Algorithm Selection in Species Distribution Modeling for the Conservation of Deciduous Eastern Temperate Trees

SUPER Program 2022

Final Report

Cristina Lucas [1] and Mollie Hendry [2]

Mentor: Josh Carrell [3]

[1] Fish, Wildlife, and Conservation Biology, Warner College of Natural Resources, Colorado State University, Fort Collins, CO

[2] Ecosystem Science and Sustainability, Warner College of Natural Resources, Colorado State University, Fort Collins, CO

[3] Forest and Rangeland Stewardship, Warner College of Natural Resources, Colorado State University, Fort Collins, CO

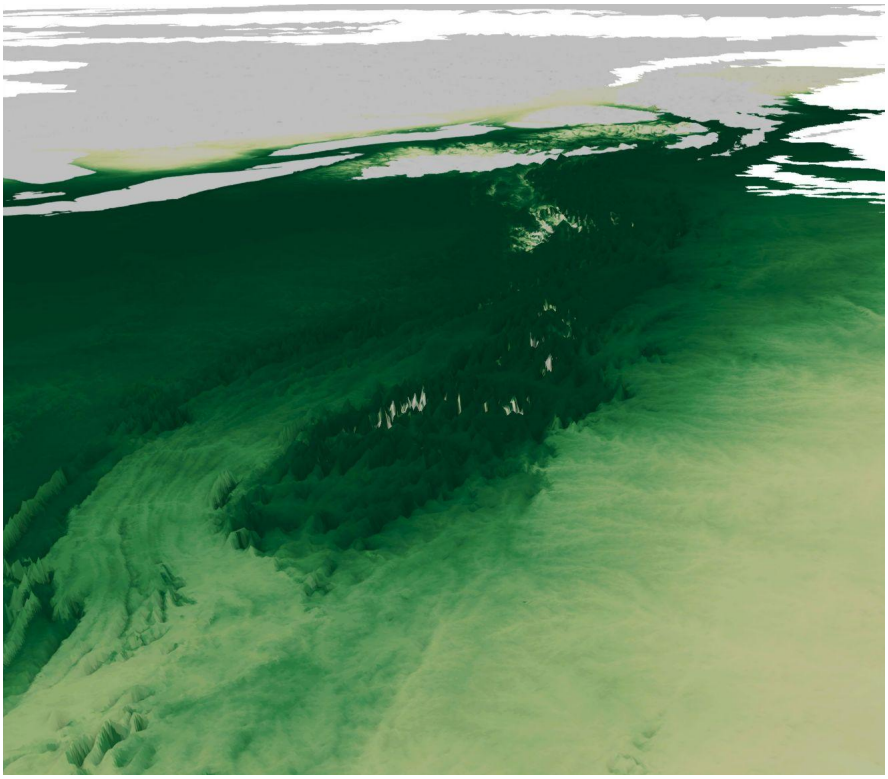

Figure 1. Image of Species Distribution Model for Red Oak trees. Photo: Mollie Hendry 2023

Abstract

Algorithm selection in Species Distribution Modeling (SDM) is essential in producing the most efficient and cost-effective areas to conserve under varying spatial and socioeconomic conditions. The purpose of this research is to identify which SDM algorithm(s) are most effective in conserving 70 deciduous tree species in Eastern Temperate forests in the continental United States. This particular study is a pilot study for 9 of the identified 70 plant species. Preexisting, public species occurrence data for these plants was sourced using an R programming script. Species distribution models were then created from several Marxan algorithms and analyzed for cost-effectiveness. The Random Forest algorithm was found to be more cost-effective in conserving 30% of biodiversity for the 9 selected Oak tree species. These findings have broad implications for ecosystem managers in a variety of fields. Our results indicate that our statistical choices as scientists can create significant differences in management plans and associated socioeconomic costs. Additional research should aim to understand if algorithm choice has varying levels of influence when used in different ecosystems, on rare flora and fauna, and when using absence vs. presence data.

1.0 Introduction

Systematic conservation planning (SCP) is a management tool that allows scientists to decide the most effective methods to conserve biodiversity while considering both the spatial and socioeconomic impacts of reserve systems (Watts et al. 2017). Making financially efficient conservation management decisions is of utmost importance when regarding the increasing human population and resource use that coincides with today's environmental challenges such as climate change and natural area depletion. SCP can be utilized to both create reserve systems or increase the reserve connectivity of existing conservation areas. This practice can also fulfill a number of different conservation goals, ranging from protecting suitable habitat, biodiversity, or individual species.

