The Condor 103:599–605

PREDICTING SPECIES' GEOGRAPHIC DISTRIBUTIONS BASED ON ECOLOGICAL NICHE MODELING

A. TOWNSEND PETERSON[1]
Natural History Museum, The University of Kansas, Lawrence, Kansas 66045

Abstract. Recent developments in geographic information systems and their application to conservation biology open doors to exciting new synthetic analyses. Exploration of these possibilities, however, is limited by the quality of information available: most biodiversity data are incomplete and characterized by biased sampling. Inferential procedures that provide robust and reliable predictions of species' geographic distributions thus become critical to biodiversity analyses. In this contribution, models of species' ecological niches are developed using an artificial-intelligence algorithm, and projected onto geography to predict species' distributions. To test the validity of this approach, I used North American Breeding Bird Survey data, with large sample sizes for many species. I omitted randomly selected states from model building, and tested models using the omitted states. For the 34 species tested, all predictions were highly statistically significant (all $P < 0.001$), indicating excellent predictive ability. This inferential capacity opens doors to many synthetic analyses based on primary point occurrence data.

Key words: ecological niche, GARP, geographic distribution, GIS.

Predicción de Áreas de Distribución de Especies con Pase en Modelaje de Nichos Ecológicos

Resumen. Avances recientes en los sistemas de información geográfica y su aplicación en la biología de conservación presentan la posibilidad de analisis nuevos y sintéticos. La exploración de estas posibilidades, de todas formas, se limita por la calidad de información disponible: la gran mayoria de datos respecto a la diversidad biológica son incompletos y sesgados. Por eso, procedimientos de inferencia que proveen predicciones robustas y confiables de distribuciones de especies se hacen importantes para los análisis de la biodiversidad. En esta contribución, se desarrollan modelos de los nichos ecológicos por medio de un algoritmo de inteligencia artificial, y los proyeccionamos en la geografía para predecir las distribuciones geográficas de especies. Para probar el método, se usan los datos del North American Breeding Bird Survey, con tamaños de muestra grande. Se construyeron modelos con base en 30 estados unidenses seleccionados al azar, y se probaron los modelos con base en los 20 estados restantes. De las 34 especies que se analizaron, todos mostraron un alto grado de significanza estadística (todos $P < 0.001$), lo cual indica un alto grado de predictividad. Esta capacidad de inferencia abre la puerta a varios analisis sintéticos con base en puntos conocidos de ocurrencia de especies.

INTRODUCTION

Many geographic applications have been developed in recent years that offer exciting new possibilities for understanding biological diversity (e.g., Scott et al. 1996). Geographic information systems (GIS) make it possible to build maps of species richness and endemism, to prioritize areas for conservation based on principles such as complementarity, and to assess the completeness of existing protected areas networks (e.g., Peterson et al. 2000). One of the most notable examples of these possibilities is that of Gap Analysis, an integrative program that links distributional information with information on land use and protection to identify priorities for conservation action (Scott et al. 1996). The success of such programs and approaches, however, depends critically on the quality of distributional information available, which has proven to be a weak link in the process (Krohn 1996).

Biodiversity information nevertheless exists in a difficult, fragmented system: sampling documents presence but rarely absence; sampling is rarely systematically planned so as to permit detailed statistical analysis; and institutional holdings separate specimens in different countries and regions (Peterson et al. 1998). Although occurrence data are now beginning to become much more available thanks to innovative, Internet-based technological developments (e.g.,

Manuscript received 20 January 2000; accepted 26 February 2001.
[1] E-mail: town@ukans.edu

Vieglais 1999), the need for development of inferential approaches to interpreting biodiversity information is clear (Soberón 1999). Hence, in this contribution, I develop detailed statistical tests of an artificial-intelligence-based approach designed to predict species' geographic distributions.

MODELING ECOLOGICAL NICHES AND PREDICTING GEOGRAPHIC DISTRIBUTIONS

The fundamental ecological niche of a species is a critical determinant of its distribution; as such, it is defined in multidimensional ecological space (MacArthur 1972). Several distinct interpretations of ecological niches exist: most relevant to the present contribution is that of Grinnell (1917), who focused on the conjunction of ecological conditions within which a species is able to maintain populations without immigration. Hutchinson (1959) provided the valuable distinction between the fundamental niche, which is the range of theoretical possibilities, and the realized niche (that part which is actually occupied, given interactions with other species such as competition). Although it can be argued that only the realized niche is observable in nature, by examining species across their entire geographic distributions, species' distributional possibilities can be observed against varied community backgrounds, and thus a view of the fundamental ecological niche can be assembled (Peterson et al. 1999).

