Bayesian Analysis (2026) **TBA**, Number TBA, pp. 1–24

Bayesian Optimal Design for Occupancy Modelling Using Integrated Presence-Only and Presence-Absence Data

Max Savery* and Stijn Luca†

Abstract. Designing surveys and collecting data for the estimation of species occupancy presents both statistical and practical difficulties. Developing approaches that make conducting such surveys more efficient and flexible are becoming increasingly important, due to the need for accurate biodiversity monitoring across many diverse environments. In this work we present a Bayesian design approach that seeks to improve designs for occupancy modelling via the integration of citizen science presence-only data and structured presence-absence data within the design optimization. Using a Nonhomogeneous Poisson Process to include prior information in a site-occupancy model, we optimize designs that minimize error in the estimation of latent occupancy status. We demonstrate that by using presence-only and presence-absence data in these designs, it is possible to design surveys that better minimize occupancy estimation error and that are more efficient than those that otherwise do not incorporate presence-only data.

MSC2020 subject classifications: Primary 62K05, 62P12; secondary 62M30.

Keywords: species occupancy, species distribution models, citizen science, experimental design, Bayesian optimal design.

1 Introduction

Optimal design is used to construct surveys that efficiently measure otherwise unknown quantities in a population of interest. In the context of monitoring species populations, designing surveys according to their theoretical potential to accurately estimate the occupancy status of a species in a study region is often an important objective (Mackenzie and Royle, 2005). A survey for estimating occupancy will include observations of both presences and absences of the species of interest. This structured data is referred to as presence-absence (PA) data (Dorazio et al., 2011). Once a survey of this type has been conducted, a site-occupancy (SO) model can be used to estimate occupancy per site, taking into account auxiliary variables such as land cover, temperature, and other relevant information (Kéry and Andrew Royle, 2016). However, because of the imposed structure of the PA data, conducting the survey is labour-intensive, expensive, and time-consuming (Sanderlin et al., 2014). Therefore, integrating other abundant or more accessible sources of data into our design algorithm is advantageous if doing so means that we do not have to spend quite as much effort conducting the structured surveys.

*Department of Data Analysis and Mathematical Modelling, Ghent University, max.savery@ugent.be

†Department of Data Analysis and Mathematical Modelling, Ghent University, stijn.luca@ugent.be

 https://doi.org/10.1214/25-BA1581

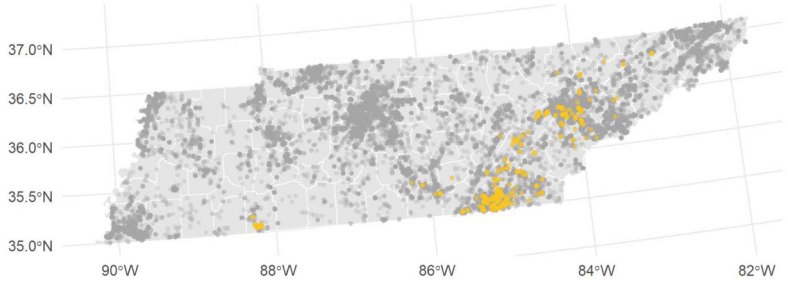

Figure 1: Brown-headed Nuthatch PO observations and checklists in Tennessee, from eBird database. PO observations are orange; checklist locations are dark grey.

One such advantageous data source is that collected by volunteer citizen scientists (Miller et al., 2019). Applications such as eBird, iNaturalist, and Observation.org allow citizen scientists to upload observations into a database. This data often falls into the category of presence-only (PO) data, which is characterized by observations that are recorded in an opportunistic or convenient fashion (Kéry and Andrew Royle, 2016). An example of this data can be found in Figure 1, showing Brown-headed Nuthatch PO data from the eBird dataset (eBird, 2021). This dataset will be used in the case-study in Section 5.

The cumulative effect of volunteers recording the presence of species has proven valuable for occupancy estimation in Species Distribution Models (SDMs). PO data can be used directly in SO models, such as in van Strien et al. (2013). Alternatively, SDMs that integrate both PA and PO data have been shown to be quite effective. This was first demonstrated in Dorazio (2014) and Fithian et al. (2015), and further elaborated on by a number of works (Koshkina et al., 2017; Pacifici et al., 2017; Peel et al., 2019; Simmonds et al., 2020; Ahmad Suhaimi et al., 2021; Dovers et al., 2024). For example, Koshkina et al. (2017) show that an integrated model can provide more accurate parameter estimates than a PO or PA SDM and the integrated model mitigates some of the issues associated with using each data source alone. Peel et al. (2019) show that using even a small amount of PO data can improve performance of a multi-species integrated model, particularly when only sparse PA data is available. Simmonds et al. (2020) also show that PO data can improve performance in the integrated model, provided that the sampling bias in the PO data model is accounted for and correctly specified. Another advantage of integrated models, discussed in Doser and Stoudt (2024), is that they can mitigate issues with single-visit surveys, useful when it may be difficult or resource-intensive to revisit every site more than once. Reviews by Miller et al. (2019) and Fletcher Jr. et al. (2019) highlight the practical challenges associated with data integration. They recognize the advantages that using integrated models offer, while pointing to the wider toolkit beyond the joint likelihood approach, such as that discussed in Pacifici et al. (2017) or a data weighting strategy. Given this rich background, integrated SDMs can be useful, but the modelling outcomes will depend on a number of factors. It is particularly important to account for the sampling bias inherent in the collection of the PO data, including spatial bias, difference in observer behaviour, and

detection bias (Johnston et al., 2023). If these sources of bias can be addressed, opportunistic data can make a valuable contribution to estimates of occupancy and species distribution.

Given the availability of citizen science PO data and its utility in integrated species distribution models, it may seem natural to use data integration to improve the design of occupancy surveys. While there is a wide variety of work about the design of surveys for occupancy estimation, there is limited previous research about the effect of data integration in the optimization of designs for occupancy, and very little about data integration for spatial design as a whole. Recent occupancy design work has investigated the effect of small sample size in the design (Guillera-Arroita et al., 2010), the optimization procedure subject to cost restraints (Sanderlin et al., 2014), design strategy and efficiency when detection is imperfect (Clement, 2016; Guillera-Arroita and Lahoz-Monfort, 2017), and the performance of occupancy design strategies in relation to asymptotic bounds (Reich, 2020). Other noteworthy design work for species distribution models includes Liu and Vanhatalo (2020), who create model-based designs when using log-Gaussian Cox processes; Mondain-Monval et al. (2024), who compare adaptive sampling approaches that direct citizen scientists to new sampling locations; and Williams et al. (2018), who provide a framework for creating designs for dynamically monitoring species populations. However, none of these works incorporate data integration into the designs.

There are few works focusing on both design and data integration. Guillera-Arroita et al. (2014), Pacifici et al. (2016), and Specht et al. (2017) present sequential, two-stage approaches using the PA data collected in the design of the first stage to inform on the PA data collection in the design of the next stage. While their goal is to efficiently allocate effort with a multi-stage approach, they do not use integrated models with PO data that immediately informs on the occupancy process. Yoo et al. (2020) is a good example of data integration outside of ecology, where model-based designs for air pollution are created using the integration of data from fixed monitoring stations, atmospheric models, and portable air sensors. Most relevant is the work of Reich et al. (2018). Here the authors integrate auxiliary information derived from citizen science data to find Bayesian optimal designs for occupancy estimation. However, they do not directly consider an integrated model (e.g., that of Koshkina et al. (2017)) or the effect of the integrated model on the design.

We advance upon the approach of Reich et al. (2018) by working directly with the integrated PA and PO model in a fully Bayesian design formulation, while still including auxiliary covariate data within the parameterized SO model. Though much work has been done on the integration of PA and PO data, no study has assessed such integrated models in the design of occupancy surveys. Studying the behaviour and complications of these models in design generation is important if multiple data sources are to be used to develop ecological surveys. Furthermore, we optimize over two objectives: The location of a set of sites of size m, and a configurable number of visits to each site, also referred to as site-specific survey effort. Optimizing over survey effort builds upon the approach of work such as Mackenzie and Royle (2005), Guillera-Arroita et al. (2014), Pacifici et al. (2016), Guillera-Arroita and Lahoz-Monfort (2017), Specht et al. (2017),

and Reich (2020). However, where these studies typically focus on designs where location and effort is allocated in stages or decision-based steps, we optimize algorithmically for site-specific survey effort and location jointly, completely within the Bayesian design framework, an approach which to the best of our knowledge has not been taken for occupancy designs.