An integral feature of SCP is creating species distribution models (SDM) that reliably present the relationship between species occurrence data and the environmental factors contributing to habitat suitability (Miller 2010). SDMs utilize mathematical algorithms to correlate presence records and predictor variables that influence species distribution (i.e. climate, precipitation, elevation) to create models reflecting probable species occurrence across space (Figure 2). Today, SDMs are often used to predict species location under shifting conditions due to the global changes in temperature and seasons. Research has shown that algorithm choice can significantly affect SDM outputs (Li and Wang 2013). Unfortunately, this results in no one algorithm being a perfect fit for all species and subsequent management challenges (Sofaer et al. 2019).

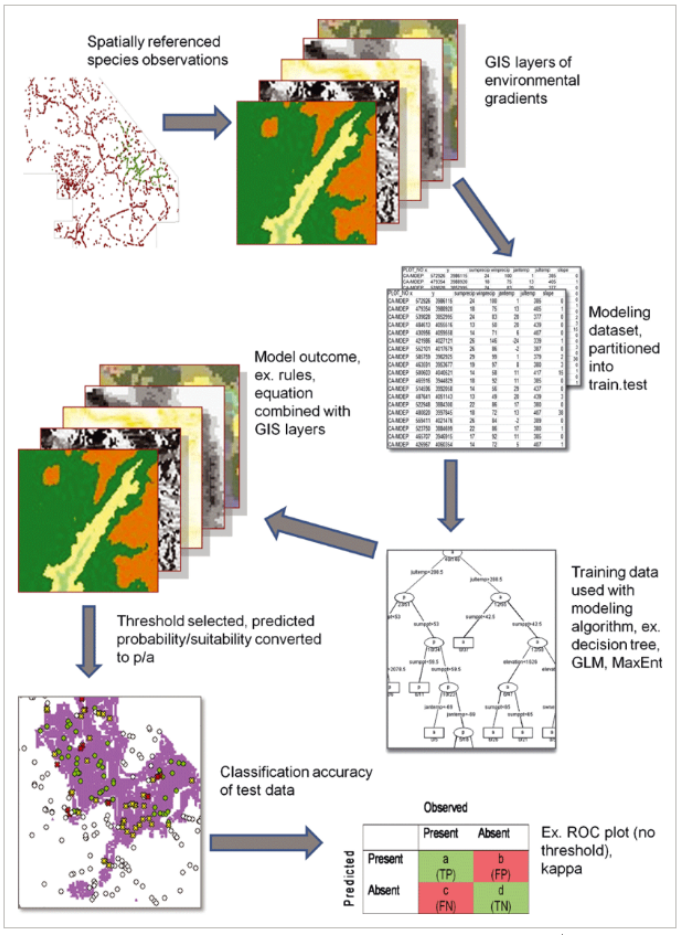

Figure 2. The species distribution modeling process. *Species Distribution Modeling*. Copyright: 2010 Jennifer Miller.

Two commonly used SDM algorithms are the Random Forest and Artificial Neural Networking algorithm. These two algorithms were selected due to past research showing significant differences among SDM outputs and accuracy (Raczko and Zagajewski 2017). With these differences in mind, this paper seeks to understand the influence SDM algorithm selection holds over the spatial and socioeconomic costs associated with conservation management decisions. To do this, we utilized a SCP software known as Marxan to create two spatial conservation plans set to conserve 30% of deciduous tree biodiversity in the eastern temperate forest region of North America (Figure 3).

