RESEARCH ARTICLE

Frontier depth modelling of species geographic distributions and a harmonic virtual species approach

Daniel Rojas Díaz[1*] | **Camilo Oberndornfer Mejía**[1*] | **Luisa Toro Villegas**[1*] | **Miguel Valencia Ochoa**[1*]

[1]Department, EAFIT University, Medellin, Antioquia, Postal Code, Colombia

Correspondence
Author One PhD, Department, Institution, City, State or Province, Postal Code, Country
Email: correspondingauthor@email.com

Present address
[†]Department, Institution, City, State or Province, Postal Code, Country

Funding information
Funder One, Funder One Department, Grant/Award Number: 123456, 123457 and 123458; Funder Two, Funder Two Department, Grant/Award Number: 123459

Virtual species are created by defining their niche as a function of environmental variables on which we simulate species occurrences on a given map. The aim of this study is to develop a nonparametric function that uses a frontier depth measure that potentially solves the bias problem while maintaining accuracy in its predictions. The method is based on the use of harmonic functions to warp the space to achieve variability in the distances to a chosen initial point. Then, the species generated using Harmonic Functions yielded a lower average accuracy for all methods, than Scalar Coefficients or Beta Distributions. The Frontier Depth average was better than expected as the accuracy averaged 87% in the commonly used virtual species generation method in a high variability map such as South America. We recommend researching to improve the method in various ways, such as finding the best hyperparameter p to construct the acceptance ranges or developing a heuristic method with the results of this method as a starting point.

KEYWORDS
ecological niche, species distribution model, presence-only, virtual species.

[*]Equally contributing authors.

1 | INTRODUCTION

Species distribution models (SDMs) explain species' spatial occurrence patterns and the environment's influence on their presence. Building SDM models requires two types of data: geospatial environmental data and species data. The latter can be found as presence-only data (registered sightings) or as a combination of presence and absence data (registered lack of sightings) which is scarce (Merow and Silander Jr, 2014). Its output is a prediction of the niche that, according to Hutchinson (1957), is a hypervolume in which each point describes the states of the environmental variables/covariates suitable for the survival of the species. Predictions of presence help make informed decisions for biodiversity research, including controlling the impacts of biological invasions (Walther et al., 2009), identifying threats to species conservation (Thuiller et al., 2005), and planning safe spaces for endangered species (Moilanen, 2005).

As of 2022, the most used SDMs (also known as niche modeling tools) are Maxent and MaxLike. Both of these use underlying parametric assumptions (Elith et al. (2011) & Merow and Silander Jr (2014)), such as fitting the likelihood to a Gibbs distribution, thus creating a bias that can cause incorrect predictions. Non-parametric methods solve these problems by avoiding distributional characteristics (IBM, 2021). This research aims to develop a nonparametric function that uses a frontier depth measure that potentially solves the bias problem while maintaining accuracy in its predictions.

One of the main issues of SDMs is validating the models with species data from real life, making virtual species crucial. According to Grimmett et al. (2021), virtual species are created by defining their niche as a function of environmental variables on which we simulate species occurrences on a given map. Austin et al. (2006) suggest performing virtual experiments on all emerging SDM methods to avoid species-related bias. The aim is to complement the literature on virtual species by thoroughly exploring the variable space, generating variability in the species simulated. The output will represent the virtual species with presence-only data.

When measuring the effectiveness of virtual species methods, we found little research done so far. Meynard et al. (2019) compare the work of two authors using a methodological approach, *virtualspecies* by Leroy et al. (2016) and *NicheA* by Qiao et al. (2016), reviewing technical details such as format and interface, as well as how they simulate the niche. Other authors test the virtual species using SDMs by determining whether they reduce their accuracy.

2 | FRONTIER DEPTH

2.1 | Landscape and species records

The ecologist's defined area is known as the landscape of interest (L). It could be constrained, for example, by geographical restrictions or knowledge of the potential dispersal range of the focus species.

