Methods in Ecology & Evolution 2015 doi: 10.1111/...

ARTICLE

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*

Summary

1. Count and occupancy surveys are often used to estimate the density, abundance, and distribution of animal populations. Recently, model-based approaches to analyzing survey data (i.e. species distribution models) have become popular because one can more readily accommodate departures from pre-planned survey routes and construct more detailed maps than one can with design-based procedures. Model-based analysis often makes use of species-covariate relationships and/or spatially autocorrelated random effects to help predict density or occurrence in unsampled locations.
2. Species distribution models often make the implicit assumption that locations chosen for sampling and animal abundance are conditionally independent given modeled covariates. However, this assumption is likely violated in many cases when survey effort is non-randomized, leading to preferential sampling.
3. We develop a hierarchical statistical modeling framework for detecting and alleviating the biasing effects of preferential sampling in species distribution models. The approach works by jointly modeling animal abundance/occurrence and the locations selected for sampling, and specifying a dependent correlation structure between the two models.
4. Using simulation, we show that under moderate levels of preferential sampling, our modeling approach reduces bias from $\approx$40% (under a naive species distribution model) to $approx 5\%$.
5. Species case study
6. When animal population surveys are conducted without a well defined sampling frame, ecologists should recognize the potential for biasing effects of preferential sampling. Joint models, such as those described in this paper, can be used to test and correct for such biases.

Word count: XXXX

Key-words: count data, preferential sampling, spatial autocorrelation, species distribution model

*Correspondence author. E-mail: paul.conn@noaa.gov

Introduction

Surveys of unmarked animal populations are often used to estimate abundance and occurrence of animal populations and to predict species distributions, enterprises central to conservation, ecology, and management. For studies of abundance, researchers historically relied on design-based statistical inference (e.g. Cochran 1977), which requires adoption of a pre-defined sampling frame (e.g. using systematic random sampling, stratified random sampling, or some variant thereof). Designing animal surveys is relatively straightforward in such applications, and unbiased point and variance estimators are available. Recently, however, there has been a surge in research describing model-based procedures for estimating abundance, density, and occupancy from surveys of unmarked animals, including N-mixture and Dail-Madsen models for repeated point counts (Royle 2004; Dail & Madsen 2011), occupancy models for presence-absence surveys (MacKenzie *et al.* 2002; Johnson *et al.* 2013), and various model-based formulations for distance-sampling data (Hedley & Buckland 2004; Miller *et al.* 2013; Johnson *et al.* 2010). In such applications, it is common to use habitat or environmental covariates together with spatial effects (e.g. via trend surfaces or spatial random effects) to predict density or distributions across the landscape. We shall refer to the amalgam of model-based approaches for making spatially explicit inference about animal populations as "Species distribution models" (SDMs; *sensu* Elith & Leathwick 2009), even though this term is more often used to refer to animal occurrence than it is to density or abundance.

One of the main advantages of using SDMs advanced in the literature is that one is no longer beholden to predetermined sampling frames, and can potentially use data gathered from non-randomized designs or platforms of opportunity to make inferences about animal populations (Johnson *et al.* 2010). However, in a recent paper, Diggle *et al.* (2010) emphasized that spatially explicit statistical models can easily provide biased estimates when sampling disproportionately targets locations where the response of interest is higher (or lower) than expected given a particular set of explanatory covariates. In the context of SDMs, this might occur if sampling disproportionately occurs in locations where animals are known to be present or of high abundance, regardless of predictive covariates. For example, if volunteer inventory participants have access to multiple sites with similar covariate values, bias might arise if they consistently choose sites where species are thought or known to be present. Bias might also arise if surveying effort is higher near bases of operations, and if animal abundance is higher (or lower) near bases of operations than elsewhere in the landscape.

Need to acknowledge that "sample selection bias" has been addressed extensively in SDM modeling with presence-only data (e.g. adjusting "background" or "availability" data to be similar to the locations chosen for sampling).

In this article, we explore potential for bias in SDMs resulting from preferential sampling (hereafter, PS), and describe several model-based approaches for detecting and correcting for such biases. We start by describing a common currency for notation and basic model structures considered in this paper. Second, we review preferential sampling bias in a mathematical light, and describe prior approaches to coping with its effects. Third, we introduce a novel generalization of previously proposed PS models, allowing the investigator to jointly model animal encounter data and the locations chosen for sampling, including possible dependence structure between these two types of observations. Fourth, we conduct a simulation study to examine the performance of traditional SDMs and our newly developed PS model when data are gathered preferentially. Finally, we demonstrate utility of our proposed modeling approach by analyzing a data set of MYSTERY SPECIES X.

