Integrating auxiliary data in optimal spatial design for species distribution modeling

Brian J. Reich[1], Krishna Pacifici, and Jonathan W. Stallings

North Carolina State University

February 13, 2018

Abstract

(1) Traditional surveys used to create species distribution maps and estimate ecological relationships are expensive and time consuming. Citizen science offers a way to collect a massive amount of data at negligible cost and has been shown to be a useful supplement to traditional analyses. However, there remains a need to conduct formal surveys to firmly establish ecological relationships and trends.

(2) In this paper, we investigate the use of auxiliary (e.g., citizen science) data as a guide to designing more efficient ecological surveys. Our aim is to explore the use of opportunistic data to inform spatial survey design through a novel objective function that minimizes misclassificaton rate (i.e. false positives and false negatives) of the estimated occupancy maps. We use an initial occupancy estimate from auxiliary data as the prior in a Bayesian spatial occupancy model, and an efficient posterior approximation that accounts for spatial dependence, covariate effects, and imperfect detection in an exchange algorithm to search for the optimal set of sampling locations to minimize misclassification rate.

(3) We examine the optimal design as a function of the detection rate and quality of the citizen-science data, and compare this optimal design with several common ad hoc designs via an extensive simulation study. We then apply our method to eBird data for the brown-headed nuthatch in the Southeast US.

(4) We argue that planning a survey with the use of auxiliary data improves estimation accuracy and may significantly reduce the costs of sampling.

Key words: Bayesian inference; Citizen science; Exchange algorithm; Geostatistics; Imperfect detection; Occupancy.[2]

[1]Department of Statistics; Campus Box 8203; Raleigh, NC 27695; bjreich@ncsu.edu

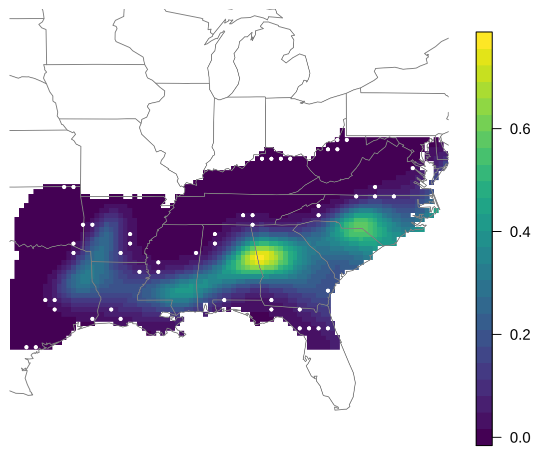
[2]Word count: Approximately 6,500; Running title: Integrating auxiliary data for optimal design

1 Introduction

Collecting data to accurately estimate the distribution of a species is a labor-intensive endeavor. For example, the Breeding Bird Survey is a network of hundreds of routes surveyed by thousands of volunteers that has been active since 1966 (Sauer et al., 2005). Even with a team of trained volunteers, sites are visited only once per year and there are large gaps in effort across years. Recent advances in statistical modeling have been applied to maximize the utility of these data and subsequently, a plethora of methods have been proposed to analyze ecological data that account for survey design and borrow strength across nearby survey sites to produce distribution maps (e.g., Royle and Wikle, 2005; Royle et al., 2007; Dorazio, 2007; Barbet-Massin et al., 2012; Johnson et al., 2013; Bailey et al., 2014; Conn et al., 2015).

An emerging line of research is to supplement systematic survey data (such as the Breeding Bird Survey) with massive auxiliary data (e.g., citizen science data) such as the Cornell Lab of Ornithology's eBird database (Sullivan et al., 2009), consisting of millions of data points from thousands of citizen scientists each year. Carefully exploiting the strengths of these two data streams can lead to improved estimates of species distributions (e.g., Dorazio, 2014; Pacifici et al., 2017). In this paper we develop a new method to optimally design a spatial survey that uses auxiliary (e.g., citizen science) data as a guide. We assume that the auxiliary data will be used in the eventual analysis of the survey data, and select the survey sample locations to minimize the expected misclassification rate (false positives and false negatives) of the joint analysis. As an example, Figure 1 plots an initial estimate of brown-headed nuthatch relative abundance based on eBird data and the recommended locations for a systematic survey of the southeast US (this map is described in detail in Section 3). Our premise is that by selecting locations based on the auxiliary

Figure 1: **The optimal sampling locations for the brown-headed nuthatch.** Initial estimate of brown-headed nuthatch relative abundance based on eBird data (background color) and the recommended locations for a systematic survey of the southeast US (white points).



data we can improve precision without additional sampling effort.

Ecological studies pose unique design challenges. Data are typically non-Gaussian (e.g., number of observed animals, latent binary occurrence state) and detection is imperfect. The non-spatial ecological design literature often focuses on the trade-off between taking few samples at a large number of sampling locations to estimate population characteristics and covariate effects, and taking many samples at a few locations to estimate and account for imperfect detection (MacKenzie and Royle, 2005; Bailey et al., 2007; Guillera-Arroita et al., 2010; Guillera-Arroita and Lahoz-Monfort, 2012). This balance is even more delicate when the survey is designed to study multiple species with potentially different detection rates (Sanderlin et al., 2014; Sliwinski et al., 2016).

Another challenge is that ecological data often exhibit spatial correlation (e.g., Johnson et al., 2013). Unfortunately, the ecological design literature has not kept pace with the developments in spatial modeling. Optimal spatial designs (Mateu and Müller, 2012) balance distributing the sampling locations uniformly over the study domain (e.g., a space-filling design) to maximize coverage for spatial interpolation, and clustering sample locations to permit estimation of the spatial correlation function (Fortin et al., 1990; Royle and Nychka, 1998; Royle, 2000; Ver Hoef, 2012; Hanks et al., 2016). Several two-stage designs have been proposed (Guillera-Arroita et al., 2014; Pacifici et al., 2016) that first analyze data from an initial sample, and then add more observations based on the interim results to target areas of uncertainty remaining after the first stage. Several authors have extended this approach to continuously monitor processes evolving over space and time, and select sites based on repeated interim analyses to minimize uncertainty about the ecological process and its evolution (Wikle and Royle, 1999, 2005; Williams et al., 2017).

In this paper, we explore spatial design guided by an initial occupancy estimate from an auxiliary data source. We use a Bayesian occupancy model that incorporates the auxiliary-data estimates in the prior while accounting for imperfect detection and spatial dependence between the occupancy status of nearby regions. The proposed design can be viewed as a two-stage design where the second-stage systematic survey is designed to reduce uncertainty remaining after the first-stage analysis of auxiliary data. However, unlike other two-stage designs, the data sources for the two stages are different and potential biases and other discrepancies between the data sources must be considered. Because of the computational challenges associated with large spatial data sets (Johnson et al., 2013) and the need to explore a wide range of design features, we develop an efficient approximation to the posterior occupancy probabilities, and use this approximation in an exchange algorithm to identify the set of sampling locations that minimize expected misclassification rate. We compare the optimal spatial design as a function of the presumed detection probability and quality of the auxiliary data, and conduct a simulation study to compare occupancy estimates from the optimal design with several ad hoc designs. We conclude by recommending a survey design for the brown-headed nuthatch (*Sitta pusilla*) in the southeast US using eBird data for initial estimates.

2 Methods and materials

2.1 Overview of optimal design

Classical experimental design (Shah and Sinha, 1989; Pukelsheim, 2006; Atkinson et al., 2007; Jones and Goos, 2011) specifies a design space (the set of possible designs) and an optimality criteria, and seeks the optimal design, defined as the member of the design space that optimizes the criteria. For example optimally estimating the regression coefficients in linear regression, we may restrict all covariates to the unit interval and select the design matrix that minimizes the average variance of the regression coefficients. In our spatial design problem, the design space constitutes all possible sampling locations and we select the subset of these location to minimize the misclassification rate.

The optimization problem can be analytically solved in some cases, but is generally intractable and optimization algorithms must be used. These algorithms typically involve some variant of adding, deleting, or exchanging the design points of an initial design (e.g. Atkinson et al., 2007, Chapter 7). Whenever the initial design is modified, the criterion measure is recalculated and the algorithm continues until no or little improvement is made. These algorithms often converge to local optima so multiple initial designs are used, reporting the best design across all initial designs considered.

In our motivating example of estimating a species distribution using occupancy surveys, the data are non-Gaussian and spatially-correlated. Optimal designs for non-normal responses have been a topic of intense research for the last few decades (see Khuri et al., 2006; Hinkelmann, 2012). Incorporating spatial correlation in the design evaluation implicitly involves spatial location in design selection. Benedetti and Palma (1995) demonstrate that correlation impacts the optimal sampling design. Müller (2007) and Mateu and Müller (2012) provide an excellent review of optimal design theory applied to spatiotemporal design. Spatially-correlated and non-Gaussian data make finding the optimal design more challenging because evaluating the criteria for a candidate design is computationally intensive and the criterion may depend on the true and unknown model parameters. In this case, an optimal design for a given value of the parameters is called a locally-optimal design, and the optimal design integrating over prior uncertainty gives the Bayesian-optimal design (Chaloner and Verdinelli, 1995; Ryan et al., 2016).

