Species Distribution Modeling to Predict the Spread of *Spartium junceum* in the Angeles National Forest

By

Jason I Martin

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

December 2019

I dedicate this to my wife, Christina Mireles Martin, who inspired me to continue my education and never give up on pursuing the dreams and goals I had set for myself when I was just a teenager. I also dedicate this to my beautiful children Elena Rose Martin and Isaiah Eusevio Martin who have helped me regain life perspective and are the driving force for me to persevere. To my amazing family, I love you all and I thank you for empowering me in ways I never knew were possible.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor Dr. Travis Longcore, for his guidance, direction, and for helping me get this process started on the right foot. I am especially grateful for his ability to re-instill my confidence in this process and the ability to understand the material in a very profound way. I would also like to thank Dr. John P. Wilson for picking up where Dr. Travis Longcore left off and helping me get through the home stretch and on to the finish line in one piece. Additionally, I would like to thank Drs. An-Min Wu and Laura C. Loyola for coming on board at the eleventh hour and agreeing to serve as my thesis committee. I would like to thank the faculty and staff of the Spatial Sciences Institute for their patience with me throughout this entire process and their willingness to work with me and accept me into this program.

List of Abbreviations

ANF	Angeles National Forest
AUC-ROC	Area Under the curve Receiver Operating Characteristic
CAL-IPC	California Invasive Plant Council
CT	Classification Trees
CART	Classification and Regression Trees
DEM	Digital Elevation Model
ENFA	Ecological Niche Factor Analysis
ENM	Environmental Niche Modeling
GARP	Genetic Algorithm for Rule-Set Prediction
GDB	Geodatabase
GIST	Geographic Information Science & Technology
GLM	Generalized Linear Model
MAXENT	Maximum Entropy
MLR	Multiple Logistic Regression
OHV	Off -highway Vehicle
SDM	Species Distribution Modeling
SSI	Spatial Sciences Institute
USC	University of Southern California
USFS	United States Forest Service
USGS	United States Geological Survey
NOAA	National Oceanic and Atmospheric Administration

Abstract

This study predicts the spreading pattern of an invasive plant species, *Spartium junceum*, using Maxent, a type of Species Distribution Model (SDM). Species Distribution Modeling estimates the relationship between species records at sites within a given study area and the environmental or spatial characteristics of those sites. This study combines environmental variables found at sites where species occurrence has been confirmed and analyzes the results to predict future spreading patterns. A subset of occurrence data is used for quality control with the intended purpose of validating the accuracy of the model and its results. This study uses ArcGIS 10.6.1 and Maxent version 3.4.1 to perform presence-only species distribution modeling of *Spartium junceum* from data collected in 2011–2016 in the Angeles National Forest (ANF), which is managed by the U.S. Forest Service. The primary study area is the ANF with an emphasis on the San Gabriel Mountains which lie in the eastern portion of the national forest. The results will increase the exposure of Maxent as a feasible, cost-effective species distribution model that Federal land management agencies can incorporate into environmental analyses and environmental impact studies that contribute to their land management plans. The Angeles National Forest does not currently employ SDMs in its invasive species management plan and Maxent modeling represents an option for studying species distribution patterns where only presence data are available. Many studies use Maxent to understand species distributions of flora and fauna, but few such efforts are updated annually and used as a key indicator in a Federal Agency's decision-making process in their land management plans and actions.

Chapter 1 Introduction

Diverse topography, ecosystems, and flora and fauna characterize California's national forests. Native vegetation diversity is considered a key indicator of forest health, but the spread of invasive species has complicated what interpreting that diversity looks like and has thus become an important issue for land management agencies such as the US Forest Service (USFS) and local communities that neighbor national forests. The ability of invasive species to compete with native vegetation for resources and alter the local ecosystem is a primary concern to the interested parties.

The purpose of this research is to better understand the underlying factors that contribute to the spread of *Spartium junceum* (Spanish Broom) in the Angeles National Forest (ANF). *Spartium junceum* is a particularly mettlesome invasive species and is very difficult to remove and therefore is emphasized as a species for removal and other mitigation efforts by the USFS (Mullin et al. 2000). The operating premise of this research is that *Spartium junceum* thrives and spreads under certain conditions and, by identifying those conditions through species distribution modeling (SDM), accurate estimates of the geographic regions to which they may spread as the weather conditions under which they will thrive in the future evolve.

Some of the conditions used in Maxent are environmental while others are non-environmental; both conditions will be referred to as variables in the remainder of the thesis. The non-environmental variables that will be considered in this research are linear distance to nearest water source, linear distance to recreation sites, linear distance to local roads, linear distance to vegetation burn areas, and linear distance to the wildland urban interface. The environmental variables that will be considered in this research are biophysical and bioclimatic with the bioclimatic variables coming from the worldclim dataset (www.worldclim.org). The biophysical variables are as follows: elevation, slope, aspect, hill shade and visibility. The bioclimatic variables are as follows: mean diurnal temperature range (mean of monthly, maximum-minimum temperature) minimum temperature of the coldest month, annual precipitation, precipitation of the driest month, precipitation seasonality, and precipitation of warmest quarter (Qin et al., 2017) Datasets associated with each variable will be accessed in Maxent and ArcGIS to perform predictive spatial analysis.

1.1 Motivation

Federal management of national public lands and the natural resources that make up the diverse array of ecosystems in those lands is a major endeavor that requires a large annual budget and workforce on the ground at the field level. With increasing budget cuts to critical natural resource programs, it has become increasingly important to be as efficient as possible with the workforce that is already in place and incorporate new technologies in the land management process. The USFS's 2018 discretionary budget, for example, decreased by 16.5% while the amount of land management either remained same or increased. The USFS discretionary budget for fiscal year 2017 was $6.2 billion and the President's approved budget for fiscal year 2018 was $4.7 billion (USDA Forest Service, 2017). The combination of shrinking budgets and increasing recreation pose a substantial challenge for the future of the agency and its land management practices.

One of the most crucial aspects of doing applied scientific research lies in its ability to affect the planet and the lives of those who live in it in a positive way and to increase the capacity and efficiency for protecting and managing natural resources. That is precisely the goal of this research. In this case, the science can benefit the USFS, which is the second largest federal land management agency in the U.S. that is tasked with caring for the largest portion of national public lands.

The USFS is responsible for managing the national forest system, including the natural resources and recreation areas which lie within its boundaries. The USFS has jurisdiction over 28.7% of the nation's federally owned land and is responsible for all the natural resource work that must be performed on that land (Vincent 2004). The ANF is one of the smaller national forests in the U.S. at 668,887 acres (USDA Forest Service, 2012) but receives some of the highest recreational use. The USFS reported recreational use on the ANF at 3,471,486 people during the 2000 calendar year (USDA Forest Service, 2001). One of the biggest explanations for the recreational use per acre is the proximity of the ANF to the second the largest metropolitan area in the U.S., Los Angeles. The estimated population of Los Angeles County in 2017 was 10,105, 722 based on information provided by the U.S. Census Bureau (2017).

The ANF is facing a modern-day management crisis. As the population grows and the size of forest stays the same, increasing levels of recreation use on the forest are approaching capacity and are creating more natural resource work when the personnel and resources available to do the work decline as budgets continue to shrink. One of the biggest problems facing the USFS today is managing the increasing number of invasive plant species found throughout the forest. This is a particularly challenging task because recreation increases on the forest and unpermitted off-trail forest users act as vectors for invasive plant seed dispersal. Invasive species compete with native species and, in the process, alter the local ecosystems, which affects natural wildlife habitat, wildlife food sources, recreational opportunities and buffer zones that are critical to the wildland-urban interface (Stein et al., 2013).

The Natural Resources Department is an integral unit to the overall structure of the forest and is responsible for surveying, documenting, reporting and treating invasive plant populations on the ANF, including *Spartium junceum.* The Forest Service Supervisor's office in Arcadia, California is the headquarters for the entire ANF, and the Natural Resources Department personnel.

The Natural Resources Department on the ANF uses GIS to analyze the data they record in the field but could greatly benefit from SDM in their many projects and programs. The lead biologist, botanist, and their subordinates use a combination of Trimble GPS units and mobile smartphones and tablets to collect and document data on the invasive species they study in the field. After collecting data in the field, each field biologist and botanist use Esri's ArcPad software for Trimble or Esri's collector mobile GIS app to digitize their data. The software they use is suitable for routine GIS field data collection and standard analysis but can greatly benefit from more robust spatial analysis tools such as SDM.

This study examines the spread of the invasive plant species *Spartium junceum* in the ANF. The high levels of vehicle and foot traffic throughout the study area are the result of its location in Los Angeles County and the presence of what is known as the wildland-urban interface, which is the zone of transition “where houses and wildland vegetation meet or intermingle” (Radeloff et al., 2018).

This project shows how the application of species distribution modeling can serve as a powerful spatial analysis tool to support land management across areas that range from small to large extents. The use of Maxent clarifies some of the variables that influence the spread of invasive species to provide a better understanding of the mechanisms that are most dominant in their distribution patterns.

Most of the data being used for this research has come internally from USFS but also from other governmental agencies including: the U.S. Geological Survey (USGS), and the National Oceanic and Atmospheric Administration (NOAA). All the government agencies that are providing the data collect that data on an annual basis and have large databases dedicated to storing and publishing that data once it is in a suitable format for the public. The data provided for this work at the time it was gathered (2017) was in a transactional phase and had yet to be made publicly available. Its use for this thesis project was approved at the time of writing by ANF resource officer Jaime Uyehara and collected and provided by Jason Martin.

The goal of this research is to use the Maxent, Esri’s ArcGIS software, and different sets of unique environmental and non-environmental variables to better understand the spatial distribution of *Spartium junceum* in the study area. In using more robust and powerful spatial analysis, a predictive approach can be taken by the Forest Service to better remove, monitor, and manage the invasive plant species problems affecting the areas they manage.

1.2 Study Area

The ANF was chosen as the primary study area due to its relatively small extent in comparison to other national forest and the high levels of recreation use found throughout the forest (Figure 1).

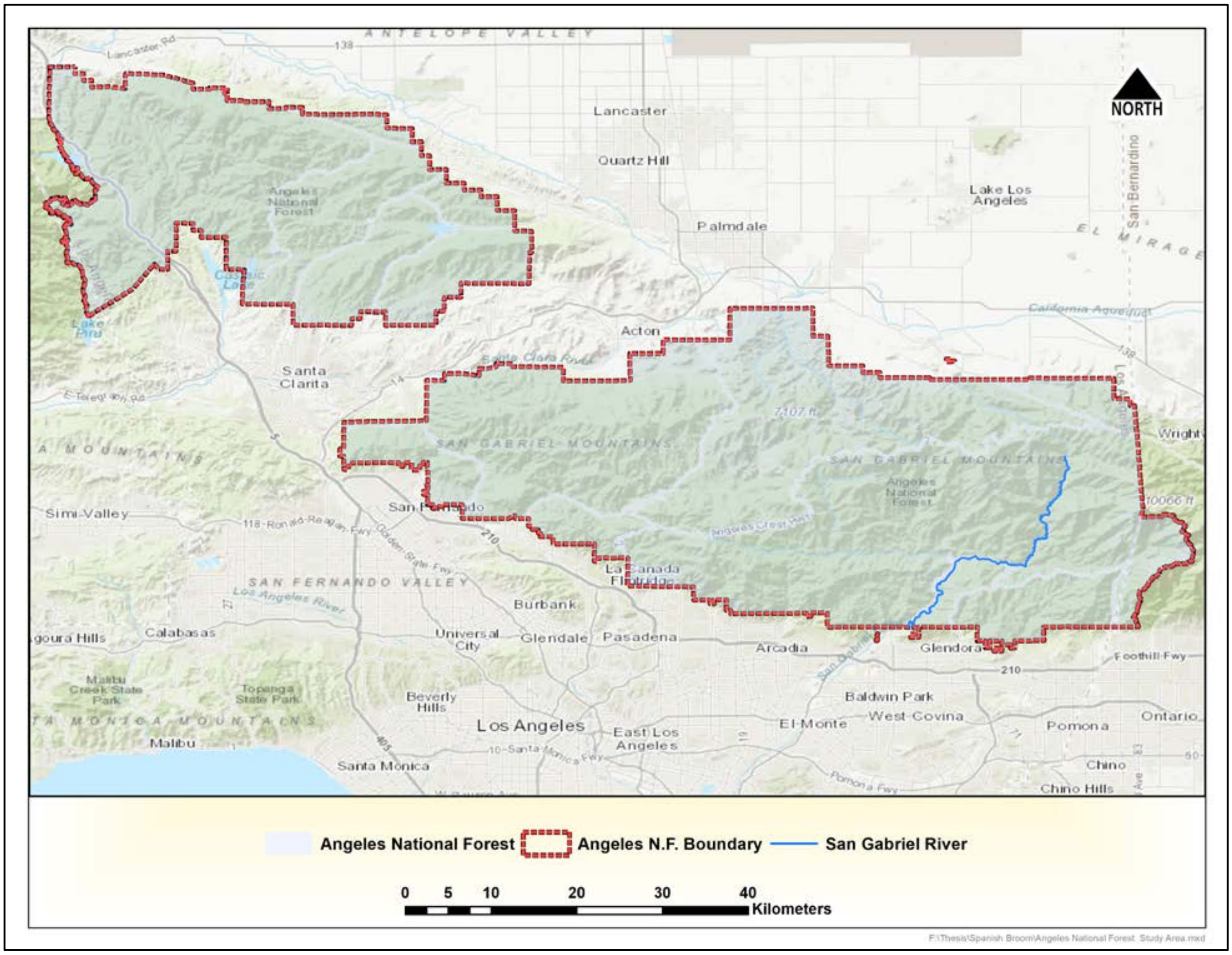


Figure 1 Angeles National Forest

This study considers many different types of environmental variables at play in the forest as well as the many recreational opportunities available to the public. The goal is to portray how those opportunities are increasing the chances and likelihood that seeds are transported and dispersed into far ranging areas where invasive species can establish themselves and compete with native plants. The ANF provides access to numerous hiking trails, off-highway vehicle (OHV) trails and access areas, scenic roadways, campgrounds, day use picnic areas and stream and river access to the many visitors who frequent the region during the hot summer months.

As a very highly frequented national forest, the ANF serves as prime location for Maxent as the area provides a plethora of variables needed for the modeling itself and an abundance of presence-based invasive plant species data which is the primary type of data needed to run this species distribution model.

1.3 Surveyed Invasive Species

An invasive plant species is defined as a non-native (or alien) species to any ecosystem that is under consideration. The ANF plays host to many different varieties of invasive species some of which fall into its noxious weeds list as well. The noxious weeds list indicates that a species is not only invasive but also harmful to the environment or to wildlife in an environment and necessitates more extreme monitoring and removal efforts than normal invasive plant species. This study focuses on one of those major invasive species. *Spartium junceum* (Spanish Broom) is the target species of this research and is one of highest priority invasive species in the ANF invasive species and noxious weeds list. *Spartium junceum* is found in many areas throughout the forest and notoriously difficult to remove. The difficulty in removing individual plants, the resilience to approved chemical applicators and the longevity of seed life has led to numerous annual exhaustive efforts to control population blooms. It has been estimated that the seeds of *Spartium junceum* remain viable for 5 years and mature plants can produce anywhere from 7,000-10,000 seeds per season (Zouhar, 2005).

This research focuses on the land contained within the administrative boundary of the ANF with an emphasis on the eastern side of the forest where the San Gabriel Mountains are located. The study area supports recreation in the form picnicking, day use camping, hiking and OHV (Off highway Vehicle) use, which makes it an ideal study area for this research.

1.4 Species Distribution Modeling and Maxent Explained

Species distribution modeling, also known as environmental niche modeling, is a modeling technique that uses two different data types to estimate the potential locations a species will likely inhabit in the future. The two different data types are integral in determining the type of SDM that is most appropriate for individual cases studies; the two data types are presence-only data and presence-absence data (Philips et al., 2004).

The required data types for a species distribution model are often determined on a case-by-case basis but one of the main determining factors is the availability of data. In many instances, insufficient records are available to support presence-absence species distribution modeling and presence-only SDMs are more suitable (Philips et al., 2004). Furthermore, a key issue with presence-absence is that a species may be declared absent from a landscape unit simply because the species was not detected using the prescribed sampling methods. The effect of this imperfect detection is that parameter estimates will be biased, and any modeling of the data provides a description of the surveyor's ability to find the species on the landscape, not where the species is on the landscape (Mackenzie, 2005).

Maximum Entropy Modeling is a "species distribution model that uses presence-only data to predict the distribution of a species based on the aforementioned theory of maximum entropy" (Qin et al., 2017). One of the constraints of species distribution modeling is the availability of data for a given study variable within a given study area. Maxent relies on large datasets to be as accurate as possible (Philips et al., 2004). The more data that is available the better the model runs as it allows for higher quality or more appropriate data to be selected and used with the model, which can help produce more accurate predictions of a probability distribution for species occurrence in each study area. Being able to obtain the necessary data is a critical component of being able to use this type of model.

The availability of historical datasets is a limiting factor in presence-absence modeling, as many of the necessary records are held in archives managed by museums or other archiving entities and are not always easily accessible or readily available. Maxent is not as burdened by this challenge as it relies on presence-only datasets (Philips et al., 2004). In this case the presence-only data needed is readily available to the public in a spatial data clearinghouse and internally via request from the USFS.

Maxent aims to predict a species' distribution based on environmental predictors (precipitation, temperature, etc.), background data and the number of individuals or samples in a data set based on the underlying theory of maximum entropy (Philips et al., 2004). The theory of maximum entropy relies heavily on the principle of being able to find what the best approximation of a probability distribution is relative to the constraints on a given study variable in a specific study area (Qin et al., 2017). Entropy itself can be looked at as a measure of how much choice is involved in the selection of an event in each system. Combining the concepts of entropy and species distribution modeling, Maxent essentially "agrees with everything that is known but carefully avoids assuming anything that is not known when trying to find the optimal probability distribution for a given study or study area (Qin et al., 2017). When using Maxent, the study area of choice is divided into a grid and from that grid, Maxent extracts a sample of background locations that it contrasts against the presence locations where a species (*Spartium junceum)* is found and uses this with the environmental predictors to predict future areas of presence (Merow 2013).

As is the case with presence-absence data, presence-only data is subject to sampling bias (Qin et al., 2017). This occurs primarily with sample selection bias where areas that are oversampled are given more consideration or weight than areas that are not. To account for this sampling bias, careful consideration will be given to the choice of the overall study area and the presence of multiple data points that fall near each to avoid spatial autocorrelation. Furthermore, presence-only data from previous years will be used to validate occurrence of the target species. The information resulting from this research will be useful in helping the USFS carry out their land management practices and serve as a basis for future studies related to species distribution as it relates to land management.

1.4.1. Multiple Maxent Models

Multiple Maxent models are run in this research in order to get a thorough understanding of the factors that play the biggest role in the study area. In taking this step a more comprehensive approach is used in understanding which variables play the biggest roles on their own and which variables when combined paint the most accurate picture in predicting habitat suitability. This is carried out by running Maxent multiple times with each unique set of environmental variables (layers) to see which instance produces the highest statistical significance and which individual variables within each unique set of layers are the most important predictors of that significance. Lastly, Maxent is run a final time with a combination of all the unique layers once they have been converted to the same spatial resolution. This final run provides a complete and thorough use of all of the variables in all of the layers and produces results that are compared to all other Maxent results which is discussed in chapter 4.

1.5 Thesis Structure

The next chapter discusses species distribution modeling, the different types of models available for use, and various case studies where species distribution modeling has been used. Chapter 3 details the procedures used to collect and process the data and the variables in Maxent. Chapter 4 details the results from this study and provides insight into how the study can be used by land management agencies such as the USFS. Chapter 5 discusses how this research can be improved and other potential land management program areas where it can be employed.

Chapter 2 Related Work

National forests across the U.S. have been dealing with a major land management crisis as a result of severe budget constraints, reoccurring furloughs and government shutdowns. The impacts of these factors have had and continue to have major implications on the availability of skilled personnel to tackle the diverse array of problems facing national forests today. The "2019 Forest Service budget for discretionary appropriations" was $4.77 billion, a decrease of $486 million from the FY 2018 budget in the annualized Continuing Resolution. It included $1.72 billion for the management of national forest system lands and $2.5 billion for wildland fire management (U.S. Forest Service, 2018). With most of the upcoming FY 2019 budget being allocated to wildfire management, other program areas must become more selective and efficient in prioritizing their approach to land management.

The aim of this research is to use Maxent, to predict future spreading locations where noxious and invasive plant species, *Spartium junceum,* may occur and use those results as a cost-effective tool to help mitigate some of the land management challenges facing the USFS and other agencies today. The purpose of this chapter is to examine and critically evaluate information on *Spartium junceum* characteristics and habitat and to highlight several studies where Environmental Niche Modeling (ENM), also referred to as SDM, was used for predictive analysis of species like the test species in this research. The chapter first examines research and reports relevant to the characteristics, habitat tendencies, and management practices for controlling *Spartium junceum* infestations. The chapter then examines some of the different types of SDM/ENM currently being used based on the data types required. Lastly, the chapter discusses Maxent as the preferred method for SDM/ENM amongst the ecological community.

2.1 *Spartium junceum* and Invasive Species

Spartium junceum, also commonly known as Spanish broom, is an invasive plant with a high-level pest rating based on the California Invasive Plant Council (CAL-IPC) pest rating system. This rating means that “these species have severe ecological impacts on ecosystems, plant and animal communities, and vegetation structure. Their reproductive biology and other attributes are conducive to moderate to high rates of dispersal and establishment. These species are usually widely distributed ecologically, both among and within ecosystems” (Warner et al., 2003, 4). In addition, *Spartium junceum* is a noxious weed and is included in the California Department of Food Agriculture’s noxious weed list. A noxious weed is a plant that has been defined as such by law or regulation (CDFA, 2009). This is consistent with how the USFS defines an invasive species and factors in the noxious weed designation.

According to the USFS, “invasive species are among the most significant environmental and economic threats facing our Nation’s forest, grassland, and aquatic ecosystems. They endanger native species and threaten ecosystem services and resources, including clean water, recreational opportunities, sustained production of wood products, wildlife and grazing habitat, and human health and safety” (USDA Forest Service, 2013). A species is invasive if it is non-native to an ecosystem that is being studied (USDA Forest Service, 2013). This criterion applies to this research and understanding the habitat characteristics is a critical component of understanding the effects this invasive species has to ecosystems.

2.1.1. Spartium junceum History

Spartium junceum was initially introduced “into the California ornamental trade in 1848 in San Francisco. Beginning in the late 1930s, it was planted along mountain highways in southern California” as a method to prevent soil erosion. Evidence of *Spartium junceum* colonies can be seen along those same access routes and highways that run through the national forest system today (Zouhar, 2005).

2.1.2. Spartium junceum Characteristics

Spartium junceum is a shrub that is 3 m tall with green stems and fragrant yellow pea-like flowers that bloom from April to June. *Spartium junceum* have very deep branched taproots that are difficult to remove and are associated with nitrogen fixing bacteria. The plant begins to produce seeds when it reaches 2-3 years of age and is indicated by its pods turning brown. Each pod can contain 10-18 seeds. The seeds can germinate readily without treatment and will produce 7,000 to 10,000 seeds per plant per season (Zouhar, 2005). Furthermore, the pods and the plants themselves can tolerate frost providing exceptional longevity and viability (CAL-IPC, 2004).

2.1.3. Spartium junceum Habitat Types and Plant Communities

Spartium junceum is a native of the southern Mediterranean region of Europe and thus thrives in a Mediterranean climate like that found in southern California. A Mediterranean climate is characterized by rainy winters and dry summers (University of California, 2019).

Spartium junceum is most commonly found in disturbed areas such as along roadsides where it was seeded in the early 1900s, but it has also spread to riparian and upland areas throughout much of the southern California national forest system (Nickerman, 2009). There are also other documented cases of Spanish Broom populations persisting in other disturbed areas such as on eroding slope, and riverbanks. *Spartium junceum* has also managed to start invading adjoining stands of chaparral that are near roadsides.

2.1.4. Impacts of *Spartium junceum* on ecosystems

Spartium junceum can rapidly colonize disturbed areas and develop thick shrub like communities that can crowd out and prevent colonization by native chaparral plant species. It also poses a particularly challenging problem during the fire season. *Spartium junceum* not only serves as additional fuel during the fire season but it can survive if its roots are not burned.

Spartium junceum is also a nitrogen-fixing plant which means it can enrich the nitrogen levels of the surrounding soils of the communities it inhabits. The ability of broom to fix nitrogen affects the way in which the local nitrogen cycles behaves. This is unlikely to benefit native plants who do not have the same nitrogen demands and may reduce species diversity in ecosystems as a result (Zouhar, 2005).

2.1.5. Spartium junceum Management and Control Methods

Established infestations are difficult to eliminate due to the large and long-lived seed banks that accumulate annually. Furthermore, the success of any control methods will vary based on the site characteristics (topography, soil, and climate) where the species are found, the condition and age of the stands or communities, and the human and technical resources available. The main methods for trying to control Spanish Broom infestations are through prevention, integrated management, chemical management, and physical/mechanical management.

2.1.5.1. Control Methods

There are the two primary control methods currently employed by the USFS for removal of invasive plants species, integrated management, and physical/mechanical management.

Integrated management refers to an effective method of combining the use of a chainsaw for larger mature plant shrubs or shears for smaller shrubs and the application of a chemical herbicide such as a glyphosate applied as a 2-3% v/v foliar spray to the cut parts of the stem.

Physical/mechanical management refers to the human use of tools to physically remove both young and mature plants. While young *Spartium junceum* can be removed by hand rather easily more mature broom must be removed with the use of a weed wrench which can be an exhausting process and is only effective for small broom infestations where the previous control method is impractical or too expensive (Zouhar, 2005). Control methods represent a crucial aspect of invasive species management alongside prevention.

2.1.5.2. Prevention

According to USFS documentation the most effective way of managing an invasive species is by preventing the establishment and spread of that species. Some of the ways this is accomplished is by limiting seed dispersal, minimizing soil disturbance, introducing more native plants into areas where infestations have been known to occur and, removing current plants to limit the spread of infestation.

While there are numerous control methods and prevention measures currently being used for removal and mitigation efforts there is no predictive analysis currently being used for *Spartium junceum* and this represents an opportunity for the incorporation or utilization of SDM as a cost-effective way to anticipate future locations of infestation and prioritize management activities.

2.2 Species Distribution Modeling

SDM is a machine learning technique that has become a popular analytical framework for predicting species distributions based on species specific geo-located occurrence information and the environmental variables that contribute to that species' survival rate (Vaclavik and Meentemeyer, 2009, 3248). There are two primary types of SDMs that are differentiated based on the types of data they require and thus cater to certain types of research and analysis. The two primary types of data required to perform any type of predictive modeling are presence-absence data and presence-only data. The merits of each data type and the availability of each data type are the primary determining factors for model selection in each community or industry field. The following sub-section details species distribution modeling that rely on presence-absence data and explores some of the different types that are commonly used.

2.2.1. Presence-Absence Data Modeling

Presence-absence species distribution modeling relies on presence-absence data, which means it is heavily dependent on the availability of current presence and absence data for the case subject and historical data for that same subject. Historical data can be from prior studies or from prior field collection provided the data is in a similar or usable format that can be converted into the format that is the same as that of the current presence data being used and the same format used by a particular model of choice.

There is a third category known as pseudo-absence data modeling which relies on modelers generating pseudo-absence data by sampling environmental conditions at locations where the species is not recorded and hence absent (Vaclavik and Meentemeyer, 2009, 3249). There is a risk of introducing false or negatives errors when engaging in this process because no actual records of absence data exist but rather, they are being created. In addition, the results from using pseudo-absence data are not as accurate as reflected in the area under the curve receiver operating characteristic (AUC-ROC) status that will be explained in the following paragraph. As a result of the type of data available and the prevalence of true presence-absence data model types used by the ecological community and suitable for comparison with presence-only data models, pseudo-absence data modeling techniques were not considered. For this research three types of presence-absence models were examined.

Three presence-absence species distribution models commonly used in the ecological community are multiple logistic regression (MLR), generalized linear models (GLM), and classification and regression trees /classification trees (CART/CT). The primary means of evaluating how effective each presence-absence model is in predicting potential future distributions of a case species is by comparing the results of each model. This can be accomplished by using similar test species amongst the three models and evaluating the results of each by using AUC-ROC, which offers a global performance measure for classification problems based on various threshold settings (Pepe et al., 2007, 928). The closer to 1.0 the AUC-ROC value is the more accurate the result is since a value of 1.0 indicates perfect discrimination whereas a value of 0.50 indicates that the model performed no better than random.

The three models are explored in more detail by referencing the work of Vaclavik and Meentemeyer (2009) on invasive species distribution modeling and Bedia et al. (2011) on predicting plant species distribution across an alpine rangeland in northern Spain as a means of further explaining the merits of this type of modeling and for providing the background for choosing presence-only data modeling for the actual research performed in this thesis. GLM and CT models were evaluated using an extensive dataset on the distribution of the invasive forest pathogen *Phytophhthora ramorum* in California and the AUC-ROC (Vaclavik and Meentemeyer, 2009, 3248). MLR and a CART model were evaluated using presence-absence data from 15 different plant species found on the Riofrio rangeland in northern Spain.

2.2.1.1. Multiple Logistic Regression

MLR is a parametric technique that relies on the assumption that the response of a species to environmental variables is consistent across space and time. This technique attempts to model a relationship between two or more independent variables and a dependent variable by fitting a linear equation to the data. MLR is considered one of the simplest and easiest SDMs to work with and is frequently used by ecologists who have access to presence-absence data.

Bedia et al. (2011) describe how MLR performed compared to the other SDMs for modeling plant distributions across northern Spain. MLR achieved mixed results when compared with the other models in the study. MLR provided high AUC-ROC values for half of the 15 samples in the study and low values for the other half with the highest AUC-ROC value being 0.92 and the lowest being 0.64. The results based on presence-absence data for non-invasive species data were somewhat inconsistent across the board for all the models. However, the MLR models produced some of the poorest resolution results amongst all the models in the study and were ultimately ruled out as an option for predicting future distributions of *Spartium junceum* (Bedia et al., 2011).

