Received Date : 16-Mar-2015

Revised Date : 09-Nov-2015

Accepted Date : 16-Nov-2015

Article type : Research Article

Editor : Robert B. O'Hara

A hierarchical distance sampling model to estimate abundance and covariate associations of species and communities

Running title: Community distance sampling

Rahel Sollmann*, North Carolina State University, Department of Forestry and Environmental Resources, Raleigh, NC 27695, USA; sollmann@izw-berlin.de, current address: US Forest Service, Pacific Southwest Research Station, Davis, CA 95618

Beth Gardner, North Carolina State University, Department of Forestry and Environmental Resources, Raleigh, NC 27695, USA

Kathryn A. Williams, Biodiversity Research Institute, Gorham, ME 04038, USA

Andrew T. Gilbert, Biodiversity Research Institute, Gorham, ME 04038, USA

Richard R. Veit, College of Staten Island, CUNY New York, Division of Humanities and Social Sciences, NY 10314, USA

*Corresponding author

Abstract

1. Distance sampling is a common survey method in wildlife studies, because it allows accounting for imperfect detection. The framework has been extended to hierarchical distance sampling (HDS), which accommodates the modeling of abundance as a function of covariates, but rare and elusive species may not yield enough observations to fit such a model.

2. We integrate HDS into a community modeling framework that accommodates multi-species spatially replicated distance sampling data. The model allows species-specific parameters, but these come from a common underlying distribution. This form of information sharing enables estimation of parameters for species with sparse data sets that would otherwise be discarded from analysis. We evaluate the performance of the model under varying community sizes with different species-specific abundances through a simulation study. We further fit the model to a seabird data set obtained from shipboard distance sampling surveys off the East Coast of the U.S.A.

3. Comparing communities comprised of 5, 15 or 30 species, bias of all community level parameters and some species-level parameters decreased with increasing community size, while precision increased. Most species-level parameters were less biased for more abundant species. For larger communities, the community model increased precision in abundance estimates of rarely observed species when compared to single species models. For the seabird application, we found a strong negative association of community and species abundance with distance to shore. Water temperature and prey density had weak effects on seabird abundance. Patterns in overall abundance were consistent with known seabird ecology.

4. The community distance sampling model can be expanded to account for imperfect availability, imperfect species identification or other missing individual covariates. The model

allowed us to make inference about ecology of species communities, including rarely observed species, which is particularly important in conservation and management. The approach holds great potential to improve inference on species communities that can be surveyed with distance sampling.

Key words: Bayesian p-value, cluster size, hierarchical model, seabirds, sparse data, wildlife surveys

Introduction

Distance sampling (Buckland 2001; Buckland *et al.* 2005) is a popular method to survey both terrestrial and marine wildlife species amenable to direct observation. In distance sampling, the probability of detecting an individual is assumed to decrease with increasing distance from the observer. This allows estimation of abundance and density while accounting for observation bias. The framework has been extended to accommodate the modeling of abundance at multiple survey sites as a function of site specific covariates (Hedley & Buckland 2004; Royle *et al.* 2004; Conn *et al.* 2012), termed hierarchical distance sampling (HDS).

HDS provides a framework to investigate factors influencing the abundance of individual species. Often, however, rare or elusive species will not yield sufficient observations to parameterize an individual model. Community modeling provides an approach to jointly analyzing multi-species data sets and sharing information across species while maintaining the ability to model species-specific parameters (Dorazio & Royle 2005; Dorazio *et al.* 2006). In this type of community modeling, information is shared across species by assuming a common underlying distribution for species-specific parameters. These distributions, in turn, are governed by hyperparameters, which reflect community-level patterns and processes. The use of collective community data allows estimation of community and species-level parameters, even for those species that are rare and elusive. This concept has been applied repeatedly in occupancy modeling (i.e., species-level detection/non-detection data, Kéry & Royle 2008; Zipkin *et al.* 2009). Multi-species distance sampling data sets have previously been analyzed using species, or species groups, as a covariate (Alldredge *et al.* 2007), but to our knowledge, no attempt has been made to combine community modeling based on shared hyperdistributions with the framework of distance sampling.

Here, we develop a community distance sampling model that estimates both community-level and species-level parameters related to detection and abundance. Specifically, we extend the HDS framework to accommodate multi-species spatially replicated distance sampling data sets. The model allows for species-specific parameters (with shared hyperdistributions) in both components describing detection probability and abundance. We evaluate the performance of the model under varying community sizes through a simulation study. We further use the model to analyze a seabird data set obtained from shipboard distance sampling surveys in the Mid-Atlantic, off the east coast of the U.S.A. This application is particularly relevant as seabirds are a highly threatened marine taxonomic group (Sydeman *et al.* 2012), and potential development of offshore wind energy facilities has raised additional concern about their conservation (e.g., Garthe & Hüppop 2004; Petersen & Fox 2007). Our analysis includes a community of 14 species, of which nine did not yield sufficient observations to be analyzed individually, and results provide important information on abundance of this community in areas of wind energy development. The method holds promise for many distance sampling applications to improve estimation of detection and abundance of species and communities.

