BIODIVERSITY RESEARCH

Bias in freshwater biodiversity sampling: the case of Iberian water beetles

David Sánchez-Fernández[1]*, Jorge M. Lobo[2], Pedro Abellán[1,2], Ignacio Ribera[2] and Andrés Millán[1]

[1]*Department of Ecology and Hydrology, University of Murcia, 30100 Espinardo, Murcia, Spain,* [2]*Museo Nacional de Ciencias Naturales, C/José Gutiérrez Abascal, 2, 28006 Madrid, Spain*

*Correspondence: David Sánchez-Fernández, Department of Ecology and Hydrology, University of Murcia, 30100 Espinardo, Murcia, Spain. E-mail: davidsan@um.es

ABSTRACT

Extensive distributional data bases are key tools in ecological research, and good-quality data are required to provide reliable conservation strategies and an understanding of biodiversity patterns and processes. Although the evaluation of data bases requires the incorporation of estimates of sampling effort and bias, no studies have focused on these aspects for freshwater biodiversity data. We used here a comprehensive data base of water beetles from the Iberian Peninsula and the Balearic Islands, and examine whether these data provide an unbiased, reliable picture of their diversity and distribution in the study area. Based on theoretical estimates using the Clench function on the accumulated number of records as a surrogate of sampling effort, about a quarter of the Iberian and Balearic 50 × 50 km Universal Transverse Mercator grid cells can be considered well prospected, with more than 70% of the theoretical species richness actually recorded. These well-surveyed cells are not evenly distributed across biogeographical and physicoclimatic subregions, reflecting some geographical bias in the distribution of sampling effort. Our results suggest that recording was skewed by relatively simple variables affecting collector activity, such as the perceived 'attractiveness' of mountainous landscapes and protected areas with recently described species, and accessibility of sampling sites (distance from main research centres). We emphasize the importance of these evaluation exercises, which are useful to locate areas needed of further sampling as well as to identify potential biases in the distribution of current biodiversity patterns.

Keywords

Completeness, data base, freshwater biodiversity, Iberian Peninsula, sampling bias, water beetles.

INTRODUCTION

Conservation assessment and biodiversity research require high-quality data on species' distributions, these usually being in the form of extensive data bases (Hortal *et al.*, 2007). Only countries with a long-standing tradition of natural history and sufficient resources are able to produce good distribution maps based on adequate sampling of a number of taxonomic groups (Lawton *et al.*, 1994; Griffiths *et al.*, 1999). This is not the case with many Mediterranean countries, in which inventories of many animal groups, particularly insects, are incomplete or non-existent (Ramos *et al.*, 2001).

Hortal *et al.* (2007) noted two general drawbacks associated with of the use of biodiversity data bases: (1) lack of survey-effort assessments (and lack of exhaustiveness in compiling data on survey effort), and (2) incomplete coverage of the geographical and environmental diversity that affects the distribution of the organisms. These problems render existing data bases and/or atlases of limited use for accurately describing patterns of biodiversity (Prendergast *et al.*, 1993; Dennis & Shreeve, 2003; Soberón *et al.*, 2007), and compromise the utility of any predictive models based on them (Hortal & Lobo, 2006; Lobo *et al.*, 2007). Therefore, it is necessary to incorporate estimates of sampling bias and measures of sampling effort in biodiversity studies to minimize their potential confounding effect (Romo *et al.*, 2006).

A number of attempts have been made to explore these issues, using data bases from a variety of regions and covering a diversity of taxonomic groups (e.g. Dennis *et al.*, 1999, 2006; Lobo & Martín-Piera, 2002; Reddy & Dávalos, 2003; Romo *et al.*, 2006). However, to date no study has focused on freshwater biodiversity, probably due to the paucity of inventory data for freshwater systems (Lévêque *et al.*, 2005), especially in Mediterranean countries. Freshwater biodiversity may be particularly at risk in many regions of the world (e.g. Allan & Flecker, 1993; Master *et al.*, 1998; Ricciardi & Rasmussen, 1999), and inland aquatic systems typically harbour a diverse biota, rich in endemic taxa.

DOI: 10.1111/j.1472-4642.2008.00474.x
© 2008 The Authors
Journal compilation © 2008 Blackwell Publishing Ltd

This is particularly the case in the Mediterranean Basin, which is considered as one of Earth's biodiversity hotspots (Quézel, 1995; Mittermeier *et al.*, 1998; Myers *et al.*, 2000). The factors affecting the quality of data bases of freshwater organisms are likely to differ from those of terrestrial ones, as the sampling and collecting of data rarely overlap. With this work, we aim to provide a case study for one of the most diverse and well-studied groups of freshwater macroinvertebrates in a highly diverse region, the aquatic Coleoptera of the Iberian Peninsula and the Balearics.

Water beetles have high species richness in the Mediterranean region, inhabiting virtually every kind of fresh- and brackishwater habitat, from the smallest ponds to lagoons and wetlands, and from streams to irrigation ditches, large rivers, and reservoirs (e.g. Ribera *et al.*, 1998; Ribera, 2000; Millán *et al.*, 2002). Water beetles have been proposed as good surrogates of biodiversity in Mediterranean aquatic ecosystems (Sánchez-Fernández *et al.*, 2006) and have been used to select priority areas for conservation in this region (Sánchez-Fernández *et al.*, 2004; Abellán *et al.*, 2005). In comparison to other groups of freshwater invertebrates in the Iberian Peninsula and the Balearic Islands, water beetles are well known in their systematic and biogeography (Ribera *et al.*, 1998; Ribera, 2000; Millán *et al.*, 2006).