Several approaches have been used to model species' fundamental ecological niches. The very simplest is BIOCLIM (Nix 1986), which involves tallying species' occurrences in categories for each environmental dimension, trimming the extreme 5% of the distribution along each ecological dimension, and taking the niche as the conjunction of the trimmed ranges to produce a decision rule. BIOCLIM suffers generally from high rates of commission error, or overprediction (Stockwell and Peterson, unpubl.). Other investigators have applied logistic regression to the challenge of combining environmental variables into predictions of presence and absence (e.g., Austin et al. 1990).

The *Genetic Algorithm for Rule-set Prediction* (GARP) includes several distinct algorithms in an iterative, artificial-intelligence-based approach (Stockwell and Noble 1992, Stockwell and Peters 1999). Here, individual algorithms (e.g., BIOCLIM, logistic regression) are used to produce component ''rules'' in a broader ruleset, and hence portions of the species' distribution may be determined as within or without its niche, based on different rules from several algorithms. As such, GARP is a superset of other approaches, and should always have greater predictive ability than any one of them. Initial testing of GARP has indicated excellent predictive ability and insensitivity to BIOCLIM's problems with dimensionality of environmental data (Peterson and Cohoon 1999, Peterson et al., in press a, b; Stockwell and Peterson, in press).

Two general types of error enter into such niche modeling and geographic prediction efforts (Fielding and Bell 1997). First, omission of areas actually inhabited represents a failure of the modeling effort to extend to all ecological conditions under which the species is able to maintain populations. Second, commission error is that of including areas actually uninhabited; this error includes two components: real commission error, in which combinations of ecological conditions not actually within the species' niche are included, and apparent commission error, which results from species' absences owing to interspecific interactions (the difference between realized and fundamental niches, MacArthur 1972), as well as to historical factors, such as limited colonization ability, speciation patterns, and local extinction (Peterson et al. 1999). In this sense, apparent commission error represents a real feature of species' distributional ecology: not all habitable areas are inhabited (Peterson et al. 1999). The purpose of the present contribution is to put the GARP algorithm to a rigorous test with North American birds.

METHODS

Distributional data for four genera of passerine birds (*Catharus, Dendroica, Toxostoma,* and *Vireo*) were selected for analysis based on their high species richness, distribution in regions well covered by the North American Breeding Bird Survey (http://www.mbr-pwrc.usgs.gov/bbs/bbs.html), and ease of detection in visual/auditory surveys. The Breeding Bird Survey data offer relatively uniform coverage of the continent, avoiding some of the challenges presented by museum specimen data in terms of uneven sampling (Peterson et al. 1998). Data points were extracted as presences (in any year) or absences (in all years) at the level of routes,

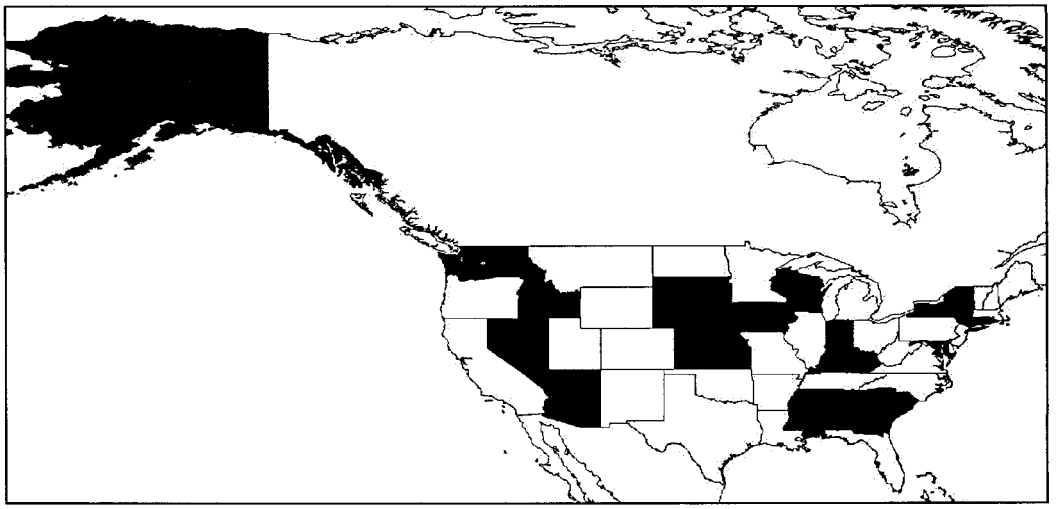

FIGURE 1. Map showing 20 states (in black) chosen randomly for testing distributional models.

