11-2022

Modeling Vegetation Cover at Mount St Helens Using Highly Correlated Vegetation Indices

Sara Livingston
University of Denver

DOI
https://doi.org/10.56902/ETDCRP.2022.3

Modeling Vegetation Cover at Mount St Helens Using Highly Correlated Vegetation Indices

Abstract

The eruption of Mount St. Helens in 1980 devastated the landscape and obliterated all ground vegetation within a 620 km2 blast zone radius. The destructive forces of the lateral blast, debris avalanche, tephra plume, and lahar flow created a complex mosaic of disturbance zones, that subsequently yielded various rates of landscape recovery. Remote sensing is an efficient method for monitoring landscape-scale changes by recording the distinct spectral reflectance of vegetation. Based on statistically significant correlations between Vegetation Indices and vegetation parameters, an empirical model can be developed for vegetation cover predictions. This capstone analysis found that NDVI holds the strongest relationship to vegetation cover when compared to other indices. Linear regression found that NDVI can account for 97.8% of vegetation cover variability when using a quadratic model (VegCover = 136.21(NDVI2) - 20.255(NDVI) - 0.1962).

Document Type

Masters Capstone Project

Degree Name

M.S. in Geographic Information Science

Department

Geography

First Advisor

Steven Hick

Second Advisor

Kristopher Kuzera

Keywords

Mount St. Helens, Vegetation index, Blast zone, Linear regression model, Empirical data model

Subject Categories

Geography | Physical and Environmental Geography | Plant Sciences

Modeling Vegetation Cover at Mount St Helens using Highly Correlated Vegetation Indices

Sara Livingston

University of Denver

Department of Geography and the

Environment MSGIS Capstone Project

November 2022

Abstract

The eruption of Mount St. Helens in 1980 devastated the landscape and obliterated all ground vegetation within a 620 km^2 blast zone radius. The destructive forces of the lateral blast, debris avalanche, tephra plume, and lahar flow created a complex mosaic of disturbance zones, that subsequently yielded various rates of landscape recovery. Remote sensing is an efficient method for monitoring landscape-scale changes by recording the distinct spectral reflectance of vegetation. Based on statistically significant correlations between Vegetation Indices and vegetation parameters, an empirical model can be developed for vegetation cover predictions. This capstone analysis found that NDVI holds the strongest relationship to vegetation cover when compared to other indices. Linear regression found that NDVI can account for 97.8% of vegetation cover variability when using a quadratic model **(VegCover = 136.21($NDVI^2$) - 20.255(NDVI) - 0.1962).**

Table of Contents

List of Tables

List of Figures

List of Acronyms

BZ	Blast Zone
DU	University of Denver
ESRI	Environmental Systems Research Institute, Inc.
EVI	Enhanced Vegetation Index
GPNF	Gifford Pinchot National Forest
GIS	Geographic Information System
GPS	Global Positioning System
GVI	Greenness Vegetation Index
MSAVI	Modified Soil-Adjusted Vegetation Index
MSH	Mount St Helens
MSH NVM	Mount St Helens National Volcanic Monument
NIR	Near-Infrared
NDVI	Normalized Difference Vegetation Index
OSAVI	Optimized Soil-Adjusted Vegetation Index
SAVI	Soil-Adjusted Vegetation Index
SR	Simple Ratio
TOPO	Topographic Map
USFS	United States Forest Service
USGS	United States Geologic Survey
VI	Vegetation Index

Introduction

When Mount St Helens violently erupted in May 1980, the cataclysmic blast exploded laterally across the landscape causing significantly more devastation to the surrounding area than what would typically occur from a vertical summit eruption. The surrounding landscape was physically reshaped and dramatically transformed by the violent volcanic eruption, with the destructive forces of the lateral blast, debris avalanche, tephra plume, and lahar flow creating a complex mosaic of varying disturbance zones (Tilling et al. 1990). In a fan-shaped area spreading out northward from the volcanic crater, roughly 153,000 acres (620 km2) became known as the blast zone (BZ) because all above-ground vegetation was obliterated, and the once thick coniferous Pacific Northwest temperate forest now resembled that of a “sterilized moonscape” (Harrington et al. 1998, 76).

In 1982 Congress established the Mount St Helens National Volcanic Monument (MSH NVM) with the primary purpose of allowing researchers an opportunity to study the ecological recovery processes without direct human interference and manipulation (Marzen et al. 2011). MSH NVM is often referred to now as a living laboratory because as life gradually returned to the varying disturbance zones surrounding the volcano, scientists were there taking measurements and documenting the details. The general problem of collecting data through field surveys and long-term study plots is that only localized revegetation processes are revealed, and in order to understand the broad landscape-scale change, other forms of data analysis are required (Marzen et al. 2011, 359-360). The technology of remote sensing offers an efficient and highly accurate method for analyzing spatial and temporal vegetation change throughout the entire BZ landscape (Xie, Sha, and Yu 2008).

Remote sensing applications can inventory, map, and monitor expansive land areas by recording the amount of electromagnetic radiation that is absorbed and reflected by landscape features. Vegetation has a unique spectral signature of reflectance values when measuring wavelengths in the visible and near-infrared (NIR) regions of the electromagnetic spectrum (Teltscher and Fassnacht 2018, 1852). Vegetation Indices (VIs) are mathematical expressions and customized algorithms that utilize various spectral combinations in order to extrapolate other vegetative parameters. The specific problem is that more than 100 different VIs have been developed to accommodate for various study objectives, satellite instrumentation, and landscape features and substrates (Xue and Su 2017, 1). With these variables in mind as they apply to the unique complexity of the volcanic blast zone, the purpose of this capstone project is to test several VIs for statistical accuracy in quantifying current vegetation cover. The vegetation index that obtains the highest correlation value to vegetation cover will then be presented in an empirical model that will allow vegetation cover to be calculated for any location within the blast zone.

Research Objective

There are two primary objectives for this capstone analysis. The first objective is to determine the vegetation index that can predict the current vegetation cover most accurately and the second, is to present those results in an empirical model for other scientists to utilize in their research. The seven vegetation indices tested in the analysis are the Normalized Difference Vegetation Index (NDVI), Soil-Adjusted Vegetation Index (SAVI), Modified Soil-Adjusted Vegetation Index (MSAVI), Simple Ratio (SR), Optimized Soil-Adjusted Vegetation Index (OSAVI), Enhanced Vegetation Index (EVI), and Green Vegetation Index (GVI), they are

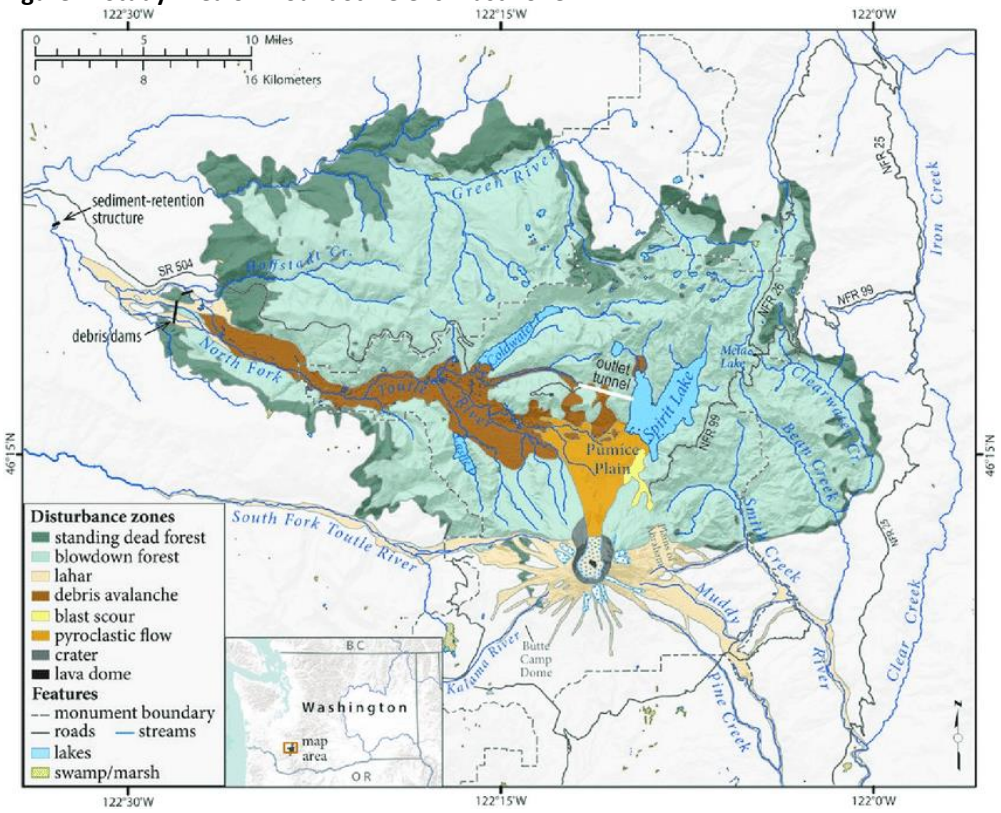
discussed with more detail in the following Literature Review section and the VIs mathematical equations are listed in Table: 1 Vegetation Index Equations. Vegetation cover, as defined as the total percentage of an area covered by vegetation, is a common attribute required for many scientific research studies, especially pertaining to the ongoing recovery process at MSH. For instance, in the field of forestry the parameter of vegetation cover is critical for understanding forest succession and how it varies across the mosaiced disturbance zones. Vegetation cover is also collected as a component for riparian surveys to assess stream and river temperatures as it relates to aquatic species like amphibians or fish. Vegetation cover is an important variable when monitoring small mammal and bird habitat as shrubs and trees are essential for their species success. By developing an empirical model constructed from the known and consistent relationship between vegetation indices and percent vegetation cover, a reliable VegCover variable can be determined for any point within the blast zone.

Study Area

Mount St Helens is located in southwest Washington state in Skamania County and is a part of the Gifford Pinchot National Forest but obtained official designation as the Mount St Helens National Volcanic Monument after the 1980 eruption. It is the youngest of the 18 stratovolcanoes that dot the Cascade Mountain Range running longitudinally from Canada to Northern California. The blast zone (BZ) is a fan-shaped 620 km^2 area that primarily spreads northward from the volcanic crater as shown in Figure 1: Study Area. All vegetation within the BZ was devastated by the volcanic eruption and it serves as the study area for this capstone analysis. The BZ can be further subdivided into regions based on the type and severity of eruption damage, such as the tree-down zone, seared zone, tree-removal zone, and the

pyroclastic-flow unit (Mazza 2010, 3). These subsections are not specifically delineated within the capstone analysis but are important components for conveying the nonhomogeneous nature of vegetation recovery and subsequent spectral variation affecting vegetation index accuracy. Another aspect that adds to the incongruent mosaic of vegetation recovery within the eruption area is that the BZ falls within 5 different management areas: the Gifford Pinchot National Forest, the MSH National Volcanic Monument, WA State land, private land, and the Weyerhaeuser Corporation; some of the private land was replanted with seedlings prior to the 1982 congressional designation of the MSH NVM (Marzen 2001, 53). Before the volcanic eruption, the mountain landscape was a dense, Pacific Northwest coniferous forest that consisted of dominantly noble and silver fir (*Abies procera* and *Abies amabilis*) above 3000 feet and mainly Douglas-fir (*Pseudotsuga menziesii*) below that elevation (Marzen et al. 2011, 362).

Figure 1: Study Area of Mount St Helens Blast Zone



(Kathryn Ronnenberg, USDA Forest Service, Pacific Northwest Research Station 2010)

Literature Review

Scientists have been using remote sensing satellite technology to monitor Earth's vegetative fluctuations since the 1960's and Carl Jordan is credited with implementing the first vegetation index named the Ratio Vegetation Index (RVI) in 1969 (Xue and Su 2017, 3). As the name implies, RVI is a ratio-based formula that divides the red wavelength by the near-infrared wavelength. RVI was renamed in subsequent literature as the Simple Ratio (SR), and it is built on the theory that vegetation foliage absorbs relatively more red light than infrared light and therefore the ratio's calculated value can estimate the amount of green biomass, with large SR values representing healthy vegetation and smaller values near zero representing soil, water, or ice (Xue and Su 2017, 3). There have been many progressive versions of the ratio-based formula over the last fifty years, but they each rely on the same principal fact: green vegetation strongly absorbs wavelengths in the red portion of the spectrum and strongly reflects in the near-infrared portion (Xie, Sha, and Yu 2008, 16).

The most commonly used and widely known ratio-based index is the Normalized Difference Vegetation Index (NDVI). Established in 1973 by Rouse, the NDVI ratio calculation takes the near-infrared radiation minus red radiation divided by near-infrared radiation plus red radiation (NIR-R)/(NIR+R), this formula is widely used because it identifies photosynthetic activity, meaning it can effectively delineate vegetation from non-vegetation (Huang et al. 2021, 2). NDVI values fall between -1 and 1 because the equation is itself normalized, with negative values generally being land characteristics of water or ice, values close to zero representing non-vegetative substrates like rock, sand, or concrete, and positive values signifying green vegetation that can include crops, grasses, shrubs, or forests (Huang et al.

