Short communication

Taxonomist survey biases and the unveiling of biodiversity patterns

Pablo Sastre, Jorge M. Lobo*

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

ARTICLE INFO

ABSTRACT

Basic and applied studies based on biodiversity data need accurate information on the distribution of species. However, several studies clearly show that this information is frequently biased, mainly as a consequence of aggregated survey patterns in which taxonomists repeatedly select localities with specific characteristics. In this study, we have constructed three different but simple virtual species richness scenarios to simulate the capacity of random, aggregated or regular survey designs to reveal the true biodiversity pattern. We are specifically interested in the effect of taxonomist insistence on surveying those localities that guarantee success in the collection of as many species as possible (species richness bias), and on the coordinated or uncoordinated character of the efforts carried out by the whole community of taxonomists. In all simulated species richness scenarios, a survey directed towards those localities that were previously recognized as having a higher species richness value is not recommended if the aim is to recover the true geographical pattern of species richness in a given territory. This aggregated process of allocating survey localities is probably caused by the primary aim of taxonomists, which is to acquire specimens of rare species and/or as many species as possible. However, an increase in taxonomist curiosity towards non-surveyed localities near those previously identified as the richest allows one to obtain better results, provided that the species richness pattern is not too patchy and the effort for discovering the true map is not too difficult. Our results suggest that planned survey designs are necessary when most of the data comes from studies not specifically designed to reveal the distribution of biodiversity. The capacity of this data to represent the real geographical pattern of biodiversity may depend on the capacity of the taxonomist community to be self-motivated.

1. Introduction

The conservation of biodiversity is an international goal. It requires a joint cooperative effort to assemble all the biodiversity information at hand in order to assess conservation priorities according to the current status of biodiversity and present and future threats. To do this, a number of taxonomic and biogeographic database programs have been developed and are currently merging (Soberón et al., 1996; Bisby, 2000; Saarenmaa and Nielsen, 2002; Godfray, 2002; GBIF, 2003; Gra-

ham et al., 2004; Hortal et al., 2007; Guralnick et al., 2007). These initiatives aim to gather all the information stored by taxonomists over more than two centuries in bibliography and natural history collections. Herein, by 'taxonomist' we refer to both professional and amateur collectors. However, we are far from knowing the whole catalogue of life (Chapman, 2005; the Linnean shortfall), not to mention the distribution of all biological diversity (Whittaker et al., 2005; the Wallacean shortfall). In countries with a long naturalist tradition, survey effort has been intense but poorly coordinated, generating a huge amount of taxonomic and distributional data dispersed in bibliography and natural history collections. In contrast, there is not enough taxonomic and biogeographical information about most parts of world's richest regions. Therefore, well-known groups and exhaustively inventoried areas are the exception rather than the rule, in spite of the massive amounts of data currently being gathered. For most of the biota and regions, only a few local inventories and some sporadic uncoordinated captures exist. This is especially true for hyperdiverse groups of organisms such as invertebrates (more than 80% of global species richness), and for global species richness hotspots (May, 1990).

