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

Locations of traps were agged 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 sh-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 di use 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 (ve 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 modi cation of the multi-species model of occurrence and detection that includes site-speci c covariates (Royle and Dorazio 2008, Kery and Royle 2009). This modi cation allows a nite set of candidate models to be speci ed and t 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 scienti cally 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 su ciently distinct that di erences in species occurrence { and possibly capture rates { are expected a priori. Furthermore, the potential covariates of occurrence di er 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 $y_{ik} \, 2 \, f0; 1; \ldots ; J_k g$ 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

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 $\mathbf{Y}_{obs}$ 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 $\mathbf{Y}_{obs}$ and the model for the observed data is appropriately expanded to analyze the augmented data matrix $\mathbf{Y} = (\mathbf{Y}_{obs}; \mathbf{0})$. 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 $\mathbf{w} = (w_1, \ldots, w_M)$ to the model to indicate whether each species is a member of the community ($w = 1$) or not ($w = 0$). The elements of $\mathbf{w}$ 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 $\mathbf{w}$ as follows: $N = \sum_{i=1}^{M} w_i$. Therefore, estimation of Ω and $\mathbf{w}$ 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 $\mathbf{Z}$, whose elements indicate the presence ($z = 1$) or absence ($z = 0$) of species i at sample site k. Although $\mathbf{Z}$ 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:

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

where ϕ_{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 $\Pr(z_{ik} = 1 | w_i = 1) = \phi_{ik}$. For purposes of computing estimates of community-level characteristics, $\mathbf{Z}$ 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 $\mathbf{Y}$ and the parameters $\mathbf{Z}$ and $\mathbf{w}$ 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 $\mathbf{Z}$ and $\mathbf{w}$ are paramount in terms of estimating measures of biodiversity. We have shown already that estimates of $\mathbf{w}$ are used to compute estimates of species richness N (a measure of gamma diversity). Similarly, $\mathbf{Z}$ may be used to estimate measures of alpha diversity, beta diversity, and other community-level characteristics. For example, summing the columns of $\mathbf{Z}$ yields the number of species present at each sample site (alpha diversity). Similarly, different columns of

Z may be compared to express differences in species composition among sites (beta diversity). For

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. Specically, 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:

$$y_{ik} | z_{ik} \quad \text{Binomial}(J_k; z_{ik} p_{ik})$$

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-specic 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 singto changeseirhangesan

species-specific estimate is shrunk in the direction of the estimated average parameter value. Of course,

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) ($\hat{} = 0.779$). This species is taxonomically unresolved and currently includes a complex of

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 ψ, 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 consensus 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 ao17

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 (= 2.482 and = 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

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 eld 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.

Kery, 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.

Kery, M., and Royle, J. A. 2009. Inference about species richness and community structure using species-speci c 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.

Kery, 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.