Materials and methods

NOTATION AND BASIC MODEL STRUCTURES

We focus here exclusively on discrete space (areal) models for animal encounter data as these seem to be the dominant form used in design and analysis of animal population surveys, although we note that preferential sampling is likely to affect analyses similarly regardless of the choice of spatial domain. 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. To formulate a "traditional" SDM, one could then write animal abundance or occurrence as a stochastic realization of a probability mass function $f()$:

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

In this example, Z_i denotes the state variable of interest (e.g., occupancy or abundance), $g()$ 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 write the 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 parameter, $\mathbf{x}_i$ is a row vector of m predictive covariates associated with site i, $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_m\}$ is a column vector of m regression parameters, and $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_S\}$ are spatially autocorrelated random effects. For occupancy, $f()$ would typically be Bernoulli, while the Poisson or negative binomial are typically choices for analysis of count data; common forms for λ_i include geostatistical specifications (Cressie 1993; Diggle *et al.* 1998), Gaussian Markov random fields (e.g. conditionally autoregressive models; Rue & Held 2005), or low rank alternatives such as predictive process (Banerjee *et al.* 2008; Latimer *et al.* 2009) or restricted spatial regression models (Reich *et al.* 2006; Hughes & Haran 2013).

The model for Z_i describes variation in the process of interest and is often described as the "process" model. However, it is usually impossible to observe the system perfectly even in locations where sampling occurs, so it is customary to include an observation model describing incomplete detection. For occupancy studies, the response variable $Y_i = 1$ if the species of interest is detected and is 0 otherwise, and is modeled with a Bernoulli distribution (Royle & Dorazio 2008)

$$Y_i \quad \sim \quad \text{Bernoulli}(Z_i p_i), \qquad \text{eqn 3}$$

where p_i is possibly a function of survey and observer specific covariates. Replicate surveys of the same sampling unit provide the necessary information to estimate p_i. For count surveys, a possible model is

$$Y_i \quad \sim \quad \text{Poisson}(Z_i A_i p_i), \qquad \text{eqn 4}$$

where the Y_i now represents the count of animals obtained while surveying unit i, A_i denotes the proportion of sample unit i that is surveyed, and p_i gives detection probability. Additional information will often be needed to estimate p_i in this context, such as data from double observers, distance observations, or double sampling (see e.g. Buckland *et al.* 2001; Royle *et al.* 2004; Borchers *et al.* 2006; Conn *et al.* 2014).

For the remainder of this treatment, we use bold symbols to denote vector-valued quantities or matrices. We also use standard bracket notation to denote probability mass and density functions. For instance $[\mathbf{Z}]$ denotes the marginal probability mass function for $\mathbf{Z}$, and $[\mathbf{Z}|\mathbf{Y}]$ represents the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$. We use $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ to denote log-scale abundance and the logit of the probability of sampling, so that $Z_i \sim f(\mu_i)$, and $Y_i \sim f(\nu_i)$.

PREFERENTIAL SAMPLING: A PRIMER

One of the appealing aspects of model-based estimation is that there is no requirement that surveys rely on a pre-planned survey design selected probabilistically from an underlying sampling frame. For instance, investigators can reallocate sampling effort if weather or logistics preclude surveying in a desired location. This can be a crucial advantage in surveys covering large areas with frequent inclement weather. It also opens the door for using platforms of opportunity, presence only, and citizen science data for estimation.

However, the manner in which effort is ultimately allocated can potentially have profound influence on SDM estimator performance. With respect to nonrandom sampling, two possible problems seem particularly likely in discrete spatial domains: coarse scale preferential sampling (CSPS), and fine scale preferential sampling (FSPS) (Fig. 1). FSPS arises when the observations taken at a particular sampling unit are non-random with respect to the density of animals within that sampling unit. For instance, when allocating line transect survey effort, it may be tempting to place the transect in a manner that targets habitat or landscape features that maximize the number of animals that will be encountered. Depending upon the interpretation of occupancy, this may or may not be reasonable. However, if trying to estimate density or abundance, this strategy will clearly lead to positive bias.

By contrast, CSPS (hereafter, PS), the primary focus of this article, arises when the locations being sampled and the process of interest (e.g. density, occupancy) are conditionally dependent given modeled covariates (Diggle *et al.* 2010). For instance, PS can occur when the investigator uses a priori knowledge or observations of the state variable obtained during sampling to allocate survey effort in places where abundance or occurrence is known to be high. Diggle *et al.* (2010) showed that this type of preferential sampling can lead to bias when this extra information is not included in models for the state variable of interest. Specifically, PS arises when we consider the set of sampled locations as stochastic and when $[\mathbf{R}, \mathbf{Z}|\mathbf{x}] \neq [\mathbf{R}|\mathbf{x}][\mathbf{Z}|\mathbf{x}]$ (Diggle *et al.* 2010), where $\mathbf{R}$ is an indicator vector taking on a value of $R_i = 1.0$ if sampling unit i is sampled and is zero otherwise. We use this definition of PS throughout the rest of the manuscript, noting that it is somewhat different than has sometimes been used in the SDM literature. For instance, Merckx *et al.* (2011) use the term "preferential sampling" to refer to the process of visiting some sites more often than others, while Manceur & Kühn (2014) define it as occurring when the locations selected for sampling are a function of an environmental covariate. Neither of these latter conditions are problematic outside of the specialized field of presence-only modelling.

Diggle *et al.* (2010) demonstrated PS with an environmental monitoring problem, whereby pollutant monitoring stations were more highly clustered around urban areas with high concentrations of pollutants than in rural areas with comparably low levels of pollutants. Fitting simple geostatistical models without fixed effects led to positively biased estimates of landscape-level pollutant concentrations. Presumably (and as noted by discussants of the article) including a fixed effect associated with a relevant covariate (e.g., an "urbanity" index) would likely reduce or eliminate bias. However, the primary point of Diggle *et al.* (2010) is well taken: inclusion of spatially autocorrelated random effects in a statistical model is insufficient to remove the potentially biasing effects of PS.