For this research, the authors consider two case studies: the map of Colombia and the map of South America, which include 19 environmental variables, which will be considered as the dimensional space. The method uses preprocessed data (via range normalization) to avoid possible biases while keeping each covariate's distributive properties and allowing for easier handling of different ranges. It also uses simulated virtual species as presence-only samples.

2.2 | Description of the model

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 at every location within $\mathcal{L}$.

Define $\mathcal{F}(z)$ as the probability density of the covariates in $\mathcal{L}$, $\mathcal{F}_1(z)$ as the probability density of the covariates across locations within $\mathcal{L}$ where the species is present. The quantity to estimate is $\mathcal{F}_1(z)$ to solve the probability of the presence of species, conditioned on the environment covariates: $Pr(Y = 1|z)$. Due to the

Bayes rule classifier:

$$Pr(Y = 1|z) = \frac{\mathcal{F}_1(z)}{\mathcal{F}(z)} Pr(Y = 1) \quad (1)$$

Thus we will focus on estimating $\mathcal{F}_1$ given that it is the probability intensity of the species' presence, also called their niche. For this, we propose a depth measure as an alternative for probability intensity, consisting of four steps. First, data dimensionality reduction. Second, the construction of a frontier that contains all the samples. Third, calculation of the distance from every sample point to every frontier vertex in the original N-dimensional space to define an acceptance range. Fourth, calculate the depth of each map pixel.

We reduced data dimensionality to 3 dimensions using Principal Component Analysis (PCA). In this space, we calculate the frontier. The sample points that make up the frontier vertices will be considered frontier points ($\mathcal{N}$), while the sample points left inside the frontier will be called acceptance points ($\mathcal{P}$).

We used a hull-shrinking method to construct the frontier. This method receives an array of points and a shrinking factor. The latter defines a maximum separation of frontier points based on the samples' variance. The frontier ranges from the convex hull to one based on the closest point according to the shrinking factor. The constructed frontier affects the acceptance points, therefore, changing the resulting depth. We experimented with multiple hulls to avoid assumptions.

To calculate the acceptance range ($\mathcal{R}$) of each acceptance point, we use the p percentile of the distances from the acceptance point to every frontier point.

$$\mathcal{R}_i = p‰ \left\{ |\mathcal{P}_i - \mathcal{N}_j| \right\}_{j \in \mathcal{N}} \quad \forall i \in \mathcal{R} \quad (2)$$

With the acceptance ranges associated with every acceptance point, the niche can be reconstructed. The depth measure which will define the probability intensity (E of the presence of the species at each point of the map ($\mathcal{M}$) is the number of acceptance ranges in which the point of the map is. To define if a point is in an acceptance range, the distance from the map pixel to an acceptance point must be smaller than the acceptance range associated with it. Let C be a set associated with the acceptance ranges a point on the map is in:

$$C_i = \left\{ \mathcal{P}_j : |\mathcal{M}_i - \mathcal{P}_j| \leq \mathcal{R}_j \right\}_{j \in \mathcal{R}} \quad \forall i \in \mathcal{M} \quad (3)$$

Finally, the probability intensity of E_i is the coefficient between the cardinality of C_i and the amount of acceptance ranges 4:

$$E_i = \frac{||C_i||}{||\mathcal{R}||} \quad (4)$$

For this research, p was considered as the $25th$ percentile. This allowed for bigger acceptance ranges, which were necessary since the reduction of the data dimensionality resulted in a loss of variability, affecting the selection of frontier points. Generally, this method tends to be quite robust against outliers.

The results accuracy was calculated with the norm and the trapezoidal rule of the non-zero pixels from the generated niche to the corresponding pixels in the estimated map.

3 | TESTING AND VALIDATING

3.1 | Virtual species

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.

Response functions explain the dependence of a specie on each of the layers, according to Leroy et al. (2016). The reaction to each layer of each specie helps estimate its environmental suitability. The suitability of each pixel is then determined with a given algorithm, forming a niche. The methods created for virtual species generation will aim to maximize the variability of response functions.

