HORTAL, J. & LOBO, J. M., 2001. A preliminary methodological approach to model the spatial distribution of biodiversity attributes. In: *Spatio-temporal modelling of environmental processes: proceedings of the 1st Spanish Workshop of Spatio-temporal Modelling of Environmental Processes*: pp. 211-239 (J. Mateu & F. Montes, eds.). Publicacions de la Universitat Jaume I, Col· lecció "Treballs d'Informàtica I Tecnologia" 10, Castelló de la Plana.

A preliminary methodological approach to model the spatial distribution of biodiversity attributes*

Joaquín Hortal[†], Jorge M. Lobo[‡]
Museo Nacional de Ciencias Naturales (C.S.I.C.). Departmento de Biodiversidad y
Biología Evolutiva. C/José Gutiérrez Abascal, 2. 28006 Madrid, Spain.
e-mail: [†]mcnjh521@mncn.csic.es, [‡]mncnj117@mncn.csic.es

Abstract

In a time when conservation strategies based on biodiversity data are needed, our knowledge about its geographical distribution is scarce and biased. Most regions and living organisms' groups are partly or completely unknown, so estimated maps of biodiversity attributes are needed. To reach this target quickly, it seems necessary to develop new survey planning methods and rapid, easy-to-use statistical methods able to forecast several biodiversity attributes on a reliable manner. Biogeographers and conservation ecologists have proposed several methods to face the latter goal, and arising Geostatistics also provide some interpolation methods, but there exists a lack of knowledge about which distributional data are good enough to produce good results, and also about which forecasting method is the best to use at a given spatial scale on each kind of territory and group of organisms. It is in this task where the joint work of mathematicians and biodiversity scientists is needed.

Many of the problems occurring when modelling the spatial distribution of biodiversity attributes are discussed, and a step-by-step heuristic modelling methodology that tries to afford many of them, based on the use of GLM with environmental and spatial variables, is also proposed and discussed. Its use has been exemplified by the developing of a model to forecast species richness scores for a group of insects (Col., Scarabaeinae) on the Iberian Peninsula.

Keywords: biodiversity attributes forecast, Generalized Linear Models, spatial distribution modelling, heuristic search

* This work is included in the PB97-1149 project of the Spanish D.G.C.Y.T. Both authors were funded by Comunidad de Madrid grants.

1 Introduction

After more than 250 years of faunistic and taxonomic data accumulation by scientists devoted to the study of life's diversity and distribution, there is no locality in the world with a complete inventory of the living organisms inhabiting it. We are unaware of its approximate spatial distribution, ever its total number [1]. Unfortunately, this knowledge is absolutely necessary to design feasible biodiversity conservation policies [2]. Due to this picture, it has been suggested that modelling the spatial distribution of biodiversity attributes is the more rational, rapid choice for biodiversity assessment [3]. However, modelling the geographic distribution of living organisms seems to be a rather impracticable task.

The geographical distribution of earth's biota is the result of a huge and dense network of factors, operating at diverse temporal and spatial scales. In spite of many decades of growth of ecology as a science, this complexity has avoided the development of a theoretical framework able to explain the causes that promote the observed distribution patterns [4-6]. Since the 1990 decade, many environmental scientists from different science fields have focused its work to explore patterns at a wider, macroecological, working scale, in order to obtain a better understanding of biodiversity patterns and its underlying processes. The joint appearance of new tools and techniques, such as personal computers (PCs), Geographic Information Systems (GIS) and Geostatistics, has allowed this new science field to use powerful spatial analysis, great computation power and a lot of georreferenced-high quality environmental information [7].

Modelling techniques can play an important role in the description of the geographical patterns of some biodiversity-related variables, such as species richness, rarity or phylogenetic diversity. Moreover, the models developed can even suggest possible mechanisms and experiments to explore the relative influence of different factors on these variables [4, 8].

Two main approaches have been used to model the distribution of biodiversity-related attributes. A first one looks for an environmental variables-based function able to account for as much variation of the used attribute as possible, by means of any kind of heuristic, iterative, search [9-33]. Here, the key criterion is to find a forecasting function from a given set of variables, whether the selected variables have a functional influence on biodiversity or not. The second approach, being strict, is a variation of the former, but, in this case, a causal relationship between predictor and response variables is first identified to, subsequently, assign a probability of occurrence for each species depending on the ranges of the environmental variables in which it occur. Both the genetic algorithms-based approximations [34-36] and those developed for the use in gap analysis techniques [37, 38] belong to this latter approach. The exhaustive search of the best function from all the possible combinations of predictor variables [39], although being a promising technique, has not been ever used for this purpose, maybe due to the general lack of communication between biogeographers and mathematicians.

Apart from this, little is known about the limitations and drawbacks implicit in the biological dependent variables modelled, which, joint with the attractive results produced by the commented new methodologies, has unfortunately leaded to an abuse of the use of them with poor quality biological data (response variable), and even ignoring the main problems inherent to real environmental datasets (predictor variables). In this paper we analyse the main shortcomings that a heuristic search must avoid when modelling biodiversity attributes. We propose a step-by-step procedure designed to overcome some of these problems, using a stepwise Generalized Linear Model-based heuristic technique. As a practical example, we create a model of species richness (number of species present in a given territory) of a family of dung-feeding beetles (Coleoptera, Scarabaeinae [40]) in the Iberian Peninsula [31].

