Received: 1 November 2023 | Accepted: 15 April 2024

DOI: 10.1111/2041-210X.14332

Methods in Ecology and Evolution

APPLICATION

spAbundance: An R package for single-species and multi-species spatially explicit abundance models

Jeffrey W. Doser[1,2] | Andrew O. Finley[2,3,4] | Marc Kéry[5] | Elise F. Zipkin[1,2]

[1]Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

[2]Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, USA

[3]Department of Forestry, Michigan State University, East Lansing, Michigan, USA

[4]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

[5]Swiss Ornithological Institute, Sempach, Switzerland

Correspondence
Jeffrey W. Doser
Email: doserjef@msu.edu

Funding information
NSF Division of Environmental Biology, Grant/Award Number: DEB-2213565; USDA U.S. Forest Service Forest Inventory and Analysis; NSF Division of Biological Infrastructure, Grant/Award Number: DBI-1954406; National Aeronautics and Space Administration, Grant/Award Number: CMS 2018 and CMS 2020; NSF Division of Mathematical Sciences, Grant/Award Number: DMS-1916395; USDA Forest Service Northern Research Station

Handling Editor: Chris Sutherland

Abstract

1. Numerous modelling techniques exist to estimate abundance of plant and animal populations. The most accurate methods account for multiple complexities found in ecological data, such as observational biases, spatial autocorrelation, and species correlations. There is, however, a lack of user-friendly and computationally efficient software to implement the various models, particularly for large data sets.
2. We developed the `spAbundance R` package for fitting spatially explicit Bayesian single-species and multi-species hierarchical distance sampling models, N-mixture models, and generalized linear mixed models. The models within the package can account for spatial autocorrelation using Nearest Neighbour Gaussian Processes and accommodate species correlations in multi-species models using a latent factor approach, which enables model fitting for data sets with large numbers of sites and/or species.
3. We provide three vignettes and three case studies that highlight `spAbundance` functionality. We used spatially explicit multi-species distance sampling models to estimate density of 16 bird species in Florida, USA, an N-mixture model to estimate black-throated blue warbler (*Setophaga caerulescens*) abundance in New Hampshire, USA, and a spatial linear mixed model to estimate forest aboveground biomass across the continental USA.
4. `spAbundance` provides a user-friendly, formula-based interface to fit a variety of univariate and multivariate spatially explicit abundance models. The package serves as a useful tool for ecologists and conservation practitioners to generate improved inference and predictions on the spatial drivers of abundance in populations and communities.

KEYWORDS
Bayesian, distance sampling, hierarchical model, imperfect detection, N-mixture model

1 | INTRODUCTION

Understanding how abundance of plant and animal populations varies across space and time is a central objective in ecology and conservation management. A variety of sampling and associated modelling techniques have been developed over the last 50 years to estimate abundance while accounting for imperfect detection (i.e. the failure to observe all individuals of a species that are present at a location during the sampling period), including distance sampling and repeated counts, among others (Nichols et al., 2009). In distance sampling, the probability of detecting an individual is assumed to decay with increasing distance to the observer, which allows for the explicit estimation of abundance/density while accommodating imperfect detection of individuals (Buckland et al., 2001). Hierarchical distance sampling (HDS; Royle et al., 2004) extends classical distance sampling to model abundance/density as a function of spatially varying covariates. Royle et al. (2004) introduced N-mixture models, which allow for estimation of abundance and effects of spatially varying covariates while accounting for detection probability using replicated count data during some period where the population is assumed to be closed, that is where no births/deaths or immigration/emigration occur. In addition to approaches that explicitly account for imperfect detection, generalized linear mixed models (GLMMs) that estimate relative abundance (i.e. ignoring imperfect detection) can be used to assess how environmental covariates influence relative changes in abundance across space and/or time (Barker et al., 2018; Goldstein & de Valpine, 2022). Multi-species (i.e. multivariate) extensions of HDS (Sollmann et al., 2016), N-mixture models (Yamaura et al., 2012) and GLMMs (e.g. Hui et al., 2015) use count data from multiple species to estimate species-specific patterns in abundance, which may also estimate correlations between species in a joint species distribution model (JSDM) framework (Warton et al., 2015).

When modelling abundance across large spatial domains and/or using a large number of observed locations, accommodating spatial autocorrelation becomes increasingly important (Guélat & Kéry, 2018). Spatial autocorrelation can arise from a variety of ecological and/or biological processes, such as additional environmental drivers not included as covariates in the model, dispersal, species interactions, and source-sink meta-population dynamics (Chapter 9; Kéry & Royle, 2021). Failing to account for residual spatial autocorrelation (i.e. remaining spatial autocorrelation after accounting for environmental covariates) can lead to overly precise estimates and inferior predictive performance. Modelling spatial dependence is commonly done via the addition of spatially structured random effects to point-referenced spatial regression models (i.e. spatially explicit models). Gaussian process-based random effects provide a flexible non-parametric approach to capture spatial patterns, offer unparalleled process parameter and predictive inference, and yield probabilistic uncertainty quantification. The hierarchical Bayesian framework is the preferred inferential framework for models developed here and in the literature due to their increased flexibility to fit models that would be infeasible with classical methods (Banerjee et al., 2014). Such models are, however, notoriously computationally intensive (Banerjee & Fuentes, 2012), as computational complexity increases in cubic order with the number of spatial locations. These computational bottlenecks make fitting spatially explicit models impractical for even moderately large data sets using Bayesian software packages such as `Stan` (Carpenter et al., 2017) and `NIMBLE` (de Valpine et al., 2017).

Many popular, formula-based `R` packages exist that can fit various combinations of distance sampling models, N-mixture models, and/or spatially explicit GLMMs for assessing spatial patterns in abundance (Supplemental Information S1: Table S1). The R package `unmarked` (Fiske & Chandler, 2011; Kellner et al., 2023) is commonly used to fit single-species distance sampling and N-mixture models, but cannot accommodate spatial autocorrelation. The `dsm` package (Miller et al., 2013) can fit spatially explicit distance sampling models using generalized additive models, the `hSDM` package (Vieilledent, 2019) can fit spatially explicit N-mixture models with an intrinsic conditional autoregessive model (Ver Hoef et al., 2018), while the `ubms` package (Kellner et al., 2021) fits both spatially explicit distance sampling and N-mixture models using restricted spatial regression (Hodges & Reich, 2010). These packages, however, cannot accommodate multiple species within a multivariate framework. A variety of R packages exist to fit spatially explicit univariate and multivariate GLMMs, such as `spBayes` (Finley et al., 2015), `Hmsc` (Tikhonov et al., 2020), and `sdmTMB` (Anderson et al., 2022). However, none of these packages can explicitly account for imperfect detection.

