Ferraz et al. 2012. How species distribution models can improve cat conservation - jaguars in Brazil. Cat News Special Issue 7, 38-42.

Supporting Online Material SOM Appendix I. Background information on Species Distribution Modeling SDM

Predicting species distribution has made enormous progress in the last decade. A wide variety of modeling techniques (see Guisan & Thuiller 2005) have been intensively explored aiming to improve the comprehension of species-environment relationships (Guisan & Zimmermann 2000, Peterson 2001, Hirzel & Lay 2008, Elith & Leathwick 2009, Franklin 2009). The species distribution modeling (SDM) relate species distribution data to information on the environmental and/or spatial characteristics of those locations. Combinations of environmental variables most closely associated to presence points can then be identified and projected onto landscapes to identify areas of predicted presence on the map (Soberón & Peterson 2005, Peterson 2006). The geographic projection of these conditions (i.e., where both abiotic and biotic requirements are fulfilled) represents the potential distribution of the species. Finally, those areas where the potential distribution is accessible to the species are likely to approximate the actual distribution of the species.

The SDMs have also been termed as ecological niche models (ENMs) or habitat models (sometimes with different emphases and meanings; Elith & Leathwick 2009, Soberón & Nakamura 2009). According to Elith & Leathwick (2009) the use of neutral terminology to describe species distribution models (SDM rather than ENM) seems preferable. Despite its extensive use, there is an enormous debate about terminology and concepts in predictive modeling and a consensus about what we are modeling – habitat, niche, environment, species distribution – does not exists until now (Soberón & Peterson 2005, Kearney 2006, Peterson 2006, Austin 2007, Soberón 2007, Hirzel & Lay 2008, Jiménez-Valverde et al. 2008, Soberón & Nakamura 2009).

The use of predictive models of species potential distribution has been increasingly used in many areas related to species ecology and conservation, such as to predict areas that could potentially be re-colonised by an expanding species, to choose the best location for reintroduction/restocking or even to indicate potential areas to be prioritized for conservation purposes, including conservation planning, management and restoration (Guisan & Zimmermann 2000, Ferrier et al. 2002a,b, Soberón & Peterson 2004, Peterson 2006, Franklin 2009, Wilson et al. 2010, Rodríguez-Soto et al. 2011). Published examples indicate that SDMs can perform well in characterizing the natural distributions of species (within their current range), particularly when well-designed survey data and functionally relevant

predictors are analyzed with an appropriately specified model (Elith & Leathwick 2009). Despite the widespread use of these models, some authors (Pulliam 2000, Soberón & Peterson 2005, Araujo & Guisan 2006, Peterson 2006, Soberón 2007, Jiménez-Valverde et al. 2008) have pointed out important conceptual ambiguities as well as biotic and algorithm uncertainties that need to be investigated in order to increase confidence in model results, such as 1) clarification of model aims; 2) clarification of niche concept, including the distinction between potential and realized distribution; 3) improved design for sampling data for building model; 4) improved model parameterization; 5) improved model selection and predictor contribution; and 6) improved model evaluation.

Modeling the species distribution

Biological data as good-quality source data

Occurrence data for species distribution models can only include presence or presence-absence data. The type of data available for modeling will determine the algorithm and model procedure selection. Species distribution data can be obtained from museum or scientific collections or by field surveys. Many scientific datasets are available for download such as Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) and SpeciesLink (http://splink.cria.org.br/). There are many problems associated to these data sets mainly related to the species identification, sampling effort bias and precision of records (Soberón & Peterson 2004). Field survey data, generally obtained by species observation, trapping or track surveys, from sampling procedure ensuring a broad environmental coverage of gradients in the species distribution range (Vaughan & Ormerod 2003), avoiding bias and pitfalls, are supposed to be good quality data for species distribution modeling. Occurrence data obtained by interviews are generally not recommended to be used in modeling as they are usually not accurate in regards to the species occurrence site.

Many problems have been faced by modelers due mainly to clustered datasets and biased sampling not covering the full range of environmental conditions (e.g., environmental heterogeneity) within the landscape, especially for wide ranging species. Clustered data, especially when provided by telemetry data, could lead to a potential bias in the final model. An option to solve this apparent problem is to subsample the dataset in order to dilute the oversampling in some parts of the species distribution range (Veloz 2009).