By analysing an exhaustive data base on Iberian water beetles, we aim to determine whether these data are able to provide an unbiased picture of the species diversity and distribution. First, we identify the most probable well-surveyed areas examining whether they effectively cover the different biogeographical and environmental subregions recognized in the Iberian Peninsula. We then examine the distribution of sampling effort, and the extent to which sampling bias can be explained by a suite of environmental variables. Finally, we identify the key areas where the effort should be concentrated in future sampling programs.

METHODS

Study area

The Iberian Peninsula and the Balearic Islands include a wide variety of biomes, relief, climates, and soil types (Fig. 1). Although the Iberian Peninsula lies in the temperate zone, its rugged topography gives rise to a great diversity of climates, from semiarid Mediterranean, to oceanic in the northern fringe, and alpine in the high mountains. Mean annual temperatures oscillate between 2.2 and 19 °C, and total annual precipitation between 203 and 2990 mm. Due to this great variety of relief and climate, the Iberian Peninsula includes an enormous diversity of vegetation types, from deciduous and coniferous forests to sclerophyllous woodlands or annual steppe grasslands (Rey-Benayas & Scheiner, 2002).

Source of biological data

This work is based on an exhaustive data base of Iberian water beetle records (ESACIB 'EScarabajos ACuáticos IBéricos') compiling all available taxonomic and distributional data from the literature (485 bibliographic references and 34,504 data base records) as well as museum and private collections, PhD theses, and other unpublished sources (16,260 records). After deletion of records with taxonomic uncertainties or doubtful or imprecise localities, ESACIB contains around 50,000 workable records belonging to 511 species or subspecies of 11 families of water beetles (Table 1).

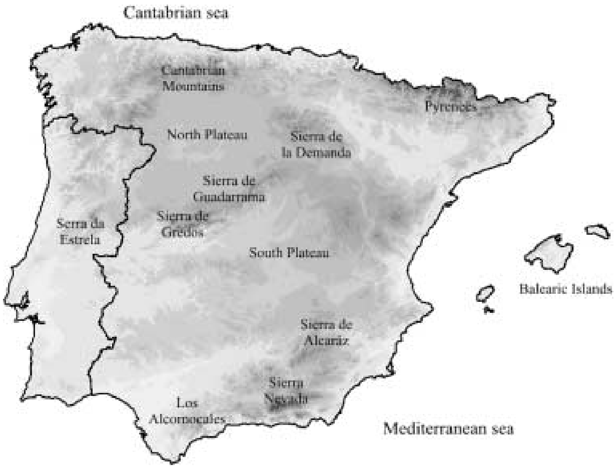

Figure 1 Study area with some locations referred to in text highlighted.

Table 1 Total species richness (S), number of data base records (DR), mean number of species (S_M), and mean number of data base records (DR_M) per 50 × 50 km Universal Transverse Mercator cell (± standard deviation) for each family of water beetles.

Family	S	DR	S_M	DR_M
Dryopidae	16	1143	4.5 ± 14.1	0.9 ± 1.3
Dytiscidae	171	22,056	85.8 ± 146.7	20.4 ± 16.3
Elmidae	32	5561	21.6 ± 52.1	4.4 ± 4.8
Gyrinidae	10	1207	4.7 ± 9	1.4 ± 1.6
Haliplidae	17	2448	9.5 ± 18.2	2.1 ± 2
Helophoridae	34	1826	7.1 ± 16.3	2.1 ± 3.1
Hydraenidae	148	8204	31.9 ± 61.3	8.5 ± 8.3
Hydrochidae	11	626	2.4 ± 10.1	0.8 ± 1.2
Hydrophilidae	68	6790	26.4 ± 66.1	6.8 ± 6.9
Hygrobiidae	1	259	1 ± 3.1	0.3 ± 0.4
Noteridae	3	644	2.5 ± 11.3	0.5 ± 0.7
Total	511	50,764	197.4 ± 408.3	48.2 ± 46.6

The data base was originally referenced at a resolution of 100 km^2 (10 × 10 km cells), although for simplicity we used 50 × 50 km Universal Transverse Mercator (UTM) squares as geographical units ($n = 257$) in this study. This loss of resolution was necessary since only 35% of the 10 × 10 km cells contain data, and only 3.7% of these have twice the number of data base records than species. Grid cells containing < 15% of land were not considered, and data base records from these cells added to the most environmentally similar neighbouring cell.

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Assessing sampling effort

The number of data base records in each cell was chosen as a surrogate of sampling effort, following Lobo & Martín-Piera (2002) and Hortal *et al.* (2004). Such an approach has been demonstrated to yield similar results to the use of other measures of sampling effort, such as data on number of individuals recorded, or number of traps employed (Hortal *et al.*, 2006). We assumed that the number of records in a grid cell is directly related to survey effort, and that the probability of species' occurrence correlates positively with the number of data base records.

