Graellsia, 59(2-3): 389-398 (2003)

NOISY DATA AND DISTRIBUTION MAPS: THE EXAMPLE OF *PHYLAN SEMICOSTATUS* MULSANT AND REY, 1854 (COLEOPTERA, TENEBRIONIDAE) FROM SERRA DE TRAMUNTANA (MALLORCA, WESTERN MEDITERRANEAN)

M. Palmer[1]*, L. Gómez-Pujol[2], G. X. Pons[2], J. Mateu[3] and M. Linde[1]

ABSTRACT

Distribution maps are key tools for environmental management and biogeographic analyses. However, success in predicting spatial distribution is limited when using noisy presence/absence data sets. Both false absences and presences can be related with local departures from equilibrium (for example, temporary extinctions or unsuccessful colonisations). Moreover, false absences can arise from limited sampling effort. Here we explore an analytical strategy to get additional information on the presence/absence pattern of one target species from the presence/absence of all other species in the community. The logic is simple: the target species should display higher probability of presence at a site if a sample from this site is faunistically very close to the samples from other sites where the species occurs. Therefore, we first model presence/absence of the target species as a function of between-sample faunistic similarity. Second, the observed data for the target species are readjusted as a function of the expected probability of presence: current presences at sites with extreme low probability of presence are interpreted as unstable presences, and are recoded as absences. Seemingly, absences at sites with high probability of presence are interpreted as false absences, and are recoded as presences. In the experimental case presented herein, the recoding procedure is based on the presence/absence of 174 species, covering a broad taxonomic scope (snails, beetles, spiders and isopods). 1 km^2 distribution maps of presence/absence of the endemic beetle *Phylan semicostatus* were modelled from these recoded data. Mapping is done using GARP based on four environmental explanatory variables. These maps seem to be more stable and less prone to fail in predicting presence than those derived directly from the observed data.

Keywords: Distribution maps, Occurrence patterns, Genetic algorithms, GARP modelling, Invertebrates, Biogeography.

RESUMEN

Datos imprecisos y mapas de distribución: el ejemplo de *Phylan semicostatus* Mulsant y Rey, 1854 (Coleoptera, Tenebrionidae) en la Serra de Tramuntana (Mallorca, Mediterráneo occidental)

Los mapas de distribución son herramientas clave para la gestión medioambiental y los análisis biogeográficos. Pero el éxito en las predicciones de distribución espacial es limitado cuando se dispone de datos imprecisos de la presencia/ausencia. Tanto falsas ausencias como falsas presencias pueden estar relacionadas con desviaciones locales del equilibrio (por ejemplo, extinciones temporales o colonizaciones no exitosas). Además, las falsas ausencias pueden surgir de un esfuerzo de muestreo limitado. Aquí se explora una estrategia analítica para obtener información adicional sobre el patron de presencia/ausencia de una especie diana a partir de la presencia/ausencia de otras especies en la comunidad. La logica es simple: la especie diana debería tener una mayor probabilidad de presencia en un punto si

* Corresponding author; ieampv@uib.es

1 IMEDEA (CSIC-UIB), Instituto Mediterráneo de Estudios Avanzados, Miquel Marquès 21, 07190 Esporles (Mallorca), Spain.

2 Department of Earth Sciences. Univ. Illes Balears. Cta. Valldemossa km 7.5, 07071 Palma de Mallorca, Spain.

3 CITTIB, Centre d'Investigació i Tecnologies Turístiques de les Illes Balears. Passatge Guillem de Torrella, 07002 Palma. Illes Balears.

una muestra de este punto es faunísticamente muy similar a las muestras de otros puntos donde la especie ha sido detectada. Por tanto, primeros se modela la presencia/ausencia de la especie diana en función de la similaridad faunística entre puntos. En segundo lugar, los datos observados para la especie diana son reajustados en función de la probabilidad esperada de presencia: las presencias observadas en puntos con probabilidad de presencia muy baja son interpretadas como presencias inestables, y recodificadas como ausencias. De manera similar, las ausencias en puntos con probabilidad de presencia muy elevada son interpretadas como falsas ausencias, y recodificadas como presencias. En el caso experimental estudiado, el procedimiento de recodificación esta basado en los datos de presencia/ausencia de 174 especies, abarcando un abanico taxonómico muy amplio (caracoles terrestres, coleópteros, arañas e isópodos). El mapa de distribución de celdas de 1 km^2 del coleóptero endémico *Phylan semicostatus* es modelado a partir de estos datos. El mapa de distribución es elaborado a partir de cuatro variables medioambientales, usando una estrategia analítica basada en algoritmos genéticos (GARP). Los mapas obtenidos con los datos recodificados parecen ser mas estables y menos susceptibles de fallar en sus predicciones que los mapas elaborados directamente con los datos originales.

Palabras clave: Mapas de distribución, patrones de presencia, algoritmo genético, modelos GARP, invertebrados.

Introduction

There is a general agreement on the low budget available for faunistic and floristic surveys; unfortunately, the near future will not bring notable increases in most of the European countries. At the same time, the current level of knowledge on the spatial distribution of most species remains still inadequate for to be considered useful tools for environmental management. The case of the Balearic Islands is an example: no reliable data exists even for the mere number of invertebrate species. This stage of knowledge is evident, for example, in a contribution by Dr. Martín-Piera, who increased by eight species the checklist of the Scarabaeidae (Coleoptera) from the Balearic Islands after only two short field surveys (Martín-Piera & Lobo, 1992).

