CONABIO

COMISIÓN NACIONAL PARA EL
CONOCIMIENTO Y USO DE LA BIODIVERSIDAD

Two approaches to species distribution modeling and how the climate change is incorporated

Juan M. Barrios

November, 1st, 2017

National Commission for the Knowledge and Use of Biodiversity (CONABIO)

Outline

About CONABIO

Species Distribution Modeling

Let’s talk about data

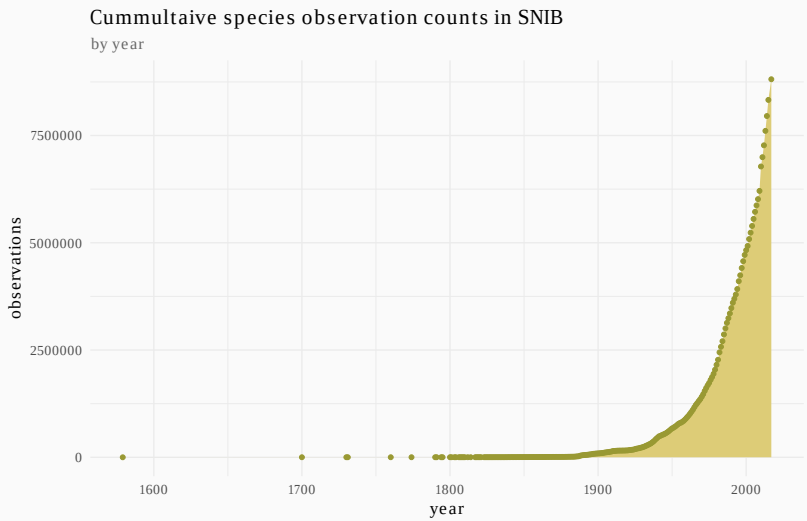
About CONABIO

About CONABIO

- CONABIO was founded on 1996. Since then it has been devoted to organize the existing information of the biota of Mexico.
- Most of that information, at a time, were recorded by academic institutions and some private people. There were some missing opportunities to analize the biota data because there was not a consolidate data facility.
- That primary goal give as a result the birth of the National Biodiversity Information System (SNIB).

SNIB

- There are approximately 8.8 M of curated data points
- Observations are as old as 1579, those registries are inferred by literature.
- A big part of CONABIO personnel and current projects are responsible to integrate and curate new information.



SNIB is more than species registries

There are about 13,000 digital maps of various subjects: vegetation maps, *species distribution*, satellite imagery, soil types, burning risk maps, water temperature maps, and much more. Most of these can be currently access on `http://www.conabio.gob.mx/informacion/gis`

Later this year we are going to release a updated tool to explore and download all these information.

There are many more projects on CONABIO

Some of them are...

- Genetic diversity project.
- MADMEX-REDD+, project to monitor deforestation and forest degradation.
- Enviromental suitability protection.
- Algal blooming and sea biodiversity programs.
- Invasive species studies.
- Mangrove monitoring, phenotypical data and chemical variables.
- *Biodiversity Monitoring Network*

There are many more projects on CONABIO

Some of them are...

- Genetic diversity project.
- MADMEX-REDD+, project to monitor deforestation and forest degradation.
- Enviromental suitability protection.
- Algal blooming and sea biodiversity programs.
- Invasive species studies.
- Mangrove monitoring, phenotypical data and chemical variables.
- *Biodiversity Monitoring Network*

A common goal:

Try to understand the biological processes with DATA

Species Distribution Modeling

What is the Species Distribution Modeling (SDM)?

- The goal is to determine where a given specie is present.
- To achieve this, most studies look a set of meaningful climate and topographic variables to find suitabily conditions based on current species observations (Ecological Niche Modeling).
- But the Species Distribution Modeling also should consider if a suitable niche can really be a habitat for a particular specie. [Peterson and Soberón, 2012]

Classical approach: MaxEnt, problem statment

Given an area of interest $\mathcal{D}$ where some environmental variables are defined $x_1, x_2, \ldots, x_m$, and a set of sites $z_1, z_2, \ldots, z_n$ where individuals of a particular specie were observed. We intend to estimate the range of specie habitat.

MaxEnt: How to do that?

Then it is assumed that the sites $\{z_i\}$ where selected independently from a unknwon probability measure p on $\mathcal{D}$.

A principle of maximum entropy states that this probability is uniform over $\mathcal{D}$. The uniform distribution is the "most random" of all.

Constraints

We have to consider that the specie might prefer some environmental features.

MaxEnt: How to do that?

Then it is assumed that the sites $\{z_i\}$ where selected independently from a unknwon probability measure p on $\mathcal{D}$.

A principle of maximum entropy states that this probability is uniform over $\mathcal{D}$. The uniform distribution is the "most random" of all.

Constraints

We have to consider that the specie might prefer some environmental features.

Then one needs to find the probability distribution $\hat{p}$ that maximize the entropy subject to: the expectation of the features $\mathbf{x}(z)$ under $\hat{p}$ matches the sample means of those features,

$$\frac{1}{n}\sum x(z_i) = \int_{\mathcal{D}} x(z)\hat{p}(z)\mathrm{d}\,z = \mathbb{E}_{\hat{p}}x(z).$$

MaxEnt

This criteria is equivalent to maximizing the likelihood of the parametric model

$$p(z) = \frac{e^{\lambda \mathbf{x}(z)}}{\int_{\mathcal{D}} e^{\lambda \mathbf{x}(u)} \mathrm{d}\, u.}$$

This likelihood is the same as a IPP with log-linear rate function.

MaxEnt is a software for modeling species distribution from presence-only data[1] using this approach.