2.2.1.2. Classification Trees/Classification and Regression Trees

CTs and CARTS offer “a non-parametric, data driven method that recursively partitions data into homogeneous groups based on identification of a specific threshold for each environmental predictor variable” (Vaclavik and Meentemeyer, 2009, 3252). In this method a hierarchical decision tree of rules is created to split data into “present and absent” categories. According to Bedia et al. (2011), CARTs are better than MLRs for predicting the distribution of oak species in California. For most species in the study, CART produced good predictive resolutions, but also low AUC-ROC values. The predictive resolution is a value that quantifies the "deviation of the prediction from the true species prevalence" (Bedia et al., 2011). The highest AUC-ROC value for a CART was 0.89 and the lowest was 0.67. This is consistent with the results from Vaclavik and Meentemeyer (2009) where CT produced an AUC-ROC of 0.89 with true presence-absence data and 0.65 when using true presence-pseudo absence data created from environmental sampling.

Both cases highlight research performed on native species distributions which does not fit invasive species distribution patterns nor address the fact that invasive species are not necessarily in equilibrium with their environments as is more typical with native plant species. When using true presence-absence data, CART/CT models produced some high AUC-ROC values but when using pseudo-absence data, the AUC-ROC values dropped significantly and as a result were ultimately ruled out as an option for predicting future distributions of *Spartium junceum* (Bedia et al., 2011).

2.2.1.3. Generalized Linear Models

A GLM "is an extension of common multiple regression that allows for the modeling of non-normal response variables" (Vaclavik and Meentemeyer, 2009, 3252). This means that GLM is well suited for SDMs focused on invasive species where there is available presence-absence data with variables that may vary tremendously such as more localized weather or climate factors that might be influenced by subtle topography changes. This was the case with the test species *Phytophhthora ramorum,* which is the invasive pathogen, studied by Vaclavik and Meentemeyer (2009). This sample species is like *Spartium junceum* in that it is invasive but differs in that it is a pathogen and not a plant species and because there was available absence data to use in the model. With their data GLM is able produces binary responses that specify the relationship between a dependent (or response) variable *Y*, and a set of predictor variables, *X*.

When using true presence-absence data GLM produced an AUC-ROC value of 0.90 which was the best out of the models tested including Maxent, which was the model used in this research (Vaclavik and Meentemeyer, 2009, 3254). There is plenty of evidence to support using CART/CT or GLM for invasive species distribution modeling but neither meet the requirements necessary for studying the invasive species in this research given the lack of absence data. There are no test cases of *Spartium junceum* being used as sample data for analysis with GLM or any of the previously mentioned models and that may be due to the lack of available absence data as well.

For a more in-depth explanation of how AUC-ROC is used to evaluate the statistical significance of the results produced by each model and thus the effectiveness of each model, the reader is referred Hanley and McNeil (1982) and Zou et al. (2007).

2.2.2. Presence-Only Data Modeling

Presence-only SDM relies on presence-only data, which means that it is not hampered by a reliance on the availability of absence data but rather only needs presence data on the test species being analyzed during the time a study is being conducted. Absence data is a very selective data type and can be very difficult to acquire. In many cases absence data records are archived by herbaria and are either not available to the public for use or are only obtainable by formal request which may or may not be granted. This may occur because researchers are unwilling to share their findings with the public or other researchers or as a result of that data being too sensitive for governmental or other reasons. Furthermore, absence data for a sample species may not even exist and this is often the case with invasive plant species. The very nature of invasive plant species makes it difficult to acquire their absence data because management and mitigation efforts require eradication and removal rather than studying patterns on a seasonal or annual basis.

One of the biggest differences between presence-absence modeling and presence-only modeling is having to account for sampling bias and an often-encountered lack of information on a species' prevalence.

Three commonly used presence only SDMs used by the ecological community are Ecological Niche Factor Analysis (ENFA), General Algorithm Rule-Set Prediction (GARP), and Maximum Entropy (Maxent) (Pearson et al., 2007). Like the examples of the presence-absence SDMs in Section 2.2.1, these three presence-only models were evaluated on how effective each one is in predicting potential future distributions of a case species and by comparing the results of each model. This was accomplished by using similar test species with the three models and evaluating the results of each by using the area under the AUC-ROC. The performance of the three models has been explored in a more detail by Vaclavik and Meentemeyer (2009). Maxent and ENFA have been evaluated by using the same invasive forest pathogen *Phytophhthora ramorum* data as the presence-absence models and AUC-ROC (Vaclavik and Meentemeyer, 2009, 3248). GARP and Maxent have also been evaluated using presence-only data from two invasive plant species, a generalist (*Bromus tectorum)* and a specialist (*Tamarix chinensi*). In this context, a generalist is one with characteristics or traits that allow them to inhabit a broad range of habitats whereas a specialist is one with traits that require specific habitats or specific ecosystem conditions for them to thrive. Lastly, additional examples of Maxent are referenced in Section *2.2.2.3*.

2.2.2.1. Ecological Niche Factor Analysis

ENFA is a model that compares the distribution of locations where a sample species was identified to a reference set describing the whole area. It does this by computing variables that explain a major part of the ecological distribution of the sample species. The model produces two factors with biological significance: marginality and specialization. Marginality describes how the sample species' optimum conditions differ from the global mean of environmental conditions in a study area. This allows for comparisons between the environmental variables that may be found at the site location and study area and the variables found at the site location but not in the study area. This helps to identify correlated variables and spatial autocorrelation, which measures the correlation between the values of one variable and the values at those locations that are closest to it. Specialization refers to factors that are sorted by the decreasing amount of explained variance, which helps describe how species variance compares to global variance.

When using presence-only *Phytophhthora ramorum* data ENFA produced an AUC-ROC value of 0.80 which was lower than any of the AUC-ROC values produced by presence-absence models which used the same presence data but also incorporated true absence data as well. Furthermore, ENFA had a poorer efficiency (commission/omission error rate) than Maxent, CT and GLM and tended to overestimate species distributions (Vaclavik and Meentemeyer, 2009, 3255). ENFA provided a good test case and positive results for presence-only modeling but was not as accurate as Maxent and does not have as many positive test case reviews as Maxent in the ecological literature. Therefore, it was not used for predicting future distributions of *Spartium junceum* in this thesis research project.

2.2.2.2. General Algorithm Rule Set Prediction

GARP models relate ecological characteristics of known species-specific occurrence locations (point data in GIS) to points that have been randomly selected from the remainder of the study region. This helps the model to develop a series of decision rules that best summarize the factors that are associated with the species' presence in a study region (Adjemian et al., 2006, 94). GARP uses presence-only data in the form of sample species occurrence data along with environmental data to enable the model to determine or predict locations where the species should be found. The output results take the form of multiple predictive distribution maps that can be used for comparison amongst each other or in combination with one another to provide a comprehensive picture of where a study sample species might be expected to be found. Like Maxent, GARP performs very well when dealing with small sample sizes and provides a cost-effective means of rapidly generating spatial data that can be analyzed in GIS.

When using the generalist species *Bromus tectorum*, GARP produced an AUC-ROC value of 0.58 for training data and a value of 0.50 for validation data. When using the specialist species *Tamarix chinensi*, GARP produced an AUC-ROC of 0.72 for training data and a value of 0.830 for validation data (Evangelista et al., 2008, 811). These values are not as high as any of the AUC-ROC values produced by presence-absence models, but these are the result of a different data set and upon closer inspection show how the choice of generalist or specialist species produces major differences in overall model performance. In general, specialist species were easier to predict.

The type of data used by Evangelista et al. (2008) is more reflective of the *Spartium junceum* data being used in this thesis research project as both species are invasive plant species and *Spartium junceum* would be considered a specialist species.

Turning next to the results from Maxent, the results were lower for *Bromus tectorum* and the same for *Tamarix chinensi*. The overall results for GARP were comparable to Maxent but not as strong and therefore was not used for predicting future distributions of *Spartium junceum* in this thesis research project.

2.2.2.3. Maximum Entropy Modeling

Maximum entropy modeling techniques use presence-only data to predict the distribution of a species based on the theory of maximum entropy (Qin et al., 2017, 140). The principle of maximum entropy states that the most appropriate "probability distribution to model a given set of data is the one with the highest degree of entropy (i.e., that which is most spread out, or closest to uniform)" among all those that satisfy the constraints of our prior knowledge (Philips et al., 2006). In information theory, entropy refers to randomness or unpredictability. The idea behind this is that the information that does not fall within the probability distribution created by the model has maximum entropy (randomness or unpredictability) and thus those results are discarded and those that fall within the probability distribution remain. The remaining information represents the best possible description of the distribution of data that was fed into the model prior to it being run (Kemp, 2012, 4). Maxent creates this probability distribution based on environmental variables spread across the entire study area (Pearson et al., 2007, 106). The modeling process in Maxent uses an iterative heuristic approach where the first instance of the model being run represents the first solution that is tested. Once the first solution is tested the results from that first test are permuted and run over again and again, sometimes through 500 iterations, until the model is trained with data that improves with each iteration. The data that does not improve the model with each successive iteration is discarded until ultimately the model moves to a stable solution that was better than the first one obtained.

Maxent offers many advantages but also a few drawbacks. In addition to utilizing easier to obtain presence-only data, Maxent can run without needing large sample datasets, it can also utilize both continuous and categorical data, and can incorporate interactions between different variables. It also has efficient deterministic algorithms that have been developed, and a probability distribution that has a concise mathematical definition. Over-fitting can be avoided by using l_1-regularization and because dependence of the Maxent probability distribution on the distribution of occurrence localities is explicit, there is the potential (in future work) to address the issue of sampling bias formally" (Philips et al., 2006, 234). Maxent requires special software in order to be run, it is relatively new statistical model in comparison to other SDMs such as GLM and GARP and thus has less guidelines and fewer methods for estimating the amount of error in a prediction. Furthermore, the amount of regularization requires more study as does the model's effectiveness in avoiding over-fitting compared to other models. When considering the use of Maxent for this thesis research project it was also important to consider another major factor in this type of modeling and that is sampling bias.

One of the biggest differences between presence-absence and presence-only modeling is having to account for sampling bias. Occurrence data are frequently biased, for example, data that is closer to access routes will be better sampled and better documented and thus better represented in the results of the model. When the bias is large, presence-only models mimic the biased sampling distribution as much as they do the species distribution. This can be corrected by having the background data reflect the same bias and thus essentially cancel it out. This is accomplished by using as a background the set of occurrence data for an entire group of species that may have been captured or observed using the same methods. In this research, this was accomplished by randomly selecting background points, but this can also lead to situations where samples are counted more than once. This problem can be mitigated by keeping track of which sites have been sampled in order to not reselect sites that have already been sampled. Lastly, another way to account for sampling bias is to keep track of which environmental variables have been better sampled and where they have been sampled in order to weight the value of those samples (Philips et al., 2008, 162). There is a large body of research performed using Maxent that shows how accurate it is in predicating species distribution patterns with the results being evaluated with the same AUC-ROC values as the models in the previous sections.

When using a generalist species such as *Bromus tectorum*, Maxent produced an AUC-ROC value of 0.55 for training data and 0.50 for validation data. When using a specialist species such as *Tamarix chinensi*, Maxent produced an AUC-ROC of 0.77 for training data and 0.830 for validation data (Evangelista et al., 2008, 811). These AUC-ROC values are very similar to those produced by GARP on the same dataset and indicate a strong predictive outcome. Maxent produced strong results predicting distributions of the endangered conifer *Thuja sutchuenensis* in southwestern China with an AUC-ROC value of 0.99. This study used sample data from a 3-year long field survey in conjunction with global BioClim data (Qin et al., 2017, 142).

When using presence-only *Phytophhthora ramorum* data, Maxent produced an AUC-ROC value of 0.85 which was lower than any of the AUC-ROC values produced by all but one of presence-absence models that used the same data (ENFA). Lastly, when Maxent was used on the following invasive species Common reed (*Phragmites australis*), musk thistle (*Carduus* *nutans*), salt cedar (*Tamarix*), and Russian olive (*Elaeagnus angustifolia*) it produced positive results. For *Phragmites australis* it produced an AUC-ROC of 0.98 for training data and 0.97 for validation data. For *Carduus nutans* it produced an AUC-ROC of 0.978 for training data and 0.965 for validation data. For *Tamarix* it produced an AUC-ROC of 0.988 for training data and 0.983 for validation data. For *Elaeagnus angustifolia* it produced an AUC-ROC of 0.976 for training data and 0.945 for validation data (Hoffman et al., 2008, 362-364).

The results from the last study which included multiple invasive plant species datasets show how Maxent produced strong AUC-ROC results. These results also show that Maxent performed well when used with both native and invasive species data. The combination of the results and the similarities of the different data types to the data being used in this research effort serve as an additional rationale for using Maxent in this thesis research project.

Maxent was ultimately chosen for this thesis research project because of the many advantages of the software, the ability to successfully address the challenges of sampling bias, the availability of presence-only data for *Spartium junceum*, and the positive AUC-ROC results generated from multiple related research studies on native and non-native invasive species in the last two decades.

Chapter 3 Methodology

The main purpose of this research is to evaluate how effective Maxent is as a tool for predicting future locations and distribution patterns of the invasive plant species *Spartium junceum*. The primary source of sample data came from the USFS, with a secondary source coming from the Cal Flora/Jepson herbarium, which was used as a quality check. The third source of data came from the WorldClim global dataset. The last source of data came from a USGS digital elevation model that was then used to derive other layers in ArcMap. Section 3.1 outlines how this research was approached and designed. Section 3.2 details the different datasets used in this research. Sections 3.2.1 and 3.2.2 discuss the different data sets in more detail and how they were obtained. Lastly, Section 3.3 details how the data was processed and formatted for use with Maxent and ArcGIS.

3.1 Research Approach and Design

The basis for this research revolves around the understanding that environmental conditions at locations where a plant species is found and considered to be thriving is consistent with the conditions that are expected to be found in other areas in the study extent where the same species might be found at some point in the future. This fundamental notion was critical to how this research was approached and designed. Multiple instances of the Maxent model are run with the same biological sample data but with different environmental layers to test this fundamental notion. Running the model with different environmental layers and comparing the results allows for a more comprehensive dive into the many factors at play including climate, topography and the scale at which the underlying environmental data is recorded and then used in the model.

This thesis research project relies on the use of ArcGIS for Desktop version 10.6 and Maxent 3.4.1 (Phillips et al., 2006). The Maxent software is available for free and can be downloaded while ArcGIS for Desktop version 10.6 needs to be purchased. Both were required for this thesis research project.

Maxent uses data in the form of raster and CSV files that have been converted from Excel spreadsheets that contain latitude and longitude coordinates for the occurrence locations of the sample data being used. Vector data can be used in Maxent, but this can be accomplished indirectly. Vector data but must be converted into rasters in ArcMap before being used in Maxent. This approach was used for the ANF features vector dataset.

Maxent uses categorical (biological samples) and continuous data (environmental and proximity variables) as inputs in order to generate a summary html file which contains several charts, images, and tables with information about AUC-ROC values, the variables that contribute the most to the predicted outcome, and a series of jackknife tests.

The AUC-ROC results are particularly important because they indicate if the model produces results that are random or are better than random, which reveals how much confidence can be had in the results. The level of confidence is indicated by the AUC value, which ranges from 0.0 to 1.0. A median AUC value of 0.5 indicates that the results from the model are no better than random while a value of 1.0 indicates the highest level of confidence. A higher AUC value in Maxent translates to higher confidence in Maxent’s ability to accurately predict the species distribution patterns. In addition, there are two different sets of results produced by Maxent in the output html file which explain the contributions each individual environmental variable has on the outcome of the model: the analysis of variable contribution and three jackknife tests.

The Jackknife of regularized training gain shows the training gain of each variable if it was run in isolation and compares that to the training gain of all of other variables (Phillips et al., 2006). This helps show which variables contribute the most individually while considering how they contribute amongst all of the other variables in the dataset. The jackknife of test gain is interpreted the same way as the jackknife of regularized training game but uses the test gain data from Maxent as opposed to the training gain data. Lastly, the jackknife of AUC uses the same jackknife test but with AUC values on the test data. The three jackknife tests combine to create the values behind the analysis of variable contribution. The analysis of variable contribution gives estimates of the relative contributions of each environmental variable used in Maxent and shows their percent contribution to the overall results in the form of a table.

Maxent's default settings produce an ASCII output file that shows the study area and displays the predicted areas of habitat suitability using values that range from 0 to 1. Warmer colors (reds, oranges, and yellows) show areas with better predicted conditions for habitat suitability while cooler colors indicate the opposite (blues and greens). The dots/points that are found scattered throughout the ASCII file represent the presence locations (sample data) that are used when training the model. The ASCII file that is produced by Maxent must be imported into ArcMap and converted from ASCII format back to its original raster format with the ASCII to Raster tool so that further analysis of the results can be done.

3.2 Research Sample Data

There are two primary data types needed by Maxent to perform this research. The first type is the sample biological species data. The second type is environmental data, which when combined with the sample data is used to train Maxent. The environmental data can be further broken up into three sub-groups: bioclimatic data, data derived from a USGS DEM, and ANF feature vector data converted to raster data. Maxent will be run a total of six times using these environmental layers in isolation at their original recorded spatial resolutions, in isolation at a resampled resolution and lastly, in combination with one another. The results of each Maxent run are discussed and compared in Chapter 4.

3.2.1. Sample Biological Data

The *Spartium junceum* sample data was obtained from the USFS. A geodatabase (GDB) containing ANF vector data was downloaded as a zip file and then extracted into a local workspace in ArcCatalog. Necessary data from the GDB in ArcCatalog for this thesis project was filtered out by using a series of queries. The results of those queries were then exported into a new project-specific file geodatabase for use in ArcMap and then Maxent.

The *Spartium junceum* sample data, which was in the project GDB, was in polygon format and had be converted into a point format. The Create Random Points tool was used to create a large set of random points in the study extent with unique x, y coordinates. Those points were then filtered out with the Select by Location tool, which allowed for the selection of all of the points that fell inside the polygon boundaries of the original *Spartium junceum* layer. Figures 2 and 3 show the *Spartium junceum* feature class shapefile before and after the Create Random Points and S elect by Location tools were run.

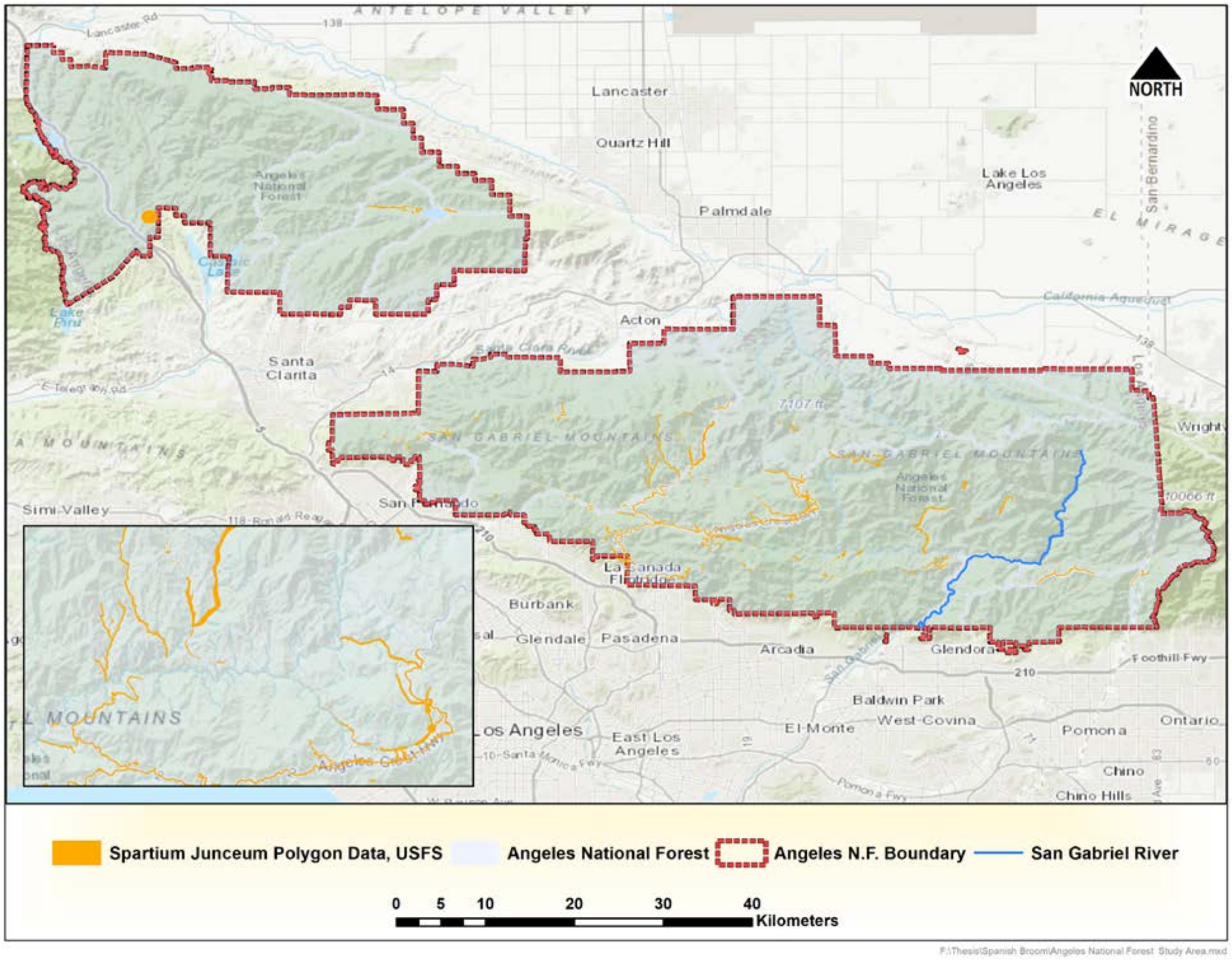


Figure 2 *Spartium junceum* polygon data

Once the *Spartium junceum* polygon feature class shapefile was converted into a point feature class shapefile and x, y coordinates were created, the data was then saved as a database file (.dbf) in the same project GDB. That database file was converted into an Excel file where only the necessary headers for species, longitude and latitude were kept. The Excel file was saved as a comma separated values (.csv) file and was then ready for use in Maxent.

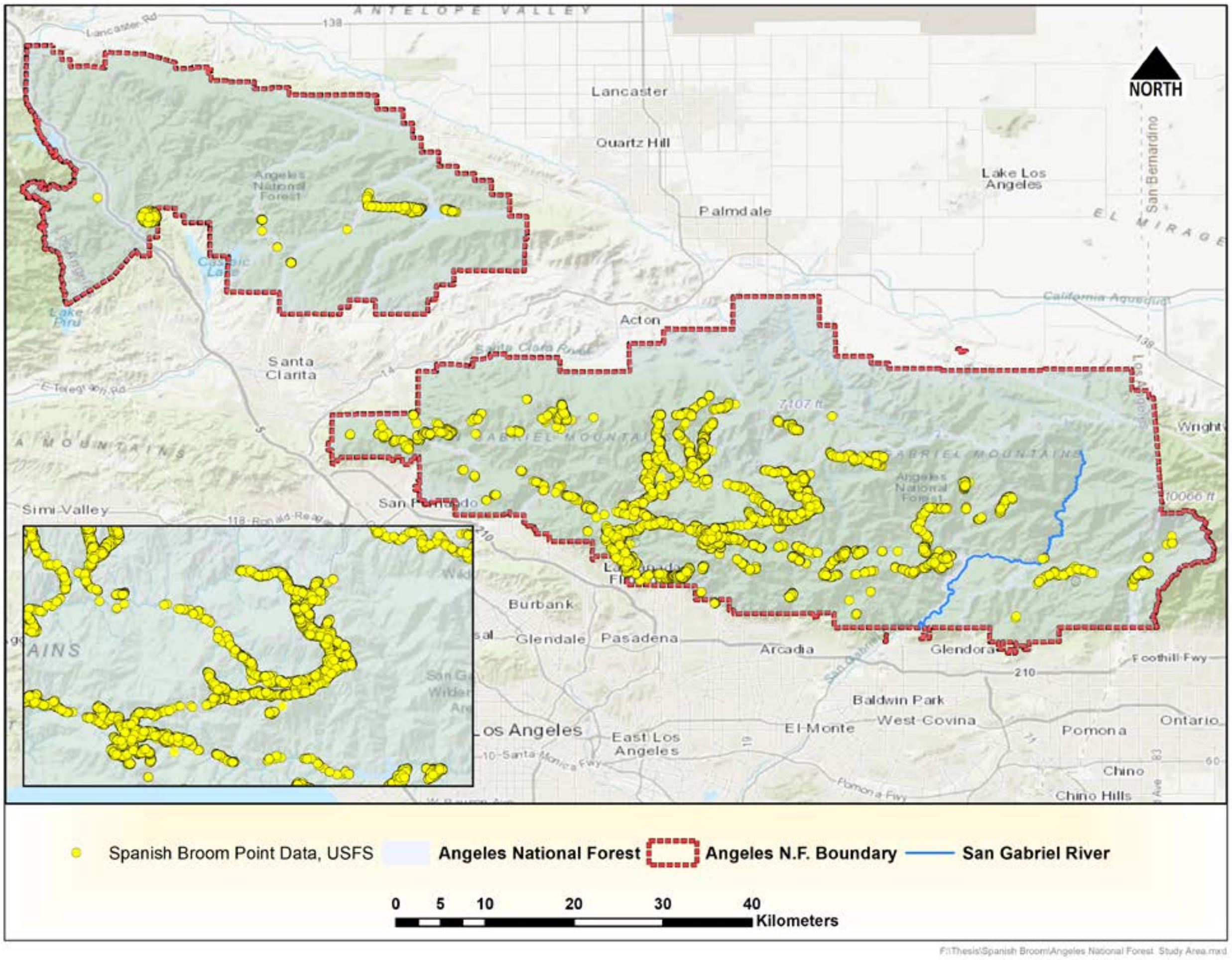


Figure 3 *Spartium junceum* point data.

3.2.2. Environmental Variables/Layers

The sample data used in Maxent must be combined with raster layers containing information on the environmental conditions that are present at locations where the sample data are found in order to train the model to predict future locations for the species of interest. For this study those environmental raster layers come in the form of three groups. The first group contains environmental layers related to climate. This study uses 19 bioclimatic variables from the WorldClim version 2.0 dataset to create a comprehensive understanding of climate factors The 19 bioclimatic variables used in this study reflect the most current data available (Table 1).

Table 1 WorldClim bioclimatic variables

Code	Variable
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

The second group of environmental variables that is used in this thesis research project are derived from a series of DEMs that were obtained from the USGS. These DEMs have a spatial resolution of 1/3 arc second (90 m) whereas the 19 bioclimatic variables have a resolution of 30 arc seconds (~1 km). The digital elevation model has a substantially higher spatial resolution than the bioclimatic data and was used to derive a secondary set of environmental variables using tools in ArcMap. Elevation, hill shade, aspect, slope and visibility rasters were generated using the Surface tool kit that is stored within the Spatial Analyst Toolbox. Table 2 lists the DEM layers that were created in ArcMap and used in the third Maxent run.

Table 2 List of DEM Rasters

Name of feature class	Type of feature class
DEM ANF Study Area (Elevation)	Raster
DEM Hillshade	Raster
DEM Slope	Raster
DEM Aspect	Raster
DEM Visibility	Raster

The third group of environmental variables used in this thesis research project were created by converting ANF features vector data into raster data by using the Euclidean distance tool, which is a type of proximity tool found in the Spatial Analyst Toolbox. Proximity tools are used to indicate the distance between presence data points (*Spartium junceum* sample data) and select features (ANF feature data) that are believed to influence species distribution patterns. Table 3 lists the ANF features data used to create the Euclidean distance rasters for use in the second run of Maxent.

Table 3 List of feature classes used to create Euclidean distance rasters

Name of feature class	Type of feature class
Angeles National Forest Service Roads	Line
Angeles National Forest Recreation Facilities	Point
Water Bodies	Polygon
California Flow lines (Hydrography)	Line
Angeles National Forest Burn Severity Dataset	Polygon

3.2.3. Data Processing and Standardization

Once all of the presence and raster data were downloaded and/or created it was necessary to standardize the spatial resolutions, projections and boundaries of these datasets. The presence-only data was already converted into csv format and needed no further changes; however, the raster data required substantial conversions. The first step was converting the rasters into the same projection, which was WGS84 for this study. This was accomplished using the Project Raster tool found in the Data Management Toolbox. The newly projected rasters were then clipped to the study boundary using the Clip tool found in the Data Management Toolbox. The clipped rasters are then converted into ASCII files, the required raster format for Maxent, by using the Raster to ASCII tool found in the Conversion Toolbox. There is one extra step in the case of the bioclimatic variables, which precedes the raster to ASCII step and that was to resample the rasters into the same format as the DEM rasters. This was accomplished with the Resample tool found in the Data Management Toolbox. This workflow is the same for all of the raster files and can be run quickly using the Model Builder function in ArcMap. Figure 4 details the conversion process from raster file to ASCII file in Model Builder.