This justifies an increasing effort toward predicting species distributions (or community descriptors as diversity) from a small sample of prospected sites (Austin, 2002). The inferential logic behind this is quite simple: a species is predicted to occur in a non-surveyed site if it displays similar environmental features to other sites where the species currently occurs. The biological basis for these inferences is founded in a number of assumptions, two of which are especially relevant here. The first one is that species are in equilibrium (or at least, some kind of quasi-equilibrium) with the environment (Austin, 2002). It is thus assumed that abundance (or probability of presence) of each species is potentially deducible from environmental data, in spite of the debate on the shape of species-response curves or on the rules controlling species turnover and species packing. However, success in prediction is limited by the effects of history or disturbance (i.e., non-equilibrium). For example, in a presence/absence survey in the context of metapopulation dynamics, it is easily imaginable that one species can be detected in a site that does not display suitable (or even unsuitable) environmental conditions. This local population is prone to becoming extinct, and its presence should be considered as temporally unstable. Another source of uncertainty is that a species currently present in a site has not been effectively observed (Zaniewski *et al*., 2002; Dennis & Shreeve, 2003). Such false positives and negatives (Fielding & Bell, 2003) can be considered as noise over the signal (species-environment relationships), and certainly they will limit the performance of prediction.

The second assumption involves the continuum concept that, widely interpreted, points to species composition changing along a continuous environmental gradient (Fig. 1, inspired, for example, in Putman & Wratten, 1984 and ter Braak & Smilauer, 1998). Figure 1 involves the additional assumption that the species response curves to environment are unimodal and symmetric. However, independently from the current shape of the species-response curve, it is clear that some information on the presence-absence of an individual species should be potentially deducible from the presence or the absence of all other species in the community. Here we explore an analytical strategy to obtain additional information on the presence/absence of one target species from the presence/absence of all other species in the community. We first model presence/absence of the target species as a function of between-sample faunistic similarity. Second, the observed data-set for the target species is readjusted as a function of the expected probability of presence: current presences at sites with extreme low probability of presence are interpreted as unstable presences, and are recoded as absences. Similarly, absences at sites with high probability of presence are interpreted as false absences, and are recoded as presences.

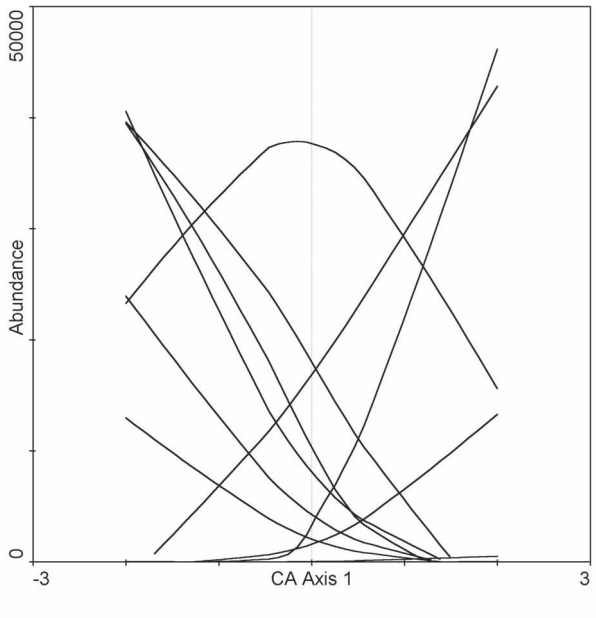


Fig. 1.— Theoretical species turnover along an ecological gradient. Species response curves are assumed to be unimodal, symmetric and defined by three parameters (optimum, tolerance and abundance at the optimum).

Fig. 1.— Reemplazamiento teórico de especies a lo largo de un gradiente ecológico. Se asume que la curva respuesta de las especies es unimodal, simétrica y definida por tres parámetros (óptimo, tolerancia y abundancia en el óptimo).

We first analyse a simulated system looking for the accurate description of the procedure using a data set corresponding to a community with a simplified and known structure. Subsequently, a macro-invertebrate community is studied. The adequacy of arthropods and other macroinvertebrate relies on their large number and their high species turnover along environmental gradients (Lawton *et al*., 1998), while most vertebrates are insensitive to fine-scale habitat heterogeneity (Mattoni *et al*., 2000). The case study is based on the presence/absence of 174 species (covering snails, beetles, spiders and isopods) on 48 sites of 1 km^2 each along a mountain range. The overall data set are used to adjust the observed presence-absence data-set of a target species (the endemic beetle *Phylan semicostatus*). The spatial distribution of this species along the whole mountain range is then modelled from the adjusted occurrence records and four environmental variables. The modelling strategy adopted is based in an heuristic search using a genetic algorithm (the Genetic Algorithm for Rule-Set Prediction, GARP; Payne & Stockwell, 2001).

Methods

Simulated gradient

In order to analyse the potential capabilities of the proposed analytical strategy, we designed a system composed of 50 sites and 254 species. For simplicity, species responses to the environment are all considered unimodal and symmetric (i.e., Gaussian, ter Braak & Smilauer, 1998; Oksanen *et al*., 2001), and the environmental gradient highly correlated with a single variable. Gaussian response curves are defined by three parameters. Namely optimum, tolerance, and abundance at the optimum. Species optimums are randomly located along the gradient. Species tolerance is randomly allowed to move between zero and the half-width of the gradient. Finally, abundance at the optimum is randomly shifted between 500 and 50000 individuals. Stochastic variability is simulated by simple addition (or subtraction) of a random proportion of the maximum species abundance. Figure 1 shows the species response of a random subsample (10) of the 254 species considered.