As in the pollution example, having good explanatory covariates may also reduce bias when fitting SDMs to animal encounter data under PS. However, in many ecological applications, predictive covariates are only able to explain a portion of variation present in the data. If the locations selected for sampling are related (intentionally or unintentionally) to some unmodelled factor related to abundance, bias may still occur. Despite the clear potential for bias in SDMs, we have been unable to find many cases where PS (*sensu* Diggle *et al.* 2010) is discussed with regard to SDMs. One such example is Chakraborty *et al.* (2010), who acknowledged the likely presence of PS when fitting SDMs to data obtained using nonrandomized designs, but did not attempt to model it.

In design-based sampling, unequal sampling intensity is often accommodated via stratification or unequal probability sampling, as with Horvitz-Thompson-like estimators where the probability of inclusion varies by sampling unit (Cochran 1977). However, in the case of PS, this inclusion probability also depends on the value of the response associated with the sampling unit. Evidently, any approach to account for PS should also account for the dependence between the state variable of interest and the locations chosen for sampling.

Several authors have attempted model-based corrections for PS in the statistical literature. For Gaussian models in a continuous spatial domain, Diggle *et al.* (2010) and Pati *et al.* (2011) jointly modeled the locations that are chosen for sampling and the underlying random field of interest. In particular, they expressed sampled locations as an inhomogeneous Poisson point process where the underlying log-scale intensity depended linearly on spatially-referenced random field values. For instance, writing observations of the spatial random field at a location i as

$$Z_i = \mu_i + \epsilon, \qquad \text{eqn 5}$$

the relative density of sampling locations at i would be written as

$$p_i \propto \exp(\xi_i + b\mu_i). \qquad \text{eqn 6}$$

Here, the parameter b describes the level of preferential sampling; $b = 0$ implies no preferential sampling, $b > 0$ implies a greater level of sampling in locations where the state variable is anomalously high, and $b < 0$ implies greater sampling where the state variable is anomalously low. Importantly, when explanatory covariates are used in models for μ_i and ξ_i, Pati *et al.* (2011) show that ". . . accounting for informative sampling is only necessary when there is an association between the spatial surface of interest and the sampling density that cannot be explained by the shared spatial covariates." Pati *et al.* (2011) also consider a simpler, plug-in based estimator, where the log of a nonparameteric estimate of sampling density (specifically, a two dimensional kernel density estimate) is used as an additional fixed effect in Eq. 5, finding that this approach helped reduce bias associated with preferential sampling, but did not perform as well as the full joint model.

A GENERALIZED PREFERENTIAL SAMPLING MODEL

The models considered by Diggle *et al.* (2010) and Pati *et al.* (2011) are a useful first step in addressing and modeling preferential sampling. However, they are somewhat limited since they are specific to continuous spatial domains, continuous data (as opposed to presence/absence or count data), and Gaussian error distributions. Also, they require the linear predictor of the preferential sampling model to be written as a simple linear function of the the spatial process model for density. In real world applications, we can envision cases where sampling is strongly preferential in certain areas of the landscape, and not in others. For instance, sampling may be more strongly preferential close to bases of operations, (e.g., landing strips in the case of aerial surveys), but less so in areas that are harder to get to.

Given these limitations, our present task is to generalize PS models to the types of data more typical of SDMs, and to allow the degree of PS to vary across the landscape. Like Diggle *et al.* (2010) and Pati *et al.* (2011), we impose a joint model for the process of interest (animal abundance or occurence) and the locations chosen for sampling. For the abundance process model, we start with eq. 1 as a general formulation for non-Gaussian data. We then write the link-scale expectation as in eq. 2. Next, recalling that R_i is a binary indicator taking on the value 1.0 if survey unit i is selected for sampling, and is 0.0 otherwise, we model R_i using a Bernoulli distribution:

$$R_i \sim \text{Bernoulli}(h^{-1}(\nu_i)), \qquad \text{eqn 7}$$

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

$$\nu_i = \beta_0^* + \mathbf{x}_i^* \boldsymbol{\beta}^* + \eta_i + \mathbf{B}\delta_i. \qquad \text{eqn 8}$$

In a similar fashion to the model for the state process, the sampling intensity model has an intercept (β_0^*), explanatory covariates ($\mathbf{x}_i^*$), fixed effect regression parameters ($\boldsymbol{\beta}^*$), spatially autocorrelated random effects (η_i and δ_i), and normally distributed error ε_i. The predictive covariates $\mathbf{x}_i$ from Eq. 2 and $\mathbf{x}_i^*$ from Eq. 8 need not be the same (although they can be). Note also that the spatially autocorrelated random effect δ_i is included in both Eqs. 2 and 8, allowing for dependency in the two models, with the matrix $\mathbf{B}$ describing the strength and type of dependence between the sampling process and underlying density. The spatially autocorrelated random effects η_i are assumed independent of the abundance process.

