

Using epidemiological survey data to infer geographic distributions of leishmaniasis vector species

Utilização dos dados de levantamentos epidemiológicos para inferir a distribuição geográfica de vetores da leishmaniose

A. Townsend Peterson¹, Ricardo Scachetti Pereira²
and Vera Fonseca de Camargo Neves³

ABSTRACT

An important aspect of tropical medicine is analysis of geographic aspects of risk of disease transmission, which for lack of detailed public health data must often be reduced to an understanding of the distributions of critical species such as vectors and reservoirs. We examine the applicability of a new technique, ecological niche modeling, to the challenge of understanding distributions of such species based on municipalities in the State of São Paulo in which a group of 5 Lutzomyia sandfly species have been recorded. The technique, when tested based on independent occurrence data, yielded highly significant predictions of species' distributions; minimum sample sizes for effective predictions were around 40 municipalities.

Key-words: Ecological niche modeling. Genetic algorithm for rule-set prediction. Lutzomyia. Leishmaniasis.

RESUMO

Um aspecto importante da medicina tropical é a análise de aspectos geográficos relacionados com o risco de transmissão de doenças. Devido à ausência de dados detalhados de saúde pública, estas análises são frequentemente reduzidas à compreensão da distribuição de espécies críticas como vetores e reservatórios. Neste trabalho, é examinada a aplicabilidade de uma nova técnica, a modelagem de nicho ecológico, no estudo da distribuição destas espécies nos municípios do Estado de São Paulo, onde um grupo de 5 mosquitos do gênero Lutzomyia foi encontrado. A técnica foi testada em conjuntos de dados independentes, resultando em previsões altamente significativas; a amostragem mínima para se obter previsões eficazes foi de cerca de 40 municípios.

Palavras-chaves: Modelagem de nicho ecológico. Genetic algorithm for rule-set prediction. Lutzomyia. Leishmaniose.

Great efforts are expended to understand the distributions of animal species relevant to disease systems, particularly vectors and reservoirs for particular diseases. For cutaneous leishmaniasis, although reservoirs remain poorly known²⁴, much energy has been dedicated to documentation of distributions of vector species, sandflies in the genus *Lutzomyia*^{4,6,9}, which are related to risk of disease transmission⁵. These efforts accumulate lists of sites or municipalities from which vector species are known to provide an idea of geographic distributions of important species.

These maps of known occurrences of species, however, present a biased picture of species' geographic distributions, mixing the

ecological needs and biogeography of a species with the geography of *sampling* of the species^{10,16}. In this sense, the known distribution of a species provides a view of its distribution that is at best incomplete, if not actually misleading. A critical step towards improving this picture is one of inference into unsampled and undersampled areas; this inference can be achieved via models of the ecological niche of species of interest²².

In this contribution, we explore the potential of ecological niche modeling techniques for interpolating into unsampled areas for understanding vector species' geographic distributions. We use multiple subsamples of available distributional points to approach

1. Natural History Museum, The University of Kansas, Lawrence, Kansas 66045 USA. 2. Centro de Referência em Informação Ambiental (CRIA), Campinas, SP, Brasil. 3. Superintendência de Controle de Endemias (SUCEN), São Paulo, SP, Brasil.

Supported by: Fundação de Amparo à Pesquisa de São Paulo

Address to : Dr. A. Townsend Peterson. Natural History Museum, The University of Kansas, Lawrence, Kansas 66045 USA.

Tel: 785 864-4540; 785 864-5335

e-mail: town@ku.edu

Recebido para publicação em: 30/12/2002

Aceito em: 19/08/2003

the question of how much sampling is needed to assemble a good distributional understanding for a vector species²⁹. In broader terms, we present the application of a method that can be generally useful in characterizing geographic distributions of vector and reservoir species based on incomplete or imprecise existing data.