Another option is to perform dimensionality reduc-

tion techniques on the environmental variables space. Principal Component Analysis (PCA) is a recurrent method in literature, according to Leroy et al. (2016). The technique then transforms a d-dimensional space into a k-dimensional space. The approach for this study is to reduce the environmental variable space and use harmonic functions to warp the space to achieve variability in the distances to a chosen initial point.

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 the threshold, then 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 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, we'll compare three different approaches: multiplying the variables by random coefficients (SC), using beta distributions (BD), and using PCA to reduce dimensionality and then create a random harmonic function, with a given order (HF). Afterwards, it will randomly choose an initial point and calculate its normalized distance from each pixel on the map. This distance is the likelihood of presence.

3.2 | A comparative study

To check which method had the highest variability, several metrics were used. All based on a comparison with the distribution with the most entropy (chaos): the uniform distribution. The tests consist of going through all the pixels in the maps, where each have a likelihood for each map compared. This gives you, at each pixel, a set of likelihoods. The empirical distribution ($\hat{F}$) of this sets is calculated and compared to the uniform CDF (F_1), using three different metrics. The first is a Gini Lorenz curve-based metric, which compares the area in common between the two CDF's:

$$L = 1 - \frac{||F_1 - \hat{F}||}{||F_1||} \tag{5}$$

The second is the two-sample Kolmogorov-Smirnov test, which uses the maximum absolute difference between the cdfs. The test statistic is:

$$D = \max_x \left(\left| \widehat{F}_1(x) - \widehat{F}_2(x) \right| \right), \tag{6}$$

where $\widehat{F}_1(x)$ is the proportion of x1 values less than or equal to x and $\widehat{F}_2(x)$ is the proportion of x2 values less than or equal to x. We consider also consider the mean of the times that the samples pass the test as an index A. The third method is the Mann-Whitney U-test measures the equality of population medians of two independent samples. The statistic used is

$$U = max(U_1, U_2) \tag{7}$$

where

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1, \quad U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$$

The decision whether to accept or reject the null hypothesis (that both come from the same distribution) is up to a table. We do the test for each pixel and, with an array containing the result of every test, we find its average. This gives us the index.

Another method is the Shannon Entropy, which measures how chaotic a random variable is, according to the following equation:

$$H = -\frac{1}{log_2(k)} \cdot \sum_i p_i \cdot log_2(p_i) \tag{8}$$

The index of each pixel is the ratio between the entropy in all the maps and the entropy of a uniform distribution. The index for the map is the average of all the indexes.

$$H_{index} = \frac{1}{N} \sum_i^N \frac{H_i}{H_{uniform}} \tag{9}$$

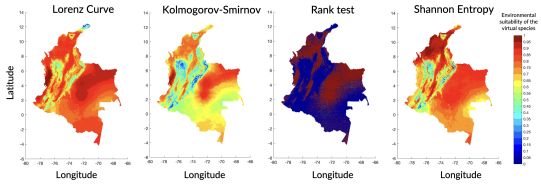
4 | RESULTS

4.1 | Virtual species variability

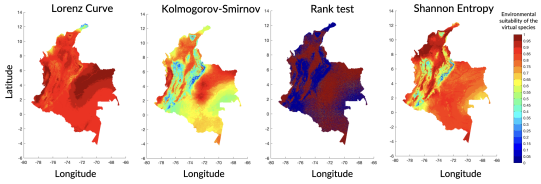
Metric	Colombia			South America		
	SC	BD	HF	SC	BD	HF
Lorenz	0.93	0.92	0.95	0.95	0.95	0.96
KS	0.71	0.71	0.75	0.78	0.79	0.77
	0.88	0.91	0.92	0.63	0.73	0.77
Rank	0.39	0.59	0.12	0.72	0.72	0.79
Entropy	0.90	0.97	0.81	0.96	0.97	0.83
Ranking	3	2	1	3	2	1

Abbreviations: SC, scalar coefficients; BD, beta distributions; HF, harmonic function; KS, Kolmogorov Smirnov.

