Indicators of Land Degradation: A Case Study of the Use of Random Forest Regression to Predict Difference in Vegetation Growth over Time

by

Ryan Littleton

A Capstone Submitted to the Faculty of

Utica College

December 2019

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in
Data Science

Abstract

Land degradation has been a concern in land use management, agriculture, and conservation as well as for policy makers in terms of evaluating and preventing the loss of vegetative fertility. With land fertility influencing food security, economics, and environmental concerns such as climate change, an enhanced understanding of land degradation is an issue of continued applicability to those fields. Despite the research applied in the past toward these goals, the issue still remains an open problem. This study utilized GIS tools and machine learning techniques, which harnessed remotely sensed vegetation data and survey-based soil data for the state of California to further the study of land degradation evaluation and prediction. This study sought to answer the question, "What factors are most predictive of land degradation and what algorithm can be trained to effectively and efficiently make such predictions?" To this end, a forest-based regression was developed to predict changes in vegetation detected by satellite imagery using soil variables: soil organic matter, sodium absorption ratio, wind erodibility group, and susceptibility to water erosion. The model had results that were small but statistically significant, indicating map areas where vegetative conditions are associated with the selected soil traits. The areas demonstrated this association then show evidence of specific cases of potential degradation that could be used to direct resources for more detailed evaluations. These results also show a specific case where machine learning can be applied to land degradation research and can be used to help refine future predictive work. Keywords: Data Science, Dr. Michael McCarthy, MODIS, STATSGO2, quantitative, California.

Acknowledgments

One thing that is known, but perhaps not spoken of as often as it ought to be is that nobody completes any important task without assistance from others. This work is no exception. As much as it was impossible to complete this work in isolation, it would also be impossible to do justice to the number and kind of support I have had during the time this was being worked on. As such, I would just like to thank the friends, family, acquaintances, and academic faculty who have been there along the way that have made this possible. It is my hope and belief that they know who they are.

Table of Contents

List of Illustrative Materials

Introduction

Concern over the quality of arable land has long been a concern in land-use management, agriculture, conservation, and public policy. In the past, the main concern was *desertification* — the spread of desert land into areas that were not previously desert (2019). The definition of "desertification" raised concerns, however, and have been the subject of some scrutiny due to imprecise definitions of what qualifies as desert land as well as the relationship of a desert to climate and the status of desert ecosystems as important in their own right (Eswaran *et al.*, 2019). With these critiques in mind, the focus of current research is on *land degradation*, broadly defined as a transformation of a geographic area to an undesirable state, usually measured by relative vegetative growth capacity (Bojórquez-Tapia *et al.*, 2013).

One of the challenges facing researchers in this area has been effectively classifying degraded land. Most historical studies relied on expert surveys, particularly of soil quality. While accurate and effective, these efforts are also costly and take considerable time. More recent developments utilized satellite-derived vegetation data along with certain climate data — in particular, precipitation levels — to classify areas of land with reduced vegetation yield as compared to other growth factors as degraded.

With this past research in mind, the focus of this study is on utilizing publicly available time-series datasets from the years of 2006 to 2018 along with modern machine learning techniques to predict areas of land that are under greater threat of future degradation. This could have future applicability as a resource identifying areas to focus field-based studies or interventions.

Because precipitation variability is known to influence plant growth in an acute fashion and such fluctuations might mask other effects that are indicative of long term trends,

precipitation will be avoided for the purposes of this research and instead data relating to land and soil conditions will be focused on for the purpose of identifying long term trends. The question addressed by this study is: What elements are most predictive of long-term land degradation and what algorithm can be trained to effectively and efficiently make such predictions?

Literature Review

To add context to the topic of land degradation and issues related to data used in its evaluation and prediction, existing literature addressing the research in the field as well as conceptual related works are examined here.

Previous Research

The body of existing research on land degradation examined in this review addresses works not only related to prediction and remote sensing of factors related to land degradation, but also methodological works examining use cases for the predictive technologies themselves. These establish the related ways that machine learning and remote sensing technologies can be applied to analyses of soil and vegetation, and how such research relates to land degradation. Research applications of machine learning technology to the analysis of land use and soil science is demonstrated in a case study wherein artificial neural networks are applied to the task of soil classification in Egypt (Amato *et al.*, 2015). This shows an application and use case for applying data analysis and machine learning to tasks fundamental to soil science and land use evaluation, areas related to this research in data selection and methodology.

Many of the previous use cases of both vegetative remote sensing data and predictive algorithms related to the mapping of land degradation. To this end, Chikhaoui and others

established an index in semi-arid regions of Morocco to map land degradation using data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), a data source related to the Moderate Resolution Imaging Spectroradiometer (MODIS) data used here that also relies on measurements of reflection to assess vegetative land coverage (Chikhoaui *et al.*, 2005). Similarly, Bojórquez-Tapia and others addressed the task of degradation mapping, with a key difference being the use of what they deemed a "knowledge-based format" rather than an index of a single spatial variable (2015). This knowledge-based format combines remote sensing and directly observed variables related to land conditions to establish a framework for mapping land degradation through a combination of these observations. Key insights drawn from this work is the inherent complexity and difficulty in defining land degradation in terms that are specific enough for mapping and also the breadth of the range of land conditions that might be considered in such a definition. Of note is the observation regarding what constitutes an undesirable change appears different based on the subjectivity of the map user. Degradation assessment and mapping might then differ among different groups such as "conservationists, emergency services, or farmers" (Bojórquez-Tapia *et al.*, 2015, p.52).

In work establishing the technique of identifying land degradation and the use of remote sensing data, Fensolt and Rasmussen use the measure of "rain use efficiency" as a proxy for land degradation in semi-arid regions (2011). This is defined by calculations of precipitation records along with remotely sensed vegetation, showing a different way this data source was used in previous research for degradation assessment (Fensolt & Rasmussen, 2011).

The research into the use of vegetation indices for remote sensing of land degradation was covered in several other areas and applications. Yengoh and others addressed much of the other research to date, describing the use of the normalized difference vegetation index (NDVI)

for remote sensing of features relevant to degradation assessment and conservation efforts (2015). Elements outlined include land use, land cover change, drought, desertification, soil erosion, vegetation fires, biodiversity monitoring, and conservation and soil organic carbon. This offers a perspective on the breadth of which specific areas have previously been the subject of research relating to the use of remote sensing in general and vegetation index data specifically for comparative evaluation (Yengoh *et al.* 2015).

Some of the other works explain the use of remotely sensed vegetation data for land degradation assessment as well as provide criticism of the practice as a sole source of degradation assessment. Higginbottom and Symeonakis offer one such discussion and critique that reviews the different techniques that have previously been used for degradation and desertification assessment by evaluation of vegetation index data (2014). Specifically, they propose metrics for evaluating the effectiveness of such systems and techniques — namely that there needs to be a standardized validation they are subjected to and that low intensity or early instances of degradation must be detected for the techniques to be of practical value. They are also critical of the idea that systems reliant on vegetation data alone can account for degradation and that they should be part of a greater methodology (Higginbottom & Symeonakis, 2014). Easdale and others likewise offer a critical view of the exclusive use of remotely sensed vegetation indices as a sole proxy for degradation (2015). Specifically, their criticism falls on the use of linear trends of these indices for modeling degradation, identifying areas where such techniques have led to the misinterpretation of actual trends. They argue for the use of more accurate and not strictly linear modeling to assess trends in vegetation for degradation assessment (Easdale, 2015).

Further criticism of the exclusive use of vegetation trends for degradation modeling using exclusively remotely detected vegetation index data occur in the work of Herrmann and Sop; they observe that in the assessment of desertification — a subset of degradation research — in the Sahelian region of sub-Saharan Africa, remotely sensed vegetation data did not accurately classify either desertification or vegetation changes (2016). This shortcoming was attributed to a failure to capture changes in the type of vegetation. As the specific plants growing in an area are closely related to the qualities of land conditions, this indicates a shortcoming of this kind of data where an important generalization goes undetected (Herrmann & Sop, 2016).

Soil and Vegetation Issues Related to Land Degradation

In the context of both the above-mentioned methodologies and the criticisms offered of linear modeling and exclusive use of remote vegetation data as a source of land degradation evaluation, research and reference texts also address the relationship between survey-gathered soils data and vegetation as well as the use of soil measures to evaluate degradation.

Shrestha demonstrated modeling of degradation in the Nepalese Himalaya using a combination of remotely sensed vegetation data and field observations of land and soil conditions (2004) with the modeling accomplished with decision trees. With both the combination of data sources and the use of decision tree algorithms for the degradation modeling, this offers a precedent for some of what was used for this current research, although in a non-predictive context (Shrestha, 2004).

Oblaum and others offer another approach to degradation modeling; this work discusses the use of soil organic matter as the only indicator for degradation evaluation, breaking from the above studies by avoiding the use of vegetation data altogether (2017). This both demonstrates

the value of this kind of data as well as more of the breadth of evaluation techniques available for this area of research (Oblaum *et al.*, 2017).

In conjunction with the other sources using soil carbon levels as a metric for land degradation evaluation and analysis, Mikhailova and others offer a cautionary take on available soil survey datasets (2016). They compare carbon estimates from the Soil Survey Geographic Database (SSURGO) dataset provided by the USDA and other field methods, identifying some of the shortcomings of that dataset in terms of accuracy without invalidating its use. This supports the notion that multiple avenues of data collection and multiple types of data are desirable for a more complete picture of degradation (Mikhailova *et al.*, 2016).

Another work that addresses the combination of data sources and the potential to increase the effectiveness of analysis is Anderson and others, who use the SSURGO soils dataset to increase the accuracy of flood forecasting (2006). This demonstrates how the use of these different kinds of data sources can be beneficial to the application of other sources. Furthermore, as runoff is one of the factors that can be of importance in degradation, flooding is a relevant topic to the other research outlined here (Anderson *et al.*, 2006).

A final work connecting research on the variables of vegetation and soils data exists in Wang and others, where soil organic matter is estimated from remotely sensed land images, slope, elevation, and vegetation index (2010). These are identified as being indicators of land degradation for that study and as such, this research offers a reverse form of some of the goals of this project, which seeks to predict vegetative growth patterns indicative of degradation using soils data including soil organic matter (Wang *et al.*, 2010).

Sociological Background

In addition to the physical variables and machine learning techniques discussed here, the issue of the socio-economic consequences of land degradation is one that has been addressed in the existing literature. Not only is the discussion one of the impacts, but per the observations in the thematic report by Nachtergaele and others, the measurements of the state of degradation also take economics, social and cultural information into account (2011). The global land degradation information system (GLADIS) maintained by the Food and Agriculture Organization of the United Nations and described in the mentioned thematic report maintains such records. Beyond the socioeconomic records being kept as data indicative of degradation states due to the subjective impact, the thematic report also specifically identifies impoverished areas as under greater impact from similar physical states (Nachtergaele *et al.*, 2011).

Other sources clarify the connection between land degradation and socio-economic factors by evaluating it in more specific contexts. Meadows and Hoffman describe the historical links between degradation in South African agricultural land and the influences of colonialism and apartheid planning (2003). This is connected to modern states of poverty in the same regions and predicted to be exacerbated by developing climate change. Especially as relates to ethical considerations of the implications of any research into land management and degradation, this work offers important context (Meadows & Hoffmann, 2003).

In further works tying degradation research to the wider sociological context, Hill and others made use of similar time series vegetation data to map conditions connected to land use patterns associated with degradation in the Mediterranean (2008). This research focuses on interactions between humans and the environment in this mapping and offers a perspective that ties together the remotely sensed vegetation data with specifically human-use related concerns (Hill *et al.*, 2008).

Rationale

Importance of Research on Land Degradation

An extensive body of research into the phenomenon of land degradation exists and holds utility for a variety of interests (Meadows & Hoffmann, 2003; Oblaum *et al.*, 2017; Hill *et al.*, 2008; Herrmann & Sop, 2016; Fensolt & Rasmussen, 2011). A loss in vegetative fertility of land over time has long-reaching consequences that span sociological, agricultural, and environmental disciplines and concerns. A loss of agricultural productivity linked to vegetative fertility loss is not only threatening to agricultural workers, but is also a concern due to issues of food security, a disproportionate economic impact on vulnerable populations, and biodiversity loss. Furthermore, as outlined in the IPBES report on land degradation, there is a bidirectional relationship between degradation and climate change (2018). While worsening climate change is a factor leading to land fertility loss, the removal of soil carbon and vegetation biomass is also a factor contributing to global climate change (Zhang *et al*., 2018).

Reasoning for this Research

The primary aim of this study was to contribute to the body of research on methods to predict land degradation. As established in the literature above, there is currently a degree of uncertainty inherent in previous methodologies for mapping and predicting degradation using a single type of data, either vegetative or soil. This uncertainty is exacerbated when limited to exclusively remotely sensed data. In this study, the use of both field survey soils data and remotely sensed vegetation data combined with a predictive approach provides a basis to build on previous research, probe methodologies that might improve future predictions, and to highlight future opportunities for research that can contribute to the ongoing challenge of land degradation prediction.

Methodology

Data Sources

This study used two sources of data. The first is a vegetation index derived from data provided by the MODIS, a satellite launched by the National Aeronautics and Space Administration (NASA) to provide image sensing data of the Earth's surface (Didan, *2015)*. The first index provided is the Normalized Difference Vegetation Index (NDVI), which is derived from light reflection data from the MODIS instrument to estimate vegetation cover levels. The second is the Enhanced Vegetation Index (EVI) which calculates estimated vegetation cover levels similarly but is calculated in a way that decreases distortion from particles in the air and also from oversaturation in areas of high vegetation. For this study, the EVI aggregated at the monthly level at a resolution of one kilometer with data collected from 2001 to 2019 (Didan, 2015). This monthly aggregation is a weighted temporal average provided as one of the data products for the MODIS instrument by NASA, including all products that overlap the month. The products offered by NASA are otherwise at a temporal resolution up to 16-day intervals. Because this index is derived from the same data source as the NDVI, and it offers technical advantages in accuracy, the EVI is used exclusively as a source of vegetation data throughout this work.

