BIODIVERSITY RESEARCH

How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time?

Jorge M. Lobo[1]*, Andrés Baselga[1], Joaquín Hortal[1,2], Alberto Jiménez-Valverde[1] and Jose F. Gómez[1]

[1]*Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales – CSIC, c/José Gutiérrez Abascal, 2, 28006 Madrid, Spain,* [2]*NERC Centre for Population Biology, Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK*

*Correspondence: Jorge M. Lobo, Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, C/José Gutiérrez Abascal, 2. 28006, Madrid, Spain. Tel.: 34 +91 +4111328; Fax: 34 +91 +5645078; E-mail: mcnj117@mncn.csic.es.

ABSTRACT

Different distribution maps can be obtained for the same species if localities where species are present are mapped at different times. We analysed the accumulation of information over time for a group of dung beetle species in the Iberian Peninsula. To do this, we used all available information about the distribution of the group as well as data on selected species to examine if the process of discovery of species distribution has occurred in a climatically or spatially structured fashion. Our results show the existence of a well-defined pattern of temporal growth in distributional information; due to this, the date of capture of each specimen can be explained by the environmental and spatial variables associated to the collection sites. We hypothesize that such temporal biases could be the rule rather than the exception in most distributional data. These biases could affect the weighting of environmental factors that influence species distributions, as well as the accuracy of predictive distribution models. Systematic surveys should be a priority for the description of species geographical ranges in order to make robust predictions about the consequences of habitat and climate change for their persistence and conservation.

Keywords

Biodiversity databases, collecting biases, distribution maps, dung beetles, Iberian Peninsula, spatial distribution, species distribution ranges.

INTRODUCTION

Although the records comprising species distribution maps often differ in spatial resolution, the overall geographical distribution of organisms is generally symbolized by dots of equal size (area of occupancy) or by continuous patches covering the area of all known presence sites (extent of occurrence; see Rapoport, 1982; Gaston, 1991; or Burgman & Fox, 2003). Different distribution maps can be obtained if the localities where a species has been recorded are represented at different times. These temporal changes in species ranges could be due to (i) the specific population dynamics of the species (Hengeveld, 1990; Parmesan *et al*., 2005) and (ii) the influence of environmental changes (Thomas & Lennon, 1999; Parmesan & Yohe, 2003). However, these changes could also be (iii) erroneous representations resulting from the existence of a defined spatial pattern in the newly added distributional information. It is well known that geographical ranges enlarge as sampling effort increases (Gaston, 2003). However, as far as we know, no study has yet analysed the variation in geographical representations of species as distributional information increases, i.e. the growth of distribution maps over time.

While new presences are being recorded during a given period, information on the distribution of a given species could accumulate either in a random or in a spatially structured fashion (assuming that its range is relatively stable within such period). If knowledge increases at will, the probability of surveying a given site would not be conditioned by the spatial position of previously sampled sites; as a result, the 'true' distribution of species would be gradually and uniformly revealed across the whole area, and discoveries of new presence sites would generally only sharpen the observed pattern rather than modify it. Of course, in the first survey stages, a random placement of sampling sites can imply an expansion of the known area of distribution, but such expansion must quickly reach a maximum. However, if surveys occur in a spatially structured fashion, the derived distribution maps would differ with time, and the 'true' distribution would not be homogenously revealed across the territory. Such a bias can be produced by the influence of the environmental and/or geographical characteristics of the sites on survey site allocation and/or by several sociological factors (see, e.g. Dennis & Thomas, 2000). In the first scenario, the information on the species range will be independent of the moment of time when the maps are created, while in the second it would not.

DOI: 10.1111/j.1472-4642.2007.00383.x

Many studies have explored the environmental and geographical determinants of species distributions (see, e.g. Rahbek & Graves, 2001; Ricklefs, 2004; Field *et al*., 2005; or Hawkins *et al*., 2005); many others have elaborated hypotheses on the potential or 'real' distribution of species based on environmental explanatory variables (see Brotons *et al*., 2004; Engler *et al*., 2004; Iverson *et al*., 2004; Soberón & Peterson, 2005). However, others have recognized that available species distribution information is influenced by uneven sampling and by recorder bias (Dennis *et al*., 1999; Dennis & Thomas, 2000; Zaniewski *et al*., 2002; Reutter *et al*., 2003; Graham *et al*., 2004; Martínez-Meyer, 2005; Romo *et al*., 2006; Hortal *et al*., in press). The potential effect of such a bias on (i) the growth of distribution knowledge, (ii) the relative weighting of environmental factors affecting species distributions, and (iii) the accuracy of predictive distribution models is yet to be examined. Indeed, current knowledge of species distributions is not usually complete; uneven discovery processes could reveal erroneous spatial patterns that might result in unreliable predictive distribution models.

