Modern Methods of Estimating Biodiversity from Presence-Absence Surveys

Robert M. Dorazio[1], U.S. Geological Survey and University of Florida, Department of Statistics, Gainesville, FL 32611, USA.
Nicholas J. Gotelli, Department of Biology, University of Vermont, Burlington, VT 05405, USA
Aaron M. Ellison, Harvard Forest, Harvard University, Petersham, MA 01366, USA

1 Introduction

Communities of species are often sampled using so-called "presence-absence" surveys, wherein the apparent presence or absence of each species is recorded. Whereas counts of individuals can be used to estimate species abundances, apparent presence-absence data are often easier to obtain in surveys of multiple species. Presence-absence surveys also may be more accurate than abundance surveys, particularly in communities that contain highly mobile species.

A problem with presence-absence data is that observations are usually contaminated by zeros that stem from errors in detection of a species. That is, true zeros, which are associated with the absence of a species, cannot be distinguished from false zeros, which occur when species are present in the vicinity of sampling but not detected. Therefore, it is more accurate to describe apparent presence-absence data as detections and non-detections, but this terminology is seldom used in ecology.

Estimates of biodiversity and other community-level attributes can be dramatically affected by errors in detection of each species, particularly since the magnitude of these detection errors generally varies among species (Boulinier et al. 1998). For example, bias in estimates of biodiversity arising from errors in detection is especially pronounced in communities that contain a preponderance of rare or difficult-to-detect species. To eliminate this source of bias, probabilities of species occurrence and detection must be estimated simultaneously using a statistical model of the presence-absence data. Such models require presence-absence surveys to be replicated at some – but not necessarily all – of the locations selected for sampling. Replicate surveys can be obtained using a variety of sampling protocols, including repeated visits to each sample location by a single observer, independent surveys by different observers, or even spatial replicates obtained by placing clusters of quadrats or transects within a sample location. Information in the replicated surveys is crucial because it allows species occurrences to be estimated without bias by using a model-based specification of the observation process, which accounts for the errors in detection that are manifest as false zeros.

Several statistical models have been developed for the analysis of replicated, presence-absence data. Each of these models includes parameters for a community's incidence matrix (Gotelli 2000, Colwell et al. 2004), which contains the binary occupancy state (presence or absence) of each species at each sample location. The incidence matrix is only partially observed owing to species- and location-specific errors in detection; however, the incidence matrix can be estimated by fitting these models to the replicated, presence-absence data. Therefore, any function of the incidence matrix – including species richness, alpha diversity, and beta diversity (Magurran 2004)– also can be estimated using these models.

Models for estimating species richness – and other measures of biodiversity – from replicated, presence-absence data were first developed by Dorazio and Royle (2005) and Dorazio et al. (2006). By including spatial covariates of species occurrence and detection probabilities in these models, Kéry and Royle (2009) and Royle and Dorazio (2008) estimated the spatial distribution (or map) of species richness of birds in Switzerland. Similarly, Zipkin et al. (2010) showed that this approach can be used to quantify and assess the effects of conservation or management actions on species richness

[1]Correspondence: University of Florida, Department of Statistics, 429 McCarty Hall C, Gainesville, FL 32611-0339, USA. E-mail: bdorazio@usgs.gov

and other community-level characteristics. More recently, statistical models have been developed to estimate *changes* in communities from a temporal sequence of replicated, presence-absence data. In these models the dynamics of species occurrences are specified using temporal variation in covariates of occurrence (Kéry et al. 2009*a*) or using first-order Markov processes (Russell et al. 2009, Dorazio et al. 2010, Walls et al. 2011), wherein temporal differences in occurrence probabilities are specified as functions of species- and location-specific colonization and extinction probabilities. The latter class of models, which includes the former, is extremely versatile and may be used to confront alternative theories of metacommunity dynamics (Leibold et al. 2004, Holyoak and Mata 2008) with data or to estimate changes in biodiversity. For example, Dorazio et al. (2010) estimated regional levels of biodiversity of butterflies in Switzerland using a model that accounted for seasonal changes in species composition associated with differences in phenology of flight patterns among species. Russell et al. (2009) estimated the effects of prescribed forest fire on the composition and size of an avian community in Washington.

In the present paper we analyze a set of replicated, presence-absence data that previously was analyzed using statistical models that did not account for errors in detection of each species (Gotelli and Ellison 2002). Our objective is to illustrate the inferential benefits of using modern methods to analyze these data. In the analysis we model occurrence probabilities in assemblages of ant species as a function of large-scale, geographic covariates (latitude, elevation) and small-scale, site covariates (habitat area, vegetation composition, light availability). We fit several models, each identified by a specific combination of covariates, to assess the relative contribution of these potential sources of variation in species occurrence and to estimate the effect of these contributions on geographic differences in ant species richness and other measures of biodiversity. We also provide the data and source code used in our analysis to allow comparisons between our results and those obtained using alternative methods of analysis.

2 Study Area and Sampling Methods

2.1 Ant sampling

The data in our analysis were obtained by sampling assemblages of ant species found in New England bogs and forests. The initial motivation for sampling was to determine the extent of the distribution of the apparent bog-specialist, *Myrmica lobifrons*, in Massachusetts and Vermont. Bogs are not commonly searched for ants, but in 1997 we had identified *M. lobifrons* as a primary component of the diet of the carnivorous pitcher plant, *Sarracenia purpurea*, at Hawley Bog in western Massachuestts. This was the first record for *M. lobifrons* in Massachusetts. At the time the taxonomic status of this species was being re-evaluated (Francoeur 1997), and it was largely unknown in the lower (contiguous) 48 states of the United States. In addition to our interest in *M. lobifrons*, we also wanted to explore whether bogs harbored a distinctive ant fauna or whether the ant faunas of bogs were simply a subset of the ant species found in the surrounding forests. Thus, at each of the sites selected for sampling, we surveyed ants in the target bog and in the upland forest adjacent to the bog (Gotelli and Ellison 2002).

At each of 22 sample sites, we established two 8×8 m sampling grids, each containing 25 evenly spaced pitfall traps. One sampling grid was located in the center of the bog; the other was located within intact forest 50-500 m away from the edge of the bog. Each pitfall trap consisted of a 180-ml plastic cup (95 mm in diameter) that was filled with 20 ml of dilute soapy water. Traps were buried so that the upper lip of each trap was flush with the bog or forest-soil surface, and left in place for 48 hours during dry weather. At the end of the 48 hours, trap contents were collected, immediately fixed with 95% ethanol, and returned to the laboratory where all ants were removed and identified to species. Traps were sampled twice in the summer of 1999, and the time between each sampling period was 6 weeks (42 days); therefore, we consider the two sampling periods as early- and late-summer replicates.

Locations of traps were flagged so that pitfall traps were placed at identical locations during the two sampling periods.

2.2 Measurement of site covariates

The geographic location (latitude (LAT) and longitude (LON)) and elevation (ELEV, meters above sea level) of each bog and forest sample site was determined using a Trimble Global Positioning System (GPS). At each forest sample site we also estimated available light levels beneath the canopy using hemispherical canopy photographs, which were taken on overcast days between 10:00 AM and and 2:00 PM at 1 m above ground level with an 8 mm fish-eye lens on a Nikon F-3 camera. Leaf area index (LAI, dimensionless) was determined from the subsequently digitized photographs using HemiView software (Delta-T, Cambridge, UK). Because there was no canopy over the bog, the LAI of each bog was assigned a value of zero.

To compute a global site factor (GSF, total solar radiation) for each forest sample site (Rich et al. 1993), we summed weighted values of direct site factor (DSF, total direct beam solar radiation) and indirect site factor (total diffuse solar radiation). GSF values are expressed as a percentage of total possible solar radiation (i.e., above the canopy) during the growing season (April through October), corrected for latitude and solar track. The GSF of each bog was assigned a value of one.

Digital aerial photographs were obtained for each sampled bog from state mapping authorities, or, when digital photographs were unavailable (five sites), photographic prints (from USGS-EROS) were scanned and digitized. Aerial photographs were used to construct a set of data layers (Arc-View GIS 3.2) from which bog area (AREA) was calculated. The area of the surrounding forests was not measured, as the forest was generally continuous for at least several km^2 around each bog.

3 Statistical Analysis

We analyzed the captures of ant species observed at our sample sites using a modification of the multi-species model of occurrence and detection that includes site-specific covariates (Royle and Dorazio 2008, Kéry and Royle 2009). This modification allows a finite set of candidate models to be specified and fit to the data simultaneously such that prior beliefs in each model's utility can be updated (using Bayes' rule) to compute the posterior probability of each model. The resulting set of posterior model probabilities can be used to select a single ("best") model for inference or to estimate scientifically relevant quantities while averaging over the posterior uncertainty of the models (Draper 1995).

To compare our results with previous analyses (Gotelli and Ellison 2002), we analyzed the data observed in bogs and forests separately. These two habitats are sufficiently distinct that differences in species occurrence – and possibly capture rates – are expected a priori. Furthermore, the potential covariates of occurrence differ between the two habitats, adding another reason to analyze the bog and forest data separately.

