Inspira Crea Transforma

Niche modeling: a non-parametric approach

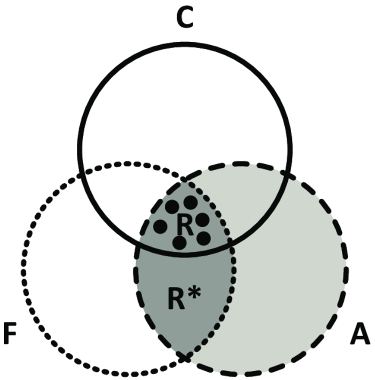
Daniel Rojas, Alexandra Catano, Jhan Carlo Carrillo, Santiago Ortiz, Nicolas Moreno Reyes

Escuela de ciencias
Universidad EAFIT
2021

Content

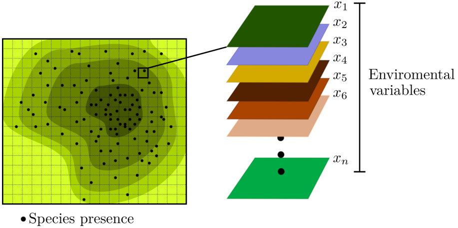
Introduction

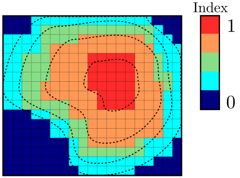
What is an ecological niche?



C = Climatic conditions
A = Accessibility
F = Fundamental niche
R = Realized niche / distribution
R* = Realizable niche
R + R* = Potential niche

What is niche modeling?





For only-presence data

Some niche models: Maxent and MaxLike

$$P(Y=1|f(\mathbf{X})) = P(Y=1)\frac{f_1(\mathbf{X})}{f(\mathbf{X})}$$

$$L(\beta) = \prod_{i=1}^{N} \frac{P(Y=1|x_i,\beta_0,\beta)}{\sum_{x\in B} P(Y=1|x_i,\beta_0,\beta)}$$

Basis of the niche modeling:

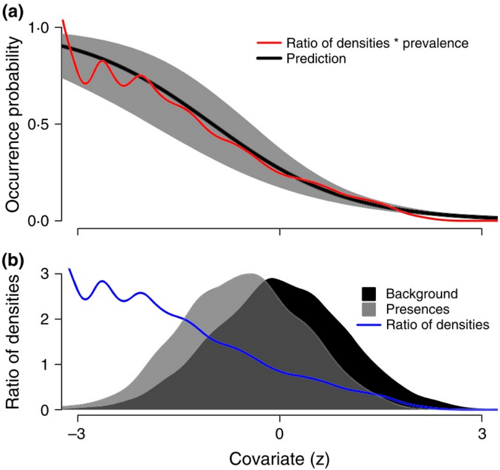
Maxlike

$$P(Y=1|x_i,\beta_0,\beta) = \frac{e^{\beta_0+\beta f(x_i)}}{1+e^{\beta_0+\beta f(x_i)}}$$

Maxent

$$P(Y=1|x_i,\beta) = e^{\beta f(x_i)}$$

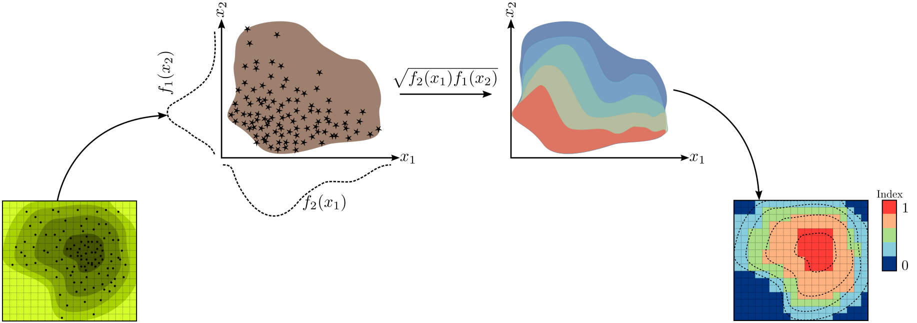
$$P(Y=1)\frac{f_1(\mathbf{X})}{f(\mathbf{X})}$$



Objective

Develop and validate a non-parametric niche modeling method based on the extraction of characteristics from presence-only data to identify and classify those suitable regions for an specie of interest.