In this paper, we study the spatial pattern of the growth of distributional information for a dung beetle family (Coleoptera, Scarabaeidae) in the Iberian Peninsula. We use all the available information about the distribution of Scarabaeidae species as whole, as well as specific information on the most recorded species. After describing the temporal variation in the number of database records and collected species, we specifically examine if the geographical variation in the year of each database record can be explained by spatial or climatic variables. In that case, the temporal growth of the information about the distribution of Iberian Scarabaeidae has occurred following a biased pattern. Our null hypothesis is that the year of collection of each database record is randomly distributed across spatial and climate gradients, so the species distribution maps produced over time provide unbiased cartographic representations of the true distribution of dung beetle species. The null hypothesis will be rejected if a meaningful and statistically significant relationship between the year of collection and the spatial or climate variables is found.

Two other different (and unrelated) processes can also explain such spatially structured collection process: (i) the tendency of taxonomists to collect rare species, and (ii) the climate change during the last century (Watson & the Core Writing Team, 2002). Since rare species are more interesting for taxonomists, the localities hosting these species could be expected to be sampled earlier and, comparatively, more thoroughly. We test this hypothesis by examining if the geographical rarity of species is related with their yearly rate of increase in the number of database records. On the other hand, if any detected spatial and/or environmental patterns were due to climate change-driven range shifts instead of sampling bias, it should be expected that geographical range shifts would differ according to the particular climatic adaptations of species. In a warming scenario, warm-adapted species would increase their ranges, while cold-adapted species would diminish them. After analysing these questions we highlight the implications of survey biases for the study of species distributions.

METHODS

The information on the distribution of Scarabaeidae species comes from BANDASCA (Lobo & Martín-Piera, 1991). This database compiles all available taxonomic and distribution information from museums, private collections, published and unpublished data for each of the 53 Scarabaeidae dung beetle species known to inhabit the Iberian Peninsula (Martín-Piera, 2000). Each database record contains information on the pool of specimens of a single species with identical database field values: site, elevation, date of capture, type of habitat and food source (see Martín-Piera & Lobo, 2003). All locality data have been correctly georeferenced at a 10 × 10 km UTM resolution. At present, this database has 13,570 records with known year of collection, which extend from 1872 to 2001. The year of recording was used as the dependent variable to describe the process of the growth in knowledge of dung beetle distribution.

Several spatial and climate variables (see below) were used as possible predictors of this process. Broad-scale spatial structure in the year of collection was described by means of the nine terms of a third degree polynomial of the central latitude (Lat) and longitude (Lon) of each 100 km^2 UTM cell (Trend Surface Analysis or TSA; see Legendre & Legendre, 1998). TSA yields an estimate of the large-scale trends in a spatially distributed dependent variable, using a regression analysis of the dependent variable to separate systematic variation (i.e. explained by the spatial variables) from random variation (due to measurement error or to the effect of other variables not included in the analysis). Latitude and longitude were standardized to 0 mean and 1 standard deviation.

The climate variables used as predictors were the seasonal precipitation scores and maximum, minimum, and mean temperatures (16 climate variables). These variables were provided by the Spanish Instituto Nacional de Meteorología and the Portuguese Instituto de Meteorologia and were handled using the IDRISI KILIMANJARO software (Clark Labs, 2003). Their scores at each 100 km^2 UTM square in the Iberian Peninsula ($n = 6063$) were extracted, normalized and standardized (to 0 mean and 1 standard deviation) and then submitted to a Principal Component Analysis (PCA) to obtain uncorrelated factors (Varimax rotation). The two first climate factors extracted with the PCA were able to explain 69% and 22% of total climate variability, respectively; the first factor characterizes a Mediterranean gradient, with high positive loadings (> 0.8) of spring, winter, and autumn temperatures. The second factor identifies an aridity gradient with high and negative loadings for spring, winter, and autumn rainfall and positive ones for maximum summer temperatures (not shown).

The broad-scale structure in the year of collection related to environmental variables was assessed by regressing it against the two above-mentioned PCA factors, selecting the statistically significant terms (p to enter and p to remove = 0.05; intercept included) by a standard stepwise backward regression procedure using generalized linear models (normal distribution of errors and a logarithmic link function between the dependent and predictors). To account for curvilinear relationships, the quadratic and cubic functions of each PCA climate factor and the interaction term between both of them were included in the regression model (see Austin, 1980). In a spatially structured environment, nearby localities tend to have similar environmental conditions, so a spatially structured pattern can be explained at the same time by different processes making it difficult to discern causation from correlation patterns (Wagner & Fortin, 2005). However, these concerns do not affect to our study; we only aim to determine if the year of collection can be described using environmental or spatial variables, without assuming any causal relationships.

