Ecology, 94(4), 2013, pp. 801–808
© 2013 by the Ecological Society of America

Spatial occupancy models for large data sets

Devin S. Johnson,[1,7] Paul B. Conn,[1] Mevin B. Hooten,[2,3,4] Justina C. Ray,[5] and Bruce A. Pond[6]

[1]*National Marine Mammal Laboratory, NOAA, 7600 Sand Point Way NE, Seattle, Washington 98115 USA*
[2]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado 80523 USA*
[3]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado 80523 USA*
[4]*Department of Statistics, Colorado State University, Fort Collins, Colorado 80523 USA*
[5]*Wildlife Conservation Society Canada, Toronto, Ontario M5S 2T9 Canada*
[6]*Wildlife Research and Development Section Ontario Ministry of Natural Resources, Peterborough, Ontario K9J 8M5 Canada*

Abstract. Since its development, occupancy modeling has become a popular and useful tool for ecologists wishing to learn about the dynamics of species occurrence over time and space. Such models require presence–absence data to be collected at spatially indexed survey units. However, only recently have researchers recognized the need to correct for spatially induced overdisperison by explicitly accounting for spatial autocorrelation in occupancy probability. Previous efforts to incorporate such autocorrelation have largely focused on logit-normal formulations for occupancy, with spatial autocorrelation induced by a random effect within a hierarchical modeling framework. Although useful, computational time generally limits such an approach to relatively small data sets, and there are often problems with algorithm instability, yielding unsatisfactory results. Further, recent research has revealed a hidden form of multicollinearity in such applications, which may lead to parameter bias if not explicitly addressed. Combining several techniques, we present a unifying hierarchical spatial occupancy model specification that is particularly effective over large spatial extents. This approach employs a probit mixture framework for occupancy and can easily accommodate a reduced-dimensional spatial process to resolve issues with multicollinearity and spatial confounding while improving algorithm convergence. Using open-source software, we demonstrate this new model specification using a case study involving occupancy of caribou (*Rangifer tarandus*) over a set of 1080 survey units spanning a large contiguous region (108 000 km^2) in northern Ontario, Canada. Overall, the combination of a more efficient specification and open-source software allows for a facile and stable implementation of spatial occupancy models for large data sets.

Key words: bias; caribou; intrinsic conditionally autoregressive; occupancy model; probit regression; Rangifer tarandus; *reduced rank; spatial regression*.

Introduction

Since the seminal work by MacKenzie et al. (2002), there has been an explosion in ecological studies designed to estimate occupancy and related parameters (see MacKenzie et al. [2005] and Long et al. [2008] for reviews). Occupancy, defined here as the probability that a focal taxa occurs (in some meaningful sense) in a survey unit, is commonly used for population monitoring and for assessing whether hypothesized covariates influence species presence or absence. The appeal of occupancy studies is undeniable, as inferences about population-level processes can be made without physically capturing and marking animals or otherwise censusing the population. Well-designed occupancy studies are particularly advantageous for assessing spatial distributions of wide-ranging, elusive species at the landscape scale, where more intensive sampling designs are costly and inefficient (e.g., Magoun et al. 2007, Karanth et al. 2011).

Typical occupancy study designs involve identification of a number of survey units or habitat patches within a larger study area. Each unit is visited by one or more trained observers, with a subset of units being visited more than once. The rationale for visiting a unit more than once is to help account for false negatives. If the species is truly absent, the probability of detecting the species is usually assumed to be zero (with the exception of a few studies, e.g., Royle and Link [2006], Hanks et al. [2011]). Observed absences are thus a mixture of true absences and nondetections; sampling a subset of units multiple times provides the information necessary to discriminate between the two. Although MacKenzie et al. (2002) is often cited as the initial paper on occupancy modeling, Hoeting et al. (2000) provides an earlier description of presence in the face of uncertain detection. Moreover, the model Hoeting et al. (2000) proposed was also spatially explicit.

Manuscript received 7 May 2012; revised 13 November 2012; accepted 14 November 2012. Corresponding Editor: E. G. Cooch.
[7] Email: devin.johnson@noaa.gov

The usual approach to estimating occupancy and related parameters (e.g., as implemented in program MARK [White and Burnham 1999] or PRESENCE [MacKenzie et al. 2002]) is through a product multinomial likelihood function. Although this approach yields unbiased inference when occupancy observations at nearby units are conditionally independent given any available covariates, residual spatial autocorrelation may lead to biases and overestimated precision. As a result, a few researchers have modeled occupancy within a hierarchical spatial framework, which explicitly acknowledges the presence of autocorrelation in the occupancy process (e.g., Hoeting et al. 2000, Hooten et al. 2003, Sargeant et al. 2005, Magoun et al. 2007, Royle and Dorazio 2008, Gardner et al. 2010, Aing et al. 2011, Chelgren et al. 2011, Karanth et al. 2011). Incorporating spatial structure also reflects the intuitive notion that adjacent survey units are likely to exhibit occupancy rates more similar than those far apart. However, a key limitation with previously proposed spatial approaches is that computing time can be prohibitive, especially for large study areas including thousands of units. As management agencies collect occupancy data over larger and larger spatial extents to address broad-scale, long-term monitoring needs, computational efficiency is becoming increasingly important. Further, recent research on spatial regression analysis (Reich et al. 2006, Hodges and Reich 2010, Paciorek 2010, Hughes and Haran 2013) has found potential for confounding between regression predictors and latent spatial processes, which may induce bias and unnecessarily over inflate uncertainty in regression coefficient estimates. This confounding is due to the fact that, in a spatial setting, regression predictor variables are themselves often spatially autocorrelated, which can produce strong cross-correlation between the predictor and a latent spatial effect. The confounding issue is relatively new to the spatial statistics literature and it has potential to impact any ecological analysis that uses a spatial regression approach for analyzing data, not just occupancy models. For example, GLMs with a spatially correlated random effect fall into this class of analyses (Hughes and Haran 2013).