and reduced to unique latitude-longitude combinations for each species. Thirty U.S. states were chosen at random for model development (''training data,'' regardless of whether the species had been recorded from the state); data from the remaining 20 states were set aside for statistical testing of models (''test data,'' Fig. 1); this ratio of training and test sample sizes was chosen so that in general more than 10 and 30 occurrence points would be available for testing and training models, respectively. This scheme is reasonable, given that the probability of detection of a particular species in one state in no way affects the probability of its detection in another state.

Four species of *Catharus*, 18 of *Dendroica*, 7 of *Toxostoma*, and 12 of *Vireo* were available in the data set, although 7 species had to be omitted because they did not occur in both training and test data sets (*Catharus minimus, Dendroica chrysoparia, Toxostoma longirostre, T. rediviyum, Vireo altiloquus, V. atricapillus,* and *V. flavoviridis*). Training data were analyzed for North America with a 50 $\times$ 50 km pixel resolution using the web-based GARP facility (http://biodi.sdsc.edu/), including coverages of (1–4) mean and standard deviation of annual mean precipitation and annual mean temperature, (5) life zones, (6) wetlands, (7) vegetation types, and (8) soil types. Variable combinations predicted present by the GARP rules were identified in the test states and used to predict species' occurrences in those states. Resulting geographic predictions were exported as ASCII raster grid files for use in ArcView (version 3.1).

In ArcView, the test occurrence data were overlain on the predictions for the 20 test states. Numbers of points correctly and incorrectly predicted by GARP were used as observed values. Expected numbers were taken as the test sample size multiplied by the proportional area predicted present in test states. A chi-square analysis for each species was used to assess model significance. To permit visualization of the ecological niche model, I crossed the eight ecological coverages (Combine option in ArcView) with the distributional prediction to produce a table of predicted presences and absences across environmental combinations.

RESULTS

Ecological niche models for each species showed restriction relative to the universe of ecological combinations available across North America. For example, the Brown Thrasher (*Toxostoma rufum*) was modeled to focus its distribution in relatively warm, yet relatively dry portions of the continent (Fig. 2). Similar visualizations of ecological distributions were developed for other ecological dimensions, and for each species.

Geographic distributions for all 34 species in the analysis were highly significantly predicted in the test states. For example, of 741 test points

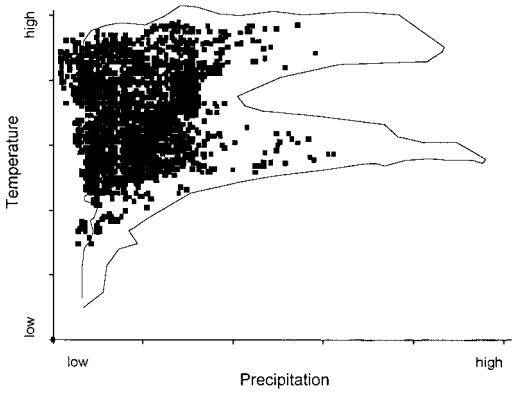

FIGURE 2. Two-dimensional visualization of the ecological niche of the Brown Thrasher (*Toxostoma rufum*), in which the availability of combinations of annual mean temperature and annual precipitation (both rescaled between 0 and 254) across North America is shown with the black outline, and the combinations predicted habitable for the species are shown as black squares.

for Brown Thrashers, 715 were correctly predicted, even though only 39% of the 20-state test area was predicted present (Fig. 3), and so only 290 points would have been correctly predicted by a random model. Although most range limits are accurately delimited in the distributional prediction, an area of overprediction runs from the southwestern extreme of the species' distribution south into northern Mexico; here, other *Toxostoma* species are present, and this is therefore another example of stability of ecological niches on evolutionary time scales (Peterson et al. 1999). This model was statistically significant at $P < 10^{-222}$, and hence it is highly likely that the model is accurately evaluating dimensions of the species' ecological requirements. Significance levels for all species ranged between 10^{-245} and 10^{-3} (Table 1).

