Journal of Animal Ecology 2004
73, 546–555

Investigating species co-occurrence patterns when species are detected imperfectly

DARRYL I. MACKENZIE*, LARISSA L. BAILEY† and JAMES. D. NICHOLS‡

**Proteus Research and Consulting Ltd, PO Box 5193, Dunedin, New Zealand; †Cooperative Fish and Wildlife Research Unit, Department of Zoology, North Carolina State University, Campus Box 7617, Raleigh, NC 27695–7617, USA; and ‡Patuxent Wildlife Research Center, 11510 American Holly Drive, Laurel, MD 20708–4017, USA*

Summary

1. Over the last 30 years there has been a great deal of interest in investigating patterns of species co-occurrence across a number of locations, which has led to the development of numerous methods to determine whether there is evidence that a particular pattern may not have occurred by random chance.
2. A key aspect that seems to have been largely overlooked is the possibility that species may not always be detected at a location when present, which leads to 'false absences' in a species presence/absence matrix that may cause incorrect inferences to be made about co-occurrence patterns. Furthermore, many of the published methods for investigating patterns of species co-occurrence do not account for potential differences in the site characteristics that may partially (at least) explain non-random patterns (e.g. due to species having similar/different habitat preferences).
3. Here we present a statistical method for modelling co-occurrence patterns between species while accounting for imperfect detection and site characteristics. This method requires that multiple presence/absence surveys for the species be conducted over a reasonably short period of time at most sites. The method yields unbiased estimates of probabilities of occurrence, and is practical when the number of species is small (< 4).
4. To illustrate the method we consider data collected on two terrestrial salamander species, *Plethodon jordani* and members of the *Plethodon glutinosus* complex, collected in the Great Smoky Mountains National Park, USA. We find no evidence that the species do not occur independently at sites once site elevation has been allowed for, although we find some evidence of a statistical interaction between species in terms of detectability that we suggest may be due to changes in relative abundances.

Key-words: detectability, Great Smoky Mountains National Park, likelihood, occupancy, *Plethodon jordani*, *Plethodon glutinosus* complex.

Journal of Animal Ecology (2004) **73**, 546–555

Introduction

One approach to ecological science seeks to draw inferences about community dynamics and function based on observed patterns (e.g. Brown 1995; Rosenzweig 1995; Marquet 2000; Hubbell 2001). One type of pattern that has attracted much attention from ecologists is the spatial occurrence of species. Indeed, a simple presence–absence matrix of species occurrence in spatial units has been termed 'the fundamental unit of analysis in community ecology and biogeography' (Gotelli 2000; also see McCoy & Heck 1987). Investigations of such matrices have led to the development of interesting ecological hypotheses (e.g. the community assembly rules of Diamond 1975) and to the identification of interesting empirical patterns (e.g. the nested subset structure of Patterson & Atmar 1986; Patterson 1987).

A key issue in the investigation of presence–absence matrices involves how to draw appropriate inferences about whether an observed matrix is unusual with respect to either random processes or processes that are neutral with respect to some purported ecological mechanism (Harvey *et al*. 1983; Gotelli & Graves 1996). For example, could a particular matrix have been generated by random species colonizations or is it more likely to have arisen as a result of interspecific competition? This issue of appropriate inferential methods has led to heated debate

Correspondence: Proteus Research and Consulting Ltd, PO Box 5193, Dunedin, New Zealand. E-mail: Darryl@proteus.co.nz

(Connor & Simberloff 1979, 1983, 1984; Diamond & Gilpin 1982; Gilpin & Diamond 1982, 1984) and continued methodological development (Kelt, Taper & Mesevre 1995; Manly 1995; Gotelli 2000; Gotelli & McCabe 2002).

In this paper we address a problem that has not received adequate attention in previous work, the assumption that all species present at a location are detected with certainty. In many, if not most, practical situations it is not realistic to obtain a census of all species. Few species are so conspicuous that they will always be detected when present at a location and in many cases, even after exhaustive searches, some species may still go undetected when present. This feature of the data collection will lead to 'false absences' in the presence–absence matrix, which may lead in turn to incorrect inferences about the patterns of species co-occurrence. Cam *et al*. (2000) presented methods that can be used to deal with species non-detection when testing hypotheses about nested subset community patterns (Patterson & Atmar 1986; Patterson 1987). The methods of Cam *et al*. (2000) are based on estimates of the fraction of species present at one location that are also present at another (Nichols *et al*. 1998). However, these estimation methods are based on groups of species and cannot be used to draw inferences about specific patterns of co-occurrence of a small number of species.

Another potential problem with attempts to draw inferences about interspecific interactions from presence–absence matrices involves other factors (e.g. habitat preferences and physiological tolerances) which are likely to result in non-random patterns of species co-occurrence, yet have nothing to do with interspecific interactions. This class of problem is inherent in all attempts to draw inferences about process based on pattern and has been recognized in previous efforts to analyse presence–absence matrices (e.g. Connor & Simberloff 1984; Gilpin & Diamond 1984; Peres-Neto, Olden & Jackson 2001). One approach to dealing with such factors is to identify them a priori and incorporate them into analyses. For example, one approach is to develop a regression model to predict detections of one species as a function of both habitat variables and detections of other species (Schoener 1974; Crowell & Pimm 1976).

Here, we present a method that deals with both problems by incorporating both non-detection and possible habitat preferences directly into the model set. This method is based on the approach of MacKenzie *et al*. (2002), who developed a single-species model for estimating the fraction of locations occupied by the species, allowing for the possible non-detection of the species when present. They considered the realistic situation where multiple surveys are conducted at the locations, over a relatively short time period. Straightforward probabilistic arguments are used to model the sequence of species detections and non-detections from the repeated surveys, enabling the probability of an observed sequence to be calculated. By combining the information from all locations, the model parameters (probability of occupancy and probability of detection given occupancy) can be estimated using maximum likelihood techniques. Importantly, the model of MacKenzie *et al*. (2002) does not require equal sampling effort across all locations, and the parameters can be functions of covariates such as habitat type.

Here we extend the work of MacKenzie *et al*. (2002) to estimate and model co-occurrence patterns between two or more species across a landscape, when species are not detected with certainty when present at a location. The likelihood-based framework detailed below enables the magnitude of interspecific interactions in probabilities of occurrence to be estimated directly, while accounting explicitly for imperfect detectability. The flexibility of our approach also enables the level of co-occurrence to be estimated, above and beyond any habitat preferences exhibited by the species. We envisage that this model could be most useful to address questions about the importance of interspecific interactions such as competition and predator–prey relationships as potential determinants of community structure.