3.1 Hierarchical model of species occurrence and capture

We summarize here the assumptions made in our analysis of the ant captures. Let $y_{ik} \in \{0, 1, \ldots, J_k\}$ denote the number of pitfall traps located at site k that contained the ith of n distinct species of ants captured in the entire sample of $R = 22$ sites. At each site 25 pitfall traps were deployed during each of 2 sampling periods (early- and late-season replicates); therefore, the total number of replicate observations per site was constant ($J_k = 50$). While constant replication among sites simplifies implementation of the model, it is not required. However, it *is* essential that $J_k > 1$ for some (ideally all) sample sites because information from within-site replicates allows both occurrence and detection

probabilities to be estimated for each species. In the absence of this replication these two parameters are confounded.

The observed data form an $n \times R$ matrix $\boldsymbol{Y}_{obs}$ of pitfall trap frequencies, so that rows are associated with distinct species and columns are associated with distinct sample sites. Note that n, the number of distinct ant species observed among all R sample sites, is a random outcome. In the analysis we want to estimate the total number of species N that are present and vulnerable to capture. Although N is unknown, we know that $n \leq N$, i.e., we know that the number of species observed in the samples provides a lower bound for an estimate of N.

To estimate N, we use a technique called parameter-expanded data augmentation (Dorazio et al. 2006, Royle and Dorazio 2011), wherein rows of all-zero trap frequencies are added to the observed data $\boldsymbol{Y}_{obs}$ and the model for the observed data is appropriately expanded to analyze the augmented data matrix $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{0})$. The technical details underlying this technique are described by Royle and Dorazio (2008, 2011), so we won't repeat them here. Briefly, however, the idea is to embed the unobserved, all-zero trap frequencies of the $N - n$ species in the community within a larger data set of fixed, but known size (say, M species, where $M > N$) for the purpose of simplifying the analysis. The conventional model for the community of N species is necessarily modified so that each of the $M - n$ rows of augmented data can be estimated as either belonging to the community of N species (and containing sampling zeros) or not (and containing structural zeros). In particular, we add a vector of parameters $\boldsymbol{w} = (w_1, \ldots, w_M)$ to the model to indicate whether each species is a member of the community ($w = 1$) or not ($w = 0$). The elements of $\boldsymbol{w}$ are assumed to be independentally and identically distributed (iid) as follows:

$$w_i \overset{iid}{\sim} \text{Bernoulli}(\Omega)$$

where the parameter Ω denotes the probability that a species in the augmented data set is a member of the community of N species that are present and vulnerable to capture. Note that the community's species richness N is not a formal parameter of the model. Instead, N is a derived parameter to be computed as a function of $\boldsymbol{w}$ as follows: $N = \sum_{i=1}^{M} w_i$. Therefore, estimation of Ω and $\boldsymbol{w}$ is essentially equivalent to estimation of N (Royle and Dorazio 2011).

The incidence matrix of the community (Gotelli 2000, Colwell et al. 2004) is a parameter of the model that is embedded in an $M \times R$ matrix of parameters $\boldsymbol{Z}$, whose elements indicate the presence ($z = 1$) or absence ($z = 0$) of species i at sample site k. Although $\boldsymbol{Z}$ is treated as a random variable of the model, each element associated with species that are not members of the community is equal to zero because z_{ik} is defined conditional on the value of w_i as follows:

$$z_{ik}|w_i \sim \text{Bernoulli}(w_i \psi_{ik}) \tag{1}$$

where ψ_{ik} denotes the probability that species i is present at sample site k. Thus, if species i is not a member of the community, then $w_i = 0$ and $\Pr(z_{ik} = 0|w_i = 0) = 1$; otherwise, $w_i = 1$ and $\Pr(z_{ik} = 1|w_i = 1) = \psi_{ik}$. For purposes of computing estimates of community-level characteristics, $\boldsymbol{Z}$ may be treated as the incidence matrix itself because the $M - N$ rows associated with species not in the community contain only zeros and make no contribution to the estimates.

The matrix of augmented data $\boldsymbol{Y}$ and the parameters $\boldsymbol{Z}$ and $\boldsymbol{w}$ may be conceptualized as characteristics of a supercommunity of M species (Table 1). This supercommunity includes N species that are members of the community vulnerable to sampling and $M - N$ other species that are added to simplify the analysis. The parameters $\boldsymbol{Z}$ and $\boldsymbol{w}$ are paramount in terms of estimating measures of biodiversity. We have shown already that estimates of $\boldsymbol{w}$ are used to compute estimates of species richness N (a measure of gamma diversity). Similarly, $\boldsymbol{Z}$ may be used to estimate measures of alpha diversity, beta diversity, and other community-level characteristics. For example, summing the columns of $\boldsymbol{Z}$ yields the number of species present at each sample site (alpha diversity). Similarly, different columns of

$\boldsymbol{Z}$ may be compared to express differences in species composition among sites (beta diversity). For example, the Jaccard index, a commonly used measure of beta diversity (Anderson et al. 2011), is easily computed from $\boldsymbol{Z}$. The Jaccard index requires the number of species from two distinct sites, say k and l, that occur at both sites. Off-diagonal elements of the $R \times R$ matrix $\boldsymbol{Z}'\boldsymbol{Z}$ contain the numbers of species shared between different sites. Therefore, the proportion of all species present at two sites, say k and l, that are common to both sites is

$$J_{kl} = \frac{\boldsymbol{z}_k'\boldsymbol{z}_l}{\boldsymbol{z}_k'\mathbf{1} + \boldsymbol{z}_l'\mathbf{1} - \boldsymbol{z}_k'\boldsymbol{z}_l}$$

where $\mathbf{1}$ denotes a $M \times 1$ vector of ones, and $\boldsymbol{z}_k$ and $\boldsymbol{z}_l$ denote the kth and lth columns of $\boldsymbol{Z}$. Note that J_{kl} is a measure of the similarity in species present at sites k and l; its complement, $1 - J_{kl}$, corresponds to the dissimilarity – or beta diversity – between sites.

In Section 4 we provide estimates of gamma diversity, alpha diversity, and beta diversity in our analyses of the ant data sets. In these analyses we assume that the community of ants contains a maximum of $M = 75$ species in the forest habitat and a maximum of $M = 25$ species in the bog habitat. The lower maximum is based on five years of collecting ants in New England bogs that yielded only 21 distinct species (Ellison and Gotelli, *personal observations*). The total number of ant species in all of New England is somewhere between 130 and 140 (Ellison et al. 2012); however, many of these species are field or grassland species, and six species, which are not indigenous to New England, are restricted mainly to warm indoors. By excluding these species and those found only in bogs, we obtain the upper limit for the number of ant species in the forest habitat.

3.1.1 Modeling species occurrence probabilities

Equation 1 implies that each element of the incidence matrix is assumed to be independent given ψ_{ik}, the probability of occurrence of species i at sample site k. Let $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})$ denote the observed value of p covariates at site k. We assume that each of these covariates potentially affects the species-specific probability of occurrence at site k. Naturally, the effects of these covariates may differ among species, so their contributions are modeled on the logit-scale as follows:

$$\text{logit}(\psi_{ik}) = b_{0i} + \delta_1 b_{1i} x_{1k} + \cdots + \delta_p b_{pi} x_{pk} \tag{2}$$

where b_{0i} denotes a logit-scale, intercept parameter for species i and b_{li} denotes the effect of covariate x_l on the probability of occurrence of species i ($l = 1, \ldots, p$). If each covariate is centered and scaled to have zero mean and unit variance, b_{0i} denotes the logit-scale probability of occurrence of species i at the average value of the covariates. This scaling of covariates also improves the stability of calculations involved in estimating $\boldsymbol{b}_i = (b_{0i}, b_{1i}, \ldots, b_{p_i})$.

The additional parameter $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ in Eq. 2 is used to specify whether each covariate is ($\delta = 1$) or is not ($\delta = 0$) included in the model. Specifically, we assume

$$\delta_l \overset{iid}{\sim} \text{Bernoulli}(0.5)$$

which implies an equal prior probability (0.5^p) for each of the 2^p distinct values of $\boldsymbol{\delta}$. This approach, originally developed by Kuo and Mallick (1998), allows several regression models to be considered simultaneously and yields the posterior distribution of $\boldsymbol{\delta}$. After all models have been considered (as described in Section 3.2), the posterior probability $\Pr(\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{X})$ of each model (vis a vis, each distinct value of $\boldsymbol{\delta}$) can be computed. In our analyses the model with the highest posterior probability is used to compute estimates of species occurrence and biodiversity.

3.1.2 Modeling species captures

We assume a relatively simple model of the pitfall trap frequencies y_{ik}, owing to the simplicity of our sampling design. Specifically, we assume that if ants of species i are present at site k (i.e., $z_{ik} = 1$), their probability of capture p_{ik} is the same in each of the J_k replicated traps. This assumption implies the following binomial model of the pitfall trap frequencies:

$$y_{ik}|z_{ik} \sim \text{Binomial}(J_k, z_{ik}p_{ik})$$

where p_{ik} denotes the conditional probability of capture of species i at site k (given $z_{ik} = 1$). Note that if species i is absent at site k, then $\Pr(y_{ik} = 0|z_{ik} = 0) = 1$. In other words, if a species is absent at sample site k, then none of the J_k pitfall traps will contain ants of that species under our modeling assumptions.

