Research

Oikos 117: 847–858, 2008
doi: 10.1111/j.2008.0030-1299.16434.x,
© 2008 The Authors. Journal compilation © 2008 Oikos
Subject Editor: Owen Petchey, Accepted 19 February 2008

Historical bias in biodiversity inventories affects the observed environmental niche of the species

Joaquín Hortal, Alberto Jiménez-Valverde, José F. Gómez, Jorge M. Lobo and Andrés Baselga

J. Hortal (j.hortal@imperial.ac.uk), NERC Centre for Population Biology, Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire, UK, SL5 7PY. – A. Jiménez-Valverde, J. F. Gómez, J. M. Lobo and A. Baselga, Depto de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), C/José Gutiérrez Abascal 2, ES–28006 Madrid, Spain.

It is well known that biodiversity data from historical inventories presents important geographic and taxonomic biases. Due to this, current knowledge on the distribution of most species could be incomplete and biased. We assess how the biases in historical biodiversity data might affect the description of the environmental niche of the species, using exhaustive data on the distribution of dung beetles in Madrid as a case study. We describe the historical process of survey and compare such historical data with the results of an exhaustive survey, identifying the environmental biases in the historical surveys during different periods, and assessing the completeness of the environmental niche of the species provided by historical data through time. Events like the Spanish Civil War affect the tempo and spread of surveys, but the exhaustive work since 1970 provides a good, though incomplete, coverage of the region by 1998. In spite of this, the biases in historical data result in a limited knowledge about the niche of an important number of species. Although nearly a half of the species had the 100% of their niche covered by data in 1998, roughly a third had less than 75%, nearly a fourth less than 50%, and 18 species had to be excluded from the analyses due to the lack of data. Our results point out that data from non-standardized inventories often provide an incomplete description of the environmental responses of most species. Due to this, we highlight that currently predictive models of species distributions present some limitations, since the results of models based in partial information about the environmental niche of the species will be compromised. Therefore, the biases in the available data must be evaluated before constructing predictive maps of species distributions, and taken into account when drawing conclusions or conservation strategies from these maps.

Species are the basic unit of biodiversity (Wilson 2002). Therefore, strategies for biodiversity conservation should be based on information about their distribution. However, the design of biodiversity conservation strategies is hampered by our partial knowledge on the geographic patterns of biodiversity (the Wallacean shortfall; Lomolino 2004, Whittaker et al. 2005). After more than 250 years of compiling distributional and taxonomic data, we have no complete inventory of the organisms inhabiting any single site in the world, nor any accurate record of the distribution of a single species (except for several endangered species with only one or a few populations). Due to this, it is generally agreed that more taxonomic and distribution information should be gathered (Saarenmaa and Nielsen 2002), to be used in the design of reserve networks that cover as much species as possible (Grenyer et al. 2006). A number of national and international initiatives are currently digitizing exhaustively all the distributional information gathered during the last centuries (e.g. GBIF; Edwards et al. 2000; <http://www.gbif.org/>).

Although some regions with a long naturalist tradition have almost complete checklists for some groups, biodiversity databases often offer an unreliable picture of the distribution of biodiversity even in these regions (Dennis and Hardy 1999, Soberón et al. 2000, Hortal et al. 2007, Soberón et al. 2007). A solution to their limited coverage could be using predictive modelling methodologies to generate coherent hypotheses of the current and future species distributions (Guisan and Zimmermann 2000, Soberón and Peterson 2005, Araújo and Guisan 2006); these hypotheses can be used to select areas for conservation, and assess their future adequacy under global change scenarios (Araújo et al. 2004, Cabeza et al. 2004). However, predictive models provide unreliable estimations of species distributions if the distributional information used to calibrate them presents important environmental and/or spatial biases (Hortal and Lobo 2006, Hortal et al. 2007, Lobo 2008a).

The distribution of biological information is generally biased to certain environmental domains as it is to certain areas. Non-systematic sampling provides biased descriptions of species geographic ranges and leads to major errors in the distribution of endangered or conservation target species (Dennis and Hardy 1999, Dennis 2001), a pattern that can be extended to the rest of species, which usually receive less attention. And there are a number of known biases in the

historical process of inventory (Hortal and Lobo 2006), such as taxonomists' home range, or the proximities of work centres and roads (Dennis et al. 1999, Dennis and Thomas 2000, Hortal et al. 2001, 2004, Lobo and Martín-Piera 2002, Martín-Piera and Lobo 2003, Reutter et al. 2003, Graham et al. 2004, Kadmon et al. 2004, Martínez-Meyer 2005, Beck and Kitching 2007). In consequence, the patterns of recording and/or description of new taxa are spatially and taxonomically structured for most groups (Gaston 1991, Gaston and Blackburn 1994, Medellín and Soberón 1999, Dolphin and Quicke 2001, Reed and Boback 2002, Cabrero-Sañudo and Lobo 2003, Collen et al. 2004, Diniz-Filho et al. 2005, Gibbons et al. 2005, Baselga et al. 2007, Guil and Cabrero-Sañudo 2007, Jiménez-Valverde and Ortuño 2007, Lobo et al. 2007).

If inventories are environmentally biased they will provide an incomplete description not only of the geographic distributions of most species, but also of their realized niche. In a recent paper, we show that the knowledge about the distribution of a given group has been historically accumulated in an environmentally-structured fashion (Lobo et al. 2007). Here, we hypothesize that such historical bias in inventory processes results in unreliable descriptions of the geographic and environmental responses of many species. To investigate this hypothesis, we describe the historical process of the inventory of a taxonomically well-known insect group (dung beetles; Coleoptera Scarabaeoidea) in a exhaustively surveyed region (Madrid, Central Iberian Peninsula; Fig. 1), analyzing how the description of the niche of these species varies through time. Thus, the specific aims of this work are:

1. to study the historical process of species recording in this region
2. to identify and describe the spatial and environmental biases of surveys in different periods, and
3. to determine if (and how much) these biases provide incomplete descriptions of the realized niche of the species, and how such niches are unravelled through time.

To do this, we use an exhaustive compilation of all the information available for the three families of Iberian dung beetles (Scarabaeidae, Aphodidae and Geotrupidae) in Madrid region and surrounding areas, comparing it with the results obtained after an extensive survey recently carried out in this region (Hortal 2004). This survey was explicitly designed to account for geographic and environmental variations in dung beetle diversity (Hortal and Lobo 2005), so we assume that current distribution information provides an accurate description of the realized niches of dung beetle species distributions in central Iberian Peninsula.