2021, 2). Countless studies have demonstrated that NDVI is highly correlated to leaf area index (LAI), green biomass, leaf chlorophyll concentration, plant productivity, vegetation cover, and overall plant health (Xue and Su 2017, 3). A niche spinoff of NDVI that could prove to be helpful as a pre-analysis classification step with MSH's imagery, is the Normalized Difference Water Index (NDWI) that utilizes the green multispectral band instead of the red band, leading to a better suited index than NDVI for distinguishing characteristics of water and ice (Asokan and Anitha 2019, 153).

The first post-eruption multispectral image analysis of MSH was published in 1998, and the chosen metric for assessing vegetation recovery and change was NDVI. Harrington et al. (1998) used Landsat Multispectral Scanner System (MSS) imagery from 1979 to 1992 to calculate NDVI values in 2-year increments, with the imagery dates indicating that a vegetation assessment prior to the blast was part of the overall NDVI analysis of vegetation change (Harrington et al. 1998, 76). While the authors did use NDVI image differencing as the primary method for determining vegetation change, they did not offer their conclusions as quantified values of vegetation change but instead presented the NDVI imagery as only a visual aid for highlighting areas of maximum vegetation change or areas of no change (Harrington et al. 1998, 80). An important observation regarding their conclusion, however, is that Harrington et al. (1998) emphasized that they had difficulty with NDVI image interpretation when deciphering between the ash-covered ground and woody log debris as well as with precise water body boundaries (Harrington et al. 1998, 78). As mentioned earlier, a possible proposal to the authors could have been to pre-apply the NDWI as a tool to decipher between water body edges when obfuscated with ash and debris. As Marzen (2011) states, the uncertainty of

substrate interpretation is a common problem faced by researchers when using remotely sensed data because vegetation reflectance values are easily contaminated by soil reflectance as the percent of vegetation cover decreases (Marzen 2001, 26). A possible remedy for soil reflectance contamination, as suggested by Xue and Su (2017) is the Soil-Adjusted Vegetation Index (SAVI) that is geared towards NDVI deficiencies when describing the spectral contamination of vegetation reflectance if given a bright soil background (Xue and Su 2017, 4).

SAVI is suggested for use when the vegetation cover is sparse. In fact, Xue and Su (2017) state that NDVI should not be used if the total vegetation cover is below 30 percent as the spectral values will be contaminated with the brightness of the soil (Xue and Su 2017, 10). As the soil background brightness increases, the NDVI values will also increase, therefore a variable called the "soil conditioning index" is added to the NDVI equation to make the SAVI formula. The adjustment factor is based on the amount of vegetation cover that exists in the landscape, from 0 for high vegetation to 1 for low vegetation, but in the absence of extrinsic knowledge of the specific landscape conditions, an intermediate adjustment factor of 0.5 is suggested (Lawrence and Ripple 1998, 11). Lawrence and Ripple (1998) hypothesize that because of the barren to low vegetation at Mt St Helens within the first decade following the volcanic blast, SAVI would theoretically make a better performing index if compared to NDVI (Lawrence and Ripple 1998, 11). Other iterations of soil indices are also proposed with varying values as the soil conditioning factor, such as the Optimized Soil-Adjusted Vegetation Index (OSAVI) for general applications, the Modified Soil-Adjusted Vegetation Index (MSAVI) with a factor represented by the vegetation inverse value, and the Transformed Soil-Adjusted Vegetation Index (TSAVI) that utilizes the slope and intercept of the specific soil line (Xue and Su

2017, 5). The most common vegetation indices referenced in this paper and utilized in the analysis are listed in Table 1: Vegetation Index Equations (Marzen 2001, 27).

Table 1: Vegetation Index Equations

Vegetation Index	Formula
NDVI	(NIR – RED)/(NIR + RED)
SR	SR = NIR/RED
SAVI	SAVI = ((1 + L) * (NIR – RED))/(NIR + RED + L)
OSAVI	OSAVI = ((1.16*(NIR – RED))/(NIR + RED + 0.16)
MSAVI	MSAVI = ((2 * NIR) + 1) – (((2*NIR) +1)$^{2)}$ - (8*(NIR – RED)))$^{-1/2}$))/2
EVI	EVI = G * ((NIR - R) / (NIR + C1 * R – C2 * B + L))
GVI	GVI = - ((0.2848*BLUE) - (0.2435*GREEN) - (0.5436*RED) + (0.7243*NIR) + (0.0840*MIR(BAND5)) - (0.1800*MIR(BAND 7))

In the four decades following the eruption, hundreds of scientific studies have been completed about Mt St Helens in a vast array of subjects from wildlife biology and botany to geology and volcanology (Mazza 2010). Despite the abundance of groundbreaking literature available, there are fewer than ten published studies that have analyzed the vegetation recovery process using multispectral imagery and vegetation indices. Of those studies, all used NDVI as at least one metric for analyzing vegetation recovery and change. The most recently published article titled *Using multispectral landsat and sentinel-2 satellite data to investigate vegetation change at Mount St. Helens*, also incorporates the Urban Index (UI) (Teltscher and Fassnacht 2018, 1855). As Teltscher and Fassnacht (2018) note, the UI formula substitutes a midinfrared (MIR) wavelength for the red wavelength and it is best suited for identifying areas of bare ground (Teltscher and Fassnacht 2018, 1855). Referring to this index as UI can be confusing because it is also referred in similar literature as the Normalized Difference Built Index (NDBI) or the Normalized Difference Soil Index (NDSI) and it excels at highlighting the urban areas or built-up areas devoid of green vegetation (Asokan and Anitha 2019, 153).

NDVI appears to be the most popular vegetation index for assessing vegetation change both in general and at Mt St Helens. For vegetation assessments Huang et al. (2021) make the case that NDVI is the most popular index because it has a long and reliable history, the equation itself is simplistic and straightforward, and it uses a spectral wavelength that is commonly recorded by most airborne or spaceborne sensors (Huang et al. 2021, 2). Marzen, who has published multiple MSH vegetation analyses (2001, 2003, and 2011) provided two reasons why he chose NDVI for his studies; first, it is considered to be a standardized approach for estimating vegetation change therefore results can be easily compared and second, NDVI reduces the adverse topographic effects (Marzen 2011, 68). Independent from Marzen's specific statement about why he chose NDVI as opposed to other common vegetation indices such as TSAVI, NDSI, or MSAVI, only one other MSH researcher explicitly states their scientific reasoning for their choice. Lawrence and Ripple (1998) state that the lack of evidence demonstrating the most appropriate vegetation index for MSH's heterogeneous landscape, was itself the primary reason for why they chose to do a comparative index analysis (Lawrence and Ripple 1998, 92). Specifically, the authors wanted to analytically test the strengths and weaknesses of various vegetation indices that are widely known (NDVI) and designed to work well with differing substrate influences (SAVI) (Lawrence and Ripple 1998, 8).

In *Comparisons among Vegetation Indices and Bandwise Regression in a Highly Disturbed, Heterogeneous Landscape*, Lawrence and Ripple (1998) used statistical regression methods to determine which vegetation indices are the most accurate at describing vegetation cover at MSH. The vegetation indices that formulated their test hypotheses consisted of SR, NDVI, SAVI, MSAVI, TSAVI, GVI, and a non-indexed Bandwise regression method (Lawrence and

Ripple 1998). Lawrence and Ripple (1998) conclude that the Bandwise regression method that utilized raw, non-indexed spectral bands in a multivariate regression approach was the most accurate method for modeling vegetation change (Lawrence and Ripple 1998, 98). Many other authors, such as Teltscher and Fassnacht (2018) and Marzen (2011), both mention Lawrence and Ripple's (1998) conclusions because the multiple Bandwise regression method they used resulted in the red and near-infrared bands being the only spectral bands achieving statistical significance, the same spectral bands that formulate the NDVI equation (Lawrence and Ripple 1998, 98). Lawrence and Ripple (1998) state that the difference between the Bandwise regression method and the NDVI algorithm was in presentation only, as the Bandwise regression method was approached as a multivariable curvilinear model whereas the ratio based NDVI model maintained a linear relationship between the spectral bands (Lawrence and Ripple 1998, 98). The final conclusion of their comparative analysis found that all vegetation indices were highly correlated to percent green vegetation cover, the Bandwise regression model using individual spectral bands was most accurate at explaining vegetation variability and the ratio-based indices of SR and NDVI outperformed the soil-lined indices of TSAVI, MSAVI, and OSAVI (Lawrence and Ripple 1998, 101).

The just discussed comparative vegetation index analysis put forth by Lawrence and Ripple (1998) offered quantifiable results that specifically address the core objective proposed for this analysis, that is, what vegetation index is best suited for assessing vegetation cover at Mount St Helens? Other comparative type analyses not centered at MSH can be just as informative. Throughout the decades of literature concerning this topic, comparative studies are often repeatedly referenced as supporting literature for other vegetation change

assessments. For instance, one article authored in 1998 that has held the scrutiny of time, has been cited more than 500 times in other vegetation change literature because it comparatively evaluates seven vegetation indices: DVI, NDVI, PVI, RVI, SARVI, SAVI, and TSAVI (Lyon et al. 1998, 144). Lyon et al. (1998) determined that NDVI was the only vegetation index among the seven tested that resulted with normally distributed histograms and it was the index least affected by topography (Lyon et al. 1998, 149). A more recent comparative study proposed by Joshi (2011) tested three vegetation indices' (NDVI, TDVI, and SAVI) ability to differentiate between a variety of land cover classes, from water and ice to sparse and dense vegetation (Joshi 2011, 234). The conclusion reached by Joshi (2011) was that NDVI gave the best results in terms of overall accuracy for identifying vegetation and it clearly classified sparse vegetation given the bright soil background (Joshi 2011, 240).

The scientific literature depicting the natural revegetation process at MSH, as assessed using remote sensing and vegetation indices, are limited in number but similar volcanic eruption events can be used as comparative tools for understanding methods and techniques. Mt Pinatubo is an active volcano on the island of Luzon in the Philippines, and it experienced a catastrophic eruption in 1991. Mt Pinatubo is a stratovolcano just like Mt St Helens, and similarly, the eruption produced large volumes of volcanic material that included pyroclastic flow, ash deposits, and lahar flow (DeRose et al. 2011, 9281). In 2011, DeRose et al. (2011) used multiband satellite imagery to investigate vegetation change using NDVI for quantification of ground cover losses and gains for a 10-to-16-year time span (DeRose et al. 2011, 9288). While DeRose et al. (2011) did not specifically state their reasoning as to why NDVI was chosen compared to other indices, two observations stick out in their analysis that could be assumptive

reasons for their certainty in NDVI. One, the multispectral imagery selected in the first image sequence was ten years after the eruption occurred, this could mean that the authors had more certainty that a barren landscape would not contaminate the NDVI calculations with brightness, as can often happen with sparsely vegetated landscapes. Two, DeRose et al. (2011) state that for NDVI to be an accurate vegetation measure, a brightness correction and an NDVI spectral calibration was completed on the imagery beforehand (DeRose et al. 2011, 9290). The brightness-correction adjustment was made to account for variations in satellite-sun geometry and the NDVI spectral calibration used a linear regression formula to fix discrepancies between image dates of similar landcover classes, such as bare soil, channel debris, and water bodies (DeRose et al. 2011, 9290).

Oldoinyo Lengai (OL) is a stratovolcano located in north Tanzania in east Africa. The volcano went through a 10-month long vulcanian-style eruption in 2007 to 2008 where the eruptive phase was characterized by repetitive and short-lived ash eruptions (DeSchutter et al. 2015, 3). The amount of ash deposited on the landscape varied significantly in all directions, but even in moderate eruptions the prolonged ash fallout can cause severe impacts to vegetation due to ash burial and overloading as well as physical or chemical changes that can affect plant growth (DeSchutter et al. 2015, 4). DeSchutter et al. (2015) completed a temporal analysis using multispectral imagery to quantify the vegetation change proceeding the eruption and similar to the research at Mt Pinatubo in the Philippines, NDVI was again the chosen metric for calculating vegetation change (DeSchutter et al. 2015, 4). The reasoning that the authors gave for choosing NDVI over a soil adjusted vegetation index was that their own literature review conducted

before the analysis indicated that SAVI had not always outperformed NDVI in relating biophysical properties to ground vegetation (DeSchutter et al. 2015, 4).