Methods

Development of the community distance sampling model

Distance sampling can be implemented along line transects or at survey points. We develop the community distance sampling model based on line transect surveys, but note how this can be adjusted to point transects. In line transect based distance sampling, for each observation the perpendicular distance of the object of interest to the transect line is recorded (for point surveys, use the radial distance) (Buckland 2001). Detection on the transect is assumed to be perfect and the detection probability *p* of an object is defined by a declining function *g* of its distance to the transect line, *x*, for example, using a half-normal detection function

$$g(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Here, *σ* is the scale parameter of the half-normal function. In reality, observations are frequently grouped into *k* = 1, 2, … *K* distance categories. This binning smoothes inaccuracies in distance estimation and reduces effects of movement of animals in response to observers. Let *L* be the length of the line transect survey, and v_k the width and A_k the area covered by the *k*-th distance category, equivalent to *Lv* (or 2*Lv* when both sides of a transect are surveyed). Further, let **b** be the *K*+1 break points of the *K* distance categories. Then, detection probability in *k*, p_k, is the integral of *g*(*x*) over the break points of *k:*

$$p_k = \frac{L\int_{b_k}^{b_{k+1}} g(x)dx}{A_k} = \frac{\int_{b_k}^{b_{k+1}} g(x)dx}{v_k}$$

Individuals are assumed to be uniformly distributed in space, so that the probability of an individual occurring in distance category *k*, ψ_k, is the proportion of the sampled area covered by the *k*-th distance interval (note that in line transect surveys with constant *v*, A_k is also constant, but in point surveys, this area increases with increasing distance from the survey point (Buckland

2001)). The vector of the number of observations in each of the *K* distance categories, **y**, is a multinomial random variable with size $n = \sum_k y_k$ and cell probabilities $\boldsymbol{\pi} = \mathbf{p}\boldsymbol{\psi} / \sum_k p_k \psi_k$, which simplifies to $\boldsymbol{\pi} = \mathbf{p} / \sum_k p_k$ for the constant-*v* line transect case. Note that this formulation of the detection model is conditional on detection (Buckland 2001), but see Royle et al. (2004) for an unconditional formulation.

We can link *n* to the true abundance *N* using the total detection probability $p.t = \sum_k p_k$:

$$n \sim Binomial(N, p.t).$$

When distance sampling surveys are carried out at *j* = 1, 2, ..., *J* survey sites, observations and model parameters are indexed by site:

$$n_j \sim Binomial\left(N_j, p.t_j\right).$$

Following Royle et al. (2004), we assume N_j to follow some probability mass function *f* (e.g., Poisson or negative binomial), and its expected value λ_j can be modeled as a function of covariates, **X**, e.g.,

$$N_j \sim f(\lambda_j)$$

$$\log\left(\lambda_j\right) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_j,$$

where α_0 is the intercept and **α** is a vector of coefficients associated with the covariates **X**. Analogously, detection parameters can be modeled as functions of site specific covariates (Marques & Buckland 2003; Oedekoven *et al.* 2013); for example, for the scale parameter of the half-normal detection function:

$$\log\left(\sigma_j\right) = \beta_0 + \boldsymbol{\beta}'\mathbf{Y}_j,$$

Where β_0 is the intercept, and **β** the vector of coefficients associated with the detection covariates in **Y**.

To expand this approach to a community model for i = 1, 2, … M species, the parameters are further indexed by species, and ascribed hyperdistributions to the resulting sets of species-specific parameters. For example, each species i has a detection intercept $\beta_{0,i}$ such that:

$\beta_{0,i} \sim Normal(\mu_{\beta 0}, \sigma^2_{\beta 0})$.

The hyperparameters of these distributions, here the mean $\mu_{\beta 0}$ and variance $\sigma^2_{\beta 0}$, constitute the community parameters shared by all species and are estimated as part of the model, as are the species-specific parameters. This is equivalent to including a normally distributed random effect for species in the model.

Simulation study

We evaluated performance of the community distance sampling model through a simulation study. We considered three community sizes with M = 5, 15 or 30 species. We simulated abundance for these species across J = 50 sites, N_{ij}, from a Poisson distribution. We generated site and species-specific Poisson means, λ_{ij}, with a normally distributed species-specific abundance intercept $\alpha_{0,i}$ (hyperparameters: $\mu_{\alpha 0} = \log(1.5)$ and $\sigma^2_{\alpha 0} = 1$), one site-specific covariate **X** (normally distributed with mean 0 and variance 1), and the respective normally distributed species-specific coefficient α_i (hyperparameters: $\mu_\alpha = 0$ and $\sigma^2_\alpha = 0.25$). This led to total abundances within communities varying from <10 to >1000.

From these abundances we simulated distance sampling detection data across a strip of width w = 10, divided into 10 distance categories of width v = 1. We generated species and site specific detection parameters σ_{ij} on the log-scale with a normally distributed species-specific intercept $\beta_{0,i}$ (hyperparameters: $\mu_{\beta 0} = \log(2.5)$, and $\sigma^2_{\beta 0} = 0.0625$), a site-specific binary covariate **Y** and an associated fixed effect, $\beta = -0.2$. The choice of hyperparameters resulted in variation in baseline σ (i.e., before covariate effects) within communities from 1 to 5 units on the real scale. For each scenario (*M* = 5, 15 or 30) we generated 100 data sets and analyzed these with the data-generating model. For community level parameters, we evaluated root mean square error, bias and confidence interval coverage across all 100 iterations for each scenario. We used the 2.5 and 97.5 percentiles as the Bayesian 95% credible interval (95BCI). For species-level parameters we were interested whether parameter estimates were influenced by species abundance. Therefore, we grouped species into abundance categories from 1-10, 11-100, 101-1000, and >1000 individuals. For each abundance category and each community scenario we calculated bias and coverage of parameters across all iterations.