[1] `https://biodiversityinformatics.amnh.org/open_source/maxent/`

Practical application

- Clean data
- Select the spatial region based on biological meaningful way
- Fit the MaxEnt models
- Given the fitted model as a map, work with some experts in the field to produce more meaningful output.

What about the climate change?

- From the model point of view, we only have a new set of values for the environmental variables. So just evaluate that new values on the fitted model.
- It rise a new problem. What happen with values never observed?
    - Just delete this regions.
    - Evaluate them and deal in some manner with those regions.

A second approach

Now we try to estimate the probability of an observation given a set of features $p(\cdot \mid \mathbf{x})$. Instead, we consider the log odds,

$$S(c \mid \mathbf{x}) = \ln\left(\frac{p(c \mid \mathbf{x})}{p(\bar{c} \mid \mathbf{x})}\right)$$

Using the Bayes theorem and assuming conditional independence of features given specie c, $p(\mathbf{x} \mid c) = \prod_i p(x_i \mid c)$, we have

$$S(c \mid \mathbf{x}) = \sum_i S(c \mid x_i) + \ln\left(\frac{p(c)}{p(\bar{c})}\right).$$

Marginal scores

Then we estimate the marginal log odds $S(c \mid x_i) = ln\left(\frac{p(x_i \mid c)}{p(x_i \mid \bar{c})}\right)$ as

$$p(x_i \mid c) = |\{c \cap x_i\}|/|\{c\}|$$
$$p(x_i \mid \bar{c}) = |\{\bar{c} \cap x_i\}|/(N - |\{c\}|)$$

in order to smooth these estimations, when $|\{c \cap x_i\}| = 0$ or $|\{c\}| = N$, we use a Laplace smoothing.

Some discussion

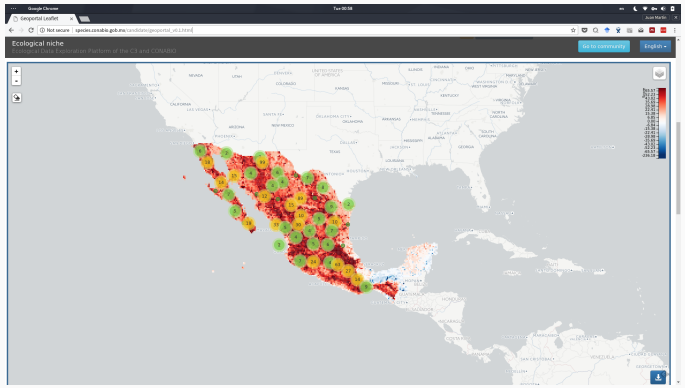
- This approach only looks for co-occurrences on a grided space.
- We can consider a space-time grid to incorporate time into the model.
- Hence it is easy to incoporate some time-dependent climate variability.
- We can also incorporate different data like other taxa. occurences.

SPECIES: Plataform to explore ecological data

We developed a platform

`http://species.conabio.gob.mx/candidate/`

- Fast prototyping
- Create reproducible experiments, You can ask for a unique link with your setup
- Incorporate some performance metrics and statistics



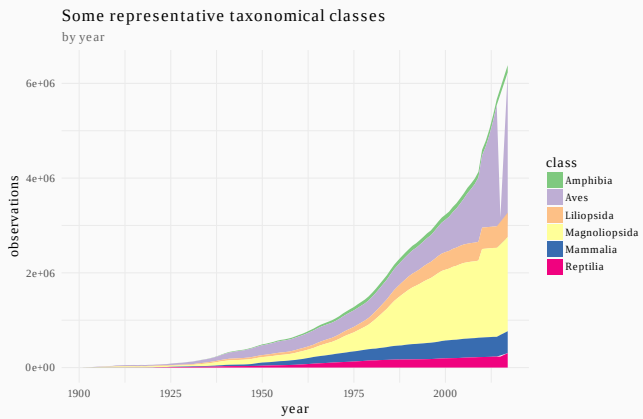
Some technical advantages of SPECIES

- There is an API that leverage all the calculations of the application... So You can use it with Python and R.
- The API also have some endpoint that gives you clean data to use on another workflow.
- Our database design was robust enough to perform cross validation in real time (2 min or less).
- We plan to release all the parts of the application as an Open Source.

Let’s talk about data

Species data is messy

- Species can be misidentified.
- We can have atypical data points... In Mexico we have observations of lions but there's no lion population in Mexico.
- Species taxonomical classificaction is variable over the time.
- There is bias: taxon groups bias and spatial sample bias.



Joint work with

SPECIES team
(UNAM-CONABIO)

C. Stephens (C3-UNAM, Phys)
C. González (UAM-Lerma, Bio)
R. Sierra (CONABIO, Math)
J.C. Salazar (CONABIO, Eng)
E. Rovredo (CONABIO, Bio)

CONABIO ecosystems
evalutation team

A. Cuervo (UNAM-CONABIO, Bio)
W. Tobon (CONABIO, Bio)
D. Ramirez (CONABIO, Bio)
J. Lopez (CONABIO, Bio)
T. Urquiza (CONABIO, Bio)


Questions?

References

W. Fithian and T. Hastie. Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, 7(4):1917, 2013.

C. González-Salazar, C. R. Stephens, and P. A. Marquet. Comparing the relative contributions of biotic and abiotic factors as mediators of species' distributions. *Ecological Modelling*, 248:57–70, 2013.

References ii

A. T. Peterson and J. Soberón. Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza & Conservação*, 10(2):102–107, 2012.

S. J. Phillips, M. Dudík, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the 21st international conference on Machine learning*, page 83. ACM, 2004.

J. Sarukhán and R. Jiménez. Generating intelligence for decision making and sustainable use of natural capital in Mexico. *Current Opinion in Environmental Sustainability*, 19: 153–159, 2016.