Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 2275-2282.

Normalized Difference Vegetation Index (NDVI) improving species distribution models: an example with the neotropical genus *Coccocypselum* (Rubiaceae)

Silvana Amaral[1]
Cristina Bestetti Costa[1]
Camilo Daleles Rennó[1]

[1]Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brazil
{silvana, bestetti, camilo}@dpi.inpe.br

Abstract. Predictive models of species distributions use occurrence records and environmental data to produce a model of the species requirements and a map of its potential geographical distribution. This work analyses the contribution of remote sensing data, specifically the NDVI, for species distribution models, based on the taxonomic revision of the Neotropical genus *Coccocypselum*. GARP were use inside the OpenModeller to develop species distribution maps of five *Coccocypselum* species in the Brazilian territory. We generated two sets of models: using climate and topographic raster grids as environmental data, and a second one adding NDVI values. Presence and absence sample points statically evaluated the models, and for all species, Kappa values were higher than 0.5, indicating a good fit of the models. The sample size influenced the Mann-Whitney test, but it proved that species distribution models including NDVI were slightly better than models without NDVI. Despite this work is not conclusive, it contributes to generate better environmental data for ecological modeling.

Keywords: NDVI, Species Distribution Models (SDM), GARP, Rubiaceae, pseudo-absence.

1. Introduction

Our poor knowledge of the number and distributions of species limits our understanding of ecological and evolutionary processes, and our ability to use this knowledge to inform biodiversity and for conservation planning. As an alternative, biodiversity studies can use ecological models to estimate potential species distributions and avoid problems arising from an incomplete sampling. There are a large number of methods for modeling wildlife distribution. They include Habitat Suitability Indices (HSI), statistical methods using Generalized Linear Models (GLM) and Generalized Additive Modeling (GAM), and spatial and inductive (cartographic and regression trees, neural networks, Bayesian and weights of evidence) models (see Guisan & Thuiller 2005, Austin 2002).

In recent years, alternative modeling techniques were developed to incorporate presence-only data without requiring random absences to be artificially generated. Examples of these techniques are environmental envelopes (BIOCLIM, e.g. Austin 1994), genetic algorithms (as Genetic Algorithm for Rule-set Production - GARP, e.g. Anderson et al. 2002) and Ecological Niche Factor Analysis (ENFA, e.g. Hirzel et al. 2002, Tole 2006). GARP has been extensive used in presence-only studies (Anderson 2003, Peterson and Kluza 2003, Peterson and Shaw 2003, Stockwell and Peterson 2003).

GARP relates ecological characteristics of known occurrence points to those of points randomly sampled from the rest of the study region, developing series of decision rules that best summarizes those factors associated with the species presence (Stockwell and Peters 1999). Extensive tests have demonstrated GARP's ability to predict geographical and ecological distributions of species in diverse ecological, geographical, temporal and taxonomic contexts (Peterson 2001, Anderson et al. 2002, Bonaccorso et al. 2006).

The availability of observational data of species, and the scope and resolution of spatially explicit environmental data, are increasing, as well as are the capabilities of the computational and analytical tools that use these information (Graham et al. 2004). Remote sensing data can contribute to the modeling process by improving the environmental data set and the niche

characterization. AVHRR/NOAA(Advanced Very High Resolution Radiometer/ National Oceanic & Atmospheric Administration) imagery, in combination with other variables, proved to have sufficient resolution to model the range of bird species (Suárez-Seoane et al. 2002). METEOSAT (Meteorological Satellite) temporal series was tested to improve climate data for wild life distribution models (Suárez-Seoane et al. 2004). Oindo and Skidmore (2002) presented that multi-temporal Normalized Difference Vegetation Index (NDVI) data when related to changes in primary productivity, can be used to model trends in species richness. Parra et al. (2004) observed that remote-sensing data, such as NDVI, should be used with discretion in topographically complex regions with heavy cloud-cover, although it has potential to improve ecological niche modeling.

The use of a vegetation index contributes to the species distribution modeling process by providing information about the canopy closure, the phenological status, and the water content variation within the different Brazilian physiognomies. We hypothesize that species distribution models that uses vegetation index (NDVI) as an additional environmental variable would improve the representation of a species spatial distribution.

This work analyses the contribution of remote sensing data, specifically the NDVI, for species distribution models, based on the taxonomic revision of the neotropical genus *Coccocypselum* P. Br. (Costa 2005).