None of the covariates observed in our samples is thought to be informative of ant capture probabilities; therefore, rather than using a logistic-regression formulation of p_{ik} (as in Eq. 2), we assume that the logit-scale probability of capture of each species is constant:

$$\text{logit}(p_{ik}) = a_{0i}$$

at each of the R sample sites.

3.1.3 Modeling heterogeneity among species

In order to estimate the occurrences of species not observed in any of our traps, a modeling assumption is needed to specify a relationship among all species-specific probabilities of occurrence and detection. Therefore, we assume that the ant species in each community are ecologically similar in the sense that these species are likely to respond similarly, but not identically, to changes in their environment or habitat, to changes in resources, or to changes in predation. The assumption of ecological similarity seems reasonable for the species we sampled owing to their overlapping diets, habitats, and life history characteristics. As a point of emphasis, we would *not* assume ecological similarity if our assemblage had included species of tigers and mice! The idea of ecological similarity has been used previously to analyze assemblages of songbird, butterfly, and amphibian species (Dorazio et al. 2006, Kéry et al. 2009*b*, Walls et al. 2011); however, this idea is not universally applicable. For example, if the occurrence of one species depends on the presence or absence of another species (as might occur between a predator and prey species or between strongly competing species), then ecological similarity would not be a reasonable assumption. In this case a model must be formulated to specify the pattern of co-occurrence that arises from interspecific interactions (MacKenzie et al. 2004, Waddle et al. 2010). The formulation of statistical models for inferring interspecific interactions in communities of species is an important and developing area of research (Dorazio et al. 2010).

In assemblages of ecologically similar species, it seems reasonable to use distributional assumptions to model unobserved sources of heterogeneity in probabilities of species occurrence and detection. For example, occurrence probabilities may be low for some species (the rare ones) and high for others, but all species are related in the sense that they belong to a larger community of ecologically similar species. By modeling the heterogeneity among species in this way, the data observed for any individual species influence the parameter estimates of every other species in the community. In other words, inferences about an individual species do not depend solely on the observations of that species because the inferences borrow strength from the observations of other species. A practical manifestation of this multispecies approach is that the estimate of a parameter (e.g., occurrence probability) of a single species reflects a compromise between the estimate that would be obtained by analyzing the data from each species separately and the average value of that parameter among all species in the community. In the statistical literature this phenomenon is called "shrinkage" (Gelman et al. 2004) because each

species-specific estimate is shrunk in the direction of the estimated average parameter value. Of course, the amount of shrinkage depends on the relative amount of information about the parameter in the observations of each species versus the information about the mean value of that parameter. An important benefit of shrinkage is that it allows parameters to be estimated for a species that is detected with such low frequency that its parameters could otherwise not be estimated. Such species are often the rarest members of the community, and it is crucial that these species be included in the analysis to ensure that estimates of biodiversity are accurate.

In the present analysis we use a normal distribution

$$\begin{bmatrix} b_{0i} \\ a_{0i} \end{bmatrix} \overset{iid}{\sim} \text{Normal}\left(\begin{bmatrix} \beta_0 \\ \alpha_0 \end{bmatrix}, \begin{bmatrix} \sigma_{b_0}^2 & \rho\,\sigma_{b_0}\sigma_{a_0} \\ \rho\,\sigma_{b_0}\sigma_{a_0} & \sigma_{a_0}^2 \end{bmatrix} \right), \tag{3}$$

to specify the variation in occurrence and detection probabilities among ant species. The parameters σ_{b_0} and σ_{a_0} denote the magnitude of this variation, and ρ parameterizes the extent to which species occurrence and detection probabilities are correlated.

We also use the normal distribution to specify variation among the species-specific effects of covariates on occurrence. Specifically, we assume $b_{li} \overset{iid}{\sim} \text{Normal}(\beta_l, \sigma_{b_l}^2)$ (for $l = 1, \ldots, p$), so that the effects of different covariates are assumed to be mutually independent and uncorrelated.

3.2 Parameter estimation

The hierarchical model described in Section 3.1 would be impossible to fit using classical methods owing to the high-dimensional and analytically intractable integrations involved in evaluating the marginal likelihood function. We therefore adopted a Bayesian approach to inference and used Markov chain Monte Carlo methods (Robert and Casella 2004) to fit the model. In the appendix (Section 7) we describe our choice of prior distributions for the model's parameters. We also provide the data and the computer code that was used to calculate the joint posterior distribution of the model's parameters. All parameter estimates and credible intervals are based on this distribution.

4 Results

4.1 Effects of covariates on species occurrence

The posterior model probabilities calculated in our analysis of forest and bog data sets are only mildly sensitive to our choice of priors for the logit-scale parameters of the model (Table 2). Recall that these parameters are of primary interest in assessing the relative contributions of geographic- and site-level covariates. Regardless of the prior distribution used (Uniform or Jeffreys' (see appendix)), the model with highest probability includes all four covariates (LAT, LAI, GSF, ELEV) in the analysis of data observed at forest sample sites and a single covariate (ELEV) in the analysis of data observed at bog sample sites. However, the model without any covariates has nearly equal probability to the favored model of the bog data, and the combined probability of these two models far exceeds the probabilities of all other models. These results suggest that occurrence probabilities of ant species found in the bog habitat are not strongly influenced by the LAT or AREA covariates, either alone or in combination with other covariates.

Each of the four covariates used to model species occurrences in the forest habitat has an average, negative effect on occurrence probabilities. Estimates of β_l and 95% credible intervals are as follows: LAT, -0.717 $(-1.217, -0.257)$; LAI, -0.850 $(-1.302, -0.440)$; GSF, -0.494, $(-0.916, -0.098)$; ELEV, -0.662 $(-1.014, -0.339)$. However, as illustrated in Figure 1, there is considerable variation among species in the magnitude of these effects . Similarly, the estimated occurrence probabilities of ants in the bog habitat decrease with ELEV ($\hat{\beta}_1 = -0.500$ $(-1.019, -0.098)$), and there is considerable variation among species ($\hat{\sigma}_{b_1} = 0.320$ $(0.014, 1.000)$) in the magnitude of ELEV effects.

4.2 Estimates of biodiversity

Our pitfall trap surveys revealed $n = 34$ distinct species of ants at the forest sample sites and $n = 19$ species at the bog sample sites. The estimated species richness of ants found in the forest habitat ($\hat{N} = 43$ (95% interval = $(37, 70)$) is nearly twice the estimated richness of ants in the bog habitat ($\hat{N} = 25$ (95% interval = $(21, 25)$); however, the estimate of forest ant richness is relatively imprecise and the estimate of bog ant richness is strongly influenced by the upper bound ($M = 25$ species).

The numbers of species found in forest and bog communities are perhaps better compared using estimates of species richness at the sample sites. These measures of alpha diversity are plotted against each site's elevation in Figure 2, which also includes the number of ant species actually captured. The estimated richness at sites in the forest habitat usually exceeds that at sites in the bog habitat when the effects of elevation on species occurrences are taken into account. Note also that a site's estimated species richness can be much higher than the numbers of species captured because capture probabilities are much lower than one for most species (Tables 3 and 4).

Site-specific estimates of beta diversity between bog and forest communities of ants are relatively high, ranging from 0.71 to 1.0 (Figure 3). These estimates also generally exceed the beta diversities between ants from different sites within each habitat (Figure 4), adding further support for the hypothesis that composition of ant species differs greatly between forest and bog habitats.

5 Discussion

5.1 Analysis of ant species

It is interesting to compare the results of our analyses with the results reported by Gotelli and Ellison (2002), who analyzed the same data but did not account for errors in detection of species. Gotelli and Ellison (2002) used linear regression models to estimate associations between the number of observed species (which was referred to as "species density") and environmental covariates. For bog ants Gotelli and Ellison (2002) reported a significant association between species density and latitude ($P = 0.041$) and a marginally significant association between species density and vegetation structure (as measured by the first principal-component score; $P = 0.081$). Collectively, these two variables accounted for about 30% of the variation in species density. In the present analysis of the bog data, the best fitting model included the effect of a single covariate (ELEV) on ant species occurrence probabilities, though a model without any covariates was a close second (Table 2). In the analysis of forest ants Gotelli and Ellison (2002) reported significant positive associations between species density and the first two principal components of vegetation structure, and they reported significant negative associations between species density and four other covariates (LAT, LAI, GSF, and ELEV). Collectively, these six regressors accounted for 83% of the variation in species density. In the present analysis of forest data, the best-fitting model included the effects of four covariates (LAT, LAI, GSF, and ELEV), and the estimated effects of these covariates were all significantly negative, which agrees qualitatively with the regression results of Gotelli and Ellison (2002), though principal components of vegetation structure were not included in the present analysis.

In comparing the results obtained using the linear regression model (Gotelli and Ellison 2002) and the hierarchical model of species occurrences and captures, we note that while both models revealed the same set of negative predictors of ant occurrence in forest habitat (Figure 1), the regression model's associations between species density of bog ants and two predictors (latitude and vegetation structure) are not supported by the hierarchical model. Part of the difference in these results may be attributed to the fact that slightly different data sets were used in the two analyses. Species detected using tuna baits, hand collections, and leaf-litter sorting (in forest habitats) were included in the regression analysis, whereas only species captured in pitfall traps were used in the present analysis. However, these differences in data are relatively minor because the alternative sampling methods used by Gotelli and

Ellison (2002) added only a few rare species to their analysis. Instead, we believe the different results stem primarily from differences in the underlying assumptions of these two models. The regression model assumes (1) that the effects of environmental covariates are identical for each species and are linearly related to species density and (2) that residual errors in species density are normally distributed and do not distinguish between measurement errors and heterogeneity among species in their response to covariates. In contrast, the hierarchical model assumes that the effects of environmental covariates differ among species (Figure 1) and that occurrence probabilities and capture probabilities can be estimated separately for each species (Tables 3 and 4) owing to the replicated sampling at each site.

