Review TRENDS in Ecology and Evolution Vol.19 No.9 September 2004

Full text provided by www.sciencedirect.com

New developments in museum-based informatics and applications in biodiversity analysis

Catherine H. Graham[1,2], Simon Ferrier[3], Falk Huettman[4], Craig Moritz[2] and A. Townsend Peterson[5]

[1]Department of Ecology and Evolution, 650 Life Sciences Building, Stony Brook University, NY 11794, USA
[2]Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building #3160, UC-Berkeley, CA 94720, USA
[3]New South Wales Department of Environment and Conservation, PO Box 402, Armidale, NSW, Australia
[4]University of Alaska-Fairbanks, Institute of Arctic Biology, Biology and Wildlife Department, 419 IRVING 1, Fairbanks, AK 99775, USA
[5]Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, KS 66045, USA

Information from natural history collections (NHCs) about the diversity, taxonomy and historical distributions of species worldwide is becoming increasingly available over the Internet. In light of this relatively new and rapidly increasing resource, we critically review its utility and limitations for addressing a diverse array of applications. When integrated with spatial environmental data, NHC data can be used to study a broad range of topics, from aspects of ecological and evolutionary theory, to applications in conservation, agriculture and human health. There are challenges inherent to using NHC data, such as taxonomic inaccuracies and biases in the spatial coverage of data, which require consideration. Promising research frontiers include the integration of NHC data with information from comparative genomics and phylogenetics, and stronger connections between the environmental analysis of NHC data and experimental and field-based tests of hypotheses.

Our sparse knowledge of the number and distributions of species limits both our understanding of ecological and evolutionary processes and our ability to use this knowledge to inform conservation planning. The availability of observational data on species, and the scope and resolution of spatially explicit environmental data, are increasing, as are the capacities of the computing and analytical tools to make use of this information. These developments enhance the relevance of natural history collections (NHCs; primarily public or university-associated museums and herbaria) because NHC collection data are increasingly being used in a variety of applied and theoretical applications [1–3]. Along with increased relevance comes responsibility; curators of NHCs have a public duty to make their data available to the scientific and broader community. Although just one of several sources of information about the distributions of a species, NHCs have a unique combination of attributes, including:

- Massive information resources on the order of 2.5 billion specimens [4], each with an associated ‘collecting event’ describing the time and place where the specimen was collected.
- Records of species backed by preserved specimens (vouchers) that enable the verification of identification and updating of species identity as nomenclature is revised.
- Records or field notes that add considerable value to the specimens themselves, and contain information that can help to address inherent limitations in the modeling of NHC data.
- Historical distributions of organisms in both the recent and distant (paleontological) past, which provide a platform for the assessment of biodiversity dynamics with and without anthropogenic influence.
- Taxonomically current data, because curators have primary expertise in taxonomy and, increasingly, biodiversity science research that draws on phylogenetics, environmental analysis and comparative genomics.

Here, we aim to increase awareness of the current effort and future promise of NHC informatics, and discuss some of the limitations of the data as they apply to documenting and predicting spatial patterns of biological diversity (Box 1). This point is particularly relevant given the increase in the number of studies of patterns of species richness and the processes that cause them, and the crucial need of conservation scientists for such information.

Development of NHC-based biodiversity informatics

The development of electronic catalogues of NHC specimen holdings began during the 1970s, immediately providing new capabilities for curating collections. Computerization of collections has proceeded slowly, but now includes 5–10% of specimens in natural history museums worldwide [4]. Computerization did not fundamentally change access to biodiversity information until the advent of web-accessible data bases. More recently, several

Corresponding author: Catherine H. Graham (cgraham@life.bio.sunysb.edu).
Available online 24 July 2004

www.sciencedirect.com 0169-5347/$ - see front matter © 2004 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2004.07.006

Box 1. Strategies and methods for ecological niche modeling

The ecological niche of a species can be defined as the set of conditions and resources necessary for an organism to maintain a viable population [57]. By integrating known occurrences of species with environmental GIS data layers that summarize meaningful niche dimensions, it is possible to determine the key suites of environmental conditions for that species (and, therefore, its approximate niche). Statistical models are used to develop relationships between environmental values and species presence (and absence, in some cases). This relationship can then be mapped spatially to predict potential geographical distributions [58,59].

