ECOGRAPHY 29: 873–885, 2006

Identifying recorder-induced geographic bias in an Iberian butterfly database

Helena Romo, Enrique García-Barros and Jorge M. Lobo

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

A database with comprehensive butterfly faunistic information from the Iberian Peninsula and the Balearic Islands was used to estimate inventory completeness as well as the environmental, spatial, and land-use effects on sampling intensities, on a 50×50 km UTM grid. The degree of sampling effort was assessed by means of accumulation curves based on the Clench function. Using the General Linear Model regression procedure, the effects of 22 variables on the estimated sampling efforts were assessed. This combination of methods is proposed as a preliminary step in biodiversity studies, in order to evaluate not only the degree of geographic coverage of existing faunistic data, but also the amount and nature of the bias on the faunistic work done throughout the last two centuries. The degree of spatial effects on the data was greater than the effects of environmental or land-use variables, although the latter two proved to be locally relevant. The results confirm previous findings that collecting is often skewed by relatively simple factors that affect collector activity, such as accessibility and attractiveness of sampling sites. With regard to Iberian and Balearic butterflies, adequate inventories on the scale investigated may probably suffice for further studies of the diversity of this insect group. Additionally, the results enabled us to develop general guide lines for the design of further faunistic work in the area.

H. Romo (helena.romo@uam.es) and E. García-Barros, Dept de Biología (Zoología), Univ. Autónoma de Madrid, Carretera de Colmenar km. 15, ES-28049 Madrid, Spain. – J. M. Lobo, Museo Nacional de Ciencias Naturales, Cl José Gutiérrez Abascal, 2, ES-28006 Madrid, Spain.

Understanding the nature and intensity of the causes that have led to present distributions is fundamental to biodiversity conservation programs, as well as biogeographic research. In theory, the patterns that underlie the distribution of organisms can be spotted reliably only when the data are ecologically and geographically balanced (Ferrier 2002, Brooks et al. 2004). However, not only is our knowledge of the geographic distribution of living beings incomplete (Gaston and Spicer 2004, and references therein); there is a growing evidence that the basis of such knowledge is itself skewed by cultural, socio-economic and policy constraints on the activity of natural historians, such as accessibility of the sampling sites (Nelson et al. 1990, Peterson et al. 1998, Parnell et al. 2003), their degree of protection (Reddy and Dávalos 2003), the distance from the place of residence of biologists (Freitag et al. 1998, Martín and Gurrea 1999, Dennis et al. 1999, Dennis and Thomas 2000), or even the degree of attractiveness and physical appearance of the species (Dennis et al. 2006). These and similar constraints can be expected to lead to unbalanced taxonomic and geographic coverage (Kress et al. 1998, Dennis and Hardy 1999, Dennis et al. 1999, Soberón et al. 2000, Parnell et al. 2003, Whittaker et al. 2005).

More often than not, the data compiled from heterogeneous sources (e.g. published reports combined with label data from public and private collections of

Accepted 1 September 2006

ISSN 0906-7590

incompletely-surveyed organisms, with insufficient sampling effort information) can be expected to suffer from the above-mentioned problem. Measurement of the extent of this imbalance might help to clarify how regional biodiversity surveys have "evolved", and to correct for bias which has generally been overlooked. This would be especially helpful in biodiversity "hotspots", where biological surveys are still in progress. Furthermore, the magnitude of the imbalance is likely to be correlated with the diversity left to be recorded. Any thorough analysis of geographic biodiversity patterns should be preceded by an effort to estimate the amount and nature of bias in the data. This research program can help to distinguish sites that appear to be reasonably well-surveyed from those that are not, to spot areas that require further sampling (Lobo et al. 1997, Lobo and Martín-Piera 2002, Parnell et al. 2003). In this context, estimation of the degree to which the data represent complete inventories on a given scale is a required first step (Petersen and Meier 2003, Magurran 2004), and can be attempted using a variety of statistical techniques (Soberón and Llorente 1993, Colwell and Coddington 1994, Gotelli and Colwell 2001, Rosenzweig et al. 2003, Koellner et al. 2004). Based on the ideas detailed above, this study aims to identify present geographic bias in the patterns of data-collection of a sample taxon, following four steps: 1) identify the area units most representative of fauna diversity; 2) check whether such units cover efficiently a reasonable range of the environmental gradients across the study area; 3) assess the degree of geographic, environmental or spatial bias on sampling efforts; 4) and, to the extent that the major part of the bias could be controlled statistically, identify cells of interest for future sampling programs.

The butterflies (Lepidoptera: Papilionoidea and Hesperioidea) of the Iberian Peninsula (continental land masses of Spain and Portugal) and the Balearic Islands were selected as the target taxon because: their geographic distribution in Europe is well known in broad terms (Kudrna 2002); they are moderately species-rich in the study area (ca 230 species: Tolman and Lewington 1997, García-Barros et al. 2004); and they may be valuable environmental indicators due to their specialized larval diets (New 1991). However, although the faunistic information from Spain and Portugal gathered over more than two centuries is of a substantial amount, it is unfortunately of a rather heterogeneous nature (as described below), and far from exhaustive on a fine-grained geographic scale (e.g. 1×1 km or 10×10 km grids: García-Barros and Munguira 1999, Garcia-Pereira et al. 1999, Romo and García-Barros 2005). While some assessments of the origins of geographic diversity patterns of Iberian butterflies have been made (Martín and Gurrea 1990), recently authors have speculated that these patterns were skewed during data collection: specifically, sampling efforts appear to be concentrated on areas appreciated by collectors as especially rewarding (García-Barros et al. 2000, 2004). This speculation, as well as the precise nature of the potential bias, has only partially been tested (Hortal et al. 2004, Romo and García-Barros 2005).

Methods

Butterfly distribution data and geographic units

An exhaustive Iberian butterfly database was compiled from a recent atlas by García-Barros et al. (2004), containing over 289 000 records with associated location data for 226 species. The sources include labelled collection specimens as well as quantitative or qualitative counts from any available source (details in García-Barros et al. 2004). In spite of the important amount of information collected, the data are skewed in several ways. This imposed limitations to the aims of the study. First, although the data base was originally referenced to 10×10 km square cells, preliminary estimates have demonstrated that the geographic coverage on this scale is poor (at most 17% of the total area), while a grid size of 50×50 km would cover a reasonably representative proportion of the Iberian and Balearic territories (García-Barros and Munguira 1999, Garcia-Pereira et al. 1999, Garcia-Pereira 2003). This determined the size of the geographic units selected for this study, i.e. cells of 50×50 km based on the UTM (Universal Transverse Mercator) projection. Grid cells containing $<15\%$ land surface were eliminated, which resulted in a total of 257 geographic units. Second, although the data covered a period of more than two centuries (from 1784 to 2003), the amount of information has grown almost exponentially through this period. As a consequence, a vast majority of the records has a relatively

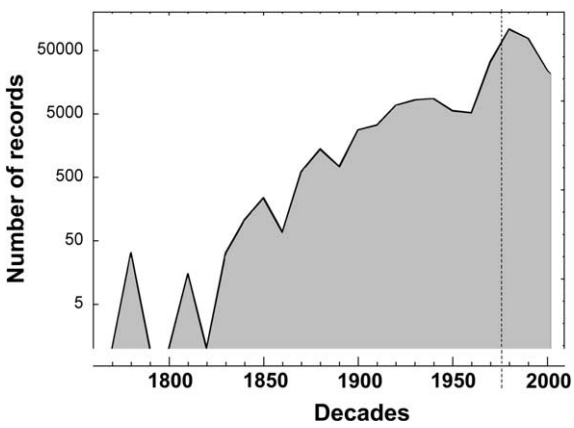


Fig. 1. Historical distribution of the database records (number of records per decade). The vertical dashed line indicates the average year (1978 ± 23.6). Note logarithmic scale in the Y-axis.