In this paper, we introduce the `spAbundance` R package for fitting Bayesian single-species and multi-species HDS models, N-mixture models, and GLMMs that may or may not include spatial autocorrelation in large data sets. We fit all spatially explicit models with Nearest Neighbour Gaussian Processes (NNGPs), a computationally efficient approach that closely approximates a full Gaussian process while drastically reducing computational run times (Datta et al., 2016; Finley et al., 2019). We designed `spAbundance` syntax to closely follow the syntax of `spOccupancy` (Doser et al., 2022), an R package that fits a variety of analogous spatially explicit occupancy models, which together provide a user-friendly and computationally efficient set of tools to model occupancy and abundance while accounting for spatial autocorrelation and imperfect detection.

2 | OVERVIEW OF MODELS IN `spAbundance`

Next we give a brief overview of the models included in `spAbundance`. See Supplemental Information S1 for details on all prior distributions and their default values.

2.1 | Single-species HDS models

The spAbundance functions DS and spDS fit non-spatial and spatially explicit single-species HDS models, respectively. Let $N(\boldsymbol{s}_j)$ denote the true abundance of a species of interest at site $j = 1, \ldots, J$ with spatial coordinates $\boldsymbol{s}_j$. We model $N(\boldsymbol{s}_j)$ using either a Poisson or negative binomial (NB) distribution following:

$$\begin{aligned} N(\boldsymbol{s}_j) &\sim \text{Poisson}(\mu(\boldsymbol{s}_j)A(\boldsymbol{s}_j)), \text{or,} \\ N(\boldsymbol{s}_j) &\sim \text{NB}(\mu(\boldsymbol{s}_j)A(\boldsymbol{s}_j), \kappa), \end{aligned} \tag{1}$$

where $\mu(\boldsymbol{s}_j)$ is the average abundance at site j, $A(\boldsymbol{s}_j)$ is an offset, and κ is a positive dispersion parameter. Smaller values of κ indicate overdispersion in the latent abundance values relative to a Poisson model, while higher values indicate minimal overdispersion in abundance. Note that as $\kappa \to \infty$, the negative binomial distribution becomes the Poisson distribution. The offset term $A(\boldsymbol{s}_j)$ can be used to convert $\mu(\boldsymbol{s}_j)$ to units of density (i.e. abundance per unit area), while if $A(\boldsymbol{s}_j) = 1$, $\mu(\boldsymbol{s}_j)$ is average abundance per site. We model $\mu(\boldsymbol{s}_j)$ using a log link function following

$$\log(\mu(\boldsymbol{s}_j)) = \boldsymbol{x}(\boldsymbol{s}_j)^{\top}\boldsymbol{\beta} + \text{w}(\boldsymbol{s}_j), \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients for a set of covariates $\boldsymbol{x}(\boldsymbol{s}_j)$ including an intercept, $\text{w}(\boldsymbol{s}_j)$ is a zero-mean spatial random effect, and the $^{\top}$ denotes transposition of column vector $\boldsymbol{x}(\boldsymbol{s}_j)$. For non-spatial HDS models, $\text{w}(\boldsymbol{s}_j)$ is removed from Equation 2. For spatially explicit HDS, we model $\mathbf{w}(\boldsymbol{s})$ using a NNGP as a computationally efficient alternative to using a full spatial GP. More specifically, we assume that

$$\mathbf{w}(\boldsymbol{s}) \sim \text{Normal}(\mathbf{0}, \widetilde{\boldsymbol{C}}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})), \tag{3}$$

where $\widetilde{\boldsymbol{C}}(\boldsymbol{s}, \boldsymbol{s}', \boldsymbol{\theta})$ is a $J \times J$ NNGP-derived spatial covariance matrix and $\boldsymbol{\theta}$ is a vector of parameters governing the spatial process according to a spatial covariance function. spAbundance supports four spatial covariance models: exponential, spherical, Gaussian and Matérn (Banerjee et al., 2014). For the exponential, spherical and Gaussian functions, $\boldsymbol{\theta} = \{\sigma^2, \phi\}$, where σ^2 is a spatial variance parameter controlling the magnitude of the spatial random effects and ϕ is a spatial decay parameter controlling the range of spatial autocorrelation, while the Matérn function additionally includes a spatial smoothness parameter, ν. See Supplemental Information S1 for statistical details on the NNGP approximation.

Suppose observers count the number of individuals of the species of interest at each site j. Our HDS software implementation in spAbundance supports two types of "sites": line transects and point count surveys. In line transects, each site j is a line transect the observer walks along and records the distance of each observed individual to the line within a set of $k = 1, \ldots, K$ distance bands. In point count surveys, each site j is the centre of an imaginary circle at which an observer stands and records the distance of each observed individual to the centre of the circle within $k = 1, \ldots, K$ circular distance bands. Note that sometimes continuous distances are recorded rather than distance bins, in which case the continuous distance measurements can then be split into K distance bins prior to analysis. Define $\boldsymbol{y}(\boldsymbol{s}_j)$ as a vector of K values indicating the number of individuals observed within each distance band k at site j. Similarly, let $\boldsymbol{y}^*(\boldsymbol{s}_j)$ be a vector of $K + 1$ values, where the first K values correspond to $\boldsymbol{y}(\boldsymbol{s}_j)$, and the last value is the number of unobserved individuals at that location (i.e. $N(\boldsymbol{s}_j) - \sum_{k=1}^{K} y_k(\boldsymbol{s}_j)$). Note the last value in $\boldsymbol{y}^*(\boldsymbol{s}_j)$ is not observed (i.e. since $N(\boldsymbol{s}_j)$ is not known). We model $\boldsymbol{y}^*(\boldsymbol{s}_j)$ according to