Conservation biologists face a conflict of interest; on the one hand the need of biological information for basic and applied studies, and on the other the need of being cautious due to the lack of quality and incompleteness of most of this information. In fact, the information compiled on species distributions presents a marked pattern of spatial and environmental biases and survey incompleteness (Dennis and Hardy, 1999; Dennis et al., 1999; Soberón et al., 2000; Lobo et al., 2007; Hortal et al., 2008; Kadmon et al., 2003, 2004; Kéry et al., 2008). Species location data suffers from temporal bias in the recording process; distribution data collected at different times generally results in different distribution maps for the same species (Lobo et al., 2007). Such spatially and environmentally structured increases in species data over time are the consequence of an expansion of the species range resulting from sociological, environmental, or sampling effort biases (Dennis and Hardy, 1999; Dennis et al., 1999, 2006; Dennis and Thomas, 2000; Zaniewski et al., 2002; Anderson, 2003; Reutter et al., 2003; Graham et al., 2004; Soberón and Peterson, 2004; Martínez-Meyer, 2005), in which species occurring in those places with a prolonged taxonomical tradition have a greater probability of being discovered than others that are generally located in equatorial, more species rich regions (Cabrero and Lobo, 2003). Therefore, site accessibility, distance from the place of residence of the collecting taxonomist, and degree of interest of taxonomists are all key parameters leading to the observed concentration in species presence (Nelson et al., 1990; Peterson et al., 1998; Dennis and Thomas, 2000; Parnell et al., 2003; Reddy and Dávalos, 2003). Hence, the efficiency of biodiversity studies is highly conditioned by the quality and spatial or environmental biases of the collected biological information (Nelson et al., 1990; Faith, 2002; Williams et al., 2002; Hortal and Lobo, 2006; Hortal et al., 2007). This needs the prior detection of the location and extent of these biases, as well as the completion of additional surveys directed at filling the current gaps in knowledge (Hortal et al., 2007; Guralnick et al., 2007).

The biases in the information stored in biodiversity databases are mainly the consequence of an aggregated survey pattern in which taxonomists repetitively select localities that are historically recognized as having greater values of biodiversity (Dennis and Thomas, 2000). For example, when species richness predictions for a whole territory are carried out using only the data from well-surveyed localities, the forecasted values tend to be higher in the most surveyed localities (Lobo and Martín-Piera, 2002; Romo et al., 2006; Sánchez-Fernández et al., 2008). This suggests that taxonomists tend to concentrate their efforts in the localities that guarantee success in the collection of as many species as possible (hereafter, species richness bias). This species richness bias highlights another feature of our current species richness distribution representations: their uncoordinated character. Although the sum of the individual efforts of the whole community of taxonomists allows us to obtain a representation of the species richness distribution for a taxonomical group, this picture is not the consequence of a collaborative, planned survey to reveal the "true" species richness map. Using three virtual scenarios, we simulate the process of species survey in order to examine the consequences of following random, aggregated or regular survey designs in the progressive discovery of the species richness distribution of a region. As these survey strategies differ in the weight of the bias towards species rich sites and in the degree of collective organization followed by the taxonomists, we subsequently discuss the implications of these forces in revealing the hidden species richness distribution that our current maps fail to show.

2. Methods

2.1. Virtual scenarios of biodiversity

First, virtual scenarios were built within a square-shaped area comprising 2500 pixels, i.e., localities (50 rows × 50 columns), assigning species richness or biodiversity values (B) from 1 to 10 to each locality. Three different simulated patterns were considered: a gradient pattern and two patchy patterns (Fig. 1). The gradient pattern was built by assigning B values in a simple latitudinal progression. Patchy patterns were built by aggregating pixels in patches by means of modal filters in which each pixel receives the most frequent value in a surrounding window. The fine-patch pattern was built by successively applying a window of 3 × 3 pixels, while the coarse-patch pattern was done by applying the modal filter to a window of 7 × 7 pixels (Fig. 1).