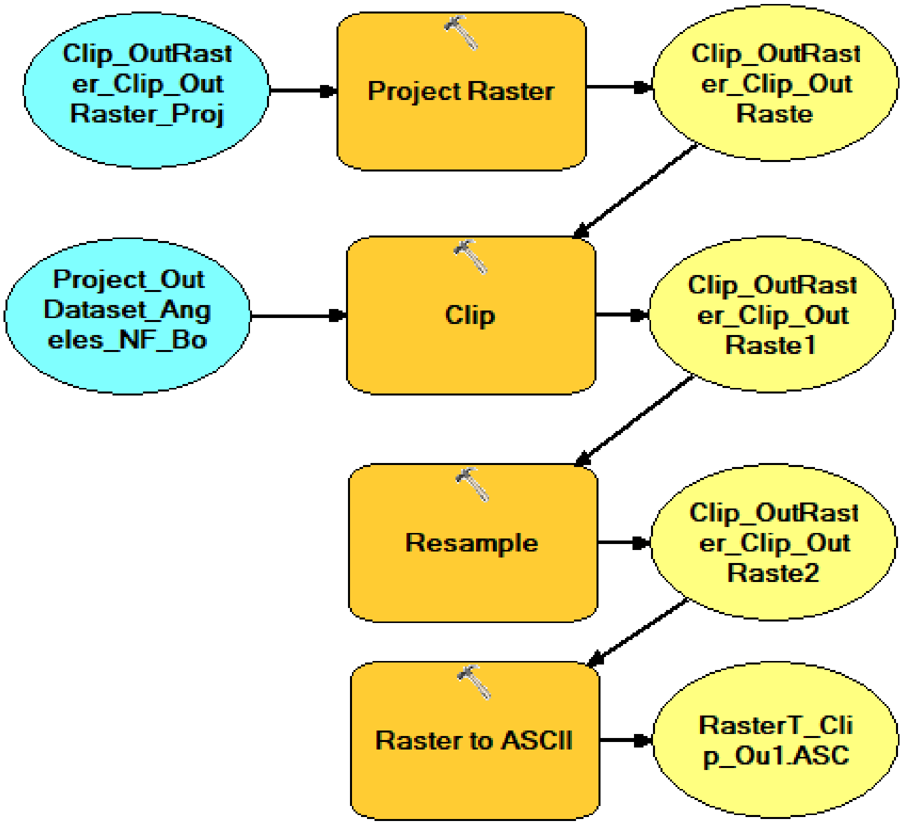
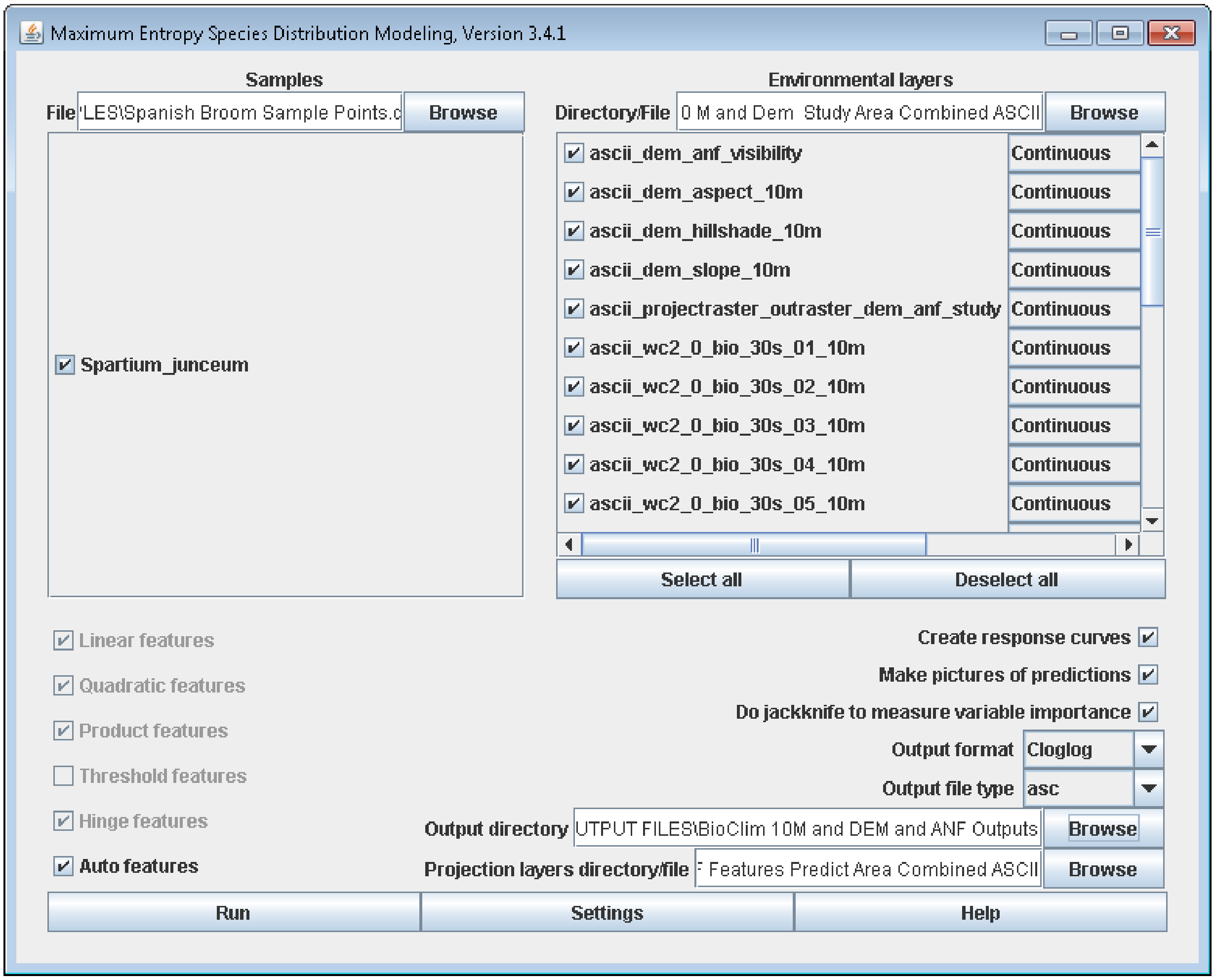

Clip_OutRast er_Clip_Out Raster_Proj
Project Raster
Clip_OutRast er_Clip_Out Raste
Project_Out Dataset_Ang eles_NF_Bo
Clip
Clip_OutRast er_Clip_Out Raste1
Resample
Clip_OutRast er_Clip_Out Raste2
Raster to ASCII
RasterT_Cli p_Ou1.ASC

Figure 4 Raster Conversion tools with Model Builder in ArcMap.

Once all of the data is converted into the necessary format it is placed into the appropriate destination windows in Maxent. The *Spartium junceum* sample data is entered in the samples window, study area versions of the BioClim, Euclidean distance and DEM rasters are entered in the environmental layers window, and study extent versions of the BioClim, Euclidean distance and DEM rasters are entered in the projection layers window. The study area versions of the environmental layers reflect the areas where the actual sample data lie within the ANF boundary and the study extent environmental layers reflect the entire area within the ANF boundary. All of the environmental layers come from the same raster source files. The outputs of the model are sent to a designated folder created by the user, which is set in the Maxent output directory window. The 'create response variables', 'make predictions', and 'do jackknife to measure variable importance' boxes are checked as well. Lastly, the 'logistic' output format was chosen per the consensus recommendations from Young et al. *(*2011) and Philips (2017). Figure 5 shows the first Maxent window with all of the sections filled out.



Maximum Entropy Species Distribution Modeling, Version 3.4.1

Samples

File 'LES\Spanish Broom Sample Points.c

☑ Spartium_junceum

Environmental layers

Directory/File 0 M and Dem Study Area Combined ASCII

☑ ascii_dem_anf_visibility	Continuous
☑ ascii_dem_aspect_10m	Continuous
☑ ascii_dem_hillshade_10m	Continuous
☑ ascii_dem_slope_10m	Continuous
☑ ascii_projectraster_outraster_dem_anf_study	Continuous
☑ ascii_wc2_0_bio_30s_01_10m	Continuous
☑ ascii_wc2_0_bio_30s_02_10m	Continuous
☑ ascii_wc2_0_bio_30s_03_10m	Continuous
☑ ascii_wc2_0_bio_30s_04_10m	Continuous
☑ ascii_wc2_0_bio_30s_05_10m	Continuous

☑ Linear features

☑ Quadratic features

☑ Product features

☐ Threshold features

☑ Hinge features

☑ Auto features

Create response curves ☑

Make pictures of predictions ☑

Do jackknife to measure variable importance ☑

Output format Cloglog

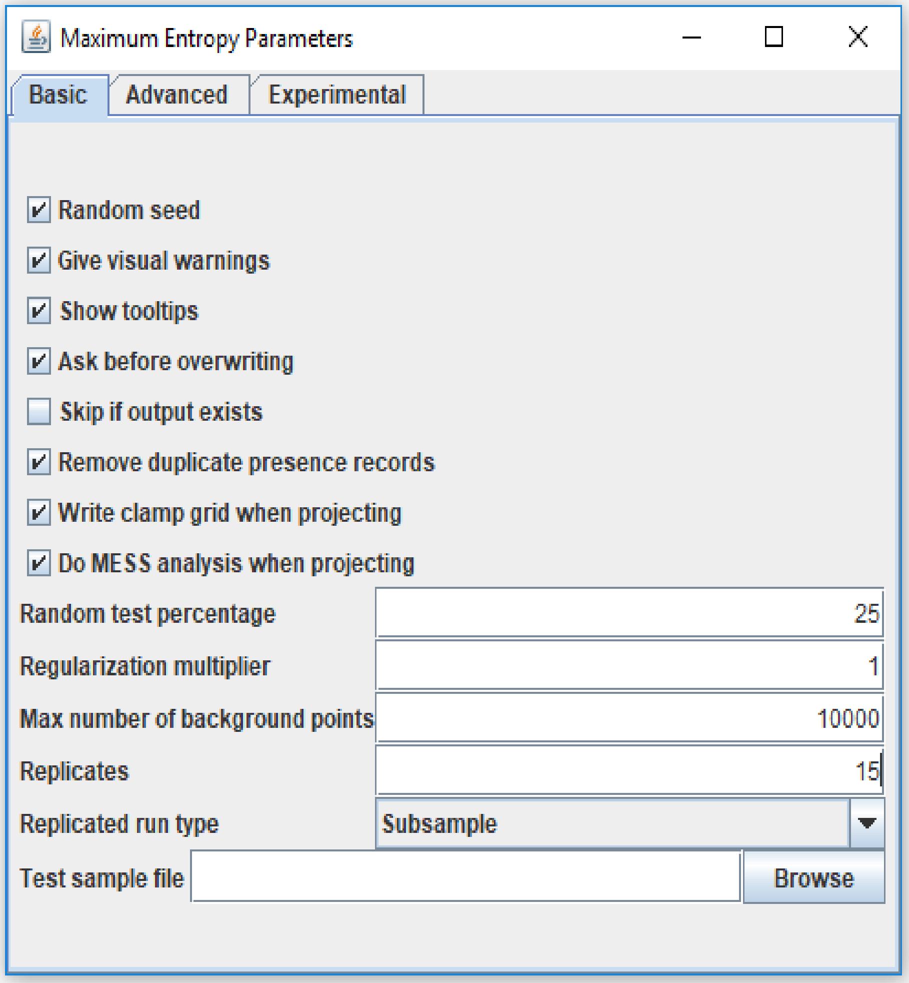
Output file type asc

Output directory UTPUT FILES\BioClim 10M and DEM and ANF Outputs

Projection layers directory/file F Features Predict Area Combined ASCII

Figure 5 Maxent window with appropriate boxes checked

In the settings section of the Maxent window there are further parameters which could be checked and/or selected. In the Basics tab (Figure 6) settings all of the boxes have been left as they are but a random test percentage of 25 has been added to the window and the number 15 has been added to the replicates window. The test percentage indicates that 25% of the sample data will be used for training the model and the other 75% will be used as the actual data. The number 15 in the replicate window signifies that the model will run through 15 replications and the results from the model will come from an average of all of the results. The replication run type selected for this thesis research project is sub-sample. These parameters in Table 6 are used each time the model is run.

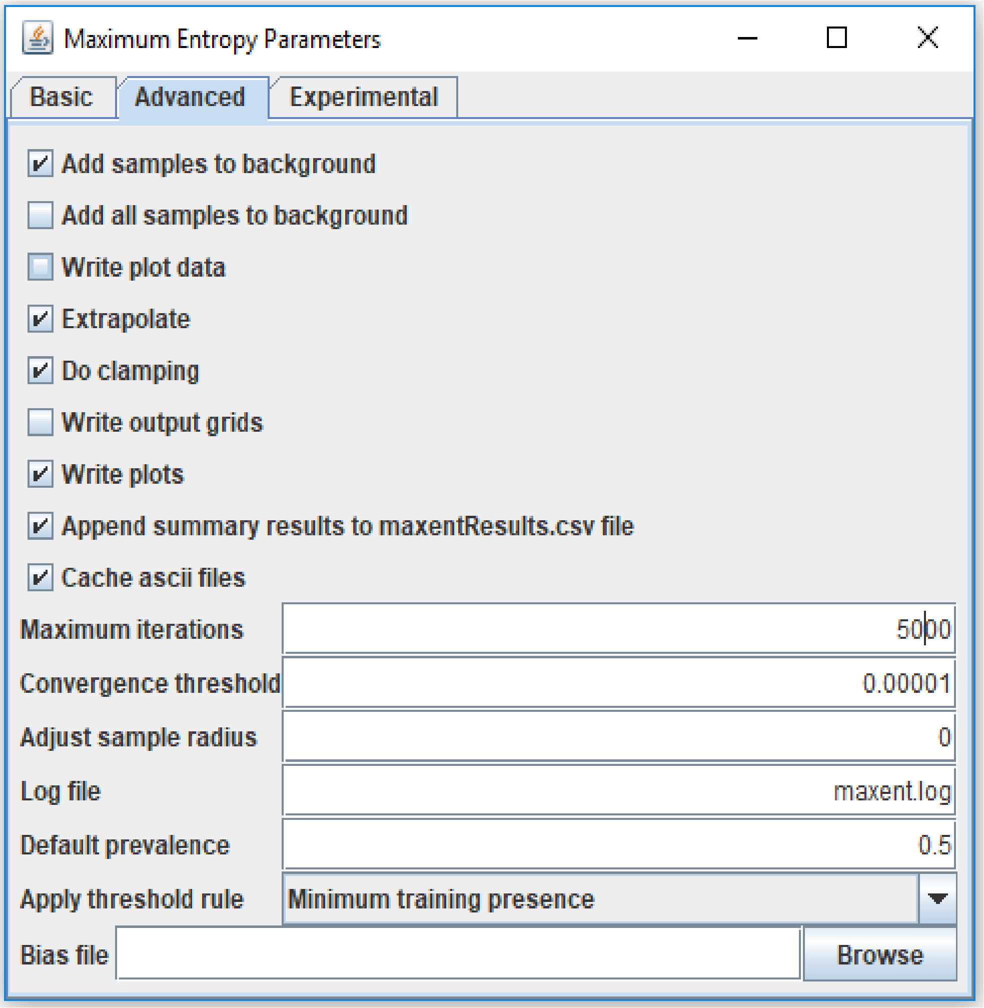


☑ Random seed
☑ Give visual warnings
☑ Show tooltips
☑ Ask before overwriting
☐ Skip if output exists
☑ Remove duplicate presence records
☑ Write clamp grid when projecting
☑ Do MESS analysis when projecting

Random test percentage	25
Regularization multiplier	1
Max number of background points	10000
Replicates	15
Replicated run type	Subsample
Test sample file	

Figure 6 Maxent Basic tab settings

All boxes in the advanced window in Figure 7 except for the 'write output grids' have been left unchanged. The 'write output grids' box has been unchecked in order to speed up model performance. This eliminates the need to produce results from each specific replication in the output folder and instead produces and exports the average results from all the replications into the folder. A value of 5,000 maximum iterations is used over the default value of 500. This allows for a more accurate model as each replication runs through, 5,000 iterations to create the best and most accurate model possible. There is no value selected for the default setting in the ‘apply threshold’ rule box but for this research, ‘minimum training presence’ is selected.



☑ Add samples to background

☐ Add all samples to background

☐ Write plot data

☑ Extrapolate

☑ Do clamping

☐ Write output grids

☑ Write plots

☑ Append summary results to maxentResults.csv file

☑ Cache ascii files

Maximum iterations	5000
Convergence threshold	0.00001
Adjust sample radius	0
Log file	maxent.log
Default prevalence	0.5
Apply threshold rule	Minimum training presence
Bias file	

Figure 7 Maxent Advanced tab settings

Maxent is run six different times with each run using a unique set of individual environmental variables or a unique combination of environmental variables. The results Maxent produces are in the form of an HTML file, which is placed in the outputs folder. The HTML file contains a graphic image that is created by the parameters set in the Maxent windows and produced an ACSII file with habitat suitability information in the form of continuous values ranging 0 to 1. These values can be changed into discrete values by using the Reclassify Tool found in the Spatial Analyst Toolbox. The output ASCII file is imported into ArcMap and converted back into a raster file using the ASCII to Raster tool. The raster is then classified into an appropriate number of breaks using a 90% threshold of sensitivity. This 90% sensitivity threshold can be found in the outputs folder in the maxentResults.csv file under the filed header “10 percentile training presence logistic threshold” (Young et al., 2011). Three breaks are used for this thesis research project to indicate low, medium and high levels of habitat suitability. The results from all six Maxent runs are presented in Chapter 4.

Chapter 4 Results

The results from running Maxent with the sample data come in the form of tables, graphic images and charts that explain which variables contribute to the model, and the overall performance of the model based on the AUC. These results are important because they allow for a well-informed evaluation of how well the model is performing with the given sample data and environmental layers.

In this thesis research project Maxent was run numerous times with different environmental layers and the same sample data. The idea behind using different environmental layers in each run is to see if one set of layers is more indicative of species presence compared to the other layers. This was accomplished by running Maxent numerous times with each set of unique environmental layers and systematically modifying the parameters based on the results. This process entails combining, or in some cases, eliminating environmental layers that are deemed insignificant contributors to the overall performance of the model and rerunning the model with those deemed significant contributors; this information is found in the ‘Analysis of variable contributions’ in the results section after Maxent is successfully run.

The results from running Maxent with each unique set of environmental layers are compared to each other as a way of evaluating which has the most significant influence on the presence of *Spartium junceum* and to see how well suited Maxent is as an SDM. The following subsections report the Maxent results for each unique set of environmental variables. The comparison of the results and the final evaluation of Maxent are discussed Chapter 5.

4.1 Current Distribution of *Spartium junceum* in the Angeles National Forest

Figure 8 represents the current distribution pattern of *Spartium junceum* in the ANF as collected by USFS natural resources staff and contracted partner, the Rancho Santa Ana Botanic Garden. The forest encompasses 655,598 acres (1,024 square miles). The aggregate acreage of all *Spartium junceum* samples collected is 3,163.23 acres or approximately 4.95 square miles. The *Spartium junceum* point data reproduced in Figure 8 does not give an area value for the samples but the original source of the *Spartium junceum* point data does. The source data is the *Spartium junceum* polygon layer that was converted to point data in ArcMap using the Create random points tool. The *Spartium junceum* polygon layer does have values for acreage and area and that is what is used to calculate the aggregate values. Sections 4.2 through 4.7 detail the results from running Maxent with different environmental layers and spatial resolutions.

4.2 Maxent results using BioClim Data at 1 km resolution

This section examines the results that come from using the 19 bioclimatic environmental layers that come from worldclim.org at a maximum spatial resolution of 30 arc seconds (~1 km). The parameters that are used for this set of environmental variables are the same as those that are displayed in Figures 5-7 in Section 3.2.3 and the same parameters are used for each ensuing set of environmental variables in the following subsections. The parameters that were ultimately selected result from trial and error and the recommendations from Young et al. *(*2011) and Philips (2017).

The results in Figure 9 indicate that the forest has roughly equal amounts of high, medium and low suitability areas for *Spartium junceum*.

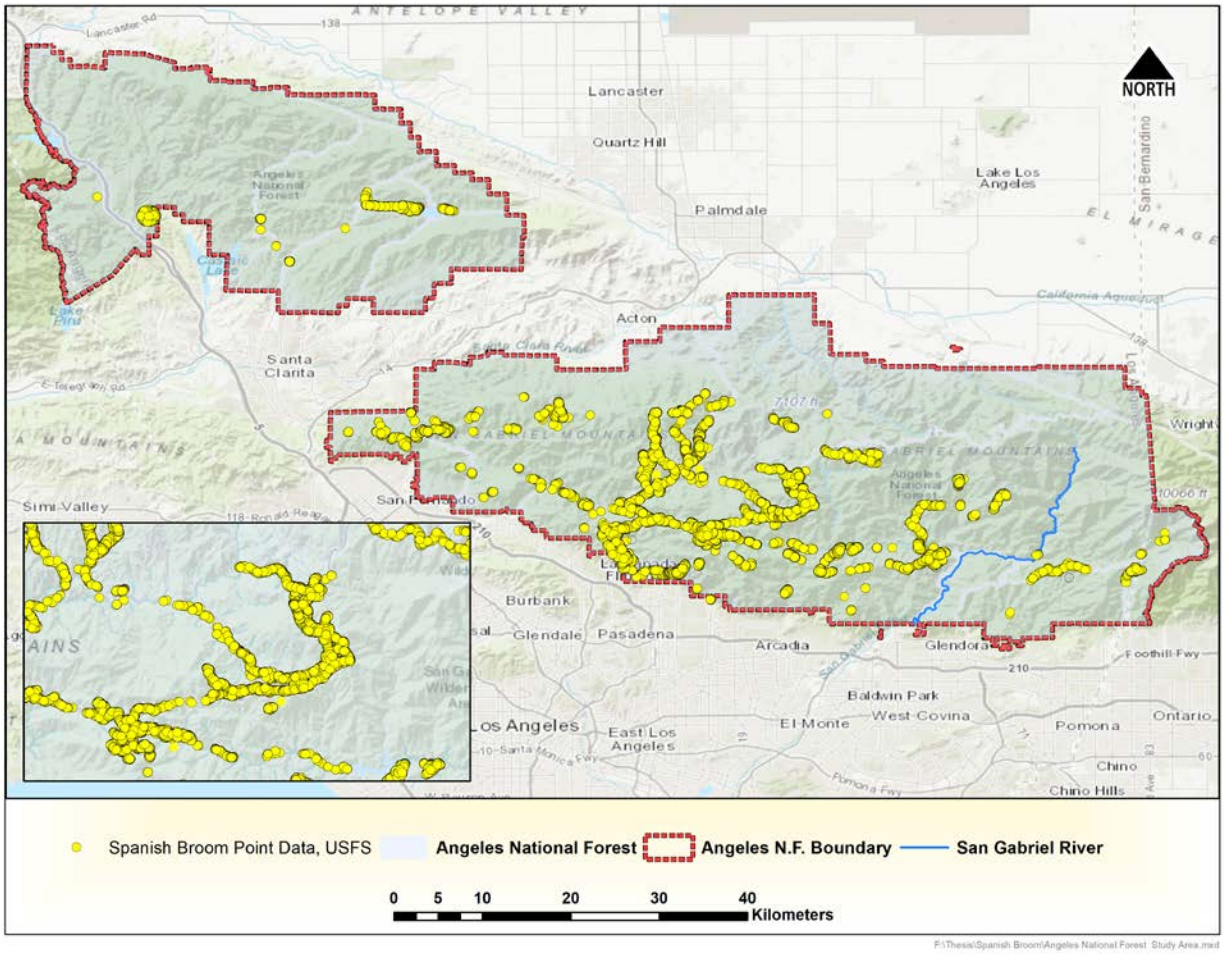


Figure 8 *Spartium junceum* point data in ANF

Figures 11-13 show the results of the jackknife test which indicate that *ascii_wc2_0_bio_30s_15* (precipitation seasonality) had the highest gain when run independently for all three of the jackknife tests of variable importance, which means it has the most useful information when run by itself.

Table 4 shows the variable contribution of each environmental layer in Maxent as a percentage. Under normal circumstances the variables that contribute very little or contribute 0% are taken out and the model is re-run but for this dataset which has such a large spatial resolution it has been deemed unnecessary to do so.

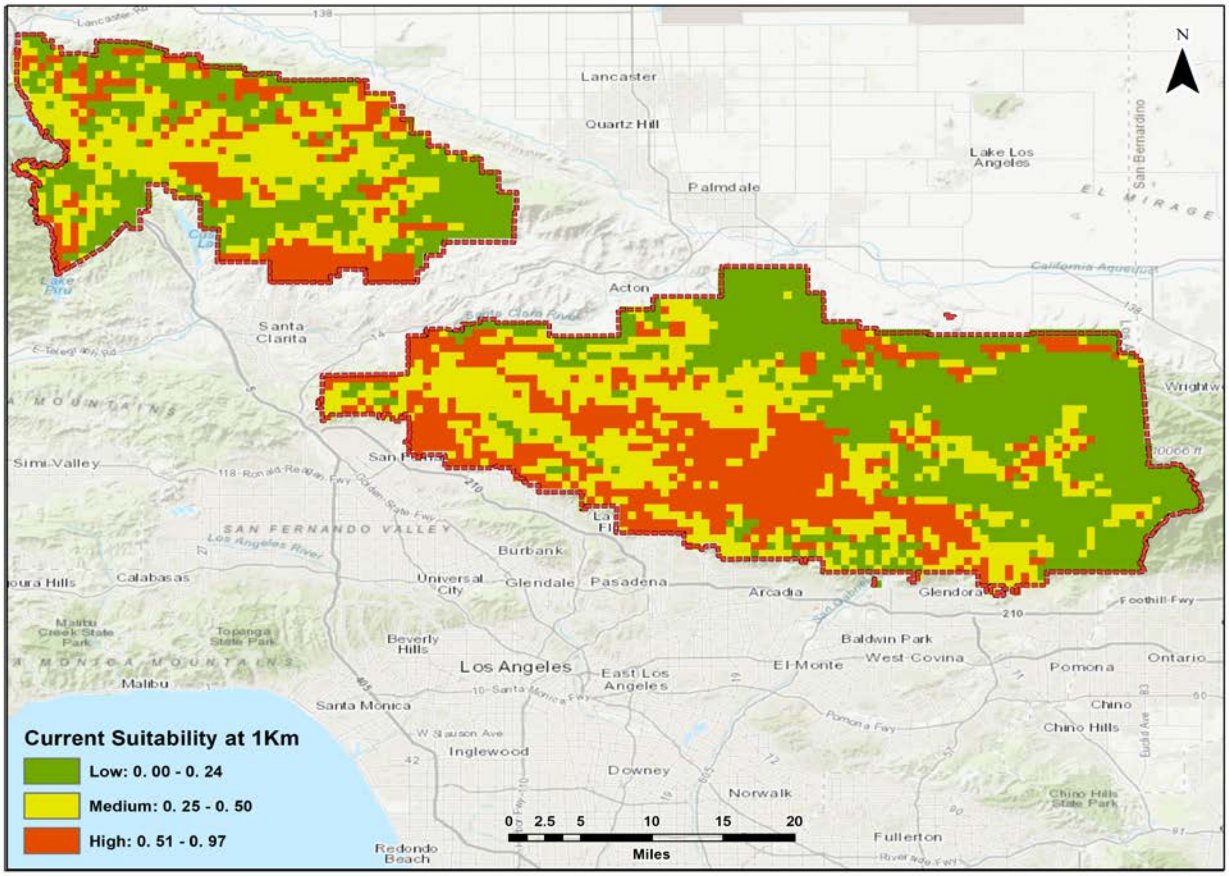


Figure 9 Maxent suitability results using BioClim 1 km spatial resolution in ANF

The world BioClim environmental layers come in a maximum spatial resolution of 1 km which is the equivalent of 0.386 mi. The entire aggregated area of all the *Spartium junceum* sample data is 4.95 mi^2. The spatial resolution of this data set is large in comparison to the study area boundary and the aggregated value for all of the sample data and thus it is not the most indicative of how much influence each variable has on the sample data set. In Section 4.3 Maxent is run with the BioClim variables after they have been resampled to a spatial resolution of 10 m.