Theoretical idea of the niche model


$f_1(x_2)$
x_2
x_1
$f_2(x_1)$
$\sqrt{f_2(x_1)f_1(x_2)}$
x_2
x_1
Index
1
0

Correlation between environmental variables

- We identify some correlation between environmental variables during the niche modeling process.
- If a *linear* correlation arises between some variables: rewrite the set of parameters $\mathbf{X}$ as new set $\mathbf{X}' = [\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \mathbf{x}_1, \ldots, \mathbf{x}_m]$ with $n \geq m + k$.
- In which each $\mathbf{Z}_i \subseteq \mathbf{Z}$, where

$$\mathbf{Z} = \{x \in X : \exists \mathbf{x_i} \in \mathbf{X} / |\rho(\mathbf{x}, \mathbf{x}_i)| \geq \alpha\}$$

- To divide $\mathbf{Z}$ in subsets chose a $\mathbf{x}_1^* \in \mathbf{Z}$ and define

$$\mathbf{Z}_1 = \{\mathbf{x} \in \mathbf{Z} : |\rho(\mathbf{x}, \mathbf{x}_1^*)| \geq \alpha\},$$

Clearly $\mathbf{x}_1^* \in \mathbf{Z}_1$, then $\mathbf{x}_1^*$ is a representative element of $\mathbf{Z}_1$.

- To obtain a second subset chose $\mathbf{x}_2^* \in (\mathbf{Z} - \mathbf{Z}_1)$ and define

$$\mathbf{Z}_2 = \{\mathbf{x} \in \mathbf{Z} : |\rho(\mathbf{x}, \mathbf{x}_2^*)| \geq \alpha\}$$

being $\mathbf{z}_2^*$ a representative element of $\mathbf{Z}_1$.

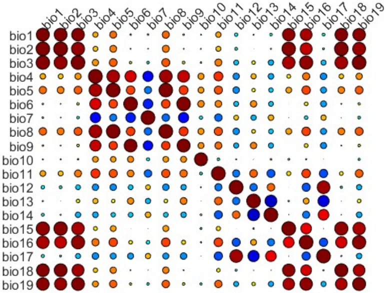
- Similarly, chose

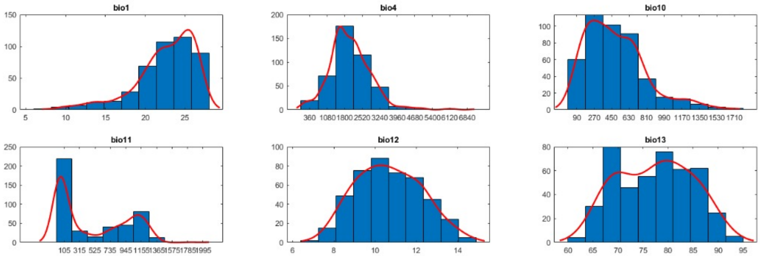
$$\mathbf{x}_k^* \in \left(\mathbf{Z} - \bigcup_{k-1}^{i=1} \mathbf{Z}_i \right)$$

and define

$$\mathbf{Z}_k = \{\mathbf{x} \in Z : |\rho(\mathbf{x},\, \mathbf{x}_k^*)| \leq \alpha\}$$