Modeling methods

The methods for distributional modeling (reviewed in [60]) vary in their applicability to natural history collections (NHC) data, and should be selected based on the nature of the question and the data [61,62], as well as statistical issues [60]. Some modeling methods (e.g. BIOCLIM, [63] and DOMAIN, [64]) require only the records of species presence; others incorporate multiple predictive approaches with varying requirements (e.g. GARP, [65]); whereas others require both the presence and absence data (e.g. general linear and additive models, and decision trees; [66]). Bayesian approaches are currently being revived because they formally and explicitly combine estimates of sampling bias, 'expert opinion', or other previous information, with observations to build posterior distributions for predictor variables and, in turn, predictions of geographical range [67].

Model evaluation

Substantial progress has been made recently with methods to evaluate models [60,61], although, again, researchers should consider the nature of the question as a guide for the choice of methodologies [61,62,68]. Some assessment has been made of the effects of sample size on the performance of NHC-based models [69]. Few studies have evaluated the predictive performance of models based on NHC data using independent, high-quality presence–absence data sets [70] and little attention has been given to the effects of data error and bias on performance (Box 3). Filling this gap is necessary to inform users of NHC data about which of the many approaches to modeling and evaluation should be applied in relation to specific questions and data sets or, perhaps, when such data should not be used for predictive modeling.

initiatives have used innovative information technology to connect multiple collections. Table 1 summarizes some examples of data bases with global coverage of a broad variety of taxonomic groups (e.g. Global Biodiversity Information Facility), regional coverage of a broad variety of taxonomic groups (e.g. European Natural History Specimen Information Network) and global coverage of specific taxonomic groups (e.g. Mammal Networked Information System; MANIS). In these initiatives, data are retained at the primary institution and connected with distributed network technology over the Internet, rather than being combined with data from multiple institutions into a single centralized data base, where they become isolated and out-of-date. The user can then conduct a single query that queries all participating institutions simultaneously (Box 2). Currently, 20–40% (~60 million data records) of existing computerized specimen information is included in distributed networks.

Once basic data records are computerized, significant steps remain, including georeferencing, standardizing,

Table 1. Examples of websites that provide data on species distributions

Full name	Key information[a]	Taxonomic coverage	Geographical coverage	Web address
Global Biodiversity Information Facility (GBIF)	64 organizations and/or collections from various countries	Broad	Global	http://www.gbif.org/
The World Information Network on Biodiversity (REMIB)	25 collections from various countries	Broad	146 Countries	http://www.conabio.gob.mx/remib_ingles/doctos/remib_ing.html
European Natural History Specimen Information Network	Seven partners from Europe	Broad	Europe	http://www.nhm.ac.uk/science/rco/enhsin/
Australian Biodiversity Information Facility (ABIF)	Three data providers representing various NHCs	Broad	Australia	http://www.deh.gov.au/biodiversity/digir/index.html
The Biota of Canada Information Network (BCIF)	40 Canadian NHC collections	Broad	Canada	http://www.durable.gc.ca/group/biota/index_e.phtml
Distributed Information Network for Biological Collections (SpeciesLink)	12 NHC in São Paulo state, Brazil	Broad	Brazil	http://splink.cria.org.br/index?&setlang=en
Instituto Nacional de Biodiversidad (INBio/Atta)	One institute	Broad	Costa Rica	http://atta.inbio.ac.cr/attaing/atta03.html
Mammal Networked Information System (MANIS)	17 NHC from North America	Mammals	World	http://dlp.cs.berkeley.edu/manis/
Fishnet	24 North American NHC of fish	Fish	North America	http://habanero.nhm.ku.edu/fishnet/
HerpNET (HerpNet)	37 NHC collections	Broad	Global	http://herpnet.org/
Missouri Botanical Garden (Tropicos)	One institute	Plants	Global	http://mobot.mobot.org/W3T/Search/vast.html

[a] Abbreviation: NHC, natural history collections.

error detection and cleaning, and enriching the data served in distributed networks. Addressing these challenges in the context of a distributed network has several advantages. For example, participants in MANIS are cooperatively georeferencing records (i.e. assigning latitude and longitude), including an assessment of location uncertainty with an online error calculator [5]. Standardization of taxonomic information (i.e. consistent naming of specimens) represents another challenge for a distributed network. Taxonomic work should remain a priority, especially in non-vertebrates, because the utility of data-basing collections rests on the accuracy of the identifications. Efforts are focused on correcting specimen data so that they correspond to global taxonomies, which continue to be developed and improved as more taxonomic research becomes available (Box 2).