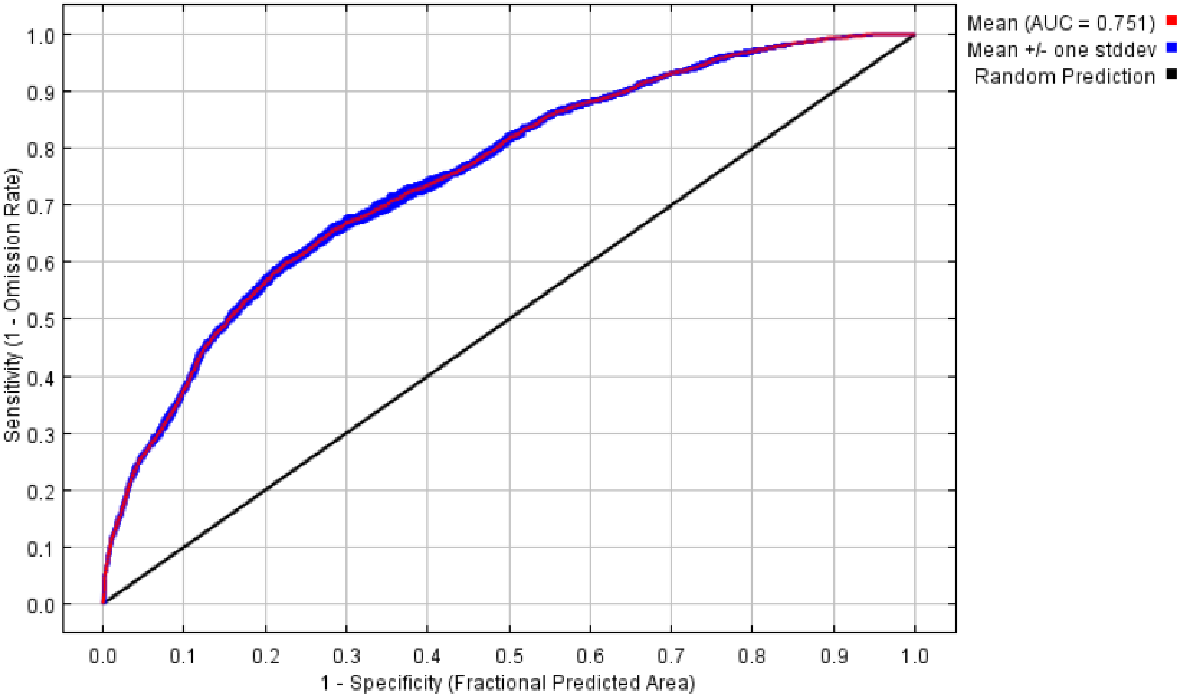


Figure 10 AUC-ROC Curve for *Spartium junceum*

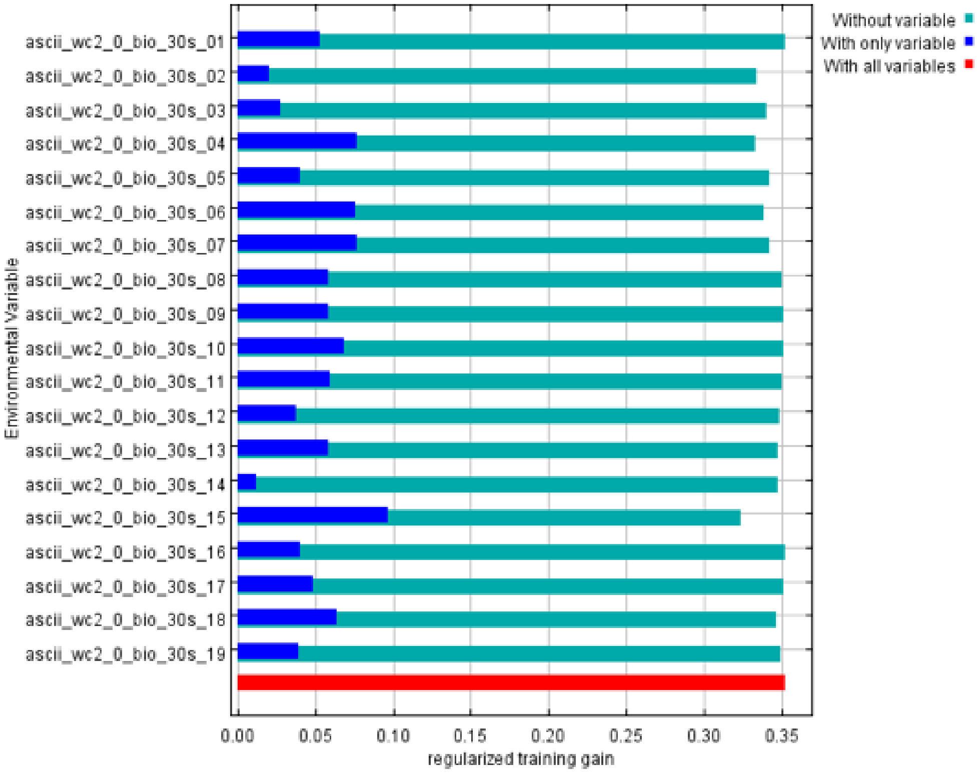


Figure 11 Jackknife of regularized training gain for *Spartium junceum*

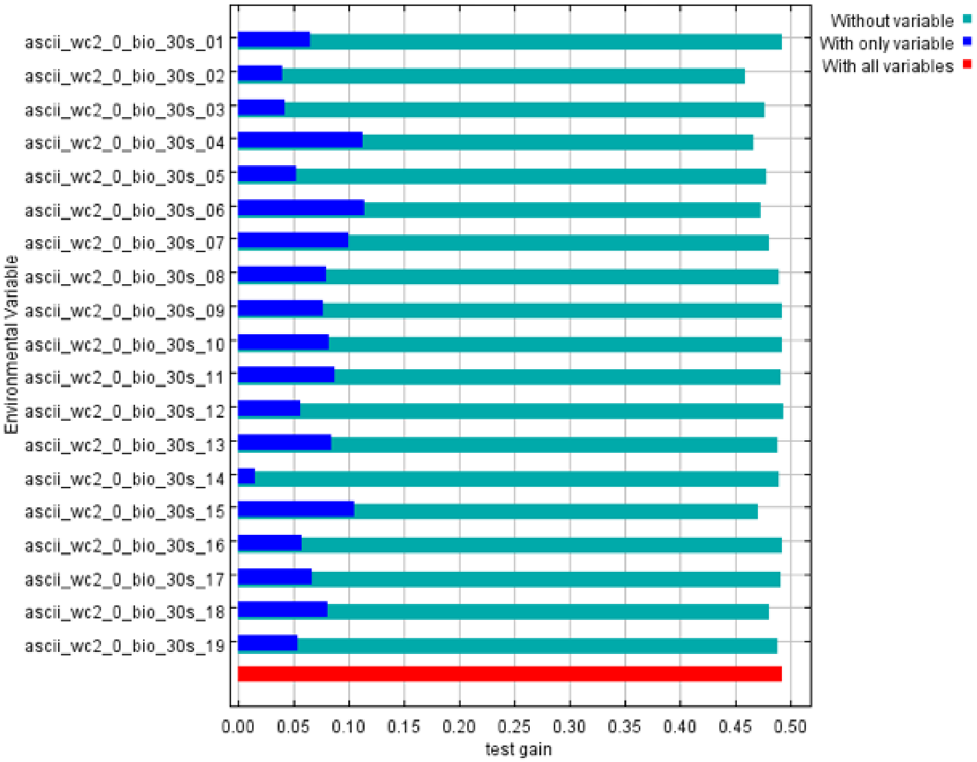


Figure 12 Jackknife for test gain for *Spartium junceum*

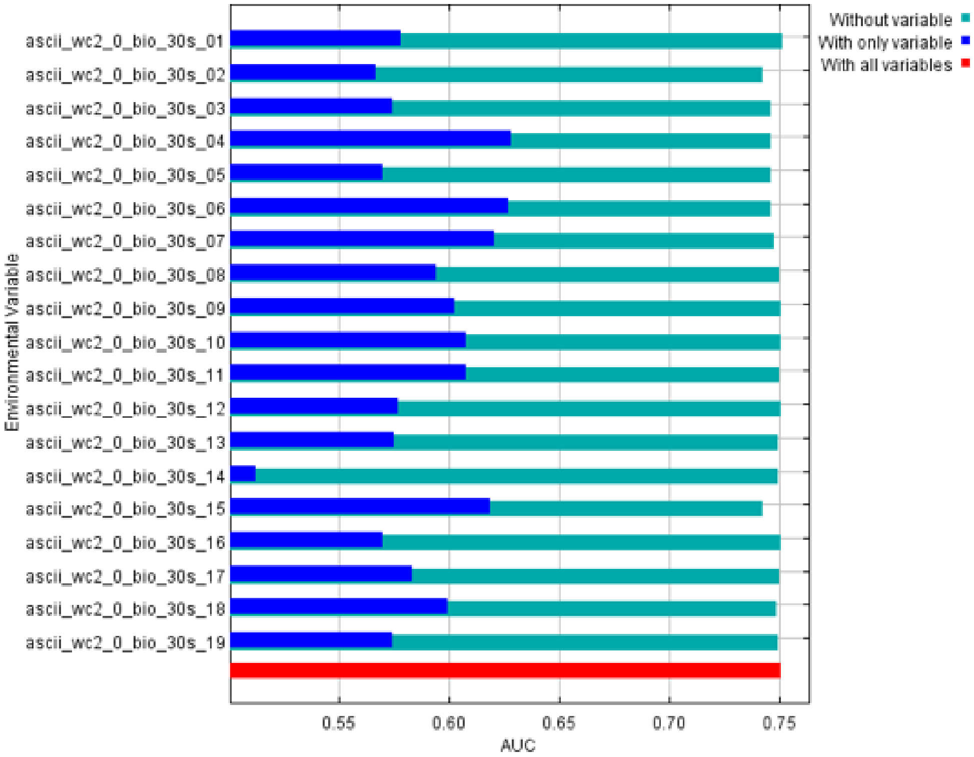
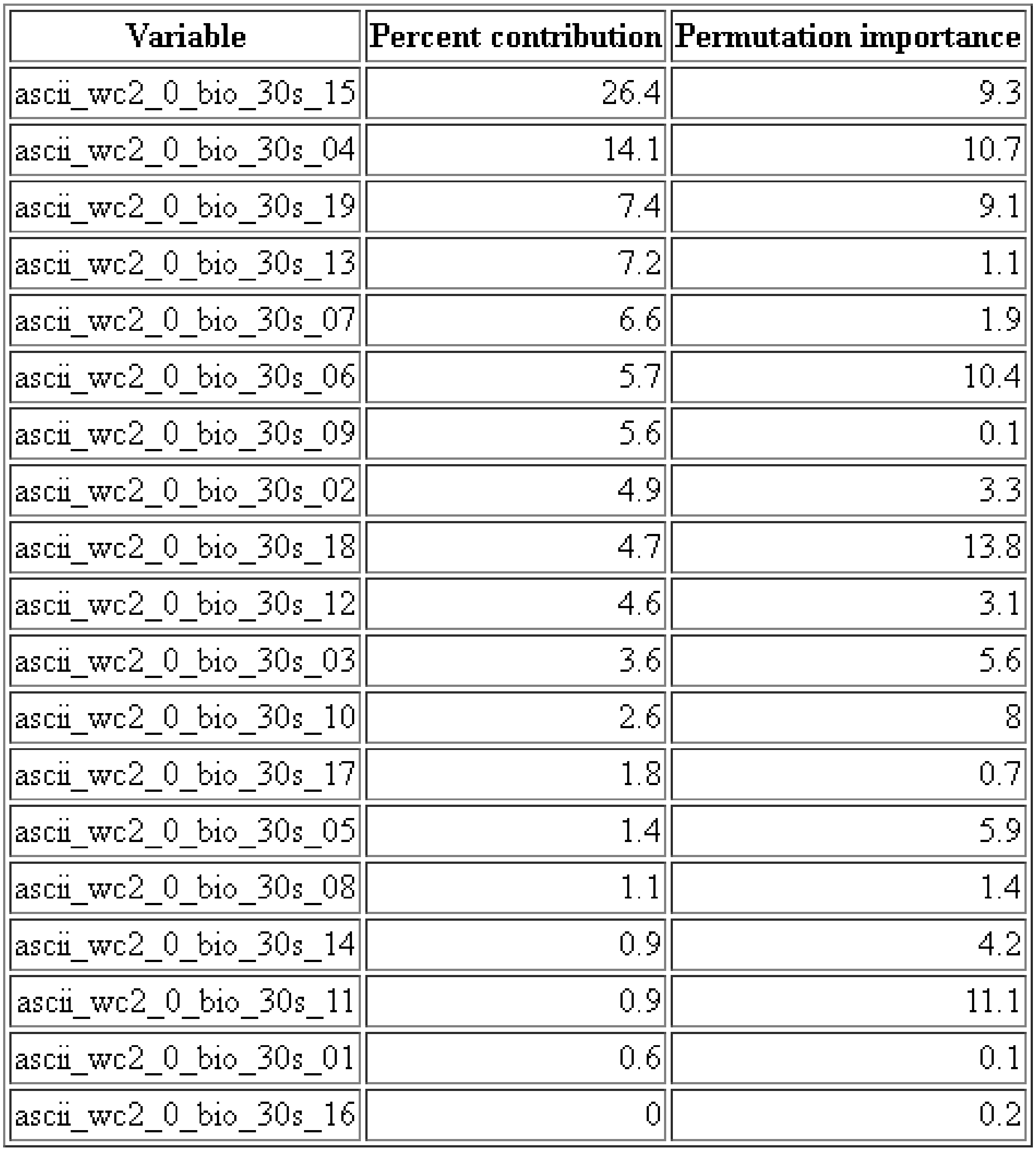


Figure 13 Jackknife of AUC for *Spartium junceum*

Table 4 Variable contributions



Variable	Percent contribution	Permutation importance
ascii_wc2_0_bio_30s_15	26.4	9.3
ascii_wc2_0_bio_30s_04	14.1	10.7
ascii_wc2_0_bio_30s_19	7.4	9.1
ascii_wc2_0_bio_30s_13	7.2	1.1
ascii_wc2_0_bio_30s_07	6.6	1.9
ascii_wc2_0_bio_30s_06	5.7	10.4
ascii_wc2_0_bio_30s_09	5.6	0.1
ascii_wc2_0_bio_30s_02	4.9	3.3
ascii_wc2_0_bio_30s_18	4.7	13.8
ascii_wc2_0_bio_30s_12	4.6	3.1
ascii_wc2_0_bio_30s_03	3.6	5.6
ascii_wc2_0_bio_30s_10	2.6	8
ascii_wc2_0_bio_30s_17	1.8	0.7
ascii_wc2_0_bio_30s_05	1.4	5.9
ascii_wc2_0_bio_30s_08	1.1	1.4
ascii_wc2_0_bio_30s_14	0.9	4.2
ascii_wc2_0_bio_30s_11	0.9	11.1
ascii_wc2_0_bio_30s_01	0.6	0.1
ascii_wc2_0_bio_30s_16	0	0.2

4.3 Maxent results using BioClim Data at 10 m resolution

The results in Figure 14 indicate that a very large portion of the forest can be deemed as suitable habitat for *Spartium junceum* with most of the suitable habitat located in the eastern portion of the San Gabriel Mountains.

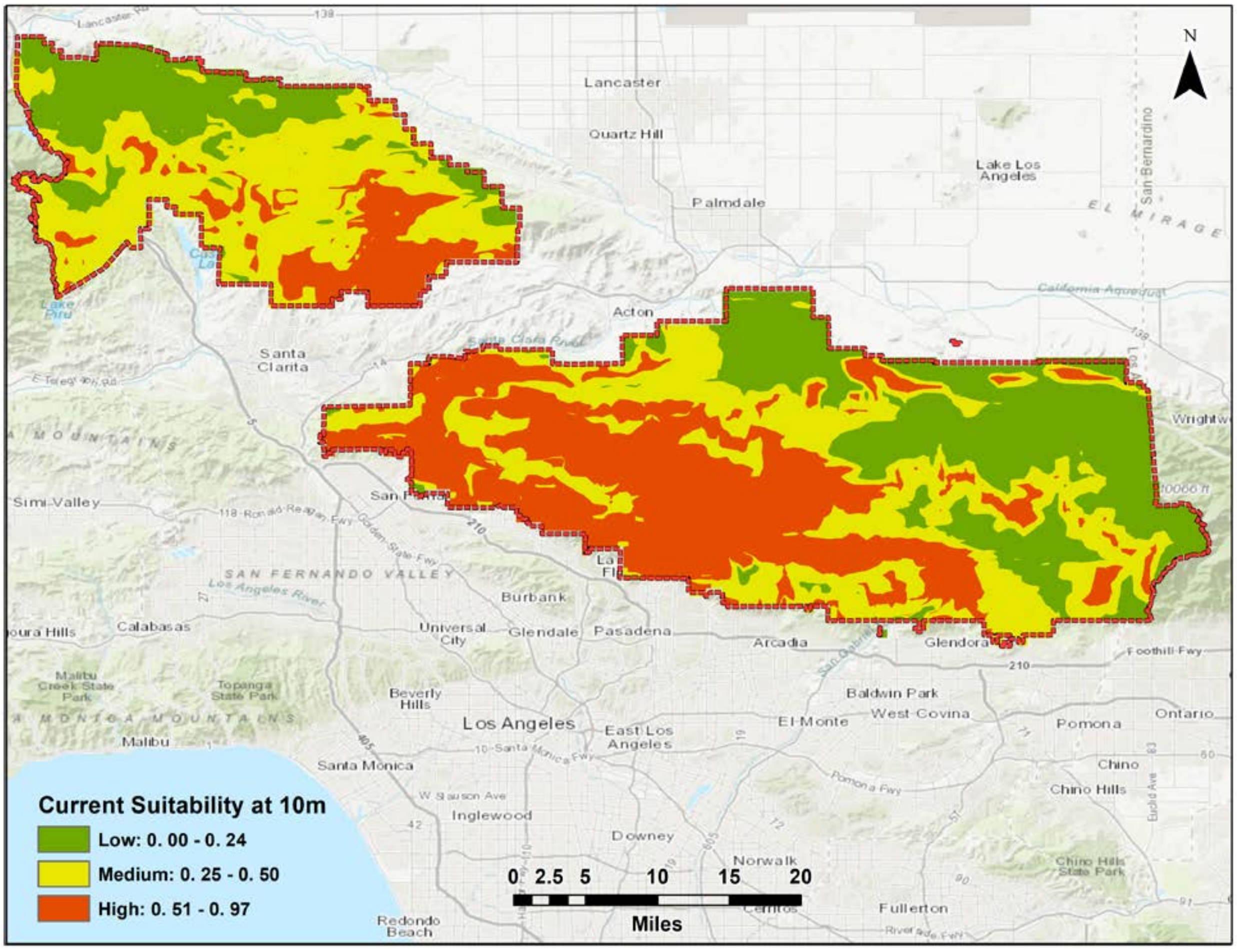


Figure 14 Maxent results using BioClim 10 m resolution in ANF

Figure 15 shows the AUC-ROC value for the BioClim environmental layers after they have been resampled to 10 m spatial resolution and have run through Maxent. The model has an AUC of 0.664 and a standard deviation of 0.006 which means it is significant, but it is slightly less significant than the BioClim data at its original spatial resolution.

Figures 16-18 show the results of the jackknife test which show that *ascii_wc2_0_bio_30s_15* (precipitation seasonality) has the highest gain when run independently for all three jackknife tests of variable importance, which means it has the most useful information when run by itself.

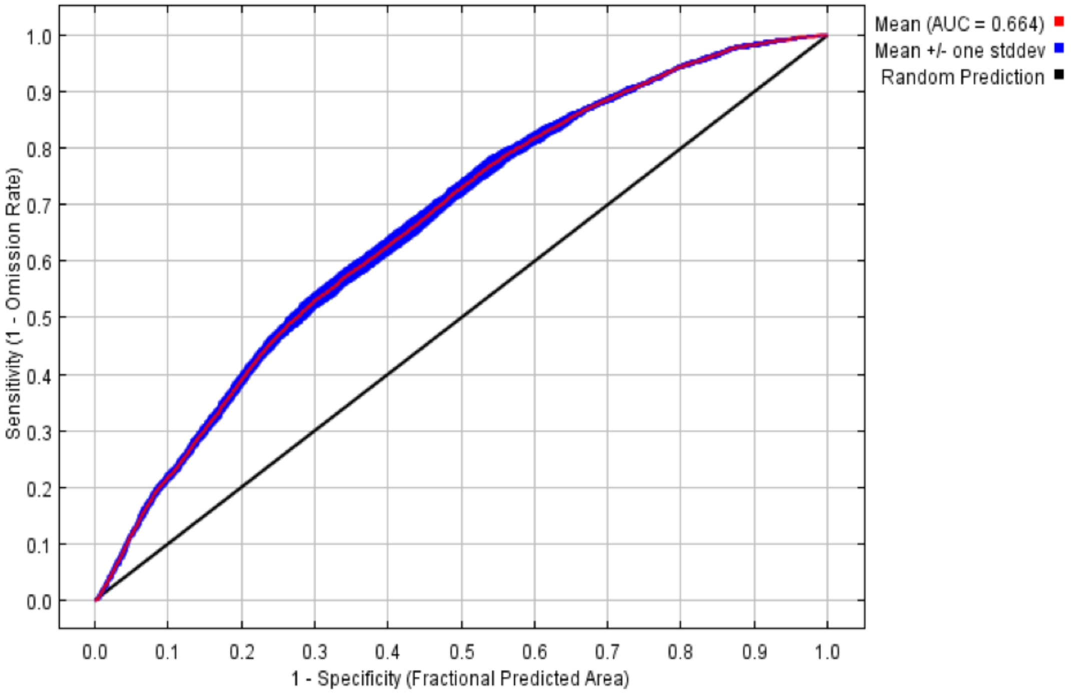


Figure 15 AUC-ROC for *Spartium junceum*

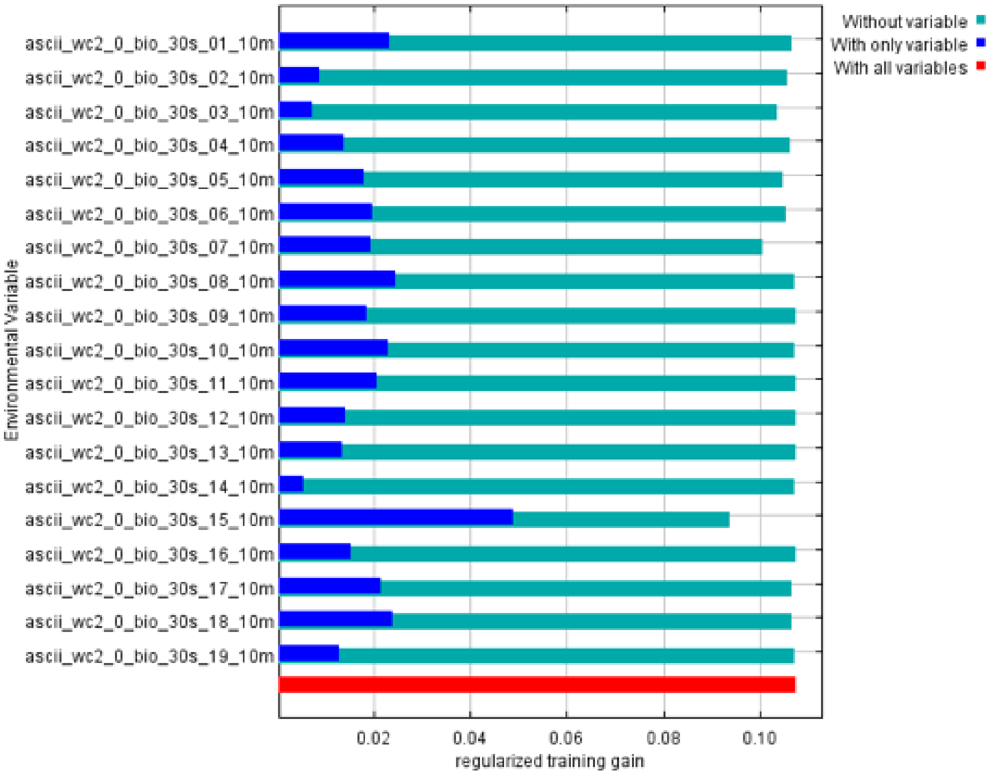


Figure 16 Jackknife regularized training gain for *Spartium junceum*

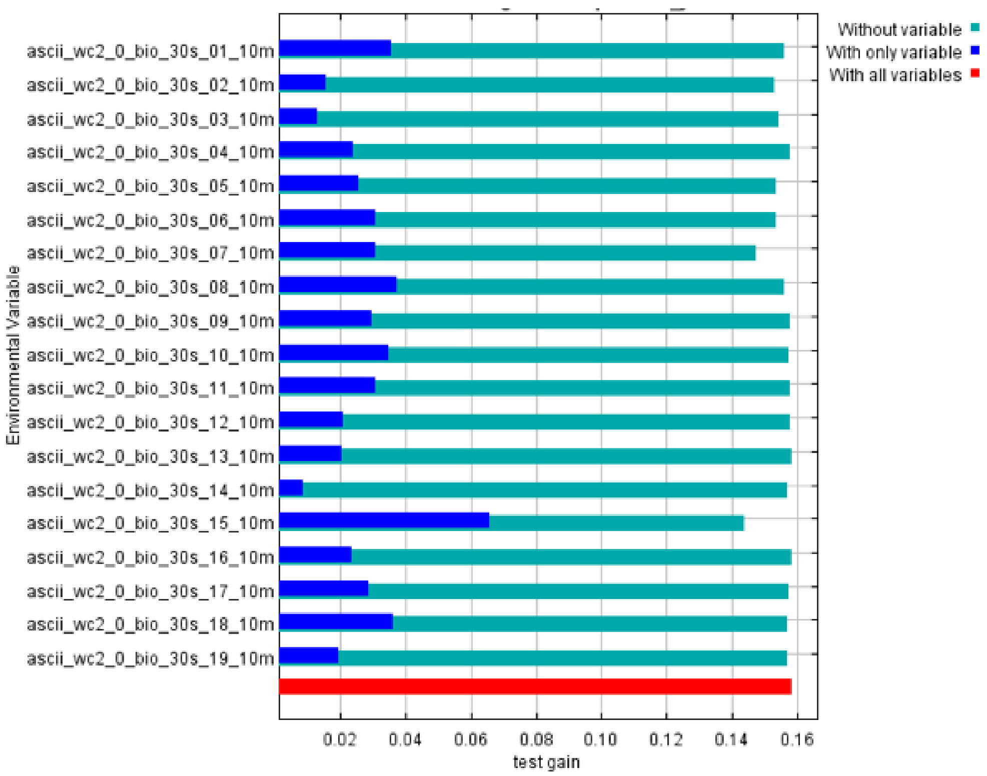


Figure 17 Jackknife of test gain for *Spartium junceum*

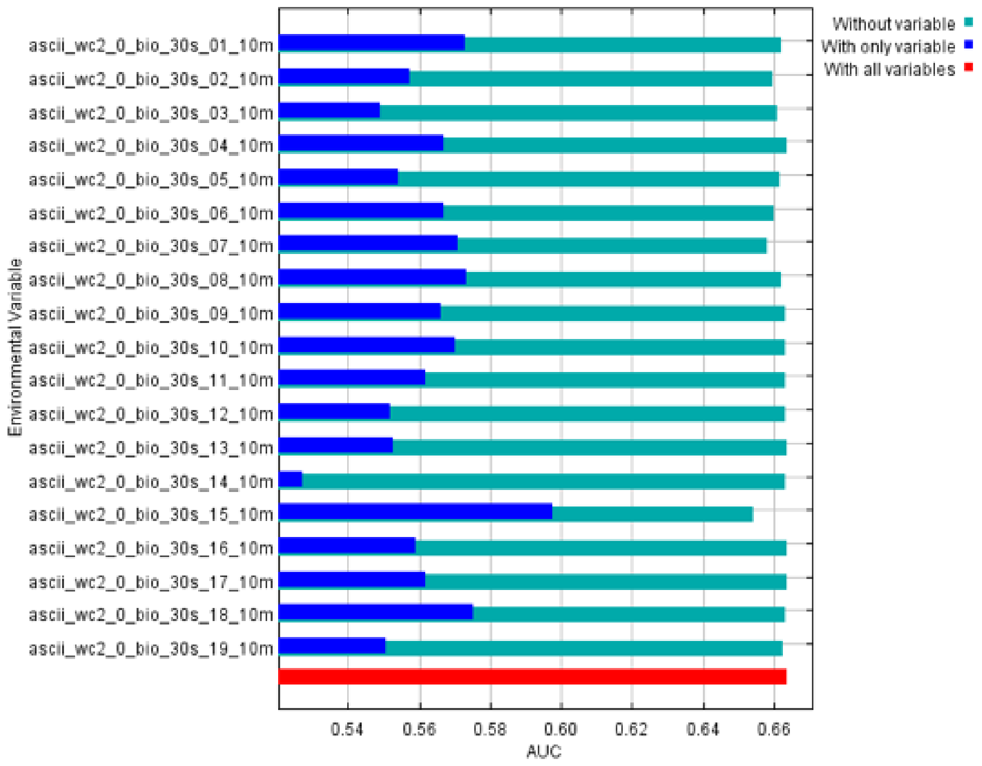
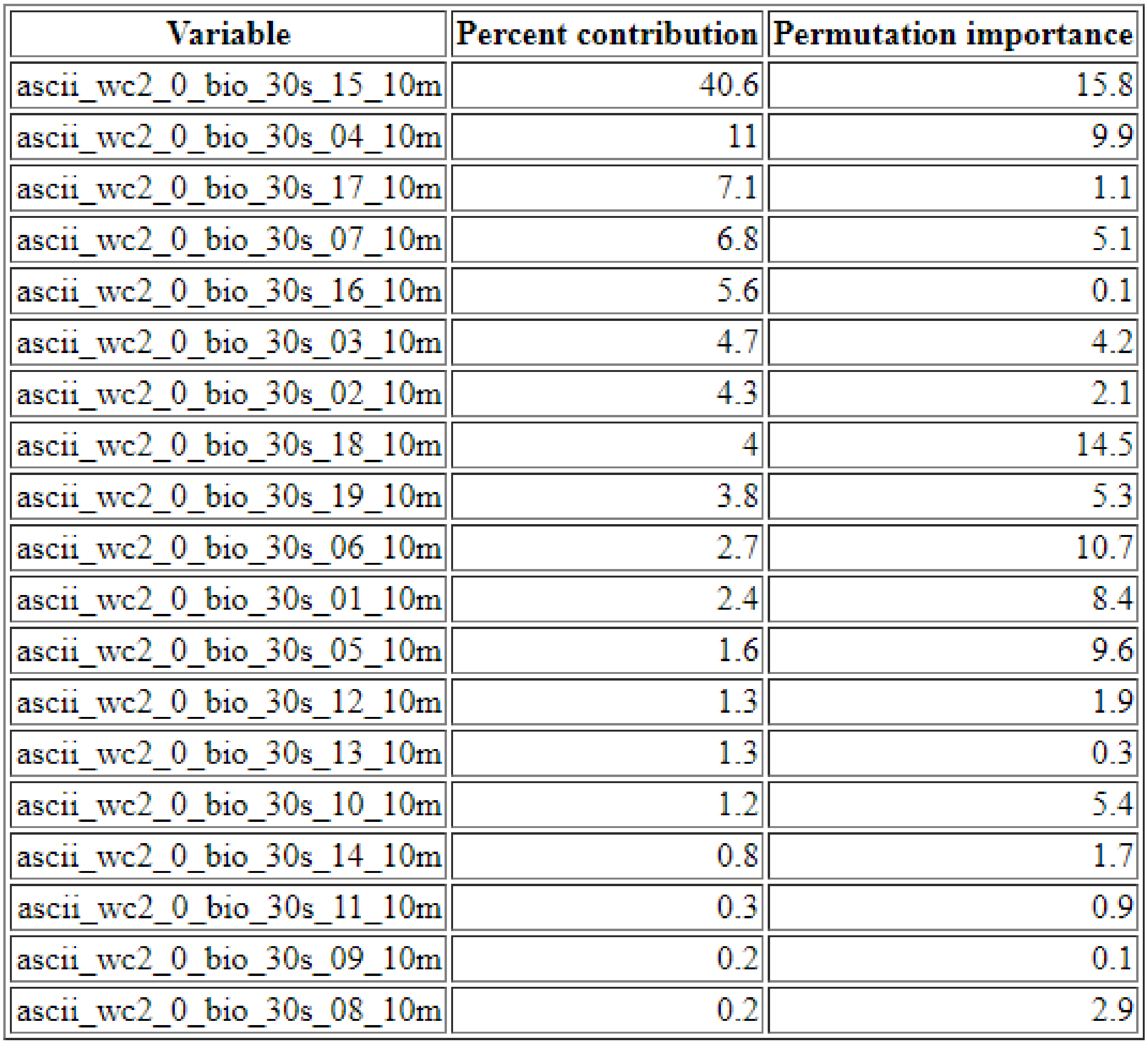


Figure 18 Jackknife of AUC for *Spartium junceum*

Table 5 shows the variable contribution for the BioClim data at 10 m spatial resolution. The results are different than those of the BioClim data at its original spatial resolution. While the variable with the highest contribution is still *ascii_wc2_0_bio_30s_15*, the percentage increased with the resampled variables from 26.4 to 40.6%. In addition, the lowest contributing variable changed from *ascii_wc2_0_bio_30s_16* (precipitation of the wettest quarter) in the original BioClim set to *ascii_wc2_0_bio_30s_08* (mean temperature of the wettest quarter) in the resampled set with both contributing variables contributing 0%.

