Ecology, 98(3), 2017, pp. 840–850

Integrating multiple data sources in species distribution modeling: a framework for data fusion*

Krishna Pacifici,[1,7] Brian J. Reich,[2] David A. W. Miller,[3] Beth Gardner,[4] Glenn Stauffer,[3] Susheela Singh,[2] Alexa McKerrow,[5] and Jaime A. Collazo[6]

[1]*Department of Forestry and Environmental Resources, Program in Fisheries, Wildlife, and Conservation Biology, North Carolina State University, Raleigh, North Carolina 27695 USA*
[2]*Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695 USA*
[3]*Department of Ecosystem Science and Management, Pennsylvania State University, University Park, Pennsylvania 16802 USA*
[4]*School of Environmental and Forest Sciences, University of Washington, Seattle, Washington 98195 USA*
[5]*U.S. Geological Survey, Core Science Systems, Biodiversity and Spatial Information Center, North Carolina State University, Raleigh, North Carolina 27695 USA*
[6]*U.S. Geological Survey, North Carolina Cooperative Fish and Wildlife Research Unit, Department of Applied Ecology, North Carolina State University, Raleigh, North Carolina 27695 USA*

Abstract. The last decade has seen a dramatic increase in the use of species distribution models (SDMs) to characterize patterns of species' occurrence and abundance. Efforts to parameterize SDMs often create a tension between the quality and quantity of data available to fit models. Estimation methods that integrate both standardized and non-standardized data types offer a potential solution to the tradeoff between data quality and quantity. Recently several authors have developed approaches for jointly modeling two sources of data (one of high quality and one of lesser quality). We extend their work by allowing for explicit spatial autocorrelation in occurrence and detection error using a Multivariate Conditional Autoregressive (MVCAR) model and develop three models that share information in a less direct manner resulting in more robust performance when the auxiliary data is of lesser quality. We describe these three new approaches ("Shared," "Correlation," "Covariates") for combining data sources and show their use in a case study of the Brown-headed Nuthatch in the Southeastern U.S. and through simulations. All three of the approaches which used the second data source improved out-of-sample predictions relative to a single data source ("Single"). When information in the second data source is of high quality, the Shared model performs the best, but the Correlation and Covariates model also perform well. When the information quality in the second data source is of lesser quality, the Correlation and Covariates model performed better suggesting they are robust alternatives when little is known about auxiliary data collected opportunistically or through citizen scientists. Methods that allow for both data types to be used will maximize the useful information available for estimating species distributions.

Key words: Brown-headed nuthatch; data fusion; multivariate conditional autoregressive; species distribution modeling.

Introduction

The last decade has seen a dramatic increase in the use of species distribution models (SDMs) to characterize patterns of species' occurrence and abundance. SDMs relate observations of a species to environmental characteristics or spatial location to better understand the processes that determine where a species occurs. The recent rise in use of SDMs has been driven by a combination of the need to forecast responses to climate and land-use change (Guisan and Thuiller 2005), greater access to data (Hochachka et al. 2012), and the availability of new estimation methods to predict distributions such as occupancy estimation (MacKenzie et al. 2002, 2003), logistic regression (Pearce and Ferrier 2000), MaxEnt (Phillips and Dudík 2008), and resource-selection functions (Manly et al. 2007). As a result, SDMs have become a key tool for both ecologists and conservation biologists.

Efforts to parameterize SDMs often create a tension between the quality and quantity of data available to fit models. Ideally SDMs would be parameterized using high quality data collected under standardized design-based sampling protocols that include randomization, consistent sampling methods, and that control for observer effort and detection uncertainty. Data of this type are rare for many species and when available often lack the sampling extent necessary to produce range-wide estimates. Instead many if not most SDMs are parameterized using data collected under non-standardized designs such as presence only data or data collection where method, effort, and sample site selection are not standardized. While use of non-standardized data is seen

Manuscript received 20 February 2016; revised 5 December 2016; accepted 14 December 2016. Corresponding Editor: Perry de Valpine.
[7]E-mail: jkpacifi@ncsu.edu

as a pragmatic solution in many cases, building models using these data often violates key assumptions of the estimation approaches used to fit SDMs (Yackulic et al. 2013).

The tension between data quality and quantity is exemplified in efforts to predict effects of land-use and climate change using SDMs. Because the amount of effort and resources needed to monitor species status across large spatial extents can be extraordinary, data collection solely by experts is often impractical. This has motivated efforts to work with the general public (i.e., citizen science) to collect data to help answer such important questions (Dickinson et al. 2010, Hochachka et al. 2012). However, citizen-science data are often limited to presence-only observations where records only include where a species is found and not where it was not found (i.e., absences). In other cases absences may be recorded, but sampling effort is not standardized and samples may be collected opportunistically. In addition, there may be greater opportunity for misidentification and false positives when data are collected by the public (e.g., Miller et al. 2013). Examples include records from museum collections, natural heritage programs, and data collected by organizations encouraging public contribution of observations through online portals (e.g., eBird or iNaturalist). While these new data sources are appealing (Phillips and Dudík 2008), there are still several looming difficulties, specifically the need for acquiring, integrating and modeling massive quantities of diverse data (Hochachka et al. 2012, Yackulic et al. 2013). Specific challenges associated with using and integrating citizen science data have been well documented (Dickinson et al. 2010, Hochachka et al. 2012). These issues are largely related to data quality (e.g., observer variation, variation in sampling effort, and measurement/recording errors) and data quantity (e.g., spatial variation in effort, ease/access of uploading data, and protocol variability). There is debate regarding the best way to analyze data that has been collected in a non-standard manner, for example presence-only data, which is often the case for citizen science projects (Dorazio 2012, Royle et al. 2012, Hastie and Fithian 2013, Phillips and Elith 2013, Renner and Warton 2013, Yackulic et al. 2013, Hefley et al. 2014). We recommend that readers explore some of the recent literature on the topic (e.g., Guillera-Arroita et al. 2015).

Our objective is to develop a framework for data integration when parameterizing SDMs that addresses the tradeoff between data quality and quantity in considering both standard and non-standard data types. We build on recent work on data integration for non-standard data types. Dorazio (2014), Fithian et al. (2015), and Giraud et al. (2016) have proposed approaches to integrate presence-only or opportunistic data across many species or in conjunction with standardized survey data (e.g., presence/absence, survey counts). These approaches are based on the use of an inhomogeneous Poisson Process (IPP) model that is thinned according to an adjustment for sampling effort. Fithian et al. (2015) and Giraud et al. (2016) independently develop similar approaches and Fithian et al. (2015) succinctly provide a clear list of limitations for both approaches, specifically the inability to account for spatial autocorrelation in occurrence, imperfect detection or spatial errors. Dorazio (2014) presents a more general case that allows for the direct estimation of imperfect detection in addition to accounting for variation in sampling effort from the presence-only data, but still does not account for spatial autocorrelation in occurrence.

Here we build on these other approaches by demonstrating a continuum of strategies that can be used to integrate data from standardized sampling protocols and non-standardized data types (e.g., opportunistic surveys, citizen science records). In addition, we show how explicitly accounting for spatial autocorrelation in estimated parameters and estimation of detection bias (both false negatives and false positives) can be used to further deal with the challenges posed by the data types. We also show the relationship between our model with a spatial prior (Multivariate Conditional Autoregressive MVCAR; Banerjee et al. 2004) and the IPP used by Dorazio (2014), Fithian et al. (2015), and Giraud et al. (2016) in Appendix S1. The general approaches we present are flexible, allowing for different data types (e.g., occurrence or abundance) and different species distribution modeling methods (e.g., occupancy estimators, Poisson point process models, and distance sampling) to be fit while using a common set of principles for integrating data sets.