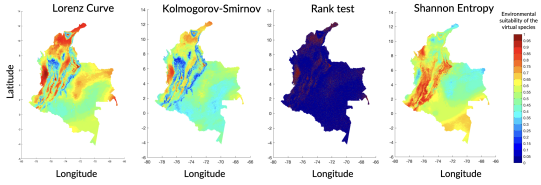
To get a complete view including all these indexes, we check which method gets a higher score in most metrics, and rank them.



(a) Scalar coefficients (SC)



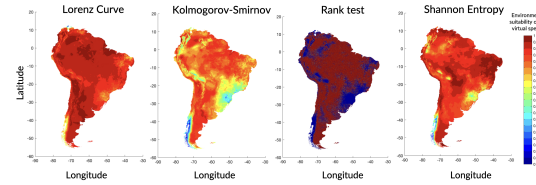
(b) Beta distributions (BD)



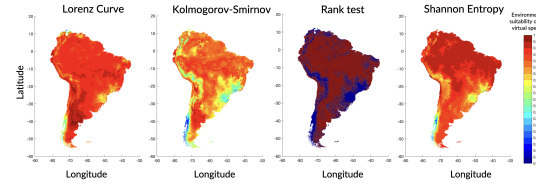
(c) Harmonic functions (HF)

FIGURE 1 Metrics at each pixel of the Colombian map using each virtual species method.

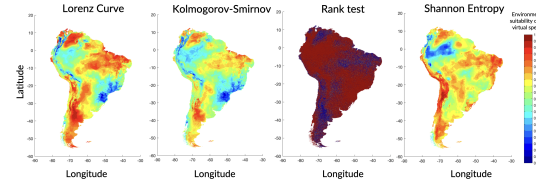
Figure 1 and Figure 2 show the metric calculated at each index of the geographical regions. In both areas with more and less variability, they exhibit coherency. When contrasting the different metrics, one could argue that the Lorenz curve metric overqualifies the variability (if compared to the other methods). Furthermore, the nature of the rank test (accepting or denying the null hypothesis that both come from the same distribution) and its results are highly dependent on the tolerance given to $p - value$ ($\alpha = 0.05$).



(a) Scalar coefficients (SC)



(b) Beta distributions (BD)



(c) Harmonic functions (HF)

FIGURE 2 Metrics at each pixel of the South America map using each virtual species method.

4.2 | Niche modelling accuracy

We performed several tests to determine and contrast how well each method predicts the species' original niche. The first was a brief tuning of the hyperparameter p. We conducted several experiments with different maps, and the median ($p = 50$) was the value that included the best results for each (more robust).

However, in many simulations, the best result, in terms of norm accuracy, was $p = 25$.

The second phase of the experiment involved extracting 50 niche maps from each virtual species method (scalar coefficients, beta distributions, and harmonic function). Then, each of them was sampled and reconstructed using both SDM methods: Frontier Depth (FD) and Frontier Depth Average (FDA), with two experimental variations, using a technique to remove outliers and leave the data as is.

TABLE 1 Accuracy SDMs

SDM		Colombia			South America		
	Outliers	SC	BD	HF	SC	BD	HF
FD	Yes	0.89	0.88	0.86	0.87	0.84	0.85
FD	No	0.90	0.89	0.86	0.89	0.86	0.88
FDA	Yes	0.87	0.89	0.85	0.86	0.85	0.83
FDA	No	0.85	0.89	0.87	0.85	0.85	0.82
Mean	-	0.88	0.88	0.86	0.87	0.85	0.85

Abbreviations: SC, scalar coefficients; BD, beta distributions; HF, harmonic function; FD, Frontier Depth; FDA, Frontier Depth Average.

