Environ Ecol Stat (2006) 13:237–245
DOI 10.1007/s10651-005-0003-3

ORIGINAL ARTICLE

Obtaining environmental favourability functions from logistic regression

Raimundo Real · A. Márcia Barbosa · J. Mario Vargas

Received: June 2005 / Revised: July 2005

Abstract Logistic regression is a statistical tool widely used for predicting species' potential distributions starting from presence/absence data and a set of independent variables. However, logistic regression equations compute probability values based not only on the values of the predictor variables but also on the relative proportion of presences and absences in the dataset, which does not adequately describe the environmental favourability for or against species presence. A few strategies have been used to circumvent this, but they usually imply an alteration of the original data or the discarding of potentially valuable information. We propose a way to obtain from logistic regression an environmental favourability function whose results are not affected by an uneven proportion of presences and absences. We tested the method on the distribution of virtual species in an imaginary territory. The favourability models yielded similar values regardless of the variation in the presence/absence ratio. We also illustrate with the example of the Pyrenean desman's (*Galemys pyrenaicus*) distribution in Spain. The favourability model yielded more realistic potential distribution maps than the logistic regression model. Favourability values can be regarded as the degree of membership of the fuzzy set of sites whose environmental conditions are favourable to the species, which enables applying the rules of fuzzy logic to distribution modelling. They also allow for direct comparisons between models for species with different presence/absence ratios in the study area. This makes them more useful to estimate the conservation value of areas, to design ecological corridors, or to select appropriate areas for species reintroductions.

Keywords Biogeographic inferences · Distribution modelling · Fuzzy logic · *Galemys pyrenaicus* · Model comparison · Presence/absence ratio · Virtual species

R. Real (✉) · A. M. Barbosa · J. M. Vargas
Laboratorio de Biogeografía, Diversidad y Conservación,
Departamento de Biología Animal, Facultad de Ciencias, Universidad de Málaga,
29071 Málaga, Spain
e-mail: rrgimenez@uma.es

1. Introduction

Logistic regression is a statistical tool that relates a binary dependent variable to a set of discrete or continuous independent variables. It is useful for making inductive inferences using a particular sample of data, in a way that the probability of occurrence of each state of the target variable may be deduced from the values of the predictor variables. Logistic regression has shown to be a powerful tool that produces robust models, and it is broadly used in the predictive modelling of species' distributions starting from presence/absence data (e.g. Romero and Real 1996; Bustamante 1997; Franco et al. 2000; Madsen and Prang 2001; Seoane and Bustamante 2001).

However, classification success using logistic regression is sensitive to the relative proportion of presences and absences in the sample, independently of the fit of the model (Hosmer and Lemeshow 1989, p. 147; Rojas et al. 2001). The logistic function is symmetric by definition, and its inflection point corresponds to a probability (P) value of 0.5. This value is commonly used as a default threshold above which to assume that the model predicts species presence. However, when the proportions of presences and absences are not equal within the sample, the logistic regression output within the function's domain is not symmetrical, but rather deviates towards the extreme that has a greater number of cases (Rojas et al. 2001). In this way, the probabilities are biased towards the state that is more frequent within the sample, but not necessarily as frequent outside it. In these situations, 0.5 is indeed the probability threshold above which presence is more likely than absence within the studied sample, but does not necessarily correspond to the environmental threshold (dependent on the predictor variables) above which presence is more likely than expected at random (i.e., than expected considering the presence/absence ratio in the sample). Consequently, if presences and absences are not evenly distributed in the studied territory, which is the most common situation, the probability values yielded by the logistic function cannot be considered to reflect actual environmental favourability.

A few strategies have been proposed to bypass this drawback when modelling species' distributions. One is to use subsets of data containing 50% presence (e.g. Brito et al. 1999), but this implies discarding valuable information and requires the choice of one specific subset among all possible subsets of data, or else leads to the production of a number of different models to account for the same distribution. Besides, for species with a restricted distribution, with a very limited number of presences, there could be not enough data to build a significant model.

Other authors equilibrate the impact of presences and absences through a weighting procedure (e.g. Teixeira et al. 2001), but this produces an alteration to the original data. The numbers of cases actually observed are artificially changed and the model produced is different from that which would have been obtained using the data as they are.

