the Pre-fire NBR and
the dNBR to get the coefficients of that best fit line.
pre.dNBR.relation <- lm(dNBR ~ NBR_pre,
na.action=na.exclude)
summary(pre.dNBR.relation) #R-squared: 0.5241

####adNBR
```

```
adNBRCoeff <- coefficients(pre.dNBR.relation,1:2)#create a
matrix of the coefficent values
adNBRCoeff #use these to populate the expected dNBR fuction
used in calculating adNBR

exptdNBR <- adNBRCoeff[1]+adNBRCoeff[2]*NBR_pre # y = b +
mx where b is the intercept and m is the slope

#Construct the adNBR
adNBR <- as.vector(dNBR-exptdNBR)

adNBR_raster <- shell
adNBR_raster
names(adNBR_raster) <- "adNBR"
adNBR_raster[] <- adNBR
writeRaster(adNBR_raster, filename="adNBR", format="GTiff",
overwrite=TRUE)


###############
#STEP 2
# BUILD THE RASTER STACK FROM ALL THE RASTER LAYERS
ASSOCIATED WITH THE DESCRIPTIVE ATTRIBUTES
###############

setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\Attribu
tes\\LANDSAT')

#read in the data from the folder
https://gis.stackexchange.com/questions/136231/importing-
several-geotiff-files-into-r
rlist=list.files(getwd(), pattern="tif$", full.names=FALSE)
for(i in rlist) { assign(unlist(strsplit(i, "[.]"))[1],
raster(i)) }
#clean up
rm(i)

# CONVERT RASTER STACK TO A DATA.FRAME

Attributes <- stack()# creates place holder stack

#loop to build the RasterStack
ptimeAttrbt <- system.time({
 for(i in 1:NROW(rlist)){
  tempraster <- raster(rlist[i])
  Attributes <- raster::stack(Attributes,tempraster)
 }
})
ptimeAttrbt #

Attributes.values <- getValues(Attributes)
```

```
LS_Attribute.dataframe <- data.frame(Attributes.values)
LS_Attribute.dataframe$order <-
seq(len=nrow(LS_Attribute.dataframe))
head(LS_Attribute.dataframe,10)

LS_Attribute.dataframe$Rim_SoilGrp <-
as.factor(LS_Attribute.dataframe$Rim_SoilGrp)#though
attributes are numbers they are really factors
LS_Attribute.dataframe$Rim_Veg <-
as.factor(LS_Attribute.dataframe$Rim_Veg)

#Merge veg ref table to change veg numbers to common name
Veg.lookup<-
as.data.frame(read.csv(file="Rim_veg_Table.csv",
header=TRUE))
head(Veg.lookup)
Veg.lookup$Rim_Veg <- as.factor(Veg.lookup$ï..Value)
merged_LS.Attribute <-
merge.data.frame(LS_Attribute.dataframe,Veg.lookup, all.x =
TRUE, sort = FALSE, by="Rim_Veg")
head(merged_LS.Attribute)

soil.lookup<-
as.data.frame(read.csv(file="Rim_Soil_Table.csv",
header=TRUE))
head(soil.lookup)
soil.lookup$Rim_SoilGrp <- as.factor(soil.lookup$Value)
merged_LS.Attribute <-
merge.data.frame(merged_LS.Attribute,soil.lookup,
all.x=TRUE, sort = FALSE, by="Rim_SoilGrp")
head(merged_LS.Attribute)
saveRDS(merged_LS.Attribute,"LandSAT_Attributes")

merged_LS.Attribute <- readRDS("LandSAT_Attributes")
colnames(merged_LS.Attribute)[which(names(merged_LS.Attribu
te) == "WHR10NAME")] <- "Land_Cover"
colnames(merged_LS.Attribute)[which(names(merged_LS.Attribu
te) == "taxgrtgrou")] <- "Soil_Type"
head(merged_LS.Attribute)
excess <- -1*c(1,2,10,11,13) #the excess data columns to be
removed
LS_Attributes <-
merged_LS.Attribute[sort.list(merged_LS.Attribute$order),ex
cess]
head(LS_Attributes)
saveRDS(LS_Attributes,"LS_Attributes")
LS_Attributes <- readRDS("LS_Attributes")

# bring in the saved Landsat Attribute table
setwd('C:\\Users\\Jessica\\Documents\\Thesis\\Data\\Attribu
tes\\LANDSAT')
LS_Attributes <- readRDS("LS_Attributes")
```

```
############################
#STEP 3
# Grow Regression Decision Tree
############################

Rattle(dataset="LS_Attributes", useGtkBuilder=TRUE)#Rattle
version 4.1.0
#using a 95% Model 5% test split of the data (so that the
sample is approximately 1/20 of population) using a seed of
42, and a complexity parameter of 0.005 to build all
models.
```

The *Rattle* software creates a log of all executed code

```
# Rattle is Copyright (c) 2006-2015 Togaware Pty Ltd.