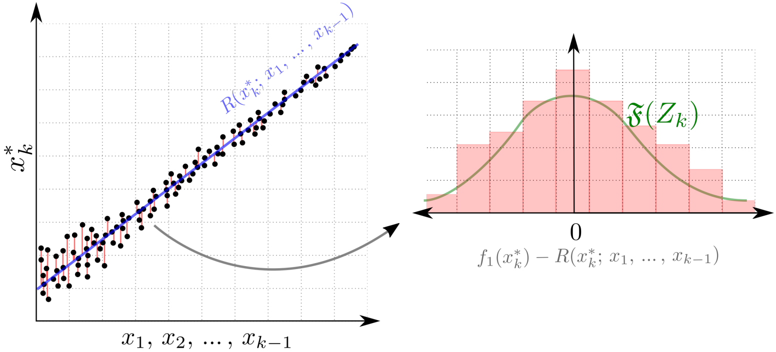
where $\mathbf{x}_k^* \in \mathbf{Z}_k$, then $\mathbf{x}_k$ is a representative element for $\mathbf{Z}_k$.

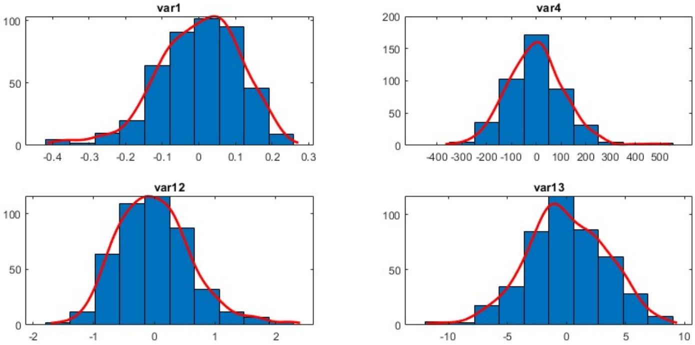

bio1
bio2
bio3
bio4
bio5
bio6
bio7
bio8
bio9
bio10
bio11
bio12
bio13
bio14
bio15
bio16
bio17
bio18
bio19


bio1
bio4
bio10
bio11
bio12
bio13

Ridge regression

$$\mathfrak{F}(\mathbf{Z}_k) := f\left(f_1(\mathbf{x}_k^*) - R(\mathbf{x}_k^*;\, \mathbf{x}_1,\, \dots,\, \mathbf{x}_{k-1})\right)$$




var1
100
50
0
-0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3
var4
200
150
100
50
0
-400 -300 -200 -100 0 100 200 300 400 500
var12
100
50
0
-2 -1 0 1 2
var13
100
50
0
-10 -5 0 5 10

Niche index

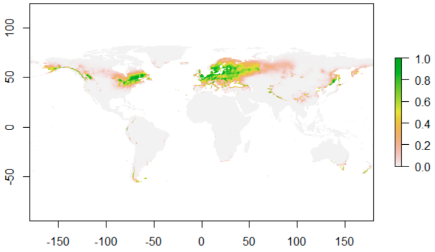
$$\mathbf{I_{ei}} = \left[\mathfrak{F}(\mathbf{Z}_1) f_1(\mathbf{x}_1^*) \ldots \mathfrak{F}(\mathbf{Z}_k) f_1(\mathbf{x}_k^*) f_1(\mathbf{x}_1) \ldots f_1(\mathbf{x}_m)\right]^{\frac{1}{m+k}} \tag{1}$$

Then, we estimate the $\mathbf{I_{ei}}$ value for each pixel and its value represents the color value on the map.

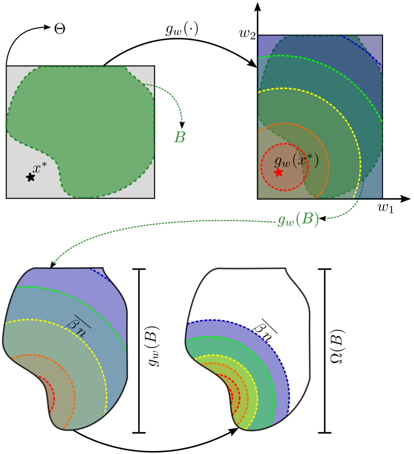
What is a virtual species?

