Limitations of Biodiversity Databases: Case Study on Seed-Plant Diversity in Tenerife, Canary Islands

JOAQUÍN HORTAL,*†‡§ JORGE M. LOBO,* AND ALBERTO JIMÉNEZ-VALVERDE*

*Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), C/José Gutiérrez Abascal, 2, Madrid 28006, Spain

†Center for Macroecology, Institute of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen O, Denmark

‡Departamento de Ciências Agrárias, CITA-A, Universidade dos Açores, Campus de Angra, Terra-Chã, Angra do Heroísmo, 9701-851 Terceira (Açores), Portugal

Abstract: *Databases on the distribution of species can be used to describe the geographic patterns of biodiversity. Nevertheless, they have limitations. We studied three of these limitations: (1) inadequacy of raw data to describe richness patterns due to sampling bias, (2) lack of survey effort assessment (and lack of exhaustiveness in compiling data about survey effort), and (3) lack of coverage of the geographic and environmental variations that affect the distribution of organisms. We used a biodiversity database (BIOTA-Canarias) to analyze richness data from a well-known group (seed plants) in an intensively surveyed area (Tenerife Island). Observed richness and survey effort were highly correlated. Species accumulation curves could not be used to determine survey effort because data digitalization was not exhaustive, so we identified well-sampled sites based on observed richness to sampling effort ratios. We also developed a predictive model based on the data from well-sampled sites and analyzed the origin of the geographic errors in the obtained extrapolation by means of a geographically constrained cross-validation. The spatial patterns of seed-plant species richness obtained from BIOTA-Canarias data were incomplete and biased. Therefore, some improvements are needed to use this database (and many others) in biodiversity studies. We propose a protocol that includes controls on data quality, improvements on data digitalization and survey design to improve data quality, and some alternative data analysis strategies that will provide a reliable picture of biodiversity patterns.*

Keywords: biodiversity patterns, biological databases, biodiversity informatics, species richness, sampling effort assessment, predictive modeling, data-quality controls

Limitaciones de las Bases de Datos de Biodiversidad: Estudio de Caso de la Diversidad de Plantas con Semillas en Tenerife, Islas Canarias

Resumen: *Las bases de datos sobre la distribución de especies pueden ser utilizadas para describir los patrones geográficos de la biodiversidad. Sin embargo, tienen sus limitaciones. Estudiamos tres de esas limitaciones: (1) inadecuación de datos crudos para describir los patrones de riqueza debido a sesgos en el muestreo, (2) falta de esfuerzo de muestreo (y falta de exhaustividad en la compilación de datos sobre esfuerzo de muestreo), y (3) falta de cobertura de las variaciones geográficas y ambientales que afectan la distribución de los organismos. Utilizamos una base de datos de biodiversidad (BIOTA-Canarias) para analizar los datos de riqueza de un grupo bien conocido (plantas con semillas) en un área muestreada intensivamente (Isla Tenerife). La riqueza de especies observada y el esfuerzo de muestreo estuvieron altamente correlacionados. Las curvas de acumulación de especies no pudieron ser usadas para determinar el esfuerzo de muestreo porque la digitalización de datos no era exhaustiva, así que identificamos sitios bien muestreados con base en la proporción riqueza – esfuerzo de muestreo. También desarrollamos un modelo predictivo basado en los datos de sitios bien muestreados y analizamos el origen de los errores geográficos en la extrapolación obtenida por*

§*email jhortal@mncn.csic.es*

Paper submitted June 8, 2006; revised manuscript accepted November 28, 2006.

Conservation Biology Volume 21, No. 3, 853–863

DOI: 10.1111/j.1523-1739.2007.00686.x

medio de la validación cruzada constreñida geográficamente. Los patrones espaciales de riqueza de especies de plantas con semillas obtenidos de datos de BIOTA-Canarias fueron incompletos y sesgados. Por lo tanto, se necesitan algunas mejoras para utilizar esta base de datos (y muchas otras) en estudios de biodiversidad. Proponemos un protocolo que incluye controles de la calidad de datos, mejoras en la digitalización de datos y diseño de muestreo para mejorar la calidad de los datos y algunas estrategias alternativas de análisis de datos que proporcionarán una descripción confiable de los patrones de biodiversidad.

Palabras Clave: bases de datos biológicos, controles de calidad de datos, evaluación de esfuerzo, informática de biodiversidad, modelo predictivo, patrones de biodiversidad, riqueza de especies

Introduction

Accurate mapping of the spatial patterns of biodiversity is needed to study the processes by which these patterns vary and to design effective regional conservation schemes. To create accurate maps, information must be gathered on the location of species and recorded in exhaustive biodiversity databases. Extensive databases are a primary tool in ecological research (Porter 2000 and other chapters in Michener & Brunt 2000). In the case of biodiversity research, the biodiversity information networks (BIN) and/or biodiversity databases under development aim to bring together the scattered information available in museum collections and herbaria and the data available in the literature from inventories developed with (or without) standardized surveys (e.g., Soberón et al. 1996). These data provide the basis for biological atlases, a number of studies on biodiversity patterns (Biodiversity Informatics; Soberón & Peterson 2004), and the development of geographically explicit conservation schemes (Systematic Conservation Planning; Margules & Pressey 2000).

The most outstanding BIN initiative is the Global Biodiversity Information Facility (GBIF) (http://www.gbif.org/). The intention of this project is to gather all information (not solely information on distribution) on known species and to make these data freely available on the Internet. Typically, database developers address four questions: why the database is needed, who will be its users, what types of questions should it help answer, and what incentives should be given to data providers (Porter 2000; http://www.gbif.org/). Nevertheless, the biodiversity data currently at hand are scarce, biased, and sometimes of poor quality. These limitations can hinder the usefulness of the databases even if all the data available are gathered exhaustively.

Adequate distribution data for many of the known species and higher taxa are lacking (the Wallacean shortfall; Whittaker et al. 2005) and are prone to taxonomic, temporal, and geographic bias (Stockwell & Peterson 2002; Soberón & Peterson 2004; Kadmon et al. 2004). Although a few databases from exhaustive survey campaigns are available for the British Isles (Prendergast et al. 1993; Griffiths et al. 1999), they are the exception, not the rule. Usually, sampling effort is limited, scattered, and not standardized, and the inventories are biased toward easily accessible sampling sites (e.g., Dennis & Thomas 2000; Kadmon et al. 2004). Gaps and biases in biodiversity data are important enough to compromise the description of biodiversity patterns from the raw information compiled in the available databases (Prendergast et al. 1993; Stockwell & Peterson 2002; Hortal & Lobo 2006).