This work lies at the intersection of the occupancy modelling and Bayesian design fields, which presents a variety of advantages and statistical and practical challenges. The approach we present here is widely applicable to any region and species for which PA data can be collected, where there is also an accessible PO citizen science database, and where there are covariates available that reasonably inform on each process. Though this is a flexible approach it depends on accounting for bias in the PO data, navigating computational complexities in the design optimization, and dealing with statistical estimation issues. But by presenting the work here, we hope to prompt further research in data integration in optimal design for species distribution modelling and general ecological monitoring.

2 Mathematical background

We now provide a brief review of the mathematical framework underlying the integrated site-occupancy model and Bayesian optimal design. We first describe how to use the Nonhomogeneous Poisson Process (NHPP) to integrate PO data into the site-occupancy model in a Bayesian sense. Then we present the optimal design procedure that takes advantage of these components in the design of surveys for occupancy estimation. Readers with a background in occupancy modelling may skip to Section 2.3. Readers with experience in optimal design but not in occupancy modelling may find the following sections useful.

2.1 Nonhomogeneous Poisson Point processes

A Nonhomogeneous Poisson process (NHPP) describes a set of points S occurring in region D. The points are indexed by $\{s_k\} \subseteq D$ for $k = 1, 2, \ldots, N(D)$, where $N(D)$ is the total number of points occurring in the region D. The process is characterized by an intensity $\lambda(s)$ which can be considered as the expected number of observations per unit area. Importantly, the NHPP has the property that the number of points $N(D)$ follows a Poisson distribution with a mean defined by the integral over the region

$$\lambda(D) = \int_D \lambda(s) ds. \tag{1}$$

We are typically interested in regions within D, e.g., a quadrat A of area $|A|$. The number of individuals occurring in that region, $N(A)$, also follows a Poisson distribution with mean $\lambda(A) = \int_A \lambda(s) ds$, such that $N(A) \sim \text{Poisson}(\lambda(A))$.

Following Banerjee et al. (2015), the nonparameteric likelihood for the PO data, L_{PO}, given the location density of the observations and the total number of points

$N(D)$ in D, is expressed as

$$L_{\text{PO}}(\lambda(s)) = \exp\left[-\lambda(D)\right] \prod_k \frac{\lambda(s_k)}{N(D)!}. \tag{2}$$

However, the integral $\int_D \lambda(s)ds$ is intractable and the likelihood is typically approximated numerically. Various approaches are discussed in the literature (Fithian and Hastie, 2013; Renner et al., 2015). It is convenient to discretize the region by imposing a grid over D and modelling the Poisson process likelihood as a product of independent Poisson variables, Each cell is referred to as site A_i, where i indexes the ith cell or site in the grid. The individual points can be aggregated to counts, so that $N(A)$ becomes N_{A_i} within their respective quadrats. This is visualized in Figure 1 of the Supplementary information (Savery and Luca, 2026).

Assuming the counts in each A_i are independent from the other quadrats, the counts are Poisson distributed with $N_{A_i} \sim \text{Poisson}(\int_{A_i} \lambda(s)ds)$, as before. Using a log-linear parameterization, we can rewrite the expected number of points in site A_i as

$$\int_{A_i} \lambda(s)ds = \int_{A_i} \exp[\alpha + \beta x(s)]ds, \tag{3}$$

where $x(s)$ refers to an environmental covariate of interest. Unfortunately, while we may want to integrate over s, we don't have information at the resolution of ds. In practice all we have is the covariate x_i for a particular site i:

$$\int_{A_i} \exp[\alpha + \beta x(s)]ds \approx |A_i| \exp[\alpha + \beta x_i] = \frac{|D|}{c} \exp[\alpha + \beta x_i], \tag{4}$$

where c is the number of pixels created by the imposition of the grid over D and $|D|$ is the total area of the region. $x(s)$ becomes x_i and $\lambda(s)$ becomes λ_i. This is known as the fine pixel approximation by Baddeley et al. (2010). We assume that covariate information is available for each site A_i and that it is relatively constant over each A_i, though if in practice the covariate is highly variable within the site or if the pixel size is too large, bias may be introduced into the model. Using this approximation, the distribution of counts in a given site i can then be written as

$$N(A_i) \sim \text{Pois}\left(\frac{|D|}{c} \exp[\alpha + \beta x_i]\right). \tag{5}$$

In cases where $|A_i| = 1$, we can disregard the offset. In this work we assume $|A_i| = 1$ unless otherwise mentioned.

Sampling bias parameterization

As reviewed in Johnston et al. (2023), the collection of citizen science PO data induces various sources of sampling bias, such as the sampling of locations that are convenient for volunteers to access or the selection of species that are appealing to observe. This

sampling bias must be accounted for in the species distribution model. The bias can be modelled in the NHPP by thinning the intensity $\lambda(s)$ by a sampling probability $b(s) \in [0, 1]$, as shown in the work of Fithian et al. (2015) and Dorazio (2014). $b(s)$ represents the proportion of the total number of individuals that the observer will record at location s. However, the choice of covariates for the parameterization of the bias depends on the modeller's knowledge of the domain and species of interest (Miller et al., 2019). When such covariates are available, $b(s)$ can be parameterized, here with a log-linear link function (Fithian et al., 2015; Dovers et al., 2024) with intercept γ and slope δ:

$$b_i = \exp[\gamma + \delta x_{2i}]. \tag{6}$$

The parameterization of the observed intensity now becomes

$$\tilde{\lambda}_i = \lambda_i b_i = \exp[\alpha + \beta x_{1i} + \gamma + \delta x_{2i}], \tag{7}$$

where x_1 and x_2 correspond to covariate data that inform upon the intensity and bias, respectively. $\tilde{\lambda}_i$ is the intensity driving the observed points, while λ_i represents the true, unobserved intensity. In the NHPP with only PO data, it is not possible to identify the parameters α and γ and estimate λ_i; only the sum $\alpha + \gamma$ can be estimated. However, in the site-occupancy model incorporating PO and PA data, the identification of these parameters becomes possible. This is notably explored by Dorazio (2014), Fithian et al. (2015), and Koshkina et al. (2017). Another related issue is that of multicollinearity between the covariates of intensity and bias, which can occur if these sets of covariates inform on both processes. Though this can lead to high uncertainty in the parameter estimates, both Simmonds et al. (2020) and Ahmad Suhaimi et al. (2021) observe that this does not necessarily transfer to poor predictive performance and the integrated model can still accurately capture the overall pattern of the species distribution.

Having parameterized the NHPP, the likelihood of the PO data becomes a function of the parameters α, β, γ, and δ: $L_{\text{PO}}(\alpha, \beta, \delta, \gamma)$. The posterior can now be defined as

$$p(\alpha, \beta, \gamma, \delta | \mathbf{y}_{\text{PO}}) \propto L_{\text{PO}}(\alpha, \beta, \gamma, \delta) p(\alpha) p(\beta) p(\gamma) p(\delta), \tag{8}$$

with independent, noninformative priors placed over the parameter space. We refer to a complete PO dataset as $\mathbf{y}_{\text{PO}}$ such that in our discretized setting $\mathbf{y}_{\text{PO}}$ contains counts for every site in the region, including sites with $N_{A_i} = 0$. The posterior from this NHPP will be used as the prior for the intensity parameters in the site-occupancy model. We elaborate on this in the next section.

