ABSTRACT

MODELING DISTRIBUTIONS OF *CANTHARELLUS FORMOSUS* USING NATURAL HISTORY AND CITIZEN SCIENCE DATA

by Zoey Nicole Armstrong

The Pacific Golden Chanterelle (*Cantharellus formosus*) is a widely sought-after mushroom most abundant in the forests of Washington and Oregon, USA. This project used the species to investigate how accurately the species distribution could be modeled using natural history (herbarium) as model training data and citizen science (iNaturalist) as validation data. To combat the potential sampling bias towards population centers an effort variable weighting scheme was used to consider observations in harder to reach areas more than those in easier to access areas. Four models were created and run using the natural history data as training points: Random Forests (RF), Maxent, General Linear Model (GLM), and Artificial Neural Network (ANN); the effort variable was only applied to the ANN and GLM models. Out of these four, RF was found to perform the best with an equitable skill score (ETS) 0.987 when tested against the iNaturalist citizen science validation points. Overall, this project provides a good proof of concept and framework for the use of herbarium and citizen science data for use in biogeographical modeling projects in the future.

MODELING DISTRIBUTIONS OF *CANTHARELLUS FORMOSUS* USING NATURAL HISTORY AND CITIZEN SCIENCE DATA

Thesis

Submitted to the

Faculty of Miami University

in partial fulfillment of

the requirements for the degree of

Master of Arts

by

Zoey Nicole Armstrong

Miami University

Oxford, Ohio

2021

Advisor: Dr. Mary C. Henry

Reader: Dr. Jessica L. McCarty

Reader: Dr. Nicholas P. Money

This thesis titled

MODELING DISTRIBUTIONS OF *CANTHARELLUS FORMOSUS* USING NATURAL HISTORY AND CITIZEN SCIENCE DATA

by

Zoey Nicole Armstrong

has been approved for publication by

The College of Arts and Sciences

and

Department of Geography

_____

Dr. Mary C. Henry

_____

Dr. Jessica L. McCarty

_____

Dr. Nicholas P. Money

Table of Contents

List of Tables

List of Figures

Dedication

Dedicated to all the fun guys out there.

Acknowledgements

I'd like to thank Dr. Mary Henry for all the support in completing this thesis, as well as to my wonderful committee members Dr. Jessica McCarty and Dr. Nicholas Money. I would also like to thank my amazing partner Yola R., my good friend Martin Ganev, and all the members of my graduate cohort for making the process all the more enjoyable. Special shoutout to all the Miami Department of Geography staff for creating a supportive and open environment for all students to study in.

Abstract:

The Pacific Golden Chanterelle (*Cantharellus formosus*) is a widely sought-after mushroom most abundant in the forests of Washington and Oregon, USA. This project used the species to investigate how accurately the species distribution could be modeled using natural history (herbarium) as model training data and citizen science (iNaturalist) as validation data. To combat the potential sampling bias towards population centers an effort variable weighting scheme was used to consider observations in harder to reach areas more than those in easier to access areas. Four models were created and run using the natural history data as training points: Random Forests (RF), Maxent, General Linear Model (GLM), and Artificial Neural Network (ANN); the effort variable was only applied to the ANN and GLM models. Out of these four, RF was found to perform the best with an equitable skill score (ETS) 0.987 when tested against the iNaturalist citizen science validation points. Overall, this project provides a good proof of concept and framework for the use of herbarium and citizen science data for use in biogeographical modeling projects in the future.

Introduction:

Understanding fungi as a group of organisms with distinct distributions is a relatively recent trend within mycological studies. It was long thought that dispersal mechanisms and barriers played no role in the distribution patterns seen within these species. Where the only obstacle to dispersal was a lack of proper conditions; if the conditions were right, the appropriate fungi would appear. These views were challenged in 1943, with the first documented case of endemism in fungi was described (Bisby, 1943; Kabir et al., 2010). At the time however, this was largely considered an outlier. In Bisby's own work, he claimed that climate affects fungal distributions only indirectly by influencing the locations of symbiotic plant species. It has only been a relatively recent development with the advancement of modern molecular studies and advancement in computational power that the distributions of fungi have taken on a more serious academic focus (Kabir et al., 2010).

Given the recent development of this field, many of the tools used to study distributions were developed for and subsequently borrowed from the sister field of botany. Looking at how these methods have been developed and implemented with plant distributions will provide a framework for understanding the differences and adaptations that have been used when applying these techniques towards macro-fungi. One method that has become popular with advances in both computational power and access to environmental data is the ecological niche model. Ecological niche modeling relies on a dataset of species observations and detailed environmental data. The environmental data serve as predictor variables which are statistically compared to determine the relative importance of each in explaining the distribution of the initial dataset (Buechling and Tobalske, 2011). Once these predictors are ranked, they can be used to interpolate the predictor surface and identify the most suitable areas for the species. This approach has been widely used to map plant distributions and within this framework there are two methodologies for compiling a starting dataset: survey data and natural history data; both of which have their strengths and weaknesses.

The traditional approach would be to run a field survey, but studies have increasingly been making use of natural history datasets like those found in herbaria. Field surveys have the distinct advantage of creating datasets with both presence and absence data (Lobo et al., 2010). Presence/absence data are only available from an in-depth and systematic field survey study, whereas herbarium data are presence only. There is no way of determining whether an area with no observations is really a gap in distribution or if it is simply an area where information has yet to be gathered. Having presence/absence data means that the accuracy of the model is of a higher standard since there are fewer assumptions being made (Lobo et al., 2010). But obtaining this type of data comes at the price of limiting the geographic scope and increasing both the financial cost and the effort of the project. These practical drawbacks have led to an increased push for using natural history data and finding ways around the limitations inherent to the methodology (Andrew et al., 2018; Lavoie, 2013).

Project Goals/Questions:

1. What is the effectiveness of using natural history data to model the distribution of fungi?
2. Which modeling procedure predicts distribution best when using herbarium datasets?
3. What is the effectiveness of adding an effort variable to the model?

Literature Review:

Herbaria have long served as a home for botanical and mycological specimens. These institutions have been compiling and storing important biological data since 1532, and now collectively house millions of specimens from across the globe (Findlen, 2017). With recent concerted digitization efforts have made the vast wealth of biological data stored by these organizations widely available. This sudden influx of potential new data sources has naturally led to questions of how it can best be utilized, with researchers tackling questions of invasive species spread, phenological changes, as well as distribution. Many records have georeferenced locations attached and can be used to populate an ecological niche model; potentially helping to predict species distributions at a lower cost than traditional methods.

This push can especially be seen in the budding field of mycological distributions. As the field has grown, it has adopted many of the methods and much of the theory behind plant distribution modeling (Guo et al., 2017; Yuan et al., 2015; Bakkestuen et al., 2008). Although, given the ephemeral nature of the above ground fruiting portion of macro-fungi, these studies inherently lend themselves towards use of natural history data (Andrew et al., 2018). This affinity towards pre-existing datasets can lead to the same pitfalls noted above when considering botanical studies, namely spatial biases, and lack of absence data, but it provides a way to examine a large landscape which is necessary given the broad ranges of many macro-fruiting fungi.

The ease of access and large scope of the results has led to impassioned pleas to make use of the growing amount of digitized natural history data to study the somewhat mysterious hidden

life of fungi (Bakkestuen et al., 2008; Andrew et al., 2018). However, as noted above, it is important to be aware of the limitations associated with this method: spatial biases, climatic biases, and lack of reliable absence data. One of the major assumptions for any good data model is random sampling of observations. Natural history data are, by their very nature, often collected in unsystematized and random ways, which can lead to heavy spatial and climatic biases (Daru et al., 2017). However, these biases in sampling can be mitigated. Syfert et al. (2013)found that use of sampling bias grids helped to lessen the effect of unequal sampling while using herbarium data. Similarly, climate biases have been found to have a small effect on overall model effectiveness, and these effects were able to be accounted for by incorporating more initial data points (Loiselle et al., 2007).

Finally, the lack of absence data has been a major concern when using natural history datasets for species distribution modeling. However, while less accurate than field surveys, pseudo-absences can be used to generate this information. Elith and Leathwick outline a procedure for producing this data from herbarium records and determine that presence only models can be "sufficiently accurate" (Elith & Leathwick, 2007). In their paper, they also find that results can be improved further by including inventory or random pseudo-absences. Inventory pseudo-absences tested better in the paper, but since this is very close to a traditional survey it could pose some difficulties when translated to fungi. It remains a potential avenue for investigation, and if any inventory data exists it should be incorporated into the model. Random absences on the other hand, were also found to improve model outputs and can be readily applied towards herbarium data. This method involves randomly sampling areas where no observations were noted and assuming them to be absence points.