The estimated probabilities of occurrence and capture of each species are of great interest in themselves and highlight differences in species compositions between ants found in bog and forest habitats. For example, the forest species with the highest occurrence probability was *Aphaenogaster rudis* (species complex) ($\hat{\psi} = 0.779$). This species is taxonomically unresolved and currently includes a complex of poorly differentiated species across its geographic range (Umphrey 1996). *Myrmica punctiventris* had the second highest occurrence probability ($\hat{\psi} = 0.739$). Both of these species are characteristic of forest ant assemblages in New England. *A. rudis* (species complex) was never captured in bogs and the occurrence probability of *M. punctiventris* in bogs was only 0.150, almost a fivefold difference between the two habitats.

In bogs the highest occurrence probabilities were estimated for the bog specialist, *Myrmica lobifrons* ($\hat{\psi} = 0.916$), and for *Dolichoderus pustulatus* ($\hat{\psi} = 0.701$), a generalist species that sometimes builds carton nests in dead leaves of the carnivorous pitcher plant *Sarracenia purpurea* (A. Ellison and N. Gotelli, personal communication). Occurrence probabilities of these species in forests were only 0.299 (*M. lobifrons*) and 0.042 (*D. pustulatus*), a 3- to 16-fold difference. These pronounced differences in the occurrence probabilities of the most common species in each habitat suggest that the two habitats support distinctive ant assemblages, a conclusion also supported by the relatively high estimates of beta diversity between habitats (Figure 3).

Although occurrence and capture probabilities were positively correlated among species (Figure 5), a few rare forest species (*Formica subintegra* and *Formica subsericea*) had relatively high capture probabilities. In the forest habitat the two species with the highest capture probabilities were *F. subsericea* ($\hat{p} = 0.248$) and *Myrmica punctiventris* ($\hat{p} = 0.248$). In bogs these species had capture probabilities of only 0.014 (*F. subsericea*) and 0.006 (*M. punctiventris*), a 17- to 41-fold difference. The two species with the highest capture probabilities in the bog habitat were *Myrmica lobifrons* ($\hat{p} = 0.559$), the bog specialist, and *Formica subaenescens* ($\hat{p} = 0.353$). In the forest habitat these species had capture probabilities of only 0.056 (*M. lobifrons*) and 0.051 (*F. subaenescens*), a 7- to 9-fold difference.

The estimated probabilities of occurrence of most species in the forest habitat decreased with latitude (Figure 1), which is consistent with previous regression analyses of species density (Gotelli and Ellison 2002, figure 1). However, the occurrence probabilities of three species (*Camponotus herculeanus*, *Lasius alienus*, and *Myrmica detritinodis*) significantly increased with latitude. Two of these species, *C. herculeanus* and *M. detritinodis*, are boreal, cold-climate specialists (Ellison et al. 2012), whereas *L. alienus* has a more widespread distribution. Under climate change scenarios of increasing temperatures at high latitudes, species whose occurrence probabilities currently increase with latitude might disappear from New England as their ranges shift northward; other species in the assemblage might show no change in distribution, or might increase in occurrence.

To summarize the comparisons between our results and those reported by Gotelli and Ellison (2002), we note that within-site replication of presence-absence surveys allowed us to estimate species-specific probabilities of capture and occurrence and species-specific effects of environmental covariates. These results represent a considerable advance over traditional regression analyses of observed species density. Using a hierarchical approach to model building, we were able to infer sources of variation in measures of biodiversity – such as the effect of elevation on site-specific species richness (Figure 2) and the

effect of habitat on beta diversity (Figure 3) – and to determine how these community-level patterns were related to differences in occurrence of individual species. Although many macroecological data sets collected at large spatial scales do not include within-site replicates, regional studies often use replicated sampling grids of traps or baits (Gotelli et al. 2011) that are ideal for the kind of analysis we have described. We therefore recommend that within-site replication be used in presence-absence surveys of communities, particularly when surveys are undertaken to assess levels of biodiversity.

5.2 Benefits and challenges of hierarchical modeling

Our analysis of the ant data illustrates the benefits of using hierarchical models to estimate measures of biodiversity and other community-level characteristics. By adopting a hierarchical approach to model building, an analyst actually specifies two models: one for the ecologically relevant parameters (or state variables) that are usually of primary interest but are not directly observable, and a second model for the observed data, which are related to the ecological parameters but are influenced also by sampling methods and sampling errors. This dichotomy between models of ecological parameters and models of data is extremely useful and has been exploited to solve a variety of inference problems in ecology (Royle and Dorazio 2008).

In our hierarchical model of replicated, presence-absence surveys, the parameter of primary ecological interest is the community's incidence matrix. This matrix is only partially observable because a species may be present at a sample location but not observed in the surveys. We use a binomial sampling model to specify the probability of detection (or capture) of each species and thereby to account for detection errors in the observed data. In this way estimates of the community's incidence matrix are automatically adjusted for the imperfect detectability of each species.

In our approach, measures of biodiversity are estimated indirectly as functions of the estimated incidence matrix of the community. Thus, species richness and measures of alpha or beta diversity depend on a set of model-based estimates of species- and site-specific occurrences. This approach differs considerably with classes of statistical models wherein species richness is treated as a single random variable – usually a discrete random variable – that represents the aggregate contribution of all species in the community. This "top-down" view of a community may yield incorrect inferences if heterogeneity in detectability exists among species or if the effects of environmental covariates on occurrence differ among species, as illustrated in our analysis of the ant data.

The inferential benefits of using hierarchical models to estimate measures of biodiversity are not free. As described earlier, the price to be paid for the ability to estimate probabilities of species occurrence and species detection is replication of presence-absence surveys within sample locations. In our opinion the improved understanding acquired in modeling the community at the level of individual species and the versatility attained by having accurate estimates of a community's incidence matrix far outweigh the cost of additional sampling. That said, there are other, perhaps less obvious, costs associated with these hierarchical models. Specifically, estimates of species richness and other community-level parameters may be sensitive to the underlying assumptions of these models, and these assumptions can be difficult to test using standard goodness-of-fit procedures. For example, the choice of distributions for modeling heterogeneity among species or sites may exert some influence on estimates of species richness. We assumed a bivariate normal distribution for the distribution of logit-scale, mean probabilities of occurrence and detection, but other distributions – even multimodal distributions – also might be useful. In single-species models of replicated, presence-absence surveys, estimates of occurrence are sensitive to the distribution used to specify heterogeneity in detection probabilities among sample sites (Royle 2006, Dorazio 2007); therefore, similar sensitivity can be expected in multispecies models, though this aspect of model adequacy has not been rigorously explored.

Another assumption of our model that is difficult to test is absence of false-positive errors in detection. In other words, if a species is detected (or captured), we assume that its identify is known

with certainty. However, in surveys of avian or amphibian communities where species are detected by their vocalizations, misidentifications of species can and do occur (Simons et al. 2007, McClintock et al. 2010*a*,*b*). These misidentifications are even more common in circumstances where surveys are conducted by volunteers whose identification skills are highly variable (Genet and Sargent 2003). If ignored, false-positive errors in detection induce a positive bias in estimates of species occurrence because species are incorrectly "detected" at sites where they are absent. While it is possible to construct statistical models of presence-absence data that include parameters for both false-positive and false-negative detection errors (Royle and Link 2006), these models are prone to identifiability problems. To reduce these problems, Royle and Link (2006) recommended that the model's parameters be constrained to ensure that estimates of misclassification probabilities are lower than estimates of detection probabilities. This constraint, though sensible, does not provide a solution when the probabilities of misclassification and detection are nearly equal (Royle and Link 2006, McClintock et al. 2010*b*). The development of statistical models of species occurrence that include both false-positive and false-negative errors in detection, as well as unobserved sources of heterogeneity in both occurrence and detection probabilities, is an active area of research owing to the difficulties associated with aural detection methods.

The conceptual framework described in this paper is broadly applicable in ecological research and in assessments of biodiversity. Hierarchical, statistical models of multispecies, presence-absence data can be used to estimate current levels of biodiversity, as illustrated in our analysis of the ant data, or to assess changes (e.g., trends) in communities over time (Kéry et al. 2009*a*, Russell et al. 2009, Dorazio et al. 2010, Walls et al. 2011). The models of community change are especially relevant in ecological research because they provide an analytical framework wherein data may be used to confront alternative theories of metacommunity dynamics (Leibold et al. 2004, Holyoak and Mata 2008). Although a few classes of statistical models have been developed to infer patterns of co-occurrence among species (MacKenzie et al. 2004, Waddle et al. 2010), models for estimating the dynamics of interacting species (e.g., competitors or predators) from replicated, presence-absence data have not yet been formulated. Such models obviously represent an important area of future research.