Relationships between model quality and sample size (Fig. 4) were strong (simple linear regression, $P < 0.05$). Small chi-square values were associated with small sample sizes in both

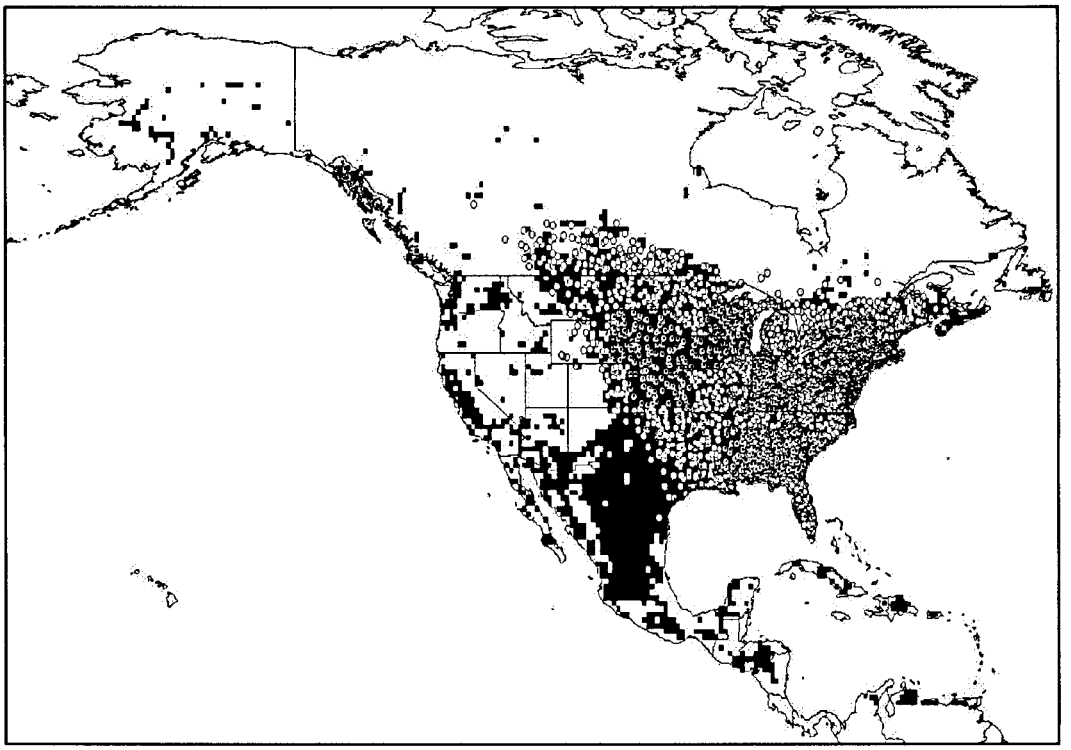

FIGURE 3. Example map of geographic predictions for Brown Thrasher (*Toxostoma rufum*). Black areas represent areas of predicted presence; open circles represent training points used to develop the model; and dotted circles represent test points used to test model accuracy.

TABLE 1. Summary of distributional predictions and significance tests for 34 species in the genera *Catharus*, *Dendroica*, *Toxostoma*, and *Vireo*. n_{train} = training sample size, n_{test} = test sample size, % correct = percentage of test points correctly predicted, $A_{present}$ = area predicted present (in 50 × 50 km pixel units), and A_{absent} = area predicted absent (in pixel units).