In this paper, we develop a novel approach for analyzing occupancy models in the face of spatial autocorrelation over large extents. Exploiting data augmentation techniques (e.g., Royle and Dorazio 2008), we cast occupancy estimation as a probit regression problem, a model choice that dramatically reduces computational burden. Spatial autocorrelation is accounted for by inclusion of a *restricted* intrinsic conditionally autoregressive (ICAR) prior distribution on a set of spatial random effects. Under this framework, issues with parameter confounding can easily be resolved by requiring orthogonality between the spatial process and linear predictors. We demonstrate the utility of our modeling approach with a case study concerned with occupancy of woodland caribou (*Rangifer tarandus caribou*) in northern Ontario, Canada.

Methods

Presence–absence occupancy data are thought to arise via an interaction of two conceptually distinct processes: (1) an underlying occupancy process and (2) an observation process. Let the true occupancy for survey unit $s = 1, \ldots, n$ be z_s, where $z_s = 0$ denotes an unoccupied unit and $z_s = 1$ denotes an occupied unit. Assuming that there are no false positives, the observed occupancy of unit s at time t, say y_{st}, will be 0 whenever $z_s = 0$. However, if $z_s = 1$, y_{st} may either be 0 (species not detected) or 1 (species detected). Observed occupancy is usually measured on a subset of units and times.

We begin by specifying a joint model with the hierarchical form $[y_{st}, z_s] = [y_{st} \,|\, z_s] \times [z_s]$ where $[\cdot \,|\, \cdot]$ generically represents a conditional distribution. A standard mixture-model specification for hierarchical occupancy models (e.g., Royle and Dorazio 2008) is

$$\begin{aligned}[y_{st} \,|\, z_s, p_{st}] &= \text{Bernoulli}(z_s p_{st}) \\ [z_s \,|\, \psi_s] &= \text{Bernoulli}(\psi_s) \end{aligned} \quad (1)$$

where p_{st} and ψ_s are, respectively, the probability of detection (given occupancy) for unit s at time t and the probability of occupancy at unit s. Typically, the number of parameters is reduced using a linear model of external covariates. For example, $\ell(\psi_s) = \mathbf{x}_s'\boldsymbol{\gamma}$, where $\mathbf{x}_s'$ is a vector of habitat related covariates for each unit, $\boldsymbol{\gamma}$ is a vector of occupancy coefficients, and $\ell(\cdot)$ is a link function to constrain ψ_s to the [0, 1] interval. The same holds true for $\ell(p_{st}) = \mathbf{x}_{st}'\boldsymbol{\beta}$, where a researcher may have covariates, $\mathbf{x}_{st}$, for unit s at time t that are thought to influence detection and $\boldsymbol{\beta}$ is a vector of regression coefficients.

A probit model for occupancy and detection

Traditionally the logit function has been chosen for the link function $\ell(\cdot)$ (i.e., $\log\{\psi_s/(1 - \psi_s)\} = \mathbf{x}_s'\boldsymbol{\gamma}$. We, however, consider the use of the probit link $\Phi^{-1}(\psi_s) = \mathbf{x}_s'\boldsymbol{\gamma}$, where Φ is the cumulative distribution function of a standard normal variable and Φ^{-1} is the quantile function. For most of the range [0,1], the probit and logit functions are linearly related, making it difficult to distinguish between the two functions in analysis of real data (McCullagh and Nelder 1999). Nevertheless, as we shall show, the probit model confers increased computational efficiency and greater flexibility through a data augmentation approach, which is not applicable when a logit link is used. Dorazio and Rodriguez (2012) independently pursued a similar approach.

The probit model can be formulated in a data augmentation framework by considering a continuous latent version, $\tilde{z}_s$ of the binary occupancy process z_s. Let $[\tilde{z}_s \,|\, \boldsymbol{\gamma}] = \mathcal{N}(\mathbf{x}_s'\boldsymbol{\gamma}, 1)$ where $\mathcal{N}(\mu, \tau)$ represents a normal distribution with mean μ and precision (inverse variance) parameter τ. Letting z_s be the indicator of $\tilde{z}_s > 0$ it

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [06/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

follows that $[z_s \mid \boldsymbol{\gamma}] = \text{Bernoulli}(\Phi(\mathbf{x}_s'\boldsymbol{\gamma}))$. This augmented data approach was initially proposed as a method for performing regression analysis of binary data with a Gibbs sampler (Albert and Chib 1993). Here we extend this approach to occupancy models where the zeros of the response variable cannot be observed due to imperfect detection. Analogous to $\tilde{z}_s$, let $\tilde{y}_{st}$ be a continuous version of y_{st}. Because $y_{st} = 0$ with probability 1 if $z_s = 0$,

$$[\tilde{y}_{st} \mid z_s, \boldsymbol{\beta}] = z_s \mathcal{N}(\mathbf{x}_{st}\boldsymbol{\beta}, 1) + (1 - z_s)\delta_0$$

where δ_0 is a degenerate distribution with point mass 1 at zero. Defining y_{st} to be the indicator of $\tilde{y}_{st} > 0$, it follows that $[y_{st} \mid z_s, \boldsymbol{\beta}] = \text{Bernoulli}(z_s\Phi(\mathbf{x}_{st}'\boldsymbol{\beta}))$.