To evaluate whether the community model improved abundance estimates over single-species models, we extracted species-specific data sets from the community data generated as described above and analyzed them with a single-species HDS model containing the same covariates on abundance and detection. Because we hypothesized that improvements should be stronger for species with fewer detections, we selected species-specific data sets with 21-60, 61-100 or 101-140 total detections (as a rule of thumb, estimating abundance from distance sampling data requires at least 60 to 80 observations (Buckland *et al.* 1993), so our criterion of >20 observations is liberal). For each observation category, we contrast community model and single species model estimates of abundance, looking at average coefficient of variation (standard error divided by estimate) and bias. We implemented this comparison separately for each of the three communities.

We further explored the ability to assess fit of the detection and the abundance component of the community distance sampling model. We tested model fit using Bayesian p-values (Gelman *et* *al.* 1996). These values are obtained by calculating some fit statistic (e.g., a residual) that depends on the model parameters and the observed data, determining the same fit statistic for a new set of data generated from the model under consideration, and then calculating the portion of time the residuals from the newly generated data are larger (or smaller) than those of the original data. If the model fits the data appropriately, the resulting Bayesian p-value will be close to 0.5 (we suggest values < 0.1 or >0.9 may indicate lack of fit). We calculated Bayesian p-values for the species and site-specific abundances, N_{ij}, to assess fit of the abundance component; and for the observations $\mathbf{y}$ to assess fit of the detection component (for details, see Appendix S1).

All N_{ij} are latent and subject to the specific assumptions of the distribution from which they are simulated. To test whether Bayesian p-values based on N_{ij} are able to distinguish between competing abundance models, we generated abundance data for a community of 15 species from a negative binomial abundance model, with the expected values generated as in the main simulation study, and a dispersion parameter of 1. We generated detection data from these simulated communities as described in the main simulation study, analyzed the resulting data with a Poisson and a negative binomial abundance model, and compared Bayesian p-values between the correctly specified and the mis-specified model. We further used these simulations to evaluate the effect of overdispersion in abundance on model performance.

Implementation

We implemented the community distance sampling model in a Bayesian framework, using the software JAGS (Plummer 2003) accessed through R version 2.15.2 (R Core Team 2014). We used vague priors on community level parameters. The model code and R script for the simulation study can be found in Appendix S2. We ran three parallel Markov chains started at different initial values with a burn-in of 500 iterations and 8,000 post burn-in iterations. For the $M = 5$ case, we ran 20,000 iterations to achieve convergence. Because of the large number of parameters to be monitored, we thinned chains by 8 to reduce the size of the model output. We tested for chain convergence using the Gelman-Rubin statistic (Gelman *et al.* 2004). This statistic is a measure of among-chain versus between chain variance, and values < 1.1 indicate convergence.

Application: Seabirds off the East Coast of the U.S.A.

Seabird line transect data were collected along 656.1 km of boat transects located off the coast of Delaware, Maryland and Virginia (Fig. S3-1), sampled over the course of four days in April 2012. Observations were restricted to one side of the boat and to the quadrant defined by the line of travel and a 90-degree angle to this line of travel. Perpendicular distance of any bird, or the center of a cluster of birds, to the transect line was recorded. Each cluster was counted as a single record and the size noted. The survey yielded a final data set that contained 632 records of 14 seabird species (Table S3-1, for details of data preparation, see Appendix S3). Preliminary analysis of the number of detections against distance for all species, both separately and combined, indicated the data conformed to the assumption of decreasing detection with increased distance (Fig. S3-2 and S3-3).

During the boat surveys, water temperature and sea state were recorded at 30-minute intervals. We used the points at which these environmental covariates were measured to divide the ship transects into 73 segments (Fig. S3-1), which constituted the survey sites. The resulting segments varied in length from 1.1 to 20.5 km (mean: 8.99 km, SD: 2.51 km). We accounted for these differences by using segment length, L_j, as an offset in the abundance component of the model.

Additionally, hydroacoustic data were collected at 500-m resolution to obtain an index of prey biomass. Details of covariate collection are given in Appendix S3. We analyzed the entire multi-species data set using the novel community distance sampling model. To contrast community model estimates with estimates from single-species models, we analyzed the 5 species in the data set with >20 observation with single-species HDS models.

Covariates

We used in situ collected water temperature (*TEMP*, °C) and an index of prey biomass density (*PD*) derived from hydroacoustic data, as well as distance to shore (*DTS*, km) as covariates on abundance. To define segment level values of *TEMP* and *DTS*, we took the mean of measurements from the start and end point of each segment. For *PD*, we averaged all measurements taken within a segment. The majority of species in our dataset are visual hunters, and are likely responding to foraging cues from the top several meters of the water column. Therefore, we used prey density in the first 3 to 5 m of the water column (the first 2 m are missed by echo sounding devices). *TEMP* is considered an inverse proxy for prey availability (Hunt *et al.* 1981; Pinaud & Weimerskirch 2002), whereas echo sounding data gives a direct index of prey availability (Wiebe *et al.* 1990; Demer & Hewitt 1995).

We explored the effect of sea state (Beaufort values recorded in the survey ranged from 1 = light air/water ripples, to 4 = moderate breeze/small waves and fairly frequent white caps; *BEAU*) on the detection parameter σ. In Appendix S3 we further present a model accounting for bird behavior in the detection component.