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., Holt, R. D.,

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.
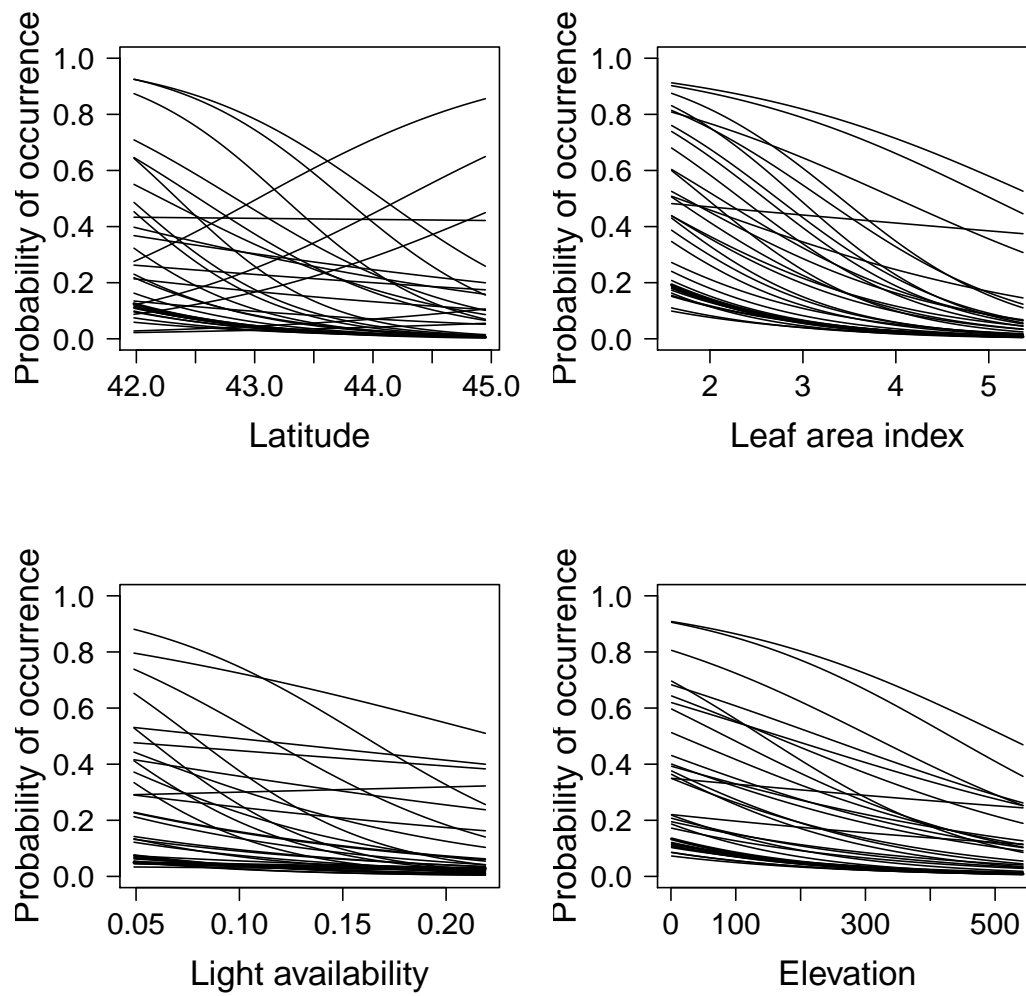
Figure 1: Estimated effects of covariates on occurrence