Virtual species generation for species distribution modelling

Luisa Toro Villegas[1]

Advisor:
Daniel Rojas Díaz[2]

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

SEPTEMBER 2022

[1]Mathematical Engineering student in EAFIT University, ltorov@eafit.edu.co
[2]Mathematical Engineering professor in EAFIT University, drojasd@eafit.edu.co

1 Introduction

Species distribution modelling (SDM) explains species' spatial distribution and the environment's influence on their presence. SDM studies the relationships between species and their environment; and the variables of the environment which affect the species' presence. Building SDM models needs maps of geospatial environmental data. Therefore, the method depends on the availability of the data to create accurate and reliable predictions, according to Elith & Franklin (2013). These predictions help make informed decisions for biodiversity research, including controlling the impacts of biological invasions (Walther *et al.*, 2009), identifying threats to the conservation of species (Thuiller *et al.*, 2005), and planning safe spaces for endangered species (Moilanen, 2005).

One of the main issues of SDM is validating the models with species data from real life, which makes virtual species crucial. Virtual species are created by defining their niche, i.e., the environmental conditions it requires to thrive, as a function of environmental variables and by simulating their occurrence on a given map, according to Grimmett *et al.* (2021). SDM has offered helpful species information, including how to scale a model made for a local map into a large-scale or global model; and what determines a species' range. Austin *et al.* (2006) and Austin (2007) suggest researchers to perform virtual experiments on all emerging SDM methods. The authors argue that issues such as bias related to the species and unknown niche dimensions frequently occur when virtual species are not involved while validating.

Previous work on virtual species has focused on three main areas. Firstly, Leroy *et al.* (2016) generated the occurrence probability of a virtual species from a spatial set of environmental conditions and the environmental suitability into presence-absence with a probabilistic approach. Secondly, Qiao *et al.* (2016) studied the Hutchinsonian method of an n-multidimensional space occupied by the species. Thirdly, De Marco & Nóbrega (2018) developed a function of the environmental variables and species Grinnellian niche. In this study, we aim to complement the literature on virtual species by providing a more thorough exploration of the variable space, generating variability in the species simulated. This aspect gives insight into species distribution model validation. This study will contribute a mathematical description of an ecological niche and its parameterization. The resulting model will generate presence-only data, which describes where the species is present but not where it is absent.

2 Statement of the problem

2.1 Statement of the problem

An ecological niche describes the position of a species within an ecosystem, including the necessary conditions for its survival and relationship with it (Polechová & Storch, 2019). Niches are areas where a species can survive by utilizing the resources available and impacting its environment. Visualizing and analyzing the species' distribution within these niches is known as species distribution modelling (SDM) or environmental niche modelling (ENM) (Elith & Franklin, 2013). Artificially designed species, known as virtual species, are commonly used

to study niches and the distributions of the species within due to scarcity of real-life data. Hirzel *et al* coined the term in 2001. To that end, virtual experiments are a set of simulations that control parameters such as sample prevalence (i.e. the frequency of the species over its entire distribution range versus within the sample) Hirzel *et al.* (2001) and the size of the spatial domain (Meynard *et al.*, 2019).

Researchers determine the scale of the problem, as SDMs have many possible uses. The use of the method depends on the purpose of the research. Variability within species is a performance indicator for any virtual species algorithm. The data needed to generate virtual species is gridded spatial data (also known as raster data). *Rasters* are maps divided into small units called pixels. Each pixel has multiple layers of values that combine to form the environmental space.

Generating virtual species is separated into two steps: defining a niche from the surrounding environment and creating presence-only data from it. The first stage entails defining a species' relationship to its environment. Some approaches include dimensionality reduction and response functions. For this research, niches can be a function that assigns a likelihood to each point in the environmental variable space. The likelihood function is then used as a probability density function to generate a sample of the species, with each pixel in the map representing a location in the environmental variable space. Researchers can model the second step using many distributions. An example is the Bernoulli or threshold distribution, where the species is either recorded as present or not according to a preestablished threshold.

