# Evaluating the effects of positioning errors on the accuracy of species distribution models using synthetic data


***Fábio Iwashita1,2, Silvana Amaral2, Antonio Miguel Vieira Monteiro2, Michael J. Friedel1,***

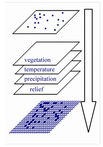
1Crustal Geophysics and Geochemistry Science Center , 2Image Processing Division, National Institute for Space Research—INPE

1{fiwashita, mfriedel}@usgs.gov; 2{miguel, silvana}@dpi.inpe.br


## Introduction

Biogeography is the science that attempts to document and understand the spatial patterns of biodiversity, studying the present and the past distributions of organisms.

Species distribution models (SDM) estimates fundamental niche for a certain specie identifying and quantifying relationships between occurrence field data and environmental variables, through statistical, mathematical and data mining methods.

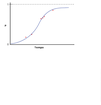


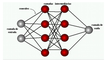

Occurrence points

Environmental variables

Predicted distribution

Modificado de Scachett (2005)



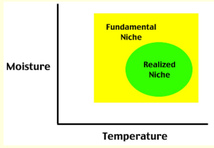


## Niche concept

The term ecological niche was popularized by Hutchinson (1957), which states it as a hypervolume n-dimensional, where each dimension represents the interval conditions or the needed resources for certain specie.

For instance, the niche for a vegetal species could include the temperature interval, the luminosity intensity required for photosynthesis, moisture regime and minimum amount of nutrients in the soil.

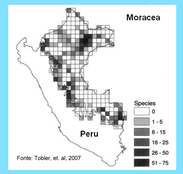
**Fundamental niche** comprises the environmental interval conditions without considering the interspecific competition or predation. The **realized niche** consist the part of the fundamental niche where the species really occupies




## Common problems with data acquisition

Difficulties in receiving the digital signal under forest cover which contributes to site coordinates inaccuracies

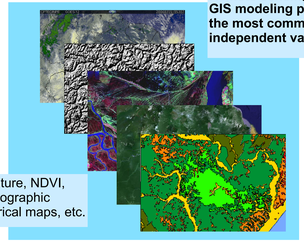
The majority of presence data, especially in tropical countries, are from herbariums and museum collections, which usually contain geographical positioning errors.




Cells with 0.5º of spatial resolution

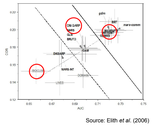
50% of the collection are within 9 cells

**Remote sensing and GIS modeling provides the most common independent variables.**



Surface temperature, NDVI, precipitation, topographic features, categorical maps, etc.

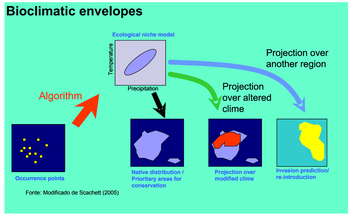
**The aim of this study is to evaluate the sensibilities of SDM to positioning error using synthetic data. Three models from different categories were assessed, bioclimatic envelope, genetic algorithm and maximum entropy.**



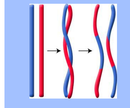

Source: Elith et al. (2006)

## Modeling fundamental niche

### Bioclimatic envelopes




Fonte: Modificado de Scachett (2005)

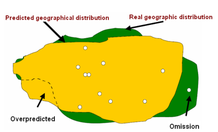


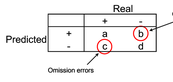
**Genetic algorithms**
GARP (Genetic Algorithm f

| Parents chromosome | Children chromosome |
|---|---|
| 11001010101010100101 | 11001010101010110110 |
| 10101000101001110110 | 10101000101001100101 |
| 10111010111000100101 | 10111000101000100100 |
| 11001000101000100100 | 11001010111000100101 |
| 11001110101011101101 | 11001110101001110110 |
| 10101000101001110110 | 10101000101011101101 |

| 'Children' chromosome | Chromosome (mutated) |
|---|---|
| 11001010101010110110 | 11101010101010000101 |
| 10101000101001100101 | 10101000101001110110 |
| 10111000101000100100 | 11111010111000100101 |
| 11001010111000100101 | 11001100101000100100 |
| 11001110101001110110 | 11001110101001110110 |
| 10101000101011101101 | 10101000101011101111 |

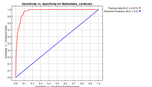
## Model errors and performance measurements

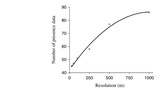


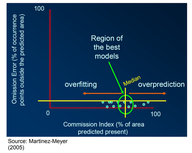




| Assessment measure | |
|---|---|
| Prevalence | (a + c) / N |
| Sensibility | a / (a + c) |
| Specificity | d / (b + d) |
| False positive rate (commission) | b / (b + d) |
| False negative rate (omission) | c / (a + c) |
| Kappa index | $\frac{(a+d)-\frac{(a+c)(a+b)+(b+d)(c+d)}{N}}{N-\frac{(a+c)(a+b)+(b+d)(c+d)}{N}}$ |

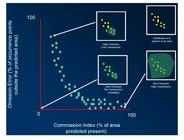
Source: (Fielding e Bell, 1997)

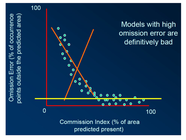







Source: Martinez-Meyer (2005)



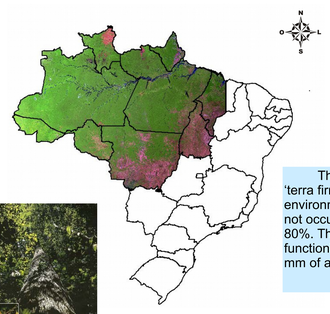

Some species are naturally rare. The number of samples may affect the models performance

**Samples will aggregated if they fall into the same cell, therefore, the number of samples will decrease, affecting SDM performance.**

| Nº Samples | |
|---|---|
| Training | 100 |
| 10 km | 100 |
| 0,25º | 66 |
| 0,5º | 39 |
| 1º | 24 |

## Study area and dataset

The artificial specie was based on a real Amazon plant, the Brazilian nut tree (*Bertholettia excelsea*).