Collector's curves were used to identify grid squares with inventories complete enough to produce reliable richness scores. Collector's curves are generally considered a good approach to evaluate the quality of inventories (Soberón & Llorente, 1993; Jiménez-Valverde & Hortal, 2003; Hortal *et al.*, 2004). These curves represent the expected accumulated number of species encountered within a certain geographical area as a function of a measure of the effort (number of records in this case) invested to collect them (Soberón & Llorente, 1993; Colwell & Coddington, 1994; Gotelli & Colwell, 2001). The slope of the collector's curve determines the rate of species accumulation at a given level of sampling effort. This slope diminishes with sampling effort and as new species are found, reaching a hypothetical value of 0 when all species are detected. As the shape of this relationship depends on the order in which individuals were recorded, this order was randomized 100 times to obtain a smoothed accumulation curve (using the EstimateS 6.0 software package; Colwell, 2000). The Clench function was fitted to the smoothed data, and the asymptotic value (i.e. the species richness predicted for an ideally unlimited sample size) was computed. The ratio of recorded to predicted species richness (the asymptotic score) was used as a measure of completeness of each cell inventory. A UTM cell was considered to be adequately sampled when the completeness values were ≥ 70% (following Jiménez-Valverde & Hortal, 2003). Completeness values measured by different estimators can provide different richness estimations that in turn depend on the sampling effort accomplished (see Chiarucci *et al.*, 2001; Hortal *et al.*, 2006). Thus, our selection of well-surveyed cells is not free of error being a compromise among the probability to choose them correctly and the number of cells able to be analysed.

Physioclimatic and biogeographical subregions

We assessed the proportion of well-surveyed 50 × 50 km cells (≥ 70% completeness) for each one of six previously delimited physicoclimatic subregions of the Ibero-Balearic area (Lobo & Martín-Piera, 2002). We also examined the degree of completeness for biogeographical subregions defined by Ribera (2000) using compositional information on Iberian water beetles, adding the Balearic Islands as a new subregion. The first regionalization allows us to consider the classic territories with different macroenvironmental characteristics, while the second one has the capacity to reflect the differences due to causal variables difficult to quantify as those due to dispersal limitation or historical factors. For each subregion (both physicoclimatic and biogeographical) we computed the species richness and the associated number of data base records.

Variables and sampling bias

We considered a total of 26 variables that could potentially explain the distribution of the sampling effort, divided into four categories: environmental, land use, spatial, and variables related to the 'attractiveness' of the sites. The environmental variables included nine climatic (minimum and maximum monthly mean temperature, mean annual temperature, total annual rainfall, summer precipitation, number of days of sun per year, aridity, annual range of temperature variation, annual precipitation variation); four topographical (minimum, maximum and mean elevation, elevation range), and four lithological (percentage of area with clay, calcareous, and siliceous substrates, lithological diversity). The four land-use variables, selected to represent the degree of human disturbance, measure the coverage of the four human-induced landscapes, which are most common in the study area: urban and industrial areas, non-irrigated croplands, irrigated croplands, and anthropic pasturelands. We also included three variables (hereafter called 'attractiveness' variables) related to the accessibility and appeal for researchers: distance from research centres (minimum distance from the central point of each square to the nearest main centre of research on water beetles in the study area, i.e. Barcelona, León, and Murcia), number of type localities (where species new to science have been found), and coverage of protected land surface. Lastly, the central latitude and longitude of each cell were also included as spatial predictor variables.

Climate data (original resolution 1 km) are courtesy of the Spanish Instituto Nacional de Meteorología and the Portuguese Instituto de Meteorologia. Topographical variables were obtained from a Digital Elevation Model (Clark University, 2000), and the land-use data (original resolution 280 m) were provided by the European Environment Agency (EEA, 2000). Data on underlying geology were extracted from 1 : 200,000 scale geological maps (IGN, 1995); these were first digitized, and then superimposed on cells through the geographical information system IDRISI (Clark University, 2003). Lithological diversity was estimated for each grid cell by applying the Shannon diversity index to the primary lithological variables.

Raw number of data base records and cell completeness values of well-surveyed cells was regressed against these 26 explanatory variables, and the importance of each subgroup of variables was assessed by using Generalized Linear Models (GLM: McCullagh & Nelder, 1989; Crawley, 1993). All variables were standardized to mean = 0 and standard deviation = 1 to eliminate the effects of differences in measurement scale. We assumed a Poisson error distribution for the dependent variables, related to the set of predictor variables via a logarithmic link function. To evaluate potentially curvilinear relationships, the dependent variable was first related separately to either a linear, a quadratic, or a cubic function of each variable (Austin, 1980). Subsequently, a stepwise procedure was used to enter the variables into the model (Nicholls, 1989; Lobo & Martín-Piera, 2002). First, the linear, quadratic or cubic function of the variable that accounted for the most