2.2 Bayesian occupancy model

For design purposes, we assume that there are N possible spatial locations of interest, $\mathbf{s}_1, ..., \mathbf{s}_N$, and that n_i sampling occasions are dedicated to site $\mathbf{s}_i$. Many locations will not be sampled and have $n_i = 0$, and for computational simplicity we assume that all M selected sites have $n \geq 1$ sampling occasions so that $n_i \in \{0, n\}$. The number of occasions where the species is observed at location i is denoted $Y_i \in \{0, 1, ..., n_i\}$; in particular, $Y_i = 0$ if $n_i = 0$. Once the data $\mathbf{Y} = (Y_1, ..., Y_N)^T$ (we use the transpose symbol to indicate that $\mathbf{Y}$ is a column vector) from the designed survey are collected they will be analyzed using the spatial occupancy model (Johnson et al., 2013;

Pacifici et al., 2017)

$$Y_i|Z_i \overset{indep}{\sim} \text{Binomial}(n_i, \pi Z_i), \tag{1}$$

where Z_i is the binary indicator that the species occupies location i and $\pi \in [0, 1]$ is the detection probability, i.e., the probability of observing the species on a sampling occasion at an occupied location. The primary objective is to estimate the species distribution (i.e. latent occurrence state) $\mathbf{Z} = (Z_1, ..., Z_N)^T$.

We model occupancy using the latent random effect θ_i so that

$$Z_i|\theta_i \overset{indep}{\sim} \text{Bernoulli}[G(\theta_i)], \tag{2}$$

where G is an inverse link function (e.g., logistic, probit or more recently Warton et al. (2010) and Zipkin et al. (2017) use the cloglog link). Spatial dependence in the random effects is captured by modeling the random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)^T$ as

$$\boldsymbol{\theta} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{3}$$

where $\mathbf{X}$ is a known $N \times p$ covariate matrix, $\boldsymbol{\beta}$ is a vector (of length p) of unknown regression coefficients, and $\boldsymbol{\Sigma}$ is an $N \times N$ spatial covariance matrix. The initial estimate of occupancy based on auxiliary data is included in the covariate matrix $\mathbf{X}$ (Pacifici et al., 2017), and the influence of this estimate on $\boldsymbol{\theta}$ is controlled by the corresponding element of $\boldsymbol{\beta}$. The user is free to use any auxiliary data and any summary of the auxiliary data as the covariate (in Section 3.3 we use smoothed eBird counts). To complete the Bayesian model we specify priors $\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\gamma}, \Lambda)$ and $\pi \sim \text{beta}(a, b)$. For design purposes, we assume $\boldsymbol{\Sigma}$ is known because estimating the spatial correlation requires cumbersome matrix operations that are prohibitively slow for large problems.

2.3 Optimal spatial design

Our objective is to optimize the design $\mathcal{D} = \{n_1, ..., n_N\}$, i.e., the number of sampling occasions at each of the N locations under consideration. The optimization depends on the spatial configuration of the N locations, the covariate matrix $\mathbf{X}$, and the true value of the unknown parameters $\Theta_0 = \{\pi_0, \boldsymbol{\beta}_0\}$. For a given data set $\mathbf{Y}$, denote the posterior occupancy probability for location i as $\bar{Z}_i = \text{Prob}(Z_i = 1|\mathbf{Y}, \mathcal{D})$. Although other metrics may be considered, we quantify the accuracy of the species distribution map using the Brier score,

$$\mathcal{C}(\mathbf{Y}, \mathbf{Z}, \mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{Z}_i)^2 = \frac{1}{N}\sum_{i|Z_i=0}\bar{Z}_i^2 + \frac{1}{N}\sum_{i|Z_i=1}(1-\bar{Z}_i)^2, \tag{4}$$

where Z_i is the true occupancy status. Other common design criteria include the average posterior variance of the Z_i, but we select the Brier score because it balances false positive probabilities (i.e., $\bar{Z}_i$ for sites with $Z_i = 0$) and false negative probabilities (i.e., $1 - \bar{Z}_i$ for sites with $Z_i = 1$). A smaller average Brier score corresponds to a better design, and so we seek the design that minimizes the expected Brier score $\mathcal{V}(\mathcal{D}) = \text{E}[\mathcal{C}(\mathbf{Y}, \mathbf{Z}, \mathcal{D})|\Theta_0]$, where the expectation is with respect to $(\mathbf{Z}, \mathbf{Y})$ given Θ_0.

2.3.1 Approximating posterior occupancy probabilities

Because we want to explore a large number of designs which all require expensive computations, we use an approximation of the posterior for quicker evaluation. Our search algorithm for the optimal design $\mathcal{D}$ requires efficient posterior evaluation. We propose a three-step approximation:

1. Approximate detection probability π in a way that is independent of $\boldsymbol{\theta}$
2. Estimate the posterior of $\boldsymbol{\theta}$ given π using a Gaussian approximation
3. Approximate $\bar{Z}_i$ by numerically integrating over $\boldsymbol{\theta}$

Each of these three steps is described in a paragraph below.

In (1), if $Y_i > 0$ then we must have $Z_i = 1$ and thus the likelihood depends only on π. Therefore, conditioning only on the observations with $Y_i > 0$, denoted $\mathbf{Y}_+$, we have

$$[\pi|\mathbf{Y}_+, \mathcal{D}] \propto b(\pi; a, b) \prod_{i; Y_i > 0} \frac{d(Y_i; n_i, \pi)}{1 - d(0; n_i, \pi)} \tag{5}$$

where b is the beta density function and d is the binomial mass function. We use the maximizor of (5), $\hat{\pi}$, as a plug-in estimate of the detection probability; this estimator does not depend on $\boldsymbol{\theta}$.

The Supplemental Materials (SM.1) derives the approximation to the posterior of $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}|\mathbf{Y}, \pi, \mathcal{D} \sim \text{Normal}[\boldsymbol{\mu}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}}), S_\theta(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})]. \tag{6}$$

The Supplemental Materials (SM.2) also includes an evaluation of the accuracy of this approximation, and suggests that it works well if $|\theta_i|$ is not too large and so the occupancy probabilities $G(\theta_i)$ are not too close to zero or one.

The posterior of $\boldsymbol{\theta}$ is used to compute the posterior mean $\bar{Z}_i$. Given $\boldsymbol{\theta}$, $\mathbf{Y}$, and $\mathcal{D}$, $Z_i = 1$ and thus $\bar{Z}_i = 1$ if $Y_i > 0$ and $Z_i|\mathbf{Y}, \boldsymbol{\theta}, \mathcal{D} \overset{indep}{\sim}$ Bernoulli$[g(n_i, \theta_i)]$ if $Y_i = 0$, where

$$g(n, \theta) = \frac{(1-\pi)^n G(\theta)}{(1-\pi)^n G(\theta) + 1 - G(\theta)}.$$

Therefore, if $Y_i = 0$

$$\bar{Z}_i = \mathrm{E}_{\boldsymbol{\theta}|\mathbf{Y},\mathcal{D}}[g(n_i, \theta_i)]$$

which is approximated using numerical integration over the univariate posterior of θ_i implied by (6).

2.3.2 Computing the optimal design

The design criteria $\mathcal{V}(\mathcal{D}) = \mathrm{E}[\mathcal{C}(\mathbf{Y}, \mathbf{Z}, \mathcal{D})|\Theta_0]$ is approximated using Monte Carlo sampling over $(\mathbf{Z}, \mathbf{Y})$ given the model parameters $\Theta_0 = (\pi_0, \boldsymbol{\beta}_0)$. We sample R datasets by first sampling R occupancy maps given $\boldsymbol{\beta}_0$, denoted $\mathbf{Z}^{(1)}, ..., \mathbf{Z}^{(R)}$, and then R complete datasets $\tilde{Y}_i^{(r)} \sim \text{Binomial}(n, \pi_0 Z_i^{(r)})$ for $r = 1, ..., R$, which are fixed throughout the optimization. The R observed datasets $\mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(R)}$ depend on the design $\mathcal{D}$ with $Y_i^{(r)}$ set to zero if $n_i = 0$ and $Y_i^{(r)}$ set to $\tilde{Y}_i^{(r)}$ if $n_i = n$. The Monte Carlo approximation is then

$$\mathcal{V}(\mathcal{D}) \approx \frac{1}{R} \sum_{r=1}^{R} \mathcal{C}(\mathbf{Y}^{(r)}, \mathbf{Z}^{(r)}, \mathcal{D}). \tag{7}$$

We use $R = 1,000$ datasets for all analyses.

We use an exchange algorithm (Royle, 2002) to optimize (7). The exchange algorithm randomly selects M initial sites to have $n_i = n$ and the remaining sites have $n_i = 0$. The algorithm then cycles through neighbor pairs and exchanges their value of n_i if this reduces $\mathcal{V}(\mathcal{D})$. This is repeated until no local moves improve $\mathcal{V}(\mathcal{D})$. We repeat this procedure for 10 random starts and retain the solution with smallest $\mathcal{V}(\mathcal{D})$.