Over 100 Vegetation Indices have been developed and implemented for remote sensing applications. VIs are effective algorithms for quantitative and qualitative evaluations of various vegetative characteristics and the use of each can be highly specific to the parameters of the landscape (Xue and Su 2017, 1). To summarize the literature just reviewed, Xue and Su (2017) outlined many indices from the first VI implemented in 1969 by Jordan to indices specific to satellite platforms, sensors, or even agricultural crop type (Xue and Su 2017). The most commonly used index is the Normalized Difference Vegetation Index (NDVI) and although it is advantageous to use for general standardized purposes, it also has deficiencies that have been documented as well. In the less than ten research articles published about vegetation change at Mount St Helens using VIs, each study used NDVI as a metric for analyzing change. Lawrence and Ripple (1998) published the most informative and comprehensive analysis comparing eight vegetation indices using multispectral imagery. The multiple Bandwise regression method using non-indexed spectral bands was their overall conclusion for most accurate at explaining variation, and the ratio-based indices of SR and NDVI outperformed the soil-adjusted indices of SAVI, OSAVI, and MSAVI (Lawrence and Ripple 1998). In a vegetation change analysis of volcano's Mt Pinatubo in the Philippines (DeRose et al. 2011) and Oldoinyo Lengai in north Tanzania (DeSchutter et al. 2015) both authors used NDVI and found that method to be highly accurate as well.

Data Acquisition

This multispectral vegetation analysis is built upon three foundational data blocks. The Landsat 8 multispectral imagery serves as the data input for calculating each vegetation index, NAIP imagery is used for determining percent vegetation cover, and two GIS vector data layers represent the sample site locations and the BZ boundary of the study area. The vector layers are produced using ArcGIS software while the imagery is acquired from outside sources.

Multispectral Imagery

NASA's Landsat Program has been recording spectral wavelengths since launching its first satellite into space in 1972. Although Landsat 9 is now the active orbiting satellite, the data for this analysis comes from the Landsat 8 Operational Land Imager (OLI). The OLI measured visible, near infrared, and shortwave infrared portions of the spectrum on a repeating cycle every 16 days (EROS 2021). The spectral band designations and wavelengths for Landsat 8 that are applied in the analysis are Band 1 (Coastal Aerosol) 0.43-0.45 µm, Band 2 (Blue) 0.45-0.51 µm, Band 3 (Green) 0.53-0.59 µm, Band 4 (Red) 0.64-0.67 µm, Band 5 (Near-Infrared) 0.85-0.88 µm, Band 6 (Shortwave Infrared 1) 1.57-1.65 µm, and Band 7 (Shortwave Infrared 2) 2.11-2.29 µm (EROS 2021). The spatial resolution is 30m in size, an indication that the resolution is too course for minute details, but the coarseness is what allows it to have global scale coverage that can accurately characterize Earth's processes of change. Landsat multispectral imagery was specifically chosen for this analysis because of the long history and expected continuity of image capture, which allows for a more seamless comparison should a temporal analysis using other multispectral images follow. There is significant variability that exists between satellite sensor calibration and preprocessing calculations that it is recommended for multitemporal

studies to maintain use of the same imagery source if available in order to ensure that any variability captured in an analysis is due to temporal land changes and not due to satellite sensor differences.

The multispectral images come from USGS Earth Explorer data portal, an expansive repository for satellite imagery and other types of landcover data. Mount Saint Helens is located in path 46, row 28. The downloaded satellite images are classified as Landsat 8 OLI/TIRS Collection 1 Level 1 and Level 2 U.S. Analysis Ready Data (ARD). Images were downloaded for various dates throughout May, June, July, and August, as image quality is a major factor in the final date selection. The acquisition date of 25 June 2021 was the concluding choice for multispectral imagery, which was also based on the availability of NAIP imagery that is discussed in the next section. Other prerequisites of image quality pertained to cloud or haze visibility over the study area and also the important factor that the image is captured during the peak vegetative growing season in order to assess vegetation accurately.

NAIP Imagery

The National Agriculture Imagery Program (NAIP) is administered by the U.S. Department of Agriculture's Farm Service Agency. NAIP imagery is collected on a 2-to-3-year cycle during the agricultural growing season called the "leaf on" conditions (USGS 2021). It is commonly referred to as aerial imagery or digital ortho photography because the images are high-resolution photographs that are orthorectified as a geographic map (USGS 2021). The resolution difference of NAIP images and Landsat multispectral images is not a close comparison. As mentioned previously, the spatial resolution or ground distance of multispectral imagery is 30 meters (900 m^2 total area per pixel) and the resolution of NAIP imagery is 0.6

meters (0.36 m^2 or 4 ft^2). The clarity provided by NAIP imagery allows for small-scaled details to be visible such as individual shrubs and trees, which is something that Landsat multispectral images cannot provide. As seen by Figure 2 the difference in spatial resolution between Landsat multispectral imagery and NAIP imagery is striking. The wispy-looking meanders shown in the river delta are visible in the NAIP imagery, but the multispectral imagery offers no clarity for that finite detail. Individual houses, buildings, roads, and agricultural fields are also visible in the NAIP imagery and not in the corresponding multispectral images. The acquisition date for the NAIP photography is June 23, 2021, an almost perfect alignment with the Landsat acquisition date.

Figure 2: Spatial Resolution of NAIP Imagery and Multispectral Imagery

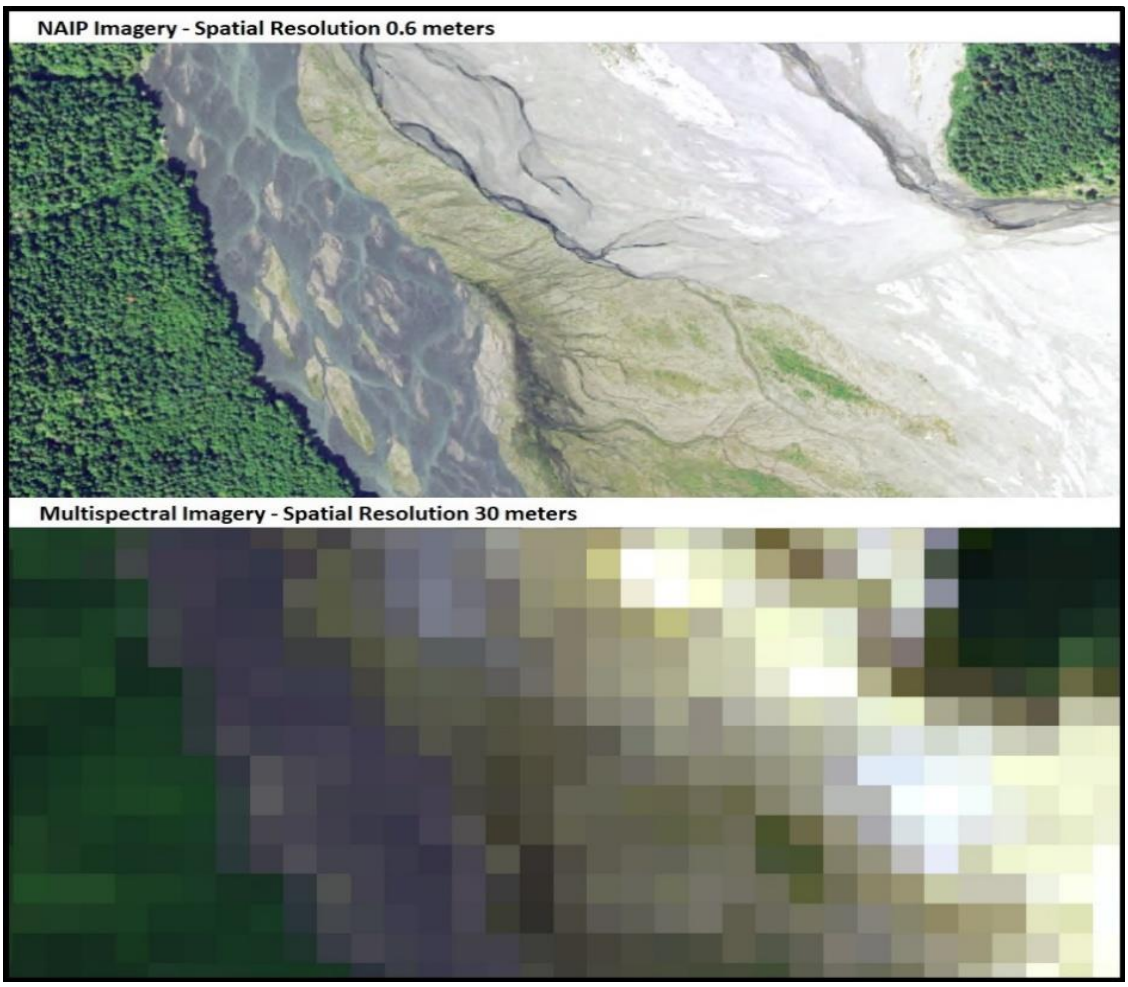
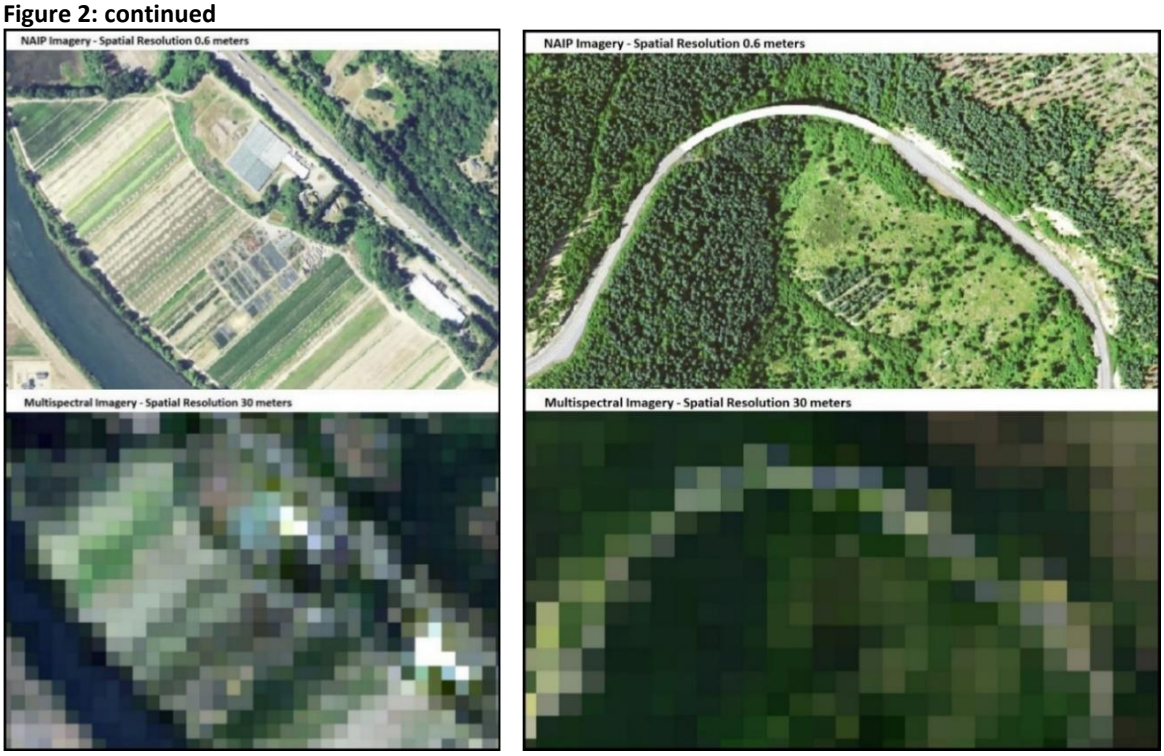