These ideas are all exemplified in three recent papers that demonstrate the application of these biogeographical principles towards mapping the distributions of fungi. Two of the papers from China used existing survey data on rare fungi as their starting dataset (Guo et al., 2017; Yuan et al., 2015). The third paper is from Norway that specifically looks at the use of herbarium records to predict distributions of several different species of fungi (Bakkestuen et al., 2008). One important similarity between these papers is the study area size. All were carried out over large geographic areas, since fungi tend to occupy larger geographic ranges. The three studies also found a high predictive value correlated to climate variables, which, as pointed out by

Bakkestuen et al., could potentially make understanding and predicting the distributions of fungi a novel local predictor of climate change effects (Bakkestuen et al., 2008).

Methodology:

Study Area and Species:

This thesis project investigated the question: "How can the distribution of *Cantharellus formosus* be accurately modelled using natural history data?". *Cantharellus formosus,* also known as the Pacific Golden Chanterelle, is a highly sought-after edible mushroom native to the coastal mountain ranges centered around Washington and Oregon, USA, though the range extends down into northern California and up into British Columbia, Canada. This species was chosen for several reasons, but one of the biggest factors was its large number of herbarium records. This makes it less susceptible to the spatial and climatic biases discussed earlier on in the literature review. In addition to this criterion the Pacific Golden Chanterelle is an economically important mushroom. According to the U.S. Department of Agriculture Forest Service, the mushroom is a key species in the "multimillion-dollar industry" of mushroom foraging (Pilz et al., 2002). Given the economic and cultural significance of the species, understanding how the species is currently distributed throughout the area can help with management efforts in the future.

The factors mentioned above make the Pacific Golden Chanterelle a good candidate for use in a distribution analysis. Performing this study may help contribute some foundational analysis to the growing field of macro-fungi distributions. In addition, this project aimed to fill a gap in the literature by performing a validation of the model using citizen science data. This approach has not been applied towards this type of study, but it has been found to be a useful supplement towards conservation biogeographic projects (Beltrame et al., 2010). Again, the Pacific Golden Chanterelle makes a good candidate for a citizen science approach, since it is often sought after and is easily recognizable for amateur mushroom hunters. By incorporating

this approach as a supplement to my methods, I expected to finish with more statistically robust results. Interpretation of these results will help add to the body of literature surrounding the effectiveness of natural history data for use within distribution modeling work.

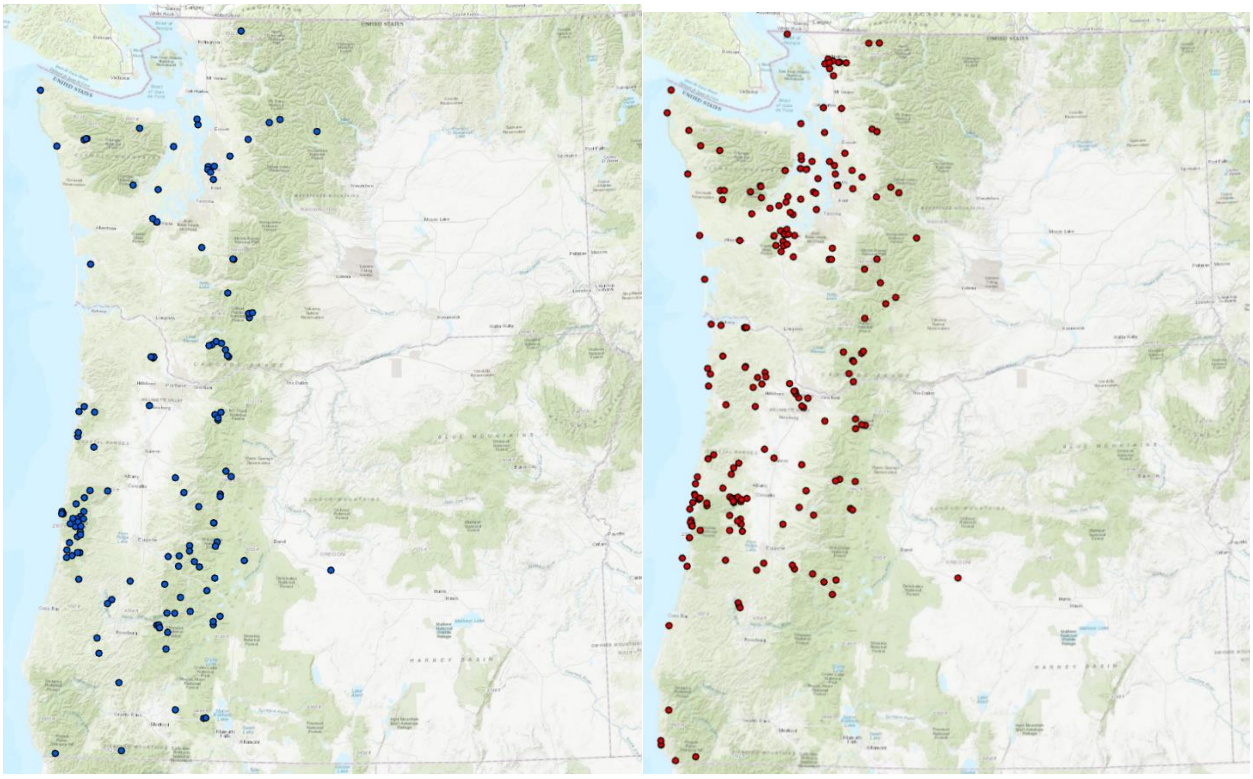


Figure 1: Left, Herbarium observation throughout the study area (n= 128), Right, iNaturalist (citizen science) observations (n = 259)

Data Collection and Clean Up:

To create and validate a distribution model amidst the ongoing uncertainty surrounding COVID-19, this project made use of information that is freely available online. Three main sources of information were needed: initial observation data, validation data, and relevant bioclimatic variables. This section will delve into the acquisition of bioclimatic variables and initial observations, and discussion of validation data will come later in the statistical comparisons section. All these datasets were cleaned up and manipulated within the statistical programming language R (R Core Team, 2017).

The initial observation data for the project were obtained from MyCoPortal (mycoportal.org), an online aggregator of digitized fungarium records. This online repository scours through hundreds of herbaria from across the globe. With over 6.4 million records in total, it serves as a huge wealth of untapped biological information. This initial dataset consisted of all the observations of the target species, *Cantharellus formosus* in the study area of Washington and Oregon. When performing this search on MyCoPortal. I found that there were 1,620 results. The data obtained from the MyCoPortal website needed to be cleaned, however, as many older observations had only approximate locations, leading to duplications of a single point, or lacking a location entirely. In total after the data cleaning there were 128 observations.

After cleaning up the initial data, predictor variables were found from online sources. Similar studies have made use of variables such as sun radiation, monthly temperature averages, monthly precipitation values, geological richness, slope, and elevation among others (Wollan et al, 2008; Yuan et al, 2016). These past studies have used broad lists of initial predictor variables that may seem unnecessary. They are included in the lists however, because a key idea within the statistical approaches to distribution modeling is the ability to take a large range of potentially relevant data and condense it to a more manageable list of those that best explain the variation seen in the initial observation points. This ability to refine from a broad range of starting predictors will allow this project to make use of a large variety of initial data sources.

It is also important to consider which variables are biologically relevant. In the case of fungi, it is important to use climate data that is associated with short time intervals, such as average temperature per month. Since fungal fruiting is highly sensitive to slight changes in climactic conditions and use of a larger timeframe could be missing patterns. In addition to this general consideration, there have been more specific studies on the life history of *Cantharellus formosus*. These have reported finding the fruiting body in areas associated with buried coarse woody debris, moss-free humus, and an open canopy (Bergemann and Largent, 2000). While these are highly site-specific factors that lead to fruiting, it shows an importance for including variables such as soils maps and land use cover or a vegetation index. Considering biologically relevant variables and looking to other similar studies provided a strong starting point for the initial variable selection. Table 1 shows the list of climate variables that were used in the study. The variables layers were imported into R where they were clipped to the study area, projected to US Pacific Northwest Albers, and resampled to 250m resolution.