The second data source used is the Digital General Soil Map of the United States (STATSGO2) soil survey database provided by the USDA (2006). It is described as a ". . . broad-based inventory of soils and non-soil areas that occur in a repeatable pattern on the landscape and that can be cartographically shown at the scale mapped of 1:250,000 in the continental U.S." (Soil Survey Staff, 2006). The dataset generalized more detailed soil maps collected by government agencies over a wide time frame and are acknowledged by the USDA

to possibly have differing levels of specificity depending on time and agency of collection. The STATSGO2 is the less detailed of the two soil surveys provided by the USDA; however, it is sufficiently detailed for this study and is available for distribution for wider geographic areas than the SSURGO, facilitating the collection of the dataset for the entire state of California as required for the task of this research. The survey itself was dated in the metadata as being last updated in 2006, and indicated that the surveys used data older than that as well; however, the complexity of the dataset leaves the detail of how old any particular selection of data is ambiguous. Nevertheless, this guarantees none of the data is newer than 2006 (Soil Survey Staff, 2006).

Exploratory Analysis

Geographic areas of interest. The intent of this research is to analyze the State of California as a case study for land degradation. The state of California was selected for the final analysis because it has several desirable characteristics for research into land growth characteristics. The diversity of land types found within the state combined with a documented history of concern over desertification and land degradation contributed to this selection.

Preprocessing of data. The EVI data required several steps to be usable for analysis. First, a scale factor was applied to be used in the exploratory analysis. This was omitted as unnecessary for the purposes of the eventual predictive algorithm application but was useful for initial mapping. The time-series nature of the vegetation data necessitated some efforts in simplification of the EVI series for the purposes of making both datasets comparable with one another. As the key indicator targeted for identification in this work is vegetation change over time, the targeted simplification created an index of change over the measurement period. Examination of the vegetation averages for each month revealed a seasonal effect of growth as shown in figures 1

and 2 below. In figure 2, an annual peak and valley are noted in the vegetation averages, demonstrating this seasonal variation. To control for this seasonality effect, difference calculations were made separately for each month from the beginning of the measurement period to the end. As the calculations were for identical parts of the year, this helped to control for seasonality. For the purposes of aligning this research with predictive goals and the knowledge that the vegetation difference was the intended dependent variable, vegetation data prior to 2006 was eliminated from the analysis, as this is the date where the soils dataset has final measurements, ensuring that the model is used predictively rather than as a classification tool. Data for 2019 was also eliminated due to being incomplete. With separate differences in EVI calculated from 2006 to 2018 for each month, an average of these was taken and used as the index of change in vegetation for the measurement period (Figure 1 for an example). Through this point of processing, the vegetation data remained in raster format.

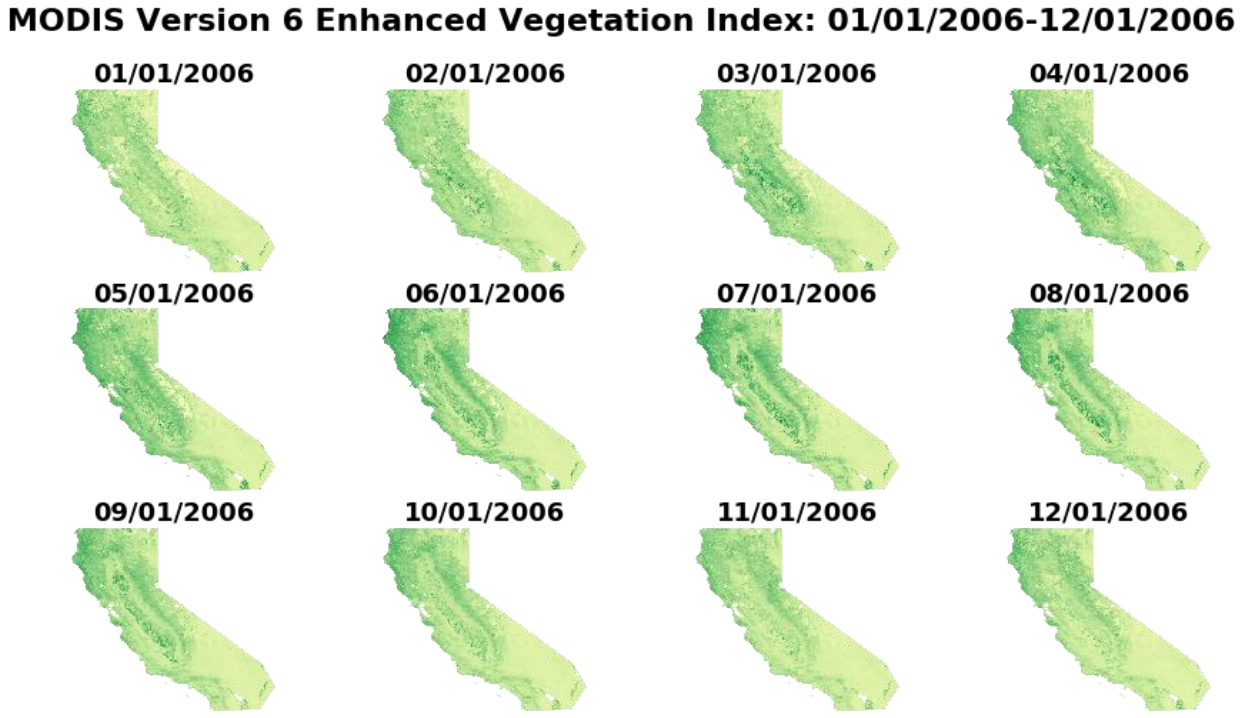

Figure 1: Monthly Vegetation Maps for Single Year (EVI)

Documentation from Esri, the producer of ArcGIS software for calculating vegetation difference, advocated assigning a break point for positive and negative vegetation change and mapping that point to categorical values to identify substantial increases or decreases in vegetation. This process is outlined in ArcGIS technical documentation (Image Change Detection, 2019); however, no literature was found in which any sort of rigorous definition of a break-point could be defined for this purpose. As such, a technique for preprocessing would be little more than preliminary classification based on arbitrary judgement, so results from these efforts were set aside. While it was not useful as preprocessing, the point maps of substantial increase and decrease in vegetation are useful visual tools in an exploratory sense and the results of attempting this step are included in appendix a.

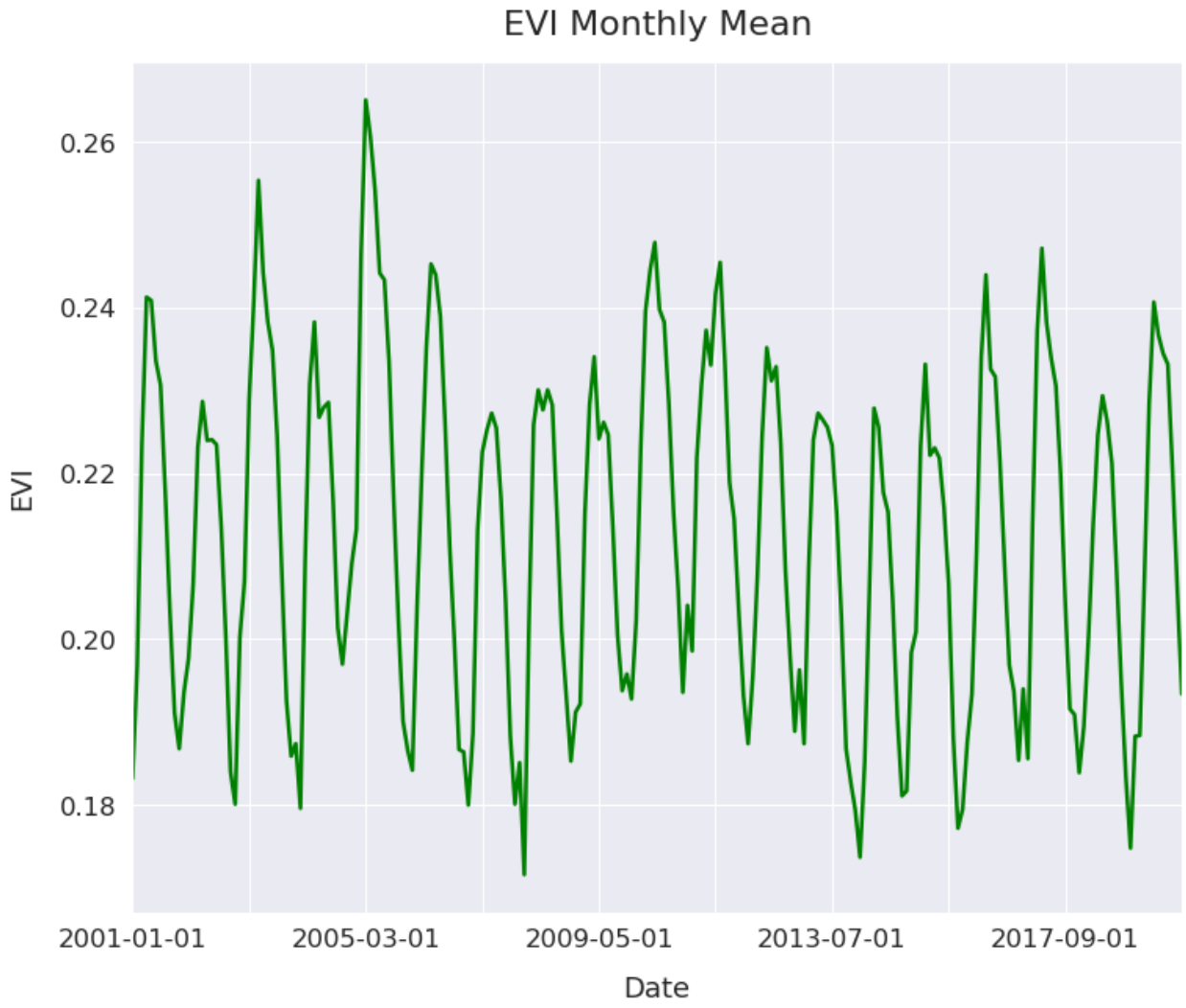

Figure 2: EVI Monthly Mean Demonstrating Seasonality of Vegetation Data

The soils dataset is packaged in a relational database including information for various soil layers and granular data regarding conditions of each layer in terms of items such as specific particles found in each soil type at each location without any aggregation of types across layers. To read this data into a format aggregated at a level that was appropriate for this analysis, the USDA soil data viewer, provided as a plugin to ArcGIS software was employed to perform initial aggregation into meaningful layers that were comparable. Each feature discussed was extracted for the surface layer only, with the rationale that surface layers and topsoil are more

variable, more susceptible to the pressures driving degradation and more relevant to variability vegetation growth.

Identification of Variables of Analysis

As established by the research topicality and previous research in the field, the targeted dependent variable was chosen to be the calculated index of change in vegetation growth. The available data within the soils dataset allow a prediction based on the conditions of soils in the evaluated region. Identifying the soils data for the predictive algorithm implementation also required some theoretical research evaluation for decision making. Soil organic content is established in the literature review as being a concern and possible indicator of soil degradation alone, so it was included in the analysis as the weighted average percent of soil composition for the surface layer (Yengoh *et al.*, 2015). This weighting was built into the USDA provided Soil Data Viewer with which the variable was aggregated, and the weighting factor is based on percent composition of all soil attributes throughout the map unit (Soil Survey Staff, 2006). Further work done by Karis and others established areas of concern relating to land degradation as involving erosion, soil salinization, water stress, forest fires, and overgrazing (Karis *et al*., 2014). Within the data available in the data sets selected, salinization and erosion were two factors that could be aggregated into several variables available with the Soil Data Viewer. For erosion, the dominant condition of the wind erodibility group and the susceptibility to water erosion (i.e., K factor), also mapped per dominant condition of the surface layer were selected. To capture the salinity, the sodium adsorption ratio was extracted as a weighted average for the surface layer.

Wind erodibility index was also considered, however the variable was exceedingly sparse, containing almost entirely missing values and zero values, as evidenced by both the variable

mean rounding to zero and the map of that variable included in the appendix. Due to these considerations, the wind erodibility index was omitted from the final analysis. Other related variables were investigated but proved extremely sparse over the geographic region and therefore unlikely to be useful for predictive purposes.

Summary statistics and correlation tables were generated for the soil variables and the calculated index of change in vegetation as part of the initial exploratory analysis as pictured in tables 3 and 4 below. It is worth noting, however, that traditional statistics in general and simple bivariate correlation in particular are of limited utility as these techniques can be undermined or compromised by spatial autocorrelation (Haining, 2015). As such, greater attention is paid to the maps generated and the eventual predictive algorithm implementation. Maps for each of the included variables are included in appendix b and basic information on the variables is summarized in table 5 below.

Table 3: Summary Statistics

	EVI Difference	OrganicMatter	SAR	WindErosionIndex	KFactor
count	4,288	4,288	4,288	2,691	3,637
mean	-84.11	2.03	0.10	0	0.23
std	345.42	2.25	0.62	0	0.09
min	-2,311	0	0	0	0.02
25%	-222	0.7	0	0	0.17
50%	-73	1.64	0	0	0.2
75%	85	2.62	0	0	0.32
max	2,457	37.13	15	0	0.55

Algorithm Implementation

For prediction, the forest-based classifier and regression tool in ArcGIS was employed. This tool is based on the random forest decision tree algorithm and has associated features such as robustness to multicollinearity. For the purposes of defining the results of this model, the null hypothesis tested was that the specified soil conditions are not predictive of differences in the vegetation index. The alternative hypothesis was that specified soil conditions are predictive of differences in the vegetation index. A significance value less than or equal to 0.05 (i.e., the alpha level) was considered sufficient to reject the null hypothesis.