Spatial regression models

We have previously noted that we can represent the occupancy probabilities as $\Phi(\psi_s) = \mathbf{x}_s'\boldsymbol{\gamma}$ where $\mathbf{x}_s$ is a vector of covariates collected at unit s that can help inform the occupancy process. However, because spatial clustering is ubiquitous in ecology, we often have reason to suspect that the occupancy process also possesses spatial autocorrelation. In order to formalize this notion, we introduce a spatial random effect into the $\tilde{z}_s$ model:

$$\tilde{z}_s = \mathbf{x}_s'\boldsymbol{\gamma} + \eta_s + \varepsilon_s \tag{2}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ is the realization of a Gaussian spatial process (see Banerjee et al. [2004] for examples and references) and $[\varepsilon_s] = \mathcal{N}(0, 1)$ The general Gaussian process implies that $[\boldsymbol{\eta} \mid \tau] = \mathcal{N}(0, \tau\mathbf{Q})$ where $\mathbf{Q}$ is a precision matrix and τ is a scalar precision parameter.

There are many forms of Gaussian process that we could employ to invoke the desired association structure. For example, to model spatially correlated unit effects, Hooten et al. (2003) and Chelgren et al. (2011) used a geostatistical model on a continuous spatial domain, whereas, Magoun et al. (2007), Gardner et al. (2010), and Aing et al. (2011) used a *conditionally autoregressive* (CAR) model on a discrete spatial domain. For the models developed here, we consider only the discrete spatial domain as occupancy study areas often represent tessellations of areas for which a prediction of the occupancy process is desired (e.g., Hooten et al. 2003, Sargeant et al. 2005, Magoun et al. 2007). The typical CAR model is described by the precision matrix $\mathbf{Q} = \mathbf{D} - \rho\mathbf{A}$, where $\mathbf{A}$ is the association matrix with the i, j entry $A_{ij} = 1$ if units i and j are neighbors and zero elsewise, ρ is a spatial association parameter in [–1, 1] (Banerjee et al. 2004), and $\mathbf{D}$ is a diagonal matrix with ith entry $D_{ii} = \sum_j A_{ij}$ The CAR model can be made substantially more computationally efficient by setting $\rho = 1$ (*intrinsic* CAR or ICAR [e.g., Rue and Held 2005]). There has been some discussion as to appropriateness of the ICAR model (e.g., Wall 2004). Besag (2002), notes, however, the main purpose of a spatial model in these settings is to absorb extra spatial variation in order to share information between adjacent locations rather than produce meaningful parameter estimates for a spatial model. To this end, the ICAR model serves its purpose as the random effect at each site is modeled using the mean of the neighboring sites. See *Discussion*, however, for additional modeling possibilities.

While Eq. 2 seems straightforward, recent studies have shown that there can be substantial confounding between the spatial process and the fixed-effects covariates in spatial regression models (Reich et al. 2006, Hodges and Reich 2010, Paciorek 2010, Hughes and Haran 2013). For ecological inference on relationships between habitat covariates and occupancy (i.e., $\boldsymbol{\gamma}$ inference), this can be problematic. In order to solve the confounding issue as well as produce a reduced dimension model that is more computationally efficient, Hughes and Haran (2013) expanded upon work by Reich et al. (2006) and Hodges and Reich (2010) to develop a *restricted spatial regression* (RSR) model. The following provides a brief description of the RSR model; complete derivations and discussion are given in Hughes and Haran (2013). To begin, we assume that $\boldsymbol{\eta}$ follows an ICAR distribution, and reparameterize Eq. 2 in vector form as

$$\tilde{\boldsymbol{z}} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{3}$$

where the rows of $\mathbf{X}$ are $\mathbf{x}_s'$, $[\boldsymbol{\alpha} \mid \tau] = \mathcal{N}(\mathbf{0}, \tau\mathbf{K}'\mathbf{Q}\mathbf{K})$, $\mathbf{Q}$ is the n-dimensional ICAR precision matrix, $[\boldsymbol{\varepsilon}] = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and finally, $\mathbf{K}$ satisfies $\mathbf{K}\mathbf{K}' = \mathbf{I}$. Now, to specify a desirable $\mathbf{K}$ consider the Moran operator matrix $\boldsymbol{\Omega} = n\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}/\mathbf{1}'\mathbf{A}\mathbf{1}$, where $\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix onto the residual space of $\mathbf{X}$. The eigenvalues λ_i and eigenvectors $\mathbf{k}_i$ of $\boldsymbol{\Omega}$ have special meaning. The λ_i represent all of the possible values for Moran's I statistic for a spatial process that is orthogonal to $\mathbf{X}$ and defined by the association matrix $\mathbf{A}$ (Boots and Tiefelsdorf 2000). The Moran's I statistic is a common measure of spatial autocorrelation that can be interpreted on the same scale as a correlation coefficient (Moran 1950). In addition, the eigenvectors $\mathbf{k}_i$ represent all of the patterns of spatial dependence residual to $\mathbf{X}$ for an ICAR model defined by the neighborhood matrix $\mathbf{A}$ (Boots and Tiefelsdorf 2000). Therefore, if one sets $\mathbf{K} = [\mathbf{k}_1, \ldots, \mathbf{k}_q]$ where $q \ll n$ is chosen such that λ_q is greater than some threshold of meaningful autocorrelation one obtains a restricted version of the ICAR model that is not confounded with $\mathbf{X}$ and does not use information contained in the data attempting to estimate spatial structure too fine-grained to be meaningful. The choice of q is up to the researcher, as $q \to n$ we obtain the original ICAR model again. Certainly, one would want to choose $\lambda_i > 0$ as negative autocorrelation is not desired for these type of models. In fact the RSR model appears robust to even high levels of dimension reduction; simulation studies that examined retaining only $q = 50$ columns illustrated little change to the estimated $\boldsymbol{\eta}$ process (Hughes and Haran 2013).

