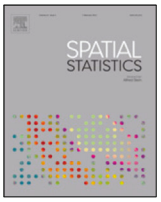
Spatial Statistics 55 (2023) 100756

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Data fusion of distance sampling and capture-recapture data

Narmadha M. Mohankumar [a,*], Trevor J. Hefley [a], Katy M. Silber [b], W. Alice Boyle [b]

[a] *Department of Statistics, Kansas State University, 1116 Mid-Campus Drive North, Manhattan, KS 66506, USA*
[b] *Division of Biology, Kansas State University, 116 Ackert Hall, Manhattan, KS 66506, USA*

ARTICLE INFO

Article history:
Received 8 March 2022
Received in revised form 29 April 2023
Accepted 1 May 2023
Available online 11 May 2023

Keywords:
Species distribution model
Hierarchical model
Inhomogeneous poisson point process
Missing data
Abundance modeling
Species–habitat relationship

ABSTRACT

Species distribution models (SDMs) are increasingly used in ecology, biogeography, and wildlife management to learn about the species–habitat relationships and abundance across space and time. Distance sampling (DS) and capture-recapture (CR) are two widely collected data types to learn about species–habitat relationships and abundance; still, they are seldomly used in SDMs due to the lack of spatial coverage. However, data fusion of the two data sources can increase spatial coverage, which can reduce parameter uncertainty and make predictions more accurate, and therefore, can be used with SDM. We developed a model-based approach for data fusion of DS and CR data. Our modeling approach accounts for two common missing data issues: 1) individuals that are missing not at random (MNAR) and 2) partially missing location information. Using a simulation experiment, we evaluated the performance of our modeling approach and compared it to existing approaches that use ad-hoc methods to account for missing data issues. Our results show that our approach provides unbiased parameter estimates with increased efficiency compared to the existing approaches. Finally, we demonstrated our approach using data collected for Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

Published by Elsevier B.V.

* Corresponding author.
E-mail address: meenu@ksu.edu (N.M. Mohankumar).

https://doi.org/10.1016/j.spasta.2023.100756
2211-6753/Published by Elsevier B.V.

1. Introduction

Species distribution models (SDMs) are widely used in ecology, biogeography, and wildlife management to learn about species–habitat relationships and estimate abundance across geographic space and time. Inference and predictions from SDMs are increasingly used to inform conservation management (Araujo and Guisan, 2006; Kéry and Royle, 2015; Hefley and Hooten, 2016; Koshkina et al., 2017). For example, conflicts between sustaining human activities and preserving biological diversity can be understood by identifying species–habitat relationships across space and time (e.g., Hefley et al., 2015). The SDMs are fitted to geo-referenced observations on species such as presence-only, presence-absence, count, distance sampling, and capture-recapture data. Spatially referenced covariates such as elevation, rainfall, soil properties, and vegetation characteristics are used in SDMs to enable statistical inference on species–habitat relationships and obtain spatially heterogeneous abundance estimates (Kéry and Royle, 2015).

Distance sampling (DS) and capture-recapture (CR) are two classic types of planned surveys that collect geo-referenced observations on species. The DS data are collected by recording distances to an individual in the study area from a point or transect (Burnham et al., 1980; Burnham and Anderson, 1984; Buckland et al., 2001). The CR data are collected by capturing an individual in the study area, which involves physically capturing the individual using a trap (e.g., mist nets) or taking a picture (e.g., camera traps; Otis et al., 1978; Seber, 1982; Pollock et al., 1990). The CR data often contain individual identification where DS data do not. There is a long history of collecting these two types of high-quality planned survey data in the field of ecology and wildlife management. However, DS and CR data are seldomly used in SDMs due to the large amount of effort and cost required to collect data that densely covers a large study area (McShea et al., 2016). These two data sources alone may suffer from the lack of spatial coverage, but fusion of the two data sources can increase spatial coverage, which can reduce parameter uncertainty (see section 25.1 in Hooten and Hefley, 2019). Therefore, it appears that a fused SDM of DS and CR can provide more precise statistical inference and predictions regarding the species distribution and abundance than using any of the data sources alone (see section 25.1 in Hooten and Hefley, 2019).

Construction of an adequate fused data SDM for DS and CR data relies upon accounting for missing data issues that are unique to each source of data. Failure to properly account for missing data issues may lead to misleading inferences and predictions from the SDMs (Little, 1992; Kéry, 2011; Dorazio, 2012; Hefley et al., 2013). For example, missing or partially missing data can produce biased parameter estimates that may invert the inferred species–habitat relationship, which is a critical consequence when making decisions in conservation management (Hefley et al., 2014, 2017). Missing data literature is well equipped with the statistical theory and tools to account for missing data issues which can be applied to SDMs (Rubin, 1976; Little, 1992; Mason et al., 2012; Little and Rubin, 2019), but such tools are rarely explicitly employed in SDM literature (Hefley et al., 2013), and approaches to properly account for missing data issues in SDMs are lacking.

Two of the common missing data issues in DS, and CR data are individuals that are missing not at random (MNAR) (Little and Rubin, 2019) and partially missing location information. The MNAR individuals can occur because of two reasons: (1) limited spatial coverage of the data due to limited accessibility, large amount of effort and cost, researcher preferences, or previous knowledge regarding the locations of individuals, or (2) the individuals in a sampled geographic region being unobserved due to the distance to the individual from the point, transect or the trap, observer's experience level, or environmental or geographical features that may obstruct detections or captures. The partially missing location information occurs when DS and CR only record partial information of the locations of individuals in contrast to complete location information (e.g., the exact geographic coordinates of the locations of the individuals). Such partially recorded location information makes spatial covariates unrecoverable because the spatial covariate values are usually obtained from a geographic information system that requires the individuals' exact locations. For example, DS surveys only record the point or the transect from which the individual was detected and the distance from the point or the transect to the individual, but do not record the exact location of the individual. As another example, CR surveys often use tools to attract the individuals to the trap, which results in the original, natural locations of the individuals being unrecoverable

because only the locations of the traps are recorded (Gerber et al., 2012; Williams and Boyle, 2018). Therefore, the spatial covariate values at the locations of the individuals that may influence the species distribution cannot be obtained.

The missing individuals that are MNAR are implicitly addressed by many DS and CR model developments using thinned point process models (e.g., Johnson et al., 2010; Borchers et al., 2015; Fletcher et al., 2019; Farr et al., 2020; Sicacha-Parada et al., 2021). Many of these developments use an inhomogeneous Poisson point process (IPPP) which can accommodate spatial inhomogeneity (Diggle et al., 1976; Cressie, 1993; Kéry and Royle, 2015) and enable inferences on the species–habitat relationship and abundance (Warton and Shepherd, 2010; Renner et al., 2015; Hefley and Hooten, 2016). However, the crux of applying existing IPPP-based approaches for DS and CR data is that they may not explicitly address the missing data issues that are unique to DS and CR data. For example, the approaches may require complete location information of the individuals; however, DS and CR data often contain only partial location information. In practice, researchers use ad-hoc methods to circumvent the limitation of partially recorded locations of individuals and fit the models. For example, Fletcher et al. (2019) transformed the DS data to presence-absence data at sampling sites using change of support and fitted the model to the transformed data. For another example, Farr et al. (2020) treated DS data as count data by defining sampling sites and counting the number of detected individuals in each site and fitted the model to count data. Both of these approaches do not require complete location information, because the models are fitted to spatially aggregated DS data, which are presence-absence or count data at sampling sites. As another example, Borchers et al. (2015) proposed an IPPP-based unified model for DS and CR data, where they used a homogeneous point process in all of their applications, but not an inhomogeneous point process. The homogeneous case contains a constant intensity function, therefore, not having complete location information of the individuals is not an issue. The (Borchers et al., 2015) model, however, is not designed to map a species distribution because of the constant intensity function. To model the species distribution, the model needs to be implemented using an inhomogeneous point process. However, the intensity function in an inhomogeneous point process typically depends on spatially referenced covariates, where the complete location information of the individuals is critical. Therefore, it is critical to account for the partially recorded location information of the individuals in the data. Hefley et al. (2020) proposed a model-based approach to account for the partially recorded location information in DS data and fit an IPPP-based model to the data. However, their model is merely constructed for DS data, and a subsequent model that accounts for the partial location information in CR data is lacking.