Three different solutions have been proposed to overcome the lack of spatial and taxonomic exhaustiveness in biodiversity data: (1) use environmental information as a surrogate for biodiversity variations (Faith & Walker 1996), (2) use of predictive modeling of species distributions (e.g., Soberón & Peterson 2004, 2005; Araújo & Guisan 2006), and (3) use of predictive modeling of biodiversity descriptors (e.g., richness, rarity, species turnover) based on the information from well-surveyed areas (e.g., Ferrier 2002; Lobo & Martín-Piera 2002; Ferrier & Guisan 2006). These three approaches present several problems and limitations. We considered the drawbacks associated with the latter. Drawbacks of the other two solutions are discussed elsewhere (Hortal & Lobo 2006). We examined the effect of two of the drawbacks associated with current use of biodiversity databases: lack of survey-effort assessment (and lack of exhaustiveness in compiling data about survey effort), and lack of coverage of the geographic and environmental variations that affect the distribution of organisms. These problems make existing databases and/or atlases less useful for describing patterns of biodiversity accurately (Prendergast et al. 1993; Johnson & Sargeant 2002; Dennis & Shreeve 2003; Soberón et al. 2007) and compromise the utility of the predictive models of biodiversity features (Hortal & Lobo 2006).

We analyzed the quality of the distributional information available at an intensively surveyed territory and predicted the distribution of species richness based on this information. Information came from BIOTA-Canarias (herein, BIOTA), a database that stores the information regarding seed plants of Tenerife (Canary Islands). This database is not exhaustive. We used the raw data in BIOTA to describe the geographic patterns of seed-plant richness in Tenerife and assessed the reliability of the resulting maps, analyzed survey effort to identify well-sampled areas, and produced a predictive model of species richness based on information from these well-sampled sites. We used the errors that resulted in each step as a

basis for consideration of the problems of using incomplete or biased information. These problems may be overcome with a protocol we propose that assesses and improves data quality and guarantees the utility of databases for the description of biodiversity patterns.

Species Richness Patterns from Raw Database Information

The database BIOTA contains data on the presence of all species in the Canary Islands (http://www.gobcan.es/medioambiente/biodiversidad/ceplam/bancodatos/bancodatos.html). The initial aim of this database was to provide the regional government and other stakeholders with information about the presence and absence of species of interest (in territorial units with a resolution [grain] of 500 × 500 m, herein called grid cells) for use in environmental impact assessment and territorial planning processes (A. Machado, personal communication). Although the exploration of the spatial patterns of biodiversity was not among the original objectives of the database, it has been used to obtain updated checklists for a large number of populations in all islands of the Archipelago (Izquierdo et al. 2005). Due to its success, the BIOTA database has been extended to the rest of the Macaronesian archipelagos (e.g., Azores, Borges et al. 2005) with the economic support of the EU (Project ATLANTICO - INTERREG III B 2000–2006).

We analyzed the spatial distribution of seed-plant species richness at Tenerife with data from BIOTA to illustrate the limitations of the data stored in this database. The Canary Islands have been intensively surveyed by European botanists since Linnaeus. In addition to systematic surveys (e.g., Voggenreiter 1974; E. Barquín Díez & V. Voggenreiter. 1987. Prodromus del Atlas Fitocorológico de las Canarias Occidentales, I. Flora autóctona y especies de interés especial. Unpublished report, ICONA, Bonn-La Laguna, Spain; E. Barquín Díez & V. Voggenreiter, 1988. Prodromus del Atlas Fitocorológico de las Cananas Occidentales [Hierro, La Palma, Gomera, Tenerife, Gran Canaria]. Unpublished report, ICONA, Bann-La Laguna, Spain.), the climatic conditions and the extraordinarily rich plant diversity of the archipelago have prompted a number of expeditions by professional and amateur botanists, resulting in a continuous sampling effort throughout the twentieth century. Although some additions are still being made to the island inventory, most seed-plant species on Tenerife are well known: 1131 species (841 [74.4%] native and 318 [28.1%] Canary endemics) were included in Tenerife's database when we extracted the data (29 November 2003).

The information compiled in the database was a geographically and taxonomically exhaustive representation of the inventory (i.e., all the species recorded in each grid cell were gathered from relevant herbaria and literature) and was the result of a huge amount of sampling effort (1,084,971 records on the above-mentioned date; approximately 960 records/species and 128 records/grid cell). Nevertheless, these surveys were spatially biased (Fig. 1a). More critically the observed species richness patterns were highly correlated with the amount of sampling effort (measured as the number of BIOTA records, see below) (Spearman rank correlation coefficient, $r_s = 0.973$, $n = 8{,}469$, $p < 0.0001$). Therefore, the observed distribution of species richness varied in accordance to the spatial bias of the historical survey effort (Fig. 1). Due to this bias, the raw information in BIOTA cannot be used directly to describe variations in seed-plant diversity at Tenerife.

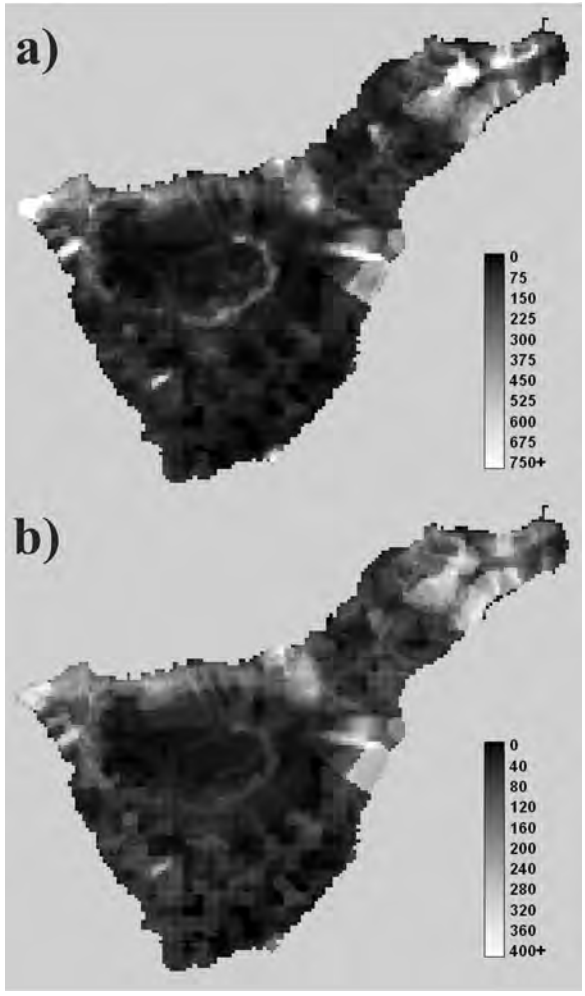

Figure 1. Number of (a) records and (b) observed seed-plant species richness in Tenerife according to the BIOTA-Canarias database (data extracted on 29 November 2003). Map resolution is 500 × 500 m grid cells, and reference system is UTM 29N.