2.2 Site-occupancy

Patterns of species occupancy are key to modelling species distribution. Site-occupancy models, such as that of MacKenzie et al. (2002), are used to estimate the occupancy probability ψ of a given species occupying a given site. To collect PA data, structured surveys must be performed, where a set of m sites is visited n_i times each. At each visit to site A_i, an absence or presence is recorded. At the end of the survey, a given species will have been detected y_i times at site A_i. We refer to $\mathbf{y}_{\text{PA}} = \{y_1, y_2, \ldots, y_m\}$ as a

PA dataset containing counts y_i, $i = 1, 2, \ldots, m$ for m total sites in the survey. We will also refer to an unobserved (latent) indicator of occupancy, $\mathbf{z} = \{z_1, z_2, \ldots, z_m\}$, where $z_i = 1$ if the site is occupied and 0 otherwise. Following MacKenzie et al. (2002) and Guillera-Arroita et al. (2014), the PA likelihood, L_{PA}, can be formulated as

$$L_{\text{PA}}(\boldsymbol{\psi}, \pi) = \prod_{i=1}^{m} \left[\psi_i \binom{n_i}{y_i} \pi^{y_i} (1-\pi)^{n_i - y_i} + (1-\psi_i) I(y_i = 0) \right], \tag{9}$$

where ψ_i is the occupancy probability for a particular site, $\boldsymbol{\psi}$ is the vector of occupancy parameters across the sites, and π is the detection probability. The indices i match the discretized regions A_i described in the previous section. Additionally, in this work we do not parameterize π and thus assume constant detection probability throughout the spatial region. While constant detection probability may be a reasonable assumption in surveys with short temporal windows, when using standardized recording equipment, or in single-observer studies (MacKenzie et al., 2018), modelling imperfect detection is an important topic of study within site-occupancy models (Guillera-Arroita et al., 2010; Dorazio, 2014). Considering the effect of detection probability on the design would be a valuable expansion to the work we present here.

It is possible to relate the intensity λ in the NHPP to the occupancy probability ψ in the site-occupancy model by the fact the probability of the number of counts in the NHPP being greater than 0 for a given site i is equivalent to occupancy at that site:

$$\psi_i = P(N_{A_i} > 0) = 1 - \exp(-\lambda_i) = 1 - \exp(-\exp[\alpha + \beta x_{1i}]). \tag{10}$$

The use of the log-linear parameterization naturally results in mapping the intensity λ to the occupancy probability ψ through the complementary log-log link function. This means that the parameters used in the NHPP to parameterize intensity can be used in the SO model to parameterize occupancy. The γ and δ parameters are not used here, as the occupancy probability does not depend on the sampling bias that is inherent in the collection of PO data and in the specification of the NHPP.

To write the Bayesian SO model with PO data as a prior, the PO data can be treated as historical data and the two models composed. The posterior of the Bayesian SO model with only PA data and independent priors can be written as

$$p(\alpha, \beta, \pi | \mathbf{y}_{\text{PA}}) \propto L_{\text{PA}}(\alpha, \beta, \pi) p(\alpha) p(\beta) p(\pi). \tag{11}$$

By using the PO posterior of (8) as a prior, the model becomes

$$p(\alpha, \beta, \gamma, \delta, \pi | \mathbf{y}_{\text{PA}}, \mathbf{y}_{\text{PO}}) \propto L_{\text{PA}}(\alpha, \beta, \pi) L_{\text{PO}}(\alpha, \beta, \delta, \gamma) p(\alpha) p(\beta) p(\gamma) p(\delta) p(\pi) \tag{12}$$

with independent, noninformative priors placed over the parameter space and assuming identifiability of α and γ. A short note about model composition is included in Section 1.3 of the Supplement. This approach allows us to combine PA and PO data via the integrated site-occupancy model described in the introduction.

By using a Bayesian formulation of the integrated SO model, it is possible to estimate the Posterior Predictive Distribution (PPD) of the unobserved occupancy status for a particular site z_i, as discussed in Dorazio et al. (2006). The PPD takes the general form

$$p(\tilde{z}_i|\mathbf{y}_{\text{PA}}) = \int_{\boldsymbol{\psi},\pi} p(\tilde{z}_i|\boldsymbol{\psi})p(\boldsymbol{\psi},\pi|\mathbf{y}_{\text{PA}})d\boldsymbol{\psi}d\pi, \tag{13}$$

where $p(\boldsymbol{\psi},\pi|\mathbf{y}_{\text{PA}})$ is the joint posterior of $\boldsymbol{\psi}$ and π. Most importantly, the use of the PPD to estimate the latent occupancy variable $\tilde{z}_i$ will make possible the development of Bayesian optimal designs.

2.3 Bayesian optimal design for spatial data

Typically, a solution to an optimal design problem seeks to find a configuration of experimental settings, referred to as a design, that minimizes a loss function or maximizes a utility function (in this work we will refer primarily to a loss function). Optimal design in the spatial setting is no different. Here we look for an optimal configuration of survey locations and optimal number of visits to each location. We call this design $\boldsymbol{d} \in \mathcal{D}$, where $\mathcal{D}$ is the design space. Using the integrated model above, it is possible to compare different configurations and move through the spatial domain, searching for the sites that minimize a loss function $\mathcal{L}(\mathbf{y}_{\text{PA}}, \mathbf{z}, \boldsymbol{\psi}, \mathbf{d})$. In this work we use a Bayesian optimal design approach, as discussed in Diggle and Lophaven (2006), Ryan et al. (2016), and Liu and Vanhatalo (2020), where the loss is a function of the posterior predictive distribution and is integrated over both the posterior distribution of the parameters and prior predictive distribution of the observed data. This stands in contrast to the "pseudo-Bayesian" approach to design where the loss is averaged over the priors placed on the parameters (Ryan et al., 2016). The expected loss can be generally written as:

$$\text{E}[\mathcal{L}(\mathbf{d})] = \int_{\mathbf{y}_{\text{PA}}} \int_{\boldsymbol{\psi}} \int_{\mathbf{z}} \mathcal{L}(\mathbf{y}_{\text{PA}}, \mathbf{z}, \boldsymbol{\psi}, \mathbf{d}) p(\mathbf{z}|\boldsymbol{\psi}, \mathbf{y}_{\text{PA}}, \mathbf{d}) p(\boldsymbol{\psi}|\mathbf{y}_{\text{PA}}, \mathbf{d}) p(\mathbf{y}_{\text{PA}}|\mathbf{d}) d\mathbf{z} d\boldsymbol{\psi} d\mathbf{y}_{\text{PA}}, \tag{14}$$

where $\mathbf{d}$ refers to a particular design, $p(\mathbf{z}|\cdot)$ is the posterior predictive distribution, $p(\boldsymbol{\psi}|\cdot)$ is the posterior, and $p(\mathbf{y}_{\text{PA}}|\mathbf{d})$ is the prior predictive distribution for the observed data. The optimal design configuration is found by minimizing the expectation:

$$\mathbf{d}^* = \text{argmin}_{d \in \mathcal{D}} \text{E}[\mathcal{L}(\mathbf{d})]. \tag{15}$$

The expectation of the loss must be approximated, as described in the next section.

2.4 Design optimization

An exchange algorithm (Royle, 2002) is used to determine the optimal site and visit configuration. The implementation of the exchange algorithm is described in Section 1.2 of the Supplementary material. The optimal design that is found from this procedure will consist of both the optimal location of sites to survey and the optimal number of times to visit each site. We fix the number of sites to survey, m, and use a fixed

set of potential number of visits to the ith site, $\mathbf{v_i} = \{n_{\min}, \ldots, n_{\max}\}$, where $n_{\max}$ is the maximum times to visit. The specific $\mathbf{v_i}$ will vary per experiment. In this work we optimize over a set of size two, $\mathbf{v_i} = \{n_{\min}, n_{\max}\}$. However, using sets larger than two is easily configurable, albeit at computational cost.

To optimize over sites and visits, a metric is required to compare design configurations. The Brier score (Brier, 1950; Gneiting and Raftery, 2007), equivalent to the Mean Squared Error for discrete estimation, is used here. The loss function of (14) becomes

$$\mathcal{L}(\mathbf{y}_{\text{PA}}, \tilde{\mathbf{z}}, \boldsymbol{\psi}, \mathbf{d}) = \frac{1}{c}\sum_{i=1}^{c}(z_i - \hat{z}_i)^2, \tag{16}$$

where $\hat{z}_i = P(z_i = 1|\mathbf{y}_{\text{PA}}, \mathbf{d})$ and is computed as the mean of the posterior predictive distribution for $\tilde{z}_i$. $\tilde{\mathbf{z}}$ refers to a set of predicted occupancy probabilities over all c sites in region D. Here we aim to directly optimize the ability of the designs to predict occupancy, which the Brier score is exactly suited for. However, there are a number of other possible utility and loss functions that could be used, as this component of the design framework is quite flexible. To adapt the framework to a different goal or application, researchers must substitute their desired optimization function and modify the output of the predictive model to match the study objective.