2. Methods

2.1. Study Species

The genus *Coccocypselum* belongs to Rubiaceae family, one of the most important families in the tropics. It comprises eighteen species of a small herbaceous genus that is widely distributed in the Neotropics, from south of Mexico to northern Argentina (Costa 2005). Fifteen species occur in Brazil and five of these species were selected for analysis (**Table 1**): *Coccocypselum capitatum* (Graham) C.B. Costa & Mamede *C. cordifolium* Nees & Mart., *C. erythrocephalum* Cham. & Schltdl., *C. lymansmithii* Standl. and *C. pulchellum* Cham. Due to the use of static distribution models, we assume that the observed species pattern is in a relative equilibrium with the environment (Guisan and Theurillat 2000, Guisan and Zimmermann 2000).

2.2. Data Sources

The species database is a simple presence data built based on the Natural History Collections (NHC, Graham et al. 2004). For this work, the occurrence information came from several institutions during the development the taxonomic revision of *Coccocypselum* (Costa 2005). Therefore, all occurrence data are based on information associated with specimen vouchers in natural history museums.

In order to compare the multivariate ecological space generated with 'presence-only' data models, 'pseudo-absent' data were generated for the species based on four criteria: i) sites that had been visited but the modeled species were not recorded; ii) sites that had been visited, but species other than the modeled ones were collected. These species usually exclude the occurrence of the modeled ones; iii) sites that had not been visited, but floristic studies did not record the species.

From the total of presence points for each species, an evaluation training set was randomly selected for the statistic analyses. Because the number of occurrence data varies according to the species, we decided to use a proportion of the total number of records available (ca. 20%). The same number of 'pseudo-absence' was aleatory selected (**Table 1**).

Table 1 – Number of training (presence) and evaluation points (presence/absence) for the *Coccocypselum* modeled species.

Studied species	Training points	Evaluation points	
	Presence	Presence	Absence
C. capitatum (Graham) C.B. Costa & Mamede	65	15	15
C. cordifolium Nees & Mart.	72	16	16
C. erythrocephalum Cham. & Schltdl.	33	8	8
C. lymansmithii Standl.	22	5	5
C. pulchellum Cham.	52	12	12

Occurrence data (presence and absence points) for the collections of five *Coccocypselum* Brazilian species were organized in geographical database, using the TerraView software (www.dpi.inpe.br/terraview). The OpenModeller (OM) was used to model the species distribution and generate Species Distribution Models (SDM). Although there are several algorithms available to model species distribution, OM works fully integrated to TerraView, facilitating data comparison and geographical analysis.

2.3. Environmental Variables

The environmental variables for the species distribution modeling were selected based on the knowledge that the general distribution of *Coccocypselum* genus is related to the conditions of humidity and altitudinal gradients (Costa 2005).

Raster grid data from temporal series of monthly temperature (maximum, mean and minimum) and precipitation, provided from the WorldClim project (Hijmans et al. 2005), characterized the climatic gradients over the Brazilian territory. Additionally, other nineteen bioclimatic variables, derived from the seasonal variation of temperature and precipitation, were used for the modeling process as well. Given that the collection of meteorological data over the Brazilian territory is very sparse, climate grids with 0.25 degrees of spatial resolution (approximately, 27.8 km at the Equator) were considered adequate to the work scale.

For the topographic characterization, Shuttle Radar Topography Mission (SRTM) images of elevation, slope and aspect were used (http://srtm.usgs.gov/index.html), with spatial resolution of 0.0083 degrees (approximately 1 km).

From AVHRR/NOAA-17 daily satellite images for the year of 2005 (from January to December), a time series of fortnightly NDVI mosaic images, were generated at Environmental Satellite Division (DSA), Weather Forecast Center and Climate Studies (CPTEC), at National Institute of Space Research (INPE). Instead of the 24 original NDVI images, we used the statistical characteristics of the image time series for the modeling: images of NDVI minimum, maximum, average and standard deviation. With this procedure, besides the volume data reduction, some noise effect from the mosaic images was minimized. An algorithm, using the Interactive Data Language (IDL) was written specifically to generate these images.

In order to evaluate the importance of the NDVI for species distribution, two sets of models were generated: one with only climatologic and topographic variables, and another with climatologic and topographic variables with NDVI images.