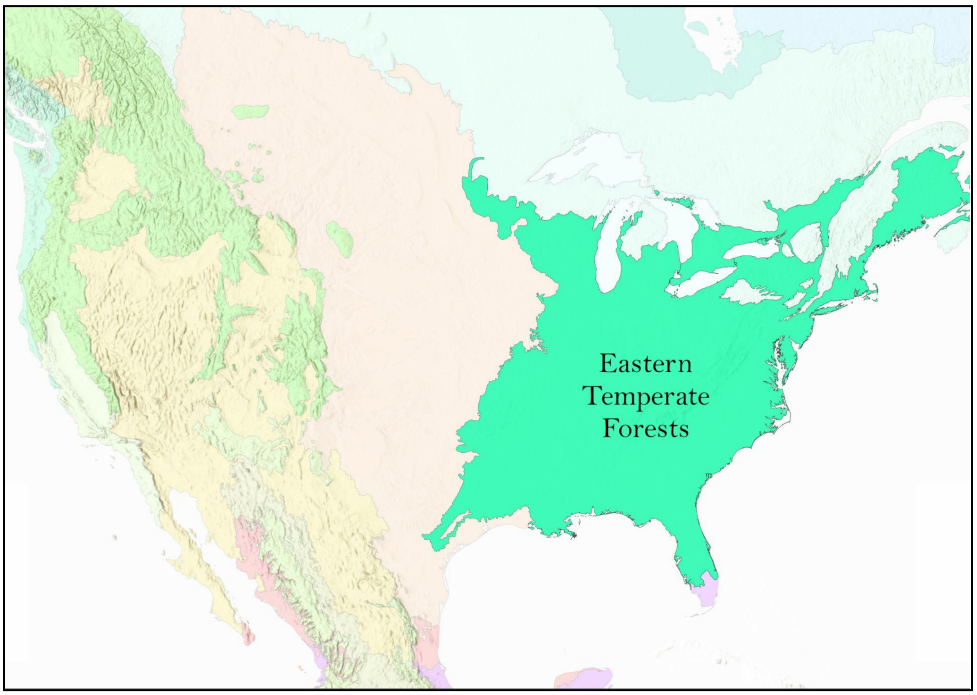

Figure 3. Project study area, ArcGIS boundary of the Eastern Temperate Forests Ecoregion in North America. Photo: Mollie Hendry, 2023.

This study acts as a pilot study to a larger project that will be analyzing significantly more species and potential SDM algorithms; as such, the research conducted here will be looking at nine common, non-threatened Oak species in our study area (Figure 4). In hopes to create a comparative study that can reach a broad scope of conservation, this paper focuses on socioeconomic differences amongst SDM algorithms without the consideration of rare or endangered species.

Figure 4. *Northern Red Oak*. Photo of the Northern Red Oak tree species collected through the iNaturalist network in June of 2018. https://www.inaturalist.org/guide_taxa/1175441

2.0 Research Questions and Hypotheses

2.1 Research question

What spatial and socioeconomic cost differences are present between Marxan outputs generated from Random Forest and Artificial Neural Network algorithms?

2.2 Expected outcome, or research (alternative) hypothesis

There will be a significant difference amongst spatial and socioeconomic costs between Marxan outputs created using SDMs that utilized Random Forest vs. Artificial Neural Network algorithms.

2.3 Emergent null hypothesis

There are no differences amongst spatial and socioeconomic costs between Marxan outputs created using SDMs that utilized Random Forest vs. Artificial Neural Network algorithms.

2.4 Explanation

Past research has shown that different algorithms have shown to work better with different geographic regions and species characteristics. As such, we are expecting to find differences among financial costs based on species distribution and algorithm choice.

