Generating process for Citizen Science data.

Kwaku Peprah Adjei [*1,2], Jorge Sicacha-Parada [†1], Robert B. O'Hara [‡1,2], and Ingelin Steinsland [§1]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology NTNU, 7491 Trondheim, Norway
[2]Centre for Biodiversity Dynamics,Norwegian University of Science and Technology NTNU, 7491 Trondheim, Norway

1 Abstract

- Citizen science has become an integral part of ecological monitoring systems in recent years. Data collected from Citizen Science projects are subject to systematic and sampling biases. Failure to account for these biases can lead to biased estimates of the parameters of the true occurrences of the species of interest.

- We propose a flexible framework that makes it possible to model all the conceivable sources of biases in the generation of Citizen science data in a probabilistic way. We do this by proposing a three-stage thinning process that accounts for sampling bias, imperfect detection and misclassification.

- As identifiability issues may arise when setting up our modelling framework, we propose a scheme based on data integration or informative prior specification to overcome these issues.

Keywords

Citizen Science, Bayesian Modelling, Multispecies modelling, Structural identifiability, point process models

[*]kwaku.p.adjei@ntnu.no
[†]jorge.sicacha@ntnu.no
[‡]bob.ohara@ntnu.no
[§]ingelin.steinsland@ntnu.no

2 Introduction

In recent years, Citizen Science (referred to as 'CS' hereafter) has become an integral part of scientific research (Bird et al. 2014). CS involves the participation of volunteers in scientific tasks such as data collection. This has led to an increase in the amount of ecological data available to researchers through the various CS databases such as iNaturalist (Matheson 2014), GBIF (Telenius 2011), Artsdatabanken, amongst many others. However, these CS data are subject to both sampling and systematic biases, thereby raising concerns of its use in scientific research (Lewandowski & Specht 2015, Crall et al. 2011). The sampling bias may be due to sampling variation in space and time and the systematic bias could arise because of misreporting the species, misidentifying the species, imperfect detection, amongst many others (Bird et al. 2014, Balázs et al. 2021).

As discussed by Bird et al. (2014), various statistical approaches have been used in analysing CS data such as the generalised linear (mixed) models, hierarchical models, Maxent, amongst many others. These methods depend on the type of response variable. Some modellers of CS data use Maxent (Phillips et al. 2006, 2009), a constrained optimisation approach that finds the optimal species density subject to constraints (Chakraborty et al. 2011). As a non-stochastic approach, it is impossible to attach any uncertainty to Maxent's predictions. CS data have also been modelled as typical geostatistic data with binary response by the inclusion of pseudo-absences (Ferrier et al. 2002, Barbet-Massin et al. 2012). However, this approach adds an arbitrary amount of data and ignores the spatial autocorrelation between absences as pointed out in Gelfand & Shirota (2019).

Other approaches propose modelling CS data as a thinned point pattern (Chakraborty et al. 2011, Fithian et al. 2015, Simmonds et al. 2020, Sicacha-Parada et al. 2021). A point pattern is a collection of points whose locations are regarded as random. That is, the locations of the points are not fixed or previously chosen. The model that determines the location and amount of points in an area is called a spatial point process. Thinning is an operation on point patterns that uses a specified rule to determine which points in the point pattern are deleted. For CS data, the thinning of the point pattern of the true species occurrences are caused by the inherent biases of CS data. Sicacha-Parada et al. (2021) proposed that a single covariate related to a source of bias (e.g. accessibility) should be included in the linear predictor as a log-linear function to determine the probability of retaining an occurrence (i.e. a variable restricted to $(-\infty, 0]$) as in Yuan et al. (2017)). Fithian et al. (2015) and Simmonds et al. (2020) however proposed the integration of

extra sources of information, such as professional surveys, as a way to account for these biases. These biases can be explicitly modelled as a function of known covariates if that information is available, or just as an extra random effect when data related to the biases is not available (Simmonds et al. 2020). These approaches do not however explicitly take into consideration how the various sources of biases affect the observed data.

CS data are typically affected by biased sampling processes. Citizen scientists usually choose where they go. This decision is frequently influenced by factors such as accessibility (Monsarrat et al. 2019) and where observers expect to find more occurrences (i.e. preferential sampling; Diggle et al. (2010)). Chakraborty et al. (2011), Simmonds et al. (2020) and Sicacha-Parada et al. (2021) have acknowledged the role of the sampling process as a thinning factor in the context of point patterns. Fithian et al. (2015) explored a model that assumes the thinning process as a log-linear function that affects the intensity of the observed point pattern. Further work by Chakraborty et al. (2011), Sicacha-Parada et al. (2021) and Simmonds et al. (2020) explored the implications of not properly accounting for the sampling bias in CS data. This was done by comparing improvements in goodness-of-fit and ecological interpretability of a model that accounts for sampling bias to approaches such as Maxent or model that does not accounting for the sampling bias at all. In both cases, modelling the sampling bias as a thinning operation on a point pattern reduced the bias in the estimated effect of ecological covariates on the spatial distribution of species occurrences. These developments, however, have not explored other sources of bias that can affect the generation of CS data.