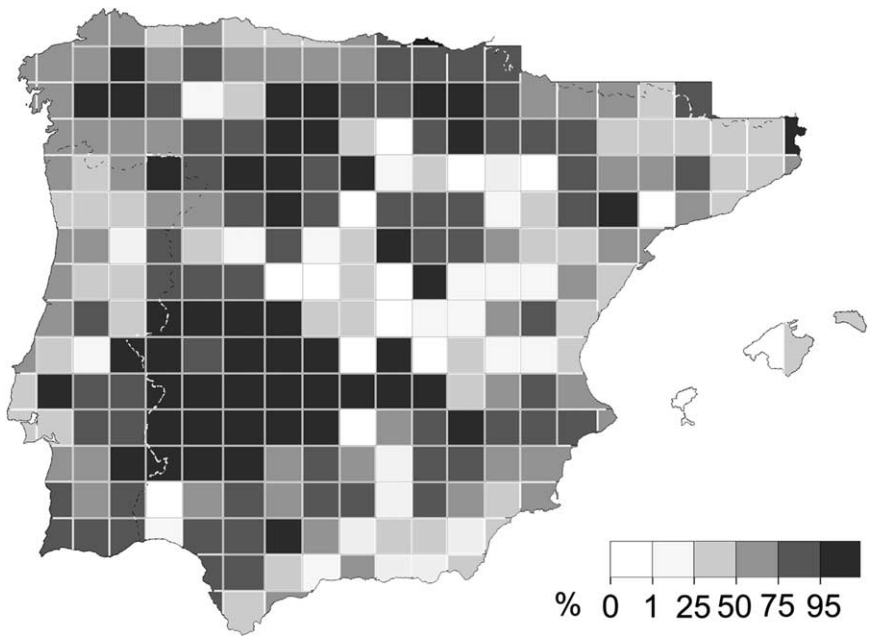


Fig. 2. Geographic distribution of the proportion of recent data. The percentage of database records dating after 1978 is represented by progressively darker grey tones.

recent origin (mean date $\pm$ SD = 1978 $\pm$ 23.6; Fig. 1). This precluded us from determining precise historical biases, to concentrate on the overall (pooled) effects.

The geographic distribution of the proportion of recent data is shown in Fig. 2

Assessing the completeness of local inventories

To determine the degree of completeness of the cell inventories, the number of database records was used as the sampling effort surrogate, and the Clench function (Soberón and Llorente 1993) was applied to estimate the cell-specific mean rate of species addition per record. The use of the raw number of database records was empirically supported by previous tests based on the same data, which demonstrated that this estimate was more strongly correlated with the actual species accumulation rates than several other candidate surrogates (namely: number of individuals, sites, publications or documental sources, dates, or combinations of site-date-collector: Romo and García-Barros 2005, see also Hortal et al. 2006).

Even though more time-consuming than other available methods, the Clench function was preferred to both parametric- and non-parametric estimators (Soberón and Llorente 1993, Colwell and Coddington 1994, Peterson and Slade 1998, Colwell 2000, Petersen and Meier 2003). This accumulation function may be best suited for data from varied sources (dating from long periods of time; Soberón and Llorente 1993, Soberón et al. 2000), does not require the probability of species-collection to be constant over time (Burnham and Overton 1979), does not depend strongly on the number of rare species or clumped distribution patterns (Petersen and Meier 2003, Petersen et al. 2003). Furthermore, it has been successfully tested for the same purpose with a very similar data set from Portugal (Hortal et al. 2004). A brief comparison of the estimated species richness obtained with this method with those derived from three popular non-parametric estimators (Jackknife 1 and 2, Chao 2: Burnham and Overton 1979, Heltshe and Forrester 1983, Chao 1984, 1987, Smith and van Belle 1984, Colwell 2000) is given at the beginning of the results section.

The Clench function is expressed by $S_n = a*n/(1 + b*n)$, where "S_n" is the cumulative number of species discovered, "a" represents the rate of increase in the number of species at the beginning of the inventory, "b" is a parameter related with the shape of the curve, and "n" is the sample effort.

This equation measures the mean increment in species richness relative to the accumulated number of records. The data were smoothed using 100 random replicates (obtained with the package EstimateS: Colwell 2000).

Table 1. Pearson correlation coefficients between different estimates of species richness for the 50×50 km UTM squares under study, as measured by three non-parametric estimators (Chao2, Jackknife1, Jackknife2) and an asymptotic estimator (Clench). In all instances n = 247 and $p < 0.0001$.

	Chao2	Jackknife1	Jackknife2
Clench	0.903	0.853	0.881
Chao2		0.923	0.953
Jackknife1			0.983

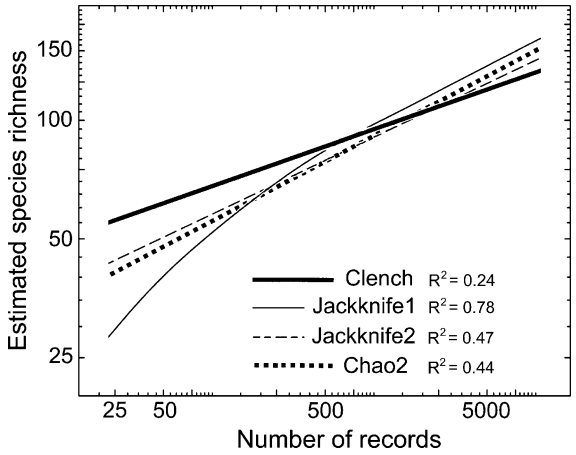


Fig. 3. Comparison of the relationship between four different estimates of species richness (obtained using four different estimators) and the number of database records. The data points have been omitted for simplicity. The coefficients of determination given (R^2) are from exponential functions fitted to the estimators values relative to the number of records except for Jackknife1 values, which fitted slightly better to a logarithmic curve ($R^2 = 0.78$) than to an exponential one ($R^2 = 0.77$). All R^2 values are significant at $p < 0.0001$ ($n = 247$).

The Clench function was then fitted to the smoothed data, and the asymptotic value (i.e. the species richness predicted for an ideally unlimited sample size) was recorded. The ratio of recorded to predicted species numbers (the asymptotic score) was used as the measure of completeness of the inventory (the completeness ratio of Soberón et al. 2000). Then, a UTM cell was considered to be adequately surveyed when the observed species-richness associated with it was equal to, or >90% of the predicted score (details in the results section).