The averaged number of individuals per site is $2.1*10^6$. A small number of individuals (1/50000) are sampled with replacement from each site, from which a presence/absence data set is built. The averaged number of species and individuals per sample are respectively 16.1 and 42.8.

Between-samples differences are analysed using correspondence analyses (CA, ter Braak & Smilauer, 1998) of the abundance and presence-absence data. The scores of the first CA axis are estimates of between-sample faunistic differences and were plotted against the (single) environmental variable for comparing the pattern extracted by abundance data against that extracted by presence-absence data.

One species was chosen at random, and its presence/absence pattern is modelled with the scores on the first CA axis. Instead of the usual Gaussian model (Oksanen *et al*., 2001) we used the full HOF model (Huisman *et al.*, 1993, Oksanen & Minchin, 2002). The main advantage of this model is that skewed (asymmetric) unimodal species response can be managed also. Two *a priori* selected thresholds allowed us to detect false absences and unstable presences from the data. A new data set (i.e., the adjusted data) is

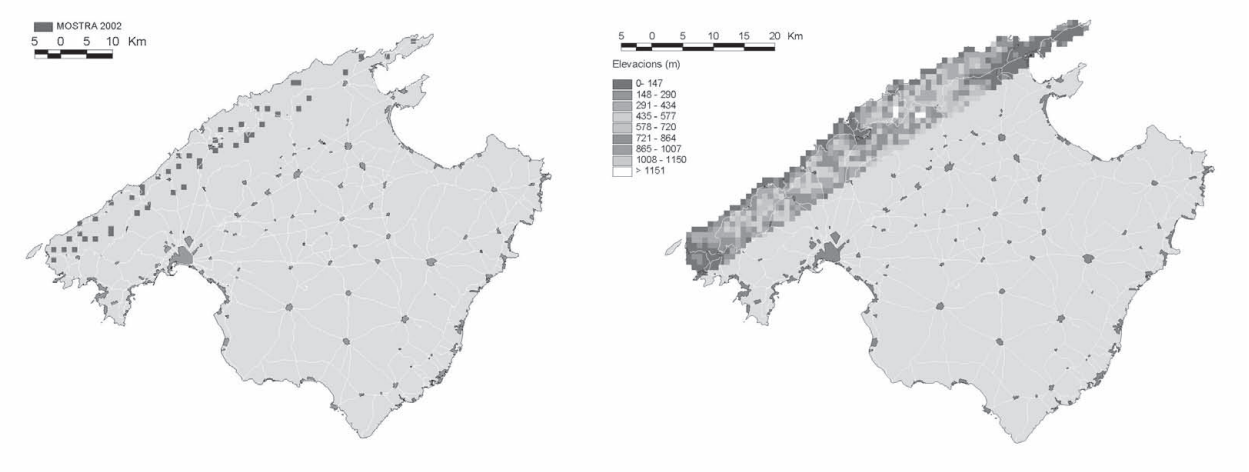


Fig. 2.— Maps showing (left) the sampling sites (n= 48) and (right) the area covered by the predictive models (n= 584), the latter displaying also the Altitude of each 1 km^2 cell. The other three predictive variables used for modelling species occurrence were aspect, distance to the nearest house and NVDI (non-linear combination of reflectance at specific lambdas extracted from a satellite image).

Fig. 2.— Mapas de los puntos muestreados (n= 48; a la izquierda) y del área cubierta por los modelos predictivos (n=584; a la derecha). En este último mapa señala también la altura de cada celda de 1 km^2. Las otras tres variables predictivas usadas para modelar la presencia de la especie diana fueron la orientación, la distancia a la casa mas cercana y el valor NVDI (se trata de una combinación no lineal de la reflectancia a longitudes de onda especificas, extraída de una imagen de satélite).

then built after recoding the putative false positives and negatives. Finally, the modelled species presence using the recoded data-set is compared with the original data set. This comparison is made by plotting the proportion of correctly predicted occurrences against an increasing threshold level (Fielding & Bell, 2003).

The case of *Phylan semicostatus*

The study area (Serra de Tramuntana) is a mountain range located to the North-West of Mallorca, rising from the seashore to 1445 m a.s.l. Forty-eight sampling sites have been selected along its main axis (100 km long; Fig. 2); Their positions GPS-determined. Five pitfall traps (10 cm wide; 2 m left between traps; detergent and salt used as preservative) were set at each site. The traps remained in the field for two moths (June to July). The sampling schedule has been complemented with one hour of direct search (e.g., under stones) at each site on a 1000 m^2 plot around the pitfall traps. We analyse presence-absence data because between-species differences in detectability and catchability were noticeable. Data from the five pitfalls and from direct search are pooled as a single sample per site, and it is assumed to represent the fauna of the corresponding 1 km^2 cell.

We focus on four groups of invertebrates, namely Coleoptera, Arachnida, Isopoda and Gastropoda. As regard Gastropoda, a species is considered to be present also when empty shells only are found. Some of the species not readily determinable are identified as morphospecies (Oliver & Beattie, 1996).

