Assessing the Fit of Site-Occupancy Models

Darryl I. MacKenzie and Larissa L. Bailey

Few species are likely to be so evident that they will always be detected at a site when present. Recently a model has been developed that enables estimation of the proportion of area occupied, when the target species is not detected with certainty. Here we apply this modeling approach to data collected on terrestrial salamanders in the *Plethodon glutinosus* complex in the Great Smoky Mountains National Park, USA, and wish to address the question "how accurately does the fitted model represent the data?" The goodness-of-fit of the model needs to be assessed in order to make accurate inferences. This article presents a method where a simple Pearson chi-square statistic is calculated and a parametric bootstrap procedure is used to determine whether the observed statistic is unusually large. We found evidence that the most global model considered provides a poor fit to the data, hence estimated an overdispersion factor to adjust model selection procedures and inflate standard errors. Two hypothetical datasets with known assumption violations are also analyzed, illustrating that the method may be used to guide researchers to making appropriate inferences. The results of a simulation study are presented to provide a broader view of the methods properties.

Key Words: Goodness-of-fit; Model fit; Patch occupancy; *Plethodon glutinosus*; *Plethodon oconluftee*.

1. INTRODUCTION

The probability a site is occupied by a target species may be of interest in many ecological settings. In a wildlife monitoring context, site occupancy may be used as a coarse surrogate for actual abundance as the methods required to collect simple presence/absence-type data are less costly in terms of time and effort than methods used for abundance estimation (MacKenzie et al. 2002), especially when multiple species are to be monitored. For example, the U.S. Geological Survey's Amphibian Research and Monitoring Initiative (ARMI; http://armi.usgs.gov) use the "proportion of area occupied" by a species as their preferred metric for mid-level monitoring efforts, reserving the use of mark-recapture techniques for key index sites. Meta-population studies are also interested in site- (or patch-) occupancy

Darryl I. MacKenzie is a Biometrician/Director for Proteus Wildlife Research Consultants, PO Box 5193, Dunedin, New Zealand (E-mail: Darryl@proteus.co.nz). Larissa L. Bailey is a Research Associate, U.S. Geological Survey, Cooperative Fish and Wildlife Research Unit, Department of Zoology, North Carolina State University, Campus Box 7617, Raleigh, NC 27695-7617, and Patuxent Wildlife Research Center, U.S. Geological Survey, Laurel, MD 20708-4022.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 9, Number 3, Pages 300–318
DOI: 10.1198/108571104X3361

probabilities as they may be used as a state variable in various metapopulation models (e.g., Levins 1969, 1970; Lande 1987, 1988; Hanski 1992, 1994, 1997). "Incidence functions" (e.g., see Diamond 1975; Hanski 1992) are used to express occupancy as a function of site characteristics, such as patch size, which under certain strong assumptions have been used to estimate population dynamic parameters such as colonization and local-extinction probabilities (Hanski 1992, 1994, 1997; Moilanen 1999). A third setting is habitat modeling, where the intent is to relate species presence/absence to characteristics of the sampling locations. Similarly, contrasts between characteristics of occupied/used sites and unoccupied/unused sites can be made using logistic regression, for example, and are sometimes referred to as "resource selection probability functions" (Manly et al. 2002). The species distribution may then be inferred from available habitat information (e.g., Boyce and McDonald 1999). However, few species are likely to be so conspicuous that they will always be detected when present at a site. Failing to account for imperfect detectability will lead to underestimates of the true occupancy probability, with the degree of bias in a naive estimate being dependent upon the probability the species is detected at least once. Furthermore, for many species detection probabilities are likely to vary with local environmental conditions, thus a comparison of naive occupancy probabilities at two (or more) points in time or space is valid only if detection probabilities are exactly equal or explicitly accounted for.

Recently, MacKenzie et al. (2002) developed a model that estimates the probability a site is occupied by a species, despite imperfect detection when the species is present. Their model offers a more flexible framework than previous efforts (Geissler and Fuller 1987; Azuma, Baldwin, and Noon 1990; Bayley and Peterson 2001), enabling relationships between occupancy/detection probabilities and potential model covariates, such as site characteristics and environmental conditions, to be investigated directly. Missing observations (occasions when a site was not surveyed) can also be accommodated by their model. A requirement, however, is that data be collected from a number of sites which are surveyed multiple times to detect the target species. Some investigators may view this as an impediment to their approach, but it is impossible to obtain an unbiased estimate of site occupancy when sites are only visited once without auxiliary information about detectability [e.g., from a previous or independent study, as in Bayley and Peterson (2001), or by making potentially restrictive assumptions].

Bailey, Simons, and Pollock (2004) use the modeling approach developed by MacKenzie et al. (2002) to estimate occupancy and detection probabilities for a suite of terrestrial salamanders in Great Smoky Mountains National Park (GSMNP), USA. The effects of both habitat (e.g., disturbance, vegetation) and seasonal covariates were explored; however, inferences are somewhat tenuous because no method was available to assess model fit. Here we reexamine data collected in GSMNP on the terrestrial salamander complex *Plethodon glutinosus* (includes the species *Plethodon glutinosus* and *Plethodon oconluftee*) with the intent of addressing the question "how accurately does the fitted model represent the data?"

In 1999, count data were collected for these species at 88 sites in the Mt. LeConte USGS Quadrangle of the GSMNP (Hyde and Simons 2001; Bailey et al. 2004). Sites were located

Table 1. Detection (1) and Nondetection (0) Data for Members of the *Plethodon glutinosus* Complex in Great Smoky Mountains National Park at Five of the Sites, Along with Measured Covariates (disturbance prior to Park formation, *D*; elevation $>$ 841m, *E*; predominantly deciduous vegetation, *V*; and stream $<$ 50m, *S*).

	Detected in survey					*Covariates*			
Site	*1*	*2*	*3*	*4*	*5*	*D*	*E*	*V*	*S*
1	1	0	1	0	0	1	1	1	1
2	0	1	0	0	0	1	1	1	0
3	0	0	0	0	0	1	0	1	1
4	1	1	1	0	1	1	0	0	1
5	0	0	1	1	0	1	0	0	0

adjacent to trails and spaced approximately 250 m apart, beginning at a random distance ($>$250m) from each trail head. At each site salamanders were detected using both a natural cover transect (50m $\times$ 3m) and a coverboard transect (consisting of 5 coverboard stations spaced at 10m intervals; see Hyde and Simons (2001) for site and sampling details. Each site was visited on four or five occasions between mid-April and late-June. For this analysis, we combined detection/nondetection data from both transects to investigate the effects of four site-specific characteristics on occupancy and detection probabilities. Dummy variables were used to indicate whether (1) a site had been disturbed by a settlement or logging prior to the Parks formation in 1934 (previously undisturbed $= 0; D$); (2) had an elevation of greater than 841m ($<$841m $= 0; E$); (3) was predominately deciduous vegetation (mixed pine $= 0; V$); and (4) proximity to a stream ($<$50m $= 1$, otherwise $0; S$). For the most global model considered here, both occupancy and detection probabilities were a function of these four covariates (with no interactions between factors), plus detection probability was also allowed to vary with survey occasion. A portion of the dataset is presented in Table 1, and the full dataset may be obtained by contacting the first author.