6 Acknowledgments

Collection of the original ant dataset was supported by NSF grants 98-05722 and 98-08504 to AME and NJG, respectively, and by contract MAHERSW99-17 from the Massachusetts Natural Heritage and Endangered Species Program to AME. Additional support for AME's and NJG's research on the distribution of ants in response to climatic change is provided by the U.S. Department of Energy through award DE-FG02-08ER64510. The statistical modeling and analysis was conducted as a part of the Binary Matrices Working Group at the National Institute for Mathematical and Biological Synthesis, sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

7 Appendix: Technical Details

7.1 Model fitting and software

Here we describe methods for fitting our hierarchical model using the Markov chain Monte Carlo (MCMC) algorithms implemented in the software package, JAGS (Just Another Gibbs Sampler), which is freely available at the following web site: `http://mcmc-jags.sourceforge.net`. This software

allows the user to specify a model in terms of its underlying assumptions, which include the distributions assumed for the observed data and the model's parameters. The latter distributions include priors, which are needed, of course, to conduct a Bayesian analysis of the data (see below). Part of the reason for the popularity of JAGS is that it allows the model to be specified and fitted without requiring the user to derive the MCMC sampling algorithms used in computing the joint posterior. That said, naive use of JAGS may yield undesirable results, and some experience is needed to ensure the accuracy of the results.

We prefer to execute JAGS remotely from R (R Development Core Team 2004) using functions defined in the R package RJAGS (`http://mcmc-jags.sourceforge.net`). In this way R is used to organize the data, to provide inputs to JAGS, and to receive outputs (results) from JAGS. However, the model's distributional assumptions must be specified in the native language of JAGS. The data files and source code needed to fit our model are provided below.

In our analysis of each data set, the posterior was calculated by initializing each of 5 Markov chains independently and running each chain for a total of 250,000 draws. The first 50,000 draws of each chain were discarded as "burn-in", and every 50th draw in the remainder of each chain was retained to form the posterior sample. Based on Gelman-Rubin diagnostics of the model's parameters (Brooks and Gelman 1998), this approach appeared to produce Markov chains that had converged to their stationary distribution. Therefore, we used the posterior sample of 20,000 draws to compute estimates of the model's parameters and 95% credible intervals.

7.2 Prior distributions

Our prior distributions were chosen to specify prior indifference in the magnitude of each parameter. For example, we assumed a Uniform(0,1) prior for Ω, the probability that a species in the augmented data set is a member of the N species vulnerable to capture. It is easily shown that this prior induces a discrete uniform prior on N, which assigns equal probability to each integer in the set $\{0, 1, \ldots, M\}$. We also used the uniform distribution for the correlation parameter ρ; specifically, we assumed a Uniform(-1,1) prior for ρ, thereby favoring no particular value of ρ in the analysis.

Each of the heterogeneity parameters $(\sigma_{a_0}, \sigma_{b_0}, \sigma_{b_l})$ was assigned a half-Cauchy prior (Gelman 2006) with unit scale parameter, which has probability density function

$$f(\sigma) = 2/[\pi(1 + \sigma^2)].$$

Gelman (2006) showed that this prior avoids problems that can occur when alternative "noninformative" priors are used (including the nearly improper, Inverse-Gamma(ϵ, ϵ) family).

Currently, there is no consenus choice of noninformative prior for the logit-scale parameters of logistic-regression models (Marin and Robert 2007, Gelman et al. 2008). To specify a prior for the logit-scale parameters of our model $(\alpha_0, \beta_0, \beta_l)$, we used an approach described by Gelman et al. (2008). Recall that the covariates of our model are centered and scaled to have mean zero and unit variance; therefore, we seek a prior that assigns low probabilities to large effects on the logit scale. The reason for this choice is that a difference of 5 on the logit scale corresponds to a difference of nearly 0.5 on the probability scale. Because shifts in the value of a standardized covariate seldom, in practice, correspond to outcome probabilities that change from 0.01 to 0.99, the prior of a logit-scale parameter should assign low probabilities to values outside the interval (-5,5). The family of zero-centered t-distributions with parameters σ (scale) and ν (degrees of freedom) can be used to specify priors with this goal in mind. For example, Gelman et al. (2008) recommended a t-distribution with $\sigma = 2.5$ and $\nu = 1$ as a "robust" alternative to a t-family approximation of Jeffreys' prior ($\sigma = 2.5$ and $\nu = 7$). However, when the logit-scale parameter (say, θ) is transformed to the probability scale ($p = 1/(1 + \exp(-\theta))$), both of these priors assign high probabilities in the vicinity of $p = 0$ and $p = 1$, which is not always desirable. As an alternative, we used a t-distribution with $\sigma = 1.566$ and $\nu = 7.763$ as a prior for each logit-scale

parameter of our model. This distribution approximates a Uniform$(0, 1)$ prior for p and assigns low probabilities to values outside the interval (-5,5).

Given our choice of priors and the amount of information in the ant data, parameter estimates based on a single model are unlikely to be sensitive to the priors used in our analysis. However, it is well known that the distributional form of a noninformative prior can exert considerable influence on posterior model probabilities (Kass and Raftery 1995, Kadane and Lazar 2004). Because these probabilities are used to select a single model for inference, we examined the sensitivity of the model probabilities to our choice of priors. In particular, we considered a t-family approximation of Jeffreys' prior ($\sigma = 2.482$ and $\nu = 5.100$) as an alternative for the logit-scale parameters of our model. As described earlier, Jeffreys' prior is commonly used in Bayesian analyses of logistic-regression models.

7.3 Data files and source code

The following files were used to fit our hierarchical model to the ant data sets.

`AntDetections1999.csv` – species- and site-specific capture frequencies of ants in bog and forest habitats (format is comma-delimited with first row as header)

`GetDetectionMatrix.R` – R code for reading capture frequencies of ants from data file and returning a species- and site-specific matrix of capture frequencies of ants collected in a specified habitat ('Forest' or 'Bog')

`GetSiteCovariates.R` – R code for reading covariates from data file

`MultiSpeciesOccModelAve.R` – R and JAGS code for defining and fitting the hierarchical model

`SiteCovariates.csv` – site-specific values of covariates (format is comma-delimited with first row as header)

8 References

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comitka, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C., and Swenson, N. G. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology Letters **14**: 19–28.

Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. 1998. Estimating species richness: the importance of heterogeneity in species detectability. Ecology **79**: 1018–1028.

Brooks, S. P., and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics **7**: 434–455.

Colwell, R. K., Mao, C. X., and Chang, J. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology **85**: 2717–2727.

Dorazio, R. M. 2007. On the choice of statistical models for estimating occurrence and extinction from animal surveys. Ecology **88**: 2773–2782.

Dorazio, R. M., Kéry, M., Royle, J. A., and Plattner, M. 2010. Models for inference in dynamic metacommunity systems. Ecology **91**: 2466–2475.

Dorazio, R. M., and Royle, J. A. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. Journal of the American Statistical Association **100**: 389–398.

Dorazio, R. M., Royle, J. A., Söderström, B., and Glimskär, A. 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. Ecology **87**: 842–854.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society, Series B **57**: 45–97.

Ellison, A. M., Gotelli, N. J., Alpert, G. D., and Farnsworth, E. J. 2012. A field guide to the ants of New England. Yale University Press, New Haven, Connecticut.

Francoeur, A. 1997. Ants (Hymenoptera: Formicidae) of the Yukon. In Insects of the Yukon, edited by H. V. Danks and J. A. Downes. Survey of Canada (Terrestrial Arthropods), Ottawa, Ontario, pp. 901–910.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). Bayesian Analysis **1**: 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2004. Bayesian data analysis, second edition. Chapman and Hall, Boca Raton.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. 2008. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics **2**: 1360–1383.

Genet, K. S., and Sargent, L. G. 2003. Evaluation of methods and data quality from a volunteer-based amphibian call survey. Wildlife Society Bulletin **31**: 703–714.

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

Gotelli, N. J., and Ellison, A. M. 2002. Biogeography at a regional scale: determinants of ant species density in New England bogs and forests. Ecology **83**: 1604–1609.

Gotelli, N. J., Ellison, A. M., Dunn, R. R., and Sanders, N. J. 2011. Counting ants (Hymenoptera: Formicidae): biodiversity sampling and statistical analysis for myrmecologists. Myrmecological News **15**: 13–19.

Holyoak, M., and Mata, T. M. 2008. Metacommunities. In Encyclopedia of Ecology, edited by S. E. Jorgensen and B. D. Fath. Academic Press, Oxford, pp. 2313–2318.

Kadane, J. B., and Lazar, N. A. 2004. Methods and criteria for model selection. Journal of the American Statistical Association **99**: 279–290.

Kass, R. E., and Raftery, A. E. 1995. Bayes factors. Journal of the American Statistical Association **90**: 773–795.

Kéry, M., Dorazio, R. M., Soldaat, L., van Strien, A., Zuiderwijk, A., and Royle, J. A. 2009*a*. Trend estimation in populations with imperfect detection. Journal of Applied Ecology **46**: 1163–1172.

Kéry, M., and Royle, J. A. 2009. Inference about species richness and community structure using species-specific occupancy models in the national Swiss breeding bird survey MHB. In Modeling demographic processes in marked populations, series: environmental and ecological statistics, volume 3, edited by D. L. Thomson, E. G. Cooch, and M. J. Conroy. Springer, Berlin, pp. 639–656.

