A comparison of three statistical methods for analysing extinction threat status

HEATHER R. TAFT[1], DEREK A. ROFF[1], ATTE KOMONEN[2] AND JANNE S. KOTIAHO[2]

[1]*Department of Biology, University of California, Riverside, California 92521, USA, and* [2]*Department of Biological and Environmental Science, University of Jyväskylä, PO Box 35, FI-40014, Finland*

SUMMARY

The International Union for Conservation of Nature (IUCN) Red List provides a globally-recognized evaluation of the conservation status of species, with the aim of catalysing appropriate conservation action. However, in some parts of the world, species data may be lacking or insufficient to predict risk status. If species with shared ecological or life history characteristics also tend to share their risk of extinction, then ecological or life history characteristics may be used to predict which species may be at risk, although perhaps not yet classified as such by the IUCN. Statistical models may be a means to determine whether there are non-threatened or unclassified species that share the characteristics of threatened species, however there are no data on which model might be most appropriate or whether multiple models should be used. In this paper, three types of statistical models, namely regression trees, logistic regression and discriminant function analysis are compared using data on the ecological characteristics of Finnish lepidopterans (butterflies and moths). Overall, logistic regression performed slightly better than discriminant function analysis in predicting species status, and both outperformed regression trees. Uncertainty in species classification suggests that multiple analyses should be performed and particular attention devoted to those species for which the methods disagree. Such standard statistical methods may be a valuable additional tool in assessing the likely threat status of a species where there is a paucity of abundance data.

Keywords: discriminant function analysis, Geometridae, Lepidoptera, logistic regression, Noctuidae, regression tree analysis, threat status

INTRODUCTION

Assessing the risk of extinction is important to determine which species are most prone to extinction and may be in need of human intervention. The International Union for Conservation of Nature (IUCN) has defined categories assessing species threat status on the basis of their risk of extinction. Classifications largely rely on quantitative information, but in practice expert opinion plays a strong role. Many different methods have been used to assess extinction risk, including population models (McCarthy *et al.* 2004), species–area correlations (Grelle 2005) and genetic analyses (Dunham *et al.* 1999). Obtaining the type of data needed to assess extinction risk can be problematic. For example, the assessment of population size and changes in population size can be difficult, particularly when assessing small populations whose individuals are not easily located. Genetic data, which may provide information about the degree of inbreeding and gene flow among populations, is costly, especially if multiple species are to be assessed, and may also be hard to obtain from small populations. However, data on simple ecological or life history characteristics can be obtained using less extensive population monitoring, or from existing knowledge of the natural history of species. In this paper, we assess the potential of such data to provide an alternative and reliable measure of extinction risk.

Among the studies that use ecological characteristics to assess extinction risk, several different types of analyses have been used, including multiple regression (Purvis *et al.* 2000; Krüger & Radford 2008), regression tree analysis (Boyer 2008; Davidson *et al.* 2009; Boyer 2010), logistic regression (Mattila *et al.* 2006, 2008; Franzen & Johannesson 2007) and risk ranking (Kotiaho *et al.* 2005). Several of these studies used multiple tests to assess extinction risk, often initially analysing single ecological or life history characters as predictor variables, followed by one of the single type of statistical tests mentioned above to analyse the complete data set. Bielby *et al.* (2010) compared decision trees and phylogenetic comparative methods. Here we compare three statistical approaches: regression tree analysis, logistic regression and discriminant function analysis. We selected these three statistical methods primarily because they are commonly found in pre-packaged statistical software programs and thus would be easier for conservation managers to access and use than more complicated programs that require writing and editing code. Our aim was to determine if one or a combination of these statistical methods could be used to predict the threat category, threatened or non-threatened, of as yet unclassified species. Such an analysis may determine whether there are unclassified species that merit immediate attention. Reclassification of a formerly non-threatened species to a threatened status is also a possible outcome, and may indicate that a species is in more immediate need of conservation management measures. The specific situation we envision is one in which most species within a particular taxon have been classified as threatened or

---

*Correspondence: Dr Heather Taft e-mail: heather.r.taft@gmail.com

non-threatened. If it can be shown that this classification is well predicted by one or more of the three statistical approaches using general ecological and life history characters (hereafter, for simplicity, LH characters) then unclassified species of the same taxon that share characteristics with threatened species can be identified to help prioritize further assessments of threat status.