3.0 Methods

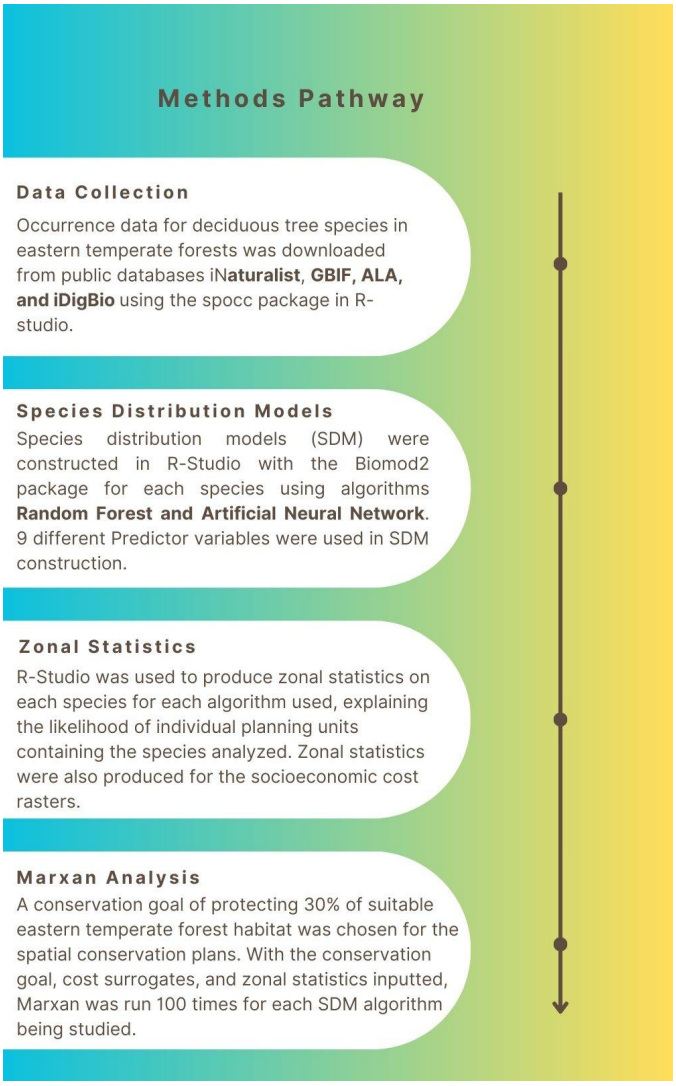

Figure 5. Infographic of the methods process for the project including steps: Data Collection, Species Distribution Models, Zonal Statistics, Marxan Analysis. Photo: Mollie Hendry

3.1 Study Area

Our study area spread across the Eastern Temperate Forest ecoregion of North America. Here, deciduous trees comprise a large portion of the area's vegetation. The ecoregion spans from the Atlantic coast into central-southern North America and is characterized by its temperate climate ranging from cool, to continental, to subtropical, humid, and dense tree cover (Gilliam et al. 2010).

3.2 Data Collection

Occurrence data for nine deciduous tree species in eastern temperate forests was downloaded from a public databases Inaturalist, GBIF, ALA, and iDigBio, using the spocc package in R. The nine deciduous tree species analyzed in this study have no current threats to population levels and are commonly found throughout the study area (Table 1). Occurrence data included location information via Latitude and Longitude as well as the time each record was made.

Table 1. Describes the number of occurrences for each Oak species analyzed in the study, along with the databases used to download presence records.

Species	# of occurrences	Sourced Databases
White Oak	25149	GBIF, ALA, iDigBio
Red Oak	54	iDigBio
Shingle Oak	7652	iNaturalist, GBIF, iDigBio
Bur Oak	14474	iDigBio, GBIF
Black Jack Oak	13981	iDigBio, iNaturalist, GBIF
Chestnut Oak	9367	GBIF, iNaturalist, iDigBio

Pin Oak	51305	ALA, iDigBio, GBIF
Post Oak	8429	iDigBio, GBIF
Black Oak	9473	ALA, iDigBio, GBIF

Abiotic factors mean annual air temperature (Celsius), mean diurnal air temperature range (celsius), isothermality (Celsius), mean daily maximum air temperature of the warmest month (Celsius), mean monthly precipitation amount of the wettest quarter (kg m-2), mean monthly precipitation amount of the driest quarter (kg m-2), and mean monthly precipitation amount of the coldest quarter (kg m-2), have been identified as factors influencing the distribution of eastern temperate deciduous tree species (Walthert and Meier 2017). Abiotic data was downloaded in raster format from Chelsea Climate (https://chelsa-climate.org/bioclim/) and were utilized as predictor variables in the 18 species distribution models created in this study.

Systematic conservation planning provides solutions that reach set conservation targets at the lowest possible cost. The complexities of providing direct financial estimates to each planning unit goes beyond the scope of this analysis; as such, a cost surrogate of agriculture use was utilized. This information was downloaded from Naidoo et. al 2008.

3.3 Species Distribution Modeling

Species distribution modeling is the act of correlating predictor variables with presence records in order to predict the probability of species occurrence across a region. A Multitude of algorithms are utilized for this correlation; however, these algorithms do not consistently produce the same distribution model even when considering identical data. Prior to creating species distribution models, a map boundary of eastern temperate forests was created in ArcGISMaps in order to isolate species presence data to the geographic region of study. This boundary was divided into 28,967 1 km^2 planning units. Given location information and predictor variables, species distribution models (SDM) were constructed in R-Studio for each species using Random

Forest and Artificial Neural Network algorithms. The R package Biomod2 was used in SDM creation for its efficacy in using presence and pseudo-absence data.