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

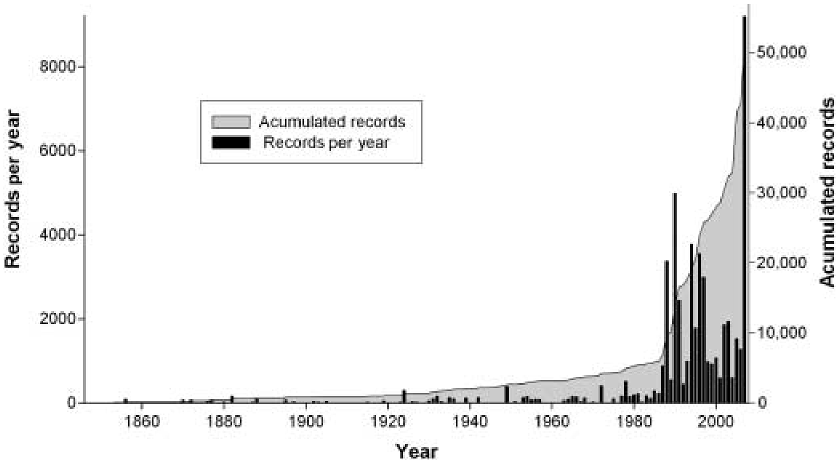

Figure 2 Historical variation in the number of data base records and accumulated number of records of Iberian water beetles.

important change in deviance was entered. The remaining variables were added to the model sequentially according to their estimated explanatory capacity. The procedure was repeated iteratively until no more statistically significant explanatory variables remained ($P = 0.05$). At each step, the significance of the terms already selected was tested by submitting the new model to a backward selection procedure. The terms that became non-significant in this step were then excluded. After examining the individual contribution of each explanatory variable, four main complete models were constructed separately for each of the four types of explanatory variables: environmental (climatic and topographical variables), land use, 'attractiveness', and spatial. Spatial variables were included as the third-degree polynomial equation of the central latitude and longitude (Trend Surface Analysis – see Legendre, 1993) in order to incorporate the influence of spatial structures arising from the effects of other historical, biotic, or environmental variables not otherwise taken into account (Legendre & Legendre, 1998). A backward stepwise regression with the nine terms of the equation as predictor variables was performed to remove non-significant spatial terms. The STATISTICA package 6.1. (StatSoft, 2004) was used for all computations.

We measured the relative importance of each type of explanatory variable using a hierarchical partitioning procedure (MacNally, 2000). First, we calculated the percentage of explained deviance for each type of variable, as well as the variability explained by all possible variable combinations in which this type of variable participates. Subsequently, we calculated the average effect of inclusion of each type of variable in all models for which this type of variable was relevant. We took such averages as estimations of the independent contribution of each type of explanatory variable. A Mann–Whitney *U*-test was used to identify the variables that differ significantly between considered well-surveyed and not well-surveyed cells.

RESULTS

Analysis of the data base

The mean value of records per 50 × 50 km cell was 197 and that of species 48 (Table 1). Most data base records (86%) published after 1987 (Fig. 2). The family Dytiscidae shows the highest number of species and records, followed by Hydraenidae and Hydrophilidae (see Table 1). Species richness scores and number of data base records were highly correlated (Spearman rank coefficient $r_s = 0.95$; $n = 257$; $P = 0.001$), showing that the observed species richness in each square depends on the sampling effort.

Geographical variation of completeness

The geographical patterns of the number of data base records and completeness are quite similar (Fig. 3); cells with higher sampling effort and completeness seem to be widespread in the Iberian Peninsula, while less surveyed cells occur mainly in central Spain (with the exception of the Sierra de Guadarrama and Sierra de Gredos) and south-central Portugal (see Figs 1 and 3). The mean value of completeness by cells was around 46% (mean ± SD; 45.4 ± 27.5). From a total of 257 cells, 56 had completeness values higher than 70% (considered as well surveyed; Fig. 3), 26 of them reached scores of 80% or more, and only three reached scores above 90%.

There are well-surveyed cells across the whole Iberian territory, although they are not evenly distributed among both biogeographical (chi-square test: $\chi^2 = 16.88$; $P = 0.005$; d.f = 5) or physioclimatic subregions ($\chi^2 = 12.37$; $P = 0.05$; d.f = 6) (Fig. 4). The percentages of considered well-surveyed cells in subregions oscillate from 10% to 50%, the best surveyed subregions being those located in the Balearic Islands, some mountain areas, northern areas close to the Cantabrian sea and the south-eastern subregion (see Table 2). The remaining subregions contained a roughly similar low proportion of well-surveyed cells.

Variables and sampling bias

The number of type localities, distance from main research centres, altitudinal range, and maximum altitude were the variables that accounted for the highest variability in the number of data base records (Table 3). Attractiveness variables seemed to be the most influential; a complete model including these three significant variables explained almost 50% of the total variability. The variation in the number of data base records was also highly linked