Figure 2: continued



Methods

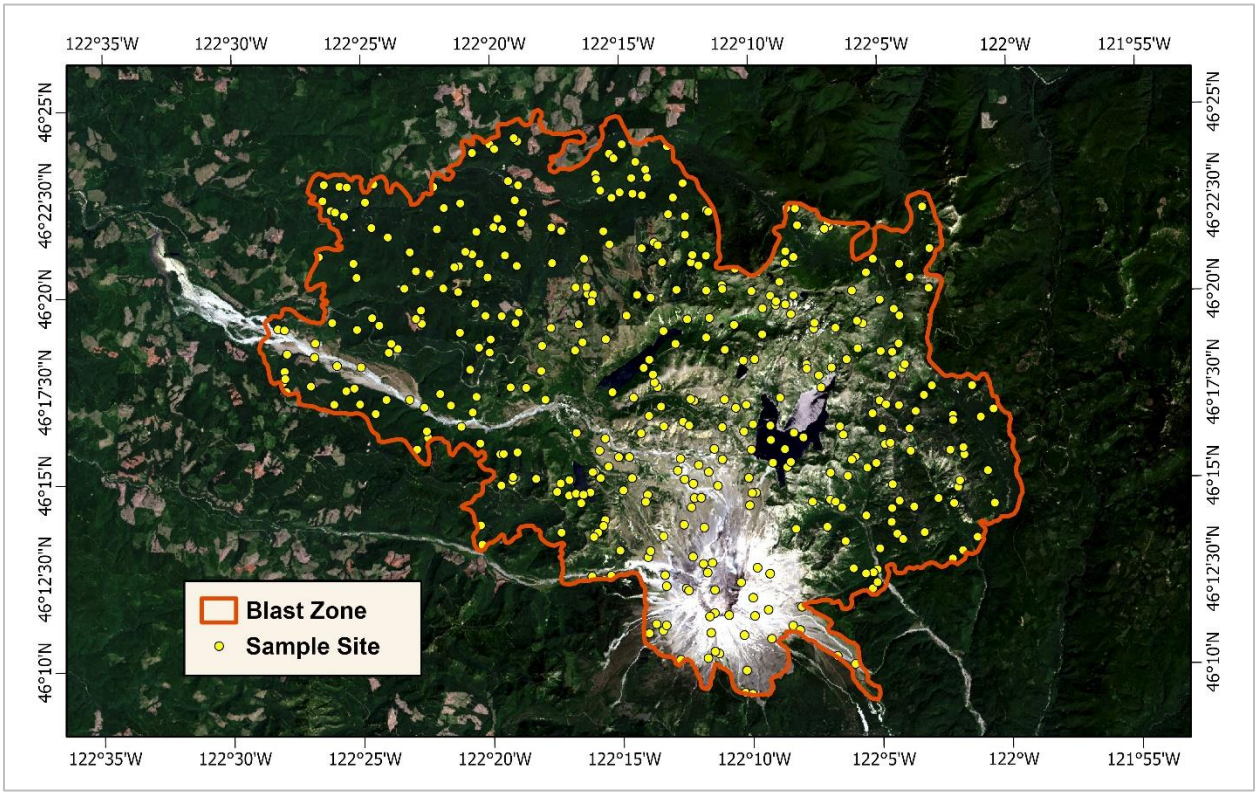
The methodological approach used in this analysis is based on the scientifically known assumptions that a consistent relationship exists between the amount of vegetation cover and the vegetation index. The relationship between the vegetation cover (VegCover) and vegetation indices (VIs) are investigated using an estimation of percent vegetation cover and the surface reflectance values recorded by the multispectral sensors. To accomplish that research objective, the following methods were applied. First, the MSH blast zone boundary is defined, then a number of random sample sites are established within that boundary. Both NAIP imagery and Landsat imagery are obtained with approximately similar time stamps. The high-resolution orthophotos are used for estimating percent vegetation cover for each pixel containing a sample site location. Seven different vegetation indices are calculated from the

individual multispectral bands and then the calculated pixel value is extracted using a GIS multipoint extraction tool. Finally, the correlated relationship between percent vegetation cover and index value is evaluated by applying a linear regression analysis technique. The regression analysis results are reported as an empirical model using the vegetation index that obtains the highest correlation value. These steps are explained with detail in the following sections and a flow chart depicting the methodological process is shown in Figure 4: Model Builder Diagram and Methodology.

Sample Site Locations

The study area as defined in this analysis, is the roughly 620 km^2 where all vegetation was obliterated, either by blow down force, searing heat, pyroclastic flow, or buried with lahar material. Using ArcGIS Pro software, the blast zone boundary was digitized from a United States Geological Survey (USGS) digital topographic map of the eruption area. It is a vector layer projected to NAD 1983 UTM Zone 10N. Once the blast zone is digitized, the total number of counted pixels and possible sample sites available are roughly 700,000. Using a 95% confidence interval and a margin of error of 5%, it is determined that at least 350 samples are required to achieve statistical significance (Equation 1, Appendix 1). In ArcToolbox a random point generator is used to create 425 random sample sites within the blast zone perimeter, an extra 75 sites are added to account for any potential exclusions if located within the volcanic crater, in a water body, or on a summit glacier. A sample site is composed of 1 pixel each and they are given X-Y coordinates and projected to NAD 1983 UTM Zone 10N. The Blast Zone and Sample Site vector layers are the only geographic features needed in the analysis other than the VI raster images. Figure 3 below shows the Blast Zone Boundary and the Sample Sites.

Figure 3: Blast Zone Boundary and Sample Sites

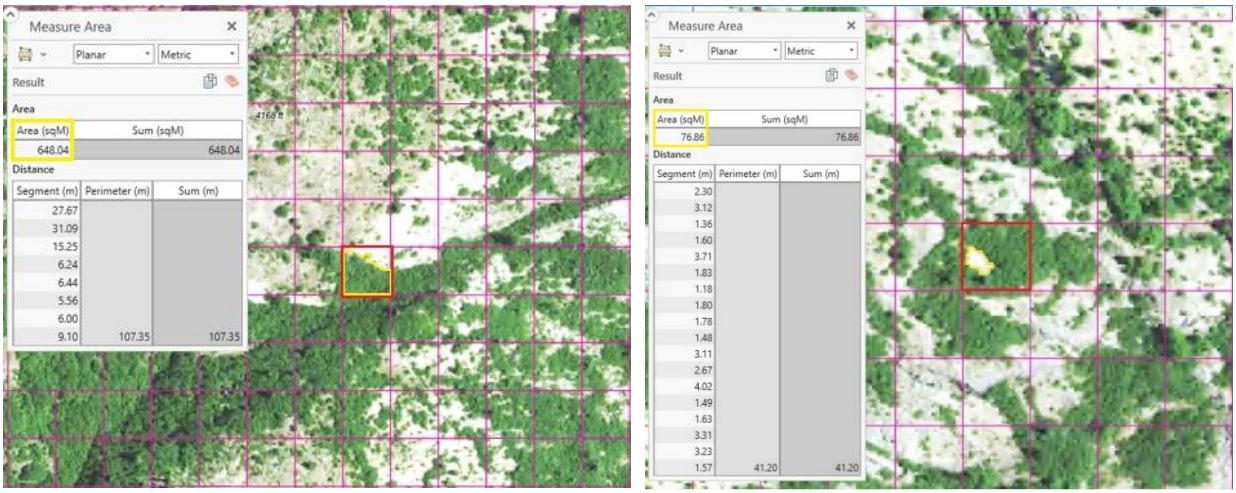


Percent Vegetation Cover

The high resolution NAIP imagery is used to estimate percent vegetation cover. A grid pattern is constructed with the same 30 m spatial dimensions as the multispectral pixel, and it serves as the guide for estimating vegetation cover for each sample site. Percent vegetation cover is estimated using two primary methods, direct polygon measurements of spatial area and the use of Vegetation Cover Density Diagrams. The most precise technique is to use ArcGIS Pro spatial tools to measure the area with or without vegetation. Patches of green vegetation contained in the 900 m^2 area of the pixel can be traced using a polygon measuring tool, with the ability to make multiple additive measurements within the same pixel. As seen by the screen captures shown in Figure 4, the red-colored square displays a yellow highlighted vegetation measurement of 648 m^2, which calculates to roughly 72% VegCover. Or, in in the

second image, the highlighted 76.8 m^2 polygon contains no vegetation and that is subtracted from the total 900 m^2 pixel area, totaling 91.5% VegCover. The second method for estimating VegCover is using established Vegetation Cover Diagrams, such as the one shown in Figure 12 in Appendix 2. The diagrams are used when the vegetation is sparsely dispersed in density and an exact polygon measurement is not feasible. Lastly, site visits were made to the mountain throughout June, July, and August 2021 with aerial photographs in order to delineate various vegetation cover amounts to understand how that translated from on-the-ground to photograph, noting vegetation density and classification of either shrub, grass, ground cover, conifer, deciduous, or mixed. The preferred method of these three options is the spatial polygon measurement using the high-resolution NAIP imagery. All precent VegCover is estimated to the nearest 5 percent increment to account for any finite visual interpretation errors. If there is no vegetation within the pixel a 0 percent cover amount is recorded and sample sites that are located on waterbodies, glaciers, or inside the volcanic crater are excluded from the regression analysis. The vegetation cover estimates at each sample site were completed prior to any vegetation index calculation in order to prevent unbiased results.

Figure 4: Polygon Area Measurement for Percent Vegetation Cover



Multispectral Processing

In most multispectral image analyses the preprocessing of the satellite images usually require radiometric, atmospheric, or topographic corrections prior to any spectral calculations. The acquisition of Landsat 8 OLI/TIRS Collection 1 Level 1 data is already radiometrically corrected by USGS before making available in the Earth Explorer portal. A few important terms to define prior to the analysis are digital number value, radiance, and reflectance - including top of atmosphere reflectance and surface reflectance. Digital number (DN) value is the generic term for pixel value and the numbers have no physically meaningful unit. If the image needs to be interpreted or compared with other images, especially between different satellites, the DN values need to be converted into a quantitative value such as surface reflectance. Most of the vegetation index equations that contain coefficients within the formula, such as SAVI or MSAVI, are based on surface reflectance coefficients and not the digital number value provided. Radiance is the amount of radiation coming from an area on the Earth's surface recorded by a satellite sensor. Information contained in an images metadata provide a *gain* and an *offset* value specific to the satellite sensor and that allows for conversion from a digital number value to a radiance value (EROS 2020). Reflectance is the proportion of the radiation striking a surface to the radiation reflected from it. There is the Top of Atmosphere reflectance (TOA) that measures the reflectance above the clouds and atmospheric aerosols, and the Surface Reflectance (SR) that have had the atmospheric components radiometrically corrected (EROS 2020). The Surface Reflectance numbers are the values needed for accurate vegetation index calculations and are used in this regression analysis.

The acquisition of Landsat 8 OLI/TIRS Collection 1 Level 1 data is provided in digital number values and would typically need the just discussed conversion to surface reflectance values. However, new imagery data is now available from USGS's Earth Explorer called Level 2 U.S. Analysis Ready Data (ARD), it has already been processed with the various atmospheric and radiometric corrections and then converted from radiance to surface reflectance. The downloadable data is available in Surface Reflectance values, but another scaling factor needs to be applied. Landsat Collection 1 Level 2 can be rescaled according to the specific values provided with the Landsat 8 Collection 1 (C1) Land Surface Reflectance Code (LaSRC) Product Guide manual (EROS 2020). The scale factor calculations are performed using ArcGIS Pro raster algebra functions, as well as removing any erroneous pixels outside the valid scale range. Mount St Helens happens to be split perfectly across the crater, and image processing requires two multispectral images to be mosaiced together.

Vegetation Index Calculations

All image preprocessing and vegetation index calculations were completed using Hexagon's ERDAS Imagine and supplemented with tools from ArcGIS Pro software. The level 2 Landsat multispectral image from 25 July 2022 covers an area of roughly one quarter of Washington state. Two images are acquired and mosaiced together to capture entire MSH blast zone area. Prior to any vegetation index calculation, the image is reprojected to NAD 1983 UTM Zone 10N and subset to a smaller sized square encompassing the BZ for concise data manageability. The seven vegetation indices included in the analysis are the Normalized Difference Vegetation Index (NDVI), Soil-Adjusted Vegetation Index (SAVI), Modified Soil-Adjusted Vegetation Index (MSAVI), Simple Ratio (SR), Optimized Soil-Adjusted Vegetation

Index (OSAVI), Enhanced Vegetation Index (EVI), and Green Vegetation Index (GVI); all equations according to spectral bands are shown in Table 1: Vegetation Index Equations. In ERDAS Imagine a layer stack including spectral bands 1 to band 7 is constructed for the purpose of simultaneous accessibility of individual spectral band values as a multifunction when preforming VI equations.

There are multiple methods available for calculating vegetation index values. In this analysis, the primary method utilized was to input the equation into ArcGIS Raster Calculator. This method required all seven SR raster layers of band 1 to band 7 be accessible as individual data layers within the GIS content menu. The equations shown in Table 1 represent the Python syntax entered into the raster calculator geoprocessing tool. Each vegetation index function produces a separate raster surface where every pixel has been calculated according to the vegetation index equations. After the seven vegetation index raster images have been calculated, the pixels that correspond to a sample site location are simultaneously extracted. Using the spatial analyst toolset, a total of fourteen new fields are added to the attribute table of the sample site vector layer as instructed using the multipoint extraction function, a pixel value for spectral band 1 to spectral band 7, and a calculated value for each index. Extracting values for band 1 to band 7 serves the purpose of being a quality assurance method if erroneous data points need to be investigated as outliers in the model. The attribute table is then exported into an Excel spreadsheet for regression modeling within Excel and SPSS software.

Linear Regression Model

Linear regression analysis is a common statistical method used when constructing prediction models. It measures the correlation that exists between two or more variables, a dependent response variable and one or more independent explanatory variables. The correlation coefficient (r), referred to as the Pearson's product-moment correlation coefficient (PPMCC) measures the strength of the correlation between two variables (Equation 2, Appendix 1). If a correlation exists, then the Coefficient of Determination (R^2) is calculated to indicate the statistical prediction power that the model offers in predicting an independent variable based on a given explanatory variable (Equation 4, Appendix 1). More specifically, the regression technique used is called Ordinary Least Squares regression (OLS) and it minimizes the likelihood of error differences between the variables of the model. The closer the R^2 value is to 1 indicates how much variability the dependent variable can be explained if given the known independent variable.