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [10/05/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Bayesian inference

As with previous occupancy models containing complex autocorrelation structures (e.g., Hoeting et al. 2000, Sargeant et al. 2005, Magoun et al. 2007, Gardner et al. 2010), we adopt a Bayesian perspective and use Markov chain Monte Carlo (MCMC) for inference. Bayesian inference is based on the joint posterior distribution, which, in our case, may be written as

$$[\tilde{\mathbf{z}}, \mathbf{z}, \tilde{\mathbf{y}}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau \,|\, \mathbf{y}] \propto [\mathbf{y} \,|\, \tilde{\mathbf{y}}]\, [\tilde{\mathbf{y}} \,|\, \boldsymbol{\beta}, \mathbf{z}]\, [\mathbf{z} \,|\, \tilde{\mathbf{z}}]\, [\tilde{\mathbf{z}} \,|\, \boldsymbol{\gamma}, \boldsymbol{\eta}]\, [\boldsymbol{\eta} \,|\, \tau]\, [\boldsymbol{\beta}]\, [\boldsymbol{\gamma}]\, [\tau]. \quad (4)$$

Here, the distributions $[\boldsymbol{\beta}]$, $[\boldsymbol{\gamma}]$, and $[\tau]$ are the prior distributions for the parameters, and $[\mathbf{y} \,|\, \tilde{\mathbf{y}}]$ and $[\mathbf{z} \,|\, \tilde{\mathbf{z}}]$ are degenerate distributions as they represent deterministic functions. Inference about relevant parameters is achieved by cyclically sampling each parameter from its full conditional distribution.

Owing to our judicious choice of link function, all the necessary full conditional distributions are available in closed form, and we are able to sample from them directly (i.e., Gibbs sampler). This is in contrast to previously published formulations based on the logit link function, which require accept/reject algorithms such as a Metropolis-within-Gibbs sampler (see Givens and Hoeting 2005) that can slow convergence and require extensive tuning. Our resulting MCMC algorithm is much more efficient and capable of fitting spatial occupancy models on larger domains than before. In order to realize this result, we have to specify the conjugate priors $[\boldsymbol{\beta}] = \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}_\beta)$, $[\boldsymbol{\gamma}] = \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}_\gamma)$ and $[\tau] = \text{Gamma}(a, b)$, where a and b are the shape and scale parameters of a gamma distribution, respectively. This is not as restrictive as it appears as these can be made sufficiently vague to impart little prior knowledge of the parameter values. Appendix A gives the details for implementing the Gibbs sampler.

Case Study: Caribou occupancy in Ontario, Canada

Here we present an application of the methodology and model we propose in the previous section on a real data set from woodland caribou surveys. Sedentary woodland caribou range across boreal forests of North America at low densities, and therefore present challenges with respect to broad-scale population monitoring. Obtaining information on population dynamics of this species is only feasible in relatively small areas; however, accurate knowledge of caribou distribution, probability of occurrence and identification of core areas (and their changes over time), can serve as practical indicators for monitoring population conservation and recovery.

Survey data and model specifics

During 2009, winter aerial surveys were undertaken in a 108 000 km^2 study area in the Boreal Shield and the Hudson Plain ecozones of northwestern Ontario (51.5°–55°N and 86°–92°W). Water bodies comprised approximately 8.4% of the study area, and other open habitat types included fen, bog, sparse forest, and recent burns. Surveys were undertaken in February and March using two types of fixed-wing aircraft: (1) two Piper PA-18 Super Cubs equipped with wheel-skis, pilot, and one observer and (2) de Havilland Turbo Beavers on skis with pilot, lead observer/navigator, and one observer on each side of the aircraft. Groundspeed averaged 117 ± 13 km/h (mean ± SD) for the Super Cubs and 150 ± 28 km/hr for the Turbo Beaver. Survey altitude was similar for both aircraft, 89 ± 32 m for the Super Cubs and 90 ± 43 m for the Turbo Beaver.

The study area was divided into a tessellation of 100-km^2 hexagonal survey units based on Magoun et al. (2007). The hexagon design allows units to be equidistant (measured at the centroid) from one another, so that neighbors are evenly spaced. Survey units were searched by air for caribou and their sign, with multiple visits to a subset (see details in Magoun et al. [2007]). Up to 50 units per day were surveyed. Overall, 729 of the 1080 units were searched an average of 2.15 times (range, 1–14 times; Appendix B: Fig. B1). Fig. 2a illustrates the units that were surveyed as well as those with confirmed occupancy (i.e., at least one $y_{st} = 1$).

The occupancy process ψ_s is the primary interest in this analysis, however, a detection model must be specified. There were several covariates that were thought to influence detection of caribou tracks in the snow. The first of these is the percentage of open habitat in a 2 km wide corridor surrounding the flight path. Second, it was thought that the average airspeed of the aircraft when flying over the unit would influence the detection. This covariate generally reflects an aircraft effect between the two aircraft types, but we used airspeed rather than a categorical variable in case it was possible to see small changes within each aircraft using an overall numeric covariate. Finally, time of day was included as a quadratic function of the hour since noon. Separate quadratic functions were used for each aircraft.