Methods

Modeling framework

In developing a framework for data integration to estimate species distributions we took into account the following three principles. First, estimation should permit sharing of information across space and among different sources of data. Second, strategies used to integrate across data sources should be motivated by an understanding of the ecological, sampling, and observational processes generating the different data sources. Finally, the general approaches should be flexible enough to accommodate different data structures and to account for sources of contamination (e.g., false positives, false negatives) and variability (e.g., effort, covariates). The set of models we consider use two general strategies for integrating data. First, similar to other authors (Dorazio 2014, Fithian et al. 2015, Giraud et al. 2016) we consider how parameters are shared when fitting models for multiple data sets. Second, we provide a unique contribution by explicitly accounting for the spatial correlation of observations and parameters by using a Multivariate Conditional Autoregressive (MVCAR) spatial model (Banerjee et al. 2004, see Heisey et al. 2010, Huston and Schwarz 2012 for ecological examples). Using a multivariate spatial model allows us to account for correlation

among detection, occupancy, and abundance in data models and borrows strength across space. Different strategies for sharing latent states and including spatial covariation allow us to encode dependence among detection, occupancy, and abundance explicitly, permitting a robust framework to capture our three principles.

We develop four general classes of models based on the MVCAR that permit varying levels of complexity in both the ecological process and the degree of contamination (e.g., false negatives and false positives) of the different data sources. We begin with a model derived from a standardized design-based data source ("Single") and consider three ways to integrate a second data source, usually of lesser quality ("Shared," "Correlation," and "Covariates"). We assume the general goal is to estimate some state variable (i.e., occurrence or abundance) across space and that in all cases that variable is imperfectly observed, even where sampling has occurred, and thus is latent. To demonstrate how the approaches may be implemented, we further develop the models below for the case where occurrence is the state variable of interest.

Brown-headed nuthatch in Southeastern U.S.

As a motivating example and to facilitate understanding, we consider modeling the breeding distribution of the Brown-headed nuthatch (BHNU) using two widely available sources of data: Breeding Bird Survey data (BBS; Sauer et al. 2005), and eBird data (Sullivan et al. 2009). Here the first data type comes from a designed large scale survey effort while the second data source includes extensive observations, but with variable effort and quality from a large scale citizen science project. BBS surveys are conducted during the breeding season throughout North America and consist of a 24.5-mile-long route with stops every 0.5-mile interval. At each stop a 3-min point count is conducted with a fixed 0.25-mile radius where every bird seen or heard is recorded. eBird data come from citizen scientists from all over the world and consist of a checklist of observed species (presence/absence of species) and counts of the species in a specific location. Data is uploaded through an online portal and observers include auxiliary information (e.g., effort in hours or distance traveled, location of sightings, no. of observers) that is georeferenced and linked to each observation. Our goal is to estimate local occurrence probabilities for the BHNU whose range occurs throughout the eastern U.S. for a grid of 0.25 by 0.25 degree lat/lon cells using BBS and eBird data. In this example, there are $J = 2$ data sources with Y_{i1} equal to the number of the N_i stops where detections occur in grid cell i for the BBS data and Y_{i2} equal to total number of eBird sightings in grid cell i (Fig. 1). We first develop the most general "Shared" model where occupancy probability is jointly estimated for both data sets and then develop two alternative models ("Correlation" and "Covariates") along with a spatial model for only one data source ("Single").

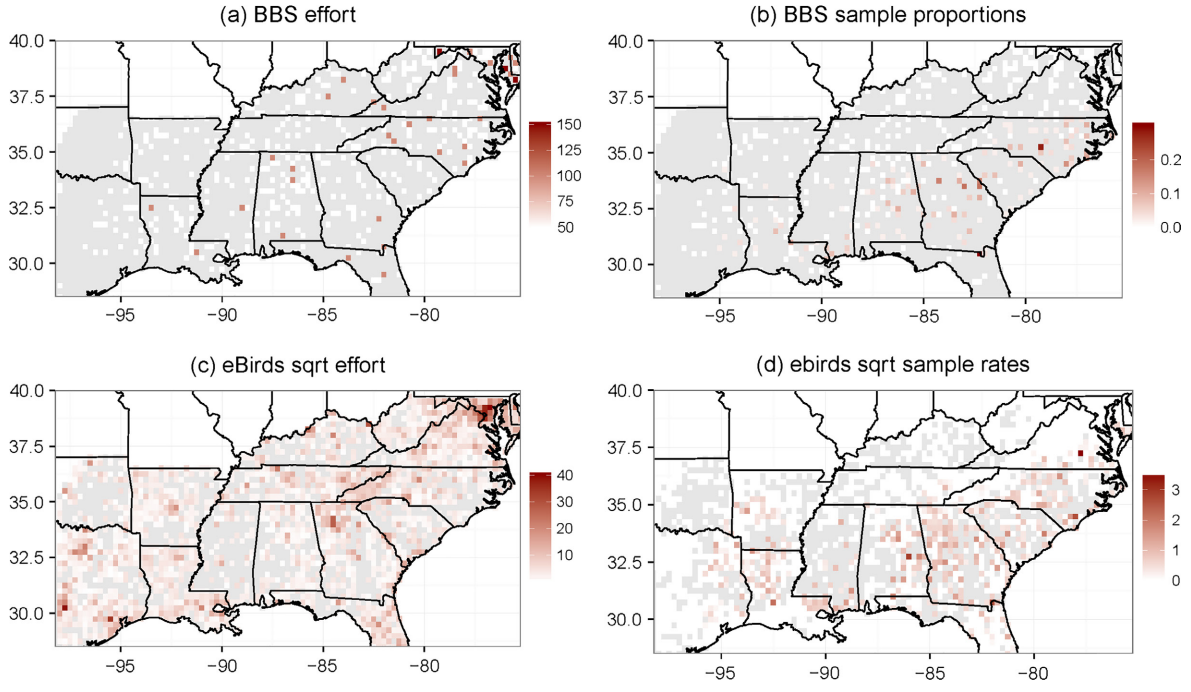

FIG. 1. Plots of the 2012 Brown-headed nuthatch data from the Breeding Bird Survey (BBS) and eBird. The top row gives the number of (a) BBS surveys (N_i) and (b) proportion of the surveys with a sighting (Y_{i1}/N_{i1}); the bottom row gives (c) the eBird effort in square root hours ($\sqrt{E_i}$) and (d) the square root of the sample proportional rate ($\sqrt{Y_{i2}/E_i}$). Locations with no sampling effort are shaded gray.

General model

Let Y_{ij} be the response for spatial location $i = 1, \ldots, n$ and data type $j = 1, \ldots, J$ after we discretize the spatial domain (note that this is not a requirement). Let's say our objective is to estimate the occupancy status $Z_i \in \{0,1\}$; if $Z_i = 1$ then the species is present in location i, and if $Z_i = 0$ then the species is not present (although this could be other population-level parameters of interest, abundance or density, without loss of generality). We assign a spatial random effect, θ_{i0}, and a random effect for observations from each data type, θ_{ij}. Conditioned on these random effects,

$$\text{Prob}\left(Z_i = 1|\theta_{i0}\right) = q(\theta_{i0}) \text{ and } Y_{ij}|Z_i,\theta_{ij} \overset{\text{indep}}{\sim} f_j\left(y|Z_i,\theta_{ij}\right) \quad (1)$$

where $q()$ is the link function, $f_j()$ is a probability density function and y is the observed data; information is potentially shared across locations and data types via the latent occurrence state Z_i and the multivariate spatial random effects $\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{i1}, \ldots \theta_{iJ})^T$.

We use a MVCAR model for the random effects $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$. The MVCAR model borrows strength across space to estimate the random effects, and also encodes dependence between the different types of random effects. The random effects are decomposed as $\theta_{ij} = \mathbf{X}_i^T\boldsymbol{\beta}_j + \alpha_{ij}$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is the vector of covariates, $\boldsymbol{\beta}_j$ is coefficient vector for random effect of data type j, and $\boldsymbol{\alpha}_i = (\alpha_{i0}, \ldots, \alpha_{iJ})^T$ are the mean-zero residuals. The multivariate spatial prior for the residuals can be defined by its full conditional distributions. Given the random effects for all other locations, for all $k \neq i$, the full conditional distribution is

$$\boldsymbol{\alpha}_i|\boldsymbol{\alpha}_k \text{ for all } k \neq i \sim \text{Normal}(\rho\bar{\boldsymbol{\alpha}}_i, 1/m_i\Sigma), \quad (2)$$

where $\rho \in (0, 1)$ controls the strength of spatial dependence; $\bar{\alpha}_i$ is the mean of the α_k over region i's m_i neighbors (e.g., rook neighbors or any neighborhood structure); and Σ is a $(J+1)$x$(J+1)$ cross-covariance matrix with $\text{cov}(\alpha_{ij},\alpha_{il}) = \Sigma_{jl}/m_i$. The cross-covariance matrix plays a central role. For example, if $\Sigma_{01} > 0$ then there is a positive correlation between the occupancy random effect θ_{i0} and the random effect for the first data type θ_{i1}.