To optimize sites, we search for the design that minimizes the expected Brier score, $E[\mathcal{L}(\mathbf{d})]$. The expectation of $\mathcal{L}(\mathbf{d})$ in (14) will need to be approximated. To do so, it is necessary to simulate R $\mathbf{y}_{\text{PA}}$ datasets from the prior predictive distribution $p(\mathbf{y}_{\text{PA}}|\mathbf{d})$ (Ryan et al., 2016; Liu and Vanhatalo, 2020). The details of this procedure are discussed in Section 3. We then average the loss over these R draws of the prior predictive PA datasets:

$$E[\mathcal{L}(\mathbf{d})] \approx \frac{1}{R}\sum_{r=1}^{R}\mathcal{L}({\mathbf{y}_{\text{PA}}}^{(r)}, \tilde{\mathbf{z}}^{(r)}, \boldsymbol{\psi}^{(r)}, \mathbf{d}), \tag{17}$$

where $\mathbf{z}^{(r)}$ refers to the set of occupancy estimates over the whole region D. Though π is technically included in the integration, we do not include it in the notation above as its estimation is not the focus of this work.

We highlight here that the Bayesian integrated SO model is used within the numerical approximation of the expected loss, where it is fit to each simulated PA dataset and the already collected, real-world PO data, in order to find an optimal design for collecting future PA data. Once the survey is conducted, the integrated model must be fit again to the collected PA data and the same PO dataset. If the PO data is not included in the post-survey analysis, the occupancy estimation will be subject to bias induced by the non-random, preferential selection of the survey sites, as the sampling locations are not selected independently of the occupancy process during the design optimization.

The complete runs of the exchange algorithm are repeated for 10 initial configurations of the m sites in order to account for the randomness in the initial choice of the sites. The first run across all experiments always uses the same manual initial configuration (shown in Figure 8 of the Supplementary material), in order to be able to compare

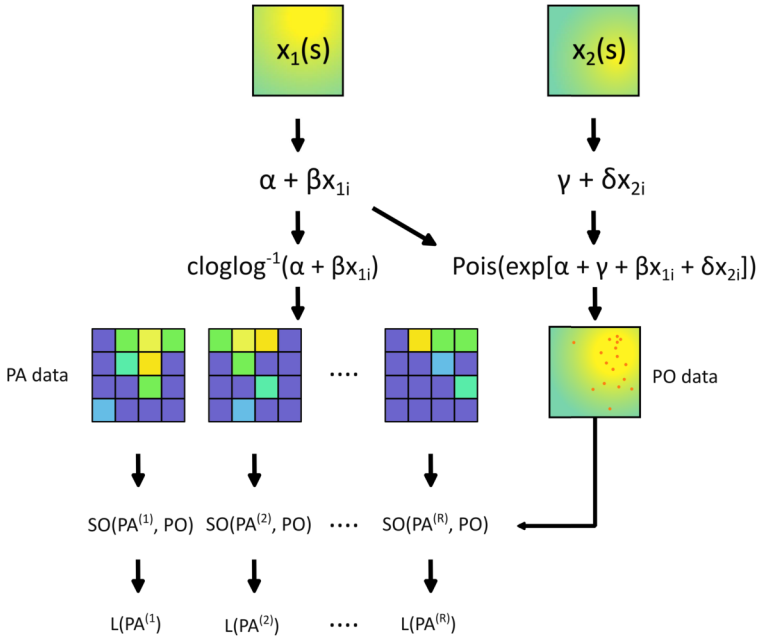

Figure 2: Simplified illustration of the data generation procedure, use of the integrated species distribution model, and optimal design approach in the simulated environment. The simulated covariate surface for $x_1(s)$ and $x_2(s)$ is shown at the top. The respective data generating models are then used to generate R PA datasets and one PO dataset. The integrated SO model takes as input a given PA dataset and the PO data. From the estimated model we can calculate the loss or utility, which will be averaged over all R repetitions to approximate the expectation.

the behaviour of the algorithm given a specific initial configuration; and the other 9 are randomly initiated. We report the average of $E[\mathcal{L}(\mathbf{d})]$ over these starts. The computation of the expected loss is the primary computational bottleneck of the Bayesian design framework. To make the approximation of the expected loss feasible, we parallelize the fitting of the R posteriors. Each posterior is fit using the CmdStanR library (Gabry et al., 2022) in R version 4.2.1 (R Core Team, 2022). An additional discussion about alternative approximations to the posterior of the integrated model and the scalability of the design framework can be found in Section 1.4 of the Supplementary material.

3 Simulated environment

We next describe the simulated environment and data generation procedures we use. We first discuss the data generation of the PA and PO datasets in the simulated environment and then consider how to generate the PA data for the real-world case-study of Section 5. A simplified summary of the entire design process is illustrated in Figure 2.

3.1 Presence-only data

To create a simulated environment, we generate grid-based covariates $\mathbf{x} = \binom{x_1}{x_2}$, where x_1 is the environmental information for intensity and x_2 for sampling bias. The functions used to define this covariate surface follow Reich et al. (2018). This simulated environment can be seen in the Supplementary material in Figures 2a and 2b, where it is described in more detail. To better control the simulation conditions in our experiments, we use the grid-based covariates to generate Poisson counts N_{A_i} for each quadrat A_i according to our parameterization of the intensity in (7). The values of α, β, γ, and δ are fixed to known values in the simulation experiments for the purpose of comparison between experimental runs.

3.2 Presence-absence data

Data simulated from the prior predictive distribution of $\mathbf{y}_{\text{PA}}$ is used to approximate the expectation of the design criterion. To generate a single dataset $\mathbf{y}_{\text{PA}}$, the parameterized intensity for site A_i is mapped to the occupancy probability using the complementary log-log link function shown in (10). Using the probability of occupancy at each site, the data generating process is then

$$\begin{aligned} z_i|\psi_i &\sim \text{Bernoulli}(\psi_i) \\ y_i|z_i &\sim \text{Binomial}(n_i, \pi z_i), \end{aligned} \tag{18}$$

where z_i is the binary indicator of occupancy, and the detection probability is held constant at $\pi = 0.2$ across sites.

To integrate over the loss, it is necessary to draw R future datasets of $\mathbf{y}_{\text{PA}}$. But to do so requires knowledge regarding α and β. In the simulated environment described above we use known values of α and β, in order to decouple the data integration from the effect of drawing from the prior predictive distribution of $\mathbf{y}_{\text{PA}}$. In real-world conditions where α and β are unknown, it is necessary to first estimate the parameters and then use these estimates, $\hat{\alpha}$ and $\hat{\beta}$, to generate the PA datasets. In the work here, we use the NHPP to pre-estimate the parameters. This does, unfortunately, raises two complications: The first is that α and γ are not identifiable in the NHPP alone but the sum is. To deal with this, we first integrate out γ by averaging multiple posterior estimates of α over draws from a uniform prior on γ. From this procedure we retrieve an $\hat{\alpha}$ that can be used to simulate from the prior predictive distribution. The second issue is that we are now using the PO data and the NHPP twice: once to pre-estimate the parameters for PA simulation, and again within the integrated model during the design optimization. This will be addressed in the fourth experiment of Section 4. Both issues are described more thoroughly in Section 5.1 of the Supplementary material. While such sequential estimation is not the main focus of our work, we are not aware of any previous work exploring the practical effect on occupancy designs of data generation via the prior predictive distribution, although the generation of such future data for adaptive Bayesian designs is discussed in various work including Leach et al. (2022), and Thilan et al. (2024).

4 Simulation experiments

To explore the extent to which integrating PO data affects optimal designs, we compare the algorithm performance with and without PO data in the SO model in a variety of simulated environments, where the number of sites in the survey, maximum number of visits, and data-generating parameters are permuted. We refer to the PA-only model from (11) as SO(ψ_i) and the PA + PO model from (12) as SO(ψ_i) + PO. The models and priors are fully described in Section 1.3 of the Supplementary material. We conduct 4 sets of experiments comparing the optimal designs that result using these models. The experiments are as follows:

1. **Fixed visits**: We first hold the number of visits constant within the exchange algorithm to observe only the effect of the data integration.
2. **Variable visits**: We then allow the number of visits to vary within the algorithm. By jointly optimizing over location and effort and including PO data, designs should more efficiently survey the region.
3. **Maximum effort** Next, we vary $n_{\max}$, the maximum allowed number of visits to each site.
4. **Model misspecification** Finally, we introduce misspecification into the design by incorporating an additional covariate into the PO data generation model that is not included in the working SO model. In this experiment we also test the sensitivity of the designs to the prior predictive PA data generation procedure.