Most dung beetle species were recorded fewer than 200 times (33 species; 61% of total species) with a mean (± SE) number of database records per species of 251 ± 39. The regression models were calculated for all species included in BANDASCA altogether and also individually for each of the 14 species with a number of database records above the upper quartile boundary (361 database records). The STATISTICA package (StatSoft, 2003) was used for all computations.

RESULTS

Temporal variation in sampling effort

The mean number of database records per year was 105 ± 19 (mean ± SE), but this figure fluctuates widely over the period of time considered. The rate of Iberian dung beetle information increases considerably from the 1970s onwards (371 ± 61 database records per year), reaching its maximum in the 1980s and subsequently declining until the present (Fig. 1). Only 343 database records were available before 1900 (2.5% of the total), and in the following 70 years, the number of database records reached only 2424 (17.9% of the total). The peak in the number of database records per year occurs in the year 1980, roughly the year when the last Scarabaeidae species was added to the Iberian inventory, which reached a plateau (Fig. 1). The number of database records per year is positively correlated with the number of collected species (Spearman rank correlation coefficient; $r_s = 0.968$, $P < 0.001$). This implies that increases in sampling effort lead to increases in the number of recorded species.

The geographical pattern of variation in the year of collection can be exemplified by the case of *Onthophagus (Palaeonthophagus) fracticornis* (Fig. 2). The observed distribution range of this species has been gradually increasing. At the 1930s, the species was only known from some isolated nuclei, mainly placed in the periphery of its actual distribution. The gaps in such distribution have been progressively filled during the twentieth century, and now the species is known to occur in most Iberian mountain ranges, showing a more continuous distribution range.

Spatial and climatic bias

The geographical variation in the year of each database record seems to be spatially structured (Fig. 3a); the oldest citations are restricted principally to the north-eastern Iberian corner, and, to a lesser extent, to the southern and western limits of the Iberian Peninsula. The existence of a spatial pattern is confirmed by TSA results (Fig. 3b). Seven of the nine terms of a third degree polynomial of latitude and longitude remained significant in a backward stepwise analysis: the three terms of the cubic function of longitude, latitude, the interaction between latitude and longitude, and the two other interaction terms (latitude^2 × longitude and longitude^2 × latitude). A function of these seven terms explains 18.6% of the variation in the year of database records. Climate variables explain around 13.1% of the variation in the year of database records (quadratic function of the first PCA climate factor and cubic function of the second PCA climate factor).

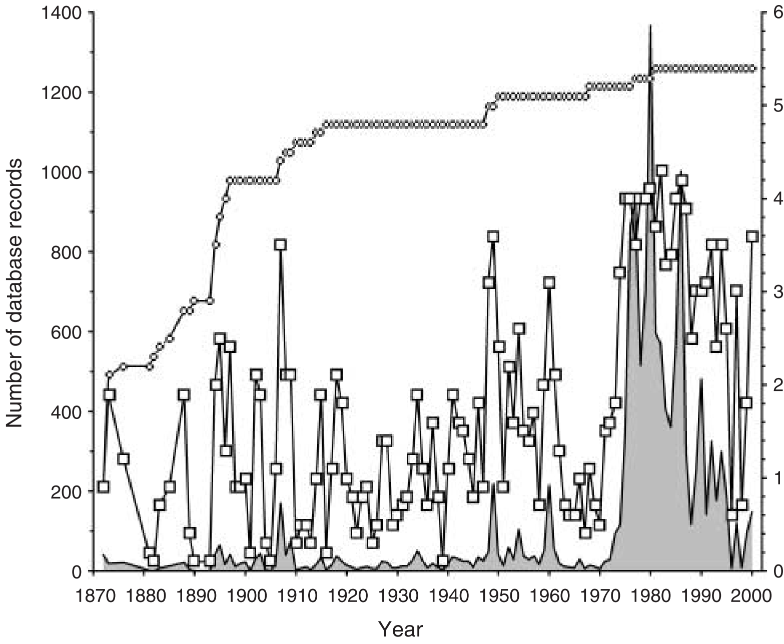

Figure 1 Yearly variation in the number of database records (shaded area) and number of collected species (squares) in the BANDASCA database, which compiles all available information on Iberian dung beetles. Circles represent the number of species accumulated in the Iberian inventory.