2.4. Species Distribution Modeling

For modeling species distribution with GARP, we used the "GARP with best subset" algorithm implemented at OpenModeller (http://openmodeller.cria.org.br/). **Table 2** presents the generic modeling parameters used for all species of *Coccocypselum* modeled. The number of occurrence data was different for each species, according to the data availability (**Table 1**).

Table 2 – GARP specific parameters.

Training Proportion	0.7
Total Runs	10
Hard Omission Threshold	100
Models Under Omission Threshold	20
Commission Threshold	50
Commission Sample Size	10000
Maximum Number of Threads	1
Max generations	400
Convergence limit	0.01

Since GARP uses a random approach for the rule sets and the predictions vary between runs (Anderson et al. 2003), ten models per species of *Coccocypselum* were generated. A final model, containing the summary of the ten models provided the potential distribution in percentage of each species.

2.5. Statistical Test

The species distribution models generated including NDVI data, were statistically compared to correspondent distribution models, generated without NDVI data. Kappa statistic (Hudson and Ramm 1987) was used to evaluate if the models were different from what would be expected from a random classification. Kappa values close to 1 suggest a good agreement between the test and the reality, and values closer to 0 indicates that the result was just by chance. Negative values of Kappa indicate that a random approach would better fit than the classification results. The species distribution models were tested considering the presence and absence evaluation points as reality (truth). Given that results from the modeling are grids with values in percentage, indicating the potential distribution of each species, the threshold of 50% was applied to convert the values to presence and absence classes. Then, the confusion matrix to calculate Kappa statistics considered only 2 classes: presence and absence, for the evaluation points (reality) and the results of the models (test), for the models with and without NDVI images as environmental data.

The number of evaluation points of presence and absence were different between species, and small in some of then, making a parametric statistical analysis unfeasible. Alternatively, the Mann-Whitney test (Mann-Whitney 1947) was applied as non-parametric statistic to compare samples, therefore avoiding the normal distribution and equal variance assumptions of other parametric methods. The Mann-Whitney test, which estimates the U statistic, is almost as powerful as the t-test, and it works by ranking the two samples to examine how they are distributed in an ordered sequence. If there is a difference in the distributions of the two samples, the data from each of the samples will tend to cluster at the ends of the common sequence. The U statistic was tested against tabled values, using significance of 5%, for each number of samples.

3. Results

From the NDVI time series, some noise was still persistent when generating the NDVI maximum image, mainly in the ocean (**Figure 1A-D**). The NDVI average image presented very homogeneous values, and it has to be interpreted together with the other images, particularly the NDVI standard deviation, when the main Brazilian biomes, as Amazon forest and “caatinga” can be distinguished.

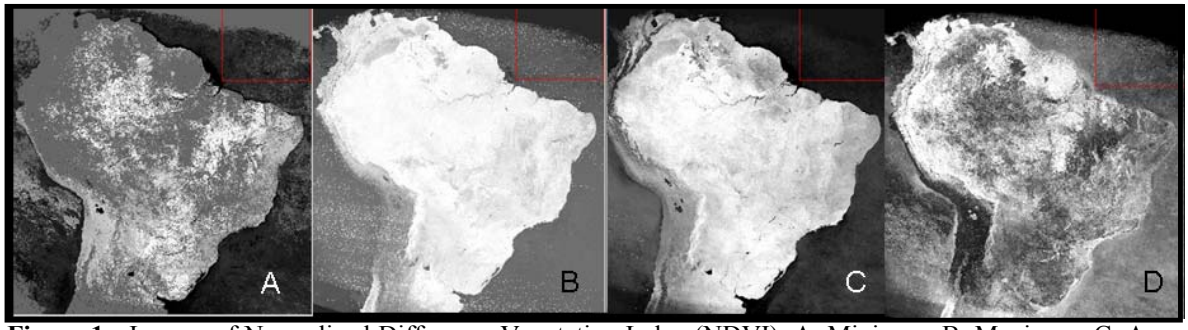

Figure 1 - Images of Normalized Difference Vegetation Index (NDVI). A. Minimum B. Maximum C. Average D. Standard deviation, for 2005.

The species of *Coccocypselum* studied are predominantly found in wet forest, with most of its population located along rivers and wet places. The NDVI data incorporated to models can predict a relation between the species and an open or more forested physiognomy.