Shared model

We use a probit link for occupancy probability $q(\theta_{i0}) = \Phi(\theta_{i0})$, where q denotes the inverse link function and Φ is the standard normal distribution function, and likelihoods:

$$\begin{aligned} Y_{i1}|Z_i,\theta_{i1} &\sim \text{Binomial}(N_i, Z_i p_i) \text{ and} \\ Y_{i2}|Z_i,\theta_{i2} &\sim \text{Poisson}[E_i\left(Z_i\lambda_i + p_0\right)] \end{aligned} \quad (3)$$

where $p_i = \Phi(\theta_{i1})$ is the BBS detection probability, $\lambda_i = \exp(\theta_{i2})$ is the eBird abundance in grid cell i, E_i is the eBird effort (auxiliary information provided with eBird data) in grid cell i, and $p_0 > 0$ is the eBird false positive rate parameter, which we treated as a constant but could vary spatially. Note that we could also include an offset (e.g., $A_i\lambda_i$, where A_i is the area of cell i) to account for variation in grid cells.

The multivariate prior for $\boldsymbol{\theta}_i$ in this context induces dependence among occupancy, detection probability, and abundance. Positive Σ_{12} denotes a positive association between detection probability and abundance where BBS data are used to model detection probability (and occurrence) and the eBird data are used to model abundance. Positive Σ_{01} denotes a positive association between occurrence and detection probability. This relationship is not well defined at a single location within a cell because detection probability is defined only conditioned on occurrence. However, this relationship is defined with data aggregated to large regions or grid cells. For example, consider the collection of cells that form the core of the species' range. These cells all likely have high occurrence and high detection probability because entire cells are occupied by the species and thus availability (i.e., probability of being available for detection) is high. In contrast, cells on the periphery of the range have low occurrence probability, and likely have low detection probability as well because if the species occupies the cell, it may only inhabit a small number of patches in the cell. Fig. 1b, d illustrates this where the non-zero counts in the center of the species' range are noticeably higher than the non-zero counts on the periphery of the species' range. A similar interpretation applies to the occupancy/abundance relationship encoded by Σ_{02}.

This "Shared" model is appropriate when all data sources are deemed reliable because each data source can directly inform the latent occurrence state and the weight given to estimate the parameters is naturally determined by their relative size and quality. However, this model may be problematic when the second data source is of poor quality. For example, consider the case where effort is high for the second data source but the resulting counts are purely noise due to unmodeled errors. In this extreme case the noisy second data source will overwhelm the first data source leading to poor estimates.

Correlation model

To add robustness to the influence of unreliable data sources, we modify (3) by removing the latent occurrence state, Z_i from Y_{i2}'s expected value:

$$\begin{aligned} Y_{i1}|Z_i,\theta_{i1} &\sim \text{Binomial}(N_i, Z_i p_i) \text{ and} \\ Y_{i2}|\theta_{i2} &\sim \text{Poisson}[E_i\lambda_i]. \end{aligned} \quad (4)$$

In (4), Y_{i2} no longer directly informs about Z_i. However, the second data source contributes information about occupancy probability indirectly through the dependence between relative abundance ($\theta_{i2} = \log(\lambda_i)$) and occupancy probability $\Phi(\theta_{i0})$. Because information is shared only through the cross-covariance Σ, we refer to the model as the "Correlation" model.

When both data sources are deemed reliable, this approach should be inferior to the "Shared" model

because the link between the second data source and occupancy probability is obscured by an additional layer of hierarchy. However, unlike (3), the link between the two data sources can be severed by simply setting $\Sigma_{02} = \Sigma_{12} = 0$. Therefore, this approach should be preferred when the quality of the second data source is less certain.

Covariate model

In (4), the second data source affects occupancy probability only through the estimated MVCAR covariance matrix. A more direct way to facilitate this information sharing is to use the second data source as a constructed covariate in the mean occupancy probability, i.e., as elements in $\mathbf{X}_i$. This allows for flexibility in exploring a wide range of constructed features that summarize the information in the second data source. Additionally, it makes no assumptions about the ecological process generating the data or the sources of error influencing this second data source. When these covariates are measured with error, estimates of their effects can be biased. Therefore we explored spatially smoothed versions of the following covariates constructed from eBird data: log abundance at site i, naïve occupancy proportion for site i, and log effort for site i, via a generalized additive model (GAM) in package mgcv (Wood 2011). After smoothing, the variables are transformed (with constants added for numerical stability) to be on the same scale as θ_i: $X_{i1} = \log(\hat{A}_i + 0.1), X_{i2} = \Phi^{-1}(0.01 + 0.98\hat{O}_i), X_{i3} = \log(\hat{E}_i)$.

Another option for a constructed covariate would be to use posterior estimates from a first-stage analysis of the second data source alone (e.g., Patton et al. 2016).

The Covariate model approach can lead to a dramatic reduction in the computational burden because there are fewer parameters to estimate and the number of data locations can be reduced. For example, in our specific case with BBS data and eBird data, the BBS data are available at far fewer sites than eBird data, and so the magnitude of the analysis is reduced significantly by only considering the BBS locations. A drawback of this model is that uncertainty in the second data source is not propagated through to the final analysis. When both data sources are reliable this approach may not make full use of the information in the second data source, but in the case when the second data source is less reliable this approach should outperform the Shared model and be similar to the Correlation model.

Single model

For comparison we include a model for only the BBS data which is a spatial MVCAR site occupancy model, $Y_{i1}|Z_i, \theta_{i1} \sim \text{Binomial}(N_i, Z_i p_i)$, where $p_i = \Phi(\theta_{i1})$ is the detection probability, Z_i is the latent occurrence state, and N_i is the number of stops in grid cell i.

We fit all four models (Shared, Correlation, Covariates, and Single) described previously to the 2012 BHNU data using Markov chain Monte Carlo to draw samples from the posterior (see Appendix S2 for description of the sampler). For all models we used uninformative priors, the elements of β_j have independent Normal $(0, 100^2)$, $\log(q) \sim \text{Normal}(0, 100^2)$, $\rho \sim \text{Uniform}(0,1)$, $\rho \sim \text{Uniform}(0,1)$, and $\Sigma \sim \text{InverseWishart}(3.1, 0.1I)$. We ran each model for 50,000 iterations and convergence was assessed visually and by monitoring trace plots and examining autocorrelation. In addition to posterior estimates of all of the parameters, we obtain a posterior mean $\hat{\mu}_i$ of the product of occupancy and detection probabilities $\hat{\mu}_i = \hat{Z}_i \hat{p}_i$ which we compare with BBS data aggregated from 2007 to 2011. Denote $\tilde{N}_i$ and $\tilde{Y}_i$ as the number of surveys and number of surveys with a sighting, respectively, for the 2007–2011 BBS data. The metrics for comparison are mean squared error, $\frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{N}_i\hat{\mu}_i)^2$, and deviance, $-2\sum_{i=1}^{n}\log[B(\tilde{Y}_i;\tilde{N}_i;\hat{\mu}_i)]$, where $B(y; N; \mu)$ is the binomial probability mass function with N trials and success probability μ; smaller scores indicate better fit for both metrics.