2 The use of real biological data

In order to build a predictive model of biodiversity spatial distribution with real distributional data of a given region, we have to deal with three main problems:

i) the existence of different data sources,

ii) the heterogeneity in data quality and origin,

iii) the lack of knowledge about how complete are the inventories found for a given territory.

2.1 Data compilation and storage

First, the information available is almost always scarce, heterogeneous, and dispersed over a lot of sources such as scientific publications, Natural History Museums and private collections. To use this information, it is necessary to develop a database able to compile it exhaustively, including data from all available sources for the group studied in the territory analysed.

To achieve this, we have to establish data storage protocols, which allow us to use this information both in qualitative and quantitative ways, to access to and share it easily, and also to compare it among different groups and regions. The database developed must include, at least, the following fields: date of capture or observation, place [and also spatial coordinates in a common-use reference system, such as Geographical (Lat/Long) or Universal Transverse Mercator (UTM)], ecological data of the capture relevant for the group studied (habitat characteristics, feeding, altitude, host species, etc.), number, and sex if available, of captured/observed specimens, capture or observation method, collector or observer's identity, taxonomic determination's responsible (in order to assess the accuracy of the data), place of storage (for specimens from Natural History Collections) or bibliographic reference, and other useful data, such as genome sequences, morphotype, etc.

In the example case of Iberian Dung Beetles, the BANDASCA database (see structure in reference 41), that at present contains 15,740 records and 101,996 individuals of the 53 Iberian dung beetle species [40], complies with this requirements. Data from this database will be used to carry out the practical example of the methodology.

2.2 Choosing the information and sampling effort units

Once all the distributional information available has been compiled, it is necessary to find a homogenous data unit which might be used to assess how completes are the inventories extracted from the database all over the studied territory. This data unit can be partitioned into two main parts: the spatial definition of the selected territorial unit, and the sampling effort measure for the information referred to each one of these units.

The territorial unit must be the minimum area possible taking into account:

i) availability and accuracy of biological information,
ii) real biogeographic meaning,
iii) spatial resolution of the available environmental information and
iv) total surface area of the studied region.

It is advisable to choose territorial units that maintain similar or comparable surface areas regardless its spatial location inside the studied region (spatial scale), independently from the total area (extent *sensu* reference 42) or the geographical situation of the region. Moreover, it is also important that these units could be easily aggregated into bigger ones (nestedness). Due to this, common reference grids, such as those based on Geographical Coordinates (i.e. one degree grid), or those derived from the Universal Transverse Mercator system (i.e. 10 Km. UTM grid), remain as the best-possible option.

To identify the main biogeographic patterns present in the spatial distribution of species richness in the Iberian Peninsula, we have chosen the 252 50X50 km Iberian UTM grid cells with more than 85% of land surface. We have decided to use the UTM grid, better than using the Lat/Long one, because, over extensive territories, its cells present the same surface area. The 50 km^2 cell was chosen because of the general purpose of the analysis, as the use of a smaller one (i.e. 10 km^2) would, maybe, obscure some general patterns, due to the 'noise' produced by local processes, such as micro-ecological and population dynamics-related ones. It also provides an enough number of cases (252), and allows us to use all the information of the environmental database (see section 2.1).

On the other hand, the sampling effort measure must be easily taken out from all the kinds of information compiled in the database, and a good surrogate, whether it is not a direct measure, of the real sampling effort (i.e. time/person or number of traps). As usually occurs, the distributional information stored in BANDASCA comes from a bunch of studies carried out with different methodologies and purposes, in which no information is provided about the sampling effort carried out. To overcome this picture, we have chosen as sampling effort unit the number of database-records contained in the database for each UTM 50X50 grid cell, on the assumption that species occurrence probability in a site positively correlates with the number of database-records. Here, the database record is defined as a pool of specimens of a single species with identical database field information (locality, altitude, date of capture, type of habitat and food resource, among others; see BANDASCA database [41]) regardless of the number of specimens. Any difference in any database field value gives rise to a new database-record. As database records have been proved to be a good surrogate of sampling effort, being its scores highly correlated with direct measures of sampling effort (i.e. number of traps or days/person; unpublished data), increments of the number of database-records provide correlative increments of the sampling effort.

2.3 Identifying the well-sampled territorial units

When mapping the geographical distribution of the sampling effort extracted from BANDASCA, a biased map appears (Figure 1). It is known that biogeographic diversity patterns for almost all insect groups reflect the distribution of the areas investigated by entomologists [43], and in the Iberian Peninsula, the geographic distribution of the entomologists themselves, rather than the geographic distribution of organisms [31, 32, 44]. So, as no assess on how complete are the inventories for each grid cell is available, where does the inventories are reliable, and where does not?

