Non Parametric Deepness Metric of Performance applied to Modelling of Niches

Camilo Oberndorfer Mejia[1]
Miguel Valencia Ochoa[2]

Advisors:
Daniel Rojas Diaz[3]
Henry Laniado[4]

Research practice 1 & 2
Research proposal
Mathematical Engineering
School of Applied Sciences and Engineering
Universidad EAFIT

August 2022

[1]Mathematical Engineering Student at EAFIT University coberndorm@eafit.edu.co
[2]Mathematical Engineering Student at EAFIT University(mvalenciao@eafit.edu.co
[3]Mathematical Engineering professor at EAFIT University, drojasd@eafit.edu.co
[4]Mathematical Engineering professor at EAFIT University, hlaniado@eafit.edu.co

1 Introduction

Species distribution models (SDMs) are numerical tools that estimate the relationship between observations of species occurrence records or abundance with environmental or spatial estimates of those sites Elith & Leathwick (2009). Often, niche models are either built with presence-only data (the registered sight of species) because absence data (the registered lack of species) is unavailable Merow & Silander Jr (2014). Presence-only models include MaxEnt and MaxLike, among others. Understanding species' spatial occurrence patterns and their dependence on environmental variables gives us insights to predict niche distributions across landscapes, sometimes requiring extrapolation in space and time. Even though the linkage between SDMs and theory is often weak, they are fundamental goals of ecology and evolution.

This research aims to create a non-parametric function that can potentially solve the problems presented by popular niche models. Functional data can represent the environmental data of each pixel of a map. Functional data is multivariate data that varies over a continuum (F.Heinrichs (2021)). Once the data of the spots of the map and the species occurrences is normalized, a specific deepness measure is applied to the data of species occurrences to create a replacement of a probability density function to color the corresponding map accordingly. The deepness metric used approximates a calculation by simplicial (Lopez-Pintado *et al.* (2014)). This method for niche modeling will then be compared to the already existing parametric methods Maxent and Max-like and checked for possible applications.

As of 2022, the most used niche modeling tool is Maxent and MaxLike. Both Maxent and MaxLike use underlying parametric assumptions, to generate the species distribution (Elith *et al.* (2010) & Merow & Silander (2014)). These assumptions create a bias that can cause incorrect predictions. Non-parametric methods offer solutions to these problems by not basing themselves on assumptions. That means species occurrence data will not use an assumed distribution (ibm (2021)).

2 Statement of the problem

2.1 Statement of the problem

Throughout centuries humans have observed and recorded consistent relationships between species and the environment (M, 2013). Systematic biological survey data is usually limited in most regions because formal biological surveys are costly and sparse. Yet, species records are available in the form of presence-only records in herbariums and museum databases. Many of these databases represent well over a century of public and private investment in biological science (Elith *et al.*, 2010) and are essential sources. Thus the more readily-available presence-only data and the desire to maximize their utility compels interest in Niche Modelling (Stockwell & Peterson, 2002).

The word niche was defined formally by G. Evelyn Hutchinson(1950), who described it as a hypervolume. Hutchinson established that for a set of environmental covariates, each variable had a range of values in which a species can survive. For example, let temperature equal x and humidity equal y. Then for x, there is a value x' above which a particular species could not survive. Similarly, there is another value x'') below which the species could not survive. Likewise, there are two identically defined points for y (Vandermeer, 1972). The region defined by these points in 2-dimensional space is called a Niche. This definition can be expanded into as many dimensions as desired, with every point in this multidimensional space describing states of the environmental variables suitable for the species' survival.

Species Distribution Modeling (SDM), also known as Environmental Niche Modeling (ENM), is the estimation of the hypervolume given by a set of covariates suitable for a species within a landscape of interest L. The objective of niche modeling is to create a colored map showing the probability intensity of the presence of a species on every pixel of the map. The estimations can give way to a better understanding of a species' correlation with its' environment and allow policymakers to make informed decisions (Elith *et al.* (2010)).

The most widely available and generally used methods for estimating a species' spatial occurrence are MaxEnt and MaxLike. Both of these methods share an underlying distribution, the Gibbs distribution, and work by minimizing the distance of the estimation to the occurrence (Merow & Silander, 2014). The main difference is that MaxEnt maximizes the relative entropy while MaxLike maximizes the likelihood. The main problem with both of these methods is the assumption of the distribution. Such suppositions can lead to bias and incorrect estimations or extrapolations. Thus a non-parametric model is a possible solution to avoid these problems.

2.2 Formalization of the problem

In this paper, we will work with presence-only data, i.e., a set of locations within L where we have observed the species. Let Y be a binary variable that expresses $Y = 0$ as absence and $Y = 1$ as presence, and let z denote a vector of environmental covariates in every location within L.