The formulation in Eq. 8 is the same as previously proposed by other authors for hierarchical multivariate models with spatial dependence (cf. Royle & Berliner 1999). There are multiple ways of structuring $\mathbf{B}$ depending on the complexity of spatial dependence desired for the preferential sampling process (Royle & Berliner 1999). For instance, setting $\mathbf{B} = \mathbf{O}_{S \times S}$ corresponds to an absence of spatial dependence (and thus no preferential sampling). Setting $\mathbf{B} = b\mathbf{I}$, where b is an estimated parameter and $\mathbf{I}$ is an $(S \times S)$ identity matrix corresponds to the linear preferential sampling model suggested by Diggle *et al.* (2010) and Pati *et al.* (2011). Alternatively, we could allow the degree of PS to vary across the landscape. For instance, one can contemplate a trend surface model for preferential sampling by specifying a diagonal matrix for $\mathbf{B}$, with entries given by $b_0 + b_1 \text{lat}_i + b_2 \text{long}_i$, where b_0, b_1, and b_2 are estimated parameters and lat_i and long_i give latitude and longitude, respectively (Royle & Berliner 1999). Theoretically, one could include more highly parameterized structures for spatial dependence, such as higher order trend surface or spline formulation (Royle & Berliner 1999), but the ability to robustly estimate the parameters of such a model is likely dependent on having a rich, spatially balanced dataset, which is often not the case in ecological applications.

A comparison of the performance of models with different sets of constraints on $\mathbf{B}$ can serve as a test of PS. In particular, if one can demonstrate that models with $\mathbf{B} = \mathbf{0}$ perform similarly or better than models with $\mathbf{B} \neq \mathbf{0}$, then PS is likely not worth modeling and inference can proceed using standard SDMs (i.e. not modeling sampling intensity).

SIMULATION STUDY

To illustrate PS and demonstrate that our proposed model has reasonable performance, we conducted a small simulation experiment. For each of 100 simulations, we generated abundance of a hypothetical species over a 25×25 grid as

$$N_i \sim \text{Poisson}(\mu_i),$$

where i indexes survey unit i, and μ_i is determined according to Eq. 2. Abundance was generated as a function of a single spatially autocorrelated landscape covariate, as well as residual spatial autocorrelation (δ_i) and overdispersion (fig. **??**). Specific details of of data generation procedures are presented in Appendix S1.

For each simulated landscape we generated three virtual count surveys using eqs. 7 and 8. Each survey had $\boldsymbol{\beta}^* = \eta_i = 0$ (that is, not covariate of spatially autocorrelated random effects), but differed in how the matrix $\mathbf{B}$ was parameterized. In the first, we set $\mathbf{B} = 0$, so that the survey was a simple random sample. For the second and third, we set $\mathbf{B}$ to be a diagonal matrix with entries $b = 1$ and $b = 5$, respectively, so that the probability of sampling a given survey unit (grid cell) was explicitly dependent on the latent abundance in that unit. We refer to these scenarios as moderate and pathological preferential sampling, respectively (see fig. 3). Simulations were configured so that $n = 50$ of the 625 survey units were sampled; each survey was set to cover half of the target cell.

We fitted two different models to each count dataset, both of which were provided the habitat covariate used (in part) to generate the data for which a log-linear coefficient β was estimated. In the first estimation model, the elements of **B** in eq. 8 were all set to zero. In this case, the abundance and sampling process submodels were independent, as is the case canonical SDMs (at lest when fitted to presence-absence or count data). In the second estimation model, we included an explicit connection between the distribution of animal abundance and the sampling process by setting $\mathbf{B} = b\mathbf{I}$, where b is an estimated parameter, and **I** is an identity matrix.

We used maximum marginal likelihood estimation to conduct statistical inference. In particular, we used Template Model Builder (TMB; Kristensen *et al.* 2015), interfaced with the R programming environment, to conduct maximization. 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) to obtain abundance estimates and standard errors that properly account for nonlinear transformations of random effects. This approach allowed for a facile implementation and fast computing times, allowing us to conduct simulation and model testing with greater efficiency than would have been possible with Bayesian simulation. In this study, we report the results of 500 simulation replicates. Further detail on statistical methods are provided in Appendix S1; 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/`.

BEARDED SEAL AERIAL SURVEYS

AVIAN POINT COUNTS

Results

SIMULATION STUDY

Estimates of cumulative animal abundance across simulated landscapes were median unbiased for both estimation methods when the sites selected for sampling were independent of animal density, though when b was estimated abundance estimates were more right skewed and had higher variance (fig. 4). Under moderate preferential sampling ($b = 1$), the jointly modeling the density and the sampling process (i.e., estimating b) led to a median bias of 5%, while the canonical SDM model had a median bias of 40%. Under pathological preferential sampling ($b = 5$), both estimation methods were extremely biased, but was even more severe for the naive model ignoring preferential sampling (fig. 4).