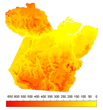


This virtual specie is found on cluster patterns, on 'terra firme' terrain, because does not tolerate acidic environments and saturated soils from floodplains. It does not occur on heights over 1600 m and slopes steeper than 80%. The temperature optimal range for physiological functioning is between 22ºC and 32ºC. It requires 1700 mm of annual precipitation and 10% to 65% soil moisture.

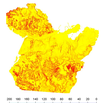
**150 samples were generated**

**Divided into calibration (100) and validation (50)**

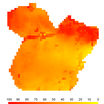
**Predictive variables**
**Spatial resolution: 1 km, 25 km and 50 km**



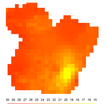
Elevation (m)



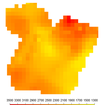
Slope (º)



Soil moisture (%)

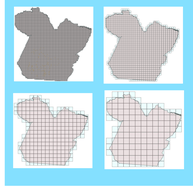


Temperature (ºC)

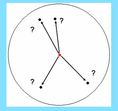


Annual Precipitation (mm)

### Positioning errors simulation

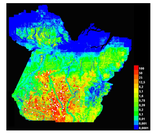


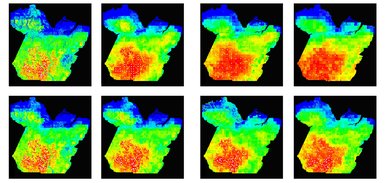
Coordinate displacements at four scales (10 km, 0.25º, 0.5º and 1º) were introduced using two methods: sample coordinates projection towards cellular center point, and errors with normal distribution in polar coordinates parameters.



The first approach reflected changes in grain size and is commonly adopted when dealing with herbarium and museum database, whereas the second approach simulated the problem in obtaining a GPS signal under the forest canopy

$$\Pr(X = x_k) = p_k \quad \text{for } k = 1, 2, \ldots \qquad H(X) = -\sum_{k \geq 1} p_k \log p_k.$$





### Omissão

| | Bioclim | | GARP | | Maxent | |
|---|---|---|---|---|---|---|
| | Centroi | Polar | Centroi | Polar | Centroi | Polar |
| | d | | d | | d | |
| 10 km | 0 | 0 | 14,67 | 10,67 | 4,14 | 3,4 |
| 0,25º | 0,67 | 0,61 | 4 | 3,33 | 5,52 | 4,76 |
| 0,5º | 4 | 1,33 | 10,67 | 2 | 4,83 | 4,76 |
| 1º | 6 | 2,11 | 6 | 4 | 2,07 | 2,04 |

### Área mínima

| | Bioclim | | GARP | | Maxent | |
|---|---|---|---|---|---|---|
| | Centroid | Polar | Centroid | Polar | Centroid | Polar |
| 10 km | 7,45 | 8,24 | 20,2 | 21,08 | 3,23 | 2,68 |
| 0,25º | 11 | 12,17 | 19,9 | 21,56 | 4,55 | 6,57 |
| 0,5º | 10,05 | 12,57 | 19,61 | 22,47 | 5,31 | 8,38 |
| 1º | 10,01 | 16,73 | 21,62 | 23,57 | 6,6 | 9,16 |

### Kappa

| | Bioclim | | GARP | | Maxent | |
|---|---|---|---|---|---|---|
| | Centroid | Polar | Centroid | Polar | Centroid | Polar |
| 10 km | 0,5036 | 0,4456 | 0,4121 | 0,4462 | 0,3994 | 0,4045 |
| 0,25º | 0,3502 | 0,3474 | 0,3909 | 0,4053 | 0,3384 | 0,3442 |
| 0,5º | 0,3051 | 0,3229 | 0,3841 | 0,3912 | 0,3381 | 0,3352 |
| 1º | 0,2916 | 0,3004 | 0,3893 | 0,395 | 0,3304 | 0,3276 |

### AUC

| | Bioclim | | GARP | | Maxent | |
|---|---|---|---|---|---|---|
| | Centroid | Polar | Centroid | Polar | Centroid | Polar |
| 10 km | 0,993 | 0,980 | 0,923 | 0,959 | 0,978 | 0,973 |
| 0,25º | 0,975 | 0,979 | 0,912 | 0,941 | 0,933 | 0,964 |
| 0,5º | 0,978 | 0,973 | 0,907 | 0,933 | 0,923 | 0,963 |
| 1º | 0,925 | 0,955 | 0,893 | 0,925 | 0,936 | 0,944 |

## Conclusions

1. More than one measure is needed to evaluate the sensibility, each model performance varied according to the assessment metric and type of error (cellular center projection/polar parameters).
2. If the error are up to 10 km, it is possible to apply any of the models without bigger performance losses;
3. When the errors have unknown spatial scale and under the premise of errors with Gaussian distributions, it is advisable to use GARP-BS which can cope more noisy data