Table 5 Variable contributions



Variable	Percent contribution	Permutation importance
ascii_wc2_0_bio_30s_15_10m	40.6	15.8
ascii_wc2_0_bio_30s_04_10m	11	9.9
ascii_wc2_0_bio_30s_17_10m	7.1	1.1
ascii_wc2_0_bio_30s_07_10m	6.8	5.1
ascii_wc2_0_bio_30s_16_10m	5.6	0.1
ascii_wc2_0_bio_30s_03_10m	4.7	4.2
ascii_wc2_0_bio_30s_02_10m	4.3	2.1
ascii_wc2_0_bio_30s_18_10m	4	14.5
ascii_wc2_0_bio_30s_19_10m	3.8	5.3
ascii_wc2_0_bio_30s_06_10m	2.7	10.7
ascii_wc2_0_bio_30s_01_10m	2.4	8.4
ascii_wc2_0_bio_30s_05_10m	1.6	9.6
ascii_wc2_0_bio_30s_12_10m	1.3	1.9
ascii_wc2_0_bio_30s_13_10m	1.3	0.3
ascii_wc2_0_bio_30s_10_10m	1.2	5.4
ascii_wc2_0_bio_30s_14_10m	0.8	1.7
ascii_wc2_0_bio_30s_11_10m	0.3	0.9
ascii_wc2_0_bio_30s_09_10m	0.2	0.1
ascii_wc2_0_bio_30s_08_10m	0.2	2.9

4.4 Maxent results using DEM Raster layers at 10 m resolution

The map reproduced in Figure 19 uses a DEM and other layers created from that DEM by using the Spatial Analyst toolbox as mentioned in Section 3.2.1. The DEM file used for this Maxent run came at a 10 m spatial resolution and did not need to be resampled. The 10 m spatial resolution is well suited to the size of the study region. As seen in the two previous results maps, most of the suitable habitat is found in the eastern portion of the mountain range. In addition, there are signs of increased occurrence and more pronounced dispersion of highly suitable areas in the study area when using the DEM variables in comparison to the BioClim variables.

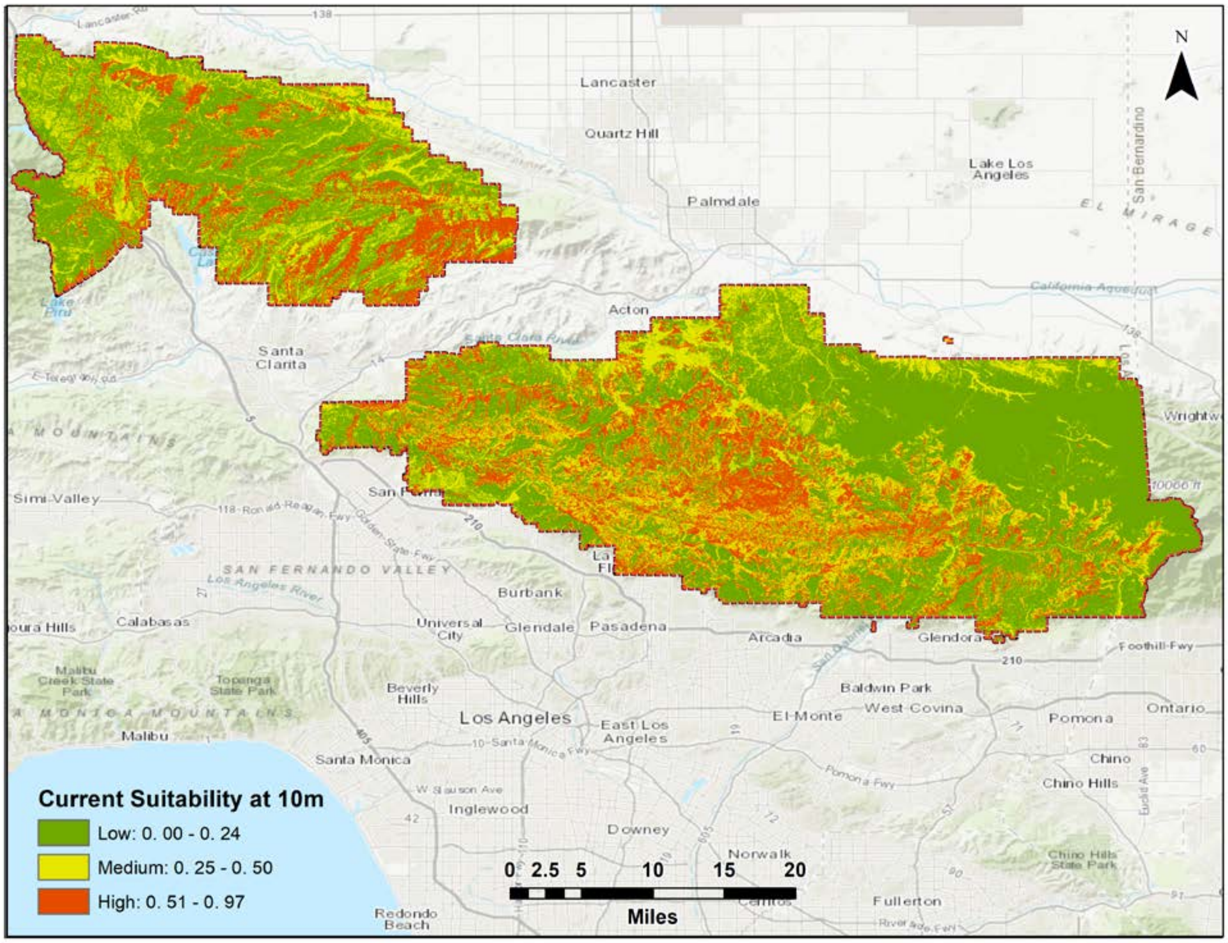


Figure 19 Maxent results using DEM at 10 m resolution in the ANF

Figure 20 shows the AUC-ROC value for the DEM environmental layers at their recorded 10 m resolution. The model has an AUC of 0.707 and a standard deviation of 0.007 which means it is significant but is slightly less significant than the AUC values (0.751) produced when running the BioClim layers at their original 1 km spatial resolution in Maxent.

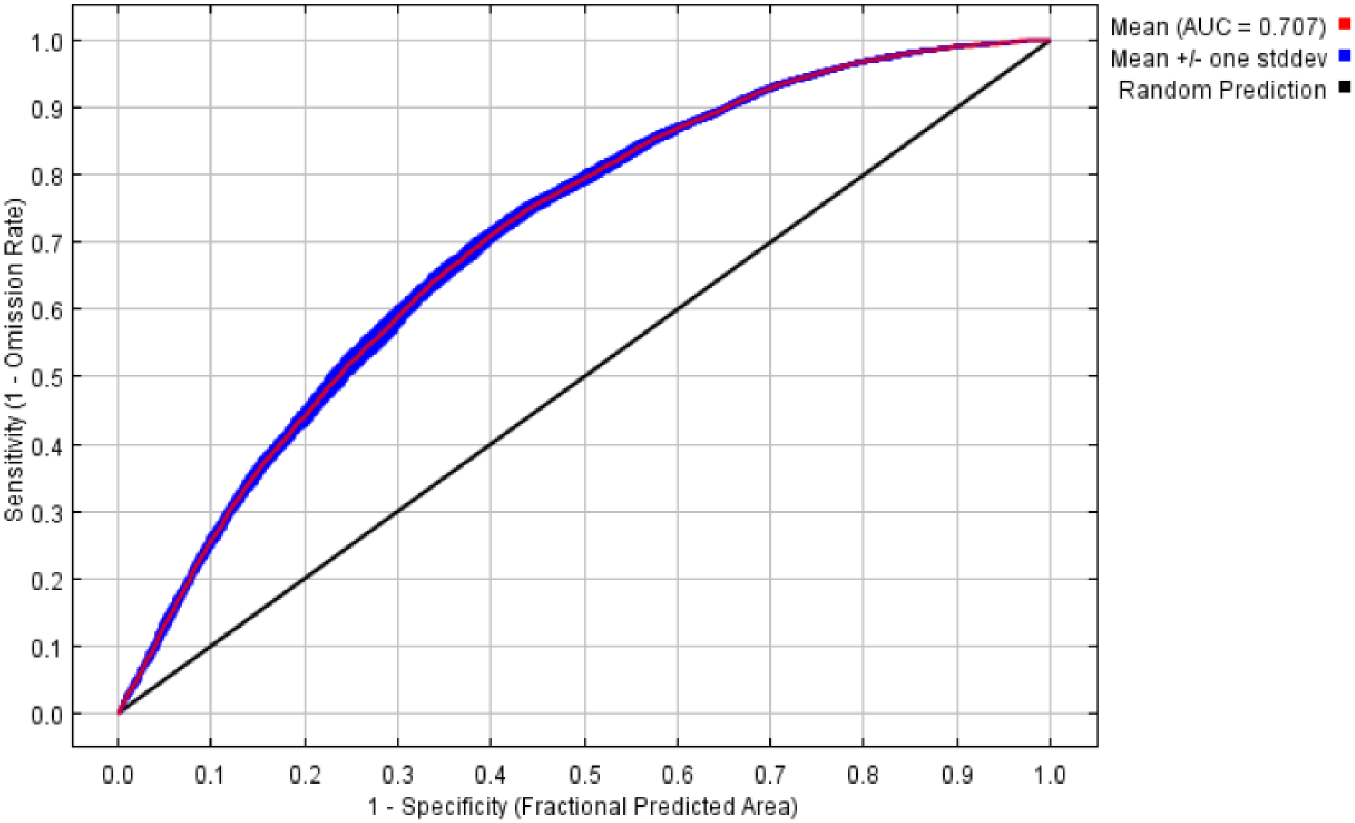


Figure 20 AUC-ROC for *Spartium junceum*

Figures 21-23 show the results of the jackknife test which shows that *ascii_dem_anf_visbility* (visibility or *Spartium junceum*) had the highest gain when run independently for all three jackknife tests for variable importance, which means it has the most useful information when run by itself.

.

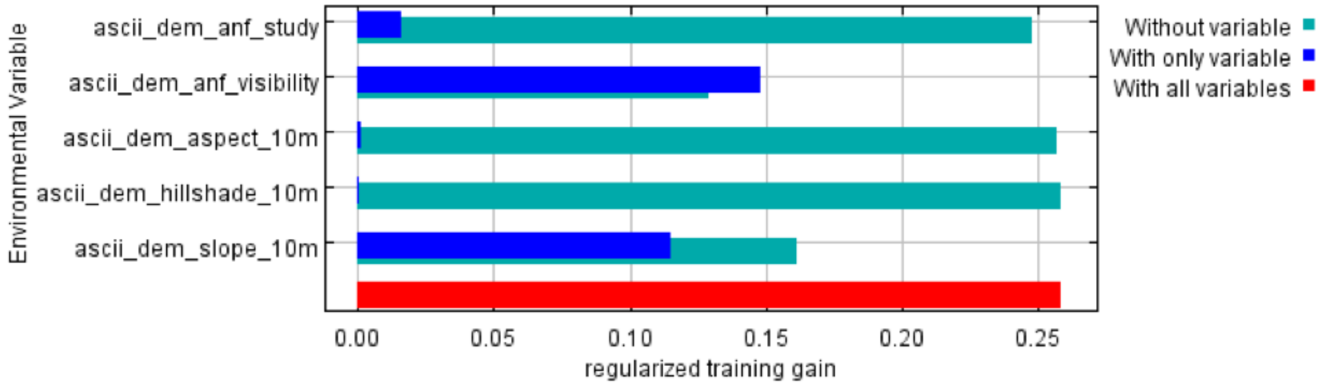


Figure 21 Jackknife of regularized training gain curve for *Spartium junceum*

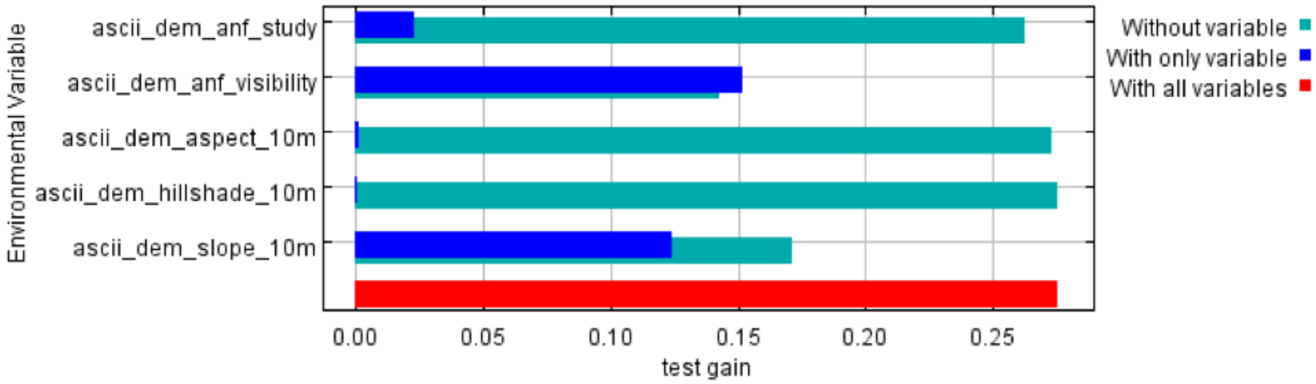


Figure 22 Jackknife of regularized training for *Spartium junceum*

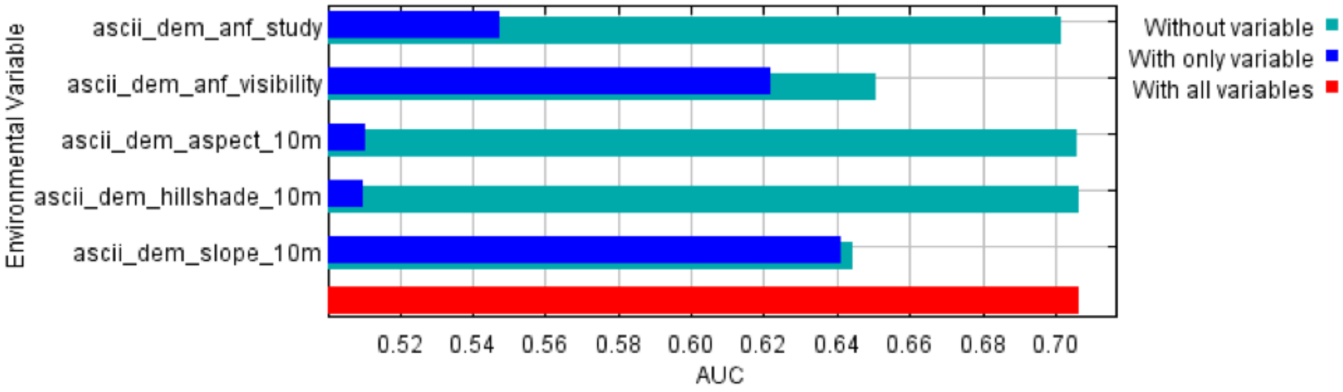
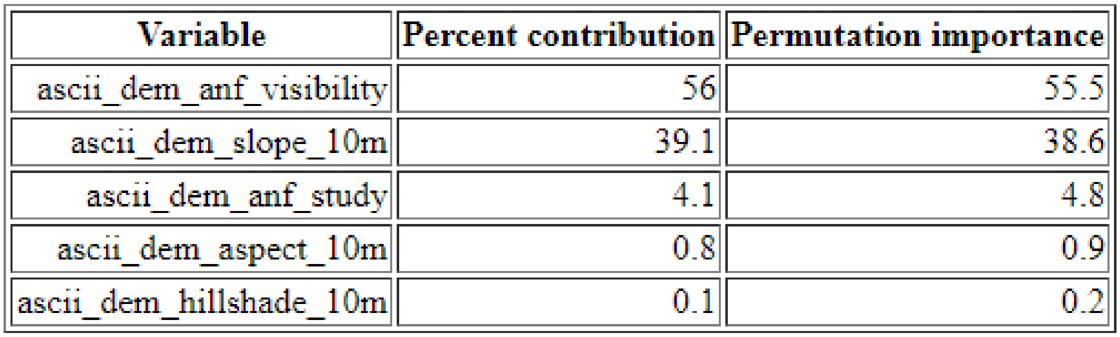


Figure 23 Jackknife of AUC for *Spartium junceum*

Table 6 shows the variable contribution for the DEM layers at a 10 m resolution. The layer with the highest variable contribution is supported by the results of the jackknife tests above.

Table 6 Variable Contribution



Variable	Percent contribution	Permutation importance
ascii_dem_anf_visibility	56	55.5
ascii_dem_slope_10m	39.1	38.6
ascii_dem_anf_study	4.1	4.8
ascii_dem_aspect_10m	0.8	0.9
ascii_dem_hillshade_10m	0.1	0.2

The DEM layers were recorded at the 10 m scale and thus the spatial resolution is more appropriate for the size of study area. The DEM layers benefit from being recorded at this scale and do not have to be interpolated whereas the BioClim layers do. While the original set of BioClim layers have a higher AUC-ROC value than the DEM layers it does not necessarily mean that they are more accurate or are better indicators of suitable habitat for the sample species. This will be discussed further in Chapter 5.

4.5 Maxent results using Euclidean Distance ANF features layers at 10 m resolution

The map reproduced in Figure 24 uses environmental layers that were created by converting USFS vector data into raster data using the Euclidean distance tool in ArcMap. The rasters were created using the same 10 m spatial resolution as the DEM environmental layers in order to maintain consistency. As is seen in the three previous maps, most of the suitable habitat is found in the eastern portion of the mountain range although there is increased suitability in the western portions of the mountain range as well. The results from running Maxent with this set of environmental layers are the most different. The predicted suitability areas are very concentrated along a visible network. This visible network is derived from the converted vector files with USFS roads producing the most visible results in Figure 24. Furthermore, the presence of USFS roads helps explain why there is a noticeable increase in suitability areas in the western portion of the study area, and similarly helps explain the large patches of unsuitable areas in the eastern portion of the study area.

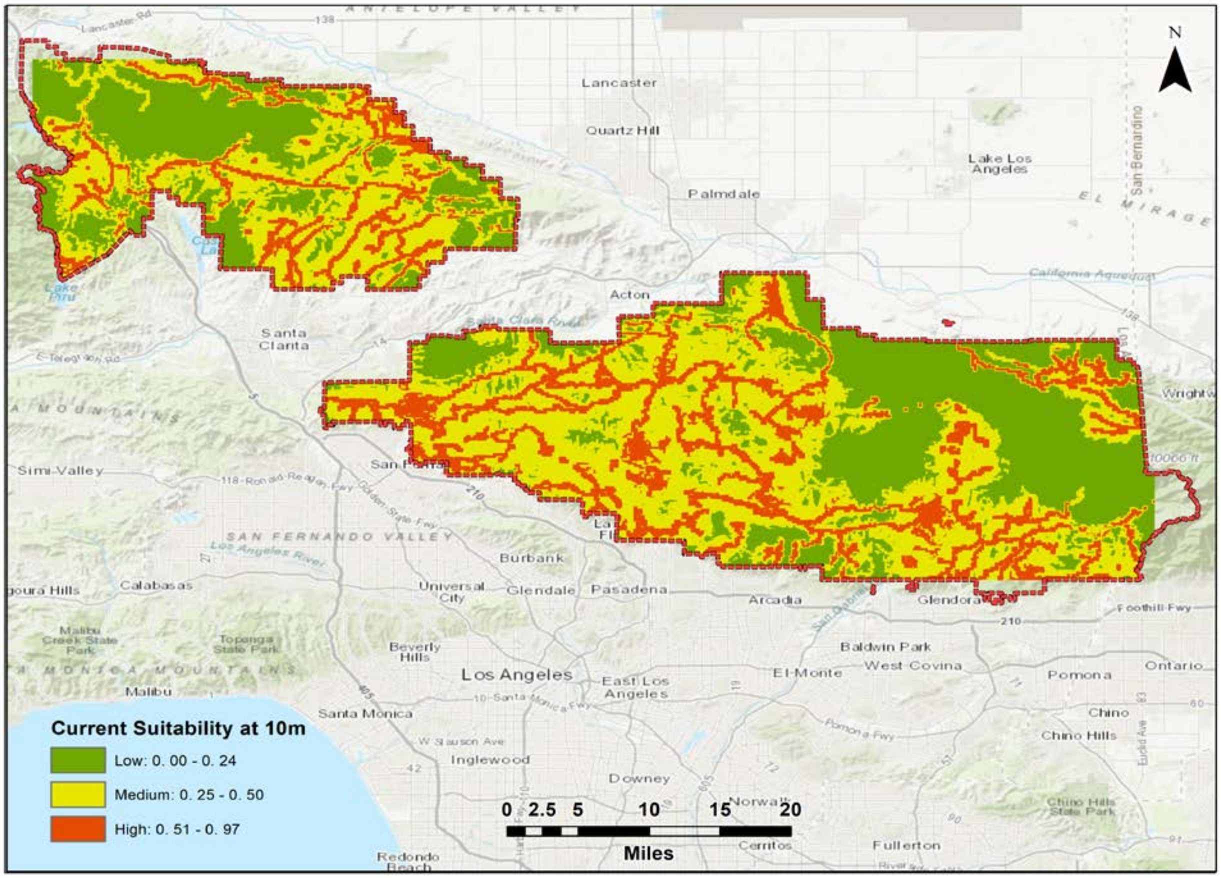


Figure 24 Maxent results using Euclidean distance ANF features and 10 m resolution in ANF

Figure 25 shows the AUC-ROC value for the Euclidean distance proximity environmental layers at their recorded 10 m spatial resolution. The model has an AUC of 0.692

and a standard deviation of 0.005 which means it is significant but less significant than the Maxent results from the DEM and original BioClim layers.

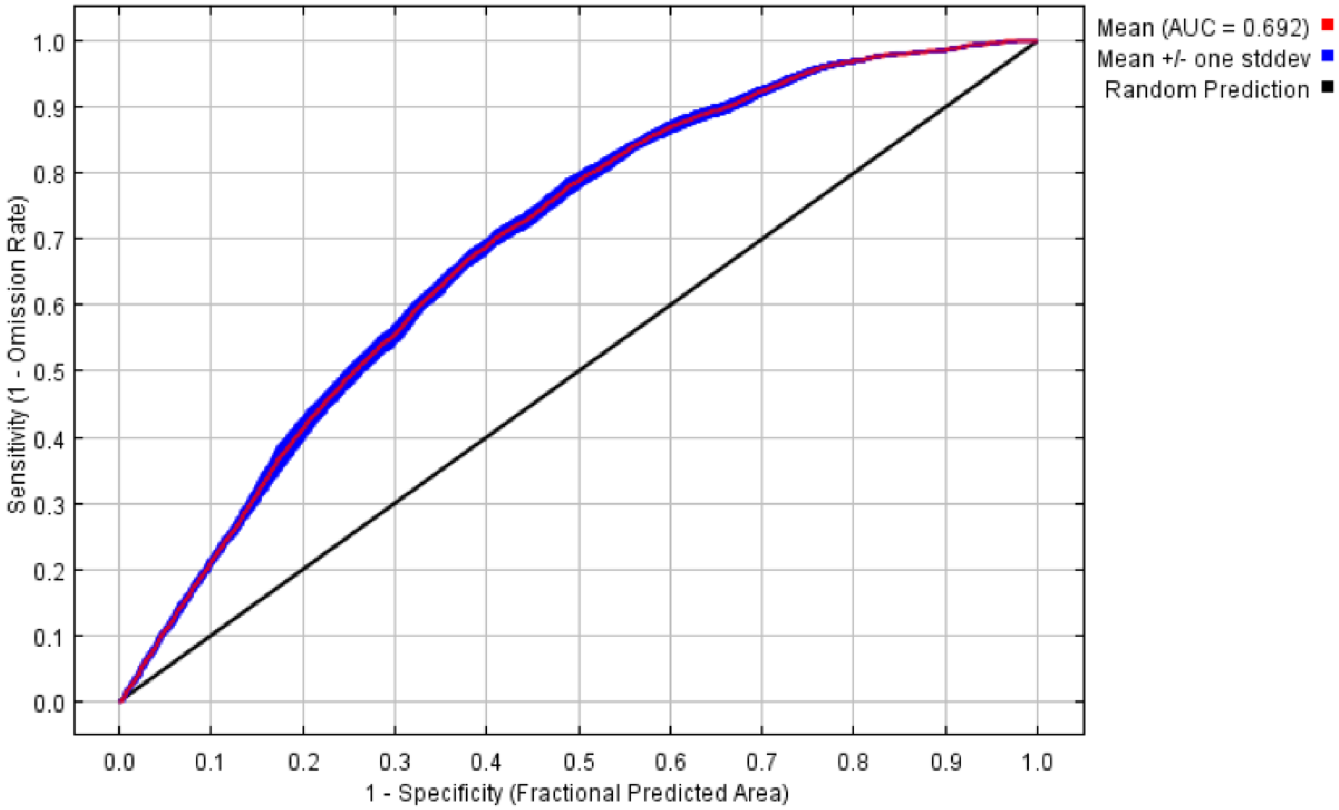


Figure 25 AUC-ROC for *Spartium junceum*

Figures 26-28 show the results of the jackknife test in which the *anf_roads* (Euclidean distance Angeles National Forest roads) had the highest gain when run independently for all three jackknife tests for variable importance, which means it has the most useful information when run by itself.

Table 7 shows the variable contribution for the Euclidean distance ANF features layers at 10 m resolution. The Euclidean distance layer of *Spartium junceum* to local forest service roads has the highest variable contribution and is supported by the results of the jackknife tests above.