Environmental variables as good predictors

Environmental data sets matter in species distribution modeling (Peterson & Nakazawa 2008). The role of a distribution model may be primarily predictive or, alternatively, may emphasize the relationship between an organism and its habitat (Vaughan & Ormerod 2003). So the environmental predictors should therefore have a biological relationship with the organism. The spatial scale should be carefully defined as it can influence the results and/or not resolve the motivated question of the study (Vaughan & Ormerod 2003). The selection of resolution and extent is a critical step in SDM building, and an inappropriate selection can yield misleading results (Guisan & Thuiller 2005). Ideally, models should examine a series of spatial scales, increasing the understanding of organism-environmental relationship (Vaughan & Ormerod 2003).

Many environmental variables, used as predictors, are available for download by many International Agencies. Some examples of frequently used environmental databases are global climate layers from Worldclim (http://www.worldclim.org/), elevation from the NASA Shuttle Radar Topography Mission (SRTM, http://www2.jpl.nasa.gov/srtm/), climate data from past, present and future from Intergovernmental Panel on Climate Change (IPCC, http://www.ipcc-data.org/), Hidro1K elevation derivative database from Earth Resources Observation and Science (EROS, http://eros.usgs.gov/), global land cover from ESA GlobCover 2009 Project (http://ionia1.esrin.esa.int/), and satellite images from MODIS (https://wist.echo.nasa.gov/api/).

Procedure of species distribution modeling

Some models are presence-only models such as DOMAIN (Carpenter et al. 1993) and BIOCLIM (Busby 1986, Nix 1986), while others demand presence and absence data, such as the GLM (Generalized Linear Model) and GAM (Additive Linear Model; Guisan & Zimmermann 2000). Others demand presence and background points such as Biomapper (Hirzel et al. 2002) and Maxent (Phillips et al. 2004, 2006) or presence and pseudoabsence such as GARP (Stockwell & Peter 1999). The latter was generated by locating sites randomly across the total geographical area, or 'domain', of interest (Ferrier et al. 2002a).

Maxent, one of the most recently used algorithm, estimates a target probability distribution by finding the probability distribution of maximum entropy (i.e., that is most spread out, or closest to uniform), subject to a set of constraints that represent our incomplete information about the target distribution (Phillips et al. 2004, 2006). When Maxent is applied to presence-only species distribution modeling, the pixels of the study area make up the space on which the Maxent probability distribution is defined, pixels with known species occurrence

records constitute the sample points, and the features are climatic variables, elevation, soil category, vegetation type or other environmental variables, and functions thereof (Phillips et al. 2006). Maxent offers many advantages performing extremely well in predicting occurrences in relation to other approaches (e.g., Elith et al. 2006, Phillips et al. 2006, Elith & Graham 2009) such as the better discrimination of suitable versus unsuitable areas for the species (Phillips et al. 2006), a good performance on small samples (Phillips & Dudik 2008), and theoretical properties that are analogous to the unbiased case when modeling presence-only data (Phillips et al. 2009), this is why it has been frequently used.

Model evaluation can be done by different approaches. One of the most common ones for model evaluation is the calculation of the Receiver Operating Curve (ROC) (DeLong et al. 1988). ROC plot is obtained by plotting all sensitivity values (true positive fraction) on the y axis against their equivalent (1 – sensitivity) values (false positive fraction) for all available thresholds on the *x* axis. The area under the ROC curve (AUC) provides a threshold-independent measure of overall model accuracy. AUC values should be between 0.5 (random) and 1.0 (perfect discrimination). Values lower than 0.5 indicates that prediction is worse than random (Fielding & Bell 1997).

Another option for model evaluation is measuring the model predictive success, which is the percentage of occurrence data correctly classified as positive, so measuring the omission error rate. This evaluation requires a specific threshold to convert continuous model predictions to a dichotomous classification of presence/absence (Hernandez et al. 2006). Optimal thresholds are presented and discussed on a comparative study by Liu et al. (2005). Also, Lobo et al. (2008) recommends that sensitivity and specificity should be also reported, so that the relative importance of commission and omission errors can be considered to assess the method performance.

References

Araujo M. B. & Guisan A. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33(10), 1677-1688.

Austin M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological Modelling 200, 1-19.

Busby J. R. 1986. A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook) Oerst. in southeastern Australia. Australian Journal of Ecology 11, 1-7.

Carpenter G., Gillison A. N. & Winter J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2, 667-680.

DeLong E. R., DeLong D. M. &, Clarke-Pearson D. L. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves. Biometrics 44(3), 837-845.