Presence/absence data are analysed in a similar way to the analysis of the simulated data set. The major differences are: (1) the ecological gradient is considered to be bivariate instead of univarite (presence/absence of *Phylan semicostatus* was modelled using the scores on the first two CA axes), and (2) the spatial distribution of *Phylan semicostatus* is deducted using a more formal modelling strategy (GARP). The occurrence pattern along 584 cells of 1 km^2 (Fig. 2) is predicted from the adjusted occurrence pattern and four environmental variables: altitude, aspect, distance to the nearest house and NVDI (non-linear combination of reflectance at specific lambdas extracted from a satellite image). The methods for estimating the environmental values corresponding to each 1 km^2 cell are fully detailed in (Palmer *et al.*, 2002).

The Genetic Algorithm for Rule-Set Prediction (GARP) is an expert system machine-learning approach to predictive modelling (Stockwell & Peters,

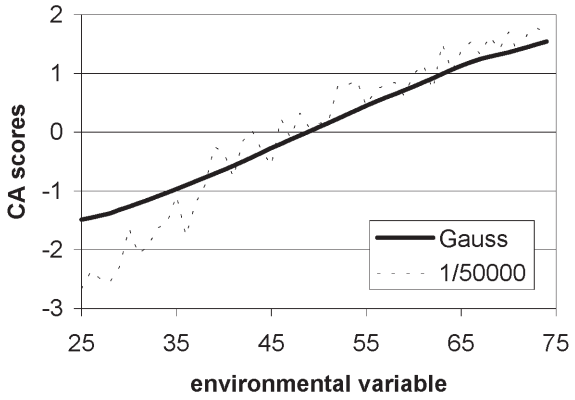


Fig. 3.— Relationship between CA sample scores determined from (1) abundance data of an hypothetical community composed of species with Gaussian response curves and (2) presence/absence data of the same community but adding an stochastic random noise, and sampling only one out of every 50000 individuals. Departures are more important at the two extremes, but the two sets of scores shown similar relationship with the environmental variable.

Fig. 3.— Relación entre los valores, de cada una de las muestras consideradas, sobre el primer eje de un análisis de correspondencias (CA), correspondiantes a (1) los datos de abundancia de una comunidad hipotética compuesta por especies con una curva de respuesta de tipo Gaussiano, y (2) los datos de presencia/ausencia de la misma comunidad, pero añadiendo al azar un nivel de incertidumbre estocástica, y muestreando solo uno de cada 50000 individuos presentes en la comunidad. Las desviaciones entre las curvas respectivas son mas importantes en los extremos, pero los dos conjuntos de datos muestran una relación similar con la variable medioambiental.

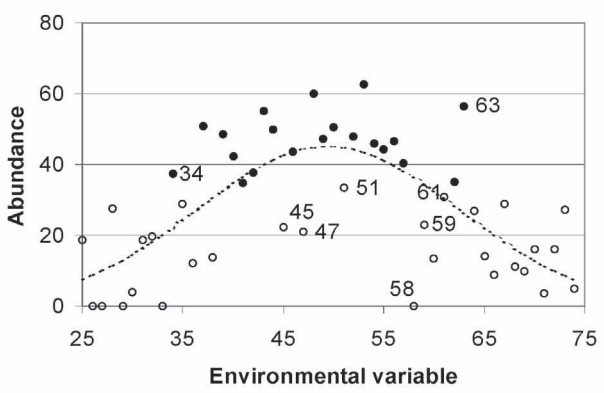


Fig. 4.— Abundance changes along the environmental gradient. The Gaussian response curve (broken line) of a specific species is compared with the "observed" abundance values resulting from adding a stochastic random noise. The solid points indicate presence of the species when sampling 1/50000 individuals and the open points indicate absence. The numbers label the recoded points.

Fig. 4.— Cambios en la abundancia a lo largo de un gradiente medioambiental. La curva de respuesta Gaussiana (línea discontinua) de una especie concreta es comparada con los valores de abundancia "observados", los cuales resultan de añadir un al azar un determinado nivel de incertidumbre estocástica a los valores esperados. Los puntos negros indican la presencia de la especie al muestrear uno de cada 50000 individuos en la comunidad. Los puntos blancos indican su ausencia. Los números indican las muestras recodificadas.

1999). Genetic algorithms are inspired in evolutionary models. They present an heuristic solution after scanning broadly across the solution space (i.e., all possible solutions), and refining solutions that show high values for the optimisation criterion. GARP has not been extensively used yet, but has proved to be a useful approach (see Anderson *et al.*, 2003 for a recent review). The ordinary strategy involves running the program a number of times to obtain several tentative models. Model selection have been carried out using the criteria suggested by Anderson *et al.*, 2003). Briefly, the models selected display both low intrinsic omission error and intrinsic commission error near the average (see Fielding & Bell, 2003 and Anderson *et al.*, 2003 for more details on error theory and model evaluation for occurrence data).

GARP (Stockwell & Peters, 1999) is downloadable at http://beta.lifemapper.org/desktopgarp/. Correspondence analyses have been carried out using CANOCO 4 (ter Braak & Smilauer, 1998), and HOF models were fitted using R (http://cran.r-project.org; nlm functions were modified from those supplied by http://cc.oulu.fi/~jarioksa/ and Oksanen & Minchin, 2002).