Data and methods

Study area

Madrid is an autonomous Spanish region, placed in the geographic centre of the Iberian Peninsula (Fig. 1). In spite of its relatively limited size (approx. 8000 km^2 and 140 km width from the northern to the southernmost points), Madrid is highly heterogeneous. Although it presents continental climate with Mediterranean influence throughout its territory, climate and topography vary, along with elevations, from 434 m a.s.l. in the Alberche valley, to the 2430 m a.s.l. of the Peñalara peak, in the Central System mountain range. Thus, annual precipitation ranges from 350 mm to 2000 mm, and important temperature

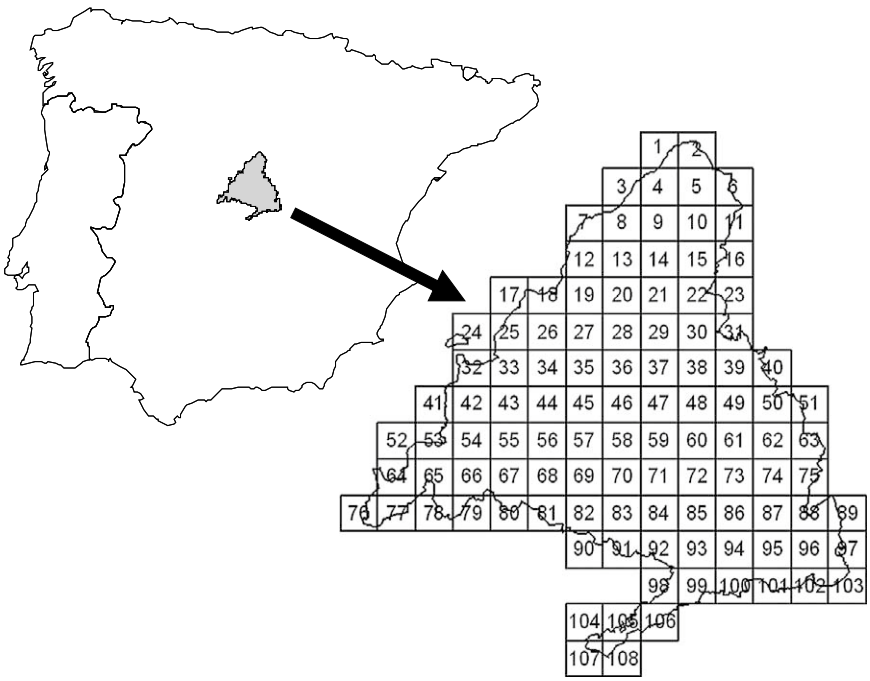

Fig. 1. Spatial location of the Madrid region within the Iberian Peninsula (left) and distribution of the 108 10×10 km UTM grid cells used as spatial units in the analyses (right) (see Table S1 for further information).

differences appear as well, so it hosts several major bioclimatic divisions (Mesomediterranean, Supramediterranean, Oromediterranean and Crioromediterranean; Rivas-Martínez 1987). Its complex geologic history has also produced an important lithologic diversity, with acidic rocks in the mountains; alluvial deposits on mountain slopes, terraces and valleys; and calcareous rocks and clays, and even gypsum soils, in the southeast. Such environmental heterogeneity, together with its positioning in the centre of the Iberian Peninsula, has made Madrid a region of transition between Mediterranean and Eurosiberian faunas (Fernández Galiano and Ramos Fernández 1987), being also an ideal region for small-scale pilot studies, as it is home to a synthesis of all inland Iberia. In addition to this, is a highly populated region, with four major universities and a number of research centres, being home of an important number of Spanish field biologists and taxonomists. Thus, its biota has been intensively surveyed since the 18th century, and is the naturalistically (and entomologically) best known area in the Iberian Peninsula, and one of the better known areas of Mediterranean Europe as well. We have divided the territory of Madrid (and its surroundings) in 108 UTM 10×10 km grid cells (herein, grid cells), which constitute the spatial units for all further analyses (Fig. 1) (Hortal and Lobo 2005).

Biological data

Historical data

We compiled all historical information on the distribution of dung beetles in the Madrid region and its surroundings available until 1998 in SCAMAD database. The structure of SCAMAD is based on a former database on Iberian Scarabaeidae (BANDASCA; Lobo and Martín-Piera 1991), and consists of 34 data fields, which include information about the taxonomy, geographic location, date, ecological/biological characteristics and origin of all database records. Data gathered in SCAMAD included all the specimens and museum vouchers existing in Natural History Museums, universities, accessible private collections and unpublished field data, as well as any additional distribution information published in the literature (see Hortal 2004 for a complete list of data sources). All individuals where assigned to valid species, according to the taxonomy of Martín-Piera (2000) for Scarabaeidae, Veiga (1998) for Aphodiinae, Baraud (1992) for the rest of Aphodiidae, and López-Colón (2000) for Geotrupidae. In total, SCAMAD 1.0 includes information about 92 741 individuals from 133 species (47 Scarabaeidae, 71 Aphodiidae and 15 Geotrupidae; Hortal 2004, Hortal and Lobo 2005). These data were used to assess variations in the observed niche of the species through time.

Validation data

Additional data were used to account for the actual realized niche of dung beetle species in Madrid (below). These data come from an extensive survey designed to complement the environmental and spatial coverage of the areas with good quality information in SCAMAD 1.0 (Hortal and Lobo 2005). Briefly, the first eight grid cells identified by Hortal and Lobo (2005) were thoroughly surveyed between 1998 and 2001. The results of these surveys and additional information from private collections and sparse captures were included in SCAMAD, resulting in the addition of another 53 163 individuals. Therefore, the current version of SCAMAD (2.1), used as validation data, includes information about 145 904 individuals from 134 species, captured from the beginning of the historical surveys until summer 2004 (Hortal 2004).

Description of survey effort

We have used the number of database records in SCAMAD (herein, records) to measure sampling effort (Martín-Piera and Lobo 2003, Hortal and Lobo 2005, Lobo 2008b). Briefly, a record is defined as each time a species has been recorded by a different method or collector, regardless of the number or sex of the captured individuals; to constitute a new record, these individuals must differ at least in one of the following database fields: capture date, place of capture, habitat type, feeding, capture or observation method and collector. This measure has the advantage of minimizing the differences in the number of recorded individuals that are caused by the different population size of the species instead of true differences in the effort devoted to their survey. Database records have been successfully used as a surrogate of sampling effort (Hortal et al. 2001, Lobo 2001, 2008b, Lobo and Martín-Piera 2002, Baselga and Novoa 2006, 2007, Carpaneto et al. 2007, González et al. 2007, Lobo et al. 2007), showing a similar behaviour to other effort measures such as individuals or traps (Hortal et al. 2006). Indeed, the number of records and the number of recorded individuals per species in SCAMAD 1.0 are also highly correlated (all species: Spearman $R = 0.917$, $t(n-2) = 26.35$, $n = 133$, $p < 0.001$; Scarabaeidae: Spearman $R = 0.938$, $t(n-2) = 18.08$, $n = 47$, $p < 0.001$; Geotrupidae: Spearman $R = 0.982$, $t(n-2) = 18.80$, $n = 15$, $p < 0.001$; Aphodiidae: Spearman $R = 0.922$, $t(n-2) = 19.81$, $n = 71$, $p < 0.001$).

All records in SCAMAD were referred to the 108 grid cells described above (Fig. 1). All information available for these grid cells but placed outside the administrative limits of Madrid was also included in the database. The accuracy of the geographic location of an important number of records was limited, since they were referred to broad areas or local councils. However, most of them could be attributed to these 108 grid cells with reliability (5237 out of the 5364 records in SCAMAD 1.0, and 6954 out of 7110 records in SCAMAD 2.1).

We plotted the 'historical species accumulation curve' (Medellín and Soberón 1999, Cabrero-Sañudo and Lobo 2003) to characterize the historical process of dung beetle inventory in Madrid until 1998. The curve describes the historic rate of species inventory, by plotting the number of observed species against the year; each species is added to the inventory the year of its first record in SCAMAD. Therefore, it can be used to identify different periods of increase in knowledge. We identified these periods, and mapped the distribution of records in the 108 cells to assess the changes in the geographic distribution of sampling effort between different periods. Whenever possible, non-dated records where assigned to these periods according to

the period of activity of the collectors indicated in the labels.