In this paper we begin by discussing the practical sampling framework required for the method; detail the straightforward probabilistic arguments used to construct the model likelihood; show, via simulation, that the estimators have reasonable properties in terms of bias and precision; and apply the method to a field study of terrestrial salamanders in Great Smoky Mountains National Park (Bailey, Simons & Pollock 2004). Throughout this paper we refer to interactions between species. When we do so, we use the term 'interaction' in the statistical sense to mean that the species are not occurring independently at sites. Our use of 'interaction' does not imply any particular biological mechanism (e.g. predation, resource competition, behavioural dominance) that could produce a lack of independence in pattern of co-occurrence.

Methods

PRACTICAL SAMPLING SITUATION

We envisage a practical situation where N locations are monitored for the presence or absence of target species. The monitoring locations may represent user-specified quadrats or sites within an area of interest, or discrete habitats such as ponds, islands or patches of vegetation. Each location is surveyed for the species on multiple (not necessarily an equal number of) occasions, and species are either detected or not detected during each survey. For the duration of the surveying the locations are closed to changes in the occupancy state with respect to each species, i.e. a species is either always present, or always absent from the location over the surveying period (this requirement may be relaxed in some situations, see the discussion).

The sequence of detections and non-detections at a location for each species may be recorded as a 'detection history': a vector of 1s (detection) and 0s (non-detection).

For example, the detection history $\mathbf{X}_i^{\mathrm{A}} = 101$ represents that location *i* was surveyed on three occasions, with species A being detected only in the first and third surveys. Similarly, the detection history $\mathbf{X}_i^{\mathrm{B}} = 000$ would represent that species B was never detected at location *i*.

STATISTICAL MODEL

We define the model here for situations involving only two species, but the approach can easily be extended to a greater number of species. However, the number of parameters in the model increases exponentially with the number of species; hence this technique could become very 'data hungry' and not all of the parameters may be estimable for a given data set. In addition it could be difficult to interpret meaningfully the interactions among a large number of species; therefore, we recommend that users focus their research questions on a small (< 4) number of target species.

A monitoring location may be considered to be in one of four mutually exclusive states of occupancy for two species (more generally there are 2^k possible states for *k* species); (1) occupied by both species A and B; (2) occupied by species A only; (3) occupied by species B only; or (4) occupied by neither species. Using the notation introduced in Table 1, we define a row vector for the probability of location *i* being in each of the four respective states as,

$$\boldsymbol{\phi}_i = [\psi_i^{\mathrm{AB}} \quad \psi_i^{\mathrm{A}} - \psi_i^{\mathrm{AB}} \quad \psi_i^{\mathrm{B}} - \psi_i^{\mathrm{AB}} \quad 1 - \psi_i^{\mathrm{A}} - \psi_i^{\mathrm{B}} + \psi_i^{\mathrm{AB}}]. \qquad \text{eqn 1}$$

Note that the elements of $\boldsymbol{\phi}_i$ sum to 1.

Conditional upon the occupancy state of the location, the probability of observing the detection histories for the two species can be stated in terms of the detection probabilities defined in Table 1. For example, the probability of observing the detection histories given in the previous section, conditional upon the location being occupied by both species, is:

$$\Pr(\mathbf{X}_i^{\mathrm{A}} = 101, \mathbf{X}_i^{\mathrm{B}} = 000 \mid \text{both species present}) = r_{i1}^{\mathrm{Ab}} r_{i2}^{\mathrm{ab}} r_{i3}^{\mathrm{Ab}}.$$

Another possibility for this example would be that the location is occupied by species A only, in which case the probability of not observing species B is 1·0. The conditional probability of observing the two detection histories in this situation would be:

$$\Pr(\mathbf{X}_i^{\mathrm{A}} = 101, \mathbf{X}_i^{\mathrm{B}} = 000 \mid \text{only species A present}) = p_{i1}^{\mathrm{A}}(1 - p_{i2}^{\mathrm{A}}) p_{i3}^{\mathrm{A}}$$

The probability of observing this combination of histories for all other occupancy states (occupied by species B only and occupied by neither species) is 0, as both states prohibit species A from being at the location, yet species A was actually observed there. Therefore, we define a column vector $\mathbf{p}_i^{\{\mathbf{X}^{\mathrm{A}}\},\{\mathbf{X}^{\mathrm{B}}\}}$ representing the probability of observing the detection histories conditional upon each state. For instance, using the above example:

$$\mathbf{p}_i^{\{101\},\{000\}} = \begin{bmatrix} r_{i1}^{\mathrm{Ab}} r_{i2}^{\mathrm{ab}} r_{i3}^{\mathrm{Ab}} \\ p_{i1}^{\mathrm{A}}(1 - p_{i2}^{\mathrm{A}}) p_{i3}^{\mathrm{A}} \\ 0 \\ 0 \end{bmatrix}. \qquad \text{eqn 2}$$

The unconditional probability for observing the two detection histories could then be calculated as:

$$\Pr(\mathbf{X}_i^{\mathrm{A}}, \mathbf{X}_i^{\mathrm{B}}) = \boldsymbol{\phi}_i \, \mathbf{p}_i^{\{\mathbf{X}^{\mathrm{A}}\},\{\mathbf{X}\mathrm{B}\}}. \qquad \text{eqn 3}$$

By using the probability vectors we account for potential uncertainties in the occupancy state of a location due to not detecting one or both of the target species during the surveys. Note that our use of different detection probability parameters for the cases of single-species and two-species occupancy is very general and permits the possibility that detection probability of one species depends on whether the site is occupied by the other species (e.g. the detection probability for a prey species may depend upon whether a predator species is also present). Some examples of detection histories and the unconditional probabilities of observing them are given in Table 2.