In order to better estimate occupancy and determine what habitat characteristics increase probability of occupancy we chose three biologically motivated habitat covariates to use in the fixed effects portion of the ψ_s model (see Appendix B: Fig. B2). The remoteness of the study area precluded the use of forest inventory data, from which caribou-relevant boreal forest habitat variables are often derived, particularly forest age. At large spatial extents in winter, boreal forest-dwelling caribou tend to select habitats (e.g., peatland complexes) that support few predators or other ungulates and avoid recent burns that destroy their principal winter food source (lichen) and result in young seral stands (Bergerud et al. 1984, Schaefer and Pruitt 1991). Accordingly, the first covariate was the proportion of each hexagon classified as bog habitat (bog), and the second covariate was the proportion of each hexagon classified as recently burned (burned between 1990–2000; fire). Both were derived from provincial landcover mapping based on

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [06/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

classified LANDSAT imagery (Spectranalysis 2000). We also included a terrain ruggedness index measured by the standard deviation of elevation in each hexagon (sd-elev) calculated from a coarse-grained (760-m) provincial digital elevation model. In addition to the habitat characteristics, the x and y coordinates were also added to account for any systematic north–south and east–west trends in occupancy.

For the spatial process portion of the model we chose a RSR model with $\lambda_i \geq 0.5$, as this provided an acceptable balance between flexibility of the spatial process and minimizing the rank of the spatial process for faster MCMC sampling. This choice resulted in a $\mathbf{K}$ matrix with $q = 169$ columns (see Appendix B: Fig. B3a–e for illustration of some frequencies). Experimentation with different q values demonstrated little sensitivity to overall results for the model we fit here. In addition to the RSR model, we also analyzed the caribou data using a traditional ICAR model as well as a nonspatial ($\boldsymbol{\eta} = \mathbf{0}$) probit model.

The model parameters in all analyses were estimated using the Gibbs sampler described in Appendix A. Flat prior distributions were chosen for both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (i.e., $\mathbf{Q}_{\boldsymbol{\beta}} = \mathbf{Q}_{\gamma} = \mathbf{0}$). A Gamma(0.5, 0.0005) distribution was chosen for [τ] (Hughes and Haran 2013). The shape and rate parameters chosen for the τ prior heavily weight large values that a priori imply less spatial autocorrelation. Therefore, any observed spatial effect in the posterior implies strong evidence of spatial autocorrelation. An add-on package stocc for the R statistical environment (R Development Core Team 2011) was developed for fitting the proposed models. The package is available on the Comprehensive R Archive Network (CRAN; *available online*).[8] The Gibbs sampler was run for 50 000 iterations following a burn-in of 10 000 iterations that were discarded. Every fifth sample was retained for inference due to storage constraints for a total posterior sample of 10 000.

Results

Concerning the comparative performance of the three models, we obtained qualitatively identical results to Hughes and Haran (2013) with respect to bias and variability of the fixed effects coefficients for the occupancy model coefficients. The ICAR model produced posterior means for $\boldsymbol{\gamma}$ coefficient values substantially different from the nonspatial and RSR models (Fig. 1). In addition, the posterior variation was much larger for the ICAR model. On a numerical note, the autocorrelation within the $\boldsymbol{\gamma}$ chain was evidently much higher for the ICAR model; providing further evidence that confounding was present in the analysis using the ICAR model (Appendix B: Fig. B4). Therefore, for the remainder of the paper we will just consider results from the RSR model. The posterior distribution of spatial variance parameter $\sigma = 1/\sqrt{\tau}$ was sufficiently far from zero (posterior 90% credible interval of 1.23–12.23 vs. prior 90% credible interval of 0.00–0.06); implying strong evidence that the spatial process was significantly contributing to the overall variability of the occupancy process across the study area.

The parameter estimates for the detection model ($\boldsymbol{\beta}$) are given in Table B1 of Appendix B. As one might expect, more open habitat around the flight path produced higher detection probabilities, while faster average speed over the unit reduced the probability of detection. The Super Cub crews had an overall higher (though not significantly) detection probability than those of the Turbo Beaver, which had significantly lower detection probabilities in the early morning and late afternoon (Appendix B: Fig. B5). There was no evidence of time-of-day effects for the Super Cub crews. It is unclear if differences in the characteristics of the platforms or differences in crew ability contributed to this result.

The parameter estimates for the occupancy portion of the model ($\boldsymbol{\gamma}$) are given in Table B2 of Appendix B. There was little evidence of an east-west trend in the data; however, there was a significant north-south trend in occupancy with higher ψ_s in the northern region of the study area. The roughness of the terrain (sd-elev) was not significantly associated with caribou occupancy. There was a significant amount of evidence to infer a positive relationship between bog and occupancy and a negative relationship between occupancy and fire. The posterior mean occupancy probability map is shown in Fig. 2b. Generally, the model fits well when visually compared to the observed occupancy (Fig. 2a). In addition, Fig. 2c shows the posterior CV of the detection probabilities. As one can see there is more uncertainty in regions that are spatially far from sampled units, as well as, regions of sparse sampling effort and unobserved occupancy. The posterior mean of the $\boldsymbol{\eta}$ process (Fig. 2d) illustrates areas for which the covariates underpredict occupancy (red) and overpredict occupancy (blue). For example, the northwest panhandle of the study area should have relatively high occupancy probabilities based on the fixed-effects portion of the model. Negative values for the $\boldsymbol{\eta}$ process in this region, however, illustrate that the covariates produce ψ_s values too high for the observed occupancy.