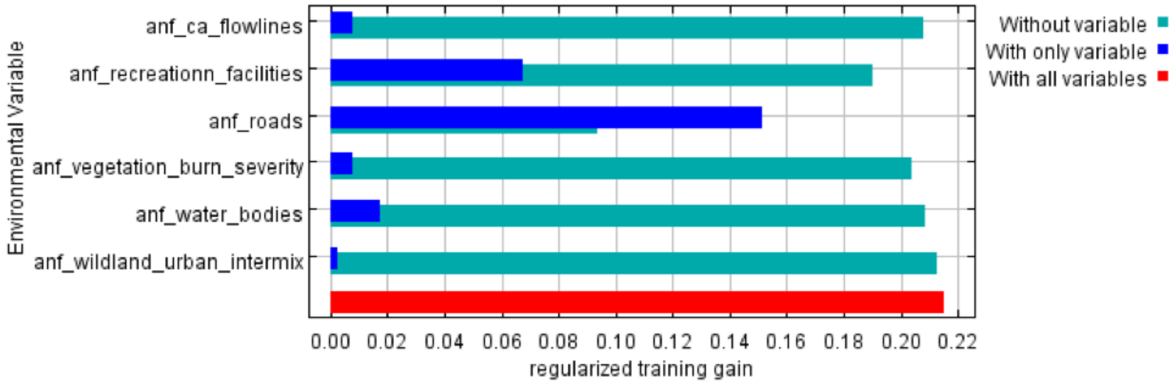


Figure 26 Jackknife of regularized training gain for *Spartium junceum*

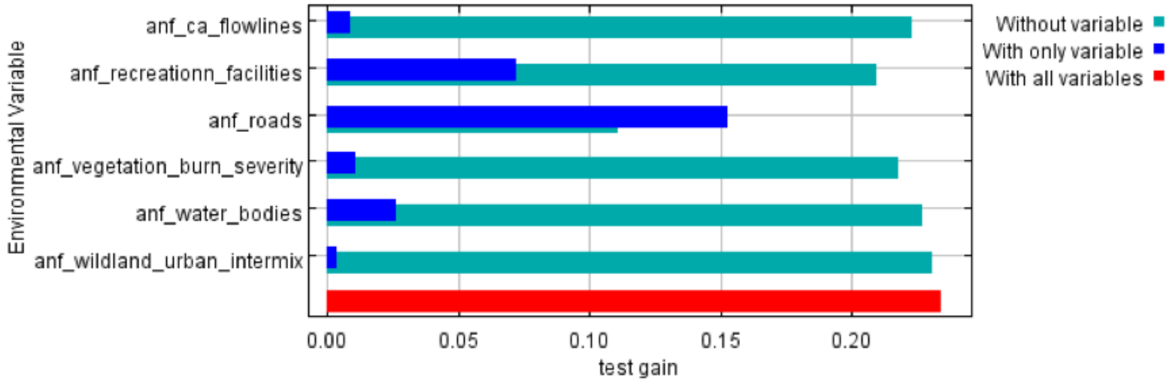


Figure 27 Jackknife of test gain for *Spartium junceum*

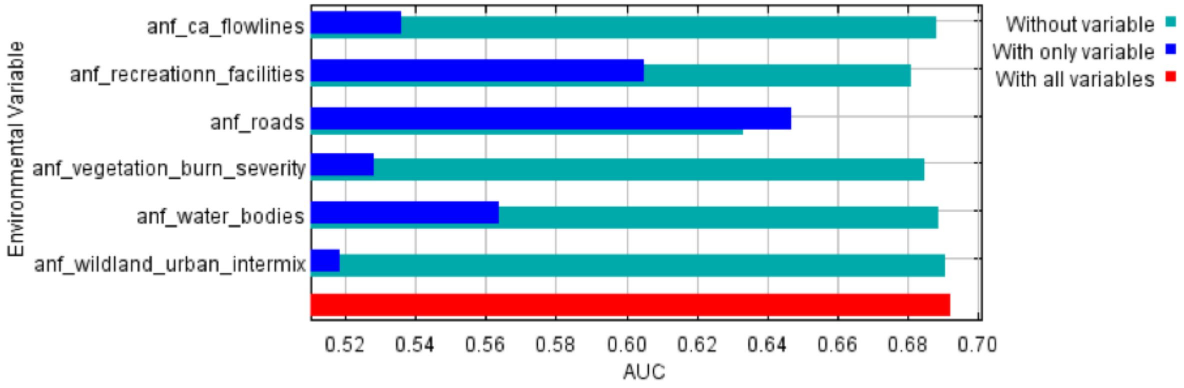
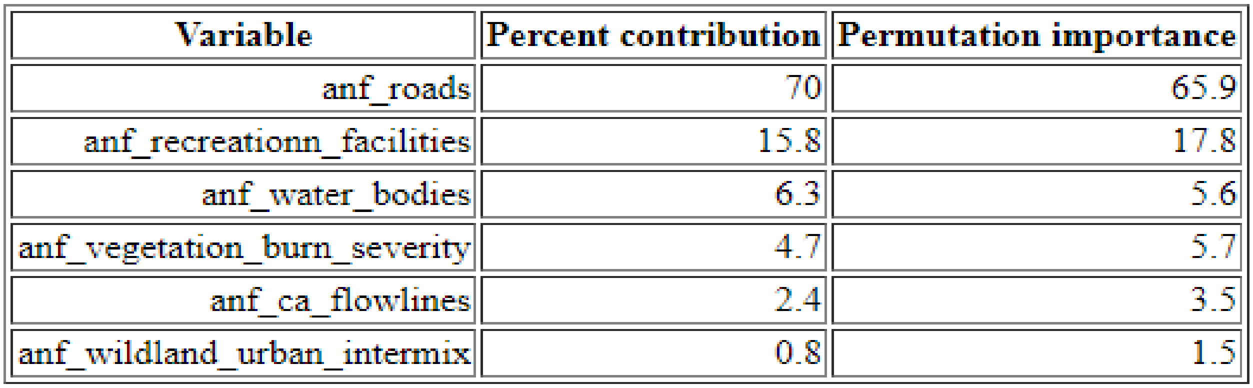


Figure 28 Jackknife of AUC for *Spartium junceum*

Table 7 Variable contributions using Euclidean distance ANF features and 10 m resolution



Variable	Percent contribution	Permutation importance
anf_roads	70	65.9
anf_recreationn_facilities	15.8	17.8
anf_water_bodies	6.3	5.6
anf_vegetation_burn_severity	4.7	5.7
anf_ca_flowlines	2.4	3.5
anf_wildland_urban_intermix	0.8	1.5

4.6 Maxent results using combined 10 m BioClim and DEM layers

The map reproduced in Figure 29 uses a combination of environmental layers from Sections 4.3 and 4.4. The spatial resolution of both sets of environmental layers is 10 m and the underlying values of the individual ASCII files are the same which allowed Maxent to run both layers combined without any issues. As seen in the four previous maps, most of the suitable habitat is again found in the eastern portion of the forest but there are some increased areas concentrated at the eastern end of the western portion of the forest. The results from running Maxent with these combined environmental layers shows higher levels of suitability located in the south facing aspects in the forest and low suitability on the north facing aspects in the forest.

Figure 30 shows the AUC-ROC value for the BioClim and DEM 10 m layers. The model has an AUC of 0.730 and a standard deviation of 0.005 which means it is significant and has a higher AUC-ROC value than the BioClim 10 m, DEM 10 m, and Euclidean distance ANF features layers when run on their own. However, it is still lower than the results produced with the original BioClim 1 km layers when run on their own.

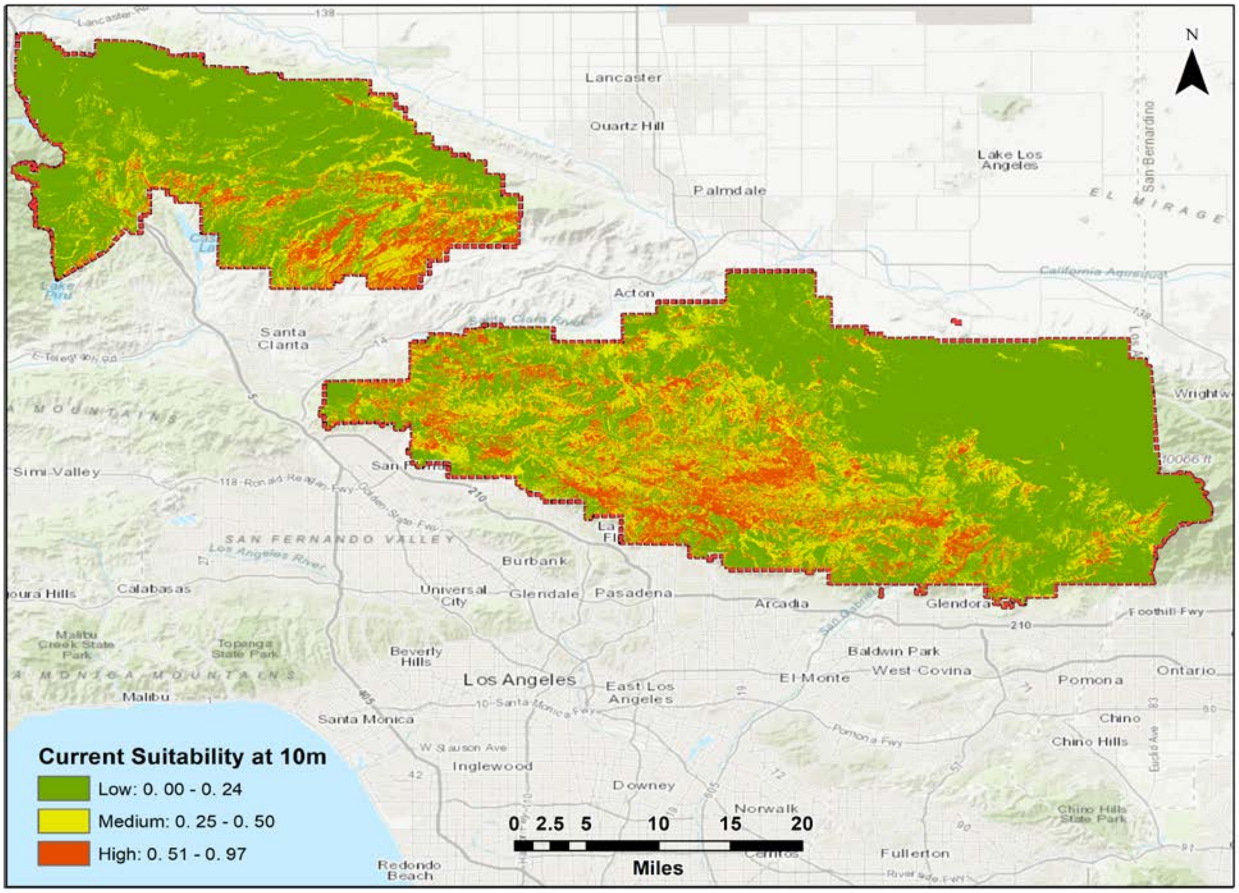


Figure 29 Maxent results using BioClim and DEM 10 m layers combined in ANF

Figures 31-33 show the results of the jackknife tests which show that *ascii_dem_anf_visbility* has the highest gain for variable importance when run independently for all three jackknife tests, which means it has the most useful information when run by itself. This is the same result found in Section 4.4, which is a further indicator of how critical this variable is to *Spartium junceum* habitat.

Table 8 shows the variable contribution of the BioClim and DEM 10 m layers. The *ascii_dem_anf_visbility* has the highest variable contribution from these input layers. This is consistent with the results of the jackknife tests above and like the results reported in Section 4.4.

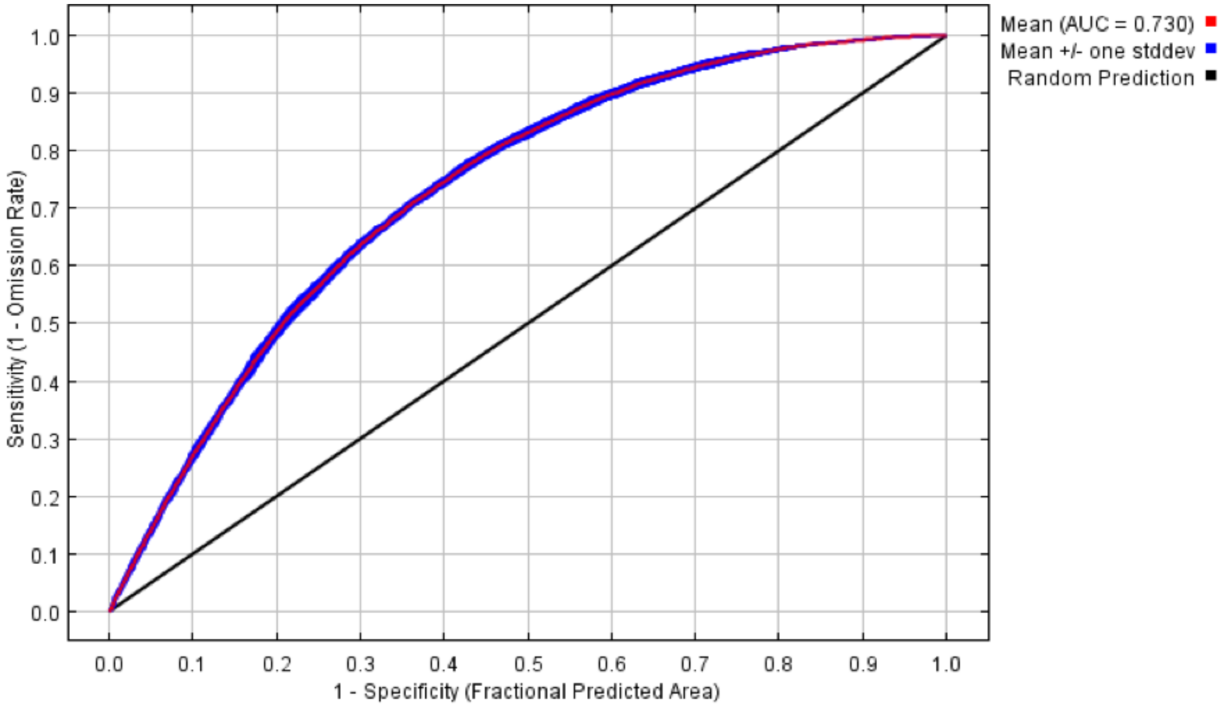


Figure 30 AUC-ROC BioClim 10 m and DEM 10 m variables combined

The BioClim layers were resampled to the 10 m scale with the intention of combining them with the DEM layers in this section in order to strengthen the number and quality of environmental layers being used in Maxent. Resampling the BioClim layers is significant because it allows these layers to be combined in a way that the BioClim 1 km layers cannot while maintaining spatial resolution and integrity. The intention of this is to provide Maxent with a set of variables that produces results from a set environmental layers that are spatially relevant and have a spatial resolution to the study area. The combined BioClim and DEM 10 m layers produce an AUC-ROC value that is higher than all but the BioClim layers at their original 1 km spatial resolution. While the combined layers in this section may have a lower AUC-ROC value than the BioClim 1 km layers this does not necessarily mean that they are less reliable or less preferred. The significance of this argument will be discussed further in Chapter 5.

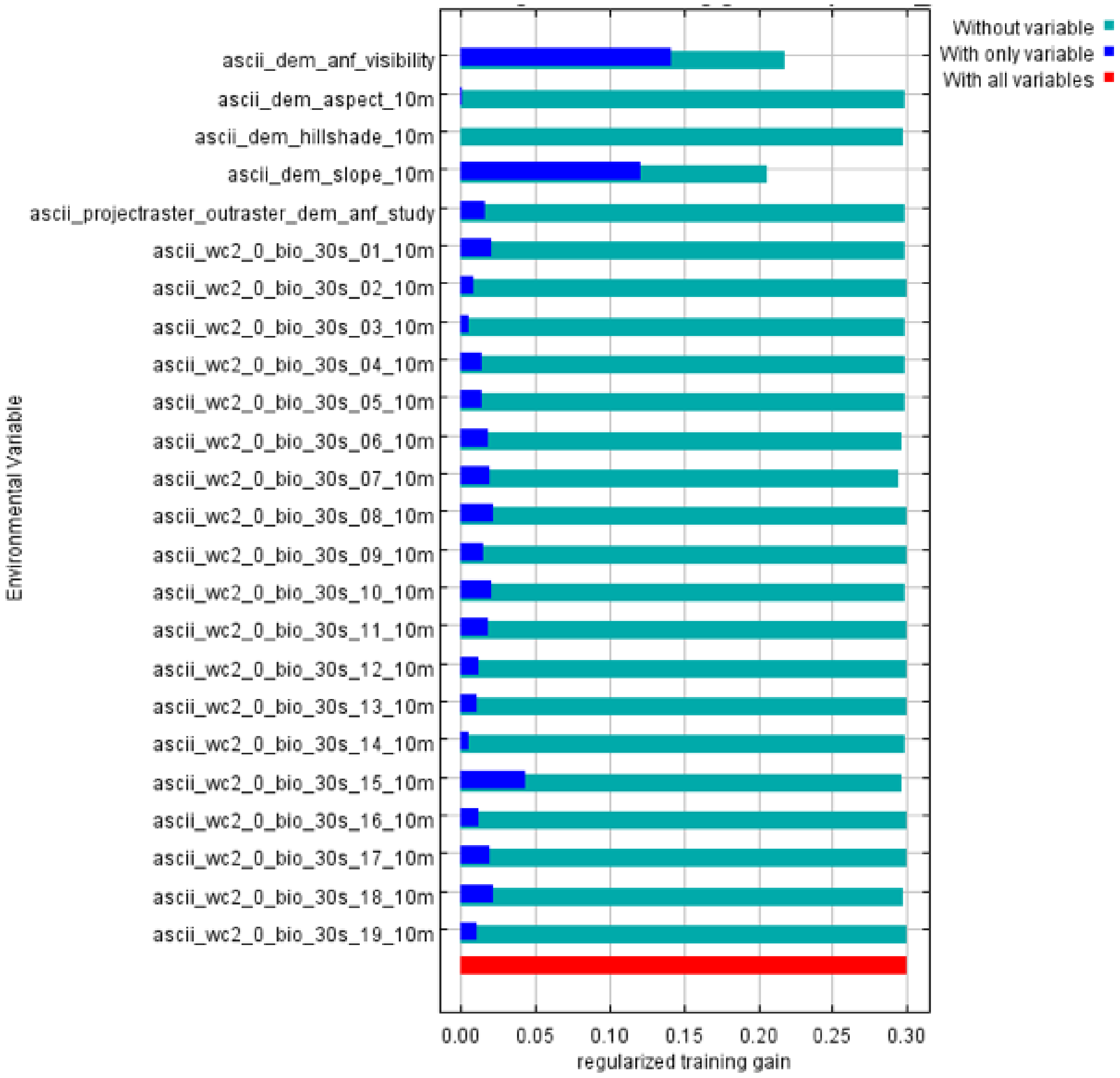


Figure 31 Jackknife of regularized training gain for *Spartium junceum*

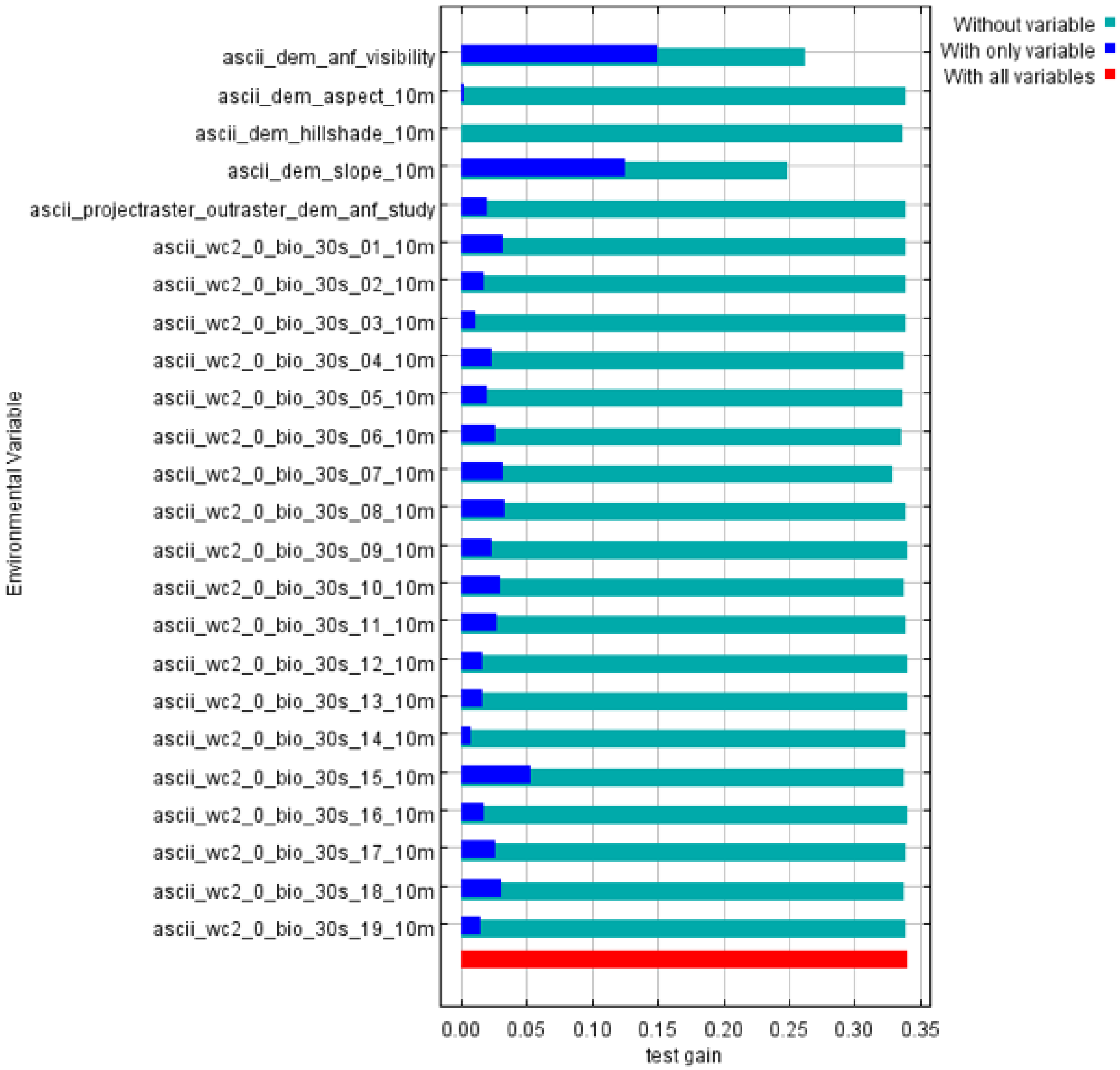


Figure 32 Jackknife of test gain for *Spartium junceum*

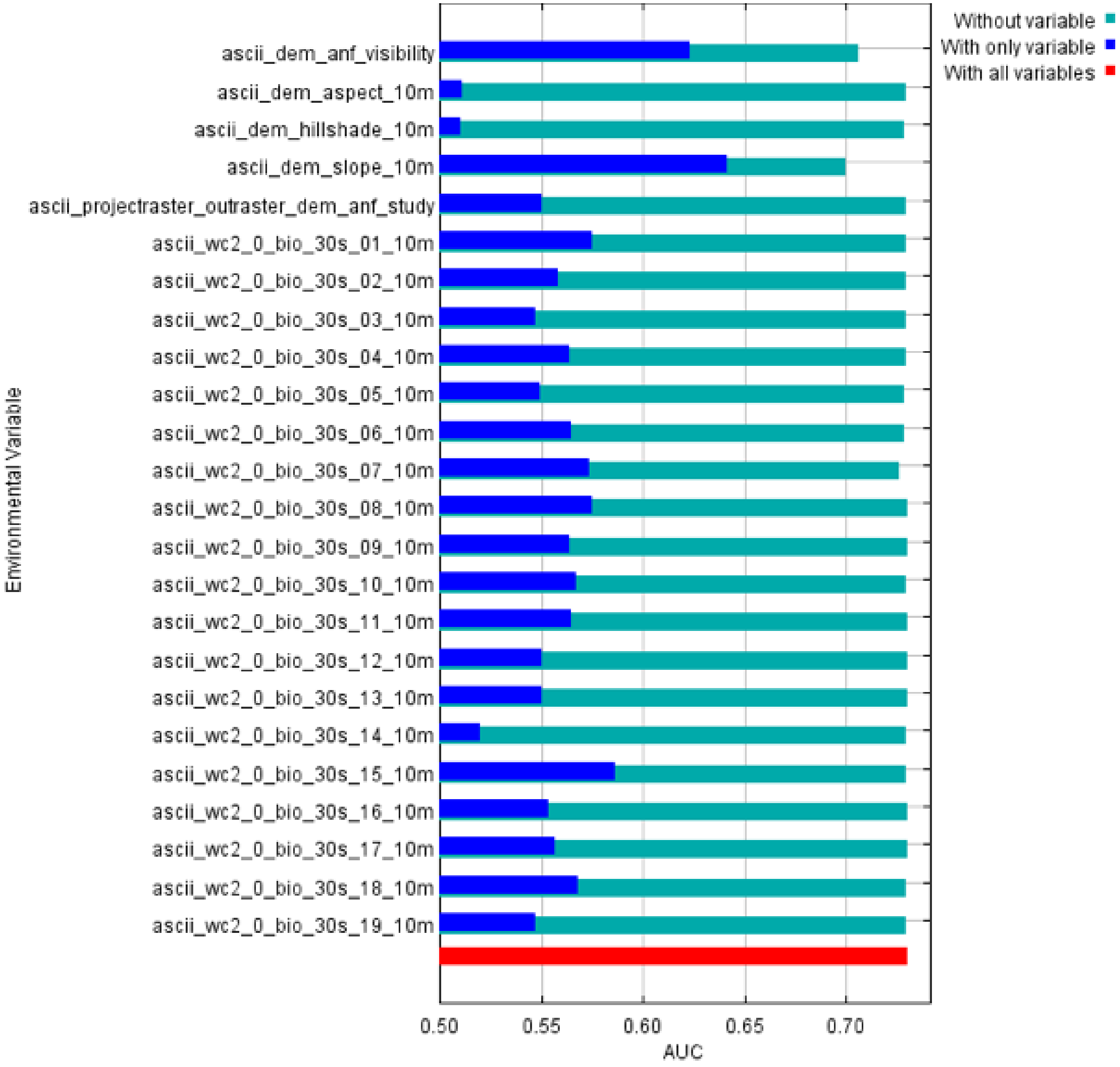
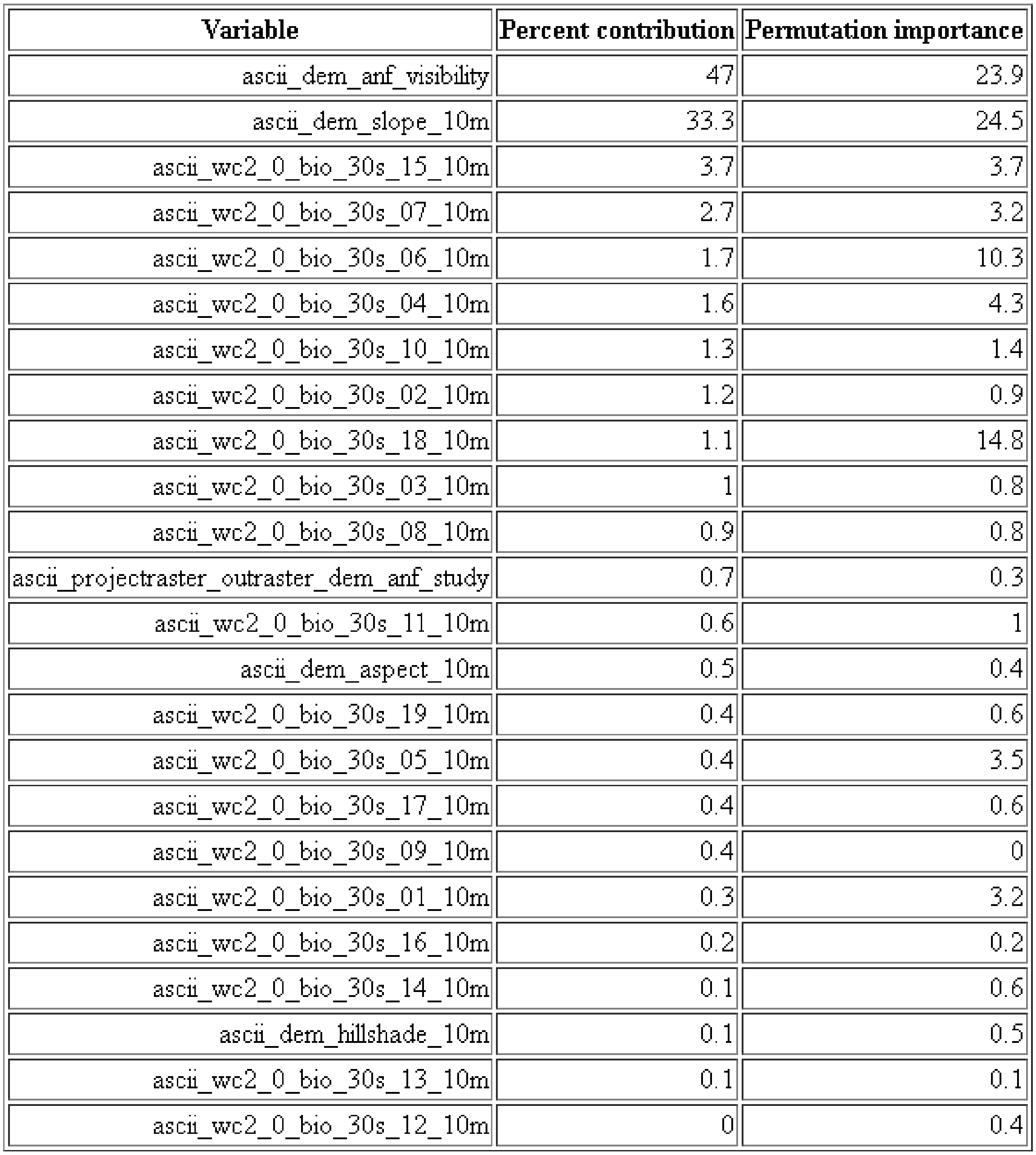


Figure 33 Jackknife of AUC for *Spartium junceum*

Table 8 Variable contributions using combined BioClim and DEM 10 m layers



Variable	Percent contribution	Permutation importance
ascii_dem_anf_visibility	47	23.9
ascii_dem_slope_10m	33.3	24.5
ascii_wc2_0_bio_30s_15_10m	3.7	3.7
ascii_wc2_0_bio_30s_07_10m	2.7	3.2
ascii_wc2_0_bio_30s_06_10m	1.7	10.3
ascii_wc2_0_bio_30s_04_10m	1.6	4.3
ascii_wc2_0_bio_30s_10_10m	1.3	1.4
ascii_wc2_0_bio_30s_02_10m	1.2	0.9
ascii_wc2_0_bio_30s_18_10m	1.1	14.8
ascii_wc2_0_bio_30s_03_10m	1	0.8
ascii_wc2_0_bio_30s_08_10m	0.9	0.8
ascii_projectraster_outraster_dem_anf_study	0.7	0.3
ascii_wc2_0_bio_30s_11_10m	0.6	1
ascii_dem_aspect_10m	0.5	0.4
ascii_wc2_0_bio_30s_19_10m	0.4	0.6
ascii_wc2_0_bio_30s_05_10m	0.4	3.5
ascii_wc2_0_bio_30s_17_10m	0.4	0.6
ascii_wc2_0_bio_30s_09_10m	0.4	0
ascii_wc2_0_bio_30s_01_10m	0.3	3.2
ascii_wc2_0_bio_30s_16_10m	0.2	0.2
ascii_wc2_0_bio_30s_14_10m	0.1	0.6
ascii_dem_hillshade_10m	0.1	0.5
ascii_wc2_0_bio_30s_13_10m	0.1	0.1
ascii_wc2_0_bio_30s_12_10m	0	0.4

4.7 Maxent results using combined 10m BioClim, DEM, and Euclidean Distance ANF Feature layers

The map reproduced in Figure 34 uses the environmental layers from Sections 4.3, 4.4 and 4.5. The spatial resolution of all three sets of environmental layers is 10 m and the underlying values of the individual ASCII files are the same which allowed Maxent once again to run all three layers combined without any issues. As seen in the results reproduced in the five previous sections, most of the suitable habitat is found in the San Gabriel Mountains in the eastern portion of the forest but some suitable habitat occurs in the western portion of the forest. In addition, the results show higher levels of suitability located on south facing aspects of the forest and low suitability on north facing aspects of the forest like the results reported in Section 4.6.