Environmental bias in the surveys

We assessed the effect that the bias in the historical surveys might have in the description of the niche of Madrid dung beetle species through time. To do this, we related the environmental conditions in each grid cell with the records in SCAMAD 1.0 available for each one of the periods defined by the historical species accumulation curve. We base our analyses in two kinds of variables that have been previously related to dung beetle distribution (Hortal et al. 2001, Lobo and Martín-Piera 2002, Hortal and Lobo 2005): climate and substrate (Table 1). Climate is described by five variables: mean, maximum and minimum temperature, total annual precipitation and total precipitation in summer (June–August). Briefly, monthly temperature and precipitation scores for 41 stations of central Iberia (means from 30-year data) were interpolated to 1 km^2 spatial-resolution maps using a moving-average procedure (Hortal 2004). Monthly 1 km^2 scores were used to calculate five climatic variables above, and these variables were re-scaled to the scale of analyses by calculating their mean scores in each UTM 10×10 km grid cell (Fig. 1). Substrate variables represent the proportion of area of each grid cell covered by each category; bedrock data come from ITGE (1988), and soil structure (horizon development) and composition come from FAO (1988). See additional details on the origin and/or conversion of all variables to GIS at Chefaoui et al. (2005) and Hortal and Lobo (2005). To avoid problems of collinearity among variables and simplify the analyses, each group of variables was reduced to two uncorrelated factors by means of a principal components analysis (varimax rotated) (Table 1). The number of factors extracted was chosen according to a broken-stick criterion, following Legendre and Legendre (1998). These four factors (two related to climate, CF1 and CF2, and two to substrate, SF1 and SF2; Table 1) are used to describe the environmental conditions on each grid cell (see scores in Supplementary material Table S1).

Table 1. Environmental factors used in the analyses. CF stands for climate factor and SF for substrate factor. PCA factor loadings are shown next to each original variable, and the eigenvalues and percentage of explained variability are included below.

a) Climate	CF1	CF2
Mean temperature	−0.784	0.584
Maximum temperature	−0.929	0.276
Minimum temperature	−0.328	0.944
Annual precipitation	0.908	−0.346
Summer precipitation	0.858	−0.382
Eigenvalue	4.20	0.52
% total variability	84.0	10.4
b) Substrate	**SF1**	**SF2**
Acid rocks	0.588	−0.758
Acid deposits	−0.858	0.086
Basic rocks and deposits	0.109	0.927
Poorly developed soils	−0.354	0.546
Soils in an early development stage	0.948	−0.192
Soils with accumulation by illuviation	−0.859	−0.145
Soils with organic matter	0.506	−0.832
Predominantly acid soils	−0.944	0.164
Soils with accumulation of bases	0.334	0.890
Eigenvalue	4.74	2.68
% total variability	52.7	29.8

We assessed the 'Environmental bias' in the distribution of the survey effort at the end of each period of time following Kadmon et al. (2004). Briefly, we compared the environmental distribution of all the records accumulated from the beginning of the survey until the end of each period with a random allocation of the same number of records in the 108 grid cells, by means of a Kolmogorov–Smirnov test. Significant differences between both data sets in the distribution of the effort across each one of the environmental factors are interpreted as a biased survey in such factor. Apart from that, we examined the 'Environmental completeness' of the survey as the percentage of coverage of the environmental conditions provided by the grid cells with more than five records accumulated at each period (Kadmon et al. 2003). We also compared visually the distribution of the environmental conditions provided by these cells to that of all the 108 grid cells. A five-records threshold was chosen to exclude the grid cells which received a few occasional captures, without restricting too much the number of cells that we consider that received some sampling (and thus inflating the possibility of finding different distributions).

Completeness of the observed niche

We used the range occupied by each species in the four environmental factors as a measure of the extent of its environmental niche. This constitutes the simplest description of the environmental niche of the species. In spite of this, we prefer to use it because a slight differences in the shape (but not in the range) of the environmental gradients provided by all the 108 grid cells and the cells with more than five records in the 2004 data used for comparison could compromise the assessment of biases if more complex descriptions of the niche are used, such as differences in the shape of the species response to the environmental gradients. Indeed, if the historical data is not able to provide a good description of the extent of the environmental conditions occupied by the species (below), it is highly likely that the description of more complex definitions of the niche would be even poorer. Therefore, we have restricted our niche analyses to the description of the range of suitable environmental conditions provided by the observed data (i.e. niche completeness sensu Kadmon et al. 2003).

We calculated the 'Environmental niche completeness' provided by the historical data by comparing the extent of the niche described at each period until 1998 (i.e. observed niche; e.g. CF1_obs_1900) with the extent of the niche provided by the SCAMAD 2.1 database, i.e. the niche observed using all the data available in 2004 (e.g. CF1_real), assuming that the inclusion of the latter survey provides a reliable picture of the realized niche of the species (above and Results). First, we estimate the percentage of the realized range of each factor that is covered by the observed data at each period (e.g. [CF1_obs_1900/CF1_real] × 100)

to assess how the observed coverage of the niche increases through time. We also assess the 'Total environmental niche completeness' of each species at each period by multiplying the relative ratios of niche coverage (i.e. percentages) of all the four environmental factors (Kadmon et al. 2003).

Results

The historical curve of the inventory of Madrid dung beetles follows a step-like pattern (Fig. 2). Such pattern is due to the overall increase in the recording rate around 1900, as well as to the pattern of the recording of Aphodiidae species (the family with smaller body size), which present two stages with important recording rates: between 1925 and 1935 (27 new species in 10 years), and between 1973 and 1982 (17 new species). Scarabaeidae and Geotrupidae species, however, accumulate in a more continuous fashion through all the twentieth century.

According to the historical curve, the effort devoted to dung beetle inventory shows four distinctive periods, roughly until 1900, from 1901 to 1935, from 1936 to 1970 and from 1970 until 1998 (Fig. 2). Until 1900 some classic entomologists start to inventory the territory, making some sporadic captures that raised the inventory to 27 species, pertaining to 472 records; these captures were scattered next to the city of Madrid, in the centre-south of the region, and the Guadarrama Mountain Range, which runs next to the edge of the region between the south-western and northernmost corners (Fig. 3). The rate of inventory increases considerably from 1901 to 1935, when 57 new species were discovered (Fig. 2) in spite of the low number of records (280) and the high similarity with the former period in terms of the geographic distribution of the surveys (Fig. 3). This increasing rate of knowledge was truncated by the Spanish Civil War (1936–1939) and the following dictatorship (1939–1975); until 1970, only 17 species were added to the inventory, thanks to some sporadic captures (245 records), mostly in the same areas sampled before. It is only in the 1970s when the rate of inventory increases again thanks to a new generation of entomologists, leaded by Fermín Martín-Piera; until 1998, they raised 3605 new records distributed throughout the entire region (Fig. 3), discovering 32 new species.