When discrimination is the main goal, one can also build the model using the whole dataset, determine the probability threshold at which most presences and absences are classified correctly, and take it as the cut-off point above which to consider that the model predicts the species to be present (Rojas et al. 2001; Barbosa et al. 2003). However, the accuracy of the threshold obtained depends on how many probability intervals are analysed. Besides, the actual values produced by the model remain unchanged, so models for species with different presence/absence ratios (hence, different probability thresholds) cannot be directly compared, as the predicted values for commoner species will be generally higher independently of the actual environmental favourability. This is a problem if we aim to contrast distribution models for related species such as a predator and its main prey, or if we want to use environmental favourability for multiple species as a basis for defining important areas for conservation, such as potential diversity hotspots or areas of rarity.

In this paper, we propose using a modification to the *logit* equation to obtain from logistic regression a favourability function whose output values, based on all the distribution data available, are independent of the proportion of presences in the dataset. We tested this function on the distribution of an imaginary species in an imaginary territory, using both the whole dataset and samples with uneven presence/absence ratios. We also compared distribution models for two other virtual species with similar responses to an imaginary independent variable but with different presence/absence ratios in the imaginary territory, using both the probability and the favourability functions. Finally, we applied the favourability function to the distribution of the Pyrenean desman (*Galemys pyrenaicus*) in Spain. We discuss the implications of this procedure for its combination with fuzzy logic and for comparative distribution modelling.

2. Theoretical approaches

The logistic regression model has the form

$$P = \frac{\mathrm{e}^{y}}{1 + \mathrm{e}^{y}} \tag{1}$$

where P is the probability of an event occurring (e.g., the probability of presence of a species), e is the basis of the natural logarithm, and y is a regression equation of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

where α is a constant and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the n predictor variables $x_1, x_2, \ldots, x_n$ (Tabachnick and Fidell 1996, p. 127). This linear regression is the *logit* or log of the odds

$$y = \ln\left(\frac{P}{1 - P}\right) \tag{3}$$

in the inductive logistic model, i.e., (2) is the (natural log of the) probability of the event occurring divided by the probability of it not occurring.

Our proposal follows two different rationales. The first one is logical. The probabilities inferred using logistic regression are composed of two elements: the random probability of presence, which is given by the proportion of presences within the dataset, and the modification of this probability caused by the values of the predictor variables. The deductive use of logistic regression modelling deals mainly with this second element, as its aim is often to deduce the probability of presence outside the sample population (e.g. in a non-sampled part of the territory, or on a finer resolution scale) only in relation to the values of the predictor variables, without assuming any proportion of presences beforehand. We should therefore eliminate the first component, which can be achieved by subtracting from (2) the value of y that corresponds to the random expectation of presences (i.e., to the species' prevalence in the studied territory). This value is, according to (3), $\ln(P_{\mathrm{r}}/1 - P_{\mathrm{r}})$, where P_{r} is the presence ratio (number of presences divided by the total number of cases) in the sample, which may also be written as $\ln(n_1/n_0)$, where n_1 is the number of presences and n_0 the number of absences. The value resulting from the new logistic function is a measure of environmental favourability for the species, as it indicates how the presence probabilities differ from those expected at random.

The second rationale is mathematical. The estimation of α in (2) differs from those of β_i in that it includes the term $\ln(n_1/n_0)$, which is independent of the predictor variables (Hosmer and Lemeshow 1989, p. 19). This is the maximum likelihood estimate of α for a model with no predictor variables included in (2), or with no covariate effects ($\beta_i = 0, \forall_i$). In this way, α can be expressed as $\alpha_0 + \alpha_1$, where $\alpha_0 = \ln(n_1/n_0)$ and α_1 is estimated iteratively according to the values of the predictor variables. Therefore, (2) may also be written as

$$y = \ln\left(\frac{n_1}{n_0}\right) + \alpha_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{4}$$

where $\ln(n_1/n_0)$ is constant throughout the iterative estimation procedure, and α_1 and β_i are the parameters to be iteratively estimated. When logistic regression is performed on samples with different presence/absence ratios, the difference between their α terms is mainly due to α_0 (provided the β_i parameters are equal between the models, i.e., assuming identical covariate effects). The term $\ln(n_1/n_0)$ links (4) to the analysed sample, so the probabilities obtained are only applicable to the same sample from which the inductive inference was made. A favourability model should not be conditioned by the presence/absence ratio in the sample, so α_0 should be eliminated a posteriori, i.e., after the regression procedure, and the new y term should be

$$y' = \alpha_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = y - \ln\left(\frac{n_1}{n_0}\right) \tag{5}$$

which would be the *logit* of the favourability model.