Sampling-Effort Assessment from Nonexhaustive Data Compilation

The first step in analyzing survey completeness was to define a measure of sampling effort that could be applied to all the data gathered in the database, so we examined the

metadata of BIOTA. Because data are generally compiled from heterogeneous sources (i.e., herbarium sheets, surveys using different methodologies), the measure of effort should take into account the detail and exhaustiveness of the information in the database. *Detail* refers to the biological resolution of the information (i.e., whether data refer to single specimens or to a series of individuals), and *exhaustiveness* refers to the proportion of information compiled (i.e., all the information available or only a subset of it).

Because detailed data usually refer to individuals or detections of the species, use of records as a measure of sampling effort has been proposed (Lobo & Martín-Piera 2002). A record refers to each time a species is detected per day, per collector, or per survey method. Records have been used successfully to assess sampling-effort completeness (e.g., Hortal & Lobo 2005), and their performance is similar to other fine-grain measures of sampling effort, such as data on number of individuals or number of traps (Hortal et al. 2006). Therefore, we used the records in BIOTA as a measure of sampling effort, assuming that the more records in a grid cell the higher the survey effort.

Common means of assessing sampling-effort success cannot be used with BIOTA because of its lack of exhaustiveness. This database contains information only on the observation of each species by decade in each grid cell of all terrestrial areas of the archipelago. For example, a single observation in a given grid cell of a species during the 1950–1959 decade provides the same information as 100 observations of the same species during the same decade. Thus, the mean number of records ($\pm$SE) per cell and seed-plant species was 1.51 ± 0.46, with a maximum of five records (i.e., the species was recorded in five different decades). This lack of exhaustiveness impedes the use of species accumulation curves (the most common sampling-effort-assessment method; e.g., Hortal et al. 2004; Hortal & Lobo 2005) to identify well-sampled grid cells. These curves describe the diminishing rate of finding new species as sampling effort increases (Soberón & Llorente 1993). If a number of records of common species (i.e., easily detected) were not included in the species accumulation curves, differences in the recording of common and rare species (i.e., difficult to detect) would not be described properly. Thus, the curves cannot saturate or be described with an asymptotic function, as needed for sampling-effort assessment.

We used an alternative three-step procedure to identify well-sampled grid cells in Tenerife. We developed a geoenvironmental regionalization of the island, plotted the regional relationships between the number of records and the number of species observed per grid cell to identify those cells where the inventories could have saturated, and determined the cells with higher probability of having reliable inventories based on the ratio between number of records and observed species. Here, we assumed that contiguous areas with similar habitat conditions host similar floristic diversity (i.e., sites placed nearby share their regional species pool), presenting similar assembly history, similar assemblages, and similar patterns of inventory saturation. Within-region differences in the richness of these assemblages (i.e., different community types) will result in different patterns of inventory saturation, which could be identified and studied separately (Hortal et al. 2001; Lobo & Martín-Piera 2002).

For other purposes, we (J. H. and J. M. L., unpublished) identified (roughly speaking) spatially contiguous areas with similar environmental conditions in Tenerife. Briefly, information on environmental similarity and spatial distance was used to classify all grid cells in Tenerife in eight regions. The relationships between database records and observed number of species per grid cell in these regions were plotted to identify the different patterns of inventory saturation (Fig. 2). Only two of the eight regions presented a single pattern of saturation (Güímar—Western Slopes and Teno), whereas in the others, two or three different patterns were identified, up to 16 patterns of saturation in all. We studied the grid cells with the most records in each of these patterns (circled in Fig. 2) and identified groups of cells with similar ratios. We assumed that within those groups the cells with higher ratios of records to observed species were well sampled. For example, from the 16 cells examined for the pattern of saturation 1a (the one with higher richness in Acentejo-La Laguna region, Fig. 2), we selected the nine that presented ratios over 2.3 (ratios from 2.39 to 3.88; more than 700 records and more than 300 species observed), leaving out the other seven (ratios from 1.7 to 2.15). We were as conservative as possible in our selection so as to ensure that all selected cells had reliable inventories. Although we risked missing some grid cells with reliable inventories (see below), such a risk was less costly than the use of unreliable data, which would have diminished reliability of the results.

Two hundred seven grid cells (2.44% of all grid cells) were identified as well-enough sampled to present reliable seed-plant inventories in Tenerife according to BIOTA. Nevertheless, these cells were spatially aggregated, clustered in several areas (Fig. 3) that roughly coincided with the localities repeatedly sampled by most botanists over time (because of their high numbers of Canarian endemics or their unique plant communities; "areas of botanical interest" in Bramwell & Bramwell 2001). Thus, the areas identified as well-sampled represent the sites that were surveyed during several decades rather than all grid cells with good inventories. Grid cells intensively sampled in a single decade could not be detected with this method (e.g., sites surveyed during fieldwork for a Ph.D. or during recent [1990s] work in phytosociological mapping, e.g., del Arco et al. 2006). Thus, the lack