The distribution of the records was environmentally biased at the end of all four periods (Table 2). In spite of this, when the cells with five or more records are considered, the surveys seem to cover relatively well the range of environmental conditions present in Madrid after 1998; all cells surveyed before that year accounted for 87% of all environmental variation, slightly less than the 91% of coverage provided by the exhaustive data (including all surveys until 2004) used for comparison (Table 3). Here, it is also remarkable the high coverage of the gradients depicted by SF1 and, especially, CF2 after 1900, although the total degree of environmental completeness stays around 45% until 1970. During all these years, the distribution of the cells with more than five records in the four environmental factors was quite different than the overall conditions in Madrid (Fig. 4). In 1998 both the range of conditions covered by the surveys and the shape of their distribution in the four environmental factors were quite similar to the regional conditions, although both this and the 2004 information present some slight differences with

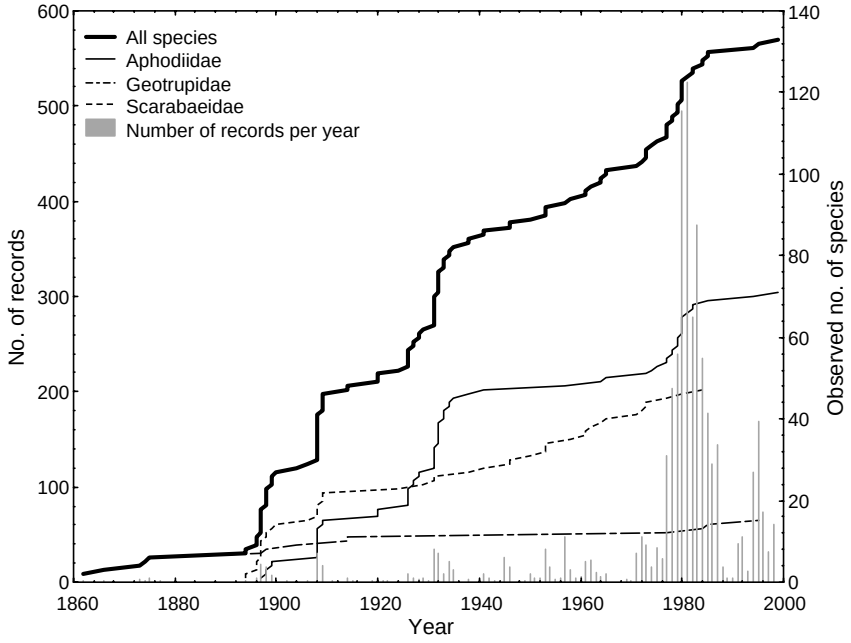

Fig. 2. Historical process of the dung beetle inventory at Madrid. The curves represent the accumulated number of species recorded in Madrid each year from 1860 to 1998, according to the data in SCAMAD 1.0, for all species (continuous thick line), and for the three dung beetle families separately: Aphodiidae (continuous thin line), Scarabaeidae (dotted line) and Geotrupidae (dot-dashed line). The grey columns correspond to the number of records gathered each year. Whenever possible, records without date where assigned to the most likely year according to the period of activity of the recorder.

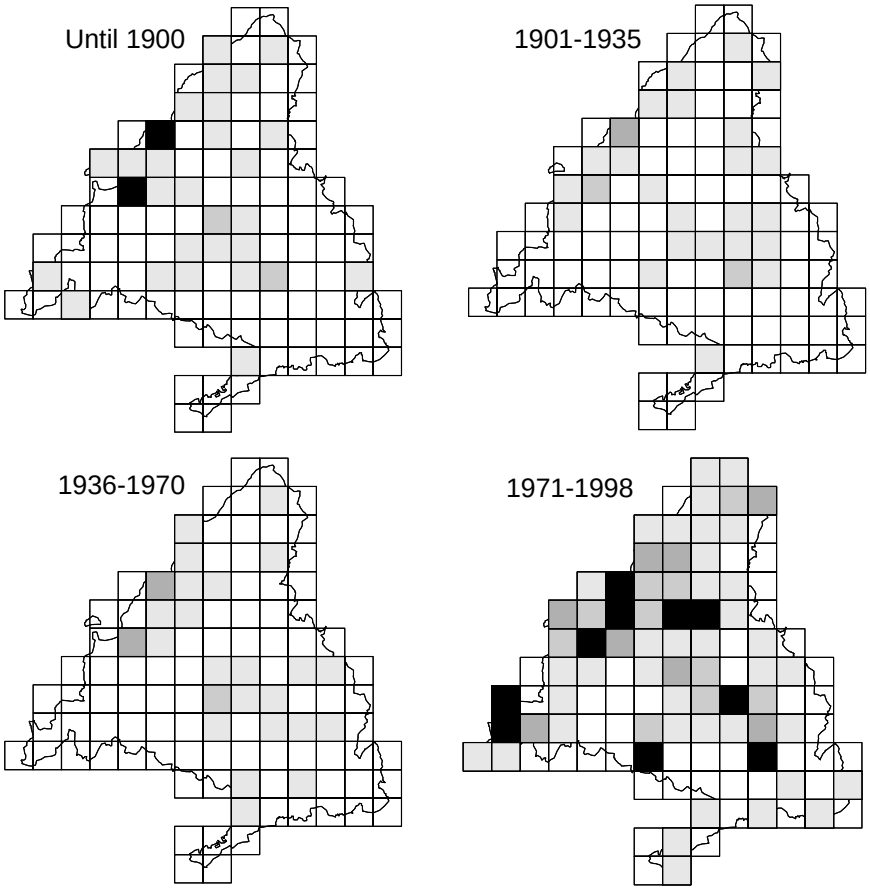

Fig. 3. Geographic distribution of the historical process of dung beetle inventory at Madrid. Each map describes the effort accumulated during each one of the four periods identified (Fig. 2), in shades of grey: no records (white), 1–20 records (light grey), 21–50 records (medium grey), 51–100 records (dark grey) and more than 100 records (deep grey) (Supplementary material Table S1).

Table 2. Environmental bias in the historical survey of Madrid dung beetles. Results from the Kolmogorov–Smirnov test (KS) comparing the distribution in the four environmental factors of the records accumulated until the end of each period, and a random sampling of the same number of records in the 108 grid cells of Madrid territory (Methods).

	Max neg difference	Max pos difference	KS p-level	Mean observed	Mean random	SD observed	SD random	Valid n observed	Valid n random
Until 1900									
FC1	0.000	0.356	$p < 0.005$	1.322	0.151	1.589	0.827	432	432
FC2	−0.373	0.153	$p < 0.001$	−0.399	0.222	1.506	0.991	432	432
FS1	−0.051	0.271	$p < 0.05$	0.090	−0.197	0.986	1.092	432	432
FS2	−0.339	0.051	$p < 0.005$	−0.442	−0.253	0.790	0.835	432	432
Until 1935									
FC1	0.000	0.378	$p < 0.001$	1.264	0.014	1.468	0.958	712	712
FC2	−0.350	0.098	$p < 0.001$	−0.642	−0.017	1.414	1.020	712	712
FS1	−0.085	0.220	$p < 0.001$	0.109	−0.056	1.027	1.011	712	712
FS2	−0.321	0.000	$p < 0.001$	−0.527	−0.028	0.722	0.969	712	712
Until 1970									
FC1	0.000	0.407	$p < 0.001$	1.230	−0.020	1.429	0.973	957	957
FC2	−0.348	0.152	$p < 0.001$	−0.509	0.005	1.454	0.994	957	957
FS1	−0.106	0.236	$p < 0.001$	0.095	−0.020	1.039	1.014	957	957
FS2	−0.348	0.000	$p < 0.001$	−0.544	0.030	0.712	1.007	957	957
Until 1998									
FC1	0.000	0.421	$p < 0.001$	1.146	−0.014	1.306	0.994	4562	4562
FC2	−0.201	0.192	$p < 0.001$	−0.127	0.007	1.375	0.987	4562	4562
FS1	−0.102	0.353	$p < 0.001$	0.422	0.007	0.879	0.997	4562	4562
FS2	−0.443	0.000	$p < 0.001$	−0.679	0.018	0.655	1.009	4562	4562