In addition to properly accounting for missing data issues, constructing a fused data SDM requires adequate model representations for DS and CR data that facilitate data fusion. A fused data SDM utilizes information from multiple data sources to reduce the uncertainty associated with limitations in each data source, hence improving the model predictions and inferences (Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017; Fletcher et al., 2019; Hooten and Hefley, 2019; Miller et al., 2019; Farr et al., 2020; Isaac et al., 2020; Martino et al., 2021). However, the model representations for DS and CR data presented in existing IPPP-based modeling approaches cannot be adequately used for data fusion of DS and CR data. For example, the unified model proposed by Borchers et al. (2015) represented the model for DS data based on the locations of the individuals and represented the model for CR data based on home range centers which are hypothetical centroids for individuals' activity. The locations of home range centers in CR data are irreconcilable with the locations of the individuals in DS data. For example, the model fitted for CR data would estimate the intensity of home range centers, and the model fitted for DS data would estimate the intensity of the locations of the individuals. Therefore, building a fused data SDM where both data sources share parameters in the underlying IPPP targeting the same inference is not achievable.

The existing IPPP-based modeling approaches for data fusion largely use spatial aggregation of individual-level data, so that explicit model representations for each data source are not required. Spatial aggregation involves partitioning the study area into nonoverlapping partitions (i.e., sampling sites) and transforming the locations of the individuals to counts or presence-absence data in each of the partitions (e.g., Fletcher et al., 2019; Farr et al., 2020). However, a significant drawback of spatial aggregation is determining the spatial resolution for the partitions. The data collected from

the surveys should be a representative sample of these partitions, and if the spatial resolution of the partitions does not adequately represent the sampled regions from the surveys, the model may yield biased parameter estimates.

We propose a hierarchical modeling framework that provides adequate model representations for DS and CR data that share parameters in the underlying IPPP targeting equivalent inference regarding species–habitat relationship and abundance; therefore facilitating data fusion of the two data sources. We use theory and tools from the missing data literature to build models for the missing data mechanism and account for the missing data issues that are unique to each data source. Our modeling framework can be viewed as a unified framework that can be extended to many other data sources (e.g., presence-only data) and a fusion of them to address critical issues with missing data. In our work, we propose two fused data SDMs for DS and CR data, one SDM incorporating the recorded distances from DS data and the other SDM without incorporating the recorded distances. We compare the two SDMs and investigate the efficiency gain of the estimated parameters by incorporating recorded distances, which provide additional information regarding the observed locations of the individuals. We conduct a simulation experiment to evaluate the performance of our two SDMs compared to existing IPPP-based approaches that use spatial aggregation. We assess the accuracy and the efficiency of the estimated parameters for the species–habitat relationship and obtain an estimate for the expected abundance in the study area. Finally, we demonstrate the approaches using data collected for Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas.

2. Materials and methods

2.1. Hierarchical modeling framework

Our proposed fused data SDM relies on a hierarchical modeling framework that is based on an IPPP. The models for the observed DS and CR data are conditioned on a common underlying IPPP that represents the underlying point pattern of individuals in the study area.

2.1.1. The underlying IPPP

The underlying IPPP describes the random number and the locations of individuals across the study area based on a continuous inhomogeneous intensity function, a function of spatially referenced covariates (e.g., elevation, temperature, soil attributes, vegetation, etc.). The intensity describes the expected number of individuals per infinitely small unit area and is usually defined as $\lambda(\mathbf{s}) = e^{\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}}$, a non-negative integrable function, where, $\mathbf{s}$ represents a vector containing coordinates of a location within the study area $\mathcal{S}$, $\mathbf{x}(\mathbf{s}) \equiv (1, x_1(\mathbf{s}), x_2(\mathbf{s}), \ldots, x_q(\mathbf{s}))'$, and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, \ldots, \beta_q)'$. The $x_1(\mathbf{s}), x_2(\mathbf{s}), \ldots, x_q(\mathbf{s})$ represent the spatial covariates at the location $\mathbf{s}$, where $x_1(\mathbf{s}), x_2(\mathbf{s}), \ldots, x_q(\mathbf{s})$ is observed for all $\mathbf{s} \in \mathcal{S}$. The β_0 represents the intercept parameter, and $\beta_1, \beta_2, \ldots, \beta_q$ represent the regression coefficients associated with the species–habitat relationship. Using the above notation, the probability distribution function (PDF) for the IPPP can be written as Cressie (1993)

$$[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N, N|\lambda(\mathbf{s})] = \frac{e^{-\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}}(\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s})^N}{N!} \times N! \prod_{i=1}^{N} \frac{\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}}, \tag{1}$$

where $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ are the locations of all N individuals (missing and observed) in the study area $\mathcal{S}$ (i.e., $\mathbf{u}_i \in \mathcal{S}$). A property of IPPP is that an estimate of the expected abundance in any sub-region $\mathcal{B}$ in the study area can be represented by $\bar{\lambda} = \int_{\mathcal{B}} e^{\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}} d\mathbf{s}$.

2.1.2. Accounting for missing individuals that are MNAR

The missing individuals that are MNAR can be accounted for by identifying and modeling the missing data mechanism. To model the missing data mechanism, we can label the random locations of all individuals in the study area as missing or observed (Gelfand and Schliep, 2018). We can define a vector $\mathbf{m} = (m(\mathbf{u}_1), m(\mathbf{u}_2), \ldots, m(\mathbf{u}_N))'$, where $m(\mathbf{u}_i)$ labels the ith individual as missing (i.e., zero)

or observed (i.e., one). Employing the missing data mechanism, we can write the distribution of $m(\mathbf{u}_i)$ as a zero-inflated Bernoulli distribution conditioned on $\mathbf{u}_i$.

$$[m(\mathbf{u}_i)|\mathbf{u}_i, q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} q(\mathbf{u}_i)^{m(\mathbf{u}_i)}(1-q(\mathbf{u}_i))^{1-m(\mathbf{u}_i)} & \text{, if } r(\mathbf{u}_i)=1 \\ 0 & \text{, if } r(\mathbf{u}_i)=0 \end{cases}, \tag{2}$$

where, $q(\mathbf{u}_i)$ denote the probability of observing the individual in a sampled region, $r(\mathbf{u}_i) = 1$ denotes that the $\mathbf{u}_i$th location is sampled within the study area, and $r(\mathbf{u}_i) = 0$ denotes that the $\mathbf{u}_i$th location is not sampled within the study area. The sampled region refers to a subset of the study area from which the data is collected. It is assumed that any individuals present in unsampled regions cannot be observed. The functional form of $q(\mathbf{s})$ and $r(\mathbf{s})$ at a location $\mathbf{s}$ can be defined based on the missing data mechanism.

By using the distribution of $m(\mathbf{u}_i)$, we can derive the PDF for the location of the ith individual conditioned on the label $m(\mathbf{u}_i)$ as

$$[\mathbf{u}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} \frac{q(\mathbf{u}_i)^{m(\mathbf{u}_i)}(1-q(\mathbf{u}_i))^{1-m(\mathbf{u}_i)}\lambda(\mathbf{u}_i)}{\int_S q(\mathbf{s})^{m(\mathbf{s})}(1-q(\mathbf{s}))^{1-m(\mathbf{s})}\lambda(\mathbf{s})d\mathbf{s}} & \text{, if } r(\mathbf{u}_i)=1 \\ 0 & \text{, if } r(\mathbf{u}_i)=0 \end{cases}. \tag{3}$$

Many recent model-based approaches based on IPPP use the so-called thinned IPPP (Diggle et al., 1976; Chakraborty et al., 2011; Cressie, 1993; Kéry and Royle, 2015), an implicit representation of the data to account for missing individuals as opposed to the complete distributional representation in (3).

2.1.3. Accounting for partially missing location information

The distributional representation in (3) accounts for the missing individuals that are MNAR; however, it does not account for the partially missing location information of observed individuals. The model requires complete location information of the individuals. We propose two models to account for the partially observed location information in data; (1) a model without incorporating the recorded distances from DS, and (2) a model incorporating the recorded distances from DS.