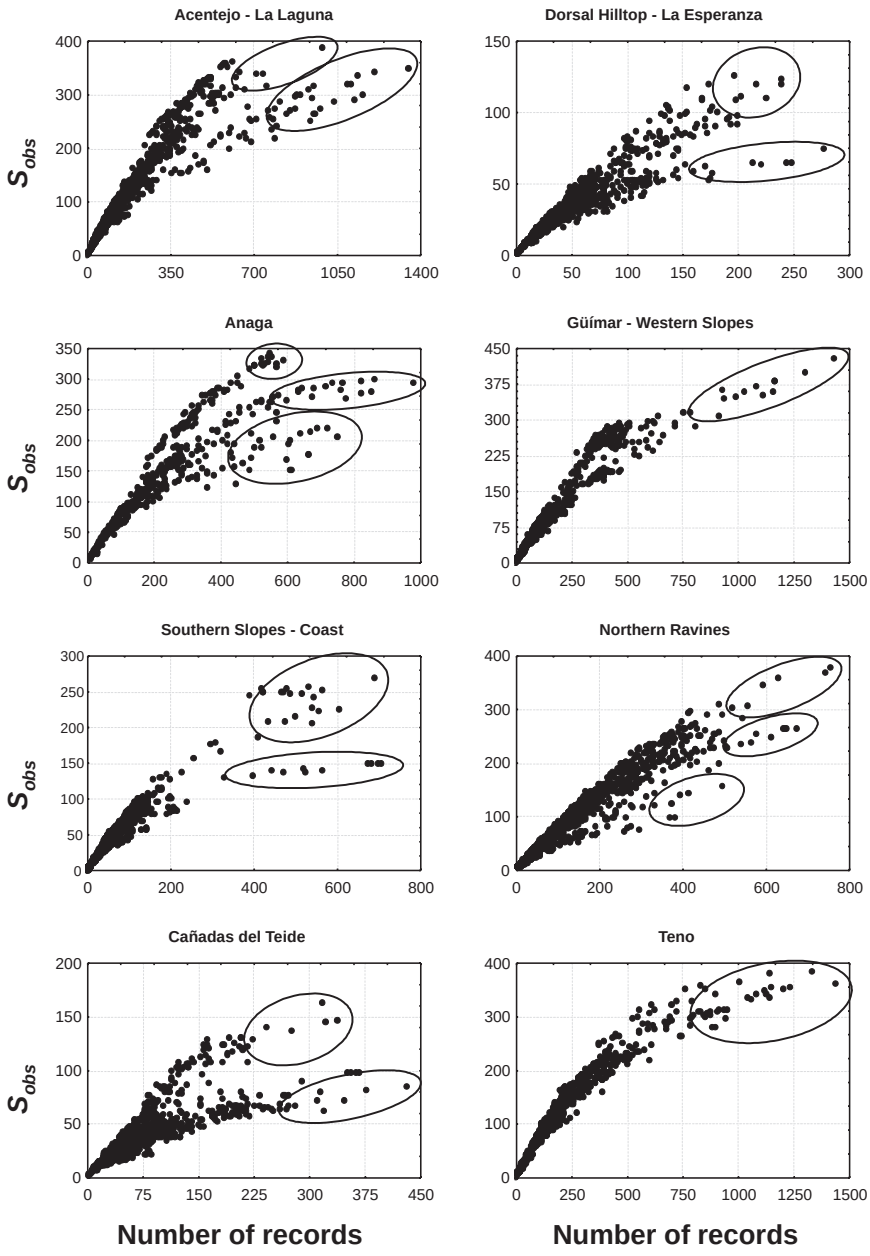

Figure 2. Relationship between the sampling effort recorded in BIOTA-Canarias (number of records) and the observed species richness (S_{obs}) at each grid cell for each of the eight geoenvironmental regions (see text). Sixteen different relationships between sampling effort and observed richness were observed. In each relationship the grid cells with higher number of records were selected (ellipses) as being well sampled.

of exhaustiveness of BIOTA impeded identification of all good-quality inventories that could be included in the information gathered in the database.

Predictive Modeling of Species Richness Based on Geographically Biased Areas

Once a set of well-sampled grid cells has been identified, a possible solution to mitigate the spatial gaps in knowledge is to model and predict the distribution of biodiversity. Predictive modeling involves building the model itself, validating the results of the model, and making improvements to the model. We used the Poisson stepwise general linear model (MacCullagh & Nelder 1989; Crawley 1993) to model species richness as a function of environmental variables (Nicholls 1989; Austin et al. 1996), following the procedure described in Hortal et al. (2001, 2004) and Lobo and Martín-Piera (2002). We validated the model by examining outliers and refining the data accordingly. Finally, we assessed its predictive power by taking into account the spatial structure of the data.

Model Building

We selected a set of environmental predictors, including the linear, quadratic, or cubic equations of these variables (or all the categories in qualitative variables) in the model, in an iterative fashion by means of a mixed forward–backward procedure until no more significant additions could be made (Austin et al. 1996). We used a set of predictors that accounted for the factors that affect

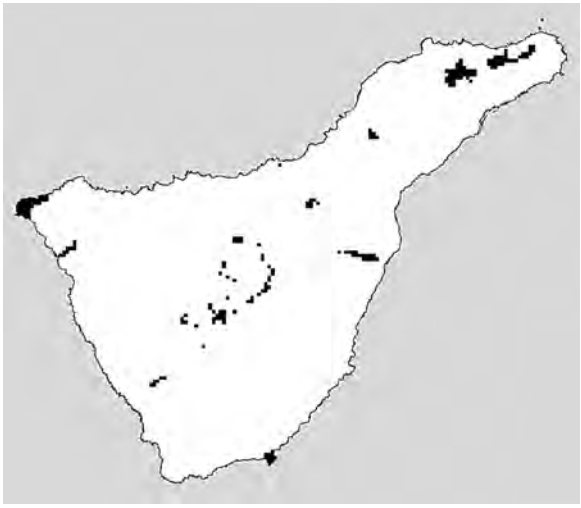

Figure 3. Grid cells (500-m width) with reliable seed-plant inventories at Tenerife according to the information gathered in BIOTA-Canarias.

variations in plant communities: climate (e.g., Fernández-Palacios 1992; Lobo et al. 2001), substrate (e.g. Lobo et al. 2001), elevation (a surrogate for environmental gradients that is related to plant distributions in Tenerife [Fernández-Palacios 1992; Fernández-Palacios & de Nicolás 1995]), and geomorphology and geomorphological diversity, which are usually related to diversity variations (e.g., Wohlgemuth 1998). Thirteen environmental predictors were used in total, 10 continuous (mean annual precipitation; precipitation in the dry and rainy seasons; mean, maximum, and minimum temperature; elevation; slope; aspect diversity; and elevational range) and three categorical (soil type, geology, and aspect [original aspect data were reclassified into nine categories]) variables. Once an environmental model was available, we included the third-degree polynomial of latitude and longitude (trend surface analysis; Legendre & Legendre 1998) in the model, eliminating the nonsignificant terms from the resulting equation. This way, we accounted for any spatial structure remaining unexplained by the model (resulting from historical effects or other unaccounted-for factors).