We have identified the adequately sampled grid cells by investigating, for each grid cell, the increase in the number of species present in the inventory when new database records are added. To eliminate the bias produced by the order each database record is included in the curve, this order was randomized 500 times [45]. The adequacy of sampling in each square was determined by a negative exponential function that describes the rate of species added to the inventory (S_r) with the increase in sampling effort (number of database records; r) [45-48]. According to Soberón and Llorente [46] and Colwell and Coddington [47] this relationship is given by

$$S_r = S_{\max} [1\text{-}\exp(-br)]$$

Where $S_{\max}$, the asymptote, is the estimated total number of species per square, and b is a fitted constant that controls the shape of the curve. The curvilinear function was fitted by the quasi-Newton method. Because 100% richness requires an infinite number of database records, the number of records required for a rate of species increment ≤ 0.01 (i.e. one added species each 100 records; $r_{0.01}$) was calculated:

$$r_{0.01} = 1/b \ln (1 + b/0.01) \text{ [46].}$$

We have considered as well sampled enough those grid cells where the total number of database records is higher than its $r_{0.01}$ score (it is necessary a sampling effort of more than 100 records to find a new species for the inventory). 82 UTM squares were selected by this method. All of them have been used to forecast Scarabaeinae species richness over the rest of the Iberian territory.

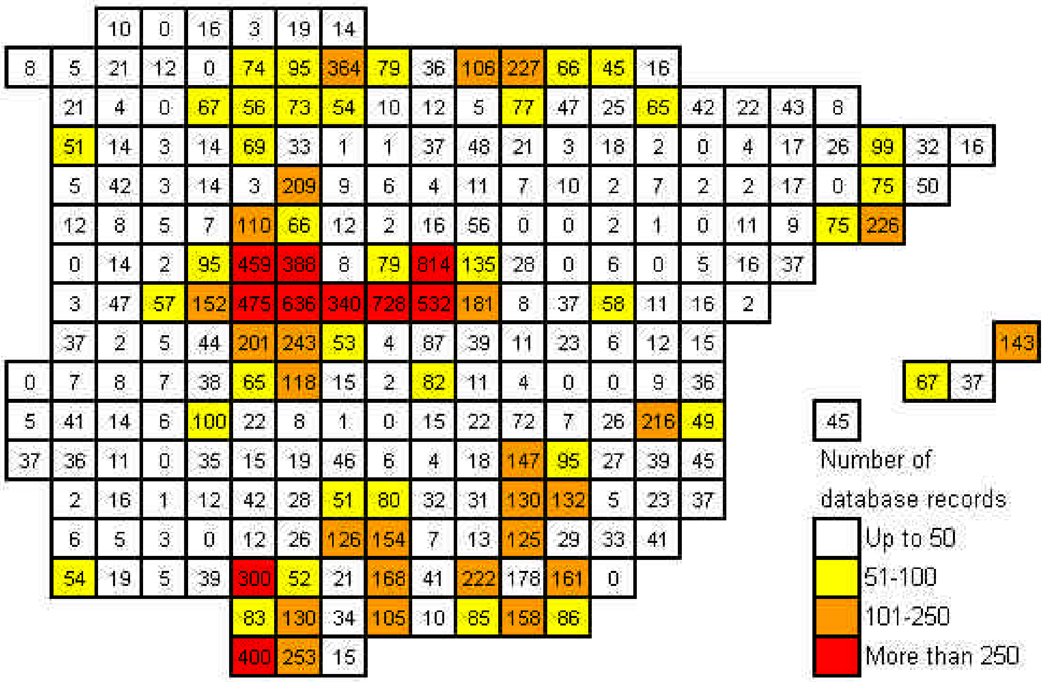

Figure 1: Number of database records compiled in BANDASCA database of Iberian Scarabaeinae for each UTM 50X50 grid cell. Modified from reference 31.

3 Data modelling

Once territorial units with reliable data have been identified, the next step is to spread this knowledge over all the territory. We must be able to extract the patterns, both environmental and spatial, from this sample, and extrapolate them by means of a modelling technique. Patterns that arise in species distribution are the result of three main causes:

i) the adequacy of the environment to their ecological requirements (ecological niche),

ii) its dispersal and colonization capabilities (vagility), and

iii) unique and contingent events (history).

While the variables that account for the first factor are easy to find and model, for the second and, specially, the latter, this goal results an almost impossible task. However, in spite of the difficulty of reconstructing the historical processes that have leaded to present species distribution, we assume that they may produce a spatial pattern [49]. Due to this, the modelling technique used must use both environmental information, which accounts for the potential distribution of the species, and that information stored in the spatial structure of the data (historical influence).

3.1 Origin of the predictor variables

Explanatory variables for the entire territory investigated can be extracted from environmental and topographic digital maps included in a GIS, overlaying this data layers with the polygons of each territorial unit [2,7]. At least, the available information must include, at the best possible resolution, the following datasets:

- Environmental Data, with three main groups:
  i) Climatic data, including common variables such as mean temperature or annual precipitation, and other variables which may be important for the group studied, such as summer precipitation or number of days with freeze.

  ii) Geomorphologic data, such as a Digital Elevation Model (DEM) and variables derived from it (slopes, aspect, watersheds, etc.).

  iii) Substrate-referred data, such as geology, soil type or hydrology.
  - Land Use/Land Cover maps, to include actual land occupancy into the estimate.
  - Other relevant information, such as variables which could produce negative impacts on the group species' distribution, socio-economic variables, Remote Sensing products, etc.