Elith J., Graham C. H., Anderson R. P., Dudík M., Ferrier S., Guisan A., Hijmans R. J., Huettmann F., Leathwick J. R., Lehmann A., Li J., Lohmann L. G., Loiselle B. A., Manion G., Moritz C., Nakamura M., Nakazawa Y., Overton J. M., Peterson A. T., Phillips S. J., Richardson K. S., Scachetti-Pereira R., Schapire R. E., Soberon J., Williams S., Wisz M. S. & Zimmermann N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129-151.

Elith J. & Graham C. H. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66-77.

Elith J. & Leathwick J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology and Evolution Systematics 40, 677-97.

Ferrier S., Watson G., Pearce J. & Drielsma M. 2002a. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New SouthWales. I. Species-level modeling. Biodiversity and Conservation 11, 2275-2307.

Ferrier S., Watson G., Pearce J. & Drielsma M. 2002b. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modeling. Biodiversity and Conservation 11, 2309-2338.

Fielding A. H. & Bell J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38-49.

Franklin J. 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge Univ. Press, Cambridge, United Kingdom. 338 pp.

Guisan A. & Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993-1009.

Guisan A. & Zimmermann N. E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147-186.

Hernandez P. A., Graham C. H., Master L. L. & Albert D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773-785.

Hirzel A. H. & Lay G. Le. 2008. Habitat suitability modelling and niche theory. Journal of Applied Ecology 45, 1372-1381.

Hirzel A. H., Hausser J., Chessel D. & Perrin N. 2002. Ecological-niche factor analysis: How to compute habitat-suitability map without absence data. Ecology 83, 2027-2036.

Jiménez-Valverde A., Lobo J. M. & Horta J. 2008. Not as good as they seem: The importance of concept in species distribution modeling. Diversity Distributions 14, 885-890.

Kearney M. 2006. Habitat, environment, and niche: What are we modeling? Oikos 115, 186-191.

Liu C., Berry P. M., Dawson T. P. & Pearson R. G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28(3), 385-393.

Lobo J. M., Jiménez-Valverde A. & Real R. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145-151.
Nix H. A. 1986. A biogeographic analysis of Australian elapid snakes. *In* Atlas of Australian Elapid Snakes. Longmore R. (Ed). Australian Government Publishing Service, Canberra, Australia, pp. 4-15.
Pearson R.G. 2007. Species' Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.
Peterson A. T. 2001. Predicting species' geographic distributions based on ecological niche modeling. Condor 103, 599-605.
Peterson A. T. 2006. Uses and requirements of ecological niche models and related distributional models. Biodiversity Informatics 3, 59-72.
Peterson A. T. & Nakazawa Y. 2008. Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. Global Ecology and Biogeography 17, 135-144.
Phillips S. J. & Dudik M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.
Phillips S. J., Anderson R. P. & Schapire R. E. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231-259.
Phillips S. J., Dudík M. & Schapire R. E. 2004. A maximum entropy approach to species distribution modeling. *In* Proceedings of the 21st International Conference on Machine Learning. 21st International Conference on Machine Learning. ACM Press, New York, pp. 655-662.
Phillips S. J., Dudík M., Elith J., Graham C. H., Lehmann A., Leathwick J. & Ferrier S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19(1), 181-197.
Pulliam H. R. 2000. On the relationship between niche and distribution. Ecology Letters 3, 349-361.
Rodríguez-Soto C. R., Monroy-Vilchis O., Maiorano L., Boitani L., Faller M. A., Briones M. A., Nunez R., Rosas-Rosas O., Ceballos G. & Falcucci A. 2011. Predicting potential distribution of the jaguar (*Panthera onca*) in Mexico: identification of priority áreas for conservation. Diversity and Distribution 17, 350-361.
Sanderson E., Redford K., Chetkiewicz C., Medellin R., Rabinovitz A. R., Robinson J. G. & Taber A. 2002. Planning to save a species: the jaguar as a model. Conservation Biology 16, 58-72.
Soberón J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecology Letters 10, 1115-1123.
Soberón J. & Peterson A. T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society B 359, 689-698.
Soberón J. M. & A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics 2, 1-10.
Soberón J. & Nakamura M. 2009. Niches and distributional areas: Concepts, methods, and assumptions. The Proceedings of the National Academy of Sciences (PNAS) 106(2), 19644-19650.
Stockwell D. R. B. & Peters D. P. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. International Journal of Geographical Information Systems 13, 143-158.
Vaughan I. P. & Ormerod S. J. 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42, 720-730.

Veloz S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36, 2290-2299.
Wilson C. D., Roberts D. & Reid N. 2010. Applying species distribution modelling to identify areas of high conservation value for endangered species: A case study using *Margaritifera margaritifera* (L.). Biological Conservation 144(2), 821-829.