All modeling exercises should demonstrate that a fitted model adequately describes the observed data, that is, a model should to be assessed for lack-of-fit (McCullagh and Nelder 1989, p. 8; Lebreton, Burnham, Clobert, and Anderson 1992). Only by examining the adequacies of the model fit can researchers demonstrate that the model(s) being considered for the data are realistic, and capture the important features of the system under study. Substantial lack-of-fit in a model(s) may lead to inaccurate inferences, either in terms of bias (point estimates may be too large or too small) or in terms of precision (reported standard errors are too large/small). Clearly, in order to place some degree of faith in the inferences resulting from an analysis of real data, it is critical that the model fit be assessed.

An increasingly popular approach for analyzing ecological data is to fit a suite or candidate set of models to the data, and use a model selection technique such as Akaike's information criterion (AIC), or similar measures, for choosing the "best" model(s). Given the rising popularity of using such techniques in the analysis of ecological data, it is important to realize that they assume that the candidate set contains at least one model that fits the data adequately (Burnham and Anderson 1998, p. 73), and are not a substitute for assessing model fit. The selection of a "best" model(s) does not guarantee the selection of a "good" model.

Motivated by the practical terrestrial salamander system in GSMNP, we develop a method to assess the fit of the MacKenzie et al. (2002) model to observed data. Importantly, the method is flexible enough to incorporate potential model covariates that may vary across sites. The overdispersion parameter $\hat{c}$ can also be estimated so that in situations where even the most global model is found to be a poor fit of the data, the quasi-likelihood version of AIC (QAIC) may be used for model selection (Burnham and Anderson 1998, p. 53) and parameter standard errors may be inflated (McCullagh and Nelder 1989, p. 125).

We begin by briefly reviewing the site-occupancy model proposed by MacKenzie et al. (2002), then present our method that can be used to assess the fit of the model to observed data. These techniques are then used to attempt accurate modeling of occupancy patterns for the terrestrial salamander data. To further demonstrate the utility of our approach to other biological situations, we assess two hypothetical datasets with known assumption violations. Simulation results are presented to illustrate how the test performs more generally under the violation of certain model assumptions.

2. METHODS

2.1 SITE-OCCUPANCY MODEL

MacKenzie et al. (2002) envisage a sampling scheme where N sites are each surveyed T times to establish the presence/absence of the species. Sites are closed to changes in the occupancy state for the duration of the surveying: no new sites become occupied nor are any vacated. On each sampling occasion, the investigators used appropriate methods to detect the species, and there is a chance that the species may go undetected even when present. The resulting sequence of detections/nondetections for site i can be summarized as a detection history $(\mathbf{X}_i)$, and probabilistic arguments may be used to describe the observed stochastic process.

For example, consider the portion of the detection data for species of the *Plethodon glutinosus* complex at five of the sites in GSMNP presented in Table 1. The first site has the history "10100" denoting that the complex was detected at the site during the first and third surveys and not detected otherwise. The probability of observing this outcome may be described as

$$\Pr(\mathbf{X}_1 = 10100) = \psi_1 p_{1,1}(1 - p_{1,2})p_{1,3}(1 - p_{1,4})(1 - p_{1,5}), \tag{2.1}$$

where ψ_1 is the probability site 1 is occupied by the complex, and $p_{1,j}$ is probability of detecting the complex, given presence, in the jth survey of site 1. The history at the third site "00000" would therefore denote that the complex was never detected during the five surveys, which may arise for one of two possible reasons. Either the complex was present but went undetected, or the complex was genuinely absent from the site. The probability of

obtaining this history for site 3 would be

$$\Pr\left(\mathbf{X}_3 = 00000\right) = \psi_3 \prod_{j=1}^{5} \left(1 - p_{3,j}\right) + \left(1 - \psi_3\right). \tag{2.2}$$

The probability of observing each of the N detection histories can be determined in such a manner, and assuming the histories are independent, the model likelihood is then

$$\mathrm{L}\left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N \mid \boldsymbol{\psi}, \mathbf{p}\right) = \prod_{i=1}^{N} \Pr(\mathbf{X}_i). \tag{2.3}$$

The model likelihood can then be maximized with respect to the parameters, either analytically or numerically, to obtain maximum likelihood estimates of the model parameters.

Generally, site-specific occupancy and detection probabilities cannot be estimated as the model would be overparameterized: containing more parameters than could be estimated, hence a basic model might assume that all probabilities are constant across sites. However, using the logistic model

$$\theta_i = \frac{\exp(\mathbf{Y}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{Y}_i\boldsymbol{\beta})}, \tag{2.4}$$

where θ_i is the probability of interest for site i, $\mathbf{Y}_i$ is the vector of covariate values for site i and $\boldsymbol{\beta}$ is the vector of logistic coefficients to be estimated) MacKenzie et al. (2002) suggested site-specific occupancy and detection probabilities may be expressed as a function of measured site-specific covariates. Detection probabilities can also be modeled as a function of covariates that may change with each survey, such as weather conditions.

Missing values (occasions when a site was not surveyed) can be easily accommodated by their model. If an observation is missing for site i at time j, the corresponding detection probability is set to zero. In effect this causes the missing observation to have no contribution to the model likelihood.

In many ways the above occupancy model is analogous to a mark-recapture model with sites being comparable to individuals; however, there are also some important differences. The most notable is that, in the current context, the "all zero" history is observable (sites at which the species was never detected), while it is not in mark-recapture. This has potentially important ramifications if relationships between probabilities and covariates are being modeled. The relationships (and hence estimates of effects) are conditional on the individual being captured at least once in mark-recapture, but are unconditional for site-occupancy studies. Second, in mark-recapture the methodology has not yet been developed enabling capture/detection probabilities to be functions of an individual covariate that varies in time (e.g., weight) as the covariate value is unknown when the individual is not recaptured. Whereas in the site-occupancy model, a site-specific, time-varying covariate—for example, air temperature—can be collected regardless of whether the species was detected at the site during that survey. Finally, the concept of missing values does not generally hold

in mark-recapture. The closest analogy would be that no effort has been made to recapture specific individuals at certain times during the study. Such a situation seems unlikely to occur in practice with the exception of a removal experiment where once an individual has been captured for the first time it is removed from the population.