Table 1: List of predictor variables along with sources and resolutions

Variable Type	**Sources**
Climate variables (temperature, precipitation, seasonality)	Worldclim standard 19 bioclimatic variables (1km), (Fick and Hijmans, 2017)
Monthly average precipitation	PRISM climate group (800 m), (PRISM Climate Group, 2010)
Land cover type and forest canopy cover	MRLC product (30m), (MRLC, 2016)
Soil class, cation exchange capacity, nitrogen, soil pH	ISRIC world soil information (250m), (Batjes et al., 2020)
Slope, aspect, and elevation	GMTED2010 DEM (250m), (Danielson and Gesch, 2010)

Effort Variable:

One method to improve the accuracy of a species distribution model using unsystematically collected data, such as herbarium or citizen science data, is to use an effort variable weighting scheme (Stolar and Nielsen, 2014). In this project the effort variable was created using four layers that were used to approximate the effort of obtaining a sample: distance to roads, distance to herbaria, terrain ruggedness index, and road density. The roads data came from the Topologically Integrated Geographic Encoding and Referencing (TIGER) dataset (census.gov, 2020), herbaria points were based of the Consortium of Pacific Northwest Herbaria website (Consortium of PNW Herbaria, 2020), and the ruggedness was derived from the 2010 Global Multi-resolution Terrain Elevation Data (GMTED2010) DEM (Danielson and Gesch, 2010). These raster layers were then run through a principal components analysis (PCA), and the primary component was used as the effort variable when running the model.

Model Creation:

Once this data was all compiled and properly formatted, the model creation was conducted. The main portion of the modeling was done using the ensemble SDM R package biomod2 (Thuiller et al, 2014). Firstly, the initial observation point locations were used to extract information from the set of predictor variables creating a new table of the environmental conditions at each observation point. This new dataset was later used in the models to determine how the predictor variables could be fit together to create the best explanation of the variability in distribution.

The models also made use of a set of pseudo-absence data. The pseudo-absence data were created within the biomod2 package making use of the "surface range envelope" (SRE) constraint. The SRE constraint means that "pseudo-absences candidates have to be selected in conditions that differs from a defined proportion of presences data" (Thuiller et al, 2014). This means pseudo-absences are selected outside of the broadly defined environmental conditions for the species. This procedure can improve the efficacy of a model compared to random sampling, but a potential drawback to this approach is that it can lead to an overfitting of the data.

However, the Pacific Golden Chanterelle's distribution over the study area makes this approach more viable. Given that the species is unlikely to be found east of the Cascade Mountains where there is a much different climate, having a pseudo-absence scheme that can reflect this reality should help the model.

Now with presences, pseudo-absences, and an effort variable weighting scheme several different model types were tested. For this project four different procedures were used: Maxent, general linear model (GLM), random forests, and artificial neural network (ANN) (Leo Breiman. 2001, McCulloch; W. S., & Pitts, W., 1943; Phillips, S. J., Dudik, M. & Schapire, R.E., 2004).

Model Tuning:

It was important to slim down the amount of predictor variables being used to reduce the chance of overfitting. The procedure used to choose the most important variables was as follows. Each model was run through a pseudo bootstrapping procedure, using all the predictor variables and a random split of 80% of the input data 10 times. From each trial the importance values of the predictor variables were recorded, summed, and averaged to get the relative importance values of the variables in the models. A graph of the relative importance values can be found in the Results section (Figure 2). These values were used to discard some variables that seemed to have little positive impact on the predictive power of the model. The monthly average precipitation layers as well as the soil class layer were put aside while the next step was run with the remaining selection.

To thin out the variables further, the package SDMtune was used on the remaining predictor variables (Vignali et al, 2020). This process compared the variables against each other. By comparing the variables against each other it checked for correlated data, and as other studies have done, data that was above a 70% threshold of correlation were removed (Botella et al, 2018). This left the following 12 variables shown in Table 2, which were then used for the final modeling scheme in the same manner as described in the model creation section.

Table 2: List of predictor variables after removing the correlated layers along with data source and resolution (Citations for sources consistent with Table 1).

Variable	Description and Source
Bio 3	Isothermality (WorldClim)
Bio 4	Temperature seasonality
Bio 5	Max Temp. of warmest month
Bio 8	Mean temperature of wettest quarter
Bio 9	Mean temperature of driest quarter
Bio 15	Precipitation seasonality

Variable	Description and Source
Cation Exchange Capacity	Cation exchange capacity at 5-15cm depths (soilsgrids.org)
DEM	Digital Elevation model (USGS)
NLCD	National land cover dataset (MRLC)
Canopy Cover	Estimated canopy cover (MRLC)
Aspect	Direction the land is facing (Derived from DEM)
Slope	Steepness of the landscape (Derived from DEM)

Statistical Comparisons:

Once the models were created and run, a quantitative comparison of output performance was carried out. Citizen science data was used to assess the accuracy of the models without needing to subset herbarium points from the initial dataset. This process meant the accuracy was assessed using data independent from what was used to create the model. Citizen science provided an easy way to get a second data set. While the data was not systematically sampled, the cost effectiveness and the scope of that was acquired made up for this shortcoming. As discussed above, the iNaturalist platform was used for this second dataset (Figure 1). These

observations are all georeferenced and there is in built community identification which helps to raise the quality of the data.

The iNaturalist data was downloaded from within the study area and cleaned up much like with the herbarium records. Whereas with the initial data needed to have duplicates removed however, the focus with citizen science data was on quality. iNaturalist already provides a grading feature which lets users easily sort by the number of other users who agree with an identification. This simplified the process and allowed extraction of only the most reliable observations in the study area. Using this narrowed list gave a total of 259 validation points. Now a comparison could be carried out between the output of the models and the citizen science points to get an independent measure of model accuracy.

Liu et al. (2011), lay out a comprehensive list of statistical tests performed when validating the accuracy of a species distribution models. Liu it al (2011) continues that when examining the effectiveness of individual models, the kappa statistic and Area Under Curve (AUC) measures are suitable for that. While for comparison between models the True Skill Statistic (TSS) and Equitable Threat Score (ETS), which adjusts for successes due to random chance, can serve as measures to compare the different modeling procedures.

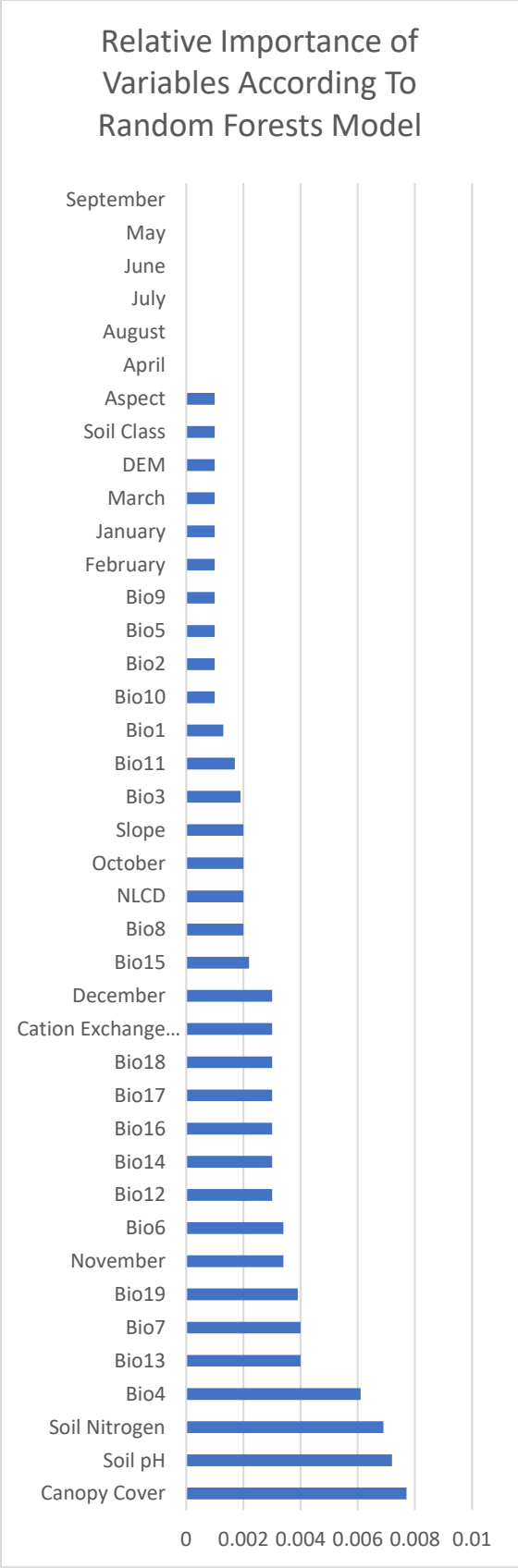
Results and Discussion:

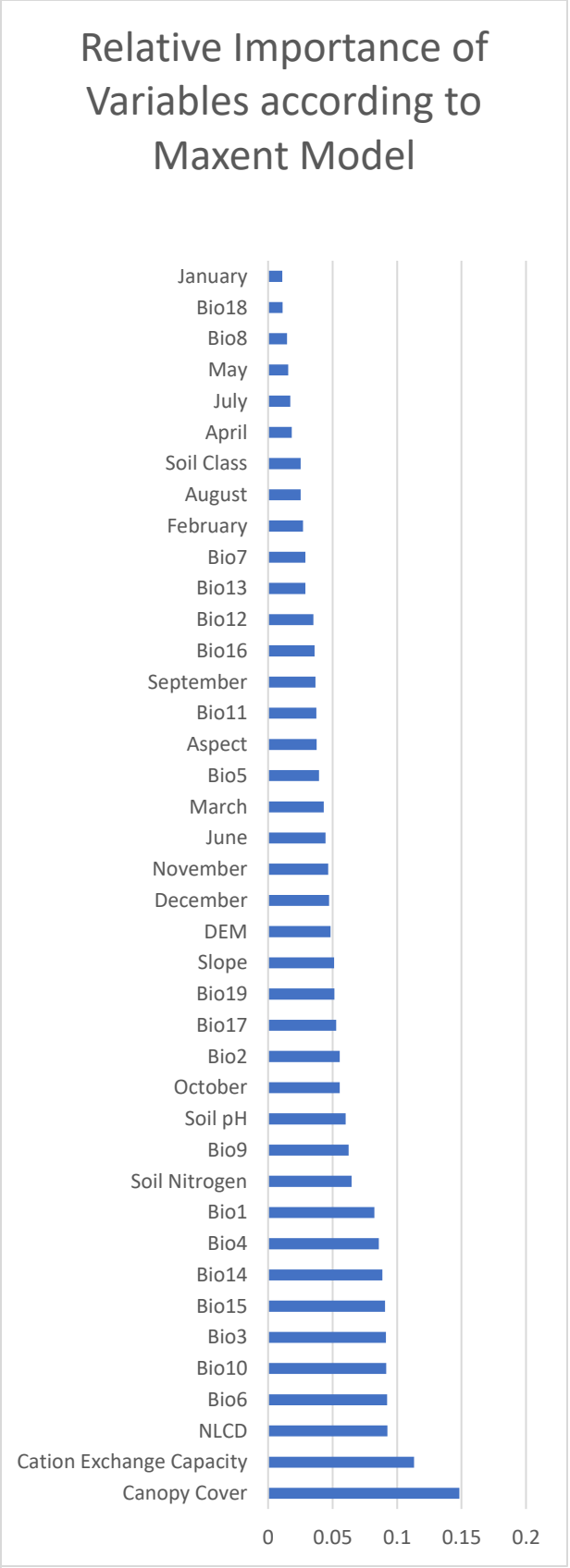
Model Tuning Results:

The results from the model tuning process are shown below in the following figures (Figure 2). These were produced in the initial testing of the models using all variables and give a broad idea about which variables are useful. From these charts some trends are visible amongst the main groupings of variables (Table 1). The soil information was rated very highly, and in particular cation exchange capacity did best. However, soil class was consistently near the bottom. The average monthly precipitation layers were also seen to be less important, with all models showing one layer of this information as the lowest rated item. This was even more drastic with the RF and ANN models with RF ranking 6 of the precipitation layers in last with no value, and ANN ranking 11 with no value. The Bioclim variables, DEM and its derivatives, and

the NLCD layers were ranked sporadically in the middle and sometimes being in the top few. The main exception being canopy cover which was consistently in the top few variables.

With the Bioclim, DEM, and NLCD layers being scattered throughout the middle of the importance rankings it makes slimming down the variable list on this information alone a challenging task. However, the monthly precipitation layers and soil class information were not seen to be adding to the model and could be taken out of the list from this information. To deal with the variables whose trends could not be easily discerned the SDMtune package was used to compare the variables against each other and produce the final set of layers (Table 2). Canopy cover and cation exchange capacity made it into this refined list and were seen to be useful information. The ten other layers may not have been top candidates according to Figure 2, but they represent unique data layers and should be capturing the majority of the information contained within the full set.


Relative Importance of Variables According To Random Forests Model
September
May
June
July
August
April
Aspect
Soil Class
DEM
March
January
February
Bio9
Bio5
Bio2
Bio10
Bio1
Bio11
Bio3
Slope
October
NLCD
Bio8
Bio15
December
Cation Exchange...
Bio18
Bio17
Bio16
Bio14
Bio12
Bio6
November
Bio19
Bio7
Bio13
Bio4
Soil Nitrogen
Soil pH
Canopy Cover
0
0.002
0.004
0.006
0.008
0.01


Relative Importance of Variables according to Maxent Model
January
Bio18
Bio8
May
July
April
Soil Class
August
February
Bio7
Bio13
Bio12
Bio16
September
Bio11
Aspect
Bio5
March
June
November
December
DEM
Slope
Bio19
Bio17
Bio2
October
Soil pH
Bio9
Soil Nitrogen
Bio1
Bio4
Bio14
Bio15
Bio3
Bio10
Bio6
NLCD
Cation Exchange Capacity
Canopy Cover
0
0.05
0.1
0.15
0.2

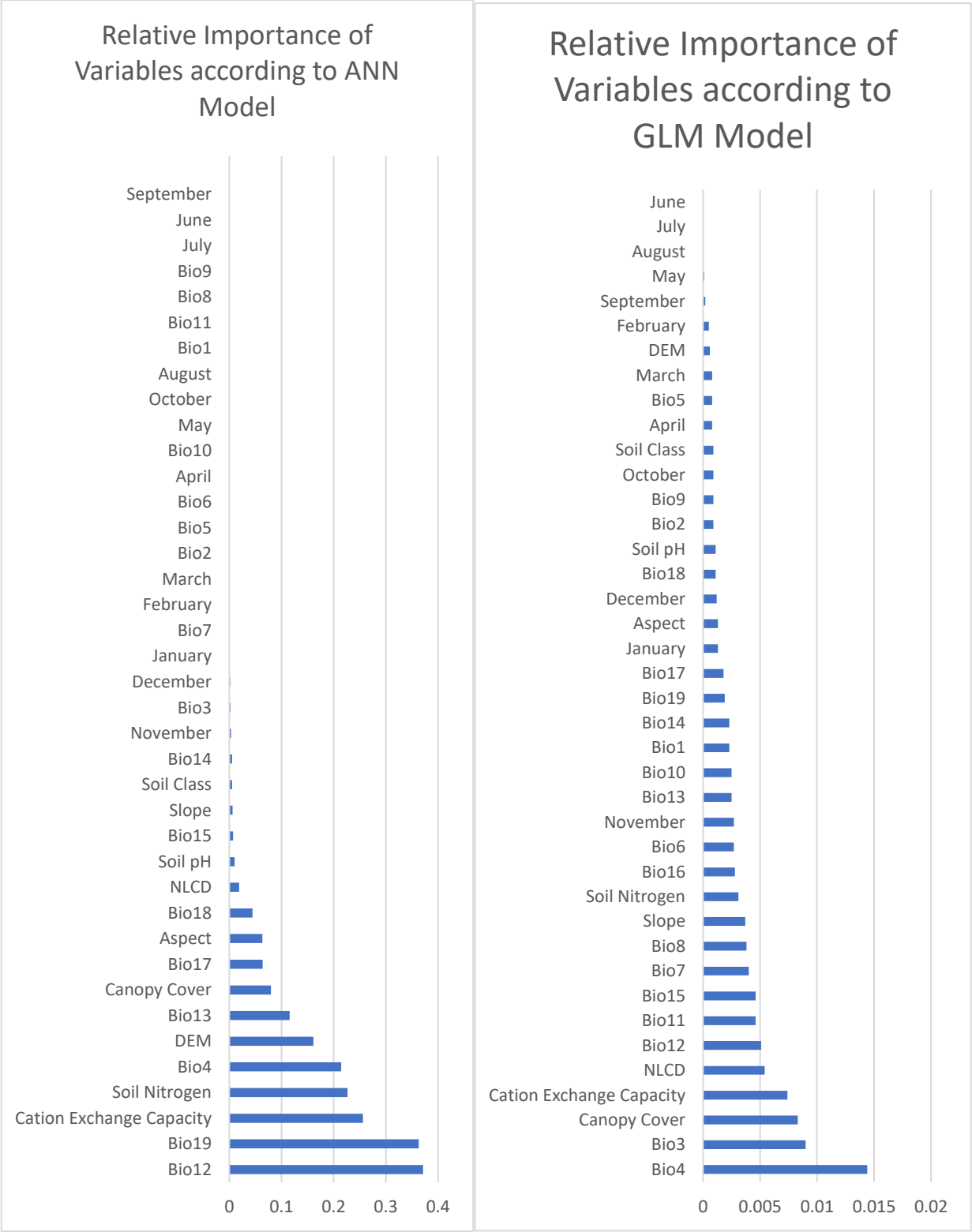


Figure 2: The outputs from running a bootstrap like procedure are best at showing which layers were consistently best and worst, it is less diagnostic of the middle layers. It does help identify the precipitation layers and soil class layer as performing poorly and the canopy cover as performing well.