It has been well documented that vegetation cover is a variable that is highly correlated to various vegetation indices. Using MSH's vegetation data from 1995 multispectral images, Lawrence and Ripple found that the Normalized Difference Vegetation Index produced an R^2 value of 0.704 and the Simple Ratio vegetation index had an R^2 value of 0.698 when constructing their predictive curvilinear regression models relating vegetation index values and vegetation cover (Lawrence and Ripple 1998). Lawrence and Ripple are the only authors over the last forty years following the eruption to construct regression models aimed at predicting vegetation cover using various vegetation indices. Countless other scientific studies have explored the known relationship between vegetation indices and vegetation cover. For

example, Purevdorj et al. (1998) found correlation coefficients ranging from r= 0.89 to 0.99 for the vegetation indices of SAVI, TSAVI, MSAVI, and NDVI to percent vegetation cover at various study sites in Mongolia. Fathoni at el. (2021) found a strong correlation of r= 0.93 when constructing an empirical model of vegetation cover using NDVI on Mount Agung in Bali after its 2017 volcanic eruption (Fathoni et al. 2021). Both Lawrence and Ripple (1998) and Purevdorj et al. (1998) mention, however, that the correlational relationship values increase significantly after incorporating polynomial terms into the empirical modeling equation. One theory for this factor is that the spectral response of vegetation saturates at a certain point and that resembles a curvilinear relationship across the spectral range (Lawrence and Ripple 1998, 99). The regression model for this capstone analysis uses similar regression techniques as demonstrated by these authors as well as statistical experimentation with including polynomial terms.

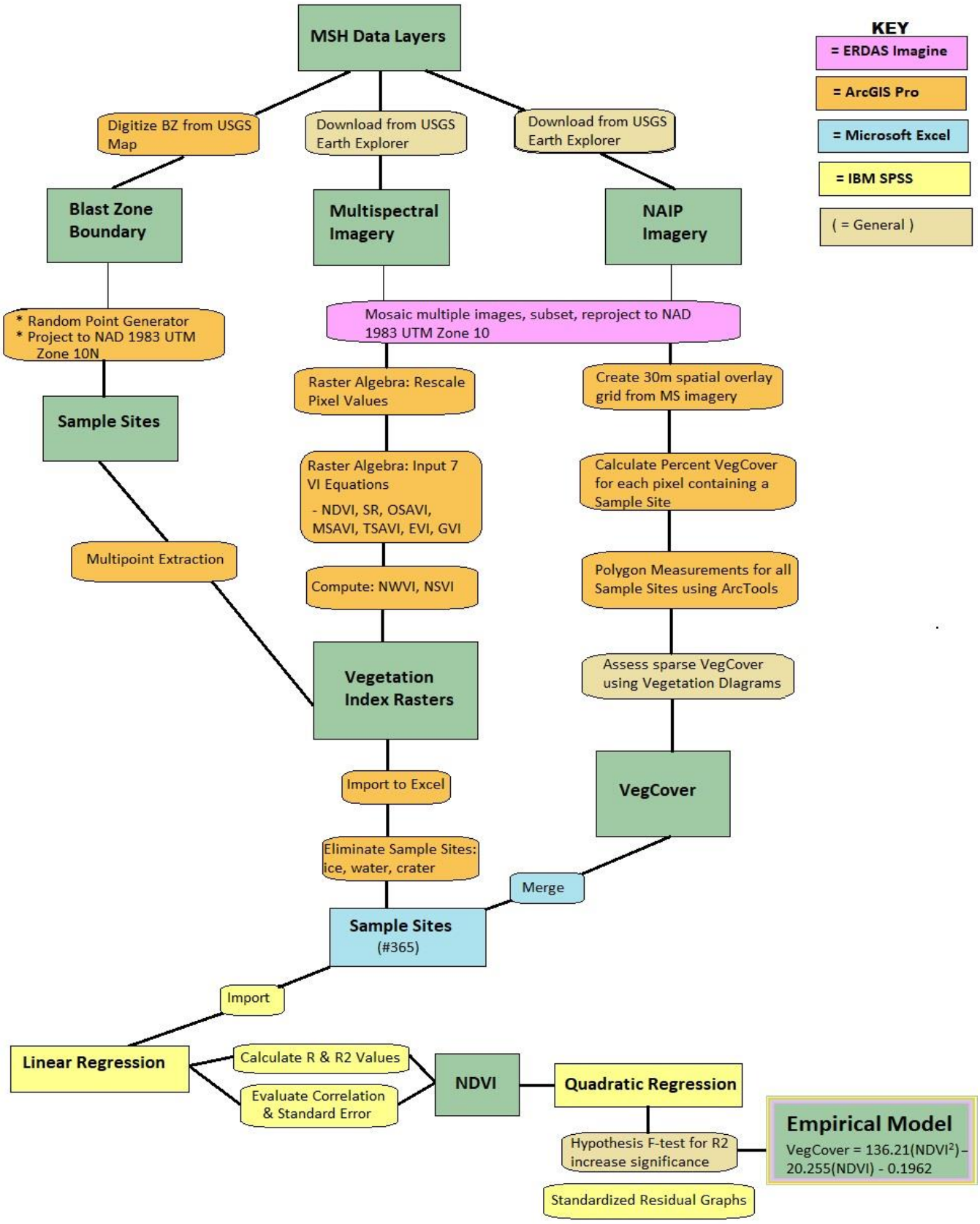
The independent variable within the blast zone is the vegetation index as calculated with spectral band equations and the dependent values are the percent vegetation cover. The correlation value can be calculated manually by entering statistical equations into an Excel spreadsheet containing the independent and dependent variables, but for the purpose of testing second-order polynomial variables, IBM's SPSS software is used. The equations used to calculate the correlation value (r) and coefficient of determination (R^2), which includes the standard error of the estimate, are listed in Appendix 1. In total, seven different vegetation indices are regressed against the estimated percent vegetation cover to determine which index most accurately models the relationship. An X-Y scatterplot of the independent and dependent sample points can be used as a rudimentary method for investigating if a second or third order

polynomial term should be incorporated within the model. The resulting linear equation can be considered an empirical model because it relies on tested data points rather than a theoretical model build on theory. The conclusion of this capstone analysis is to construct an empirical model that can compute the vegetation cover percent for any point within the blast zone if given a vegetation index value. Each Landsat satellite repeats its orbital pattern every 16 days, allowing for the ability to always have access to current vegetation cover estimates for any scientific study within the blast zone.

Methods Flowchart

The following graphic (Figure 5) is a visual representation that outlines the methods just described in the preceding paragraphs. The flowchart begins with the top square labeled MSH Data Layers, leading to the acquired primary layers of Multispectral Imagery and NAIP Imagery, as well as the digitized Blast Zone Boundary and randomly created Sample Sites. There are four major software packages utilized in the analysis that are color coded for flowchart processes simplicity. The pink-colored steps were completed with Hexagon's ERDAS Image v.2018, the orange color represents ArcGIS Pro 2.3, blue signifies Microsoft Excel, yellow is IBM's SPSS, and the tan color represents generalized nonspecific steps. As the flowchart works its way down the page, the final step displays the Empirical Data Model equation, and it represents the current vegetation cover at Mount St Helens and the conclusive foundation for the value of VegCover.

Figure 5: Methods Flowchart



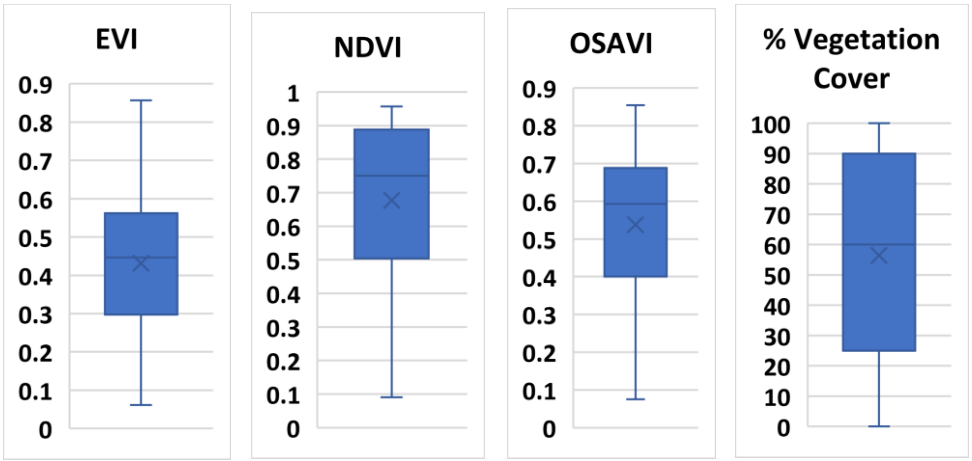
Results

Each vegetation index was calculated from the surface reflectance values of mosaiced Landsat 8 multispectral images captured on 25 July 2021. Using surface reflectance values instead of digital number values for vegetation index calculations, improves the ability to compare multiple images if warranted, as it accounts for atmospheric components of aerosol scattering and thin clouds. Surface reflectance is the amount of light reflected by the Earth's surface and it is a unitless ratio of surface radiance to surface irradiance, producing values between 0 and 1. The vegetation index equations used in the analysis are listed in Table 1: Vegetation Index Equations. Each pixel in the subset multispectral image obtained a calculated result, but only the pixels containing a sample site location were extracted to the attribute table and exported to Excel for regression analysis. A total of 365 sample sites were included in the final analysis, originally 425 sample sites were acquired but 60 sample sites were eliminated due to location on a water body, glacier near the summit, or inside the crater. The eliminated sample sites were confirmed using the Normalized Difference Water Index (NDWI) and Normalized Difference Snow Index (NDSI) to verity that the pixel's reflectance value captured water, ice, or snow.

The first step of result interpretation is to construct boxplots graphing the data distribution of the 365 sample sites for each vegetation index. The purpose for viewing boxplots is verify that there is normal data distribution with no deviations in symmetry caused by outliers, or if the data is skewed indicating that a log transformation would be an appropriate data transformation prior to linear regression. Examples of boxplot graphs are shown in Figure 5: Boxplot Sample Point Distribution for VIs. Each vegetation index appears to have normal data

distribution. The importance of data having normal distribution is to reduce the uncertainty that a random sample deviates significantly from the population, that is the random sampling of sample sites will be an accurate representation of the entire blast zone in the regression model. There will always be a margin of error present when drawing conclusions from the population parameters within the blast zone, but the purpose of constructing a statistical model is to reduce that uncertainty. The Central Limit Theorem states that if sufficient samples are drawn from a population, then the distribution of the sample data will be approximately normal in distribution (Burt, Barber, and Rigby, 2009, 275-276).

Figure 6: Boxplot Sample Point Distribution for Vegetation Cover, EVI, NDVI, & OSAVI



After validating that each index has normally distributed data, the correlation coefficient will indicate whether a linear relationship exists between the index and vegetation cover variables. Using the Pearson's Product-Moment Correlation Coefficient, an r value between 0.2-0.4 indicates a low correlation, an r value between 0.4-0.7 indicates a moderate correlation, an r value of 0.7-0.9 indicates a high correlation and strong relationship, and anything between 0.9-1.0 indicates a very high correlation and a dependable relationship (Burt, Barber, and Rigby, 2009, 168-170). The results of the Correlation Coefficient calculations, also referred to as

Multiple R, are summarized in Table 2: Linear Regression Multiple R & R^2 Results. According to the regression analysis SR, NDVI, SAVI, MSAVI, and OSAVI have very high positive correlations to vegetation cover and the variable's relationship can be considered dependable. The vegetation indices of EVI and GVI both show high positive correlations indicating a strong relationship to vegetation cover. These correlation results indicate that all seven vegetation indices would be suitable candidates for use in an empirical regression model, but the correlation calculations show that NDVI has the highest correlation when compared to vegetation cover.