Assuming that the detection histories collected at the *N* locations are independent, we can define the model likelihood as:

$$L = \prod_{i=1}^{N} \Pr(\mathbf{X}_i^{\mathrm{A}}, \mathbf{X}_i^{\mathrm{B}}). \qquad \text{eqn 4}$$

Table 1. Notation for the parameters used in the model

Parameter	Description
ψ_i^{AB}	Probability of both species being present at location *i*
ψ_i^{A}	Probability of species A being present at location *i*, regardless of occupancy status of species B
ψ_i^{B}	Probability of species B being present at location *i*, regardless of occupancy status of species A
p_{ij}^{A}	Probability of detecting species A during the *j*th survey of location *i*, given only species A is present
p_{ij}^{B}	Probability of detecting species B during the *j*th survey of location *i*, given only species B is present
r_{ij}^{AB}	Probability of detecting both species during the *j*th survey of location *i*, given both species are present
r_{ij}^{Ab}	Probability of detecting species A, but not B, during the *j*th survey of location *i*, given both species are present
r_{ij}^{aB}	Probability of detecting species B, but not A, during the *j*th survey of location *i*, given both species are present
r_{ij}^{aB}	Probability of detecting neither species during the *j*th survey of location *i*, given both species are present; $= 1 - r_{ij}^{\mathrm{AB}} - r_{ij}^{\mathrm{Ab}} - r_{ij}^{\mathrm{aB}}$

Table 2. Example detection histories ($\mathbf{X}^A$, $\mathbf{X}^B$) and the probabilities of observing them ($\Pr(\mathbf{X}^A, \mathbf{X}^B)$)

$\mathbf{X}^A$	$\mathbf{X}^B$	$\Pr(\mathbf{X}^A, \mathbf{X}^B)$
101	001	$= \begin{bmatrix} \psi^{AB} \\ \psi^{A} - \psi^{AB} \\ \psi^{B} - \psi^{AB} \\ 1 - \psi^{A} - \psi^{B} + \psi^{AB} \end{bmatrix}^T \begin{bmatrix} r_1^{Ab}(1 - r_2^{AB} - r_2^{Ab} - r_2^{aB}) r_3^{AB} \\ 0 \\ 0 \\ 0 \end{bmatrix}$ $= \psi^{AB}\{r_1^{Ab}(1 - r_2^{AB} - r_2^{Ab} - r_2^{aB}) r_3^{AB}\}$
000	110	$= \begin{bmatrix} \psi^{AB} \\ \psi^{A} - \psi^{AB} \\ \psi^{B} - \psi^{AB} \\ 1 - \psi^{A} - \psi^{B} + \psi^{AB} \end{bmatrix}^T \begin{bmatrix} r_1^{aB} r_2^{aB}(1 - r_3^{AB} - r_3^{Ab} - r_3^{aB}) \\ 0 \\ p_1^{B} p_2^{B}(1 - p_3^{B}) \\ 0 \end{bmatrix}$ $= \psi^{AB}\{r_1^{aB} r_2^{aB}(1 - r_3^{AB} - r_3^{Ab} - r_3^{aB})\} + (\psi^{B} - \psi^{AB})\{p_1^{B} p_2^{B}(1 - p_3^{B})\}$
000	000	$= \begin{bmatrix} \psi^{AB} \\ \psi^{A} - \psi^{AB} \\ \psi^{B} - \psi^{AB} \\ 1 - \psi^{A} - \psi^{B} + \psi^{AB} \end{bmatrix}^T \begin{bmatrix} \prod_{j=1}^{3}(1 - r_j^{AB} - r_j^{Ab} - r_j^{aB}) \\ \prod_{j=1}^{3}(1 - p_j^{A}) \\ \prod_{j=1}^{3}(1 - p_j^{B}) \\ 1 \end{bmatrix}$ $= \psi^{AB} \prod_{j=1}^{3}(1 - r_j^{AB} - r_j^{Ab} - r_j^{aB}) + (\psi^{A} - \psi^{AB}) \prod_{j=1}^{3}(1 - p_j^{A})$ $+ (\psi^{B} - \psi^{AB}) \prod_{j=1}^{3}(1 - p_j^{B}) + (1 - \psi^{A} - \psi^{B} + \psi^{AB})$

This can then be maximized numerically to obtain the maximum likelihood estimates (MLEs) of the parameters.

For generality, we have presented the above model using location-specific parameters (as denoted by the subscript *i*); however, there is never sufficient information in the type of data considered here to estimate a different parameter for each location. Constraints are required in order to obtain MLEs (e.g. require all or groups of locations to have a common parameter). Another approach is to let the location-specific parameters be defined by some function of the features that characterize a location, i.e. habitat type, patch size, etc. We consider how these covariates could be accommodated by the model in a later section.

TESTING AND QUANTIFYING INTERACTIONS BETWEEN SPECIES

The general likelihood-based framework presented above provides the opportunity to both test for, and quantify, the level of interaction between two species. There are two mechanisms through which we can investigate species interactions that may reflect different questions of biological interest; species occupancy probabilities or in terms of detection probabilities given the species are present. In both cases we can determine whether the occupancy (or detection events) for one species appear to be occurring independently of the presence (or detection) of the other species, i.e. do the species both occur at a site (or similarly, do we detect both species in a survey) more/less often than expected under an assumption of independence. In addition, we can determine whether there is evidence that the probability of detecting one species changes in the presence of the other species, i.e. if species B is also present at a site, we are more/less likely to detect species A (regardless of detecting species B).

To investigate potential interactions between species, one has the choice of using hypothesis testing or a model selection approach, depending upon the goals of the research. Standard likelihood ratio tests (LRT) could be used to test for independence of the species with respect to either occupancy or detection. For example, if species occupy sites independently then, based upon the statistical definition of independence, it would be expected that $\psi^{AB} = \psi^{A} \times \psi^{B}$. A LRT could be constructed by comparing the likelihood values from two models; a full model where ψ^{AB}, ψ^{A} and ψ^{B}, and are each estimated; and a reduced model where only ψ^{A} and ψ^{B} are estimated, with ψ^{AB} being calculated as the product of ψ^{A} and ψ^{B} (note that the structure for all other parameters is unchanged between the full and reduced models). By conducting the test it is possible to determine whether there is sufficient evidence to reject the null hypothesis of independence. Examples of the constraints that could be imposed are given in Table 3. Alternatively, it may be appropriate to explore the data using information-theoretic model selection approaches (e.g. Akaike's

Table 3. Examples of the constraints that should be imposed for testing the independence of occupancy and detection probabilities, where r_{ij}^{A} and r_{ij}^{B} are the marginal detection probabilities for the respective species in survey j, given both species are present

Testing	Constraints
Occupancy	$\psi_i^{AB} = \psi_i^{A} \times \psi_i^{B}$
Detection	$r_{ij}^{AB} = r_{ij}^{A} \times r_{ij}^{B}$ $r_{ij}^{AB} = r_{ij}^{A} \times (1 - r_{ij}^{B})$ $r_{ij}^{aB} = (1 - r_{ij}^{A}) \times r_{ij}^{B}$ $r_{ij}^{ab} = (1 - r_{ij}^{A}) \times (1 - r_{ij}^{B})$

information criterion, AIC), where the intent is to find a set of parsimonious models upon which inferences about the species biology could be made (e.g. Burnham & Anderson 2002).