Model Outputs and Stats:

The four different modeling procedures were then run using refined list of 12 layers selected from the initial of 40 variables, all the herbarium points, and 200 pseudo-absence points. For two of the models, artificial neural network (ANN) and general linear model (GLM), the effort variable weighting scheme was implemented. The Maxent and random forest (RF) models could not make use of the effort variable because this functionality was not available for those models when the biomod2 package was being developed. Differences between the models with and without the effort variable will be shown in the next section.

The outputs of all four models shown below each show similar patterns (Figure 3). With the highest probability of suitable habitat in the Oregon Coast Range and Cascade Mountains of Oregon. There is less predicted habitat in the valleys which are more heavily populated and have less suitable canopy cover. Looking at the statistics, RF and Maxent scored much higher in all measures (Table 3). In particular, the True Skill Statistic (TSS) and Equitable Threat Score (ETS) which are being used to compare between models are exceptionally high with RF getting a near perfect score of 0.987. Also, the RF and Maxent models scored perfectly with respect to specificity meaning they correctly avoided areas of non-habitat. From these results RF was the best performing model, followed by Maxent, ANN, and GLM. Although there may be some issues with this analysis that will be discussed in the following sections.

Another thing to note is the large difference in ETS and Kappa values between the models that did not utilize the effort variable (Maxent and RF) versus those that did make use of it (ANN and GLM). The two models that incorporated the effort variable scored lower, but still had respectable values of 0.571 (ANN) and 0.511 (GLM). This suggests there is a moderating effect from the inclusion of the effort variable. More on this in the following sections.

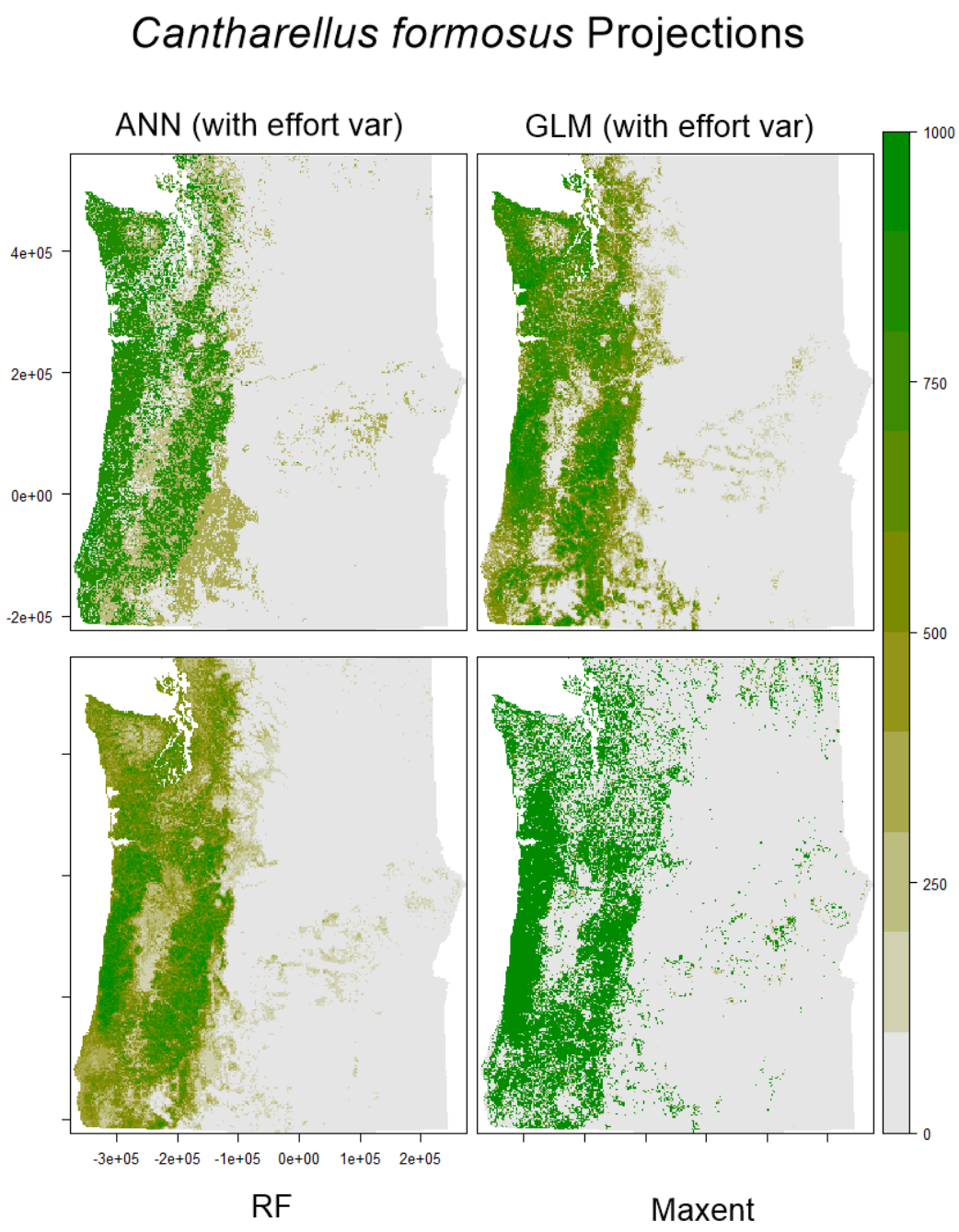


Figure 3: The four model projections, two of which (ANN, GLM) make use of the effort variable. They all display similar trends.

Table 3: The statistical outputs of the models seen in Figure 2. RF scored the best with the near perfect ETS score of 0.987. There also seems to be a trend between the effort var models (ANN, GLM) and those without the effort var. With the effort var models scoring much lower.

	Maxent		
	Test Result	Sensitivity	Specificity
Kappa	0.909	89.06	100.000
TSS	0.891	89.06	100.000
ROC	0.951	89.06	100.000
ETS	0.832	89.06	100.000

	RF		
	Test Result	Sensitivity	Specificity
Kappa	0.994	99.22	100.000
TSS	0.992	99.22	100.000
ROC	0.999	99.22	100.000
ETS	0.987	99.22	100.000

	ANN		
	Test Result	Sensitivity	Specificity
Kappa	0.727	79.69	92.000
TSS	0.757	97.66	78.000
ROC	0.916	97.66	78.000
ETS	0.571	79.69	92.000

	GLM		
	Test Result	Sensitivity	Specificity
Kappa	0.676	92.97	78.000
TSS	0.714	98.44	73.000
ROC	0.917	92.97	78.500
ETS	0.511	92.97	78.000

Figure 4 shows a breakdown of how each of the models utilized the variables. Bio 4, temperature seasonality, was consistently rated very highly. As was canopy cover and Bio 15, precipitation seasonality. Compared to the earlier bootstrapping results, there are some similarities (Figure 2). Canopy cover and cation exchange capacity remained as the most important of the physical variables, and Bio 4 which was one of the more important bioclim variables earlier is seen as top used layer in most of the models. It is also interesting to see that on the whole all the models except for ANN made much more use of the bioclimatic variables compared to the physical measures. While fungi fruiting is undeniably linked to the climate, local knowledge of where to find the Pacific Golden Chanterelle tends to focus more on physical characteristics such as aspect, slope, and canopy cover (OregonDiscovery). In the case of this project, these site-specific characteristics may be lost given the scale and resolution of the data layers and that could be reflected in the importance values.