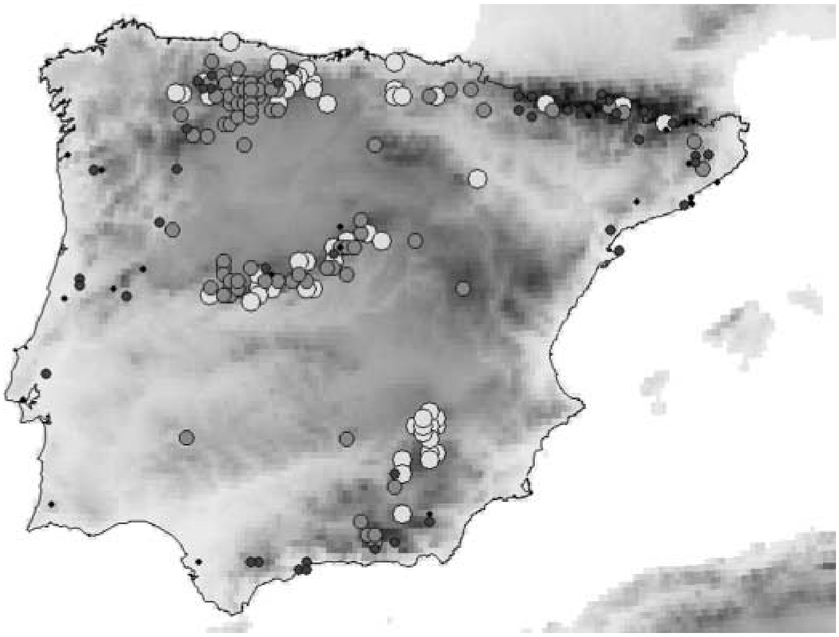

Figure 2 Increase over time and geographical changes in the known distribution area of *Onthophagus fracticornis* in the Iberian Peninsula. The database records gathered in four consecutive time periods are represented by plots of decreasing shades of grey and increasing sizes: before 1930 (black dots), from 1931 to 1960 (dark grey, small circles), from 1960 to 1980 (medium grey, bigger circles), and from 1981 to 2000 (light grey, biggest circles).

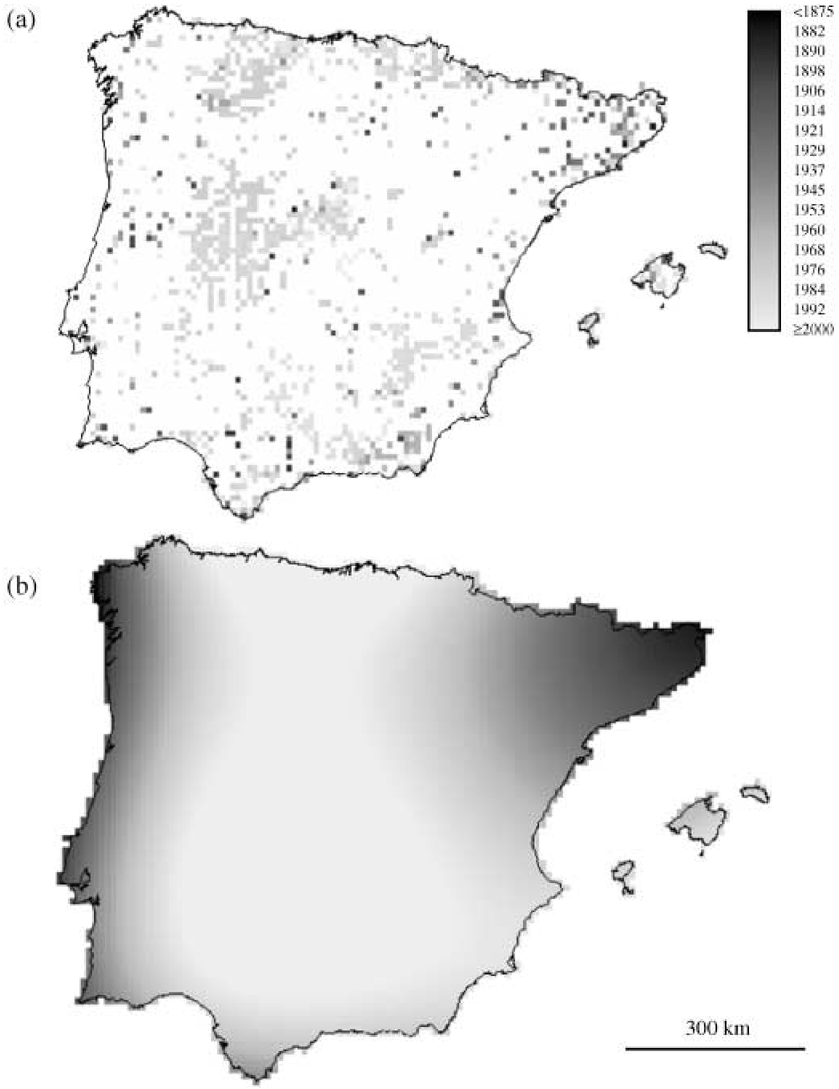

Figure 3 Spatial distribution of the year of collection in BANDASCA database records, according to (a) the original data (only the first year of collection at each 100 km^2 square is shown), and (b) the scores predicted by a Trend Surface Analysis.