The magnitude of the interaction between species could be estimated from the parameter estimates of the full model (e.g. as $\hat{\gamma} = \hat{\psi}^{AB}/\hat{\psi}^{A}\hat{\psi}^{B}$), which we term a species interaction factor (SIF). Values of $\hat{\gamma} < 1$ would suggest species avoidance (i.e. the species co-occur less frequently than if they were distributed independently), while values > 1 would suggest contagion, or a tendency to co-occur more frequently than expected under independence. Note that $\hat{\gamma} = 1$ would suggest the species occur independently. However, often it may be advantageous to reparameterize the model so that the SIF is estimated directly, i.e. $\psi^{AB} = \psi^{A} \times \psi^{B} \times \gamma$. Similarly, we can redefine the detection probabilities r^{AB} as $r^{AB} = r^{A} \times r^{B} \times \delta$ where r^{A} and r^{B} are the overall probabilities of detecting species A and B during a survey, given both species are present, and δ is the SIF for the detection probabilities.

To consider whether the probability of detecting species A during a survey is different when species B is also present, we could compare models where the constraint $r_j^{A} = p_j^{A}$ is used (and similarly for species B when species A is also present). Note that this issue is distinct from the question of whether detections of the two species occur independently given that both species are present (i.e. does $\delta = 1$?).

INCORPORATING COVARIATE INFORMATION

Potentially, the probability that a species occupies a location may be affected by characteristics of the location. For example, some species may prefer particular habitat types over other available habitats; have a higher occupancy rate at locations near permanent water sources; require a minimum patch size for a sustainable population; or show reduced probability of occurrence in isolated patches (e.g. Verner, Morrison & Ralph 1986; Scott *et al.* 2002). Similarly, the probability of detecting species at the location may also be affected by location-specific covariates (e.g. old growth forest vs. rejuvenating forest). Detection probabilities may also be affected by conditions at the time of the survey, such as air temperature, cloud cover, or time since a rain event.

One method for incorporating such covariates is to use the multinomial logistic model (eqn 5).

$$\theta_i^k = \frac{\exp(\mathbf{Y}_i\boldsymbol{\beta}_k)}{1 + \sum_{k=1}^{m-1} \exp(\mathbf{Y}_i\boldsymbol{\beta}_k)}, \text{ for } k = 1, 2, \ldots, m-1, \quad \text{eqn 5}$$

where θ_i is the probability of interest, $\mathbf{Y}_i$ is a row vector of the covariate values for the ith location, $\boldsymbol{\beta}_i$ is the column vector of coefficients to be estimated and m is the number of discrete outcomes. For example, when modelling the probabilities for detecting/not-detecting both species at a survey occasion, e.g. $\{r_{i1}^{AB}, r_{i1}^{Ab}, r_{i1}^{aB}, r_{i1}^{ab}\}$ there are four discrete outcomes. Three of these probabilities could be modelled using eqn 5, with the final probability being obtained by subtraction. Note that when $m = 2$ (i.e. only two discrete outcomes), eqn 5 reduces to the more familiar binomial logistic model that could be used, for instance, for modelling the p_{ij}^{A}s or p_{ij}^{B}s where the individual species may be either detected or not detected.

For modelling the occupancy probabilities, one could use the multinomial logistic model on the elements of $\boldsymbol{\phi}_i$, although the results may not be biologically meaningful, i.e. interpreting the effect of a covariate on $\psi_i^{A} - \psi_i^{AB}$. Another approach would be to use the SIFs, so that modelling of ψ_i^{A} and ψ_i^{AB} is achieved using separate binomial logistic models, while ψ_i could be modelled as:

$$\gamma_i = \exp(\mathbf{Y}_i\boldsymbol{\beta}_\gamma). \quad \text{eqn 6}$$

However, when using such an approach, users must be mindful of the natural relationship among ψ_i^{AB}, ψ_i^{A} and ψ_i^{B}, which restricts the values that ψ_i^{AB}, hence, γ_i can possibly take, reflecting limits to the degree of overlap that is possible between the two species, i.e.:

$$\max(\psi_i^{A} + \psi_i^{B} - 1, 0) \le \psi_i^{AB} \le \min(\psi_i^{A}, \psi_i^{B}). \quad \text{eqn 7}$$

For example, if ψ_i^{A} and $\psi_i^{B} = 0{\cdot}6$, then the two species must both occur at a minimum of 20% of the locations, while if they exactly co-occur then it can only be at 60% of sites at most. This restriction must be enforced when using SIFs (which may cause numerical problems), but when using the first approach the restriction is automatically imposed because of the different parameterization of the covariate relationship. Similar reasoning applies when using the SIF parameterization with respect to the r parameters.

MISSING OBSERVATIONS

A probable feature of many wildlife studies is that occasionally not all locations will be surveyed for the target species. This may be due either to logistical constraints (it is simply not possible to survey all locations virtually simultaneously); study design; or unforeseen circumstances such as a vehicle breakdown en route. We define

such occasions as a missing observation. The flexible modelling framework presented above can be modified easily to accommodate missing observations. As in MacKenzie *et al*. (2002), for occasions when the location was not surveyed, the respective detection probability (or probabilities) is set to zero, effectively removing it from the probabilistic statement about the observed detection history for that location.

An important point is that by being able to accommodate missing observations, the model does not require equal sampling effort across all locations. This provides a great deal of flexibility for study design. For example, under certain conditions it may be appropriate to survey a subsample of locations more frequently to gain adequate information about the detection probabilities, and elsewhere survey only once or twice.