Table 4: Correlations Table

	EVI Difference	OrganicMatter	SAR	KFactor
EVI Difference	1	0.06	0.03	-0.06
OrganicMatter	0.06	1	0.12	-0.07
SAR	0.03	0.12	1	0.13
KFactor	-0.06	-0.07	0.12	1

Table 5: Variables

Variable Name	Definition	IV/DV	Unit	Preprocessing	Variable Type	Source
EVI Difference	Vegetation Difference	IV	Index Value	Calculated Difference Over Time	Interval	NASA/MODIS
Organic Matter %	% of soil that is organic matter	DV	Percent	None	Ratio	STATSGO2
WEG	Wind Erodibility Group	DV	Group ID	None	Ordinal	STATSGO2
SAR	Sodium Adsorption Ratio	DV	Ratio	None	Ratio	STATSGO2
K Factor	Succeptibility to Water Erosion	DV	Index Value	None	Ratio	STATSGO2

Because the vegetation data and soils data were not mapped to identical coordinate points, the raster layers of the maps were converted to polygon forms for classification. This allowed the predictor and target variables to be directly comparable. The forest-based regression was employed with 20% of the dataset withheld for validation and 100 trees set as the parameter.

Results

Predictive Results

The forest-based regression successfully yielded the boundary map of predicted change in vegetation based on the soils data for the measurement period (map in Figure 6). The bin intervals for the map were determined by the break point detection in the regression tool. For the training data, the regression yielded an R squared of 0.177, a p-value < 0.001, and a standard error of 0.004. For the validation data, the results were an R squared of 0.081, a p-value < 0.001, and a standard error of 0.009. This p-value for the validation data was sufficient to reject the null hypothesis that the soil variables were not predictive of changes in vegetation over time. The R squared corresponds to the model output's report of 8.148 percent of variation explained by the model. The regression modeled the relative importance of the variables as follows: Organic matter at 45%, susceptibility to water erosion at 27%, wind erodibility group at 21%, and sodium adsorption ratio at 7%, each of these figures indicating the percent of the variation explained by the model accounted for by the variable. Repeating the regression model to get different validation samples provided similar results, indicating consistency of the model for this data. This indicates a small but statistically significant effect from the independent variables of this model. Included in appendix C is the complete text results output.

Viewing these results in the context of degradation prediction, it is important to note that many potential variables that can influence differences in vegetation growth over time were not

included, and this certainly influenced the size of the effect of this model's prediction. For instance, precipitation was not included as a variable in this predictive analysis, as issues such as a state of drought, while useful to degradation analysis, could have a masking effect on

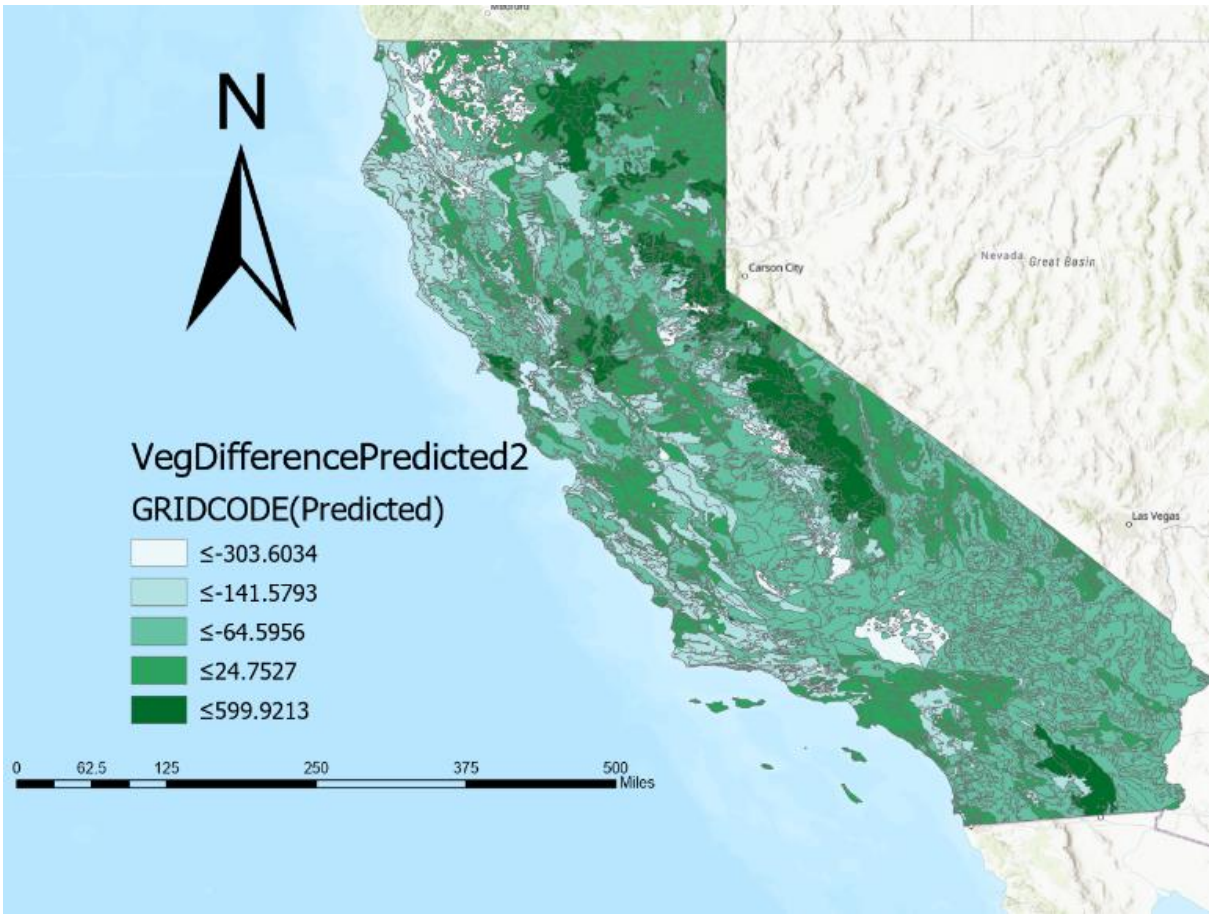

Figure 6: Map of Model Results

evaluation of the specific influence of soil states on degradation in general and vegetation growth differences in particular With this in mind, these results provide evidence that soil states at a point in time are predictive of a certain degree of difference in vegetation growth and can be used in conjunction with other metrics as one of many indicators of land degradation in an area. The areas found to have losses in vegetation growth ascribable by this model to soil conditions denote areas with evidence of degradation from multiple types of data sources.

Given the previous research indications that reliance on remote sensing alone to characterize land conditions is inconclusive, and that site collected data is also insufficient, this offers a demonstration of the combination of these two factors for assessment (Higginbottom & Symeonakis, 2014). Furthermore, by predicting remote sensing data with field survey data from a past point, the usefulness of field surveys past the time of collection for this task is reinforced. The combination and development of these techniques along with other evaluation metrics can lead to better and earlier predictions of degradation conditions, improving both research into the phenomenon and also intervention for preventative purposes. The understanding of these variables as indicators of land degradation is enhanced, and also a model is generated of what areas are of greatest concern for future investigations into degradation in this geographic area.

Immediate next steps then include the investigation into the specific regions where soil conditions were predictive of negative vegetation with other data sources and observations, either new field observations or more detailed data. The use of vegetation growth difference can also be enhanced by the development of a rigorous definition of a break-point for notable change quantities and also the evaluation of it in terms of precipitation data, which would allow the concept of rain use efficiency to be compared against these soils data as well. Other immediate future research would involve the analysis of other early indicators including how early a prediction must be to be useful, as well as what sorts of patterns degradation processes such as erosion follow over time.

Ethical Considerations and Bias

As all research into land degradation is inherently connected to land management, natural resources, food security, and climate, a number of ethical considerations arise when it is engaged

in. This study is no exception, and it is important to consider the greater social and ecological context of the research to avoid potential harm or misuse of the results.

Sociological Implications

As discussed in the review of the existing research literature on land degradation, not only does land degradation vary in its effect on populations due to socio-economic conditions, the definition can be phrased in terms of subjective human consequences due to different interest groups. Landowners, conservationists, lawmakers, and residents are all groups with different considerations regarding what level of changes in a landscape might be negative enough to regard as degradation. As such, when evaluating ethical concerns relating to any research in land degradation, it is vital to examine which individuals are most at risk from improper action. *The IPBES Assessment Report on Land Degradation and Restoration* identifies those in poverty as being the most impacted, largely due to their disproportionate involvement in and reliance on agriculture (2018). For instance, the adoption of no-till methods involving constant vegetative land cover, as suggested in the IPBES report, has associated labor, equipment and adoption costs that would not be easily absorbed by those already under economic stress (Zhang *et al*., 2018). Such considerations make it clear that this research not only has implications for those who work in agriculture or land management, but also that those who are the most susceptible to potential harm from improper research or application of the results are the populations who are already the most vulnerable.

Possibility of Unintended Consequences and Relationship to Land Management

Considering the social context, those factors that might arise from the misapplication of this research must be considered. First, under the above-mentioned assumption that land degradation is a special concern for the impoverished and for agricultural workers in general, the accuracy of

research results and rigorous methodology is an essential ethical consideration. Both false positive and false negative results have the potential for causing harm. False negatives can result in the failure to identify degrading land early enough to intervene, allowing ecological harm to continue, threatening the above-mentioned populations. However, since those populations are also identified as disproportionately impoverished, and therefore economically vulnerable, false-positive results could, if used in policymaking decisions, lead to interventions and regulatory actions that are unnecessary for ecological protection that still have negative economic consequences for those who often cannot afford any further hardship.

Limitations

Among the factors introducing a degree of uncertainty to these results is a lack of indication in the data of the influences of human activity on vegetation growth in the area. Land degradation is acknowledged in the literature earlier reviewed as being due to a combination of human activity and natural processes, exacerbated by such factors as global climate change, and as such, ignorance of the influences of human intervention has the potential to cause a misjudgment of analysis (Zhang *et al.*, 2018). If, for example, intensive agriculture is undertaken in a region, a greater degree of vegetation than the soil would otherwise support might be present in an area. It is entirely possible that new farming activity might result in an artificial increase in vegetation in a particular geographic area. Such possibilities weaken the predictive strength of vegetation difference as an indicator, as land-use decisions might impact the vegetation index difference more than the variables relating to soil conditions.

Another limitation of this research lies in the nature of the soils data. As mentioned in the overview of the dataset, the soil survey is dated from 2006 but with indications that it contains samples from older surveys to complete the geographic area. This indicates that while no data of

the soil survey is from after 2006, substantial portions are from older surveys which might have had less rigorous methodologies or — perhaps more concerning — dated from times when soil conditions were different. If conditions have declined or improved since the measurement of a segment of soils data prior to the beginning of the vegetation measurement dates used, the predictive power might also be weakened.

It is also of note that while this study is designed in a way that is repeatable and has theoretical justification why it would be effective in another geographic area, that the data sampled for this case study is geographically limited to California, and due to this fact, generalizations should not be applied to other regions without further iteration of the process in different geographies. Furthermore, as the state itself covers such a wide variety of biomes where variables outside the measured soil traits are not uniform across the area, conclusions across even the region measured in this study are comparatively weaker than they would be in a more focused geography.

Sources of Bias

The data sources themselves have sources of bias and quality limitations inherent to the ways they were collected. The vegetation data, being machine collected, does not contain any bias in its collection, however the use of the EVI as an evaluation procedure for light reflection as a proxy for vegetation growth requires acceptance of the premise that the equation yields an accurate image of vegetation in an area. As the equation was described in the metadata as being designed to account for factors like oversaturation and ground canopy conditions, the possibility exists of an overcompensation or incorrect judgment in this design. Furthermore, the data collection, while satellite-derived, is still to some degree susceptible to interference from atmospheric conditions — a factor that is also acknowledged by the data source. The soil survey

data, by contrast, holds biases inherent by the collection of data by human agents. While the survey data is provided by a government agency and intended to offer as accurate an image as possible of soils in each geographic region, the USDA acknowledges in the metadata that the surveys that comprise the source of this dataset were collected by numerous professionals at various times, and the standards of collection and evaluation of survey data might not have been uniform across these timeframes. A potential lack of uniformity of data collection has the potential to create reliability issues for analysis done over large landmasses such as used in this study.

In addition to the aforementioned limitations and biases inherent in the data and methodology, the author also acknowledges several sources of personal bias as relates to the field of study here. As this study builds on previous work the author was engaged in, there is a possibility for confirmation bias due to a priming effect of previous research into desertification connecting to existing beliefs regarding the state of land degradation and how widespread it is. This also introduces the propensity to ascribe greater value to the results than is due as a form of effort justification. Without a careful detachment, an assessment of the conclusions of the results could easily be stretched further than their actual implications. Even from a methodological standpoint, biases exist that can interfere with this work. The information bias could lead to the use of more variables than is parsimonious, possibly even including variables that would have functionally been noise for this particular analysis. Furthermore, human pattern-seeking behavior, such as falling victim to the clustering illusion could have led to conclusions being drawn from the data here that was not merited.

Conclusion

To address the question of what factors are most predictive of land degradation, a model was specified that gave a small but statistically significant prediction of the vegetation growth difference from the time of collection of the model data to the last full year of the measurement period for the vegetation data. The model avoided factors that the literature indicated were likely to result in acute differences in vegetation growth unrelated to degradation, such as precipitation. The variables used for the model found to be predictive of this vegetation growth difference were soil organic content, the sodium adsorption ratio, the wind erodibility group, and susceptibility to water erosion as specified by the k-factor. All of these variables had a role in the prediction of the model, but the soil organic content had the largest influence of the variables on the model, at 45%, and the sodium adsorption ratio had the smallest influence, with 7% of model prediction relying on that variable. The model itself was found to have a small but statistically significant predictive power, indicating a possibility for these variables to be used with this or related techniques for predictive assessment of land degradation.