TABLE 2 Trapezoidal Accuracy SDMs

SDM		Colombia			South America		
	Outliers	SC	BD	HF	SC	BD	HF
FD	Yes	0.79	0.78	0.75	0.79	0.78	0.76
FD	No	0.81	0.79	0.77	0.80	0.80	0.79
FDA	Yes	0.79	0.79	0.74	0.77	0.79	0.73
FDA	No	0.77	0.79	0.76	0.77	0.81	0.73
Mean	-	0.79	0.79	0.76	0.79	0.80	0.75

Abbreviations: SC, scalar coefficients; BD, beta distributions; HF, harmonic function; FD, Frontier Depth; FDA, Frontier Depth Average.

Table 1 and Table 2 show the results of the experiments on average. We can observe that, in general, species generated with Harmonic Functions yielded a lower average accuracy for all methods than Scalar Coefficients or Beta Distributions.

5 | CONCLUSIONS

The Frontier Depth performance was better than expected, as the accuracy averaged 87% in the commonly used virtual species generation method on a high variability map such as South America. Therefore, the method is consistent and capable of handling a variety of scenarios, also shown by its performance in the more complex virtual species methods that have a lower bound of 84% accuracy on average. Despite the positive results, future researchers could use this method as a rough estimate for a more accurate heuristic algorithm. Frontier Depth and Frontier Depth Average methods were insensitive to outliers, as the average accuracy didn't change by more than 3% in the worst cases. They can benefit from outlier removal in almost all cases. We recommend researching to improve the method in various ways, such as finding the best hyper-parameter p to construct the acceptance ranges or developing a heuristic method with the results of this method as the starting point. This is due to the outlier insensitivity and the negligible performance difference between FD and FDA. In all of the experiments conducted, the FD method had better results than FDA and a shorter run time. Therefore, we suggest to keep on investigating with the Frontier Depth method. Based on the mean of all the indexes obtained from the different metrics, it's hard to conclude which virtual species method is better. The results rely heavily on the environmental data used and the samples created. The Harmonic Functions method showed better average results on both maps, but the difference is not big enough to reach a solid conclusion. For Colombia and South America, the Harmonic function method is visually less variable than other methods, and the Scalar Coefficients show more. It would be necessary to use the same experimental conditions, such as the environmental variables, the maps produced, and the metrics, to compare these virtual species methods to the literature methods.

references

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

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., J, Y. E. C. C. and Yates (2011) A statistical explanation of maxent for ecologists. *Diversity and distributions*, **17**, 43–57.

Grimmett, L., Whitsed, R. and Horta, A. (2021) Creating virtual species to test species distribution models: the importance of landscape structure, dispersal and population processes. *Ecography*, **44**, 753–765.

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

IBM (2021) Statistics - parametric and nonparametric. URL: `https://www.ibm.com/docs/en/db2woc?topic=procedures-statistics-parametric-nonparametric`.

Leroy, B., Meynard, C. N., Bellard, C. and Courchamp, F. (2016) virtualspecies, an r package to generate virtual species distributions. *Ecography*, **39**, 599–607.

Merow, C. and Silander Jr, J. A. (2014) A comparison of maxlike and m axent for modelling species distributions. *Methods in Ecology and Evolution*, **5**, 215–225.

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

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

Qiao, H., Peterson, A. T., Campbell, L. P., Soberón, J., Ji, L. and Escobar, L. E. (2016) Nichea: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, **39**, 805–813.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T. and Prentice, I. C. (2005) Climate change threats to plant diversity in europe. *Proceedings of the National Academy of Sciences*, **102**, 8245–8250.

Walther, G.-R., Roques, A., Hulme, P. E., Sykes, M. T., Pyšek, P., Kühn, I., Zobel, M., Bacher, S., Botta-Dukát, Z., Bugmann, H. et al. (2009) Alien species in a warmer world: risks and opportunities. *Trends in ecology & evolution*, **24**, 686–693.