Simulation study

To assess the performance of the above modelling a simulation study was conducted, with four basic patterns in species occupancy being investigated. Two species were given equal probabilities for occupying sites at a moderate and a high level. The species were then assumed to either exhibit a strong association or disassociation. The four combinations of $\{\psi^A, \psi^B, \psi^{AB}\}$ used in the simulations were; (i) {0·4, 0·4, 0·08}; (ii) {0·4, 0·4, 0·24}; (iii) {0·7, 0·7, 0·4}; and (iv) {0·7, 0·7, 0·6125}. In addition, the effects of three other factors were varied to assess their influence on the estimation of the model parameters; (1) total number of locations surveyed (N) = 50, 100 or 200; (2) number of repeat surveys (T) = 3 or 5; and (3) probability of detecting each species during a survey, given presence (p) = 0·214 or 0·5. For simplicity, the detection of each species was assumed to be independent of detection of the other ($\delta = 1$), detection probabilities were made constant across time, equal for both species ($p^A = p^B$), and equal regardless of whether one or both species were present ($r = p$). The values of p used were chosen such that the probability of never detecting the species given it was actually there, i.e. $(1 - p)^T$, was approximately 0·5 and 0·3 when $p = 0{\cdot}214$ (for $T = 3$ and 5, respectively); and 0·125 and 0·03 when $p = 0{\cdot}5$.

For each scenario, 1000 sets of simulated data were generated and a model with the following parameters was fitted to the data, ψ^A, ψ^B, ψ^{AB}, p^A, p^B, r^{AB}, r^{Ab} and r^{aB}. This represents a model where neither the occupancy nor the detection probabilities are assumed to be independent between the two species, and detection probabilities are constant across time (and locations). From each set of data, parameter estimates were obtained and their standard errors approximated by inverting the matrix of second partial derivatives (a standard numerical technique). The average of the 1000 parameter estimates was used to assess unbiasedness, while the standard deviation of the 1000 parameter estimates was compared to the average of the 1000 standard errors to ensure that the approximated standard errors fairly reflected the true level of uncertainty in the parameter estimates.

In approximately 8·5% of the simulations (on average) the matrix of second partial derivatives could not be inverted. This was not unexpected and is a common feature of likelihood-based methods when parameters are estimated very close to the bounds of allowable values (e.g. 0 or 1). These simulation results were discarded, which may introduce a small bias, but our results and further investigations suggest any such bias is negligible.

The results of the simulations suggest the parameter estimates are virtually unbiased for most scenarios considered, and have a reasonable level of precision. The standard errors are generally in good agreement with the true level of uncertainty. Figure 1 presents the percentage bias for the estimated joint probability of occupancy (ψ^{AB}) and its standard error. In this instance, the bias is minimal except for when $N = 50$, $T = 3$, $p = 0{\cdot}5$ and occupancy for both species was moderate, with a strong disassociation ($\psi^A, \psi^B, \psi^{AB}$) = {0·4, 0·4, 0·008}, in which case ψ^{AB} tended to be overestimated and its standard error underestimated. Full results for the simulation study can be obtained by contacting the corresponding author.

Example: terrestrial salamanders in Great Smoky Mountains National Park

We illustrate the utility of this approach using monitoring data collected on terrestrial salamanders at 88 sites within the Roaring Fork Watershed, Great Smoky Mountains National Park (GSMNP, Mt LeConte USGS Quadrangle). Sites were located adjacent to trails and spaced approximately 250 m apart (see Hyde & Simons 2001 for sampling details). Two parallel transects were sampled at each site: a natural cover transect (50 m long × 3 m wide) and coverboard transect consisting of five stations placed 10 m apart (see Hyde & Simons (2001) for details). Sites were sampled five times between 4 April 1999 and 27 June 1999, with approximately 2 weeks between successive sampling occasions. Relative abundance information was collected for each species but here we consider only detection/non-detection data (pooled for both transects) for two species: the red cheek variation of Jordan's salamander (*Plethodon jordani* Blatchley; PJ) and members of the *Plethodon glutinosus* complex including *Plethodon glutinosus* (Green) and *Plethodon oconaluftee* (Hairston; PG). We stress that the following analyses are presented only as an example of the above method, and they should not be used to draw definitive conclusions about co-occurrence patterns between these two species.

Several previous studies have sought to document the spatial distributions of these two species and explain geographical variation in their altitudinal overlap. Hairston (1980) found that competitive interactions were stronger in areas of little altitudinal overlap (GSMNP and Black Mountains, NC) than in areas of broad altitudinal overlap (Balsalm Mountains, NC). Other studies

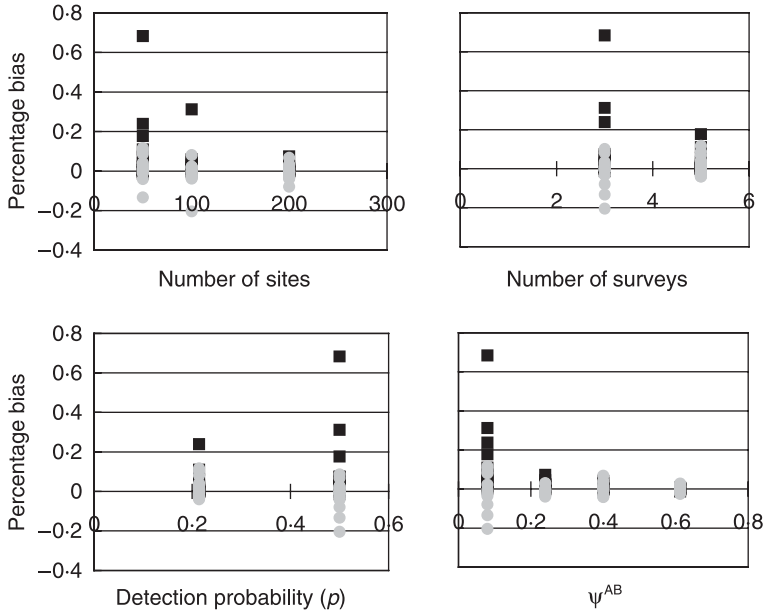

Fig. 1. Approximate percentage bias of estimated joint probability of occupancy ($\hat{\psi}^{AB}$) and its associated standard error ( ), obtained from a simulation study, plotted against the factors; number of sites, number of surveys, detection probability per survey (p); and true value of ψ^{AB}.