Parameterization of the community distance sampling model for the seabird dataset

Based on the observed distances, we set the maximum observation distance *w* at 1000 m and binned observations into K = 10 100-m distance categories. We used a negative binomial distribution (with species and site-specific mean λ_{ij} and constant dispersion parameter *r*) for abundance and included segment length (as offset) and all abundance covariates in the predictor. We included a random species-specific intercept and random species-specific coefficients for these covariates in the abundance component:

$N_{ij} \sim Negative\ Binomial(\lambda_{ij}, r)$

$$\log(\lambda_{ij}) = \log(L_j) + \alpha_{0,i} + \alpha1_i TEMP_j + \alpha2_i PD_j + \alpha3_i DTS_j.$$

We assumed that the detectability of different species is influenced in a similar way by sea state and therefore estimated fixed coefficients β for all species in our model for the detection parameter σ. Differences in detectability among species were accounted for by a random species specific intercept:

$\log\ (\sigma_{ij}) = \beta_{0,i} + \boldsymbol{\beta} BEAU_j,$

(note that *BEAU* is categorical). We used Normal hyperdistributions for all random species-specific parameters and estimated the respective community means and variances.

Accounting for clusters of birds

When objects are observed in clusters, then individuals are not observed independently, and clusters should be used as the unit of observation. In this situation, N_{ij} is no longer the number of individuals of species *i* at site *j*, but the number of clusters. To estimate total abundance, we augmented the above described model with a component describing cluster size of observation *m*, C_m, to be a zero-truncated negative binomial variable, with a mean, μ_C, and dispersion parameter, ρ, shared by all species:

$$C_m \sim zt\ Negative\ Binomial(\mu_C, \rho).$$

Note that C_m is partially observed, i.e., known for observed clusters and unknown for $\sum_i \sum_j N_{ij} - n_{ij}$ unobserved clusters. In general, it may be more appropriate to have a species-specific mean cluster size; however, in the seabird dataset, 74% of all observations were of single individuals; 95% of all observations were of 4 or less individuals. We therefore decided against the additional complexity of a species-specific cluster size model, and also refrained from adding cluster size into the detection model as a covariate. Many seabird studies, however, report observations of large aggregations of birds. In these situations, the effect of cluster size on detectability (e.g., Smith *et al.* 1995; Pearse *et al.* 2008) can be included as a covariate on the log-linear predictor of σ. We calculated total abundance for a species at a site as the sum of all clusters for that species at that site, and total abundance in the survey area by summing over all clusters across all sites. The survey area is equivalent to a 1000-m strip along the combined boat transects.

Implementation

We implemented the analysis in JAGS (Plummer 2003), with three parallel Markov chains, a burn-in of 1,000 iterations and 50,000 post burn-in iterations, thinned by 20. We report results as posterior means and standard deviations, as well as 95BCI. We considered covariate effects as strong/significant if their 95BCI did not overlap 0. Posterior distributions of total abundance estimates across all sites for the less abundant species tended to be right-skewed. Therefore, we provide the mode and the mean in our summary statistic for species level abundances. Data, R script and JAGS model code to implement this analysis are in Appendix S2.

Results

Simulation study – general performance

Community means for the three species-level random effects distributions (abundance intercept and coefficient, and detection intercept), as well as the fixed effect coefficient of the detection covariate had low to moderate bias (-2% to 14%); for the abundance intercept, bias declined with increasing community size, from -13% and 14% for M = 5 and 15, to 2% for M = 30.

Bias for the standard deviations of these three species-level random effects distributions was moderate to high (15 – 66%) for M = 5 and declined to moderate-low (4 – 12%) and low (-0.2 – 3%) for M = 15 and M = 30, respectively. Root mean square errors for all parameters decreased as community size increased. For details, see Table S4-1.

Bias in species-specific abundance intercepts α_0 was moderate to low and overall decreased with increasing abundance and increasing community size (Table S4-2). Bias in the species-specific habitat coefficient α was moderate in communities of all sizes and showed no discernable relationship with abundance (Table S4-3). For both parameters we do not present relative bias because true values are often close to 0, and division by these values leads to very large numbers, even when absolute bias is low. Relative bias in the intercept of σ was low to moderate across communities, and declined with increasing abundance (Table S4-4). Confidence interval coverage for all parameters was nominal or close to nominal.

Relative bias in total abundance estimates for the smallest abundance category (0 – 10 individuals) was high (90 – 145%) for communities of all sizes; absolute bias was only 6 – 8 individuals and the high relative bias is a function of the low true abundances encountered in this category. Relative bias was moderate (9 – 11%) for the 11 – 100 abundance category, and low (2 – 6 %) for the two highest abundance categories. Confidence interval coverage was (close to) nominal across categories. These patterns held true for communities of all sizes (Table 1).

With the negative binomial distribution for abundance data generation and model fitting (*M* = 15 species only), species-specific parameters showed very similar magnitude and patterns in bias and confidence interval coverage compared to those under a Poisson distribution (Tables S4-2 – S4-4). Bias in community level parameters was also very similar, but these parameters had much higher root mean squared errors (Table S4-1).

Simulation study – Community versus single species models

For small communities (*M* = 5), bias in abundance estimates under the community model was considerably higher for rarely observed species (20 – 60 detections) and moderately higher for the intermediate (61 – 100 detections) group, than under a single-species model (34% versus 4%, and 9% versus 3%, respectively). For larger communities (*M* = 15 or 30), bias in abundance estimates was comparable and low (<2%) under single species and community models, but community models resulted in more precise estimates for the lowest observation class (Figure S4-1).