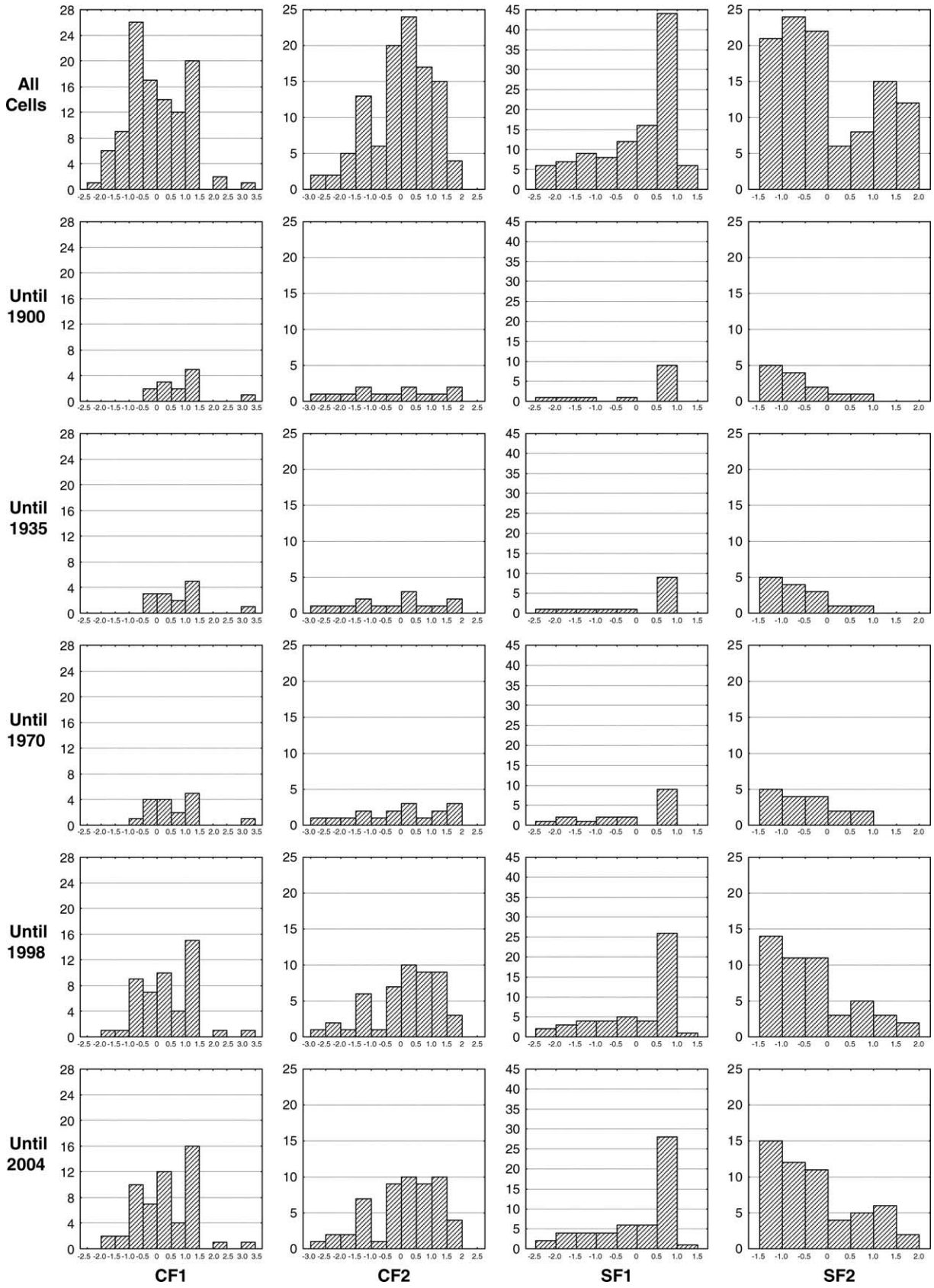

Fig. 4. Distribution in the four environmental factors of the UTM 10 km cells with more than five records accumulated at the end of each period and of all cells.

Table 3. Percentage of coverage of the range of environmental conditions present in Madrid provided by the cells with more than five records accumulated at the end of each period. Total percentage coverage was calculated as the climatic completeness provided by these cells.

Date	CF1	CF2	SF1	SF2	Total
Until 1900	71.92	99.96	85.13	70.35	43.06
Until 1935	74.83	99.96	85.13	70.35	44.80
Until 1970	74.83	99.96	85.13	70.88	45.14
Until 1998	90.60	99.96	97.01	98.93	86.92
Until 2004	94.54	99.96	97.01	98.93	90.69

the distribution of all the cells, especially in the two climatic factors (Fig. 4).

The coverage of the niche of dung beetle species increases highly during the historical process of inventory, especially between 1971 and 1998 (Fig. 5, 6). To avoid problems derived from considering species with highly restricted distributions, we excluded the 18 species that were recorded from only one cell in the validation data (i.e. the 2004 inventory; see Supplementary material Table S2 for detailed information). The number of species with significant percentages of their niche covered shows a spectacular increment (Fig. 5); while in 1900 no species had more of 50% of their niche represented, nearly half of the species considered had the 100% of their niche observed by 1998 (54 out of 113). This pattern is also apparent when the mean percentage of representation is considered; a big step forward in the representation of dung beetle niches is reached between 1971 and 1998, both in general and when considering each environmental factor separately (Fig. 6). The 23 species that were recorded before 1901 (and are present in more than one cell) follow this general pattern, although the knowledge of their niches is, in general, greater, and its increase with time is steeper. In sum, the high recording effort invested from 1970 onwards resulted in a mean representation of 70–80% of the environmental niche of Madrid dung beetles (Fig. 6).

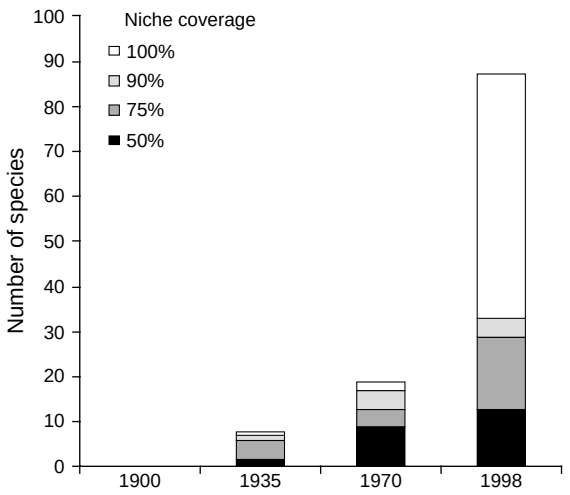

Fig. 5. Variation of the environmental completeness of the niche of Madrid dung beetle species during the historical inventory. Columns represent the number of species with increasing levels of coverage of their niche at the end of each period (Supplementary material Table S2).

However, in spite of such a high mean representation, and of the high similarity between the distribution of the environmental conditions covered by the historical data until 1998 and the validation data (Fig. 4), an important number of species presented significant misrepresentations of their niches. According to the niche completeness measure used, only two thirds (74 species) presented more than 75% of their niches represented by the historical surveys, and nearly a fourth (26 species) had less than 50% of their niche represented (Fig. 5), apart from the 18 species excluded from these analyses (Supplementary material Table S2).

Discussion

Our results show that the gaps and biases of the biodiversity data historically gathered in a non-systematic way are enough to limit the reliability of the observed relationship between species and the environment, even when good taxonomic knowledge and an important amount of distributional data are available.

In spite of its limited size, Madrid synthesizes the biota of all inland Iberia for most groups due to its geographic location and high environmental heterogeneity (Chefaoui et al. 2005, Hortal and Lobo 2005; above). Due to this heterogeneity, the regional inventory and the environmental responses of the species were misrepresented by the non-systematic surveys of classical entomologists, characterized by the repeated sampling of easily accessible places and some classical localities. From the 1970s, the new generation of dung beetle specialists start a 'modern' process of survey, where the aim is to inventory the whole variety of habitats in the region (Fig. 3). As a consequence, the regional inventory was almost complete in 1998, as indicated by the low rates of new species discovery after 1985 (Fig. 2). In fact, after the survey of the first eight cells identified by Hortal and Lobo (2005) between 1998 and 2001 only a new Aphodiidae species (*Aphodius* (*Agrilinus*) *ater*) was added to the previous inventory (Hortal 2004).