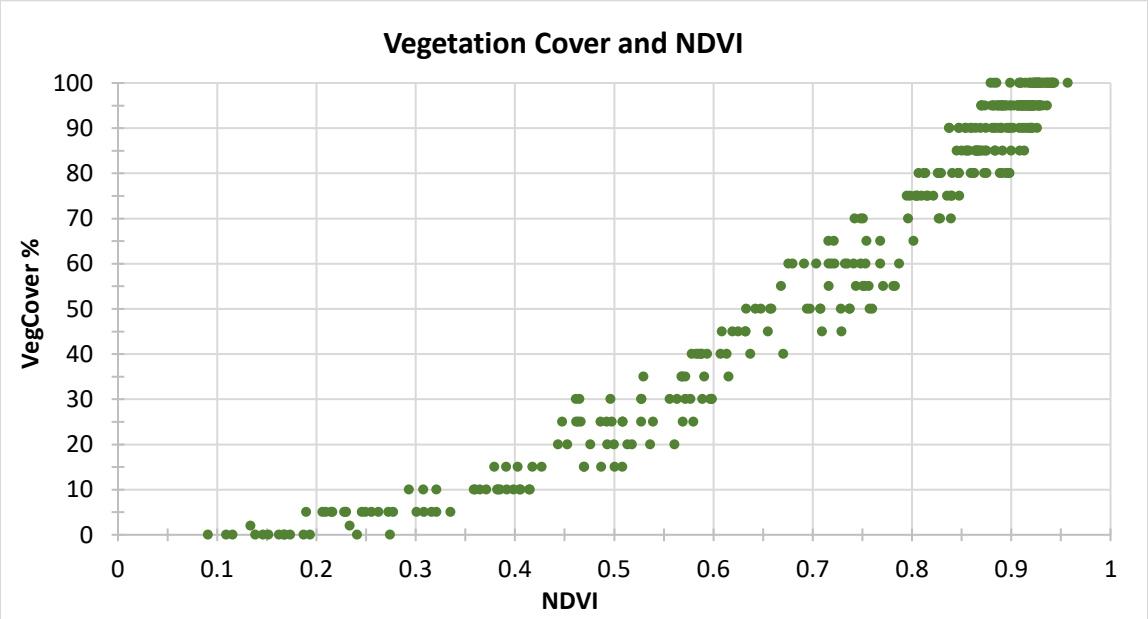
Table 2: Linear Regression Multiple R & R^2 Results

Vegetation Index	*Multiple R (r)*	*R Square (R^2)*
SR	0.9330	0.8701
NDVI	0.9701	0.9411
SAVI	0.9032	0.8158
MSAVI	0.9330	0.8705
OSAVI	0.9482	0.8990
EVI	0.8837	0.7810
GVI	0.8544	0.7301

The empirical model takes the form of a linear regression line that is commonly referred to as the Line of Best Fit. Using the specific method of Ordinary Least Squares, it is a straight line that aims to minimize the distance between it and the data points that are represented by the dependent and independent variables in the model, often called the sum of squares and can represent the variance. The linear regression line, represented as $Y = \beta_0 + \beta_1 X + \varepsilon$ is the model's format and can then be used to make predictions that fall within the defined data range. The components of the equation are as follows: β_1 is the slope of the equation, β_0 is the y-intercept of the regression line, X represents the independent variable, and Y is the

dependent variable (Equation 5, Appendix 1). In this analysis, the performance of the vegetation index for capturing vegetation cover are evaluated based on the standard error of the estimate (ε) of percent vegetation cover. The specific equation for the standard error of the estimate (ε) is shown in Equation 3, Appendix 1 and it represents the accuracy of VegCover as predicted by the vegetation index parameter, with smaller values signifying less uncertainty and a better fitting model.

Figure 7: X-Y Scatterplot NDVI plotted with VegCover %



Each of the vegetation indices tested in the regression analysis had high correlation values when evaluating their relationship to vegetation cover (Table 2). The Normalized Difference Vegetation Index obtained the highest correlation value and will therefore be the selected index for constructing the empirical model. Prior to calculating the regression coefficients that form the model components, viewing an X-Y scatterplot of the sample points can show if the variables display a positive linear trend as the high correlation results would indicate. The data points depicted above in Figure 7: X-Y Scatterplot of NDVI and Percent

VegCover validates that the graphed trend of the sample sites exhibits a positive linear formation. The scatterplot also reveals that the linear trend exhibits a slight concave upward, signifying that polynomial terms should be considered as an option in the regression analysis as well, but the standardized residual plot is more accurate at deciphering that data trend.

Although NDVI presents the strongest correlation value when regressed with percent vegetation cover, an automated regression analysis function is completed in Excel for the other indices in order to compare the regression statistics. The Coefficient of Determination (R^2) value for each index is listed in Table 2 and indicates that most of the vegetation indices could serve as an acceptable model base, especially since each VI's regression calculations indicate statistical significance. An important aspect to mention is that the other vegetation indices had much larger standard error values when compared to NDVI, ranging from roughly 12.4 (MSAVI) to 17.8 (GVI), which could lead to some considerably inaccurate vegetation cover prediction results. The standard error of the estimate for NDVI is about 8.3, meaning that the observed values of vegetation cover could vary from the model predictions of VegCover by an average of 8.3%, a much better estimate than the other indices.

The predicting variable utilized in the empirical model is NDVI and the regression calculations are carried out with full form equations as opposed to using the simple Excel regression function that gives quick stats. The total variation in the model is calculated and given further context by obtaining the explainable and unexplainable variation values within the regression model. The R^2 value represents how much the VegCover model variation can be explained by the mathematical relationship of the independent and dependent variables. The R^2 value is a ratio of the amount of explained variation divided by the total variation, with the

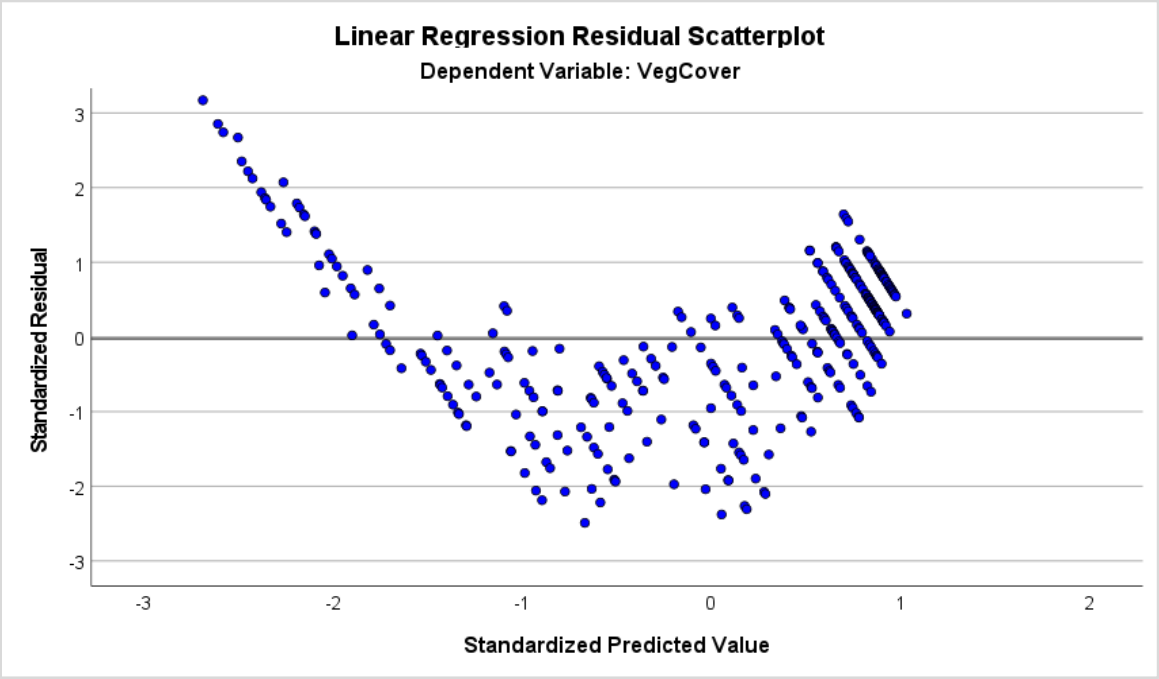
unexplainable proportion signifying the residuals that constitute the unexplainable. If a hypothesis test was applied using a calculated t-statistic and critical value, the null hypothesis stating the relationship is not significant could be rejected, meaning that the correlation between percent vegetation cover and NDVI is statistically significant (p-values < 0.001).

The NDVI linear model has a coefficient of determination of $R^2 = 0.9411$ and the model equation is scripted as $Y_p = 142.75 (X_p) - 39.307$. The regression coefficient signifies a positive relationship between the variables and for every 0.1 increase in NDVI value, there is an accompanied 14.27% increase in vegetation cover. Based on the X-Y scatterplot showing the sample sites and the fitted linear equation, it appears that the model consistently overpredicts the percent VegCover variable when NDVI values range between 0.35 to 0.65 and the model predictions appear most consistent with observed VegCover values when NDVI is greater than 0.70 (Figure 9). Possible regression model remedies for capturing the displayed curvature present in the graphed data points would be to include another independent variable, transform the current dependent variable, or to add polynomial terms of the independent variable.

Formulating a curvilinear regression model could provide the best option for fitting a dataset with concave line trajectories. This specific data trend is also captured by viewing a scatterplot showing the residuals of each data point when comparing the predicted and observed values with relation to the linear regression line (Figure 8: Linear Regression Residual Scatterplot). One statistical assumption required for a linear data model is that the errors should be normally distributed with a mean value of zero and that the residuals are evenly dispersed with no distinct pattern or trend. Assumption violations can indicate that the variable

relationship is nonlinear, or a variable transformation is needed to equalize the nonnormal variance. As expected, the residual scatterplot indicates that the linear model is biased and heteroscedastic which violates a key homoscedastic assumption required for linear modeling. Ordinary Least Squares regression assumes that the population from which a sample is drawn from has a constant variance, but impure heteroscedasticity can cause a non-constant variance if important variables are left out of the model.

Figure 8: Linear Regression Residual Scatterplot



A regression equation expressed with polynomial terms can model a curvilinear relationship of the variables. The regression line trajectory of a polynomial model resembles a parabola for a quadratic term (X^2) or an S-shape for a cubic term (X^3) or higher, and it can be interpreted with some similarities to linear regression. Although this type of data model indicates that there is a nonlinear relationship between the *X* and *Y* variable, the variables coefficients can still present as having a linear relationship (β_1, β_2, … β_h). From the modeling

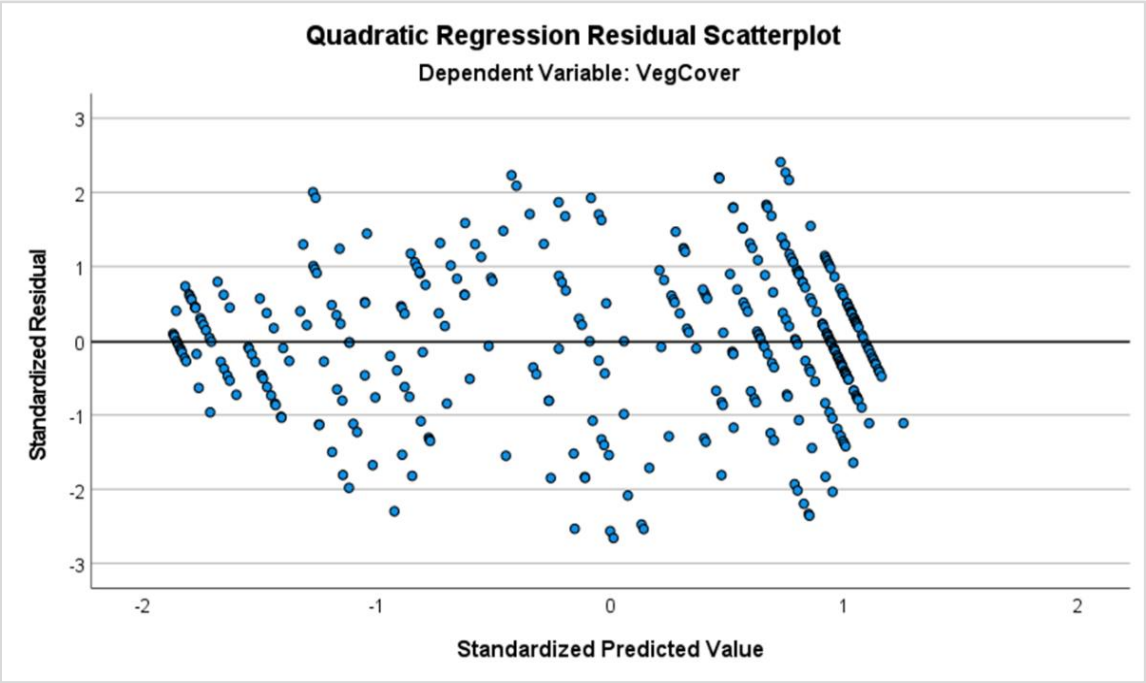
perspective, the added polynomial term can be viewed as an added independent variable within a multiple regression format. The regression equation format is expressed as $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \varepsilon$ where the added polynomial terms are represented as independent variables that are squared or cubed ($\beta_2 X^2$ or $\beta_3 X^3$) (Equation 6, Appendix 1).