Eco-physiographic regions

To estimate the degree to which the selection of cells would cover the most evident environmental gradients across the study area, the proportion of well-surveyed 50×50 km cells in each eco-physiographic region was assessed. For this assessment, recent classification by Lobo and Martín-Piera (2002) was adopted. It is based on a 50×50 km grid and on hierarchical classification methods of the distribution of the most relevant physiographic variables across the study area, an independent and objective, but manageable, synthesis. The proportion of well-surveyed cells within each subregion, their associated species richness, and the associated number of database records, were determined.

Environmental and spatial variables

Cell completeness values and the raw number of database records were regressed on 22 continuous, primarily environmental (non-human induced), land use (human-induced), or spatial variables. The environmental variables included four topographic (minimum, maximum and mean elevation, plus elevation range); four lithology (percentage of area with clay, calcareous, and siliceous substrates, plus lithologic diversity); and eight climate variables (minimum and maximum monthly mean temperature, mean annual temperature, total annual rainfall, summer precipitation, number of days of sun per year, annual range of temperature variation, and annual precipitation variation). The land-use variables, selected to represent the degree of human disturbance, measure the coverage of the four human-induced landscapes which are presently most

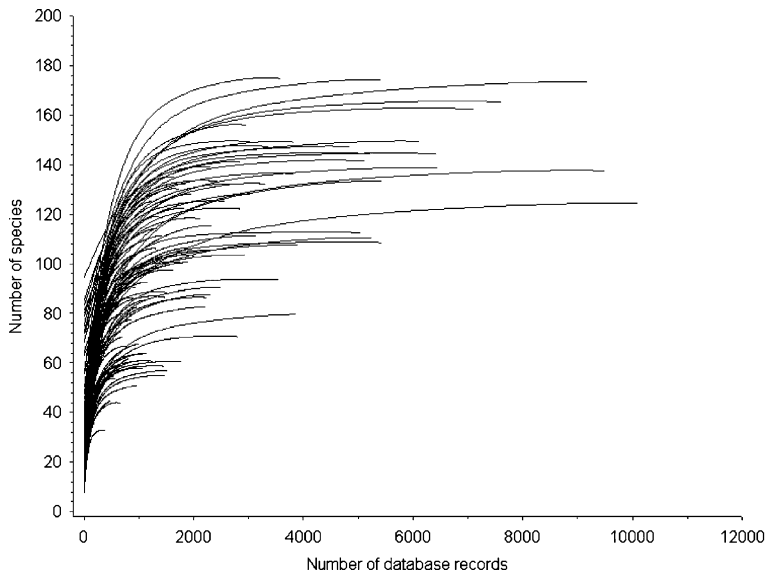


Fig. 4. Accumulation curves (Clench function) for the 95 50×50 km UTM grid cells with percentage of observed to predicted species richness equal to or higher than 90%. The data of each cell were randomized 100 times.

common in the study area: 1) urban and industrial areas, 2) non irrigated croplands, 3) irrigated croplands, and 4) anthropic pasturelands. The central latitude and longitude of each cell were used as spatial variables.

The climate data (original resolution 1 km) are courtesy of the Spanish Instituto Nacional de Meteorología and the Portuguese Instituto de Meteorologia. The topographic variables were obtained from a Digital Elevation Model (Anon. 2000a), and the land-use data (original resolution 280 m) were provided by the European Environment Agency (Corine Programme 1985–1990, Anon. 2000b). The dominant compositions of the ground substrate were from printed geologic charts (scale 1:200 000, Anon. 1995); these were first digitized, and then superimposed on the polygons of cells through the geographic information system IDRISI (Anon. 2003). Lithology diversity was estimated for each grid cell by applying the Shannon diversity index (Magurran 2004) to the primary lithology variables.

Fig. 5. Geographic distribution of sampling intensity of butterfly faunistic studies across the Iberian Peninsula and the Balearic Islands. The varying diameter of solid dots is proportional to sampling intensities on a scale of 4 categories (quartiles) in each range of values. (A) Sampling effort estimated as the ratio of the number of species actually recorded to the number of species predicted by the accumulation curve (Clench function). (B) Raw number of database records per cell. White squares indicate "well-surveyed cells", those where the proportion (number of species recorded/number of species predicted) equals or exceeds 90%.

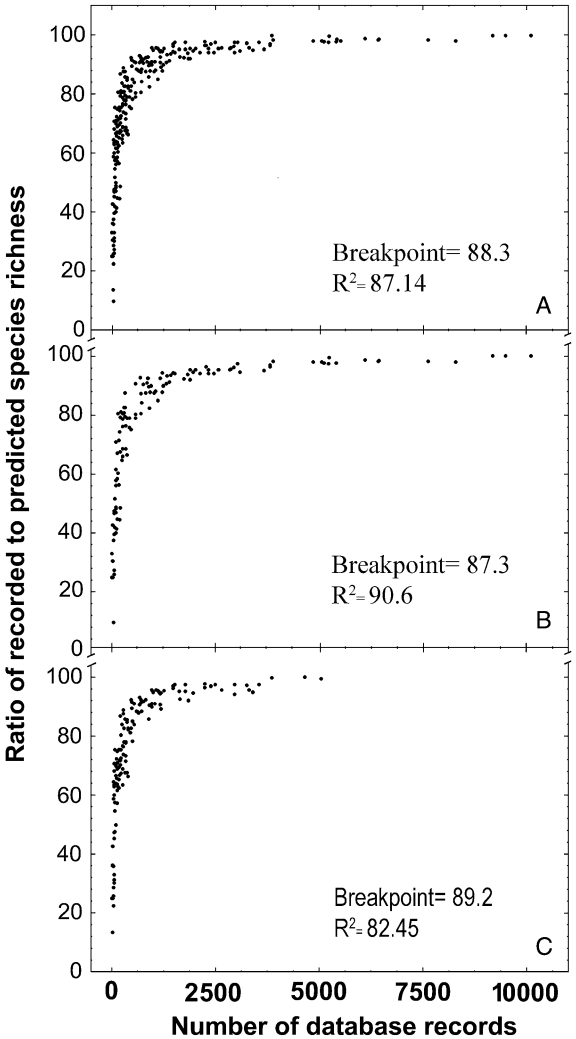


Fig. 6. Relationship between the number of raw database records and the ratio of recorded to predicted species richness for the full set of area units (A), and the subsets of squares with mean elevation higher than (B), and lower than (C) the average. The respectively breakpoints and coefficients of determination (R^2) are shown.

Assessing the effects of environmental and spatial variables on sampling effort

The relationship between cell completeness and test variables was assessed by means of the Generalised Linear Model approach (GLM: McCullagh and Nelder 1989, Crawley 1993), which allows for non-linearity in

Table 2. Number of 50 × 50 km UTM grid cells and species belonging to each physioclimate subregion, database records, and species by cell, at two levels of estimated sampling effort: >75%, >90%. All means are given ±1 SE.