Results

Simulated gradient

As expected, the scores on the first axis from a Correspondence Analysis (CA) of the original data are clearly correlated with the environmental variable (Fig. 3). Moreover, and more interestingly, the corresponding scores derived from a presence/absence data set that simulate a low sampling effort (1 of each 50000 individuals) show a very close trend. The importance of this result relies on the fact that CA scores on the first few axes obtained using real-world presence/absence data probably reflect the main betwe-

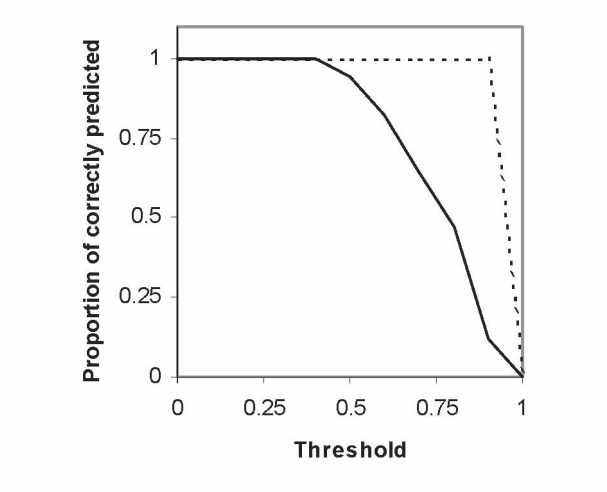


Fig. 5.— Proportion of correctly predicted occurrences of the original (non-noisy) data from "observed" data (solid line) and from the recoded data (dashed line).

Fig. 5.— Proporción de presencias de los datos originales (sin incertidumbre) correctamente predichas a partir de los datos "observados" (línea continua) y de los datos recodificados (línea discontinua).

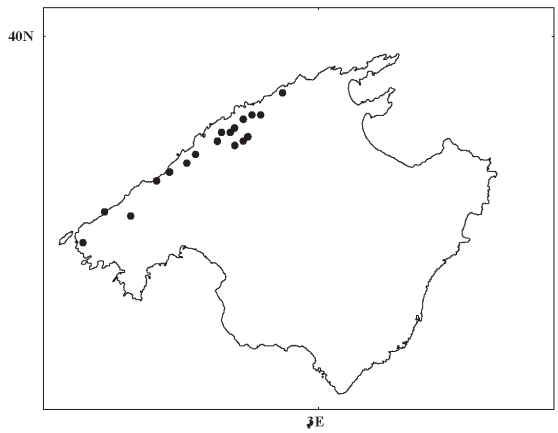


Fig. 6.— Observed occurrences of *Phylan semicostatus*.

Fig. 6.— Presencias observadas para *Phylan semicostatus*.

en-sample similarity trends, even in cases of noisy data sets.

The results of modelling presence/absence data of a target species using the scores on the first CA axis are shown in Figure 4. The observed presences/absences are derived from the noise-added data. The crux of the strategy is the following recoding procedure: The species was not detected at the sites labelled as open points because of the low relative abundance of the species at these sites. However, some of these sites (# 45, 47, 51, 58, 59 and 61) showed CA scores that are very close to all the other sites where the specie was present because they share a similar faunistic composition (except for the case of the target species). After modelling probability of presence as a function of the scores on the first two CA axes, the probability values of presence for the sites listed above is always higher than a prefixed threshold (here, 0.5); consequently the absences of the target species on the sites are considered to be false absences and are recoded as presences.

Similarly, the species is detected at the sites labelled as closed points, but two of them (#34 and 63) display a low probability of presence (below 0.3). Consequently, these two occurrences are considered to be unstable and are recoded as absences. The latter threshold is selected to include 90% of the observed occurrences (i.e., sites 34 and 63 are those with more deviant CCA scores).

The consequences of the recoding procedure on the prediction of presence/absence from the (single) environmental variable are shown in Figure 5. The overall performance (total missclassifications) of the two data sets (observed *versus* recoded) in predicting the occurrence profile of the original (noise-free, Gaussian) data is similar. However, occurrences are better predicted by the recoded data, which is exactly the final objective of the recoding procedure.

The case of *Phylan semicostatus*

The target species occurs along the length of the Serra de Tramuntana mountain range, but occurrences seem to be more frequent at the central area (Fig. 6). This pattern coincides with field experience pointing to this species being characteristic of places at high-moderate altitude (in relation to the observed maximum altitude, 1445 m). This endemic beetle is easily found at sites with low vegetation cover and mountain-type shrubs. However, it is noteworthy that the species is also frequent in a number of small islets around Mallorca, demonstrating that altitude itself has no biological relevance (i.e., indirect effects; Palmer, 1994; Palmer, 1997). The current research does not cover littoral habitats.

Presence/absence data of the target species is modelled using the scores on the two first axes of