Simulation study – Model fit

Bayesian p-values for the correctly identified Poisson abundance model for communities with *M* = 15 species were close to 0.5, for both the detection and the abundance component (abundance: mean = 0.518, SD = 0.072, range 0.364 - 0.732; detection: mean = 0.511, SD = 0.052, range

0.402 - 0.638). When using a Poisson model to analyze data generated under a negative binomial abundance model, the Bayesian p-value for the abundance model fell to 0.081 (SD 0.065, range 0 – 0.270); when analyzing these same data with the correctly specified model, the Bayesian p-value was 0.475 (SD 0.025, range 0.386 - 0.527), indicating that the Bayesian p-value based on the latent *N* was useful for investigating fit of the abundance component.

Seabirds off the East Coast of the U.S.A.

The community distance sampling model provided estimates of abundance of 14 seabird species (Table S3-4), and identified covariates influencing their detectability and distribution. Bayesian p-values indicated the community distance sampling model fit the data appropriately (Table S3-2). Sea states 3 and 4 had a strong negative effect on detectability of seabirds, relative to sea state 1 (Fig. S3-5, Table S3-3). Detailed results of this and the model considering bird behavior as a covariate on detection can be found in Appendix S3.

Distance to shore had a strong negative effect on seabird abundance across the entire community, with a mean, $\mu_{\alpha 1}$, of -0.999 ± 0.252 (Table S3-3). The effect was significantly negative for nine species (Fig. S3-4). The mean effect of water temperature and prey density on the seabird community were negligible (-0.001 ± 0.170 and -0.018 ± 0.159, respectively, Table S3-3), but temperature had a significantly positive effect on two species (Figure S3-4).

The (back-transformed) mean detection parameter across species, exp($\mu_{\beta 0}$), was 216.269 ± 21.104 m at sea state = 1 and declined to 159.415 ± 18.680 m at sea state = 4. Among species, σ at sea state = 1 varied between 183.943 and 271.566 m (Figure S3-5). Mean cluster size for all species was 1.927 ± 0.107 individuals. Total abundance across all survey sites was highest for

Common Loons (mode = 1677) and lowest for Forster's Terns and Surf Scoters (mode = 2) (Table S3-4).

Comparing single-species and community model estimates for 5 species with >20 records, we found that most parameter estimates were consistent across the two modeling approaches, and confidence intervals for all parameters overlapped (Table S5). In general, differences in parameter estimates were strongest for those parameters fixed for the entire community in the community model. There were some considerable differences in estimates of total abundance (e.g., Laughing Gull single species model: 376; community model: 608). These stemmed from species-specific mean cluster size estimates under single-species models being quite different for some species from the community-wide mean adopted in the community model. Estimates of number of clusters, *N*, were largely consistent across approaches.

Discussion

The hierarchical framework of community modeling as implemented here, has been applied mostly within occupancy models (Kéry & Royle 2008; Zipkin *et al.* 2009). We developed a community distance sampling model to estimate relationships between abundance and environmental covariates for multiple species, using species-level random effects. Random effects have been used in distance sampling in other contexts (e.g., modeling variation in σ across sites Oedekoven *et al.* 2015), but not to accommodate multi-species data sets. Estimation of the hyperparameters governing these random effects distributions should be influenced by the number of species in the community, which correspond to the number of levels of the random effects. Indeed, we observed reduced bias with increasing community size for some, and improved precision for all community level parameters. In addition, we had to run models for small communities considerably longer to achieve convergence. The strongest effect of small community size came into play when comparing the performance of single-species versus community HDS models. In communities with only 5 species, for species with 20 to 100 observations, single-species models provided abundance estimates with much lower bias than community models (Fig. S4-1). In larger communities, both approaches performed similarly. This suggests that the community distance sampling model may not work well for communities with 5 or fewer species, though further investigation of the effect of number of observations on model performance would provide more insight on these specific scenarios. In such small communities, more suitable modeling approaches might be to share parameters across species, or to group species and use these groups as fixed effects in the model (e.g., Alldredge *et al.* 2007). Species specific parameters tended to be less biased for more abundant species, rather than being primarily influenced by community size, which is intuitive, as these also tend to be the species with more detections, and hence, more data to inform parameter estimates (e.g., Pacifici *et al.* 2014).

The main benefit of using a community model over a single-species model lies in the ability to obtain abundance estimates of species observed so rarely they cannot be modeled individually. Previous analyses of multi-species data sets have relied on analyzing species separately, excluding very rarely detected species (e.g., Studeny *et al.* 2013). The community model performed well in providing estimates of total abundance of species, including rare ones, but abundance estimates exhibited some degree of “shrinkage”, where abundance of very rare and very abundant species is pulled towards the community mean (i.e., over and underestimated, respectively). In real applications, parameter estimates can also be influenced by how one defines a community. Pacifici *et al.* (2014) showed for occupancy models that (a) combining all species into a single community potentially masks differences among sub-communities in their response to covariates; and (b) especially for rarely observed species, parameter estimates are sensitive to group membership.

We found that Bayesian p-values based on the latent site and species specific abundances N_{ij} were useful in detecting mis-specified abundance models (in this case, they indicated lack of fit when analyzing data generated from a negative binomial distribution with a Poisson model). It is noteworthy, though, that amount and patterns in bias in parameter estimates under the mis-specified model were very similar to those under the correctly specified models. Thus, parameter estimates appeared to be fairly robust to our specific mis-specification of the abundance model.