	Eurosiberian	West Mediterranean	East Mediterranean	Islands	North Plateau	South Plateau	Montane	Total
Number of 50 × 50 km UTM cells	37	26	32	5	46	76	35	257
More than 75% of predicted species	18 (48.6%)	19 (73.1%)	20 (62.5%)	1 (20%)	32 (69.6%)	42 (55.3%)	26 (74.3%)	158 (61.5%)
More than 90% of predicted species	11 (29.7%)	7 (26.9%)	12 (37.5%)	1 (20%)	21 (45.7%)	27 (35.5%)	16 (45.7%)	95 (37.0%)
Number of database records	30532	18156	36265	752	53884	63674	85809	289072
Mean number of records by square	825.2±202.9	698.3±186.9	1133.3±273.4	150.4±81.7	1171.4±201.5	837.8±173.8	2451.7±481.9	1124.8±107.8
Mean number of records by square >75%	1555.6±342.8	911.3±238.2	1705.2±385.8	452±0	1616.9±252.2	1455.2±281.2	3250.6±570.7	1754.7±155.7
Mean number of records by square >90%	2250.5±442.8	1867.6±464.4	2511±527.5	452±0	2218.7±307.2	2052.0±393.6	4821.7±672.0	2605.9±217.9
Number of species by square	66.0±6.0	50.9±3.6	84.1±6.5	22±6.2	84.4±5.6	54.3±3.5	116.1±6.8	226
Mean number of species by square >75%	89.6±7.5	58.53±3.1	99±7.7	35±0	98.5±5.8	73.8±3.7	134.6±4.9	91.7±2.9
Mean number of species by square >90%	107.6±6.9	66.1±5.9	115.7±9.3	35±0	113.0±5.1	77.8±4.4	145.4±5.6	103.9±3.6
Total number of species in the region	176	113	190	39	183	157	223	226

Fig. 7. Relationships between the estimated degree of completeness of the inventories (ratio of observed to predicted species richness) and the observed number of species ($R = 0.771$, $p < 0.0001$), and the last one with the number of database records ($R = 0.717$, $p < 0.0001$). Note different scales in the left and right Y-axes. The relationship between the number of records and the estimated species richness can be appreciated in Fig. 3.

the data, as well as for a wide range of model specification distributions other than the normal distribution of the random component. A Poisson error distribution was assumed, since the relationship between the dependent and the explanatory variables (the link function) proved to be logarithmic for the number of database records, and linear for the percentage of observed species.

To evaluate potentially curvilinear relationships within the well-surveyed cells, the dependent variable was first related separately to either a linear, quadratic, or cubic function of each environmental variable (Austin 1980). Subsequently, a forward-stepwise procedure was used to enter the variables into the model (Nicholls 1989, Austin et al. 1996). First, the linear, quadratic or cubic function of the variable that accounted for the most important change in deviance was entered. The remaining variables were then tested for significance, and added to the model sequentially according to their estimated weight. The procedure was iteratively repeated 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. The final model was built separately for each of the three types of explanatory variables (environmental, land use, or spatial). The percentage of variability explained by the different possible combinations of these types of explanatory factors was retained. In the case of spatial variables, the third-degree polynomial equation of the central latitude and longitude was included in the model (Trend surface analysis: Legendre 1993). This is useful to

Table 3. Summary of the regression models selected to estimate the dependence of sampling effort (measured as the raw number of database records) on environmental, geographic and spatial variables. The data are from 257 UTM 50×50 km cells. Of the environmental and land use variables, only those with a statistically significant effect ($p < 0.05$) and a percentage of deviance >1.0 are represented. Dev: deviance explained by each variable; % Dev: percentage of deviance explained.

Explanatory variables	Dev	% Dev	function	sign
Environmental variables				
Maximum altitude	349982	24.59	quadratic	+−
Altitude range	367873	20.74	quadratic	+−
Annual mean temperature	389606	16.06	quadratic	−+
Minimum mean temperature	394044	15.10	quadratic	−+
Mean altitude	397251	14.41	linear	+
Summer precipitation	401710	13.45	quadratic	++
Maximum mean temperature	406156	12.49	quadratic	−+
Annual days of sun	426741	8.05	quadratic	−−
Minimun altitude	430675	7.21	quadratic	+−
Lithologic diversity	440320	5.13	quadratic	++
Calcareous soils	443705	4.40	quadratic	+−
Siliceous soils	447798	3.52	quadratic	−−
Clay soils	450583	2.92	quadratic	+−
Annual temperature range	451376	2.75	quadratic	+−
Annual mean precipitation	453411	2.31	quadratic	+−
Land use variables				
Urban land use	428706	7.63	quadratic	−+
Anthropic pasturelands	432497	6.81	quadratic	+−
Non-irrigated crops	453716	2.24	quadratic	−−
Irrigated crops	457075	1.52	quadratic	−+
Environmental model (E)	277985	40.11		
Land use model (L)	376839	18.81		
Spatial model (S)	301827	34.97		
E+L	223904	51.76		
E+S	197234	57.50		
L+S	217658	53.10		
E+L+S	141284	69.56		

incorporate the influence of spatial structures arising from the effects of other historic, biotic or environmental variables not otherwise taken into account (Legendre and Legendre 1998). A backward stepwise regression with the nine terms of the equation as predictor variables was performed to remove the non-significant spatial terms.

All the variables were standardized to mean = 0 and standard deviation = 1. The Statistica package 6.1. (Anon. 2004) was used for all computations.