$$\boldsymbol{y}^*(\boldsymbol{s}_j) \sim \text{Multinomial}\left(N(\boldsymbol{s}_j), \boldsymbol{\pi}_j^*\right), \tag{4}$$

where $\boldsymbol{\pi}_j^*$ is a vector of cell probabilities with the first K values denoted as $\boldsymbol{\pi}_j$ and the final value $\pi_{j,K+1} = 1 - \sum_{k=1}^{K} \pi_{j,k}$. More specifically, $\pi_{j,k}$ is the probability of detecting an individual in the kth distance band at site j. We define $\pi_{j,k}$ as

$$\pi_{j,k} = \bar{p}_{j,k}\psi_k, \tag{5}$$

where $\bar{p}_{j,k}$ is the probability of detecting an individual in distance band k, given the individual occurs in distance band k, and ψ_k is the probability an individual occurs in distance band k. The definitions of $\bar{p}_{j,k}$ and ψ_k are different depending on whether the distance bands are linear (as in line transects) or circular (as in point count surveys). Following the standard distance sampling assumption that animals are uniformly distributed in space, for line transects we have

$$\psi_k = \frac{b_{k+1} - b_k}{B}, \tag{6}$$

where b_{k+1} and b_k are the upper and lower distance limits for band k, and B is the line transect half-width (i.e. the maximum distance within which individuals are counted). Further, for distance x we have

$$\bar{p}_{j,k} = \frac{1}{b_{k+1} - b_k}\int_{b_k}^{b_{k+1}} g(\text{x})d\text{x}. \tag{7}$$

For point count surveys, we have

$$\psi_k = \frac{b_{k+1}^2 - b_k^2}{B^2}, \tag{8}$$

where b_{k+1} and b_k are similarly the upper and lower distance limits for band k, and B is the radius of the full point count circle. We then define $\bar{p}_{j,k}$ as

$$\bar{p}_{j,k} = \frac{1}{b_{k+1}^2 - b_k^2}\int_{b_k}^{b_{k+1}} g(\text{x})2\text{x}d\text{x}. \tag{9}$$

For both line transects and point count surveys, $g(\text{x})$ is some declining function of distance x from the transect line/point count survey centre. We approximate the integrals in Equation 7 and 9 using numerical integration. Our software implementation in spAbundance currently supports two detection functions: half-normal and negative exponential (see Supplemental Information S1 for their

definitions). Both of these functions are governed by a scale parameter, σ_j, which can be modelled as a function of covariates to allow detection probability to vary across sites. More specifically, we have

$$\log(\sigma_j) = \mathbf{v}_j^\top \boldsymbol{\alpha}, \tag{10}$$

where $\boldsymbol{\alpha}$ is a vector of regression coefficients for covariates $\mathbf{v}_j$ (including an intercept).

2.2 | Multi-species HDS models

Now consider the case where distance sampling data, $\mathbf{y}_i(\mathbf{s}_j)$, are collected for multiple species $i = 1, \ldots, I$ at each survey location j with coordinates $\mathbf{s}_j$. We are now interested in estimating the abundance of each species i at each location j, denoted as $N_i(\mathbf{s}_j)$. We model $N_i(\mathbf{s}_j)$ analogous to Equation 1, with expected abundance now varying by species and site according to

$$\log(\mu_i(\mathbf{s}_j)) = \mathbf{x}\left(\mathbf{s}_j^\top\right)\boldsymbol{\beta}_i + \mathrm{w}_i^*(\mathbf{s}_j), \tag{11}$$

where $\boldsymbol{\beta}_i$ are the species-specific effects of covariates $\mathbf{x}(\mathbf{s}_j)$ (including an intercept) and $\mathrm{w}_i^*(\mathbf{s}_j)$ is a species-specific random effect. When $N_i(\mathbf{s}_j)$ is modelled using a negative binomial distribution, we estimate a separate dispersion parameter κ_i for each species. We model $\boldsymbol{\beta}_i$ as random effects arising from a common, community-level normal distribution, which leads to increased precision of species-specific effects compared to single-species models (Sollmann et al., 2016). For example, the species-specific abundance intercept $\beta_{0,i}$ is modelled according to

$$\beta_{0,i} \sim \mathrm{Normal}\left(\mu_{\beta_0}, \tau^2_{\beta_0}\right), \tag{12}$$

where μ_{β_0} is the community-level abundance intercept, and $\tau^2_{\beta_0}$ is the variance of the intercept across all I species. The observation portion of the multi-species distance sampling model is identical to the single-species model and follows Equations 4–10, with all parameters indexed by species, and the species-specific coefficients $\boldsymbol{\alpha}_i$ modelled hierarchically analogous to the species-specific abundance coefficients $\boldsymbol{\beta}_i$ (Equation 12). Given that species-specific effects are treated as random effects, such an approach requires at least 5–6 species in the modelled community in order for reliable estimation of variance parameters.

`spAbundance` fits three types of multi-species models that differ in how they incorporate the species-specific random effect $\mathrm{w}_i^*(\mathbf{s}_j)$ (if included). The function `msDS` fits the non-spatial multi-species distance sampling model of Sollmann et al. (2016) in which we remove the random effect $\mathrm{w}_i^*(\mathbf{s}_j)$ from Equation 11. The function `sfMsDS` fits spatial multi-species distance sampling models using a spatial factor model (Hogan & Tchernis, 2004), which simultaneously accommodates spatial autocorrelation and residual species correlations in a spatial JSDM framework. Briefly, we decompose $\mathrm{w}_i^*(\mathbf{s}_j)$ into a linear combination of q latent variables (i.e. factors) and their associated species-specific coefficients (i.e. factor loadings). More specifically, we have

$$\mathrm{w}_i^*(\mathbf{s}_j) = \boldsymbol{\lambda}_i^\top \mathbf{w}(\mathbf{s}_j), \tag{13}$$

where $\boldsymbol{\lambda}_i^\top$ is the ith row of factor loadings from an $I \times q$ loadings matrix $\boldsymbol{\Lambda}$, and $\mathbf{w}(\mathbf{s}_j)$ is a vector of length q of independent spatial factors at site j. By setting $q \ll I$, we achieve dimension reduction to efficiently model communities with a large number of species (Doser et al., 2023; Taylor-Rodriguez et al., 2019). The approach accounts for residual species correlations via their species-specific responses to the q spatial factors, which results in a residual interspecies covariance matrix that can be derived from the model as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$. We model each spatial factor using an independent NNGP according to Equation 3, except we fix the spatial variance parameter to 1 to ensure identifiability (Lopes & West, 2004). As an alternative, the function `lfMsDS` models $\mathrm{w}_i^*(\mathbf{s}_j)$ identical to Equation 13, except assumes each of the q factors in $\mathbf{w}(\mathbf{s}_j)$ arises from an independent standard normal distribution. This model does not account for spatial autocorrelation but does allow for the estimation of species correlations. The models fit by `sfMsDS` and `lfMsDS` can be thought of as abundance-based JSDMs that account for imperfect detection (Tobler et al., 2019; Chapter 8 in Kéry & Royle, 2021).