2.2 ASSESSING MODEL FIT

One classical measure of model fit is to use a Pearson chi-square statistic

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{2.5}$$

where O_i and E_i are the observed and expected numbers of observations for class i, and n is the total number of classes defined by the current model.

X^2 would then be compared to the chi-square distribution with appropriate degrees of freedom to determine whether there was any evidence of poor fit. Traditionally in log-linear modeling, and similarly in mark-recapture, individuals would be partitioned into homogeneous cohorts according to measured covariates or factors, for example, gender or age. The "significance" of a factor could be determined by the change in the value of X^2, or alternatively using the change in deviance, due to collapsing the cohorts and model across the factor of interest.

Therefore, the widespread opinion among most mark-recapture experts is when testing for model fit, all individuals that have a unique combination of covariate values must be treated as a separate cohort. As a result, when a continuous covariate is used in a model that varies between individuals (such as weight) a large number of cohorts may be required, to the extreme point where each cohort contains only one individual. In such situations, the above approach is no longer appropriate because data becomes too sparse. In the following, we show that solely for the purpose of assessing model fit, all individuals can be treated as a single cohort and the Pearson chi-square statistic can be used as a suitable metric.

Let O_h be the number of sites observed to have detection history h, and E_h be the expected number of sites with history h according to the current model. Generally, E_h equates to the sum of the estimated probabilities of observing h

$$E_h = \sum_{i=1}^{N} \Pr(\mathbf{X}_i = h). \tag{2.6}$$

obtained by substituting the estimated parameter values into equations similar to (2.1) and (2.2). These probabilities may be site-specific depending upon the model that has been fit to the data.

Assuming no missing observations, there are 2^T possible detection histories that may be observed. The test statistic for assessing the fit of the model can be calculated as

$$X^2 = \sum_{h=1}^{2^T} \frac{(O_h - E_h)^2}{E_h}. \tag{2.7}$$

Note that because we are pooling across all sites in (2.6), the usual distributional arguments used to justify the chi-square distribution of X^2 are unlikely to hold unless the estimated probabilities are constant across all sites. Furthermore, many of the E_h may be relatively small (< 2) for even moderate values of T (say ≥ 5), again suggesting that X^2 will not have a chi-square distribution. As an alternative, the parametric bootstrap may be used to determine whether the observed value is unusually large.

By using the parametric bootstrap, we assume that the fitted model is correct; hence it is an ideal technique for assessing the model's structure. In this setting, parametric bootstrapping may be implemented as follows:

1. Fit model to the observed data and estimate parameters $\hat{\psi}_i$ and $\hat{p}_{ij}$ (which may be functions of covariates).
2. Calculate the test statistic for the observed data, X^2_{Obs}, using the model fit in Step 1.
3. For each site generate a pseudo-random number (r) between 0 and 1. If $r \leq \hat{\psi}_i$, then the site is occupied, hence generate T further pseudo-random numbers (r_j) between 0 and 1. If $r_j \leq \hat{p}_{ij}$, then the species was "detected" and the corresponding bootstrapped observation is a "1", otherwise "0". If $r > \hat{\psi}_i$, then the site is unoccupied and the bootstrapped observations will all be "0" for that site.
4. Fit a model with the same structure as in Step 1 to the bootstrapped dataset.
5. Calculate the test statistic for the bootstrapped data, X^2_B, using the model fit in Step 4, and store the result.
6. Repeat Steps 3–5 a large number of times to approximate the distribution of the test statistic, given the fitted model is correct.
7. Compare X^2_{Obs} to the values of X^2_B to determine the probability of observing a larger value (the p value).

We note that this general testing procedure has strong similarities with a parametric bootstrap test for model fit in the mark-recapture context, suggested by White, Burnham, and Anderson (2002) and implemented in Program MARK. However, there are two important differences between our method and the test in Program MARK: (1) individuals (or sites in the present setting) with unique combinations of covariates are pooled; and (2) a Pearson's chi-square statistic is used rather than the deviance statistic. We defer further discussion of these differences to later in the article.

Following White et al. (2002), the overdispersion parameter $\hat{c}$ may be estimated using

$$\hat{c} = X^2_{\text{Obs}} / \overline{X}^2_B, \tag{2.8}$$

where $\overline{X}^2_B$ is the average of the test statistics obtained from the parametric bootstrap. The estimate of $\hat{c}$ may then be used as a variance inflation factor, using quasi-likelihood theory, to adjust our model selection procedures and standard errors.

If the model is an adequate description of the data, then $\hat{c}$ should be approximately 1. Values greater than 1 suggest there is more variation in the observed data than expected by the model, with values less than 1 suggesting less variation.

To accommodate missing values in the detection/nondetection data, we suggest that it is necessary to treat sites with different combinations of missing observations as separate cohorts, as each cohort will have a different set of possible detection histories. For example, in our terrestrial salamander data, sites were usually surveyed 5 times, thus there are a total of 32 possible detection histories, assuming no missing observations. However, 12 of the 88 sites were not sampled on the first occasion, therefore these sites have only 16 possible detection histories. If a site was not sampled on the last occasion, it would have a different set of 16 possible histories. An extra level of summation would therefore be required in (2.7) to aggregate the test statistic over cohorts.

3. FIELD DATA COLLECTED ON TERRESTRIAL SALAMANDERS

We wish to investigate the potential effects of the four covariates (disturbance prior to Park formation, D; elevation $>$ 841m, E; predominantly deciduous vegetation, V; and stream $<$ 50m, S) on occupancy and detection probabilities, and in addition we wish to allow detection probability to vary with survey occasion (t). A model selection procedure will be used to rank the models, and our inference will be based upon the most parsimonious models. Considering only models with no interactions between factors, there are $2^4 \times 2^5 = 512$ possible models that could be fit to the data. Usually, with such a large number of possible models we would restrict the set of candidate models before attempting some form of model selection, however, here we do not believe there is a sound biological reason for doing so, therefore we considered all 512 models.

Performing the above test for model fit on the most global model considered, $\psi(D + E + V + S)p(t + D + E + V + S)$, there is some evidence of poor fit ($X^2 = 63.1$, p value $= 0.056$, $\hat{c} = 1.43$) hence QAIC was used for the model selection procedure and standard errors inflated by a factor of $\sqrt{\hat{c}} = 1.20$. By considering the contribution of each observed detection history to the test-statistic, it appears the poor fit is caused by an unusually large number of sites where the salamanders were detected on each sampling occasion. This may be due to an unmeasured site characteristic that also affects detection probabilities, or possibly caused by the species occurring at higher densities at those sites (probability of detecting at least one member of the species could be higher at sites where the species is more abundant). We suggest that this should be kept in mind when drawing conclusions about the affects of the available covariates from this analysis.