Seabirds off the East Coast of the U.S.A.

One advantage in the approach developed here is the flexibility related to species-specific differences in parameters. In the seabird dataset, the detection parameter σ showed variation across species: the large and predominantly white Northern Gannet had a significantly larger σ than the community average (Fig. S3-5). Similarly, the community mean effect of *TEMP* on abundance was close to 0 (Table S3-3), yet *TEMP* had a significantly positive effect on abundance for two species (Fig. S3-4). We expected the consistent negative effect of distance to shore on species abundance. DTS can be a limiting factor for foraging activities (Weimerskirch 2007; Fauchald 2009) and correlate positively with ocean depth, another important predictor of seabird foraging activity and abundance (e.g., Freeman *et al.* 2010; Nur *et al.* 2011). Contrary to our expectations, we observed weak effects of *TEMP* and *PD* on the abundance of seabird species (Fig. S3-4). Lower water temperatures are generally associated with higher primary productivity. There are, however, several intermediate trophic levels between primary production and top marine predators like seabirds (Barnes & Hughes 1988), which can lead to spatio-temporal lags in the response of seabirds to changes in these covariates. It is also conceivable that the in-situ measures of *PD* and *TEMP*, taken immediately under the boat, do not adequately represent the environmental conditions in the 1000-m strip sampled.

Overall, Common Loon, Northern Gannet and Laughing Gull had the highest estimated abundances in the study area. In contrast, Surf Scoters and Forster's Terns were extremely rare during the April survey. These patterns are in agreement with what is known about the seasonal ecology and foraging behavior of these species (for details, see Appendix S3).

For the more frequently observed species, analysis with single-species HDS models indicated that there was some variation among species in the parameters we considered fixed across all species in the community model. Most notably, average cluster size was variable, leading to differences between the two modeling approaches in total abundance estimates for some species. This indicates that different parameterizations of the cluster size model itself may be required to adequately describe the observed data, for example in the form of species-specific means, finite mixtures, or distributions allowing for more variability in counts (Zipkin *et al.* 2014). Estimates in numbers of clusters, *N*, were largely consistent across both modeling approaches, but differences were larger than those observed in the simulation study (Fig. S4-1). This is likely the result of higher levels of variability in parameters among species than the normally distributed random effects of the community model allowed for.

Modeling detection probability and missing individual covariates

We focused our simulations and application on a fully observed, site-level categorical covariate, but we also explored additional detection covariates in our analysis of the seabird data set

(Appendix S3). Particularly, we investigated whether bird behavior, classified as “on the water” or “in the air”, had an impact on detection distances of seabirds. Behavior is an observation-level covariate, and in order to accommodate this kind of covariate, we estimated abundance for the two behavioral categories separately, using behavior-specific intercepts in the abundance model. Although we found no significant effect of behavior on the detection parameter σ, we believe that the conceptual set-up of the “behavior” model allows for some interesting ecological insight into the percentage of individuals in a population performing certain behaviors (Fig. S3-6). The approach of estimating abundance separately for two behavioral categories circumvents the issue of unknown behavioral category of unobserved individuals/clusters, but likely performs poorly with an increasing number of categories (due to low sample size per category), and breaks down completely for continuous individual covariates. In these cases, a different approach is to treat individual covariates of unobserved clusters as missing data, and specify a parametric model to estimate missing covariate values (e.g., Conn *et al.* 2014). This approach is equivalent to how the seabird application deals with cluster size of unobserved seabirds (see *Accounting for clusters of birds*), and could also be used to accommodate observations with uncertain species identification and/or missing distance-to-transect information (excluded in the seabird application). Specifically, the present model could be augmented with a species identification model as developed by Conn *et al.* (2013, 2014), where species identity is treated as a latent multinomial variable. Knowledge about species-specific identification probabilities (e.g., from double-observer surveys or experiments with known species identity) can be used to formulate informative priors on these multinomial cell probabilities (with vague priors, unidentified observations will be distributed among species according to their proportion in the identified observations). In our model description, missing distances could be sampled from the multinomial model specified for the observations **y** (see *Development of the community distance sampling model*) by assuming the probability of not recording a distance occurs at random across distance bins.

Certain behaviors can render animals unavailable for detection – diving in seabirds and other aquatic species, temporary emigration from the survey area (Chandler *et al.* 2011), or failure to sing in song-based bird surveys (e.g., Diefenbach *et al.* 2007). Failure to take into account availability <1 will lead to negative bias in abundance estimates. At present, our model assumes that individuals are always available to be detected. Availability can be estimated separately, for example from intensive observation studies or telemetry studies that allow inference on animal behavior (e.g., Diefenbach *et al.* 2007; Conn *et al.* 2014), and can be incorporated into the estimator of abundance (Buckland 2001) so that Eq. 1 becomes

$$n_{ij} \sim Binomial(N.a_{ij}, p.t_{ij}).$$

Here, $N.a_{ij}$ is the number of individuals of species i at site j that are available for detection, $p.t_{ij}$ is the total detection probability, and

$$N_{ij} = \frac{N.a_{ij}}{p.a_{ij}},$$

where $p.a_{ij}$ is the probability of species i at site j being available.