Our factor modelling approach to fitting spatially explicit multi-species models in `spAbundance` implicitly assumes species are correlated through latent factors $\mathbf{w}(\mathbf{s}_j)$. If there is no interest in residual species correlations, we could imagine a multi-species model that includes a separate spatial process for each species. However, we do not include such models in `spAbundance` because they are computationally infeasible when working with even a moderate number of species (e.g. 10). Further, in the context of occupancy models, the spatial factor modelling approach has been shown to perform equally as well as a model that estimates a separate spatial random effect for each species even when there are no residual correlations between the species in the community (Doser et al., 2023).

2.3 | Single-species N-mixture models

The functions `NMix` and `spNMix` fit non-spatial and spatial N-mixture models in `spAbundance`. Following the N-mixture model structure of Royle (2004), we assume observers count the number of individuals of a target species at each site j over a set of multiple surveys $k = 1, \ldots, K_j$, denoted as $y_k(\mathbf{s}_j)$. Note the number of surveys can vary by site, but at least some sites must be surveyed more than once to ensure identifiability. We model $y_k(\mathbf{s}_j)$ conditional on the true abundance of the species at site j, $N(\mathbf{s}_j)$, following:

$$y_k(\mathbf{s}_j) \sim \mathrm{Binomial}(N(\mathbf{s}_j), p_{j,k}), \tag{14}$$

where $p_{j,k}$ is the probability of detecting an individual given it is present at the site. We model $p_{j,k}$ using a logit link function in which we can

allow detection probability to vary over space and/or surveys. More specifically, we have

$$\text{logit}(p_{j,k}) = \boldsymbol{v}_{j,k}^{\top}\boldsymbol{\alpha}, \quad (15)$$

where $\boldsymbol{\alpha}$ is a vector of effects of covariates $\boldsymbol{v}_{j,k}$ (including an intercept). The model for abundance $N(\boldsymbol{s}_j)$ is identical to the single-species distance sampling model, which can include covariates and/or spatial random effects (Equations 1–3).

2.4 | Multi-species N-mixture models

Analogous to HDS models, we can extend single-species N-mixture models to model abundance of a community of I total species (Yamaura et al., 2012). In multi-species N-mixture models, we estimate the abundance of species i at spatial location j, $N_i(\boldsymbol{s}_j)$. Our model for $N_i(\boldsymbol{s}_j)$ follows that of multi-species HDS models, such that expected abundance can be modelled as a function of species-specific effects of spatially varying covariates (Equation 11) and species-specific random effects that can accommodate residual species correlations and residual spatial autocorrelation using a factor modelling approach (Equation 13). Species-specific covariate effects are modelled hierarchically following Equation 12. The observation portion of the multi-species N-mixture model is identical to the single-species model, now with species-specific covariate effects modelled hierarchically analogous to the abundance coefficients. `spAbundance` provides functions to fit non-spatial multi-species N-mixture models with (`lfMsNMix`) and without (`msNMix`) residual species correlations, as well as spatial multi-species N-mixture models that account for residual species correlations and spatial autocorrelation (`sfMsNMix`).

2.5 | Single-species GLMMs

The functions `abund` and `spAbund` fit single-species (i.e. univariate) non-spatial and spatial GLMMs in `spAbundance` using abundance and related (e.g. biomass) data. As opposed to HDS and N-mixture models, GLMMs do not explicitly account for imperfect detection via an additional hierarchical component to the model, and instead directly model the observed abundance at site j, $y(\boldsymbol{s}_j)$ to provide inference on relative abundance (e.g. Chapter 1 in Kéry & Royle, 2021). Observed abundance $y(\boldsymbol{s}_j)$ is modelled using some probability distribution with mean $\mu(\boldsymbol{s}_j)$. `spAbundance` currently supports Poisson and negative binomial for use with count data and the Gaussian distribution for use with continuous abundance data, such as biomass. Mean relative abundance $\mu(\boldsymbol{s}_j)$ is modelled according to Equation 2 for the Poisson and negative binomial cases, while the log link function is removed for the Gaussian case. Note that variables thought to influence detection probability can be incorporated in the model for $\mu(\boldsymbol{s}_j)$ to improve estimates of relative abundance (e.g. random observer effects, Link & Sauer, 2002).

2.6 | Multi-species GLMMs

Now consider the case where we have count data for multiple species I at each survey location j, denoted $\boldsymbol{y}_i(\boldsymbol{s}_j)$. We jointly model relative abundance of each species using a multivariate GLMM (e.g. Hui et al., 2015; Warton et al., 2015), in which expected abundance for each species i at site j, $\mu_i(\boldsymbol{s}_j)$, is modelled analogous to Equations 11–13. Note the log link function is removed from Equation 11 when modelling abundance using a Gaussian distribution. As with HDS and N-mixture models, `spAbundance` provides functions to fit non-spatial multivariate GLMMs with (`lfMsAbund`) and without (`msAbund`) residual species correlations. Multivariate spatial GLMMs with residual species correlations are fit using the `sfMsAbund` function.

3 | `spAbundance` FUNCTIONALITY

Here we highlight the five main tasks performed by `spAbundance` (see Table 1 for function names).