We assess this approach using data on Finnish lepidopterans previously analysed by Komonen *et al.* (2004), Kotiaho *et al.* (2005) and Mattila et al. (2006, 2008). Komonen *et al.* (2004) used analysis of variance and linear regressions on subsets of LH characters to assess butterfly mobility, but did not relate this to IUCN threat status. Mattila *et al.* (2006, 2008) used similar analyses, running logistic regressions of IUCN threat status on single LH characteristics and then using multiple LH characteristics in a multinomial logistic regression to determine the ability to classify species into their correct IUCN threat status. Kotiaho *et al.* (2005) primarily used t-tests and logistic regressions with single LH characteristics to compare differences between threatened and non-threatened species. The four LH characteristics (dispersal ability, larval specificity, adult habitat breadth and length of flight period) that they found to be significantly related to IUCN threat status were used to create an ecological risk ranking of all the species by ranking the species according to each LH characteristic and then summing the four rankings. This summed ranking was used in a logistic regression and compared to the actual IUCN threat status of the species.

In the present analysis, we address the question of whether the LH characteristics of a species may be used directly en masse to predict the probability of unclassified species being threatened or non-threatened, and identify non-threatened species that may need reclassification of their threat status.

METHODS

Regression and classification trees

Roff and Roff (2003) initially suggested regression tree analysis to determine factors contributing to the risk of extinction. Several studies have since used a regression tree analysis to assess extinction probability (Jones *et al.* 2006; Boyer 2008, 2010; Davidson *et al.* 2009). In our case study the predicted category consisted of two states, 'threatened' ($= 1$) and 'non-threatened' ($= 0$), and thus regression and classification trees were identical. In principal, classification trees could be used to identify each different IUCN category. We restricted the analysis to the two states and used the regression tree approach, owing to insufficient data.

Logistic regression

Our aim was to determine the accuracy of models created using logistic regression at predicting the correct assignment of a species as threatened or non-threatened based on LH characteristics. As done with the regression trees, species in the non-threatened category were coded as 0 and those

in the threatened category as 1. To use the model as a mechanism for placement of a species into the threatened or non-threatened category, we required a threshold for the fitted value, for example 0.5, above which species were placed into the threated category and below which they were placed into the non-threatened category. Alternatively, we could have selected two thresholds, such as 0.25 and 0.75, and placed species below the lower threshold in the non-threatened category and species above the upper threshold in the threatened category, classifying species lying between the two thresholds as 'uncertain.' We explored the consequences of both approaches.

An important point to note in this method of analysis is that, whereas the stopping point for the stepwise regression is defined by a metric such as the Akaike information criterion (AIC), the adequacy of the model in the present context was measured by the assignment to the two categories: because of this, it was possible for the 'best' model to contain more or fewer variables than that specified by the 'best' stepwise logistic model.

Discriminant function analysis

There are several covariance structures that can be identified when performing a discriminant function analysis: homoscedastic, spherical, proportional, group spherical, equal correlation and heteroscedastic. Principal components can also be specified for analysis in a discriminant function analysis. In the analyses here, to compare statistical tests, we used the covariance structure that correctly assigned the most threatened species based on LH characteristics.

Determining the probability of correct assignment by chance alone

An important consideration is the probability of assigning a species to the correct category, threatened or non-threatened, by chance alone. To determine this we used a simulation model (coding in R given in Appendix 1, see supplementary material at Journals.cambridge.org/ENC). First, we generated a vector, **V**, of length N, where N was the total number of species in the sample with ones in the first n_1 rows and zeros in the remaining $N - n_1$ rows, the former being the number of threatened species and the latter the number non-threatened species. These zeros and ones were rearranged at random in the vector. The number of correct assignments in the threatened category, N_1, was given by

$$N_1 = \sum_{i=1}^{n_1} V_i,$$

that is, the number of ones in the first n_1 rows, and the number of correct assignments in the non-threatened category, N_0 was

$$N_0 = N - n_1 - \sum_{i=n_1+1}^{N} V_i,$$

namely the number of zeros in the remaining $N - n_1$ rows. The total number of correct assignments, N_T, was thus $N_T = N_1 + N_0$. These three numbers were stored and the process repeated to generate three new samples. The whole process was replicated 10 000 times, generating a matrix with three columns (total correctly assigned, correctly assigned to threatened, correctly assigned to non-threatened) and 10 000 rows. From this matrix we calculated the distribution of correct assignments. For each column we determined the probability of correctly assigning n_j species ($j = 0$ to N, $j = 0$ to n_1, $j = 0$ to $N - n_1$) as $S_J/10000$, where S_J was the number of times n_j appeared in the relevant column.