MATERIAL AND METHODS

Input occurrence data. Ecological niche models were based on 366 unique occurrence records for the 5 most dominant *Lutzomyia* species in São Paulo state, Brazil, with overall sample sizes ranging 40-112 points per species. Distributional data for these species (*Lutzomyia fischeri*, *L. intermedia*, *L. migonei*, *L. pessoai*, and *L. whitmani*) were drawn from previous, intensive sampling in municipalities across the state⁶. All occurrence points - perforce given how the data were collected and also to mimic many of similar datasets available in other similar situations - were georeferenced to the centroids of the municipalities. To provide independent data sets for model building (input data) and model testing (extrinsic test data), and to assess sample size needs for modeling these species in this region, we randomly selected municipalities to create input training data sets representing 10%, 30%, 50%, 70%, and 90% of available points - remaining points were used for testing model quality.

Ecological niche modeling. Ecological niches were modeled using the *Genetic algorithm for rule-set prediction* (GARP)^{25,26,27}, a machine-learning software package now available for public download (<http://www.beta.lifemapper.org/desktopgarp/>). In general, the procedure focuses on modeling ecological niches (the conjunction of ecological conditions within which a species is able to maintain populations without immigration)⁸. Specifically, GARP relates ecological characteristics of known occurrence points to those of points randomly sampled from the rest of the study region, seeking to develop a series of decision rules that best summarize those factors associated with the species' presence^{25,26,27}.

Within GARP, input data are further divided randomly and evenly into training and intrinsic testing data sets. GARP works in an iterative process of rule selection, evaluation, testing, and incorporation or rejection: a method is chosen from a set of possibilities (e.g., logistic regression, bioclimatic rules), applied to the training data, and a rule is developed or evolved^{25,26,27}. Predictive accuracy is then evaluated based on 1250 points resampled from the test data and 1250 points sampled randomly from the study region as a whole. Rules may evolve by a number of means that mimic DNA evolution: point mutations, deletions, crossing over, etc. The change in predictive accuracy from one iteration to the next is used to evaluate whether a particular rule should be incorporated into the model, and the algorithm runs either 1000 iterations or until convergence.

All modeling in this study was carried out on a desktop implementation of GARP that offers much-improved flexibility in choice of predictive environmental/ecological GIS data coverages. In this case, we used 15 data layers summarizing aspects of topography [elevation, slope, aspect, flow accumulation, flow direction, and topographic index (tendency to pool water) from the U.S. Geological Survey's (<http://edcdaac.usgs.gov/gtopo30/hydro/>) Hydro-1K data set]; aspects of climate including daily temperature

range, mean annual precipitation, maximum, minimum, and mean annual temperatures, vapor pressure, and wet days (annual means over the period 1960-1990 from the Intergovernmental Panel on Climate Change (<http://www.ipcc.ch/>); and aspects of land use and land cover including an overall land cover classification and a tree cover map (based on AVHRR satellite imagery for 1992-1993, University of Maryland Global Land Cover Facility (<http://glcf.umd.edu/index.shtml>) for an area consisting of all of São Paulo state, Brazil. GARP's predictive abilities have been tested and proven under diverse circumstances^{1,2,7,12,13,14,15,17,18,19,21,22,23,28,29}.

We developed multiple replicate models of each species' ecological niche. Unlike previous applications, which either used single models to predict species' distributions^{12,13} or summed multiple models to incorporate model-to-model variation²³, we used a new procedure³ for choosing best subsets of models. The procedure is based on the observations that 1) models vary in quality, 2) variation among models involves an inverse relationship between error of omission (leaving out true distributional area) and commission (including areas not actually inhabited), and 3) best models (as judged by experts blind to error statistics) are clustered in a region of minimum omission of independent test points and moderate area predicted (an axis related directly to commission error). The relative position of the cloud of points relative to the two error axes provides an assessment of the relative accuracy of each model. To choose best subsets of models, we 1) produced replicate models until we had produced 20 models with omission error of <5% based on independent intrinsic test points, 2) calculated the median area predicted present among these minimum-omission points, 3) identified the 10 models closest to the overall median area predicted, and 4) summed these 'best subsets' models.