probabilities of ant species in forest habitat.
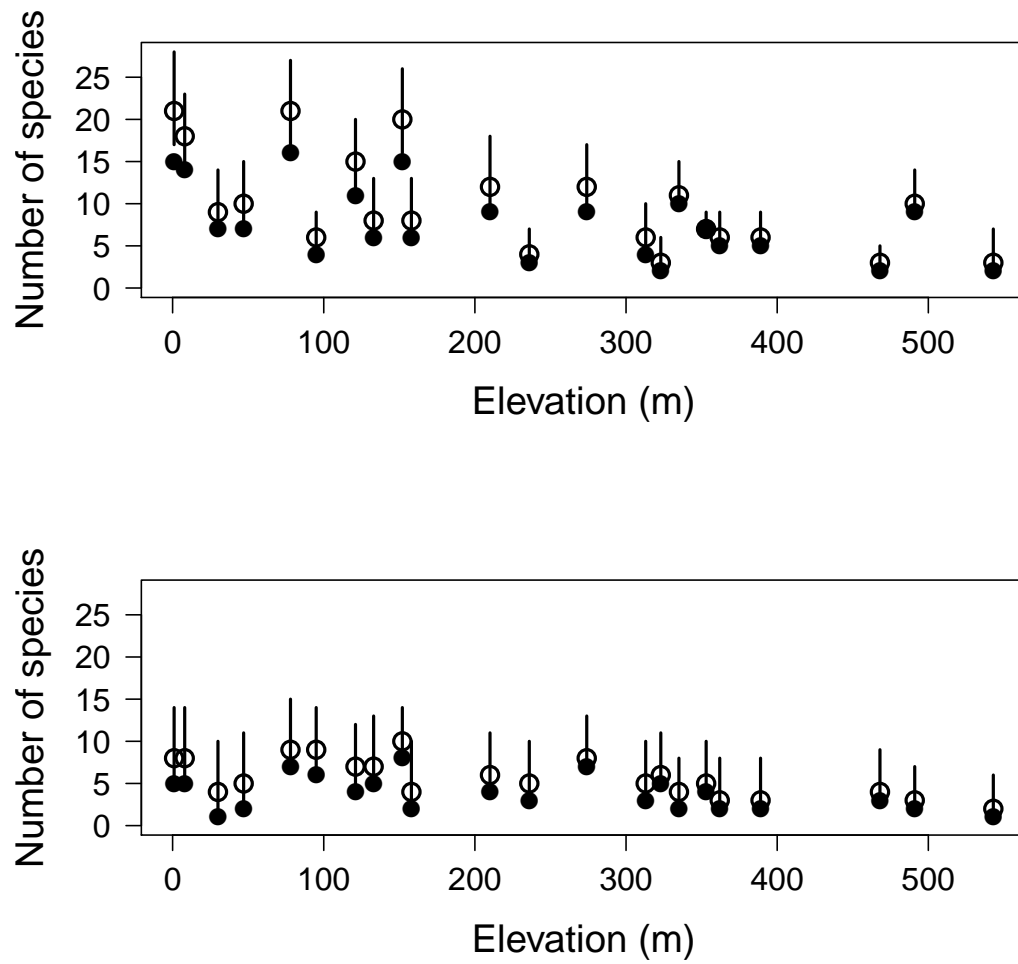
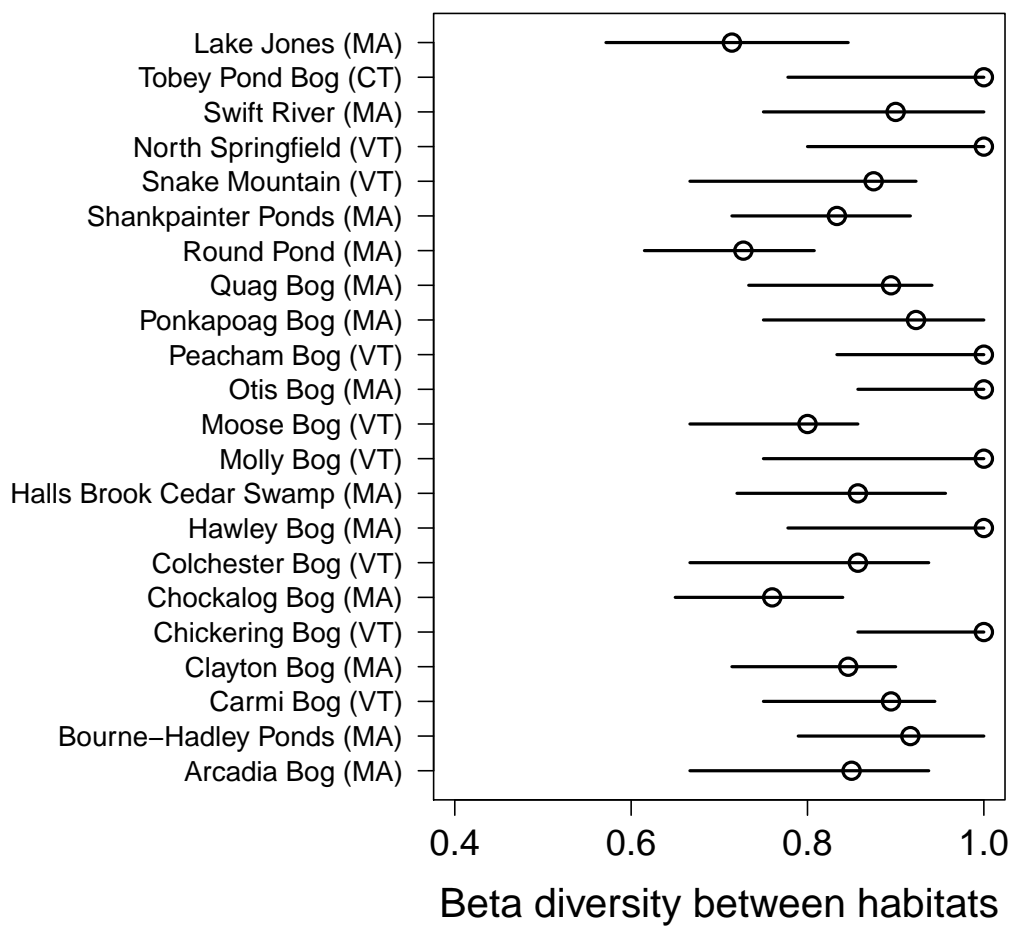Figure 2: Estimates of site-specific species richness (open circles with 95% credible intervals) for ants