1. *Data simulation*: The functions `simDS`, `simMsDS`, `simNMix`, `simMsNMix`, `simAbund`, and `simMsAbund` simulate data under the single-species and multi-species HDS, N-mixture and GLMM frameworks for use in simulation studies or power analyses.
2. *Model fitting*: Model fitting functions were described in Section 2. All models are implemented in a Bayesian framework using custom Markov chain Monte Carlo (MCMC) algorithms written in `C/C++` using `R`'s foreign language interface. `spAbundance` uses standard `R` formula syntax to specify abundance and detection probability models, with options to include random intercepts and random slopes using `lme4` syntax (Bates et al., 2015). Users can specify initial values for the MCMC algorithm as well as each parameter's prior distribution to yield vague or informative priors as desired (Supplemental Information S1).
3. *Model validation and comparison*: The function `ppcAbund` performs posterior predictive checks on `spAbundance` model objects to assess model Goodness of Fit. The function `waicAbund` computes the conditional version (Millar, 2018) of the Widely Applicable Information Criterion (WAIC; Watanabe, 2010) for model selection and assessment.
4. *Posterior summaries*: We include `summary` functions for `spAbundance` model objects that display concise summaries of the posterior distributions for estimated parameters as well as the potential scale reduction factor ($\hat{R}$; Gelman & Rubin, 1992) and effective sample size for convergence diagnostics. Simple `plot` functions allow for further convergence diagnostics via visual assessment of traceplots. The complete posterior samples are returned as `coda::mcmc` objects (Plummer et al., 2006).
5. *Prediction*: `predict` functions for all `spAbundance` model objects provide predictions of abundance across a user-specified set of locations, given covariate values and spatial coordinates. The resulting posterior predictive distributions can be used to generate abundance-based species distribution maps with associated

TABLE 1 List of core functions in the `spAbundance` package. Model fitting function name components correspond to: `DS` (hierarchical distance sampling), `NMix` (N-mixture model), `abund` (abundance-based GLMM), `sp` (spatial), `ms` (multi-species), `lf` (latent factor) and `sf` (spatial factor).

Functionality	Description
Data simulation	
`simDS`	Simulate single-species distance sampling data
`simMsDS`	Simulate multi-species distance sampling data
`simNMix`	Simulate single-species repeated count data with imperfect detection
`simMsNMix`	Simulate multi-species repeated count data with imperfect detection
`simAbund`	Simulate single-species count data with perfect detection
`simMsAbund`	Simulate multi-species count data with perfect detection
Model fitting	
`DS`	Single-species HDS model
`spDS`	Single-species spatial HDS model
`msDS`	Multi-species HDS model
`lfMsDS`	Multi-species HDS model with species correlations
`sfMsDS`	Multi-species spatial HDS model with species correlations
`NMix`	Single-species N-mixture model
`spNMix`	Single-species spatial N-mixture model
`msNMix`	Multi-species N-mixture model
`lfMsNMix`	Multi-species N-mixture model with species correlations
`sfMsNMix`	Spatial multi-species N-mixture model with species correlations
`abund`	Single-species GLMM
`spAbund`	Single-species spatial GLMM
`msAbund`	Multi-species GLMM
`lfMsAbund`	Multi-species GLMM with species correlations
`sfMsAbund`	Multi-species spatial GLMM with species correlations
Model assessment	
`ppcAbund`	Posterior predictive check using Bayesian *p*-values
`waicAbund`	Compute Widely Applicable Information Criterion

uncertainty or to obtain population size estimates across the study region or within smaller areal units of interest. Users can also predict detection probability for HDS and N-mixture models to yield insight on how detection probability varies across a user-specified range of covariate values.

4 | WORKED EXAMPLES AND ONLINE RESOURCES

We demonstrate `spAbundance` functionality with three worked examples and three vignettes. Complete details for all worked examples are provided in Supplemental Information S1, along with associated code and data available on Zenodo (Doser et al., 2024). The vignettes are provided in Supplemental Information S2–S4 as well as on the package website (https://www.jeffdoser.com/files/spabundance-web/). Here we provide a short overview of the worked examples and vignettes.

4.1 | Case study 1: Bird density in central Florida

This case study demonstrates `spAbundance` functionality to fit spatial and nonspatial multi-species HDS models. We estimated density of 16 bird species in 2018 in the Disney Wilderness Preserve (48.5 km^2) in central Florida, USA. Distance sampling data were collected as part of the National Ecological Observatory Network landbird monitoring program (NEON, 2024). We compared the performance of the three multi-species model variants in `spAbundance` using WAIC. The spatial model substantially outperformed the non-spatial model with species correlations (ΔWAIC = 86) and the non-spatial model without species correlations (ΔWAIC = 155). Effects of forest cover on species-specific density varied across the community (Figure 1a), resulting in clear spatial variation in density of the 16 species (Supplemental Information S1: Figure S1). Detection probability quickly decayed with increasing distance from the observer (Figure 1b).

4.2 | Case study 2: Black-throated blue warbler abundance in Hubbard Brook Experimental Forest

In this case study, we showcase how to fit spatial and non-spatial single-species N-mixture models. We estimated abundance of black-throated blue warblers (*Setophaga caerulescens*) in the Hubbard Brook Experimental Forest (31.8 km^2) in New Hampshire, USA using repeated count data from 2015 (Rodenhouse & Sillett, 2021; Supplemental Information S1). We found minimal support for overdispersion and residual spatial autocorrelation, with a non-spatial Poisson N-mixture model performing best according to WAIC among multiple candidate models. A strong negative quadratic relationship with elevation revealed that abundance peaked at mid-elevations in the forest (Supplemental Information S1: Figure S2).