The X-Y scatterplot of NDVI and VegCover data points show a distinct curvilinear shape within the trendline, that was also reinforced by viewing the residual plot. A squared NDVI data column is added to the Excel table containing the 365 sample data points and a multiple regression analysis function is implemented. The Adjusted R Square value increased from R^2=0.937 to R^2=0.978 with the change from a linear to a quadratic model. Typically, the R^2 values always increase when a higher order term is added, but the specific increase can be tested to determine if it is statistically greater than zero. Using Equation 7 listed in Appendix 1, the incremental change in R^2 can be factored into a formula with the corresponding degrees of freedom associated with a quadratic to a linear model based on the total number of 365 sample sites. The calculated F-statistic is greater than the corresponding F-table value listed for $F_{(0.05, 2, 362)}$, meaning that with 95% confidence the null hypothesis stating no significance pertaining to the increase of R^2 can be rejected.

The statistically significant quadratic model explains more of the variation between the two variables and the standard error decreased from 8.3 percent to 5.0 percent. Based on the coefficient of determination, the new quadratic equation is better suited at explaining more variance between the variables and the curved regression line accurately resembles the graphed data points. A second residual plot reinforces those statistics as now the residuals appear evenly dispersed around the mean and it resembles an unbiased and homoscedastic

residual plot (Figure 9: Quadratic Regression Residual Scatterplot). The residual plot resembles a normally distributed dataset with all residual points falling within the expected three standard deviations from the mean. A cubic NDVI term was also tested with comparison to the quadratic formula but the increased R^2 value was small and not statistically significant and the standard error for both models remained the same.

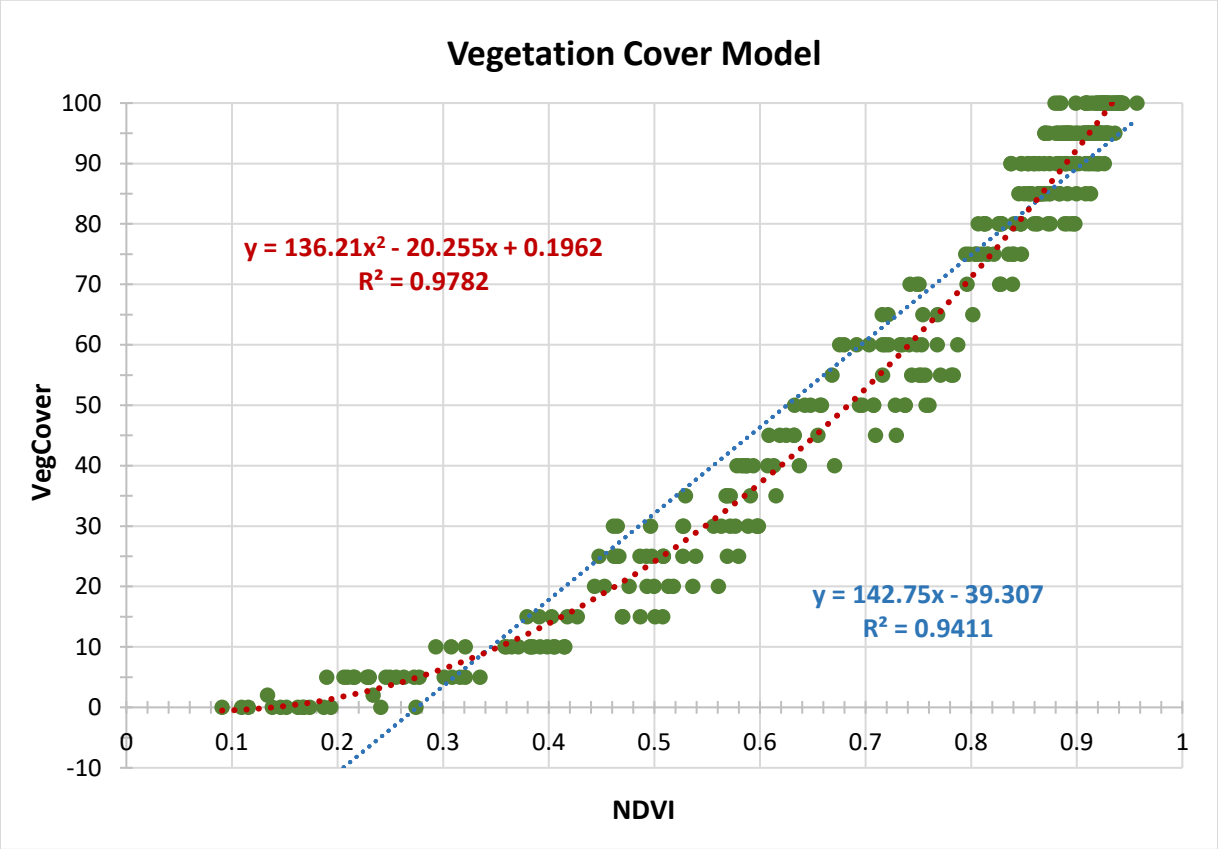
Figure 9: Quadratic Regression Residual Plot



The final regression model is statistically significant ($p<0.001$), and the empirical model fits a quadratic equation formatted as $\mathbf{Y_p = 136.21(X_p^2) - 20.255(X_p) - 0.1962}$. This model can be utilized as a VegCover prediction tool for any point within the blast zone, given a pixel's known NDVI value. Implementing the model for visualization purposes can be demonstrated by using ArcGIS Pro spatial analyst map algebra. The quadratic regression model is entered as the syntax function and the NDVI raster layer serves as the independent variable inputs within the equation (X_p^2 and X_p). Both the quadratic and linear regression models are shown in Figure 10

and the final Empirical Vegetation Cover Model just constructed in the capstone analysis is shown in Figure 11: Empirical Data Model Map.

Figure 10: X-Y Scatterplot of Linear and Quadratic Regression Models



Discussion

The primary objective of this capstone analysis was to construct an Empirical Data Model that could accurately depict current vegetation cover at MSH. This was accomplished by analyzing statistical correlations between applicable vegetation indices and percent vegetation cover. In total, seven vegetation indices were analyzed: EVI, GVI, NDVI, MSAVI, TSAVI, and OSAVI, and each resulted in statistically significant correlation values ranging between r= 0.854 to r= 0.970 (Table 2: Linear Regression Multiple R & R^2 Results). The results were considerably

higher than what was expected based on previous Mount St Helens studies but were in line with similar analyses referenced earlier in the capstone paper (Purevdorj et al. 1998 and Fathoni at el. 2021). The Normalized Difference Vegetation Index demonstrated the strongest relationship to percent vegetation cover and was therefore selected as the independent variable in the Empirical Data Model. The linear regression model is able to explain 93.7% of the variable's variation with an 8% standard error, whereas the more complex quadratic regression model can explain 97.8% of the variation with only a 5% standard error. These statistics indicate that using NDVI as a prediction tool for estimating vegetation cover will produce dependable and highly accurate results.

There is only one previously published analysis that compared various vegetation indices to vegetation cover at Mount St Helens in the years since the eruption. Using Landsat multispectral imagery from August 1995, Lawrence and Ripple (1998) statistically compared multiple VIs to vegetation cover within the blast zone. The purpose for mentioning this published article in the closing statement is to highlight the overdue necessity for another vegetation index evaluation given the dramatic vegetation changes that have occurred over the past twenty-five years. While it may not be scientifically reasonable to make a direct numerical comparison between the analysis results, especially given the length of time and differing methodology, but a comparison interpreting broad conclusions is just as important. Similar to this capstone analysis, Lawrence and Ripple (1998) found that all vegetation indices had statistically significant correlation results. This similarity supports an inference that vegetation indices are in fact an appropriate and highly accurate parameter for modeling vegetation cover. Another comparable conclusion reached by both studies is that out of all the vegetation indices

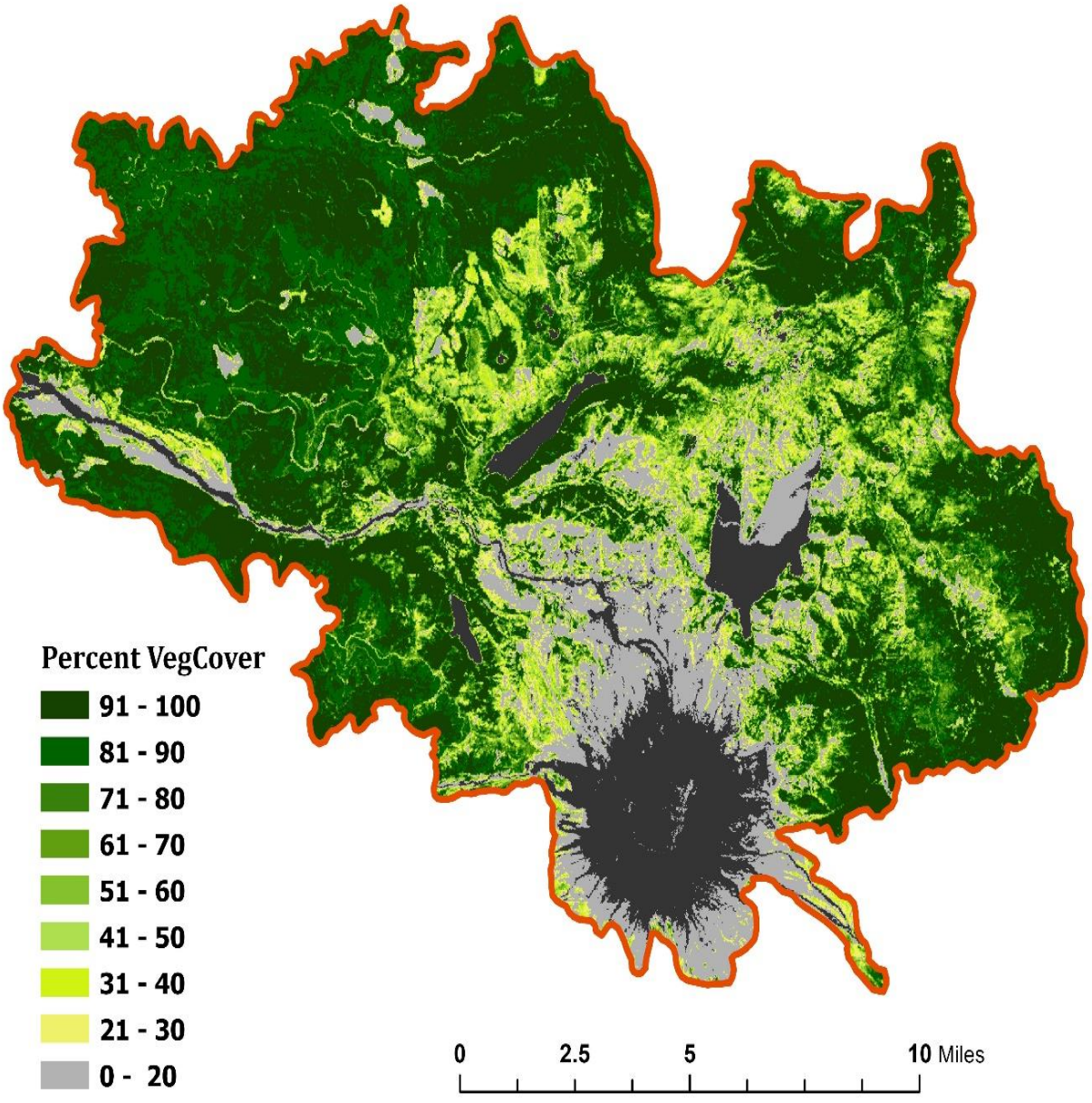
statistically compared, the Normalized Difference Vegetation Index obtained the highest correlation coefficient value. This indicates that NDVI is accurate at describing both sparse and dense vegetation and using the index as a predictor for vegetation cover will produce dependable results. Lastly, the linear regression models formulated by both studies were significantly improved with the addition of polynomial terms. This trend is commonly cited throughout vegetation index literature, and it implies that although indices are significantly correlated to percent vegetation cover, the linear relationship is not static and therefore requires an advanced model to predict.

By interpreting this capstone's final results through the lens of the 1995 MSH analysis, provides an important perspective that would otherwise not be known. Ecological variables in recovering landscapes can change dramatically but research longevity gives the change a relative context. Drawing conclusions based on prior vegetation index studies is an important contribution that this capstone adds to the ongoing understanding of Mount St Helens. Much like the broad principles of statistics - the more data points that are gathered for an analysis, the more confidence the final conclusion provides. This capstone analysis proposes a highly accurate Quadratic Regression Model that can be used at Mount St Helens to predict vegetation cover, $VegCover = 136.21(NDVI^2) - 20.255(NDVI) - 0.1962$. Both the Linear Regression Model and the Quadratic Regression Model (Figure 9) are provided together with the statistical analysis and an example of the Empirical Data Model in action (Figure 10).