Validation and Improvement of the Models

We measured the explanatory capacity of the models based on changes in deviance and assessed significance with the *F* statistic (see McCullagh & Nelder 1989). Nevertheless, these measures only assessed how well the models fit the data we used, not the true reliability of model predictions. A measure of their true reliability should account for the accuracy of predicted scores when the model is extrapolated to an unknown territory. Therefore, model extrapolations have to be validated with independent data (i.e., not used to calibrate the model). If independent data are lacking, one should use a cross-validation procedure in which data are split n times into two sets, one used to calibrate the model and other to validate its results (Fielding & Bell 1997; Fielding 2002). In our case we (1) subsampled data eight times, extracting a mean of 27 squares each (12.5%); (2) adjusted model parameters based on data from the rest of the grid cells; (3) extrapolated the models to the cells extracted in step 1; and (4) calculated prediction errors at each grid cell as the scaled difference between observed and predicted scores (prediction error = [observed richness − predicted richness] × 100 / observed richness).

The overall predictive performance of the model was the mean of all prediction errors obtained in step 4, and the predictive power of the models was the inverse of this mean (100 − mean prediction error; Hortal et al. 2001). A good predictive model will provide reliable extrapolations in most cases. Therefore, to decide if model extrapolations are good enough, it is necessary to set up cutoffs for both mean predictive power and the variability in prediction errors. We considered that if a large number of prediction errors were higher than 30%, model extrapolations would not be reliable. Thus, we considered good models those with 85% or higher predictive power (15% mean prediction error) and with <15 SD in prediction errors (i.e., most errors within 0 and 30% in the case of a 15% mean prediction error). These cutoffs were necessarily arbitrary because they referred to the error in model extrapolations that we considered acceptable.

Given the clumped distribution of well-sampled areas, we used a geographically constrained procedure for data splitting in step 1. Using random selection of sites to folds for cross-validation in our data could give an erroneous measure of the ability of the model to extrapolate species richness scores because its performance would be tested against squares placed near the cells used to adjust it. Therefore, we included a spatial restriction in the selection of the extracted cells in step 1 to assess the accuracy of model predictions outside the bounds of the data used (Boyce et al. 2002). Well-sampled cells placed nearby were extracted each time to extract all data from spatially contiguous areas, so model predictions were tested in areas where no data points were used to adjust model parameters.

Following the modeling protocol described above, we developed an environmental model in which soil type, elevation, aspect, and the quadratic equation of minimum temperature accounted for 84.3% of species richness variability in the data (Table 1). The inclusion of the spatial terms resulted in a model that explained 89.1% of data variability, including soil type, elevation, the quadratic equation of minimum temperature, and four spatial terms (lat, lat^2, long, and lat × long). Despite their high explained variability, the predictions from both models were highly unreliable (Table 1). All prediction errors from the former were below 55%, although the mean predictive

Table 1. Goodness of fit and reliability of the general linear models of seed-plant species richness.

Models	*df*	*Dev*[a]	ΔDev[b]	F[c]	*% Dev*[d]	*PP*[e]
Null	206	2,396,312				
environmental	190	374,403.3	2,021,908.9	1026.1	84.38	72.68 ± 26.14
environmental + spatial	192	262,338.1	2,133,974.1	1561.8	89.05	43.81 ± 130.86
Without outliers	180	2,148,301				
environmental	161	170,947.6	1,977,353.3	1862.3	92.04	76.27 ± 27.78
environmental + spatial	157	141,585.4	2,006,715.5	2225.2	93.41	48.72 ± 48.44

[a] *Deviance of each model.*
[b] *Variation of deviance with respect to the null model (quoted above the statistics of each kind of model).*
[c] *Fischer's* F *score.*
[d] *Proportion of variability in the data explained by each model.*
[e] *Predictive power, calculated from the results of the cross-validation subsampling of the data (see text).*

power was 72.7%. Therefore, predictions were relatively good in some areas but highly biased in others. This pattern was more striking in the latter model (< 45% of mean predictive power and a high variability) (Table 1).

We investigated model residuals to find outliers and discarded them to improve model performance. We assessed their residual value and potential leverage (i.e., the importance of the observation in the adjustment of model parameters) to determine their importance in the final configuration of the model (Nicholls 1989). We eliminated those outliers that were likely to correspond to erroneous data or be irrelevant for model parameters according to their potential leverage (Hortal et al. 2001). Deleting these outliers could contribute to overfit training data (and thus diminish the predictive power of the models). Nevertheless, some erroneous richness scores could be present in the data due to inaccuracies caused by uneven sampling effort, given the impossibility of developing a proper assessment of sampling effort. Therefore, we eliminated dubious (and potentially inaccurate) data (see a discussion on the deletion of outliers due to differences in sampling effort in Lobo & Martín-Piera 2002). Twenty-six outliers were identified and eliminated from model calculations. The models developed with the corrected data set were more explanatory (92% and 93.4%; Table 1). They predicted higher species richness scores along the northern and, more importantly, eastern coasts and decreasing scores toward the higher elevations of the Teide (Fig. 4a). Despite these slight increases in model reliability, the inaccuracy in model predictions was also quite high (Table 1). Important numbers of prediction errors per case where higher than 50% (76 cases, 42% of all cases) or 75% (51 cases, 28% of all).

The inaccuracy of model predictions was also spatially biased (Fig. 4b). One key issue when assessing the reliability of models is to determine whether they can explain the spatial structure in data and whether some unexplained spatial structure remains in the errors of the model (Diniz-Filho et al. 2003). The final model underpredicted species richness especially in the southern and western coastal areas of, for example, Teno or Montaña Roja, and overpredicted species richness especially in some areas near the Teide and in some points near the northern coast (Fig. 4b). Interestingly, predicted richness scores were unrealistically high in some coastal areas of the Anaga Peninsula (northeast of the island) (>500 species, whereas the highest observed richness was 428). Spatial bias in the results of the model with only environmental variables was similar, although the magnitude of the bias was, in general, less dramatic (not shown). The higher magnitude of the biases in the spatioenvironmental model could be attributable to "geographical overfitting" (i.e., a tight fit to the scores present in the areas covered by the training data that diminish the predictive power of the model).