The high level of completeness of the regional inventory and the high geographic and environmental coverage of the data compiled during the last period of historical survey up to 1998 (Fig. 3, 4) might lead to the conclusion that the quality of such data would be good enough to represent the patterns in species distributions within the region. However, the additional coverage provided by the post 1998 extensive survey reveals that the knowledge on an important number of species presented significant gaps even in 1998. Ninety-two species increased their known range of distribution (measured as 10 km grid cells) between 1998 and 2004, and three of them even doubled it (*Onthophagus* (*Onthophagus*) *illyricus*, *O.* (*Palaeonthophagus*) *opacicollis* and *A.* (*Bodilus*) *longispina*) (Hortal 2004). Such misrepresentation compromises the reliability of the observed niche for a number of species: leaving out the 18 species that had to be excluded from the analyses due to the lack of data, mean niche completeness was 74.6% (78% for the 23 species already recorded in 1900), roughly a third of the species had less than their 75% of their niches covered by

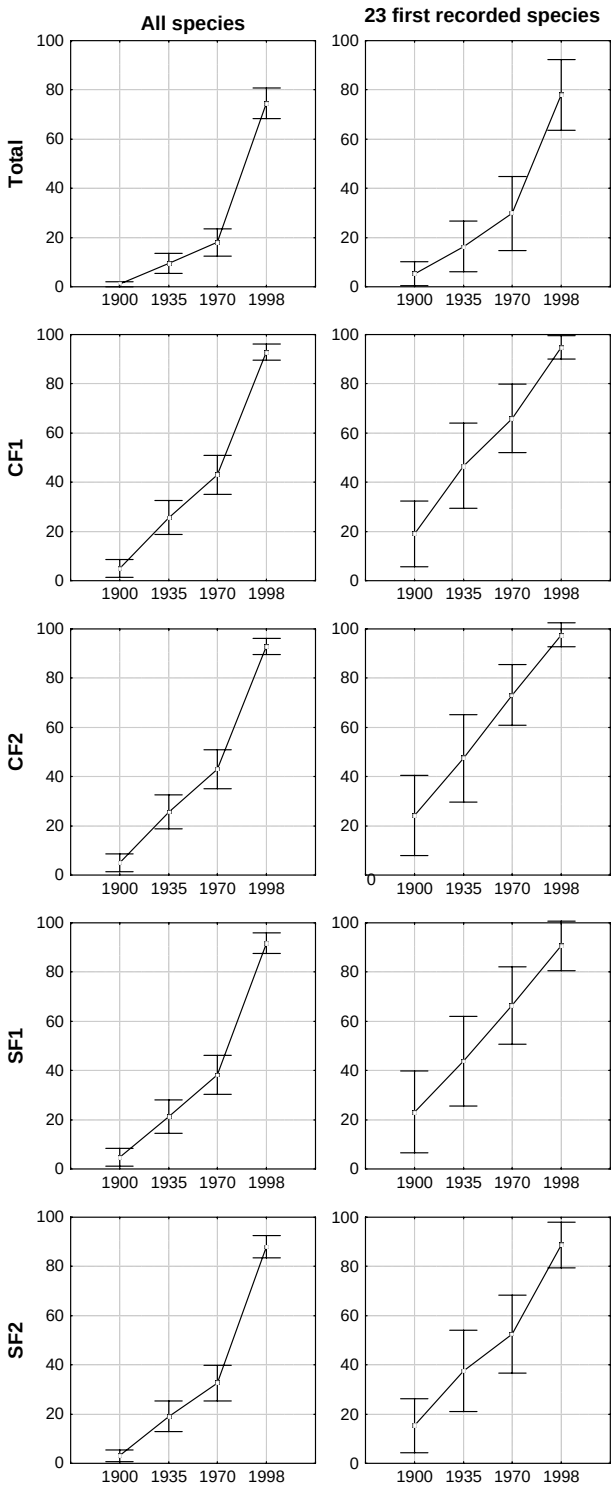

Fig. 6. Variation of the environmental completeness of the niche of Madrid dung beetle species during the historical inventory. Plots represent the mean (±95% CI) of the percentage of coverage of the total niche (and of the four environmental factors considered) for all the species with more than 1 cell observed in 2004, and of the 23 species recorded before 1901 that fulfil such condition.

the data, and nearly a fourth had less than 50% (Fig. 5, 6; Supplementary material Table S2). These effects of environmental bias in the data are likely to be worst if the goal would have been to describe the shape (instead of the range) of the response of the species to environmental gradients, due to the deviation from the shape of these gradients described by the data (Fig. 4).

It could be argued that the increase in observed environmental niche through time is partly due to range shifts produced by climate change. If this hypothesis was true, the increase in range of species in a small region like Madrid should be consistently directed towards higher altitudes (placed in the northwest). However, changes in observed ranges were commonly directed to the southeastern low plains (which were sampled more thoroughly in recent times), or showed no altitudinal trend at all (not shown; see Fig. 3 for the geographic pattern of collecting bias). In addition, although some small changes in mean climatic conditions might have happened during the two centuries of the historical surveys, these changes have been relatively small in comparison with the effect of land use changes in the highly developed region of Madrid. The progressive urbanization of agricultural areas has resulted in the exclusion of some species from parts of their potential distribution (Hortal 2004). If these absences of non-environmental origin have any effect on our results, this effect will be an artificial inflation of the niche completeness measured through time, for the species would not be detected in places where it could potentially host populations even after the exhaustive survey we use for comparison. Indeed, no species were observed to increase their observed niches as a consequence of land use changes (unpubl.). All these facts, together with the relatively large grain size of the geographic units used, make unlikely that the lack of coverage of species niches we report is due to range shifts instead of to a true misrepresentation of their environmental responses. This is especially true for the last period of historical surveys, from which we draw most of our conclusions, given the small temporal difference with the surveys used for comparison.

Although we use a particular case as an example of the effects of the limitations of biodiversity data, they are likely to be the rule rather than the exception. This is especially true for most hyperdiverse groups, which in turn represent most of the species richness of any region. It could be argued that our results are contextual, since they refer only to a particular insect group in a particular region. However, dung beetles are eye-catching, easy-to-collect insects, and there is a long tradition of specialists in the highly-populated region of Madrid, where they have been extensively captured and studied. Thus, they constitute a good example of what could be expected from the information about a taxonomically well-known group in a well-known region. This is certainly not the case for most groups and regions (van Jaarsveld et al. 1998, Dennis et al. 1999, Dennis and Thomas 2000, Hortal and Lobo 2006, Hortal et al. 2007, Soberón et al. 2007), where no effort comparable to the one made from 1970 to 1998 in Madrid has been conducted yet.