Simulation study

To further explore the properties of the different data-fusion approaches presented we conducted a simulation study. We generated data from two sources under the Shared model and the Correlation model with $n = 400$ cells arranged on a 20×20 square grid, with $N_i = 5$ (no. of observations in cell i) and $E_i = 10$ (effort in cell i) for all i. The parameters are set such that the mean of $\theta_{i0} = 0$, so that occupancy probability is centered around 0.5, and mean of $\theta_{i2} = -1$, $\Sigma_{00} = 2^2$, $\Sigma_{11} = 0.2^2$, $\Sigma_{22} = 2^2$, $\rho = 0.95$, and $\text{Cor}(\theta_{ij}, \theta_{il}) = \phi$. The simulations vary by whether occupancy probability is shared or not, the strength of correlation between the two data sources ($\phi \in \{0.0, 0.8\}$), which also dictates the quality of the second data source (shared: both are high quality, $\phi = 0.8$: second data source is of okay quality, $\phi = 0$: second data source is poor quality), the proportion of sites for which the first data source is included in the training data for model fitting (0.25 or 0.5), and the average detection probability ("low" sets mean of $\theta_2 = -0.5$ so that average detection probability is about 0.3, and "high" sets mean of $\theta_2 = 0$ so that average detection is about 0.5). This approach allowed us to simulate the ecological process (i.e., latent occurrence probability and spatial correlation) and the sampling process separately (sampling the unobserved latent state) by including different levels of imperfect detection (average detection probability = 0.3 or 0.5) which mimicked an observers inability to observe truth (i.e., Zurell et al. 2010). We then explored the effects of this bias along with different generating models (Shared or Correlation) and the degree of information in the second data source on the overall performance of the data fusion methods presented.

To do so we fit all four models (Shared, Correlation, Covariates, and Single) to each simulated data set. With the exception of the Covariates model, we use the intercept

TABLE 1. Percent agreement (Monte Carlo standard error) for estimating occupancy probability in the simulation study. The column labels give the data generating model and the model fit to each data set, and the rows specify the conditions of the simulation (% training is the percentage of the first data source used to fit the model, Detection represents the detection probability with low ~ 0.3 and high ~ 0.5, and ϕ is the MVCAR cross-correlation parameter).

Generating model	% training	Detection	ϕ	Single	Covariate	Shared	Correlation
Shared	25	Low	0	52.8 (0.6)	63.9 (0.8)	92.6 (0.3)	89.2 (0.5)
	25	Low	0.8	53.0 (0.6)	66.3 (1.0)	97.1 (0.1)	93.2 (0.9)
	25	High	0	52.5 (0.7)	64.6 (0.5)	92.6 (0.4)	90.0 (0.5)
	25	High	0.8	52.1 (0.7)	66.9 (0.5)	97.1 (0.1)	94.4 (0.3)
	50	Low	0	52.4 (0.7)	64.8 (0.6)	92.7 (0.4)	90.2 (0.4)
	50	Low	0.8	53.1 (0.6)	66.4 (0.5)	97.3 (0.2)	93.8 (0.3)
	50	High	0	53.8 (0.6)	64.8 (0.6)	92.8 (0.4)	91.2 (0.4)
	50	High	0.8	53.1 (0.7)	66.8 (0.5)	97.3 (0.2)	94.2 (0.3)
Correlation	25	Low	0	52.8 (0.7)	52.7 (0.6)	51.1 (0.8)	52.4 (0.6)
	25	Low	0.8	52.7 (0.7)	58.5 (0.7)	54.6 (1.1)	67.9 (0.8)
	25	High	0	52.2 (0.7)	52.0 (0.7)	50.9 (0.7)	53.7 (0.6)
	25	High	0.8	52.3 (0.7)	58.7 (0.5)	58.1 (1.3)	70.4 (0.4)
	50	Low	0	52.4 (0.7)	53.6 (0.7)	51.4 (0.8)	54.5 (0.7)
	50	Low	0.8	53.2 (0.6)	59.7 (0.7)	54.3 (1.2)	69.5 (0.6)
	50	High	0	53.7 (0.6)	53.7 (0.7)	51.4 (0.8)	58.6 (0.8)
	50	High	0.8	53.6 (0.6)	59.7 (0.6)	58.2 (1.3)	70.2 (0.5)

in $\mathbf{X}_i$. For the Covariates model we include columns in $\mathbf{X}_i$ for the smoothed covariates from the GAM. For all models we used uninformative priors, the elements of β_j have independent N(0, 100^2) priors, $\log(q) \sim$ N(0, 100^2), $\rho \sim$ Uniform(0, 1), and $\Sigma \sim \text{InvWishart}(3.1, 0.1I)$.

We simulate 100 data sets for each data generation (Shared or Correlation) scenario. The four models are fit using all n observation of the second data source and a randomly-selected subset of either 25% or 50% of the cells for the first data source. Methods are evaluated by comparing true and estimated occupancy probability to the out-of-sample hold out cells (cells that were not used with the first data source to fit the models). Denote Z_i as the true occupancy status of test site i, $\bar{Z}_i$ as the posterior mean of Z_i, and $\hat{Z}_i = I(\bar{Z}_i > 0.5)$ as the predicted value. Table 1 gives the average (over data sets) classification accuracy, i.e., the percentage of test-set regions with $\hat{Z}_i = Z_i$.

Simulation results

Not surprisingly the Shared model performs the best when the data is generated from the Shared model (Table 1). However, the Correlation model provides comparable accuracy, and the simpler Covariates model is substantially more accurate than the Single model. For data generated from the Correlation model, the agreement is generally lower (Table 1) because the contribution of the second data source is less direct than the Shared model. In fact, with correlation equal to zero (no information in the second data source), the two data sources are independent and thus the Single model is sufficient (Table 1). In these cases, the Shared model is the least accurate model because it assumes a strong relationship between data sources. The Correlation model is the most accurate when data are generated with nonzero correlation (i.e., some information in second data source), and again the Covariates model improves the accuracy compared to the Single model. Sensitivity and specificity were highest for the Shared model when it was also the generative model (Appendix S1: Tables S1 and S2) and the lowest values were for the Shared model with the Correlation model as the generative model and the two data sources were independent. We hypothesize that this occurs because observed non-zero values from the second data source are misconstrued as true occurrences by the Shared model that incorrectly assumes the true underlying occupancy state is shared by the second data source.

In summary, the simulation study confirms that the Shared model is the most powerful when both data sources provide high-quality information about occupancy probability; the Correlation model is slightly less powerful, but more robust to contamination of the second data source (e.g., low quality of information); and the Covariates model is a simple and useful addition to the Single data source model.

RESULTS

All three models that incorporated eBird data had smaller MSE and deviance than the single (BBS only) model (Table 2). The three models that incorporated eBird data had similar MSEs and the Shared occupancy model had the lowest deviance (Table 2). The most striking difference between predicted occupancy probabilities is that the single (BBS only) model leads to higher occupancy probabilities and nonzero probabilities over a much larger range, e.g., in Central Texas, Northern

TABLE 2. Mean squared error (MSE) and deviance comparing estimates of occurrence probabilities for the Brown-headed nuthatch based on 2012 breeding bird survey and eBird data to the observed 2007–2011 breeding bird survey data for the Brown headed nuthatch.

	Single	Covariate	Shared	Correlation
MSE	9.28	8.86	8.67	8.64
Deviance	3,705	3,875	3,362	3,388

Arkansas, and Tennessee, compared to the other models (Fig. 2). In contrast, the models that incorporated eBird data clearly delineated a boundary of the species range (Fig. 2). Table 3 shows the spatial variability in the performance of the models by state. It is clear that the models have similar performance in a lot of states in the central portion of the BHNU range (e.g., AL, GA, LA, NC, SC), but there is a large amount of discrepancy at the periphery of the range. The Shared model and Correlation model perform very similarly in almost all cases and the Single model has substantial differences in several states (e.g., AR, MD, TN).