Discussion

Bias attributed to PS may seem counterintuitive, especially given the maxim in survey sampling to allocate more effort to strata for which animal density is high. For instance, in large scale line transect surveys under stratified sampling, the optimal amount of effort that should be allocated to stratum s is $A_s D_s^{0.5}$, where A_s is the area of s and D_s is the anticipated density (Buckland *et al.* 2001;

eqn 7.7). Thus, there are theoretical reasons to sample more in high density areas than in low density areas. The obvious solution in this instance is to compensate for PS in model-based inferences by accounting for variation in sampling intensity with explanatory covariates or post hoc stratification. However, this does not always work when effort is allocated in a subjective manner.

Extension to continuous space - sampling locations as point process similar to Warton and Shepherd (2010)

Presumably, trends less biased when pref sampling ignored.

Acknowledgements

Later.

References

Banerjee, S., Gelfand, A.E., Finley, A.O. & Sang, H. (2008) Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society B*, **70**, 825–848.

Borchers, D.L., Laake, J.L., Southwell, C. & Paxton, C.G.M. (2006) Accomodating unmodeled heterogeneity in double-observer distance sampling surveys. *Biometrics*, **62**, 372–378.

Buckland, S., Anderson, D., Burnham, K., Laake, J., Borchers, D. & Thomas, L. (2001) *Introduction to Distance Sampling: Estimating the abundance of biological populations*. Oxford University Press, Oxford, U.K.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M. & Silander Jr, J.A. (2010) Modeling large scale species abundance with latent spatial processes. *The Annals of Applied Statistics*, 1403–1429.

Cochran, W. (1977) *Sampling Techniques, 3rd Edition*. Wiley, New York.

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 multi-species abundance using automated detection systems: ice-associated seals in the eastern Bering Sea. *Methods in Ecology and Evolution*, **5**, 1280–1293.

Cressie, N.A.C. (1993) *Statistics for spatial data, revised edition.* Wiley, New York.

Dail, D. & Madsen, L. (2011) Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, **67**(2), 577–587.

Diggle, P.J., Tawn, J.A. & Moyeed, R.A. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**(3), 299–350.

Diggle, P.J., Menezes, R. & Su, T.l. (2010) Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**(2), 191–232, doi:10.1111/j.1467-9876.2009.00701.x, URL http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

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

Hughes, J. & Haran, M. (2013) Dimension reduction and alleviation of confounding for spatial generalized mixed models. *Journal of the Royal Statistical Society B*, **75**, 139–159.

Johnson, D.S., Conn, P.B., Hooten, M., Ray, J. & Pond, B. (2013) A probit approach for spatio-temporal modeling of ecological occupancy data. *Ecology*, **94**, 801–808.

Johnson, D., Laake, J. & Ver Hoef, J. (2010) A model-based approach for making ecological inference from distance sampling data. *Biometrics*, **66**, 310–318.

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

Latimer, A.M., Banerjee, S., Sang, H., Moshner, E.S. & Silander Jr., J.A. (2009) Hierarichical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northern United States. *Ecology Letters*, **12**, 144–154.

MacKenzie, D., Nichols, J., Lachman, G., Droege, S., Royle, J. & Langtimm, C. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

Manceur, A.M. & Kühn, I. (2014) Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. *Methods in Ecology and Evolution*, **5**(8), 739–750.

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M. & Vanaverbeke, J. (2011) Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, **222**(3), 588 – 597, doi:http://dx.doi.org/10.1016/j.ecolmodel.2010.11.016, URL http://www.sciencedirect.com/science/article/pii/S0304380010006216.

Miller, D.L., Burt, M.L., Rexstad, E.A. & Thomas, L. (2013) Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution*, **4**, 1001–1010.

Pati, D., Reich, B.J. & Dunson, D.B. (2011) Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, **98**(1), 35–48.

Reich, B., Hodges, J. & Zadnik, V. (2006) Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**, 1197–1206.

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.

Royle, J. (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60**, 108–115.

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

Royle, J. & Dorazio, R. (2008) *Hierarchical Modeling and Inference in Ecology*. Academic Press, London, U.K.

Rue, H. & Held, L. (2005) *Gaussian Markov Random Fields*. Chapman & Hall/CR, Boca Raton, Florida, USA.

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*.

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.