In the modeled distribution maps (**Figure 2**), light grey and yellow (class of 10 to 50%) represent unsuitable areas for the species, regions where it is very unlikely that the species in question would be found. In the zones in orange and red (class of 60 to 100%) the occurrence of the species is expected. The predicted distribution map resulting from the climatic and topographic variables and NDVI data reveals a more restricted pattern for all species of *Coccocypselum* studied. Two species (*C. capitatum* and *C. cordifolium*) present a wider distributional pattern. The other three species have a more restricted range and the models seem to be more accurate.

Although the usefulness of the environmental niche modeling when applied to biogeographical and conservation approaches has been contested (Araújo and Guisan 2006), the species distribution models shows a result consistent to the distributional observation found in the taxonomic study of the genus *Coccocypselum* (Costa 2005). As expected, the SDM showed a wider distributional pattern. However, the SDM can be improved by restricting the predicted ranges using expert drawn range maps and biogeographical regions information (Graham and Hijmans 2006).

For the Kappa analysis (**Table 3**), all species presented values higher than 0.5, indicating a relative good fit between the evaluation points (presence and absence) and the modeled species distribution (values higher or lower than 50% of potential distribution of occurrence). In other words, the models are better than a random classification of the species distribution. However, models fitted with NDVI data proved not to be superior to those fitted with climatic and topographical predictors only.

On the other hand, observing the results of the Mann-Whitney test (**Table 3**), only the models developed to *C. pulchellum*, *C. capitatum* and *C. cordifolium* differentiated statistically between the presence and absence occurrence data (significative at 5%). For *C. lymansmithii* and *C. erythrocephalum* the distribution model was not conclusive. This can be explained by the confined distribution of theses species and the low number of samples points, less than ten (see Stokwel & Peterson (2002) estimative for an ideal sample size for GARP analysis).

Comparing the species distribution models generated including NDVI data, they presented better results than the distribution models generated without NDVI data (smaller U statistic means better models). Despite the values were very similar, this results suggested an improvement when using NDVI as environmental variables in the modeling process.