Specimen data can be categorized broadly into three dimensions: (i) identity (with associated ancillary data regarding the natural history, biology, phenotype and genotype of individual organisms); (ii) space, and (iii) time (Box 3). Identity and space are the most error prone [6,7]. Error-detection modules designed to detect incorrect species identifications, mistaken georeferences and other data problems are being developed to detect and flag data records that require inspection, assessment and, perhaps, correction [5]. These detection modules are more effective when applied to large data sets because outliers are more easily detected. Such efforts have the potential to improve not only the data resource available to all, but also the quality and information content of the specimens

Box 2. Challenges and promises of distributed networks

Some of the major challenges of creating a distributed network include: developing data common schematas that establish which data fields are served; addressing taxonomic incongruence across collections; and connecting data nodes from a diverse array of natural history collections (NHC) via the Internet. There is a commonality in the content of collection and observation data bases that has been exploited by several schemes, including the Darwin Core (http://speciesanalyst.net/docs/dwc/) and Task Group on Access to Biological Collection Data (ABCD; http://www.bgbm.org/TDWG/CODATA/Schema/default.htm) to perform ordered search and retrieval from diverse data sets. Taxonomic information in distributed networks is being standardized by several initiatives, including the Integrated Taxonomic Information System (http://www.itis.usda.gov/), Species2000 (http://www.sp2000.org/) and the Electronic Catalogue of Names of Known Organisms (ECAT; http://www.gbif.org/prog/ecat/prog). These initiatives represent collaborative endeavors among data providers and taxonomists.

Open-source software, developed with the Distributed Generic Information Retrieval (DiGIR; http://digir.sourceforge.net/) in the case of the Darwin Core and the BioCASE data transmitter protocol (http://www.biocase.org/) in the case of ABCD schema, is used to serve the data over the Internet. These client-server protocols provide a single source of access for retrieving data from a series of data sources. To establish the distributed network, each participating NHC, or other data provider, agrees to develop a computerized node for sharing data. Advantages of a distributed network system over a centralized system include: participant control over which data are served and what restrictions are placed on these data; continual updating of data by the providers; ready tracking to the actual specimen documentation upon which data are based; commitment from participants to long-term data management; and the ability to add new data providers easily.

Box 3. Caveats and limitations for the use of NHC specimen data

There are three major issues surrounding the utility of natural history collections (NHC) data for spatial modeling: (i) error, including error in taxonomic identification and spatial error; (ii) bias, primarily the geographical and environmental biases associated with *ad hoc* data collection; and (iii) presence only versus presence–absence data, which influences the type of modeling algorithm that can be used.

Error

The identification of species can be: correct (no error), incorrect (misidentification), correct but based on incomplete knowledge (cryptic species), or correct but based on outdated knowledge (synonyms). Identification errors can be detected based on conflicting name usage across collections, or distribution records that are suspect because they exist in different geographical or environmental space than the rest of the records of a given species. To avoid such errors, NHC data should be used in the context of a thorough knowledge of the taxonomic and systematic history of the group under study [71], in many cases requiring physical examination of the specimens themselves. Spatial error includes georeferencing error, imprecision of location of a record, and error in the original location of a record. Records with these types of error can often be detected because they represent outliers in geographical or environmental space or because discrepancies exist between the georeferenced location and the collector field notes. Spatial errors can often be corrected by checking the specimens themselves and filed notes, eliminating or down-weighting suspect records, and including precision estimates in georeferencing.

Bias

Biases exist because collectors tend to sample along roads and rivers, and near towns or biological stations [72]. Nonrepresentative sampling in environmental space remains the most difficult source of error to detect and correct [73]. To detect bias, records can be mapped in geographical and/or environmental space to determine which regions or environmental contexts have been poorly sampled. If bias is detected, it can be incorporated into inference and records can be subsampled to reduce the bias. Furthermore, new surveys can be undertaken in underrepresented areas to also reduce bias.

Presence versus absence data

In NHC data, 'presence' indicates that a species was present at a given locality at the time of its collection. Limitations associated with these data include species that might no longer be present at a historical collection site, or presence locations might represent a demographic sink for the species. 'Absence' indicates that a species was not found at time of collection. Absence might indicate that: a particular species was truly absent at a site; there was a lack of collecting effort; or there was a failure to detect the species. In general, NHC data should be regarded as 'presence only', unless collecting intent and effort can be reconstructed from specimen tags or collectors' field notes. For modeling techniques requiring absence data for species, surrogate 'pseudo-absence' points can be created using several approaches:

- Sampling of locations from which collections have been made, but the species is not recorded (with reference to field notes);
- Sampling of habitat types or regions judged not to include the species in question;
- Sampling across the region, but excluding sites with presence records.

Although possibly including some undetected true presences, pseudo-absence points can serve to increase the range and statistical power of applicable methods [60].

themselves, a real and tangible benefit to participating NHC data providers.