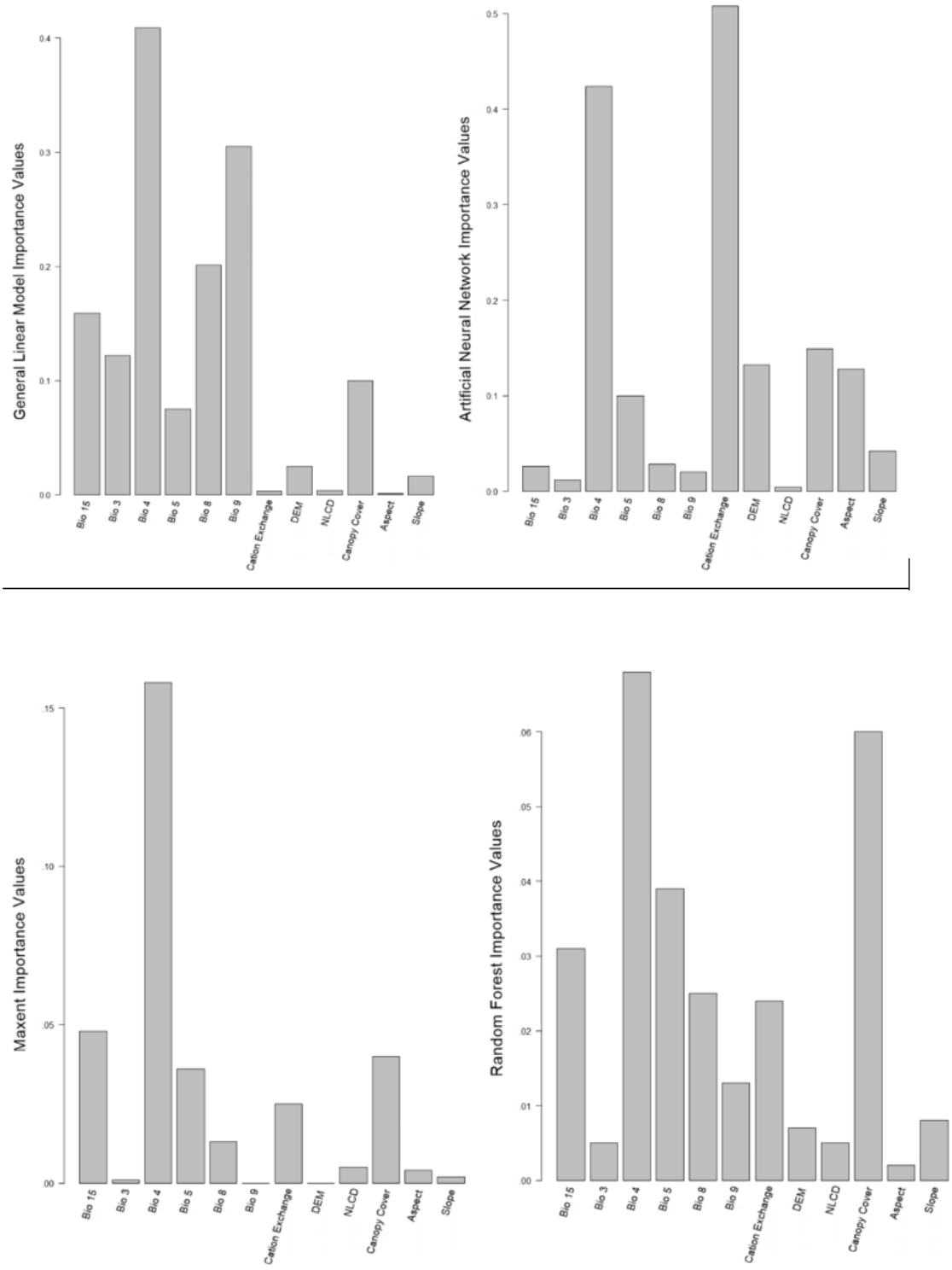


Figure 4: A breakdown of how each variable was used within the models. The bioclimatic variables were overall the more important variables for all the models besides ANN. However, canopy cover was found to be the most important non-bioclimatic measure.

Effectiveness of the Effort Variable:

The inclusion of an effort variable into the ANN and GLM models was to help combat sampling bias present in the herbarium observation points. Since the observations were collected from herbarium datasets in an unsystematic fashion, there ended up being a bias towards population centers. Looking at the map outputs it is clear to see that the effort variable help to limit the scope of the projections (Figure 5). This limiting effect may also explain the lower ETS and Kappa scores compared to the RF and Maxent models, since with the addition of the effort variable there seems to be less overfitting of the data ending up in a blanketing of the study area. However, it is surprising to see that the statistics for the non-effort variable models do not change significantly, given the noticeable visual change between the models with and without the weighting (Table 4).

Table 4: Statistical outputs for the models seen in Figure 5. There was not much difference between the outputs, the biggest change was in the sensitivity vs. specificity measures for with and without the effort var.

	ANN with effort var		
	Test Result	Sensitivity	Specificity
Kappa	0.727	79.69	92.000
TSS	0.757	97.66	78.000
ROC	0.916	97.66	78.000
ETS	0.571	79.69	92.000

	GLM with effort var		
	Test Result	Sensitivity	Specificity
Kappa	0.676	92.97	78.000
TSS	0.714	98.44	73.000
ROC	0.917	92.97	78.500
ETS	0.511	92.97	78.000

	ANN without Effort Var		
	Test Result	Sensitivity	Specificity
Kappa	0.732	96.88	80.000
TSS	0.769	96.88	80.000
ROC	0.911	96.88	80.000
ETS	0.577	96.88	80.000

	GLM without effort var		
	Test Result	Sensitivity	Specificity
Kappa	0.672	94.53	76.000
TSS	0.710	94.53	76.000
ROC	0.913	94.53	76.500
ETS	0.507	94.53	76.000

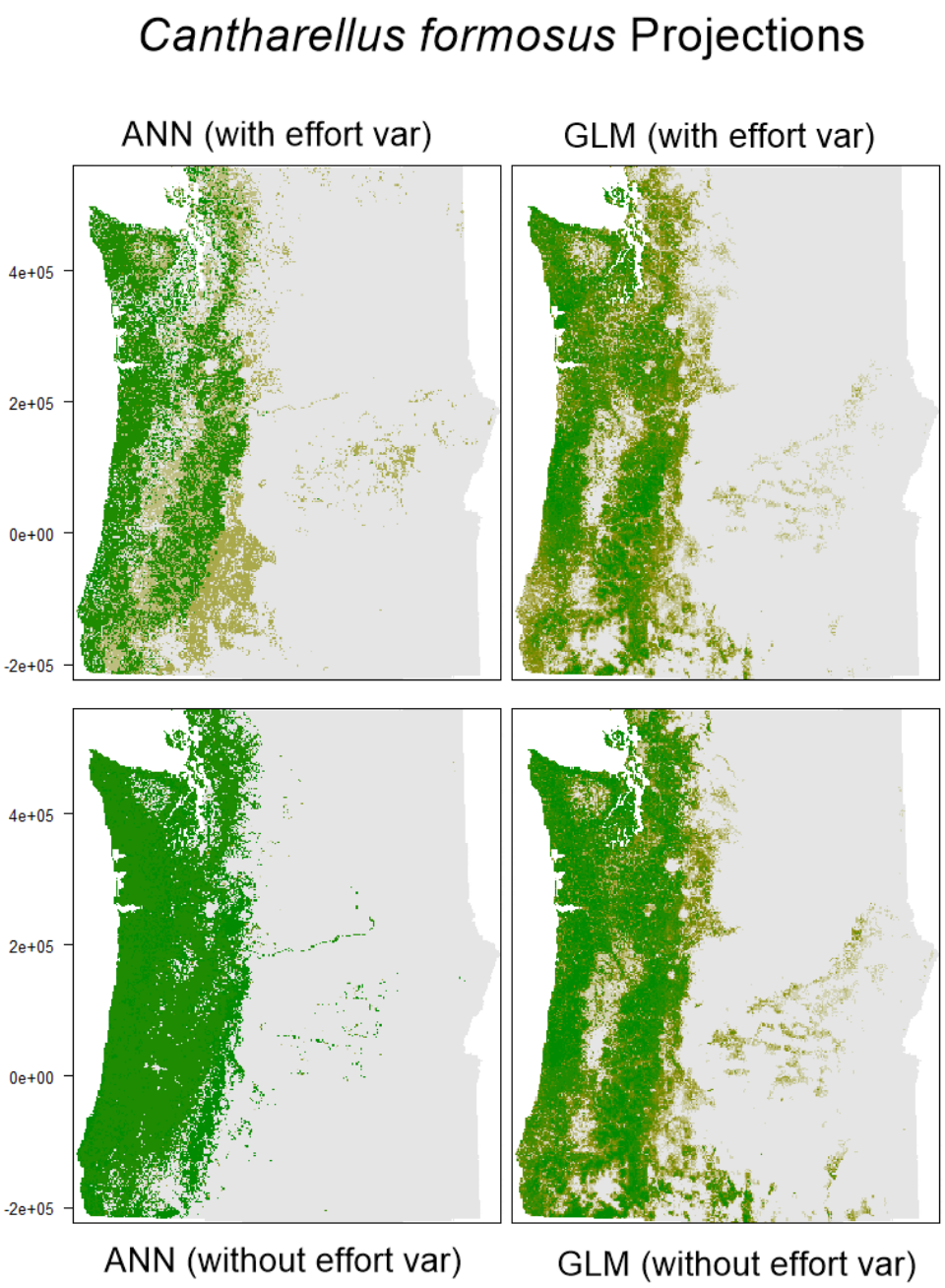


Figure 5: A comparison between the model projections with and without the effort variable. The effort variable seems to help avoid overfitting the study area for both model types.