In the example case, for the adequately sampled grid cells 24 continuous variables were extracted using the GIS software Idrisi 2.0 [50]: two spatial variables (central latitude and central longitude); two geographic (distance from Pyrenees and sea area); three topographic (minimum, maximum and mean elevation); two geologic (calcareous and acid rock surface); six climate (minimum and maximum monthly mean temperature, annual mean temperature, total annual and summer precipitation, and annual days of sun); four land use (cultivated and urban area, forest, scrub and grassland area); and five environmental diversity variables (altitude range, annual temperature precipitation variation, land use and geologic diversity).

The climate data for each square are courtesy of W. Cramer (CLIMATE database v.2; http://www.pik-potsdam.de/~cramer/climate.htm). Annual temperature variation and annual precipitation variation are taken as the difference between the most extreme monthly values. The topographic and spatial data were extracted from a DEM of the Iberian Peninsula (grid cell size of 1 km) with the polygons of grid cells using the GIS [50]. Land use data come from a reclassification of the raster information (282 meter resolution) on the 44 land cover categories present in Spain and Portugal provided by the European Environment Agency (CORINE Programme 1985-1990) [51]. The different land cover categories have been grouped into four: forest (all types of forests), scrub and grassland area surface (either natural or artificial), and the surface of areas with strong anthropic influence (urban, industrial and cultivation zones). Geologic data were obtained by digitising, from an Iberian map [52], soils on calcareous rocks, on acid rocks (siliceous), and soils on clay (spatial resolution of 1 km). Land use and geologic diversity in each grid square was estimated using the Shannon diversity index [53]:

$$H' = - \sum pi \log_2 pi$$

where *pi* is the relative frequency of each one of the 44 land cover categories [51], and of each geologic category, respectively.

Predictor variables were standardised to zero means and unit variances to eliminate the effect of differences in the measurement scale for the different independent variables. The algorithm used for the standardization was:

$$\text{Std. Value} = (\text{raw value} - \text{mean}) / \text{std. deviation}$$

except for the case of LAT and LON, that were standardized as recommended by Legendre and Legendre [49]:

$$\text{Std. Value} = \text{raw value} - \text{mean}$$

3.2 Model building

There are four main problems with the use of environmental variables to build species richness predictive models: 1) the collinearity, and thus interdependence, of predictor variables used; 2) the spatial autocorrelation of variables; 3) the usual non-linear relationship between the dependent and independent variables; and 4) the frequently complex interactions among explanatory variables.

Environmental variables are usually correlated among themselves, making them collinear. This collinearity may bias model parameter estimation, but, if the aim is to forecast, maximizing the explained variance of the data, with no ecological inference, collinearity of the explanatory variables is not a concern [49].

As spatial heterogeneity in nature is the result of non-random processes, spatial autocorrelation is also an intrinsic propriety of biological and environmental variables [49]. Autocorrelation in one variable implies the spatial dependence of observations, invalidating the assumption on which classical statistical tools are based, since the observed values of variables at any given locality are influenced by those of the neighboring localities. What can be done in this case? The removal of spatial autocorrelation, a consequence of the processes that lead to species richness spatial patterns, would diminish the impact of spatially structured factors, thus reducing the forecasting ability of the model [49, 54]. Thus, the key criterion when developing regression models with spatially autocorrelated data is to check if function errors (residual scores) from the final model are spatially autocorrelated. In this case, at least one spatially structured variable was not included in the analysis [55, 56]. Spatial variables have been included in the modelling procedure (see below) in order to include these hypothetically ignored variables, as well as spatial patterns produced by historical processes. To check that function errors are not spatially structured Moran's I and Geary's C autocorrelation tests were used [57], dividing all the possible UTM50 pairwise comparisons in eight distance classes.

We have used a Generalized Linear Model (GLM) stepwise, heuristic procedure to model variation in Scarabaeinae species richness as a function of the most significant environmental and spatial explanatory variables [58, 59] (see references 12, 19, 24 and 60-62 for a description of the method and some examples). A Poisson error distribution for the number of Scarabaeinae species was assumed, and was linked to the set of predictor variables by means of a logarithmic link function [63].

The adequacy of the models developed was tested by means of the change from a null model in which the number of parameters is equal to the total number of observations (n=82) and the species richness is modelled alone (with no explanatory variables; see reference 59). The goodness-of-fit of the models was measured by the deviance statistic and the change in deviance F-ratio tested [58, 59], with a change in deviance significance level of 0.05.

A forward stepwise procedure was used to enter the variables into the model. In order to account for non-linear relationships, in a first step the total number of Scarabaeinae species registered on each adequately-sampled UTM50 was related, one-by-one, with each predictor variable's linear, quadratic and cubic function. The function that accounted for the highest reduction in deviance from that of the null model was selected [10, 19, 60]. Next, from the functions selected before, the one that accounted for the most important change in deviance was chosen. Then all the remaining functions were added one-by-one to the model and tested again for significance in the change of deviance. The one which accounted for the most significant change was included in the model. After each significant inclusion, the new model was submitted to a backward selection procedure, in order to eliminate those terms that had become non-significant. This procedure was repeated iteratively until no more statistically significant changes remained.

As interactions between variables are often highly predictive [10], subsequently the importance of all the interaction terms between explanatory variables (including spatial) was tested by adding them sequentially one by one to the previously obtained model. Again, a backward procedure was used after each forward inclusion.