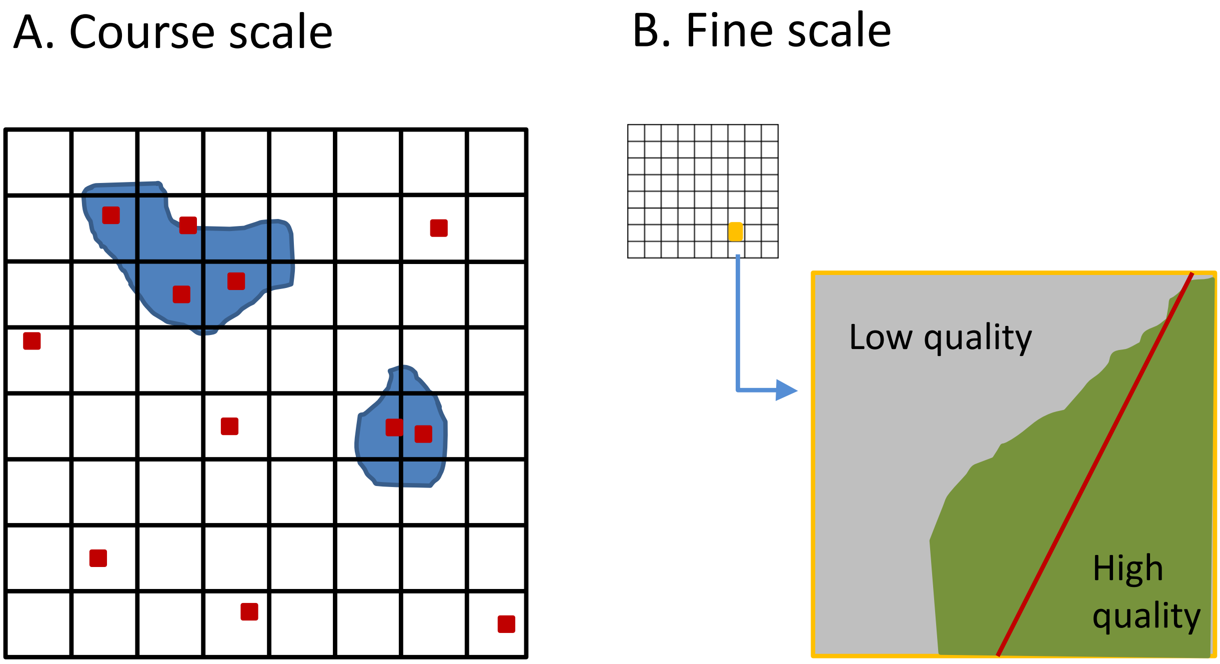

Fig. 1. A depiction of two types of preferential sampling. In (A), an investigator preferentially places point transects (red squares) within regions of high known animal density (blue polygons). This can cause bias in abundance or occupancy estimators unless this a priori knowledge about density is explicitly modeled. In (B), a fine scale version of preferential sampling occurs when a line transect (red line) is intentionally placed across a region of high quality habitat. If a landscape is discretized into homogeneous survey units (as in a grid), it is essential that the habitat surveyed within each survey unit be randomly determined when estimating abundance. If not, bias (usually positive) can be expected.