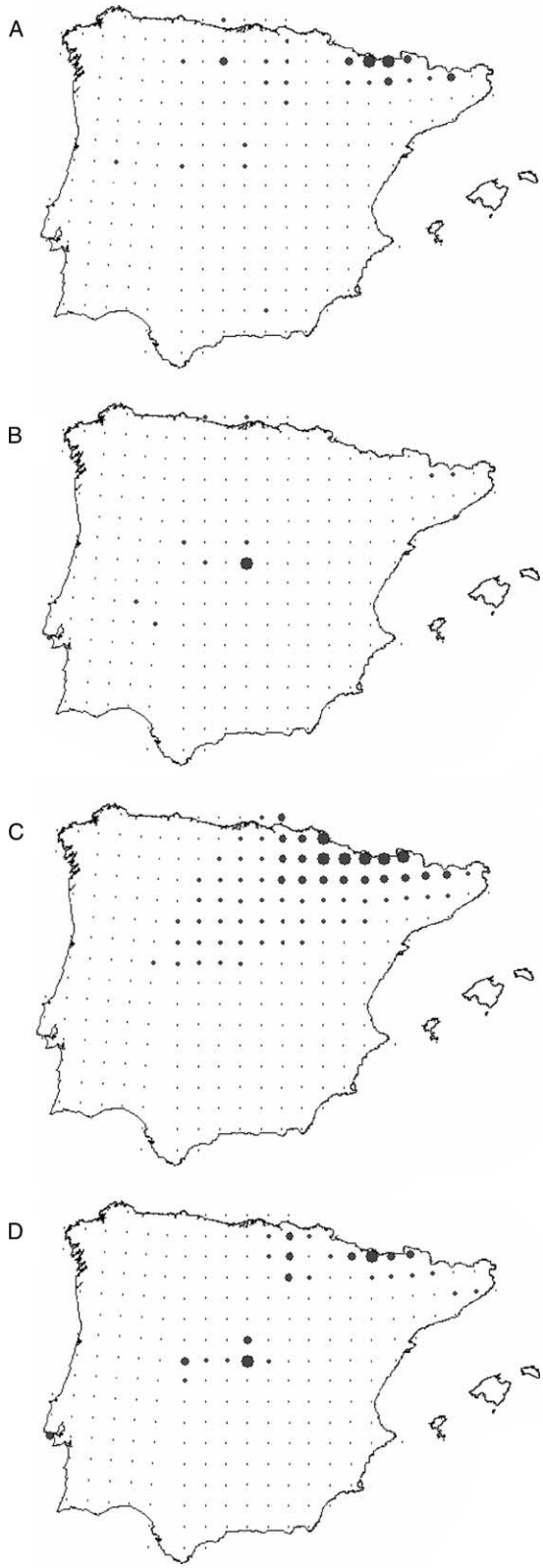
Results

Estimated species richness

Although different estimators of species richness were well correlated to each other (Table 1), those obtained from the Clench function were less directly related to the number of database records in the UTM cells. Further, this effect was more marked at low number of records and more conservative at high numbers of records (Fig. 3). Although this does not demonstrate that this function is more efficient than other methods to determine the actual number of species, it suggests that it is more sensitive to the structure of the data than to the number of independent observations (i.e. the number of records).

Selecting thoroughly prospected cells

On average, the asymptotic number of database records was approached at around 2000 records (Fig. 4). From a total of 257 cells, 158 had completeness values >75%, and 95 of them reached scores of 90% or more (Fig. 4 and 5). A discontinuity in the relationship between the observed and the estimated values was detected, using piecewise linear regression, at a completeness percentage of around 88% (Fig. 6). Taking this percentage as the breakpoint would lead to a classification of 101 UTM cells as well-surveyed (Table 2), hence the very close (and slightly more conservative) figure of 90% was selected. Throughout the remaining sections, cells with completeness ratios equal to or > 90% are simply termed "well-surveyed". Although there was a generally correlated increase in both the actual and the predicted species richness, the number of species in well-surveyed cells varied widely (35–177; Fig. 7). The effects of some environmental variables on sampling intensity could be detected at this stage. For instance, dividing the area units into those with mean elevation above or below the mean suggests that some sites in mountainous areas are prospected repeatedly after their inventories are completed, while this does not happen in lowland areas.



Distribution of well-surveyed cells

The percentage of well-surveyed cells in each of the areas defined by Lobo and Martín-Piera (2002) ranged from roughly 27 to 46% (Table 2). The proportion was lower in the insular area, where only one out of five insular cells was classified as well-surveyed. The remaining subregions contained an acceptable and roughly similar proportion of well-surveyed cells (Chi-square test: $\chi^2 = 3.82$, $p = 0.70$, $DF = 6$). On this basis, no important geographic bias in the distribution of estimated sampling effort should be expected a priori. However, the significant correlation of the number of database records with most of the environmental variables (Table 3) suggests that a more thorough analysis might demonstrate a degree of spatial heterogeneity in sampling effort underlying the general pattern. This is addressed in the next section.

Measuring environmental, geographic and spatial bias in the distribution of well-surveyed areas

The backward stepwise regression explained almost 70% of the variation in the distribution of the number of database records. The three subsets of variables (environmental, land use and spatial) accounted for significant amounts of the variation (Table 3) following a well-defined geographic pattern. The complete model predicts records more densely concentrated in the north-east (in the Pyrenean mountains), as well as in the neighbourhood of Madrid (Fig. 8D). The latter is the area where greater survey effort correlated more closely with environmental and spatial variables, and with an additional, relevant, land use variables (Fig. 8). Entering the sets of variables in the model in the order: 1) environmental, 2) land use, 3) spatial, produced significant progressive increments of the percentage of explained variability (Table 3), thus demonstrating the general relevance of spatial variable addition to land use or environmental variables.

Completeness values were also explained by the selected variables, although their explanatory power was considerably lower (27%). Spatial variables were again the most relevant ones (Table 4). The predicted scores for the complete model fall into a bimodal geographic pattern along a central-western/north-eastern gradient (Fig. 9D). The environmental variables were

Fig. 8. Geographic distribution of bias in the number of raw database records, in relation to environmental, land use, and spatial variables. The dots represent the number of records predicted by four regression models with different sets of independent variables: (A) environmental variables, (B) land use variables, (C) spatial variables, and (D) complete model with all the variables considered. The diameters of the dots are proportional to the predicted cell values, on a scale of 4 categories (quartiles).

Table 4. Summary of the final regression models selected to explain the dependence of sampling bias on environmental, geographic and spatial variables. The sampling bias in the dependent variable was measured as the ratio between the observed and predicted species richness (as estimated from the Clench function). Of the environmental and land use variables, only those with a statistically significant effect ($p < 0.05$) and a percentage of deviance > 1.0 are given. Dev: deviance explained by each variable; % Dev: percentage of deviance explained.

Explanatory variables	Dev	% Dev	function	sign
Environmental variables				
Altitude range	1867	2.98	quadratic	++
Maximum altitude	1873	2.66	quadratic	++
Summer precipitation	1887	1.96	linear	+
Maximum mean temperature	1896	1.49	linear	−
Annual temperature range	1901	1.21	quadratic	−−
Lithologic diversity	1902	1.18	quadratic	++
Land use variables				
Urban land use	1877	2.45	linear	+
Anthropic pasturelands	1886	2.00	quadratic	+−
Environmental model (E)	1835	4.67		
Land use model (L)	1833	4.76		
Spatial model (S)	1551	19.40		
E+L	1761	8.50		
E+S	1488	22.68		
L+S	1475	23.41		
E+L+S	1412	26.61		

able to explain the northeastern maximum (Fig. 9A), while land use accounted for high completeness scores around the most important cities (Madrid, Barcelona and Lisbon).

These environmental and spatial patterns in the collection effort were detected even when only the well-surveyed cells were considered. Spatial variables were still able to account for almost 13% of the percentage of completeness. This percentage increased to 19% when all the variables were included in the model (Table 5). As the proportion of variation accounted for by land use and environmental variables was much lower than that explained by spatial effects, a low correlation between spatial patterns and other sources of variation was evident. The geographic distribution of the scores predicted by the model followed the same bimodal pattern described above (Fig. 10).

Discussion

The statistical procedure we presented here has previously been used as a tool to estimate species richness in incompletely surveyed areas (Lobo and Martín-Piera 2002, Carrascal et al. 2002, Lobo et al. 2002, Hortal et al. 2004). However, rather than focusing on species numbers, our approach stresses the potential use of GLM for detecting and explaining the ways in which the information itself has been modelled.

No doubt, the results can be criticized from several points of view. Those include the relatively large grid size, which certainly does not contribute to a highly resolved explanation. It is doubtful, however, that a more finely grained scale would really show a different pattern, because of the substantially lower number of well-prospected squares on a 10×10 km basis, and because virtually all of these would be located within the large (50 km cells) units that were determined as sufficiently prospected.

Second, a source of bias caused by shifting patterns of human activity during the last two centuries can not be derived directly from the results, because such potential shifts have not been controlled for. From this point of view, and given the time distribution of the data analysed, the patterns detected probably represent the dominating trends in butterfly collection during the second half of the 20th century.