Table 2 contains the top ten ranked models upon the basis of QAIC, which account for 53% of the total QAIC model weights (Burnham and Anderson 1998). Based upon these top ten models, we could approximate the model averaged covariate coefficients (on the logistic scale) by adjusting the QAIC model weights such that the ten modified weights sum to 100%. The approximate values ($\pm$ 1 standard error; inflated for poor model fit) are given in the parentheses below.

Common features of these models are that disturbance (3.60 ± 0.92), elevation $(1.6\text{e}-3 \pm 0.6\text{e}-3)$ and vegetation type (-3.02 ± 1.58) seem to be important covari-

Table 2. Top Ten Ranked Models According to QAIC for the Terrestrial Salamander Example. ΔQAIC is the relative difference in QAIC values from the model with the smallest QAIC value; *w* is the QAIC model weight; and *K* is the number of parameters.

Model	*ΔQAIC*	*w*	*K*
$\psi(D+E+V)p\,(E+S)$	0.00	0.13	7
$\psi(D+E+V)p\,(D+S)$	0.53	0.10	7
$\psi(D+E+V)p(D+E+S)$	0.97	0.08	8
$\psi(D+V+S)p\,(E+S)$	2.30	0.04	6
$\psi(D+E+V)p\,(S)$	2.46	0.04	6
$\psi(D+E+V+S)p\,(E+S)$	2.47	0.04	8
$\psi(D+E+V)p\,(E+V+S)$	2.78	0.03	8
$\psi(D+V)p\,(D+S)$	3.13	0.03	6
$\psi(D+E+V)p\,(D+V+S)$	3.17	0.03	8
$\psi(D+E+V+S)p\,(D+S)$	3.22	0.03	8

ates for the occupancy probability, while stream proximity (-0.79 ± 0.38) appears to be the most important covariate for detection probability, with some suggestion that elevation (-0.40 ± 0.44) and/or disturbance (0.53 ± 0.73) may also be important. Interpreting the coefficient values suggests that the odds of a site being occupied were: 36 times larger for a disturbed site; marginally (but significantly) larger for sites above 841m; and 20 times larger for site with mixed pine vegetation. The odds of detecting a member of the *P. glutinosus* complex were: 2.2 times smaller at sites within 50m of a stream; 1.5 times smaller at higher elevations ($>$ 841m); and 1.7 times smaller at undisturbed sites. It should be noted that there is some degree of correlation between the disturbance and elevation covariates, with disturbed sites being more common at lower elevations. This may cause multicollinearity, explaining why there is no clear outcome, with respect to detection probabilities, as to which of these variables is preferential in a model.

4. HYPOTHETICAL EXAMPLES

To explore how well the proposed method for assessing model fit performs when the data contain specific, known violations of the model assumptions, we now consider two hypothetical datasets. The first hypothetical example represents the type of data that might arise from a metapopulation study (for instance) where sites are discrete patches of habitat, and at least one of the models in the candidate set is a good fit for the observed data. We envisage the second hypothetical example as a situation where sites represent study quadrats within a less fragmented landscape (such as a bird point count study) and sites are located such that the detection of the species is no longer independent between sites. In such a case, all of the fitted models describe the observed data poorly; hence model selection procedures and parameter standard errors need to be adjusted. Contemplating hypothetical datasets is useful here as we can introduce known violations of the model assumptions, and evaluate how well our test for model fit identifies these violations.

Table 3. Summary of Model Selection Procedure for the First Hypothetical Example. ΔAIC is the relative difference in AIC values from the model with the smallest AIC value; *w* is the AIC model weight; *K* is the number of parameters; $\overline{\hat{\psi}}$ is the estimated overall occupancy probability; SE($\overline{\hat{\psi}}$) is the associated standard error for the estimate; X^2 is the test statistic for model fit; *p* value is the probability of observing a test statistic $\geq X^2$ based upon 999 parametric bootstraps; and $\hat{c}$ is the estimated overdispersion parameter.

Model	ΔAIC	*w*	*K*	$\overline{\hat{\psi}}$	SE($\overline{\hat{\psi}}$)	X^2	*p value*	$\hat{c}$
$\psi(\cdot)p$(Size)	0.00	0.68	3	0.64	0.11	23.8	0.694	0.81
ψ(Size)p(Size)	1.70	0.29	4	0.70	0.15	23.0	0.719	0.80
$\psi(\cdot)(t+$ Size)	6.66	0.02	7	0.64	0.11	24.5	0.503	0.93
ψ(Size)$p(t+$ Size)	8.37	0.01	8	0.70	0.15	23.7	0.483	0.94
ψ(Size)$p(\cdot)$	53.15	0.00	3	0.32	0.06	59.0	0.002	2.07
$\psi(\cdot)p(\cdot)$	59.58	0.00	2	0.32	0.07	59.4	0.003	2.04
ψ(Size) $p(t)$	60.32	0.00	7	0.32	0.06	55.5	0.006	2.15
$\psi(\cdot)$ $p(t)$	66.75	0.00	6	0.32	0.07	55.7	0.008	2.19

4.1 HYPOTHETICAL DATASET 1

Data were generated such that occupancy and detection probabilities were a function of a site-specific covariate, patch size (say). Both probabilities increased with increasing patch size, representing a mechanism where the species prefers larger habitat patches and within such a patch the species is more detectable, possibly because of higher abundances or densities of individuals.

Fifty patches were each surveyed on five successive days, with the species being detected at least once at 16 of the patches, giving a naive occupancy estimate of 0.32. Table 3 summarizes the results of the model selection procedure for eight candidate models, ranging from the most global model in which occupancy and detection probabilities are functions of patch size, and detection probabilities also have an additive time effect (ψ(Size)$p(t+$Size)), to the most restrictive model where occupancy and detection probabilities are both constant ($\psi(\cdot)p(\cdot)$). The test of model fit for the most global model, ψ(Size)$p(t+$ Size), provides no cause for concern, suggesting that model selection using AIC should be reasonable. Based upon AIC, the most parsimonious model is $\psi(\cdot)p$(Size) suggesting patch occupancy is constant (0.64) and detection probabilities are a function of patch size. Overall there is very strong evidence that patch size is an important covariate for detection probabilities as all models involving such a term have substantially smaller AIC values than those without. In addition, for all models without the patch size covariate for detection probabilities there is very strong evidence of poor model fit. In the face of uncertainty about which model provides the best description of the data, a model averaged estimate of overall patch occupancy is 0.66 with standard error 0.12.