3 Results

3.1 Exploring the optimal design

In this section, we explore how the locally-optimal design varies with the quality of the auxiliary information, sampling density, and the detection probability. All scenarios assume the model in Section 2.2 with $N = 400$ potential locations on a 20×20 grid with grid spacing 1. The spatial covariance in (3) is fixed as $\text{Cov}(\theta_i, \theta_j) = \exp(-||\mathbf{s}_i - \mathbf{s}_j||/3)$. The number of sampling occasions at each of the $M = 36$ sampling locations is $n = 5$ and we use the probit link $G(\theta) = \Phi(\theta)$, where Φ is the standard normal distribution function. There are $p = 2$ covariates, $\mathbf{X}_i^T = [1, X_{i1}]$, where X_{i1} is a function of initial occupancy probability at site i. We consider two hypothetical initial surfaces estimated by the auxiliary information (Figure 2),

$$\begin{aligned} x_i &= 0.01 + 0.49[\cos(s_{i1}) + \cos(s_{i2})]_+ && \text{(“Hot Pockets”)} \\ x_i &= \exp\left[-10\left(\frac{r_i - 7}{5}\right)^2\right] && \text{(“Donut”)} \end{aligned}$$

where $[x]_+ = \max\{0, x\}$ and $r_i = ||\mathbf{s}_i - (10, 10)^T||$ is the distance from location $\mathbf{s}_i$ to the center of the grid. These were chosen to explore variability in the spatial surface. The initial estimate enters the statistical model as $X_{i1} = \Phi^{-1}(0.98x_i + 0.01)$ to match the scale of the random effects. The quality of this auxiliary information is either “poor” with $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01})^T = (0.0, 0.5)^T$ or “good” $\boldsymbol{\beta}_0 = (0.0, 2.0)^T$. The proportion of random effect variance explained by the auxiliary information, $\text{Var}(X_{1i}\beta_{01})/[1+\text{Var}(X_{1i}\beta_{01})]$, is 0.24 and 0.33 for the “poor” case for “Hot Pockets” and “Donut” covariate, respectively, compared to 0.83 and 0.89 for the “good” case for the “Hot Pockets” and “Donut” covariate, respectively. We also vary the true detection probability $\pi_0 \in \{0.3, 0.7\}$.

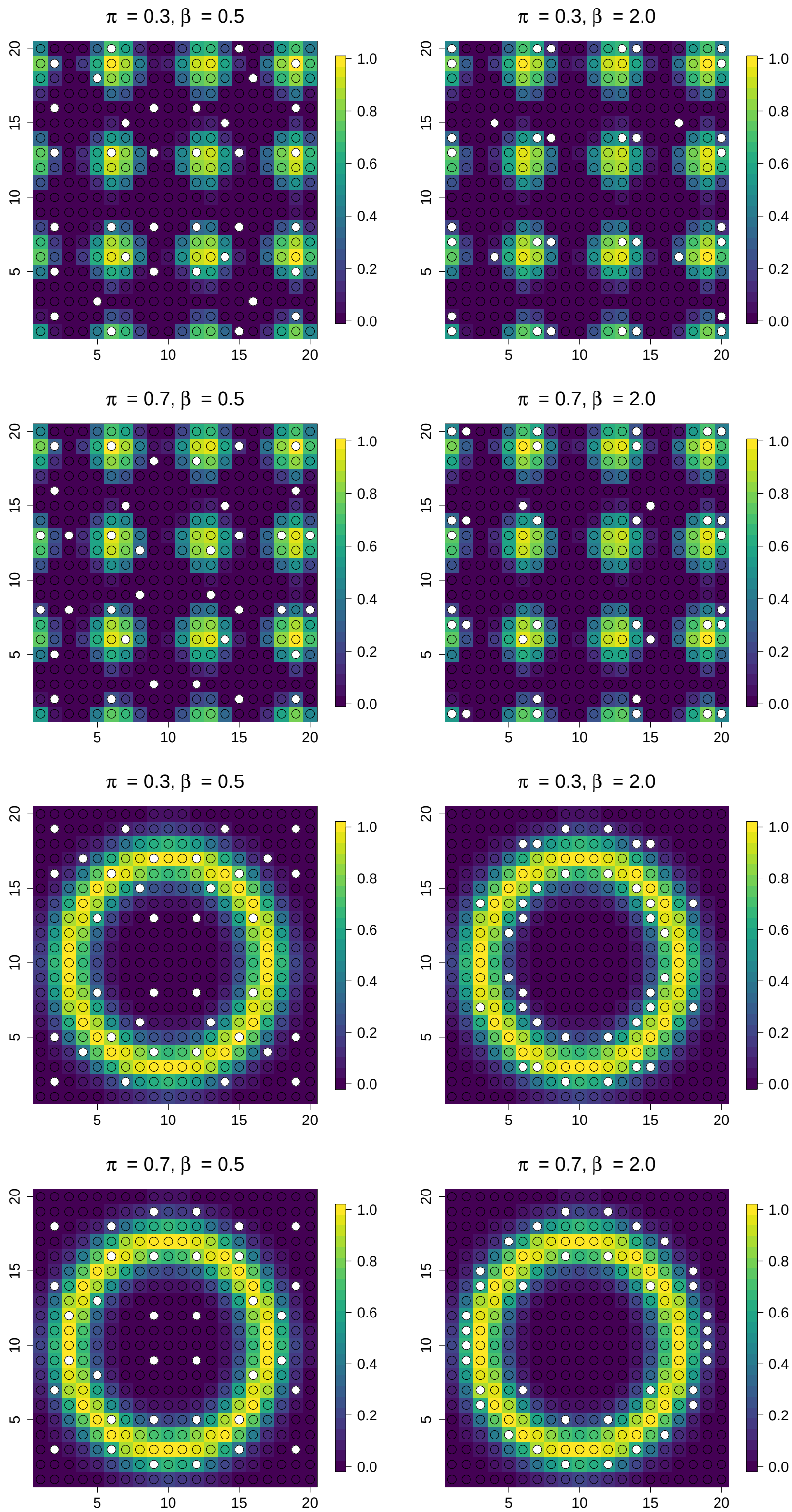
Table 1: **Summaries of the optimal designs for the simulation study.** The eight optimal designs vary by the strength of the spatial pattern of the auxiliary information (X), the detection probability (π), and the quality of the auxiliary information (β_1). The designs are summarized using the average distance between each sampling location and its nearest neighbor, the average and standard deviation of the initial estimates (X) at the sample locations, and the proportion of sampling locations with initial estimates less than 0.25 and greater than 0.75.

X	π	β_1	Mean Dist to Neighbor	Ave X	SD X	$X < 0.25$	$X > 0.75$
Hot Pocket	0.3	0.5	2.56	0.29	0.34	0.56	0.17
	0.3	2.0	1.24	0.40	0.25	0.31	0.08
	0.7	0.5	2.54	0.31	0.35	0.53	0.19
	0.7	2.0	1.09	0.42	0.28	0.33	0.14
Donut	0.3	0.5	2.47	0.35	0.38	0.56	0.25
	0.3	2.0	1.57	0.47	0.25	0.33	0.17
	0.7	0.5	2.63	0.42	0.36	0.44	0.25
	0.7	2.0	1.84	0.45	0.25	0.36	0.14

The Gaussian approximation for the optimal design is centered on $\tilde{\theta}_i = 1$ if $Y_i > 0$ and $\tilde{\theta}_i = -1/[(1-\hat{\pi})^{n_i} + 1]$ if $Y_i = 0$. To improve computational speed and stability, we assume that $\mathcal{D}$ is symmetric in the four quadrants and optimize over only $M/4$ locations in the first quadrant (which reduces computational time by roughly a factor of 4). This is reasonable in these cases because the covariates exhibit this symmetry. With this assumption, computing the optimal design for one random start takes approximately 41 minutes on a standard PC. The time increases to 92 minutes with $N = 400$ and $M = 60$, 97 minutes with $N = 900$ and $M = 36$, and 199 minutes with $N = 900$ and $M = 60$ (the optimal design for these cases are not shown).

The optimal designs in each scenario are plotted in Figure 2 and summarized numerically in Table 1. For these examples the assumed quality of the auxiliary information (β_1) influences the design more than the detection probability (π). In most cases, the designs avoid sampling in regions where the initial estimates are near one, especially when auxiliary data quality is assumed to be

Figure 2: **The optimal sampling locations for several combinations of model parameters.** The recommended sampling locations are the white circles and the auxiliary data (x_i) is the background color. The designs vary by the detection probability (π), quality of the auxiliary information (β), and covariate structure ("Hot Pockets" in the top two rows, "Donut" in the bottom two rows).



good ($\beta = 5$). These regions do not warrant sampling effort because they will almost certainty be estimated to be occupied by extrapolation based on the regression relationship between the auxiliary data and occupancy. Because the spatial occupancy model allows for false negatives but not false positives, more sampling locations are placed in areas with low values of the initial occupancy estimate where extensive sampling may reveal isolated occupied areas, especially when the detection probability is low. Also, the average distance between points increases (Table 1), and thus the sampling locations fill the space more uniformly, as reliability of the auxiliary data (quantified by β_1) decreases because in the absence of other information the optimal design reverts to the space-filling design with sampling location spread uniformly to cover the spatial domain.