In each of these experiments, we compare up to 4 intensity landscapes, where the parameters used for the data generation procedure are set so as to generate a particular PO pattern. We permute between $\beta \in \{0.5, 1\}$ and $\delta \in \{0.25, 2\}$, and leave α and γ fixed, in order to consider only the strength of the site-specific covariates. In all experiments, $\alpha = -2$ and $\gamma = 1$, which were chosen to balance the datasets and prevent the PO data from overwhelming the PA data in every scenario. The datasets these parameter values create are shown in Supplementary Figure 4.

Table 1 shows the outcome of the first experiment, where PO data is included in the design and the visits are fixed. Each entry of the table refers to the expected loss of the optimal design, averaged over the 10 random starts of the exchange algorithm.

	$\beta = 0.5$		$\beta = 1$	
	$\delta = 0.25$	$\delta = 2$	$\delta = 0.25$	$\delta = 2$
SO(ψ_i), n = 1	0.1338	0.1338	0.1199	0.1199
SO(ψ_i), n = 3	0.1313	0.1313	0.1129	0.1129
SO(ψ_i) + PO, n = 1	0.1022	0.1021	0.1000	**0.0940**
SO(ψ_i) + PO, n = 3	0.1171	0.1173	0.1073	0.1026

Table 1: Comparison of $E[\mathcal{L}(\mathbf{d})]$ between optimal designs, when PO data is and is not included, holding visits constant at $n = 1$ and $n = 3$, $m = 5$. SO(ψ_i) + PO refers to the integrated model with PO data; SO(ψ_i) refers to the model without PO data. The best-performing setting (lowest) is shown in bold.

In all tables, the most important comparison is between models and within the same experimental settings (n, β, γ). The results show that in all situations, designs generated using SO(ψ_i) + PO lead to a smaller average $E[\mathcal{L}(\mathbf{d})]$ than compared to the designs that do not integrate PO data, SO(ψ_i). The difference between the models is largest in the single-visit $n = 1$ scenario, indicating that the integrated approach may be particulary useful when replicating visits at survey sites is prohibitively expensive. However, the performance also depends on the strength of β. When $\beta = 1$, the performance of SO(ψ_i) approaches the integrated model. Even though $\beta = 1$ leads to more PO data, it also leads to a clearer PA pattern, such that the PO data becomes less necessary in the selection of sites. Note that SO(ψ_i) will always generate the same results between different values of δ because δ only affects the PO data generation.

The second experiment assesses the effect of jointly optimizing both the location of each site and the number of visits. These results are shown in Table 2 for both β settings and $\delta = 2$. The results over all parameter values are shown in Table 1 of the Supplementary material. In all settings integrating PO data results in lower average loss. Furthermore, across both models and all settings, allowing the visits to vary during the optimization leads to lower average loss. This effect is less pronounced when $\beta = 1$ and $n_{\max} = 2$.

Table 3 shows the average number of visits per site of each design when optimizing site-specific survey effort. In all situations integrating PO data results in fewer average structured visits to each location. For example, in the $\beta = 1$, $\delta = 2$, and $n = 3$ setting,

	Constant visits		Varying visits	
	$\beta = 0.5$, $\delta = 2$	$\beta = 1$, $\delta = 2$	$\beta = 0.5$, $\delta = 2$	$\beta = 1$, $\delta = 2$
SO(ψ_i), n = 2	0.1316	0.1067	0.1245	0.1064
SO(ψ_i), n = 3	0.1313	0.1129	0.1248	0.1036
SO(ψ_i) + PO, n = 2	0.1123	0.0974	0.1027	0.0958
SO(ψ_i) + PO, n = 3	0.1173	0.1026	0.1027	**0.0930**

Table 2: Comparison of $E[\mathcal{L}(\mathbf{d})]$ with and without varying visits, with and without PO data, $m = 5$, $\delta = 2$. When allowing the visits to vary, $n = 2$ and $n = 3$ refers to $n_{\max}$, where the sets of potential visits are $\mathbf{v_i} = \{1, 2\}$ and $\mathbf{v_i} = \{1, 3\}$ respectively.

	$\beta = 0.5$		$\beta = 1$	
	$\delta = 0.25$	$\delta = 2$	$\delta = 0.25$	$\delta = 2$
SO(ψ_i), n = 2	1.68	1.68	1.66	1.66
SO(ψ_i), n = 3	2.20	2.20	2.12	2.12
SO(ψ_i) + PO, n = 2	1.34	1.34	1.48	1.42
SO(ψ_i) + PO, n = 3	1.56	1.52	1.64	1.72

Table 3: Average visits per site when optimizing for effort, across values of β and δ. To compare to the experiments that do not optimize visits, consider that the average number of visits will be equal to n for a given run.

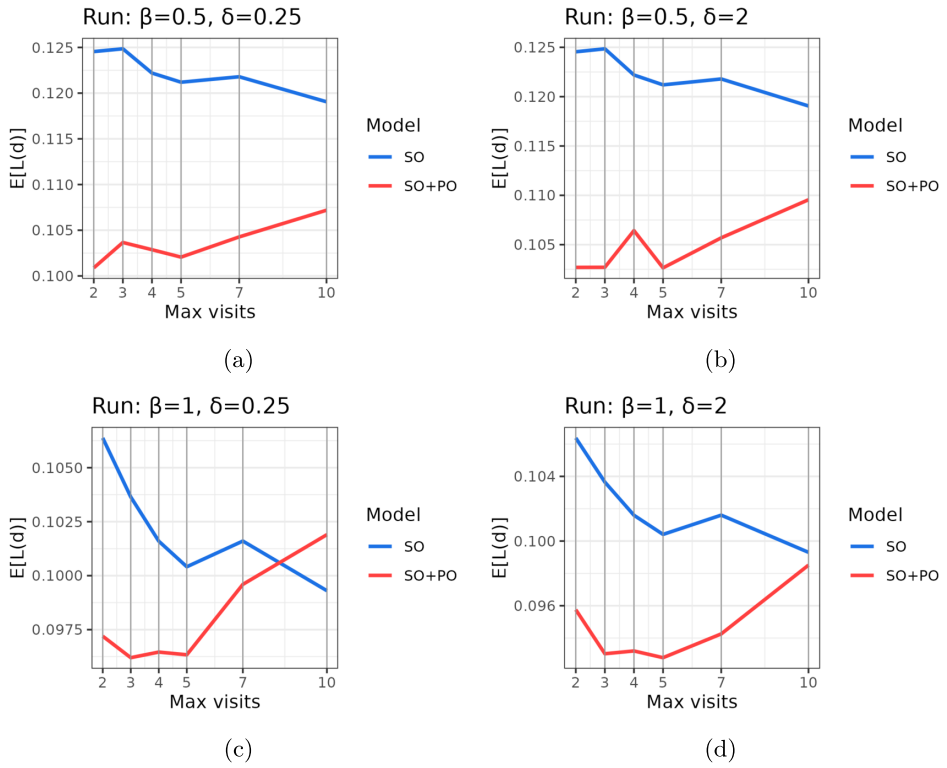

Figure 3: Comparison of $E[\mathcal{L}(d)]$ between models in different data generating environments, with PO data (red) and without PO data (blue), as the maximum number of visits to each site is increased. Runs with only PA data and the same β setting result in the same loss. Vertical dark grey lines indicate the maximum number of visits for each run: $n_{\max} \in \{2, 3, 4, 5, 7, 10\}$.

SO(ψ_i) + PO needs on average only 1.72 visits to each site, whereas SO(ψ_i) needs 2.12 visits. This results in, on average, 4 visits saved for every 10 location surveyed, an 18.9% reduction in effort compared to the PA-only designs. And, as indicated by Table 2, the integrated designs still result in more accurate estimates of occupancy. Even in the settings where SO(ψ_i) nears SO(ψ_i) + PO performance ($\beta = 1$), we require less structured PA data when PO data is included. This result will be corroborated in the case-study as well.

In the third experiment we compare designs for increasing values of $n_{\max}$. The results are shown in Figure 3. In these results, the SO(ψ_i) + PO designs perform better than SO(ψ_i) in nearly all cases, but the performance of SO(ψ_i) designs improve as more survey effort is allowed. We can also observe that in each figure, though particularly in the $\beta = 1$ setting, the loss in the integrated model increases at $n_{\max} = 10$. An additional analysis shown in Figure 5 of the Supplementary material indicates that this is due to

the low detection probability used in these experiments ($\pi = 0.2$) and the influence of detection on the estimate of α. When the detection probability is low, it is more likely that the PA data conflicts with the PO dataset. In Figure 6 of the Supplement we also show the effect of increasing the number of sites, where we observe similar results as Figure 3.