Kéry, M., Royle, J. A., Plattner, M., and Dorazio, R. M. 2009*b*. Species richness and occupancy estimation in communities subject to temporary emigration. Ecology **90**: 1279–1290.

Kuo, L., and Mallick, B. 1998. Variable selection for regression models. Sankhya **60B**: 65–81.

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., Holt, R. D., Shurin, J. B., Law, R., Tilman, D., Loreau, M., and Gonzalez, A. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters **7**: 601–613.

MacKenzie, D. I., Bailey, L. L., and Nichols, J. D. 2004. Investigating species co-occurrence patterns when species are detected imperfectly. Journal of Animal Ecology **73**: 546–555.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford.

Marin, J.-M., and Robert, C. P. 2007. Bayesian Core. Springer, New York.

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simon, T. R. 2010*a*. Experimental investigation of observation error in anuran call surveys. Journal of Wildlife Management **74**: 1882–1893.

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simon, T. R. 2010*b*. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. Ecology **91**: 2446–2454.

R Development Core Team. 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rich, P. M., Clark, D. B., Clark, D. A., and Oberbauer, S. F. 1993. Long-term study of solar radiation regimes in a tropical wet forest using quantum sensors and hemispherical photography. Agricultural and Forest Meteorology **65**: 107–127.

Robert, C. P., and Casella, G. 2004. Monte Carlo Statistical Methods (second edition). Springer-Verlag, New York.

Royle, J. A. 2006. Site occupancy models with heterogeneous detection probabilities. Biometrics **62**: 97–102.

Royle, J. A., and Dorazio, R. M. 2008. Hierarchical modeling and inference in ecology. Academic Press, Amsterdam.

Royle, J. A., and Dorazio, R. M. 2011. Parameter-expanded data augmentation for Bayesian analysis of capture-recapture models. Journal of Ornithology **123**: in press.

Royle, J. A., and Link, W. A. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology **87**: 835–841.

Russell, R. E., Royle, J. A., Saab, V. A., Lehmkuhl, J. F., Block, W. M., and Sauer, J. R. 2009. Modeling the effects of environmental disturbance on wildlife communities: avian responses to prescribed fire. Ecological Applications **19**: 1253–1263.

Simons, T. R., Alldredge, M. W., Pollock, K. H., and Wettroth, J. M. 2007. Experimental analysis of the auditory detection process on avian point counts. Auk **124**: 986–999.

Umphrey, G. 1996. Morphometric discrimination among sibling species in the *fulva - rudis - texana* complex of the ant genus *Aphaenogaster* (Hymenoptera: Formicidae). Canadian Journal of Zoology **74**: 528–559.

Waddle, J. H., Dorazio, R. M., Walls, S. C., Rice, K. G., Beauchamp, J., Schuman, M. J., and Mazzotti, F. J. 2010. A new parameterization for estimating co-occurrence of interacting species. Ecological Applications **20**: 1467–1475.

Walls, S. C., Waddle, J. H., and Dorazio, R. M. 2011. Estimating occupancy dynamics in an anuran assemblage from Louisiana, USA. Journal of Wildlife Management **75**: in press.

Zipkin, E., Royle, J. A., Dawson, D. K., and Bates, S. 2010. Multi-species occurrence models to evaluate the effects of conservation and management actions. Biological Conservation **143**: 479–484.

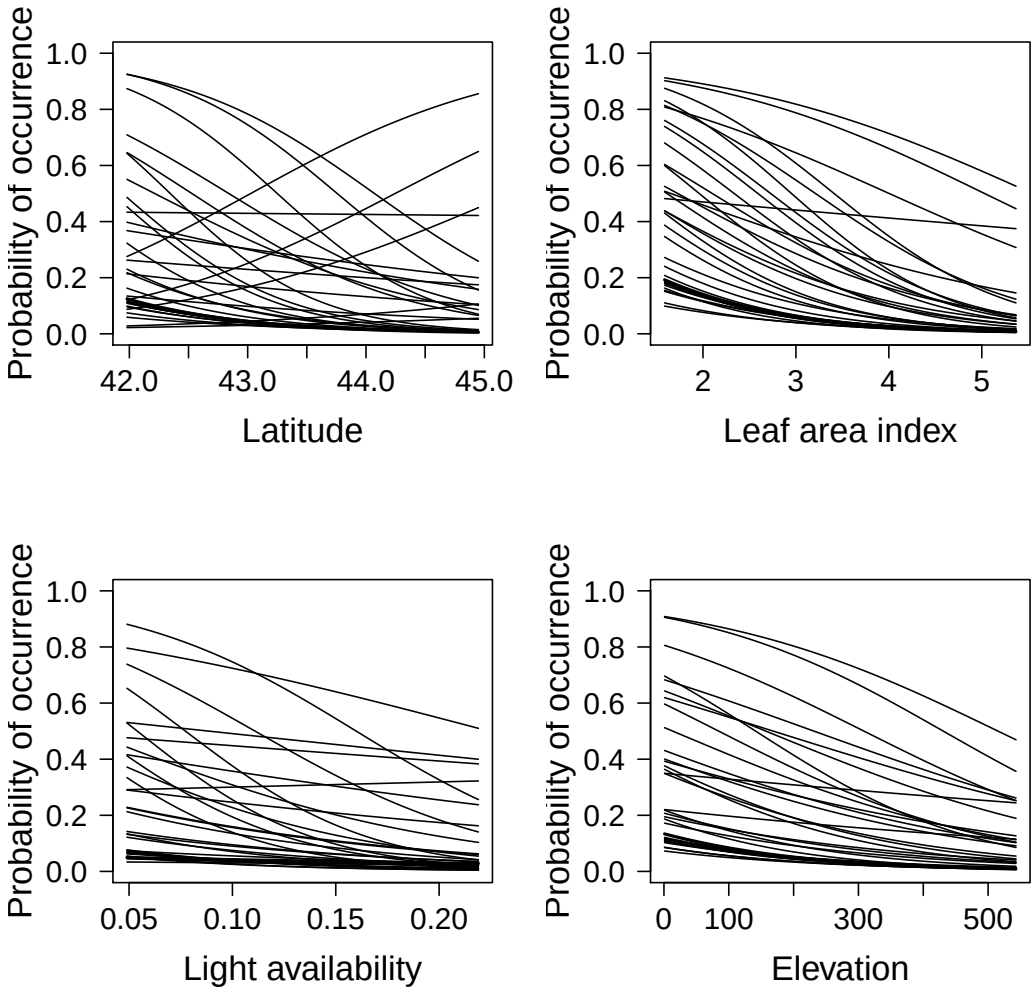

Figure 1: Estimated effects of covariates on occurrence probabilities of ant species in forest habitat.

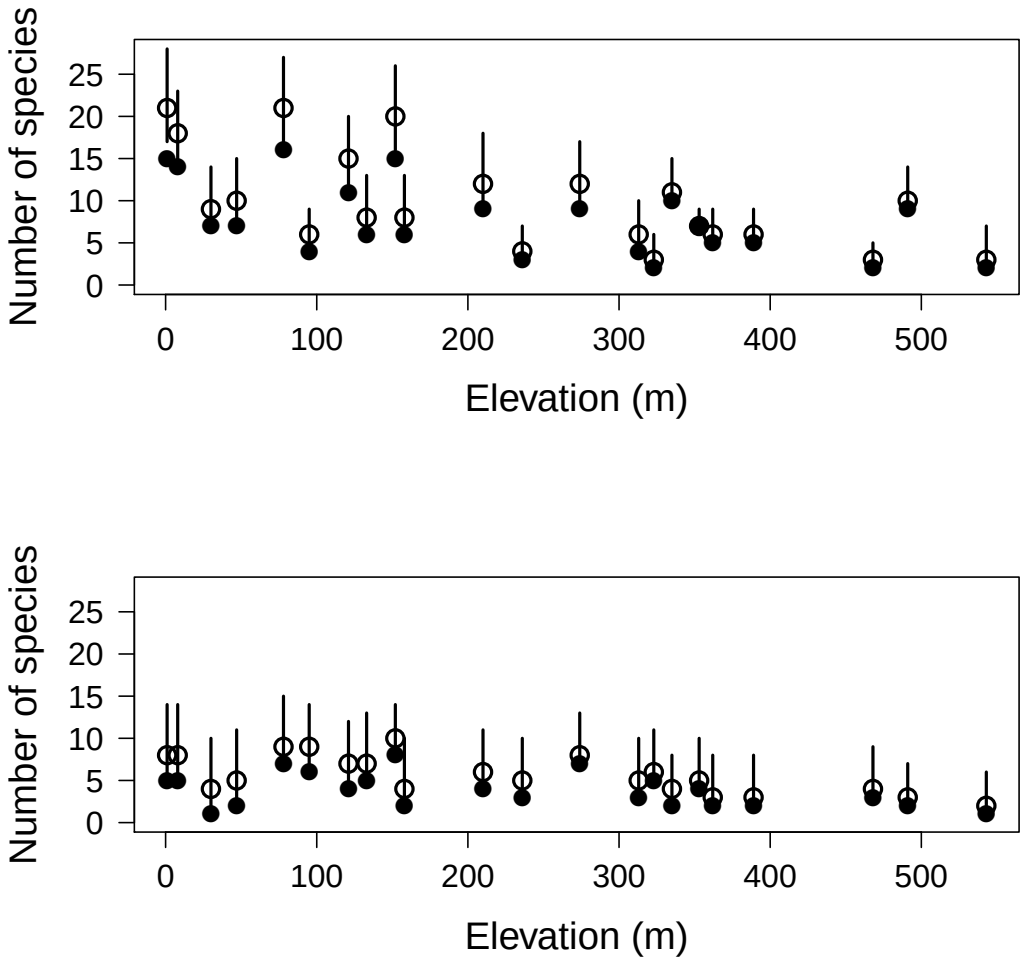

Figure 2: Estimates of site-specific species richness (open circles with 95% credible intervals) for ants in forest habitat (upper panel) and bog habitat (lower panel) versus elevation. Number of species captured at each site (closed circles) is shown for comparison.