Applications of the spatial analysis of web-based NHC data

There are increasing and diverse applications of NHC data, but all require assessment of data accuracy. Moreover, the researcher must decide if the NHC data at hand are sufficient to conduct the study of interest or if it is necessary to conduct new appropriately designed surveys. In the case of conducting new surveys, analysis of existing information can provide distributional information that can facilitate new surveys [8]. That said, there are geographical regions and taxa where NHC data represent the only data available, with further surveys being impractical because of loss of habitat or logistic and political constraints.

Ecological and evolutionary processes

Research integrating distribution records with phylogenetic hypotheses [9,10] and ecological data [11,12] has provided new insights into factors influencing patterns of the geography and evolution of species. This integration enhances our ability to ask fundamental questions, including: (i) how do current abiotic and biotic factors, together with biogeographical history, influence geographical limits of species and how do contributions of these factors vary with spatial scale [13,14]? and (ii) can patterns of phenotypic or genetic variation be better understood by incorporating information about the environmental niche of the species and the geographical extent of this niche [15,16]? Niche models, especially when combined with physiological and ecological data, can be used to address range limits by testing whether hypothetical ranges predicted using abiotic variables either coincide with the observed range limits or extend beyond the known range [17]. Coincidence with observed limits might indicate that broad-scale range limits are set by abiotic predictor variables (e.g. temperature, rainfall, seasonality and soils) used (or other correlated factors), whereas extension beyond the known range suggests either undersampling of true range, or range limitations resulting from other factors, such as competition (e.g. [11,18]), local extinction (e.g. [19]), or barriers to dispersal (e.g. [20]). For example, in a natural experiment, Anderson *et al.* [11] used ecological niche modeling to demonstrate that the pocket mouse *Heteromys australis* was competitively excluding its sister species *H. anomalus* in areas of predicted sympatry. When *H. australis* was not present (probably for historical reasons), the distribution of *H. anomalus* extended to the limit predicted by the distribution model.

An approach to the second question is to combine environmental analyses of NHC distribution records with phylogenies to test whether either speciation or genetic variation is associated with, or independent of, divergence in the ecological niches [10,20,21]. Finally, insights into the effects of biogeographical history on patterns of diversity can come from predictions of the potential species range under various paleoclimates, particularly when such predictions are combined with analyses of molecular phylogeography [21,22]. For example, Hugall *et al.* [22] used distributional models of paleoclimate in the wet tropics of Australia to identify regions that have been climatically stable since the last glacial maximum. Each region of climatic stability corresponded to a discrete phylogeographical lineage of the snail *Gnarosophia bellendenkerensis*, providing a more mechanistic understanding of the observed patterns of molecular diversity.

Conservation assessment and planning

NHC data contribute significantly to conservation efforts that are directed toward species of concern, prediction of the spread of invasive species, multi-species conservation prioritization schemes and predictions of biodiversity consequences of climate change. Niche-based predictions often include areas in which a species is currently not known to occur, these being potential targets for additional surveys [23,24] or candidate sites for reintroduction programs (subject to constraints imposed by biotic interactions). Historical distributional data offer unique opportunities to track distributional changes in relation to threatening processes [2,25,26] and thereby anticipate future impacts.

Using records from the native range, the environmental niche of an invasive species can be assessed and projected spatially to the newly occupied landscape to estimate potential range. The numerous successful applications of niche modeling in invasive species systems [27–29] indicate that species in non-native distributional areas obey the same suites of ‘rules’ in relation to their environments as they do on their native areas. For example, the environmental niche of the aquatic plant *Hydrilla verticillata*, an aggressive invader, was estimated from NHC distribution data in its native range in Southeast Asia and projected to the invaded range within North America, where the model accurately predicted the regions of successful invasion [29].

Increasingly, global conservation efforts have been broadening in focus from the management of individual species to the conservation of entire communities and ecosystems [30,31]. One common approach to addressing limitations of sparse data about the distribution of biodiversity is to use surrogate taxa that have dense information about distributions (e.g. birds or butterflies; [32], but see [33]). Spatial distribution models can be created for several species within a taxonomic group, and the resulting spatial representation of biodiversity can then serve as a basis for determining which areas to protect to optimize the conservation of biodiversity in a region [31,34]. This approach is more effective for identifying conservation areas than are methods using umbrella or flagship species to delineate conservation areas [35]. An alternative strategy is to model emergent properties of biodiversity, such as local richness or spatial turnover in community composition, directly from NHC data [8].