The DS and CR surveys each contain a sampled region in the study area which is a region surrounding the points, transects, or traps where the probability of detection or capture is greater than zero. We denote this region as the detection/capture region. In our first proposed model, we assume that the observed location of an individual is uniformly distributed in the detection/capture region that surrounds the point, transect, or trap the individual was detected or captured. Under this assumption, we can write the PDF of the observed location of the ith individual conditioned on the actual location of the individual as

$$[\mathbf{y}_i|\mathbf{u}_i] = \begin{cases} |A_{u_i}|^{-1} I(\mathbf{y}_i \in A_{u_i}) & \text{, if } m(\mathbf{u}_i)=1 \\ 0 & \text{, if } m(\mathbf{u}_i)=0 \end{cases}, \tag{4}$$

where, $\mathbf{y}_i$ denote the observed location of the ith individual, $\mathbf{u}_i$ is the actual location of the ith individual, A_{u_i} is the detection/capture region surrounding the point, transect or the trap where the individual was detected or captured, and $|A_{u_i}|$ is the area of the detection/capture region. Here, $\mathbf{y}_i$ and $\mathbf{u}_i$ are different because there is uncertainty associated with the observed locations of the individuals, and DS and CR data do not have complete location information of the individuals.

We then propose a second model by incorporating the recorded distances from DS data. We expect that adding additional information regarding the observed locations of the individuals may increase the efficiency of the model parameter estimates. Hefley et al. (2020) account for the partial location information in DS data by incorporating the recorded distances. Based on their approach, and under the assumption that the distances are recorded perfectly, we can assume that the observed location of an individual from a transect is uniformly distributed along the parallel lines to the transect (L_{u_i}) with a perpendicular distance that is equal to the recorded distance d_i. Under this assumption, we can write the PDF of the observed location of the ith individual conditioned on the actual location of the individual as

$$[\mathbf{y}_i|\mathbf{u}_i] = \begin{cases} |L_{u_i}|^{-1} I(\mathbf{y}_i \in L_{u_i}) & \text{, if } m(\mathbf{u}_i)=1 \\ 0 & \text{, if } m(\mathbf{u}_i)=0 \end{cases}. \tag{5}$$

For a point, L_{u_i} is the perimeter of the circle, where the radius is equal to the recorded distance, d_i. The $|L_{u_i}|$ is the length of the lines or the length of the perimeter of the circle.

2.2. Model implementation

The distributions in (4) and (5) represent the observed location of the ith individual conditioned on the actual location of the observed individual, $\mathbf{u}_i$; however, the actual location of the observed individual is of little interest in our study. Therefore, we can remove $\mathbf{u}_i$ from the model by integrating the joint PDF of $\mathbf{y}_i$ and $\mathbf{u}_i$. For all $\mathbf{s} \in \mathcal{S}$, the resulting PDFs representing the observed location of the ith individual are

$$[\mathbf{y}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} \frac{\int_{A_{u_i}} |A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)d\mathbf{u}_i}{\int_{\mathcal{S}} \lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} & , \text{if } r(\mathbf{u}_i) = 1 \ \& \ m(\mathbf{u}_i) = 1 \\ 0 & , otherwise \end{cases}, \tag{6}$$

$$[\mathbf{y}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} \frac{\int_{L_{u_i}} |L_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)d\mathbf{u}_i}{\int_{\mathcal{S}} \lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} & , \text{if } r(\mathbf{u}_i) = 1 \ \& \ m(\mathbf{u}_i) = 1 \\ 0 & , otherwise \end{cases}. \tag{7}$$

Moreover, our objectives in the study do not focus on estimating the locations of the unobserved individuals. Therefore, we can retain the PDF for the observed individual locations from (6) and (7) by setting $m(\mathbf{u}_i) = 1$. The resulting PDF is a simple marginal distribution that can be fitted using a likelihood-based or Bayesian approach. If practitioners are interested in estimating the locations of unobserved individuals, they can fit the model using a Bayesian hierarchical modeling approach from (3–5). Details associated with deriving our models are provided in the Supplementary Material.

2.3. Fused data SDM

The distributional representations in (6) and (7) can be used to construct a fused data SDM for DS and CR data. Our proposed distributional representations represent both DS and CR data based on observed locations of the individuals; therefore, the models share parameters in the underlying IPPP that target the same inference. We assume that the observed locations in the DS and CR data are independent across points, transects, and traps within and between the surveys. Representing DS and CR data using our proposed distributional representations and jointly modeling them leads to the following two fused data SDMs. The distribution in (8) does not incorporate the recorded distances from DS data, and the distribution in (9) incorporates the recorded distances.

$$\begin{aligned}[\mathbf{y}_1, \ldots, \mathbf{y}_{n_{ds}}, \mathbf{y}_{n_{ds}+1}, \ldots, \mathbf{y}_{n_{ds}+n_{cr}}, n_{ds}, n_{cr}|\lambda(\mathbf{s}), q_{ds}(\mathbf{s}), r_{ds}(\mathbf{s}), q_{cr}(\mathbf{s}), r_{cr}(\mathbf{s})] = \\ e^{-\int_{\mathcal{S}} \lambda(\mathbf{s})q_{ds}(\mathbf{s})I(r_{ds}(\mathbf{s})=1)d\mathbf{s} - \int_{\mathcal{S}} \lambda(\mathbf{s})q_{cr}(\mathbf{s})I(r_{cr}(\mathbf{s})=1)d\mathbf{s}} \times \\ \prod_{i=1}^{n_{ds}} \int_{A_{u_i}} |A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q_{ds}(\mathbf{u}_i)I(r_{ds}(\mathbf{u}_i) = 1)d\mathbf{u}_i \times \\ \prod_{i=n_{ds}+1}^{n_{ds}+n_{cr}} \int_{A_{u_i}} |A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q_{cr}(\mathbf{u}_i)I(r_{cr}(\mathbf{u}_i) = 1)d\mathbf{u}_i,\end{aligned} \tag{8}$$

$$\begin{aligned}[\mathbf{y}_1, \ldots, \mathbf{y}_{n_{ds}}, \mathbf{y}_{n_{ds}+1}, \ldots, \mathbf{y}_{n_{ds}+n_{cr}}, n_{ds}, n_{cr}|\lambda(\mathbf{s}), q_{ds}(\mathbf{s}), r_{ds}(\mathbf{s}), q_{cr}(\mathbf{s}), r_{cr}(\mathbf{s})] = \\ e^{-\int_{\mathcal{S}} \lambda(\mathbf{s})q_{ds}(\mathbf{s})I(r_{ds}(\mathbf{s})=1)d\mathbf{s} - \int_{\mathcal{S}} \lambda(\mathbf{s})q_{cr}(\mathbf{s})I(r_{cr}(\mathbf{s})=1)d\mathbf{s}} \times \\ \prod_{i=1}^{n_{ds}} \int_{L_{u_i}} |L_{u_i}|^{-1}\lambda(\mathbf{u}_i)q_{ds}(\mathbf{u}_i)I(r_{ds}(\mathbf{u}_i) = 1)d\mathbf{u}_i \times \\ \prod_{i=n_{ds}+1}^{n_{ds}+n_{cr}} \int_{A_{u_i}} |A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q_{cr}(\mathbf{u}_i)I(r_{cr}(\mathbf{u}_i) = 1)d\mathbf{u}_i,\end{aligned} \tag{9}$$

where, n_{ds} and n_{cr} are the number of detected and captured individuals from DS and CR respectively, $q_{ds}(\cdot)$ is the probability of detection from a point or transect which depends on the distance from the point or transect to the individual, $q_{cr}(\cdot)$ is the probability of capture from a trap, $r_{ds}(\mathbf{s})$ and $r_{cr}(\mathbf{s})$ are indicator functions defining the detection/capture regions of the DS and CR data respectively,

and $n = n_{ds} + n_{cr}$ is the total number of observed individuals from surveys. In our study, we define the probability of detection for DS data by a half-normal function, that is $q_{ds}(\mathbf{u}_i) = e^{-d_i^2/\phi}$, where, d_i is the distance between the point or transect and the ith detected individual, and ϕ is a scale parameter. To ensure the identifiability of the parameter estimates, the detection function is constrained on probability of detection being equal to one at distance of zero. The indicator function truncating the detection region from a point or transect is defined as, $r_{ds}(\mathbf{u}_i) = I(\mathbf{u}_i \in A_{ds})$, where A_{ds} is the detection region surrounding a point or transect where probability of detection is greater than zero. We define the probability of capture from a trap as $q_{cr}(\mathbf{u}_i) = \theta$. The indicator function truncating the capture region of a trap is defined as $r_{cr}(\mathbf{u}_i) = I(\mathbf{u}_i \in A_{cr})$, where A_{cr} is the capture region surrounding a trap where probability of capture is greater than zero.