Definition:

"Virtual species are simulated by defining their niche as a function of environmental variables and simulating their occurrence in a map" [1].



Generation of virtual species



For data sampling, we have that

$$\frac{\Omega^\alpha(b)}{\sum_{b\in B}\Omega^\alpha(b)}, \quad \alpha \in [1, 4]$$

is a pdf for each b.

Partial results

Design of experiments with virtual species

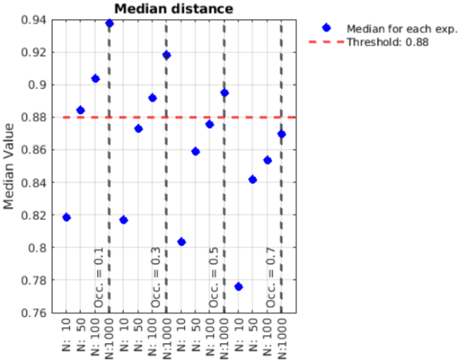
- Creation of a experiment grid with the following parameters for the virtual species:
  - Occupation: $\{0.1, 0.3, 0.6, 0.7\}$
  - Maximum number of samples: $\{10, 50, 100, 1000\}$
  - **Sampling weight:** $\{1, 2, 3, 4\}$ **(To do)**
- 16 experiments with 2000 repetitions each one.
- A total of 32000 virtual species.

Evaluation criteria

$$Similarity = 1 - \frac{\|M_1 - M_2\|_1}{|M_1|}$$

- M_1 : Niche map as a vector
- M_2 : Model map as a vector
- $|M_1|$: Cardinality of M_1

Best and worst scenarios



Best scenario:

$N = 1000$ & $Occ = 0.1$

Worst scenario:

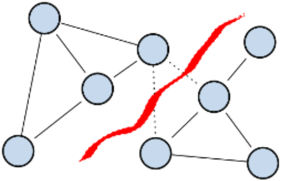
$N = 10$ & $Occ = 0.7$

Improving the model

Clustering approach

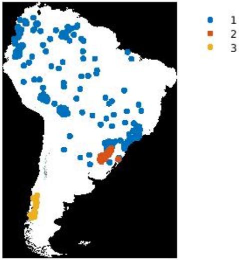
Spectral clustering:

- Graph-based algorithm for clustering observations
- It constructs a graph, finding its Laplacian matrix
- It uses the matrix to find k eigenvectors to split the graph k ways



Clusters approach

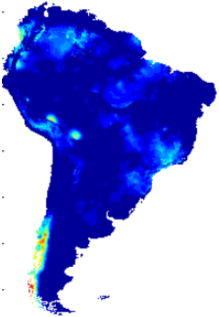
(a) Sampling map



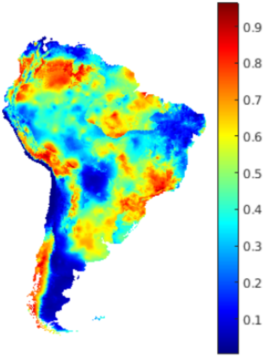
(b) Clustering according environmental variables