Finally, spatial variables were added to the model. As commented before, this would include in our analysis the influence of ignored spatially structured variables, also diminishing the probability of occurrence of spatially autocorrelated residuals. The nine terms of the third-degree polynomial equation of latitude and longitude (Trend Surface Analysis; $b_1LAT + b_2LON + b_3LAT^2 + b_4LAT\times LON + b_5LON^2+b_6LAT^3 + b_7LAT^2\times LON + b_8LAT\times LON^2 + b_9LON^3$) have been added to the former model and submitted to the backward stepwise selection procedure in order to remove the non-significant terms (see references 49 and 64).

3.3 Residual analysis

Once a preliminary model was chosen, the residual analysis recommended by Nicholls [12] was carried out to identify outliers, those grid squares in which the residual absolute value is greater than the standard deviation of the predicted values. Points with high scores of Potential Leverage (PLV) were also selected. The PLV is a measure of the distance of each observation from the centroid of the multi-dimensional space defined by the variables included in the model. Each of the outliers and observations with high PLV was explored in order to ascertain if it were due to erroneous data, or if it included part of the environmental variability of the investigated territory different from the rest of the observations. While the former should be deleted, the latter kind of observations may remain in the model in order to include as much environmental heterogeneity as possible. The final model parameters were then estimated after the deletion of the real outliers.

3.4 Goodness-of-fit and predictive power of the models

The best method to test model reliability is empirical: an inventory taken in the poorly sampled zones to check if predicted and real scores are similar. However, this method is, in most cases, too much expensive and slow. Several statistical tools can be used on the dataset to reach this objective. To check the final model, a Jackknife test was carried out. With a data set of *n* grid cells the model was recalculated *n* times, leaving out one square in turn. Each one of the regression models based on the *n*-1 grid squares was then applied to that excluded square, to predict species richness score in each territorial unit. Then observed and estimated values were checked for correlation using the Pearson correlation coefficient.

The percentage of explained deviance for each model was calculated to obtain an estimation of the total variability of the data explained by each model [59]. Moreover, to estimate the predictive power of the model, the relative distance between the predicted value for case i when excluded in the model building (P_i) and the observed score (O_i) is used as a prediction error (E_i) for that observation [65]. The percentage error for case i is:

$$E_i = \frac{|O_i - P_i|}{O_i} \times 100$$

The mean of all the error estimations (Mean Prediction Errors; MPE) was used as a measure of the prediction error associated with the model, and the inverse of this measure (MPE^{-1}=100–MPE) as an estimation of the predictive power of it.

3.5 Modelling Iberian dung beetle richness

After the ascertainment of all adequately sampled 50 x 50 km grid cells (n = 82), a preliminary model was developed. Residuals and PLV scores derived from this model where investigated. Seven grid cells, located in the most intensively surveyed regions were supposed to be real outliers [31]. Therefore, to build the final predictive model of species richness these seven grid cells were removed (n = 75).

When tested separately as either linear, quadratic, or cubic functions, 8 environmental variables (distance from Pyrenees, minimum, maximum, and annual mean temperature, total annual precipitation, annual days of sun, grassland area, and land-use diversity) and the quadratic and cubic terms of latitude were significant. As the linear function of annual days of sun accounted for the most important change in deviance, this variable was the first included in the model; then, the linear function of maximum elevation was selected, followed by the quadratic function of grassland area, although only its quadratic term accounted for a significant change in deviance. In the next step, the cubic function of land-use diversity was the only environmental diversity variable added to the model, being its quadratic term deleted from the model. From the interaction terms, the joint effect of forest area and geologic diversity, terrestrial area in grid square x maximum elevation, latitude x annual precipitation variation, and calcareous rocks x geologic diversity entered iteratively in the model, but their inclusion removed the cubic term of land use diversity. Finally, significant spatial terms (the quadratic and cubic terms of latitude) were added to the previous model, although the linear term of annual days of sun and the interaction between latitude and annual precipitation variation were removed during the selection procedure. The final model was:

$$S = \exp [c + \text{Maximum elevation} + (\text{Grassland area})2 + \text{Land use diversity} + (\text{Forest area} \times \text{Geologic diversity}) + (\text{Terrestrial area in grid square} \times \text{Maximum elevation}) + (\text{Calcareous rocks} \times \text{Geologic diversity}) + \text{Latitude2} + \text{Latitude3}],$$

where S is the total number of dung beetles and c is the intercept. This model explains 62.41% of the total deviance [31].

Residuals from this new model were normally distributed, the plot of residuals versus predicted values form a homogeneous cloud around the center, and the standard errors of the coefficients were low. None of Moran's I values in the different lag classes were significant at a 0.05 significance level with the Bonferroni correction (they were not spatially autocorrelated). However, plotting predicted versus observed values showed an overestimation of the number of species at low species richness squares, and an underestimation at high species richness ones. This drawback of the model remains even if we eliminate all observations that could potentially be considered outliers. The results of a jackknife test on the final model show strong correspondences between observed species richness scores and those predicted by the jackknife procedure for the 75 squares. The correlation between observed and values was positive and significant, being the MPE^{-1} = 84.1%, showing that the predictive model was reasonably good in spite of its tendency to reduce the difference between the lowest and highest species richness scores [31].