Table 1 Number of database records (*N* records) in BANDASCA (a database on the distribution of Iberian dung beetles), number of 10 × 10 km UTM cells in which each species has been observed, and percentage of deviance explained by the third-degree polynomial of central latitude and longitude (Trend Surface Analysis, Legendre & Legendre, 1998), and by the two main Principal Component Factors that represent the climate variation across the Iberian 100 km^2 UTM cells. AMT is the annual mean temperature of the UTM cells in which each species was observed (± 95% confidence interval), while the last two columns are the yearly rate (± SD) of database increase before (< 1975) and after 1975 (> 1975).

Species	*N* records	*N* UTM cells	Spatial variables	Climate variables	AMT	< 1975	> 1975
Onthophagus (Onthophagus) taurus (Schreber, 1759)	1239	170	17.11%	3.51%	13.61 ± 0.26	2.57 ± 0.11	36.68 ± 2.01
Onthophagus (Palaeonthophagus) similis (Scriba, 1790)	1121	128	15.15%	6.31%	11.95 ± 0.29	1.21 ± 0.04	38.80 ± 2.54
Onthophagus (Palaeonthophagus) vacca (Linnaeus, 1767)	903	148	24.02%	7.14%	13.31 ± 0.29	0.76 ± 0.03	31.53 ± 2.22
Onthophagus (Furconthophagus) furcatus (Fabricius, 1781)	870	134	30.46%	11.10%	13.46 ± 0.27	2.57 ± 0.11	36.68 ± 2.01
Euoniticellus fulvus (Goeze, 1777)	801	121	29.48%	5.78%	13.32 ± 0.34	0.92 ± 0.03	24.50 ± 1.97
Bubas bubalus (Olivier, 1811)	596	101	34.01%	17.22%	14.31 ± 0.32	0.58 ± 0.01	16.87 ± 2.05
Caccobius schreberi (Linnaeus, 1767)	529	110	23.11%	4.32%	12.88 ± 0.34	0.94 ± 0.02	15.78 ± 1.48
Copris lunaris (Linnaeus, 1758)	512	86	32.71%	8.57%	11.94 ± 0.30	1.22 ± 0.03	12.10 ± 1.61
Copris hispanus (Linnaeus, 1764)	511	109	20.27%	9.57%	15.45 ± 0.25	0.74 ± 0.01	16.71 ± 1.25
Euonthophagus amyntas (Olivier, 1789)	482	108	43.91%	11.36%	13.20 ± 0.36	1.57 ± 0.07	10.96 ± 1.31
Bubas bison (Linnaeus, 1767)	436	81	20.88%	10.85%	15.96 ± 0.23	0.48 ± 0.01	16.57 ± 1.23
Onthophagus (Palaeonthophagus) fracticornis (Preyssler, 1790)	429	76	44.63%	32.63%	10.42 ± 0.43	1.28 ± 0.05	12.03 ± 1.05
Onthophagus (Parentius) punctatus (Illiger, 1803)	363	86	38.14%	21.46%	13.71 ± 0.41	0.62 ± 0.01	11.90 ± 0.73
Onthophagus (Palaeonthophagus) opacicollis (Reitter, 1892)	361	77	21.94%	11.83%	14.47 ± 0.41	0.17 ± 0.01	15.25 ± 0.56

Adding these climate factors to the spatial terms does not lead to a better explanatory function.

The year of database records of the 14 most recorded species can also be jointly explained by spatial and climate variables, accounting for 22.0% and 5.4% of total variability, respectively. The explanatory capacity of spatial variables for each one of these 14 species ranges from 15.1% to 44.6% (28.3% ± 2.5), and from 3.5% to 32.6% in the case of climate variables (11.5% ± 2.1; see Table 1). Thus, spatial variables seem to account for a significantly higher percentage of temporal variability (Wilcoxon matched pairs test = 3.29; $n = 14$; $P < 0.001$), although the percentages of variability explained by both kinds of variables are positively correlated (Spearman rank correlation test; $r_s = 0.67$; $n = 14$; $P < 0.01$).