Projection of the rule-sets for these models back onto geography provided distributional predictions for each species. We tested model quality via the independent extrinsic test sets of occurrence points in two ways: one using all available test data points to permit best estimation of levels of omission error, and the other using yet another random subsetting down to 10% of the total occurrence points available for the species to avoid differences in statistical power owing to different sampling densities. The χ^2 tests were used to compare observed success in predicting distributions of test points with those expected under random models (proportional area predicted present provides an estimate of the proportion of occurrence points correctly predicted were the prediction to be random with respect to the distribution of the test points). Predicted presence was conservatively defined as the area in which all best-subsets models agreed in predicting presence.

RESULTS AND DISCUSSION

Ecological niche models and predictions of geographic distributions of species predictably improved in their performance as training sample size increased. For example, in models for *Lutzomyia fischeri*, 1) omission error was quite high at 10% training data density, moderate at 30% training data density, and lower thereafter; and 2) statistical significance was unevenly related to sample size and omission error (Figure 1). Results were similar across all species, with low model significance at the smallest training sample sizes; interestingly, model significance was also lower at the largest training sample sizes (Table 1 and Figure 1).

Inspecting levels of omission error across different data densities for each species, trends are remarkably coincident (Figures 2 and 3). At smallest training sample sizes (<10 points), all models for all species showed high omission error rates (80-100%). In each case, at intermediate sample sizes (40-50 points), omission rates reached minima, and were relatively constant or slightly higher thereafter (Figure 3).

Overall, model predictivity was maximal at intermediate sample sizes, when both training and test data sample sizes are substantial, and neither is small. When training sample sizes are too small, model parameters are not estimated accurately. On the other hand, when testing sample sizes are small, statistical power is too low to detect good models. These results coincide with the results of

Table 1 - Summary of results of predictions for five *Lutzomyia* species across five data densities (10-90% of available occurrence data for training).

Species	Training				
	10%	30%	50%	70%	90%
<i>Lutzomyia fischeri</i>	95.3 ***/-	72.0 -/-	55.6 -/-	36.4 **/*	42.9 +/+
<i>Lutzomyia intermedia</i>	85.4 ***/*	46.9 ***/-	28.6 ***/*	38.2 ***/*	41.7 -/-
<i>Lutzomyia migonei</i>	82.8 **/-	69.6 ***/-	36.4 ***/*	35.0 +/-	28.6 +/+
<i>Lutzomyia pessoai</i>	100 -/-	85.7 **/**	55.0 +/-	50.0 +/-	66.7 -/-
<i>Lutzomyia whitmani</i>	98.5 -/-	64.2 ***/-	31.6 ***/*	26.1 **/*	37.5 +/+

Numbers presented are percent omission values, as measured with independent test data (- no significance, (+) 0.1 > P > 0.05, * 0.05 > P > 0.01, ** 0.01 > P > 0.001, *** P < 0.001).


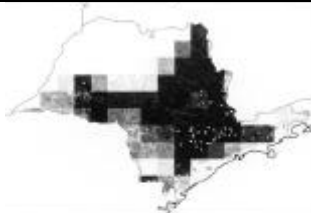
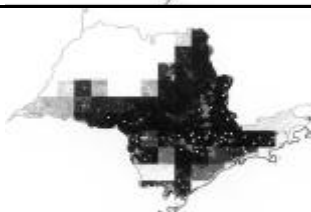
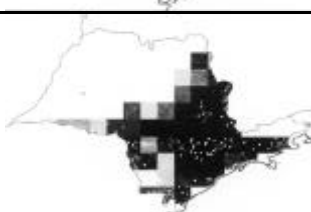
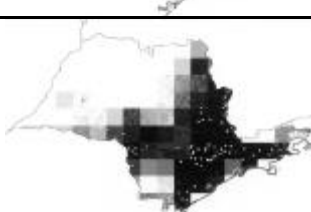
Pewrcent of points for building models	Number of points for building models	Significance (all test points / 10% for test)	No omission	Result
10	7	*** / -		
30	22	- / -		
50	37	- / -	X	
70	52	** / *	X	
90	67	++ / ++	X	

Figure 1 - Summary of entire spectrum of training data density for model-building for *Lutzomyia fischeri*, showing model improvement in terms of avoidance of omission (X's indicate models that do not omit heavily) and in terms of model significance (- no significance, + 0.1 > P > 0.05, * 0.05 > P > 0.01, ** 0.01 > P > 0.001, *** P < 0.001).

previous tests of the effects of sample size on model predictivity^{13,29}; in the present case, sample sizes considered adequate (40-50 unique points) were somewhat higher than in previous studies, probably owing to the imprecise georeferencing involved. More generally, the unreliability and unpredictable behavior of significance tests regarding model predictions coincides with the results of previous comparative tests¹⁷.

