Confronting preferential sampling in count and occupancy surveys: diagnosis and model-based triage

Paul B. Conn[1,*], **James T. Thorson**[2], **Devin S. Johnson**[1]

[1] National Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA National Marine Fisheries Service, 7600 Sand Point Way NE, Seattle, WA 98115 USA; [2] Fisheries Resource Assessment and Monitoring Division (FRAM), Northwest Fisheries Science Center, National Marine Fisheries Service (NMFS), NOAA, 2725 Montlake Boulevard E, Seattle, WA 98112, USA

[*]paul.conn@noaa.gov

Appendix S1: Further details on preferential sampling model development and simulations

In this appendix, we provide further details on model development, implementation, and simulation testing for the preferential sampling model outlined in the main article text. Our modeling approach works by constructing (a) a species distribution model (SDM) describing abundance or presence/absence that includes spatially autocorrelated random effects, and (b) a model for the probability of site selection. Both submodels can be expressed in terms of habitat or landscape covariates. However, site selection probability can also be written as a function of the same spatially autocorrelated random effects in (a). Including the same set of random effects in both submodels is the mechanism for allowing dependence betwen the two models, and allows us to estimate and control for the potentially biasing effects of preferential sampling.

Preferential sampling model development

We focus here on a discrete spatial domain; extensions to point process models in continuous space are relatively straightforward. In particular, we suppose that the investigator intending to fit a SDM to animal encounter data breaks their study area up into S survey units (label these $U_1, U_2, \ldots, U_S$), of which n are selected for sampling (call the set of sampled locations $\mathcal{S}$). Each survey unit i is assigned a vector of covariates, $\mathbf{x}_i$, and an indicator R_i that takes on the value 1.0 if survey unit i is sampled (i.e. if $U_i \in \mathcal{S}$), and is 0 otherwise. Let Z_i denote the state variable of interest in cell i (e.g., occupancy, abundance), and let Y_i denote observations made in cell i. In some cases, observations may be equivalent to the state variable in sampled cells (i.e. when detection probability is 1.0). In others, an observation model $[Y_i|Z_i, \boldsymbol{\theta}]$ relating Y_i to the underlying process of interest will be required, where $\boldsymbol{\theta}$ are nuisance parameters describing detection. For instance, for survey counts in our simulation study and in the bearded seal example, we set

$$[Y_i|Z_i, \boldsymbol{\theta}] \quad = \quad \text{Binomial}(Z_i, A_i p_i),$$

where A_i gives the proportion of survey unit i that is sampled and p_i is a detection probability (p_i is set to 1.0 for the simulation study and an informative prior distribution is specified in the bearded seal example).

We express variation in Z_i generically as

$$Z_i \quad \sim \quad f(g^{-1}(\mu_i)), \qquad \text{eqn 1}$$

where $f()$ is a probability mass function (e.g., Poisson for count data, Bernoulli for occupancy), $g()$ is a link function (e.g. probit or logit for occupancy, log for count data), and μ_i is a link-scale intensity value. In applications described in this paper, we specify this intensity as

$$\mu_i \quad = \quad \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + \lambda_i, \qquad \text{eqn 2}$$

where β_0 is an intercept, $\mathbf{x}_i$ is a vector of covariates used to explain variation in abundance (here indexed by the site, i), and $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_S\}$ are spatially autocorrelated random effects.

We model the site selection probabilities in the same manner. First, we write R_i, as

$$R_i \sim \text{Bernoulli}(h^{-1}(\nu_i)), \quad \text{eqn 3}$$

where $h()$ denotes a link function appropriate for binary data (e.g. logit, probit). We then write the intensity for this model as

$$\nu_i = \beta_0^* + \mathbf{x}_i^* \boldsymbol{\beta}^* + \eta_i + \mathbf{B}\delta_i. \quad \text{eqn 4}$$

Here, β_0^* is an intercept, $\mathbf{x}_i^*$ is a vector of covariates (in practice, these can be the same or different from $\mathbf{x}_i$), and $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_S\}$ are spatially autocorrelated random effects. The connection between the two models (eqs. 2 and 4) is the reliance of both on δ_i. In eq. 4, the matrix $\mathbf{B}$ describes this relationship. Previous preferential sampling models in the statistical literature have set $\mathbf{B} = b\mathbf{I}$, where $\mathbf{I}$ is an identity matrix and b is a parameter to be estimated. This is the approach we take in simulation study and in several of the models fit to real data. Another approach is to construct $\mathbf{B}$ so it changes smoothly in space, as in (Royle & Berliner 1999). In this manner, the degree of preferential sampling could vary across the landscape; presumably, one would need fairly rich data to be able to model a spatially varying relationship.

We have yet to specify a model for spatial autocorrelation. For analyses in this paper, we model mean-zero spatial random effects (i.e., the λ_i and η_i) for each site i as multivariate normal:

$$\begin{aligned} \boldsymbol{\lambda} &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\lambda), \text{ and} \\ \boldsymbol{\eta} &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\eta), \end{aligned}$$

where $\boldsymbol{\Sigma}_\lambda$ and $\boldsymbol{\Sigma}_\eta$ are covariance matrices. As is typical of spatial models in statistics, we allow the magnitude of spatial autocorrelation to decrease as a function of the distance between survey units. In particular, we use the well known Matérn covariance function to model covariance as a function of distance. In addition to the distances between survey units (which we calculate relative to survey unit centroids), the Matérn covariance function uses a range (or scale) parameter κ, a precision parameter τ (inverse marginal variance), and a smoothness parameter to model covariance. For analyses in this paper, we set the smoothness parameter to 1.0 and treat κ and τ as parameters to be estimated. To improve computational efficiency, we use a Gaussian Markov random field (GMRF) representation of the underlying Gaussian field (Lindgren *et al.* 2011), as described by Thorson *et al.* (2015) (Appendix B).