2.2. Surveying strategies

The three virtual species richness scenarios were sampled using different survey strategies so that each successive locality selection (trial) was analogous to the choice of a locality by a taxonomist. These survey procedures differ in whether there exists species richness bias or not, and also in their coordinated or uncoordinated character, depending on whether the probability of selection of each taxonomist is influenced or not by past selections.
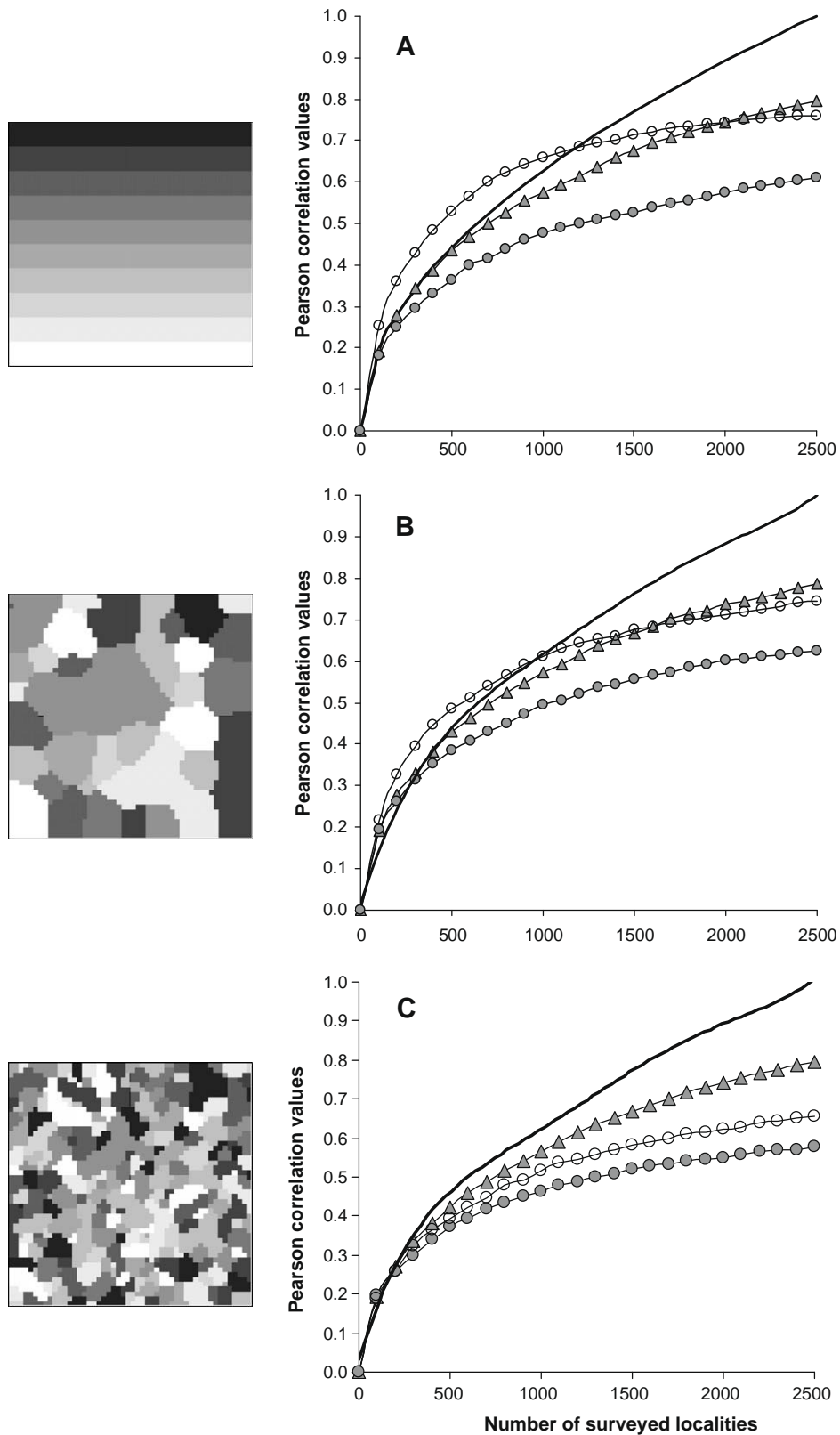
Fig. 1 – Left: the three simulated biodiversity scenarios (50 × 50 pixels): a gradient pattern, a coarse-patch pattern, and a fine-patch pattern. In each pattern, local species richness values vary from 1 to 10. Patchy patterns were built by using modal filters (windows of 3 × 3 pixels or 7 × 7 pixels; see methods). Right: variation of the Pearson correlation values between the whole species richness in the three simulated biodiversity scenarios and the species values revealed at different levels of survey effort (number of surveyed localities) according to the four survey strategies considered: regular survey (continuous line), random survey (triangles), directed survey (black circles), and distance-directed survey (open circles). A = gradient pattern, B = coarse-patch pattern, and C = fine-patch pattern.