Discussion

The spatial representation of species richness obtained with the predictive models we developed was spatially biased. It is extremely unlikely that the representations reflect accurately the spatial patterns of richness variation outside the bounds of the grid cells used to fit them (i.e., the well-sampled cells). An incorrect model structure or the lack of some important explanatory variables may have affected the occurrence of biased predictions. Although some overdispersion existed in the data, varying the overdispersion parameter did not affect model building or model parameters (McCullagh & Neldner 1989; Crawley 1993) and changed only slightly the explained variability of the models (<0.5%). The predictors used are also unlikely to cause poor model performance; all known determinants of plant species richness (including those previously related to plant distribution in Tenerife) were included in the modeling process, and the variables selected during the cross-validation were quite stable.

On the contrary, spatial biases in data are a plausible explanation for the poor predictive performance of our models. Spatially biased data result in spatially biased models and predictions. Provided that Tenerife is a highly

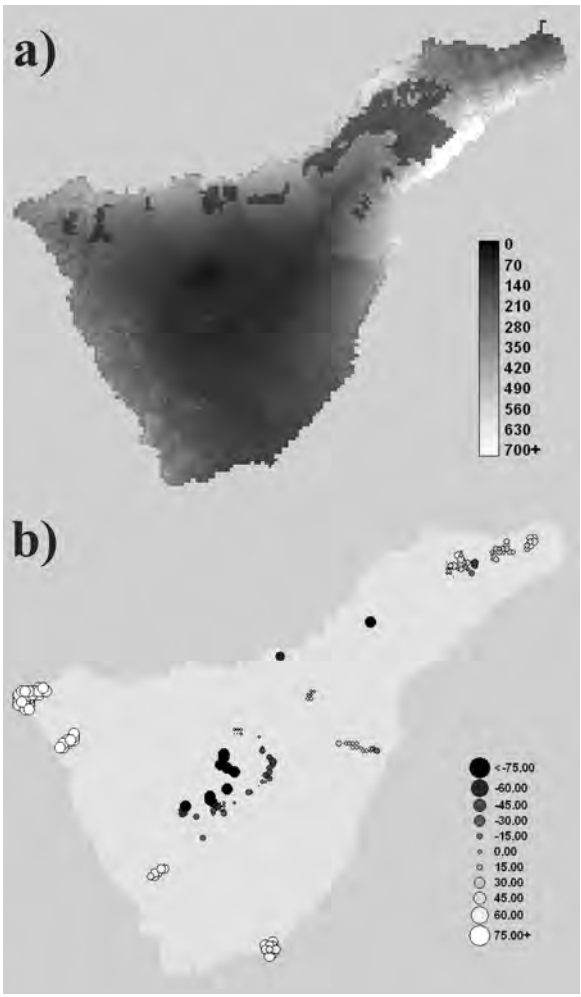

Figure 4. Results of the predictive model of the spatial distribution of seed-plant species richness in Tenerife (including native and introduced species) from the grid cells identified as well sampled after the elimination of outliers (Table 1): (a) species richness predicted from the model including environmental and spatial variables and (b) prediction errors (%) in each data cell based on the results of the spatially explicit cross-validation (see text) (circle size, magnitude of the error; negative scores [black to gray], predicted scores higher than the observed richness; positive scores [gray to white], predicted richness figures lower than the observed). Map resolution is 500 × 500 m grid cells, and reference system is UTM 29N.

heterogeneous island (e.g., del Arco et al. 2006), the lack of survey coverage (apparent in Fig. 3) resulted in predictive errors that we were unable to detect in absence of additional complete inventories coming from less-surveyed parts of the island. This poor representation of plant diversity is not the case in other published examples (e.g., Wohlgemuth 1998), but we are aware that it might be the rule rather than the exception for most regions and living groups worldwide, especially for fine-grain data.

Enhancing the Utility of Biological Databases

Conservation assessment processes and biodiversity research need good-quality data to provide reliable (and long-lasting) conservation strategies and scientific information on biodiversity pattern and process. Nevertheless, most times these data are lacking, and it is necessary to

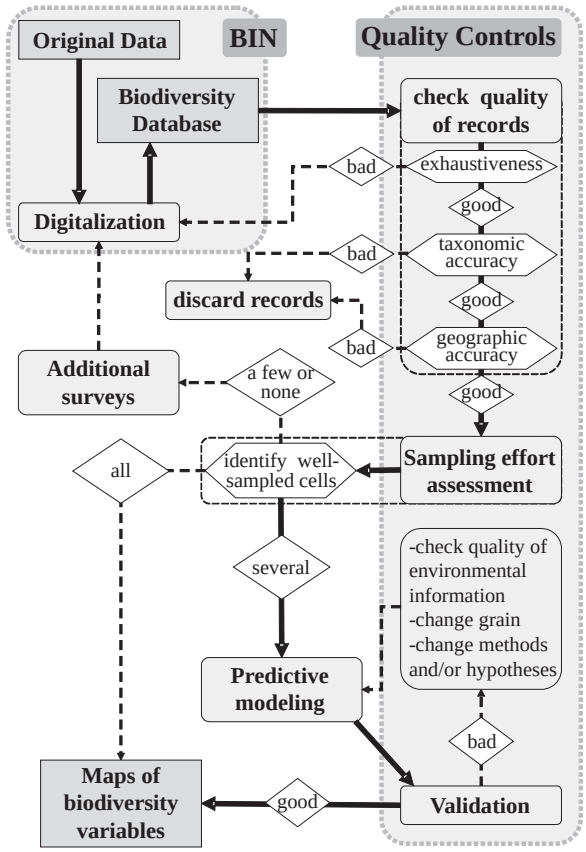

Figure 5. Schematized protocol to obtain reliable maps of biodiversity variables from biodiversity databases (dark gray rectangles, physical objects; light gray rectangles with round corners, processes; white hexagons, specific parts of processes that lead to two or more options [represented by rhombuses]; dashed light gray rectangles, domain of operation of biodiversity information networks (BIN) such as GBIF (see text) and the quality-control processes needed to ensure the reliability of the final product. Thick, continuous-line arrows represent the process when all quality controls are satisfactory, and dashed-line arrows represent secondary processes occurring when the result of some process or quality control is not satisfactory (or very rare, such as all cells being well sampled).

obtain surrogates for the distribution of biodiversity. We summarized the protocol needed to assess data quality and obtain reliable information on the geographic distribution of biodiversity features (Fig. 5).