Similar analysis of Scarabaeinae species richness distribution carried out over Portuguese (0.01 confidence level; explained deviance = 85.4%; MPE-1 = 90.8%) [30] and French (0.05 confidence level; explained deviance = 86.2%; MPE-1 = 82.3%) [33] territories obtained better scores both for explained deviance and MPE^{-1}. This may be due to the higher environmental homogeneity of both territories with respect to the entire Iberian Peninsula, pointing out that, as may be expected, patterns arising from heterogeneous processes are more difficult to model.

4 Discussion

From the presented results it can be ascertained that it is possible to produce a reliable geographical estimation of the species richness of a given group by means of an exhaustive compilation of distributional data, an assess of the sampling level of the territory studied, and a GIS-based environmental and land-use dataset in which the spatial structure of the response variable must be taken into account. The forecasted maps may allow us to identify major spatial patterns of many biodiversity attributes [4], and also can be used in conservation policies.

However, dung beetles are one of the taxonomically best-known insect groups, and, due to their attractiveness and easy capture, one of the more collected by professional and amateur entomologists, just one step beyond butterflies. In the Iberian Peninsula, all the present species are known and can be identified [40], and there exists a long tradition of scientific study over them. Unfortunately, this picture is not the same for the great majority of living groups and world regions.

With so much inventorying work to be done for nearly all biota worldwide, and the pressing need to identify the species richness geographical distribution due to the so called biodiversity crisis, how can the spatial distribution of biodiversity be described within a reasonable time? There are too many unstudied groups, too many unexplored regions, very few resources and very few taxonomists in the world. As sampling all poorly explored or unexplored territories, and identifying and classifying all the unknown species, seems to be impractical in the short term [32, 66-72], other strategies must be used. Hence, it is necessary to design procedures to quickly and cheaply identify areas of greatest diversity, so it seems to be more reasonable to:

– develop new methodologies of sampling design at regional scales which allow us to maximize our inventorying capability (recover more species with less economic and time costs),

– build extensive databases of species distributional data, with general sharing protocols which allow its general and easy use and comparison via WWW,

– and find easy-to-use predictive modelling techniques based on environmental and spatial variables.

To deal with the latter task, a joint collaboration between biologists and mathematicians is needed, in order to test all the different methodologies able to produce forecasted maps of diversity attributes such as species richness (heuristic modelling, niche modelling, exhaustive model search, spatial interpolation techniques such as kriging and co-kriging, etc.) at as much spatial scales and over as much regions and groups as possible. The strengths and weaknesses of each technique at each spatial scale on each kind of region and each group must be analysed, allowing an accurate assessment of which method and spatial scale must be used in each case, in terms of simplicity, accuracy and time consume. Once this information is known, regional surveys could be designed to reach faunistic levels of knowledge enough to allow a reliable forecast with the less possible cost.

References

[1] A. Purvis and A. Hector. Getting the measure of biodiversity. *Nature*, 405(6783): 212-219, 2000.

[2] R. I. Miller. *Mapping the Diversity of Nature*. Chapman and Hall, 1994.

[3] K. J. Gaston. Species richness: measure and measurement. pp. 77-113 in K. J. Gaston (ed.) *Biodiversity. A biology of numbers and difference.* Blackwell Science, 1996.

[4] S. A. Levin. The problem of pattern and scale in ecology. *Ecology*, 73(6): 1943-1967, 1992.

[5] J. H. Brown. *Macroecology*. University of Chicago Press, 1995.

[6] J. H. Lawton. Are there general laws in ecology? *Oikos*, 84(2): 177-192, 1999.

[7] C. A. Johnston. *Geographic Information Systems in Ecology*. Blackwell Science, 1998.

[8] J. M. Lobo, I. Castro and J. C. Moreno. Spatial and environmental determinants of vascular plant species richness distribution in the Iberian Peninsula and Balearic Islands. *Biological Journal of the Linnean Society,* 73: 233-253, 2001.

[9] M. P. Austin, R. B. Cunninghamand and P. M. Fleming. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Vegetatio,* 55: 11-27, 1984.

[10] C. R. Margules, A. O. Nicholls and M. P. Austin. Diversity of *Eucalyptus* species predicted by a multi-variable environment gradient. *Oecologia,* 71: 229-232, 1987.

[11] C. R. Margules and J. L. Stein. Patterns in the distribution of species and the selection of nature reserves: an example from Eucalyptus forest in south-eastern New South Wales. *Biological Conservation,* 50: 219-238, 1989.

[12] A. O. Nicholls. How to make biological surveys go further with generalised linear models. *Biological Conservation* 50: 51-75, 1989.

[13] M. P. Austin, A. O. Nicholls and C. R. Margules. Measurement of the realised qualitative niche: environmental niches of five *Eucalyptus* species. *Ecological Monographs,* 60: 161-177, 1990.

[14] P. A. Walker. Modelling wildlife distribution using a geographic information system: kangaroos in relation to climate. *Journal of Biogeography,* 17: 279-289, 1990.

[14] S. Ferrier. and P. A. Smith. Using geographical information systems for biological survey design, analysis and extrapolation. *Australian Biologist*, 3: 105-116, 1990.