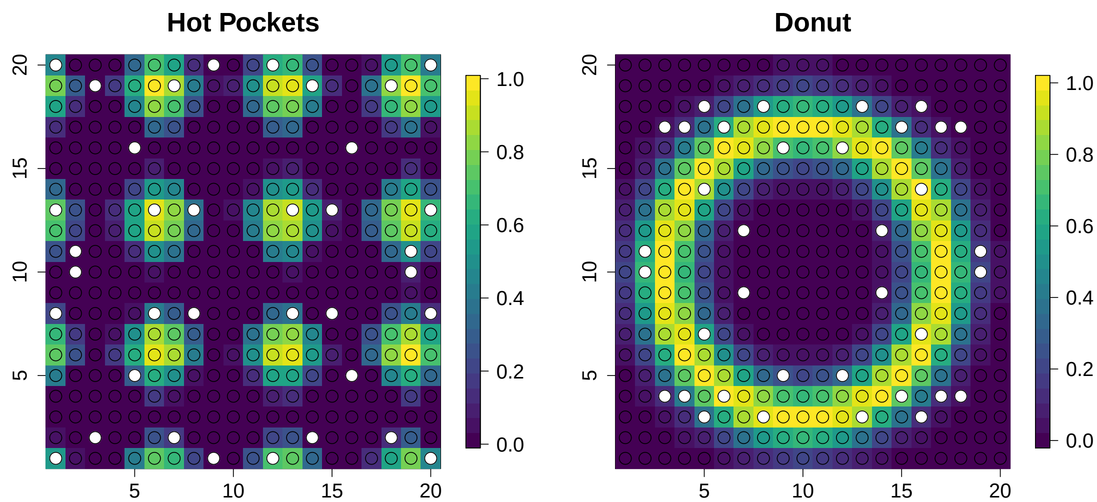
3.2 Simulation study

We conduct a simulation study to compare the performance of the optimal design with other ad hoc designs. Data are generated as Section 3.1. We compare five designs with $M = 36$ sampling locations:

1. **Grid**: a complete regular grid of 6×6 sites
2. **Low X**: a regular grid 5×5 grid and the sites with lowest X_{i1}
3. **Medium X**: a regular grid 5×5 grid and the sites with X_{i1} closest to the median x_{i1}
4. **High X**: a regular grid 5×5 grid and the sites with highest X_{i1}
5. **Optimal-Prior**: The design selected by the proposed method

For the final design, rather than picking a locally-optimal design for specific values of the unknown parameters, we specify design priors $\pi_0 \sim \text{Beta}(2, 2)$ and $\boldsymbol{\beta}_0 \sim \text{Normal}[(0, 1)^T, 0.5 I_2]$ and average

Figure 3: **The optimal sampling design with priors on model parameters**. The recommended sampling location are white circles and the auxiliary covariate structure (x_i) is the background color.



$\mathcal{V}(\mathcal{D})$ over prior uncertainty by sampling π_0 and $\boldsymbol{\beta}_0$ from the prior for each of the R simulated datasets in (7). The design priors are chosen to cover the range of values for which the optimal design is displayed in Figure 2. This Bayesian-optimal design (Chaloner and Verdinelli, 1995) is plotted for both covariate structures in Figure 3.

For each combination of surface ("Hot Pockets" or "Donut"), detection probability ($\pi = 0.3$ or $\pi = 0.7$), and quality of auxiliary data ($\beta_1 = 0.5$ or $\beta_1 = 2.0$), we generate $S = 500$ complete datasets $\tilde{Y}_i$ (independent of those used to compute the optimal design). Each complete dataset is analyzed using data at the design points for each of the five designs using the Bayesian spatial occupancy model given in Section 2.2 and the MCMC algorithm and uninformative priors in Supplemental Materials (SM.2). To determine the benefit of including auxiliary information, we also include the regular grid design except fit to the data excluding the covariate X_{i1}. For each dataset and each method we record the posterior mean of Z_i and classify a site as occupied when the pos-

Table 2: **The average Brier score ($\times$ 100) of occupancy status for the simulation study (smaller is better).** The data are generated with different spatial initial occupancy estimates (X, see Figure 2), detection probability (π), and quality of auxiliary information ($\boldsymbol{\beta}$). The six designs are a regular 6×6 grid (fit with and without using X as a covariate), a smaller 5×5 grid supplemented with sites with low, medium, or high X, and the estimated optimal spatial design. The final column gives the maximum standard error for the values in the row, and the method with the best performance in each case is in bold.

			Complete grid		Small grid +				
X	π	β_1	No X	X	Low X	Medium X	High X	Optimal	Max SE
Hot Pocket	0.3	0.5	22.1	21.1	**21.0**	**21.0**	21.5	21.3	0.2
		2	16.6	7.6	8.2	8.2	8.4	**7.3**	0.2
	0.7	0.5	20.3	18.8	19.4	19.4	18.9	**18.7**	0.1
		2	14.9	7.1	7.9	7.9	7.3	**6.6**	0.1
Donut	0.3	0.5	23.0	20.9	21.1	20.8	21.1	**20.7**	0.1
		2	20.8	7.5	8.4	8.4	7.6	**7.1**	0.1
	0.7	0.5	21.9	19.0	19.9	19.4	19.0	**18.7**	0.1
		2	20.6	6.9	7.8	7.7	7.5	**6.4**	0.1

terior mean exceeds 0.5. Tables 2 and 3 reports the average (over the N locations and S simulated datasets) Brier score and classification accuracy, i.e., the proportion (over the N locations and S simulated datasets) of agreement between the true and estimated occupancy status.

In all cases with reliable auxiliary information (large β_1), including this information in the statistical model provides a substantial improvement regardless of the spatial design. Overall, the complete grid that includes the covariate outperforms any of the ad-hoc designs formed by a small grid supplemented with locations determined by the auxiliary information. In all cases except the first with Hot Pocket covariate, low detection, and unreliable auxiliary information the optimal design gives the best performance. The largest reduction in Brier score compared to the regular grid that includes the auxiliary covariate is for the final case with high detection and large regression coefficient. The reduction in Brier score is 7% for the Hot Pocket (6.6 compared to 7.1) and Donut (6.4 compared to 6.9) covariates. The Brier score is mean squared error applied to binary data, and

Table 3: **Classification accuracy (%) for the simulation study (larger is better).** The data are generated with different spatial initial occupancy estimates (X, see Figure 2), detection probability (π), and quality of auxiliary information ($\boldsymbol{\beta}$). The six designs are a regular 6×6 grid (fit with and without using X as a covariate), a smaller 5×5 grid supplemented with sites with low, medium, or high X, and the estimated optimal spatial design. The final column gives the maximum standard error for the values in the row, and the method with the best performance in each case is in bold.

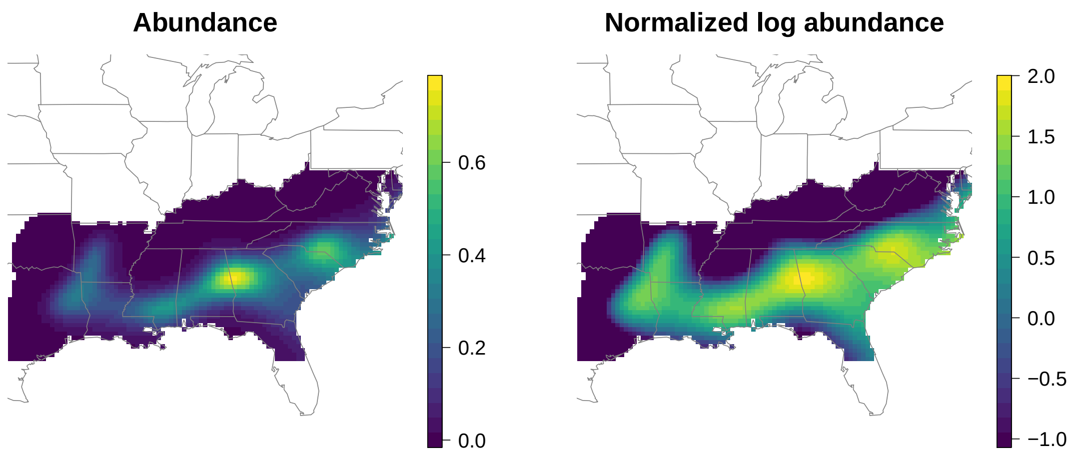
			Complete grid		Small grid +				
X	π	β_1	No X	X	Low X	Medium X	High X	Optimal	Max SE
Hot Pocket	0.3	0.5	63.9	67.0	**68.2**	**68.2**	67.0	67.2	0.5
		2	78.8	89.0	88.0	88.0	88.0	**89.4**	0.5
	0.7	0.5	68.2	71.5	70.9	70.9	71.6	**72.0**	0.3
		2	81.3	89.6	88.3	88.3	89.4	**90.3**	0.1
Donut	0.3	0.5	62.5	68.0	67.4	68.4	68.1	**68.5**	0.4
		2	71.0	89.2	87.8	87.9	89.2	**90.1**	0.3
	0.7	0.5	65.0	71.3	70.1	71.0	71.6	**71.9**	0.3
		2	72.0	90.0	88.8	88.8	89.5	**91.1**	0.1

so these reductions can be interpreted as reductions in mean squared error.

3.3 Application to the brown-headed nut hatch

We use eBird data (Sullivan et al., 2009) to construct an initial estimate of the species distribution map of the brown-headed nuthatch (BHNU). The Southeast US is partitioned into 0.25 degree $\times$ 0.25 degree grid cells (Figure 4). Denote O_i as the number of sightings of the brown-headed nuthatch in cell i and E_i as the corresponding sampling effort, defined as the self-reported number of sampling hours; both O_i and E_i are aggregated over all surveys in 2012. We fit a generalized additive model (GAM) $O_i \sim \text{Poisson}(E_i\lambda_i)$ where λ_i is the relative abundance. Abundance is estimated using the `gam` function in the `R` package `mgcv`. The GAM model assumes that $\log(\lambda_i)$ is a smooth function of the cell's latitude and longitude with smoothness determined by generalized cross-validation. Figure 4 (left) plots the GAM estimates of λ_i. We use as the constructed covariate

Figure 4: **Initial estimates of abundance of the brown-headed nuthatch derived from the 2012 eBird data**. The left panel plots the generalized additive model's initial estimates of relative abundance, and the right panel plots the normalized log abundance which is used as the covariate in the spatial design.



the normalized (to have mean zero and variance one) log estimated relative abundance plotted in Figure 4. To avoid numerical problems, estimated relative abundances less than 0.01 were set to 0.01. We intend to make both the raw data and log relative abundance data available on the first author's personal webpage.