Discussion

Ecological data often include spatial structure (i.e., patchiness). There is no reason to suspect that occupancy data sets are an exception, and thus, spatial correlation will often be present. However, despite this expectation, there have been relatively few studies that have formally accounted for this dependence in a statistical sense. This is notable because the failure to account for spatial autocorrelation can result in overstated measures of precision. Conversely, haphazard correction for spatial autocorrelation can produce bias and uncertainty overinflation of estimated parameters. In addition to habitat inference there are

[8] http://cran.r-project.org/

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [10/05/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

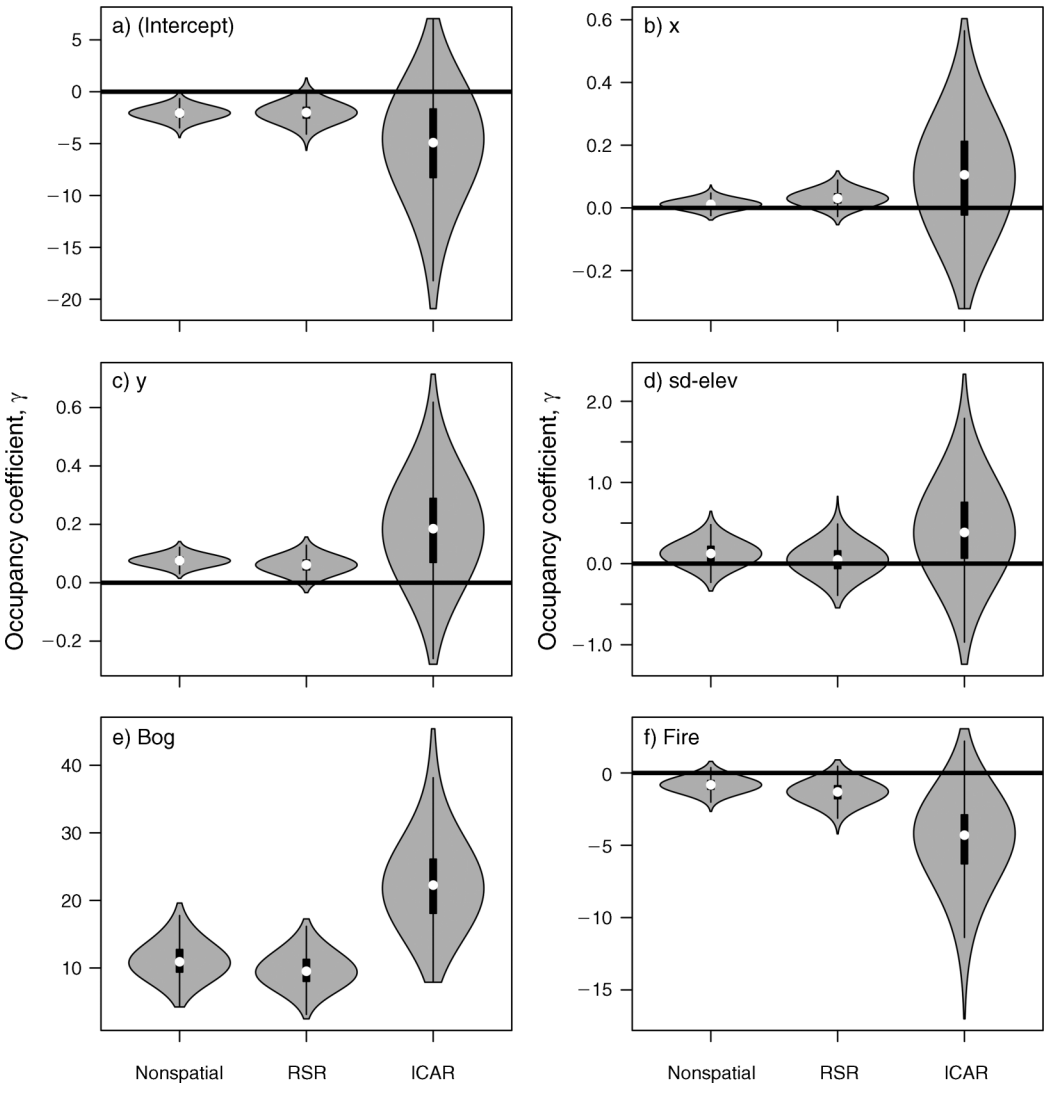

FIG. 1. Violin plots of the Gibbs sampler output for the occupancy coefficients (γ; no units), on the y-axis, from each of the three analyses. A violin plot is a vertical mirror image of the kernel density plot for a sample that gives slightly more information on distributional shape than a boxplot. The box and bar in the middle of the plot are the standard boxplots, with the white dot representing the posterior median, the box is drawn to the posterior quartiles, and the whiskers are drawn to the last datum within the median $\pm$ 1.5 times the interquartile range. Horizontal lines are drawn at zero on the y-axis. Each of the panels (a–f) represent the coefficient for each of the habitat covariates used in the occupancy portion of the caribou survey model; for panel (d) "sd-elev" indicates the standard deviation of elevation (see *Case study: caribou occupancy in Ontario, Canada: Survey data and model specifics*).

implications for occupancy study design. By exploiting autocorrelation when present, one can potentially decrease the number of visits to specific survey units within a study area, instead relying on spatial dependence to take the place of some temporal replication.

Of existing occupancy studies that have explicitly accounted for spatial autocorrelation, many have assumed a formulation that makes analyzing large data sets a challenge, has poor MCMC mixing properties, or has the potential for confounding between fixed effects and the spatial process. In contrast, we have proposed a unifying spatial occupancy model that suffers from none of these drawbacks. In particular, the simple switch from a logit to a probit link function yielded full conditional distributions that are in closed form, where MCMC was exceedingly efficient, and where extensions such as restricted spatial regression (RSR) could easily be employed to resolve parameter confounding. With a few exceptions (e.g., Hooten et al. 2003, Dorazio and Rodriguez 2012), it is surprising that this technique has not been explored further in Bayesian analysis of binary data in ecological settings.