Species distribution model accuracy is not only affected by the spatial and sample bias, but also by imperfect detection (Mugumaarhahama et al. 2022, Kéry & Schmid 2004, Metz et al. 2020). Imperfect detection occurs when the citizen scientist fails to detect the species even though it was present. Detectability must be accounted for in estimating species trend and abundance because some species will be overlooked and the detection of the species will depend on its behaviour (Kéry & Schmid 2004). The detection and identification of these species are influenced by the observers attention to a particular species as well as place, time and factors determining visibility such as weather conditions (Arazy & Malkinson 2022).Failure to account for the imperfect detection when analysing CS data makes the parameter estimates more difficult and statistical inference less reliable (Mugumaarhahama et al. 2022). Integrated species distribution models have been proposed

to analyse the CS data with repeated site occupance data (Koshkina et al. 2017) or replicated point count data (Dorazio 2014) by using a thinned point pattern for the CS data and a N-Mixture model or occupancy model for the point count data and site occupancy data respectively; or combining data-level versus model-level bias correction methods (Erickson & Smith 2021).

During CS data collection, more than one species are reported. The report of one species can be misclassified as another. For example, iNaturalist reports potential misclassifed species when a species of interest is queried on their website (www.inaturalist.org). It is therefore important that we treat CS data as multispecies data with possible misclassifications and model it as such (Adjei, O'Hara, Finstad & Koch 2022). Misidentification, misreporting of species, and other sources of false positives (collectively known as misclassification) are one of the critical issues to be considered in CS data. Various methods have been developed to model these false positives such as those proposed by Miller et al. (2011), Chambert et al. (2015), Wright et al. (2020), Strickfaden et al. (2020), Kéry & Royle (2020) and Adjei, O'Hara, Finstad & Koch (2022) . These methods can be both model-based, which includes taking a subset of the data with verifiable certainty, instructing observers only to record observations they are sure about and increasing observer experiences (Strickfaden et al. 2020). These methods can also be design-based, for example, dependent double-observer method (Strickfaden et al. 2020). Failure to account for misclassification in CS data modelling can increase bias and decrease the precision of the parameter estimates (Royle et al. 2007, Bird et al. 2014, Strickfaden et al. 2020), leading to accidental culling of endangered species and assessments of the population status and incorrect conservation decisions (Austen et al. 2016).

Accounting for the biases existing in CS data in biodiversity is of paramount importance for users of these data. However, there is no consensus on how this can be done, and it is indeed a growing research field within both statistics and ecology (Isaac et al. 2014, Sicacha-Parada et al. 2021, Adjei, O'Hara, Finstad & Koch 2022). In this paper, therefore, we propose a multispecies model of the generating process of CS data in biodiversity. This model accounts simultaneously for multiple sources of bias and specifies each of them probabilistically. As far as we are aware, no work has been done in modelling the generating process of CS data in a probabilistic way. Proposing it is necessary as it can be used as a general framework for modelling how different sources of biases affect CS data. It relies on a straightforward specification of the observed CS data as a thinned point pattern. This point pattern is affected by common biases such as sampling bias,

imperfect detection and misclassification. We propose three stages of thinning, one for each type of bias considered. Each stage produces a probability of retaining a point from the previous stages. This novel approach is flexible and can accommodate for more biases in CS data beyond the ones discussed in this paper. This paper is meant to only introduce our modelling framework and its properties. Fitting and estimation of the parameters of this model is out of the scope of this paper.

This paper is organised as follows: in section 2, we introduce the multispecies data generating process we propose. In section 3, we explore properties of the proposed model for CS data by exploring potential issues of identifiability that this model can face and possible ways to solve them. Finally, in section 4 we make concluding remarks and suggest future work.

3 Methods

3.1 Motivating Example

We assume that in our study region (defined by mountains, roads, parks, forests) with a closed system (no immigration and emigration), there are two species, each with average number of observations (mean intensity) $\lambda_{i,true}$ for $i = 1, 2$ species. This expected number of observations can be explained by some environmental covariates. We assume that species 1 is more abundant at the parks, forests and roadsides than the mountainous areas, whereas the opposite is true for species 2, as presented in Figure 1a).

Citizen scientists are likely to collect data at the most accessible places such as roads and parks, leaving the less accessible sites such as mountains and forests with fewer data collected. We assume that the CS would collect data on both species at the same sampling location and thus the same probability of sampling for both species, as presented in Figure 1b).

Once citizen scientists have visited a site, there is a chance that some species are not observed. This brings about imperfect detection which reduces the species distribution at the sampled locations. The probability of detecting each species is different for the sampling location, and in this example we assume that the detection of species 1 and 2 is higher at the parks and roadside than the mountains and forests, as can be observed from Figure 1c).