As can be deduced from (1), (3) and (5), favourability values can be obtained directly from the probability values produced by logistic regression, using the formula

$$F = \frac{e^{y'}}{1 + e^{y'}} = \frac{e^{\left(\ln\frac{P}{1-P} - \ln\frac{n_1}{n_0}\right)}}{1 + e^{\left(\ln\frac{P}{1-P} - \ln\frac{n_1}{n_0}\right)}} = 1 - \frac{1}{1 + e^{\left(\ln\frac{P}{1-P} - \ln\frac{n_1}{n_0}\right)}} \tag{6}$$

or, in another form,

$$F = \frac{\frac{P}{(1-P)}}{\frac{n_1}{n_0} + \frac{P}{(1-P)}} = \frac{e^y}{\frac{n_1}{n_0} + e^y} \tag{7}$$

In this way, when the number of presences equals the number of absences, $F{=}P$ (cf. Eqs. (1) and (7)); if the number of presences is lower than that of absences, $F{>}P$; if, on the contrary, there are more presences than absences, $F{<}P$. The environmental threshold $F{=}0.5$ corresponds to the species' prevalence averaged over the entire sample, i.e., the amount "expected" under equal favourability everywhere. This procedure is equivalent to assigning the value $F = 0.5$ to the environmental conditions with which P is equal to the proportion of presences in the studied sample. In this way, the output value of 0.5 will always correspond to the same environmental threshold, whatever the proportion of presences in the sample.

3. Practical applications

3.1. Tests with virtual species

Our third approach is a demonstration based on the modelling of three virtual species' distributions. We made up a territory, which we divided into 200 equal-area squares, an imaginary independent variable *x*, and a species that shows a logistic response to that variable, implying that it is present in 50% of the squares (species 50). We then divided the study area into two complementary samples, one comprising 80 presences and 20 absences and the other comprising the remaining 20 presences and 80 absences of this species. We designed the samples so that they included nearly the same range of *x* values found in the whole territory. In Fig. 1 we show the imaginary territory, its division into two samples, and the distribution of species 50. We performed logistic regressions of this species' presence/absence data on *x* in the whole territory and in each of the two samples separately, and compared the results of applying these three models to the whole territory (Fig. 2). When compared with the model obtained from the whole dataset (Fig. 2c), we see the impact of relative prevalence on the estimated probability (Fig. 2a, b). However, when we applied the favourability function, the three models yielded practically the same results, which in this case coincided with those shown in Fig. 2c.

In an analogous way, we made up two other species with a similar trend to be more frequent towards higher values of *x*, but one occurring in only 20% of the squares (species 20) and the other covering 80% (species 80) (Fig. 3). The presence probability values obtained directly from logistic regression of their distributions on *x* were quite different, due to their different prevalence in the territory. Nonetheless, the *F* values obtained from their favourability functions were virtually identical, so reflecting their common trend to be more frequent than expected at random in the same areas of the territory (Fig. 3).

3.2. Application to a real species

Finally, we illustrate the consequences of this procedure with a practical example using a distribution model for a real species in Spain. We used *Galemys pyrenaicus* presence/absence data on UTM 10 × 10-km squares, obtained from Palomo and Gisbert (2002), to build a multiple logistic regression model using a set of geographical,