Although we have chosen to implement RSR via an ICAR formulation, several other choices are possible. Wikle (2010) provide a general description of reduced rank spatial models, some of which could possibly be

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [06/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

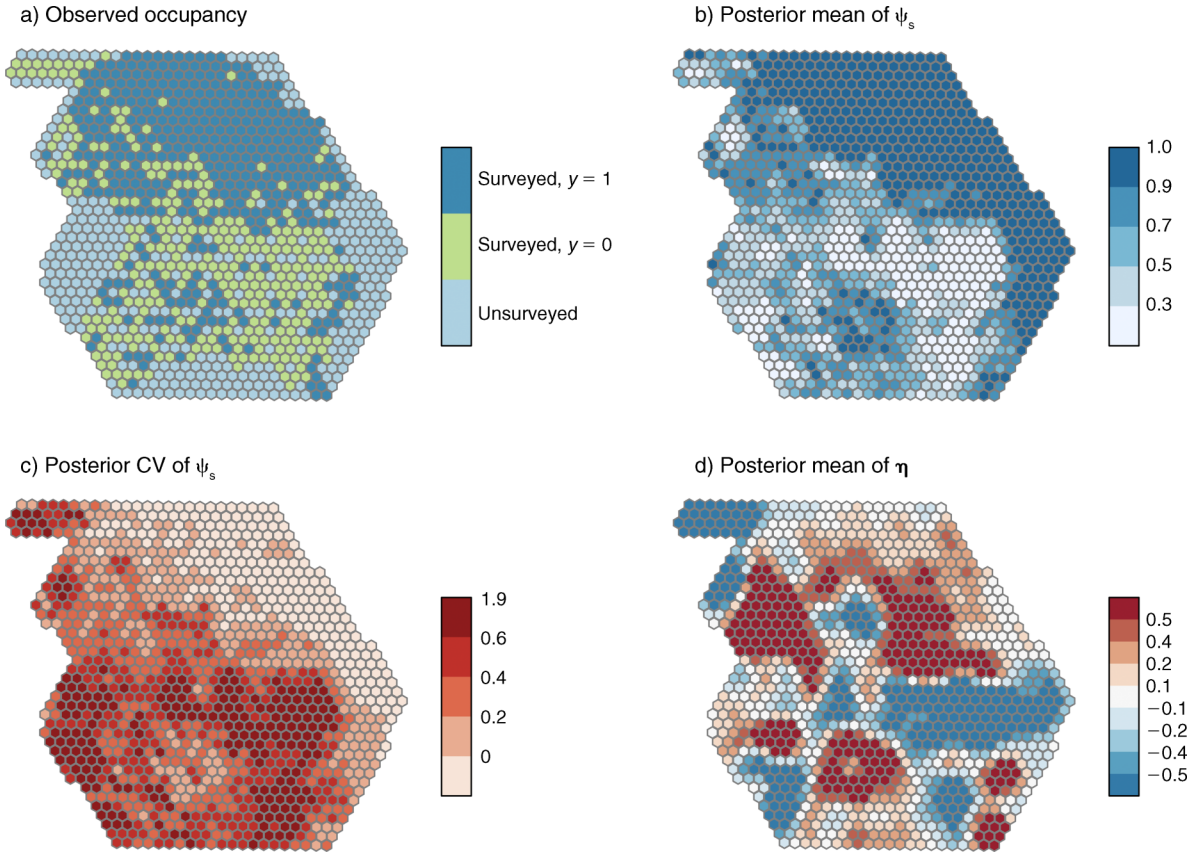

FIG. 2. Maps of the survey results and estimated caribou occupancy. Panel (a) illustrates the sampling process where light blue hexagons are unsurveyed, green hexagons were surveyed but caribou tracks were never observed, and dark blue hexagons were surveyed at least once and caribou tracks were detected (i.e., $y = 1$ for at least one track). Panel (b) represents the posterior mean of ψ_s (the occupancy probability) for each unit. Panel (c) is the posterior coefficient of variation (CV) for ψ_s at all survey units. Panel (d) illustrates the posterior mean of the $\boldsymbol{\eta}$ process (spatially correlated random effect).

used to extend occupancy models to the continuous spatial domain. It is also conceivable that given sufficient sampling over time that a spatiotemporal process might be warranted. Wikle (2010) provide some extensions to reduced-rank spatiotemporal models that could be utilized for occupancy processes that evolve over time. Similarly, it may be worth considering extensions in which parameters of the detection function have spatial or temporal autocorrelation. However, in such a setup care would be needed to avoid statistical confounding between occupancy and detection parameters.

The issue of spatial confounding is relatively new in the statistics literature. As Hodges and Reich (2010) state, it is a "rich area of research." Experienced spatial modelers might say that confounding is desirable as it allows an adjustment in fixed effects inference due to unmeasured (unknown), spatially correlated covariates. However, Hodges, Hodges and Reich (2010) and Paciorek (2010) illustrate bias and variance inflation depend on the structure of these unmeasured variables. Because the structure of these latent spatial variables is unknown, a spatial model may not account for them the way the researcher intends. Hodges and Reich (2010) note that without knowledge of the missing structure, purposefully adding spatial confounding is a haphazard adjustment that may bias the known fixed effects in unknown ways. There may be some middle ground, however, that is worth exploring (Paciorek 2010) and could be the topic of future research in the context of occupancy modeling and spatial modeling in general.

ACKNOWLEDGMENTS

The authors thank J. Hughes, J. Hodges, M. Haran, B. Reich, J. Ver Hoef, J. Laake, and M. Kery for helpful discussion. We are grateful to Ontario Ministry of Natural Resources (OMNR) pilots and field and science unit staff, to Supercub pilots H. McMahan and G. Lee, L. Yesno who acted as an observer, and to G. Woolmer, C. Stratton, and F. Hofmann for data processing. We are indebted to the OMNR Species-at-Risk Research Fund for Ontario for financial support. The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names or products does not imply endorsement by the National Marine Fisheries Service, NOAA or the U.S. Government.