However, previous approaches for virtual species have some drawbacks, such as spatial autocorrelation and spatial non-stationarity. The former occurs because most probability-based techniques focus on exogenous processes like abiotic conditions while ignoring endogenous processes like biotic conditions. These factors cause spatial autocorrelation, which violates the assumption of independence used for most statistical procedures in the method. The latter is because the relationships between species and their environments do not remain constant over time.

2.2 Formalization of the problem

Let

$$H = \frac{1}{\sum w_i} \sum w_i H_i + \varepsilon \tag{1}$$

be the environmental suitability or niche of a pixel with $H \in [0, 1]$, H_i the value of the ith partial niche coefficient, w_i the weight assigned to the ith partial niche coefficient, and ε a random variable.

The partial niche coefficients are the habitat suitability determined by a response function. These can be modelled by using a Gaussian function, a linear function, or others.

Then, let

$$T(p) = \begin{cases} 1 & if \quad H \geq p \\ \\ 0 & if \quad H < p \end{cases} \tag{2}$$

be a threshold function, where the environment is considered suitable according to a predetermined threshold p. Then, the species is present; if it is not, no data will be recorded.

3 Objectives

3.1 General objective

Design and parameterize virtual species generation through functions that warp the shape of the environmental variable space to generate diverse sets of virtual species.

3.2 Specific objectives

- Propose a method of virtual species generation with much variability that makes sense from a mathematical and biological standpoint.
- Implement the method for virtual species generation, ensuring a low computational cost and a user-friendly interface.
- Assess the potential of the method for generating virtual species that are unique from one another.

4 Justification

Growing concerns regarding the reduction of biodiversity, particularly as a direct effect of human behaviour, have led to an interest in estimating any future potential damage. As a result, SMS's ability to predict species distributions, particularly in response to potentially fatal factors like global changes and invasive species, has increased its popularity. For policymakers, understanding which areas to protect is an indispensable tool. Therefore, methodological advances are spreading, but it then raises the issue of validating whether the results obtained from a new method are cohesive with reality.

A traditional approach for validation is to use real-life data obtained by experts in the field, where they go to the habitat of a species and record any sightings or evidence indicating their presence. However, this endeavour has many drawbacks when upscaling or replicating the results: the funding required for the sampling trips, the time it takes to obtain statistically relevant data (a sufficient number of samples), the sampling methods used (sometimes traps are set up, endangering the species), or the human cost in and of itself (sampling often needs professionals).

Studying virtual species offers some insights into evaluating SDM model performance. Uniqueness amongst virtual species is a good test for the precision and robustness of SDMs. For example, some models may perform well under species that behave similarly but may be confused by ones that do not. It is beneficial when designing validation tests to control the variability in the different species, which is impossible using real-life samples. It also permits

studying species that would otherwise be extremely hard to study, either because of their habitat or lack of resources.

Personally, protecting species and biodiversity is a great interest of mine. By helping validate a robust and precise SDM method using virtual species, I also hope to help ensure endangered species preservation. Another factor of interest for this project lies in the fact that this is a new field for me, and the knowledge required and the numerical difficulty it yields are valuable to my education. It allows me to integrate the knowledge acquired in my career so far with a personal interest of mine, and it provides an opportunity to explore new methods.

5 Scope

The main aim of this study is to develop and parameterize a method for generating virtual species. We hope to achieve this by exploring different functions that warp the environmental variables' space, such as probability distributions and harmonic functions. In addition, we will examine the variability attained among each virtual species to determine the method's performance; the greater the difference, the better the virtual species. A further measurement is that it has ecological realism (that is, the simulated relationships between virtual species and their environments are similar to the ones found in species), with different levels of complexity determined by the user.

The development of a new SDM model is beyond the scope of this study. However, there are some practical considerations. The first is the computation capacity and efficiency while still attempting to maintain a significant difference between species. The second requirement is for the method to be simple enough for users to use.