© 2008 The Authors

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

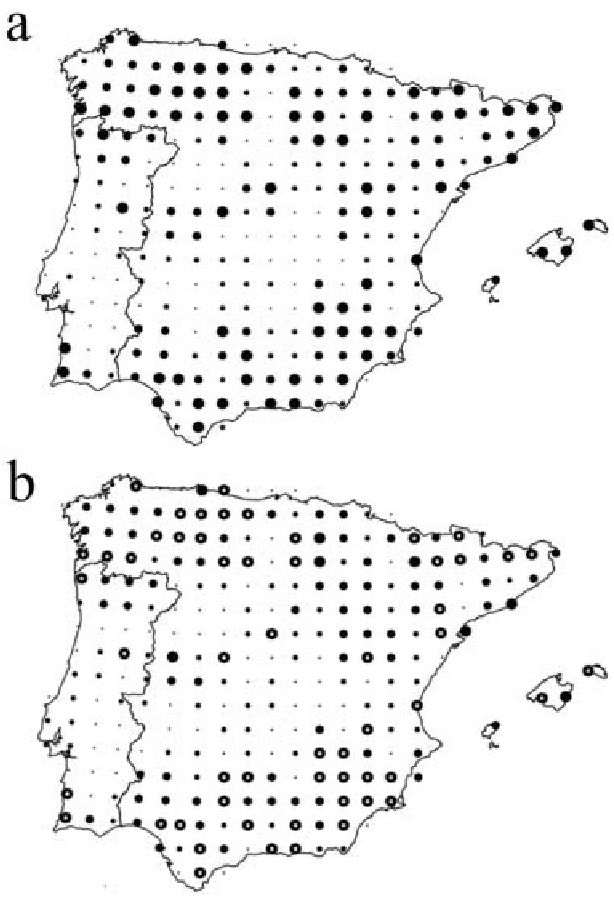

Figure 3 (a) Raw number of data base records per cells, and (b) sampling completeness. Sampling completeness was estimated as the proportion of species actually recorded, to the number of species predicted by the accumulation curve adjusted by the Clench function (the asymptotic score of the relationships between the accumulated number of species and the increase in the number of data base records). The varying diameter of symbols is proportional to sampling intensity on a scale of four categories (quartiles) in each range of values. White dots indicate those well-surveyed cells where the proportion (number of species recorded/number of species predicted) equals or exceeds 70%.

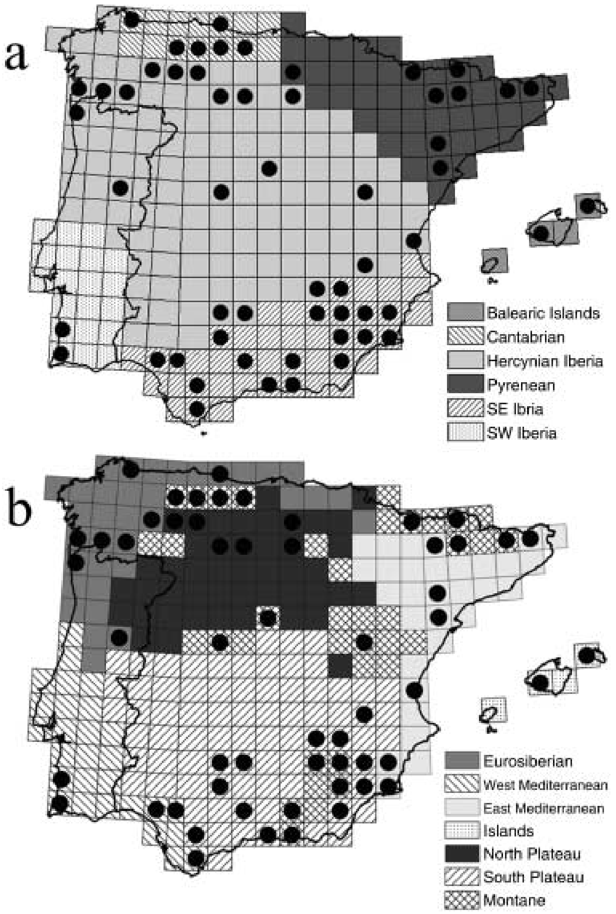

Figure 4 Distribution of well-surveyed 50 × 50 km Universal Transverse Mercator cells within (a) the biogeographical subregions defined by Ribera (2000) and (b) the physicoclimatic subregions defined by Lobo & Martín-Piera (2002).

to environmental variables (around 39% of total variability), while spatial and land-use variables were less relevant. A stepwise regression model with all the significant variables accounted for almost 56% of the total variability in the number of data base records (Table 3).

The results of the hierarchical partitioning demonstrated that the attractiveness variables had the highest average effect after inclusion in all models (around 23.5%), followed by the averaged effect of environmental variables (14.6%), spatial variables (8.3%), and land-cover variables (5.3%). However, none of these variables are statistically significant when regressed against the completeness values in the previously considered well-surveyed cells ($n = 56$).

Well-surveyed cells significantly differed from the rest in a number of variables: they had a higher number of type localities, wider altitude range, larger protected surface, higher maximum altitude, and higher annual and summer precipitation. They were also closer to the main research centres, had less surface of non-irrigated crops, and a lower maximum mean temperature and aridity index (Table 4).

DISCUSSION

How complete is the water beetle data base for the Iberian Peninsula and Balearic Islands?

To date, approximately half of the territory remains characterized by a remarkable scarcity of water beetles records (with < 50% of the predicted species recorded). The data base ESACIB is less complete than other comparable data bases of Iberian insects, probably due to the larger number of species, which are in general small and inconspicuous, with a limited appeal for amateur entomologists. Thus, in a data base for butterflies (226 species), more than 68% of squares had completeness values of > 75%, and a third of the territory reached scores of 90% or more (Romo *et al.*, 2006). For dung beetles (56 species), although the survey

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Number (N) of 50 × 50 km Universal Transverse Mercator cells, species (S), considered well-surveyed cells (WSC), data base records (DR), and mean number (MN) of species and data base records for each of the biogeographical and physicoclimatic Iberian subregions (see Ribera, 2000 and Lobo & Martín-Piera, 2002).