3.4 Marxan Analysis

Marxan is a systematic conservation planning tool that provides financially efficient spatial solutions to specific conservation goals. These goals can range from protecting specific species, supporting biodiversity, or increasing the connectivity of current reserve systems. Here, a conservation goal of protecting 30% of eastern temperate forest deciduous tree biodiversity was chosen for the spatial conservation plans. Conserving 30% of biodiversity or suitable habitat has shown to be a standard in maintaining long-lasting healthy population levels (Carrell et al. 2022). The costs we chose to analyze included those associated with agriculture. As a result, coordinates that fall within areas commonly used for agriculture will be deemed pricier to conserve. Marxan utilizes the average costs and species distributions within each planning unit when selecting priority areas. To accomplish this, zonal statistics were carried out as a table function within ArcGIS Pro. With the conservation goal, cost surrogates, and zonal statistics inputted, Marxan was run 100 times for each SDM algorithm being studied. The best solution from every 100 runs was chosen based on the required number of planning units needed to fulfill the set conservation goal maintaining 30% deciduous tree diversity and the costs associated with how those planning units negatively affected agriculture use.

4.0 Results

Random Forest (RF) Species Distribution Models were observed to be more economically efficient than Artificial Neural Network (ANN) Species Distribution Models when utilized in spatial conservation planning software, Marxan. The conservation plan developed using RF models required 4940 planning units and created a cost value of 394,038.5 to fulfill the set conservation goal of conserving 30% of biodiversity. In comparison, ANN models required 7416 planning units and created a cost value of 526,687.8. As shown in **Figure 6,** the planning units prioritized from the RF

models were considerably more concentrated and did not travel as far south as the planning units prioritized from the ANN models.