Figure 5 plots the optimal design with $M = 50$ sampling locations for various values of the parameters, and Table 4 provides numerical summaries of the design points and eBird relative abundance estimates at the design points. In this optimal design calculation, we fix the intercept at $\beta_0 = 0$ and the spatial variance at $\sigma^2 = 1$ and assume exponential correlation $\text{Cor}(\theta_i, \theta_j) = \exp(-||\mathbf{s}_i - \mathbf{s}_j||/\rho)$. The simulations vary the detection probability $\pi \in \{0.3, 0.7\}$, spatial range $\rho \in \{0.5, 2.0\}$ (corresponding to correlation 0.14 and 0.61 for adjacent sites, respectively), and strength of the auxiliary information $\beta_1 \in \{1, 5\}$, so that $X_{i1}\beta_1$ explains either 50% or 96% of $\boldsymbol{\theta}$'s prior variance. Supplemental Materials (SM.5) tests for convergence of the exchange algorithm for these data. Although the estimated optimal design varies a bit depending on the starting values, the Brier scores are fairly consistent across starting values, suggesting that results may be robust to small deviations from the optimal design.

The most glaring difference between the optimal designs in Figure 5 is that the sampling locations cluster on the periphery of the eBird distribution map when the auxiliary information is assumed to be reliable ($\beta_1 = 5$), and the sampling locations are more evenly distributed when the auxiliary information is less reliable ($\beta_1 = 1$). The former design feature is intuitive because if we trust the eBird relative abundance estimates then there is no new information to be gained by sampling in high abundance areas that are almost certainly occupied. It is more efficient to focus sampling effort on the edge of the distribution map to resolve lingering uncertainties. On the other hand, when the auxiliary information is assumed to be less reliable, the optimal sampling locations are spread out over space and include a wider range of initial estimates to hedge against faulty initial estimates.

The influence of detection and spatial correlation are more apparent in the numerical summaries in Table 4 than the maps in Figure 5, especially when $\beta_1 = 0.5$. In this case, the mean distance to the nearest neighbor increases with spatial correlation, i.e., when there is strong spatial dependence in the occupancy status the optimal design resembles a space-filling design. The mean distance to the nearest neighbor also increases with detection probability. With low detection probability there is value in sampling two points close to each other because a single sample is less likely to provide definitive information on the occupancy status of the region. If we did not restrict all sample locations to have n sampling occasions, then two nearby sites might be replaced by a single site

Figure 5: **Optimal spatial design for the brown-headed nuthatch for different model parameters.** The recommended sampling locations are white dots and the background color is the eBird relative abundance estimate. The designs vary by the strength of the auxiliary estimates (β), the spatial range in degrees (ρ), and the detection probability (π). The estimated design values $\mathcal{V}(\mathcal{D})$ (times 100) is given at the bottom of each plot ("V="); the standard error of these values is less than 0.01 in all cases.

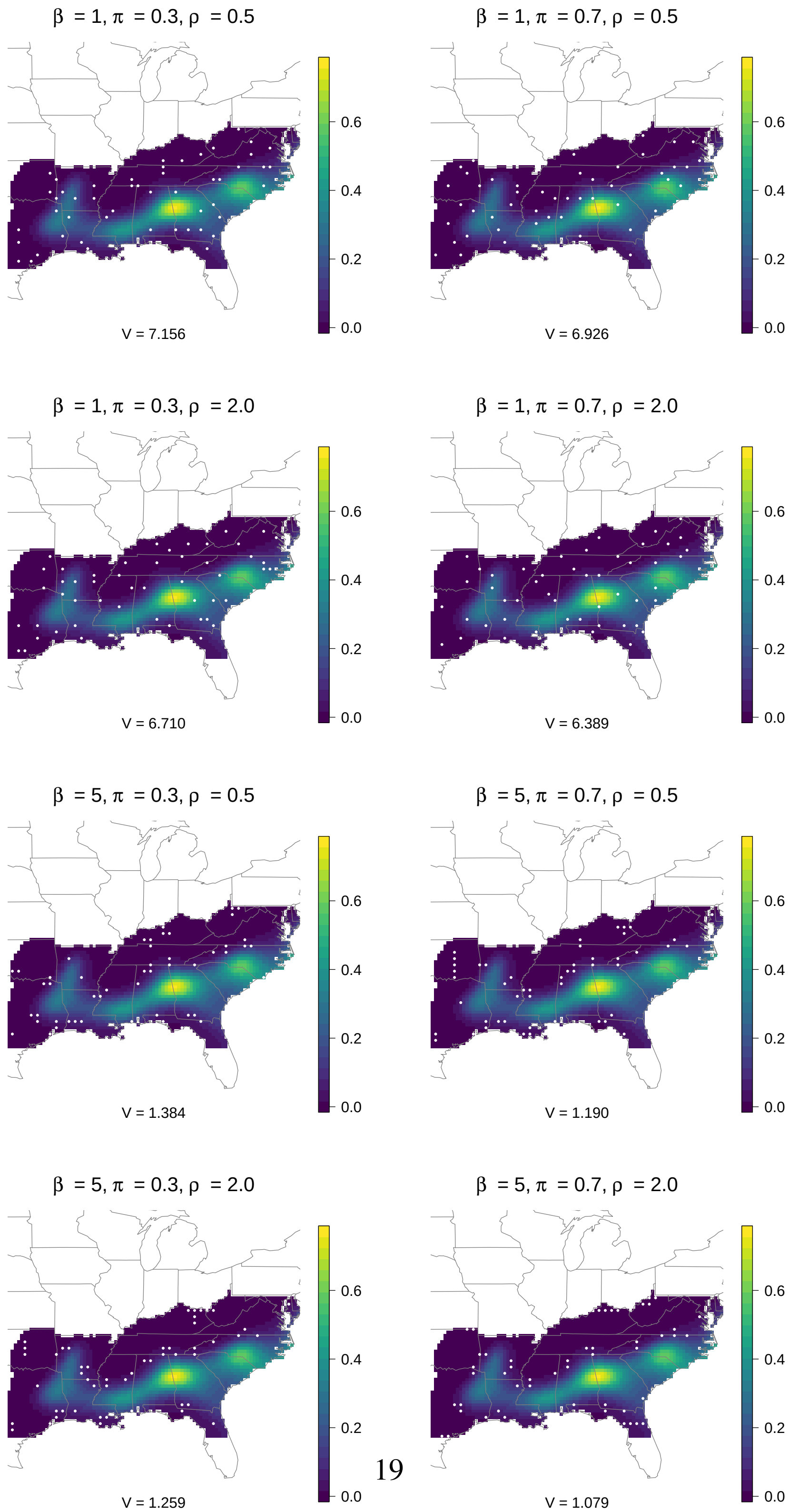


Table 4: **Summaries of the optimal designs for the brown-headed nuthatch.** The eight optimal designs vary by the strength of the auxiliary estimate (β_1), the spatial range in degrees (ρ), and the detection probability (π). The designs are summarized using the average distance (km) between each sampling location and its nearest neighbor, the average and standard deviation of the initial eBird estimates (X) at the sample locations, and the proportion of sampling location with initial estimates less than 0.05 and greater than 0.20.

β_1	ρ	π	Mean Dist to Neighbor	Ave X	SD X	$X < 0.05$	$X > 0.20$
1	0.5	0.3	59.7	0.082	0.102	0.52	0.18
		0.7	74.6	0.094	0.134	0.54	0.16
	2.0	0.3	67.0	0.086	0.110	0.60	0.20
		0.7	75.7	0.094	0.124	0.54	0.14
5	0.5	0.3	46.6	0.045	0.064	0.62	0.02
		0.7	47.4	0.034	0.035	0.72	0.00
	2.0	0.3	47.6	0.038	0.036	0.60	0.00
		0.7	41.3	0.037	0.038	0.70	0.00

with many sampling occasions.

4 Discussion

In this paper, we have proposed a new approach to designing a survey to estimate a species distribution map that incorporates auxiliary data both at the design stage and the analysis stage. The statistical model fit to the data accounts for spatial dependence, imperfect detection, and potential bias in the initial estimate. The design minimizes the expected misclassification rate, which is very difficult to compute for this comprehensive model, and we propose approximations to make this evaluation computationally feasible. Our Bayesian-optimal design accounts for prior uncertainty in the quality of the auxiliary data and other parameters such as the detection rate. We show with simulation studies that the two-stage design leads to lower misclassification rates than several ad-hoc designs.

In addition to studying the performance of the model and proposing an optimal design for the brown-headed nuthatch, some general design principles emerged. In summary, the least informative sampling locations appear to be those with high *a priori* occupancy probability. Because we are assuming no false positive observations, only a few samples in high *a priori* occupancy regions are sufficient to confirm that these regions are indeed occupied. The optimal design often places more sample locations on the periphery and exterior of the species' domain to refine the map in these areas of uncertainty. Another general trend is that a purely space-filling design is reasonable when detection is high and/or spatial correlation is strong, and simultaneously sampling nearby sites (or presumably adding more replications at one site) is justifiable only when detection is low and it is possible that a region is occupied even if a few observations fail to detect the species. Although the maps in Figure 5 are specific to this application, we feel these design principles can be generalized to other settings.