In principle, including additional information regarding the observed individual locations ought to increase the efficiency of parameter estimates from a model. Therefore, we expect the fused data SDM in (9) to provide more efficient parameter estimates than the fused data SDM (8) since the SDM in (9) incorporates the recorded distances from DS data. We investigate this conjecture in both the simulation experiment and the data example that follows.

3. Simulation experiment

We conducted a simulation experiment to evaluate the performance of our two proposed fused data SDMs and compare them to standard IPPP-based approaches that use spatial aggregation. We assessed the performance of the models using the five scenarios listed below. Note that all the models in the scenarios below account for MNAR individuals and we assess the performances of the models in accounting for the partial location information.

1. The model from (3) is fitted to DS and CR data containing complete location information of the individuals.
2. The model proposed by Farr et al. (2020) for spatially aggregated data is fitted to DS and CR data containing partial location information of the individuals.
3. The model from (3) transformed for spatially aggregated data using change of support is fitted to DS and CR data containing partial location information of the individuals.
4. Our proposed fused data SDM from (8) that do not incorporate recorded distances is fitted to DS and CR data containing partial location information of the individuals.
5. Our proposed fused data SDM from (9) that incorporates recorded distances is fitted to DS and CR data containing partial location information of the individuals.

In our simulation experiment, we simulated a single spatial covariate, $x(\mathbf{s})$ using a reduced rank Gaussian process on a unit square study area (i.e., $\mathcal{S} = [0, 1] \times [0, 1]$, where $\mathbf{s} \in \mathcal{S}$). We generated the reduced rank Gaussian process by approximating the exponential covariance function with a low-rank approximation, using a variance of 1 and a length scale of 0.05 as parameters. We simulated the actual locations of the individuals using the IPPP represented by (1) with the intensity $\lambda(\mathbf{s}) = e^{\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}}$. We set the parameter values as $\beta_0 = 9, \beta_1 = 1, \theta = 0.2$, and $\phi = 0.025$. We placed 15 points and 65 traps in the study area to obtain DS and CR data, respectively (Fig. 1; panel a). We set non-overlapping detection/capture regions to ensure the independence of the observed data across surveys and within surveys (Fig. 1; panel c). We defined the detection region surrounding each point by assuming that the individual has to be within a maximum distance of 0.04 from the point to be detected. We defined the capture region surrounding each trap by defining that the individual has to be within a maximum distance of 0.02 from the trap to be attracted to the trap and get captured. We obtained spatially aggregated data required to fit the models in scenario 2 and scenario 3 by dividing the study area into 100 non-overlapping partitions and counting the number of observed individuals in each partition (Fig. 1; panel b). If a partition does not consist of a survey point or a trap, we defined the partition as an unsampled partition.

We fitted the models described in scenarios 1–5 to 1000 simulated data sets. We used the complete location information of the individuals in scenario 1, and the partial location information of the individuals in scenarios 2–5. Scenario 1 acts as the benchmark scenario since the data with

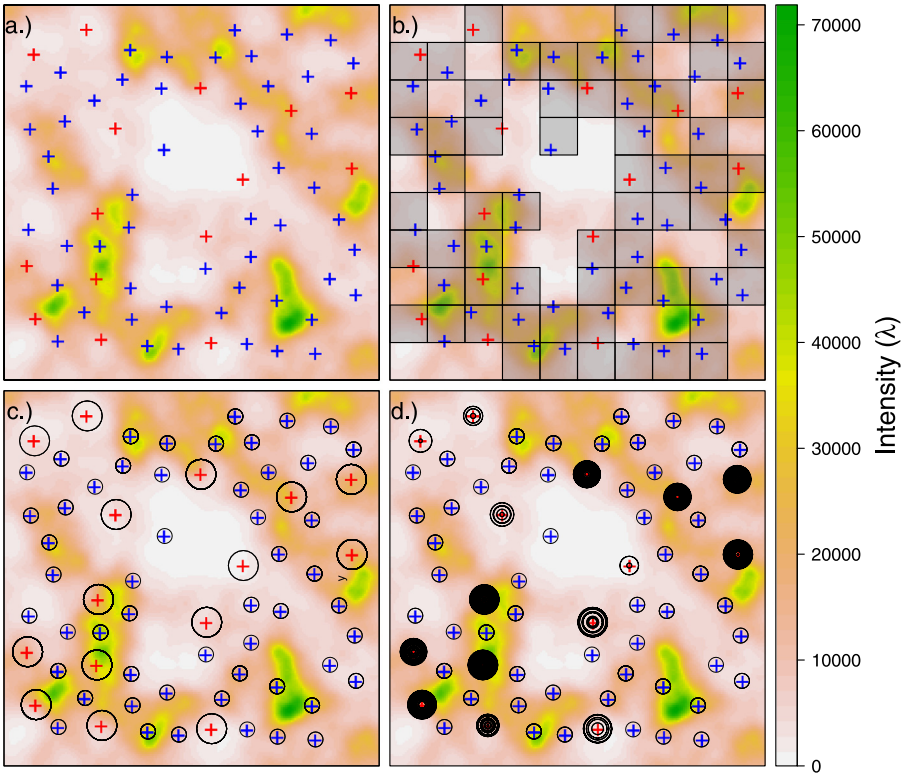


Fig. 1. Panel (a) displays the points (red +) and traps (blue +) placed in the study area to collect DS and CR data. Panel (b) shows the partitioning of the study area to obtain spatially aggregated DS and CR data (for scenario 2 and scenario 3). Spatially aggregated data are obtained by dividing the study area into 100 non-overlapping partitions and choosing the partitions that include a point or a trap. Panel (c) displays the detection and capture regions of DS and CR data (for scenario 4). Panel (d) displays the circle's perimeter surrounding the points, where the radius is equal to each individual's recorded distance (for scenario 5). Panel (d) also displays the capture regions of the traps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

complete location information matches the process described by the fitted model. We evaluated the performance of the models in scenarios 2–5 for data containing partial location information and compared them to the benchmark scenario 1. For each simulated data set, we obtained the parameter estimates for the intercept (β_0), the relationship to the spatial covariate (β_1), and the expected abundance ($\bar{\lambda}$). We assessed the reliability of the parameter estimates by calculating the coverage probabilities of the 95% Wald-type confidence intervals (CIs). We included side-by-side box plots to visually compare the empirical distributions of the parameter estimates. We obtained the relative efficiency of the parameter estimates under scenarios 2–5 with reference to the efficiency of parameter estimates obtained under benchmark scenario. The relative efficiency is calculated by dividing the standard deviation of the respective empirical distribution of the estimates by the standard deviation of the empirical distribution of the estimates under scenario 1.

The integrals in the likelihood functions and the integrated intensity function are approximated using numerical quadrature. We used the Nelder–Mead algorithm in R to numerically maximize the likelihoods and obtain the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The estimate for the expected abundance is obtained using $\hat{\bar{\lambda}} = \int_S e^{\mathbf{x}(\mathbf{s})'\hat{\boldsymbol{\beta}}} d\mathbf{s}$. We inverted the Hessian matrix to approximate the standard errors of the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and then calculated the 95% Wald-type CIs for $\hat{\beta}_0$ and $\hat{\beta}_1$. We approximated the standard error of the parameter estimate $\hat{\bar{\lambda}}$ using the delta method under

first-order Taylor expansion and then calculated 95% Wald-type CI for $\hat{\bar{\lambda}}$. We provide the annotated R code associated with the simulation experiment in the Simulation.R file in the supplementary material.

4. Grasshopper Sparrows at Konza Prairie Biological Station, Kansas

We illustrated our proposed models and the existing IPPP-based approaches using data on Grasshopper Sparrows (*Ammodramus savannarum*) from Konza Prairie Biological Station (KPBS). The KPBS is a long-term ecological research site in northeastern Kansas, comprised of native tallgrass prairie (Knapp et al., 1998; Williams and Boyle, 2018, 2019). Grasshopper Sparrows are a migratory grassland songbird species that winter in the southern United States and northern Mexico and breed throughout grasslands in the United States and southern Canada. However, the loss of prairie habitat has contributed to a long-term population decline in Grasshopper Sparrows (Herse et al., 2018). Therefore, identifying suitable habitats and investigating the abundance of Grasshopper Sparrow populations is essential for directing conservation efforts.