The process of gathering the information into biodiversity databases or BIN (Fig. 5) cannot be understood as a process limited to digitizing the original data. Some quality controls are needed to assess the reliability and utility of records (Fig. 5), namely assessing their taxonomic accuracy (e.g., Valdecasas & Camacho 2003; Dillon & Fjeldsa 2005), the accuracy of their geographic allocation (e.g., Chefaoui et al. 2005), and the exhaustiveness of records stored in the database. Regarding the latter, some BIN initiatives of academic origin are as exhaustive as possible (e.g., GBIF). Nevertheless, other biodiversity databases contain limited information, recording only the presence of each species in a cell (i.e., biological atlases) or storing nonexhaustive compilations of all information (this work). Our results show that such a lack of exhaustiveness complicates the assessment of survey completeness. Therefore, we recommend checking whether all the available information has been included in the database. If there is some information remaining to be entered, it is worth it to continue its digitalization until the database is complete and sampling effort can be analyzed with reliability.

Once the quality of all records to be used has been ensured, an adequate assessment of sampling effort is the second quality control (Fig. 5). The completeness of regional or local inventories can be assessed reliably with species accumulation curves if the information gathered is exhaustive (Hortal & Lobo 2005). This analysis can determine the likelihood of the absence of species from some places and identify the places where measures of other biodiversity features (e.g., species richness, species composition) can be obtained accurately (i.e., from complete inventories). In addition, exhaustive data can be used to estimate the scores of these variables at each territorial unit, rather than using raw, observed data (Hortal et al. 2004).

Once well-sampled cells (or areas) are identified, their coverage in the studied territory can be assessed to determine whether it is necessary to carry out additional surveys (Fig. 5). Unavoidably, survey data from a given region constitute only a group of samples, not a complete inventory (Nicholls & Margules 1993). If all areas are well sampled, maps can be obtained directly from the database. If several (but not a few) areas are well sampled, predictive modeling can fill in the gaps in knowledge. Nevertheless, even if a number of well-surveyed areas can be identified accurately, unevenness in sampling effort could result in a partial (and biased) description of biodiversity variations (Dennis 2001). Therefore, it is necessary to determine the degree of environmental and geographic coverage of the studied region provided by the well-sampled areas (Hortal & Lobo 2005) and the possible bias produced by the coverage of regional environmental conditions obtained with these areas (Kadmon et al. 2004).

If well-sampled areas are not sufficient to describe the environmental and geographic variability of the region or are biased toward a limited number of the environmental domains present, additional surveys should be conducted. Well-surveyed areas should cover the entire spectrum of environmental conditions in the region and the entire geographic extent (i.e., represent all combinations of environmental conditions and areas that are environmentally similar but separated spatially). Araújo and Guisan (2006) stress the importance of improving sampling designs used for the prediction of species distributions. Additional surveys should target previously unsampled areas that improve the spatial and environmental coverage of well-sampled areas and diminish their bias (Hortal & Lobo 2005; Funk et al. 2005).

Once a sufficient number of areas are well surveyed (and well distributed), predictive modeling techniques can be used to obtain maps of biodiversity features (Fig. 5). There are many examples of reliable predictive maps of the distribution of biodiversity variables (e.g., Wohlgemuth 1998; Lobo & Martín-Piera 2002). Alternatively, modeling techniques can be used to predict the distribution of all species in the database one by one. This approach presents some intrinsic problems, especially when data are biased, if reliable absence data are lacking, and when modeling rare species (Hortal & Lobo 2006). The last quality control required is the validation of model predictions (Fig. 5).

Concluding Remarks

The BIOTA example shows the potential magnitude of the effects of two of the limitations inherent in a number of biodiversity databases: lack of exhaustiveness and lack of geographical (and environmental) coverage. The extent to which the effects of the limitations in this example can be extrapolated to other studies depends on each particular case, especially if one takes into account that the original purpose of BIOTA was not to develop analyses of the distribution of biodiversity. Such limitations exist in most databases. Lack of exhaustiveness or large gaps in their spatial and/or environmental coverage could compromise their future utility, in the same way that data gathered in the past have limited utility because the data lack detail and geographical coverage is not exhaustive. Therefore, we stress the importance of making an assessment of (and, eventually, improving) the quality of biodiversity databases to account for the effects of both problems. We encourage exhaustive compilation of all the available information with sufficient quality and detail.

Acknowledgments

We are indebted to M. Arechavaleta, I. Izquierdo, M. C. Marrero, N. Zurita, and J. L. Martín Esquivel, who provided access to and support in working with the data and promoted our use of BIOTA-Canarias. The paper has benefited from the critical review and fruitful advice of A. Machado, the comments of two anonymous referees and some discussion with J. de la Torre. This work was supported by a Spanish Ministerio de Educación y Ciencia Project (CGL2004-04309) and by a Portuguese Fundação para a Ciência e a Tecnología postdoctoral grant (BPD/20809/2004) to J.H. and a MNCN/CSIC/CM Ph.D. grant to A.J.-V.

Literature Cited

Araújo, M. B., and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography **33:**1677–1688.

Austin, M. P., J. G. Pausas, and A. O. Nicholls. 1996. Patterns of tree species richness in relation to environment in southeastern New South Wales, Australia. Australian Journal of Ecology **21:**154–164.

Borges, P. A. V., R. Cunha, R. Gabriel, A. F. Martins, L. Silva, and V. Vieira 2005. A list of the terrestrial fauna (Mollusca and Arthropoda) and flora (Bryophyta, Pteridophyta and Spermatophyta) from the Azores. Direcção Regional de Ambiente and Universidade dos Açores, Horta, Angra do Heroísmo and Ponta Delgada, Portugal.

Boyce, M. S., P. R. Vernier, S. E. Nielsen, and F. K. A. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling **157:**281–300.

Bramwell, D., and Z. Bramwell. 2001. Flores silvestres de las Islas Canarias. 4th edition. Editorial Rueda, Madrid.