Define $f(z)$ to be the probability density of covariates across L, $f_1(z)$ to be the probability density of covariates across locations within L where the species is present. The quantity to estimate is $f_1(z)$ to solve the probability of the presence of the species, conditioned on environment covariates: $Pr(Y = 1|z)$. Because of the Bayes rule:

$$Pr(Y = 1|z) = \frac{f_1(z)}{f(z)} Pr(Y = 1) \tag{1}$$

Due to the lack of capability to assure the distribution of $f_1(z)$, we will focus on estimating the distribution.

3 Objectives

3.1 General objective

Propose a non-parametric niche modeling technique through depth measures that allow assigning an appropriate probability intensity to a map of environmental covariates from a set of registers of the presence of a species.

3.2 Specific objectives

- Develop a non-parametric method for niche modeling based on a simplicial like depth measure.
- Implement the proposed technique in a Matlab toolbox that creates a colored map based on the probability intensity of a species occurring in each pixel of the map.
- Asses the performance of the proposed method with a diverse set of virtual species results including accuracy and computation time.

4 Justification

The traditional process for studying species' spatial occurrence patterns is a slow and expensive process that can be invasive to some ecosystems; this makes niche modeling a better alternative for conducting these studies. The traditional method consists of excursions to the place of interest with biologists to set traps, kill animals in the ecosystem, and revise them to establish the area's biodiversity. For more information on traditional methods, Hoel (1943) describes the other conventional methods and their disadvantages. Niche modeling, in contrast, lets biologists check for the probability intensity of a species being in an area using the sampled data of similar places in terms of environmental characteristics. Niche modeling presents several practical advantages over traditional sampling methods, making it a crucial tool for the efficiency of these studies and environment conservation.

Some implemented parametric models have issues with a high sampling bias and computation time; for example, the MaxLike model results depicted in Merow & Silander (2014) show poor accuracy with a high computation time. The most explored flaw in parametric models is the sampling bias; this means the model is usually wrong when analyzing uncommon species that are incredibly relevant in biodiversity studies. Non-parametric models can reduce this bias by not depending on parameter estimation but on information about all samples.

A niche modeling method has some possible benefits on price, facility, and time to entities that need or perform studies of species' spatial occurrence. The theoretic qualities of a non-parametric method include a simpler algorithm, a low computation time, and a more robust metric for the map coloring function; These qualities improve the practicality of niche modeling methods as a tool for studying species distributions. These improvements together lead to a different way to tackle map coloring problems based on probability intensity and set the ground for better methods, or at least slight advancement for the future.

This research enriches the literature on niche models as it opens a new investigation path using a non-parametric estimation of probability intensity. Non-parametric estimation is an important step towards replacing old models as new parametric estimation-based models are usually worse than Maxent created in 2004. The used estimator for replacing simplicial calculation of deepness can be applied in similar problems with functional data where the final product is a probability intensity. These added benefits from this research open a whole lot of opportunities to improve the tools for map coloring problems, especially for niche modeling.

5 Scope

The tools for the research will include Matlab toolboxes and functions, reference papers about niche modeling, and premade data in the form of virtual species and ASCII maps containing environmental information of different areas. The main Matlab tools for the research will be the Matlab shapes toolbox (mat, 2022) and a Matlab function developed by assessor Daniel Rojas Diaz and student Luisa Toro for virtual species generation. Also, we will use the Maxent software to check the accuracy of the results of the map coloring function. These tools give the method an easy-to-access and easy-to-use algorithm for comparisons with future work.

Although the proposed method for the research can solve some of the issues with the parametric estimation of $f_1(z)$, it can still share or even create problems of its own. For example, Its' accuracy has a high dependence on the accuracy of the sampled data of environmental characteristics and species spatial occurrences as it uses this data to create the colored map. Its dependency on the sampled data also brings an issue on accurately predicting species occurrences in places with not very explored niches. Either way, Using sampled data from niches instead of sampled areas has the advantage of being a better estimator for habitats that share characteristics with the more explored areas. There is little information about non-parametric niche modeling methods, making a comparison to a more similar model virtually impossible. As Elith & Leathwick (2009) says, the factors that influence the robustness and accuracy of the model are the extents of the extrapolation, the interplay between environmental variables, and the consideration of scale. These factors leave room for improvement of the method in future research.

With the mentioned limits and tools, the research's expected results are the following: A map coloring function for probability intensity of species spatial occurrence in a Matlab toolbox for future applications and modifications of the method. The results contrast the colored maps generated by Maxent software to check for the hypothesized accuracy and computation time. A paper detailing the process and tools ultimately used for the research and recommendations for future improvement. These deliverables aim to set a clear ground base for future non-parametric methods that attempt to create colored maps based on probability intensity for niche modeling or other similar subjects.