Determining the preferred method

As indicated below, regression tree analysis was not satisfactory for either of the two data sets (butterflies or moths) and, therefore, our further analysis focused upon logistic regression versus discrimination function analysis. We compared the ability of these two methods to correctly classify species into the threatened or non-threatened category with a χ^2 analysis. Of particular interest were those species which were incorrectly classified according to one or both methods: we plotted the predicted values from the logistic regression against the predicted values from the discriminant function analysis to see whether the species that were classified differently fell near the 0.5 cutoff. It is safer to classify non-threatened species as threatened than it is to classify threatened species as non-threatened, because in the former case a species will receive attention, but in the latter case a threatened species that needs attention will be overlooked. Therefore, we used the number of correctly classified threatened species in a final comparison of methods to determine which method was to be preferred, at least for the data sets assessed here.

Data sets

The Kotiaho *et al.* (2005) butterfly data set consisted of 94 species and 13 predictor variables: family, genus, species, abundance, distribution, distribution change, resource distribution, extent of range, larval specificity, female size, length of flight period, mobility and habitat breadth (see Komonen *et al.* 2004 and Kotiaho *et al.* 2005 for variable definitions). Because the primary criterion for listing these species according to IUCN threat status is a function of three 'distribution' variables (distribution, distribution change, and extent of range), we included these variables as predictor variables to assess whether any of the other variables were better at predicting IUCN threat status. After initial analysis, these distribution variables were excluded from subsequent analyses to determine whether any more easily accessible variables (such as those obtainable from published natural history descriptions) could be used to predict threat status. One species, *Glaucopsyche alexis*, was listed using only abundance as the criterion, and so this species was not used

in the analyses. Thirteen other species were excluded due to missing data. The analysed data consisted of 18 threatened and 62 non-threatened species. The rest of the variables, excluding resource distribution due to lack of data, were used to predict IUCN threat status as given in the 2000 Finnish Red List (Rassi *et al.* 2001). One of the principal aims of the analysis was to investigate the ability of variables that are readily available from published data on the natural history of a species to determine threat status. Therefore, we ran the analyses with and without the variable 'abundance,' which might typically be difficult to estimate and, in many cases, unavailable. However, due to the similarity of the results, only the analyses excluding abundance are reported here (Appendix 2 provides a table of the analyses including abundance, see supplementary material at Journals.cambridge.org/ENC).

Two data sets on moths were used, one on noctuids (Mattila *et al.* 2006) and one on geometrids (Mattila *et al.* 2008). These two data sets consisted of 284 and 306 species, respectively, and each had the same eight predictor variables: genus, species, male size, length of flight period, larval specificity, resource distribution, overwintering stage, and distribution change. After we ran the analyses with each data set, we combined them into a single data set with the added variable 'taxonomic family' to increase the power of the analyses and avoid issues of non-independence by using closely related species. In total, we excluded 40 species from the data sets due to missing data. The analysed data consisted of 68 threatened and 482 non-threatened species. Distribution change was the only distribution variable for these data sets and, after initial analysis, was, as before, excluded to determine which other variables may be important for predicting IUCN threat status. As previously noted, the response variable for all data sets was binomial, threatened or non-threatened. It included all species listed as near threatened or higher according to the IUCN threat status as threatened and the rest as non-threatened.

RESULTS

Among the three distribution variables used in the butterfly data set (distribution, range position, and distribution change), only distribution and range position were highly correlated ($r = 0.64$; Table 1). Correlations between any of the distribution variables and the other variables was highest for distribution and mobility ($r = 0.74$) and distribution and length of flight ($r = 0.64$), although these values were not high enough to cause problems with collinearity since they were < 0.90 (Tabachnick & Fidell 2007, p. 89–90). None of the correlations among variables for the geometrid or noctuid data sets exceeded an absolute value of 0.26 and thus did not pose problems with collinearity.

We first examined the ability of the distribution variables by themselves to classify species status. After this we considered the ability of the other LH characteristics to classify species status.