The single (BBS only) model identifies strong spatial dependence and correlation between occupancy and detection probabilities, as reflected in the strong and similar spatial patterns (Figs. 2a and 3a). Both the spatial dependence and correlation between occupancy probability and detection probability slightly decrease when eBird covariates are added to the model although their credible intervals still overlap (Table 4). The eBird naïve occupancy covariate from the GAM has a statistically significant effect on both occupancy probability (1.93; posterior 90% interval [1.57, 2.35]) and detection probability (0.43; posterior 90% interval [0.15, 0.79]), and this covariate explains much of the spatial variation and dependence between these parameters. Both the Shared model and Correlation model reveal a strong dependence between occupancy probability and abundance, which leads to similar predictions to the covariate model (Fig. 2). False positive rate (*p0*) in the Shared model was estimated to be very small (0.0002, variance 4.89E-09) as there were very few cases with an eBird observation occurring without BBS observations in a cell.

DISCUSSION

We demonstrate a range of approaches that can be used to combine data, all of which improved predictions relative to single data set method in both simulated data sets and our real world example. For many species, data available for fitting distribution models will be a combination of designed surveys that employ standardized protocols and non-standard data including incidental or opportunistic observations, indices, and records from museums or citizen science programs. Methods that allow for both data types to be used will maximize the useful information available for estimating species distributions. At the same time approaches that blindly combine data with no attention to differences in quality and design are likely to induce biases and overestimate confidence in predictions.

Admittedly, the use of eBird data as a second data source provides unique advantages for our joint modeling. Specifically, the auxiliary information provided

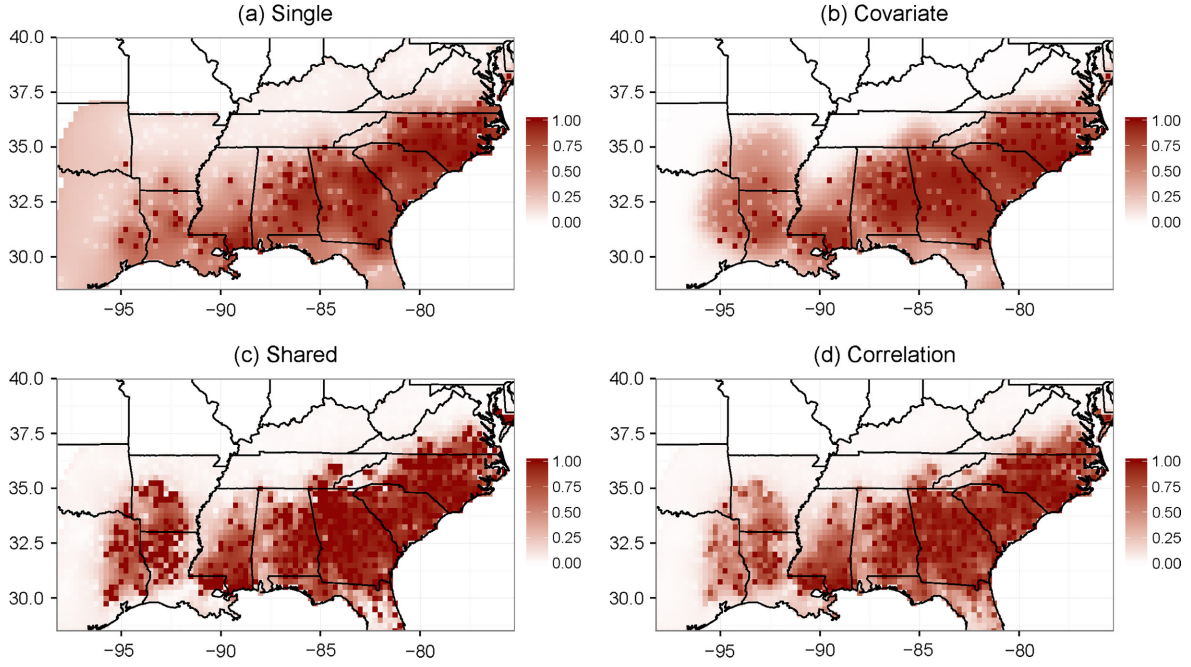

FIG. 2. Posterior mean occupancy probabilities (i.e., the posterior mean of Z_i) for the four models ([a] Single, [b] Covariate, [c] Shared, [d] Correlation) applied to the 2012 Brown-headed nuthatch data (BBS and eBird).

TABLE 3. Percent of cells (0.25 × 0.25 degrees lat/long) that have a predicted occurrence probability >0.5 for each state by each of the four models.

State	Single	Covariates	Shared	Correlation
AL	73.1	62.9	70.1	70.6
AR	0.5	17.9	28.3	17.5
FL	52.8	57.6	56.8	53.6
GA	96.5	97.8	98.3	97.8
LA	63.4	57.4	44.8	43.7
MD	5.4	5.4	16.2	16.2
MS	39.5	34.2	51.1	50.5
NC	80.6	87.7	80.6	80.6
OK	0.6	0.6	2.4	1.8
SC	95.1	100	98.4	98.4
TN	0.5	4.3	10.8	6.5
TX	11.1	16.5	24.5	11.4
VA	9.9	18.0	24.2	24.2

with eBird data (e.g., effort, survey time, distance traveled), allows us to estimate many sources of variation (false positives and false negatives) and to have completely identifiable parameters (see Appendix S1). As noted by Dorazio (2014) and Fithian et al. (2015), without this information one must rely on strict assumptions to ensure parameter identifiability or to use the condition number of the Fisher information matrix to assess the identifiability of parameters (Dorazio 2014). Being able to incorporate information directly on sampling effort allows us greater flexibility in the models we can fit and the sources of variation we can explore. We suggest users look to capitalize on auxiliary information that is often included with presence-only and opportunistic survey records.

Each of the three approaches to data integration that we outline has advantages and disadvantages and each will be most useful in different situations. The Shared model may be seen as the most appropriate when both data sets are of high quality. In this model both data sets directly inform the distribution model through shared parameters in a jointly estimated likelihood. Information from all data sources is fully utilized and combining data may improve identifiability of parameters allowing one to fit models in situations where a single data type may be prohibitive (e.g., Dorazio 2014). At the same time, this approach places all data on similar footing and in cases when quality of data is not fully known or sampling effort or observational uncertainty is difficult to model, the approach may not be warranted. In our analysis, we relax the assumption of no false positives for the eBird data, which may alleviate some of the potential problems that could occur due to misidentification and poor quality data. Miller et al. (2013) and Pillay et al. (2014) provide examples of combining presence/absence datasets using a similar approach that allows for false positive detection errors in data collected from non-experts.

Alternatively, both the Correlation and Covariate approaches should limit bias that may come from using non-standardized data sources. As shown in our simulations, in cases where sampling effort is hard to recover or misidentification errors are more prominent, estimates using these approaches can be more accurate than the Shared approach. Our Correlation and Covariate approaches still take advantage of information from