[15] D. B. Lindenmayer, H. A. Nix, J. P. McMahon, M. F. Hutchinson and M. T. Tanton. The conservation of Leadbeater’s Possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography,* 18: 371-383, 1991.

[15] P. A. Walker and K. D Cocks. HABITAT: A procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*, 1: 108-118, 1991.

[16] P. E. Osborne and B. J. Tigar. Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. *Journal of Applied Ecology*, 29: 55-62, 1992.

[17] S. T. Buckland and D. A. Elston. Empirical models for the spatial distribution wildlife. *Journal of Applied Ecology*, 30: 478-495, 1993.

[18] G. Carpenter, A. N. Gillison and J. Winter. DOMAIN: A flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2: 667-680, 1993.

[19] M. P. Austin, G. J. Pausas and A. O. Nicholls. Patterns of tree species richness in relation to environment in south-eastern New South Wales, Australia. *Australian Journal of Ecology*, 21: 154-164, 1996.

[20] G. E. Austin, C. J, Thomas, D. C. Houston and D. B. A. Thompson. Predicting the spatial distribution of buzzard *Buteo buteo* nesting areas using a geographical Information system and remote sensing. *Journal of Applied Ecology*, 33: 1541-1550, 1996.

[21] S. T. Buckland, D. A. Elston and S. J. Beaney. Predicting distributional change, with application to bird distributions in northeast Scotland. *Global Ecology and Biogeography Letters*, 5: 66-84, 1996.

[22] J. B. Kirkpatrick and M. J. Brown. A comparison of direct and environmental domain approaches to planning reservation of forest higher plant communities and species in Tasmania. *Conservation Biology*, 8: 217-224, 1994.

[23] R. K. Heikkinen. Predicting patterns of vascular plant species richness with composite variables: a meso-scale study in Finnish Lapland. *Vegetatio*, 126: 151-165, 1996.

[24] R. K. Heikkinen and S. Neuvonen. Species richness of vascular plants in the subarctic landscape of northern Finland: modelling relationships to the environment. *Biodiversity and Conservation*, 6: 1181-1201, 1997.

[25] F. Skov and F. Borchsenius. Predicting plant species distribution patterns using simple climatic parameters: a case study of Ecuadorian palms. *Ecography*, 20: 347-355, 1997.

[26] L. R. Iverson and A. M. Prasad. Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs*, 68: 465-485, 1998.

[27] V. Parker. The use of logistic regression in modelling the distributions of bird species in Swaziland. *South African Journal of Zoology*, 34: 39-47, 1999.

[28] K. J. Wessels, S. Freitag and A. S. van Jaarsveld. The use of land facets as biodiversity surrogates during reserve selection at a local scale. *Biological Conservation*, 89: 21-38, 1999.

[29] S. M. Lenton, J. E. Fa, and J. Pérez del Val. A simple non-parametric GIS model for predicting species distribution: endemic birds in Bioko Island, West Africa. *Biodiversity and Conservation*, 9: 869-885, 2000.

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

[31] J. M. Lobo and F. Martín-Piera. Searching for a predictive model for Iberian dung beetle species richness based on spatial and environmental variables. *Conservation Biology*, in press.

[32] F. Martín-Piera and J. M. Lobo. Database records as a sampling-effort surrogate to predict spatial distribution of insects in either poorly or unevenly surveyed areas. *Acta Zoologica Ibérica e Macaronésica*, in press.

[33] J. M. Lobo, J. P. Lumaret and P. Jay-Robert. Modelling the species richness distribution of French dung beetles and delimiting the predictive capacity of different groups of explanatory variables (Coleoptera, Scarabaeidae). *Global Ecology and Biogeography*, in press.

[34] N. D. Mitchell. The derivation of climate surfaces for New Zealand, and their application to the bioclimatic analysis of the distribution of kauri (*Agathis australis*). *Journal of the Royal Society of New Zealand*, 21: 13-24, 1991.

[35] D. R. B. Stockwell and D. Peters. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, 13(2): 143-158, 1999.

[36] A. T. Peterson, J. Soberón and V. Sánchez-Cordero. Conservatism of ecological niches in evolutionary time. *Science*, 285: 1265-1267, 1999.

[37] J. M. Scott, F. W. Davis, B. Csuti, R. Noss, B. Butterfield, C. Groves, H. Anderson, S. Caicco, F. D'Erchia, T.C. Edwards, J. Ulliman and R.G. Wright. Gap analysis: a geographic approach to protection of biological diversity. *Wildlife Monographs*, 123: 1-41, 1993.

[38] J. M. Scott and M. D. Jennings. A description of the National GAP Analysis Program. U.S. Geological Survey technical report, published at http://www.gap.uidaho.edu/About/Overview/GapDescription/default.htm.

[39] M. Cortés, Y. Villacampa, J. Mateu and J. L. Usó. A new metholology for modelling highly structured systems. *Journal of Environmental Modelling and Software,* 15: 461-470, 2000.

[40] F. Martín-Piera. Familia Scarabaeidae. In F. Martín-Piera and J.I. López-Colón *Fauna Ibérica 14. Coleoptera, Scarabaeoidea I.* Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, 2000.