Table 1 Correlations among the continuous variables used in the analyses of the butterfly data (sample size = 18 threatened species and 62 non-threatened species). $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; probabilities not corrected for multiple test.

Characteristic	Abundance	Distribution	Range position	Mobility	Length of flight	Female size
Distribution	–0.42**					
Range position	–0.48***	0.64***				
Mobility	–0.55***	0.74***	0.42***			
Length of flight	–0.24*	0.64***	0.29*	0.57***		
Female size	–0.10	0.07	–0.11	0.31**	0.19	
Distribution change	–0.33	0.32**	0.17	0.35***	0.21*	–0.10

Regression trees

Distribution variables only

A significant regression tree, with two nodes was obtained ($p = 0.0002$) using only the distribution variables from the butterfly data set. This split in the regression tree was based on distribution and correctly classified 94% of threatened species and 95% of non-threatened species. When we used the variable distribution change in a regression tree analysis on the geometrid data, the pruned regression tree was significant ($p = 0.0006$) and had two nodes, but all threatened species were misclassified as non-threatened. The threatened species did not have any defining range of distribution change to use to divide the data. When we used only distribution change in a regression tree analysis on the noctuid data set, the pruned regression tree was significant ($p = 0.0002$) and had five nodes with a misclassification rate of 11%. Using distribution change in a regression tree analysis on the combined moth data set, we found that the pruned regression tree was significant ($p = 0.0002$) and had four nodes, but all threatened species were misclassified, again indicating no defined range for the threatened species.

Other LH variables

Using variables from the butterfly data set not explicitly used for determining IUCN threat status (family, mobility, larval specificity, habitat breadth, female size, and flight length), a pruned tree could not be created because only one terminal node was produced during the cross-validation. Thus, in this case, regression tree analysis could not discriminate threatened from non-threatened species.

When all the variables (male size, length of flight period, larval specificity, and overwintering stage) except distribution change were used from the geometrid data set, a non-significant pruned tree with two nodes resulted ($p = 0.2826$), but no threatened species were correctly classified because both nodes were classified as non-threatened in the regression tree. Thus, breaking the data down into these two nodes based on length of flight period did not allow for enough subdivision of the data to correctly assign IUCN threat status. When all the variables from the noctuid data set or the combined data set (male size, length of flight period, larval specificity, and overwintering stage) except distribution change were used, a pruned tree could not be created because only one terminal node was produced during the cross-validation.

We conclude that for these data sets regression tree analysis did not result in a satisfactory prediction of threat status.

Logistic regression

Distribution variables only

Species with probabilities greater than 0.5 were classified as threatened and those below as non-threatened. A stepwise logistic regression using the distribution variables from the butterfly data set retained the variables distribution and distribution change and was able to correctly classify 94% of threatened species and 95% of non-threatened species. The logistic regression on distribution change in the geometrid data set correctly classified 7% of threatened species and 99% of non-threatened species. Using the noctuid data set or the combined data set, the logistic regression was unable to correctly classify any threatened species.

Other LH variables

We restricted the logistic regression analysis to additive models only (interactions were excluded) because the butterfly data set was limited. Incorporating all the variables except the distribution variables and abundance produced a model that correctly assigned 67% of threatened and 95% of non-threatened species when the fitted value cutoff point was 0.50 (Table 2). This cutoff point (0.5) for the fitted values produced the highest correct assignment for both threatened and non-threatened species. The alternate criterion of 0.25 as the cutoff point for the non-threatened species and anything greater than 0.75 as a cutoff point for the threatened species correctly assigned 39% of the threatened species and 79% of the non-threatened species, with 21 ambiguous species. Moth data produced a similar response to the cutoff values (results not shown), so we opted to use the 0.5 cutoff for all reported classification assignments for the logistic regressions.

The stepwise logistic regression on the geometrid data set using all variables except distribution change included interactions due to a larger sample size and retained all variables and their interaction terms except the four-way interaction. This regression correctly assigned 47% of the threatened species (and was able to correctly classify all but one of the non-threatened species ($F_{47,263} = 2.84$, $p < 0.001$; AIC = 168.11; Table 3). The stepwise logistic regression using all the variables from the noctuid data set except distribution change retained all the variables and their interaction terms, did poorly at correctly assigning threatened species, but well at

Table 2 Classification of threatened and non-threatened species by logistic regression and discriminant function analysis for the butterfly data, when the distribution variables and abundance are excluded from the analysis. Regression tree analysis was excluded because a pruned tree could not be created. When multiple analyses were performed, such as when different structures were used for discriminant analysis, only the analysis with the best result is given. [‡]Probability of correctly predicting by chance alone at least as many as observed by a given method.