Future Research Indications

Following this work, several avenues are evident where future research would expand knowledge related to the questions posed. First, an investigation into the variables and assumptions of this model could yield more accurate predictions. For instance, the soils data only included a relatively small number of candidate features from that dataset as well as only the surface layer defined by the USDA. Other features and soil layers might have the potential to produce different insights. Some variables that might be of interest such as drought conditions or precipitation variability in general, were excluded in this study. Including them in a future model or controlling for them more explicitly could improve future predictions. The iteration of this or

related methodologies in other geographic regions or with different focus is another important step. This can include both more focused model building, such as in specific areas known to have problems related to degradation, and also locations not included. Outside the United States, this would necessitate access to different data sources for soils, which in some regions might have different levels of quality, however testing the conclusions here in various geographic would probe to what degree they are generalizable. Delving further into the soils dataset is another avenue for future research. A relatively small subset of the soil data was used here and investigating different soil features or including soil layers beyond the surface layer to determine whether they improve predictions would also be a valuable investigation. Furthermore, the USDA offers more detailed SSURGO soil survey dataset that offers even more granular soils data, but greater challenges with acquiring samples for larger land areas. The use of different resolutions of vegetation data in follow-up research is another avenue of future investigation that might improve conclusions. The NDVI and EVI are also available at a resolution up to a 250m scale. In research requiring physical methodologies rather than the use of existing data, the utility of mixing field observation with remotely sensed data mentioned in the review of literature on degradation indicate that more manually derived soil surveys or human observation of vegetative conditions would also be a valuable source for new data for future research.

References

Amato, F., J. Havel, A-A. Gad, & A. El-Zeiny (2015). Remotely Sensed Soil Data Analysis Using Artificial Neural Networks: A Case Study of El-Fayoum Depression, Egypt. *ISPRS International Journal of Geo-Information*, 4(2), 677–696. doi: 10.3390/ijgi4020677

Anderson, R. M., V.I. Koren, & S. M. Reed (2006). Using SSURGO data to improve Sacramento Model a priori parameter estimates. *Journal of Hydrology*, 320(1-2), 103–116. doi: 10.1016/j.jhydrol.2005.07.020

Bojórquez-Tapia, L. A., G. M. Cruz-Bello, & L. Luna-González (2013). Connotative land degradation mapping: A knowledge-based approach to land degradation assessment. *Environmental Modelling & Software*, 40, 51–64. doi: 10.1016/j.envsoft.2012.07.009

Chikhaoui, M., F. Bonn, A. I. Bokoye, & A. Merzouk (2005). A spectral index for land degradation mapping using ASTER data: Application to a semi-arid Mediterranean catchment. *International Journal of Applied Earth Observation and Geoinformation*, 7(2), 140–153. doi: 10.1016/j.jag.2005.01.002

Didan, K. (2015). MOD13A1 MODIS/Terra Vegetation Indices 16-Day L3 Global 500m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2019-11-03 from https://doi.org/10.5067/MODIS/MOD13A1.006

Easdale, MH, O. Bruzzone, P. Mapfumo, P. Tittonell(2018). Phases or regimes? Revisiting NDVI trends as proxies for land degradation. *Land Degrad Dev*. 2018; 29: 433– 445. https://doi.org/10.1002/ldr.2871

Eswaran, H., R. Lal, & P. Reich (2019). Land degradation: An overview. *Response to Land Degradation*, 20–35. doi: 10.1201/9780429187957-4

Fensholt, R., & K. Rasmussen (2011). Analysis of trends in the Sahelian 'rain-use efficiency' using GIMMS NDVI, RFE and GPCP rainfall data. *Remote Sensing of Environment*, 115(2), 438–451. doi: 10.1016/j.rse.2010.09.014

Haining, R. (2015). International encyclopedia of the social & behavioral sciences. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 105–110). Amsterdam: Elsevier.

Herrmann, S. M., & T. K. Sop (2016). The Map Is not the Territory: How Satellite Remote Sensing and Ground Evidence Have Re-shaped the Image of Sahelian Desertification. The End of Desertification? *Springer Earth System Sciences*, 117–145. doi: 10.1007/978-3-642-16014-1_5

Higginbottom, T., & E. Symeonakis (2014). Assessing Land Degradation and Desertification Using Vegetation Index Data: Current Frameworks and Future Directions. *Remote Sensing*, 6(10), 9552–9575. doi: 10.3390/rs6109552

Hill, J., M. Stellmes, Th. Udelhoven, A. Röder, S. Sommer (2008). Mediterranean desertification and land degradation: Mapping related land use change syndromes based on satellite observations, *Global and Planetary Change*,Volume 64, Issues 3–4,2008, Pages 146-157,ISSN 0921-8181,https://doi.org/10.1016/j.gloplacha.2008.10.005.

Image Change Detection. (2019). Retrieved December 3, 2019, from https://solutions.arcgis.com/defense/help/image-change-detection/workflows/image-differencing/.

Kairis, O., C. Kosmas, C. Karavitis, C. Ritsema, L. Salvati, S. Acikalin, V. Fassouli (2014). Evaluation and Selection of Indicators for Land Degradation and Desertification Monitoring: Types of Degradation, Causes, and Implications for Management. *Environmental Management*, 54(5), 971–982. https://doi-org.ezproxy.utica.edu/10.1007/s00267-013-0110-0

Meadows, M. E., T. M. & Hoffman (2003). Land degradation and climate change in South Africa. *Geographical Journal*, 169(2), 168–177. https://doi-org.ezproxy.utica.edu/10.1111/1475-4959.04982

Mikhailova, E., A. Altememe, A. Bawazir, R. Chandler, M. Cope, C. Post, M.Schlautman (2016). Comparing soil carbon estimates in glaciated soils at a farm scale using geospatial analysis of field and SSURGO data. *Geoderma*, 281, 119–126. doi: 10.1016/j.geoderma.2016.06.029

Nachtergaele, F., M. Petri, & R. Biancalani (2011). Land degradation. SOLAW background thematic report, 3.

Obalum, S., G. Chibuike, S. Peth, & Y. Ouyang (2017). Soil organic matter as sole indicator of soil degradation. *Environmental Monitoring and Assessment*, 189(4). doi: 10.1007/s10661-017-5881-y

Shrestha, D. (2004). Modelling land degradation in the Nepalese Himalaya. *Catena*, 57(2), 135–156. doi: 10.1016/s0341-8162(03)00241-8

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture (2006). U.S. General Soil Map (STATSGO2). Available online. Accessed [Nov/14/2019].

Wang, J., T. He, C. Lv, Y. Chen, & W. Jian (2010). Mapping soil organic matter based on land degradation spectral response units using Hyperion images. *International Journal of Applied Earth Observation and Geoinformation*, 12. doi: 10.1016/j.jag.2010.01.002

Yengoh, G. T., D. Dent, L. Olsson, A. E. Tengberg, & C. J. Tucker (2015). Applications of NDVI for Land Degradation Assessment. Use of the Normalized Difference Vegetation Index (NDVI) to Assess Land Degradation at Multiple Scales SpringerBriefs in Environmental *Science*, 17–25. doi: 10.1007/978-3-319-24112-8_3

Zhang, B., Pan, Y., Xu, J., & Tian, Y. (2018). IPBES thematic assessment on land degradation and restoration and its potential impact. *Biodiversity Science*, *26*(11), 1243–1248. doi: 10.17520/biods.2018117

Appendices

Appendix A – Point Maps of Extent of Vegetation Change Per Month

All figures in this appendix contain point maps of the vegetation difference for individual months between the 2006 and 2018 years, filtered and divided into positive and negative values that exceed a threshold of a 0.2 change in either direction of the EVI. Decreases are mapped in red points, with increases mapped in dark green points.