LITERATURE CITED

Aing, C., S. Halls, K. Oken, R. Dobrow, and J. Fieberg. 2011. A Bayesian hierarchical occupancy model for track surveys

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [10/05/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

conducted in a series of linear, spatially correlated, sites. Journal of Applied Ecology 48(6):1508–1517.
Albert, J., and S. Chib. 1993. Bayesian-analysis of binary and polychotomous response data. Journal of the American Statistical Association 88(422):669–679.
Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical modeling and analysis for spatial data. Number 101 in Monographs on statistics and applied probability. Chapman and Hall/CRC, Boca Raton, Florida, USA.
Bergerud, A. T., R. D. Jakimchuk, and D. R. Carruthers. 1984. The buffalo of the north: caribou (*Rangifer tarandus*) and human developments. Arctic 37:7–22.
Besag, J. 2002. What is a statistical model? Discussion. Annals of Statistics 30:1267–1277.
Boots, B., and M. Tiefelsdorf. 2000. Global and local spatial autocorrelation in bounded regular tessellations. Journal of Geographical Systems 2:319–348.
Chelgren, N. D., M. J. Adams, L. L. Bailey, and R. B. Bury. 2011. Using multilevel spatial models to understand salamander site occupancy patterns after wildfire. Ecology 92:408–421.
Dorazio, R. M., and D. T. Rodriguez. 2012. A Gibbs sampler for Bayesian analysis of site-occupancy data. Methods in Ecology and Evolution 3(6):1093–1098
Gardner, C. L., J. P. Lawler, J. M. V. Hoef, A. J. Magoun, and K. A. Kellie. 2010. Coarse-scale distribution surveys and occurrence probability modeling for wolverine in interior Alaska. Journal of Wildlife Management 74(8):1894–1903.
Givens, G., and J. A. Hoeting. 2005. Computational Statistics. Wiley, New York, New York, USA.
Hanks, E. M., M. B. Hooten, and F. A. Baker. 2011. Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence. Ecological Applications 21:1173–1188.
Hodges, J. S., and B. J. Reich. 2010. Adding spatially-correlated errors can mess up the fixed effects you love. American Statistician 64(4):325–334.
Hoeting, J. A., M. Leecaster, and D. Bowden. 2000. An improved model for spatially correlated binary responses. Journal of Agricultural, Biological, and Environmental Statistics 5:102–114.
Hooten, M., D. Larsen, and C. Wikle. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. Landscape Ecology 18(5):487–502.
Hughes, J., and M. Haran. 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. Journal of the Royal Statistical Society, B. dx.doi.org/10.1111/j.1467-9868.2012.01041.x
Karanth, K. U., A. M. Gopalaswamy, N. S. Kumar, S. Vaidyanathan, J. D. Nichols, and D. I. MacKenzie. 2011. Monitoring carnivore populations at the landscape scale: occupancy modelling of tigers from sign surveys. Journal of Applied Ecology 48:1048–1056.
Long, R., P. MacKay, W. J. Zielinski, and J. C. Ray, editors. 2008. Noninvasive survey methods for carnivores. Island Press, Washington, D.C., USA.
MacKenzie, D., J. Nichols, G. Lachman, S. Droege, J. Royle, and C. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.
MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollack, L. L. Bailey, and J. E. Hines. 2005. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, San Diego, California, USA.
Magoun, A. J., J. C. Ray, D. S. Johnson, P. Valkenburg, F. N. Dawson, and J. Bowman. 2007. Modeling wolverine occurrence using aerial surveys of tracks in snow. Journal of Wildlife Management 71(7):2221–2229.
McCullagh, P., and J. A. Nelder. 1999. Generalized linear models. Second edition. Chapman and Hall/CRC, Boca Raton, Florida, USA.
Moran, P. 1950. Notes on continuous stochastic phenomena. Biometrika 37:17–23.
Paciorek, C. J. 2010. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. Statistical Science 25(1):107–125.
R Development Core Team. 2011. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org
Reich, B. J., J. S. Hodges, and V. Zadnik. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics 62(4):1197–1206.
Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology. Academic Press/Elsevier, New York, New York, USA.
Royle, J., and W. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835–841.
Rue, H., and L. Held. 2005. Gaussian Markov random fields: theory and applications. Volume 104 of monographs on statistics and applied probability. Chapman and Hall, London, UK.
Sargeant, G., M. Solvada, C. Slivinski, and D. Johnson. 2005. Markov chain Monte Carlo estimation of species distributions: a case study of the swift fox in western Kansas. Journal of Wildlife Management 69:483–497.
Schaefer, J. A., and W. O. Pruitt, Jr. 1991. Fire and woodland caribou in southeastern Manitoba. Wildlife Monographs 116:1–39.
Spectranalysis. 2000. Introduction to the Ontario land cover data base, second edition: outline of production methodology and description of 27 land cover classes. Unpublished report to Ontario Ministry of Natural Resources. Spectranalysis, Oakville, Ontario, Canada.
Wall, M. M. 2004. A close look at the spatial structure implied by the CAR and SAR models. Journal of Statistical Planning and Inference 121:311–324.
White, G. C., and K. P. Burnham. 1999. Program mark: survival estimation from populations of marked animals. Bird Study 46(Supplement):120–138.
Wikle, C. 2010. Low-rank representations for spatial processes. Pages 107–118 *in* A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, editors. Handbook of spatial statistics. CRC Press, Boca Raton, Florida, USA.

Supplemental Material

Appendix A

Detailed description of the Gibbs sampler used for the spatial occupancy model presented in the paper (*Ecological Archives* E094-069-A1).

Appendix B

Additional figures and results for the spatial occupancy analysis of the caribou data set (*Ecological Archives* E094-069-A2).

19399170, 2013, 4, Downloaded from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/12-0564.1 by Brian Gerber - Colorado State University , Wiley Online Library on [06/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License