Analysis	Threatened		Non-threatened	
	Predicted	$p^{\ddagger}$	Predicted	p
Logistic regression, 0.5 cutoff	12/18	<0.0001	59/62	<0.0001
Stepwise logistic regression	10/18	0.0005	57/62	<0.0001
Discriminant function analysis – homoscedastic	10/18	0.0005	56/62	<0.0001

assigning non-threatened species (21% and 96% respectively; $F_{61,244} < 0.001$, $p = 1$; AIC = 2887.41; Table 3). The stepwise logistic regression on all variables from the combined data set except distribution change retained all the variables and their four-way interaction terms except the interaction involving family, male size, larval specificity and overwintering stage, correctly assigned one-third of threatened species (29%), and was able to correctly assign almost all the non-threatened species (96%; $F_{112,477} < 0.001$, $p = 1$; AIC = 4687.5; Table 3).

Discriminant function analysis

Distribution variables only
Discriminant function analysis produced results that were not quite as good as the logistic regression at classifying threatened and non-threatened species (88% and 89%, respectively) using only the distribution variables from the butterfly data set. For the geometrid, noctuid and combined data sets 90%, 100% and 96% of threatened species, respectively, were correctly classified, and 52%, 46% and 53% of non-threatened species were correctly classified using just distribution change.

Other LH variables
The discriminant function analysis on the butterfly data set including all variables except the distribution variables and abundance was significant (Table 2), correctly classifying 56% of threatened species and 90% of non-threatened species. Weighting was highest for family and habitat breadth when classifying threatened species.

The discriminant function analysis for the geometrid data set including all the variables except distribution change performed best using the principal components model, and correctly classified 27% of the threatened species and 95% of the non-threatened species (Table 3). The best discriminant analysis including all the variables from the noctuid data set except distribution change used the equal correlation model and correctly assigned 88% of non-threatened species and 32% of threatened species. The best discriminant analysis on the combined data set including all variables except distribution change used the equal correlation model also, and assigned 94% of non-threatened species, but only 19% of threatened species.

Which method is best?

Logistic regression and discriminant function analysis were able to correctly classify a significant number of threatened and non-threatened species for most analyses. Using the butterfly data set, the logistic regression and the discriminant function analysis did not differ in the number of correctly classified threatened and non-threatened species ($\chi^2 = 0.078$, df = 1, $p = 0.7804$). However, using the combined moth data, the logistic regression and the discriminant function analysis did differ significantly in the number of correctly classified threatened and non-threatened species ($\chi^2 = 6.416$, df = 1, $p = 0.0113$).

Logistic regression and discriminant function analysis agreed on the classification (Table 4), rightly or wrongly, of all but two of the threatened butterfly species, with the logistic regression classifying the species 'correctly' according to the published IUCN red list. Interestingly, the incorrectly classified butterfly species did not cluster about the intersection of the 0.5 cutoff (vertical and horizontal lines in Fig. 1) demarking the transition from threatened to non-threatened species for each analysis (the moth data could not be plotted this way because the fitted values for the logistic regression were all 0s and 1s). The two species that were correctly classified by the logistic regression, but not the discriminant function analysis, were not notable outliers. In one case, the species lies close to the 'decision' boundary for the discriminant function analysis, while in the other case, the species lies close to the 'decision' boundary for the logistic regression. For such species the discrepancy in the analyses invites closer inspection. While the logistic regression correctly classified more species than the discriminant function analysis, the difference overall is relatively minor, and we recommend that both methods be used, with additional attention being paid to those species classified differently.

DISCUSSION

All of the analyses on the butterfly data and most of the analyses on the moth data sets were able to correctly assign a significant number of threatened and non-threatened species. Regression tree analysis was not very helpful in classifying species in the

Table 3 Classification of threatened and non-threatened species by regression tree analysis, logistic regression and discriminant function analysis for the moth data when distribution change was excluded from the analysis. Data for the regression tree analysis was excluded for the noctuid and combined data sets because a pruned tree could not be created.