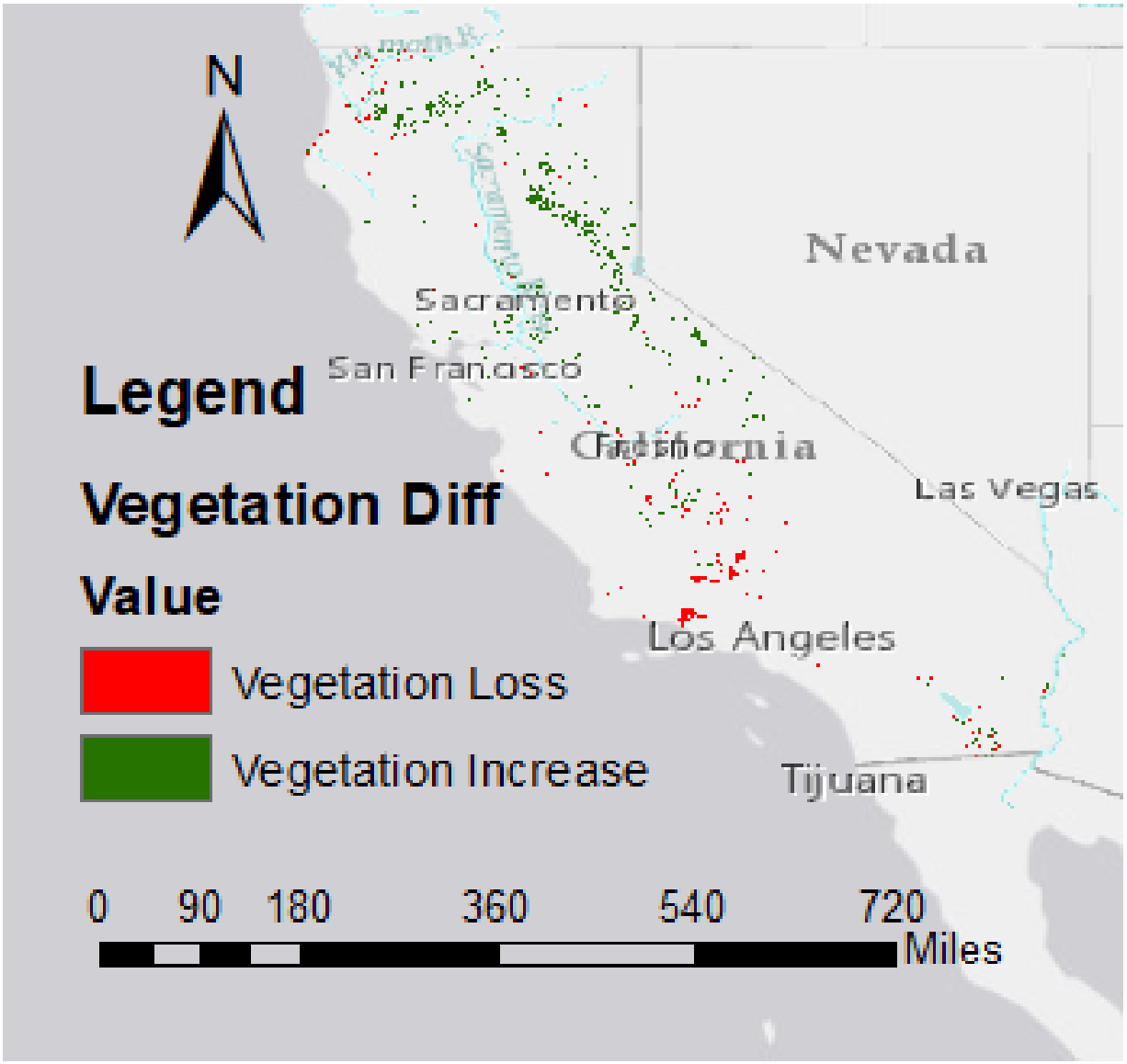

Figure 7: Vegetation Difference for January

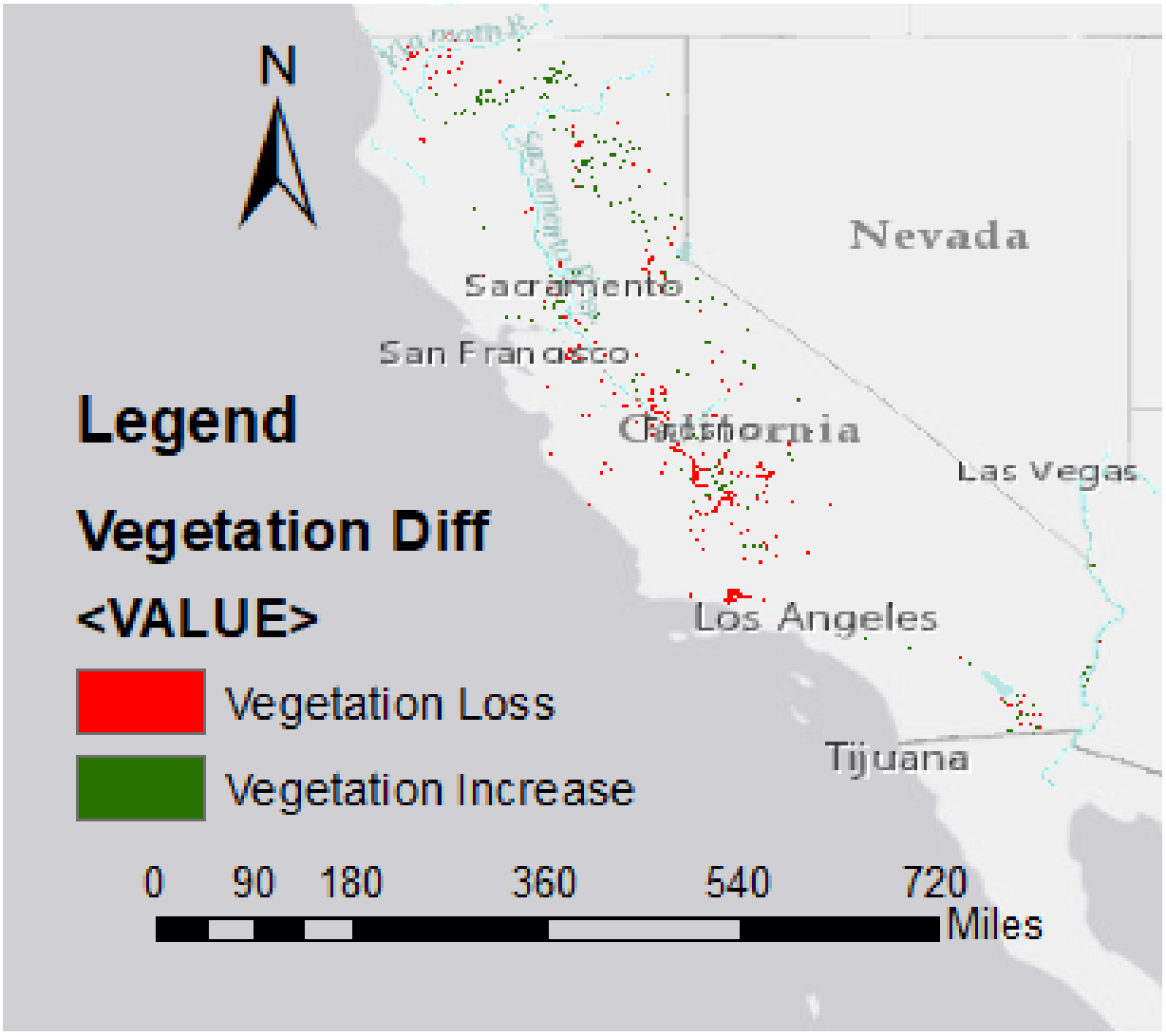

Figure 8: Vegetation Difference for February

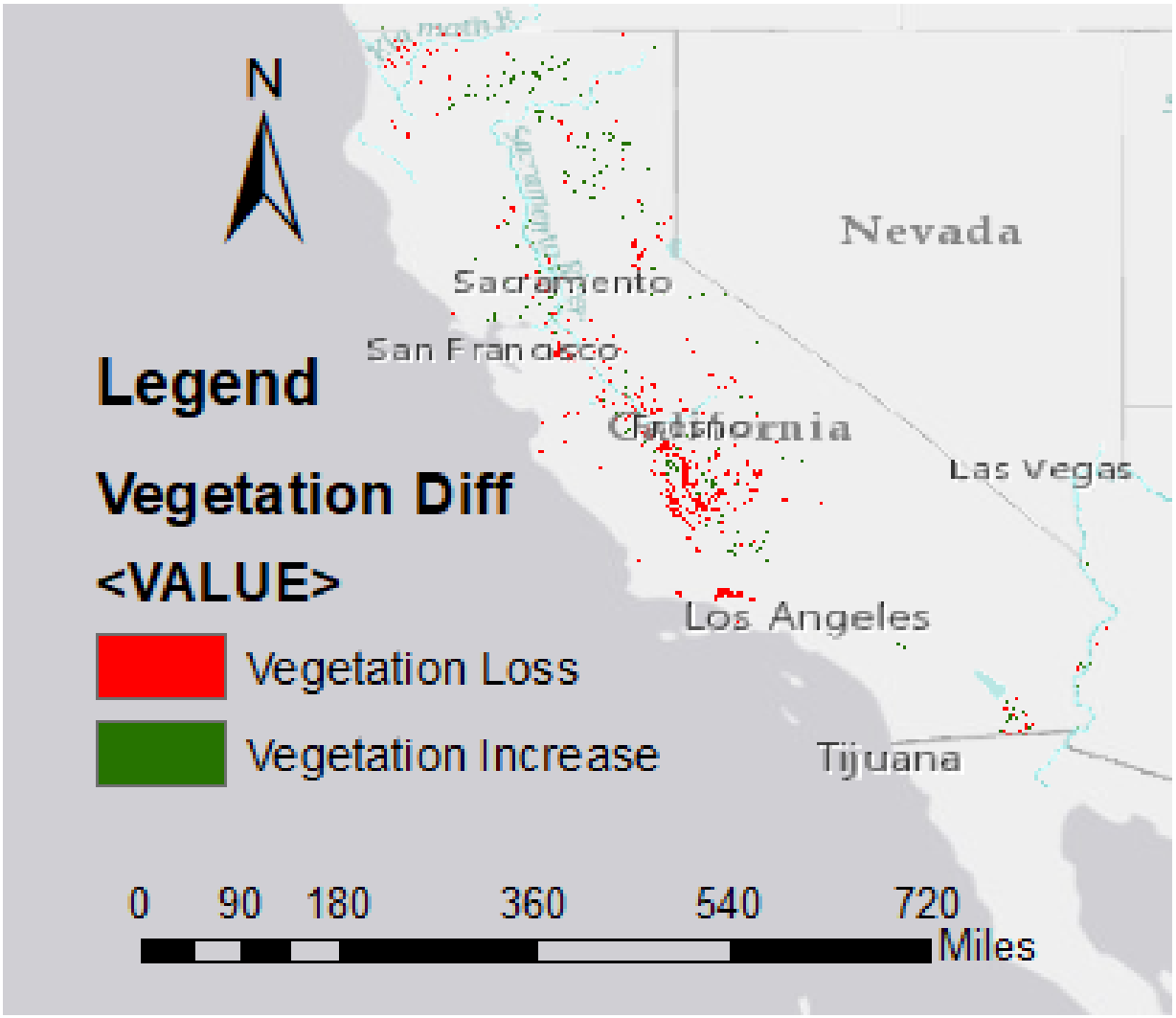

Figure 9: Vegetation Difference for March

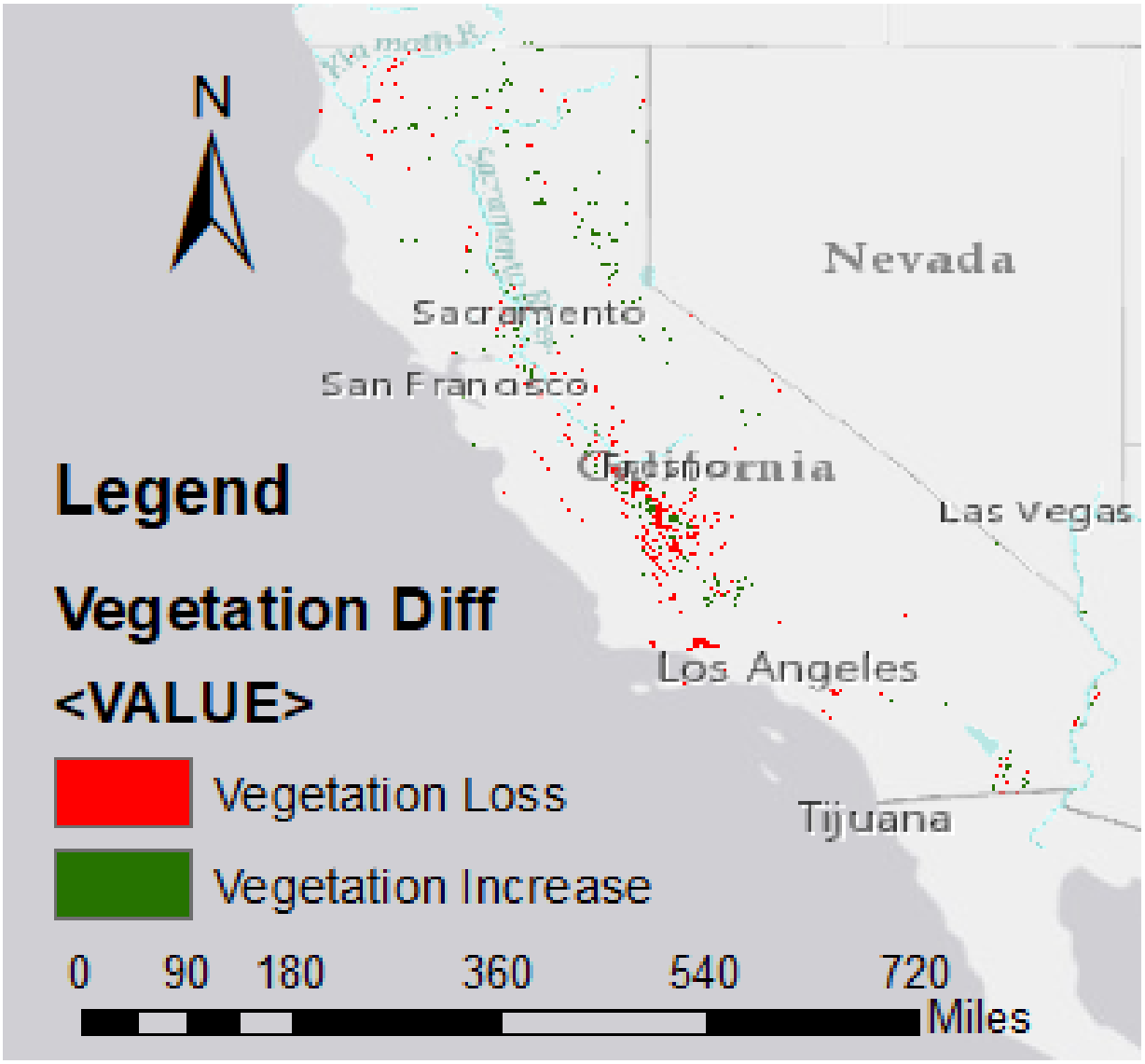

Figure 10: Vegetation Difference for April

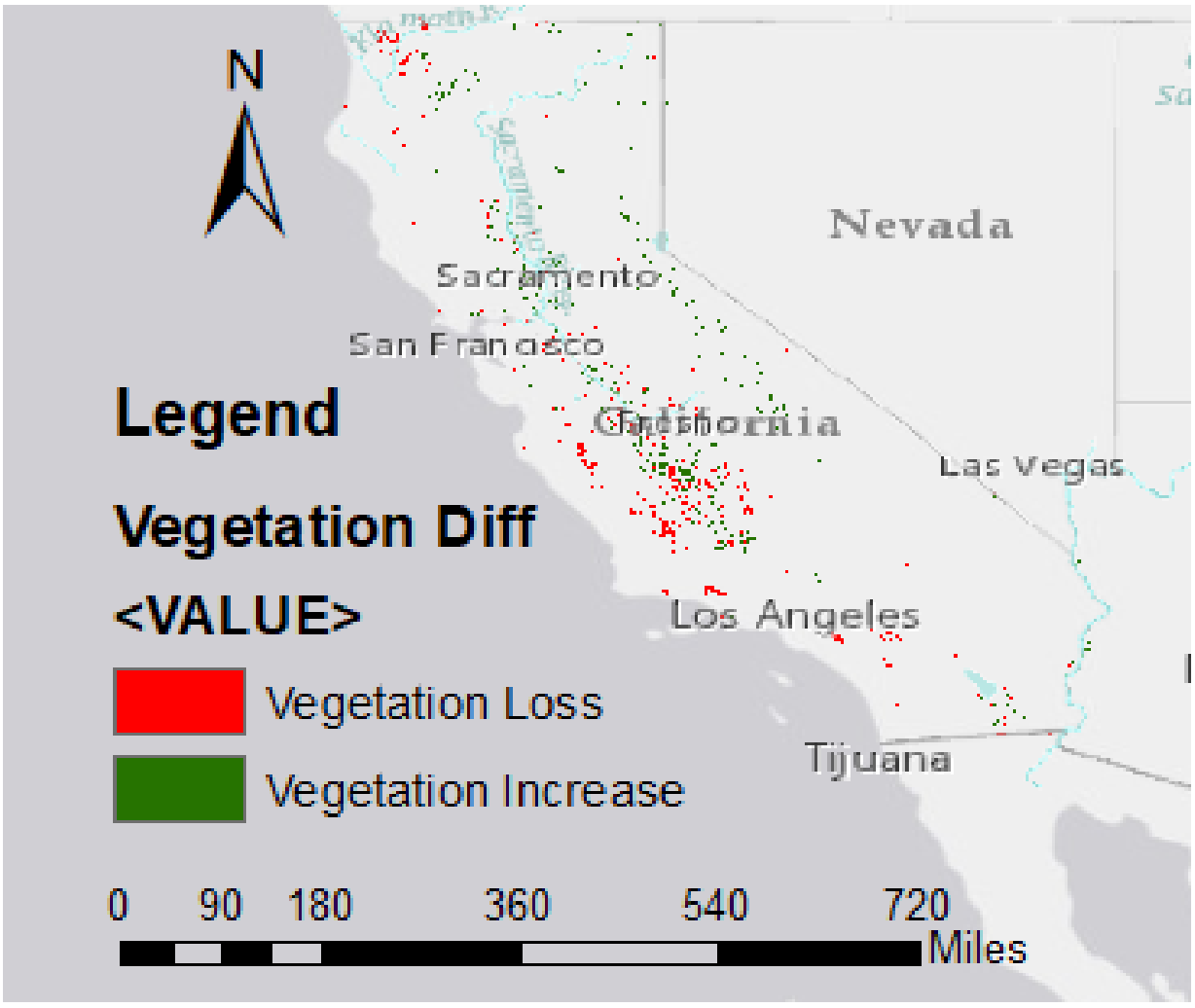

Figure 11: Vegetation Difference for May

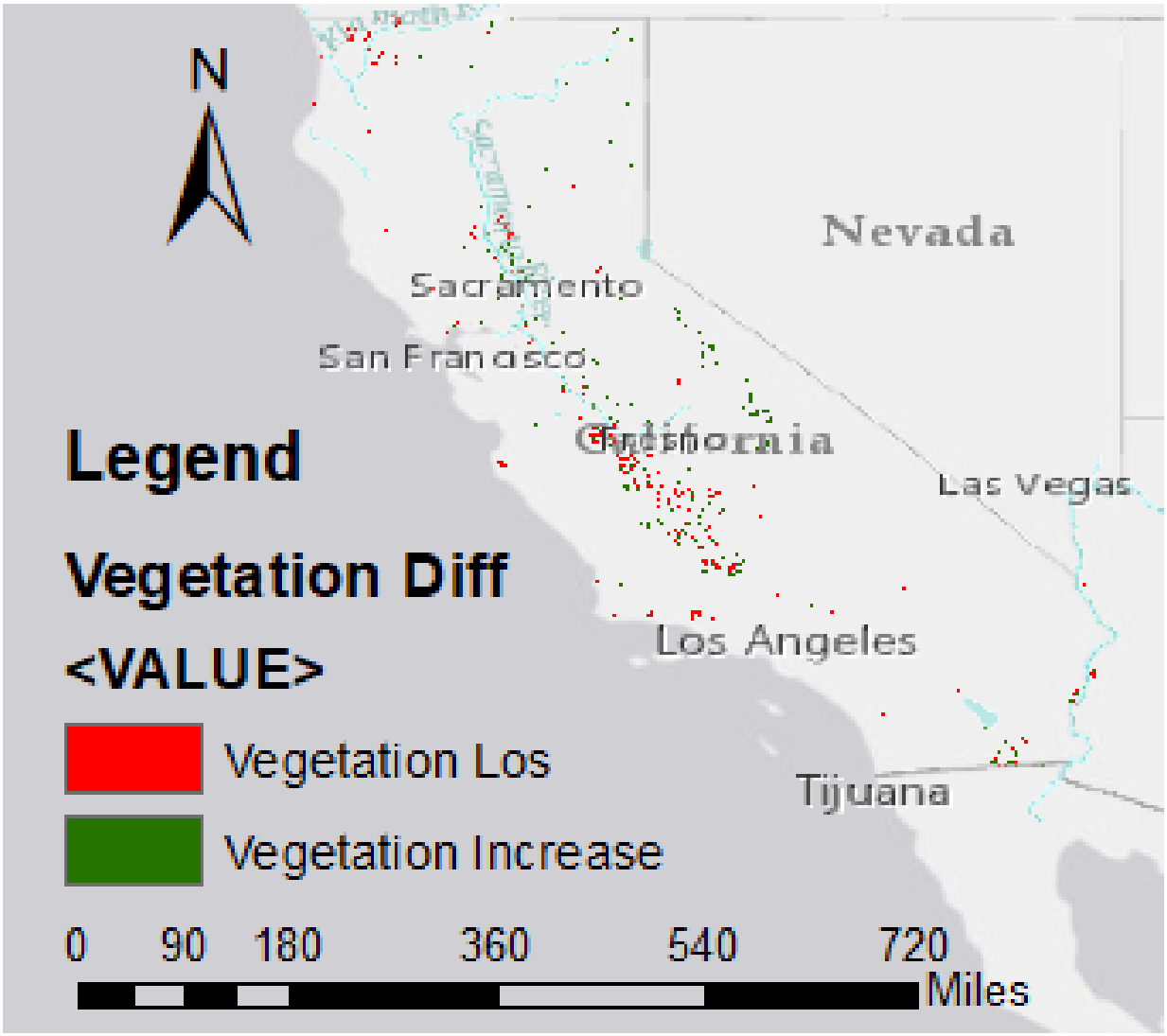

Figure 12: Vegetation Difference for June

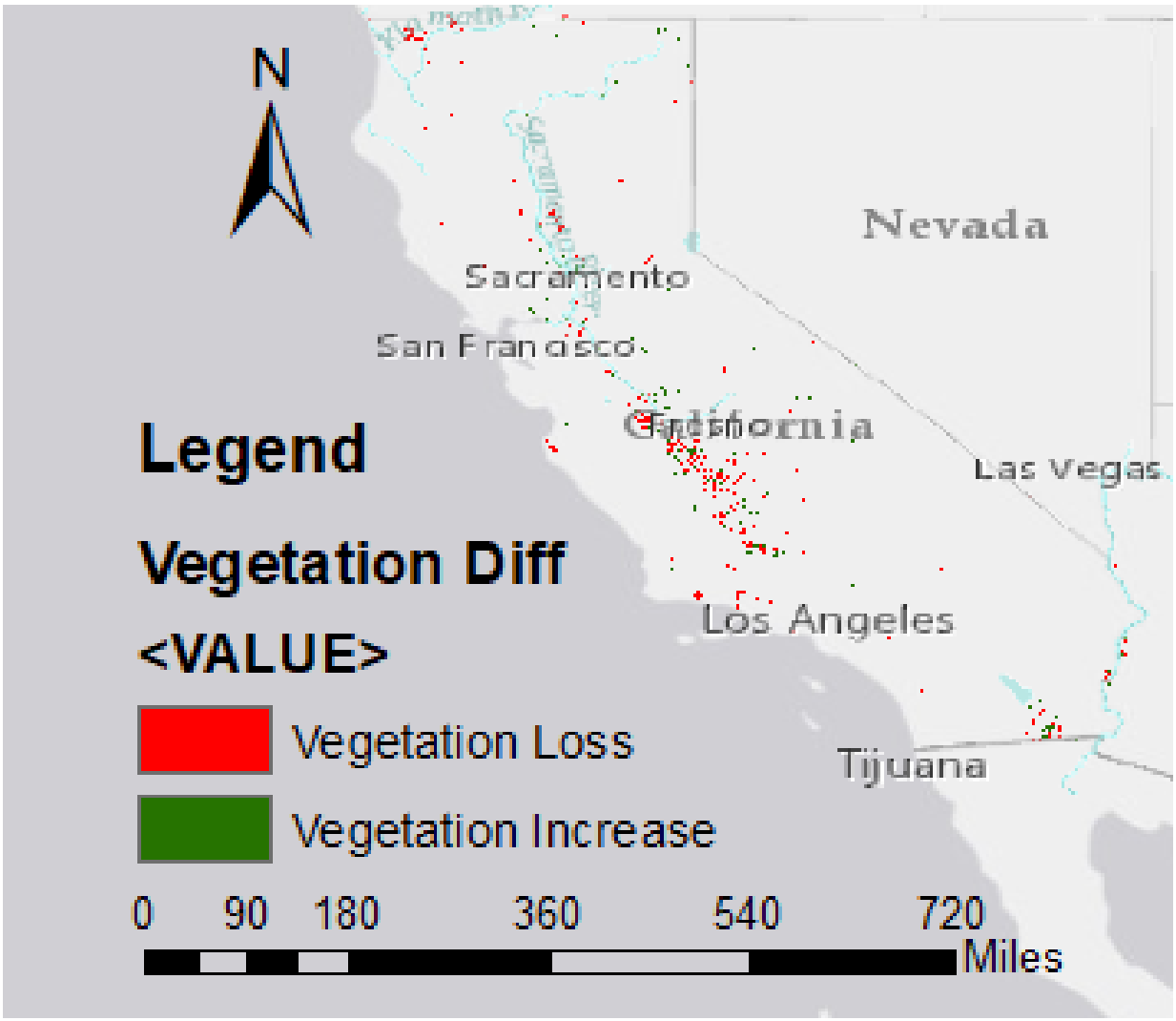

Figure 13: Vegetation Difference for July