If historical surveys provide an unbiased representation of biodiversity variations within the region, the errors in the description of the niche made with these data would be randomly distributed, and accurate predictive maps of species distributions could be obtained. However, this is not the case for most biodiversity data, as exemplified by the environmental bias in the historical survey of dung beetles

in Madrid. Unfortunately, the few databases of outstanding quality available for the British Isles (Prendergast et al. 1993, Griffiths et al. 1999) are an extremely rare case (Hortal et al. 2007). On the contrary, it is well known that historic biodiversity data is taxonomically and geographically biased (Lomolino 2004, Soberón and Peterson 2004, Whittaker et al. 2005). Apart from the obvious differences in the survey effort devoted to different groups (e.g. vertebrates vs arthropods or butterflies vs true bugs), there are important biases within taxonomic groups due to differences in body size and/or showiness (Gaston 1991, Gaston and Blackburn 1994, Dennis and Hardy 1999, Cabrero-Sañudo and Lobo 2003, Diniz-Filho et al. 2005, Gibbons et al. 2005, Jiménez-Valverde and Ortuño 2007). Moreover, species with restricted ecological requirements, such as narrow trophic ranges, are usually described later (Baselga et al. 2007). In addition to this, the geographic patterns of historical inventory are also highly biased across scales: at global extent there are great differences in the effort devoted to different regions and/or countries (Medellín and Soberón 1999, Reed and Boback 2002, Collen et al. 2004, Gibbons et al. 2005, Guil and Cabrero-Sañudo 2007); within these territories, inventories have been spatially biased towards easily-accessible or environmentally attractive areas, some hotspots or the surroundings of the places where taxonomists live and/or work (Dennis et al. 1999, Dennis and Thomas 2000, Martín-Piera and Lobo 2003, Kadmon et al. 2004, Diniz-Filho et al. 2005). As a result, the knowledge on species distributions increases on an environmentally and spatially biased fashion throughout time (Lobo et al. 2007).

The lack of coverage of the environmental responses of species in biodiversity databases can compromise the reliability of the predictive maps of their distributions obtained from these data. Predictive modelling methodologies are widely used to generate predictive maps of current and future species distributions (Guisan and Zimmermann 2000, Guisan and Thuiller 2005, Soberón and Peterson 2005, Araújo and Guisan 2006, Araújo and Rahbek 2006). It is generally thought that current level of development of methods for the predictive modelling of species distributions is enough to provide good predictive maps in spite of the noise in the distributional data used, provided that a minimum set of presence points are available and some specific (and computationally powerful) techniques are used (Elith et al. 2006). Most works assume that presence data offers a good coverage of the environmental response of the species without any quantitative or qualitative assessment (but see Wintle et al. 2005). However, differences in the coverage and distribution of environmental and geographic gradients provided by the data of origin have important effects in the performance of these methods (Kadmon et al. 2003, Segurado and Araújo 2004, Hortal et al. 2007, Lobo 2008a). Due to this, systematic biases in biological data can result in a spatial aggregation of model errors across species (Thuiller et al. 2004a, 2004b, Araújo et al. 2005a, 2005b, Soberón and Peterson 2005, Hortal and Lobo 2006). These errors will concentrate in the areas of the environmental and geographic spectrum that have been less surveyed within the region, so these areas could be underrepresented in the conservation policies developed using predictive maps (Hortal and Lobo 2006). In sum, if models are based in a partial coverage of the environmental niche of the species, their results will be compromised (Araújo et al. 2005a, Hortal et al. 2007, Lobo 2008a).

Our results point out that the use of predictive models might not be the panacea for the incompleteness of the biodiversity data currently available, since the biases in historical surveys result in incomplete descriptions of species niches. Therefore, the quality and biases of the available data must be evaluated before constructing predictive maps of species distributions (Hortal et al. 2007). If these biases are strong, predictive models will not be a substitute for additional surveys (Guisan and Thuiller 2005, Hortal and Lobo 2005, Araújo and Guisan 2006). This point is critical nowadays, when an increasing amount of distributional information is being made widely (and freely) available through the internet, together with some free analytical tools (Rangel et al. 2006, Guralnick et al. 2007). The new GBIF data portal is operative at <http://data.gbif.org/> (from 11 July 2007), providing access to more than 130 million data records from more than 200 institutions scattered over more than 30 countries around the world (see <http://www.danbif.dk/News/news829546/>, posted 1 August 2007). In the light of our results, the potential utility of this huge amount of information for biodiversity conservation might be compromised if the appropriate quality controls are not used.

Acknowledgements – This work is devoted to the memory of Fermín Martín-Piera (1954–2001), mentor and friend, without whose work this one would never have been possible. We will always remember him with a smile. Fermín led the study of Iberian dung beetles until his death in 2001, and encouraged a new generation of biologists to continue their study, including JH and JML. We are also indebted to all the professional and amateur entomologists that captured dung beetles in Madrid and left the vouchers accessible in natural history collections, and in particular to Francisco J. Cabrero-Sañudo, Jose Ignacio López Colón, Carlos M. Veiga, Arturo Baz, Juan J. de la Rosa and Miguel Corra. JH was supported by the UK Natural Environment Research Council. JML and AJ-V were also supported by the Spanish MEC project CGL2004-0439/BOS, a Fundación BBVA Project, and The European Distributed Institute of Taxonomy (EDIT) project.

References

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

Araújo, M. B. and Rahbek, C. 2006. How does climate change affect biodiversity? – Science 313: 1396–1397.

Araújo, M. B. et al. 2004. Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. – Global Change Biol. 10: 1618–1626.

Araújo, M. B. et al. 2005a. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. – Global Ecol. Biogeogr. 14: 17–30.

Araújo, M. B. et al. 2005b. Reducing uncertainty in projections of extinction risk from climate change. – Global Ecol. Biogeogr. 14: 529–538.

Baraud, J. 1992. Coléoptères Scarabaeoidea d'Europe. – Féd. Fr. Soc. Sci. Nat.

Baselga, A. and Novoa, F. 2006. Diversity of Chrysomelidae (Coleoptera) in Galicia, Northwest Spain: estimating the completeness of the regional inventory. – Biodiv. Conserv. 15: 205–230.