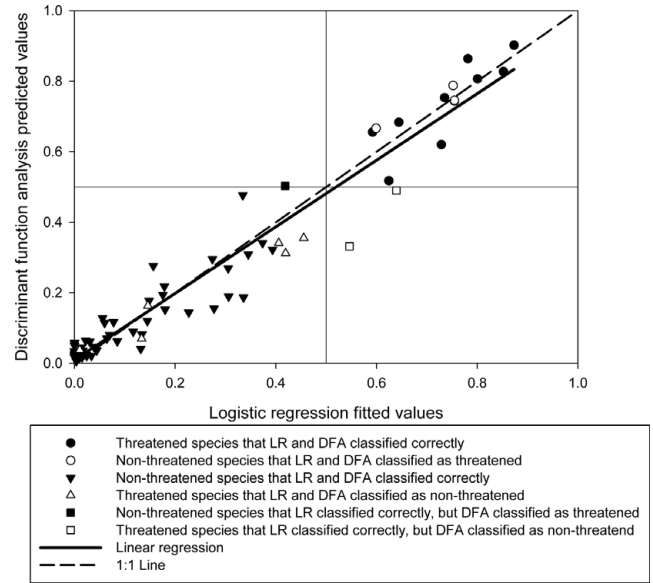| | Geometrids | | | | Noctuids | | | | All Moth Data | | | |
| | Threatened | | Non-threatened | | Threatened | | Non-threatened | | Threatened | | Non-threatened | |
Analysis	Predicted	p	Predicted	p	Predicted	p	Predicted	p	Predicted	p	Predicted	p
Regression tree	0/30	1	238/238	<0.0001								
Logistic regression, 0.5 cutoff	14/30	<0.0001	237/238	<0.0001	8/38	0.1166	234/244	<0.0001	27/68	<0.0001	404/482	1
Stepwise logistic regression	14/30	<0.0001	237/238	<0.0001	8/38	0.1166	234/244	<0.0001	20/68	0.0001	465/482	<0.0001
Discriminant function	8/30	0.0087	226/238	<0.0001	12/38	0.0013	214/244	0.1166	13/68	0.0566	455/482	<0.0001

Figure 1 Plot of the logistic regression (LR) fitted values versus the discriminant function analysis (DFA) predicted values for the butterfly data, indicating the two methods have approximately a 1:1 relationship in their prediction of threat status.

two data sets, although previous analyses have suggested it may be a promising tool (Roff & Roff 2003; Jones *et al.* 2006; Boyer 2008, 2010; Davidson *et al.* 2009). Importantly, the approach here did not fail because it incorrectly classified species. For the present data sets, it failed because no trees could be produced. Thus, we recommend that this approach still be tried for other data sets.

There was a decrease in the percentage of correctly classified species when a stepwise logistic regression was used on the butterfly data set and on several moth regressions. As noted above, this is not unexpected, as the criterion for the best fitting model is not the same as the criterion for the stopping point in the stepwise regression. Thus, for logistic regression analysis, a second step, namely a comparison of the models predicting correct assignment is called for, which can be done using simulation (Appendix 1, see supplementary material at Journals.cambridge.org/ENC). Since we were primarily focused on the category the species fell within, and not necessarily how good individual variables were at classifying the species, a case could be made for leaving all the variables in the model, because including variables will not increase the misclassification rate. Comparing the full model and the stepwise model may then be a useful way to identify the ecological characteristics that have the strongest correlation with IUCN threat status, and which variables increase the number of correctly assigned species, even if the difference is not significant.

The overall correct classification rate can be highly misleading. For example, suppose that 90% of species were classified as non-threatened and the statistical analysis classified all species as non-threatened, then the overall correct

Table 4 Comparison of classifications by logistic regression and discriminant function analysis for threatened butterfly and moth species.

Logistic regression classification	Discriminant function analysis classification	Number of threatened butterfly species classified as indicated	Number of threatened moth species classified as indicated
Threatened	Threatened	10	8
Non-threatened	Threatened	0	5
Threatened	Non-threatened	2	19
Non-threatened	Non-threatened	6	36

classification rate would be 90%, which appears to be very good but is actually of little use. This issue was particularly evident in the moth data sets, and illustrates the importance of analysing the ability of the statistical analysis to classify species into each category, as done here. When methods of comparison such as these are used, classification to both categories should always be reported.