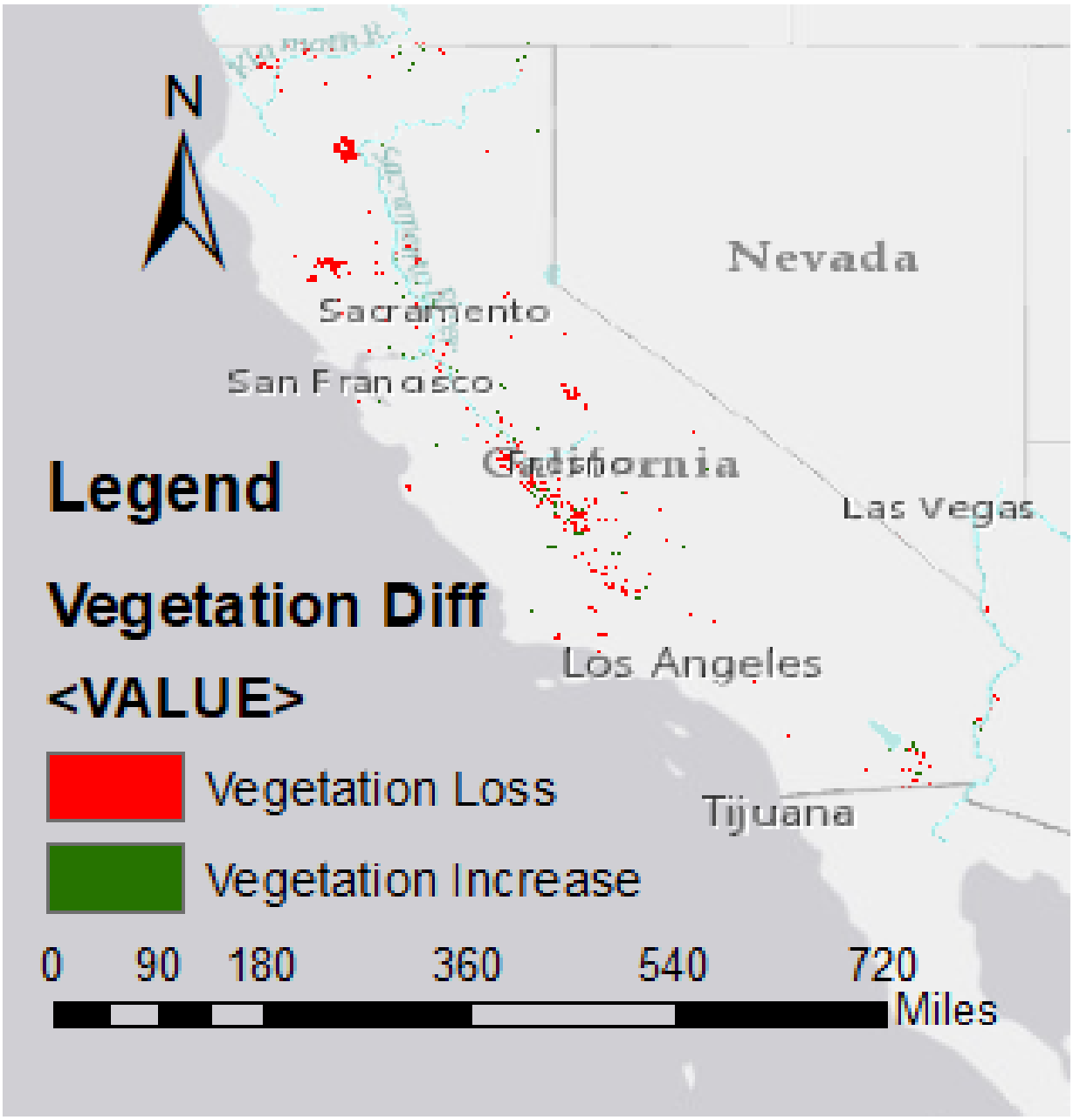

Figure 14: Vegetation Difference for August

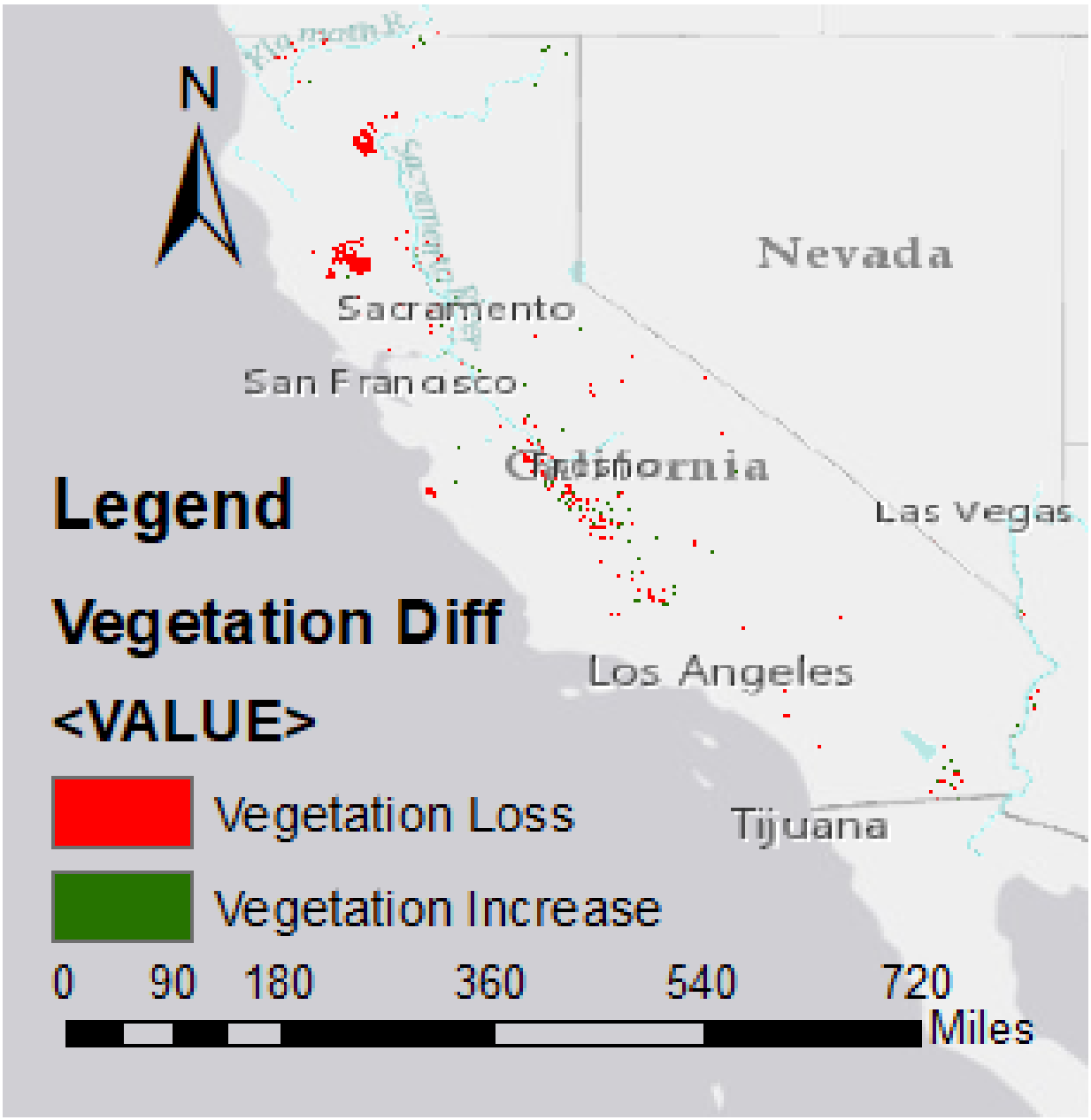

Figure 15: Vegetation Difference for September

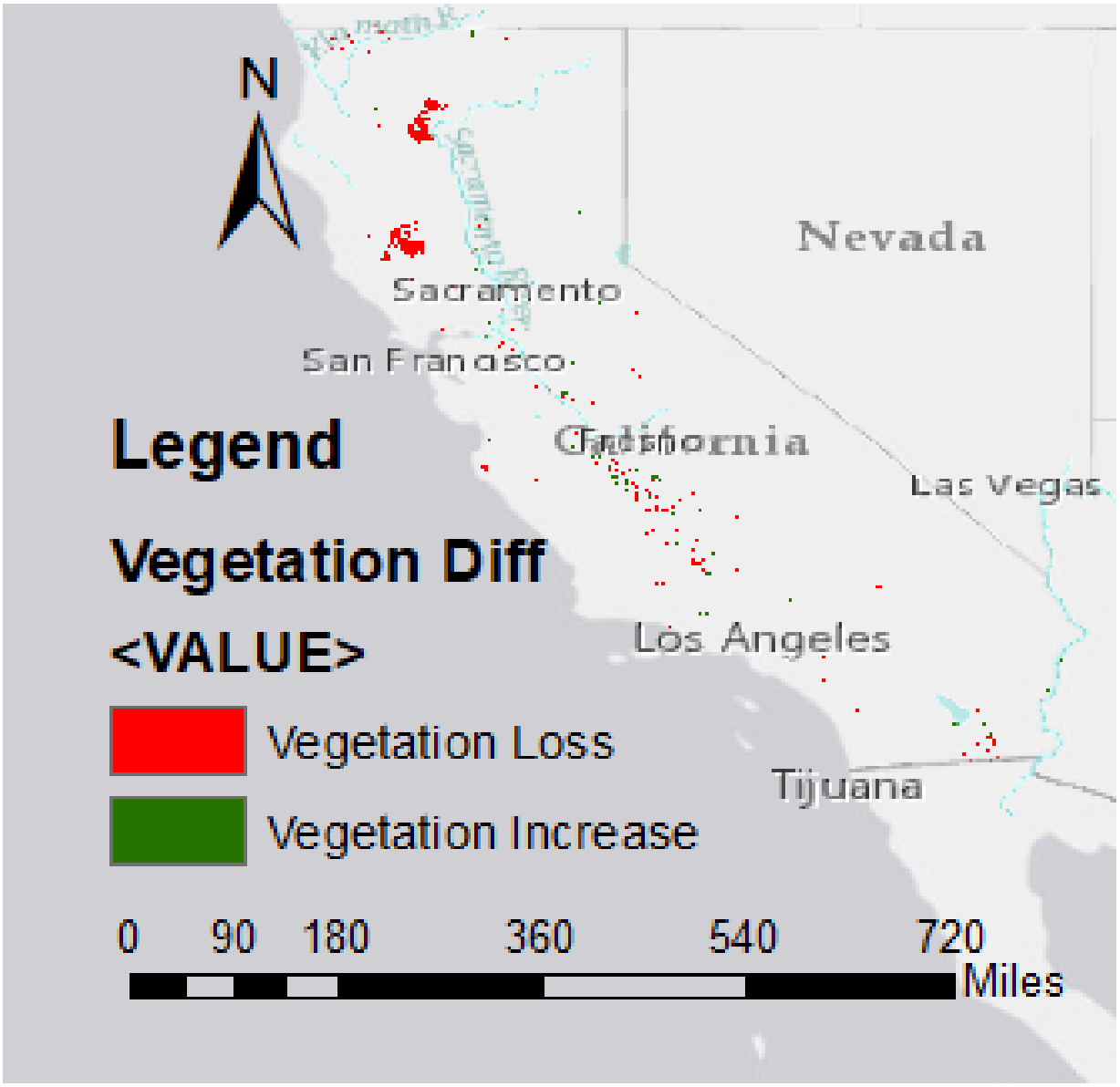

Figure 16: Vegetation Difference for October

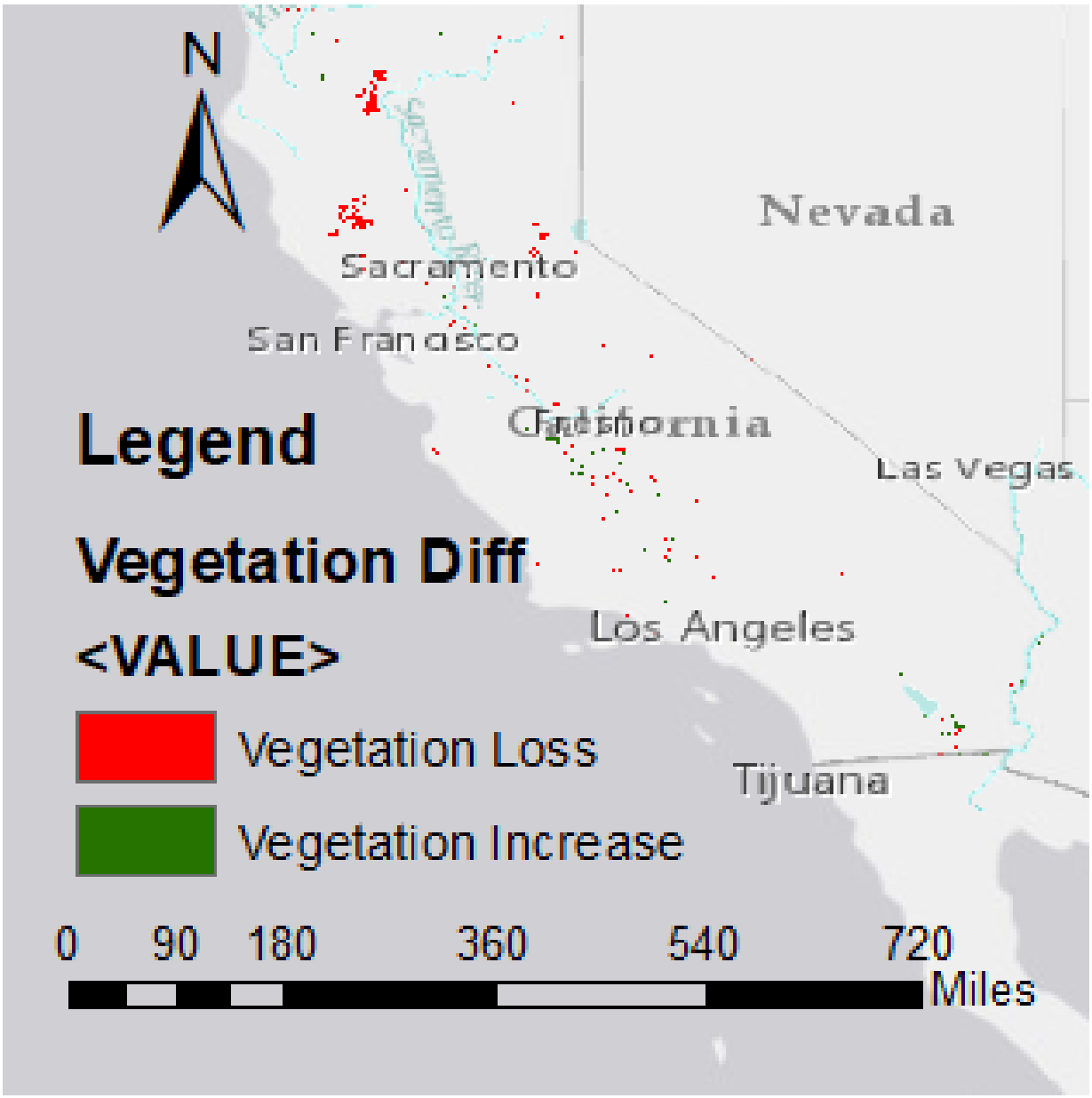

Figure 17: Vegetation Difference for November

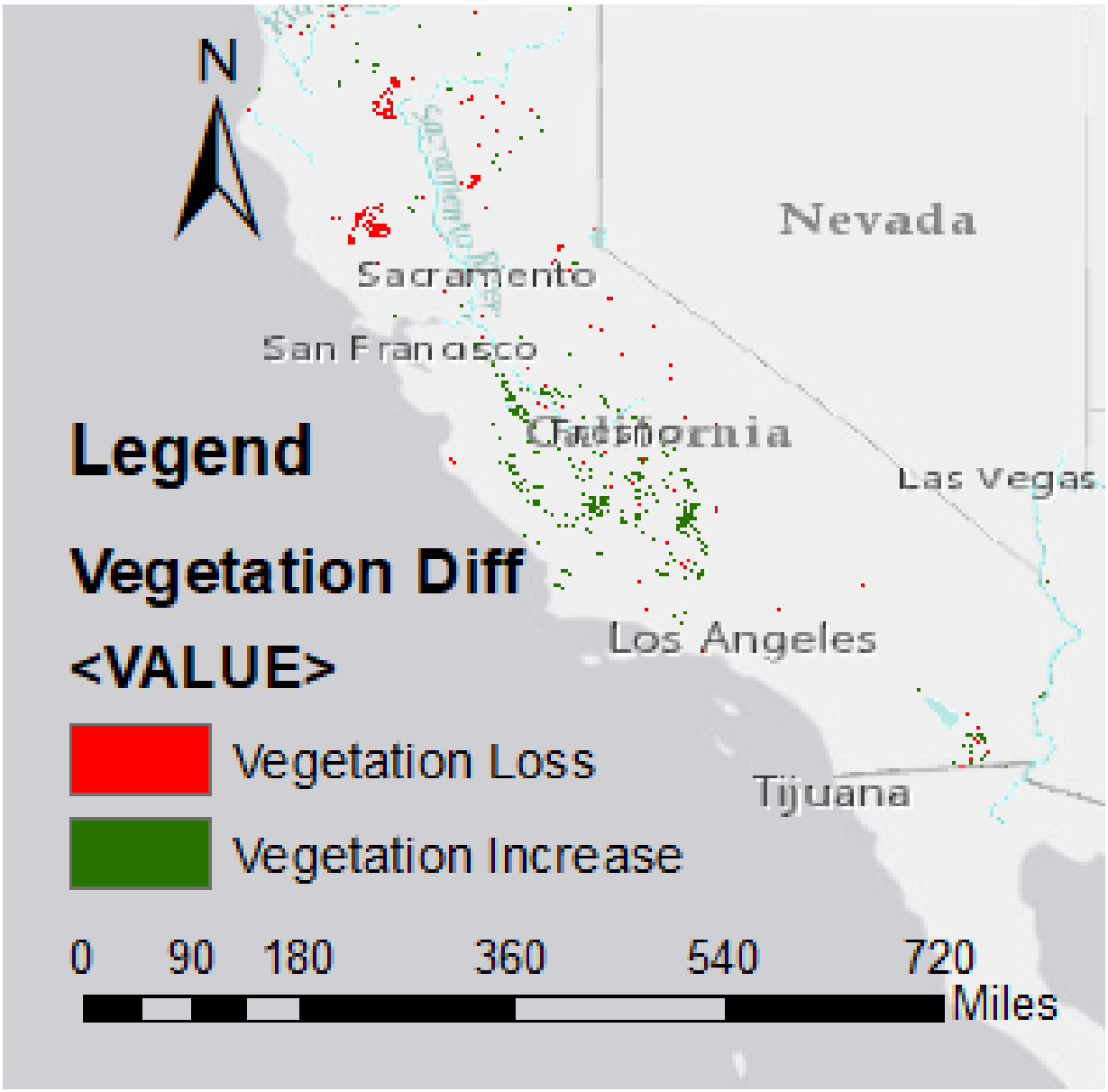

Figure 18: Vegetation Difference for December

Appendix B – Generated Maps of Each Independent Variable

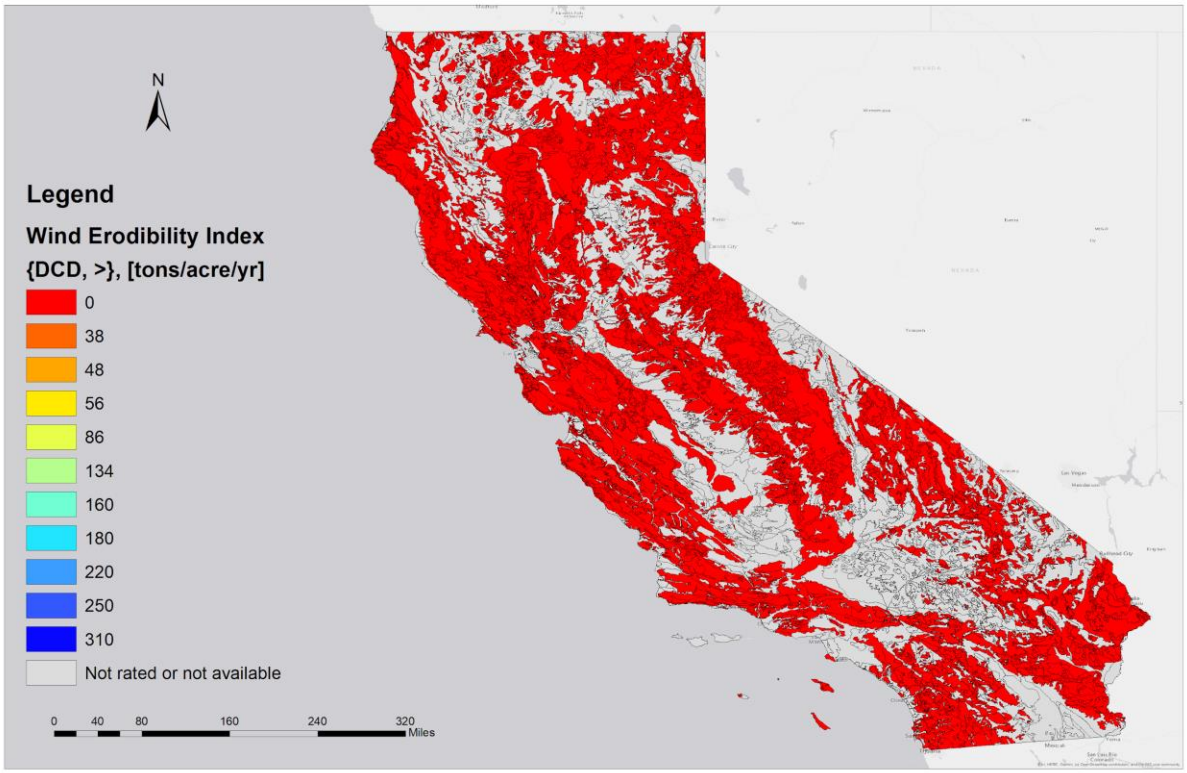

Figure 19: Wind Erodibility Index