4.3 | Case study 3: Forest biomass across the continental USA

Our final case study demonstrates how `spAbundance` can be used to fit models using "big data". We estimated forest above-ground biomass across the continental US (~ 7.8 million km^2) using data from J = 86,933 forest inventory plots (Figure 2a) collected via the US Forest Service Forest Inventory and Analysis Program (Bechtold & Patterson, 2005). We fit a spatially explicit univariate GLMM using a Gaussian distribution with an ecoregion-specific random slope of

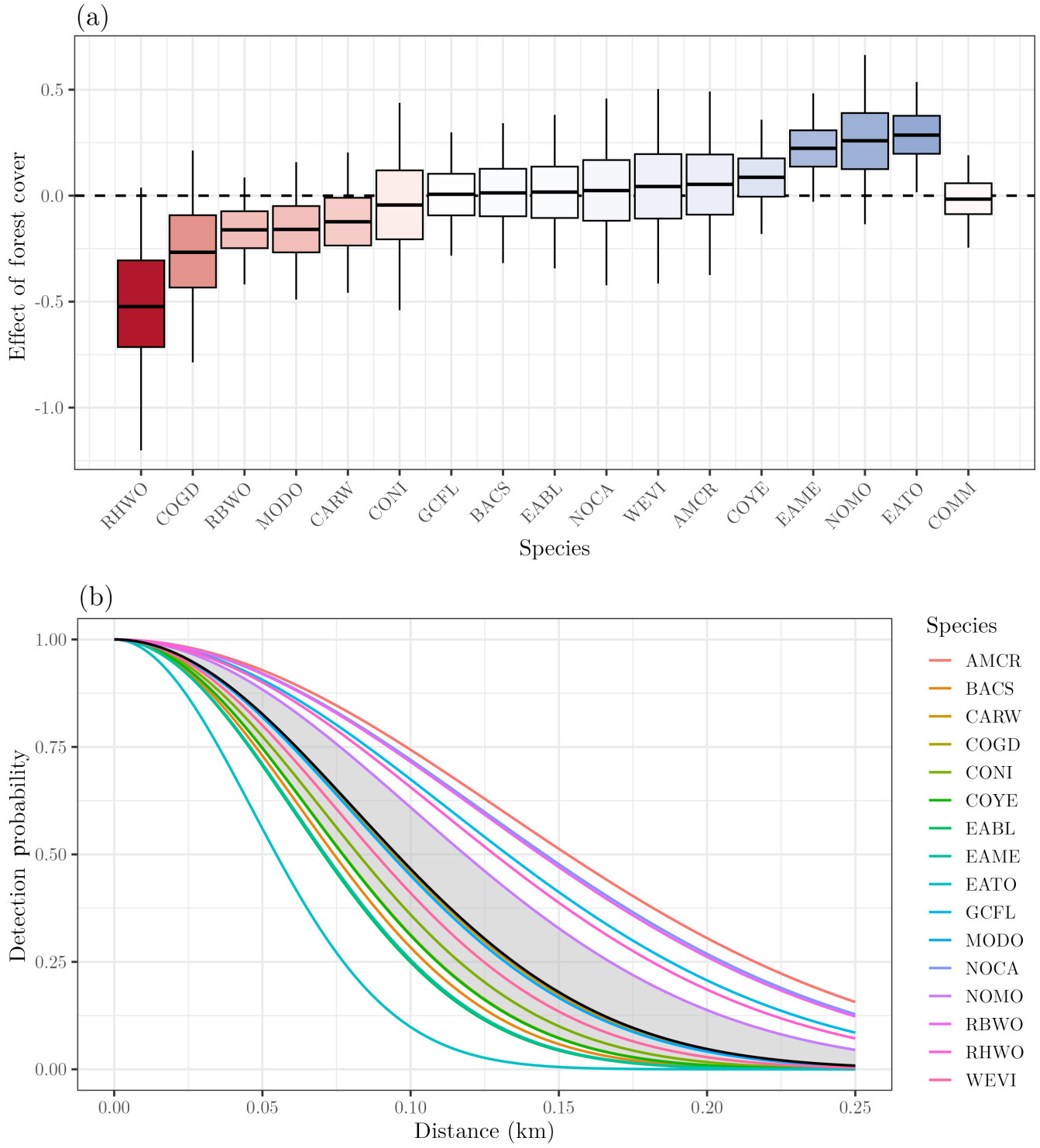

FIGURE 1 Species-specific effects of forest cover on density (a) and relationship between detection probability and distance from the observer (b) in the central Florida bird case study. Panel (a) shows the estimated mean (dark line), 50% credible interval (box), and 95% credible interval (whiskers) for the effect of forest cover on the overall community (COMM) and 16 individual species. In Panel (b), lines show the posterior mean detection probabilities for each species. The black line represents the average across the community (i.e. the community-level effect), and the grey region is the associated 95% credible interval. See Supplemental Information S1 for species code definitions.

tree canopy cover to reflect potential spatial variation in the relationship between canopy cover and biomass across different forest types. We found an overall positive relationship between tree canopy cover and biomass (median = 0.54, 95% credible interval 0.43–0.66), but clear variation in the magnitude of the effect across ecoregions (Figure 2b). Biomass predictions across the US aligned with expectations, with highest biomass predicted in the Pacific Northwest (Figure 2c,d).