Within the limits stated above, the results demonstrate the interest of incorporating estimates of sampling bias in biodiversity studies. Preliminary, intuitive evaluations of Iberian and Balearic butterfly data suggested that the geographic coverage of the data might be interpreted as representative of the expected diversity of butterflies (García-Barros et al. 2000, Garcia-Pereira 2003, Romo and García-Barros 2005). This view would be reinforced by the roughly homogeneous spread of well-surveyed area units across the main ecological regions. However, further statistical treatment detected several sources of bias in the recording intensity; this was evident even when only the well-surveyed cells were analysed.

A relevant proportion of the variation in the data is not explained by the regression models. It is likely that one part of such variation is due to shifting patterns of human activity (namely, changes in land use). It has been suggested that such effects can be indirectly detected by the third degree polynomial equation of latitude and longitude (Legendre and Legendre 1998, Lobo and Martín-Piera 2002). Our results suggest that this may be the case for Iberian butterflies. Since this requires a more specific analytic approach, the remaining discussion has

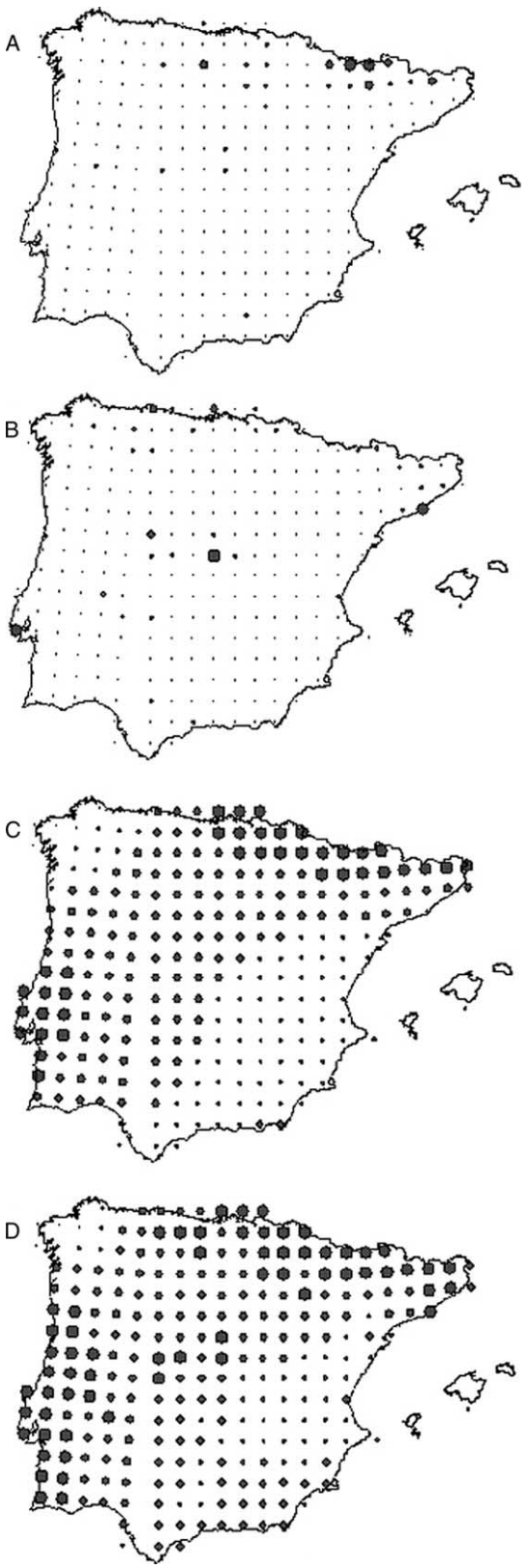


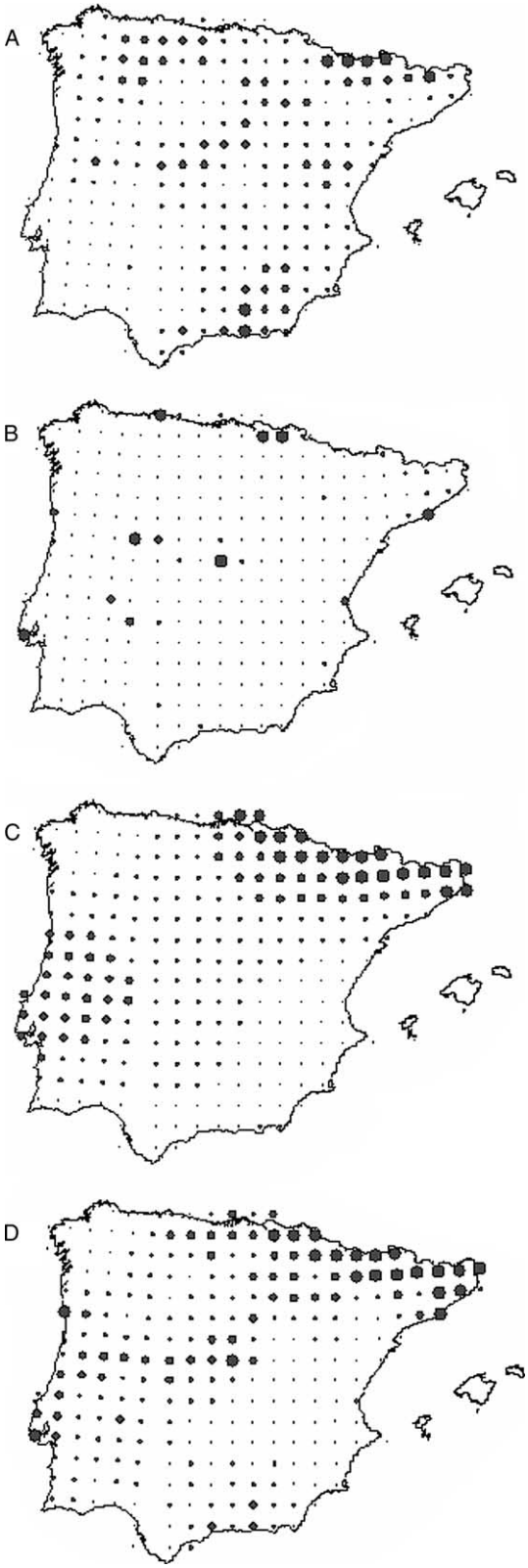
Table 5. Summary of the final regression models selected to explain the dependence of sampling bias on environmental, geographic and spatial variables. Sampling bias (the dependent variable) was measured as the ratio of recorded to predicted species richness in the well-surveyed cells (n = 95; those in which the observed species-richness was equal to, or > 90% of the predicted score). Of the environmental and land use variables, only those with a statistically significant effect ($p < 0.05$) and a percentage of deviance > 1.0 are represented. Dev: deviance explained by each variable; % Dev: percentage of deviance explained.

Explanatory variables	Dev	%Dev	function	sign
Environmental variables				
Maximum altitude	329	2.88	linear	–
Altitude range	330	2.49	linear	–
Annual mean temperature	333	1.68	linear	+
Minimum mean temperature	334	1.41	linear	+
Maximum mean temperature	334	1.41	linear	+
Mean altitude	335	1.14	linear	–
Summer precipitation	335	1.08	linear	–
Land use variables				
Urban land use	332	1.97	linear	–
Anthropic pasturelands	333	1.59	linear	–
Environmental model (E)	329	2.88		
Land use model (L)	327	3.54		
Spatial model (S)	296	12.66		
E+L	318	6.14		
E+S	288	15.02		
L+S	287	15.14		
E+L+S	274	19.15		

to be interpreted in terms of "pooled effects that are vastly attributable to the dominating environmental circumstances during the last 50 or 60 yr".