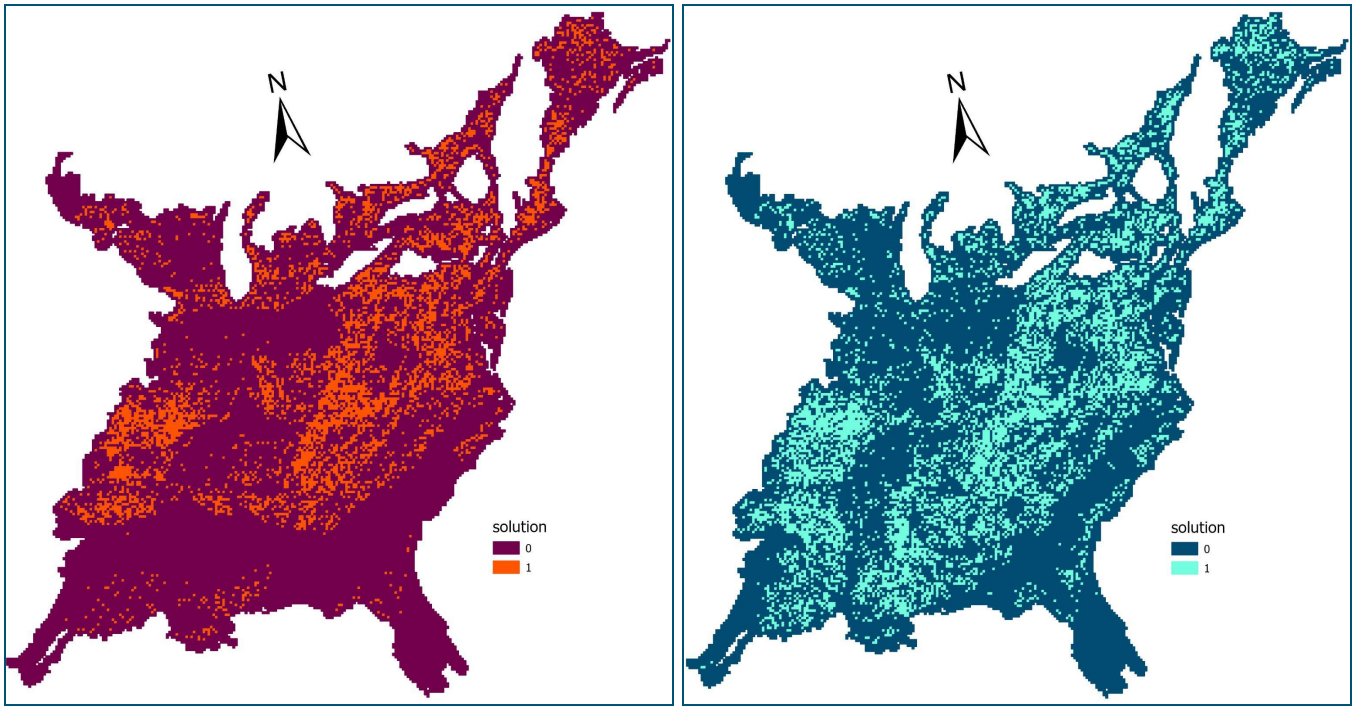

Figure 6. Maps show the best Marxan solutions for the Random Forest algorithm SDMs (left) and Artificial Neural Networking algorithm SDMs (right) to reach a conservation goal of conserving 30% deciduous tree biodiversity. Bright coloration signifies planning units selected in the Marxan best solutions; whereas, dark colors reflect areas left out of the best solutions.

Figure 7 shows the mean SDMs when using RF models and ANN models respectively. The RF model produced a species-averaged SDM that has significantly more concentrated priority habitat areas than the ANN Model. In addition to having less succinct priority areas, the ANN model's habitat prioritization showed a greater amount of high priority habitat, and had low priority habitat that was more expansive across the study region. These results indicate a significant difference in spatial and socioeconomic costs between the Marxan solutions produced by the RF and ANN algorithm SDMs.