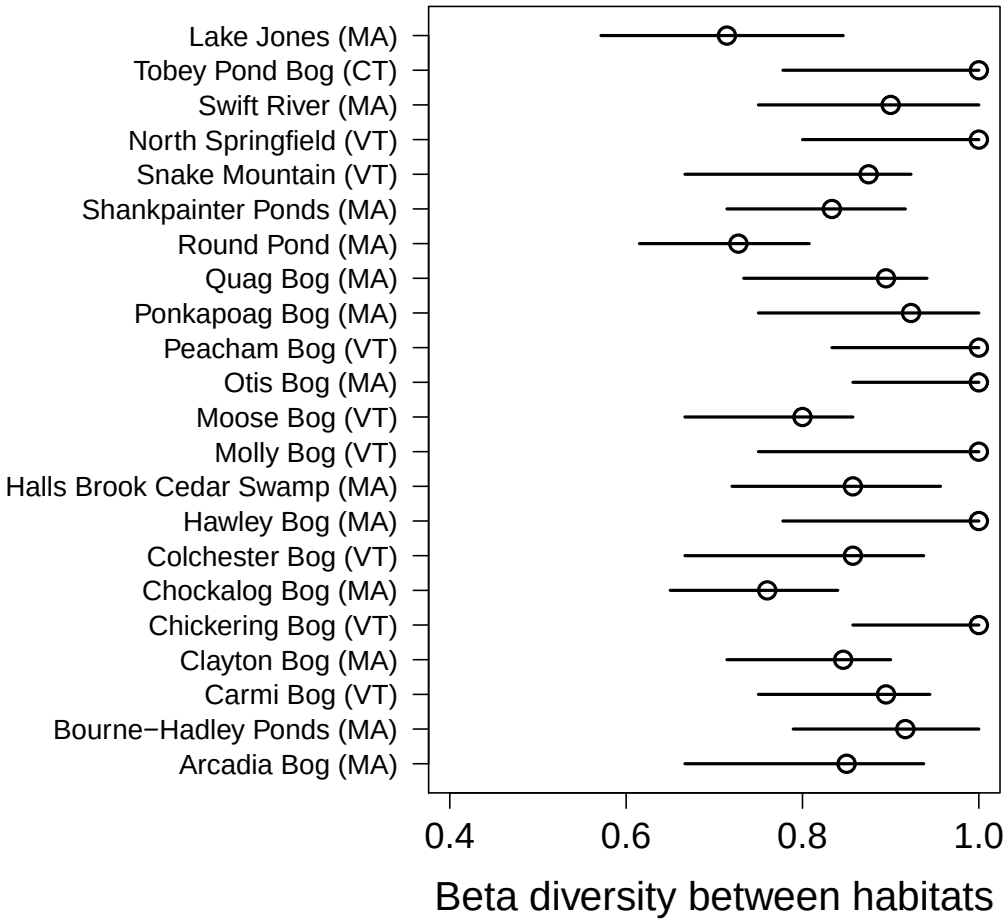

Figure 3: Estimates of beta diversity (open circles with 95% credible intervals) between ant communities present in bog and forest habitats at each sample location.

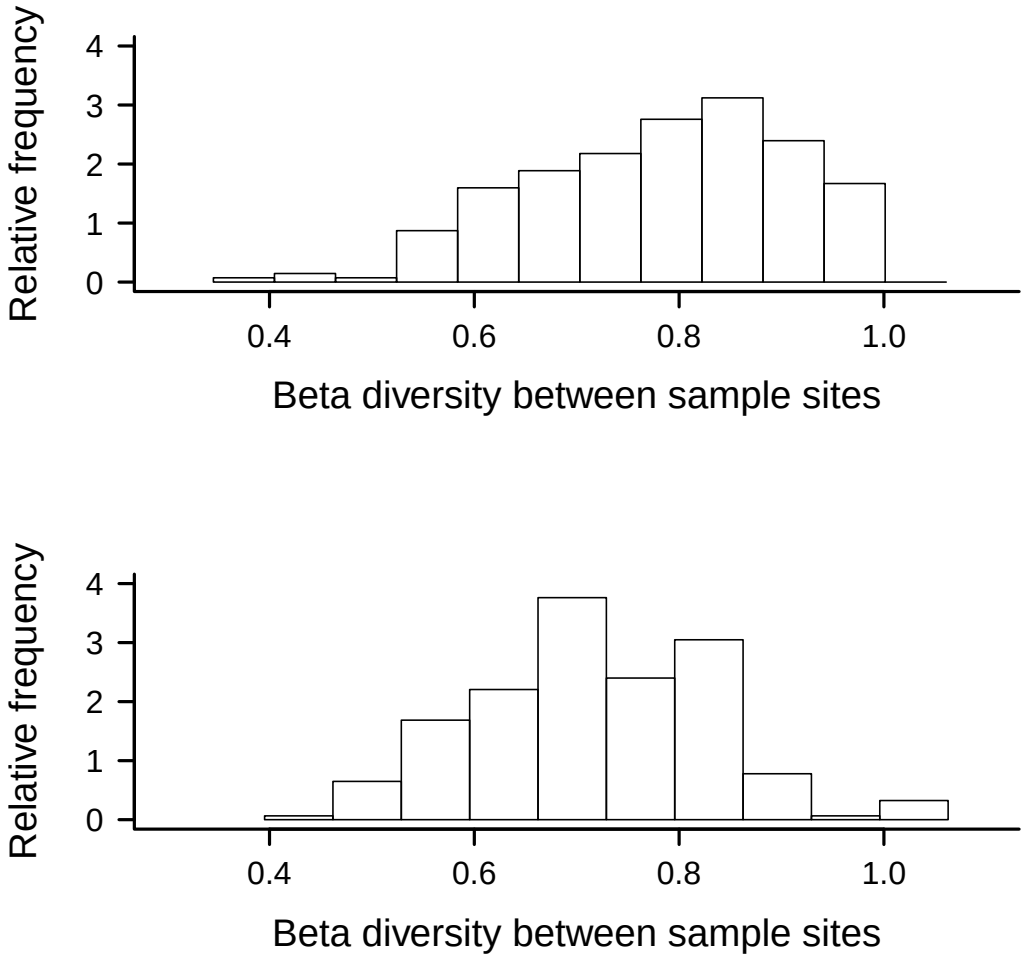

Figure 4: Distribution of estimates of beta diversity computed for all pairwise combinations of samples collected in forest habitat (upper panel) or bog habitat (lower panel).