Species	n_{train}	n_{test}	% correct	$A_{present}$	A_{absent}	χ^2	P
Catharus fuscescens	526	277	93	281	285	210	2.4×10^{-46}
Catharus guttatus	599	302	75	183	383	250	6.3×10^{-55}
Catharus ustulatus	394	246	68	173	393	162	7.9×10^{-36}
Dendroica caerulescens	486	97	62	201	364	176	7.5×10^{-39}
Dendrica castanea	266	14	95	164	402	34	3.6×10^{-8}
Dendroica cerulea	267	127	89	155	411	281	1.2×10^{-61}
Dendroica discolor	576	353	95	185	381	691	8.0×10^{-151}
Dendroica dominica	410	251	88	114	451	886	5.3×10^{-193}
Dendroica fusca	532	111	83	225	340	144	6.9×10^{-32}
Dendroica graciae	24	20	100	83	482	40	1.8×10^{-9}
Dendroica magnolia	610	102	85	240	326	121	6.5×10^{-27}
Dendroica nigrescens	237	72	88	118	447	185	6.5×10^{-41}
Dendroica occidentalis	108	15	100	165	400	30	3.4×10^{-7}
Dendroica palmarum	141	7	100	218	348	11	3.8×10^{-3}
Dendroica pensylvanica	769	220	94	341	224	110	1.1×10^{-24}
Dendroica pinus	699	347	98	251	314	365	5.6×10^{-80}
Dendroica striata	239	75	96	144	422	10	6.5×10^{-3}
Dendroica tigrina	300	22	96	78	488	123	1.6×10^{-27}
Dendroica townsendi	174	107	65	127	439	152	1.2×10^{-33}
Dendroica virens	697	150	96	371	195	76	3.3×10^{-17}
Toxostoma bendirei	33	28	86	82	484	127	2.3×10^{-28}
Toxostoma crissale	33	32	93	90	476	170	1.1×10^{-37}
Toxostoma curvirostre	147	46	100	79	487	148	8.5×10^{-33}
Toxostoma locontei	39	14	95	115	451	55	1.2×10^{-12}
Toxostoma rufum	1611	741	95	221	345	1026	1.7×10^{-222}
Vireo bellii	242	145	95	116	450	154	4.5×10^{-34}
Vireo flavifrons	944	535	72	220	346	705	7.3×10^{-154}
Vireo gilvus	1680	627	99	257	308	468	2.3×10^{-102}
Vireo griseus	825	385	41	131	435	113	2.7×10^{-245}
Vireo huttoni	147	33	89	88	478	131	3.0×10^{-29}
Vireo olivaceus	1884	743	100	222	343	581	5.9×10^{-127}
Vireo philadelphicus	266	12	76	128	438	41	1.2×10^{-9}
Vireo solitarius	1135	284	100	200	366	310	4.6×10^{-68}
Vireo vicinior	37	17	96	91	475	66	5.3×10^{-15}

training and test data sets. In this sense, although all models were highly statistically significant, building truly predictive models may often require 100 or more occurrence points in this particular geographic scenario and with these particular ecological coverages.

DISCUSSION

GARP modeling approaches were able to predict species' occurrences at high levels of statistical significance in every species tested. This result parallels those obtained for 25 species of tropical birds in Mexico (Peterson et al., in press a), and suggests the generality of this tool. Comparisons with other algorithms are under development, but GARP appears to outperform each quite consistently (Stockwell and Peterson, unpubl.). For example, a BIOCLIM model of *Toxostoma rufum* distribution omitted more than twice as many of the test points as the GARP model discussed above, and overpredicted severely in the northwestern portion of the species' distribution (Peterson, unpubl.), making for a model that is clearly less predictive.

Development of robust algorithms for predicting species' geographic distributions based on point occurrence data opens doors to many exciting approaches and analyses. Although the present paper focuses on the well-known bird fauna of North America, applications are feasible for any taxon in any region on Earth. On the most basic level, then, given certain requirements as to sample size (Peterson et al. 1998), locality information from the approximately 3 ×

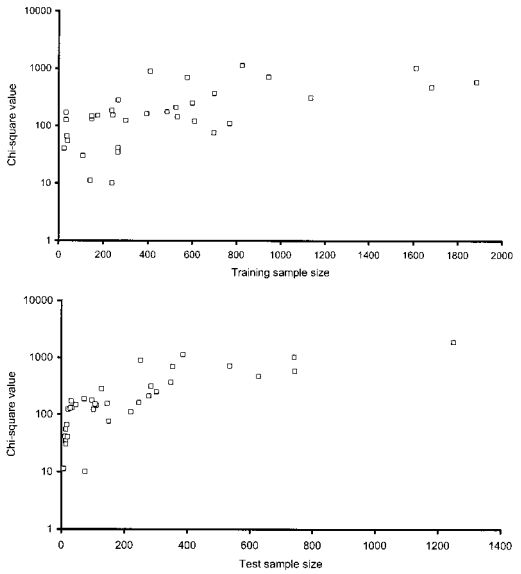

FIGURE 4. Graph of chi-square values used to evaluate statistical significance of predicted geographic distributions versus sample sizes used to (top) train ($r^2 = 0.38$, $n = 34$) and (bottom) test ($r^2 = 0.78$, $n = 34$) the predictive models.

10^9 specimens in the world's natural history museums can be put to use in developing distributional hypotheses for many tens of thousands of species, providing a first view of species' distributions worldwide. Diverse tests have confirmed that GARP is able to build highly predictive models even given the spatial biases inherent in specimen data (Peterson et al. 1999, Peterson et al., in press a, b; Stockwell and Peterson, in press).