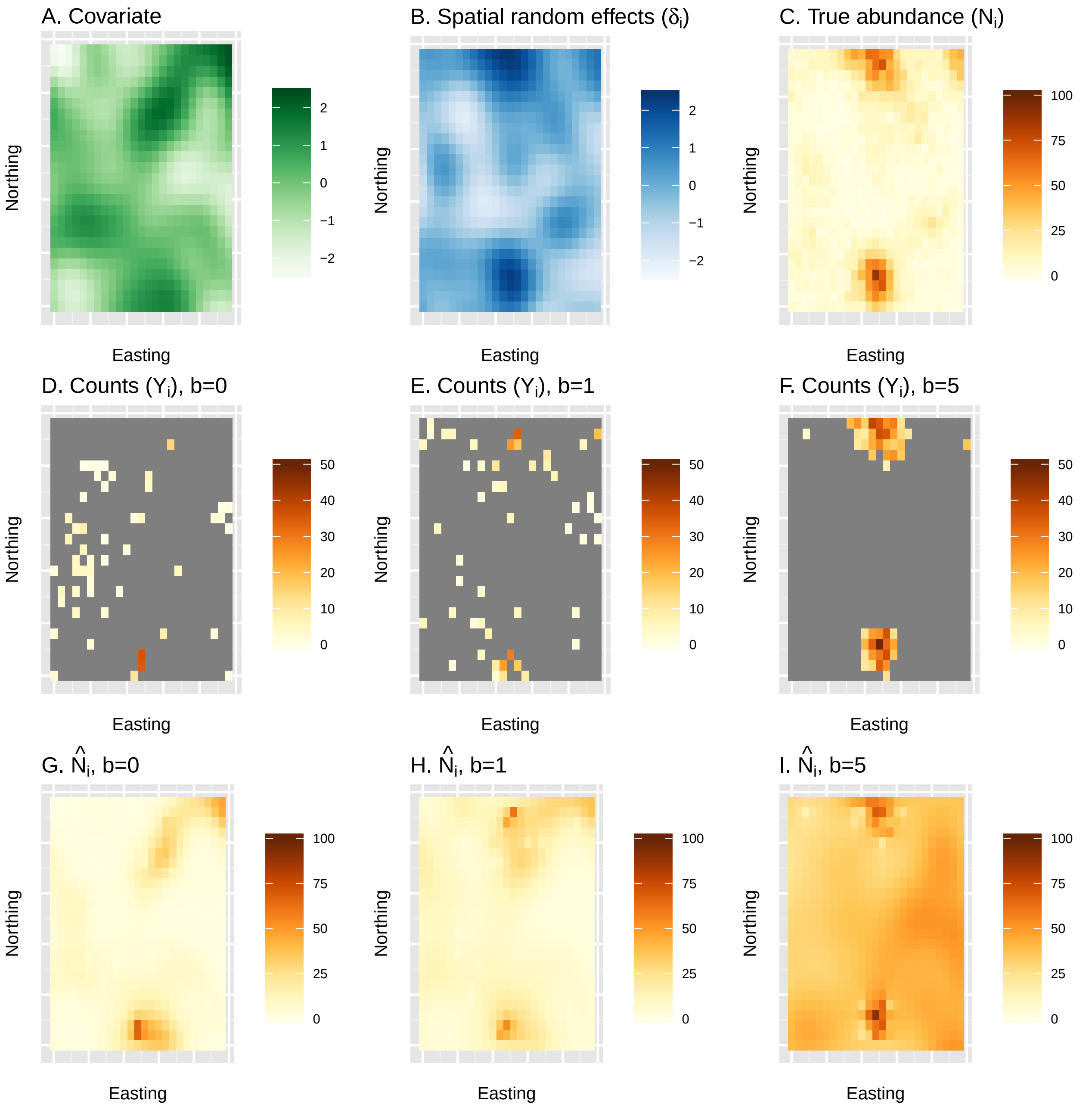

Fig. 2. An example of a single simulation replicate examining estimates of abundance from a naive species distribution model under preferential sampling. First, true abundance (C) is generated as a function of a spatially autocorrelated covariate (A) and a spatially autocorrelated random effect (B). Second, counts are generated for three different types of surveys, including a simple random sample ($b = 0$; D) and surveys with moderate ($b = 1$; E) or pathological ($b = 5$; F) levels of preferential sampling. Finally, spatially explicit estimates of abundance are generated using a traditional SDM (with b set to 0.0) to each of the count datasets (G-I). In this particular simulation replicate, cumulative abundance was underestimated by 18% when $b = 0$, overestimated by 17% when $b = 1$, and overestimated by 293% when $b = 5$. For a summary of bias over 500 simulation replicates, see fig. 4.

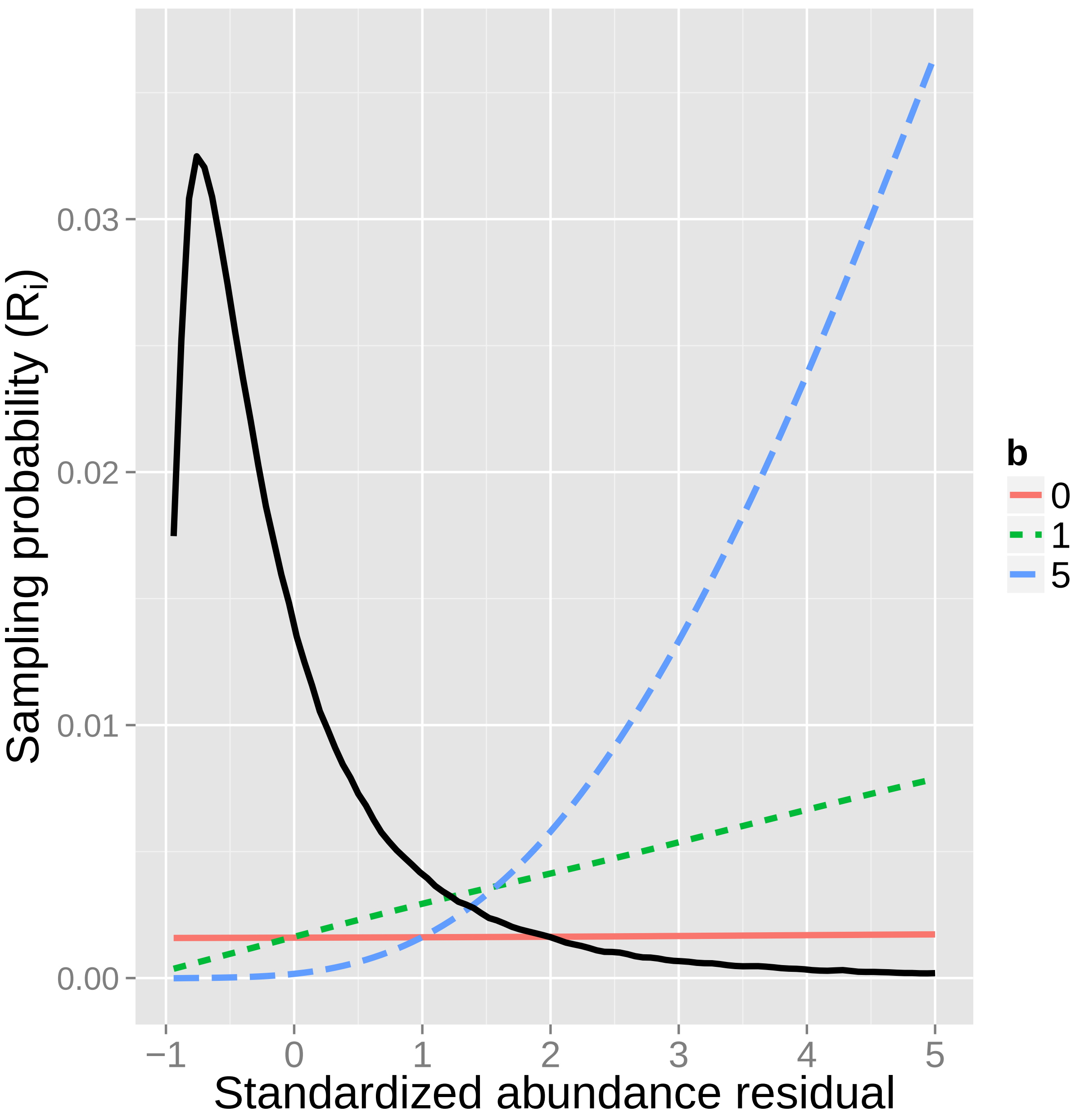

Fig. 3. Expected relationship between the probability of a survey unit being selected for sampling and its abundance residual in the simulation study. The base case $b = 0$ represents simple random sampling, while $b = 1$ and $b = 5$ represent moderate and pathological levels of preferential sampling, respectively. Also shown are is the realized distribution (smoothed histogram) of abundance residuals among survey units in the simulation study, scaled to fit in the plot margins (solid black line).