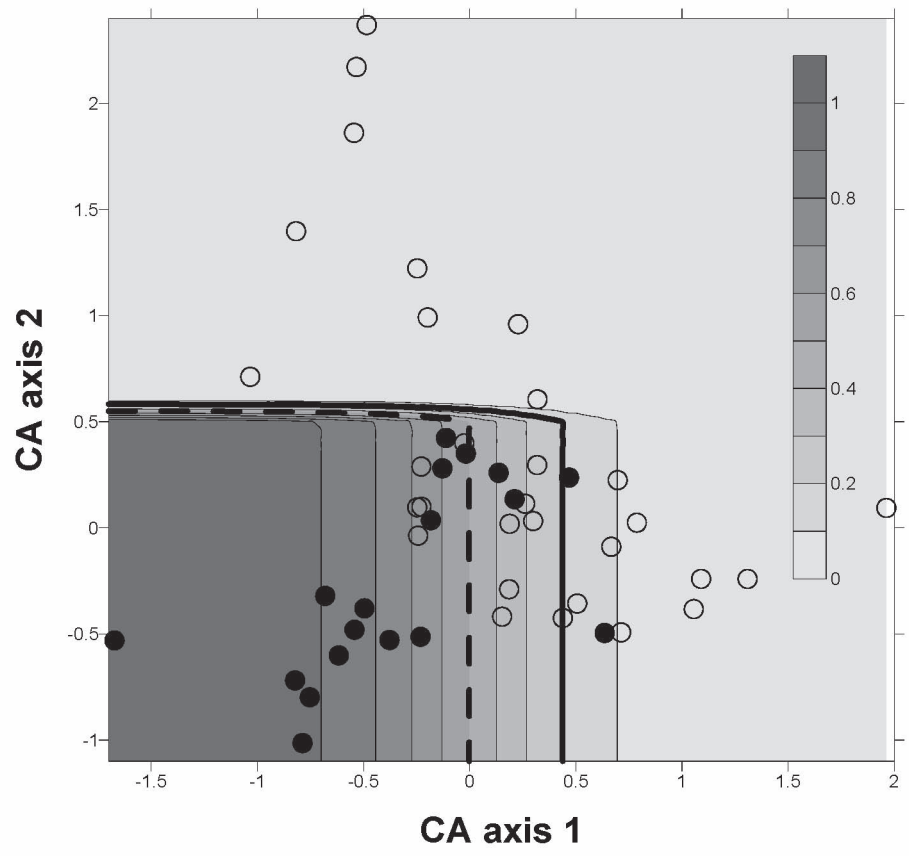
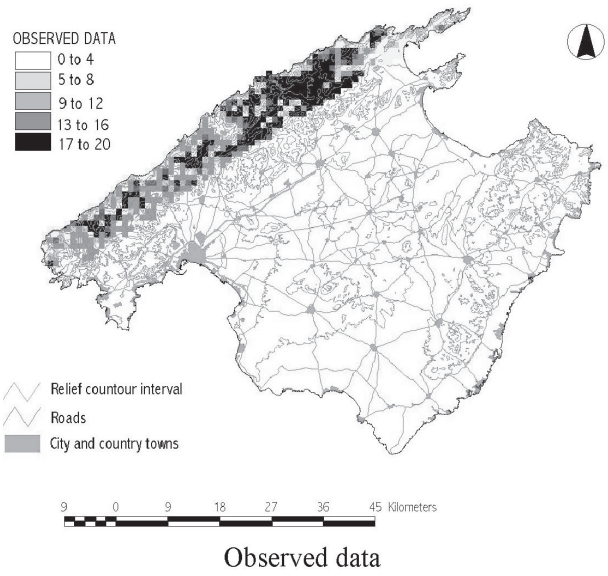


Fig. 7.— Standard plot of the sample scores on the first two axes of a Correspondence Analysis (CA). Solid and open points denote respectively samples with and without *Phylan semicostatus*. Probability of presence of *Phylan semicostatus* (HOF model) is indicated by a grey scale (darkness indicate higher probability values). The wider isoline indicates the threshold for considering presence as unstable. The dashed isoline indicates the threshold for considering absence as false absence.

Fig. 7.— Representación estándar de la posición de las muestras con respecto a los dos primeros ejes de un análisis canónico de correspondencias. Los puntos negros y blancos indican respectivamente las muestras con y sin *Phylan semicostatus*. La probabilidad de presencia de *Phylan semicostatus* es indicada mediante una escala de grises (tanto mas oscura al aumentar la probabilidad de presencia). La línea ancha continua indica el umbral para considerar una presencia como inestable. La línea ancha discontinua indica el umbral para considerar una ausencia como falsa ausencia.

a CA (174 species and 48 sites) as predictors. The fitted model is a full HOF model (four parameters for each of the two predictors involved; more details are provided by Oksanen & Minchin, 2002). The probability of presence predicted by the model is shown in Figure 7. The threshold for detecting a false absence is 0.5 (i.e., all sites with expected P-values higher than 0.5 are assumed to be occupied by the species), while unstable presence threshold is defined as the probability level that includes 90% of the occurrences (i.e., the occurrences that display the more extreme values on CA1 and CA2 axes are considered to be unstable; in the case of *Phylan semicostatus* this value was 0.2).

Twenty optimal (Anderson *et al.*, 2003) GARP-predicted spatial distribution maps are obtained with both recoded and observed data. The median values for each of the 584 pixels are compared in Fig. 8. The maps corresponding to the observed and recoded data consistently show similar intrinsic omission rates (close to 0 %). However, the between-models variability is higher among the maps built with the observed data (Fig. 9). The high rate of pixels for which the model does not offer clear-cut prediction is also noteworthy (Fig. 8).