We used observations from the 2019 breeding season for our analysis. The data consist of 72 observations from 53 transects and 160 observations from 137 mist-net locations (Fig. 2; panel a). The transects were surveyed during the month of June as part of the long-term monitoring efforts of birds at the Konza Prairie. Within 24 experimentally-managed pastures, one to four 300 m long transects bisect the topographic gradients within the sampling site. A single observer slowly walks the transect, recording the individuals seen or heard on either side of the transect, with the distance to each individual (Boyle, 2019). The mist-nets were used to capture individuals during the entire breeding season from shortly after the adult male birds arrive in April until nests complete in August. The mist net locations were selected to maximize chances of capturing the adult male birds within their territories, and the birds were attracted to nets using a small speaker broadcasting a territorial song (Williams and Boyle, 2018).

Male adult birds sing territorial songs from conspicuous perches in suitable habitats and actively defend 0.5 ha territories from other male birds (Winnicki et al., 2020). Female birds select and build nests within the territories of male birds. Their behavior is very secretive, making them difficult to detect. Thus, both detections and captures consist of male adult birds only. Upon arrival, the male adult birds establish breeding territories at the site. These individual male adult birds may select territories based on many environmental cues such as vegetation, topography, location of conspecifics, and land management (Andrews et al., 2015; Shaffer et al., 2021). To illustrate our approach, we use elevation as the spatial covariate.

We illustrate our approach for DS and CR data using the detections from transects and captures from mist-nets. We assume that the individual has to be within a maximum distance of 150 m from the transect to be detected, which is realistic given the topography, song attenuation, and realized distance values (Fig. 2; panel c). For captures from mist-nets, we assume that the individual has to be within a maximum distance of 25 m to elicit a response and be attracted to the mist-net, a distance reasonable given the speaker volume and observed behavior of the species (Fig. 2; panel c). Furthermore, we assume that the observations from the transects and the mist-nets are independent within and between the surveys.

As in scenarios 2–5 in the simulation experiment, we fit the four models to the observed data: (1) the model proposed by Farr et al. (2020) for spatially aggregated data, (2) the model from (3) transformed for spatially aggregated data using change of support, (3) our proposed fused data SDM from (8) that do not incorporate recorded distances, and (4) our proposed fused data SDM from (9) that incorporates recorded distances. We obtain the spatially aggregated data by dividing the study area into non-overlapping partitions and counting observed individuals in each partition. The partitions are selected in a way that ensures they closely correspond to the sampled regions from transects and traps. If a partition does not consist of a transect or a mist net, we define the partition as an unsampled partition which led to 66 non-overlapping sampled partitions (Fig. 2; panel b).

Finally, we fit the models to the data and compare the maximum likelihood estimates and the corresponding 95% Wald-type CIs for β_0, β_1, and $\bar{\lambda}$. We provide the annotated R code associated with the data analysis in the Grasshopper_sparrows_data_example.R in the Supplementary Material.

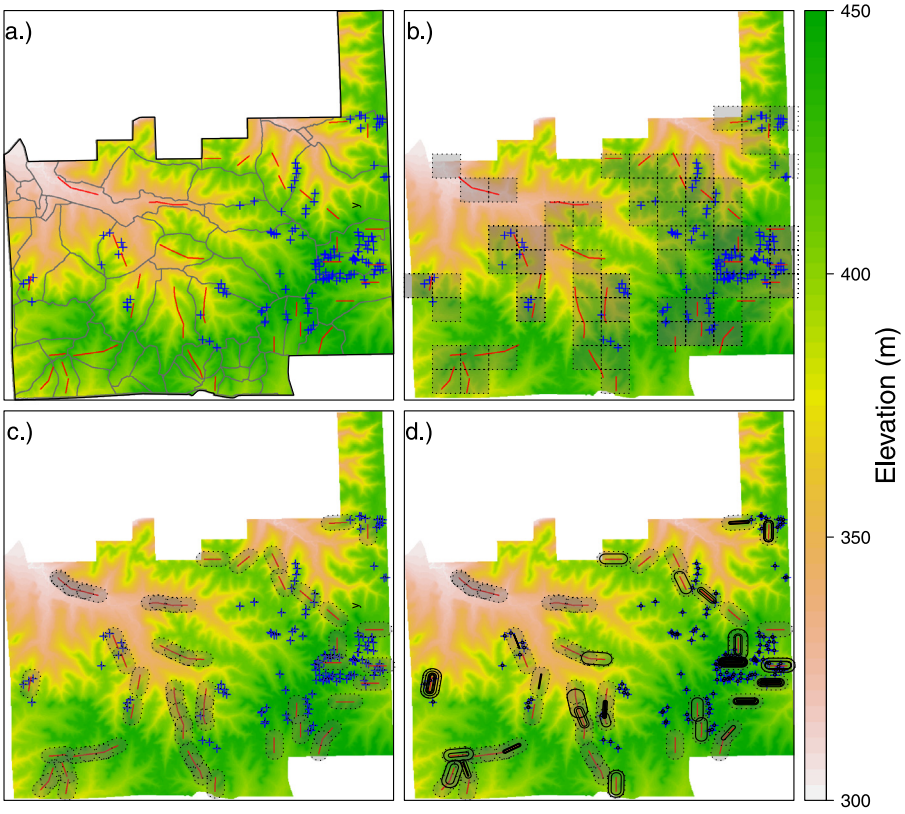


Fig. 2. Panel (a) displays the transects (red –) and mist nets (blue +) that are used to collect data on Grasshopper Sparrows at Konza Prairie Biological Station (KPBS). The surveys are conducted at watershed-level (gray – in panel (a)). Panel (b) shows the partitioning of the study area (66 partitions) to obtain spatially aggregated data (dashed line) to fit the two models; the model proposed by Farr et al. (2020) for spatially aggregated data, and the model from (3) transformed for spatially aggregated data using change of support. Panel (c) displays the detection and capture regions of transects and traps (dashed line) used for our proposed fused data SDM from (8) that do not incorporate recorded distances. Panel (d) displays the parallel lines to the transect with a perpendicular distance equal to each individual's recorded distance, which is used for our proposed fused data SDM from (9) that incorporates recorded distances. Panel (d) also displays the capture regions of the traps (dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Results

5.1. Simulation experiment

As expected, the benchmark scenario (i.e., scenario 1) yielded an unbiased estimate for β_0, with a coverage probability of the 95% CIs of 0.942. When the data contained partial location information, scenario 2 and scenario 3 yielded biased estimates for β_0, whereas scenario 4 and scenario 5 yielded unbiased estimates (see Fig. 3 for graphical comparison). The coverage probabilities of the 95% CIs for β_0 under scenarios 2–5 were 0.190, 0.180, 0.761, and 0.925, respectively. The relative efficiencies of estimates for β_0 obtained from scenarios 2–5 were 23.204, 15.949, 13.907, and 1.007, respectively. We noticed that the efficiency of the estimate for β_0 under scenario 5, almost reaches the efficiency obtained under the benchmark scenario 1 (see Table 1).

Similar to the parameter estimate for β_0, scenario 1 yielded an unbiased estimate for β_1 with a coverage probability of the 95% CIs, 0.948. However, when the data contained partial location

Table 1
Estimated coverage probability (CP) for the 95% confidence interval (CI) and the relative efficiency (RE) for the parameters β_0, β_1, and expected abundance ($\bar{\lambda}$) obtained under scenario 1, scenario 2 , scenario 3, scenario 4, and scenario 5 in the simulation experiment. The parameter estimates are obtained by fitting the models to 1000 simulated data sets.

Scenarios	β_0		β_1		$\bar{\lambda}$	
	CP	RE	CP	RE	CP	RE
Scenario 1	0.942	–	0.948	–	0.944	–
Scenario 2	0.190	23.204	0.749	1.891	0.343	265.921
Scenario 3	0.180	15.949	0.838	1.394	0.430	285.819
Scenario 4	0.761	13.907	0.942	1.089	0.783	141.896
Scenario 5	0.925	1.007	0.942	1.041	0.944	1.038

Table 2
Parameter estimates and the width of the 95% CIs for the intercept (β_0), the relationship between the abundance and elevation (β_1), and the log of the expected abundance ($\bar{\lambda}$) for Grasshopper Sparrows at Konza Prairie Biological Station, Kansas. The parameter estimates are obtained from the model proposed by Farr et al. (2020) for spatially aggregated data (Spatially aggregated: FARR), the model from (3) transformed for spatially aggregated data using change of support (Spatially aggregated: from (3)), our proposed fused data SDM from (8) that do not incorporate recorded distances (Fused SDM: from (8)), and our proposed fused data SDM from (9) that incorporates recorded distances (Fused SDM: from (9)).