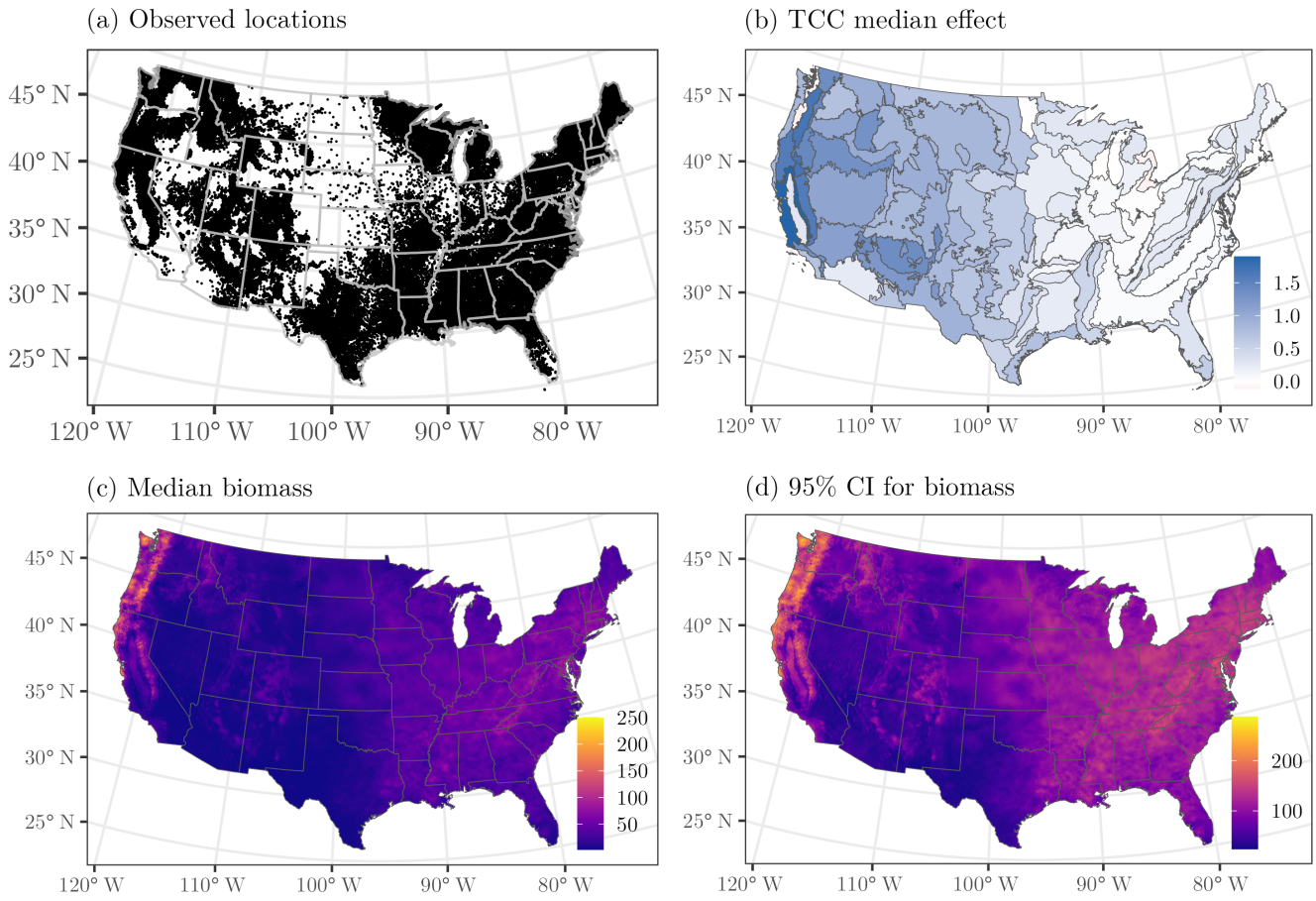

FIGURE 2 Data and predictions from the forest biomass case study. Panel (a) shows the observed locations of the 86,933 Forest Inventory and Analysis plots. Note these are the publicly available perturbed locations in which FIA adds a small amount of random noise to the true plot locations. Panel (b) shows the estimated random effect of tree canopy cover on forest biomass within distinct ecoregions. Panel (c) shows predicted biomass (posterior median) across the continental USA (tons per acre), with associated uncertainty (95% credible interval [CI] width) depicted in Panel (d).

4.4 | Vignettes

The three package vignettes provide complete details and examples for fitting all single-species and multi-species model types for HDS models (Supplemental Information S2), N-mixture models (Supplemental Information S3) and GLMMs (Supplemental Information S4). We provide extensive details on the required data formats for implementing the models in `spAbundance` and all function arguments including their default values. We additionally provide code to manipulate resulting objects after fitting models to generate a variety of plot types and summary figures.

5 | CONCLUSIONS AND FUTURE DIRECTIONS

We envision numerous extensions to existing `spAbundance` functionality and associated statistical methodology. We are currently working on functionality for zero-inflated models and spatiotemporal models, including "generalized" HDS and N-mixture models that account for imperfect availability (Chandler et al., 2011). We encourage future simulation studies to better identify the potential impacts of spatial confounding on inference in spatially explicit abundance models. Spatial confounding occurs when the spatial random effect is correlated with covariates included in the model, thus leading to difficulties in making inference on the covariate effects (Hodges & Reich, 2010). While approaches exist to reduce such confounding (e.g. restricted spatial regression; Hodges & Reich, 2010), they do not always provide more accurate inferences than standard approaches (Zimmerman & Ver Hoef, 2022). We echo the guidelines of Mäkinen et al. (2022) to assess the potential impacts of spatial confounding given the specific characteristics of the covariates of interest.

The aim in developing `spAbundance` is to provide ecologists and conservation practitioners with a user-friendly tool to quantify and understand spatial variation in the abundance of plant and animal species. This `R` package fits Bayesian spatially explicit single-species and multispecies versions of three of the most common modelling frameworks for "unmarked" data types: HDS models, N-mixture models and generalized linear mixed models. By using efficient statistical algorithms implemented in `C/C++` via R's foreign language interface, `spAbundance` is capable of handling datasets with a large

number of species (e.g. >100) and locations (e.g. 100,000). Together, the package vignettes (Supplemental Information S2–S4), code to implement the three case studies (Doser et al., 2024), and the package website (https://www.jeffdoser.com/files/spabundance-web/) provide full details and thorough exposition of `spAbudance` model objects.

AUTHOR CONTRIBUTIONS

Jeffrey W. Doser developed the package with insights from Andrew O. Finley; Jeffrey W. Doser wrote the package vignettes with insights from Marc Kéry; Jeffrey W. Doser performed analyses and led writing of the manuscript with critical insights from Elise F. Zipkin, Marc Kéry and Andrew O. Finley. All authors gave final approval for publication.

ACKNOWLEDGEMENTS

We thank J. Andrew Royle for helpful comments on the hierarchical distance sampling functionality and vignette. This work was supported by: Elise F. Zipkin NSF grants DBI-1954406 and DEB-2213565; Andrew O. Finley NASA CMS grants Hayes (CMS 2020) and Cook (CMS 2018), NSF grant DMS-1916395, joint venture agreements with the USDA Forest Service Forest Inventory and Analysis, USDA Forest Service Region 9 Forest Health Protection Northern Research Station.

CONFLICT OF INTEREST STATEMENT

We declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14332.

DATA AVAILABILITY STATEMENT

The package `spAbundance` is available on the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/web/packages/spAbundance/index.html). Data and code used in the examples are available on GitHub (https://github.com/zipkinlab/Doser_et_al_2024_MEE) and Zenodo (https://doi.org/10.5281/zenodo.10841651; Doser et al., 2024).