Overall, the logistic regression gave the best results based on the number of correctly predicted threatened species, followed by the discriminant function analysis. All variables can be used in these analyses, so deciding among variables is not an issue. Given the ease with which the analyses can be performed, we suggest that multiple analyses be undertaken to identify species that may not be consistently classified as threatened.

Suggested reassessments based on the current analysis

The butterfly species *Boloria frigga*, *B. freija*, and *B. thore*, were classified as threatened by logistic regression and discriminant function analysis, although the IUCN does not list them as threatened. These species may be at an increased risk of extinction. Kotiaho *et al.* (2005) assigned these species ecological risk rankings of 4, 11, and 15, respectively, and our results support a reassessment of their IUCN threat status. Discriminant function analysis classified *Pyrgus centaureae* as threatened; Kotiaho *et al.* (2005) assigned an ecological risk ranking of 8 to *P. centaureae*, and thus its threat status may also merit reassessment, although the first three species would be a priority.

For the moth data sets, five non-threatened species were classified as threatened by both logistic regression and discriminant function analysis. All the misclassified moth species fall in the family Noctuidae. *Cucullia gnaphalii*, *Dryobotodes eremita* and *Orthosia populeti* were classified as threatened when the noctuid data set was used. *D. eremita*, *Abrostola triplasia* and *A. tripartita* were classified as threatened when the combined data set was used. We suggest that the threat status of all these species merits re-evaluation; an IUCN threat listing would attract appropriate conservation management, thus enhancing these species' chance for recovery.

We have shown that a variety of statistical analyses can produce useful assessments of threat status and that readily-available data on ecological and LH characteristics may be used to identify species that merit reassessment of their threat status. Extending the analyses beyond the geographical area of interest should be undertaken cautiously, as important variables may change in different locations and at different scales (Nylin & Bergstrom 2009). In particular, if abundance is to be used as a variable, understanding its relationship to the species being assessed is essential to determine whether a positive or negative relationship to distribution can be generalized (see the following articles for a debate on positive and negative density-distribution relationships: Paivinen *et al.* 2005; Blackburn *et al.* 2006; Blackburn & Gaston 2009; Komonen *et al.* 2009, 2011; Kotiaho *et al.* 2009; Selonen & Helos 2010). These caveats notwithstanding, the present results suggest that 'off-the-shelf' statistical methods such as logistic regression and discriminant function analysis can be extremely valuable in determining the IUCN threat status of a species in areas where there is only limited abundance data.

CONCLUSIONS

Standard statistical analyses may be applied to ecological and life history characteristic data to produce an assessment of threat status where there is limited abundance data. Applying these methods to Finnish butterfly and moth species datasets produced results that were highly consistent with present IUCN threat listings, and identified a few additional candidates that probably deserve increased attention and monitoring. Identifying species that merit reassessment is crucial to allocating conservation management efforts appropriately.

References

Bielby, J., Cardillo, M., Cooper, N. & Purvis, A. (2010) Modelling extinction risk in multispecies data sets: phylogenetically independent contrasts versus decision trees. *Biodiversity and Conservation* **19**(1): 113–127.

Blackburn, T.M. & Gaston, K.J. (2009) Sometimes the obvious answer is the right one: a response to 'Missing the rarest: is the positive interspecific abundance–distribution relationship a truly general macroecological pattern?' *Biology Letters* **5**(6): 777–778.

Blackburn, T.M., Cassey, P. & Gaston, K.J. (2006) Variations on a theme: sources of heterogeneity in the form of the interspecific relationship between abundance and distribution. *Journal of Animal Ecology* **75**(6): 1426–1439.

Boyer, A.G. (2008) Extinction patterns in the avifauna of the Hawaiian islands. *Diversity and Distributions* **14**(3): 509–517.

Boyer, A.G. (2010) Consistent ecological selectivity through time in pacific island avian extinctions. *Conservation Biology* **24**(2): 511–519.

Davidson, A.D., Hamilton, M.J., Boyer, A.G., Brown, J.H. & Ceballos, G. (2009) Multiple ecological pathways to extinction in mammals. *Proceedings of the National Academy of Sciences USA* **106**(26): 10702–10705.