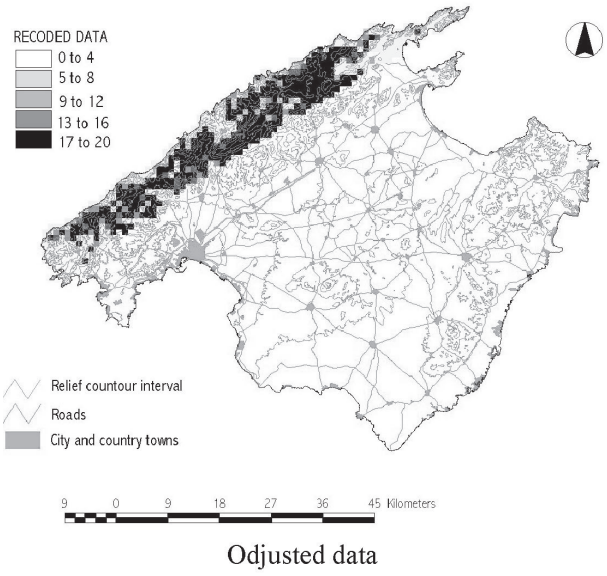
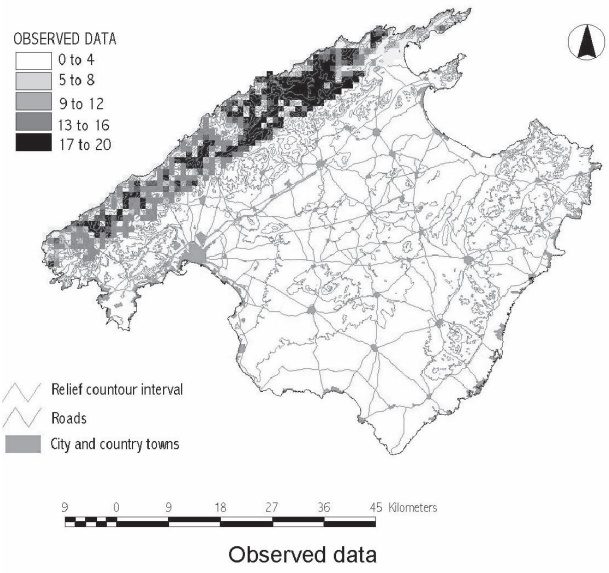


Fig. 8.— Maps of the modelled spatial distribution of *Phylan semicostatus* in the study area. These maps show for each pixel (1 km^2) the median value of twenty optimal maps. The map corresponding to the recoded data seems to be more stable, and show fewer points with undefined prediction.

Fig. 8.— Mapas de la distribución espacial modelada para *Phylan semicostatus* en el área de estudio. Estos mapas muestran para cada pixel (1 km^2) el valor de la mediana de 20 mapas originales (i.e., modelados independientes). El mapa correspondiente a los valores recodificados parece ser mas estable, y muestra menos puntos sin una predicción definida.



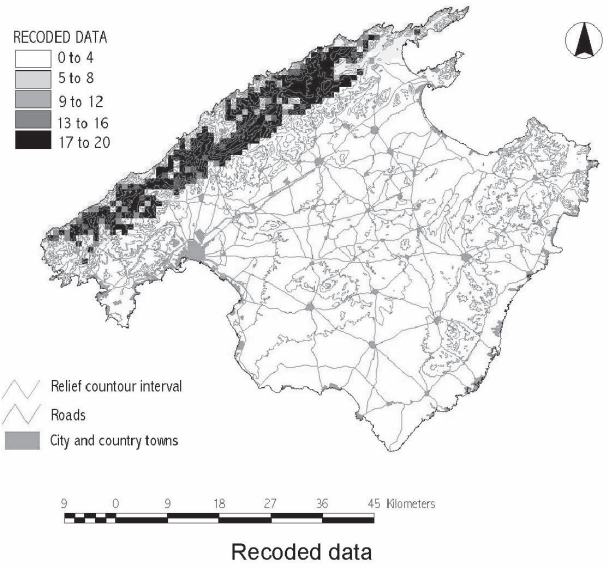


Fig. 9.— Maps of the modelled spatial distribution of *Phylan semicostatus* in the study area. These maps show for each pixel (1 km^2) the number of models predicting presence. Between-model differences in the range of potential distribution of the target species are higher in the case of the observed data.

Fig. 9.— Mapas de la distribución espacial modelada para *Phylan semicostatus* en el área de estudio. Estos mapas muestran para cada pixel (1 km^2) el numero presencias en 20 modelados independientes. Para la especie diana, las diferencias entre modelados para las predicciones son mas grandes en el caso de los datos observados que en el caso de los datos recodificados.

Discussion

The results produced by the analysis using the simulated community data-set show that faunistic differences between sites are well described with presence/absence data even in cases of small sampling effort and noisy data. Canonical correspondence analyses of the original abundance data and presence/absence data would render similar patterns if the number of species considered is high enough. Simulated data illustrate also the type of noise that potentially affects the species-specific response curves to ecological gradients, and reveals ways for filtering this noise. Recoded data are less prone to fail in predicting presence (Fig. 5). The counter-part of recoding is that these data show increased risk of failing to predict absence. However, the latter is an advantage when sampling is not very exhaustive, because the high prevalence of false absences.

The results derived from the experimental data suggest that the solutions obtained using the recoded data are more stable in the sense that they are more similar to each other. However, two potential limitations should be noticed. First, the method is not applicable to species from environments very different to a single and general environmental gradient. The reason is that the first axis from correspondence analyses (or any other multivariate approach) account for the faunistic differences between these marginal sites and the rest of sites, instead of being correlated with the target environmental gradient. An extreme example of this would be to include some samples from ponds or other aquatic environments: these samples are easily detectable by plotting the CA scores, and it has no sense to include them in any further analysis that focuses on non-aquatic fauna. Another limitation of the method is it implicitly assumes similar sampling effort. This is the case of the data analysed here, but most of the data sources used for mapping do not meet this assumption (e.g., museum data) and provides biased images of species distribution (Dennis, 1999; Dennis, 2000).