Table 4. Summary of model fit and selection statistics for models without elevation as a covariate, where K is the number of estimated parameters in the model and ΔAIC is the absolute difference in AIC values relative to the model with the smallest AIC. The terms in parentheses represent the factors in the model for the respective parameter; with 'S' denoting that species has been used as a factor and "·" indicating that the parameter is constant. For example, $\psi(S)$ indicates that the occupancy probability has been estimated separately for both species, whereas $\gamma(\cdot)$ indicates that this parameter has a constant value to be estimated. Absence of the parameter in the model notation implies $\gamma(\cdot)$ and absence of $r(S)$ implies $r(S) = p(S)$

Model	Log-likelihood	K	ΔAIC
$\psi(S)\gamma(\cdot)p(S)r(S)$	736·6	7	0·0
$\psi(S)p(S)r(S)$	747·0	6	8·3
$\psi(S)\gamma(\cdot)p(S)$	761·4	5	20·7
$\psi(S)p(S)$	776·0	4	33·4

Table 5. Summary of model fit and selection statistics for models with elevation as a covariate, where K is the number of estimated parameters in the model and ΔAIC is the absolute difference in AIC values relative to the model with the smallest AIC. The terms in parentheses represent the factors in the model for the respective parameter; with 'S' denoting that species has been used as a factor, 'E' indicating use of elevation as a factor, and '·' indicating a parameter set equal across species and elevation. The best model from Table 4, $\psi(S)\gamma(\cdot)p(S)r(S)$ has been included to show how including elevation as a covariate substantially improves the fit of the models. Absence of the γ parameter in the model notation implies $\gamma(\cdot)$ and absence of $r(S)$ implies $r(S) = p(S)$

Model	Log-likelihood	K	ΔAIC
$\psi(S\times E)p(S\times E)r(S\times E)$	617·3	12	0·0
$\psi(S\times E)p(S)r(S\times E)$	623·6	10	2·3
$\psi(S\times E)p(S\times E)r(S)$	660·1	10	38·8
$\psi(S\times E)p(S\times E)$	675·6	8	50·2
$\psi(S\times E)p(S)r(S)$	676·1	8	50·8
$\psi(S)p(S\times E)r(S\times E)$	673·2	10	51·8
$\psi(S)\gamma(\cdot)p(S\times E)r(S\times E)$	671·8	11	52·5
$\psi(S)\gamma(\cdot)p(S)r(S)$	736·6	7	109·3

have found no evidence of competitive exclusion, suggesting the species' distributions are either independent (Rissler, Barber & Wilbur 2000) or determined by habitat or environmental factors (Dakin 1978).

Here we are interested in determining whether there is any evidence that the two species exhibit strong co-occurrence patterns after allowing for any elevational gradient in occupancy probabilities. Throughout the following analysis we use the SIF parameterization of the model and assume that $\delta = 1$, i.e. the species are detected independently when both are present. We feel this is a reasonable assumption to make given the field design and known biology of these species.

Table 4 shows the model fit and selection statistics for models that do not acknowledge a potential elevational gradient in occupancy and detection probabilities. Based upon AIC, the most parsimonious model among those considered for the data is $y(S)y(\cdot)p(S)r(S)$, which suggests that the detection probability for each species is different if the other species is also present (for PG: $\hat{p} = 0{\cdot}54$ and $\hat{r} = 0{\cdot}48$; for PJ: $\hat{p} = 0{\cdot}91$ and $\hat{r} = 0{\cdot}55$), and that there is very strong evidence that the two species avoid each other $\hat{\gamma}(S\hat{E}) = 0{\cdot}67$ (0·11).

However, once we allow these parameters to vary with elevation, we obtain models that provide much better descriptions of the data (Table 5). Unfortunately we were not able to obtain models that included a γ term to

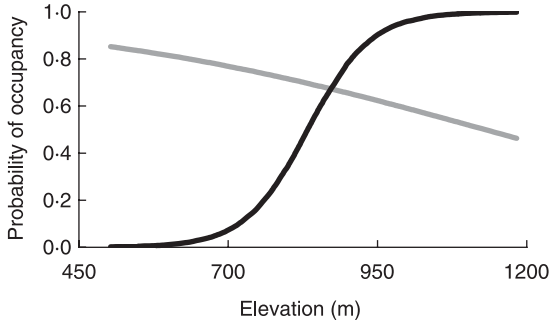

Fig. 2. Estimated probability of occupying a site for *Plethodon jordani* (▬) and members of the *Plethodon glutinosus* complex (▬) as a function of elevation according to the model.

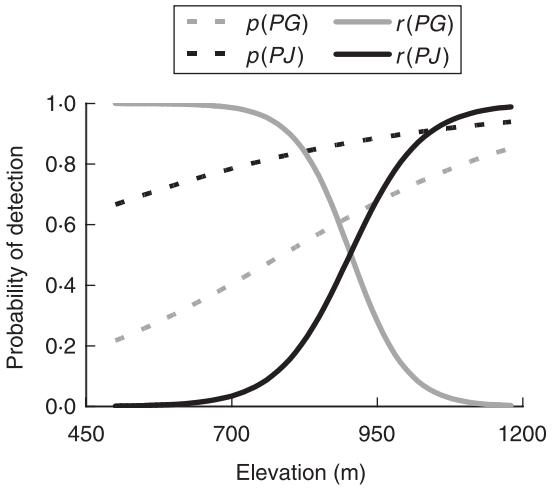

Fig. 3. Estimated probability of detecting the species *Plethodon jordani* (*PJ*) and *Plethodon glutinosus* (*PG*) in a survey, as a function of elevation according to the model $\psi(S \times E)p(S \times E)r(S \times E)$.

converge satisfactorily, because once the probability of occupancy for PJ was modelled as a function of elevation, the predicted occupancy probability was < 0·15 for elevations below 750 m and > 0·80 for elevations above 902 m (Fig. 2). At lower elevations, this means there are very few data on which to judge whether the species were acting independently, while at the higher elevations there is a very small range of allowable values for γ, implying that there is little scope to evaluate nonindependent behaviour in terms of occupancy for these species (i.e. the lower and upper bounds on allowable values for γ, from eqn 7, both tend to 1·0 as elevation increases). The most parsimonious model we were able to fit to the data, $\psi(S \times E)p(S \times E)r(S \times E)$ was indicated clearly as the 'best' model in terms of AIC. Figures 2 and 3 illustrate how the various factors are affected by elevation. While we have no evidence of an interaction between the species in terms of occupancy probabilities, there is strong evidence that the detection functions are different if both species are present at a site. For instance, the detection probability for PG increases with elevation when PJ is not present but decreases when PJ is present, whereas for PJ the effect of elevation is much larger when both species are present than when PJ is present alone. From an observational study such as this it is difficult to suggest exactly what may be the cause for this phenomenon, but a plausible explanation involves effects of relative abundance, a potentially important determinant of species detection probability. For example, it may be that PG becomes more abundant as elevation increases until PJ is also reasonably abundant. At that point, the abundance of PG starts to decrease while PJ continues to become even more abundant (perhaps through competition for resources). This reasoning is consistent with other field studies, which conclude that while PG and PJ have shown no tendency to be mutually exclusive, PG is more tolerant of dry locations found usually at lower elevations (Grover 2000; Rissler *et al*. 2000) and PJ seems to have a numerical advantage in moist microhabitats common at higher elevations (Hairston 1951; Dakin 1978). While this reasoning is supported by published studies it is speculative, and we caution against inferring ecological process from spatial patterns without the support of experimental studies. In addition, in this example analysis we have not considered the potential effects of other habitat variables through more complicated models.