Uncertainty about these estimates could readily be incorporated by treating availability as a parameter, rather than fixing it, and formulating an informative prior based on available information. Alternatively, Chandler *et al.* (2011) developed an approach to account for availability in distance sampling studies that requires repeated sampling of survey sites, which could also readily be incorporated into the community distance sampling framework.

Conclusion

Distance sampling is employed in the study of a variety of taxa, and often, data on multiple species are collected (e.g., Jathanna *et al.* 2003; Somershoe *et al.* 2006; Williams & Thomas 2007). The present approach allows such studies to investigate community ecology and distribution of many species from within a flexible and coherent modeling framework. In the context of conservation and management, rare and listed species are often of particular interest, and the ability to incorporate rare species into analyses provides important information about their abundance and distribution.

Acknowledgments

We thank Holly Goyert, Nicholas Flanders, David L. Miller, Paul Conn and three anonymous reviewers for helpful comments on an earlier draft of this manuscript. Data were collected by Christopher Vogel, Holly Goyert, and Emily Connelly, with support from boat captain Brian Patteson. This material is based upon work supported by the Department of Energy's Office of Energy Efficiency & Renewable Energy under Award Number DE-EE0005362. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed

herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Data Accessibility

The data to implement the analysis of the seabird case study are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.gb905.

Literature Cited

Alldredge, M.W., Pollock, K.H., Simons, T.R. & Shriner, S.A. (2007). Multiple-species analysis of point count data: a more parsimonious modelling framework. *Journal of Applied Ecology*, **44**, 281–290.

Barnes, R.S.K. & Hughes, R.N. (1988). *An Introduction to Marine Ecology, 2nd edn*. Blackwell Scientific Publications, Oxford, UK.

Buckland, S.T. (2001). *Introduction to distance sampling: estimating abundance of biological populations*. Oxford University Press, Oxford, UK.

Buckland, S.T., Anderson, D.R., Burnham, K.P. & Laake, J.L. (2005). *Distance sampling*. Wiley Online Library.

Buckland, S.T., Anderson, D.R., Burnham, K.P. & Laake, J.L. (1993). *Distance sampling: estimating abundance of biological populations.* Chapman & Hall, London, UK.

Chandler, R.B., Royle, J.A. & King, D.I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology*, **92**, 1429–1435.

Conn, P.B., Ver Hoef, J.M., McClintock, B.T., Moreland, E.E., London, J.M., Cameron, M.F., Dahle, S.P. & Boveng, P.L. (2014). Estimating multispecies abundance using automated detection systems: ice-associated seals in the Bering Sea. *Methods in Ecology and Evolution*, **5**, 1280–1293.

Conn, P.B., Laake, J.L. & Johnson, D.S. (2012). A hierarchical modeling framework for multiple observer transect surveys. *PloS one*, **7**, e42294.

Conn, P.B., McClintock, B.T., Cameron, M.F., Johnson, D.S., Moreland, E.E. & Boveng, P.L. (2013). Accommodating species identification errors in transect surveys. *Ecology*, **94**, 2607–2618.

Demer, D.A. & Hewitt, R.P. (1995). Bias in acoustic biomass estimates of Euphausia superba due to diel vertical migration. *Deep Sea Research Part I: Oceanographic Research Papers*, **42**, 455–475.

Diefenbach, D.R., Marshall, M.R., Mattice, J.A., Brauning, D.W. & Johnson, D.H. (2007). Incorporating availability for detection in estimates of bird abundance. *The Auk*, **124**, 96–106.

Dorazio, R.M. & 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. & Glimskär, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.

Fauchald, P. (2009). Spatial interaction between seabirds and prey: review and synthesis. *Marine Ecology Progress Series*, **391**, 139–151.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Flanders, N.P., Gardner, B., Winiarski, K.J., Paton, P.W.C., Allison, T. & O'Connell, A.F. (2015). Key seabird areas in southern New England identified using a community occupancy model. *Marine Ecology Progress Series*, **533**, 277–290.

Freeman, R., Dennis, T., Landers, T., Thompson, D., Bell, E., Walker, M. & Guilford, T. (2010). Black Petrels (Procellaria parkinsoni) Patrol the Ocean Shelf-Break: GPS Tracking of a Vulnerable Procellariiform Seabird. *PLoS ONE*, **5**, e9236.

Garthe, S. & Hüppop, O. (2004). Scaling possible adverse effects of marine wind farms on seabirds: developing and applying a vulnerability index. *Journal of Applied Ecology*, **41**, 724–734.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian data analysis, second edition.* CRC/Chapman & Hall, Bocan Raton, Florida, USA.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–760.

Hedley, S.L. & Buckland, S.T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 181–199.

Hunt, G.L., Gould, P.J., Forsell, D.J. & Peterson Jr, H. (1981). Pelagic distribution of marine birds in the eastern Bering Sea. *The eastern Bering Sea shelf: oceanography and resources* (eds D.W. Hood & Caulder), pp. 689–718. National Oceanic and Atmospheric Administration, Rockville, MD.

Jathanna, D., Karanth, K.U. & Johnsingh, A.J.T. (2003). Estimation of large herbivore densities in the tropical forests of southern India using distance sampling. *Journal of Zoology*, **261**, 285–290.

Johnson, D.S., Ream, R.R., Towell, R.G., Williams, M.T. & Leon Guerrero, J.D. (2013). Bayesian Clustering of Animal Abundance Trends for Inference and Dimension Reduction. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**, 299–313.

Kéry, M. & Royle, J.A. (2008). Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. *Journal of Applied Ecology*, **45**, 589–598.