ORCID

Jeffrey W. Doser https://orcid.org/0000-0002-8950-9895
Elise F. Zipkin https://orcid.org/0000-0003-4155-6139

REFERENCES

Anderson, S. C., Ward, E. J., English, P. A., & Barnett, L. A. (2022). sdmTMB: An R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *bioRxiv*.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Banerjee, S., & Fuentes, M. (2012). Bayesian modeling for large spatial datasets. *WIREs Computational Statistics*, *4*(1), 59–66.

Barker, R. J., Schofield, M. R., Link, W. A., & Sauer, J. R. (2018). On the reliability of N-mixture models for count data. *Biometrics*, *74*(1), 369–377.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bechtold, W. A., & Patterson, P. L. (2005). *The enhanced forest inventory and analysis program–national sampling design and estimation procedures*. Number 80. USDA Forest Service, Southern Research Station.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2001). *Introduction to distance sampling: Estimating abundance of biological populations*. Oxford University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Chandler, R. B., Royle, J. A., & King, D. I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology*, *92*(7), 1429–1435.

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514), 800–812.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*, 403–413.

Doser, J. W., Finley, A. O., & Banerjee, S. (2023). Joint species distribution models with imperfect detection for high-dimensional spatial data. *Ecology*, *104*(9), e4137.

Doser, J. W., Finley, A. O., Kéry, M., & Zipkin, E. F. (2022). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, *13*(8), 1670–1678.

Doser, J. W., Finley, A. O., Kéry, M., & Zipkin, E. F. (2024). Code and data for spAbundance: An R package for single-species and multi-species spatially-explicit abundance models. Zenodo https://doi.org/10.5281/zenodo.10841651

Finley, A. O., Banerjee, S., & Gelfand, A. E. (2015). spBayes for large univariate and multivariate point-referenced Spatio-temporal data models. *Journal of Statistical Software*, *63*(13), 1–28.

Finley, A. O., Datta, A., Cook, B. D., Douglas, C., Morton, H. E. A., & Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, *28*(2), 401–414.

Fiske, I., & Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Goldstein, B. R., & de Valpine, P. (2022). Comparing N-mixture models and GLMMs for relative abundance estimation in a citizen science dataset. *Scientific Reports*, *12*(1), 12276.

Guélat, J., & Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, *9*(6), 1614–1625.

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, *64*(4), 325–334.

Hogan, J. W., & Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, *99*(466), 314–324.

Hui, F. K., Taskinen, S., Pledger, S., Foster, S. D., & Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*(4), 399–411.

Kellner, K. F., Fowler, N. L., Petroelje, T. R., Kautz, T. M., Beyer, D. E., Jr., & Belant, J. L. (2021). ubms: An R package for fitting hierarchical occupancy and N-mixture abundance models in a Bayesian framework. *Methods in Ecology and Evolution*, *13*(3), 577–584.

Kellner, K. F., Smith, A. D., Royle, J. A., Kéry, M., Belant, J. L., & Chandler, R. B. (2023). The unmarked R package: Twelve years of advances in occurrence and abundance modelling in ecology. *Methods in Ecology and Evolution*, *14*(6), 1408–1415.

Kéry, M., & Royle, J. A. (2021). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance, and species richness in R and BUGS: Volume 2: Dynamic and advanced models*. Academic Press.

Link, W. A., & Sauer, J. R. (2002). A hierarchical analysis of population change with application to cerulean warblers. *Ecology*, *83*(10), 2832–2840.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, *14*, 41–67.

Mäkinen, J., Numminen, E., Niittynen, P., Luoto, M., & Vanhatalo, J. (2022). Spatial confounding in Bayesian species distribution modeling. *Ecography*, *2022*(11), e06183.

Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, *28*(2), 375–385.

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*(11), 1001–1010.

NEON. (2024). Breeding landbird point counts (dp1.10003.001). Dataset accessed from https://data.neonscience.org/data-products/DP1.10003.001/RELEASE-2024 on March 5, 2024.

Nichols, J. D., Thomas, L., & Conn, P. B. (2009). Inferences about landbird abundance from count data: Recent advances and future directions. In *Modeling demographic processes in marked populations* (pp. 201–235). Springer.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Rodenhouse, N. L., & Sillett, T. S. (2021). Valley-wide bird survey, Hubbard brook experimental forest, 1999–2016 (ongoing). https://doi.org/10.6073/pasta/faca2b2cf2db9d415c39b695cc7fc21

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, *60*(1), 108–115.

Royle, J. A., Dawson, D. K., & Bates, S. (2004). Modeling abundance effects in distance sampling. *Ecology*, *85*(6), 1591–1597.

Sollmann, R., Gardner, B., Williams, K. A., Gilbert, A. T., & Veit, R. R. (2016). A hierarchical distance sampling model to estimate abundance and covariate associations of species and communities. *Methods in Ecology and Evolution*, *7*(5), 529–537.

Taylor-Rodriguez, D., Finley, A. O., Datta, A., Babcock, C., Andersen, H.-E., Cook, B. D., Morton, D. C., & Banerjee, S. (2019). Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. *Statistica Sinica*, *29*, 1155.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, *11*(3), 442–447.

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P., & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, *100*(8), e02754.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., & Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, *88*(1), 36–59.

Vieilledent, G. (2019). *hSDM: Hierarchical Bayesian species distribution models*. R package version 1.4.1.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12), 3571–3594.

Yamaura, Y., Royle, J. A., Shimada, N., Asanuma, S., Sato, T., Taki, H., & Makino, S. (2012). Biodiversity of man-made open habitats in an underused country: A class of multispecies abundance models for count data. *Biodiversity and Conservation*, *21*, 1365–1380.

Zimmerman, D. L., & Ver Hoef, J. M. (2022). On deconfounding spatial confounding in linear models. *The American Statistician*, *76*(2), 159–167.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Supplementary Information S1. `spAbundance`: An R package for single-species and multi-species spatially explicit abundance models.

Supplementary Information S2. Fitting hierarchical distance sampling models in `spAbundance`.

Supplementary Information S3. Fitting N-mixture models in `spAbundance`.

Supplementary Information S4. Fitting generalized linear mixed models in `spAbundance`.

How to cite this article: Doser, J. W., Finley, A. O., Kéry, M., & Zipkin, E. F. (2024). `spAbundance`: An R package for single-species and multi-species spatially explicit abundance models. *Methods in Ecology and Evolution*, *15*, 1024–1033. https://doi.org/10.1111/2041-210X.14332