Models	β_0		β_1		$\log(\bar{\lambda})$	
	$\hat{\beta}_0$	Width of 95% CI	$\hat{\beta}_1$	Width of 95% CI	$\log(\hat{\bar{\lambda}})$	Width of 95% CI
Spatially aggregated: FARR	−4.767	6.034	0.022	0.015	12.766	6.033
Spatially aggregated: from (3)	−4.751	5.616	0.012	0.015	12.669	5.619
Fused SDM: from (8)	−11.669	0.486	0.011	0.015	5.742	0.463
Fused SDM: from (9)	−11.663	0.484	0.010	0.015	5.743	0.463

information, scenario 2 and scenario 3 yielded biased estimates for β_1, whereas scenario 4 and scenario 5 yielded unbiased estimates for β_1 (see Fig. 3 for graphical comparison). The coverage probabilities of the 95% CIs for β_1 under scenarios 2–5 were 0.749, 0.838, 0.942 and 0.942, respectively. The relative efficiencies of estimates for β_1 obtained from scenarios 2–5 were 1.891, 1.394, 1.089, and 1.041, respectively (see Table 1).

Scenario 1 yielded an unbiased estimate for $\bar{\lambda}$ with a coverage probability of the 95% CIs, 0.944. When the data contained partial location information, scenario 4 and scenario 5 yielded unbiased estimates for $\bar{\lambda}$. The coverage probabilities of the 95% CIs for $\bar{\lambda}$ under scenarios 2–5 were 0.343, 0.430, 0.783, and 0.944, respectively. The relative efficiencies of the estimates for $\bar{\lambda}$ obtained from scenarios 2–5 were 265.921, 285.819, 141.896, and 1.038, respectively. We noticed that scenario 5 provides the most efficient parameter estimate for $\bar{\lambda}$, which nearly reaches the efficiency obtained under benchmark scenario 1 (see Table 1).

5.2. Grasshopper Sparrows at Konza Prairie Biological Station, Kansas

The estimates obtained for the intercept parameter (β_0) under our two proposed models were similar, with narrow 95% CIs. The models that use spatially aggregated data yielded similar estimates for β_0 but with approximately 12 times wider CIs than our proposed models (see Fig. 4; panel a, and 95% CIs in Table 2). The estimates obtained for β_1 under all four models yielded similar inference regarding the relationship between species abundance and elevation; however, the estimate for β_1 under the model proposed by Farr et al. (2020) was twice as large as the estimates obtained from the other models (see Fig. 4; panel b, and 95% CIs in Table 2). The crucial outcome from our fitted models is the estimates obtained for $\bar{\lambda}$. The models that use spatially aggregated data yielded unrealistic estimates for $\bar{\lambda}$ with an approximate 163000 times wider 95% CIs than our proposed models (see Fig. 4; panel c, and 95% CIs in Table 2). Altogether, the parameter estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\bar{\lambda}}$ from our proposed two models were similar and yielded narrower 95% CIs.

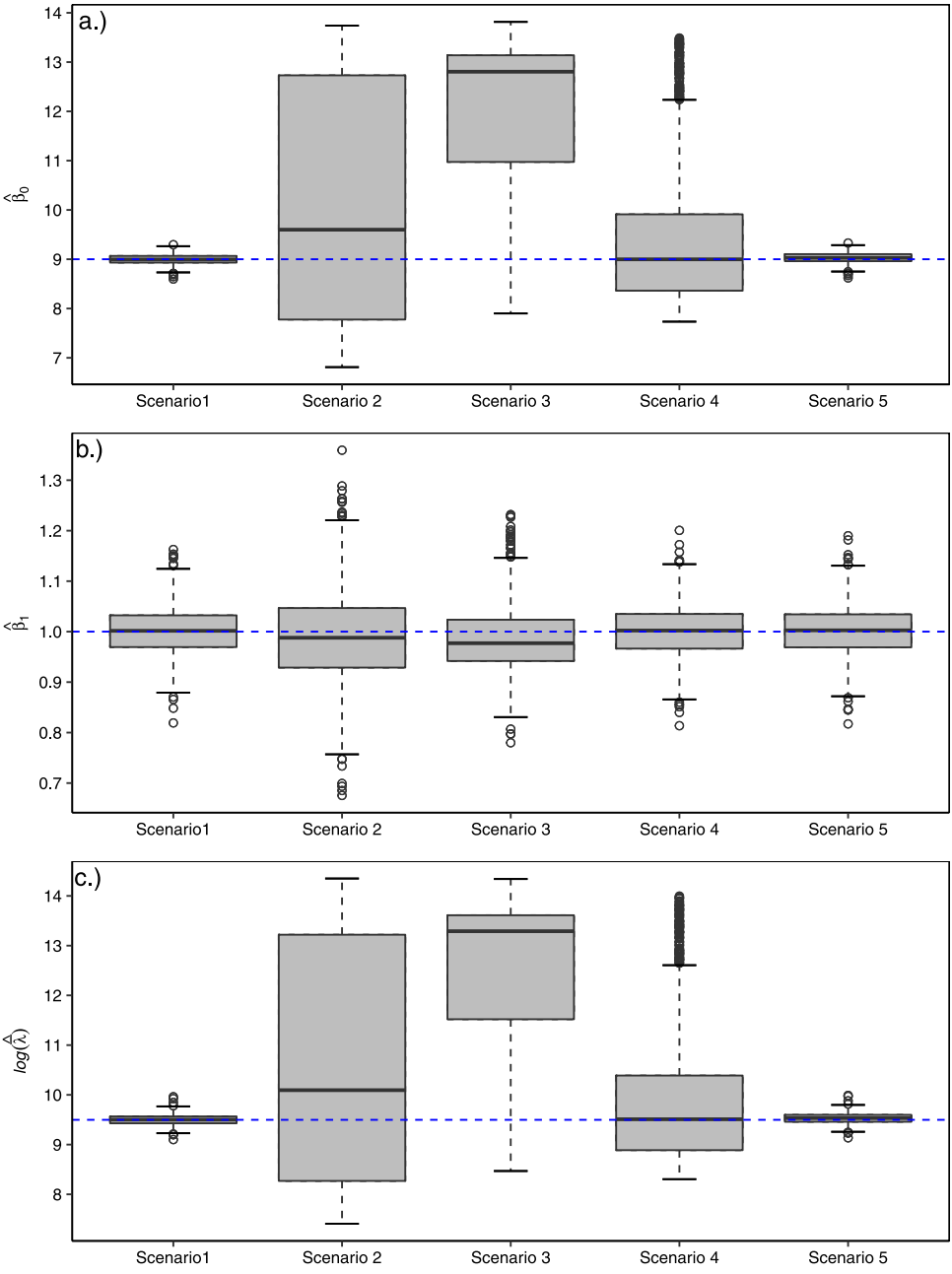


Fig. 3. The box plots display the estimates of parameters β_0 (panel a), β_1 (panel b), and $\log(\bar{\lambda})$ (panel c) obtained under scenarios 1–5 for 1000 simulated data sets. The true values of the parameters (β_0 = 9, β_1 =1, $\log(\bar{\lambda}) = 9.5$) are shown by the blue dash line (–). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