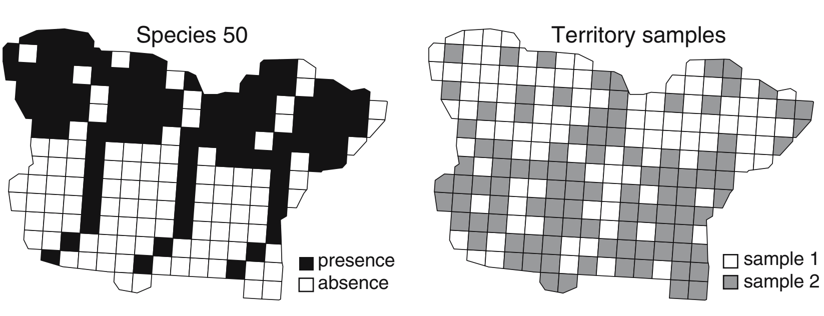
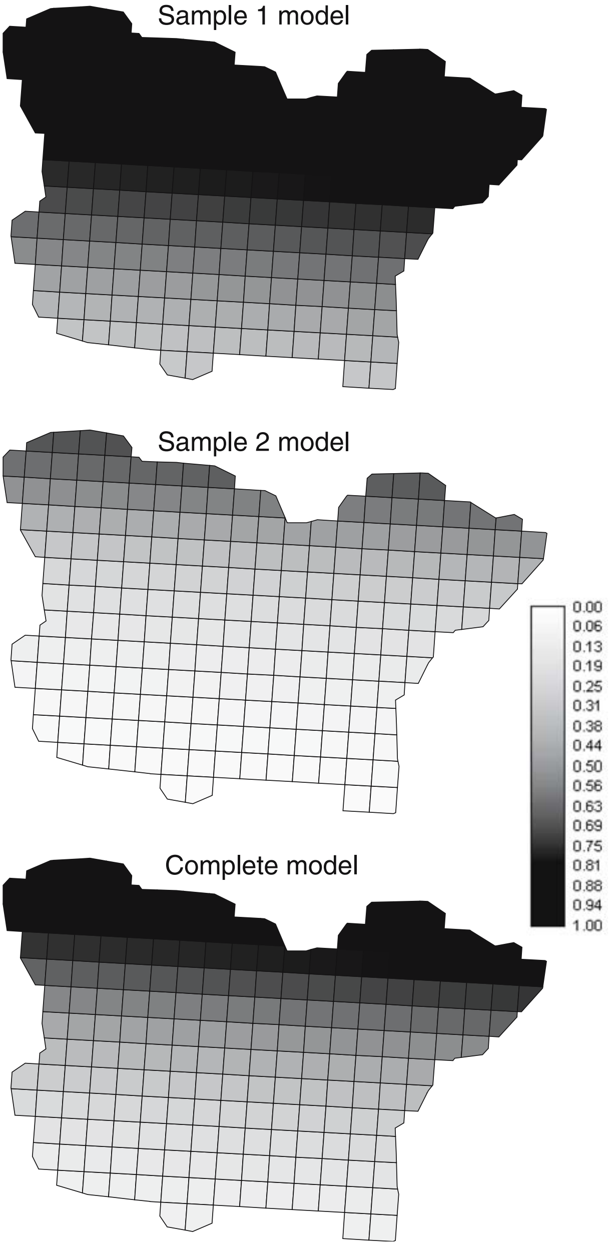


Fig. 1 Imaginary territory divided into equal-area squares, distribution of imaginary species 50, and distribution of the two uneven samples. The values of imaginary variable *x* increase from left to right and from bottom to top

Fig. 2 Potential distribution of species 50 as given by logistic regression of presence/absence data on imaginary variable *x*, using data from each of the two samples and from the whole territory. Using the favourability function, the three datasets produced results indistinguishable from the ones on the third map



environmental and human variables. The variables and the methodology used are similar to those described by Barbosa et al. (2003) for the modelling of otter (*Lutra lutra*) distribution in the same territory. Using both the presence probabilities given by the logistic regression model, and the values resulting from our favourability function, we built potential distribution maps taking 0.5 as the threshold above which presence is more likely than absence, and high favourability maps using 0.8 as the threshold