Two major emerging applications using NHC data that inform future persistence of biodiversity in a region are: (i) detecting recent changes in geographical ranges associated with climate change [36,37]; and (ii) using niche modeling in conjunction with global future climate scenarios to anticipate future changes in the distributions of species and patterns of species richness [38–41]. For the second application, available occurrence records are used to define the ecological niche, which is then projected onto predicted surfaces of future climate conditions, resulting in a before-and-after view of the potential geographical range of the species. Such predictions of the responses of species can be used to evaluate the suitability of existing protected areas [42], evaluate potential modifications of existing conservation strategies to incorporate areas that are likely to be important in the future, or place monitoring systems in regions with strong predicted effects. The resulting predictions are sensitive to assumptions about dispersal potential (e.g. [43]) and minor differences among modeling methods under current conditions can be amplified in projections of future responses [44]. Furthermore, experimental evidence suggests that contributions of, and interactions among, current predictive variables can change in different climatic regimes [45]. These limitations and others [45,46] emphasize the need to test the validity of such predictive models using experimental approaches and historical evidence.

Agriculture and human health

Agricultural systems represent complex sets of interacting species, and yet very simple quantities (e.g. crop yields) can be informative response variables. Sánchez-Cordero and Martínez-Meyer used NHC records to model the distributions of 17 rodent species that are known to cause crop damage, and observed a relationship between inferred rodent species richness and estimates of damage for five out of seven crops in the 207 municipalities in the state of Veracruz, Mexico [47]. This pilot exploration suggests that NHC data for rodents, as well as for insect herbivores and pollinators, birds, and other taxa can help predict and plan agricultural strategies to maximize production [48].

Human, livestock, and wildlife diseases have been modeled and forecasted using point locality information about species and remotely sensed data. Disease transmission systems frequently involve multiple species, including vectors and reservoirs, and their properties can be predicted from the distribution and ecology of the component species. For example, West Nile Virus, which is native to Europe, Asia and eastern Africa, recently invaded North America, with songbird populations acting as a principal reservoir and ornithophilous mosquitoes (e.g. *Culex*) as vectors [49]. Its rapid spread south and west across North America is somewhat unexpected based upon mosquito dispersal abilities alone. A recent study [50] used NHC records of mosquitoes to determine patterns of vector suitability across the continent; simulations of spread combining these patterns with patterns of bird migration pointed to long-distance migratory birds as a crucial vector for the disease in North America.

Future improvements and new directions

NHC data provide a valuable resource for ecologists, evolutionary biologist and conservationists and future data improvements and innovative research programs continue to extend the value of NHC data. Further improvements associated with NHC data include: increased participation by NHCs; methods to objectively plan future survey work; incorporation of collectors' field notes; and methodological research on the limits of distributional modeling with NHC data [51]. Perhaps one of the most exciting research directions for the use of NHC data is using them in conjunction with other types of data, including ecological, physiological and genetic data, to continue to address questions relating to the spatial patterns of diversity and some of the mechanisms underlying these patterns.

Given the frequent complementarity of geographical and taxonomic focus among NHCs [52], an immediate benefit of connecting of NHC collections will be to increase the density and geographical coverage of available data. The increased geographical coverage of any given organism, a crucial result of connecting NHC collections, will promote more rational decision making regarding species that span jurisdictional borders and will enhance repatriation of data from NHCs in developed nations to developing nations [7].

Even with full and effective integration of NHC collections, it remains clear that the ranges of relatively few species will be known thoroughly and that most species will be poorly sampled [53]. Therefore, it is vital that ongoing effort is devoted to expanding NHC collections through strategic sampling in previously unsurveyed areas. Museum data can be used to identify gaps in the environmental and geographical coverage of existing collections, thereby providing an objective basis for directing future collection effort [8,23,24]. For example, Raxworthy *et al.* [23] surveyed intersecting areas of overprediction for chameleons in Madagascar (effectively, areas meeting the niche requirements of key genera but not known to hold representatives of those genera) and discovered seven chameleon species that were new to science. A recent multi-species application was built around a new analytical technique: generalized dissimilarity modeling (GDM, [8,31]), which models the spatial pattern in turnover of species in both environmental and geographical space. GDM can be used to predict the probable dissimilarity in species composition between any two localities within a region, using only the geographical and environmental properties of these localities. In turn, such predictions can provide a basis for identifying new survey or collection localities that will best complement those already sampled [8,31], and will maximize the probability of discovering new species in the region of interest. This general strategy can be applied iteratively, using data acquired in new surveys to refine the underlying model of spatial pattern in biodiversity, thereby providing an improved basis for designing any subsequent surveys.