Something that might be driving the similarity between the main ETS and TSS scores is the difference in specificity and sensitivity between the tests (Table 4). The effort variable runs had a higher degree of sensitivity, meaning they were better at locating areas where the ground truth iNaturalist points were, but struggled with specificity indicating some degree of overfitting areas with no validation points. This difference is most pronounced in the two ANN tests which also had a greater visual difference, but it also can be seen in the GLM tests to a slight degree as well. In the case of the ANN tests, running the model without the effort variable leads to slightly higher TSS and ETS scores. While the scores are not as high as the Maxent and RF runs this increase follows the trend of higher scores when there is a higher degree of overfitting.

Discussion of Model Performance:

There are a few potential problems with the measures of accuracy and model performance. As noted in the sections above, the RF and Maxent models greatly outscored the GLM and ANN models. One factor that may be leading to this trend is the inclusion of the effort variable in the GLM and ANN models. The goal of the effort variable was to discourage fitting the model in lower effort areas and it was able to achieve this (Figure 5). This may have had an adverse effect with respect to accurately measuring model performance using the separate iNaturalist validation dataset. The iNaturalist data had an underlying trend skewing the data towards low effort areas. When compared to the herbarium observations, it is clear to see this divide in the observation types (Table 5, Figure 6). On average the herbarium observations were twice the distance away from major roads than iNaturalist observations. Fitting the model on the higher effort herbarium points while simultaneously discouraging fitting of low effort areas could be leading to an artificial decrease in accuracy when measuring with the lower effort validation dataset. This decrease in the accuracy measure would be more pronounced for the ANN and GLM models which made use of the effort variable. With this potential limitation in mind, future studies may be able to assess model performance more accurately by combining the herbarium and iNaturalist datasets and using a random subset of the combined points as a validation set. This methodology may lessen the effects of this spatial biasing issue.

Table 5: The mean and median distance to a major road for each dataset. The iNaturalist data were much closer to the major roads than the herbarium points.

Herbarium Observations		iNaturalist Observations	
Mean	8.34 (Km)	Mean	4.54 (Km)
Median	8.56 (Km)	Median	2.65 (Km)

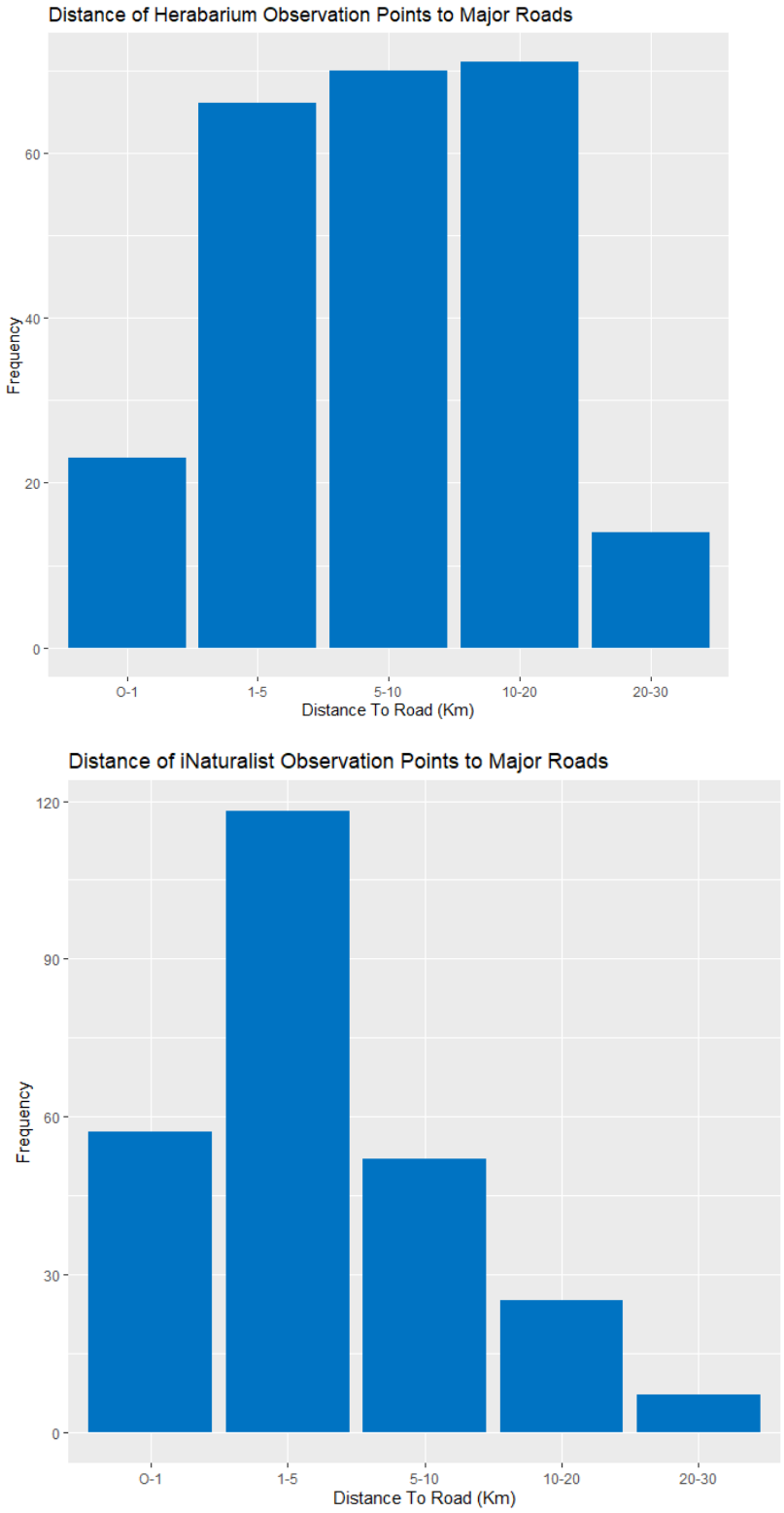


Figure 6: Distribution of distance to roads for the observation points of each dataset. The herbarium data is less skewed creates a more normal distribution pattern.

With this limitation in assessing the accuracy of the models it is difficult to determine which model truly performed the best. Going purely based on the main test values may be misleading, but by looking at the specificity and sensitivity values along with visually assessing the model projections some conclusions can be made. Visually assessing the models shows that the RF and GLM models present the smoothest surfaces with a gradient of habitat suitability. These outputs may be more useful for management decisions compared to the discontinuous output of ANN and all or nothing output of Maxent. While it is unclear whether the RF and Maxent model accuracies can be trusted, the ANN accuracy scores are still fairly high and have a higher focus on specificity. Meaning while the ANN model may not show the full range of habitat, the habitat it finds is likely trustworthy potentially making it a useful output as well.

Conclusion:

This project has sought to demonstrate and explore four main questions surrounding the mapping of the Pacific Golden Chanterelle:

- Whether natural history data could be used to successfully model distributions of fungi?
- Which modeling procedure best predicts distribution when using these herbarium datasets?
- How the use of citizen science could enhance the project?
- And how the inclusion of an effort variable would affect the models?

Overall, the project was met with moderate success in mapping the distribution of the species while solely making use of natural history data. One thing that was demonstrated well through this project, is that there is a wealth of natural history data available to conduct studies such as this one. In this study, a total of 387 observations were obtained and more could be obtained in the future as digitizing efforts within herbaria expand. Given the difficulties with applying traditional sampling techniques towards macro-fungi, making use of this data is possibly the best way to examine the natural histories of these species on a large scale. So, it’s

hopeful to see that the study yielded results which held true to local knowledge and performed reasonably well under statistical analysis.