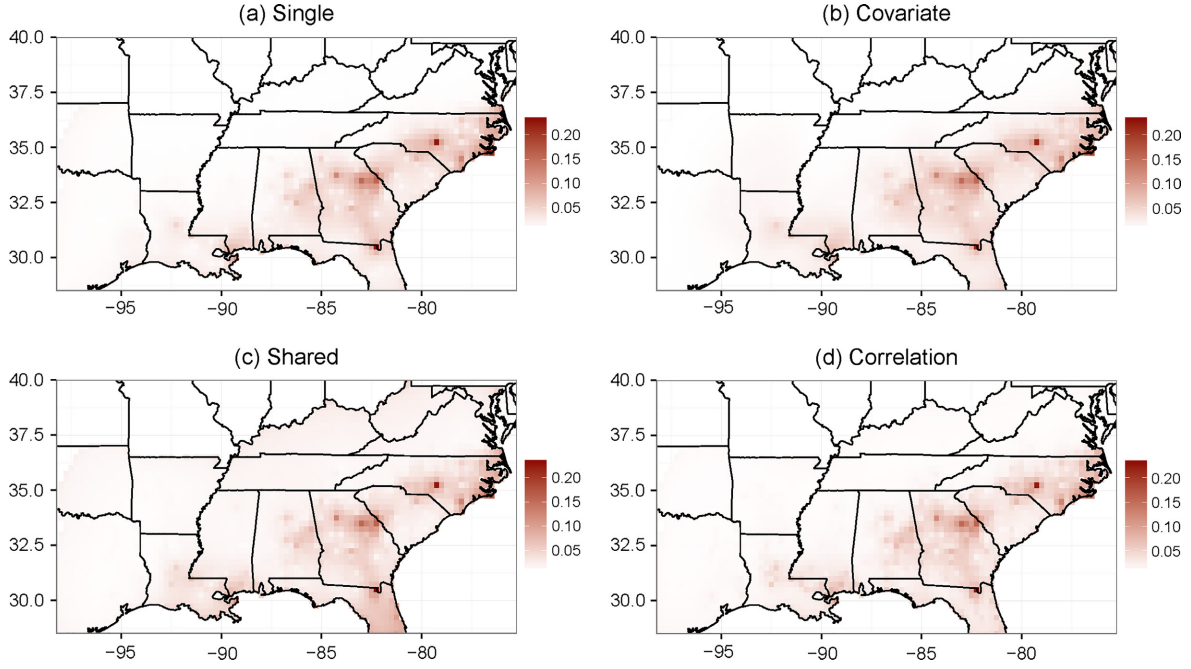

FIG. 3. Posterior mean detection probabilities ($\Phi(\theta_{i1})$) for the four models ([a] Single, [b] Covariate, [c] Shared, [d] Correlation) applied to the 2012 Brown-headed nuthatch data (BBS and eBird).

TABLE 4. Posterior median (90% interval) of the MVCAR correlation (elements of the correlation matrix corresponding to covariance Σ) and MVCAR spatial dependence parameter (ρ) for each model fit to the 2012 Brown-headed nuthatch data (breeding bird survey and eBird).

Model	Occupancy-Detection	Occupancy-Abundance	Detection-Abundance	Spatial dependence
Single	0.82 (0.45, 0.93)	–	–	1.000 (0.998, 1.000)
Covariate	0.97 (0.93, 0.98)	–	–	0.997 (0.991, 1.000)
Shared	0.08 (−0.36, 0.65)	0.84 (0.49, 0.93)	0.19 (−0.13, 0.46)	1.000 (0.999, 1.000)
Correlation	0.50 (−0.02, 0.82)	0.91 (0.75, 0.97)	0.36 (−0.06, 0.66)	1.000 (0.999, 1.000)

non-standard data. This comes by specifying a relationship through covariance or covariate structure to the underlying state process estimated from data collected under a standard design. In both our main analysis and simulations these improve fit compared to models fit using only a single standard data source.

Combining data using the Correlation model has an additional advantage in that it provides an ability to share information when it is not possible to specify shared parameters between data types. One example of when this may occur is when state variables are measured at different scales. For example, the design used in our BBS-eBird example could be modified to measure processes at different scales for each of the data types although a change of support prior may be needed (Banerjee et al. 2004). One could imagine a scenario where intensity of eBird observations is still measured at a large-scale, such as the grid size we use in this analysis, but more detailed fixed-radius point count data (e.g., that allowed for detection to be estimated) were used to estimate occurrence at smaller units within the larger cells. It would still be logical to assume intensity of eBird observations at the larger scale would inform local occurrence probabilities of points within each cell and would require an offset to account for the variable sized areas. Similarly, using the BBS-eBird example again, if both processes had been measured at fine scales, linking eBird data to occurrence probability through zero-inflation is unlikely to work well since non-occurrence in many cases will be explained by low intensity in the Poisson process model. However, specifying a covariance structure between occupancy probability and intensity would allow for the two data types to be linked in the analysis.

Perhaps the most appealing aspect of the Covariate approach is the ease of implementation. The only requirement is the ability to generate a set of predictors from the non-standard data. These predictors can then be integrated within any existing SDM modeling framework that allows for covariates. This opens up the approach to use in most standard approaches and statistical packages used to estimate species distributions (e.g., Package unmarked, Fiske and Chandler 2011). In both our example data set and simulations the Covariate model was a distinct improvement over models that only used the standard data type when estimating species distributions.

One is not limited to the specific covariates we considered in this study. For example, it might be useful to generate model based estimates of relative occurrence or abundance from the non-standard data and then use the predicted occurrence or abundance values as covariates in the model for the standard data set. Or one might use covariates for other species expected to be positively (or negatively) correlated with the focal species to predict the focal species distribution (Patton et al. 2016). This would be a flexible and easy to implement an approach for including among-species covariance in predictions. Recent work has demonstrated the value of species co-occurrence in predicting species distributions (Ovaskainen et al. 2010, Clark et al. 2014, Pollock et al. 2014, Fithian et al. 2015).

Modeling of spatial variation using spatial random effects was an important component of the modeling framework we present but not necessarily a requirement. However, spatial random-effects are likely to be especially useful to improve estimates by allowing information to be shared across data sets in cases where the locations of observations do not necessarily align, but are in close proximity or when improving upon standardized designs (Pacifici et al. 2016). Some caution is needed, however, when interpreting additional covariates when spatial random effects are included because the inclusion of such effects can inflate the variance of the fixed effects (see Hodges and Reich 2010, Johnson et al. 2013).

Here we focused on two specific data types (BBS – standardized fixed effort point counts; eBird – citizen science observations indexed by effort) and modeling frameworks (BBS – presence/absence models with spatial replication; eBird – a discretized Poisson point process model). However, the approaches we have outlined can readily be applied to many data types and modeling frameworks. Species distribution models are generally used to predict occurrence or abundance, with a wide range of methods available for estimating each. Much recent work has gone into developing hierarchical frameworks for estimating spatial variation in abundance using repeat counts (N-mixture; Royle 2004, Kéry et al. 2005), mark-recapture (Otis et al. 1978, Royle et al. 2014), and distance sampling (Buckland et al. 2005, Sillett et al. 2012). Similarly a wide range of approaches are used for estimating occurrence, including methods for presence-only data that can be used for non-standardized data types (Dorazio 2012). We see the three general approaches we outline as a flexible framework that can be used as a starting point, regardless of the data being used to parameterize models.

Acknowledgments

We would like to acknowledge SAMSI for organizing a session on Mathematical/Statistical Ecology. Funding was provided by the U.S. Geological Survey through North Carolina Cooperative Fish and Wildlife Research Unit Research Work Order 215. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. government.

Literature Cited

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical modeling and analysis for spatial data. CRC Press, CRC Press, Boca Raton, Forida, USA.

Buckland, S. T., D. R. Anderson, K. P. Burnham, and J. L. Laake. 2005. Distance sampling. Encyclopedia of Biostatistics 2.

Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecological Applications 24:990–999.

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter. 2010. Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41:149–172.

Dorazio, R. M. 2012. Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68:1303–1312.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23:1472–1484.

Fiske, I., and R. Chandler. 2011. Unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. Journal of Statistical Software 43:1–23. https://cran.r-project.org/web/packages/unmarked/index.html

Fithian, W., J. Elith, T. Hastie, and D. A. Keith. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6:424–438.

Giraud, C., C. Calenge, C. Coron, and R. Julliard. 2016. Capitalizing on opportunistic data for monitoring relative abundances of species. Biometrics 72:649–658.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24:276–292.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Hastie, T., and W. Fithian. 2013. Inference from presence-only data; the ongoing controversy. Ecography 36: 864–867.

Hefley, T. J., D. M. Baasch, A. J. Tyre, and E. E. Blankenship. 2014. Correction of location errors for presence-only species distribution models. Methods in Ecology and Evolution 5:207–214.

Heisey, D. M., E. E. Osnas, P. C. Cross, D. O. Joly, J. A. Langenberg, and M. W. Miller. 2010. Linking process to pattern: estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data. Ecological Monographs 80:221–240.

Hochachka, W. M., D. Fink, R. A. Hutchinson, D. Sheldon, W. K. Wong, and S. Kelling. 2012. Data-intensive science applied to broad-scale citizen science. Trends in Ecology and Evolution 27:130–137.