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.

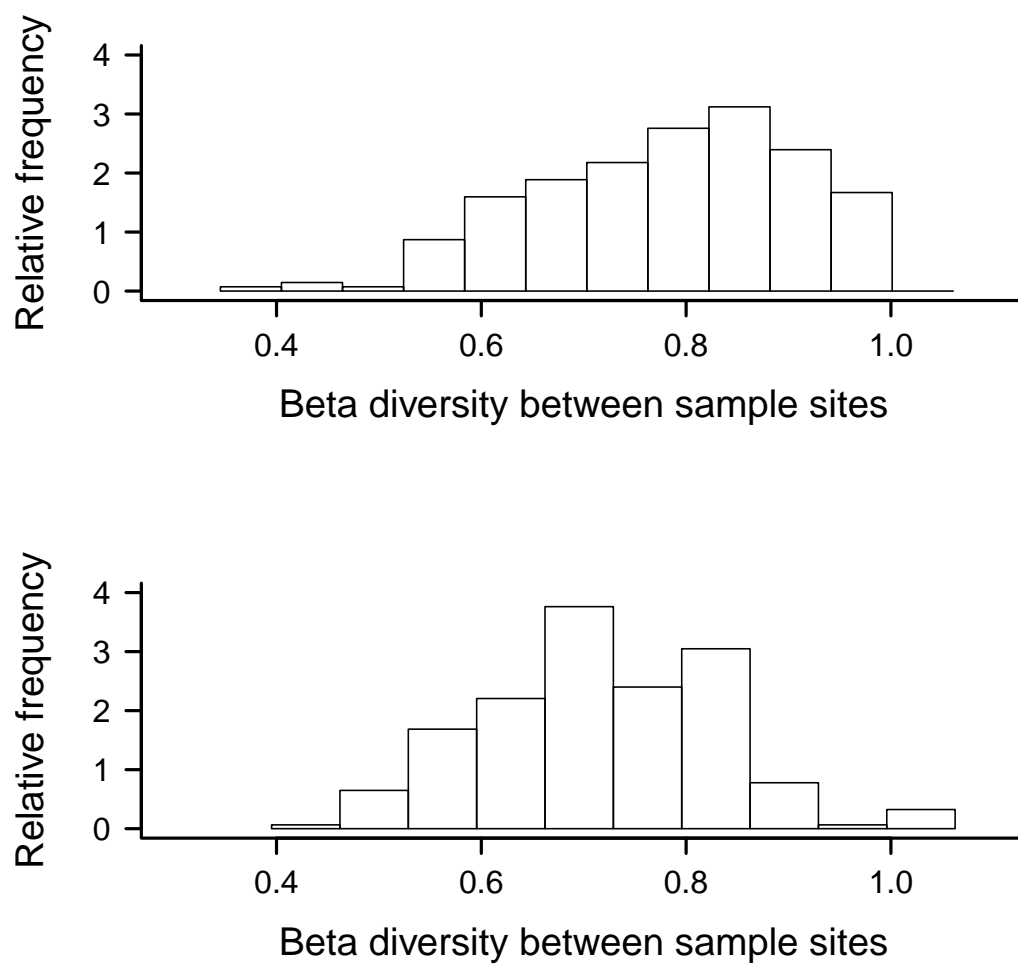Figure 4: Distribution of estimates of beta diversity computed for all pairwise combinations of samples collected in forest habitat (upper panel) or bog habitat (lower panel).
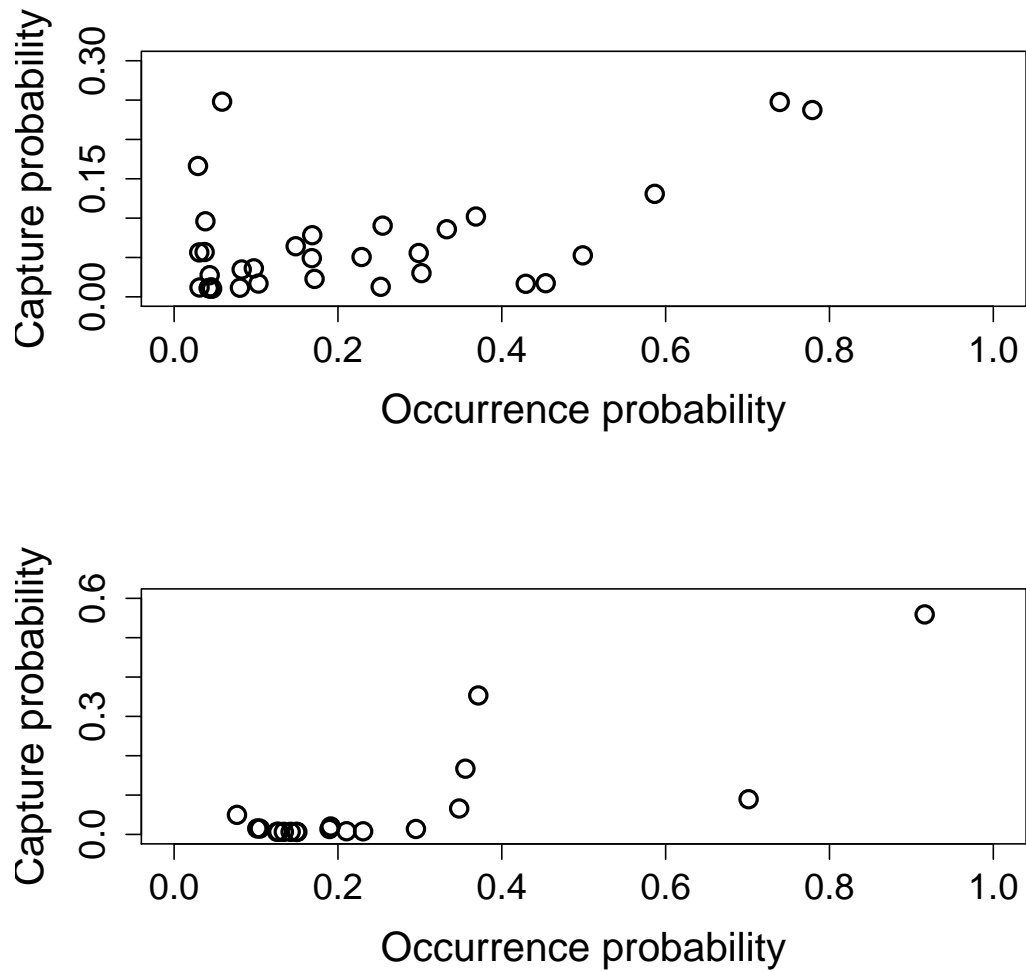
Figure 5: Estimates of species-specific capture probability versus occurrence probability for ants in forest habitat (upper panel) and bog habitat (lower panel). Note difference in scale between ordinates of upper and lower panels.

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

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

Habitat	Covariates	Posterior probability	
		Uniform prior	Je reys' 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 di erent 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.

Species	Capture probability			Occurrence probability		
	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 avus	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.

Species	Capture probability			Occurrence probability		
	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 neoru barbis	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

Dollch744nusFpUStU08islit)28288435 Tf 163.74448u60a[(0.353)-1052(0.071)-1550(0.037)-2845(0.295)-1582(0.