As in most experimental design problems, the user must specify a prior distribution for the quality of the auxiliary data (as measured by β_1). In the simulation study and BHNU example we use the proportion of variance in the spatial random effects that is explained by the auxiliary data as a guide to selecting the prior. For the BHNU example we varied the proportion from 0.50 to 0.96. Values much smaller than 0.5 suggest that the auxiliary data is not useful, and values much higher than 0.96 suggest that the formal survey may not be needed. But of course, this step requires either user input about the true parameters or a pilot dataset to provide initial estimates.

A potential concern with the proposed design is bias in the resulting predictions and parameter estimates caused by preferential sampling (Diggle et al., 2010; Pati et al., 2011; Reich and Fuentes, 2013; Conn et al., 2017), i.e., selecting the sample locations based on prior knowledge about the true process. The proposed design is determined by the auxiliary data and the design priors for the parameters. If this information is not used in the analysis then likely this would lead to bias. Using the auxiliary information in the analysis should minimize the effects of preferential sampling. In our analysis, we used uninformative priors instead of the design priors used to select the sampling locations, so one might question whether this could cause bias. However, the simulation results in Supplemental Methods Section SM.4 show that the optimal design does not lead to bias for the scenarios considered here.

While the proposed approximation to the expected misclassification rate is much faster than an MCMC approximation, the computing required to find the optimal design remains cumbersome. Currently, when one sampling location is moved to a nearby location the entire posterior is recomputed to determine whether this move improves the design. An area of future work is to use only sites in a window around the location in question to approximate the effect of a local move. Another intriguing possibility proposed by Overstall and Woods (2017) is to build a statistical emulator for the expected misclassification rate of a candidate design and then select the design that optimizes the emulated design criteria. Emulating the expected misclassification rate for the complex system considered here would be challenging, but given a reasonable emulator the optimization step would be straightforward.

Finally, we have focused entirely on constructing species distribution maps. Another important problem is to design a survey with high power to detect the effects of environmental covariates on occupancy. In this case, rather than approximate the posterior distribution of the latent occupancy indicators we could approximate the posterior of the regression coefficients, which should follow similar steps. The optimality criteria would also need to be changed from expected misclassification rate to a function of the posterior covariance matrix of the regression coefficients. Changing the optimality criterion does not fundamentally change our algorithm because it is approximated using Monte Carlo simulation. It would also be interesting to consider optimizing the design for multiple objectives, e.g., both prediction and covariate estimation.

References

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum Experimental Design with SAS*. New York: Oxford University Press Inc.

Bailey, L. L., Hines, J. E., Nichols, J. D. and MacKenzie, D. I. (2007) Sampling design trade-offs in occupancy studies with imperfect detection: Examples and software. *Ecological Applications*, **17**, 281–290.

Bailey, L. L., MacKenzie, D. I. and Nichols, J. D. (2014) Advances and applications of occupancy models. *Methods in Ecology and Evolution*, **5**, 1269–1279.

Barbet-Massin, M., Jiguet, F., Albert, C. H. and Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Benedetti, R. and Palma, D. (1995) Optimal sampling designs for dependent spatial units. *Environmentrics*, **6**, 101–114.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: A review. *Statistical Science*, **10**, 277–304.

Conn, P. B., Johnson, D. S., Hoef, J. M. V., Hooten, M. B., London, J. M. and Boveng, P. L. (2015) Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts. *Ecological Monographs*, **85**, 235–252.

Conn, P. B., Thorson, J. T. and Johnson, D. S. (2017) Confronting preferential sampling when analyzing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*.

Diggle, P. J., Menezes, R. and Su, T.-l. (2010) Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 191–232.

Dorazio, R. M. (2007) On the choice of statistical models for estimating occurrence and extinction from animal surveys. *Ecology*, **88**, 2773–2782.

— (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**, 1472–1484.

Fortin, M.-J., Drapeau, P. and Legendre, P. (1990) Spatial autocorrelation and sampling design in plant ecology. In *Progress in Theoretical Vegetation Science*, 209–222. Springer.

Guillera-Arroita, G. and Lahoz-Monfort, J. J. (2012) Designing studies to detect differences in species occupancy: power analysis under imperfect detection. *Methods in Ecology and Evolution*, **3**, 860–869.

Guillera-Arroita, G., Ridout, M. and Morgan, B. (2014) Two-stage bayesian study design for species occupancy estimation. *Journal of Agricultural, Biological & Environmental Statistics*, **19**.

Guillera-Arroita, G., Ridout, M. S. and Morgan, B. J. (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, **1**, 131–139.

Hanks, E. M., Hooten, M. B., Knick, S. T., Oyler-McCance, S. J., Fike, J. A., Cross, T. B., Schwartz, M. K. et al. (2016) Latent spatial models and sampling design for landscape genetics. *The Annals of Applied Statistics*, **10**, 1041–1062.

Hinkelmann, K. (2012) *Design and Analysis of Experiments: Volume 3 Special Designs and Applications*. New Jersey: Wiley.

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

Jones, B. and Goos, P. (2011) *Optimal Design of Experiments: A Case Study Approach*. New Jersey: Wiley.

Khuri, A. I., Mukherjee, B., Sinha, B. K. and Ghosh, M. (2006) Design issues for generalized linear models: A review. *Statistical Science*, **21**, 376–399.

MacKenzie, D. I. and Royle, J. A. (2005) Designing occupancy studies: General advice and allocating survey effort. *Journal of applied Ecology*, **42**, 1105–1114.

Mateu, J. and Müller, W. G. (2012) *Spatio-temporal Design: Advances in Efficient Data Acquisition*. John Wiley & Sons.

Müller, W. G. (2007) *Collecting Spatial Data (3rd Edition*. New York: Springer.

Overstall, A. M. and Woods, D. C. (2017) Bayesian design of experiments using approximate coordinate exchange. *Technometrics*, 1–13.

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

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. and Collazo, J. A. (2017) Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, **98**, 840–850.

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

Pukelsheim, F. (2006) *Optimal Design of Experiments*. New York: Wiley.

Reich, B. J. and Fuentes, M. (2013) Accounting for design in the analysis of spatial data. *Spatio-Temporal Design: Advances in Efficient Data Acquisition*, 131–141.

Royle, J. (2002) Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, **100**, 121–134.

Royle, J. A. (2000) Optimal spatial network design for temporal trend estimation.

Royle, J. A., Kéry, M., Gautier, R. and Schmid, H. (2007) Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecological Monographs*, **77**, 465–481.

Royle, J. A. and Nychka, D. (1998) An algorithm for the construction of spatial coverage designs with implementation in splus. *Computers & Geosciences*, **24**, 479–488.

Royle, J. A. and Wikle, C. K. (2005) Efficient statistical mapping of avian count data. *Environmental and Ecological Statistics*, **12**, 225–243.

Ryan, E. G., Drovandi, C. C., McGree, J. M. and Pettitt, A. (2016) A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, **86**, 128–154.

Sanderlin, J. S., Block, W. M. and Ganey, J. L. (2014) Optimizing study design for multi-species avian monitoring programmes. *Journal of Applied Ecology*, **51**, 860–870.

Sauer, J. R., Hines, J. E. and J, F. (2005) *The North American Breeding Bird Survey, results 557 and analysis 19662005*. Version 6.2.2006. USGS Patuxent Wildlife Research Center, Laurel, Maryland, USA.

Shah, K. R. and Sinha, B. K. (1989) *Theory of Optimal Designs: Lecture Notes in Statistics*. New York: Springer.

Sliwinski, M., Powell, L., Koper, N., Giovanni, M. and Schacht, W. (2016) Research design considerations to ensure detection of all species in an avian community. *Methods in Ecology and Evolution*, **7**, 456–462.

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

Ver Hoef, J. M. (2012) Practical considerations for experimental designs of spatially autocorrelated data using computer intensive methods. *Statistical Methodology*, **9**, 172–184.

Warton, D. I., Shepherd, L. C. et al. (2010) Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *The Annals of Applied Statistics*, **4**, 1383–1402.

Wikle, C. K. and Royle, J. A. (1999) Space-time dynamic design of environmental monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, 489–507.

— (2005) Dynamic design of ecological monitoring networks for non-gaussian spatio-temporal data. *Environmetrics*, **16**, 507–522.

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G. and Bower, M. R. (2017) Monitoring dynamic spatio-temporal ecological processes optimally. *arXiv preprint arXiv:1707.03047*.

Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J. and Grant, E. H. C. (2017) Integrating count and detection–nondetection data to model population dynamics. *Ecology*, **98**, 1640–1650.