The detected individuals are then reported to the CS database, but are subject to misclassifications due to some factors noted in the introduction. Since we have a closed system, misclassification does not reduce the total number of observed individuals, however it alters the observed amount of individuals of each species reported. This explanation is presented in Figure 1d) with an equal chance of being classified as either species 1 or 2 when true species identity is species 2 and no probability of misclassification when the true identity of species is Species 1.

The observed intensity of species 1 and 2 are very low and almost the same at mountainous areas, although species 2 has a high true intensity at that region. Species 1 has a high intensity in the forest, but has approximately the same observed intensity as species 2. Along the road and in the parks, the observed intensity is lower than the true species intensity. This explanation shows that the observed intensity of each species distribution is a thinning of the true species distribution by the sampling process, detection and misclassification.

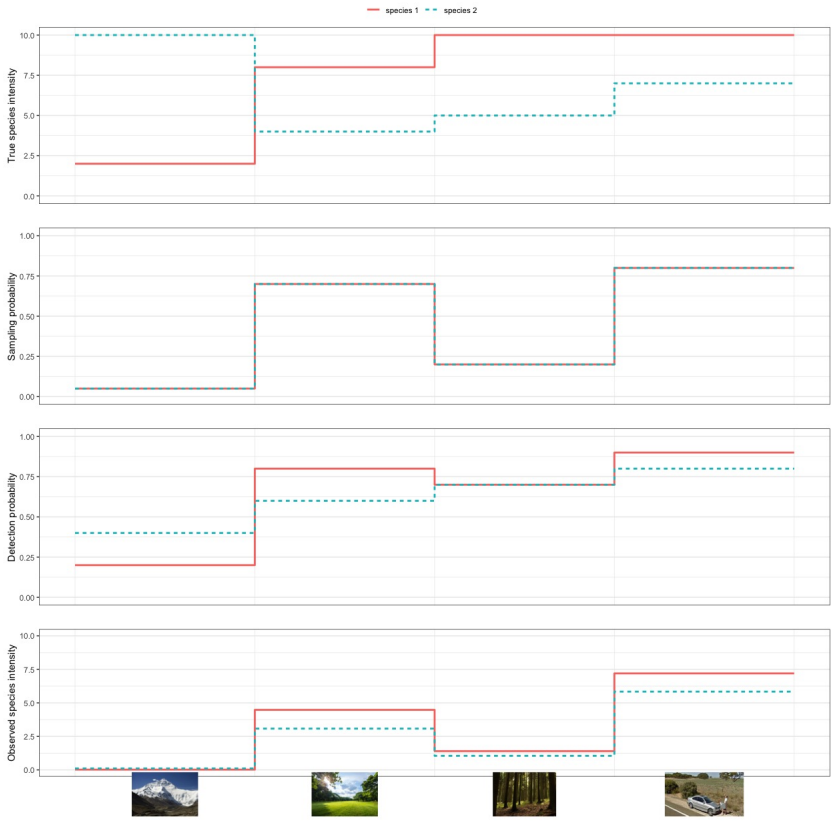

Figure 1: The observed intensity of two species after thinning the true species intensity with the probability of sampling, detection and misclassification in four regions depicted by the four images: mountains, parks, forests and roadside.

3.2 Definition of terminology

For the rest of this paper, we define the following parameters to be used in the presentation of the model:

- $\lambda_{i,obs}(\mathbf{s})$ be the observed intensity (expected number of observations) of the species i at location $\mathbf{s} \in D \subset \mathbb{R}^2$.
- $\lambda_{i,true}(\mathbf{s})$ be the true intensity (true expected number) of species i at location $\mathbf{s}$.
- $\zeta(\mathbf{s})$ represent the probability of sampling by citizen scientists (assumed to be the same for all the species).
- $\psi_i(\mathbf{s})$ be the probability of detection of species i at location $\mathbf{s}$ given it was sampled by citizen scientists during the sampling process .
- Ω_{ji} be the probability of reporting species i as species j. The whole matrix $\mathbf{\Omega}$ defines the classification probabilities

3.3 Generating process for Citizen science data Proposed Modelling Framework

We propose that CS data is the result of a thinning process on the true point pattern of the occurrences of species. This generating process for CS data involves more than one thinning stages, with each thinning stage being a source of bias in the CS data. The various stages of thinning are discussed in the following subsections and summarised in Figure 2.

The CS observations as can be obtained from any CS database can be represented in the modelling framework defined from the generating process. Having a conceptual understanding of this generating process can help in modelling all potential sources of biases that can be perceived in the generating process. A graphical representation of each of the thinning stages in the flowchart are displayed in Figure 3.