Hodges, J. S., and B. J. Reich. 2010. Adding spatially-correlated errors can mess up the fixed effects you love. American Statistician 64:325–334.

Huston, C., and C. Schwarz. 2012. Hierarchical Bayesian strategy for modeling correlated compositional data with observed zero counts. Environmental and Ecological Statistics 19:327–344.

Johnson, D. S., P. B. Conn, M. B. Hooten, J. C. Ray, and B. A. Pond. 2013. Spatial occupancy models for large data sets. Ecology 94:801–808.

Kéry, M., J. A. Royle, and H. Schmid. 2005. Modeling avian abundance from replicated counts using binomial mixture models. Ecological Applications 15:1450–1461.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84:2200–2207.

Manly, B. F. J., L. L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erickson. 2007. Resource selection by animals: statistical designs and analysis for field studies. Kluwer Academic Publishers, New York, New York, USA.

Miller, D. A. W., J. D. Nichols, J. A. Gude, L. N. Rich, K. M. Podruzny, J. E. Hines, and M. S. Mitchell. 2013. Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. PLoS One 8:e65808.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62:1–135.

Ovaskainen, O., J. Hottola, and J. Siitonen. 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91:2514–2521.

Pacifici, K., B. J. Reich, R. M. Dorazio, and M. J. Conroy. 2016. Occupancy estimation for rare species using a spatially-adaptive sampling design. Methods in Ecology and Evolution 7:285–293.

Patton, P. T., K. Pacifici, and J. Collazo. 2016. Inferring habitat quality and habitat selection using static site occupancy models. arXiv:1607.05175 [q-bio.PE] preprint.

Pearce, J., and S. Ferrier. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological Modeling 128:127–147.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175.

Phillips, S. J., and J. Elith. 2013. On estimating probability of presence from use-availability or presence-background data. Ecology 94:1409–1419.

Pillay, R., D. A. W. Miller, J. E. Hines, A. A. Joshi, and M. D. Madhusudan. 2014. Accounting for false positives improves estimates of occuapncy from key informant surveys. Diversity and Distributions 20:223–235.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modeling species simultaneously with a joint species distribution model. Methods in Ecology and Evolution 5:397–406.

Renner, I. W., and D. I. Warton. 2013. Equivalence of MAXENT and Poisson process models for species distribution modeling in ecology. Biometrics 69:274–281.

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

Royle, J. A., R. B. Chandler, C. Yackulic, and J. D. Nichols. 2012. Likelihood analysis of species occurrence probability from presence-only data for modeling species distributions. Methods in Ecology and Evolution 3:545–554.

Royle, J. A., R. B. Chandler, R. Sollmann, and B. Gardner. 2014. Spatial capture-recapture. Academic Press, Waltham, Massachusetts, USA.

Sauer, J. R., J. E. Hines, and J. Fallon. 2005. The North American breeding bird survey, results and analysis 1966–2005. Version 6.2.2006. USGS Patuxent Wildlife Research Center, Laurel, Maryland, USA.

Sillett, T. S., R. B. Chandler, J. A. Royle, M. Kéry, and S. A. Morrison. 2012. Hierarchical distance-sampling models to estimate population size and habitat-specific abundance of an island endemic. Ecological Applications 22:1997–2006.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society B 73:3–36.

Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant, and S. Veran. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution 4: 236–243.

Zurell, D., et al. 2010. The virtual ecologist approach: simulating data and observers. Oikos 119:622–635.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.1710/suppinfo

Appendix S1

Reformulation of the Shared model as a thinned inhomogeneous Poisson Process (IPP) with spatial autocorrelation

In this appendix we derive our Shared model as a series of thinned inhomogeneous Poisson processes to show the connection between our model and those of Dorazio (2014), Fithian et al. (2015), and Giraud et al. (2016) with the exception that we include explicit spatial autocorrelation via a Multivariate CAR model (MVCAR). We begin with a process model at the individual level to lay out the assumptions needed to arrive at our final Shared model.

Assume the locations of individuals in the population follow an inhomogeneous Poisson process with intensity $\lambda(\mathbf{s}) \geq 0$. Each data source is imperfect, and so define $\omega_j(\mathbf{s}) \in [0, 1]$ as the thinning process for data source $j = 1, 2$ so that sightings from data source j follow a Poisson process with intensity $\lambda_j(\mathbf{s}) = \theta_j(\mathbf{s})\lambda(\mathbf{s})$. The observed data are aggregated by regions $B_1, ..., B_n$, and for identifiability we model only the integrated intensities over these regions with the assumption that the thinning processes and the intensity are constant within these regions: $\omega_j(\mathbf{s}) = \exp(\gamma_{ij})$ and $\lambda_i = Z_i \exp(\gamma_{i0})$ for all $\mathbf{s} \in B_i$, where Z_i is the binary indicator that the species occupies region i. The integrated intensities over region i for response j is then $\lambda_{ij} = \int_{B_i} \lambda_j(\mathbf{s}) d\mathbf{s} = Z_i |B_i| \exp(\delta_{0i} + \delta_{ji})$.

In our application, the eBird counts in cell i are then distributed as $Y_{i2} \sim$ Poisson$[Z_i \exp(\delta_{0i} + \delta_{2i} + \log |B_i|)]$ and the BBS counts are distributed as Poisson$[Z_i \exp(\delta_{0i} + \delta_{1i} + \log |B_i|)]$. The BBS data are given as presence/absence data with the probability of occurrence in sampling occasion $l = 1, ..., N_i$ equal to Prob$(Y_{il1} = 1) = Z_i\{1 - \exp[\exp(\delta_{0i} + \delta_{1i} + \log |B_i|)]\}$. Without external information it is not possible to identify all three latent processes δ_{0i}, δ_{1i}, and δ_{2i}, and so we directly model $\theta_{ji} = \delta_{0i} + \delta_{ji} + \log |B_i|$ for $j = 1, 2$. After these assumptions we have $Y_{ij} \sim f_j(Z_i, \theta_{ji})$ as in (1). This is similar to the thinned IPP that others have developed. We then extend their work by directly incorporating spatial autocorrelation by specifying an MVCAR model for the θ_{ji}, thus completing the Shared model (see Pacifici et al. for full details).

1 References

Dorzio, R.M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23:1472-1484.

Fithian, W., J. Elith, T. Hastie, and D.A. Keith. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6:424-438.

Giraud, C., Calenge, C., Coron, C. and R. Julliard. 2016. Capitalizing on opportunistic data for monitoring relative abundances of species. Biometrics 72: 649-658.

Table S1. Sensitivity (Monte Carlo standard error) for estimating occupancy probability in the simulation study. The column labels give the data generating model and the model fit to each data set, and the rows specify the conditions of the simulation (% training is the percentage of the first data source used to fit the model, Detection represents the detection probability with low ~ 0.3 and high ~ 0.5, and ϕ is the MVCAR cross-correlation parameter).