6 State of the art

The study of species spatial occurrence is widely used as a decisive factor when establishing agricultural centers, roads, or constructions (Gaston & Fuller, 2009). Government and non-government organizations have adopted these methodologies for large-scale, real-world biodiversity mapping applications, like humming-birds distribution prediction for conservation planning (Tinoco *et al.*, 2009), environmental correlates of the European Wildcat (Monterroso *et al.*, 2009) or ants potential invasive expanding distribution (Ward, 2006).

The idea of correlating a species distribution with environmental/geographical features has long been studied. Early work includes The geographical distribution of mammals, by Murray (1866), and Pflanzen-geographie auf physiologischer Grundlage by Schimper (1898). Early quantitative approaches used multiple linear regression and linear discriminant function analyses to associate species and habitat (Capen, 1981). Application of generalized linear models (GLMs) allowed for non-normal error distributions, additive terms, and nonlinear fitted functions (Beery *et al.*, 2021).

The distribution of a species is typically represented as a map indicating the spatial extent of the species. These are classified into three categories according to Beery *et al.* (2021). Firstly is raw species observation data, which shows all the locations where a species is known to be present or absent. Secondly are statistical models, which encapsulate SDMs and are concerned with the correlation of survey data with environmental characteristics. Thirdly are expert range maps, based on various information sources, including personal statements, understanding of the species' habitat preferences, and local knowledge, among multiple other sources. This paper will focus on the map coloring of statistical models.

In equation 1 we can see three terms, these are $f_1(z)$, $f(z)$ and $Pr(Y = 1)$. Some methods take $Pr(Y = 1)$ as a constant since it corresponds to the species prevalence. A rough estimate for this can be the proportion of occupied sites by the species (Fernandez-Manjarrés, 2018). Naturally, different ways of estimating $Pr(Y = 1)$ have arisen, like MaxEnt and MaxLike.

As previously mentioned, both MaxEnt and MaxLike are states of the art regarding niche modeling. MaxEnt produces a suitability index because of the way $f(z)$ and $f_1(z)$ are defined. MaxEnt assumes $f_1(z)$ can be found through the Gibbs distribution, as $f_1(z) = f(z)e^{\eta(z)}$ where $\eta(z) = \alpha + \beta * h(z)$. α is a normalizing constant that ensures $f_1(z)$ sums to 1, and β is a vector of coefficients applied to the different terms of the model. Finally $h(z)$ is the constrained covariates (Fernandez-Manjarrés, 2018). Yet, since MaxEnt does not produce estimates of occurrence probability (Elith *et al.*, 2010), it is not suitable for making explicit predictions of an actual state variable or testing hypotheses about factors that influence occurrence probability.

7 Proposed methodology

Species spatial distribution pattern recognition problems usually are solved via three alternative processes. Yet, the traditional methods are underdeveloped, being the most invasive and expensive among the options. Furthermore, neural networks trained via individual samplings of the species have a high computational cost and require sizeable samples for training. Lastly, there are probability intensity models that don't need as much data, nor are they as invasive ecosystems as traditional methods.

Probability intensity models share a data pre-processing procedure that considers this research. This process starts by organizing data in a matrix comparing the values of environmental variables for all samples of the presence of a species. Then the data is normalized to prevent a scale bias for certain variables. The normalization selected is range normalization as it is simple to apply and has little computation time. It also transforms values from a big scale to values between zero and one, representing the position in the interval from its minimum sampled value to the maximum sampled value. Alam (2020) gives a complete explanation on range normalization. After this process, the model in itself begins.

The non-parametric and parametric methods differ mainly in the probability intensity calculations. Parametric methods try to find parameters from the normalized data for an underlying distribution. The proposed alternative for probability intensity with non-parametric models is a depth measure. Simplicials are the most common non-parametric method for calculating depth, but it has a high computational time. To replace simplicial calculations, we propose a different approach described ahead.

A Simplicial based calculation for depth can be beneficial for a non-parametric model as it changes the computation time and can make the model more simple or complex to apply. The alternative proposed is to; firstly reduce the data dimensionality with Principal component analysis (PCA), then construct a frontier that encloses all samples while reducing the space lacking occurrences. Abdi & Williams (2010) provides a good explanation of the PCA technique. With the calculated boundary, we measure the distance R_i between it and every sample i. Then we calculate the frontier and construct a sphere with the center on the occurrence and radius R_i. Finally, the sphere for every sample represents its influence.