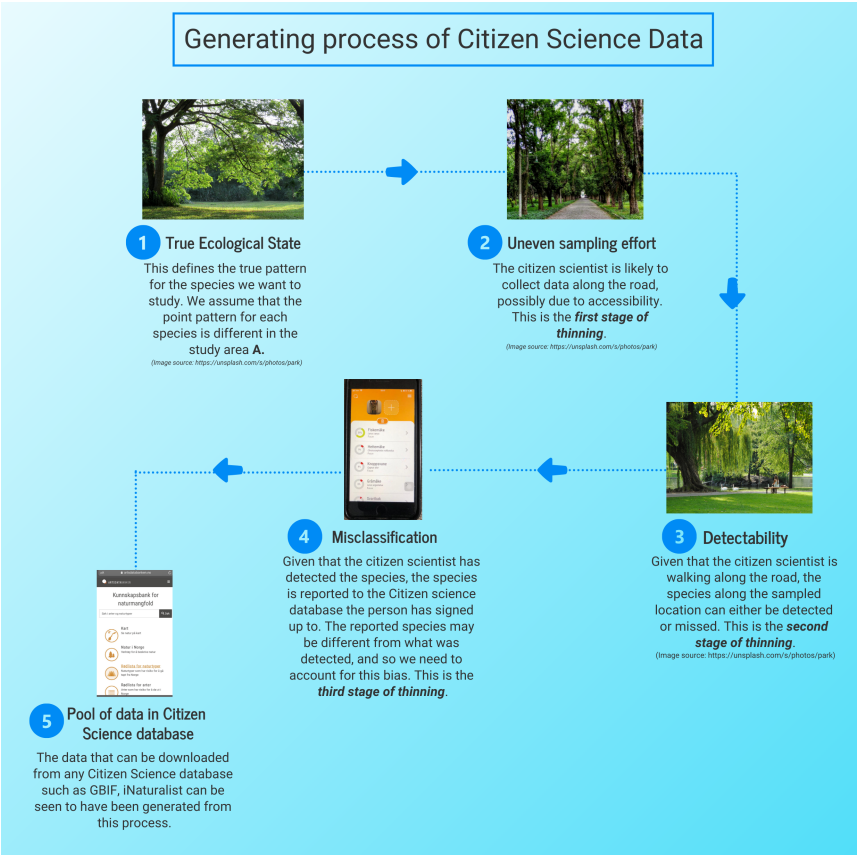

Figure 2: Flowchart of the generating process of citizen science data described in this paper. The generating process assumes that the true ecological state of the species is thinned by the sampling probability, detection probability and the classification probability. The arrows indicate the thinning process and the dependency in the Bayesian framework used for the modelling framework.

$$\text{Observed intensity of species j} = \sum_{i=1}^{\text{nspecies}} \text{True intensity of species i}$$
$$\times \text{P(species i was sampled given its true intensity)}$$
$$\times \text{P(species i was detected given it was sampled)}$$
$$\times \text{P(species j was reported given the true species was species i).}$$

Mathematically,

$$\lambda_{j,obs}(\mathbf{s}) = \sum_{i=1}^{\text{nspecies}} \lambda_{i,true}(\mathbf{s}) \times \zeta(\mathbf{s}) \times \psi_i(\mathbf{s}) \times \Omega_{ji}. \tag{1}$$

In matrix notation, the model can be represented as:

$$\begin{bmatrix} \lambda_{1,obs}(\mathbf{s}) \\ \lambda_{2,obs}(\mathbf{s}) \\ \vdots \\ \lambda_{nspecies,obs}(\mathbf{s}) \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{21} & \dots & \Omega_{nspecies,1} \\ \Omega_{12} & \Omega_{22} & \dots & \Omega_{nspecies,2} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{1,nspecies} & \Omega_{2,nspecies} & \dots & \Omega_{nspecies,nspecies} \end{bmatrix}^T \begin{bmatrix} \lambda_{true,1}(\mathbf{s}) \times \zeta(\mathbf{s}) \times \psi_1(\mathbf{s}) \\ \lambda_{true,2}(\mathbf{s}) \times \zeta(\mathbf{s}) \times \psi_2(\mathbf{s}) \\ \vdots \\ \lambda_{true,nspecies}(\mathbf{s}) \times \zeta(\mathbf{s}) \times \psi_{nspecies}(\mathbf{s}) \end{bmatrix} \tag{2}$$

□

From equations (1) and (2), the true ecological state is affected by the sampling process, the detection probability as well as the classification probabilities and results in species-specific point patterns with intensities $\lambda_{i,obs}, i = \{1, \dots, nspecies\}$. In equation 2, it is noteworthy that $\lambda_{i,obs}$ are weighted sums of the observed intensities before misclassification ($\lambda_{true,j}(\mathbf{s}) \times \zeta(\mathbf{s}) \times \psi_j(\mathbf{s})$) with weights given by the classification probabilities for species i ($\Omega_{\cdot i} = [\Omega_{i1}, \dots, \Omega_{i,nspecies}]$).