Supplemental materials for "Integrating auxiliary data in optimal spatial design for species distribution modeling"

SM.1 – Second-order Approximation

Marginally over the occupancy indicators Z_i, the likelihood is

$$[Y_i|\theta_i, \pi, \mathcal{D}] = G(\theta_i)\binom{n_i}{Y_i}\pi^{Y_i}(1-\pi)^{n_i - Y_i} + [1 - G(\theta_i)]I(Y_i = 0), \quad (1)$$

independent over i. We use the second-order approximation around $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$,

$$-2\sum_{i=1}^{N} \log[Y_i|\theta_i, \hat{\pi}, \mathcal{D}] \approx c - 2\mathbf{M}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})]^T\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{V}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})\boldsymbol{\theta} \quad (2)$$

where c is a constant that does not depend on $\boldsymbol{\theta}$, and the vector $\mathbf{M}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ and the diagonal matrix $\mathbf{V}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are given below. Combined with priors $\boldsymbol{\theta}|\boldsymbol{\beta} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\gamma}, \Omega)$, the approximate posterior (marginal over $\boldsymbol{\beta}$) is

$$\boldsymbol{\theta}|\mathbf{Y}, \pi, \mathcal{D} \sim \text{Normal}[\boldsymbol{\mu}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}}), S_\theta(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})] \quad (3)$$

where

$$\begin{aligned} S_\theta(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})^{-1} &= \mathbf{V}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}}) + \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \Lambda^{-1})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1} \quad (4) \\ \mu_\theta(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}}) &= S_\theta(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})\left[\mathbf{M}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}}) + \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \Lambda^{-1})^{-1}\Lambda^{-1}\boldsymbol{\gamma}\right]. \end{aligned}$$

Finally, we construct the elements of vector $\mathbf{M}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ and $\mathbf{V}(\mathbf{Y}, \hat{\pi}, \mathcal{D}, \tilde{\boldsymbol{\theta}})$. The negative log-likelihood corresponding to (1) for one term (hence dropping the subscript i) is (dropping constants that do not depend on θ)

$$-\log[p(Y|\theta, \pi, \mathcal{D})] = \begin{cases} 0 & \text{if } n = 0 \\ -\log[G(\theta)] & \text{if } n > 0, Y > 0 \\ -\log[G(\theta)q + 1], & \text{if } n > 0, Y = 0 \end{cases} \tag{5}$$

for $q = (1 - \pi)^n - 1$. Therefore, for terms with $n = 0$, corresponding elements of $\mathbf{M}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ and $\mathbf{V}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are zero. For terms with $n > 0$ and $Y > 0$, the corresponding elements of $\mathbf{M}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are

$$\frac{G'(\tilde{\theta})^2 - G''(\tilde{\theta})G(\tilde{\theta})}{G(\tilde{\theta})^2}\tilde{\theta} + \frac{G'(\tilde{\theta})}{G(\tilde{\theta})}$$

and the corresponding diagonal elements of $\mathbf{V}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are

$$\frac{G'(\tilde{\theta})^2 - G''(\tilde{\theta})G(\tilde{\theta})}{G(\tilde{\theta})^2},$$

where G' and G'' are the first and second derivatives of G, respectively. For terms with $n > 0$ and $Y = 0$, the corresponding elements of $\mathbf{M}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are

$$\frac{q^2 G'(\tilde{\theta})^2 - qG''(\tilde{\theta})[G(\tilde{\theta})q + 1]}{[G(\tilde{\theta})q + 1]^2}\tilde{\theta} + \frac{qG'(\tilde{\theta})}{G(\tilde{\theta})q + 1}$$

and the corresponding diagonal elements of $\mathbf{V}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are

$$\frac{q^2 G'(\tilde{\theta})^2 - qG''(\tilde{\theta})[G(\tilde{\theta})q + 1]}{[G(\tilde{\theta})q + 1]^2}.$$

For the probit link, $G(\theta) = \Phi(\theta)$, $G'(\theta) = \phi(\theta)$ and $G''(\theta) = -\theta\phi(\theta)$ where Φ and ϕ are the standard normal distribution and density functions, respectively.

SM.2 – Evaluation of the posterior approximation

To evaluate the quality of Section 2.3.1's approximation to the posterior of $\boldsymbol{\theta}$, we compare the posterior mean and variance of $\boldsymbol{\theta}$ obtained from the usual MCMC approximation versus the second-order approximation in three cases. The three datasets are generated as in Section 3.1 with "Donut" covariate structure and parameters fixed at $\boldsymbol{\beta} = (0, 0.5)^T$, $n = 5$, and $\text{Cov}(\theta_i, \theta_j) = \sigma^2 \exp(-d_{ij}/3)$. The data are collected at $M = 100$ randomly selected locations and we compare the fidelity of the approximation for spatial standard deviations $\sigma \in \{1, 3\}$ and detection probabilities $\pi = \{0.3, 0.7\}$. Both the MCMC and Gaussian approximations assume the parameters to be fixed and known at their true values except for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which have the same priors as Section 3.1.

Figure S1 plots the posterior mean and variance of each θ_i for the three simulated datasets. The approximations are similar in the first and third cases with small spatial standard deviation except for sites with small θ_i and thus occupancy probability near zero. In the second case with large σ, the second-order approximation to the posterior mean is shrunk to zero and the posterior variance is underestimated. For data generated with large σ the occupancy probabilities are often close to zero or one and the normal approximation is inaccurate. However, even in the cases where the approximation is poor on the absolute scale, the ordering of sites in terms of occupancy probability remains reasonably accurate. In terms of computation, the MCMC approximation takes around 9 minutes while the second-order approximation takes less than 0.1 seconds.

SM.3 – Priors and MCMC details

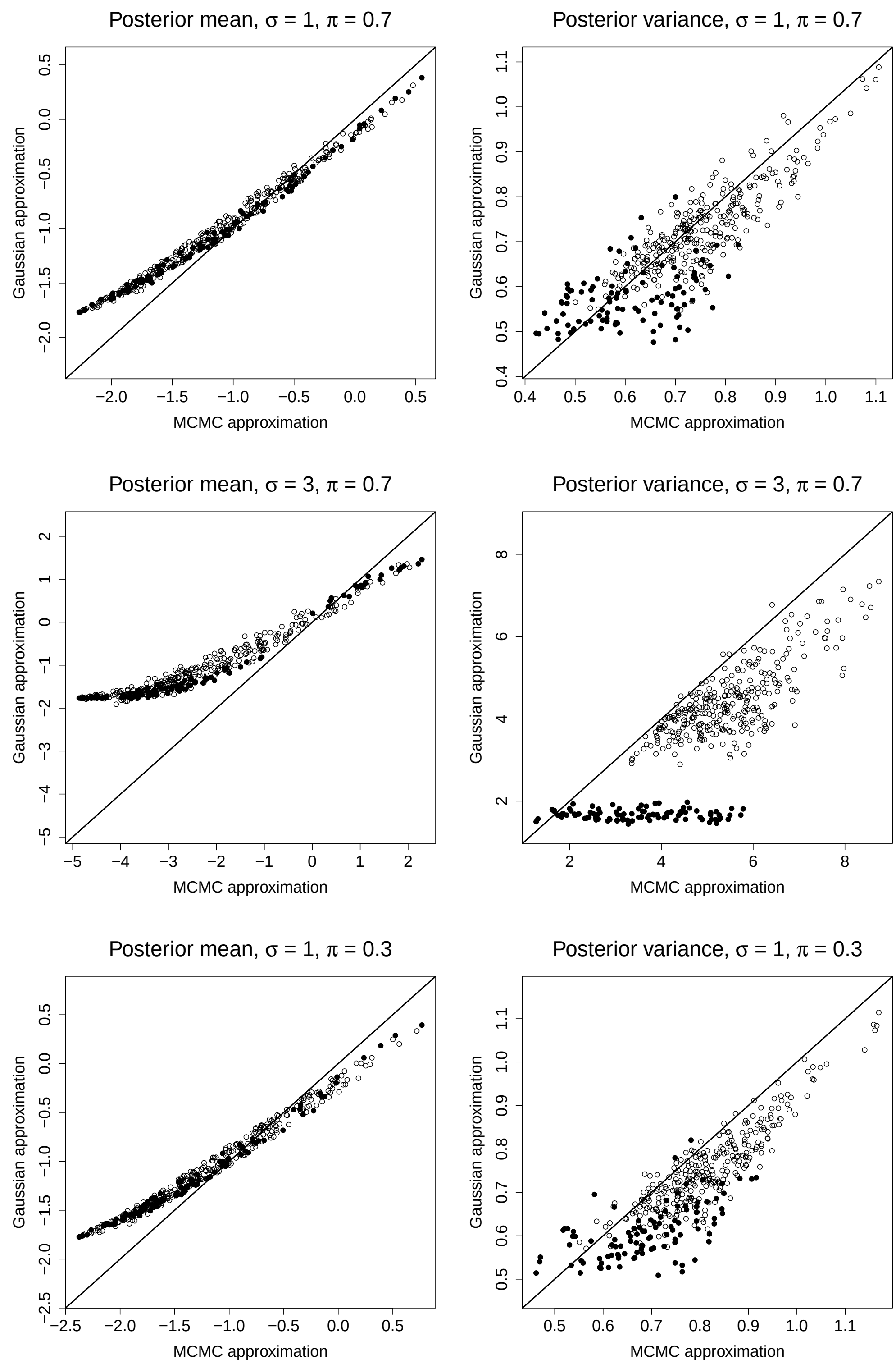
The spatial occupancy model described in Section 2.2 with probit link $G(\theta) = \Phi(\theta)$ and exponential spatial correlation can be written

$$
\begin{aligned}
Y_i|Z_i &\overset{indep}{\sim} \text{Binomial}(n_i, \pi Z_i) \\
Z_i &= I(\tilde{Z}_i > 0) \\
\tilde{Z}_i &\overset{indep}{\sim} \text{Normal}(\theta_i, 1) \\
\boldsymbol{\theta} &\sim \text{Normal}[\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\rho)]
\end{aligned}
$$

where σ^2 is the spatial variance and the (i, j) element of spatial correlation matrix $\mathbf{C}(\rho)$ is $\text{Cor}(\theta_i, \theta_j) = \exp(-d_{ij}/\rho)$. The priors are $\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\gamma}, \Lambda)$, $\pi \sim \text{Beta}(a, b)$, $\sigma^2 \sim \text{InvGamma}(c, d)$, and $\log(\rho) \sim \text{Normal}(m, s^2)$. We use uninformative priors by selecting hyperparameters $a = b = 1$, $c = d = 0.1$, $m = 0$, $s = 2$, $\boldsymbol{\gamma} = 0$ and $\Lambda = 10^2 I_p$.