Matlab, a programming and numeric computing platform used primarily for data analysis, algorithm development, and model creation, will be the primary tool for this research. The expected outcome is to create a novel approach to characterizing and parameterizing niches.

6 State of the art

Researchers have explained species distributions using environmental and geographical features for a long time, with some early works including *The geographical distribution of mammals*, by Murray (1866) and *Plant Geography Upon a Physiological Basis*, by Schimper (1898), which attempted to explain the expansion of plants based on geography. Afterwards, the concept of an ecological niche appeared, introduced by Hutchinson (1957) and later formalized by Hutchinson *et al.* (1978), which he called biotopes. Species distribution models combined natural history with recent developments in statistics and numerical methods. On another note, population ecology studies showed that modelling individuals of a species instead of a community provide a better understanding of the population behaviour (e.g. MacArthur, 1958).

Modern modelling of species distribution started with two main branches of research. The first is field-based studies of the relationships between a species and its habitat (e.g.

Capen, 1981) using techniques such as multiple linear regression. In the same vein, the use of Generalized Linear Models (GLM) allowed for non-normal error distributions and nonlinear fitted functions (Austin, 1985), features used when implementing more recent methods, such as Resource Selection Functions (RSFs) by Manly *et al.* (2002). The second branch was physical geography, more specifically, geospatial data. Digital models of the earth's surface conditions, such as Geographic Information Systems (GIS), have greatly helped advance SDM estimations. Early work on integrating both branches includes Ferrier *et al.* (1984). They used presence-only data collected by listening to the songs the males of the *Rufous scrub-bird* made and with environmental variables of Australia such as the time of the year, the weather, and the forest type to determine the distribution of the bird.

The last two decades have brought many developments to the area, mainly propelled by the vast amount of data available and the increase in computation capacity. There are two categories of methods: machine learning methods which require data with presence and absence records (such as logistic regression and fitted regression models, Pearce & Ferrier, 2000), or SDM methods which require presence-only data. Maxent is one such method, a maximum-entropy technique that handles many variables and is still competitive with the newest methods available. Governmental organizations have adopted it for real-world biodiversity mapping applications, such as the prediction feature in the Atlas of Living Australia (Belbin, 2012), where the Maxent software predicts where a species might be observed (Elith *et al.*, 2011). However, Maxent does not estimate the probability of occurrence because it assumes that the method cannot determine it with the available data and instead calculates a suitability index. The index is a ratio between the conditional density of the covariates at the present sites and the marginal density of covariates across the study area. Later, Maxlike, a maximum-likelihood approach, asserts that the probability of occurrence can be estimated and builds the model upon a maximum likelihood estimator (Merow & Silander Jr, 2014).

Hirzel *et al.* (2001) proposed simulating virtual species to test the performance of the new methods. They claimed that the new technique allows researchers to control the input data and assess models. In their work, they used virtual species to test the performance of Ecological Niche Factor Analysis (ENFA) compared to GLM, which they later published in Hirzel *et al.* (2002). They found ENFA to be more robust in three different scenarios: when the species spread, stayed at equilibrium, and became overabundant. Meynard & Quinn (2007) replicated their work by comparing statistical models using artificial species. They simulated 18 species with different response functions (e.g. Gaussian, linear, random), from three layers of environmental data and with response interactions such as additive or multiplicative.

Current work includes an R package called *virtualspecies* by Leroy *et al.* (2016). Their contribution was the variability of the method in possible response functions, its probabilistic approach for conversion into presence-absence, its dispersal limitations, and the introduction of biases into the samples. Similarly, Qiao *et al.* (2016) developed a toolkit called *NicheA* that simulates virtual species based on the Hutchinsonian approach of an n-multidimensional space occupied by the species. They integrate five methods: minimum-volume ellipsoids, convex polyhedrons, physiological ranges, and linear and logistic functions. The resulting method was more robust than other existing tools. Finally, De Marco & Nóbrega (2018) defined

a *niche's suitability* as a Gaussian multivariate distribution function of the environmental variables using a spatially explicit automata cellular model. Then, they sampled the presence-only data as a percentage of the suitability area based on the representativeness of a niche.