Rarity and survey effort

The temporal variation in the accumulated number of database records for these 14 species (Fig. 4) indicates that the yearly rate of increase in the number of database records is generally slow before 1975, increasing noticeably after this date (see Table 1). However, both rates are uncorrelated ($r_s = 0.10$; $n = 14$; $P = 0.74$). Before 1975, the number of UTM squares where each species was observed is only slightly correlated with the rate of increase in database records ($r_s = 0.45$; $n = 14$; $P = 0.10$); however, after that date these two figures are highly and positively correlated ($r_s = 0.78$; $n = 14$; $P = 0.001$). Interestingly, although the number of UTM squares in which each species was observed is not correlated with the variability explained by spatial bias ($r_s = –0.44$; $n = 14$; $P = 0.12$), it is negatively correlated with the climatic bias ($r_s = –0.72$; $n = 14$; $P = 0.003$).

Climatic adaptations and survey effort

The mean annual temperature for the UTM cells in which each species was observed (Table 1) is not significantly correlated with the magnitude of the spatial or climatic bias ($r_s = –0.32$ $n = 14$; $P = 0.26$ and $r_s = 0.17$; $n = 14$; $P = 0.55$, respectively). Therefore, the observed biases are not related to possible climate change-driven shifts in species distributions.

DISCUSSION

Temporally changing biases in Iberian dung beetle data

According to our results, survey effort seems to follow a defined pattern both in time and in space. The amount of collection effort varies over time; most collections have been made once taxonomic knowledge is almost complete (around the year 1975), probably because this knowledge facilitates the gathering of faunistic information. The surveys of Iberian dung beetles also follow a well-defined spatial pattern; the earliest faunistic studies were carried out in Portugal and Catalonia, while the inner Iberian Peninsula was sampled later. Such temporal patterns in the areas chosen by taxonomists produce an evident spatial structure in collection dates (see Fig. 2). Nearby sites usually yield similar collection dates, probably because each taxonomist tends to sample only in a limited territory (see Dennis *et al.*, 1999). Interestingly, when the information available for each species is analysed separately, sampling bias is more evident than when data on all species are studied together. We suspect that this fact is due to the masking effect of overlaying and analysing information

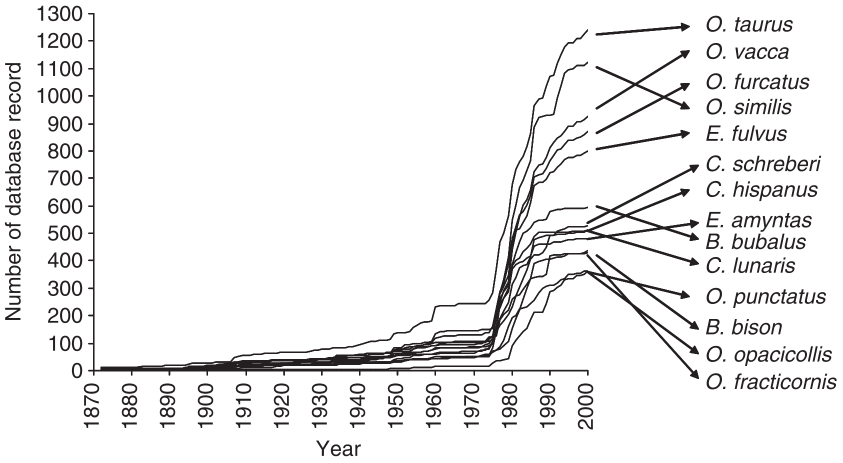

Figure 4 Accumulated increase over time of the number of database records in BANDASCA for the 14 most recorded species (i.e. those with a number of database records above the upper quartile boundary, 361 database records). The species are ordered in the right column according to the total number of 10 × 10 km UTM cells in which each species was observed (see Table 1).

about species with different kinds of climatic tolerances and distributions (see Table 1).

The spatial structure in collection dates is a consequence of the variation over time of the geographical and/or climatic conditions of the areas chosen by taxonomists for their surveys. Due to these temporal biases, the differences in collection dates between coastal and inland localities are also reflecting one of the main environmental gradients of the Iberian Peninsula (i.e. precipitation is lower and temperatures much warmer at the hinterland than at the coast). Despite this, climate variables explain on average a lower percentage of variability in the year of collection than spatial variables, and they do not add any explanatory power to the TSA function. This indicates that differences in sampling effort over time are not a response to climate conditions *per se*; instead, they are the result of sociological and scientific processes related to variations in the preferences and goals of specialists over time (Martín-Piera & Lobo, 2003).

The inclination of taxonomists to collect rare species could also stand as a partial explanation of the observed temporal bias. The most geographically rare species have received a more intense survey effort after 1975; as a result, the effort devoted to each individual species before and after such date is uncorrelated. Recent taxonomists have changed their collection trends, seeking rarer species in places that are climatically and spatially far away from classic localities. As a result, the known distribution range of all species has increased. Here, we can discard the alternative hypothesis of climate change having a significant effect on the detected temporal trend, since the temporal biases in survey effort are uncorrelated with the climatic adaptations of the species); warm-adapted species do not increase their ranges more than cold-adapted ones.