Further Research

NAIP orthoimages have a spatial resolution that is hard to surpass in finite detail and that specifically allows for precise vegetation cover assessments. The fundamental issue preventing their use as the primary imagery tool for scientific research, is that orthophotos are not readily available. NAIP imagery is expensive to produce due to the time-consuming acquisition methods required and are therefore collected in two-year cycles, sometimes with an added year of processing before a public release. By developing an empirical regression model based on Landsat multispectral imagery, accurate vegetation cover values can be calculated at any time for an up-to-date ground data reflection. Multispectral imagery is captured by Landsat satellites every 16-days, and that allows for a consistently changing vegetation cover measurements if needed. The empirical model can be tested yearly with similar research or a sample dataset to determine if the model continues to stay accurate at predicting vegetation cover.

Figure 11: Empirical Data Model representing Vegetation Cover using an NDVI raster



References

Asokan, Anju, and J. Anitha. 2019. “Change detection techniques for remote sensing applications: A survey.” *Earth Science Informatics* 12, no. 2 (March 8, 2019): 143–160. https://doi.org/10.1007/s12145-019-00380-5

Dale, Virginia and Crisafulli, Charlie. 2018. “Ecological Responses to the 1980 Eruption of Mount St. Helens: Key Lessons and Remaining Questions.” In *Ecological Responses at Mount St. Helens: Revisited 35 years after the 1980 Eruption*. New York: Springer Press. https://doi.org/10.1007/978-1-4939-7451-1_1

DeRose, Ronald C., Takashi Oguchi, Wataru Morishima, and Mario Collado. 2011. “Land cover change on Mt. Pinatubo, the Philippines, monitored using ASTER VNIR.” *International Journal of Remote Sensing* 32, no. 24 (January 2019): 9279-9305. https://doi.org/10.1080/01431161.2011.554452

De Schutter, Ann., Matthieu Kervyn, Frank Canters, Sonja Bosshard-Stadlin, Marjura A Songo, and Hannes Mattsson. 2015. “Ash fall impact on vegetation: a remote sensing approach of the Oldoinyo Lengai 2007–08 eruption.” *Journal of Applied Volcanology* 4, no. 15 (2015): 1-18. https://doi.org/10.1186/s13617-015-0032-z

Earth Resources Observation and Science (EROS) Center. 2020. *Landsat 8 Collection 1 (C1) Land Surface Reflectance Code (LaSRC) Product Guide*. No LSDS 1368 Version 3. Reston, VA: Department of the Interior U.S. Geological Survey

Fathoni, Mohammad N., Pramaditya Wicaksono, and Syamsul Bachri. 2021. “Estimated change in the percentage of vegetation cover after the eruption of Mount Agung, Bali in 2017.”

Proceedings of SPIE 12082, Seventh Geoinformation Science Symposium (22 December 2021). https://doi.org/10.1117/12.2617334

Harrington, Lisa. M., John A. Harrington Jr, and Peter M Frenzen. 1998. "Vegetation change in the Mount St. Helens (U.S.A.) blast zone, 1979–1992." *Geocarto International* 13, no. 1 (1998): 75–82. https://doi.org/10.1080/10106049809354631

Huang, Sha, Lina Tang, Joseph P. Hupy, Yang Wang, and Guofan Shao. 2021. "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing." *Journal of Forestry Research* 32, no. 1 (2021): 1–6. https://doi.org/10.1007/s11676-020-01155-1

Joshi, Prem C. 2011. "Performance evaluation of vegetation indices using remotely sensed data." *International Journal of Geomatics and Geosciences* 2, no. 1 (2011): 231-240.

Lawrence, Rick L., and William J. Ripple. 1998. "Comparisons among Vegetation Indices and Bandwise Regression in a Highly Disturbed, Heterogeneous Landscape: Mount St. Helens, Washington." *Remote Sensing of Environment* 64, no. 1 (1998): 91-102. https://doi.org/10.1016/S0034-4257(97)00171-5

Lawrence, Rick L., and William J. Ripple. 2000. "Fifteen Years of Revegetation of Mount St. Helens: A Landscape-Scale Analysis." *Ecology* 81, no. 10 (October 2000): 2742-2752. https://doi.org/10.2307/177338

Lyon, John G., Ding Yuan, Ross S. Lunetta, and Chris D. Elvidge. 1998. "A change detection experiment using vegetation indices." *Photogrammetric Engineering and Remote Sensing* 64, no. 2 (February 1998): 143-150.

Marzen, Luke. J., Zoltan Szantoi, Lisa M. B. Harrington, and John A Harrington. 2011. "Implications of management strategies and vegetation change in the Mount St. Helens blast zone." *Geocarto International* 26, no. 5 (2011): 359–376. https://doi.org/10.1080/10106049.2011.584977

Mazza, Rhonda. 2010. "Mount St. Helens 30 years later: A landscape reconfigured." *Science Update* 19 (Spring 2010): 1-11. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. https://www.fs.fed.us/pnw/pubs/science-update-19.pdf

Purevdorj, T.S., R. Tateishi, T. Ishiyama, and Y. Honda. 1998. "Relationships between percent vegetation cover and vegetation indices." *International Journal of Remote Sensing* 19, no. 18 (1998): 3519-3535. https://doi.org/10.1080/014311698213795

Teltscher, Katharina, and Fabian E. Fassnacht. 2018. "Using multispectral landsat and sentinel-2 satellite data to investigate vegetation change at Mount St. Helens since the great volcanic eruption in 1980." *Journal of Mountain Science* 15, no. 9 (2018): 1851–1867. https://doi.org/10.1007/s11629-018-4869-6

Tilling, Robert I., Lyn J. Topinka, and Dale Swanson. 1990. *Eruptions of Mount St. Helens: past, present and future*. Reston, VA: U.S. Dept. of the Interior, Geological Survey. https://doi.org/10.3133/7000008

Xie, Yichun, Zongyao Sha, and Mei Yu. 2008. "Remote sensing imagery in vegetation mapping: a review." *Journal of Plant Ecology* 1, no. 1 (March 2019): 9–23. https://doi.org/10.1093/jpe/rtm005

Xue, Jinru, and Baofeng Su. 2017. “Significant remote sensing vegetation indices: A review of developments and applications.” *Journal of Sensors* 2017, no. 1 (2017): 1–17. https://doi.org/10.1155/2017/135369

Appendix 1

Equations

Equation 1: Sample Size

Sample Size = ((Z-score^2) * Standard Deviation * (1-Standard Deviation)) / (margin of error^2)

Equation 2: Correlation Coefficient (r)

r = ∑i (Xi - Xm)*(Yi - Ym)/((n-1)*(Sx*Sy))

r= SSxy/((n-1)*(Sx*Sy))

r= SSxy/(SQRT(SSx*SSy))

Equation 3: Standard Error of the Estimate

SE_r = SQRT((1-r^2)/(n-2))

Equation 4: Coefficient of Determination (R^2)

R^2 = SSR / SSy

R^2 = (Yp-Ym)² / (Yi-Ym)²

Equation 5: Linear Regression Model

$Y_p = \beta_1 (X_p) + \beta_0$

where:

Slope β_1 = SSxy/SSx or β_1 = r*(Sy/Sx)

Y-Intercept $\beta_0 = Y_p - \beta_1 * X_p$

Equation 6: Quadratic Regression Model

Format: $Y_p = \beta_0 + \beta_1 X_p + \beta_2 X_p^2 + \ldots + \beta_h X^h + \varepsilon$

$Y_p = \beta_2 X_p^2 + \beta_1 X_p + \beta_0$

Equation 7: **F-Statistic** (R^2 Significance from Linear to Quadratic)

R^2_i is the R^2 for the ith order

R^2_j is the R^2 for the jth order

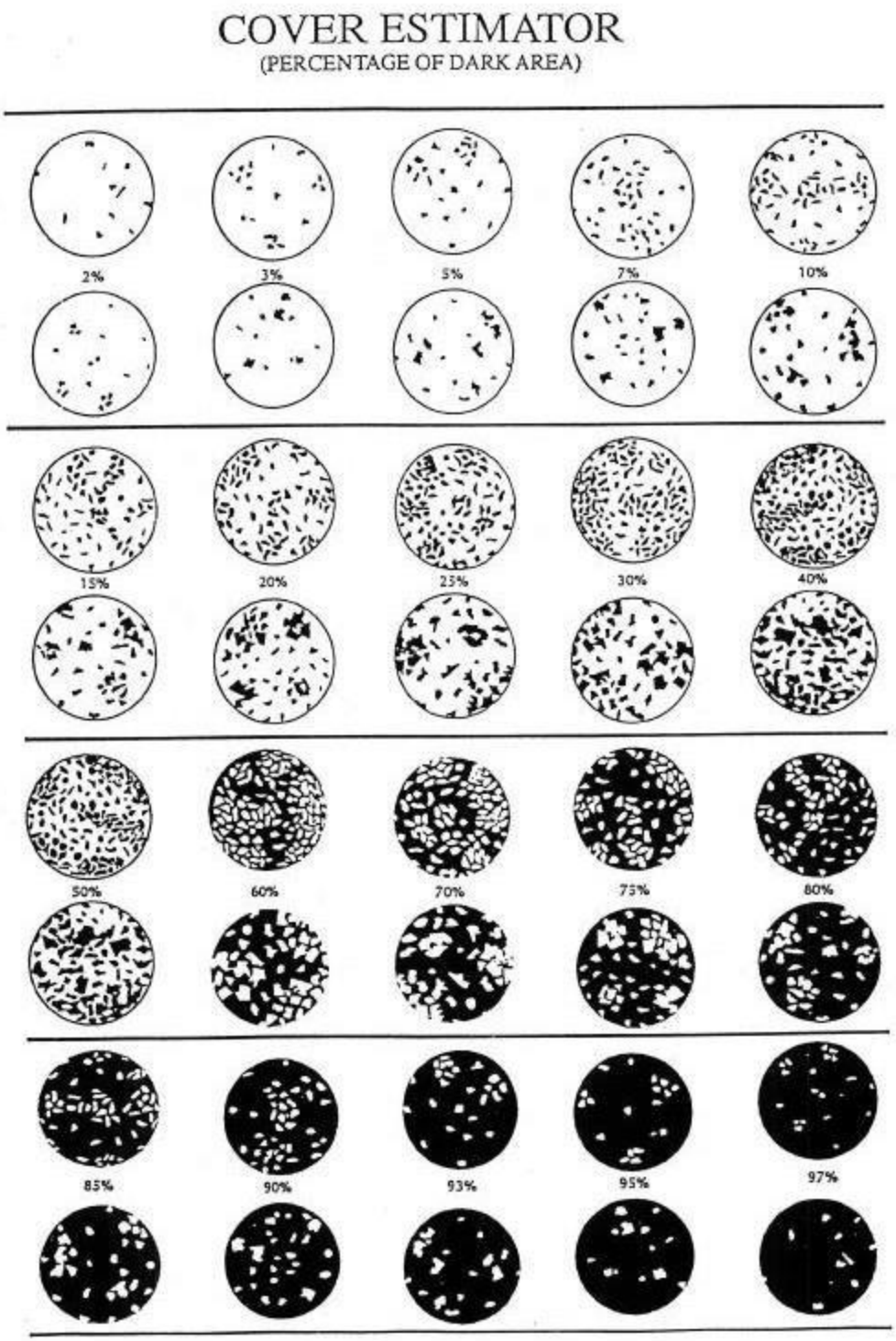
$d.f._j$ = degrees of freedom for the jth order

F-statistic = $d.f._j \times (R^2_j - R^2_i)/(1 - R^2_j)$

F-table degrees of freedom = j numerator and $d.f._j = n - j - 1$ denominator

Appendix 2

Figure 12: Visual Guide to Estimate Vegetation Cover



Appendix 3

Metadata

Multispectral Imagery
Earth Explorer. U.S. Geological Survey. NASA. Landsat 4-9 C2 U.S. Analysis Ready Data. Accessed August 2022. https://earthexplorer.usgs.gov/

Title ID: LC08_CU_003002_20210725_20210806_C01_V01_SR
Title ID: LC08_CU_003003_20210725_20210806_C01_V01_SR
Acquisition Date: 2021-07-25
Horizontal: 003
Vertical: 002
Spacecraft Identifier: Landsat 8
Sensor Identifier: OLI_TIRS
Datum: WGS84
Map Projection: Albers Equal Area
Collection Number: 2
Units: Meters
Latitude: 44.68808

NAIP Imagery
2021 National Agriculture Imagery Program (NAIP)
Source ID: USDA FPAC-BC-GEO
NAIP Entity ID:FGDC-STD-001-1998
Acquisition Date: 2021-07-26
State: WA
Agency: USDA
Map Projection: UTM
Projection Sone 10N
Datum: NAD83
Resolution:0.600
Units: Meters
Number of Bands: 4
Sensor Type: CNIR
Project Name: “201911_WASHINGTON_NAIP_0X6000M_UTM_CNIR”