The geographic bias in the sampling intensities attributable to different sets of environmental variables are broadly coincident irrespective of the estimate of sampling intensity, although some differences arise when the spatial correlations are tested. The regression models accounted for a much higher proportion of the variation in the number of raw records than in the estimated sampling effort. This suggests that statistical artefact effects were largely smoothed in the completeness values, highlighting the fact that the predicted richness values depend more on the quality of database records than on their number, so a high percentage of predicted richness can be obtained from different levels of database records, depending on their quality. For instance, published materials may or may not include detailed information such as specimen counts. This often depends on the specific journal, or on whether the data were published in periodical journals or concentrated in regional

Fig. 9. Geographic distribution of the amount of bias in sampling effort (measured as the ratio of observed to predicted species richness) attributable to environmental, land use, and spatial variables. The values plotted are the scores predicted for the dependent variable in four multiple regression models based on four different sets of predictor variables: (A) environmental variables, (B) land use variables, (C) spatial variables, and (D) complete model with all the variables. The diameters of the dots are proportional to the predicted cell values, on a scale of 4 categories (quartiles).



distribution atlases; the latter are available for only some regions, and often consist of dot maps devoid of any further information.

Land use variables have only locally relevant effects, but these are evident in the cells surrounding the cities of Barcelona, Madrid and Lisbon. These cities, major human settlements, have present populations within the range of 2–5 million inhabitants (depending on whether the suburbs are considered). Faunistic reports in these cells have been made more or less regularly for nearly 150 yr, while human population in the same areas has increased fourfold during the last five decades. Furthermore, the data on land use date from 1990, and hence reflect a "modern" pattern of land exploitation. This should lead to locally overestimated present species diversities. This might be worth considering when the causes of butterfly population losses are evaluated, such as in the case of the Lycaenid *Lycaena tityrus*, attributed to climate warming (Parmesan et al. 1999).

Even if marginal, the environmental effects on estimated sampling effort prove that collecting has been concentrated in mountainous areas. The geographic distribution of sampling effort falls into a mixed pattern, the highest densities spreading across: 1) the main mountain chains (Pyrenees, Cantabrians, central Iberian and Baetic mountain ranges), and 2) the southern Mediterranean and the western Atlantic coasts (plus a few inland patches along the Portuguese-Spanish border). This reinforces former findings, drawn from British butterflies (Dennis and Thomas 2000) and several Iberian insect taxa (Martín and Gurrea 1999), of two co-existing trends: 1) the local lepidopterists overexploit accessible sites near their headquarters, and 2) when collection implies a long trip from the entomologist's place of residence (or the collectors travel from other countries), the sites selected are more frequently scattered across mountain ranges. This may reasonably be expected because in a dry, Mediterranean environment, high elevations exist in topographically varied areas with comparatively pleasing summer temperatures and more varied landscapes. Because the peak diversity of these insects through most of the study area occurs in July, this pattern of site selection is difficult to distinguish from the "diversity tracking" selection that might be expected from a traditional entomologist (mountain ranges host comparatively high numbers of butterfly species in the

Fig. 10. Geographic distribution of bias in sampling effort (ratio of observed to predicted species richness), in relation to environmental, land use, and spatial variables, measured using only the subset of well surveyed cells. The regression models were fitted to the data of cells with ratios of recorded to predicted species richness higher than 90% (see Fig. 5), and the values were predicted for the full set of squares. The four maps show the predicted scores of (A) environmental variables, (B) land use variables, (C) spatial variables, and (D) environmental plus land use and spatial variables. The diameters of the dots are proportional to the predicted cell values, on a scale of 4 categories (quartiles).

Mediterranean countries: Munguira 1995, Hawkins and Porter 2003). Perhaps paradoxically, wide patches along the eastern Mediterranean and south-west Atlantic shorelines remain under-surveyed in spite of their relevance as current tourism resources. This may be due to the low "entomological appeal" of some rather heavily-developed tourist coastal areas, which may induce the collectors to travel inland from their residences to visit the coastal sierras. Thus the apparently low diversity of these areas needs to be re-assessed, which might lead to controversy on the management of coastal habitats in the Iberian Peninsula south.

The results indicate that environmental, spatial and land-use related biases are largely independent of each other. However, only the spatial variables had a relatively widespread effect on the data. Interestingly, their effects suggest a bipolar axis of NE-SW orientation. This trend is in the same direction described for the dominant gradient of butterfly species richness in Iberia (Martín and Gurrea 1990), hence it would be interesting to re-assess the geographic trends in species richness while controlling for spatial effects in sampling effort.

Thus with regard specifically to butterfly faunistics, at least one third of the areal units in each physiographic region seems to have been adequately prospected. This may suffice for broad-scale prospective studies, but confirms that the data are not yet suitable for analyses on a more fine-grained scale (only ca 40% of the cells were classified as well-surveyed, although the percentage would certainly increase with a less-demanding completeness criterion, Hortal et al. 2001). Future faunistic work should specifically be concentrated in Galicia, southern Portugal, the southern plateau (La Mancha), the Mediterranean coast, and the Balearic Islands. The effort should focus on sites featuring typical Mediterranean climate, with high mean yearly temperature, moderate to low elevations, distant from large human settlements scattered regularly across the aforementioned areas.

Acknowledgements – The authors thank Tom Brereton for comments on a draft version of this work, James Cerne for kindly reviewing the English, and also to the members of the project ATLAMAR (REN2000-0466 GLO): P. Garcia-Pereira, E. Maravalhas, J. Martín, and M. L. Munguira. H.R. was supported by a PhD grant from the Spanish Ministry of Education and Culture (reference AP2002-0147).

References

Anon, 1995. Atlas nacional de España, 16. – Inst. Geográfico Nacional, Centro Nacional de Información.

Anon, 2000a. Global change data archive, Vol. 3. 1 km. – Global Elevation Model, CD-Rom, Clark Univ.

Anon, 2000b. Natural Resources CD-Rom. CORINE Land Cover. Technical guide. European Environment Agency, <http://www.eea.eu.int >.

Anon, 2003. Idrisi Kilimanjaro. – GIS software package. Clark Labs.

Anon, 2004. STATISTICA (data analysis software system), computer program manual. – StatSoft, ver. 6, <www.statsoft.com >.

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

Austin, M. P. et al. 1996. Patterns of tree species richness in relation to environment in south-eastern New South Wales, Australia. – Aust. J. Ecol. 21: 154–164.

Brooks, T. et al. 2004. Species, data, and conservation planning. – Conserv. Biol. 18: 1682–1688.

Burnham, K. P. and Overton, W. S. 1979. Robust estimation of the size of population size when capture probabilities vary among animals. – Ecology 60: 927–936.