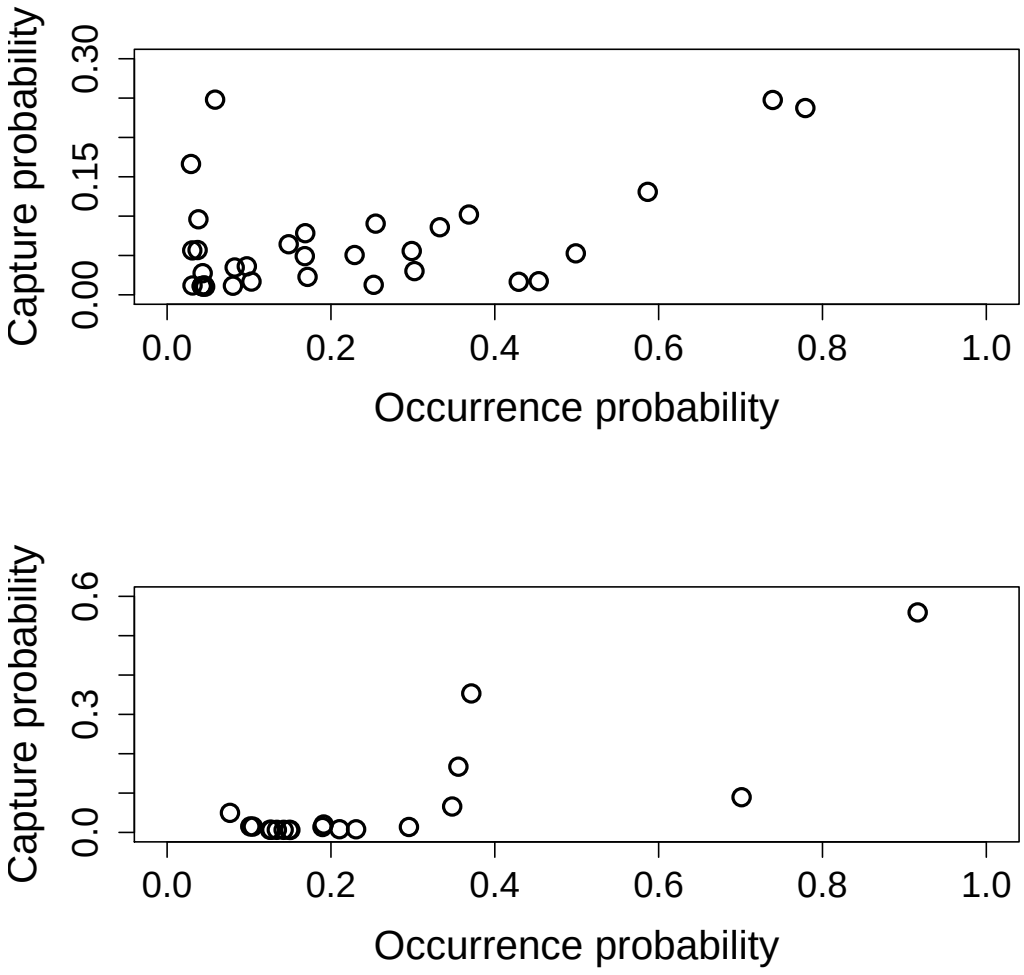

Figure 5: Estimates of species-specific capture probability versus occurrence probability for ants in forest habitat (upper panel) and bog habitat (lower panel). Note difference in scale between ordinates of upper and lower panels.

	Site k								
	Observed				Partially observed				
species i	1	2	$\cdots$	R	1	2	$\cdots$	R	w_i
1	y_{11}	y_{12}	$\cdots$	y_{1R}	z_{11}	z_{12}	$\cdots$	z_{1R}	w_1
2	y_{21}	y_{22}	$\cdots$	y_{2R}	z_{21}	z_{22}	$\cdots$	z_{2R}	w_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nR}	z_{n1}	z_{n2}	$\cdots$	z_{nR}	w_n
$n+1$	0	0	$\cdots$	0	$z_{n+1,1}$	$z_{n+1,2}$	$\cdots$	$z_{n+1,R}$	w_{n+1}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
N	0	0	$\cdots$	0	z_{N1}	z_{N2}	$\cdots$	z_{NR}	w_N
$N+1$	0	0	$\cdots$	0	$z_{N+1,1}$	$z_{N+1,2}$	$\cdots$	$z_{N+1,R}$	w_{N+1}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
M	0	0	$\cdots$	0	z_{M1}	z_{M2}	$\cdots$	z_{MR}	w_M

Table 1: Conceptualization of the supercommunity of M species used in parameter-expanded data augmentation. $\boldsymbol{Y}$ comprises a matrix of n rows of observed trap frequencies and $M - n$ rows of unobserved (all-zero) trap frequencies. $\boldsymbol{Z}$ denotes a matrix of species- and site-specific occurrence parameters. $\boldsymbol{w}$ denotes a vector of parameters that indicate membership in the community of N species vulnerable to sampling.

		Posterior probability	
Habitat	Covariates	Uniform prior	Jeffreys' prior
Forest	LAT, LAI, GSF, ELEV	0.818	0.767
Forest	LAT, LAI, ELEV	0.177	0.229
Forest	LAT, ELEV	0.005	0.003
Forest	LAT, GSF, ELEV	< 0.001	0.001
Bog	ELEV	0.424	0.416
Bog	None	0.342	0.412
Bog	LAT	0.082	0.070
Bog	AREA, ELEV	0.060	0.034
Bog	LAT, ELEV	0.045	0.029
Bog	AREA	0.038	0.036
Bog	LAT, AREA	0.006	0.003
Bog	LAT, AREA, ELEV	0.004	0.001

Table 2: Posterior probabilities of models containing different covariates of species occurrence probabilities. Covariates include latitude (LAT), leaf area index (LAI), light availability (GSF), elevation (ELEV), and bog area (AREA). Models with less than 0.001 posterior probability are not shown.

	Capture probability			Occurrence probability		
Species	Median	2.5%	97.5%	Median	2.5%	97.5%
Amblyopone pallipes	0.028	0.008	0.073	0.043	0.005	0.237
Aphaenogaster rudis (species complex)	0.237	0.209	0.269	0.779	0.539	0.927
Campnnotus herculeanus	0.090	0.062	0.123	0.255	0.104	0.482
Campnnotus nearcticus	0.035	0.013	0.074	0.083	0.014	0.316
Campnnotus novaeboracensis	0.017	0.008	0.037	0.454	0.121	0.897
Campnnotus pennsylvanicus	0.131	0.107	0.158	0.587	0.322	0.819
Dolichoderus pustulatus	0.011	0.002	0.053	0.042	0.003	0.389
Formica argentea	0.011	0.001	0.053	0.044	0.003	0.411
Formica glacialis	0.012	0.002	0.055	0.045	0.003	0.413
Formica neogagates	0.096	0.049	0.163	0.038	0.005	0.166
Formica obscuriventris	0.010	0.001	0.051	0.046	0.003	0.448
Formica subaenescens	0.051	0.029	0.081	0.229	0.085	0.476
Formica subintegra	0.166	0.083	0.284	0.029	0.003	0.140
Formica subsericea	0.248	0.184	0.320	0.059	0.009	0.218
Lasius alienus	0.053	0.035	0.075	0.499	0.260	0.761
Lasius flavus	0.011	0.002	0.051	0.043	0.003	0.397
Lasius neoniger	0.036	0.013	0.076	0.097	0.020	0.333
Lasius speculiventris	0.012	0.003	0.040	0.080	0.009	0.502
Lasius umbratus	0.017	0.007	0.037	0.429	0.109	0.931
Myrmecina americana	0.011	0.002	0.052	0.042	0.003	0.398
Myrmica detritinodis	0.078	0.049	0.117	0.169	0.055	0.378
Myrmica lobifrons	0.056	0.036	0.082	0.299	0.118	0.568
Myrmica punctiventris	0.248	0.218	0.279	0.739	0.474	0.911
Myrmica species 1 ("AF-scu")	0.102	0.078	0.131	0.368	0.152	0.642
Myrmica species 2 ("AF-smi")	0.064	0.039	0.097	0.148	0.036	0.385
Prenolepis imparis	0.012	0.002	0.054	0.031	0.002	0.334
Stenamma brevicorne	0.017	0.005	0.046	0.103	0.014	0.526
Stenamma diecki	0.030	0.014	0.056	0.302	0.097	0.725
Stenamma impar	0.049	0.026	0.081	0.168	0.052	0.396
Stenamma schmitti	0.013	0.005	0.030	0.252	0.046	0.753
Tapinoma sessile	0.023	0.010	0.047	0.171	0.035	0.552
Temnothorax ambiguus	0.056	0.015	0.138	0.031	0.003	0.150
Temnothorax curvispinosus	0.057	0.022	0.113	0.037	0.005	0.169
Temnothorax longispinosus	0.086	0.062	0.114	0.333	0.141	0.587

Table 3: Estimated probabilities of capture and occurrence (with 95% credible intervals) for ant species captured in forest habitat. Probabilities are estimated at the average value of the covariates observed in the sample.

	Capture probability			Occurrence probability		
Species	Median	2.5%	97.5%	Median	2.5%	97.5%
Camponotus herculeanus	0.014	0.002	0.050	0.190	0.040	0.731
Camponotus novaeboracensis	0.066	0.043	0.094	0.348	0.172	0.571
Camponotus pennsylvanicus	0.007	0.001	0.040	0.134	0.017	0.723
Dolichoderus plagiatus	0.015	0.002	0.073	0.105	0.016	0.515
Dolichoderus pustulatus	0.090	0.071	0.112	0.701	0.491	0.863
Formica neorufibarbis	0.007	0.001	0.040	0.126	0.015	0.691
Formica subaenescens	0.353	0.308	0.402	0.371	0.194	0.580
Formica subsericea	0.014	0.004	0.037	0.295	0.083	0.774
Lasius alienus	0.020	0.006	0.054	0.191	0.051	0.550
Lasius speculiventris	0.050	0.010	0.138	0.077	0.014	0.263
Lasius umbratus	0.008	0.001	0.034	0.210	0.037	0.766
Leptothorax canadensis	0.007	0.001	0.039	0.142	0.018	0.764
Myrmica lobifrons	0.559	0.529	0.589	0.916	0.748	0.984
Myrmica punctiventris	0.006	0.001	0.039	0.150	0.018	0.783
Myrmica species 1 ("AF-scu")	0.015	0.002	0.073	0.102	0.015	0.486
Myrmica species 2 ("AF-smi")	0.008	0.001	0.034	0.231	0.041	0.826
Stenamma brevicorne	0.007	0.001	0.041	0.149	0.019	0.772
Tapinoma sessile	0.167	0.133	0.207	0.356	0.184	0.561
Temnothorax ambiguus	0.007	0.001	0.042	0.127	0.017	0.697

Table 4: Estimated probabilities of capture and occurrence (with 95% credible intervals) for ant species captured in bog habitat. Probabilities are estimated at the average value of the covariates observed in the sample.