Three basic survey strategies were applied to the virtual biodiversity scenarios discussed above: regular, random and directed. The regular survey represents a planned sampling in which the whole territory to be surveyed is divided between the different taxonomists and the effort carried out at each locality is enough to detect the species inhabiting it. This regular survey is organized and free of species richness bias, and is equivalent to a random selection without replacement (i.e., one deliberately avoids choosing any locality more than once). The random survey implies sampling one locality each time with replacement, so the same locality can be selected more than once by chance during the whole survey process. This survey procedure would be equivalent to one generated by a completely uncoordinated community of taxonomists that select the localities without any species richness bias.

The directed survey process is one in which the probability of selection of a locality depends on its known B value, and so generates an aggregated survey pattern. After a first random selection, the subsequent probability of selection of the same locality is provided by a "biodiversity correction score" (BCS) which depends on the number of species discovered in this locality. Thus, those surveyed localities with B values higher than average values for the whole territory will have a higher probability of being selected again than those localities still un-surveyed. BCS was calculated as:

$$\frac{\text{Number of localities with a } B \text{ value} \leq B \text{ value of locality } i}{\text{Total number of localities}}$$

In contrast to the random survey strategy, the directed survey is a coordinated procedure since each taxonomist takes advantage of the information previously compiled by other colleagues. In order to estimate the effect of a relaxation in the species richness bias, we finally simulated a directed survey procedure in which a criterion of spatial distance is adopted for the allocation of higher survey probabilities (distance-directed survey). Thus, the BCS probability is applied both to each selected locality and to their neighbours (higher probability of sampling in the zones near localities of greater biodiversity). To include this spatial distance criterion, a buffer zone is defined around each surveyed locality assigning similar values to all the localities within it. The buffer zone includes 20 localities around the surveyed one (located at a distance less than 2.5 times the size of the side; see Fig. 2). When a locality was included in the buffer zones of two or more sampled localities, a corresponding average value is assigned to it. This distance-directed survey is also coordinated, but assumes that, once taxonomists have visited a locality with high values of biodiversity (i.e., a historical collection place), they have enough curiosity to survey in nearby but unknown localities.

2.3. Survey evaluation

In order to evaluate the capacity of the different types of surveys to reveal the artificially created biodiversity patterns, the similarity between the species richness values obtained in all localities (2500) and the "true" species richness picture was assessed by calculating the Pearson correlation coefficient at each 100 runs or surveyed localities. As the standard devia-
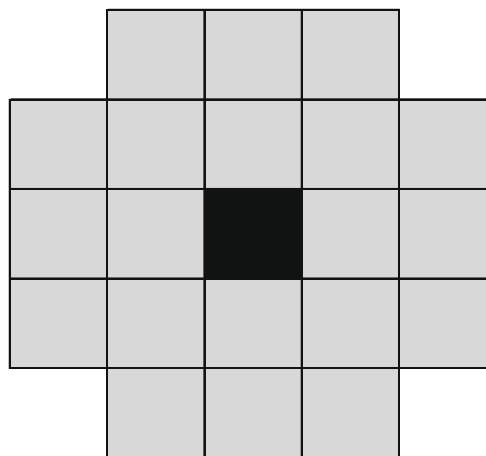

Fig. 2 – Definition of the buffer zone (grey) around the surveyed locality (black) used to define the distance-directed survey procedure.

tion of the correlation coefficients is lower than 0.01 when we performed several survey trials (not shown), we have decided to describe the changes in these values without using any statistical test. We decide to end the survey when the number of sampled localities equals the whole number of localities (2500).