	N of cells	*N* of species	N-WSC	N-DR	N-DR in WSC
Biogeographical subregions					
Hercynian Iberia	132	403	19	17,229	8917
Cantabrian	14	221	6	4256	3783
Pyrenean	44	352	9	10,915	6776
Southeastern Iberia	43	341	18	15,435	12,489
Southwestern Iberia	20	150	2	988	625
Balearic Islands	4	145	2	1941	1491
Physioclimatic subregions					
Eurosiberian	35	256	8	5709	3655
West Mediterranean	28	228	5	3412	2141
East Mediterranean	31	309	7	8791	5810
Islands	4	145	2	1941	1491
North Plateau	46	305	6	6316	3257
South Plateau	76	345	12	10,991	7462
Montane	37	377	16	13,604	10,265
Total	257	511	56	50,764	34,081

effort was low (15,740 records), 33% of the Iberian Peninsula could be considered well surveyed (Lobo & Martín-Piera, 2002). Our results highlight the lack of complete and extensive inventory data for aquatic taxa (Lévêque *et al.*, 2005), as water beetles could be probably be considered one of the best studied groups of freshwater invertebrates in the region. Despite the general incompleteness of the data, a quarter of the study area can already be considered well prospected (completeness values > 70%).

Bias in the sampling of freshwater biodiversity

Even when a number of well-surveyed areas have been identified, unevenness in sampling effort may result in a partial (and biased) description of biodiversity variation (Dennis, 2001). In our case, well-surveyed cells are not evenly distributed across biogeographical or physicoclimatic subregions. Furthermore, the proportion of considered well-surveyed cells was also higher on islands and in montane areas (principally in Cantabria and southeastern Iberia), and lower in central Spain (principally both plateaus) and southwestern Iberia (Fig. 4).

As in other studies (Dennis & Thomas, 2000; Romo *et al.*, 2006), our results demonstrate that the geographical variation in sampling effort is mainly related to attractiveness and environmental variables, and that these same variables significantly differ among well- and not enough surveyed cells. Thus, although the sampling effort carried out within those cells considered as well-surveyed does not seem to be biased, simple factors affecting the activity of collectors, such as perceived attractiveness of landscapes and accessibility of sampling sites would have deeply influenced the apparent observed species richness pattern. Two coexisting trends appear to occur: researchers tend to sample more intensely in accessible sites near their research centres, and select the study sites based on the presence of interesting species, and/or mountainous and protected areas.

Some of the sites located in mountainous areas with the highest survey effort were also the areas identified by Ribera (2000) as those with the highest conservation value for Iberian aquatic Coleoptera. These include medium altitude streams and freshwater lagoon of the pre-Pyrenees; Sierra de Alacaraz in southeastern Spain; streams in the Parque Natural de los Alcornocales (South Spain); Sierra de la Demanda (northeastern Spain, in Hercynian Iberia); eastern part of the provinces of Lugo and Orense (northwestern Spain); Serra da Estrela (northern Portugal); Sierra de Guadarrama (central Spain); some national parks such as Picos de Europa in the Cantabrian mountains, or Sierra Nevada in the south-east; and some saline streams in southeastern Spain. The future use of distributional predictive models (see Hortal *et al.*, 2001; Lobo & Martín-Piera, 2002) can help us to assess if the higher comparative species richness of more surveyed territories is real or some not enough surveyed cells become species-rich.

Where to sample next?

Results show that additional surveys of islands or mountainous areas would only further increase the unbalanced distribution of well-sampled areas across the region. Future surveys should be concentrated in currently recognized undersampled subregions. As discussed above, well-sampled and undersampled cells differ significantly in a number of variables, and these differences should be borne in mind to diminish environmental bias in future surveys (Funk *et al.*, 2005; Hortal & Lobo, 2005). To reduce further bias in collecting effort and improve both spatial and environmental coverage, further work should be concentrated in grid cells with completeness values close to 70%, located in

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3 Individual explanatory capacity of each one of the considered variables on the sampling effort variation in the 257 50 × 50 km Universal Transverse Mercator cells (measured as the raw number of data base records), and explanatory capacity of the obtained models with all the considered variables of the same type (environmental, land-use, spatial, and attractiveness variables). Only those variables with a statistically significant effect ($P < 0.05$) and a percentage of deviance > 1.0 are represented.