Data-basing of collectors' field notes will contribute significantly to the detection and correction of errors, improve the interpretation of data, add significant new information (habitats, local conditions, etc.) about the time of collection and, together with increased density of specimen data, enhance the use of NHC evidence to detect changes in historical versus current properties of the distributions of species. Finally, given that NHC data represent significant and important sources of information for many regions of the world, a continued emphasis on methodological research is needed to evaluate the influence of the uncertainty and bias, which are inherent to NHC data.

Linking specimens (including the expanding NHC tissue banks) to genomic data is important to overcome taxonomic errors that are common in existing genomic data bases [54], to broaden the scope of genomic data available (e.g. including microsatellite and single nucleotide polymorphism profiles as well as DNA sequences), and to improve the interface between biodiversity analysis and comparative genomics. In a similar vein, the connection between studies of physiology, species interactions and natural history on the one hand, and statistical modeling on the other, is crucial to the proper and productive use of NHC evidence. For example, Thomas *et al.* [41] assumed range shifts of species in response to climate change can be predicted based on deriving the environmental parameters of where a given species currently exists and projecting these onto future climate surfaces. Although this approach can generate hypotheses about how species ranges will shift with climate change, it ignores ecological interactions, dispersal limitation and the plasticity of physiological limits [17,55]. Mechanistic studies investigating these factors are essential if we are to infer factors that limit species distribution (i.e. the relative contributions of biotic and abiotic factors). Ideally, extrapolations to novel situations, such as climate change, should be based on a mechanistic understanding of the processes involved, rather than only on a descriptive understanding of the niche [17,45,56].

Summary

We see a rich promise for novel research and conservation results based on NHC data. Future research should extend the utility of NHC data by further developing methods and statistics to use the data themselves, and for integrating NHC data into other research programs, such as those related to biogeography, ecology and evolution. Such transdisciplinary approaches will continue to provide novel insights into how current and historical environmental, geographical and ecological factors have influenced the distribution of biodiversity and how best to conserve this diversity in the face of rapid anthropogenic change.

Acknowledgements

This work was conducted as part of the 'Testing alternative methodologies for modeling species' ecological niches and predicting geographic distributions' Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #DEB-0072909), the University of California and the Santa Barbara campus. Members include: Thomas Edwards, Jane Elith, Antoine Guisan, Chrissy Howell, Bette Loiselle, William Michener, Miguel Nakamura, Jake Overton, Karen Richarson, Ricardo Scachetti Pereira, Jorge Soberon Mainero, Stephen Williams and Mary Wisz. We also thank Robert Hijmans, Brad Shaffer, Antoine Guisan, Stephen Baines and an anonymous reviewer for providing constructive comments about the article.

References

1 Krishtalka, L. and Humphrey, P.S. (2000) Can natural history museums capture the future? *Bioscience* 50, 611–617

2 Ponder, W.F. *et al*. (2001) Evaluation of museum collection data for the use in biodiversity assessment. *Conserv. Biol.* 15, 648–657

3 Suarez, A.V. and Tsutsui, N.D. (2004) The value of museum collections for research and society. *Bioscience* 54, 66–74

4 Duckworth, W.D. *et al*. (1993) *Preserving Natural Science Collections: Chronicle of our Environmental Heritage*, National Institute for the Conservation of Cultural Property

5 Wieczorek, J. *et al.* The point-radius method for georeferencing point localities and calculating associated uncertainty. *Int. J. Geo. Inf. Sci.* (in press)

6 Chapman, A.D. (1999) Quality control and validation of point-sourced environmental resource data. In *Spatial Occurrence Assessment: Land Information Uncertainty in Natural Resources* (Lowell, K. ed), pp. 409–418, Ann Arbor Press

7 Soberón, J. *et al*. (2002) Issues of quality control in large, mixed-origin entomological databases. In *Towards a Global Biodiversity Information Infrastructure* (Saarenmaa, H. and Nielsen, E.S. eds), pp. 13–22, European Environment Agency

8 Ferrier, S. *et al*. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiv. Conserv.* 11, 2309–2338

9 Barraclough, T.G. *et al*. (1998) Revealing the factors that promote speciation. *Philos. Trans. R. Soc. Lond. Ser. B* 353, 241–249

10 Graham, C.H. *et al.* Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution* (in press)

11 Anderson, R.P. *et al*. (2002) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* 98, 3–16

12 Johnson, N.K. and Cicero, C. (2002) The role of ecologic diversification in sibling speciation of *Empidonax* flycatchers (Tyrannidae): multigene evidence from mtDNA. *Mol. Ecol.* 11, 2065–2081