3. Results

Except in the case of the regular survey, all the remaining survey strategies provide species richness maps that are progressively more similar to the real pattern in a quasi-asymptotic form. The directed survey is always biased according to the detected species richness values showing the worst results (Fig 1).

In the gradient biodiversity scenario (Fig. 1A), the directed survey strategy begins to generate worse results than an uncoordinated random survey when the survey effort is around 200 localities (8% of total). However, the distance-directed survey reveals the true species richness map better than a random survey until 70% of total localities have been surveyed, but is also better than a regular survey strategy until reaching a survey effort of around 40% of total localities (Fig. 1A). Once half of the territory was surveyed, only a coordinated regular strategy, free of species richness bias, is able to progressively increase the accuracy of the derived species richness map beyond a Pearson correlation value of 0.80.

The different survey processes seem to provide similar results in the case of the coarse-patch species richness scenario (Fig. 1B), although in this case the advantage of the distance-directed survey over the random and regular surveys disappears at lower levels of survey effort (around 50% and 30% of total localities, respectively). However, when the patchiness of the simulated species richness scenario increases, the distance-directed survey protocol cannot improve the revelation of the true species richness map carried out by regular or random survey processes (Fig. 1C). In these circumstances, the coordinated regular survey seems to always be the most successful strategy to discover the distribution of species richness.

4. Discussion

Taxonomist survey efforts generally possess an unknown degree of coordination, often unintentional. Although a taxonomist can decide to survey an un-prospected region after considering the location of former studies, many taxonomists frequently go to the same locality or region (species rich or not). Both our random-uncoordinated and regular-coordinated survey processes should be considered extreme and unrealistic situations. The lack of information about the localities formerly surveyed by other taxonomists together with the inaccessibility of collected specimens are factors that promote the uncoordinated and repeated survey of the same localities (as in our random survey), while data accessibility would have the advantage of inevitably promoting a more efficient use of previous survey efforts. Of course, well-designed methodologies able to maximize the amount of variation covered by the localities to be surveyed (Austin and Heyligers, 1989; Hortal and Lobo, 2005) can lead to better results than the regular protocol by using a minimum survey effort. However, these approaches are adequate for the compilation of completely new biological information on a nearly un-surveyed territory, but not for representing the survey process generated by the successive accumulation of information provided by taxonomists over a long period of time.

A survey directed towards localities previously recognized as having higher species richness values is unpromising for discovering the true species richness distribution of a territory. Of all the simulated species richness scenarios, this kind of survey process is the worst one, finishing without surveying many localities even when a high level of sampling effort has been carried out on the whole territory. Paradoxically, this aggregated process of allocating survey localities is caused by data accessibility; previously generated taxonomical and distributional data can direct taxonomist efforts towards species rich localities. The frequent coincidence of rarity and species richness hotspots (Blackburn and Gaston, 1996; Lamoreux et al., 2005; but see also Grenyer et al., 2006) would favour the universality of this bias. In our opinion this aggregated survey pattern happens because the primary intention of a taxonomist is not to reveal the complete biodiversity distribution pattern of a territory, but to acquire specimens of rare species and/or as many species as possible in order to accumulate a taxonomic collection useful for comparative and genealogic purposes. However, and interestingly, an increase in taxonomist curiosity towards non-surveyed localities near those previously recognized as the richest could increase the success of revealing the true species richness pattern, provided that: (i) this pattern is not too patchy, and (ii) the effort of discovering the true map is not too exigent. More powerful survey designs (De Gruijter et al., 2006) should be used if the patchiness in the distribution of species richness is high or when we aim to completely reveal the species richness pattern.