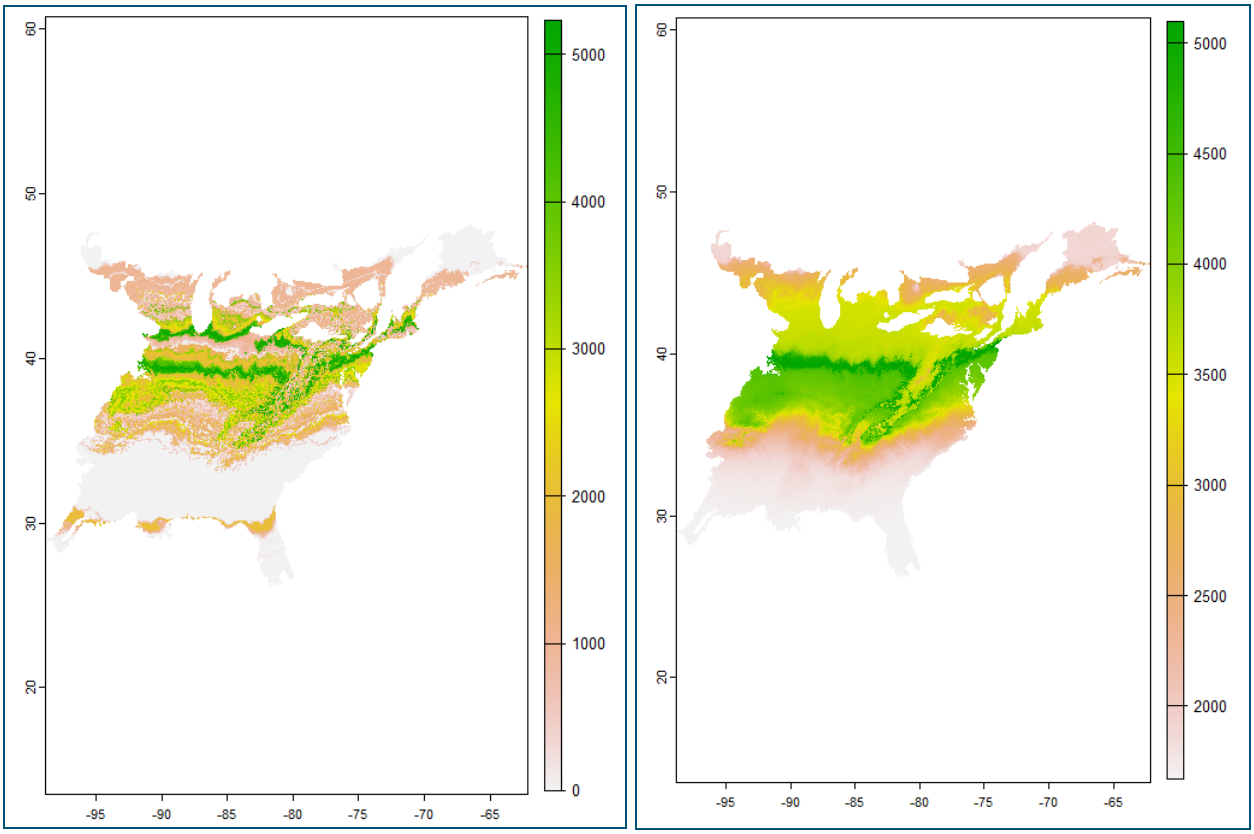

Figure 7: Maps show the average SDMs produced from the Random Forest algorithm (left) and the Artificial Neural Network Algorithm (right). Bright green coloration reflects high intensities of suitable habitat. Suitable habitat decreases in intensity as coloration travels from bright green to light pink.

5.0 Discussion

Our results indicate that our statistical choices as scientists can create significant differences in management plans and associated socioeconomic costs. The SDMs built from RF models, shown in **Figure 7**, provided more succinct priority habitat and resulted in less planning units utilized in Marxan as shown in **Figure 6**; therefore, a more cost efficient solution. This result may be due to the fact that the RF SDM produced far less priority habitat while also having a harsher cut off for low priority habitat. In comparison, the SDMs produced by the ANN model, shown in **Figure 7**, had a larger proportion of high priority habitat as shown in **Figure 6**. Additionally, the low priority habitat created from the ANN model spread further South and was increasingly less concentrated the closer it got to the southern U.S. border as demonstrated in **Figure**

6. Interestingly, Marxan selected a large proportion of planning units that fell into this low priority habitat when using the ANN Model. This could show that SDM Model choice may not only affect the socioeconomic cost of conservation planning but the environmental robustness of management actions. Further research should focus on understanding what factors of the SDM prompts Marxan to choose low priority habitats. Additionally, past studies have found that RF performance has varied in research between default programming and the use of techniques to manage class imbalance (Valavi et al. 2021). The gap in scientific understanding around the strengths and weaknesses of algorithms with varying programming techniques is especially prevalent today. A study analyzing SDMs created for Western tree species found significant differences in algorithm extrapolation performance when applied across environmental space (Charney et al. 2021). Comparative studies should be conducted to assess performance differences between algorithms in order to bridge this knowledge gap. A growing need to understand the strengths and weaknesses of different SDM algorithms is present due to environmental shifts caused by climate change and other anthropogenic pressures.

This pilot study was instrumental in allowing us to develop a relevant process for data collection, SDM development code, zonal statistics development code, and Marxan procedures for the methodology of the larger research study. We were successful in producing results that provided the information we hoped to obtain about the differences between SDM algorithms. We plan to expand the methodology for the larger study to include variations in coding/development of each of the SDM algorithms to understand the abilities of algorithms when produced with default code versus other programming to manage varying conditions and environments. We also plan to expand the number of SDM algorithms to be evaluated in the larger study.

6.0 Conclusion

Our study aimed to analyze how SDM algorithm choice affected the costs associated with spatial conservation planning. Our findings suggest that algorithm choice can have significant impacts on the socioeconomic costs associated with

systematic conservation planning. More research should be conducted to determine if these findings continue to be supported when analyzing rare or endangered species.

References

Charney ND, Record S, Gerstner BE, Merow C, Zarnetske PL. 2021. A Test of Species Distribution Model Test of Species Distribution Model Transferability Acr ability Across Environmental and Geographic Space for 108 Western North American Tree Species. Frontiers in Ecology and Evolution. 9:689295. https://doi.org/10.3389/fevo.2021.689295

Carrell, J. D., Hammill, E., & Edwards, T. C. (2022). Balancing rare species conservation with extractive industries. Land, 11(11), 2012. https://doi.org/10.3390/land11112012

Gilliam, F. S., Goodale, C. L., Pardo, L. H., Geiser, L. H., & Lilleskov, E. A. (2010). *Assessment of Nitrogen Deposition Effects and Empirical Critical Loads of Nitrogen for Ecoregions of the United States: Chapter 10: Eastern Temperate Forests* (General Technical Report NRS; pp. 99–116). United State Forest Service. https://www.nrs.fs.usda.gov/pubs/gtr/gtr-nrs-80chapters/10-gilliam.pdf

McCain, C. M., King, S. R., & Szewczyk, T. M. (2021). Unusually large upward shifts in cold-adapted, montane mammals as temperature warms. *Ecology*, *102*(4), 12.