The results are better defined here than in any other section and that is the result of having so many environmental layers available for Maxent to use. Figure 35 shows the AUC-ROC value for the BioClim, DEM, and Euclidean Distance ANF features 10 m layers combined. The model has an AUC of 0.762 and a standard deviation of 0.004 which means it is significant and has the highest AUC-ROC value of all of the Maxent runs, indicating it the best predictor of *Spartium junceum* habitat suitability.

Figures 36-38 show the results of the jackknife tests which show that *anf_roads* have the highest gain when run independently for all three jackknife tests for variable importance, meaning it has the most useful information by itself. The variable that decreases the gain the most when it is omitted is *ascii_dem_slope*, which indicates it has the most information that is not present in any of the other variables. The results from this set of jackknife tests indicate that the proximity of *Spartium junceum* to forest service roads is a key indicator of habitat suitability which is like the results in Section 4.5.

Table 9 shows the variable contributions of the combined BioClim, DEM, and Euclidean Distance ANF features 10 m layers. The *ascii_dem_anf_visbility* has the highest variable contribution at 31.9% which is only slightly higher than the variable contribution of *anf_roads* at 30.1 %. The next highest variable contribution (19.65%) is *ascii_dem_slope and* the remaining variable contributions from the remainder of the layers are 2.4% or less.

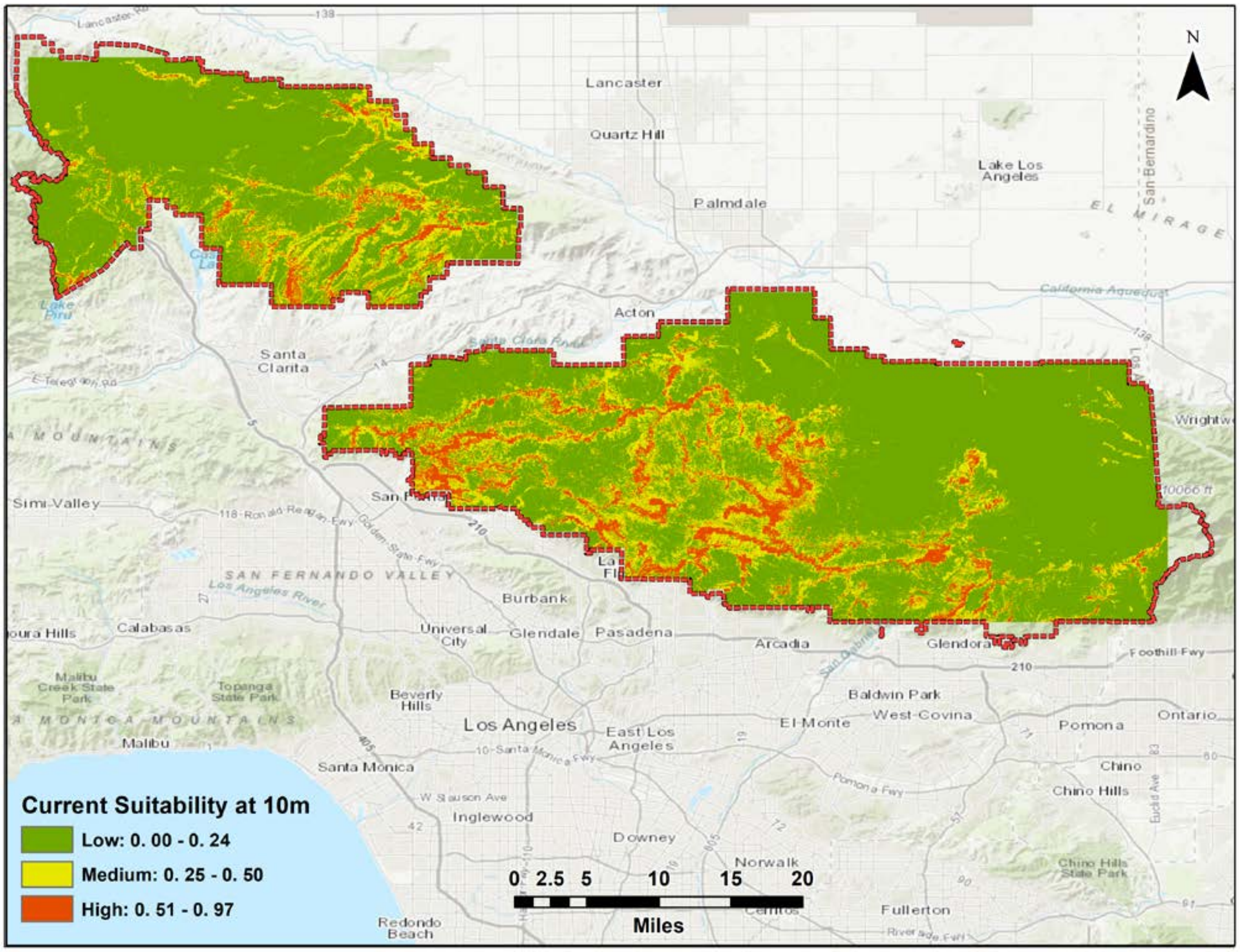


Figure 34 Maxent results using the BioClim, DEM, and Euclidean distance ANF features 10 m layers combined in ANF

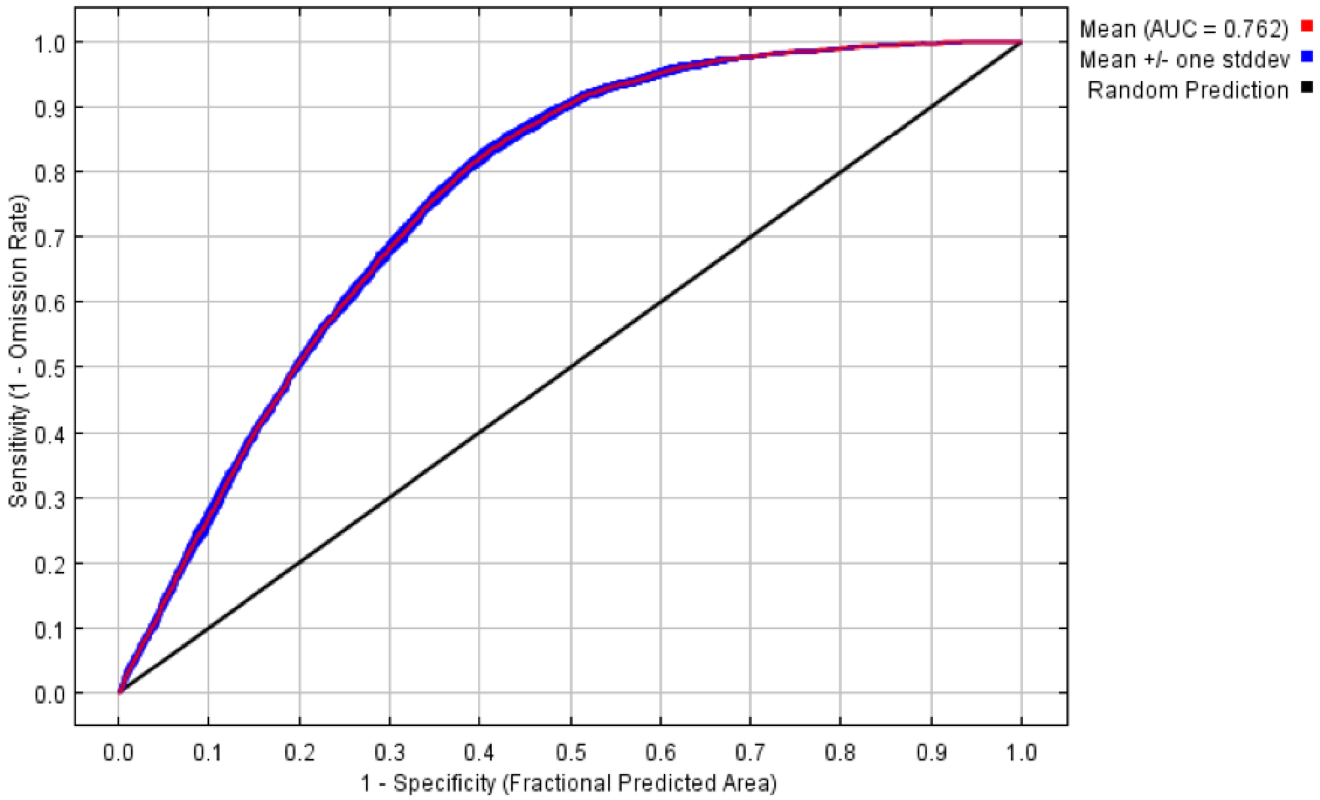


Figure 35 AUC for *Spartium junceum*

The BioClim layers that were resampled at a 10 m scale as described in Section 4.6 have been combined with the Euclidean distance ANF features layer, which were created at the same 10 m scale. The Euclidean distance ANF features were created by first converting the original source vector data into raster data and then using the Euclidean distance tool with a 10 m spatial resolution. This is accomplished with the intention of combining all the layers to provide the most complete set of environmental layers possible for use in Maxent.

The combined BioClim, DEM, and Euclidean distance ANF features layers at a 10 m resolution produce the highest AUC-ROC values of all of the Maxent results in this research and that is possible because all of the layers have the same spatial resolution. The significance of this result is discussed in the next chapter.

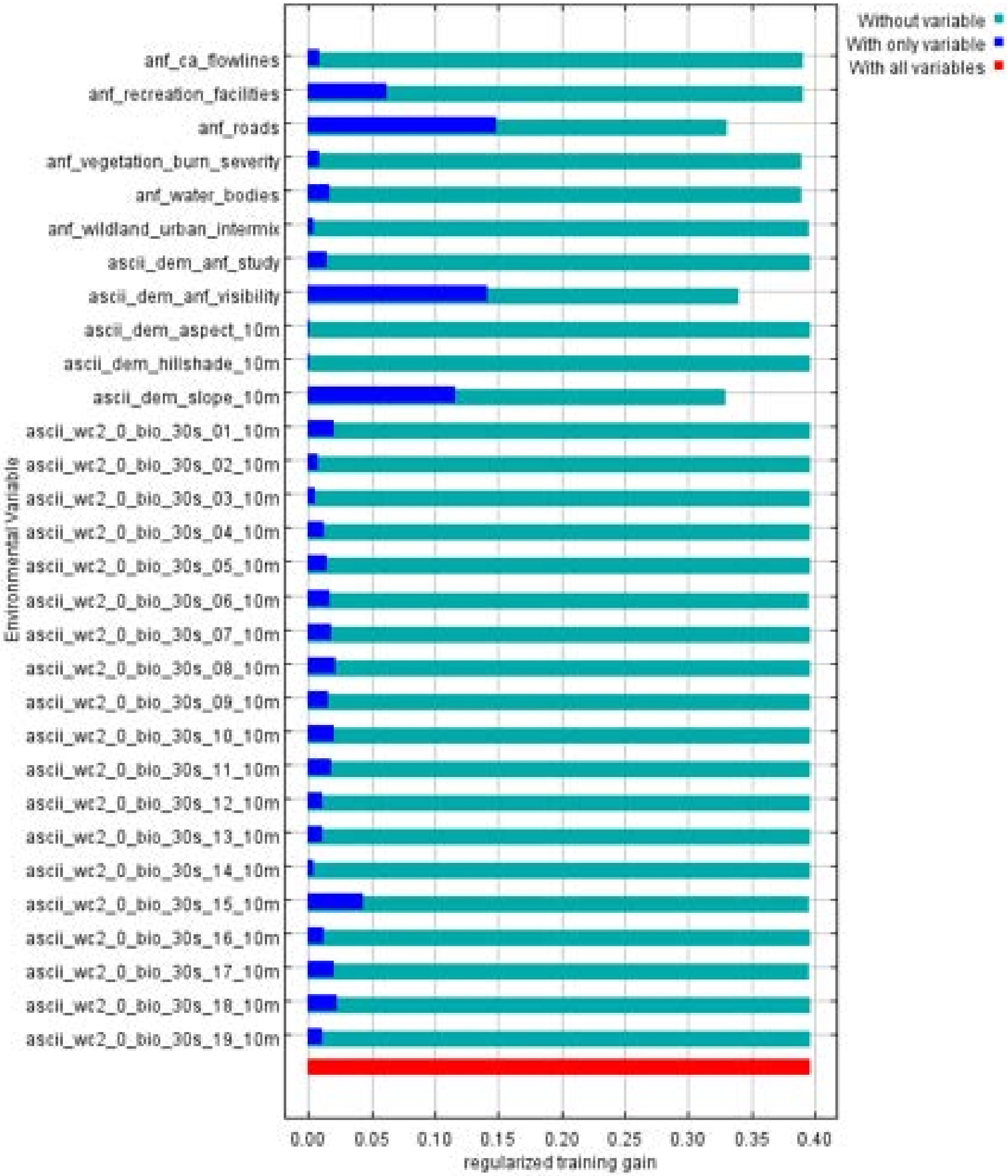


Figure 36 Jackknife of regularized training gain for *Spartium junceum*

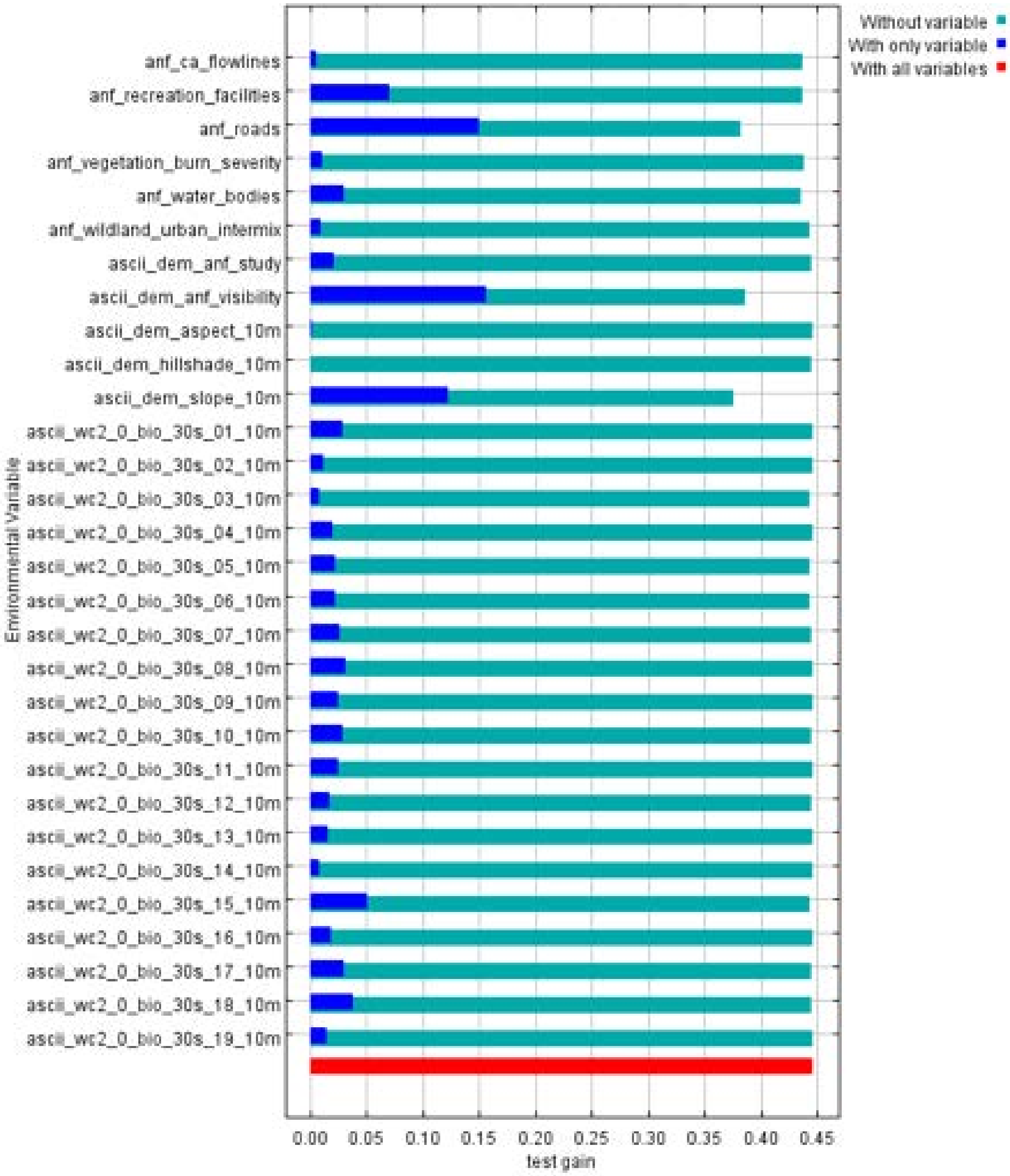


Figure 37 Jackknife of test gain for *Spartium junceum*

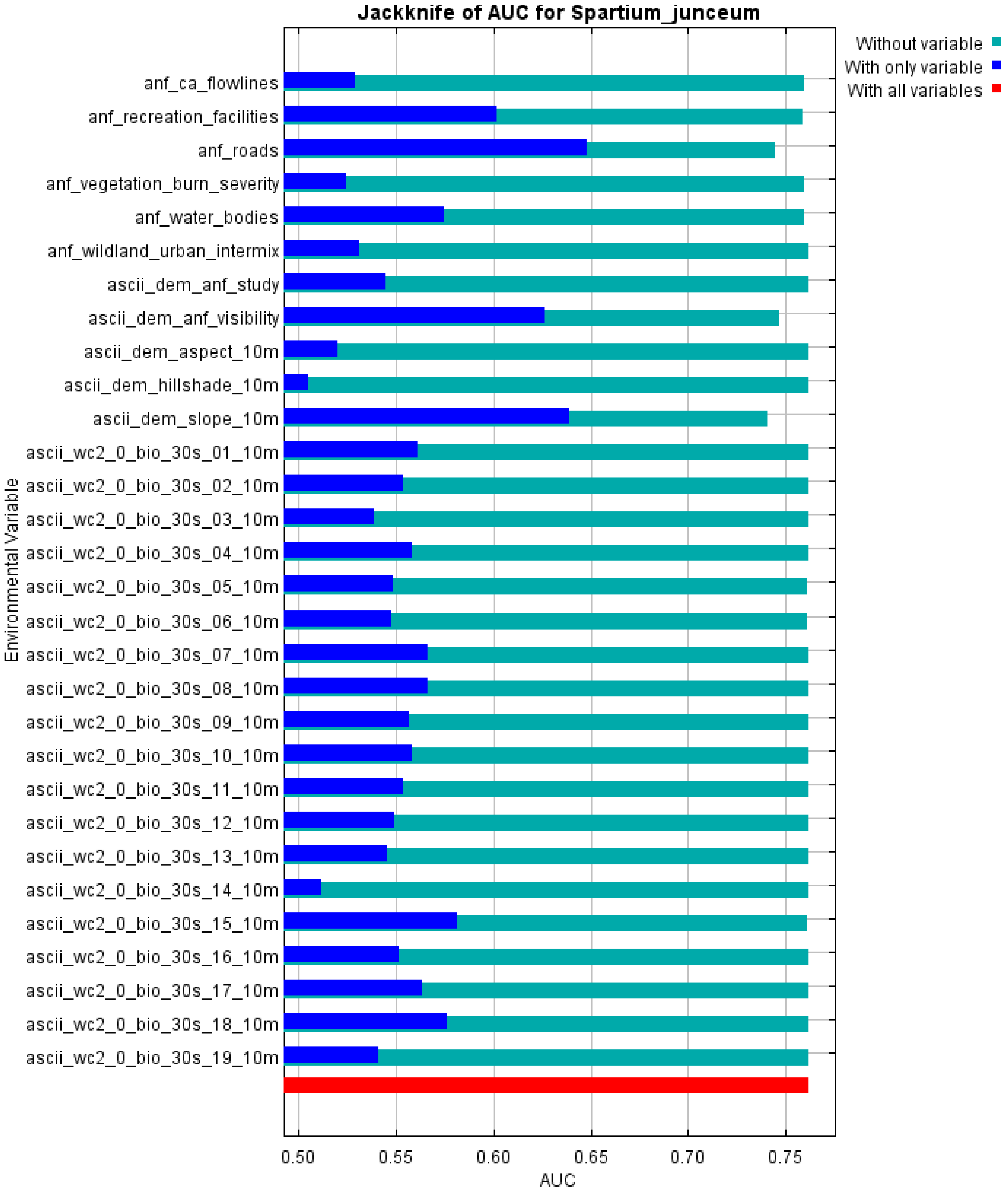
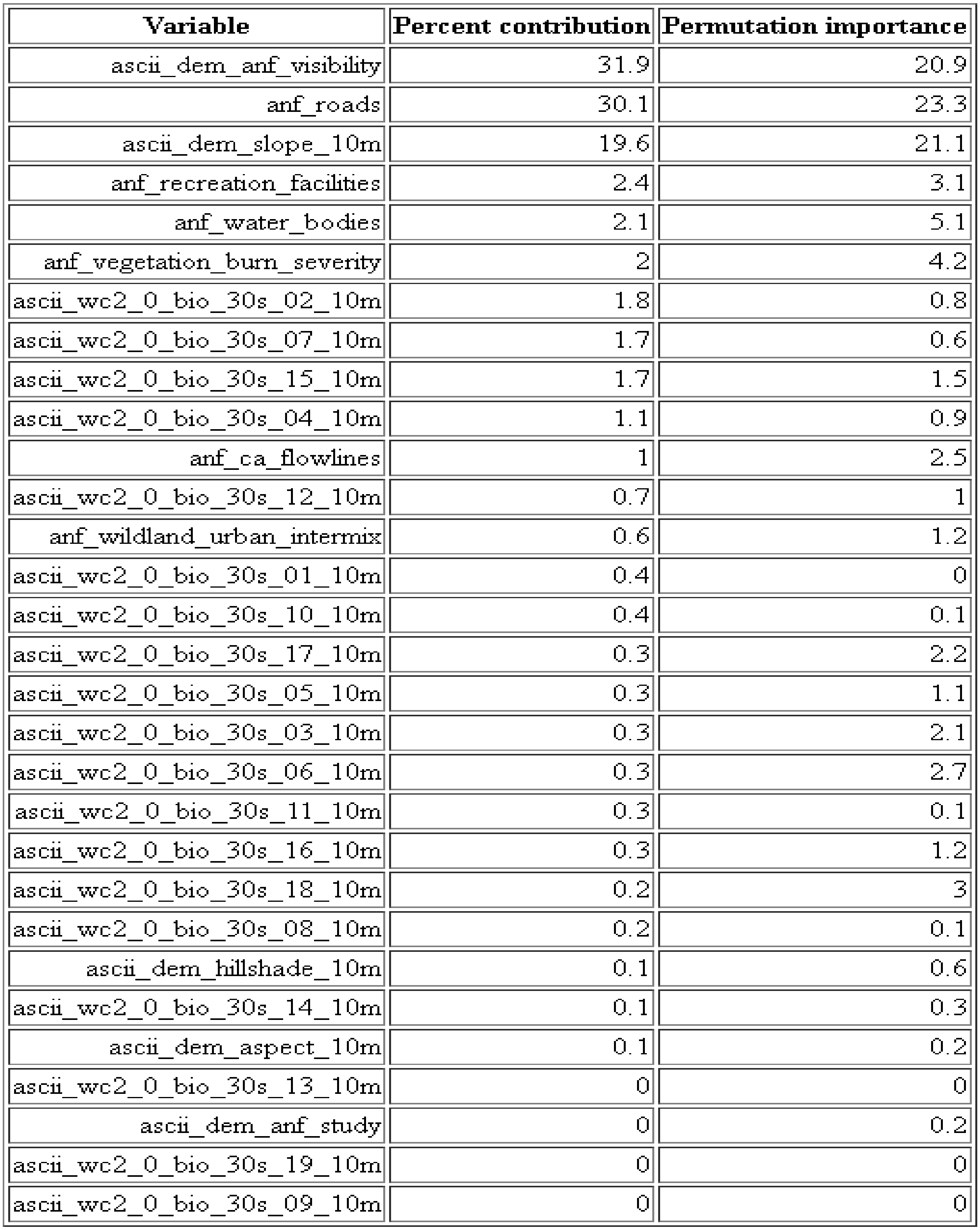


Figure 38 Jackknife of AUC for *Spartium junceum*

Table 9 Variable contributions from the combined BioClim, DEM, and Euclidean Distance ANF features layers at 10 m resolution



Variable	Percent contribution	Permutation importance
ascii_dem_anf_visibility	31.9	20.9
anf_roads	30.1	23.3
ascii_dem_slope_10m	19.6	21.1
anf_recreation_facilities	2.4	3.1
anf_water_bodies	2.1	5.1
anf_vegetation_burn_severity	2	4.2
ascii_wc2_0_bio_30s_02_10m	1.8	0.8
ascii_wc2_0_bio_30s_07_10m	1.7	0.6
ascii_wc2_0_bio_30s_15_10m	1.7	1.5
ascii_wc2_0_bio_30s_04_10m	1.1	0.9
anf_ca_flowlines	1	2.5
ascii_wc2_0_bio_30s_12_10m	0.7	1
anf_wildland_urban_intermix	0.6	1.2
ascii_wc2_0_bio_30s_01_10m	0.4	0
ascii_wc2_0_bio_30s_10_10m	0.4	0.1
ascii_wc2_0_bio_30s_17_10m	0.3	2.2
ascii_wc2_0_bio_30s_05_10m	0.3	1.1
ascii_wc2_0_bio_30s_03_10m	0.3	2.1
ascii_wc2_0_bio_30s_06_10m	0.3	2.7
ascii_wc2_0_bio_30s_11_10m	0.3	0.1
ascii_wc2_0_bio_30s_16_10m	0.3	1.2
ascii_wc2_0_bio_30s_18_10m	0.2	3
ascii_wc2_0_bio_30s_08_10m	0.2	0.1
ascii_dem_hillshade_10m	0.1	0.6
ascii_wc2_0_bio_30s_14_10m	0.1	0.3
ascii_dem_aspect_10m	0.1	0.2
ascii_wc2_0_bio_30s_13_10m	0	0
ascii_dem_anf_study	0	0.2
ascii_wc2_0_bio_30s_19_10m	0	0
ascii_wc2_0_bio_30s_09_10m	0	0

Chapter 5 Discussion and Conclusions

The Maxent results for each run indicate larger areas of suitable habitat than actual surveys of documented locations where *Spartium junceum* has been found at this time. While the results from each run are different, they do reflect an increased habitat suitability range where *Spartium junceum* may spread or may already be colonizing. This reflects a trend that USFS should monitor.

This research helps fill in some of the gaps as it pertains to *Spartium junceum* and invasive species management options in the national forest system by adding to the body of knowledge that is currently available. The results from this research show how SDMs can serve as a very low-cost option for surveying and monitoring invasive species that is worth exploring further in national forests.

The remainder of this chapter is divided into two sections. The first discusses the strengths and weaknesses of the Maxent modeling approach, and the second discusses the opportunities for future work

5.1 Strengths and Weaknesses

Maxent appears to be a capable modeling tool for evaluating invasive species habitat suitability over large areas. The results produced by running the model with different environmental variables showed that the model can handle layers from different sources and at different resolutions without incurring any issues. The model was able to produce results that were consistent in terms of the type of trends that were highlighted. The model's ability to use presence-only data is one of its biggest strengths because there is no real way to accumulate true presence-absence data for this invasive species. There is no real way to survey areas where the species is absent without documenting areas where it is present first and then removing the species and continually monitoring where it pops up again as a control measure for comparison which runs counter to the primary land management goal of removing the species altogether.

The study area for this research was selected at the forest level whereas much of the research that was conducted previously has been at either the regional or the continental level. The bulk of SDM research has used the bioclimatic data layers at a 1 km resolution (Elith e t al., 2010) or layers recorded at a more regionally friendly 30 m resolution (e.g. Vaclavik and Meentemyer, 2009). Maxent was able to handle all of the environmental layers at their different resolutions and produce results that reflected an overall trend of increased habitat suitability. The AUC value for each model run was over 0.60 which means that each was statistically significant with results that are better than random (0.50 or less). There is some debate about how accurate AUC is as a measure of statistical significance in SDMs but the variable results from the DEM layers and the Euclidean Distance ANF features layers are consistent with what was expected to be the most influential factors in predicting suitable habitat (i.e., proximity to forest service roads, elevation, slope and visibility).

There were no real challenges in terms of obtaining the sample data from the Forest Service as I was able to obtain the data directly while I was employed by them as an intern when this thesis project research was launched. However, there were significant issues with the format of the sample data, the introduction of uncertainty from using the Create Random Points tool in ArcMap, the collection methods of the sample data, sampling bias, and the resolution of the environmental variables from the BioClim dataset.

The sample data and x, y coordinates associated with them were in polygon form and needed to be converted into point form in ArcMap using the Create random points tool. The results from that step produced a large dataset that needed to be filtered by selecting only those points that fell within the original recorded polygon locations and eliminating all those that did not. Using the Create Random Points tool was a necessary step in order to get the sample data into a usable format in Maxent but as a result almost certainly would have introduced uncertainty into the data set. That is because the original source polygons included numerous 10 m grid cells and therefore the random points produced by the tool may have acquired one of a variety of values for one or more of the pertinent co-variates.

There were 1,285 records in the original *Spartium junceum* polygon file and 6,245,000 records after running the create random points tool. This large number was necessary to ensure that enough point features fell within every polygon in the study area. That number was reduced to 222,614 after selecting point features that fell within the *Spartium junceum* polygons from the source feature class. The two important things to understand is that the number of records substantially increased because of the need to have individual x, y coordinates for each occurrence of *Spartium junceum* and that uncertainty was almost certainly introduced into the dataset as a result. Each instance of a point feature that was assigned an x, y coordinate pair corresponded to a location where *Spartium junceum* should be found but did not necessarily reflect the amount of plants actually found at each location (introduced uncertainty) nor the number of coordinate pairs that were required to define the entire polygon boundary from the source file.