4.2 HYPOTHETICAL DATASET 2

In this example, data were generated such that the detection of the species at pairs of sites was no longer independent. If the species was present at the first site, then the second site was also occupied, and if the species was detected at the first site there was a 90% chance

Table 4. Summary of Model Selection Procedures for the Second Hypothetical Example, With Models Being Ranked According to AIC or QAIC. ΔIC is the relative difference in information criterion values (IC; AIC or QAIC) from the model with the smallest respective IC value; *w* is the model weight based on the respective IC's; *K* is the number of parameters; $\overline{\hat{\psi}}$ is the estimated overall occupancy probability; and SE($\overline{\hat{\psi}}$) is the associated standard error for the estimate.

Ranking method	*Model*	ΔIC	*w*	*K*	$\overline{\hat{\psi}}$	*SE*($\overline{\hat{\psi}}$)
AIC	$\psi(\cdot)p(t)$	0.00	0.64	5	0.66	0.15
	$\psi(\cdot)p(\cdot)$	1.16	0.36	2	0.69	0.16
QAIC	$\psi(\cdot)p(\cdot)$	0.00	0.73	2	0.69	0.22
	$\psi(\cdot)p(t)$	1.95	0.27	5	0.66	0.20

of also detecting the species at the second site. Such a situation may result in a study where sites are placed too close together with respect to the territorial patterns of the target species (e.g., the home range of one or more individuals overlap with more than one monitoring sites), or when two nearby sites are being surveyed simultaneously and observers are likely to record the same event (e.g., bird call) as a detection of the species at both sites.

Fifty sites were each monitored on four occasions for the presence of the species, with the species being detected at least once at 22 sites. No covariates were available for the analysis resulting in only two possible models that could be considered, the most global model $\psi(\cdot)p(t)$ and $\psi(\cdot)p(\cdot)$. From the test described above, there is some evidence of lack-of-fit for the model $\psi(\cdot)p(t)$($X^2 = 17.3$, p value $= 0.079$, $\hat{c} = 1.77$) suggesting QAIC should be used for model selection. Table 4 presents the results of the model selection using both AIC and QAIC. Note that by using QAIC, the $\psi(\cdot)p(\cdot)$ model is now ranked as the most parsimonious and the adjusted standard errors are substantially larger, reflecting that due to the lack of independence, less information has actually been gathered about the model parameters. Using model averaging, based upon AIC an overall estimate of the occupancy probability is 0.67 with standard error 0.16, while based upon QAIC the overall estimate is 0.68 with standard error 0.21. Here we can see that while accounting for lack-of-fit has had little effect on our point estimate of occupancy, the associated uncertainty in the estimate has greatly increased.

5. SIMULATION STUDY

To investigate the properties of the test more generally, a Monte Carlo simulation study was undertaken using a wide variety of scenarios. The effect of up to four factors were considered (N, T, ψ, and p) within four broad scenarios: (a) all sites had common values of ψ and p; (b) the sites were comprised of two groups with different ψ or p values; (c) p varied between sites according to a site-specific covariate; and (d) detection histories for the sites were not independent. For each combination of factors 2,000 simulated sets of data were generated, and the proportion of datasets where the fitted model displayed a "significant" lack of fit (at an α level of 5%), was recorded. When the fitted model is correct, the proportion of simulations with "significant" lack of fit should be equal to the nominal size of the test (5%), while if the fitted model is incorrect, the proportion should be

something larger and is referred to as the power of the test. All tests were performed using 199 bootstrap resamples, which should be an adequate number for simulation studies of this type, but for the analysis of real data one should consider using many more (generally at least 999) depending upon the complexity of the model being analyzed. The average value of $\hat{c}$ was also recorded, and for all scenarios p was kept constant across time.

5.1 SCENARIO A

The following factor levels were investigated.

- $N = 50, 100,$ or 200
- $T = 5$ or 10
- $\psi = 0.5$ or 0.9
- $p = 0.3$ or 0.8.

A site-occupancy model with ψ and p constant across sites ($\psi(\cdot)p(\cdot)$) was fit to each dataset. For this scenario, the $\psi(\cdot)p(\cdot)$ model is correct, hence the size of the test (probability of a Type I error) should be at the nominal level of $\alpha = 5\%$ and $\hat{c}$ should be estimated at close to 1.0, which was confirmed by the simulations.

5.2 SCENARIO B

The same values of N and T as above were used in this scenario, where study sites comprise of two groups of equal size. Each group had either (1) different occupancy probabilities; or (2) different detection probabilities.

5.2.1 Scenario B(i)

Occupancy and detection probabilities were set at the following levels (where subscripts denote groups).

- $(\psi_1, \psi_2) = (0.5, 0.2)$ or $(0.9, 0.7)$
- $p = 0.3$ or 0.8.

Occupancy probabilities were chosen such that the odds of a site being occupied for group 1 was approximately 4 times greater than the odds of a site being occupied for group 2. When the incorrect $\psi(\cdot)p(\cdot)$ model, which ignores group membership, is fit to each simulated dataset, the test has no power to detect the poor fit of the model with the power remaining at the nominal level of 5% and $\hat{c}$ estimated at close to 1.0. Not presented are the results of fitting the correct $\psi(G)p(\cdot)$ model, where occupancy probabilities are group specific, which confirmed the test had the desired properties when the correct model was fit to the data.