Discussion

A number of previous authors have suggested various methods to test the null hypothesis of independence of species occurrence and to provide related interaction metrics both for two-species systems (Forbes 1907; Dice 1945; Cole 1949; Pielou 1977; Hayek 1994) and for more complex multispecies systems (Connor & Simberloff 1979, 1984; Gilpin & Diamond 1982, 1984; Kelt *et al*. 1995; Manly 1995; Gotelli 2000; Gotelli & McCabe 2002). However, with the exception of the work of Cam *et al*. (2000) directed at specific questions about nested subset structures (Patterson & Atmar 1986), we believe that the approach presented here is the first attempt to account explicitly for the imperfect detectability of species while modelling species co-occurrence data. Failure to allow for the fact that a species may have been present, but not detected, can result in misleading conclusions about species associations and interactions, as some species may have been classified falsely as absent. The flexible likelihood-based modelling framework we present is based on simple probabilistic arguments that are used commonly in other areas of statistical ecology such as mark–recapture (Lebreton *et al*. 1992), and are used widely in many statistical disciplines. Hence there is already a vast body of literature supporting the general approach. The modelling of the different occupancy states involves the same kind of thinking that has been used to develop previous approaches to testing for independence in the case of perfect detection (Forbes 1907; Dice 1945; Cole 1949; Pielou 1977; Hayek 1994). Thus, our approach to modelling and inference unites two approaches that are themselves very familiar to ecologists.

Initial investigations into the different possible methods for incorporating covariates into the occupancy probabilities suggest that using the multinomial logistic model on the elements of is the most numerically robust approach. However, as suggested earlier, this may give results that are difficult to interpret biologically. Our preference is for the use of the species interaction factors (SIFs), as they provide a meaningful interpretation for the strength of a covariate relationship on the nonindependence of two species.

While in the terrestrial salamander example we were unable to get convergence for models that involved both γ and occupancy as a function of elevation (hence we were unable to investigate possible species interactions in this respect after allowing for the effects of elevation), the example does highlight the importance of considering factors that may affect the marginal probabilities of species occurring at study sites when exploring patterns of species co-occurrences. When we did not use elevation as a covariate in our models, there was very strong evidence that the species were less likely to both occupy a site than they would have been if they were acting independently (Table 4). However, once we began to consider models that included elevation as a covariate, this strong evidence of an interaction disappeared. For example, consider the models $\psi(S)p(S \times E)r(S \times E)$ and $\psi(S)\gamma(\cdot)p(S \times E)r(S \times E)$ in Table 5. Here we have only allowed the detection probabilities to be functions of elevation, yet already there is little indication that by including γ we have a better model, given that both models have similar AIC values. By ignoring potential factors that may affect a researcher's ability to detect target species or factors that may affect whether a species occupies a particular location (such as habitat variables), erroneous conclusions may be reached concerning patterns of co-occurrence.

Above we have presented the estimation of model parameters in terms of maximizing the likelihood. However, another approach would be to assign appropriate prior distributions on the model parameters, representing current knowledge (or ignorance), and use the likelihood within a Markov chain Monte Carlo framework to obtain posterior distributions for the parameters. Such an approach may provide some benefits, enabling models to be explored that are intractable using standard maximum likelihood theory.

In some circumstances it may be appropriate to relax the assumption that all locations are closed to any changes with respect to occupancy for the duration of the surveying. If the species move in and out of the study locations in a completely random manner, such as for a highly mobile species, then based upon the results of Kendall (1999) in a closely related mark–recapture context we believe that parameter estimates will still be valid, although their interpretation should change. What we have referred to as 'occupied locations' above should be interpreted as 'used locations', and 'probability of detection' is now 'probability species is present and detected'. However, parameter estimates are no longer valid if the changes in occupancy are non-random, i.e. if animals move to a location during the middle of the seasonal survey period or vacate the location before the sampling has been completed.

Finally, although we believe that the methods proposed here can yield strong inferences about species co-occurrence using presence–absence data from multiple locations at a single point (e.g. season) in time, we warn that this does not imply strong inference about the processes that generated any observed patterns of co-occurrence. Despite the popularity of inferring process from pattern in ecology, strong inference about process requires typically some sort of manipulative experimentation. Although not generally as powerful as experimentation, it is often useful to observe system dynamics over time. MacKenzie *et al*. (2003) presented a model structure for estimating the vital rates associated with occupancy dynamics (local probabilities of extinction and colonization) based on multiple seasons or years of detection/nondetection data. It might be useful to extend this dynamic modelling approach to the multispecies case in order to estimate effects of one species on the vital rates of another. Thus, we believe that the methods presented in this paper will be very useful in drawing inferences about species co-occurrence, and we believe that such inferences can be combined with other kinds of studies and analyses in order to investigate mechanisms underlying community dynamics.

This approach to modelling detection/nondetection data for two species has been implemented in program PRESENCE, which may be downloaded freely from http://www.proteus.co.nz/.

Acknowledgements

We would like to thank Evan Cooch and an anonymous referee for their helpful comments on an earlier draft of this paper.

References

Bailey, L.L., Simons, T.R. & Pollock, K.H. (2004) Estimating site occupancy and species detection probability parameters for terrestrial salamanders. *Ecological Applications*, in press.

Brown, J.H. (1995) *Macroecology*. University of Chicago Press, Chicago, IL.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Inference – a Practical Information–Theoretic Approach*, 2nd edn. Springer-Verlag, New York.