3.3.1 True Ecological state

The intensity of the true ecological state of each species, $\lambda_{i,true}(\mathbf{s})$, is modelled as a Log Gaussian Cox Process (Illian et al. 2008):

$$\log(\lambda_{i,true}(\mathbf{s})) = X(\mathbf{s})\boldsymbol{\beta_i} + \omega_{1i}(\mathbf{s}) \qquad \text{for } i = \{1, \dots, nspecies\}, \tag{3}$$

where $\omega_{1i}(\mathbf{s})$ is a zero-mean GRF that differs for each species, $\boldsymbol{\beta}_i$ are the fixed effects that describe the mean of the GRF for each species and $\mathbf{X}(\mathbf{s})$ is the set of covariates that affects the true intensity

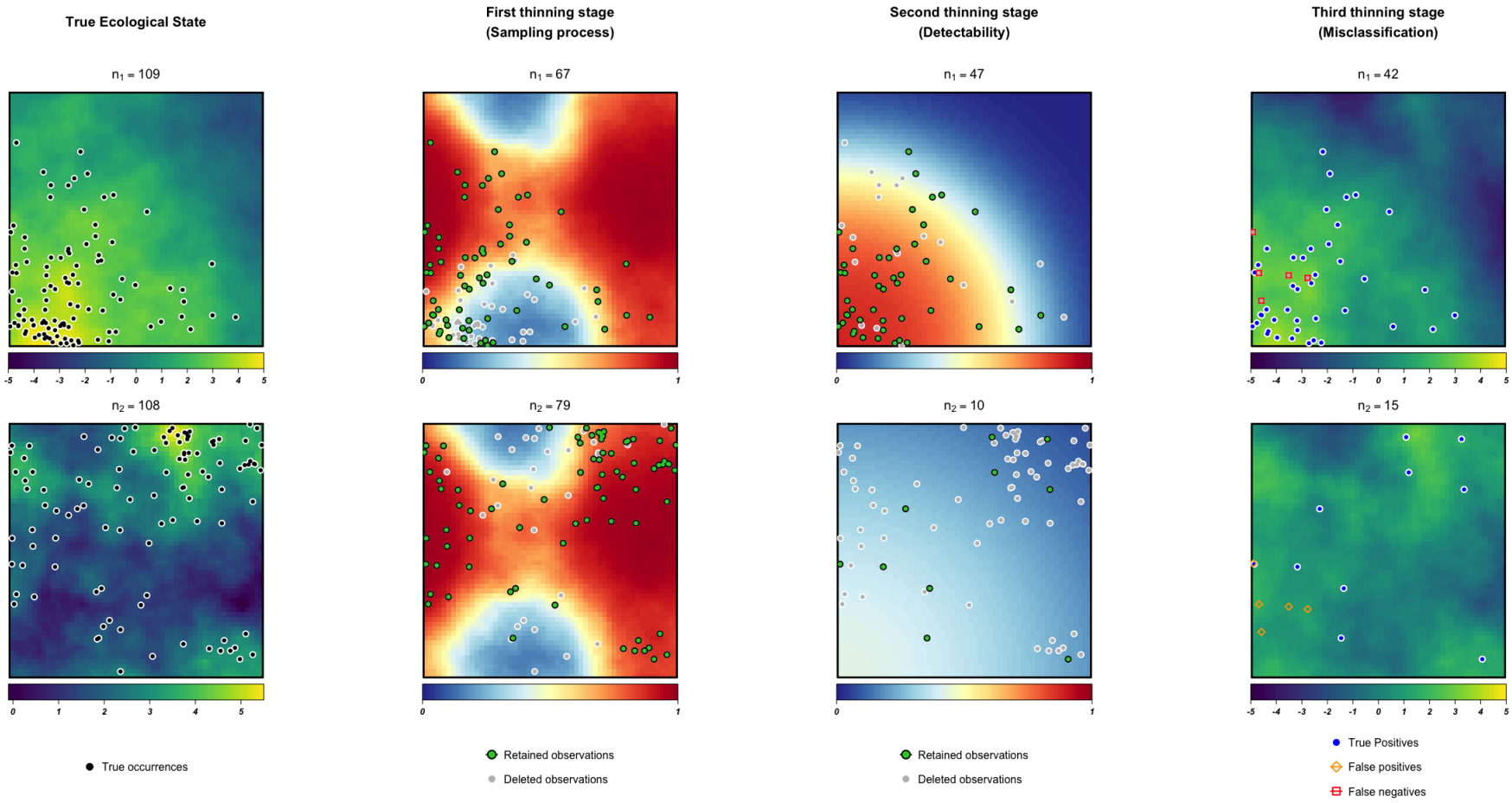

Figure 3: The reported occurrences for each species as can be downloaded from a Citizen Science Database without verification. The blue circles are the true species 1 observations, the green squares are the true species 2 observations, the yellow triangles are the true species 3 observations and the red rhombus are the true species 4 observations.

of the species. Given the same study region, the covariate matrix $\mathbf{X}$ is assumed, in principle, to be the same for all the species. However, it can be species-specific. We assume a different GRF for each species as we assume the spatial autocorrelation is varies from species to species, as presented in the left-side panel of Figure 3 .