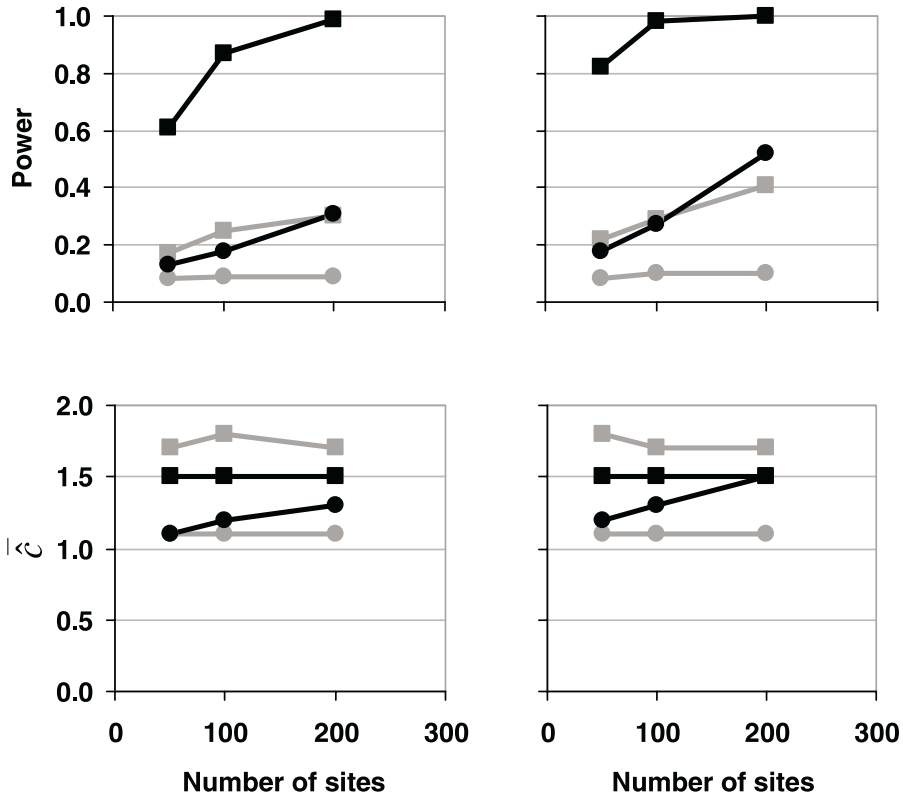

Figure 1. Power of the test for model fit (at $\alpha = 5\%$) and average of the overdisperison parameter estimate ($\bar{\hat{c}}$) from fitting the incorrect $\psi(\cdot)p(\cdot)$ model to 2,000 simulated datasets for Scenario B(ii), where $\psi(\cdot)p(G)$ was the correct model. Gray symbols represent situations where $(p_1, p_2) = (0.3,0.1)$ and black symbols represent situations where $(p_1, p_2) = (0.8,0.5)$, with circles denoting scenarios with $T = 5$ and squares where $T = 10$.

5.2.2 Scenario B(ii)

For this scenario occupancy and detection probabilities were set at the following levels.

— $\psi = 0.5$ or 0.9

— $(p_1, p_2) = (0.3, 0.1)$ or $(0.8, 0.5)$.

Similar to above, detection probabilities were chosen such that the odds of the species being detected was approximately 4 times greater for group 1 than for group 2. Figure 1 summarizes the results for this scenario, where the incorrect $\psi(\cdot)p(\cdot)$ model is fit to each simulated set of data. The test clearly has the ability to detect model lack-of-fit with respect to detection probabilities. Power is low to moderate for most of the factor combinations considered, and high when detection probabilities are higher with 10 surveys of each site. On average, estimates of $\hat{c}$ are generally much greater than 1.0, even when the test has only moderate power. Not presented are the results of fitting the correct $\psi(\cdot)p(G)$ model, where detection probabilities are group specific, which confirmed the test had the desired properties when the correct model was fit to the data.

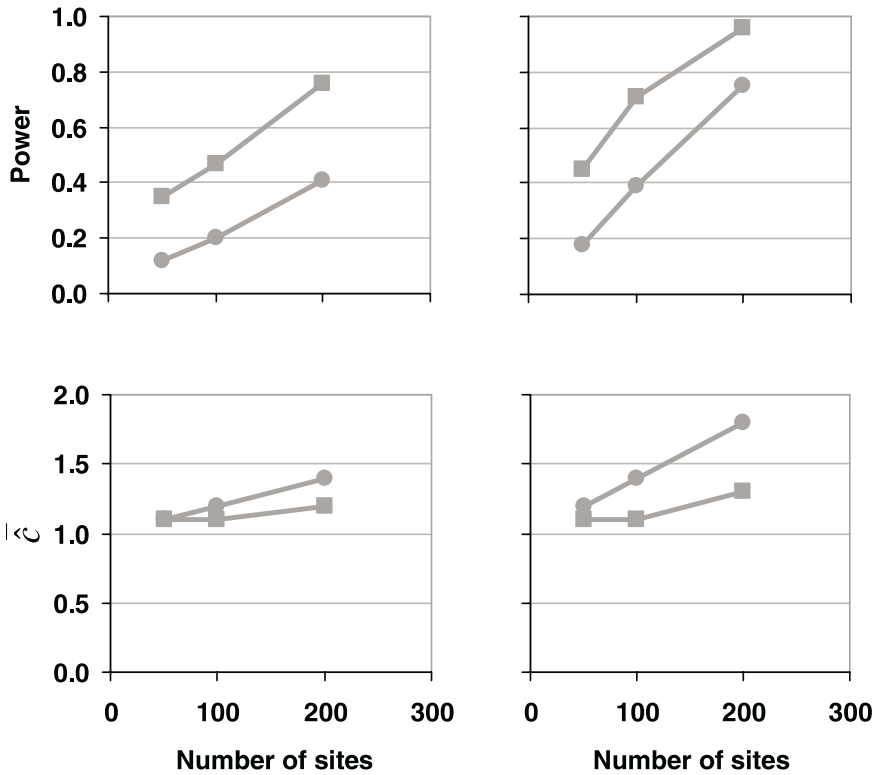

Figure 2. Power of the test for model fit (at $\alpha = 5\%$) and average of the overdisperison parameter estimate ($\bar{\hat{c}}$) from fitting the incorrect $\psi(\cdot)p(\cdot)$ model to 2,000 simulated datasets for Scenario C, where $\psi(\cdot)p(\text{cov})$ was the correct model. Circles denote scenarios with $T = 5$ and squares where $T = 10$.

5.3 SCENARIO C

Factors N, T, and ψ were maintained at the same levels as those used in Scenario A, while p was a function of a site-specific covariate. For each site a random value was generated from the standard normal distribution to represent a measured covariate, and p was calculated using the logistic model with slope and intercept terms of 1.0 and 0.0, respectively. This created a distribution of p's where the central 95% of values lay between 0.12 and 0.88. New random covariate values were generated for each simulated set of data. Figure 2 presents the results for this scenario where the incorrect $\psi(\cdot)p(\cdot)$ model has been fit to the data. The test clearly has moderate to good power to detect that the fitted model is inadequate for the observed data. Interestingly, in most cases, the average estimate of $\hat{c}$ does not vary substantially from 1.0, even when the test has high power. Also, $\hat{c}$ tends to be estimated with a smaller value when 10 rather than 5 surveys are conducted at each site, even though the test tends to have greater power in the former situation. Again, the results of fitting the correct model $\psi(\cdot)p(\text{cov})$, where detection probabilities are a function