13 Kirkpatrick, M. and Barton, N.H. (1997) Evolution of species ranges. *Am. Nat.* 150, 1–23

14 Case, T.J. and Taper, M.L. (2000) Interspecific competition, environmental gradients, gene flow, and the coevolution of species' borders. *Am. Nat.* 155, 583–605

15 Schluter, D. (2001) Ecology and the origin of species. *Trends Ecol. Evol.* 16, 372–380

16 Turelli, M. *et al*. (2000) Theory and speciation. *Trends Ecol. Evol.* 16, 330–343

17 Kerney, M. and Porter, W.P. Calculating and mapping the fundamental niche: physiology, climate and the distribution of nocturnal lizards across Australia. *Ecology* (in press)

18 Leathwick, J.R. and Austin, M.P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology* 82, 2560–2573

19 Leathwick, J.R. *et al*. (1998) Environmental correlates of tree alpha-diversity in New Zealand primary forests. *Ecography* 21, 235–246

20 Peterson, A.T. *et al*. (1999) Conservation of ecological niches in evolutionary time. *Science* 285, 1265–1267

21 Rice, A. *et al*. (2002) Ecological niche differentiation in the *Aphelocoma* jays: a phylogenetic perspective. *Biol. J. Linn. Soc.* 80, 369–383

22 Hugall, A. *et al*. (2002) Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail *Gnarosophia bellendenkerensis* (Brazier 1875). *Proc. Natl. Acad. Sci. U. S. A.* 99, 6112–6117

23 Raxworthy, C.J. *et al*. (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature* 426, 837–841

24 Funk, V.A. and Richarson, K.S. (2002) Systematic data in biodiversity studies: use it or lose it. *Syst. Biol.* 51, 303–316

25 Shaffer, P. *et al*. (1998) The role of natural history collections in documenting species declines. *Trends Ecol. Evol.* 13, 27–30

26 Drost, C.A. and Fellers, G.M. (1996) Collapse of a regional frog fauna in Yosemite area of the California Sierra Nevada USA. *Conserv. Biol.* 10, 414–425

27 Scott, J.K. and Panetta, F.D. (1993) Predicting the Australian weed status of southern African plants. *J. Biogeogr.* 20, 87–93

28 Zalba, S.M. *et al*. (2000) Using a habitat model to assess the risk of invasion by an exotic plant. *Biol. Conserv.* 93, 203–208

29 Peterson, A.T. (2003) Predictability of the geography of species' invasions via ecological niche modeling. *Q. Rev. Biol.* 78, 419–433

30 Margules, C.R. and Pressey, R.L. (2000) Systematic conservation planning. *Nature* 405, 243–253

31 Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Syst. Biol.* 51, 331–363

32 Howard, P.C. *et al.* (1998) Complementarity and the use of indicator groups for reserve selection in Uganda. *Nature* 394, 472–475

33 Moritz, C. *et al.* (2001) Biogeographic concordance and efficiency of taxon indicators for establishing conservation priority for a tropical rainforest biota. *Proc. R. Soc. Lond. Ser. B* 268, 1875–1881

34 Peterson, A.T. *et al.* (2000) Geographic analysis of conservation priorities using distributional modeling and complementarity: endemic birds and mammals in Veracruz, Mexico. *Biol. Conserv.* 93, 85–94

35 Williams, P.H. *et al.* (2000) Flagship species, ecological complementarity and conserving the diversity of mammals and birds in sub-Saharan Africa. *Anim. Conserv.* 3, 249–260

36 Parmesan, C. and Yohe, G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature* 421, 37–42

37 Root, T.L. *et al.* (2003) Fingerprints of global warming on wild animals and plants. *Nature* 421, 57–60

38 Busby, J.R. (1988) Potential impacts of climate change on Australia's flora and fauna. In *Greenhouse: Planning for Climate Change* (Pearman, G.I. ed), pp. 387–398, CSIRO

39 Williams, S.E. *et al.* (2003) Climate change in the Australian tropical rainforests: an impending environmental catastrophe. *Proc. R. Soc. Lond. Ser. B* 270, 1887–1892

40 Midgley, G.F. *et al.* (2003) Developing regional and species-level assessments of climate change impacts on biodiversity in the Cape Floristic Region. *Biol. Conserv.* 112, 87–97

41 Thomas, C.D. *et al.* (2004) Extinction risk from climate change. *Nature* 427, 145–148

42 Burns, C.E. *et al.* (2003) Global climate change and mammalian species diversity in U.S. national parks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 11474–11477

43 Peterson, A.T. *et al.* (2001) Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecol. Mod.* 144, 21–30

44 Thuiller, W. (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global climate change. *Glob. Change Biol.* 9, 1353–1362