In summary, the proposed analytical strategy (i.e., using the information from other species in order to improve the observed pattern of occurrence of a target species) seems to work reasonably well. This improvement, far from being considered as data manipulation, should be better defined as noise filtering. Obviously, further research should be done regarding a number of points. For example, enlarging the number of target species (and covering a wide range of prevalence), and improving the selection procedure of the thresholds to detect false positives and negatives, which are reasonably but subjectively defined. However, this or similar analytical procedures opens the possibility of making full use of all the information obtained with extensive invertebrate trapping surveys. The unbiased image of species distribution can be the input for a variety of additional analyses. One interesting approach is to integrate presence-absence data with knowledge on interspecific associations and multivariate ordination, using distance-dependent variance-covariance matrices (Wagner, 2003). The same idea is extrapolable from spatially to temporally autocorrelated data, opening new ways to estimate the relative importance of historical *versus* environmental forces in determining species composition.

ACKNOWLEDGEMENTS

The authors thanks the Conselleria de Medi Ambient del Govern de les Illes Balears and "Sa Nostra" Caixa de Balears for their financial support, to the land owners of the sampling sites for allowing us to install pitfall traps on their proprieties, to Jordi Monterde for his invaluable help during the field work, and to an anonymous referee for his comments.

References

ANDERSON, R. P., LEW, D. & PETERSON, A. T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, 166: 287-293.

AUSTIN, M. P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling,* 157: 101-118.

DENNIS, R. L. H., 1999. Bias in butterfly distribution maps: The effects of sampling effort. *Journal of Insect Conservation,* 3: 33-42.

DENNIS, R. L. H., 2000. Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. *Journal of Insect Conservation,* 4: 73-77.

DENNIS, R. L. H. & SHREEVE, T. G., 2003. Gains and losses of French butterflies: tests of predictions, underrecoding and regional extinction from data in a new atlas. *Biological Conservation,* 110: 131-139.

FIELDING, A. H. & BELL, A. F., 2003. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation,* 24: 38-49.

HUISMAN, J., OLFF, H. & FRESCO, L. F. M., 1993. A hierarchical set of models for species response analysis. *Journal of Vegetation Science,* 4: 37-46.

LAWTON, J. H., BIGNELL, D. E., BOLTON, B., BLOEMERS, G. F., EGGLETON, P., HAMMOND, P. M., HODDA, M., HOLT, R. D., LARSEN, T. B., MAWDSLEY, N. A., STORK, N. E., SRIVASTAVA, D. S. & WATT, A. D., 1998. Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature,* 391: 72-76.

MARTÍN-PIERA, F. & LOBO, J. M., 1992. Los Scarabaeoidea Laparosticti del Archipiélago Balear (Coleoptera). *Nouvelle Revue d'Entomologie,* 9: 15-28.

MATTONI, R., LONGCORE, T. & NOVOTNY, V., 2000. Arthropod monitoring for fine-scale habitat analysis: a case study of the El Segundo sand dunes. *Environmental Management,* 25: 445-452.

OKSANEN, J., LÄÄRÄ, E., TOLONEN, K. & WARNER, B. G., 2001. Confidence intervals for the optimum in the Gaussian response function. *Ecology,* 82: 1191-1197.

OKSANEN, J. & MINCHIN, P. R., 2002. Continuum theory revisited: what shape are species response along ecological gradients? *Ecological Modelling,* 157: 119-129.

OLIVER, I. & BEATTIE, A. J., 1996. Invertebrate morphospecies as surrogates for species: a case study. *Conservation Biology,* 10: 99-109.

PALMER, M., 1994. *Aspectes filogenètics i biogeogràfics dels Tenebrionidae (Coleoptera) de les Illes Balears*. Ph.D.Universitat de les Illes Balears. Palma de Mallorca.

PALMER, M., 1997. Diversity in small Western Mediterranean islets: effects of rats on beetles communities. *Acta Oecologica,* 17: 297-305.

PALMER, M., GÓMEZ-PUJOL, L., PONS, G. X., MATEU, J., MCMINN, M. & RODRÍGUEZ, A., 2002. *Cartografia de la distribució d'espècies endèmiques i bioindicadores a la Serra de Tramuntana: una aproximació des de la teledetecció i la geoestadística*. Conselleria de Medi Ambient, Govern Balear. Palme de Mallorca. 58 pp. (and four appendix).

PAYNE, K. & STOCKWELL, D. R. B., 2001. *GARP modelling system user's guide and technical refeence*. BIODI Advanced computational approaches to environment and biodiversity information. 52 pp.

PUTMAN, R. J. & WRATTEN, S. D., 1984. *Principles of Ecology*. University of California Press. Berkeley. 388 pp.

STOCKWELL, D. & PETERS, D., 1999. The GARP modelling system: problems and solutions to automated spatial predictions. *International Journal of Geographical Information Science,* 13: 143-158.

TER BRAAK, C. J. F. & SMILAUER, P., 1998. *CANOCO reference manual and user's guide to Canoco for Windows: Software for Canonical Community ordination (version 4)*. Microcomputer Power. Ithaca, NY, USA. 352 pp.

WAGNER, H. H., 2003. Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. *Ecology,* 84: 1045-1057.

ZANIEWSKI, A. E., LENHMANN, 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.