#============================================================
==
# Rattle timestamp: 2017-06-16 01:52:38 x86_64-w64-mingw32

# Rattle version 4.1.0 user 'Jessica.'
http://kamanja.org/forums/topic/r-pmml-generation-currently-has-a-bug-for-the-sv...

# This log file captures all Rattle interactions as R
commands. https://rdrr.io/cran/rattle/src/R/log.R
http://kamanja.org/forums/topic/r-pmml-generation-
currently-has-a-bug-for-the-svm-algorithm/

# We begin by loading the required libraries.

library(Rattle)  # To access the weather dataset and
utility commands.
library(magrittr) # For the %>% and %<>% operators.

building <- TRUE
scoring <- ! building


# A pre-defined value is used to reset the random seed so
that results are repeatable.
http://kamanja.org/forums/topic/r-pmml-generation-
currently-has-a-bug-for-the-svm-algorithm/

crv$seed <- 42

#============================================================
==

# Load an R data frame.

crs$dataset <- LS_Attributes
```

```
# Display a simple summary (structure) of the dataset.

str(crs$dataset)

#=============================================================
==

# Note the user selections.

# Build the training/validate/test data sets.

set.seed(crv$seed)
crs$nobs <- nrow(crs$dataset) # 2083200 observations
crs$sample <- crs$train <- sample(nrow(crs$dataset),
0.95*crs$nobs) # 1979040 observations
crs$validate <- sample(setdiff(seq_len(nrow(crs$dataset)),
crs$train), 0.05*crs$nobs) # 104160 observations
crs$test <- NULL

# The following variable selections have been noted.

crs$input <- c("adNBR", "cos_aspect", "elevation",
"flow_accum",
   "sin_aspect", "slope", "Land_Cover", "Soil_Type")

crs$numeric <- c("adNBR", "cos_aspect", "elevation",
"flow_accum",
   "sin_aspect", "slope")

crs$categoric <- c("Land_Cover", "Soil_Type")

crs$target <- "ReGreen"
crs$risk  <- NULL
crs$ident  <- NULL
crs$ignore <- NULL
crs$weights <- NULL

#=============================================================
==

# Decision Tree

# The 'rpart' package provides the 'rpart' function.

library(rpart, quietly=TRUE)

# Reset the random number seed to obtain the same results
each time.

set.seed(crv$seed)
```

```
# Build the Decision Tree model.

crs$rpart <- rpart(ReGreen ~ .,
  data=crs$dataset[crs$train, c(crs$input, crs$target)],
  method="anova",
  parms=list(split="information"),
   control=rpart.control(cp=0.005000,
    usesurrogate=0,
    maxsurrogate=0))

# Generate a textual view of the Decision Tree model.

print(crs$rpart)
printcp(crs$rpart)
cat("\n")

# Time taken: 35.04 secs

#=============================================================
==
# Evaluate model performance.

# RPART: Generate a Predicted v Observed plot for rpart
model on LS_Attributes.

crs$pr <- predict(crs$rpart, newdata=crs$dataset)

# Obtain the observed output for the dataset.

obs <- subset(crs$dataset, select=crs$target)

# Handle in case categoric target treated as numeric.

obs.rownames <- rownames(obs)
obs <- as.numeric(obs[[1]])
obs <- data.frame(ReGreen=obs)
rownames(obs) <- obs.rownames

# Combine the observed values with the predicted.

fitpoints <- na.omit(cbind(obs, Predicted=crs$pr))

# Obtain the pseudo R2 - a correlation.

fitcorr <- format(cor(fitpoints[,1], fitpoints[,2])^2,
digits=4)

# Plot settings for the true points and best fit.

op <- par(c(lty="solid", col="blue"))

# Display the observed (X) versus predicted (Y) points.
```

```
plot(fitpoints[[1]], fitpoints[[2]], asp=1, xlab="ReGreen",
ylab="Predicted")

# Generate a simple linear fit between predicted and
observed.

prline <- lm(fitpoints[,2] ~ fitpoints[,1])

# Add the linear fit to the plot.

abline(prline)

# Add a diagonal representing perfect correlation.

par(c(lty="dashed", col="black"))
abline(0, 1)

# Include a pseudo-R-square on the plot

legend("bottomright", sprintf(" Pseudo R-square=%s ",
fitcorr), bty="n")

# Add a title and grid to the plot.

title(main="Predicted vs. Observed
 Decision Tree Model
 LS_Attributes",
  sub=paste("Rattle", format(Sys.time(), "%Y-%b-%d
%H:%M:%S"), Sys.info()["user"]))
grid()

#============================================================
==
# Score a dataset.

# Obtain predictions for the Decision Tree model on
LS_Attributes.

crs$pr <- predict(crs$rpart, newdata=crs$dataset)

# Extract the relevant variables from the dataset.

sdata <- subset(crs$dataset, select=c("ReGreen"))

# Output the combined data.

write.csv(cbind(sdata, crs$pr),
file="C:\Users\Jessica\Documents\Thesis\LS_Attributes_score
_idents.csv", row.names=FALSE)
```