Dunham, J., Peacock, M., Tracy, C.R., Nielsen, J. & Vinyard, G. (1999) Assessing extinction risk: integrating genetic information. *Conservation Ecology* **3**(1): 2 [www document]. URL http://www.consecol.org/vol3/iss1/art2/

Franzen, M. & Johannesson, M. (2007) Predicting extinction risk of butterflies and moths (Macrolepidoptera) from distribution patterns and species characteristics. *Journal of Insect Conservation* **11**(4): 367–390.

Grelle, C.E.V. (2005) Predicting extinction of mammals in the Brazilian Amazon. *Oryx* **39**(3): 347–350.

Jones, M.J., Fielding, A. & Sullivan, M. (2006) Analysing extinction risk in parrots using decision trees. *Biodiversity and Conservation* **15**(6): 1993–2007.

Komonen, A., Grapputo, A., Kaitala, V., Kotiaho, J.S. & Paivinen, J. (2004) The role of niche breadth, resource availability and range position on the life history of butterflies. *Oikos* **105**(1): 41–54.

Komonen, A., Paivinen, J. & Kotiaho, J.S. (2009) Missing the rarest: is the positive interspecific abundance-distribution relationship a truly general macroecological pattern? *Biology Letters* **5**(4): 492–494.

Komonen, A., Paivinen, J. & Kotiaho, J.S. (2011) Varying definitions of abundance and incomplete assemblages challenge the generality of the interspecific abundance-distribution relationships. *Annales Zoologici Fennici* **48**(3): 161–166.

Kotiaho, J.S., Kaitala, V., Komonen, A. & Paivinen, J. (2005) Predicting the risk of extinction from shared ecological characteristics. *Proceedings of the National Academy of Sciences USA* **102**(6): 1963–1967.

Kotiaho, J.S., Komonen, A. & Paivinen, J. (2009) On the obvious positive interspecific relationship between abundance and

distribution: a reply to Blackburn and Gaston. *Biology Letters* **5**(6): 779–780.

Krüger, O. & Radford, A.N. (2008) Doomed to die? Predicting extinction risk in the true hawks Accipitridae. *Animal Conservation* **11**(2): 83–91.

Mattila, N., Kaitala, V., Komonen, A., Kotiaho, J.S. & Paivinen, J. (2006) Ecological determinants of distribution decline and risk of extinction in moths. *Conservation Biology* **20**(4): 1161–1168.

Mattila, N., Kotiaho, J.S., Kaitala, V. & Komonen, A. (2008) The use of ecological traits in extinction risk assessments: a case study on geometrid moths. *Biological Conservation* **141**(9): 2322–2328.

McCarthy, M.A., Keith, D., Tietjen, J., Burgman, M.A., Maunder, M., Master, L., Brook, B.W., Mace, G., Possingham, H.P., Medellin, R., Andelman, S., Regan, H., Regan, T. & Ruckelshaus, M. (2004) Comparing predictions of extinction risk using models and subjective judgement. *Acta Oecologica-International Journal of Ecology* **26**(2): 67–74.

Nylin, S. & Bergstrom, A. (2009) Threat status in butterflies and its ecological correlates: how far can we generalize? *Biodiversity and Conservation* **18**(12): 3243–3267.

Paivinen, J., Grapputo, A., Kaitala, V., Komonen, A., Kotiaho, J., Saarinen, K. & Wahlberg, N. (2005) Negative density-distribution relationship in butterflies. *BMC Biology* **3**(1): 5.

Purvis, A., Gittleman, J.L., Cowlishaw, G. & Mace, G.M. (2000) Predicting extinction risk in declining species. *Proceedings of the Royal Society B. Biological Sciences* **267**(1456): 1947–1952.

Rassi, P., Alanen, A., Kanerva, T. & Mannerkoski, I. (2001) *The Red List of Finnish Species*. Helsinki, Finland: Ministry of the Environment and Finnish Environment Institute.

Roff, D.A. & Roff, R.J. (2003) Of rats and Maoris: a novel method for the analysis of patterns of extinction in the New Zealand avifauna before European contact. *Evolutionary Ecology Research* **5**(5): 759–779.

Selonen, V. & Helos, I. (2010) Positive density-distribution relationship in Finnish butterflies. *Annales Zoologici Fennici* **47**(5): 347–350.

Tabachnick, B.G. & Fidell, L.S. (2007) *Using Multivariate Statistics*. Boston, CT, USA: Pearson Education, Inc.