If dung beetles are one of the most-studied insect groups in the Iberian Peninsula, why is the information about their distribution so poor and biased yet? The most obvious explanation is that taxonomists do not have the resources to survey everywhere and also that their efforts will always be biased in some way. The community of Iberian dung beetle specialists has not given a high priority to the development of accurate atlases while ‘planning’ their faunistic effort. On the contrary, they tend to survey areas close to their residences or in particularly interesting territories (see Dennis & Thomas, 2000), as they are more interested in collecting remarkable or rare species than in offering a comprehensive picture of the distribution of a group of organisms in a territory (Soberón & Peterson, 2004). Some sources of bias in historical surveys (e.g. recorder’s home-range or repeated sampling in several classical localities) are likely to be common in most groups. We suspect that the drawbacks in the distributional information currently at hand highlighted by the example of the Iberian dung beetles are likely to be the rule rather than the exception for other groups and regions. Therefore, the process of unravelling species distributions over time should be considered as an unstandardized sampling protocol that is conducted by several specialists with often unrelated aims, and that results in spatially structured information.

The effect of survey biases on distributional hypotheses

Species distribution ranges are difficult to define because they change spatially and temporally, particularly at their borders (Parmesan *et al*., 2005). Furthermore, it is sometimes difficult to decide when a species is present at a site, not to mention when it is truly absent. Should those unsuitable sites where the species has been collected but is unable to persist in the absence of continued immigration be considered presence plots? (Pulliam, 1988, 2000). How should these records be classified and incorporated into distribution maps? If some collected specimens of a species can be accurately designated as ‘dead alive’ or ‘vagrants’ (Gaston, 2003), is it worth to incorporate their information as an indication of the location of its distribution edge? In the same way, although we can define the sites in which a species is truly absent, it is necessary to consider that only a fraction of suitable sites are occupied due to metapopulational dynamics and dispersal limitations (Pulliam, 2000). As a result, mapped distributions should be considered as probabilistic maps with an unknown degree of uncertainty.

These problems are inherent to the delimitation of geographical ranges, but are aggravated as consequence of the gaps in the data.

Many different modelling techniques have been proposed to fill in these gaps by interpolating and extrapolating the known distribution to the territories without enough biological information (see Elith *et al*., 2006). Typically, these techniques provide statistically derived distribution hypotheses using climatic and spatial predictors. Although these hypotheses could be used in the (common) absence of exhaustive data, modelling techniques need data of relatively good quality to produce reliable hypotheses. Could the temporally structured bias in the recording process be affecting the reliability of predicted distributions?

Our results show that the differences between distribution maps taken at different time points are climatically and spatially biased. Therefore, regardless of the origin of sampling biases, the representation of the environmental gradients by the surveyed localities will differ over time. This compromises the reliability and comparability of the distributional hypotheses developed from these variables (e.g. predictive models or range shift assessments). Therefore, ignoring the effects of the spatial bias in historical sampling could lead to misleading conclusions about the role of climate change in distributional changes; spurious increases in species distribution ranges produced by new surveys in previously unsampled areas could be erroneously attributed to climate-related changes, complicating the task of assessing the effects of global warming on biodiversity. Temporally changing spatial and environmental biases in survey effort also hinder the process of forecasting the distribution of a species using the fragmentary data at hand. If the absence of a species is erroneously assigned to places where the species is present (i.e. false absences) within specific environmental conditions it is possible to obtain apparently reliable (but erroneous) distribution hypotheses, given that false absences will be spatially and climatically structured. Therefore, if the biases in biological data are related to the spatial and environmental variables used as predictors, the resulting predictive models will present a good fit to the data but lack biological accuracy.

Given that the spatial structure of the temporal growth in the information about the distribution of species may affect the reliability of distributional hypotheses, is necessary to incorporate some methods to assess and enhance data quality in the study of species distributions; this stands for both range shift assessments and predictive models. Unfortunately, the environmentally and spatially biased presence data compiled in most of the exhaustive databases currently available present important gaps and biases (see Hortal *et al*., 2007). Therefore, an assessment of the quality of the information is central to the use of data coming from historical surveys and collections (Hortal *et al*., 2001, 2004; Lobo & Martín-Piera, 2002; Hortal & Lobo, 2005; Romo *et al*., 2006). In most cases, it would be necessary to gather supplementary data in a standardized fashion, with the aim to represent all environmental and spatial variability of the territory (see Kadmon *et al*., 2004; Hortal & Lobo, 2005); comprehensive analyses of previous survey efforts can help to identify well-surveyed sites as well as to locate additional sites to be surveyed (Margules & Pressey, 2000; Hortal & Lobo, 2005). Apart from additional surveys, the reliability of predictive models could also be improved from some additional work on the detection and effects of false absences. On the one hand, absence sites with a high degree of certainty can be derived from the sites identified as well-surveyed (i.e. sites in which the amount of sampling effort is so high that it is unlikely that the species is present but remains unnoticed). On the other, specific studies of the effect of spatially structured false absences would help to identify the effects produced by biased unstandardized surveys on the estimation of geographical determinants of species distributions and the accuracy of predictive models.