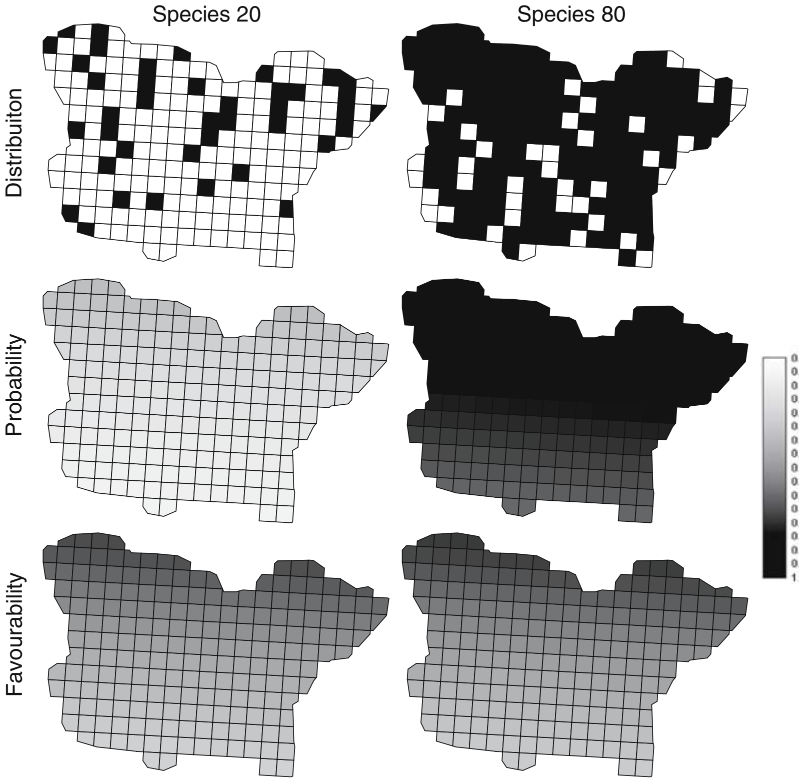


Fig. 3 Distribution of imaginary species 20 and 80, respectively; presence probability values given by the function of logistic regression on *x*; and favourability values given by our favourability function

above which to assume that the environment is highly favourable, since the odds are higher than 4:1. The distribution of *Galemys pyrenaicus* in Spain and the potential distribution and high favourability maps produced by the probability and favourability models, respectively, are shown in Fig. 4. The species' presence/absence ratio in the studied area is 463/4,704 ($\approx$9% presences) and, consequently, the logistic regression-based predicted probabilities are visibly dominated by the low baseline prevalence of *Galemys pyrenaicus*. The favourability model, on the other hand, more directly identifies spatial variations in the presence/absence ratio.

4. Discussion

Spatial modelling is a typical tool for assessing ecological responses of species to environmental conditions and predicting the evolution of their geographical distributions. The values of *F* do not reflect presence probability, but rather environmental favourability values, which are what distribution modellers are usually going after. Logistic regression models probabilities, but favourability is not simply the probability of presence but rather a description of local deviations from the overall probability of presence, that is to say, from prevalence.

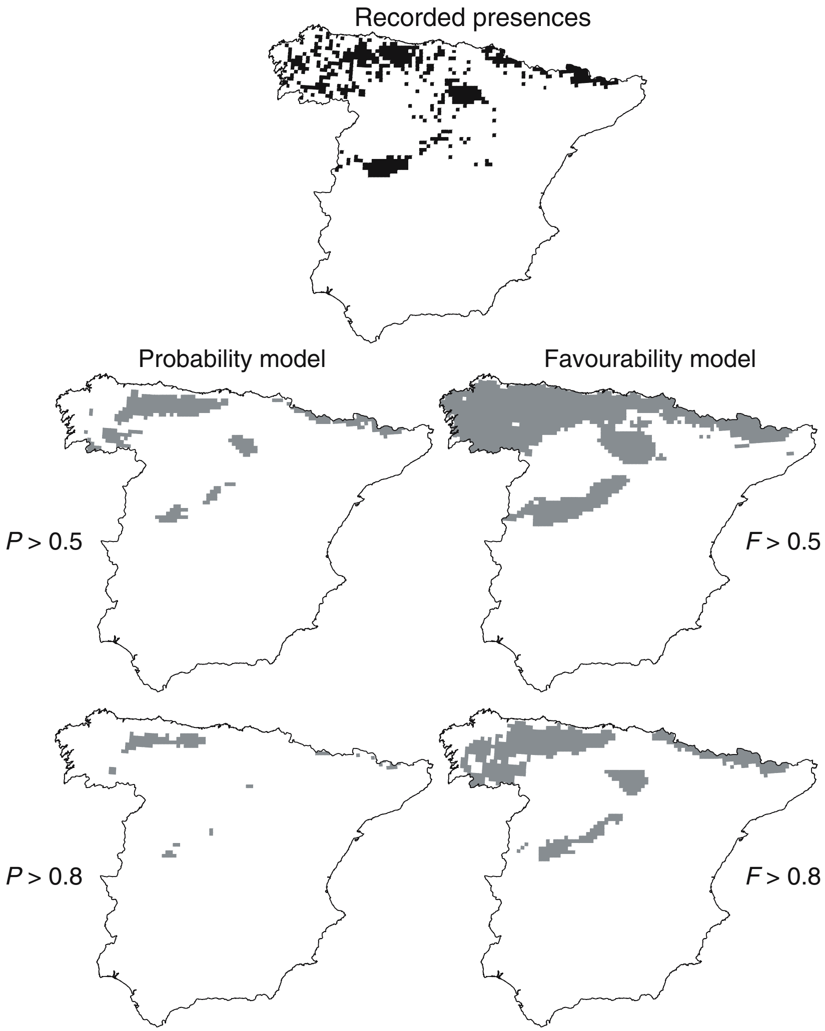


Fig. 4 Recorded presences of *Galemys pyrenaicus* in Spain on 10 × 10-km UTM squares (Palomo and Gisbert 2002), and predicted presences according to the probability (P) function derived directly from logistic regression, and to the favourability (F) function, considering thresholds of 0.5 and 0.8