MCMC proceeds by setting initial values for all parameters and updating them in sequence from their full conditional posterior distributions. The occupancy indicators at location i $(Z_i, \tilde{Z}_i)$

Figure S1: **Approximate posterior mean and variance of θ_i for three datasets using MCMC versus second-order approximations.** Each point is the estimated posterior mean (left) or variance (right) of θ_i from the two approximation, and is shaded for sites with $n_i > 0$ and empty for sites with $n_i = 0$. The datasets were generated with different values of the spatial standard deviation (σ) and detection probability (π).



are drawn simultaneously from their full conditional distribution $Z_i, \tilde{Z}_i|$rest as

$$Z_i|\text{rest} \sim \text{Bernoulli}\left[g(n_i, \theta_i)\right] \quad \text{and} \quad \tilde{Z}_i|Z_i, \text{rest} \sim \begin{cases} \text{TN}_{(-\infty,0)}(\theta_i, 1) & Z_i = 0 \\ \\ \text{TN}_{(0,\infty)}(\theta_i, 1) & Z_i = 1 \end{cases}$$

where $g(n, \theta) = \frac{(1-\pi)^n \Phi(\theta)}{(1-\pi)^n \Phi(\theta) + 1 - \Phi(\theta)}$ and TN is the truncated normal distribution. The full conditional distributions for all $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, σ^2, and π are

$$\begin{aligned}
\boldsymbol{\theta}|\text{rest} &\sim \text{Normal}\left[(\Omega + I_N)^{-1}(\Omega \mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}), (\Omega + I_N)^{-1}\right] \\
\boldsymbol{\beta}|\text{rest} &\sim \text{Normal}\left[(\mathbf{X}^T \Omega \mathbf{X} + \Lambda^{-1})^{-1}(\mathbf{X}^T \Omega \boldsymbol{\theta} + \Lambda^{-1}\boldsymbol{\gamma}), (\mathbf{X}^T \Omega \mathbf{X} + \Lambda^{-1})^{-1}\right] \\
\pi|\text{rest} &\sim \text{Beta}\left[a + \sum_{i=1}^{N} Z_i Y_i, b + \sum_{i=1}^{N} Z_i(n_i - Y_i)\right] \\
\sigma^2|\text{rest} &\sim \text{InvGamma}\left[c + N/2, d + (\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{C}(\rho)^{-1}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right]
\end{aligned}$$

where $\Omega = \sigma^{-2}\mathbf{C}(\rho)^{-1}$ and $\tilde{\mathbf{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_N)^T$. The spatial range parameter is transformed to $\rho^* = \log(\rho)$. The full conditional for ρ^* is proportional to

$$|\mathbf{C}(\rho)|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{C}(\rho)^{-1}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right] \phi\left(\frac{\rho^* - m}{x}\right)$$

where $\mathbf{C}(\rho) = \mathbf{C}[\exp(\rho^*)]$ and ϕ is the standard normal density function. The log range is updated using Metropolis sampling with random-walk normal candidate distribution tuned to have acceptance probability around 0.4. In all analyses we generate 10,000 MCMC samples and discard the first 2,000 as burn-in. Convergence is monitored by visual inspection of the chains.

SM.4 – Simulation results for parameter estimation

Figure S2 summarizes the simulation results for estimating the parameters $\boldsymbol{\beta}$, π and ρ under the optimal design and the regular design (that includes X). Coverage for all parameters and all designs is at or near the nominal level. There is a bias for the intercept and the spatial range. This bias is consistent across the two designs and therefore likely not attributed to design issues, but rather simply an artifact of fitting a fairly complex model to a small datasets.

SM.5 – Sensitivity to starting values

For each of the eight combinations of β, π, and ρ we ran the exchange algorithm 10 times with different randomly-selected starting values for the m sampling locations; Figure 5 shows the best of the 10 solutions for each of the eight parameter settings. Figure S3 plots nine of the ten solutions for the scenario with high-quality auxiliary data, $\beta = 5$, high detection, $\pi = 0.7$, and strong spatial dependence in the true occupancy, $\rho = 2.0$. The 10 solutions are of course not identical, but all 10 place most of the sampling locations on the periphery of the distribution map estimated by eBird, and a few sampling locations on the edge of the domain where the eBird abundance estimate is low.

Figure S2: **Simulation study results for parameter estimation.** The boxplots summarize the sampling distributions of the posterior mean of the intercept (β_0), slope (β_1), detection probability (π), and spatial range parameter (ρ) for the simulation study. Results for the regular grid (white boxes) are compared with the optimal design (gray boxes). The horizontal dashed lines are the true values, the numbers along the top are the coverage percentages for the optimal design, and the numbers along the bottom are the coverage percentages for the regular grid. The first four simulation designs use the hot pocket covariate, the remaining four use the donut covariate, and the others are distinguished by the true values of the slope and detection parameters given by the dashed lines.

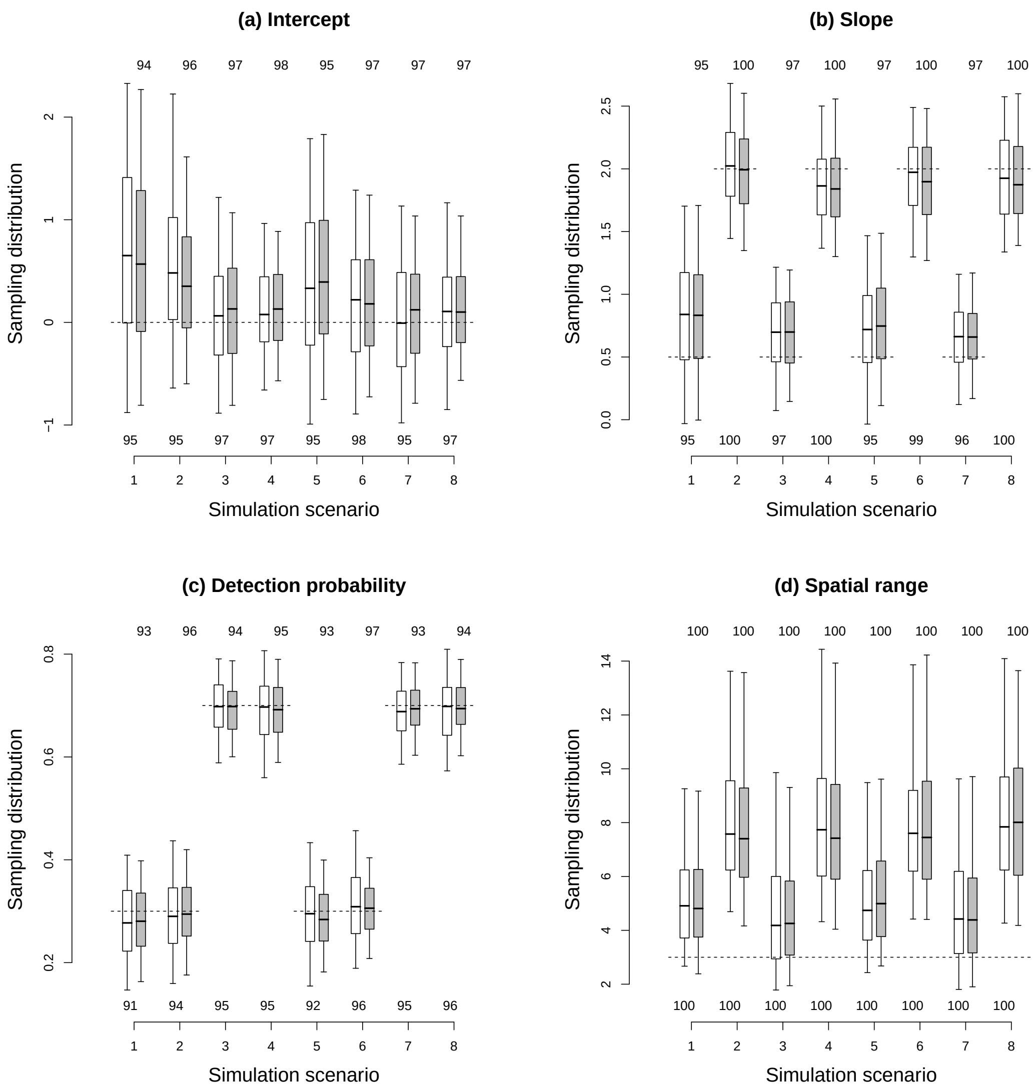


Figure S3: **Optimal spatial design for the brown-headed nuthatch for 9 different starting values.** The recommended sampling locations are white dots and the background color is the eBird abundance estimate. The designs are all for the case with true values $\beta = 5$, $\pi = 0.7$ and $\rho = 2.0$, but vary by initial configuration of sampling locations. The estimated design values $\mathcal{V}(\mathcal{D})$ (times 100) is given at the bottom of each plot ("V="); the standard error of these values is less than 0.01 in all cases.