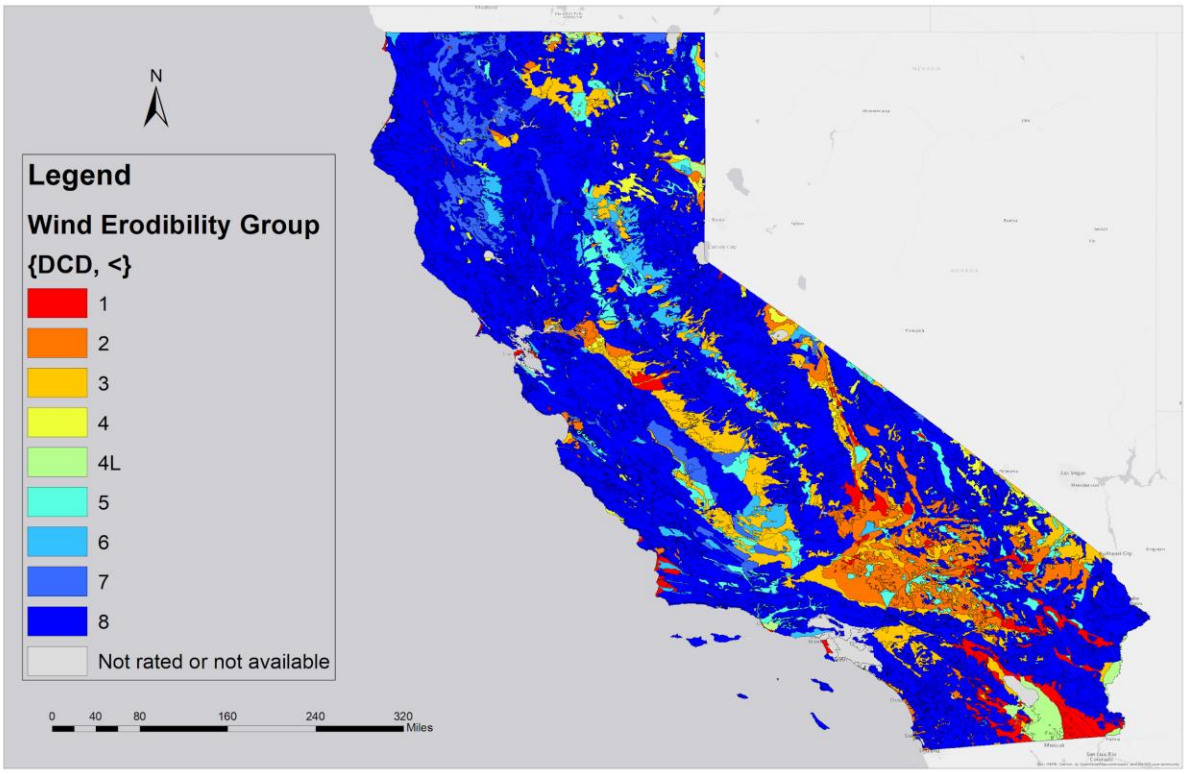

Figure 20: Wind Erodibility Group

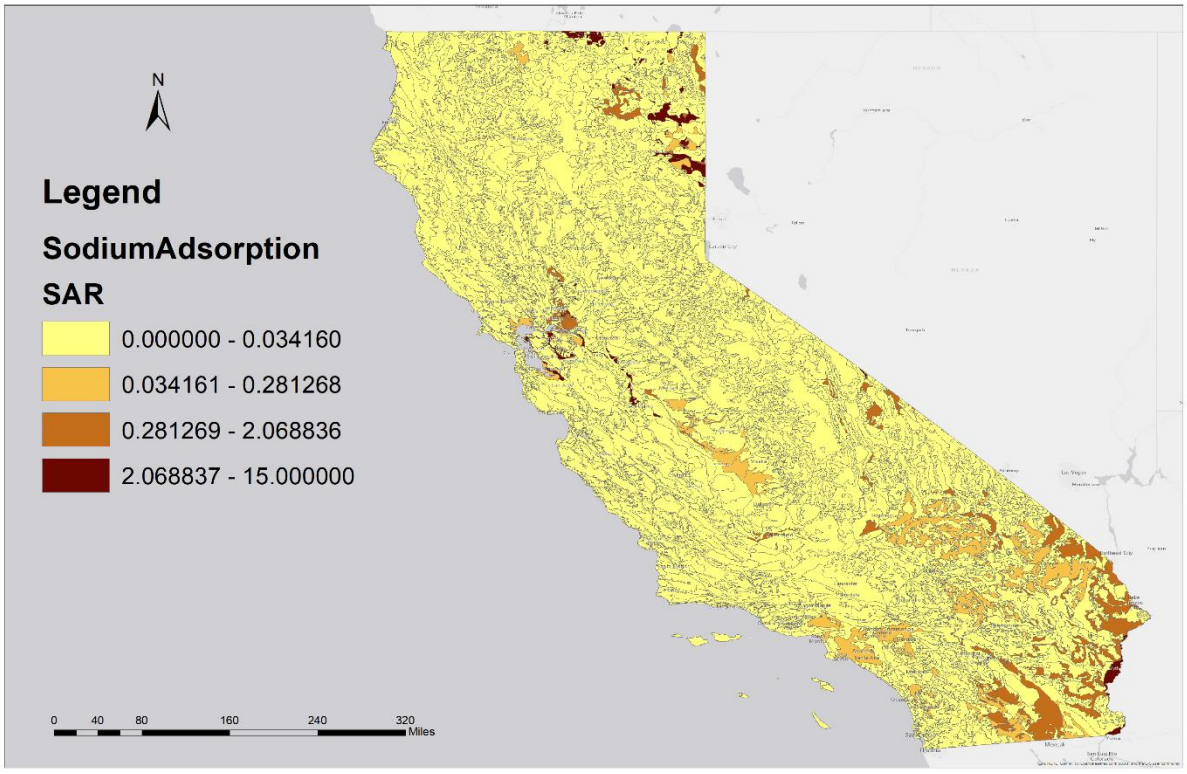

Figure 21: Sodium Adsorption Ratio (Break Points Assigned by Geometrical Interval)

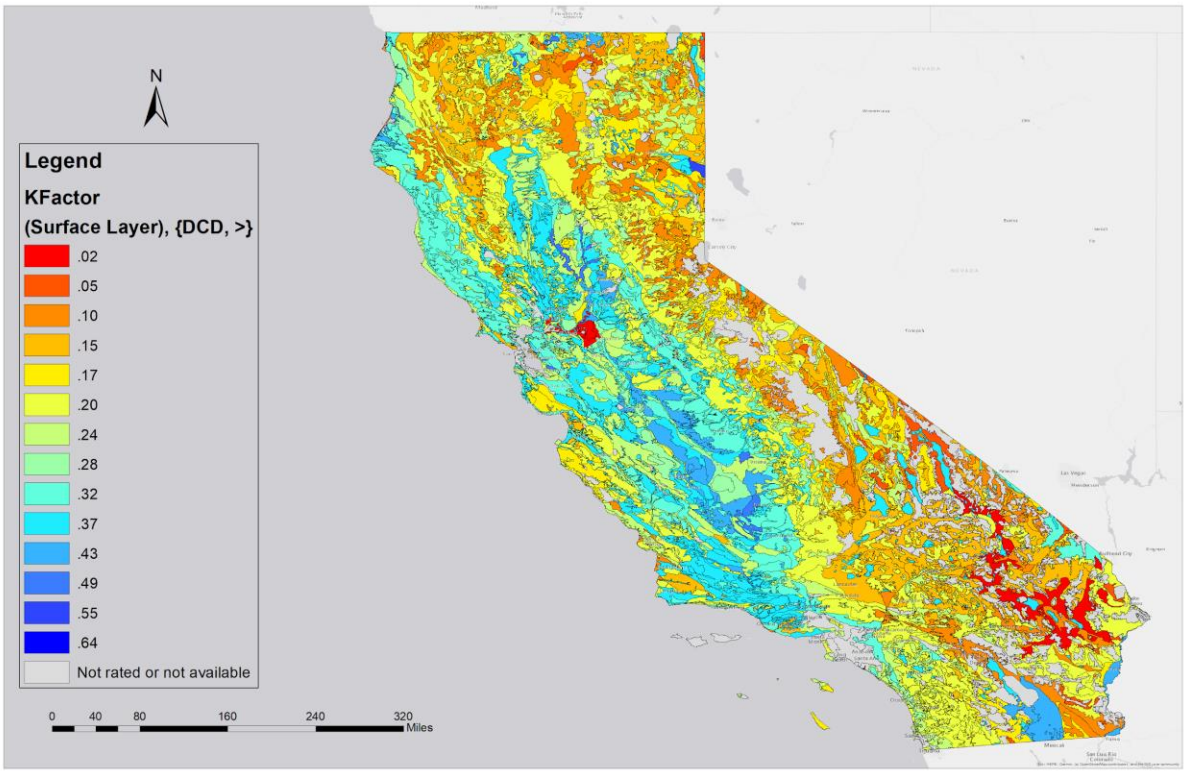

Figure 22: K Factor (Susceptibility to Water Erosion)

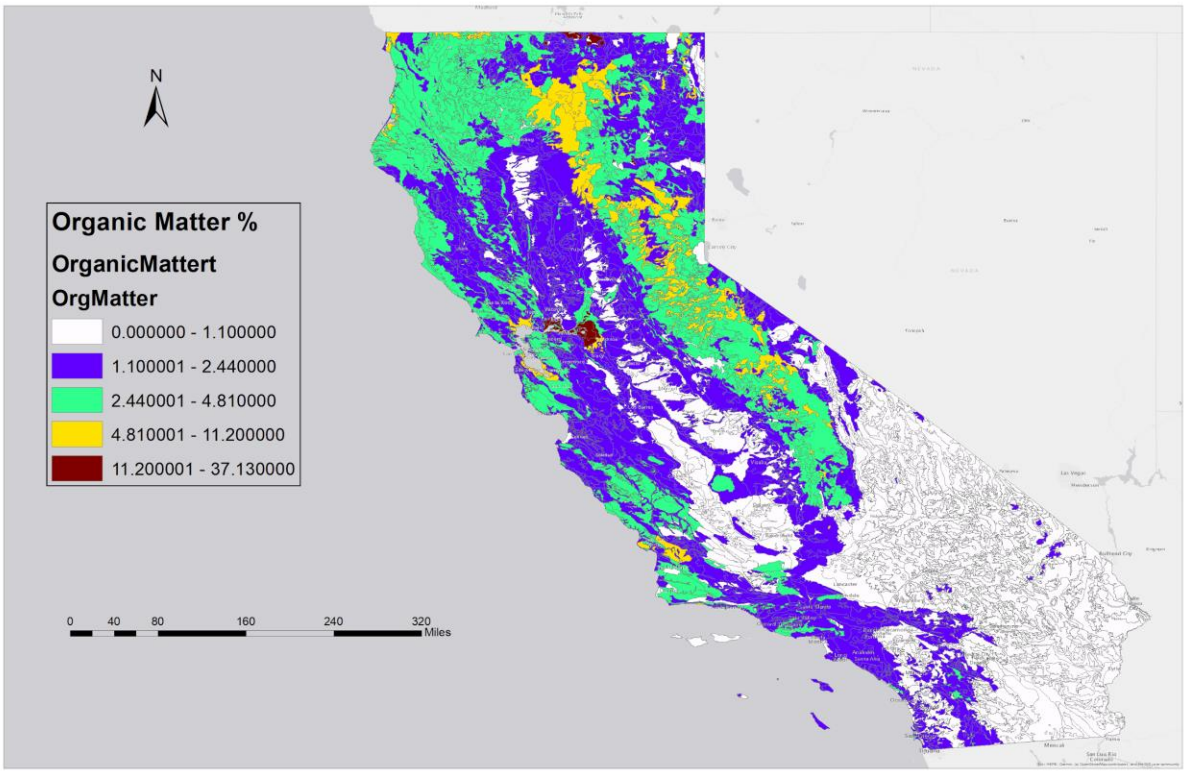

Figure 23: Soil Organic Matter %

Appendix C: Model Output Table

Table 24: Text Output of Model

Model Characteristics	
Number of Trees	100
Leaf Size	5
Tree Depth Range	7-20
Mean Tree Depth	13
% of Training Available per Tree	100
Number of Randomly Sampled Variables	1
% of Training Data Excluded for Validation	20

Model Out of Bag Errors		
Number of Trees	50	100
MSE	113930.897	113667.690
% of variation explained	7.936	8.148

Top Variable Importance		
Variable	Importance	%
OrgMatter	27911715.59	45
KfactWS	16634767.31	27
WEG	13012098.79	21
SAR	4114318.38	7

Training Data: Regression Diagnostics	
R-Squared	0.177
p-value	0
Standard Error	0.004

*Predictions for the data used to train the model compared to the observed categories for those features

Validation Data: Regression Diagnostics	
R-Squared	0.081
p-value	0
Standard Error	0.009