Cam, E., Nichols, J.D., Hines, J.E. & Sauer, J.R. (2000) Inferences about nested subset structure when not all species are detected. *Oikos*, **91**, 428–434.

Cole, L.C. (1949) The measurement of interspecific association. *Ecology*, **30**, 411–424.

Connor, E.F. & Simberloff, D. (1979) The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

Connor, E.F. & Simberloff, D. (1983) Interspecific competition and species co-occurrence patterns on islands: null models and the evaluation of evidence. *Oikos*, **41**, 455–465.

Connor, E.F. & Simberloff, D. (1984) Neutral models of species' co-occurrence patterns. *Ecological Communities: Conceptual Issues and Evidence* (eds D.R. Strong Jr, D. Simberloff, L.G.Abele & A.B. Thistle), pp. 316–331. Princeton University Press, Princeton, NJ.

Crowell, K.L. & Pimm, S.L. (1976) Competition and niche shifts of mice introduced onto small islands. *Oikos*, **27**, 251–258.

Dakin, S.F. (1978) The influence of inter-specific competition on the microhabitat distribution of terrestrial lungless salamanders in the Southern Appalachian Mountains. Duke University, Durham, NC.

Diamond, J.M. (1975) Assembly of species communities. *Ecology and Evolution of Communities* (eds M.L. Cody & J.M. Diamond), pp. 342–444. Harvard University Press, Cambridge, MA.

Diamond, J.M. & Gilpin, M.E. (1982) Examination of the 'null' model of Connor and Simberloff for species co-occurrences on islands. *Oecologia*, **52**, 64–74.

Dice, L.R. (1945) Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

Forbes, S.A. (1907) On the local distribution of certain Illinois fishes. An essay in statistical ecology. *Bulletin of the Illinois State Laboratory of Natural History*, **7**, 273–303.

Gilpin, M.E. & Diamond, J.M. (1982) Factors contributing to nonrandomness in species co-occurrences on islands. *Oecologia*, **52**, 75–84.

Gilpin, M.E. & Diamond, J.M. (1984) Are species co-occurrences on islands non-random, and are null hypotheses useful in community ecology. *Ecological Communities: Conceptual Issues and Evidence* (eds D.R. Strong Jr, D. Simberloff, L.G.Abele & A.B. Thistle), pp. 297–315. Princeton University Press, Princeton, NJ.

Gotelli, N.J. (2000) Null model analysis of species co-occurrence patterns. *Ecology*, **81**, 2606–2621.

Gotelli, N.J. & Graves, G.R. (1996) *Null Models in Ecology*. Smithsonian Institution Press, Washington, DC.

Gotelli, N.J. & McCabe, D.J. (2002) Species co-occurrence: a meta-analysis of J.M. Diamond's assembly rules model. *Ecology*, **83**, 2091–2096.

Grover, M.C. (2000) Determinants of salamander distributions along moisture gradients. *Copeia*, **2000**, 156–168.

Hairston, N.G. (1951) Interspecies competition and its probable influence upon the vertical distributions of Appalachian salamanders of the genus Plethodon. *Ecology*, **32**, 266–274.

Hairston, N.G. (1980) The experimental test of an analysis of field distributions – competition in terrestrial salamanders. *Ecology*, **61**, 817–826.

Harvey, P.H., Colwell, R.K., Silverton, J.W. & May, R.M. (1983) Null models in ecology. *Annual Review of Ecology and Systematics*, **14**, 189–211.

Hayek, L.-A.C. (1994) Analysis of amphibian biodiversity data. *Measuring and Monitoring Biology Diversity: Standard Methods for Amphibians* (eds W.R. Heyer, M.A. Donnelly, R.W. McDiarmid, L.-A.C. Hayek & M.S. Foster), pp. 207–273. Smithsonian Institution Press, Washington, DC.

Hubbell, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.

Hyde, E.J. & Simons, T.R. (2001) Sampling plethodontid salamanders: sources of variability. *Journal of Wildlife Management*, **65**, 624–632.

Kelt, D.A., Taper, M.L. & Mesevre, P.L. (1995) Assessing the impact of competition on community assembly: a case study using small mammals. *Ecology*, **76**, 1283–1296.

Kendall, W.L. (1999) Robustness of closed capture–recapture methods to violations of the closure assumption. *Ecology*, **80**, 2517–2525.

Lebreton, J.D., Burnham, K.P., Clobert, J. & Anderson, D.R. (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, **62**, 67–118.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.D. (2003) Estimating site occupancy, colonization and local extinction probabilities when a species is not detected with certainty. *Ecology*, **84**, 2200–2207.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

Manly, B.F.J. (1995) A note on the analysis of species co-occurences. *Ecology*, **76**, 1109–1115.

Marquet, P.A. (2000) Invariants, scaling laws, and ecological complexity. *Science*, **289**, 1487–1488.

McCoy, E.D. & Heck, K.L. Jr (1987) Some observations on the use of taxonomic similarity in large-scale biogeography. *Journal of Biogeography*, **14**, 79–87.

Nichols, J.D., Boulinier, T., Hines, J.E., Pollock, K.H. & Sauer, J.R. (1998) Inference methods for spatial variation in species richness and community composition when not all species are detected. *Conservation Biology*, **12**, 1390–1398.

Patterson, B.D. (1987) The principle of nested subsets and its implications for biological conservation. *Conservation Biology*, **1**, 323–334.

Patterson, B.D. & Atmar, W. (1986) Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biology Journal of the Linnean Society*, **28**, 65–82.

Peres-Neto, P.R., Olden, J.D. & Jackson, D.A. (2001) Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, **93**, 110–120.

Pielou, E.C. (1977) *Mathematical Ecology*. John Wiley, New York, NY.

Rissler, L.J., Barber, A.M. & Wilbur, H.M. (2000) Spatial and behavioral interactions between a native and introduced salamander species. *Behavioral Ecology and Sociobiology*, **48**, 61–68.

Rosenzweig, M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, UK.

Schoener, T.W. (1974) Competition and the form of habitat shift. *Theoretical Population Biology*, **6**, 265–307.

Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Rafael, M.G., Wall, W.A. & Samson, F.B., eds (2002) *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press, Washington, DC.

Verner, J., Morrison, M.L. & Ralph, C.J., eds (1986) *Wildlife 2000: Modeling Habitat Relationships of Terrestrial Vertebrates*. University of Wisconsin press, Madison, WI.

Received 5 August 2003; accepted 14 November 2003