3.3.2 Sampling bias

CS data is affected by where citizen scientists prefer to go and collect observations as discussed in the motivation example. This preference in sampling introduces a bias and is regarded as a thinning source on the true ecological state. This thinning source depends on another GRF, $\omega_2(\mathbf{s})$, which we assume the same for all the species because we assume the sampling process does not depend on the species distribution. In scenarios where the target species influence the sampling process, species-specific GRFs can be defined. Moreover, if the presence of the species influences where CS go to collect observations (i.e. preferential sampling (Diggle et al. 2010)), the GRF of the true ecological state ($\omega_1(\mathbf{s})$) can be shared with the sampling process.

The probability of retaining a point in the true ecological state is given by:

$$logit(\zeta(\mathbf{s})) = U(\mathbf{s})\boldsymbol{\alpha} + \omega_2(\mathbf{s}) \tag{4}$$

where $U(\mathbf{s})$ are the covariates that explain the sampling process of citizen scientists, $\boldsymbol{\alpha}$ are the parameters associated to the fixed effect and $\omega_2(\mathbf{s})$ is a zero-mean GRF. An example of this thinning stage can be seen in the second column of Figure 3. For both species most of the true occurrences in the blue area are not reported. Hence, the number of occurrences reported after the first stage of thinning is about 61% of the true abundance for species 1 and 73% for species 2.

3.3.3 Detection Probability

Ecological data are subject to imperfect detection of the species. Given the true ecological state and the sampling process, citizen scientists may not be able to detect all the species that are available. Hence, this model framework incorporates the detection probability by assuming a generalised linear model for the detection. This model does not include a GRF, but it has the covariates varying in space. This is because the GRF for the detection of the species are already captured by the sampling process in the preceding subsection. That is, the detection can only be made along the the roads, for example, and nowhere else. The model is given as:

$$logit(\psi_i(\mathbf{s})) = Z(\mathbf{s})\boldsymbol{\gamma_i}, \tag{5}$$

where $Z(\mathbf{s})$ are the covariates that affect the detection of the species, such as the rainfall, etc and γ are the fixed effects that describe the covariate effects for each species. In the third column of figure 3, we can see how the detectability of each species affects the resulting point pattern. For example, species 2 has high detectability almost everywhere in our region of study, whereas species 1 has low detectability in the upper right corner, which means the loss of many occurrences because they could not be detected. Consequently the number of occurrences after the second stage of thinning is about 31% of the sampled species 1 occurrences and 23% of the sampled species 2.

3.3.4 Classification

Citizen scientists report the observations they have detected to a database they have signed up to. Due to various reasons such as the experience of the citizen scientists, easy confusion of similarly looking species, among others, the species detected can be reported as something else. This leads to a lot of false positives in the CS observations. These reported observations are modelled as:

$$Observation_i \sim Multinomial(\Omega_{i.}) \tag{6}$$

where $\Omega_{ij} = P$(observing species j|true species i). Given the definition of Ω, the rows of this matrix should sum to 1. The last column of figure 3 shows the reported species with the potential misclassifications that citizen scientists can make.

4 Identifiabilty issues

The model framework we have defined in section 3.3 is overparameterized (ie depends on many parameters). Overparameterized models are most likely to have identifiability issues (Wieland et al. 2021). We therefore assessed the structural identifiability of our model. Structural identifiabilty occurs when different set of parameter values are solutions to the same equation under study, in our case Equations (1) and (2). For more details on the theory of this kind of identifiability, the reader is referred to Wieland et al. (2021).

The structural identifiability happens in this model when, for example, there are no misclassifications and the true intensity, $\lambda_{true}(\mathbf{s})$ is multiplied by a constant $K \in \mathbb{R}$, and either the sampling

retaining probabilities $\zeta(\mathbf{s})$ or the detection probabilities $\psi_i(\mathbf{s})$ are multiplied by a constant $1/K$. That is, either

$$\begin{aligned} \lambda_{i,obs} &= K \cdot \lambda_{i,true}(\mathbf{s}) \times \left(\frac{1}{K} \cdot b(\mathbf{s})\right) \times \psi_i(\mathbf{s}) \quad \text{or} \\ \lambda_{i,obs} &= K \cdot \lambda_{i,true}(\mathbf{s}) \times b(\mathbf{s}) \times \left(\frac{1}{K} \cdot \psi_i(\mathbf{s})\right) \end{aligned} \tag{7}$$

would give the same observed intensity λ_{obs}.

We summarise some instances where structural identifiability can occur using two species in the Table 1. When there are no misclassifications ($\Omega_{12} = \Omega_{21} = 0$), it can be observed that increasing the true intensity of species 1 ($\lambda_{1,true}$) by a factor of 100 and dividing the sampling probability (ζ) by 100 as well as increasing the true intensity of species 1 ($\lambda_{1,true}$) by a factor of 100 and dividing its detection probability (ψ_1) by 100 gives the same observed intensity ($\lambda_{1,obs}$) as the baseline scenario. When there are misclassifications (Ω_{12} and Ω_{21} are different from zero), then increasing the true intensity of species 1 ($\lambda_{1,true}$) by a factor of 100 and dividing its detection probability (ψ_1) by 100 as well as increasing the true intensity of both species 1 ($\lambda_{1,true}$) and 2 ($\lambda_{2,true}$) by a factor of 100 and dividing the sampling probability (ζ) by 100 gives the same observed intensity as the baseline scenario for species 1 ($\lambda_{1,obs}$) in both instances and ($\lambda_{2,obs}$) for the latter.