As can be seen in (6), F is also a logistic function, so it assumes values ranging from 0 to 1. Thus, the F value may be considered as the degree of membership of the fuzzy set of areas favourable for the species, so that it may be used to apply the concepts, operations and rules of fuzzy logic to environmental modelling: for example, $1 - F$ corresponds to the degree of membership of the complementary fuzzy set of sites whose environmental conditions are unfavourable to the species; intersection can be used to identify areas simultaneously favourable to several species; and union can identify areas favourable to either of several species. These values also allow for directly comparing the degree of favourability for, for instance, a rare predator and a more common prey, which is more difficult to achieve using the original logistic functions, as the different proportions of presences for the two species bias their random expectations in opposite directions. A region may be equally favourable for both species, even if one of them is much less frequent due to its biology or behaviour. Besides, it makes little sense to obtain a probability of occurrence for a species in an area where it has in effect been recorded, whereas it is perfectly logical to label even as unfavourable an area where the species does occur. The favourability model is useful to elucidate biogeographical trends, as well as for practical purposes such as the design of ecological corridors, the assessment of the conservation value of territories, or the selection of the most suitable locations for species reintroductions.

Acknowledgements We thank Ramón Hidalgo (SCI-University of Málaga) and an anonymous editor for useful comments that helped improve this manuscript. A.M.B.'s research is supported by a doctoral grant (SFRH/BD/4601/2001) from the Fundação para a Ciência e a Tecnologia, Portugal. This work was partially financed by project REN2003-02471/GLO (Ministry for Science and Technology, Spain).

References

Barbosa AM, Real R, Olivero J, Vargas JM (2003). Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. Biol Conserv 114:377–387.

Brito JC, Crespo EG, Paulo OS (1999). Modelling wildlife distributions: logistic multiple regression vs overlap analysis. Ecography 22:251–260.

Bustamante J (1997). Predictive models for lesser kestrel Falco naumanni distribution, abundance and extinction in southern Spain. Biol Conserv 80:153–160.

Franco AMA, Brito JC, Almeida J (2000). Modelling habitat selection of common cranes *Grus grus* wintering in Portugal using multiple logistic regression. Ibis 142:351–358.

Hosmer DW, Lemeshow S (1989). Applied logistic regression. John Wiley and Sons, Inc., New York, p 19 and 147.

Madsen AB, Prang A (2001). Habitat factors and the presence or absence of otters *Lutra lutra* in Denmark. Acta Theriologica 46:171–179.

Palomo LJ, Gisbert J (2002). Atlas de los mamíferos terrestres de España. Dirección General de Conservación de la Naturaleza-SECEM-SECEMU, Madrid.

Rojas AB, Cotilla I, Real R, Palomo LJ (2001). Determinación de las áreas probables de distribución de los mamíferos terrestres en la provincia de Málaga a partir de las presencias conocidas. Galemys 13(NE):217–229.

Romero J, Real R (1996). Macroenvironmental factors as ultimate determinants of the distribution of common toad and natterjack toad in the south of Spain. Ecography 19:305–312.

Seoane J, Bustamante J (2001). Modelos predictivos de la distribución de especies: una revisión de sus limitaciones. Ecología 15:9–21.

Tabachnick BG, Fidell LS (1996). Using multivariate analysis, 3rd edn. HarperCollins College Publishers, Northridge, California.

Teixeira J, Ferrand N, Arntzen JW (2001). Biogeography of the golden-striped salamander, *Chioglossa lusitanica*: a field survey and spatial modelling approach. Ecography 24:618–623.

Biographical sketches

Raimundo Real and J. Mario Vargas are professors, and A. Márcia Barbosa is a Portuguese PhD student at the Department of Animal Biology of the University of Málaga (Spain). Their main research interests include the study and spatial modelling of species distribution and diversity, and their applications for conservation. They apply ecological statistics to the study of various taxonomic groups, with an emphasis on terrestrial vertebrates.