The outputs of the four model types serve as a good proof of concept for the techniques. Of the four model types, the random forest model preformed best in the validation testing, with an ETS score of 0.987. The score was near perfect however, and visually assessing its model output shows that it blankets most of the study area raising some questions of overfitting and model evaluation issues. Following the random forest model, was Maxent (ETS = 0.832), Artificial Neural Network (ETS = 0.571), and lastly the General Linear Model (ETS = 0.511). All the outputs however, showed similar trends with the most suitable areas within the located in the Coast and Cascade Mountains in Oregon, with a trailing off towards the northern parts of Washington. Overall, the models found the bioclimatic variables, especially Bio 4 (Temperature Seasonality), to be of most value when computing suitable habitat. The ANN model was the exception to this trend, as it utilized the physical measures of the environment to a greater degree. The only physical predictor consistently used by all the models was canopy cover.

The model's heavier usage of bioclimatic variables indicates that the species could come under threat as the climate changes. Under current climate change projections (CMIP6) the Pacific Northwest is expected to get warmer and drier, with more abundant fires (College of the Environment University of Washington, 2017). The changes in climate will impact the bioclimatic variables identified as important in this study and may lead to very different suitable habitat areas in the future. Similarly, the Pacific Golden Chanterelle relies on a mycorrhizal connection with the evergreen trees of the region, represented by the canopy cover layer in the model. A recent study looking at the effects of climate change in the Pacific Northwest found that the areas with the greatest increased risk of fires are in the Cascade Mountains of Oregon (Davis et al, 2017). This area was found to be a highly suitable region for the species in all the models, as well as according to local knowledge. With this area under increased threat, the habitat may shrink down to focus more on the Coast Mountain Range and the Olympic Peninsula. Losing this large swath of suitable habitat may make managing the species more necessary, as it would likely still be heavily sought after. More studies are needed to examine how changes in the region could affect the species range.

In addition to creating the four previously discussed distributions, the study also made use of use an effort variable and iNaturalist citizen science data. The inclusion of the effort variable was meant to help limit the spatial bias towards population centers that was present within the herbarium datasets by weighting observations in higher effort areas more heavily. This weighting scheme could only be applied to the GLM and ANN models, which may explain why they had much lower ETS scores than the other two. There is a noticeably large visual difference between the two outputs of the models with and without the weighting scheme. This difference seems to indicate that the effort variable did indeed help the models avoid overfitting lower effort areas, however this did not change the statistical accuracy scores of the models very significantly. The lack of statistical change between the two tests could be the result of using iNaturalist data to measure accuracy. The method was initially expected to lead to more statistically robust results since the training and testing data was independent, but it may have created some uncertainty in the interpretation of these results. The inherent differences in sampling effort between the two datasets may have led to validation scores that were less diagnostic. A future study could improve upon this process by combining the data before modeling or by using the iNaturalist data in association with the effort variable for the model. Doing so may have helped avoid this uncertainty in the results.

Despite these complications, the study was completed relatively successfully and yielded some promising results. As fungi foraging becomes a larger hobby or as climate change disrupts the natural environment conducting and understanding distributions of at risk or sought-after fungi could become more important. Hopefully, this project can serve as a framework for similar future species distribution modeling studies on macro fungi.

References:

Batjes N.H., Ribeiro E, and van Oostrum Ad (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data* doi: 10.5194/essd-12-299-2020

Bergemann, S. E., & Largent, D. L. (2000). The site specific variables that correlate with the distribution of the Pacific Golden Chanterelle, Cantharellus formosus. *Forest Ecology and Management*, *130*(1), 99–107. https://doi.org/10.1016/S0378-1127(99)00177-2

Bisby, G. R. (1943). Geographical Distribution of Fungi. *Botanical Review*, *9*(7), 466–482. JSTOR.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. (2018). Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences*, *6*(2). doi:10.1002

Buechling, A., & Tobalske, C. (2011). Predictive Habitat Modeling of Rare Plant Species in Pacific Northwest Forests. *Western Journal of Applied Forestry*, *26*(2), 71–81. https://doi.org/10.1093/wjaf/26.2.71

Bureau, U. C. (n.d.). *TIGER Data Products Guide*. The United States Census Bureau. Retrieved March 16, 2021, from https://www.census.gov/programs-surveys/geography/guidance/tiger-data-products-guide.html

College of the Environment University of Washington (2017), Pacific northwest Climate PROJECTION TOOL. https://cig.uw.edu/resources/analysis-tools/pacific-northwest climate-projection-tool/#

Danielson, J. J., & Gesch, D. B. (n.d.). *Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)*. 34.

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M., & Davis, C. C. (2018). Widespread

sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, *217*(2), 939–955. https://doi.org/10.1111/nph.14855

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, *16*(3), 354–362. https://doi.org/10.1111/j.1472-4642.2009.00615.x

Elith, J., & Leathwick, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, *13*(3), 265–275. https://doi.org/10.1111/j.1472-4642.2007.00340.x

Fick, S.E. and R.J. Hijmans, 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37 (12): 4302-4315. /aps3.1029

Findlen, P. (2017). The death of a naturalist: Knowledge and community in late Renaissance Italy. In G. Manning & C. Klestinec (Eds.), *Professors, Physicians and Practices in the History of Medicine* (pp. 127–167). New York, NY: Springer.

Herbaria—Consortium of PNW Herbaria. (n.d.). Retrieved March 16, 2021, from https://www.pnwherbaria.org/herbaria.php

Lavoie, C. (2013). Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics*, *15*(1), 68–76. https://doi.org/10.1016/j.ppees.2012.10.002

Leo Breiman. 2001. Random Forests. Mach. Learn. 45, 1 (October 1 2001), 5–32. DOI:https://doi.org/10.1023/A:1010933404324

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, *34*(2), 232–243. https://doi.org/10.1111/j.1600-0587.2010.06354.x

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, *33*(1), 103–114. https://doi.org/10.1111/j.1600-0587.2009.06039.x

Loiselle, B. A., Jørgensen, P. M., Consiglio, T., Jiménez, I., Blake, J. G., Lohmann, L. G., & Montiel, O. M. (2008). Predicting species distributions from herbarium collections: Does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, *35*(1), 105–116. https://doi.org/10.1111/j.1365-2699.2007.01779.x

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Multi-Resolution Land Characteristics Consortium (U.S.). National land cover dataset (NLCD), USFS Tree Canopy Cover. [Research Triangle Park, NC] :[Multi-Resolution Land Characteristics Consortium].

Peay, K. G., Bidartondo, M. I., & Elizabeth Arnold, A. (2010). Not every fungus is everywhere: Scaling to the biogeography of fungal–plant interactions across roots, shoots and ecosystems. *New Phytologist*, *185*(4), 878–882. https://doi.org/10.1111/j.1469-8137.2009.03158.x

Phillips, S. J., Dudik, M. & Schapire, R.E. (2004). A maximum entropy approach to species distribution modeling. Pages 655-662 *in* Proceedings of the 21st International Conference on Machine Learning. ACM Press, New York

Pilz, D., Norvell, L., Danell, E., & Molina, Randy. (2003). *Ecology and management of commercially harvested chanterelle mushrooms.* (PNW-GTR-576; p. PNW-GTR-576). U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. https://doi.org/10.2737/PNW-GTR-576

Pilz et al. - 2003—Ecology and management of commercially harvested c.pdf. (n.d.). Retrieved February 27, 2020, from https://www.fs.fed.us/pnw/pubs/gtr576.pdf

Prediction of the potential geographic distribution of the ectomycorrhizal mushroom Tricholoma matsutake under multiple climate change scenarios | Scientific Reports. (n.d.). Retrieved February 26, 2020, from https://www.nature.com/articles/srep46221

PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created 4 Feb 2004

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLOS ONE*, *8*(8), e71218. https://doi.org/10.1371/journal.pone.0071218

Stolar, J., & Nielsen, S. (2014). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, *21*. https://doi.org/10.1111/ddi.12279

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0055158

Thuiller, W., Georges, D., Engler, R. (2014). biomod2: Ensemble platform for species distribution modeling. R package version 3.1-64. http://CRAN.R-project.org/package=biomod2

Vignali S, Barras A, Arlettaz R, Braunisch V (2020). "SDMtune: An R package to tune and evaluate species distribution models." *Ecology and Evolution*, **00**, 1-18. doi: 10.1002/ece3.6786.

Wollan, A. K., Bakkestuen, V., Kauserud, H., Gulden, G., & Halvorsen, R. (2008). Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography*, *35*(12), 2298–2310. https://doi.org/10.1111/j.1365-2699.2008.01965.x

Yuan, H.-S., Wei, Y.-L., & Wang, X.-G. (2015). Maxent modeling for predicting the potential distribution of Sanghuang, an important group of medicinal fungi in China. *Fungal Ecology*, *17*, 140–145. https://doi.org/10.1016/j.funeco.2015.06.001