The available studies on the increase in survey effort suggest that taxonomists tend to survey repeatedly in some localities and landscape types (Dennis and Hardy, 1999; Funk and Richardson, 2002; Reddy and Dávalos, 2003; Soberón and Peterson, 2004; Diniz-Filho et al., 2005; Moerman and Esta-brook, 2006; Pautasso and McKinney, 2007; Schulman et al., 2007). Such biases can seriously diminish the accuracy of the geographical species richness representations generated by using the taxonomical and distributional information individually compiled by taxonomists over the last two centuries (see for example Gu and Swihart, 2004). Our results suggest that planned survey designs are necessary when most of the data come from studies not specifically designed to reveal the distribution of biodiversity, and also that the capacity of this data for representing the real geographical biodiversity pattern may depend on the capacity of the taxonomists' community to be self-motivated. If we assume that "taxonomic curiosity" is higher when the collected specimens come from a diverse community of taxonomists, the geographical biodiversity representations of those regions prospected by a group of unrelated taxonomists that pursue different purposes would have a high likelihood of being accurate.

Acknowledgements

REFERENCES

Anderson, R.P., 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. Journal of Biogeography 30, 591–605.

Austin, M.P., Heyligers, P.C., 1989. Vegetation survey design for conservation: gradsect sampling of forest in North-eastern New South Wales. Biological Conservation 50, 13–32.

Bisby, F.A., 2000. The quiet revolution: biodiversity informatics and the internet. Science 289, 2309–2312.

Blackburn, T.M., Gaston, K.J., 1996. Spatial patterns in the species richness of birds in the new world. Philosophical Transactions of the Royal Society of London series B 351, 897–912.

Cabrero, F.J., Lobo, J.M., 2003. Estimating the number of species not yet described and their characteristics: the case of western Palaearctic dung beetle species (Coleoptera, Scarabaeoidea). Biodiversity and Conservation 12, 147–166.

Chapman, A.D., 2005. Uses of primary species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

De Gruijter, J., Brus, D., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer-Verlag, Berlin.

Dennis, R.L.H., Hardy, P.B., 1999. Targeting squares for survey: predicting species richness and incidence for a butterfly atlas. Global Ecology and Biogeography 8, 443–454.

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

Dennis, R.L.H., Shreeve, T.G., Isaac, N.J.B., Roy, D.B., Hardy, P.B., Fox, R., Asher, J., 2006. The effects of visual apparency on bias in butterfly recording and monitoring. Biological Conservation 128, 486–492.

Dennis, R.L.H., Sparks, T.H., Hardy, P.B., 1999. Bias in butterfly distribution maps: the effects of sampling effort. Journal of Insect Conservation 3, 33–42.

Diniz-Filho, J.A.F., Bastos, R.P., Rangel, T.F.L.V.B., Bini, L.M., Carvalho, P., Silva, R.J., 2005. Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. Global Ecology and Biogeography 14, 469–477.

Faith, D., 2002. Those complementarity analysis do not reveal extent of conservation conflict in Africa. Science dEbate. <http://www.sciencemag.org/cgi/eletters/293/5535/1591/#381>.

Funk, V.A., Richardson, K.S., 2002. Systematic data in biodiversity studies: use it or lose it. Systematic Biology 51, 303–316.

GBIF, 2003. Global biodiversity information facility strategic plan. <http://www.gbif.org/GBIF_org/documents>.

Godfray, C., 2002. Challenges for taxonomy. Nature 417, 17–19.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19, 497–503.

Grenyer, R., Orme, C.D.L., Jackson, S.F., Thomas, G.H., Davies, R.G., Davies, T.J., Jones, K.E., Olson, V.A., Ridgely, R.S., Rasmussen, P.C., Ding, T.-S., Bennett, P.M., Blackburn, T.M., Gaston, K.J., Gittleman, J.L., Owens, I.P.F., 2006. Global distribution and conservation of rare and threatened vertebrates. Nature 444, 93–96.

Gu, W., Swihart, R.K., 2004. Absent or undetected? effects of non-detection of species occurrence on wildlife-habitat models. Biological Conservation 116, 195–203.

Guralnick, R.P., Hill, A.W., Lane, M., 2007. Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters 10, 663–672.