7 Proposed methodology

The generation of virtual species usually consists of two stages, according to Grimmett *et al.* (2021). Firstly, raster data defines the species' relationship to its environment. It creates a map of the spatial distribution of environmental suitability. Secondly, the suitability is either converted into presence-only or presence-absence data. Because most SMDs use this format, this study will return presence-only data.

Leroy *et al.* (2016) determined the dependence of a specie on each of the layers or environmental variables using response functions to generate virtual species. An example of a response function is the Gaussian distribution, mainly due to its low computational cost. The reaction to each layer of each specie helps estimate its environmental suitability. Then, the algorithm colors the map by assigning a value for each pixel. That is known as a virtual species distribution or a niche. Researchers have used response functions such as linear, quadratic, and logistic. Because the importance of each layer varies between species, it is critical to consider how response functions define niches.

Another option is to perform dimensionality reduction techniques on the environmental variables space. Principal Component Analysis (PCA) is a recurrent method in literature, according to Leroy *et al.* (2016). PCA is an unsupervised linear transformation technique used for feature extraction. It identifies patterns based on the correlation between layers. Mathematically, the principal components are orthogonal axes that represent the direction of the maximum variance of two layers. The technique then transforms a d-dimensional space into a k-dimensional space. The approach for this study is to modify the environmental variable space and find metrics that uniquely define the niche. Similar to PCA, correlations between variables are measured to reduce dimensionality.

Once the niche is defined, the method will convert it into presence-only data. Leroy *et al.* (2016) use a straightforward approach by establishing an environmental suitability threshold. If the species does not meet it, it is absent from the pixel. If it does, the species is present. However, this is an overly simplistic way of looking at it, and it may skew the results of the SDM tests. Researchers solve this by employing a probability approach and calculating the likelihood of the species' presence in terms of suitability. The logistic distribution is an example of a distribution they use for this conversion because it yields more realistic results than a linear or threshold method when tested. It uses two parameters: inflection point β and slope α. Because it is a probability-based method, repetitions will produce different data sets, which will all come from the same distribution.

The methodology of this project will start by using raster data from 19 environmental variables to create a map of normalized climatic variables using the *readgeoraster* method from Matlab. Then, where other methods might have multiplied these variables by a given coefficient to warp the variable space, this project will use beta distributions and harmonic

functions to achieve this. Afterwards, it will randomly choose an initial point and calculate and normalize its distance from each pixel on the map. The next step will be to define an occupation percentage that determines a limit to how close an area is where the species is present. This normalized distance is the probability of presence.

8 Schedule, commitments, and deliverables

Meetings will be held every Monday at 2:00 p.m. with the advisor. The deliverables expected are an improvement to a Matlab toolbox and a paper detailing the procedure.

Table 1: Schedule

Activity	*Weeks*																	
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Literature review	■	■	■	■	■													
Theoretical and numerical studies						■	■	■										
Simulation design									■	■	■	■						
Simulation testing													■	■	■			
Article writing																■	■	■

9 Intellectual property

According to the internal regulation on intellectual property within Universidad EAFIT, the results of this research practice are product of Luisa Toro Villegas and Daniel Rojas Díaz.

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

Austin, Michael Phillip. 1985. Continuum concept, ordination methods, and niche theory. *Annual review of ecology and systematics*, 39–61.

Austin, Mike. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological modelling*, **200**(1-2), 1–19.

Austin, MP, Belbin, L, Meyers, J a A, Doherty, MD, & Luoto, M. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *ecological modelling*, **199**(2), 197–216.

Belbin, Lee. 2012 (Jan). *Predict*.