The asymptotes of the omission error X training sample size curves are relatively high - that is, even the best models are still plagued by error rates of 30-40%. This seemingly poor performance results from the use of municipality centroids for georeferencing occurrence records - a sizeable proportion of centroids may fall in areas of predicted absence, even though some portion of the

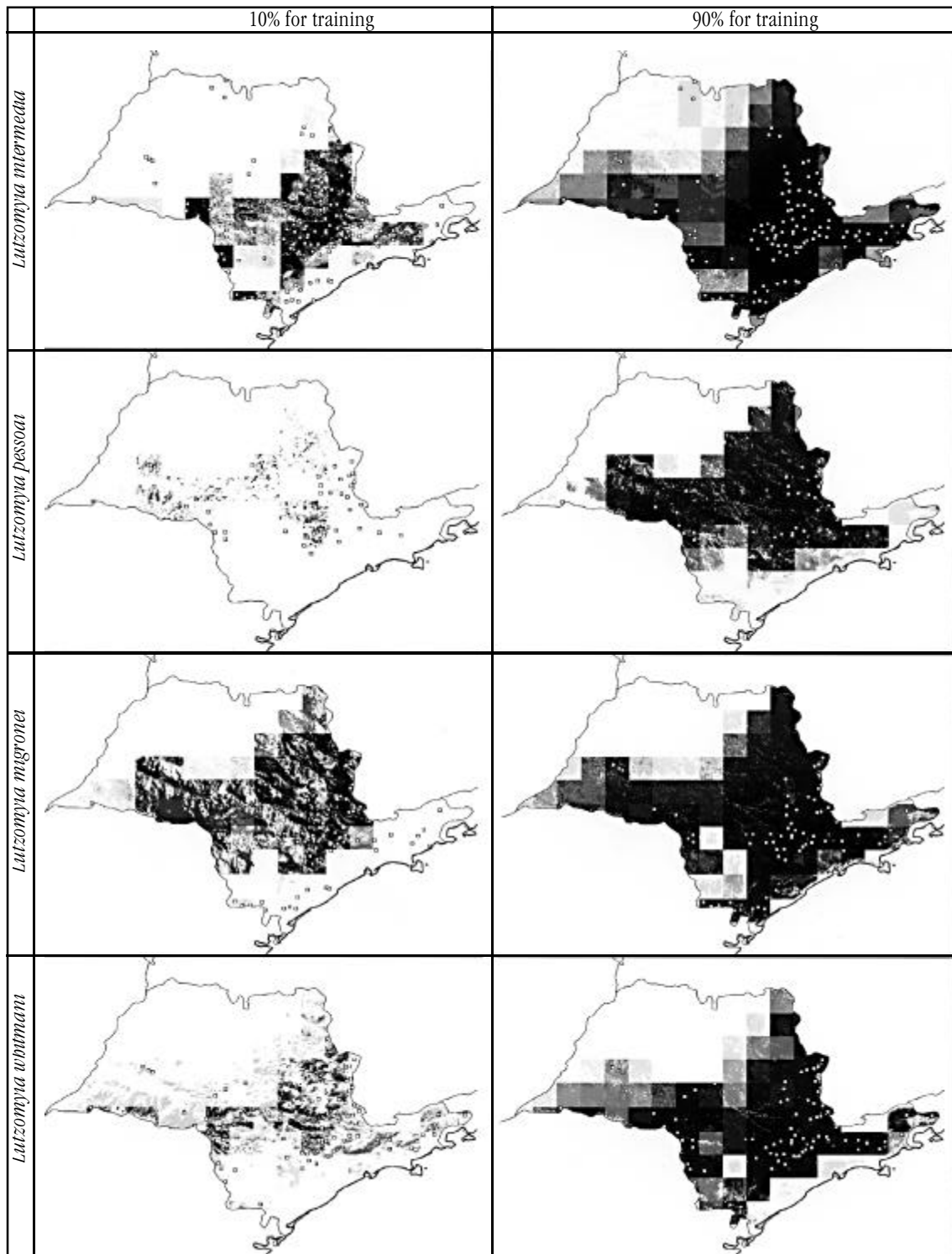


Figure 2 - Comparison of models for *Lutzomyia* species in São Paulo State based on 10% of points for training models versus 90% of points for training models.