Carrascal, L. M. et al. 2002. Patrones de preferencias de hábitat y de distribución y abundancia invernal de aves en el centro de España. Análisis y predicción del efecto de factores ecológicos. – Anim. Biodiv. Conserv. 25: 7–40.

Chao, A. 1984. Non-parametric estimation of the number of classes in a population. – Scand. J. Stat. 11: 265–270.

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. – Biometrics 43: 783–791.

Colwell, R. K. 2000. EstimateS: statistical estimation of species richness and shared species from samples. – Software and user's guide, ver. 6.0b1, <http://viceroy.eeb.uconn.edu/estimates >.

Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity through extrapolation. – Phil.Trans. R. Soc. B 354: 101–118.

Crawley, M. J. 1993. GLM for ecologists. – Blackwell.

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

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

Dennis, R. L. H. et al. 1999. Bias in butterfly distribution maps: the effects of sampling effort. – J. Insect Conserv. 3: 33–42.

Dennis, R. L. H. et al. 2006. The effects of visual apparency on bias butterfly recording and monitoring. – Biol. Conserv. 128: 486–492.

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

Freitag, S. et al. 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. – Anim. Conserv. 1: 119–127.

García-Barros, E. and Munguira, M. L. 1999. Faunística de mariposas diurnas en España peninsular. Áreas poco estudiadas: una evaluación en el umbral del Siglo XXI (Lepidoptera: Papilionoidea & Hesperiidae). – SHILAP Revta. Lepid. 27: 189–202.

García-Barros, E. et al. 2000. The geographic distribution and state of butterfly faunistic studies in Iberia (Lepidoptera Papilionoidea Hesperioidea). – Belg. J. Entomol. 2: 111–124.

García-Barros, E. et al. 2004. Atlas de las mariposas diurnas de la Península Ibérica e islas Baleares (Lepidoptera: Papilionoidea & Hesperioidea). – Monografías de la SEA, Vol. 11, Zaragoza.

Garcia-Pereira, P. N. C. 2003. Mariposas diurnas de Portugal continental: faunística, biogeografía y conservación. – Ph.D. thesis, Univ. Autónoma de Madrid, Madrid.

Garcia-Pereira, P. et al. 1999. Evaluación del conocimiento de la fauna de mariposas de Portugal continental (Lepidoptera: Papilionoidea, Hesperioidea). – SHILAP Revta. Lepid. 27: 225–231.

Gaston, K. J. and Spicer, J. I. 2004. Biodiversity. An introduction, 2nd ed. – Blackwell.

Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – Ecol. Lett. 4: 379–391.

Hawkins, B. A. and Porter, E. E. 2003. Water-energy balance and the geographic pattern of species richness of western Palearctic butterflies. – Ecol. Entomol. 28: 678–686.
Heltshe, J. and Forrester, N. E. 1983. Estimating species richness using the Jackknife procedure. – Biometrics 39: 1–11.
Hortal, J. et al. 2001. Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeinae). – Biodiv. Conserv. 10: 1343–1367.
Hortal, J. et al. 2004. Butterfly species richness in mainland Portugal: predictive models of geographic distribution patterns. – Ecography 27: 68–82.
Hortal, J. et al. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. – J. Anim. Ecol. 75: 274–287.
Koellner, T. et al. 2004. Rarefaction method for assessing plant species diversity on a regional scale. – Ecography 27: 532–544.
Kress, W. J. et al. 1998. Amazonian biodiversity: assessing conservation priorities with taxonomic data. – Biodiv. Conserv. 7: 1577–1587.
Kudrna, O. 2002. The distribution atlas of european butterflies. – Oedippus 20: 1–342.
Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – Ecology 74: 1659–1673.
Legendre, P. and Legendre, L. 1998. Numerical ecology, 2nd English ed. – Elsevier.
Lobo, J. M. and Martín-Piera, F. 2002. Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. – Conserv. Biol. 16: 158–173.
Lobo, J. M. et al. 1997. Les atlas faunistiques comme outils d'analyse spatiale de la biodiversité. – Ann. Soc. Entomol. France 33: 129–138.
Lobo, J. M. et al. 2002. Modelling the species richness distribution of French dung beetles (Coleoptera, Scarabeidae) and delimiting the predictive capacity of different groups of explanatory variables. – Global Ecol. Biogeogr. 11: 265–277.
Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.
Martín, J. and Gurrea, P. 1990. The peninsular effect in Iberian butterflies (Lepidoptera: Papilionoidea and Hesperioidea). – J. Biogeogr. 17: 85–96.
Martín, J. and Gurrea, P. 1999. Áreas de especiación en España y Portugal. – Boletín de la AeE 23: 83–103.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
Munguira, M. L. 1995. Conservation of butterfly habitats and diversity in European Mediterranean countries. – In: Pullin, A. S. (ed.), Ecology and conservation of butterflies. Chapman and Hall, pp. 277–289.
Nelson, B.W. et al. 1990. Endemism centers, refugia and botanical collection density in Brazilian Amazonia. – Nature 345: 714–716.
New, T. R. 1991. Butterfly conservation. – Oxford Univ. Press.
Nicholls, A. O. 1989. How to make biological surveys go further with generalised linear models. – Biol. Conserv. 50: 51–75.
Parmesan, C. et al. 1999. Poleward shifts in geographic ranges of butterfly species associated with regional warming. – Nature 399: 579–583.
Parnell, J. A. N. et al. 2003. Plant collecting spread and densities: their potential impact on biogeographic studies in Thailand. – J. Biogeogr. 30: 193–209.
Petersen, F. T. and Meier, R. 2003. Testing species-richness estimation methods on single-sample collection data using the Danish Diptera. – Biodiv. Conserv. 12: 667–686.
Peterson, A. T. and Slade, N. A. 1998. Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. – Div. Distrib. 4: 95–105.
Petersen, J. F. T. et al. 2003. Testing species richness estimation methods using museum label data on the Danish Asilidae. – Biodiv. Conserv. 12: 687–701.
Peterson, A. T. et al. 1998. The need for continued scientific collecting: a geographic analysis of Mexican bird specimens. – Ibis 140: 288–294.
Reddy, S. and Dávalos, L. M. 2003. Geographic sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Romo, H. and García–Barros, E. 2005. Distribución e intensidad de los estudios faunísticos sobre mariposas diurnas en la península Ibérica e islas Baleares (Lepidoptera, Papilionoiea y Hesperioidea). – Graellsia 61: 37–50.
Rosenzweig, M. L. et al. 2003. Estimating diversity in unsampled habitats of a biogeographic province. – Conserv. Biol. 17: 864–874.
Smith, E. P. and van Belle, G. 1984. Nonparametric estimation of species richness. – Biometrics 40: 119–129.
Soberón, J. and Llorente, J. 1993. The use of species accumulation functions for the prediction of species richness. – Conserv. Biol. 7: 480–488.
Soberón, J. M. et al. 2000. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. – Biodiv. Conserv. 9: 1441–1466.
Tolman, T. and Lewington, R. 1997. Collins field guide. Butterflies of Britain and Europe. – Harper Collins.
Whittaker, R. J. et al. 2005. Conservation biogeography: assessment and prospect. – Div. Distrib. 11: 3–23.

Subject Editor: Andrew Liebhold.