Explanatory variables	*Dev*	*% Dev*	Function
Environmental variables			
Altitude range	68,866.8	18.67	Cubic
Maximum altitude	70,741.0	16.46	Cubic
Mean altitude	78,136.2	7.72	Cubic
Summer precipitation	78,803.4	6.94	Cubic
Annual mean hours of sun	78,954.2	6.76	Cubic
Lithologic diversity	80,584.5	4.83	Cubic
Annual mean temperature	81,017.1	4.32	Cubic
Clay soils	81,151.6	4.16	Cubic
Aridity index	81,552.0	3.69	Cubic
Minimum mean temperature	81,617.2	3.61	Cubic
Siliceous soils	81,850.2	3.34	Cubic
Annual mean precipitation	82,368.3	2.73	Cubic
Maximum mean temperature	82,589.2	2.46	Cubic
Calcareous soils	82,672.7	2.37	Cubic
Temperature range	82,993.1	1.99	Cubic
Land-use variables			
Anthropic pasturelands	77,494.4	8.48	Cubic
Non-irrigated crops	81,628.5	3.60	Cubic
Irrigated crops	83,481.3	1.41	Cubic
Attractiveness variables			
Number of type localities	57,157.1	32.50	Cubic
Distance from research centres	69,556.6	17.86	Cubic
Protected surface	77,756.5	8.17	Cubic
Full environmental model	51,463.4	39.22	
Full land-use model	70,408.4	16.85	
Full spatial model	68,354.3	18.22	
Full attractiveness model	42,719.0	49.55	

Dev: Deviance of each variable;% *Dev*: percentage of explained deviance on total variability in the number of data base records.

central Spain, specially those on the north and south plateaus (Castilla León and Castilla La Mancha) and in south Portugal (Beja, Setubal, Portalegre, and Santarém). This will increase the number of well-sampled squares in areas currently undersampled, allowing the possibility of less biased analyses. Efforts should focus on water bodies on arid, low-elevation areas with a temperate Mediterranean climate and a high coverage of non-irrigated crops, which are in general far from research centres. The scarcity of water bodies in these areas, together with their low perceived attractiveness, could explain the low sampling effort invested on them to date.

Concluding remarks

We emphasize the importance of the evaluation of data quality as a preliminary step in biodiversity studies (see Reddy & Dávalos, 2003), assessing the degree of geographical coverage of existing faunistic data, and the amount and nature of any bias in its collection. Our results demonstrate that, as happens with other taxa and territories (Dennis *et al.*, 1999; Dennis & Thomas, 2000; Zaniewski *et al.*, 2002; Reutter *et al.*, 2003; Graham *et al.*, 2004; Soberón & Peterson, 2004; Mart'nez-Meyer, 2005; Romo *et al.*, 2006; Hortal *et al.*, 2007; Lobo *et al.*, 2007; Soberón *et al.*, 2007), the available distributional information on Iberian water beetles is the result of a biased collection process influenced by sociological and environmental factors. More survey effort needs to be carried out to obtain a detailed and reliable representation of the diversity of Iberian water beetles. Despite this, our results reveal that all regions contain well-surveyed cells at the coarse resolution considered here, and both the number and the lack of bias in the completeness percentages of these well-surveyed cells facilitate their use in biogeographical and conservation studies. In this sense, our results provide a basis for the design of efficient future field campaigns, since they allow the identification of genuinely undersampled regions. The identification of well-surveyed cells is invaluable in modelling species' distributions, since taxa not recorded from these are likely to be genuinely absent (see Lobo, 2008). The discrimination of those localities with good quality distributional information, together with the use of these modelling techniques, will allow us to obtain a better picture of the distribution of water beetle diversity in the Iberian Peninsula in the future.

Table 4 Used explanatory variables with significant differences among well-surveyed cells (WSC) and remaining squares (RS) by using a Mann–Whitney *U*-test ($n_1 = 56$; $n_2 = 201$). The last two columns represent if the median score of each one of these variables is higher (+) or lower (–) for each one of the two groups of cells.

	U	*P*	WSC	RS
Number of type localities	3002	< 0.00001	+	–
Altitude range	3350	< 0.00001	+	–
Distance from research centres	3670	0.00007	–	+
Protected surface	3806	0.0002	+	–
Maximum altitude	3740	0.0001	+	–
Annual mean precipitation	4385	0.01	+	–
Summer precipitation	4400	0.01	+	–
Non-irrigated crops	4422	0.01	–	+
Maximum mean temperature	4501	0.02	–	+
Aridity index	4532	0.03	–	+

ACKNOWLEDGEMENTS

This work was partially supported by projects CGL2006-04159 and CGL2004-00028 of the Spanish MEC. We thank P. Aguilera, R. Angus, D.T. Bilton, A. Castro, H. Fery, J. Fresneda, G. Foster, J. Garrido, C. Hernando, and L.F. Valladares for unpublished records of Iberian water beetles. We also thank the Ecología Acuática group (University of Murcia) their help in the compilation of the data and D.T. Bilton and J. Velasco for multiple comments on previous stages of the manuscript.

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

REFERENCES

Abellán, P., Sánchez-Fernández, D., Velasco, J. & Millán, A. (2005) Conservation of freshwater biodiversity: a comparison of different area selection methods. *Biodiversity and Conservation*, **14**, 3457–3474.

Allan, J.D. & Flecker, A.S. (1993) Biodiversity conservation in running waters. *Bioscience*, **43**, 32–43.

Austin, M.P. (1980) Searching for a model for use in vegetation analysis. *Vegetatio*, **42**, 11–21.

Chiarucci, A., Maccherini, S. & De Dominicis, V. (2001) Evaluation and monitoring of the flora in a nature reserve by estimation methods. *Biological Conservation*, **101**, 305–314.

Clark University (2000) *Global change data archive, Vol. 3. 1 km*. Global Elevation Model, Worcester, MA, USA.