Capen, David E. 1981. *The use of multivariate statistics in studies of wildlife habitat*. Vol. 87. Rocky Mountain Forest and Range Experiment Station, Forest Service, US ....

De Marco, Paulo, & Nóbrega, Caroline Corrêa. 2018. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PloS one*, **13**(9), e0202403.

Elith, Jane, & Franklin, Janet. 2013. Species Distribution Modeling. *Pages 692–705 of:* Levin, Simon A (ed), *Encyclopedia of Biodiversity (Second Edition)*, second edition edn. Waltham: Academic Press.

Elith, Jane, Phillips, Steven J, Hastie, Trevor, Dudík, Miroslav, J, Yung En Chee Colin, & Yates. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, **17**(1), 43–57.

Ferrier, Simon, Ford, Hugh A, & Jarman, Peter. 1984. The Status of the Rufous Scrub-Bird'Atrichornis Rufescens': Habitat, Geographical Variation and Abundance. *University of New England*.

Grimmett, Liam, Whitsed, Rachel, & Horta, Ana. 2021. Creating virtual species to test species distribution models: the importance of landscape structure, dispersal and population processes. *Ecography*, **44**(5), 753–765.

Hirzel, Alexandre H, Helfer, Véronique, & Metral, F. 2001. Assessing habitat-suitability models with a virtual species. *Ecological modelling*, **145**(2-3), 111–121.

Hirzel, Alexandre H, Hausser, Jacques, Chessel, Daniel, & Perrin, Nicolas. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**(7), 2027–2036.

Hutchinson, G Evelyn, *et al.* 1978. *Introduction to population ecology*. Yale University Press.

Hutchinson, GE. 1957. Concluding remarks cold spring harbor symposia on quantitative biology, 22: 415–427. *GS SEARCH*.

Leroy, Boris, Meynard, Christine N, Bellard, Céline, & Courchamp, Franck. 2016. virtualspecies, an R package to generate virtual species distributions. *Ecography*, **39**(6), 599–607.

MacArthur, Robert H. 1958. Population ecology of some warblers of northeastern coniferous forests. *Ecology*, **39**(4), 599–619.

Manly, Bryan FJ, McDonald, Lyman L, Thomas, Dana L, McDonald, Trent L, & Erickson, Wallace P. 2002. Introduction to resource selection studies. *Resource selection by animals: statistical design and analysis for field studies*, 1–15.

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

Meynard, Christine N, & Quinn, James F. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34**(8), 1455–1469.

Meynard, Christine N, Leroy, Boris, & Kaplan, David M. 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? *Ecography*, **42**(12), 2021–2036.

Moilanen, Atte. 2005. Reserve selection using nonlinear species distribution models. *The American Naturalist*, **165**(6), 695–706.

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

Pearce, Jennie, & Ferrier, Simon. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological modelling*, **128**(2-3), 127–147.

Polechová, Jitka, & Storch, David. 2019. Ecological Niche. *Pages 72–80 of:* Fath, Brian (ed), *Encyclopedia of Ecology (Second Edition)*, second edition edn. Oxford: Elsevier.

Qiao, Huijie, Peterson, A Townsend, Campbell, Lindsay P, Soberón, Jorge, Ji, Liqiang, & Escobar, Luis E. 2016. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, **39**(8), 805–813.

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

Thuiller, Wilfried, Lavorel, Sandra, Araújo, Miguel B, Sykes, Martin T, & Prentice, I Colin. 2005. Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences*, **102**(23), 8245–8250.

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

Walther, Gian-Reto, Roques, Alain, Hulme, Philip E, Sykes, Martin T, Pyšek, Petr, Kühn, Ingolf, Zobel, Martin, Bacher, Sven, Botta-Dukát, Zoltán, Bugmann, Harald, *et al.* 2009. Alien species in a warmer world: risks and opportunities. *Trends in ecology & evolution*, **24**(12), 686–693.