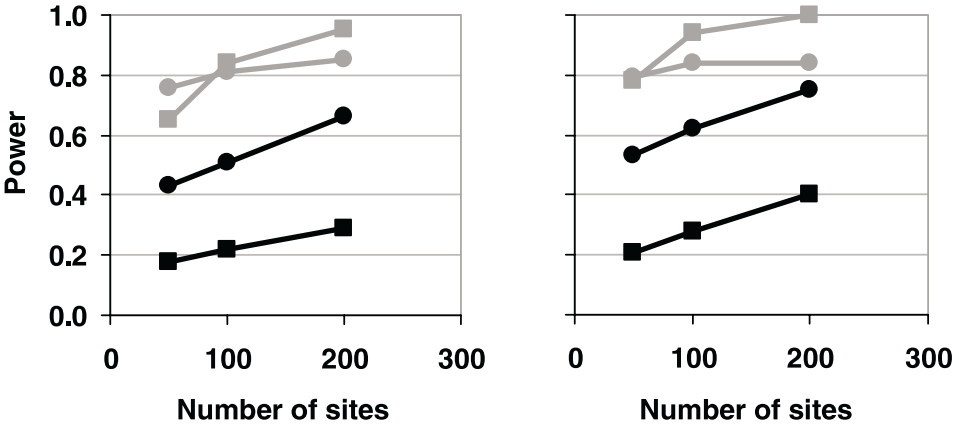

Figure 3. Power of the test for model fit (at $\alpha = 5\%$) from fitting the incorrect $\psi(\cdot)p(\cdot)$ model to 2,000 simulated datasets for Scenario D, where pairs of sites had identical detection histories. Gray symbols represent situations where $p = 0.3$ and black symbols for $p = 0.8$, with circles denoting scenarios with $T = 5$ and squares where $T = 10$. The average of the overdisperison parameter estimate ($\bar{\hat{c}}$) is not presented because it was near the true value of 2.0 in all scenarios.

of a covariate, are not presented here, although they confirm that the test has the desired properties (power $\approx 5\%$ and $\hat{c} \approx 1.0$) when the correct model is fit to the data.

5.4 Scenario D

All factors were maintained at the same levels as in Scenario A; however, here every second site was given a detection history that was an exact copy of the one previous. In effect this creates a dataset with $N/2$ independent detection histories rather than N, and $\hat{c}$ is therefore known to be 2.0. The $\psi(\cdot)p(\cdot)$ is not correct in this scenario as now there is some correlation structure between the "twin" sites. Figure 3 shows that in this situation, the above test has moderate to excellent power to detect this lack of independence between sites. Results for the estimation of $\hat{c}$ are not presented because (on average) it was estimated to be very near 2.0.

6. DISCUSSION

Terrestrial salamanders have gained increasing attention in conservation and management arenas because they possess unique attributes (e.g., longevity, sensitivity to environmental perturbation) which are believed to make them good indicators of forest biodiversity and integrity (Welsh and Droege 2001). These claims, together with apparent worldwide declines in numerous amphibian fauna, have prompted studies to explore the distribution and habitat associations of various salamander species (e.g., Hyde and Simons 1999). Models such as those developed by MacKenzie et al. (2002) are vital to studies of salaman-

der habitat requirements because salamander detection probabilities are imperfect (<1), species-specific, and vary across time, space, and sampling method (Bailey et al. 2004). Here, we enhanced the utility of the MacKenzie model by providing a method that allows researchers to determine if the model(s) being considered is a realistic representation of observed detection/nondetection data.

Our analysis of the *Plethodon glutinosus* complex in GSMNP showed evidence of lack-of-fit for the most global model containing all covariates, suggesting model selection should be evaluated using QAIC values. Our method allows us to estimate a variance inflation factor, $\hat{c}$, by which we adjusted standard errors. Members of the *Plethodon glutinosus* complex had higher occupancy probabilities on previously disturbed sites with mixed pine vegetation, and lower detection probabilities on sites near streams. Interestingly, an increase in overall occupancy or detection probability for the *Plethodon glutinosus* complex may actually represent a degradation of forest habitat as the relationship of these covariates to model parameters indicate that the *Plethodon glutinosus* complex favors drier, disturbed forest habitat. The model set we explored here was much more comprehensive than the one presented by Bailey et al. (2004) (512 models vs. 10 models, respectively), but qualitatively both analyses identified the same covariates important in both occupancy and detection probabilities for the *Plethodon glutinosus* complex. By assessing the fit of our global model and adjusting model weights and ranks accordingly, we are more confident in both the model selection procedure and resulting model-averaged estimates, which form the basis of our biological inferences. Any presentation that focuses on estimates of occupancy should use model-averaged estimates based on the full model set presented in this article. Our analysis revealed that the simple models explored by Bailey et al. (2004) are not as likely as the more complex models in the larger model set.

We have shown that the test of model fit proposed above has some power to detect lack-of-fit in site-occupancy models caused by an incorrectly specified structure for detection probabilities or by detection histories that are not independent. Even though all sites are treated as a single cohort, the test has the desired properties when a model that includes site covariates is fit to the data. This is an important finding as the general consensus of many mark-recapture experts is that separate cohorts are required for each unique combination of covariate values. However, for the scenarios considered in the simulation study, the power of the test tends to be low, with only five surveys per site and 50 sites. This may limit the usefulness of the test for small datasets, although we stress that while the results of the simulation study give some overall view of the tests properties, assessing the fit of a model for a given small set of data may still be informative.

Failure of the test to detect poor model fit caused by occupancy probabilities is not unexpected. When the model is misspecified with respect to detection probabilities, then some sites will have an unusually large (or small) number of species detections. However, for occupancy probabilities there is no such outward indication that the model may be inadequately describing the data. This is similar to the problem of assessing the fit of logistic regression models for binary data, where the actual binary observation conveys little

information regarding model fit. In logistic regression this has led to the development of the Hosmer-Lemeshow Test (Hosmer and Lemeshow 1989, p. 140) which uses the predicted probabilities of a success to classify the observations into k groups. Such an approach could possibly be modified to detect poor model fit with respect to occupancy probabilities.

It is generally recommended (e.g., Lebreton et al. 1992; Burnham and Anderson 1998, p. 54) that the fit of the most global model be assessed first and that any estimate of $\hat{c}$ be based upon those results. In some cases, however, if the number of parameters in the global model is large, then poor precision of the estimates may hamper the tests ability of detect lack-of-fit. In such a case one may also wish to test the fit of a more parsimonious model.

In practice, one would not assess the fit of all models as done in the first hypothetical example, but this was done to illustrate that those models with large ΔAIC values also showed very strong evidence of lack-of-fit. This lack-of-fit is undoubtedly caused by the exclusion of the patch size covariate for detection probabilities, which accounted for the heterogeneity in detectability. Note that by not accounting for this heterogeneity, the occupancy estimates were substantially underestimated.