We next address two additional challenges that arise when using the integrated model. First, to consider the sensitivity of the design procedure to misspecification of the sampling bias, we introduce an additional covariate in the PO data generating model that is absent in the integrated working model. This is in the spirit of experiments in Simmonds et al. (2020) and Ahmad Suhaimi et al. (2021), where integrated models are fit without the covariate accounting for bias. This misspecified covariate is shown in the Supplementary material in Figure 7. The influence of the covariate is controlled by varying its rate of (radial) decay, ρ, across the spatial region.

The second issue is that of PA data generation, used for the approximation of the expectation of the utility. In a real-world setting α and β must be estimated for the conditional simulation from the prior predictive distribution of $\mathbf{y}_{\text{PA}}$. In the experiments below, we analyze the sequential procedure of Section 3.2. There are three versions compared: (1) The "oracle" procedure used in the above experiments where α is known; (2) the sequential procedure for estimating α without integrating over γ; and (3) the sequential procedure integrating over γ.

The results shown in Table 4 indicate that the oracle approach (column 1) is robust to misspecification and the sequential runs were more sensitive, particularly without integration over γ (columns 2 and 4). In these runs, the estimate of α used to generate the PA data will be highly variable and this effect appears to propagate to the designs. When using the sequential method with integration, the design performance is much

	Oracle	Sequential, SO(ψ_i) + PO		Sequential, SO(ψ_i)	
		w/out int.	w/ int.	w/out int.	w/int.
$\beta = 0.5, \delta = 0.25, \rho = 0.01$	0.1085	0.0980	0.0167	0.1193	0.0484
$\beta = 0.5, \delta = 0.25, \rho = 0.05$	0.1089	0.1678	0.0380	0.1622	0.0702
$\beta = 0.5, \delta = 0.25, \rho = 0.1$	0.1084	0.1210	0.0333	0.1268	0.0608
$\beta = 0.5, \delta = 0.25, \rho =$ None	0.1094	0.1301	0.0235	0.1358	0.0553
$\beta = 0.5, \delta = 2, \rho = 0.01$	0.1071	0.0697	0.0134	0.0916	0.0475
$\beta = 0.5, \delta = 2, \rho = 0.05$	0.1076	0.1857	0.0318	0.1685	0.0598
$\beta = 0.5, \delta = 2, \rho = 0.1$	0.1070	0.0640	0.0161	0.0790	0.0517
$\beta = 0.5, \delta = 2, \rho =$ None	0.1051	0.0773	0.0142	0.0805	0.0421

Table 4: Evaluation of misspecification and strategy for generation of PA data. For each setting $E[\mathcal{L}(d)]$ is shown, visits $n_{\max} = 3$. The optimal designs were found using the three described strategies to generate PA data: Oracle, sequential with integration, and sequential without integration. For the sequential approach, the SO(ψ_i) + PO model and the SO(ψ_i) model is used. The level of spatial decay is permuted in each environment: $\rho \in \{0.01, 0.05, 0.1\}$. $\rho =$ None refers to the environment without misspecification.

lower and less variable than the runs without integration. Importantly, the SO(ψ_i) + PO runs with integration over γ (column 3) perform better than the SO(ψ_i) runs (column 5). However, compared to the oracle runs, the runs with integration over γ are much better (for example 0.1085 vs 0.0167 in the first row of the table). This indicates there may be an issue with overfitting during the design stage due to the double use of PO data—once to estimate α and β and again in the integrated model (for SO(ψ_i) + PO).

Finally, to fully validate the designs we mimic a real survey scenario: The designs are used to select future PA survey locations. New PA data is "collected" (simulated) at these locations and SO models are fit to this new data. This result is shown in Table 3 of the Supplementary material. The SO(ψ_i) + PO designs from the sequential approach with integration are comparable to that of the oracle, and consistently outperformed the SO(ψ_i) designs. While the optimization criterion is apparently overconfident, the designs integrating PO data are still more effective once the new PA data is collected.

In first three experiments presented here, we demonstrate that by integrating PO data in the SO model, it is possible to develop more efficient and accurate occupancy surveys. While we found that the designs will be sensitive to misspecification and the data generation procedure, all of which must be carefully considered by practitioners, the data integration approach consistently outperformed its PA-only counterpart.

5 Case-study

5.1 Data collection

In this case-study we focus on designs for occupancy estimation of the Brown-headed Nuthatch (BHN). We emulate a scenario where researchers face resource constraints, such as limited budget, expensive labour, or other logistical challenges, and therefore must leverage a more abundant citizen science data source to maximize the resources used for the monitoring effort. As citizen science data, we use PO data available from the eBird database (Sullivan et al., 2009; eBird, 2021). The eBird database contains detailed observations records that are amenable to PO modelling. We use observations of the Brown-headed Nuthatch collected in 2019 within Tennessee. Tennessee is on the edge of the range map of the BHN and thus presents an interesting case-study for the development of designs. The data is processed following Johnston et al. (2021). Briefly, we use only checklists in which the sampling protocol is listed as "Stationary", or "Travelling". We filter out checklists in which the duration is greater than 5 hours and the distance travelled is greater than 5 km. Checklists with more than 10 observers are also removed. These steps are taken to reduce variability within the observations.

Land cover, elevation, and Enhanced Vegetation Index (EVI) are used as ecological covariates. These are further described in Section 5.2 of the supplement. More detailed instructions for recreating the ecological covariate data and eBird data are described in the documentation of the code that accompanies this paper.

5.2 Modelling

The modelling and exchange algorithm here proceeds as described in the simulated environment, with a few changes to accommodate the additional data sources. The intensity and bias are parameterized as

$$\begin{aligned} \lambda_i &= \exp\left[\alpha + \beta_1 x_{\text{EVI},i} + \beta_2 x_{\text{elevation},i} + \beta_3 x_{111,i} + \beta_4 x_{115,i} + \beta_5 x_{124,i}\right] \\ b_i &= \exp\left[\gamma + \beta_6 x_{50,i}\right], \end{aligned} \tag{19}$$

where the subscripts refer to the covariate types by land cover map codes shown in Figure 10 of the Supplementary material. In the cloglog link the intensity is offset by the area of the buffers. The buffers are analogous to the sites A_i from the simulation experiments. The intensity is therefore corrected by the area of each site, such that $\psi_i = 1 - \exp(-|A_i|\lambda_i)$. Finally, we follow the sequential approach with integration over γ as described in Section 5.1 of the Supplementary material to pre-estimate α and β for the PA data generation.

5.3 Results

We now compare the SO(ψ_i) + PO and SO(ψ_i) approaches for optimizing BHN occupancy designs. The runs are shown in Table 5. We pair relatively few, resource-intensive PA surveys (5 to 80 total visits) with an already available, abundant PO dataset of 1208 observations, mimicking a scenario where researchers integrate citizen science data in order to avoid wasting resources and maximize the information gained from the survey. We evaluate the effect of adding increasingly more PO into the integrated model, ranging from 5% to 100% of the PO dataset, randomly sampling without replacement the given proportion of the full PO dataset.

Table 5 shows the extent to which the SO(ψ_i) + PO runs lead to better designs for BHN occupancy. The benefit of PO data is most pronounced when there is less structured sampling effort, such as in the case of $n = 1$ and $n = 2$. As the structured sampling effort increases, the advantage diminishes, in some cases leading to the SO(ψ_i) designs overtaking the data integration approach.