Hortal, J., Lobo, J.M., 2006. Towards a synecological framework for systematic conservation planning. Biodiversity Informatics 3, 16–45.

Hortal, J., Lobo, J.M., Jiménez-Valverde, A., 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). Conservation Biology 21, 853–863.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed realized niche of the species. Oikos 117, 847–858.

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

Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13, 853–867.

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

Kéry, M., Royle, J.A., Schmid, H., 2008. Importance of sampling design and analysis in animal population studies: a comment on Sergio et al. Journal of Applied Ecology 45, 981–986.

Lamoreux, J.F., Morrison, J.C., Ricketts, T.H., Olson, D.M., Dinerstein, E., McKnight, M.W., Shugart, H.H., 2005. Global tests of biodiversity concordance and the importance of endemism. Nature 440, 212–214.

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

Lobo, J.M., Baselga, A., Hortal, J., Jiménez-Valverde, A., Gómez, J.F., 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? Diversity and Distributions 13, 772–780.

Martínez-Meyer, E., 2005. Climate change and biodiversity: some considerations in forecasting shifts in species potential distributions. Biodiversity Informatics 2, 42–55.

May, R.M., 1990. How many species? Philosophical Transactions of the Royal Society of London 330, 171–182.

Moerman, D.E., Estabrook, G.F., 2006. The botanist effect: counties with maximal species richness tend to be home to universities and botanists. Journal of Biogeography 33, 1969–1974.

Nelson, B.W., Ferreira, C.A.C., da Silva, M.F., Kawasaki, M.L., 1990. Endemism centers, refugia and botanical collection density in Brazilian Amazonia. Nature 345, 714–716.

Parnell, J.A.N., Simpson, D.A., Moat, J., Kirkup, D.W., Chantaranothai, P., Boyce, P.C., Bygrave, P., Dransfield, S., Jebb, M.H.P., Macklin, J., Meade, C., Middleton, D.J., Muasya, A.M., Prajaksood, A., Pendry, C.A., Poomar, R., Suddee, S., Wilkin, P., 2003. Plant collecting spread and densities: their potential impact on biogeographical studies in Thailand. Journal of Biogeography 30, 193–209.

Pautasso, M., McKinney, M.L., 2007. The botanist effect revisited: plant species richness, county area, and human population size in the United States. Conservation Biology 21, 1333–1340.

Peterson, A.T., Navarro-Sigüenza, A.G., Benítez-Díaz, H., 1998. The need for continued scientific collecting: a geographic analysis of Mexican bird specimens. Ibis 140, 288–294.

Reddy, S., Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30, 1719–1727.

Reutter, B.A., Helfer, V., Hirzel, A.H., Vogel, P., 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. Journal of Biogeography 30, 581–590.

Romo, H., García-Barros, E., Lobo, J.M., 2006. Identifying recorder-induced geographic bias in an Iberian butterfly database. Ecography 29, 873–885.

Saarenmaa, H., Nielsen, E.S. (Eds.), 2002. Towards a global biological information infrastructure. Challenges, Opportunities, Synergies, and the Role of Entomology. European Environment Agency, Copenhagen.

Sánchez-Fernández, D., Lobo, J.M., Abellán, P, Ribera, I., 2008. Bias in freshwater biodiversity sampling: the case of Iberian water beetles. Diversity and Distributions, DOI: 10.1111/j.1472-4642.2008.00474.x.

Schulman, L., Toivonen, T., Ruokolainen, K., 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species rage estimation. Journal of Biogeography 34, 1388–1399.

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

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

Soberón, J.M., Llorente, J., Oñate, L., 2000. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. Biodiversity and Conservation 9, 1441–1466.

Williams, P.H., Margules, C.R., Hilbert, D.W., 2002. Data requirements and data sources for biodiversity priority area selection. Journal of Bioscience 27, 327–338.

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

Zaniewski, A.E., Lehmann, A., Overton, J.M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157, 261–280.