This raises an important point. If a model is found to fit the data poorly, it may be caused by an inadequate model structure (as in hypothetical example 1) or by a violation of the model assumptions such as independence of the sampling units (as in hypothetical example 2). When lack-of-fit is caused by a lack of independence, parameter estimates remain unbiased but standard errors are too small (McCullagh and Nelder 1989, p. 125); however, a structural lack-of-fit may cause parameter estimates to be badly biased. Expert knowledge of the system under study should be used to determine what may be causing poor model fit.

One of the most topical subjects in mark-recapture research at present is how to assess the fit of models with individual covariates, and because of the basic similarities between mark-recapture and site-occupancy models, it seems likely that the above test should perform admirably with such models. Currently, a similar test for model fit has been implemented in Program MARK (White et al. 2002). The parametric bootstrap is used to determine whether the observed deviance statistic is unusually large with respect to the values obtained from a large number of bootstrap samples, and the variance-inflation factor $\hat{c}$ is calculated as suggested by White et al. (2002; as in Equation (2.8)), although based upon the deviance rather than a Pearson's chi-square statistic. However, there are some important differences. First, the test in Program MARK subscribes to the philosophy that each combination of individual covariate values should be treated as separate cohorts. If groups have been defined in the input data (such as gender or age), then each group is treated as a separate cohort, whereas the test described above would pool across such groupings. Also, the test in Program MARK is not available for models with individual covariates. Second, the value of $\hat{c}$ based upon the deviance statistic and parametric bootstrap has been found to be biased (Gary White, personal communication). Subsequent investigations of mark-recapture data suggest that using the Pearson's chi-square statistic, as we have above, may give unbiased estimates of $\hat{c}$. Although, a large number of bootstraps is required ($> 10{,}000$) to adequately approximate

the distribution of the test statistic if there are a large number of capture occasions (> 10), due to the distribution having a very long right-hand tail in such situations. This is an area of ongoing research.

All analyses in this article were conducted using Program PRESENCE, software specifically developed for analyzing site-occupancy data. Program PRESENCE may be downloaded from http://www.proteus.co.nz. The datasets used here and detailed results from the simulation study may be obtained by contacting the first author.

ACKNOWLEDGMENTS

The authors thank Richard Barker for the conversations that helped form our thinking on this topic; Gary White for his useful comments on an earlier draft of this article; and Erin Hyde for supplying the terrestrial salamander data. The comments from two anonymous referees and the associate editor greatly improved earlier drafts of this article. This research was conducted as part of the contract *Estimating Site Occupancy and Related Parameters* with the USGS, as part of ARMI (http://armi.usgs.gov).

[Received September 2002. Revised November 2003.]

REFERENCES

Azuma, D. L., Baldwin, J. A., and Noon, B. R. (1990), "Estimating the Occupancy of Spotted Owl Habitat Areas by Sampling and Adjusting Bias," USDA Forest Service General technical Report, PSW-124.

Bailey, L. L., Simons, T. R., and Pollock, K. H. (2004), "Estimating Site Occupancy and Species Detection Probability Parameters for Terrestrial Salamanders," *Ecological Applications*, 14, 692–702.

Bayley, P. B., and Peterson, J. T. (2001), "An Approach to Estimate Probability of Presence and Richness of Fish Species," *Transactions of the American Fisheries Society*, 130, 620–633.

Boyce, M. S., and McDonald, L. L. (1999), "Relating Populations to Habitats Using Resource Selection Functions," *Trends in Ecology and Evolution*, 14, 268–272.

Burnham, K. P., and Anderson, D. R. (1998), *Model Selection and Inference*, New York: Springer-Verlag.

Cox, D. R., and Snell, E. J. (1989), *Analysis of Binary Data* (2nd ed.), New York: Chapman and Hall.

Diamond, J. M. (1975), "Assembly of Species Communities," in *Ecology and Evolution of Communities*, eds. M. L. Cody and J. M. Diamond, Cambridge, MA: Harvard University Press, pp. 342–444.

Geissler, P. H., and Fuller, M. R. (1987), "Estimation of the Proportion of area Occupied by an Animal Species," in *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, Alexandria, VA: American Statistical Association, pp. 533–538.

Hanski, I. (1992), "Inferences From Ecological Incidence Functions," *American Naturalist*, 139, 657–662.

——— (1994), "A Practical Model of Metapopulation Dynamics," *Journal of Animal Ecology*, 63, 151–162.

——— (1997), "Metapopulation Dynamics: From Concepts and Observations to Predictive Models," in *Metapopulation Biology: Ecology, Genetics, and Evolution*, eds. I. A. Hanski and M. E. Gilpin, New York: Academic Press, pp. 69–91.

Hosmer, D. W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: Wiley.

Hyde, E. J., and Simons, T. R. (2001), "Sampling Plethodontid Salamanders: Sources of Variability," *Journal of Wildlife Management*, 65, 624–632.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992), "Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach with Case Studies," *Ecological Monographs*, 62, 67–118.

Lande, R. (1987), "Extinction Thresholds in Demographic Models of Territorial Populations," *American Naturalist*, 130, 624–635.

——— (1988), "Demographic Models of the Northern Spotted Owl (*Strix occidentalis caurina*)," *Oecologia*, 75, 601–607.

Levins, R. (1969), "Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control," *Bulletin of the Entomological Society of America*, 15, 237–240.

——— (1970), "Extinction," in *Some Mathematical Questions in Biology* (Vol. II), ed. M. Gerstenhaber, Providence, RI: American Mathematical Society, pp. 77–107 .

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A. (2002), "Estimating Site Occupancy Rates When Detection Probabilities are Less Than One," *Ecology*, 83, 2248–2245.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L., and Erickson, W. P. (2002), *Resource Selection by Animals: Statistical Design and Analysis for Field Studies* (2nd ed.), Dordrecht: Kluwer Academic Publishers.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, (2nd ed.), New York: Chapman and Hall.

Moilanen, A. (1999), "Patch Occupancy Models of Metapopulation Dynamics: Efficient Parameter Estimation Using Implicit Statistical Inference," *Ecology*, 80, 1031–1043.

Seber, G. A. F. (1982), *The Estimation of Animal Abundance and Related Parameters* (2nd ed.), London: Charles Griffin and Company.

Welsh, H. H., and Droege, S. (2001), "A Case for Using Plethodontid Salamanders for Monitoring Biodiversity and Ecosystem Integrity of North American Forests," *Conservation Biology*, 15, 558–569.

White, G. C., Burnham, K. P., and Anderson, D. R. (2002), "Advanced Features of Program MARK" in *Integrating People and Wildlife for a Sustainable Future: Proceedings of the Second International Wildlife Management Congress*, ed. R. Fields, Bethesda, MD: The Wildlife Society.