With these potential identifiability issues, additional data is needed to constrain the parameters of each thinning stage in our model. This data integration method helps provide unique solutions to equations 1 and 2. For example, professional surveys can give information about the detection probability, massive collections of CS data can inform about the sampling process and the Machine Learning as well as verified data classifications could provide necessary information for the classification probabilities (McDonald & Hodgson 2021). Alternatively, informative priors on the parameters of the model can be proposed as a way to overcome potential structural identifiability issues (Gelman et al. 1996, Zyphur & Oswald 2015, Lemoine 2019). However, in practice it is hard to come up with such prior distributions (Lemoine 2019).

It however has to be commented that practical identifiability and other structural identifiability issues can be tested after the fitting and estimating the parameters in the model. The fitting and estimation of the parameters is out of the scope of this paper.

		$\lambda_{1,true}(s)$	$\lambda_{2,true}(s)$	$\zeta(s)$	$\psi_1(s)$	$\psi_2(s)$	Ω_{11}	Ω_{12}	Ω_{21}	Ω_{22}	$\lambda_{1,obs}(s)$	$\lambda_{2,obs}(s)$
No misclassification	Baseline scenario	10	7	0.5	0.2	0.5	1	0	0	1	1	1.75
	$K \times \lambda_{1,true}$ and ζ/K	1000	7	0.005	0.2	0.5	1	0	0	1	1	0.0175
	$K \times \lambda_{1,true}$ and ψ_1/K	1000	7	0.5	0.002	0.5	1	0	0	1	1	1.75
Misclassification	Baseline scenario	10	7	0.5	0.2	0.5	0.5	0.5	0.2	0.8	0.85	1.9
	$K \times \lambda_{1,true}$ and ψ_1/K	1000	7	0.5	0.002	0.5	0.5	0.5	0.2	0.8	0.85	1.9
	$K \times \lambda_{1,true}$, $K \times \lambda_{2,true}$ and ζ/K	1000	700	0.005	0.2	0.5	0.5	0.5	0.2	0.8	0.85	1.9

Table 1: Structural identifiability scenarios with and without misclassification for two species. In red: Equal observed intensities λ_{obs} with different parameter values with $K = 100$.

5 Conclusion

Biodiversity data from CS data have lots of data quality issues (Dickinson et al. 2012, Balázs et al. 2021). These issues include but are not limited to sampling bias, imperfect detection and misclassification. When these biases are left unaccounted for, they can affect the inferences made about the species distribution. In this paper, we have proposed an approach that makes it possible to account for these common sources of bias in CS data for more than one species in a single model framework. This general model framework assumes these sources of biases act as a thinning process on the true ecological state, with similar approaches used by Sicacha-Parada et al. (2021), Simmonds et al. (2020) for accounting for varying sampling effort in CS data.

By trying to model the sources of biases in CS data, we also described the generating process of CS data. Existing frameworks have used the reporting process the citizen scientists go through (Kéry & Schmid 2004, Kelling et al. 2015), and decision-making approach by treating observers monitoring activity as a series of decisions (Arazy & Malkinson 2022). In our proposed framework, CS data are generated probabilistically by defining a hierarchical structure that describes how the sources of biases affect the true species distribution. Such structure makes it straightforward to account for more sources of biases in CS data than the ones considered in this paper. This novel approach adds up to the existing frameworks that provides knowledge of the true ecological state of species.

Structural identifiability arises naturally in our modelling framework as the observed intensity depends on many parameters. That is, multiple inputs produce the same observed intensity. The framework therefore needs additional data to be integrated with the CS data to obtain parameter estimates at the various stages of thinning. Camera traps data and telemetry can be integrated to reduce spatial bias (Cretois et al. 2021, Sicacha-Parada et al. 2021, Santangeli et al. 2020), automated identification of species reports for misclassification and professional surveys for detection probability (Kelling et al. 2015). The data integration opportunity presented by this framework allows data from different survey schemes to be used in estimating the parameters. For example, questions regarding the location of hotspots of collision of birds due to the placement of power lines can be addressed with this modelling framework as it uses both CS data and professional surveys.

A natural extension of this work is the fitting of the proposed models. As our models may include spatial random effects and several parameters, computationally efficient methods are required to fit them.

6 Acknowledgements

This work is part of the Transforming Citizen Science for Biodiversity project, funded by the NTNU digital transformation initiative .