Clark University (2003) *Idrisi Kilimanjaro*. GIS software package. Worcester, MA, USA.

Colwell, R.K. (2000) *EstimateS: statistical estimation of species richness and shared species from samples*. Software and user's guide, version 6.0b1. Available at http://viceroy.eeb.uconn.edu/estimates.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, **354**, 101–118.

Crawley, M.J. (1993) *GLM for ecologists*. Blackwell Science, Oxford, UK.

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. & Shreeve, T.G. (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. & 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., Sparks, T.H. & Hardy, P.B. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation.*, **3**, 33–42.

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 butterfly recording and monitoring. *Biological Conservation*, **128**, 486–492.

EEA (2000) *NATLAN. Nature/land cover information package*. European Environment Agency, Luxembourg.

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

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

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.

Griffiths, G.H., Eversham, B.C. & Roy, D.B. (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. & Lobo, J.M. (2005) An ED-based protocol for the optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913–2947.

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

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

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

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

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

IGN (1995) *Atlas Nacional de España*, 16. Instituto Geográfico Nacional, Madrid, Spain.

Jiménez-Valverde, A. & Hortal, J. (2003) Las curvas de acumulación de especies y la necesidad de evaluar la calidad de los inventarios biológicos. *Revista Ibérica de Aracnología*, **8**, 151–161.

Lawton, J.H., Prendergast, J.R. & Eversham, B.C. (1994) The numbers and spatial distributions of species: analyses of British data. *Systematics and conservation evaluation* (ed. by P.L. Forey, C.J. Humphries and R.I. Vane-Wright), pp. 177–195. Clarendon Press, Oxford, UK.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edn. Elsevier, Amsterdam.

Lévêque, C., Balian, E.V. & Martens, K. (2005) An assessment of animal species diversity in continental waters. *Hydrobiologia*, **542**, 39–67.

Lobo, J.M. (2008) More complex distribution models or more representative data? *Biodiversity Informatics*, **5**, 15–19.

Lobo, J.M. & Martín-Piera, F. (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., 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.

MacNally, R. (2000) Regression and model-building in conservation biology biogeography and ecology: the distinction between and reconciliation of 'predictive' and 'explanatory' models. *Biodiversity and Conservation*, **9**, 655–587.

Martínez-Meyer, E. (2005) Climate change and biodiversity:

© 2008 The Authors

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

some considerations in forecasting shifts in species potential distributions. *Biodiversity Informatics*, **2**, 42–55.

Master, L.L., Flack, S.R. & Stein, B.A. (1998) *Rivers of life: critical watersheds for protecting freshwater biodiversity*. The Nature Conservancy, Arlington, VA, USA.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*. Chapman & Hall, London.

Millán, A., Moreno, J.L. & Velasco, J. (2002) *Estudio faunístico y ecológico de los coleópteros y heterópteros acuáticos y semiacuáticos de la provincia de albacete*. Instituto de Estudios Albacetenses, Albacete, Spain.

Millán, A., Abellán, P., Ribera, I., Sánchez, D. & Velasco, J. (2006) The Hydradephaga (Coleoptera) of the Segura basin (SE Spain): twenty five years studying water beetles. *Memorie Della Società Entomologica Italiana*, **85**, 137–158.

Mittermeier, R.A., Myers, N., Thomsen, J.B., da Fonseca, G.A.B. & Olivieri, S. (1998) Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. *Conservation Biology*, **12**, 516–520.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A.B. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

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

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

Quézel, P. (1995) La flore du basin mediterranéen en: origine, mise en place, endemisme. *Ecologia Mediterranea*, **21**, 19–39.

Ramos, M.A., Lobo, J.M. & Esteban, M. (2001) Ten years inventorying the Iberian fauna: results and perspectives. *Biodiversity and Conservation*, **10**, 19–28.

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.

Rey-Benayas, J.M. & Scheiner, S.M. (2002) Plant diversity, biogeography and environment in Iberia: patterns and possible causal factors. *Journal of Vegetation Science*, **13**, 245–258.

Ribera, I. (2000) Biogeography and conservation of Iberian water beetles. *Biological Conservation*, **92**, 131–150.

Ribera, I., Hernando, C. & Aguilera, P. (1998) An annotated checklist of the Iberian water beetles (Coleoptera). *Zapateri*, **8**, 43–111.

Ricciardi, A. & Rasmussen, J.B. (1999) Extinction rates in North American freshwater fauna. *Conservation Biology*, **13**, 1220–1222.

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

Sánchez-Fernández, D., Abellán, P., Velasco, J. & Millán, A. (2004) Selecting areas to protect the biodiversity of aquatic ecosystems in a semiarid Mediterranean region. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **14**, 465–479.

Sánchez-Fernández, D., Abellán, P., Mellado, A., Velasco, J. & Millán, A. (2006) Are water beetles good indicators of biodiversity in Mediterranean aquatic ecosystems? The case of the Segura river basin (SE Spain). *Biodiversity and Conservation*, **15**, 4507–4520.

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

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

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

StatSoft (2004) *STATISTICA (data analysis software system)*, Version 6. www.statsoft.com.

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.

Editor: Anthony Ricciardi

14724642, 2008, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00474.x by Universidad De Murcia, Wiley Online Library on [01/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License