To conduct statistical analysis of the preceding models, we use maximum marginal likelihood as implemented in Template Model Builder (TMB; Kristensen *et al.* 2015). Interfaced with the R programming environment, the TMB software uses a Laplace approximation to integrate out random effects, and a bias correction algorithm (Tierney *et al.* 1989; Thorson & Kristensen In Press) is used to obtain abundance estimates and standard errors that properly account for nonlinear transformations of random effects. Requisite R and TMB code will be published to a publicly accessible repository upon acceptance, and is also available at `https://github.com/NMML/pref_sampling/`.

Simulation study details

For each of 500 simulations, we performed the following steps:

1. Generate a spatially autocorrelated covariate over a 25×25 grid, where the covariate at grid cell i, X_i, is generated via a multivariate normal distribution:

$$\mathbf{X} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}). \quad \text{eqn 5}$$

An isotropic, Gaussian covariance structure is used for $\boldsymbol{\Sigma}$, where the ith row and jth column of $\boldsymbol{\Sigma}$ is set to $\Sigma_{ij} = \sigma^2 \exp -(d_{ij}/s)^2$, where d_{ij} is the distance between the centroid of grid cell i and grid cell j, σ is a standard deviation (here, set to 1.0), and s is a scale parameter (here, set to 5.0).

2. Generate true abundance in each survey unit (grid cell), N_i, as a function of the spatially generated covariate from step 1, together with additional spatially autocorrelated error. We generate a vector of such abundances as

$$N_i \quad \sim \quad \text{Poisson}(\exp(\beta_0 + \mathbf{X}\beta_1 + \boldsymbol{\delta})), \qquad \text{eqn 6}$$

where $\boldsymbol{\delta}$ are drawn in an identical manner to $\mathbf{X}$ in step 1. We set $\beta_0 = 2.0$ and drew $\beta_1 \sim \text{Uniform}(-0.5, 0.5)$ for each simulation.

3. For each simulation, we selected 50 grid cells for sampling (i.e., set $R_i = 1.0$) for according to 3 alternative surveying configurations. Each surveying configuration draws grid cells randomly without replacement, where the probability of sampling cell i is proportional to

$$\exp(X_i\beta_1^* + \eta_i + b * delta_i).$$

Here, η_i is a spatially autocorrelated random effect drawn in an identical manner to $\mathbf{X}$ in step 1, and b describes the level of preferential sampling. Three levels are used: $b = 0$, which makes the sampling model conditionally independent from the abundance process (i.e., no preferential sampling; $b = 1$, which induces some dependence between the abundance and sampling process (i.e., moderate preferential sampling); $b = 5$ which makes cells of high abundance much more likely to be sampled than low abundance cells (i.e., pathological preferential sampling). An example of these alternative configurations is provided in the main manuscript. We drew $\beta_1^* \sim \text{Uniform}(-0.5, 0.5)$ for each simulation replicate.

4. For each of the 3 sampling configurations, we generated a count dataset. In each case, we generated a count Y_i for each cell selected for sampling (i.e., for cells where $R_i = 1.0$) according to

$$Y_i \quad \sim \quad \text{Binomial}(N_i, A).$$

Here, A represents the portion of the grid cell that is surveyed, and was set to 0.5 for all simulations and grid cells.

5. For each count data set, we fitted 2 different estimation models in TMB (see *Preferential sampling model development*. In the first, the preferential sampling parameter b was fixed to 0.0 (thus in effect assuming no preferential sampling). In the second, b was estimated (allowing a potential for preferential sampling).

6. For each combination of simulation replicate, survey configuration, and estimation model, we recorded relative bias with respect to total estimated abundance, $\hat{N} = \sum \hat{N}_i$, where a bias correction algorithm (Tierney *et al.* 1989; Thorson & Kristensen In Press) is used to obtain abundance estimates and standard errors that properly account for nonlinear transformations of random effects.

References

Kristensen, K., Nielsen, A. & Berg, C.W. (2015) Template Model Builder tmb. *Journal of Statistical Software*, In Press.

Lindgren, F., Rue, H. & Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(4), 423–498.

Royle, J.A. & Berliner, L.M. (1999) A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 29–56.

Thorson, J.T. & Kristensen, K. (In Press) Implementing a generic method for bias correction in statistical models using random effects, with spatial and popualtion dynamics examples. *Fisheries Research*.

Thorson, J.T., Skaug, H.J., Kristensen, K., Shelton, A.O., Ward, E.J., Harms, J. & Benante, J.A. (2015) The importance of spatial models for estimating the strength of density dependcence. *Ecology*, **96**, 1202–1212.

Tierney, L., Kass, R.E. & Kadane, J.B. (1989) Fully exponential Laplace approximations to expectations and vvariance of non positive functions. *Journal of the American Statistical Association*, **84**, 710–716.