Generating Model	% training	Detection	ϕ	Single	Covariate	Shared	Correlation
Shared	25	Low	0	47.8 (4.9)	72.2 (1.8)	85.2 (0.4)	79.8 (0.7)
	25	Low	0.8	61.1 (4.7)	72.1 (1.5)	94.4 (0.2)	87.9 (0.5)
	25	High	0	48.3 (4.8)	64.4 (1.6)	85.3 (0.4)	81.4 (0.6)
	25	High	0.8	51.4 (4.8)	66.2 (1.3)	94.4 (0.2)	88.7 (0.4)
	50	Low	0	52.5 (4.8)	66.4 (1.5)	85.3 (0.5)	80.9 (0.6)
	50	Low	0.8	53.8 (4.6)	68.7 (1.3)	94.8 (0.3)	87.7 (0.4)
	50	High	0	55.1 (4.7)	65.2 (1.4)	85.4 (0.5)	82.7 (0.5)
	50	High	0.8	58.2 (4.7)	66.9 (1.2)	94.8 (0.3)	88.3 (0.4)
Correlation	25	Low	0	53.1 (4.9)	61.5 (3.5)	98.3 (0.5)	74.2 (2.9)
	25	Low	0.8	56.9 (4.8)	69.1 (2.5)	95.6 (0.8)	73.2 (1.4)
	25	High	0	46.7 (4.9)	50.1 (3.7)	98.7 (0.5)	62.9 (3.1)
	25	High	0.8	48.9 (4.8)	59.0 (2.7)	92.4 (0.9)	68.2 (1.1)
	50	Low	0	48.9 (4.8)	59.0 (3.6)	99.5 (0.1)	77.1 (2.5)
	50	Low	0.8	50.6 (4.7)	65.0 (2.2)	95.4 (0.7)	69.9 (1.0)
	50	High	0	54.6 (4.7)	52.1 (3.6)	99.0 (0.2)	65.7 (2.6)
	50	High	0.8	56.6 (4.7)	61.6 (2.1)	92.3 (0.7)	68.0 (0.9)

Table S2. Specificity (Monte Carlo standard error) for estimating occupancy probability in the simulation study. The column labels give the data generating model and the model fit to each data set, and the rows specify the conditions of the simulation (% training is the percentage of the first data source used to fit the model, Detection represents the detection probability with low ~ 0.3 and high ~ 0.5, and ϕ is the MVCAR cross-correlation parameter).

Generating Model	% training	Detection	ϕ	Single	Covariate	Shared	Correlation
Shared	25	Low	0	52.5 (4.9)	54.2 (2.1)	99.6 (0.1)	96.9 (1.2)
	25	Low	0.8	39.5 (4.7)	59.0 (1.9)	100.0 (0.0)	97.5 (1.3)
	25	High	0	52.4 (4.7)	63.7 (1.5)	99.5 (0.2)	98.2 (0.6)
	25	High	0.8	49.2 (4.8)	66.5 (1.2)	100.0 (0.0)	99.8 (0.1)
	50	Low	0	47.9 (4.9)	61.5 (1.6)	99.5 (0.1)	99.7 (0.1)
	50	Low	0.8	47.7 (4.7)	62.9 (1.4)	100.0 (0.0)	99.9 (0.0)
	50	High	0	46.5 (4.7)	62.7 (1.4)	99.4 (0.1)	99.4 (0.2)
	50	High	0.8	42.5 (4.7)	65.3 (1.2)	100.0 (0.0)	99.9 (0.1)
Correlation	25	Low	0	47.4 (4.9)	40.3 (3.5)	1.5 (0.5)	27.6 (2.9)
	25	Low	0.8	43.7 (4.8)	45.2 (2.7)	13.6 (2.1)	62.2 (2.1)
	25	High	0	53.6 (4.9)	51.3 (3.7)	1.5 (0.5)	40.7 (3.1)
	25	High	0.8	51.9 (4.8)	56.4 (2.6)	23.6 (2.5)	70.8 (1.1)
	50	Low	0	51.8 (4.8)	44.2 (3.7)	0.6 (0.2)	27.6 (2.8)
	50	Low	0.8	50.6 (4.7)	51.9 (2.3)	15.5 (2.0)	69.1 (1.1)
	50	High	0	46.4 (4.7)	50.7 (3.6)	1.1 (0.2)	46.1 (2.7)
	50	High	0.8	44.6 (4.8)	55.1 (2.1)	26.4 (2.2)	71.5 (0.9)

Appendix S2

Integrating multiple data sources in species distribution modeling: a framework for data fusion

1 MCMC Details

We use Markov chain Monte Carlo to draw samples from the posterior. Specifically, we use Metropolis within Gibbs sampling (Carlin and Louis (2008)). The algorithm is described below for the shared-occupancy model; the algorithms for other models are either special cases of this algorithm (the single and covariate models take $E_i = 0$) or a minor modification of this algorithm (correlation model). The algorithm begins by setting initial values for all parameters, $\{Z_i, \boldsymbol{\theta}_i, \boldsymbol{\beta}_j, \Sigma, p_0, \rho\}$, and then sequentially updating each parameter conditioned on all other parameters. The parameters Z_i, $\boldsymbol{\beta}_j$, and $\boldsymbol{\Sigma}$ have conditionally-conjugate priors and are thus updated by sampling from their full conditional distributions; $\boldsymbol{\theta}_i$, p_0, and ρ do not have conditionally conjugate prior and are updated using a Metropolis step. We sample 20000 iterations and discard the first 5000 as burn-in. The individual updates are described below.

The full conditional distribution of Z_i is

$$\text{Prob}(Z_i = 1|\cdot) = \left[1 + \frac{B(Y_{i1}; N_i, 0)P(Y_{i2}; E_i p_0)[1 - \Phi(\theta_{i0})]}{B[Y_{i1}; N_i, \Phi(\theta_{i1})]P(Y_{i2}; E_i[\exp(\theta_{i2}) + p_0])\Phi(\theta_{i0})}\right]^{-1}, \tag{S1}$$

where B and P are the binomial and Poisson mass functions, respectively. The full conditional distribution for $\boldsymbol{\beta}_j$ and $\boldsymbol{\Sigma} = \Omega^{-1}$ are

$$\begin{aligned} \boldsymbol{\beta}_j|\cdot &\sim \text{Normal}\left[\mathbf{V}_j\mathbf{X}^T\mathbf{Q}\left(\Omega_{jj}\mathbf{G}_j + \sum_{k\neq j}\Omega_{jk}(\mathbf{G}_k - \mathbf{X}\boldsymbol{\beta}_k)\right), \mathbf{V}_j\right] \\ \boldsymbol{\Sigma}|\cdot &\sim \text{InvWishart}\left[n + 3 + \epsilon, (\mathbf{H} + \epsilon I)^{-1}\right] \end{aligned} \tag{S2}$$

where the priors are $\boldsymbol{\beta}_j \sim \text{Normal}(0, c^2 I)$ and $\boldsymbol{\Sigma} \sim \text{InvWishart}(3 + \epsilon, \epsilon I)$, the $n \times n$ CAR inverse covariance matrix is $\mathbf{Q} = \mathbf{M} - \rho\mathbf{A}$, $\mathbf{A}$ is the spatial adjacency matrix with (i, j) element $I(i \sim j)$ (zero on the diagonal) and $\mathbf{M}$ is the diagonal matrix with diagonal elements $m_1, ..., m_n$, $\mathbf{V}_j = \left[\Omega_{jj}\mathbf{X}^T\mathbf{Q}\mathbf{X} + c^{-2}I\right]^{-1}$, $\mathbf{H}$ is the 3×3 matrix with (j, k) element $(\mathbf{G}_j - \mathbf{X}\boldsymbol{\beta}_j)^T Q(\mathbf{G}_k - \mathbf{X}\boldsymbol{\beta}_k)$, and $\mathbf{G}_j = (\theta_{1j}, ..., \theta_{nj})^T$.

The random effect at spatial location i, $\boldsymbol{\theta}_i$, is updated using Metropolis sampling. We use the full conditional normal MCAR prior distribution, and thus the acceptance probability is $\min\{1, R\}$ where

$$R = \frac{B[Y_{i1}; N_i, Z_i\Phi(\theta^*_{i1})]P(Y_{i2}; E_i[Z_i \exp(\theta^*_{i2}) + p_0])B[Z_i; 1, \Phi(\theta^*_{i0})]}{B[Y_{i1}; N_i, Z_i\Phi(\theta'_{i1})]P(Y_{i2}; E_i[Z_i \exp(\theta'_{i2}) + p_0])B[Z_i; 1, \Phi(\theta'_{i0})]},$$

$(\theta^*_{i0}, \theta^*_{i1}, \theta^*_{i2})^T$ is the candidate and $(\theta'_{i0}, \theta'_{i1}, \theta'_{i2})^T$ is the current value. The remaining parameters ρ and $\log(p_0)$ are also updated using Metropolis sampling, with beta and Gaussian candidate distributions, respectively, adaptively tuned during burn-in to give acceptance probability near 0.3.

References

Carlin, B. P. and Louis, T. A. (2008) *Bayesian methods for data analysis*. CRC Press.