45 Dunne, J.A. *et al.* (2004) Integrating experimental and gradient methods in ecological climate change research. *Ecology* 85, 904–916

46 Schmitz, O.J. *et al.* (2003) Ecosystem responses to global climate change: moving beyond color mapping. *Bioscience* 53, 1199–1205

47 Sánchez-Cordero, V. and Martínez-Meyer, E. (2000) Museum specimen data predict crop damage by tropical rodents. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7074–7077

48 Jones, P.G. and Thornton, P.K. (2003) The potential impacts of climate change on maize production in Africa and Latin America in 2055. *Glob. Environ. Change* 13, 51–59

49 Petersen, L.R. and Roehrig, J.T. (2001) West Nile virus: a reemerging global pathogen. *Emerg. Infect. Dis.* 7, 1–10

50 Peterson, A.T. *et al.* (2003) Migratory birds as critical transport vectors for West Nile Virus in North America. *Vector Borne Zoonotic Dis.* 3, 39–50

51 Soberon, J. and Peterson, A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philos. Trans. R. Soc. Lond. Ser. B* 359, 689–698

52 Navarro-Sigüenza, A.G. *et al.* (2002) A Mexican case study on a centralized database from world natural history museums. *CODATA J.* 1, 45–53

53 Peterson, A.T. *et al.* (1998) Distribution and conservation of birds of northern Central America. *Wilson Bull.* 110, 534–543

54 Ruedes, L.A. *et al.* (2001) The importance of being earnest: what, if anything, constitutes a specimen examined? *Mol. Phylogenet. Evol.* 17, 129–132

55 Davis, A.J. *et al.* (1998) Making mistakes when predicting shifts in species range in response to global warming. *Nature* 391, 783–786

56 Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Mod.* 157, 101–118

57 Grinnell, J. (1924) Geography and evolution. *Ecology* 5, 225–229

58 Austin, M.P. (1985) Continuum concept, ordination methods, and niche theory. *Annu. Rev. Ecol. Syst.* 16, 39–61

59 Peterson, A.T. (2001) Predicting species' geographic distributions based on ecological niche modeling. *Condor* 103, 599–605

60 Guisan, A. and Zimmermann, N.E. (2000) Predictive habitat distributional models in ecology. *Ecol. Mod.* 135, 147–186

61 Fielding, A.H. and Bell, J.F. (1997) A review of methods for assessment of predictive errors in conservation presence/absence models. *Environ. Conserv.* 24, 38–49

62 Loiselle, B.A. *et al.* (2003) Identifying conservation priorities: sensitivities to model selection. *Conserv. Biol.* 17, 1591–1600

63 Nix, H.A. (1986) A biogeographic analysis of Australian Elapid Snakes. In *Atlas of Elapid Saneds of Australia* (Longmore, R., ed.), pp. 5–15, Australian Flora and Fauna Series Number 7. Australian Government of Publishing Service

64 Carpenter, G. *et al.* (1993) DOMAIN: a flexible modelling procedure for mapping potential distribution of plants and animals. *Biodiv. Conserv.* 2, 667–680

65 Stockwell, D.R.B. and Peters, D.B. (1999) The GARP modeling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inf. Syst.* 13, 143–158

66 Huettmann, F. and Diamond, A.W. (2001) Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic. *Ecol. Mod.* 141, 261–298

67 Gelfand, A.E. *et al.* (2003) Explaining species distribution patterns through hierarchical modeling. *Bayesian Anal.* 1, 1–47

68 Parra, J.L. *et al.* (2004) Evaluating alternative datasets for environmental niche models of birds in the Andes. *Ecography* 27, 350–360

69 Stockwell, D.R.B. and Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecol. Mod.* 148, 1–13

70 Ferrier, S. and Watson, G. (1997) *An Evaluation of the Effectiveness of Environmental Surrogates and Modeling Techniques In Predicting the Distribution of Biological Diversity*, Environment Australia (http://www.ea.gov.au/biodiversity/publications/technical/surrogates)

71 Peterson, A.T. and Navarro-Sigüenza, A.G. (1999) Alternate species concepts as bases for determining priority conservation areas. *Conserv. Biol.* 13, 427–431

72 Hijmans, R.J. *et al.* (2000) Assessing the geographic representation of genebank collections: the case of the Bolivian wild potatoes. *Conserv. Biol.* 14, 1755–1765

73 Williams, P.H. *et al.* (2002) Data requirements and data sources for biodiversity priority area selection. *J. Biosci.* 27(Suppl. 2), 327–338