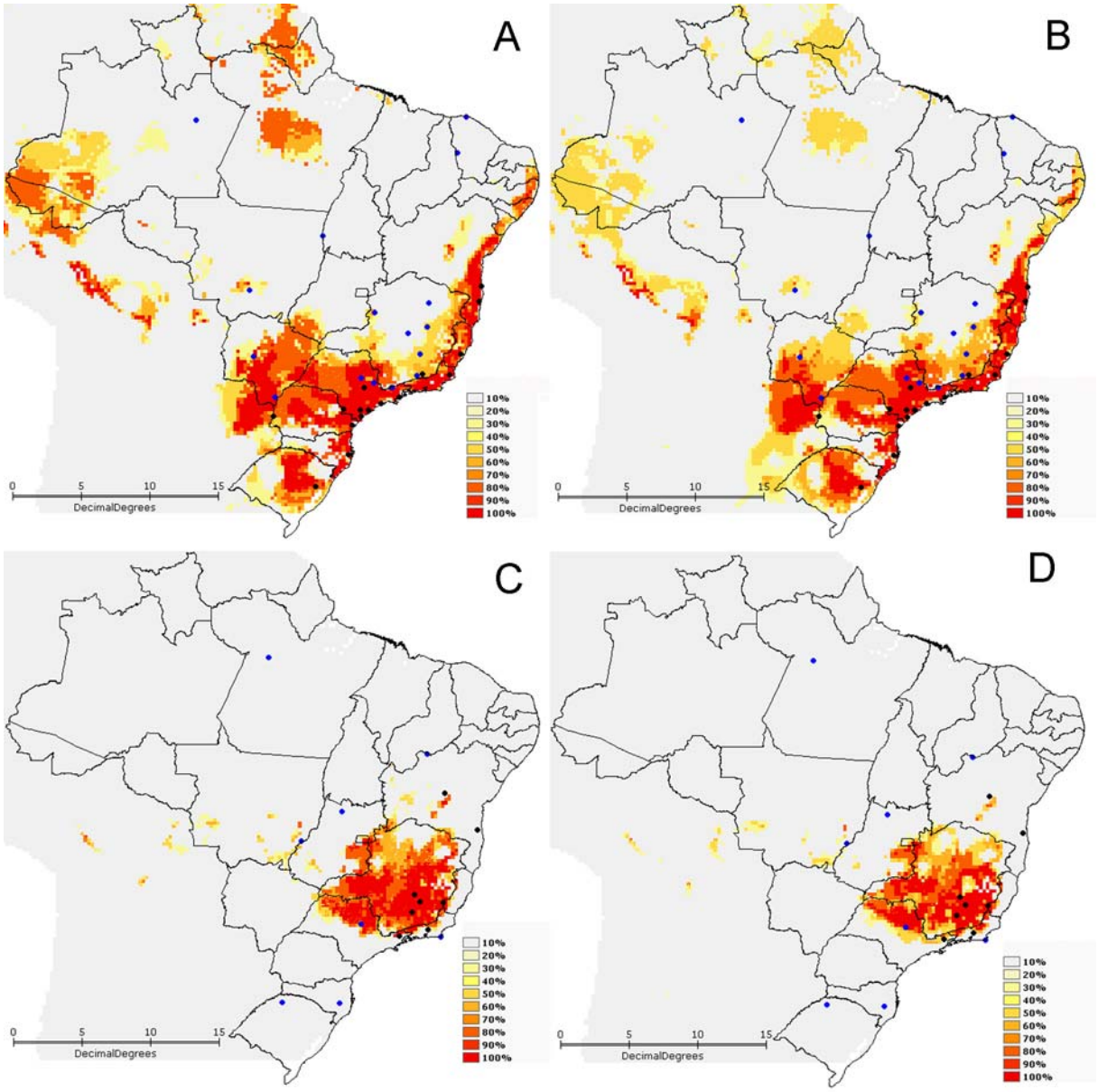

Figure 2 – Map of predictive probabilities modeled with GARP with best subset. **A.** *C. cordifolium* modeled with climatic and topography data. **B.** *C. cordifolium* modeled with climatic, topography and NDVI data. **C.** *C. erythrocephalum* modeled with climatic and topography data. **D.** *C. erythrocephalum* modeled with climatic, topography and NDVI data. Presence (black) and absence (blue) evaluation data points

Table 3 – Mann-Whitney and Kappa statistic for the species distribution models with and without NDVI images as environmental variable.

Specie	Kappa		Mann-Whitney (U Statistic)			
	no NDVI	NDVI	no NDVI	NDVI	Critical value (α =5%)	N
C. lymansmithii	0.8	0.8	7.5	7.5	2	5
C. erythrocephalum	0.5	0.5	21.5	22.5	12	8
C. pulchellum	0.83	0.83	16*	15*	37	12
C. capitatum	0.6	0.53	55*	48.5*	64	15
C. cordifolium	0.75	0.56	46.5*	36*	75	16

* Significative value at 5%

4. Final comments

This work provides environmental data of higher quality than usually is available for distribution modeling of plant species in Brazilian region. It presents a very robust statistical analysis of the models, since absent data were generated exclusively for this purpose. This approach is not frequent in the species distribution modeling analysis by the difficulty of having absence data.

The species distribution models generate by GARP indicates a potential presence of the species studied closely related to the known distribution of the species studied. The models for wide distribution species (*C. capitatum* and *C. cordifolium*), were more consistent to the real distribution then the restrict ones (*C. lymansmithii, C. erythrocephalum* and *C. pulchellum*).

This study illustrated the potential of incorporating NDVI data into large-scale models of plant species distribution. The NDVI data showed to have a potential application in ecological modeling approaches since slightly statistical differences were observed between the species distribution models. Even tough, the same approach should be applied over other species with higher sample size for more accurate analysis.

Acknowledgements

The authors are particularly grateful to André de Souza (CPTEC/INPE) for the NOAA/NDVI images time series, to Sidnei Sant'anna (INPE/DPI) for the IDL algorithms to generate the NDVI statistic imagery, and Laércio Namikaua (INPE/DPI) for the English revision.

References

Anderson, R. P., Gómez-Laverde, M., Peterson, A. T. Geographical distributions of spiny pocket mice in South America: insights from predictive models. **Global Ecology & Biogeography**, v. 11, p. 131–14, 2002.

Anderson, R. P., Lew, D., Peterson, A. T. Evaluating predictive models of species' distributions: criteria for selecting optimal models. **Ecological Modelling**, v. 162, p. 211–232, 2003.

Araújo, M. B. & Guisan, A. Five (or so) challenges for species distribution modeling. Journal of Biogeography, v. 33, p. 1677–1688, 2006.

Austin, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. **Ecological Modeling**, v. 157, p. 101-118, 2002.

Austin, M. P., Nicholls, A. O., Doherty, M. D., Meyers, J. A. Determining species response functions to an environmental gradient by means of a b-function. **Journal of Vegetation Science**, v. 5, p. 215-228, 1994.