Marques, F.F. & Buckland, S.T. (2003). Incorporating covariates into standard line transect analyses. *Biometrics*, **59**, 924–935.

Nur, N., Jahncke, J., Herzog, M.P., Howar, J., Hyrenbach, K.D., Zamon, J.E., Ainley, D.G., Wiens, J.A., Morgan, K., Ballance, L.T. & Stralberg, D. (2011). Where the wild things are: predicting hotspots of seabird aggregations in the California Current System. *Ecological Applications*, **21**, 2241–2257.

Oedekoven, C.S., Buckland, S.T., Mackenzie, M.L., Evans, K.O. & Burger, L.W. (2013). Improving distance sampling: accounting for covariates and non-independency between sampled sites. *Journal of Applied Ecology*, **50**, 786–793.

Oedekoven, C.S., Laake, J.L. & Skaug, H.J. (2015). Distance sampling with a random scale detection function. *Environmental and Ecological Statistics*, 1–13.

Pacifici, K., Zipkin, E.F., Collazo, J.A., Irizarry, J.I. & DeWan, A. (2014). Guidelines for a priori grouping of species in hierarchical community models. *Ecology and Evolution*, **4**, 877–888.

Pearse, A.T., Gerard, P.D., Dinsmore, S.J., Kaminski, R.M. & Reinecke, K.J. (2008). Estimation and Correction of Visibility Bias in Aerial Surveys of Wintering Ducks. *The Journal of Wildlife Management*, **72**, 808–813.

Petersen, I.K. & Fox, A.D. (2007). *Changes in bird habitat utilization around the Horns Rev 1 offshore wind farm, with particular emphasis on Common Scoter*. Vattenfall A/S, DK. National Environmental Research Institute, Kalø, Denmark.

Pinaud, D. & Weimerskirch, H. (2002). Ultimate and proximate factors affecting the breeding performance of a marine top-predator. *Oikos*, **99**, 141–150.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pp. 20–22.

R Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Royle, J.A., Dawson, D.K. & Bates, S. (2004). Modeling abundance effects in distance sampling. *Ecology*, **85**, 1591–1597.

Smith, D.R., Reinecke, K.J., Conroy, M.J., Brown, M.W. & Nassar, J.R. (1995). Factors affecting visibility rate of waterfowl surveys in the Mississippi Alluvial Valley. *Journal of Wildlife Management*, **59**, 515–527.

Somershoe, S.G., Twedt, D.J. & Reid, B. (2006). Combining breeding bird survey and distance sampling to estimate density of migrant and breeding birds. *The Condor*, **108**, 691–699.

Studeny, A.C., Buckland, S.T., Harrison, P.J., Illian, J.B., Magurran, A.E. & Newson, S.E. (2013). Fine-tuning the assessment of large-scale temporal trends in biodiversity using the example of British breeding birds. *Journal of Applied Ecology*, **50**, 190–198.

Sydeman, W.J., Thompson, S.A. & Kitaysky, A. (2012). Seabirds and climate change: roadmap for the future. *Marine Ecology Progress Series*, **454**, 107–117.

Weimerskirch, H. (2007). Are seabirds foraging for unpredictable resources? *Deep Sea Research Part II: Topical Studies in Oceanography*, **54**, 211–223.

Wiebe, P.H., Greene, C.H., Stanton, T.K. & Burczynski, J. (1990). Sound scattering by live zooplankton and micronekton: Empirical studies with a dual-beam acoustical system. *The Journal of the Acoustical Society of America*, **88**, 2346–2360.

Williams, R. & Thomas, L. (2007). Distribution and abundance of marine mammals in the coastal waters of British Columbia, Canada. *Journal of Cetacean Research and Management*, **9**, 15.

Zipkin, E.F., DeWan, A. & Andrew Royle, J. (2009). Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. *Journal of Applied Ecology*, **46**, 815–822.

Zipkin, E.F., Leirness, J.B., Kinlan, B.P., O’Connell, A.F. & Silverman, E.D. (2014). Fitting statistical distributions to sea duck count data: Implications for survey design and abundance estimation. *Statistical Methodology*, **17**, 67–81.

Table 1: Summary results for estimates of species-specific total abundance, *N*, from 100 iterations of a community distance sampling model, split by abundance categories. *M* = number of species in the community, # cases = number of species, across all iterations, in abundance category; Mean true = average input value for category; Mean estimate = average posterior mean estimate of parameter for category; Bias = average relative bias, CI coverage = percentage of time, out of # cases, true value fell within 2.5th and 97.5th percentile of the posterior of the estimate.

M	Abundance category	# cases	Mean true	Mean estimate	Bias	CI coverage
5	1 - 10	11	6.909	14.819	144.545	0.818
	11 - 100	276	52.181	54.584	10.83	0.931
	101 - 1000	212	236.83	231.748	-2.603	0.948
	> 1000	1	2537	2480.776	-2.216	1
15	1 - 10	28	7.607	13.667	90.027	0.964
	11 - 100	785	51.343	53.518	9.211	0.958
	101 - 1000	670	249.221	243.178	-2.113	0.942
	> 1000	17	1518.235	1443.983	-5.748	0.882
30	1 - 10	79	7.304	14.725	124.232	0.911
	11 - 100	1614	51.889	54.447	9.538	0.962
	101 - 1000	1271	239.629	232.853	-2.844	0.947
	> 1000	36	1420.389	1391.255	-2.177	0.944