Data collection methods serve as a major limitation for Maxent because the sample data need to have one x, y coordinate pair per record for the model to be run. The ANF does not currently have a clear and established protocol for capturing presence data in the field that relies on point sampling which limits the confidence in the accuracy and truth of the data. Given this lack of confidence, we cannot be certain how many actual plants are found at an actual location where a polygon boundary was created, nor can we be certain of the actual coordinate pair count that would define the entire boundary for each polygon with sample data. Figure 39 illustrates this point.

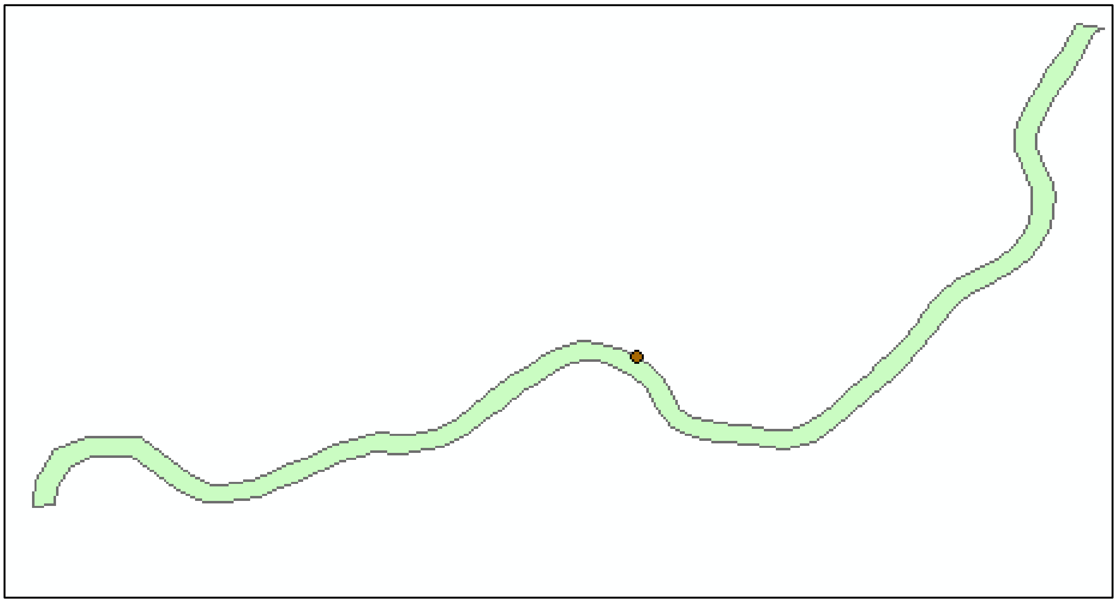


Figure 39 *Spartium junceum* polygon data and point data after conversion

There also existed a realistic and likely possibility that the *spartium junceum* data included sampling bias due the increased likelihood that mapped instances where *spartium junceum* was found was closer to Forest Service roads and thus more likely to be included in the USFS dataset than those instances of *spartium junceum* that were further away from the Forest Service road network.

There were some difficulties when working with the BioClim data due to the resolution at which it was recorded, and the efforts needed to resample the data. There was significant loading time required to build the pyramids needed to display the BioClim raster data in ArcMap and significant processing time needed to clip the raster to the study boundary. This proved to be a very limiting factor in terms of using the model because any changes or errors in the final versions of the rasters necessitated reloading and reprocessing of the files, which required substantial time. The time required increased significantly when the BioClim data was resampled at a 10 m spatial resolution as that required the tool to create substantially more pixels than were in the original file.

Running Maxent with the higher resolution raster files takes substantial memory and required more than the 524 MB that is preset in the Maxent bat File. The available RAM was increased to 2,096 MB to allow Maxent to run using the larger raster files without any issues. This is an important factor to consider when working with data on thumb or hard drives which will be used on multiple computers or machines. The amount of RAM available will need to be the same on all machines that run the software. Systems resources are a crucial factor when working with Maxent and that is something the USFS will need to consider if they want to use the software even if on a trial basis only. Furthermore, storage space is a critical issue that should be considered. The outputs from running the model for this thesis project using 15 replications consumed 80 GB of storage space. If larger files are used or more replications are needed or desired this can become a significant factor that would need to be accounted for.

5.2 Interpretation of Maxent results for each set of layers

When looking at the results from running Maxent with each set of environmental layers it is important to consider two major factors, the resolution at which the data was captured and the overall trend of the results.

The overall trend from each Maxent run is the increase in the areas that can be deemed as suitable habitat for *Spartium junceum*; however, there were significant differences in the actual size of those areas predicted by the various runs. This can be attributed to the resolution at which the data was recorded and the appropriateness of that data relative to the size of the study area. The spatial resolution of the BioClim layers is 1 km, which is the equivalent of 247.11 acres in a study area spanning 655,387 acres. The resolution does not seem to be as much of an issue when looking at the entire forest, but it was when selection was focused on the area occupied by *Spartium junceum*. At this resolution, it takes just only 12.8 cells of BioClim 1 km data to cover the entire *Spartium junceum* habitat. The BioClim data were recorded at the global scale to track global climate trends. The BioClim layers are better suited for predicting habitat suitability across a larger study area such as the entire national forest system rather than one forest by itself.

To account for the problems caused by the spatial resolution the original BioClim layers were resampled in ArcMap to 10 m to match the DEM layers which were recorded at that spatial resolution. At this resolution, a 10 m cell is the equivalent of 0.00247 acres which is well suited to the thesis research given the total area occupied by *Spartium junceum* (3,168 acres). When looking at the results from Figures 9 and 14 the predicted areas of habitat suitability are smoother and more defined in the resampled set of layers than in the original raster layer and match the habitat area in the eastern and western parts of the forest better than the original BioClim layers.

The DEM environmental layers seem to be the best suited for the study area and that is most likely because they were initially recorded at a 10 m spatial resolution. The study area is relatively small, and the DEM layers are better suited to account for the sizes of both the study area and the area occupied by *Spartium junceum*. The results from running the DEM layers with Maxent reflect this as they show habitat suitability areas that are much more detailed than those produced with the BioClim layers. The AUC value 0.707 shows strong statistical significance and reflects how elevation, slope, aspect and visibility tend to influence fauna in various ways in forest ecosystems. The AUC values for the BioClim variables and resampled BioClim variables were 0.751 and 0.664, which offered justification for combining layers from those pair of sources. It is important to emphasize that each cell or pixel in the original BioClim layers accounts for 1/13 of the sample species occurrence area, which is misleading with such a small study area and does raise questions about the efficacy of the 0.751 AUC value in this instance.

Maxent was run a fourth time with a set of vector data feature classes that were converted into raster data using the Euclidean Distance tool in order to indicate the proximity of those geographic features to each other. The reason for using this tool was to account for spatial autocorrelation and to indicate the influence that the proximity of each layer has on the sample data. The results in Figure 24 look very similar to the sample species locations shown in Figure 9. The suitable habitat area is smaller and more delineated in this instance and that can be attributed to the underlying vector data being the most detailed and spatially relevant in terms of the scale at which it was recorded. The AUC value when running Maxent with the Euclidean Distance rasters was 0.692 which means it is statistically significant but not as significant as the results generated using the DEM layers or the BioClim 1 km layers. This can be attributed to the fact that the data in these layers were originally recorded as vector features and needed to be converted to raster features and then Euclidean distances. This major factor was considered when performing the final Maxent run with the combination of all three of the 10 m spatial resolution layers.

The fifth Maxent run combined the BioClim and DEM 10 m layers. This was done because both layers were either recorded or converted to 10 m spatial resolution and had AUC values that were significant. The AUC value for this run was 0.730, which was higher than the AUC values produced when the BioClim and DEM layers when run on their own. This AUC value reflects a high degree of statistical significance indicating that the result from running Maxent with these two data sets is significantly better than random. This can be attributed to the strength of combining both layers together to provide a stronger understanding of the many variables at play that influence habitat suitability. The variable contribution percentages are critical indicators of which variables play the largest role in habitat suitability and hence potential future occupancy. The two biggest contributors from each data set are in line with some of the key variables found in the historical background information for this species (Zouhar, 2005). Section 5.2.1 discusses the variable contributions in a little more depth.

The sixth and final Maxent run incorporates the BioClim 10 m layer, the DEM 10 m layers, and the Euclidean Distance ANF features layers. This was done because all layers were either recorded or converted to 10 m spatial resolution and had AUC values that were significant. The AUC value for this run was 0.762, which was the highest AUC value produced in all of the Maxent runs. This can be attributed to two primary factors. The first being that in combining the three layers a robust final layer set was created which can better account for the many variables and factors at play that influence habitat suitability. The second factor is that in combining the three different layers some the inconsistencies of a given individual layer, such as a source being a vector layer, or a layer being recorded at a different spatial resolution, can be accounted for by combining the three sets of layers.

The AUC values were used as the primary means of evaluating how well a model performed at predicting habitat suitability. The AUC is “a common metric for assessing the predictive ability and hence utility of a habitat suitability model” (Glover-Kapfer (2015). The AUC assesses whether model predictions are better than random by interpreting the value scores on a 0 to 1 scale. AUC values with a score of 1 indicate the model predicts presence perfectly and a value of 0.5 indicates model predictions are equivalent to random guesses (Philips et al., 2006). Interpreting the results from the six Maxent runs indicates that all models are better than random and the BioClim, DEM and ANF features 10 m combination layers present a good level of performance in terms of habitat suitability. Although there is some criticism of using AUC as the primary means of evaluating how well SDMs such as Maxent perform (e.g., Lobo, Jimenez-Valverde & Real, 2008) it still "remains the most common means of measuring model performance to date" (Glover-Kapfer, 2015)

5.2.1 Interpretation of Variable Results

The relevancy of the BioClim variables in this study needs to be carefully assessed. The original BioClim variables were recorded at a scale that is too large to be considered a good predictor of suitable habitat. This is because the resolution reflects large-scale climate trends rather than localized weather factors, which are more appropriate for the size of the study area. Resampling the BioClim data did help adjust the resolution and make the data more applicable to the study region but did not necessarily represent localized weather or climate as well as it could have had it been modeled at a 10 m resolution at the onset. It is worth mentioning that "precipitation seasonality" had the highest variable contribution in Maxent for both the original and resampled BioClim data. This could indicate that even though the data was resampled it is likely that it still does not accurately depict the local climate or microclimate expected or experienced in the study area.

The most reliable indicators of habitat suitability in this study are the DEM and Euclidean Distance ANF features layers because they were recorded at a scale that is appropriate for the study area. The largest variable contributors to the DEM layers are *ascii_dem_anf_visbility* at 56% and *ascii_dem_slope_10 m* at 39.1%. The largest variable contributors for the Euclidean distance ANF features layer are *anf_roads* at 70% and *and_recreation_facilities* at 15.8%. For the combined the BioClim and DEM 10 m layer data set, the two highest variable contributions are *ascii_dem_anf_visbility at* 47% and *ascii_dem_slope_10 m* at 33.3%. The next highest variable contribution came from *ascii_wc2_0_bio_30s_15* (precipitation seasonality) at 3.7%. This is a significant result because this was the same variable that had the highest variable contribution in both BioClim Maxent runs. This is another indication that BioClim on its own may not be a good indicator of habitat suitability at the local scale.

In the combined BioClim, DEM, and Euclidean Distance ANF features combined 10 m layer dataset, the top three variable contributors are *ascii_dem_anf_visbility at* 31.9%, *anf_roads* at 30.1%, and *ascii_dem_slope_10 m* at 19.6%. These three variables account for 81.6% of the total variable contribution while the remaining 27 variables account for only 18.4%. The top three variables come from the DEM and Euclidean distance ANF features layers while 19 of remaining 27 variables come from the BioClim data. This is another strong indication that despite the high AUC values found in Sections 4.2 and 4.3, the results are not as a strong an indicator as one might initially think. This further supports the merits of creating a more comprehensive variable dataset by combining the three 10 m layer data sets. The top three variables can be considered the most critical elements that need to be present when assessing an area for *Spartium junceum* habitat suitability. The remainder of the variables should be considered but not with the same weight as the top three. Furthermore, while the BioClim variables should not be relied on in this case, they do represent a key set of variables that could and should be incorporated in future models if they are captured at finder resolutions than 1 km.

5.3 Opportunities for future research and Model improvement

The overall goal of this study was to test the feasibility of using Maxent as a tool for aiding land management agencies such as the USFS in their invasive species management plans. There is no real cost associated with using the software as it is free to use and comes with tutorial information. This makes Maxent a great tool for land management use but based on the results of this research there are a few more steps that need to be taken before Maxent can be confidently used in everyday land management.

While this research did demonstrate that with presence-only data and some initial reading and training, agencies can easily get started with Maxent, it also revealed some issues that need to be addressed before confidently using Maxent as an everyday tool.

The USFS should work to establish clear protocols for capturing presence data in the field that rely on point sampling to avoid some of the aforementioned problems. Additionally, the USFS can ensure the enforcement of a standardized workflow for data entry into invasive species data dictionaries. Furthermore, one or more environmental variables focused on geology and/or soil should be added to Maxent to account for *spartium junceum's* characteristically large root system that enables it to penetrate substrate which can promote long life and further spreading (invasion) to new habitat areas. Lastly, resampling of the BioClim data represents another key area for improvement.

Resampling of the BioClim data can be improved by incorporating the use of a more precise resampling scheme that includes some form of interpolation such as block kriging or by locating regional weather and/or climate data sets that were recorded at finer resolutions. There are a large number of studies that now utilize the ANUCLIM software (Xu and Hutchinson, 2013) to generate finer scale representations of climate (e.g. solar radiation, precipitation, daily maximum temperature, isothermality), which can be used to improve on the methods and results of this research.

There is no shortage of opportunities for future research with the abundance of different plant and animal species found throughout the forest. Maxent is not only applicable to invasive species management but also to other areas such as native species management, forest health assessment and forecasting, vector born disease prediction, and regional weather and climate forecasting (Antoine and Thuiller, 2005). The USFS could benefit greatly from incorporating SDMs such as Maxent into their everyday land management plans but this research only serves as a starting point for having those conversations. The issues of sampling bias introduced uncertainty from using the Create Random Points tool, data collection and data entry methods, incorporation of substrate variables, and selecting appropriately scaled climate data would all need to be addressed first before Maxent would be ready for everyday use.

REFERENCES

Adjemian, Jennifer C Z, Girvetz, Evan H, Beckett, Laurel, and Foley, Janet E. "Analysis of Genetic Algorithm for Rule-Set Production (GARP) Modeling Approach for Predicting Distributions of Fleas Implicated as Vectors of Plague, Yersinia Pestis, in California." *Journal of medical entomology* 43, no. 1 (January 2006): 93–103.

Bedia, J., J. Busqué, and J.M. Gutiérrez. "Predicting Plant Species Distribution across an Alpine Rangeland in Northern Spain. A Comparison of Probabilistic Methods." *Applied Vegetation Science*14, no. 3 (2011): 415-32. Accessed Feburary 2, 2018. Doi:10.1111/j.1654-109x.2011.01128.x.

CAL-IPC. Match 9, 2004. "Plant Assessment Form *Spartium junceum*." Accessed December 13, 2017. https://www.cal-uipc.org/plants/profile/spartium-junceum-profile/

CDFA. n.d. "Encyclopedia: Weed Ratings" Accessed January 06, 2018. https://www.cdfa.ca.gov/plant/IPC/encycloweedia/weedinfo/winfo_table-sciname.html

Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions*17, no. 1 (2011): 43-57. Doi:10.1111/j.1472-4642.2010.00725.x.

Evangelista, Paul H., Sunil Kumar, Thomas J. Stohlgren, Catherine S. Jarnevich, Alycia W. Crall, John B. Norman Iii, and David T. Barnett. "Modelling Invasion for a Habitat Generalist and a Specialist Plant Species." *Diversity and Distributions* 14, no. 5 (2008): 808-17. Doi:10.1111/j.1472-4642.2008.00486.x.

Glover-Kapfer, Paul. 2015. "A training manual for habitat suitability and connectivity modeling using tigers (Panthera tigris) in Bhutan as example". 10.13140/RG.2.2.34804.86409. Accessed August 26, 2019.

Hanley, J. A., and B. J. McNeil. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology*143, no. 1 (1982): 29-36. Doi:10.1148/radiology.143.1.7063747.

Hoffman, Justin D., Sunil Narumalani, Deepak R. Mishra, Paul Merani, and Robert G. Wilson. "Predicting Potential Occurrence and Spread of Invasive Plant Species along the North Platte River, Nebraska." *Invasive Plant Science and Management*1, no. 04 (2008): 359-67. Accessed January 12, 2019. Doi:10.1614/ipsm-07-048.

Kemp, Karen. 2012."The Hawai`i Island Crop Probability Map: An Update of the Crop Growth Parameters for the Hawai`i County Crop Model." Accessed December 18, 2018. https://spatial.usc.edu/wp-content/uploads/2017/09/CropProbabilityMap-Final-Report.pdf

Lobo, J., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151.Accessed June 12, 2018. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Mackenzie, Darryl I., and Royle, J. Andrew. “Designing Occupancy Studies: General Advice and Allocating Survey Effort.” *Journal of Applied Ecology* 42, no. 6 (December 2005): 1105–1114.

Merow, Cory, Smith, Matthew J., and Silander, John A. “A Practical Guide to MaxEnt for Modeling Species’ Distributions: What It Does, and Why Inputs and Settings Matter.” *Ecography* 36, no. 10 (October 2013): 1058–1069.

Mullin, Barbara H., Lars W.J. Anderson, Joseph M. DiTomaso, Robert E. Eplee, and Kurt D. Getsinger. 2000. Invasive Plant Species. *Council for Agricultural Science and Technology*, no. 13. Accessed September 14, 2016. http://www.iatp.org/files/Invasive_Plant_Species.html

Nickerman, Janet., and Joanna Clines. September 2009. “Noxious and Invasive, non-native weeds Specialist Report and Noxious Weed Detection Survey Plan.” Accessed December 13, 2017. https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb5167066.pdf

Pearson, Richard G., Raxworthy, Christopher J., Nakamura, Miguel, and Townsend Peterson, A. “ORIGINAL ARTICLE: Predicting Species Distributions from Small Numbers of Occurrence Records: a Test Case Using Cryptic Geckos in Madagascar.” *Journal of Biogeography* 34, no. 1 (January 2007): 102–117.

Pepe, Margaret S., Holly Janes, and Jessie Wen Gu. “Letter by Pepe Et Al Regarding Article, “Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction”.” *Circulation*116, no. 6 (2007). Accessed October 23, 2018 doi:10.1161/circulationaha.107.709253.

Phillips, S. J. “A Brief Tutorial on Maxent.” 2017. Accessed on October 27, 2018. http://biodiversityinformatics.amnh.org/open_source/maxent/.

Phillips, Steven J., and Dudík, Miroslav. “Modeling of Species Distributions with Maxent: New Extensions and a Comprehensive Evaluation.” *Ecography* 31, no. 2 (April 2008): 161–175.

Phillips, Steven J., Miroslav Dudík, and Robert E. Schapire. “A Maximum Entropy Approach to Species Distribution Modeling.” *Twenty-first International Conference on Machine Learning – ICML 04*, 2004. Doi:10.1145/1015330.1015412.

Phillips, Steven J., Anderson, Robert P., and Schapire, Robert E. “Maximum Entropy Modeling of Species Geographic Distributions.” *Ecological Modelling* 190, no. 3 (2006): 231–259.

Qin, Aili, Bo Liu, Quanshui Guo, Rainer W. Bussmann, Fanqiang Ma, Zunji Jian, Gexi Xu, and Shunxiang Pei. “Maxent Modeling for Predicting Impacts of Climate Change on the Potential Distribution of Thuja Sutchuenensis Franch., an Extremely Endangered Conifer from Southwestern China.” *Global Ecology and Conservation*10 (2017): 139-46. Doi:10.1016/j.gecco.2017.02.004.

Radeloff, Volker C., David P. Helmers, H. Anu Kramer, Miranda H. Mockrin, Patricia M. Alexandre, Avi Bar-Massada, Van Butsic, Todd J. Hawbaker, Sebastián Martinuzzi, Alexandra D. Syphard, and Susan I. Stewart. “Rapid Growth of the US Wildland-urban Interface Raises Wildfire Risk.” PNAS. March 27, 2018. Accessed October 13, 2018. http://www.pnas.org/content/115/13/3314.

Stein, Susan M., James P. Menakis, Mary A. Carr, Sara J. Comas, Susan I. Stewart, Helene Cleveland, Lincoln Bramwell, Volker C. Radeloff. “Wildfire, wildlands, and people: understanding and preparing for wildfire in the wildland-urban interface—a Forests on the Edge report. January 2013. Accessed September 30, 2017. https://www.fs.fed.us/openspace/fote/reports/GTR-299.pdf

Thuiller, Wilfried, Araújo, Miguel B., and Lavorel, Sandra. “Generalized Models Vs. Classification Tree Analysis: Predicting Spatial Distributions of Plant Species at Different Scales.” *Journal of Vegetation Science* 14, no. 5 (October 2003): 669–680.

University of California. n.d. “Mediterranean Climate.” UC Rangelands Archive. Accessed January 11, 2019. http://rangelandarchive.ucdavis.edu/Annual_Rangeland_Handbook/Mediterranean_Climate/.

US Census Bureau. 2017. “ACS Demographics and Housing Estimates”. Accessed September 28, 2017. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

USDA Forest Service. February 2018. “FY 2019 Budget Justification.” Budget & Performance. Accessed January 5, 2019. https://www.fs.fed.us/sites/default/files/usfs-fy19-budget-justification.pdf

USDA Forest Service. May 2017. “Fiscal Year 2018 Budget Overview”. Accessed October 6, 2017. https://www.fs.fed.us/sites/default/files/usfs-fy18-budget-overview.pdf

USDA Forest Service. August 2013. “Forest Service National Strategic Framework for Invasive Species Management.” Accessed January 10, 2019. https://www.fs.fed.us/foresthealth/publications/Framework_for_Invasive_Species_FS-1017.pdf

USDA Forest Service. n.d. “History & Culture.” Accessed October 13, 2018. https://www.fs.usda.gov/main/angeles/learning/history-culture.

USDA Forest Service. January 2012. “Land Areas of the National Forest System”. Accessed September 25, 2016. http://www.fs.fed.us/land/staff/lar/LAR2011/LAR2011_Book_A5.pdf.

USDA Forest Service. 2001. “National Visitor Use Monitoring Results August 2001 USDA Forest Service Region 5.” Accessed September 25, 2016. https://www.fs.fed.us/recreation/programs/nvum/reports/year1/R5_Angeles_final.%09htm#_Toc522596885

Václavík, Tomáš, and Ross K. Meentemeyer. “Invasive Species Distribution Modeling (iSDM): Are Absence Data and Dispersal Constraints Needed to Predict Actual Distributions?” *Ecological Modelling* 220, no. 23 (2009): 3248-258. doi:10.1016/j.ecolmodel.2009.08.013.

Vincent, Carol Hardy. 2004. “Federal Land Management Agencies: Background on Land and Resources Management”. Accessed September 30, 2017. http://nationalaglawcenter.org/wp-content/uploads/assets/crs/RL32393.pdf.

Warner, Peter J., Carla C. Bossard, Matthew L. Brooks, Joseph M. DiTomaso, John A. Hall, Ann M. Howald, Douglas W. Johnson, John M. Randall, Cynthia L. Roye, Maria M. Ryan, and Alison E. Stanton. “Criteria for Categorizing Invasive Non-Native Plants that Threaten Wildlands.” 2003. Accessed January 10, 2019.

1.

Xu, Tingbao, Hutchinson, Michael F. New developments and applications in the ANUCLIM spatial climatic and bioclimatic modelling package. *Environmental Modelling and Software*. 2013;40(C):267-279. doi:10.1016/j.envsoft.2012.10.003

Young, Nick, Lane Carter, Paul Evangelista. “A MaxEnt Model v3.3.3e Tutorial (ArcGIS v10).” 2011. Accessed February 17, 2019. http://ibis.colostate.edu/webcontent/ws/coloradoview/tutorialsdownloads/a_maxent_model_v7.pdf

Zou, Kelly H., A. James O’Malley, and Laura Mauri. "Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models." *Circulation*115, no. 5 (2007): 54-57. Accessed January 10, 2019. doi:10.1161/circulationaha.105.594929.

Zouhar, Kris. "Fire Effects Information System (FEIS).” 2005. Accessed October 12, 2018. https://www.fs.fed.us/database/feis/plants/shrub/spajun/all.html.

Appendix A: List of Layers

BioClim 1 km Raster Layers				
Short Name	**Description**	**Projection**	**Source URL**	**Spatial Resolution**
BIO1	Annual Mean Temperature	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO3	Isothermality (BIO2/BIO7) (* 100)	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO4	Temperature Seasonality (standard deviation *100)	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO5	Max Temperature of Warmest Month	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO6	Min Temperature of Coldest Month	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO7	Temperature Annual Range (BIO5-BIO6)	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO8	Mean Temperature of Wettest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO9	Mean Temperature of Driest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO10	Mean Temperature of Warmest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO11	Mean Temperature of Coldest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO12	Annual Precipitation	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO13	Precipitation of Wettest Month	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO14	Precipitation of Driest Month	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO15	Precipitation Seasonality (Coefficient of Variation)	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO16	Precipitation of Wettest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO17	Precipitation of Driest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO18	Precipitation of Warmest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km
BIO19	Precipitation of Coldest Quarter	WGS 1984	https://www.worldclim.org/bioclim	1km

BioClim 10 m Raster Layers				
Short Name	**Description**	**Projection**	**Source URL**	**Spatial Resolution**
BIO1	Annual Mean Temperature	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO3	Isothermality (BIO2/BIO7) (* 100)	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO4	Temperature Seasonality (standard deviation *100)	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO5	Max Temperature of Warmest Month	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO6	Min Temperature of Coldest Month	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO7	Temperature Annual Range (BIO5-BIO6)	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO8	Mean Temperature of Wettest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO9	Mean Temperature of Driest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO10	Mean Temperature of Warmest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO11	Mean Temperature of Coldest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO12	Annual Precipitation	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO13	Precipitation of Wettest Month	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO14	Precipitation of Driest Month	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO15	Precipitation Seasonality (Coefficient of Variation)	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO16	Precipitation of Wettest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO17	Precipitation of Driest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO18	Precipitation of Warmest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m
BIO19	Precipitation of Coldest Quarter	WGS 1984	https://www.worldclim.org/bioclim	10 m

Angeles National Forest Digital Elevation Model Raster Layers

Short Name	Description	Projection	Source	Source URL	Spatial Resolution
DEM ANF Study	Angeles National Forest Elevation Data	WGS 1984	United States Geological Survey	https://viewer.nationalmap.gov/basic/	10m
DEM ANF Visibility	Angeles National Forest Spanish Broom Visibility data derived from DEM and Spanish Broom Sample Points	WGS 1984	Created using Arc Map from source DEM file via the United States Geological Survey	N/A	10m
DEM ANF Aspect	Angeles National Forest Aspect/ downslope direction of the maximum rate of change derived from DEM file	WGS 1984	Created using Arc Map from source DEM file via the United States Geological Survey	N/A	10m
DEM ANF Hill shade	Angeles National Forest Hill shade/shaded relief derived from DEM file considering the illumination source angle and shadows	WGS 1984	Created using Arc Map from source DEM file via the United States Geological Survey	N/A	10m
DEM ANF Slope	Angeles National Forest Slope/ gradient derived from DEM	WGS 1984	Created using Arc Map from source DEM file via the United States Geological Survey	N/A	10m

Angeles National Forest features converted to Rasters from Vector Layers

Short Name	Description	Projection	Source	Spatial Resolution
ANF CA Flow lines	California Hydrography Data clipped to Angeles National Forest Boundary	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	10m
ANF Recreation Facilities	Designated recreation facilities/sites located through the Angeles National Forest	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	10m
ANF Roads	Angeles National Forest Service roads managed by the US Forest Service	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	10m
ANF Vegetation Burn Severity	Areas burned in the Angeles National Forest by wildfire including the year of the fire and the name of the fire.	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	10m
ANF Water Bodies	Standing water bodies (Lakes, Small Lakes, ponds, etc...) in the Angeles National Forest.	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	10m
ANF Wildland Urban Intermix	Buffered feature class detailing areas where Wildland habitat, flora and fauna mix with Urban interface and human interaction.	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	10m

Angeles National Forest Vector layers list				
Short Name	**Description**	**Projection**	**Source**	**Source URL**
ANF CA Flow lines	California Hydrography Data clipped to Angeles National Forest Boundary	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
ANF Recreation Facilities	Designated recreation facilities/sites located through the Angeles National Forest	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
ANF Roads	Angeles National Forest Service roads managed by the US Forest Service	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
ANF Vegetation Burn Severity	Areas burned in the Angeles National Forest by wildfire including the year of the fire and the name of the fire.	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
ANF Water Bodies	Standing water bodies (Lakes, Small Lakes, ponds, etc...) in the Angeles National Forest.	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
ANF Wildland Urban Intermix	Buffered feature class detailing areas where Wildland habitat, flora and fauna mix with Urban interface and human interaction.	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/

Angeles National Forest Vector layers list (cont.)				
Short Name	**Description**	**Projection**	**Source**	**Source URL**
Spanish Broom Sample Points	Spanish Broom features class created in ArcMap using random point generating tool and filtered to .5 meters distance of actual locations of Spanish Broom from Spanish Broom polygon file	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
Spanish Broom Polygon	Spanish Broom invasive species data in the form of a polygon feature class resulting from field crew collection	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/
ANF Boundary	Administrative Boundary of the Angeles National Forest as administered by the US Forest Service	WGS 1984	Initial internal data request via Jason Martin, GIS Intern USFS. Data now currently available on USFS Geospatial Data Clearinghouse website.	https://data.fs.usda.gov/geodata/