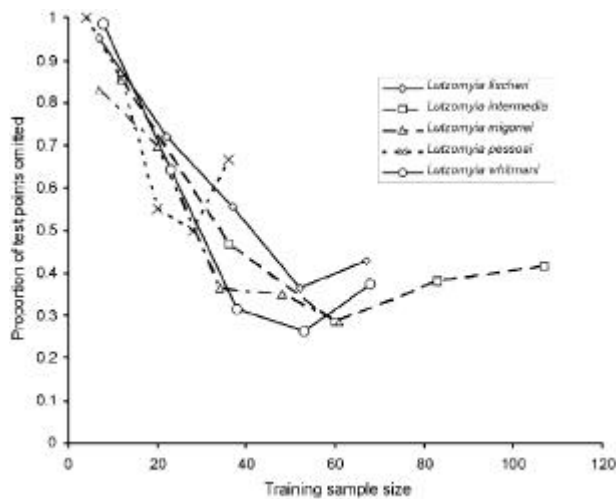


Figure 3 - Summary of model quality as a function of training point sample size for *Lutzomyia* species in São Paulo State. Model quality is measured as proportional omission of independent test points.

municipality is predicted present. Hence, given this imprecise level of georeferencing, a certain base level of omission error is to be expected.

On a much more positive note, this study demonstrates that moderate sampling densities - at sample sizes that likely characterize many epidemiological surveys of vector or reservoir distributions - are sufficient to produce excellent summaries of species' geographic distributions. That is, even with moderate sample sizes, it is possible to interpolate into unsampled or poorly sampled areas, and produce reliable and predictive maps of species' geographic distributions. This capacity permits development of geographic predictions for poorly known species important in understanding the geography of disease systems, which have important implications for human health issues¹¹.

ACKNOWLEDGMENTS

This study was funded by the U.S. National Science Foundation. ATP's work in Brazil, as well as a significant portion of RSP's time was supported by a grant from the Fundação de Amparo à Pesquisa do Estado de São Paulo.

REFERENCES

1. Anderson RP, Gomez M, Peterson AT. Geographical distributions of spiny pocket mice in South America: Insights from predictive models. *Global Ecology and Biogeography* 11:131-141, 2002a.
2. Anderson RP, Laverde M, Peterson AT. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* 93:3-16, 2002b.
3. Anderson RP, Lew D, Peterson AT. Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecological Modelling* 162:211-232, 2003.
4. Azevedo ACR, Monteiro FA, Cabello PH, Souza NA, Rosa-Freitas MG, Rangel EF. Studies on populations of *Lutzomyia longipalpis* (Lutz & Neiva, 1912) (Diptera: Psychodidae: Phlebotominae) in Brazil. *Memórias do Instituto Oswaldo Cruz* 95:305-322, 2000.
5. Camargo-Neves VLF, Gomes AC, Antunes JLF. Correlation of the presence of phlebotominae species (Diptera: Psychodidae) with records of American tegumentary leishmaniasis cases in the state of São Paulo, Brazil. *Revista da Sociedade Brasileira de Medicina Tropical* 35:299-306, 2002.