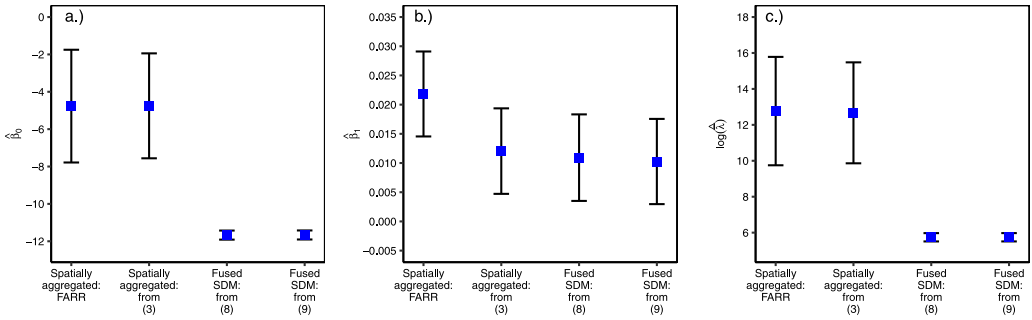


Fig. 4. Panel (a), panel (b), and panel (c) display the parameter estimates and the 95% CIs for the intercept (β_0), the relationship between the abundance and elevation (β_1), and the expected abundance ($\bar{\lambda}$) for Grasshopper Sparrows at Konza Prairie Biological Station, Kansas. The parameter estimates are obtained from the model proposed by Farr et al. (2020) for spatially aggregated data (Spatially aggregated: FARR), the model from (3) transformed for spatially aggregated data using change of support (Spatially aggregated: from (3)), our proposed fused data SDM from (8) that do not incorporate recorded distances (Fused SDM: from (8)), and our proposed fused data SDM from (9) that incorporates recorded distances (Fused SDM: from (9)). The parameter estimates are shown by the blue square (■), and the 95% CIs are shown by whisker ends. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Discussion

6.1. IPPP generalization for DS and CR data that enables data fusion

A critical aspect of data fusion of multiple data sources is providing model representations for the data types that target the same inference (i.e., have equivalent parameters). The existing point process models for DS data use individual location information to infer about species–habitat relationship and abundance (e.g., Johnson et al., 2010). In contrast, the existing point process models for CR explicitly use home range centers (e.g., Borchers et al., 2015). Modeling the distribution of home ranges of individuals using a particular area as its home range is different from modeling the individual locations. Therefore, the shared parameters in the underlying point process for the two data sources do not target the same inference. This incompatibility in the underlying process model may explain the lack of approaches for data fusion of DS and CR data.

Our proposed approach provides a generalization of Borchers et al. (2015)'s IPPP-based model with model representations for both DS and CR data based on the locations of the individuals. The focus of our study is on large-scale patterns of species–habitat relationships at the population level and the expected abundance in the study area, and not on individual movement patterns and habitat use within a home range. Therefore, the movement of individuals within a home range is less relevant and can be safely ignored. Our model representations for DS and CR data based on the locations of the individuals allowed the models to share parameters in the underlying process that target the same inference. Therefore, our approach facilitated data fusion enabling the use of these two types of high-quality planned survey data to obtain useful statistical inference regarding the species–habitat relationship, more accurate estimates for the expected abundance, and more accurate spatial maps for species distributions.

6.2. Improvement of inference regarding species–habitat relationship and estimate for the expected abundance by properly accounting for missing data issues

Efficiently acquiring reliable parameter estimates for both β_0 and β_1 is of utmost importance. However, many recent studies only attempt to improve the estimate of β_1, focusing on species–habitat relationships or relative abundance which is a measure of expected abundance relative to other species within a community. These approaches do not improve estimates of β_0. In contrast

to relative abundance, expected abundance plays a vital role in studying the dynamics of species populations, however, estimating the expected abundance depends on both β_0 and β_1. It is also important to note that a small deviation of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the true parameter value would significantly affect the estimate for the expected abundance due to the exponential function in the intensity function (i.e., $\hat{\lambda}(\mathbf{s}) = e^{\mathbf{x}(\mathbf{s})'\hat{\boldsymbol{\beta}}}$).

Our study shows that obtaining reliable, more efficient parameter estimates for β_0 and β_1 crucially relies upon properly accounting for the missing data issues. Our modeling framework explicitly acknowledges and accounts for the missing data issues unique to DS and CR data using theory and tools from missing data literature. Our results show that when the data contain partial location information, ad-hoc approaches such as spatial aggregation result in biased parameter estimates with poor efficiency, whereas, our proposed models provide reliable, more efficient parameter estimates than existing approaches that use spatial aggregation (see Table 1). Furthermore, our simulation experiment led to an important finding: the inclusion of additional information regarding individual locations into the model, such as recorded distances, led to significant efficiency gain in the parameter estimates. In fact, the efficiency surprisingly reaches the efficiency of the parameter estimates obtained under the benchmark scenario which contains complete location information.

In this paper, we present the simulation experiment for the parameter specifications $\beta_0 = 9, \beta_1 = 1, \theta = 0.2, \phi = 0.025$ and evaluate the performance of our modeling approach. However, we conducted the simulation experiment using other parameter choices as well and irrespective of the choice of the parameter values, our proposed models provided reliable, more efficient parameter estimates than existing approaches that use spatial aggregation.

6.3. A spatio-temporal fused data SDM

Our simulation experiment utilized non-overlapping detection/capture regions to ensure the independence of observations both within and across surveys. The independence assumption is often valid in real-world SDM applications for DS and CR data, as these typically involve DS and CR data collected over large, sparsely sampled spatial extents. A fused SDM involves combining such DS data and CR data from large different study areas that are often located hundreds of miles apart satisfying the independence assumption both within and across surveys. In our Grasshopper Sparrows data example involving KPBS ecological research site in northeastern Kansas, we assumed that the observations are independent across and within the surveys. However, it is worth noting that KPBS is likely a special case as one of the most intensively studied areas on earth with a relatively small spatial extent, and thus the observations may contain some lack of independence. We can strengthen the independence assumption by extending our model to a spatio–temporal model. A spatio–temporal model enables the modeling of species abundance patterns across both time and space. By using a continuous-space discrete-time model with short time periods, we can strengthen the independence assumption. The dependence of observations can be further addressed by adding a spatio–temporal random effect to the spatio–temporal model accounting for the spatio–temporal autocorrelation. Moreover, in cases where there are spatial and temporal patterns that are not explained by the covariates, adding a spatio–temporal random effect to the intensity function can improve the accuracy of predictions and inferences from the model. Numerous methods have been developed in the literature on SDM to model the spatial and spatio–temporal autocorrelation (e.g., Chakraborty et al., 2011; Renner et al., 2015; Mohankumar and Hefley, 2021). These methods can be leveraged to incorporate a spatial or a spatio–temporal random effect and expand our proposed modeling framework.

6.4. Detection and capture functions

In our study, we defined the probability of detection by a half-normal function of the distance between the point or the transect and the location of the individual. We defined the probability of capture as a constant parameter. However, the probability of detection can be defined by other functions such as uniform, hazard-rate, negative exponential, etc. Similarly, the probability

of capture can be defined as a function of covariates such as the observer's experience level or environmental or geographical features. Such extensions of the model enable identifying the factors that influence the probability of detection or capture.

It is possible that the parameters in the detection function or capture function are confounded with the parameters in the intensity function. For example, in a model in which the underlying intensity and the probability of capture are both functions of the same spatial covariate, the underlying point process is confounded with the capture process. For another example, if the underlying intensity function is a function of the distance from the transect, the underlying point process is confounded with the detection process. Accounting for such confounding of the underlying intensity and the detection/capture probability is an area that needs further research. In most situations, we can avoid such confounding during the design of the surveys.

6.5. Inclusion of the spatial and non-spatial covariates

The intensity function, probability of detection, and probability of capture can depend on many covariates that are spatial or non-spatial. For instance, in our Grasshopper sparrow data example, the practitioners may want to include "effort" to define the probability of detection, which is a non-spatial covariate, or they may want to include "vegetation", which is a spatial covariate. A non-spatial covariate that is measured during the survey can be easily incorporated into our model. However, for the spatial covariate, our approach requires the spatial covariate values for the entire study region. In most cases, they can be obtained from a geographical information system. However, obtaining the spatial covariate values in the entire study region can be trivial in some situations. In such situations, we can employ an auxiliary model (e.g., kriging) to utilize the available data to predict the spatial covariate values for the entire region and use the predicted values as the input values for the spatial covariate in our models.