To calculate the colored map. We calculate the depth of every pixel on the map as the number of spheres it is inside. This depth measure is a first step in making non-parametric models useful for niche modeling. Still, it can be experimented with by using different distances of the samples or even using alternative metrics in future research.

Table 1: Schedule

Activity	*Weeks*																	
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Define workflow and methodology	■	■	■															
Literature check				■	■	■	■	■										
Toolbox developing									■	■	■	■						
Validation													■	■	■			
Discussion																■	■	■

8 Schedule, commitments, and deliverables

Meetings will be held every week Monday at 2:00 pm with both students and Advisor Daniel Rojas Diaz and occasional meetings with advisor Henry Laniado. The expected deliverable for the research is a Matlab toolbox improvement and a paper detailing the results of it tested with multiple data.

9 Intellectual property

According to the internal regulation on intellectual property within Universidad EAFIT, the results of this research practice are product of *Miguel Valencia Ochoa*,*Camilo Oberndorfer Mejia* and *Daniel Rojas Diaz*.

In case further products, besides academic articles, could be generated from this work, the intellectual property distribution related to them will be directed under the current regulation of this matter determined by Universidad EAFIT (2017).

References

2021 (Aug). *Statistics - parametric and nonparametric.*

2022. *Draw Shapes and Lines - MATLAB Simulink - MathWorks América Latina.*

Abdi, Hervé, & Williams, Lynne J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, **2**(4), 433–459.

Alam, Rizwan. 2020. Normalization vs Standardization Explained. *Medium*, May.

Beery, Sara, Cole, Elijah, Parker, Joseph, Perona, Pietro, & Winner, Kevin. 2021. Species Distribution Modeling for Machine Learning Practitioners: A Review. *Page 329–348 of: ACM SIGCAS Conference on Computing and Sustainable Societies*. COMPASS '21. New York, NY, USA: Association for Computing Machinery.

Capen, David E. 1981. *The use of multivariate statistics in studies of Wildlife Habitat*. Rocky Mountain Forest and Range Experiment Station, Forest Service, U.S. Dept. of Agriculture.

Elith, Jane, & Leathwick, John R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**(1), 677–697.

Elith, Jane, Phillips, Steven J., Hastie, Trevor, Dudík, Miroslav, Chee, Yung En, & Yates, Colin J. 2010. A statistical explanation of Maxent for ecologists. *Diversity and Distributions*, **17**(1), 43–57.

Fernandez-Manjarrés, Juan. 2018. Using Ecological Modelling Tools to Inform Policy Makers of Potential Changes in Crop Distribution: An Example with Cacao Crops in Latin America. *Economic Tools and Methods for the Analysis of Global Change Impacts on Agriculture and Food Security*, 11–23.

F.Heinrichs. 2021. Functional Data Analysis(FDA). *towards science*, **1**(1), 1–3.

Gaston, Kevin J, & Fuller, Richard A. 2009. The sizes of species' geographic ranges. *Journal of applied ecology*, **46**(1), 1–9.

Hoel, Paul G. 1943. The Accuracy of Sampling Methods in Ecology. *The Annals of Mathematical Statistics*, **14**(3), 289–300.

Lopez-Pintado, Sara, Sun, Ying, Lin, Juan, & Genton, Marc. 2014. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, **8**(09), 321–338.

M, Basvarajaiah D. 2013. *Biodiversity Modeling and Tribal Livelihood Status in Western Ghats*.

Merow, Cory, & Silander, John A. 2014. A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution*, **5**(3), 215–225.

Merow, Cory, & Silander Jr, John A. 2014. A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution*, **5**(3), 215–225.

Monterroso, P., Brito, J. C., Ferreras, P., & Alves, P. C. 2009. Spatial ecology of the European wildcat in a Mediterranean ecosystem: dealing with small radio-tracking datasets in species conservation. *Journal of Zoology*, **279**(1), 27–35.

Murray, Andrew. 1866. *The geographical distribution of mammals*. Day and Son.

Schimper, Andreas Franz Wilhelm. 1898. *Pflanzen-geographie auf physiologischer Grundlage*. Vol. 2. G. Fischer.

Stockwell, David R.B, & Peterson, A.Townsend. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**(1), 1–13.

Tinoco, Boris A., Astudillo, Pedro X., Latta, Steven C., & Graham, Catherine H. 2009. Distribution, ecology and conservation of an endangered Andean hummingbird: the Violet-throated Metaltail (Metallura baroni). *Bird Conservation International*, **19**(1), 63–76.

Universidad EAFIT. 2017. *Reglamento de propiedad intelectual*.

Vandermeer, John H. 1972. Niche Theory. *Annual Review of Ecology and Systematics*, **3**, 107–132.

Ward, Darren F. 2006. Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions*, **9**(6), 723–735.