6. Ferreira AL, Sessa PA, Varejão JBM, Falqueto A. Distribution of sand flies (Diptera: Psychodidae) at different altitudes in an endemic region of American cutaneous leishmaniasis in the State of Espírito Santo, Brazil. *Memórias do Instituto Oswaldo Cruz* 96:1061-1067, 2001.
7. Godown ME, Peterson AT. Preliminary distributional analysis of U.S. endangered bird species. *Biodiversity and Conservation* 9:1313-1322, 2000.
8. Grinnell J. Field tests of theories concerning distributional control. *American Naturalist* 51:115-128, 1917.
9. Marcondes CB, Lozovei AL, Vilela JH. Geographic distribution of phlebotomine sandflies of the *Lutzomyia intermedia* (Lutz & Neiva, 1912) complex (Diptera, Psychodidae). *Revista da Sociedade Brasileira de Medicina Tropical* 31:51-58, 1998.
10. Navarro-Sigüenza AG, Peterson AT, Gordillo-Martínez A. A Mexican case study on a centralised database from world natural history museums. *CODATA Journal* 1:45-53, 2002.
11. Nissen MD, Walker JC. Bancroftian filariasis. *In: e-medicine*, <http://www.emedicine.com/ped/topic205.htm>, 2002.
12. Peterson AT. Predicting species' geographic distributions based on ecological niche modeling. *Condor* 103:599-605, 2001.
13. Peterson AT, Ball LG, Cohoon KC. Predicting distributions of tropical birds. *Ibis* 144: E27-E32, 2002a.
14. Peterson AT, Cohoon KC. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling* 117:159-164, 1999.
15. Peterson AT, Egbert SL, Sánchez-Cordero V, Price KP. Geographic analysis of conservation priorities using distributional modelling and complementarity: Endemic birds and mammals in Veracruz, Mexico. *Biological Conservation* 93:85-94, 2000.
16. Peterson AT, Navarro-Sigüenza AG, Benítez-Díaz H. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* 140:288-294, 1998.
17. Peterson AT, Ortega-Huerta MA. Modeling ecological niches and predicting geographic distributions: A test of five methods. *Ecological Modelling*, Submitted.
18. Peterson AT, Ortega-Huerta MA, Bartley J, Sánchez-Cordero V, Soberón J, Buddemeier RH, Stockwell DRB. Future projections for Mexican faunas under global climate change scenarios. *Nature* 416:626-629, 2002b.
19. Peterson AT, Sánchez-Cordero V, Soberón J, Bartley J, Buddemeier RH, Navarro-Sigüenza AG. Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling* 144:21-30, 2001.
20. Peterson AT, Scachetti-Pereira R, Hargrove WW. Potential distribution of Asian longhorned beetles (*Anoplophora glabripennis*) in North America. *American Midland Naturalist*, 2004.
21. Peterson AT, Soberón J, Sánchez-Cordero V. Conservatism of ecological niches in evolutionary time. *Science* 285:1265-1267, 1999.
22. Peterson AT, Stockwell DRB, Kluza DA. Distributional prediction based on ecological niche modeling of primary occurrence data. *In: Scott JM, Heglund PJ, Morrison ML (eds) Predicting Species Occurrences: Issues of Scale and Accuracy*. Island Press, Washington, DC, pp. 617-623, 2002c.
23. Peterson AT, Vieglais DA. Predicting species invasions using ecological niche modeling. *BioScience* 51:363-371, 2001.
24. Shaw JJ. New World leishmaniasis: The ecology of leishmaniasis and the diversity of leishmanial species in Central and South America. *In: J Farrell J (ed) World Class Parasites: Leishmania*, 4. Kluwer Academic Publishers, Boston, pp. 11-31, 2002.
25. Stockwell DRB. Genetic algorithms II. *In, AH Fielding Machine learning methods for ecological applications*. Kluwer Academic Publishers, Boston, pp. 123-144, 1999.
26. Stockwell DRB, Noble IR. Induction of sets of rules from animal distribution data: A robust and informative method of analysis. *Mathematics and Computers in Simulation* 33:385-390, 1992.
27. Stockwell DRB, Peters DP. The GARP modelling system: Problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems* 13:143-158, 1999.
28. Stockwell, DRB, Peterson AT. Controlling bias in biodiversity data. *In: JM Scott, Heglund PJ, Morrison ML. Predicting Species Occurrences: Issues of Scale and Accuracy*. Island Press, Washington, DC, pp. 537-546, 2002a.
29. Stockwell DRB, Peterson AT. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148:1-13, 2002b.