Acknowledgments

We thank all individuals, including R. Donnelly, J. Gresham, K. Kersten, A. Mayers, E. Smith, and M. Winnerman, who contributed to the Grasshopper Sparrows data used in our data example. This material is based upon work supported by the National Science Foundation, United States under Grant No. 1754491 and the Konza Prairie Long-Term Ecological Research (LTER) Grant No. 1440484. This work was permitted by the US Geological Survey's Bird Banding Lab (23836), the Kansas Department of Parks, Wildlife, and Tourism, and the Kansas State University Institutional Animal Care and Use Committee (IACUC) (protocol 4250).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2023.100756. The supplementary material includes details associated with deriving our models. Annotated R codes, data, and additional files (e.g., shapefiles) that can be used to reproduce all results and figures associated with the simulation experiment, and the Grasshopper Sparrows data example are available in the supplementary material.

References

Andrews, J.E., Brawn, J.D., Ward, M.P., 2015. When to use social cues: Conspecific attraction at newly created grasslands. Condor Ornithol. Appl. 117, 297–305.

Araujo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.

Borchers, D.L., Stevenson, B., Kidney, D., Thomas, L., Marques, T.A., 2015. A unifying model for capture–recapture and distance sampling surveys of wildlife populations. J. Amer. Statist. Assoc. 110, 195–204.

Boyle, A., 2019. CBP01 variable distance line-transect sampling of bird population numbers in different habitats on Konza Prairie. http://dx.doi.org/10.6073/pasta/053fe6a82e54394a70ff22b4794c0489, (Accessed 16 December 2021).

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., Thomas, L., 2001. Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press, Oxford, United Kingdom.

Burnham, K.P., Anderson, D.R., 1984. The need for distance data in transect counts. J. Wildl. Manage. 48, 1248–1254.

Burnham, K.P., Anderson, D.R., Laake, J.L., 1980. Estimation of density from line transect sampling of biological populations. Wildl. Monogr. 72, 3–202.
Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M., Silander, J.A., 2011. Point pattern modelling for degraded presence-only data over large regions. J. R. Stat. Soc. Ser. C. Appl. Stat. 60, 757–776.
Cressie, N., 1993. Statistics for Spatial Data. John Wiley & Sons, New York.
Diggle, P.J., Besag, J., Gleaves, J.T., 1976. Statistical analysis of spatial point patterns by means of distance methods. Biometrics 32, 659–667.
Dorazio, R.M., 2012. Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68, 1303–1312.
Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecol. Biogeogr. 23, 1472–1484.
Farr, M.T., Green, D.S., Holekamp, K.E., Zipkin, E.F., 2020. Integrating distance sampling and presence-only data to estimate species abundance. Ecology 102, e03204.
Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: Pooling survey and collection data for multiple species. Methods Ecol. Evol. 6, 424–438.
Fletcher, R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A., Dorazio, R.M., 2019. A practical guide for combining data to model species distributions. Ecology 100, e02710.
Gelfand, A.E., Schliep, E.M., 2018. Bayesian inference and computing for spatial point patterns. NSF-CBMS Regional Conf. Ser. Probab. Statist. 10, 1–125.
Gerber, B.D., Karpanty, S.M., Kelly, M.J., 2012. Evaluating the potential biases in carnivore capture–recapture studies associated with the use of lure and varying density estimation techniques using photographic-sampling data of the Malagasy civet. Popul. Ecol. 54, 43–54.
Hefley, T.J., Baasch, D.M., Tyre, A.J., Blankenship, E.E., 2014. Correction of location errors for presence-only species distribution models. Methods Ecol. Evol. 5, 207–214.
Hefley, T.J., Baasch, D.M., Tyre, A.J., Blankenship, E.E., 2015. Use of opportunistic sightings and expert knowledge to predict and compare whooping crane stopover habitat. Conserv. Biol. 29, 1337–1346.
Hefley, T.J., Boyle, W.A., Mohankumar, N.M., 2020. Accounting for location uncertainty in distance sampling data. arXiv: 2005.14316.
Hefley, T.J., Brost, B.M., Hooten, M.B., 2017. Bias correction of bounded location errors in presence-only data. Methods Ecol. Evol. 8, 1566–1573.
Hefley, T.J., Hooten, M.B., 2016. Hierarchical species distribution models. Curr. Landsc. Ecol. Rep. 1, 87–97.
Hefley, T.J., Tyre, A.J., Baasch, D.M., Blankenship, E.E., 2013. Nondetection sampling bias in marked presence-only data. Ecol. Evol. 3, 5225–5236.
Herse, M.R., With, K.A., Boyle, W.A., 2018. The importance of core habitat for a threatened species in changing landscapes. J. Appl. Ecol. 55, 2241–2252.
Hooten, M.B., Hefley, T.J., 2019. Bringing Bayesian Models to Life. Chapman & Hall/CRC Press, Florida.
Isaac, N.J., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., et al., 2020. Data integration for large-scale models of species distributions. Trends Ecol. Evol. 35, 56–67.
Johnson, D.S., Laake, J.L., Ver Hoef, J.M., 2010. A model-based approach for making ecological inference from distance sampling data. Biometrics 66, 310–318.
Kéry, M., 2011. Towards the modelling of true species distributions. J. Biogeogr. 38, 617–618.
Kéry, M., Royle, J.A., 2015. Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS. In: Prelude and Static Models, vol. 1, Academic Press, London, United Kingdom.
Knapp, A.K., Briggs, J.M., Hartnett, D.C., Collins, S.L., 1998. Grassland Dynamics Long-Term Ecological Research in Tallgrass Prairie. Oxford University Press, New York.
Koshkina, V., Wang, Y., Gordon, A., Dorazio, R.M., White, M., Stone, L., 2017. Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. Methods Ecol. Evol. 8, 420–430.
Little, R.J.A., 1992. Regression with missing X's: A review. J. Amer. Statist. Assoc. 87, 1227–1237.
Little, R.J.A., Rubin, D.B., 2019. Statistical Analysis with Missing Data. John Wiley & Sons, New Jersey.
Martino, S., Pace, D.S., Moro, S., Casoli, E., Ventura, D., Frachea, A., Silvestri, M., Arcangeli, A., Giacomini, G., Ardizzone, G., et al., 2021. Integration of presence-only data from several sources: A case study on dolphins' spatial distribution. Ecography 44, 1533–1543.
Mason, A., Richardson, S., Plewis, I., Best, N., 2012. Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. J. Off. Stat. 28, 279–302.
McShea, W.J., Forrester, T., Costello, R., He, Z., Kays, R., 2016. Volunteer-run cameras as distributed sensors for macrosystem mammal research. Landsc. Ecol. 31, 55–66.
Miller, D.A., Pacifici, K., Sanderlin, J.S., Reich, B.J., 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods Ecol. Evol. 10, 22–37.
Mohankumar, N.M., Hefley, T.J., 2021. Using machine learning to model nontraditional spatial dependence in occupancy data. Ecology 103, e03563.
Otis, D.L., Burnham, K.P., White, G.C., Anderson, D.R., 1978. Statistical inference from capture data on closed animal populations. Wildl. Monogr. 62, 3–135.
Pollock, K.H., Nichols, J.D., Brownie, C., Hines, J.E., 1990. Statistical inference for capture-recapture experiments. Wildl. Monogr. 107, 3–97.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods Ecol. Evol. 6, 366–379.
Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.
Seber, G.A.F., 1982. The Estimation of Animal Abundance and Related Parameters. Macmillan, New York.
Shaffer, J.A., Igl, L.D., Johnson, D.H., Sondreal, M.L., Goldade, C.M., Nenneman, M.P., Wooten, T.L., Euliss, B.R., 2021. The Effects of Management Practices on Grassland Birds—Grasshopper Sparrow (Ammodramus Savannarum). Tech. Rep., U.S. Geological Survey, http://dx.doi.org/10.3133/pp1842GG.
Sicacha-Parada, J., Steinsland, I., Cretois, B., Borgelt, J., 2021. Accounting for spatial varying sampling effort due to accessibility in citizen science data: A case study of moose in Norway. Spatial Stat. 42, 100446.
Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Ann. Appl. Stat. 4, 1383–1402.
Williams, E.J., Boyle, W.A., 2018. Patterns and correlates of within-season breeding dispersal: A common strategy in a declining grassland songbird. Auk Ornithol. Adv. 135, 1–14.
Williams, E.J., Boyle, W.A., 2019. Causes and consequences of avian within-season dispersal decisions in a dynamic grassland environment. Anim. Behav. 155, 77–87.
Winnicki, S.K., Munguía, S.M., Williams, E.J., Boyle, W.A., 2020. Social interactions do not drive territory aggregation in a grassland songbird. Ecology 101, e02927.