Table 6 shows only SO(ψ_i) runs, for $m \in \{5, 10, 15, 20\}$. These additional runs highlight the data inefficiency of the PA surveys without PO data. For example, when $m = 20, n = 4$ we achieve a similar loss (0.1760) as when $m = 5, n = 1$ for the SO(ψ_i) + PO = 1 run (0.1746), the same performance at 16 times the effort. Alternatively, the loss of SO(ψ_i) at $m = 20, n = 4$ is similar to the SO(ψ_i) + PO = 0.1 run at $m = 10, n = 4$. By including only 10% the total PO dataset, the survey effort is cut in half and the design still achieves similar performance. It is certainly possible to develop accurate and effective high-effort surveys with only PA data. But integrating PO data enables the design of surveys that achieve this accuracy with less structured sampling effort.

	m = 5 n = 1	m = 5 n = 2	m = 5 n = 4	m = 10 n = 1	m = 10 n = 2	m = 10 n = 4
SO(ψ_i)	0.2259	0.2083	0.1923	0.2144	0.2009	0.1838
SO(ψ_i) + PO = 0.05	0.1846	0.1964	0.1984	0.1952	0.1940	0.1818
SO(ψ_i) + PO = 0.1	0.1825	0.1972	0.1954	0.1903	0.1899	0.1753
SO(ψ_i) + PO = 0.2	0.1837	0.1958	0.2037	0.1899	0.1918	0.1751
SO(ψ_i) + PO = 0.3	0.1779	0.1876	0.1995	0.1861	0.1997	0.1790
SO(ψ_i) + PO = 0.5	0.1831	0.1896	0.1990	0.1880	0.1942	0.1800
SO(ψ_i) + PO = 0.7	0.1808	0.1928	0.1922	0.1841	0.1875	0.1771
SO(ψ_i) + PO = 1	**0.1746**	0.1853	0.1992	0.1860	0.1958	0.1856

Table 5: Comparison of SO model performance for surveys for the BHN in terms of $E[\mathcal{L}(d)]$, with and without PO data. The results are shown as the proportion of PO data, number of sites, and number of visits per site is increased.

	m = 5	m = 10	m = 15	m = 20
SO(ψ_i), n = 1	0.2259	0.2144	0.2095	0.2022
SO(ψ_i), n = 2	0.2083	0.2009	0.1912	0.1861
SO(ψ_i), n = 4	0.1923	0.1838	0.1786	**0.1760**

Table 6: Effect of varying number of sites and visits when no PO data is included.

6 Discussion

In this paper we have presented a Bayesian optimal design for occupancy estimation that integrates PO and PA data. Through both simulations and a case-study, we established that by integrating the PO data in the SO model, it is possible to achieve designs that better optimize the expected loss and are more efficient in the number of visits when compared to respective designs based on PA data alone. In our analysis of model misspecification and the procedure for PA data generation, the designs tended to be robust to misspecification but sensitive to the procedure for PA generation. We take these results as evidence in support of the argument that, with careful implementation, the data integration approach presented here can have real-world benefit for creating occupancy surveys that save on cost, labour, time, or other resources. The approach will be most useful when collecting PA data is highly resource-intensive.

However, the effectiveness of this approach depends on the balance of PO and PA data and the behaviour of underlying environmental process. In situations where PA data is simple to collect or otherwise available, the benefit of PO data diminishes. But in those same settings, it is quite possible that PO data will also be readily accessible. Such a situation is similar to the $\beta = 1, \delta = 0.25$ setting in the simulated environment, where the strong effect of the environmental covariate increases the probability of both types of observations. The PO and PA integration approach is therefore best suited for situations in which PA data is prohibitively inefficient to acquire and in which PO data usefully informs on the PA data generating process. Future research could explore the

effect of PO filtering techniques and noise reduction (such as that of Sicacha-Parada et al. (2021) and Van Eupen et al. (2021)) on the effectiveness of integrating PO data in the designs.

Additionally, it is essential to properly account for the sampling bias in the integrated model, as discussed in work such as Pacifici et al. (2017), Ahmad Suhaimi et al. (2021), Peel et al. (2019) and Simmonds et al. (2020). If the PO data is collected with high amounts of bias, or if the bias is misspecified, PO data integration can have a detrimental effect on the occupancy estimation. In our case, this will in turn bias both the generation of PA data for numerical integration and the integrated model itself. Our work primarily assumed bias in the PO data model is correctly specified in order to provide a foundational framework for data integration in Bayesian design. While our misspecification experiment indicates that the designs have some robustness to misspecification of the sampling bias, exploring misspecification and the quality of PO data for specific species and environments is required to fully adapt this approach to a real-world sampling environment. We also did not explore other cases in which the PO data can be detrimental to the design, such as the occurrence of false positives in the PO data. Approaches to mitigate the effect of the double-use of PO data may also be necessary, such as introducing regularization or informative priors from previous studies into the PA generation. Finally, here we assumed constant detection probability in order to focus on the properties of the design and data integration, but a necessary extension to our work is the parameterization of the detection probability in the integrated model. This will require exploring the behaviour of the designs as the detection probability varies across the environment, per observer or per species, and will also require covariates independent of those used to model the sampling bias in the PO data.

The computational demand of the Bayesian design framework is an important consideration for practical applications. Its use may be most justified when surveys are expensive, logistically complex, or intended for long-term monitoring programs where the efficiency gained from an optimal design outweighs computational cost. If one's goals are adequately served by a simpler approach to incorporating prior information into a survey, then site-selection via model-based predictive uncertainty or environmental variation are faster alternatives. Future work should compare the trade-offs of using such faster heuristic methods to the more robust Bayesian design approach, as well as more efficient optimization algorithms, such as Overstall and Woods (2017), in order to apply the Bayesian design framework to larger spatial scales. In this work we also did not incorporate spatial correlation structure in the integrated model. To do so will require the replacement of the Markov chain Monte Carlo (MCMC) posterior approximation with alternatives, e.g., the Integrated nested Laplace approximation or template model builder, such as in Dovers et al. (2024) or Belmont et al. (2024). We discuss these possible extensions in the Supplementary material, Section 1.4.

While we have studied strictly data integration in occupancy surveys, the framework we use can be applied to other ecological monitoring settings. These include designs for predicting relative intensity surfaces or abundance, designs for parameter estimation and power, and more complex scenarios such as capture-recapture surveys, multi-species monitoring, or biodiversity trends. Further promising applications include integrating

citizen science data into designs for monitoring spatiotemporal dynamic processes, such as designs for animal movement models (Williams et al., 2018; Leach et al., 2022), designs for monitoring trends in coral reefs (Thilan et al., 2024), or integrating other data sources into adaptive designs for citizen scientists (Mondain-Monval et al., 2024). Moreover, the decision to use citizen science data depends on the resources, goals, and mandates of the surveying agency. It is necessary to assess how the data can positively or negatively contribute to designs for particular species and regions under the guidelines and objectives of specific biodiversity monitoring policies.

Acknowledgments

Thanks are given to the members of the Biostat group in the Faculty of Bioscience Engineering at Ghent University for their helpful conversations related to the statistical topics in this work.

Funding

This research was funded by the Flemish Government (AI Research program).

Supplementary Material

Supplementary Material for "Bayesian Optimal Design for Occupancy Modelling Using Integrated Presence-Only and Presence-Absence Data" (DOI: 10.1214/25-BA1581SUPP; .pdf). Supplementary results. This supplementary document includes additional figures and tables that may be helpful for readers to understand the work in further detail. Code and data. All code, data, and instructions to recreate the work are available at the GitHub repository https://github.com/saverymax/optimal-design-data-integration.

References

Ahmad Suhaimi, S. S., Blair, G. S., and Jarvis, S. G. (2021). "Integrated species distribution models: A comparison of approaches under different data quality scenarios." *Diversity and Distributions*, 27(6): 1066–1075. 2, 6, 15, 19

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., et al. (2010). "Spatial logistic regression and change-of-support in Poisson point processes." *Electronic Journal of Statistics*, 4: 1151–1201. MR2735883. doi: https://doi.org/10.1214/10-EJS581. 5

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC. MR3362184. 4

Belmont, J., Martino, S., Illian, J., and Rue, H. (2024). "Spatio-temporal occupancy models with INLA." *Methods in Ecology and Evolution*, 15(11): 2087–2100. 19

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability." *Monthly Weather Review*, 78(1): 1–3. 9

Clement, M. J. (2016). "Designing occupancy studies when false-positive detections occur." *Methods in Ecology and Evolution*, 7(12): 1538–1547. 3

Diggle, P. and Lophaven, S. (2006). "Bayesian Geostatistical Design." *Scandinavian Journal of Statistics*, 33(1): 53–64. MR2255109. doi: https://doi.org/10.1111/j.1467-9469.2005.00469.x. 8

Dorazio, R. M. (2014). "Accounting for imperfect detection and survey bias in statistical analysis of presence-only data." *Global Ecology and Biogeography*, 23(12): 1472–1484. 2, 6, 7