One application with a virtual species

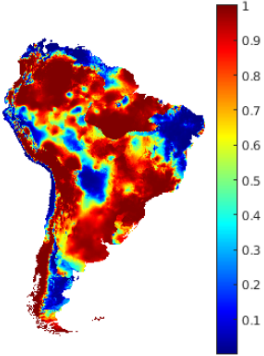
Virtual species generated



Maxent and MaxLike comparison



(a) *Maxent: 0.63 Similarity*



(b) *Maxlike: 0.38 Similarity*

New model comparison: $f_1(x)$

(a) Niche

(b) Model

Figure 4: New model: 0.63 Similarity

New model comparison: $f_1(x)/f(x)$

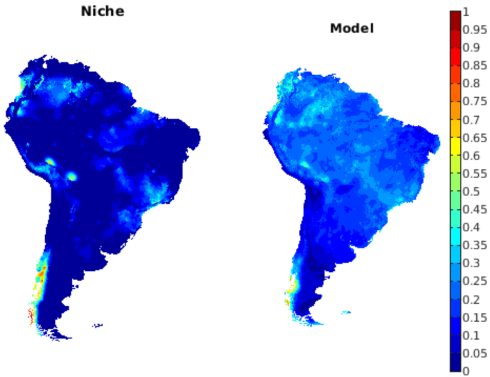


Figure 5: New model: 0.85 Similarity

Future work

Compare the new algorithm performance with other common niche algorithms as:

- Maxent
- Maxlike
- Bioclim
- BRT
- CART
- Domain
- FDA
- GAM
- GML
- MARS
- MDA
- RF
- SVM

Using SMD package in R.

Conclusions

- The method we proposed is robust to outliers.
- The method we proposed is computationally cheap.
- The method we proposed performs well in usual and new metrics.
- We managed to develop a novel niche modeling approach.
- Dependence among variables is the weakness of the method. Should we use multivariate kernel density estimation?
- Replacing correlation with a measure of dependence do not affect the most of our approach.
- The use of $f(x)$ must rely on strong *a priori* information.

References

[1] L. Grimmett, R. Whitsed, and A. Horta, "Creating virtual species to test species distribution models: the importance of landscape structure, dispersal and population processes," *Ecography*, vol. 44, no. 5, pp. 753–765, Feb. 2021. [Online]. Available: https://doi.org/10.1111/ecog.05555

[2] D. Nania, M. Flecks, and D. Rödder, "Continuous expansion of the geographic range linked to realized niche expansion in the invasive mourning gecko lepidodactylus lugubris (duméril & bibron, 1836)," *PLOS ONE*, vol. 15, no. 7, p. e0235060, Jul. 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0235060

[3] D. L. Warren and S. N. Seifert, "Ecological niche modeling in maxent: the importance of model complexity and the performance of model selection criteria," *Ecological Applications*, vol. 21, no. 2, pp. 335–342, Mar. 2011. [Online]. Available: https://doi.org/10.1890/10-1171.1

[4] J. A. Royle, R. B. Chandler, C. Yackulic, and J. D. Nichols, "Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions," *Methods in Ecology and Evolution*, vol. 3, no. 3, pp. 545–554, Jan. 2012. [Online]. Available: https://doi.org/10.1111/j.2041-210x.2011.00182.x

[5] M. Kéry, B. Gardner, and C. Monnerat, "Predicting species distributions from checklist data using site-occupancy models," *Journal of Biogeography*, pp. no–no, Jun. 2010. [Online]. Available: https://doi.org/10.1111/j.1365-2699.2010.02345.x

[6] C. B. Yackulic, R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. C. Grant, and S. Veran, "Presence-only modelling using MAXENT: when can we trust the inferences?" *Methods in Ecology and Evolution*, vol. 4, no. 3, pp. 236–243, Nov. 2012. [Online]. Available: https://doi.org/10.1111/2041-210x.12004

[7] W. Fithian, J. Elith, T. Hastie, and D. A. Keith, "Bias correction in species distribution models: pooling survey and collection data for multiple species," *Methods in Ecology and Evolution*, vol. 6, no. 4, pp. 424–438, Oct. 2014. [Online]. Available: https://doi.org/10.1111/2041-210x.12242

[8] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*, 1992.

[9] M. Kéry, "Towards the modelling of true species distributions," *Journal of Biogeography*, vol. 38, no. 4, pp. 617–618, Feb. 2011. [Online]. Available: https://doi.org/10.1111/j.1365-2699.2011.02487.x