Miller, J. (2010). Species Distribution Modeling. *Geography Compass*, *4*(6), 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x

LI, X., & WANG, Y. (2013). Applying various algorithms for species distribution modelling. Integrative Zoology, 8(2), 124–135. https://doi.org/10.1111/1749-4877.12000

Naidoo, R., Balmford, A., Costanza, R., Fisher, B., Green, R. E., Lehner, B., Malcolm, T. R., & Ricketts, T. H. (2008). Global mapping of Ecosystem Services and Conservation Priorities. Proceedings of the National Academy of Sciences, 105(28), 9495–9500. https://doi.org/10.1073/pnas.0707823105

Walthert, L., & Meier, E. S. (2017). Tree species distribution in temperate forests is more influenced by soil than by climate. Ecology and Evolution, 7(22), 9473–9484. https://doi.org/10.1002/ece3.3436

Raczko, E., & Zagajewski, B. (2017). Comparison of support vector machine, Random Forest and neural network classifiers for tree species classification on airborne hyperspectral apex images. European Journal of Remote Sensing, 50(1), 144–154. https://doi.org/10.1080/22797254.2017.1299557

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., Edwards, T. C., Jr, Guala, G. F., Howard, T. G., Morisette, J. T., & Hamilton, H. (2019). Development and Delivery of Species Distribution Models to Inform Decision-Making. *BioScience*, *69*(7), 544–557. https://doi.org/10.1093/biosci/biz045

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. 2021. Modelling species presence-only data with random forests. Ecography. 44: 1731-1742. https://doi.org/10.1111/ecog.05615\

Watts, M. E., Stewart, R. R., Martin, T. G., Klein, C. J., Cawardine, J., & Possingham, H. P. (2017). Systematic Conservation Planning with Marxan. In *Learning Landscape Ecology* (pp. 211–227). Springer, New York, NY. https://link.springer.com/chapter/10.1007/978-1-4939-6374-4_13

Acknowledgements

Thank you to Dr. Stacy Lynn of the Ecosystem Science and Sustainability department as our wonderful professor. This project was completed through the SUPER (Skills for Undergraduate Participation in Ecological Research) program at Colorado State University.

Appendix 1

1. **Methods Outline**
    - **Data Collection**
        - The data we will be using is occurrence data for species that includes time and location information. This data is publicly available.
    - **Data Entry and Processing**
        - Download data public occurrence data.
    - **Data Analysis**
        - Create a boundary in GISMaps so that we can attach species presence data within that boundary.
        - Create species distribution models using a variety of algorithms. These will be chosen shortly and created in R-Studio.
        - Create Cost surrogates and conservation goals in Marxan and plug in species distribution data.
        - Run Marxan to create several spatial conservation plans, each using a species distribution model created from the different modeling algorithms.
        - Analyze data for significant differences among socioeconomic costs in R-Studio.
    - **Data Interpretation**
        - We will be looking at variations in conservation plans by their spatial and financial costs, which will be produced by Marxan. We will analyze the conservation plans to determine which algorithms are best at producing conservation plans with the lowest spatial and financial costs for the species we have chosen to conserve.

Appendix 2

General Parameters For Marxan

VERSION 0.1

BLM 25

PROP 0.5

RANDSEED -1

NUMREPS 100

Annealing Parameters

NUMITNS 1000000

STARTTEMP -1

NUMTEMP 10000

COOLFAC 0.00000000000000E+0000

Cost Threshold

COSTTHRESH 0.00000000000000E+0000

THRESHPEN1 1.40000000000000E+0001

THRESHPEN2 1.00000000000000E+0000

Input Files

INPUTDIR input

PUNAME pu.dat

SPECNAME spec.dat

PUVSPRNAME puvsp2.dat

BOUNDNAME bound.dat