Bonaccorso, E., Koch, I., Peterson, A. T. Pleistocene fragmentation of Amazon species' ranges. **Diversity and Distributions**, v. 12, 157–164. 2006.

Congalton, R. G. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing of Environment**, v. 37, n. 1, p. 35-46, 1991.

Costa, C. B. **Revisão taxonômica de *Coccocypselum* P. Br. (Rubiaceae).** 2004. 133p. Ph.D. dissertation, University of São Paulo, São Paulo, Brazil. 2005.

Graham, C. H., Hijmans R. J. A comparison of methods for mapping species ranges and species richness. **Global Ecology and Biogeography**, 2006.

Graham, C. H.; Ferrier, S.; Huettman, F.; Moritz, C.; Peterson, A. T. New developments in museum-based informatics and applications in biodiversity analysis. **TRENDS in Ecology and Evolution** v. 19, n. 9, p. 497-503, 2004.

Guisan, A., Theurillat, J.-P. Equilibrium modeling of alpine plant distribution: how far can we go? **Phytocoenologia**, v. 30, p. 353–384, 2000.

Guisan, A., Thuiller, W. Predicting species distribution: offering more than simple habitat models. **Ecology Letters**, v. 8, p. 993–1009, 2005.

Guisan, A., Zimmermann, N. E. Predictive habitat distribution models in ecology. **Ecological Modeling,** v. 135, p. 147- 186, 2000.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, p. 1965-1978, 2005.

Hirzel, A. H., Hausser, J., Chessel, D., Perrin, N. Ecological-Niche factor analysis: how to compute habitat-suitability maps without absence data? **Ecology**, v. 83, n.7, 2027–2036, 2002.

Hudson, P. E. & Ramm, R. S. Correct formulation of the kappa coefficient of agreement. **Photogrammetric Engineering and Remote Sensing,** v. 53, n. 4, p. 421-422, 1987.

Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Annals of Mathematical Statistics**, v. 18, n. 1, p. 50-60, 1947.

Oindo, B. O.; Skidmore, A. K. Interannual variability of NDVI and species richness in Kenya. **International Journal of Remote Sensing**, v. 23, p. 285-298, 2002.

Parra, J. L., Graham, C. C., Freile, J. F. Evaluating alternative data sets for ecological niche models of birds in the Andes. **ECOGRAPHY**, v. 27, p. 350- 360, 2004.

Peterson, A. T. Predicting Species' Geographic Distributions Based on Ecological Niche Modeling. **The Condor**, v. 103, p. 599–605, 2001.

Peterson, A. T., Kluza, D. A. New distributional modeling approaches for gap analysis. **Animal Conservation**, v. 6, n. 1, 47-54, 2003.

Peterson, A. T., Shaw, J. *Lutzomyia* vectors for cutaneous leishmaniasis in Southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. **International Journal for Parasitology**, v. 33, n. 9, p. 919-931. 2003.

Peterson, A. T., Stockwell, D. R. B., Kluza, D. A. Distributional prediction based on ecological niche modeling of primary occurrence data (ed. by J. M. Scott, P. J. Heglund, M. L. Morrison), **Predicting species occurrences: Issues of scale and accuracy**, Island Press, Washington DC., 2002. pp. 617–623.

Stockwell, D. & Peterson, A. T. Comparison of resolution of methods used in mapping biodiversity patterns from point-occurrence data. **Ecological Indicators**, v. 3, p. 213–221, 2003.

Stockwell, D., Peters, D. The GARP modeling system: problems and solution to automated spatial prediction. **International Journal of Geographical Information Science**, v. 13, n. 2, 143-158, 1999.

Stockwell, D. R. B., Peterson, A. T. Effects of sample size on accuracy of species distribution models. Ecological Modelling, v. 148, p. 1–13, 2002.

Suárez-Seoane, S.; Osborne, P. E.; Alonso, J. C. Large scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalizes additive models. **Journal of Applied Ecology**, v. 39, p. 755-771, 2002.

Suárez-Seoane, S.; Osborne, P. E.; Rosema, A. Can climate data from METEOSAT improve wildlife distribution models? **Ecography**, v. 27, p. 629-636, 2004.

Tole, L. Choosing reserve sites probabilistically: a Colombian Amazon case study. **Ecological Modeling**, v. 194, p. 344–356, 2006.