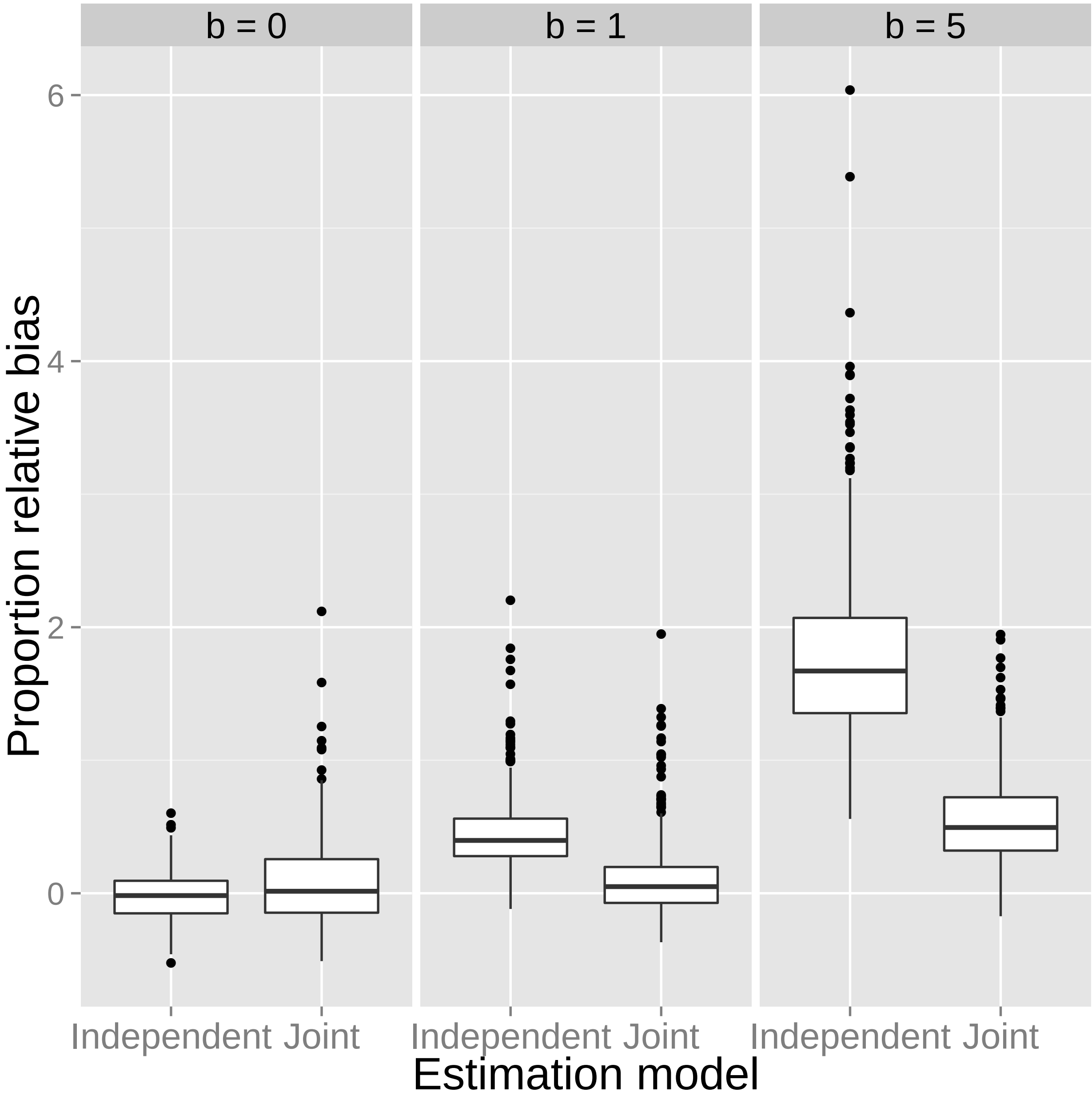

Fig. 4. Relative proportional error in abundance from the simulation experiment as computed with respect the posterior mode with a bias correction. Each boxplot summarizes the distribution of relative proportional error as a function of the type of sampling, including simple random sampling ($b = 0$), moderate preferential sampling ($b = 1$), and pathological preferential sampling ($b = 5$). Results vary by the type of estimation model; in the "independent" model, b is set to 0.0; in the "joint" model, b is estimated. Lower and upper limits of each box correspond to first and third quartiles, while whiskers extend to the lowest and highest observed bias within 1.5 interquartile range units from the box. Points denote outliers outside of this range. Horizontal lines within boxes denote median bias.