The value of biodiversity studies and applications of the information historically gathered by naturalists is well understood (see Soberón & Peterson, 2004; Graham *et al*., 2004; Suarez & Tsutsui, 2004). However, the biases in this information could be compromising the conservation assessments and hypotheses developed using these data. Therefore, analyses of the quality, temporal bias and spatio-environmental coverage of the information should be a preliminary step in any protocol intended to take advantage of the information gathered in biodiversity databases. Additional surveys that fill in the environmental and spatial gaps in data will be frequently needed. Predictive models coming from data improved after these assessment and sampling processes will constitute much more reliable hypotheses of species distributions, though coming from incomplete data.

ACKNOWLEDGEMENTS

This paper was supported by a Fundación BBVA project and a MEC Project (CGL2004-04309). AJ-V was supported by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant. JH was supported by a Portuguese FCT (Fundação para a Ciência e Tecnologia) grant (BPD/20809/2004) and the UK Natural Environment Research Council. The comments of an anonymous referee highly improved a former version of this manuscript.

REFERENCES

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

Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004) Presence–absence versus presence-only based habitat suitability models for bird atlas data: the role of species ecology and prevalence. *Ecography*, **27**, 285–298.

Burgman, M.A. & Fox, J.C. (2003) Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation*, **6**, 19–28.

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

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

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

Elith, J., Graham, C.H., Anderson, R.P., Dudı́k, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Field, R., O'Brien, E. & Whittaker, R.J. (2005) Global models for predicting woody plant richness from climate: development and evaluation. *Ecology*, **86**, 2263–2277.

Gaston, K.J. (1991) How large is a species' geographic range? *Oikos*, **61**, 434–438.

Gaston, K.J. (2003) *The structure and dynamics of geographic ranges*. Oxford University Press, Oxford.

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

Hawkins, B.A., Diniz-Filho, J.A.F. & Soeller, S.A. (2005) Water links the historical and contemporary components of the Australian bird diversity gradient. *Journal of Biogeography*, **32**, 1035–1042.

Hengeveld, R. (1990) *Dynamic biogeography*. Cambridge University Press, Cambridge.

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

Hortal, J. & Lobo, J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2013–2047.

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

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

Iverson, L.R., Schwartz, M.W. & Prasad, A.M. (2004) How fast and far might tree species migrate in the eastern United States due to climate change? *Global Ecology and Biogeography*, **13**, 209–219.

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

Legendre, P. & Legendre, L. (1998) *Numerical ecology*. Elsevier, Amsterdam.

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

Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–253.

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

Martín-Piera, F. (2000) Familia scarabaeidae. *Coleoptera, Scarabaeoidea I* (ed. by F. Ramos, *et al.*), pp. 207–242. Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid, Spain.

Martín-Piera, F. & 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 Entomológica Ibérica e Macaronésica*, **1**, 23–35.

Parmesan, C., Gaines, S., Gonzalez, L., Kaufman, D.M., Kingsolver, J., Peterson, A.T. & Sagarin, R. (2005) Empirical perspectives on species borders: from traditional biogeography to global change. *Oikos*, **108**, 58–75.

Parmesan, C. & Yohe, G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Pulliam, H.R. (1988) Sources, sinks and population regulation. *American Naturalist*, **132**, 652–661.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

Rahbek, C. & Graves, G.R. (2001) Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences of the USA*, **98**, 4534–4539.

Rapoport, E.H. (1982) *Aerography: geographic strategies of species*. Pergamon Press, Oxford.

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

Ricklefs, R.E. (2004) A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, **7**, 1–15.

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

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

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

StatSoft, Inc. (2003) *STATISTICA (data analysis software system)*, Version 6. www.statsoft.com.

Suarez, A.V. & Tsutsui, N.D. (2004) The value of museum collections for research and society. *Bioscience*, **54**, 66–74.

Thomas, C.D. & Lennon, J.J. (1999) Birds extend their ranges northwards. *Nature*, **399**, 213.

Wagner, H.H. & Fortin, M.J. (2005) Spatial analysis of landscapes: concepts and statistics. *Ecology*, **86**, 1975–1987.

Watson, R.T. & the Core Writing Team (eds.) (2002) *Climate change 2001: synthesis report*. Cambridge University Press, Cambridge, UK.

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