Chefaoui, R. M., J. Hortal, and J. M. Lobo. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. Biological Conservation **122:**327–338.

Crawley, M. J. 1993. GLIM for ecologists. Blackwell Scientific Publications, Oxford, United Kingdom.

del Arco, M., P. L. Pérez de Paz, J. R. Acebes, J. M. González-Mancebo, J. A. Reyes-Betancort, J. A. Bermejo, S. de Armas, and R. González-González. 2006. Bioclimatology and climatophilous vegetation of Tenerife (Canary Islands). Annales Botanici Fennici **43:**167–192.

Dennis, R. L. H. 2001. Progressive bias in species status is symptomatic of fine-grained mapping units subject to repeated sampling. Biodiversity and Conservation **10:**483–494.

Dennis, R. L. H., and T. G. Shreeve. 2003. Gains and losses of French butterflies: test of predictions, under-recording and regional extinction from data in a new atlas. Biological Conservation **110:**131–139.

Dennis, R. L. H., and C. D. Thomas. 2000. Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. Journal of Insect Conservation **4:**73–77.

Dillon, S., and J. Fjeldsa. 2005. The implications of different species concepts for describing biodiversity patterns and assessing conservation needs for African birds. Ecography **28:**682–692.

Diniz-Filho, J. A. F., L. M. Bini, and B. A. Hawkins. 2003. Spatial autocorrelation and red herrings in geographical ecology. Global Ecology and Biogeography **12:**53–64.

Faith, D. P., and P. A. Walker. 1996. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. Biodiversity and Conservation **5:**399–415.

Fernández-Palacios, J. M. 1992. Climatic responses of plant species on Tenerife (Canary Islands). Journal of Vegetation Science **3:**595–602.

Fernández-Palacios, J. M., and J. P. de Nicolás. 1995. Altitudinal pattern of vegetation variation on Tenerife. Journal of Vegetation Science **6:**183–190.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology **51:**331–363.

Ferrier, S., and A. Guisan. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology **43:**393–404.

Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure? Pages 271–280 in J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall, and F. Samson, editors. Predicting species occurrences: Issues of accuracy and scale. Island Press, Covelo, California.

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.

Funk, V. A., K. S. Richardson, and S. Ferrier. 2005. Survey-gap analysis in expeditionary research: where do we go from here? Biological Journal of the Linnean Society **85:**549–567.

Griffiths, G. H., B. C. Eversham, and D. B. Roy. 1999. Integrating species and habitat data for nature conservation in Great Britain: data sources and methods. Global Ecology and Biogeography **8:**329–345.

Hortal, J., and J. M. Lobo. 2005. An ED-based protocol for the optimal sampling of biodiversity. Biodiversity and Conservation **14:**2913–2947.

Hortal, J., and J. M. Lobo. 2006. A synecological framework for systematic conservation planning. Biodiversity Informatics **3:**16–45.

Hortal, J., J. M. Lobo, and F. Martín-Piera. 2001. Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeinae). Biodiversity and Conservation **10:**1343–1367.

Hortal, J., P. Garcia-Pereira, and E. García-Barros. 2004. Butterfly species richness in mainland Portugal: predictive models of geographic distribution patterns. Ecography **27:**68–82.

Hortal, J., P. A. V. Borges, and C. Gaspar. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology **75:**274–287.

Izquierdo, I., J. L. Martín, N. Zurita, and M. Arechavaleta, editors. 2005. Lista de especies silvestres de Canarias (hongos, plantas y animales terrestres). 2nd edition. Consejería de Política Territorial y Medio Ambiente, Gobierno de Canarias, La Laguna, Tenerife, Spain.

Johnson, D. H., and G. A. Sargeant. 2002. Toward better atlases: improving presence-absence information. Pages 391–397 in J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall, and F. Samson, editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Covelo, California.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications **14:**401–413.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam.

Lobo, J. M., and F. Martín-Piera. 2002. Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. Conservation Biology **16:**158–173.

Lobo, J. M., I. Castro, and J. C. Moreno. 2001. Spatial and environmental determinants of vascular plant species richness distribution in the Iberian Peninsula and Balearic Islands. Biological Journal of the Linnean Society **73:**233–253.

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

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Chapman & Hall, London.

Michener, W. K., and J. Brunt. 2000. Ecological data: design, management and processing. Blackwell Science, London.

Nicholls, A. O. 1989. How to make biological surveys go further with generalised linear models. Biological Conservation **50:**51–75.

Nicholls, A. O., and C. Margules. 1993. An upgraded reserve selection algorithm. Biological Conservation **64:**165–169.

Porter, J. H. 2000. Scientific databases. Pages 48–69 in W. K. Michener and J. Brunt, editors. Ecological data: design, management and processing. Blackwell Science, Oxford, United Kingdom.

Prendergast, J. R., S. N. Wood, J. H. Lawton, and B. C. Eversham. 1993. Correcting for variation in recording effort in analyses of diversity hotspots. Biodiversity Letters **1:**39–53.

Soberón, J., and J. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology **7:**480–488.

Soberón, J., and T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London B **359:**689–698.

Soberón, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distribution areas. Biodiversity Informatics **2:**1–10.

Soberón, J., J. Llorente, and H. Benítez. 1996. An international view of national biological surveys. Annals of the Missouri Botanical Garden **83:**562–573.

Soberón, J., R. Jiménez, J. Golubov, and P. Koleff. 2007. Assessing completeness of biodiversity databases at different spatial scales. Ecography **30:**152–160.

Stockwell, D. R. B., and A. T. Peterson. 2002. Controlling bias in biodiversity data. Pages 537–546 in J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall, and F. Samson, editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Covelo, California.

Valdecasas, A. G., and A. Camacho. 2003. Conservation to the rescue of taxonomy. Biodiversity and Conservation **12:**1113–1117.

Voggenreiter, V. 1974. Geobotanischen Untersuchungen an der natürlichen Vegetation der Kanariensiel Teneriffa. Dissertationes Botanicae **26:**1–718.

Whittaker, R. J., M. B. Araújo, P. Jepson, R. J. Ladle, J. E. M. Watson, and K. J. Willis. 2005. Conservation biogeography: assessment and prospect. Diversity and Distributions **11:**3–23.

Wohlgemuth, T. 1998. Modelling floristic species richness on a regional scale: a case study in Switzerland. Biodiversity and Conservation **7:**159–177.