Dorazio, R. M., Gotelli, N. J., Ellison, A. M., Dorazio, R. M., Gotelli, N. J., and Ellison, A. M. (2011). "Modern Methods of Estimating Biodiversity from Presence-Absence Surveys." In *Biodiversity Loss in a Changing Planet*. IntechOpen. 1

Dorazio, R. M., Royle, J. A., Söderström, B., and Glimskär, A. (2006). "Estimating Species Richness and Accumulation by Modeling Species Occurrence and Detectability." *Ecology*, 87(4): 842–854. 8

Doser, J. W. and Stoudt, S. (2024). ""Fractional replication" in single-visit multi-season occupancy models: Impacts of spatiotemporal autocorrelation on identifiability." *Methods in Ecology and Evolution*, 15(2): 358–372. 2

Dovers, E., Popovic, G. C., and Warton, D. I. (2024). "A fast method for fitting integrated species distribution models." *Methods in Ecology and Evolution*, 15(1): 191–203. 2, 6, 19

eBird (2021). "eBird: An online database of bird distribution and abundance." 2, 16

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). "Bias correction in species distribution models: pooling survey and collection data for multiple species." *Methods in Ecology and Evolution*, 6(4): 424–438. 2, 6

Fithian, W. and Hastie, T. (2013). "Finite-sample equivalence in statistical models for presence-only data." *The Annals of Applied Statistics*, 7(4): 1917–1939. MR3161707. doi: https://doi.org/10.1214/13-AOAS667. 5

Fletcher Jr., R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M. (2019). "A practical guide for combining data to model species distributions." *Ecology*, 100(6): e02710. 2

Gabry, J., Češnovar, R., and Johnson, A. (2022). "cmdstanr: R Interface to 'CmdStan'." 10

Gneiting, T. and Raftery, A. E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, 102(477): 359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 9

Guillera-Arroita, G. and Lahoz-Monfort, J. J. (2017). "Species occupancy estimation and imperfect detection: shall surveys continue after the first detection?" *AStA Advances in Statistical Analysis*, 101(4): 381–398. MR3712405. doi: https://doi.org/10.1007/s10182-017-0292-5. 3

Guillera-Arroita, G., Ridout, M. S., and Morgan, B. J. T. (2010). "Design of occupancy studies with imperfect detection." *Methods in Ecology and Evolution*, 1(2): 131–139. MR4435306. 3, 7

Guillera-Arroita, G., Ridout, M. S., and Morgan, B. J. T. (2014). "Two-Stage Bayesian Study Design for Species Occupancy Estimation." *Journal of Agricultural, Biological, and Environmental Statistics*, 19(2): 278–291. MR3257915. doi: https://doi.org/10.1007/s13253-014-0171-4. 3, 7

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., et al. (2021). "Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions." *Diversity and Distributions*, 27(7): 1265–1277. 16

Johnston, A., Matechou, E., and Dennis, E. B. (2023). "Outstanding challenges and future directions for biodiversity monitoring using citizen science data." *Methods in Ecology and Evolution*, 14(1): 103–116. 3, 5

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., and Stone, L. (2017). "Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection." *Methods in Ecology and Evolution*, 8(4): 420–430. 2, 3, 6

Kéry, M. and Andrew Royle, J. (2016). *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS*, volume 1. Academic Press, 1 edition. MR3616659. 1, 2

Leach, C. B., Williams, P. J., Eisaguirre, J. M., Womble, J. N., Bower, M. R., and Hooten, M. B. (2022). "Recursive Bayesian computation facilitates adaptive optimal design in ecological studies." *Ecology*, 103(2): e03573. 11, 20

Liu, J. and Vanhatalo, J. (2020). "Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process." *Spatial Statistics*, 35: 100392. MR4024741. doi: https://doi.org/10.1016/j.spasta.2019.100392. 3, 8, 9

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). "Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One." *Ecology*, 83(8): 2248–2255. 6, 7

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. (2018). *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, 2 edition. 7

Mackenzie, D. I. and Royle, J. A. (2005). "Designing occupancy studies: general advice and allocating survey effort." *Journal of Applied Ecology*, 42(6): 1105–1114. 1, 3

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). "The recent past and promising future for data integration methods to estimate species' distributions." *Methods in Ecology and Evolution*, 10(1): 22–37. 2, 6

Mondain-Monval, T., Pocock, M., Rolph, S., August, T., Wright, E., and Jarvis, S. (2024). "Adaptive sampling by citizen scientists improves species distribution

model performance: A simulation study." *Methods in Ecology and Evolution*, 15(7): 1206–1220. 3, 20

Overstall, A. M. and Woods, D. C. (2017). "Bayesian Design of Experiments Using Approximate Coordinate Exchange." *Technometrics*, 59(4): 458–470. MR3740963. doi: https://doi.org/10.1080/00401706.2016.1251495. 19

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(3): 285–293. 3

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., et al. (2017). "Integrating multiple data sources in species distribution modeling: a framework for data fusion*." *Ecology*, 98(3): 840–850. 2, 19

Peel, S. L., Hill, N. A., Foster, S. D., Wotherspoon, S. J., Ghiglione, C., and Schiaparelli, S. (2019). "Reliable species distributions are obtainable with sparse, patchy and biased data by leveraging over species and data types." *Methods in Ecology and Evolution*, 10(7): 1002–1014. 2, 19

R Core Team (2022). "R: A language and environment for statistical computing." manual, R Foundation for Statistical Computing, Vienna, Austria. 10

Reich, B. J., Pacifici, K., and Stallings, J. W. (2018). "Integrating auxiliary data in optimal spatial design for species distribution modelling." *Methods in Ecology and Evolution*, 9(6): 1626–1637. 3, 11

Reich, H. T. (2020). "Optimal sampling design and the accuracy of occupancy models." *Biometrics*, 76(3): 1017–1027. MR4151868. doi: https://doi.org/10.1111/biom.13203. 3, 4

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., et al. (2015). "Point process models for presence-only analysis." *Methods in Ecology and Evolution*, 6(4): 366–379. 5

Royle, J. (2002). "Exchange algorithms for constructing large spatial designs." *Journal of Statistical Planning and Inference*, 100(2): 121–134. MR1877182. doi: https://doi.org/10.1016/S0378-3758(01)00127-6. 8

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). "A Review of Modern Computational Algorithms for Bayesian Optimal Design." *International Statistical Review*, 84(1): 128–154. MR3491282. doi: https://doi.org/10.1111/insr.12107. 8, 9

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(4): 860–870. 1, 3

Savery, M. and Luca, S. (2026). "Supplementary material for "Bayesian optimal design for occupancy modelling using integrated presence-only and presence-absence data"." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1581SUPP. 5

Sicacha-Parada, J., Steinsland, I., Cretois, B., and Borgelt, J. (2021). "Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway." *Spatial Statistics*, 42: 100446. MR4233261. doi: https://doi.org/10.1016/j.spasta.2020.100446. 19

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., and O'Hara, R. B. (2020). "Is more data always better? A simulation study of benefits and limitations of integrated distribution models." *Ecography*, 43(10): 1413–1422. 2, 6, 15, 19

Specht, H. M., Reich, H. T., Iannarilli, F., Edwards, M. R., Stapleton, S. P., Weegman, M. D., et al. (2017). "Occupancy surveys with conditional replicates: An alternative sampling design for rare species." *Methods in Ecology and Evolution*, 8(12): 1725–1734. 3

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(10): 2282–2292. 16

Thilan, A. W. L. P., Peterson, E., Menéndez, P., Caley, J., Drovandi, C., Mellin, C., et al. (2024). "Bayesian design methods for improving the effectiveness of ecosystem monitoring." *Environmental and Ecological Statistics*, 31(4): 893–919. 11, 20

Van Eupen, C., Maes, D., Herremans, M., Swinnen, K. R., Somers, B., and Luca, S. (2021). "The impact of data quality filtering of opportunistic citizen science data on species distribution model performance." *Ecological Modelling*, 444: 109453. 19

van Strien, A. J., van Swaay, C. A., and Termaat, T. (2013). "Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models." *Journal of Applied Ecology*, 50(6): 1450–1458. 2

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G., and Bower, M. R. (2018). "Monitoring dynamic spatio-temporal ecological processes optimally." *Ecology*, 99(3): 524–535. 3, 20

Yoo, E.-H., Zammit-Mangion, A., and Chipeta, M. G. (2020). "Adaptive spatial sampling design for environmental field prediction using low-cost sensing technologies." *Atmospheric Environment*, 221: 117091. 3