[41] J. M. Lobo and F. Martín-Piera. La creación de un banco de datos zoológico sobre los Scarabaeidae (Coleoptera: Scarabaeoidea) íbero-baleares: una experiencia piloto. *Elytron*, 5: 31-38, 1991.

[42] R. J. Whittaker, K. J. Willis and R. Field. Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography*, 28: 453-470, 2001.

[43] R. L. H. Dennis and P. B. Hardy. Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. *Global Ecology and Biogeography,* 8: 443-454, 1999.

[44] J. Martín and P. Gurrea. Áreas de especiación en España y Portugal. *Boletín de la Asociación Española de Entomología* 23: 83-103, 1999.

[45] R. K. Colwell. *EstimateS, statistical estimation of species richness and shared species from samples. Version 5.* User's Guide and application available from http://viceroy.eeb.uconn.edu/estimates, 1997.

[46] M. J. Soberón and B. J. Llorente. The use of species accumulation functions for the prediction of species richness. *Conservation Biology,* 7: 480-488, 1993.

[47] R. K. Colwell and J. A. Coddington. Estimating terrestrial biodiversity through extrapolation. pp. 101-118 in D. L. Hawksworth (ed.) *Biodiversity, measurement and estimation*. Chapman and Hall, 1995.

[48] W. F. Fagan and P. M. Kareiva. Using compiled species list to make biodiversity comparisons among regions: a test case using Oregon butterflies. *Biological Conservation,* 80: 249-259, 1997.

[49] P. Legendre and L. Legendre. *Numerical Ecology, 2nd edition*. Elsevier, 1998.

[50] Clark Labs. *Idrisi 2.0.* Clark University, 1998.

[51] European Environment Agency. *Natural Resources CD-Rom*. European Environment Agency, 1996.

[52] Instituto Geográfico Nacional. *Atlas nacional de España. Volume 1 and 2.* Centro Nacional de Información, 1995.

[53] A. E. Magurran. *Ecological diversity and its measurement*. Princeton University Press, 1988.

[54] P. A. Smith. Autocorrelation in logistic regression modelling of species' distributions. *Global Ecology and Biogeography Letters*, 4: 47-61, 1994.

[55] A. D. Cliff and J. K. Ord. *Spatial Processes. Models and Applications*. Pion Limited, 1981.

[56] J. Odland. *Spatial Autocorrelation*. Sage Publications Inc., 1988.

[57] P. Legendre and P. Vaudor. *The R Package: Multidimensional analysis, spatial analysis*. Département de sciences biologiques, Université de Montréal, 1991.

[58] P. McCullagh and J. A. Nelder. *Generalized Linear Models (2nd ed).* Chapman and Hall, 1989.

[59] A. Dobson. *An introduction to Generalized Linear Models*. Chapman and Hall/CRC, 1999.

[60] M. P. Austin. Searching for a model for use in vegetation analysis. *Vegetatio,* 42: 11-21, 1980.

[61] A. O. Nicholls. Examples of the use of Generalised Linear Models in analysis of survey data for conservation evaluation. pp 54-63 in C. R. Margules and M. P. Austin (eds.) *Nature Conservation: Cost-effective Biological Surveys and Data Analysis*. CSIRO, 1991.

[62] T. Tonteri. Species richness of boreal understorey forest vegetation in relation to site type and successional factors. *Annali Zoologi Fennici,* 31: 53-60, 1994.

[63] M. J. Crawley. *GLIM for Ecologists*. Blackwell Scientific Publications, 1993.

[64] P. Legendre. Spatial autocorrelation: trouble or new paradigm? *Ecology,* 74: 1659-1673, 1993.

[65] M. A. Pascual and O. O. Iribarne. How good are empirical predictions of natural mortality? *Fisheries Research,* 16: 17-24, 1993.

[66] R. May. Taxonomy as destiny. *Nature*, 347: 129-130, 1990.

[67] P. R. Ehrlich. Population biology of checkersport butterflies and the preservation of global biodiversity. *Oikos,* 63: 6-12, 1992.

[68] Systematics Agenda 2000. *Systematics Agenda 2000: Charting the Biosphere.* Technical Report. American Society Plant Taxonomist, Society of Systematics Biologist, Willi Hennig Society, Association of Systematics Collection, 1994.

[69] F. P. D. Cotterill. Systematics, biological knowledge and environmental conservation. *Biodiversity and Conservation*, 4: 183-205, 1995.

[70] P. H. Williams and C. J. Humphries. Comparing character diversity among biotas. pp 54-76 in K. J. Gaston (ed.) *Biodiversity: A biology pf numbers a difference*. Blackwell Science, 1996.

[71] S. Blackmore and D. Cutler (eds.). *Systematic Agenda 2000. The challenge for Europe. Proceedings of Workshop organized by the European Science Foundation Systematic Biology Network, the Linnean Society of London, the Rijksherbarium/Hortus Botanicus, Leiden University, and the Systematics Association.* Linnean Society of London, 1996.

[72] F. Martín-Piera. Apuntes sobre biodiversidad y conservación de insectos: Dilemas, ficciones y ¿soluciones?. *Boletín de la Sociedad entomológica aragonesa*, 20: 25-55, 1997.