Such an understanding has much to offer to workers in organismal biology and species conservation. Species' distributions can be modeled to produce first-pass hypotheses that may be the only usable information for many rare and poorly known taxa. Intensively managed species' potential distributions can identify sites at which reintroduction programs could be focused. Overlaying many species' predicted distributions allows prediction of community composition for any site in the region analyzed, and such cross-species predictions allow identification of conservation priorities (Peterson et al. 2000) or assessment of environmental impacts.

ACKNOWLEDGMENTS

Research for this contribution was assisted greatly by advice and assistance from David R. B. Stockwell, and provision of data by Bruce Peterjohn. Financial support was provided by the National Science Foundation.

LITERATURE CITED

AUSTIN, M. P., A. O. NICHOLLS, AND C. R. MARGULES. 1990. Measurement of the realized quantitative niche: environmental niches of five *Eucalyptus* species. Ecological Monographs 60:161–177.

FIELDING, A. H., AND J. F. BELL. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38–49.

GRINNELL, J. 1917. Field tests of theories concerning distributional control. American Naturalist 51: 115–128.

HUTCHINSON, G. E. 1959. Homage to Santa Rosalia, or why are there so many kinds of animals? American Naturalist 93:145–159.

KROHN, W. B. 1996. Predicted vertebrate distributions from Gap Analysis: considerations in the designs of statewide accuracy assessments, p. 147–169. *In* J. M. Scott, T. H. Tear, and F. W. Davis [EDS.], Gap analysis: a landscape approach to biodiversity planning. American Society for Photogrammetry and Remote Sensing, Bethesda, MD.

MACARTHUR, R. A. 1972. Geographical ecology. Princeton University Press, Princeton, NJ.

NIX, H. A. 1986. A biogeographic analysis of Australian elapid snakes, p. 4–15. *In* Bureau of Flora and Fauna [ED.], Atlas of Australian elapid snakes. Bureau Flora Fauna, Canberra, Australia.

PETERSON, A. T., L. G. BALL, AND K. C. COHOON. In press a. Predicting distributions of tropical birds. Ibis.

PETERSON, A. T., AND K. C. COHOON. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. Ecological Modelling 117:159–164.

PETERSON, A. T., S. L. EGBERT, V. SÁNCHEZ-CORDERO, AND K. P. PRICE. 2000. Geographic analysis of conservation priorities using distributional modelling and complementarity: endemic birds and mammals in Veracruz, Mexico. Biological Conservation 93:85–94.

PETERSON, A. T., A. G. NAVARRO-SIGÜENZA, AND H. BENÍTEZ-DÍAZ. 1998. The need for continued scientific collecting: a geographic analysis of Mexican bird specimens. Ibis 140:288–294.

PETERSON, A. T., J. SOBERÓN, AND V. SÁNCHEZ-CORDERO. 1999. Conservatism of ecological niches in evolutionary time. Science 285:1265–1267.

PETERSON, A. T., D. R. B. STOCKWELL, AND D. A. KLUZA. In press b. Distributional prediction based on ecological niche modeling of primary occurrence data. *In* J. M. Scott [ED.], Predicting species occurrences: issues of scale and accuracy. Island Press, Washington, DC.

SCOTT, J. M., T. H. TEAR, AND F. W. DAVIS [EDS.] 1996. Gap Analysis: a landscape approach to biodiversity planning. American Society for Photogrammetry and Remote Sensing, Bethesda, MD.

SOBERÓN, J. 1999. Linking biodiversity information sources. Trends in Ecology and Evolution 14:291.

STOCKWELL, D. R. B., AND I. R. NOBLE. 1992. Induc-

tion of sets of rules from animal distribution data: a robust and informative method of data analysis. Mathematics in Computers and Simulation 33: 385–390.

STOCKWELL, D. R. B., AND D. PETERS. 1999. The GARP modeling system: problems and solutions to automated spatial prediction. International Journal of Geographic Information Science 13:143–158.

STOCKWELL, D. R. B., AND A. T. PETERSON. In press. Comparison of methods of mapping biodiversity using museum data. *In* J. M. Scott [ED.], Predicting species occurrences: issues of scale and accuracy. Island Press, Washington, DC.

VIEGLAIS, D. A. [ONLINE]. 1999. The Species Analyst <http://habanero.nhm.ukans.edu/TSA/> (12 February 2001).