Baselga, A. and Novoa, F. 2007. Diversity of Chrysomelidae (Coleoptera) at a mountain range in the limit of the Eurosiberian region, northwest Spain: species richness and beta diversity. – Entomol. Fenn. 18: 65–73.
Baselga, A. et al. 2007. Which leaf beetles have not been yet described? Determinants of the description of Western Palaearctic Aphthona species (Coleoptera: Chrysomelidae). – Biodiv. Conserv. 16: 1409–1421.
Beck, J. A. N. and Kitching, I. J. 2007. Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods. – J. Appl. Ecol. 44: 672–681.
Cabeza, M. et al. 2004. Combining probabilities of occurrence with spatial reserve design. – J. Appl. Ecol. 41: 252–262.
Cabrero-Sañudo, F. J. and Lobo, J. M. 2003. Estimating the number of species not yet described and their characteristics: the case of Western Palaearctic dung beetle species (Coleoptera, Scarabaeoidea). – Biodiv. Conserv. 12: 147–166.
Carpaneto, G. M. et al. 2007. Inferring species decline from collection records: roller dung beetles in Italy (Coleoptera, Scarabaeidae). – Div. Distr. 13: 903–919.
Collen, B. et al. 2004. Biological correlates of description date in carnivores and primates. – Global Ecol. Biogeogr. 13: 459–467.
Chefaoui, R. M. et al. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. – Biol. Conserv. 122: 327–338.
Dennis, R. L. H. 2001. Progressive bias in species status is symptomatic of fine-grained mapping units subject to repeated sampling. – Biodiv. Conserv. 10: 483–494.
Dennis, R. L. H. and Hardy, P. B. 1999. Targeting squares for survey: predicting species richness and incidence of species 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.
Diniz-Filho, J. A. F. et al. 2005. Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. – Global Ecol. Biogeogr. 14: 469–477.
Dolphin, K. and Quicke, D. L. J. 2001. Estimating the global species richness of an incompletely described taxon: an example using parasitoid wasps (Hymenoptera: Braconidae). – Biol. J. Linn. Soc. 73: 279–286.
Edwards, J. L. et al. 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. – Science 289: 2312–2314.
Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.
FAO 1988. Soil map of the World. – FAO/UNESCO.
Fernández Galiano, E. and Ramos Fernández, A. (eds) 1987. La Naturaleza de Madrid. – Comunidad de Madrid, Consejería de Agricultura y Ganadería.
Gaston, K. J. 1991. Body size and probability of description: the beetle fauna of Britain. – Ecol. Entomol. 16: 505–508.
Gaston, K. J. and Blackburn, T. M. 1994. Are newly described bird species small-bodied? – Biodiv. Lett. 2: 16–20.
Gibbons, M. J. et al. 2005. What determines the likelihood of species discovery in marine holozooplankton: is size, range or depth important? – Oikos 109: 567–576.
González, J. et al. 2007. Diversity of water beetles (Coleoptera: Gyrinidae, Haliplidae, Noteridae, Hygrobiidae, Dytiscidae and Hydrophilidae) in Galicia, northwest Spain: estimating the completeness of the regional inventory. – Coleopt. Bull. 61: 95–110.
Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.
Grenyer, R. et al. 2006. Global distribution and conservation of rare and threatened vertebrates. – Nature 444: 93–96.
Griffiths, G. H. et al. 1999. Integrating species and habitat data for nature conservation in Great Britain: data sources and methods. – Global Ecol. Biogeogr. 8: 329–345.
Guil, N. and Cabrero-Sañudo, F. 2007. Analysis of the species description process for a little known invertebrate group: the limnoterrestrial tardigrades (Bilateria, Tardigrada). – Biodiv. Conserv. 16: 1063–1086.
Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Modell. 135: 147–186.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Guralnick, R. P. et al. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. – Ecol. Lett. 10: 663–672.
Hortal, J. 2004. Selección y Diseño de Áreas Prioritarias de Conservación de la Biodiversidad mediante Sinecología. Inventario y modelización predictiva de la distribución de los escarabeidos coprófagos (Coleoptera, scarabaeoidea) de Madrid. – Depto de Biología, Facultad de Ciencias, Univ. Autónoma de Madrid, p. 333.
Hortal, J. and Lobo, J. M. 2005. An ED-based protocol for the optimal sampling of biodiversity. – Biodiv. Conserv. 14: 2913–2947.
Hortal, J. and Lobo, J. M. 2006. Towards a synecological framework for systematic conservation planning. – Biodiv. Inform. 3: 16–45.
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.
Hortal, J. et al. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). – Conserv. Biol. 21: 853–863.
ITGE 1988. Atlas Geocientífico y del Medio Natural de la Comunidad de Madrid. – Inst. Tecnológico GeoMinero de España.
Jiménez-Valverde, A. and Ortuño, V. M. 2007. The history of endemic Iberian ground beetle description (Insecta, Coleoptera, Carabidae): which species were described first? – Acta Oecol. 31: 13–31.
Kadmon, R. et al. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecol. Appl. 13: 853–867.
Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.
Legendre, P. and Legendre, L. 1998. Numerical ecology. – Elsevier.
Lobo, J. M. 2001. Decline of roller dung beetle (Scarabaeinae) populations in the Iberian peninsula during the 20th century. – Biol. Conserv. 97: 43–50.

Lobo J. M. 2008a. More complex distribution models or more representative data? – Biodiv. Inform. 5: 14–19.
Lobo, J. M. 2008b. Database records as a surrogate for sampling effort provide higher species richness estimations. – Biodiv. Conserv. 17: 873–881.
Lobo, J. M. and Martín-Piera, F. 1991. La creación de un banco de datos zoológico sobre los Scarabaeidae (Coleoptera: Scarabaeoidea) íbero-baleares: una experiencia piloto. – Elytron 5: 31–38.
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. 2007. How does the knowledge on the spatial distribution of species increase? – Div. Distr. 13: 772–780.
Lomolino, M. V. 2004. Conservation Biogeography. – In: Lomolino, M. V. and Heaney, L. R. (eds), Frontiers of biogeography: new directions in the geography of nature. Sinauer, pp. 293–296.
López-Colón, J. I. 2000. Familia Geotrupidae. – In: Martín-Piera, F. and López-Colón, J. I. (eds), Coleoptera, Scarabaeoidea I. Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, pp. 107–183.
Martínez-Meyer, E. 2005. Climate change and biodiversity: some considerations in forecasting shifts in species potential distributions. – Biodiv. Inf. 2: 42–55.
Martín-Piera, F. 2000. Familia Scarabaeidae. – In: Martín-Piera, F. and López-Colón, J. I. (eds), Coleoptera, Scarabaeoidea I. Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, pp. 207–432.
Martín-Piera, F. and Lobo, J. M. 2003. Database records as a sampling effort surrogate to predict spatial distribution of insects in either poorly or unevenly surveyed areas. – Acta Entomol. Iber. Macaronesica 1: 23–35.
Medellín, R. A. and Soberón, J. 1999. Predictions of mammal diversity on four land massess. – Conserv. Biol. 13: 143–149.
Prendergast, J. R. et al. 1993. Correcting for variation in recording effort in analyses of diversity hotspots. – Biodiv. Lett. 1: 39–53.
Rangel, T. F. L. V. B. et al. 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. – Global Ecol. Biogeogr. 15: 321–327.
Reed, R. N. and Boback, S. M. 2002. Does body size predict dates of species description among North American and Australian reptiles and amphibians? – Global Ecol. Biogeogr. 11: 41–47.
Reutter, B. A. et al. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. – J. Biogeogr. 30: 581–590.
Rivas-Martínez, S. 1987. Memoria del Mapa de Series de Vegetación de España. – Ministerio de Agricultura, Pesca y Alimentación.
Saarenmaa, H. and Nielsen, E. S. (eds) 2002. Towards a global biological information infrastructure. Challenges, opportunities, synergies, and the role of entomology. – Eur. Environ. Agency.
Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.
Soberón, J. and Peterson, T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. – Philos. Trans. R. Soc. Lond. B 359: 689–698.
Soberón, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distribution areas. – Biodiv. Inf. 2: 1–10.
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: 1441–1446.
Soberón, J. et al. 2007. Assessing completeness of biodiversity databases at different spatial scales. – Ecography 30: 152–160.
Thuiller, W. et al. 2004a. Uncertainty in predictions of extinction risk. – Nature 430: doi:10.1038/nature02716.
Thuiller, W. et al. 2004b. Effects of restricting environmental range data to project current and future species distributions. – Ecography 27: 165–172.
van Jaarsveld, A. S. et al. 1998. Biodiversity assessment and conservation strategies. – Science 279: 2106–2108.
Veiga, C. M. 1998. Los *Aphodiinae* (*Coleoptera*, *Aphodiidae*) Ibéricos. PhD thesis. – Facultad de Ciencias Biológicas. Depto de Biología Animal I. Univ. Complutense de Madrid.
Whittaker, R. J. et al. 2005. Conservation biogeography: assessment and prospect. – Div. Distr. 11: 3–23.
Wilson, E. O. 2002. The future of life. – Alfred A. Knopf.
Wintle, B. A. et al. 2005. Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. – Austral Ecol. 30: 719–738.

Supplementary material (available online as Appendix O16434 at www.oikos.ekol.lu.se/appendix). Table S1: Information on the 108 UTM 10 km grid cells of Madrid. Table S2: Environmental completeness of the observed niche of all Madrid dung beetle species during the historical surveys.