7 Data accessibility

The Rcode used in generating the plots in the paper can be accessed openly on Adjei, Sicacha-Parada, O'Hara & Steinsland (2022)

References

Adjei, K. P., O'Hara, R. B., Finstad, A. G. & Koch, W. (2022), 'Accounting for misclassification in multispecies distribution models'.
URL: *https://arxiv.org/abs/2204.03708*

Adjei, K. P., Sicacha-Parada, J., O'Hara, R. B. & Steinsland, I. (2022), 'Rcode and supplementary information for data generating process of citizen science data', *Dyrad, Dataset* .
URL: *https://doi.org/10.5061/dryad.0rxwdbs2m*

Arazy, O. & Malkinson, D. (2022), 'A framework of observer-based biases in citizen science biodiversity monitoring: Semi-structuring unstructured biodiversity monitoring protocols', *Citizen Science for Future Generations* .

Austen, G. E., Bindemann, M., Griffiths, R. A. & Roberts, D. L. (2016), 'Species identification by experts and non-experts: comparing images from field guides', *Scientific Reports* **6**(1), 1–7.

Balázs, B., Mooney, P., Nováková, E., Bastin, L. & Arsanjani, J. J. (2021), 'Data quality in citizen science', *The science of citizen science* p. 139.

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

3(2), 327–338.

URL: *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2011.00172.x*

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F. et al. (2014), 'Statistical solutions for error and bias in global citizen science datasets', *Biological Conservation* **173**, 144–154.

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', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**(5), 757–776.

Chambert, T., Miller, D. A. & Nichols, J. D. (2015), 'Modeling false positive detections in species occurrence data under different study designs', *Ecology* **96**(2), 332–339.

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J. & Waller, D. M. (2011), 'Assessing citizen science data quality: an invasive species case study', *Conservation Letters* **4**(6), 433–442.

Cretois, B., Simmonds, E. G., Linnell, J. D., van Moorter, B., Rolandsen, C. M., Solberg, E. J., Strand, O., Gundersen, V., Roer, O. & Rød, J. K. (2021), 'Identifying and correcting spatial bias in opportunistic citizen science data for wild ungulates in norway', *Ecology and evolution* **11**(21), 15191–15204.

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T. & Purcell, K. (2012), 'The current state of citizen science as a tool for ecological research and public engagement', *Frontiers in Ecology and the Environment* **10**(6), 291–297.

Diggle, P. J., Menezes, R. & Su, T.-l. (2010), 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2), 191–232.

URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2009.00701.x*

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.

Erickson, K. D. & Smith, A. B. (2021), 'Accounting for imperfect detection in data from museums and herbaria when modeling species distributions: combining and contrasting data-level versus model-level bias correction', *Ecography* **44**(9), 1341–1352.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002), 'Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling', *Biodiversity & Conservation* **11**(12), 2309–2338.
URL: *https://doi.org/10.1023/A:1021374009951*

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 in ecology and evolution* **6**(4), 424–438.

Gelfand, A. E. & Shirota, S. (2019), 'Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data', *Ecological Monographs* **89**(3), e01372.
URL: *https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1372*

Gelman, A., Bois, F. & Jiang, J. (1996), 'Physiological pharmacokinetic analysis using population modeling and informative prior distributions', *Journal of the American Statistical Association* **91**(436), 1400–1412.

Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. (2008), *Statistical analysis and modelling of spatial point patterns*, Vol. 70, John Wiley & Sons.

Isaac, N. J., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. (2014), 'Statistics for citizen science: extracting signals of change from noisy ecological data', *Methods in Ecology and Evolution* **5**(10), 1052–1060.

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F. A., Moore, T., Wiggins, A. et al. (2015), 'Can observation skills of citizen scientists be estimated using species accumulation curves?', *PloS one* **10**(10), e0139600.

Kéry, M. & Royle, J. A. (2020), *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 2: Dynamic and Advanced Models*, Academic Press.

Kéry, M. & Schmid, H. (2004), 'Monitoring programs need to take into account imperfect species detectability', *Basic and applied ecology* **5**(1), 65–73.

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 in Ecology and Evolution* **8**(4), 420–430.

Lemoine, N. P. (2019), 'Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses', *Oikos* **128**(7), 912–928.

Lewandowski, E. & Specht, H. (2015), 'Influence of volunteer and project characteristics on data quality of biological surveys', *Conservation Biology* **29**(3), 713–723.

Matheson, C. A. (2014), 'inaturalist', *Reference Reviews* .

McDonald, J. L. & Hodgson, D. (2021), 'Counting cats: The integration of expert and citizen science data for unbiased inference of population abundance', *Ecology and Evolution* **11**(9), 4325–4338.

Metz, M. C., SunderRaj, J., Smith, D. W., Stahler, D. R., Kohl, M. T., Cassidy, K. A. & Hebblewhite, M. (2020), 'Accounting for imperfect detection in observational studies: modeling wolf sightability in yellowstone national park', *Ecosphere* **11**(6), e03152.

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L. & Weir, L. A. (2011), 'Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification', *Ecology* **92**(7), 1422–1428.

Monsarrat, S., Boshoff, A. F. & Kerley, G. I. H. (2019), 'Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records', *Ecography* **42**(1), 125–136.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.03944*

Mugumaarhahama, Y., Fandohan, A. B., Mushagalusa, A. C., Sode, I. A. & Glèlè Kakaï, R. L. (2022), 'Inhomogeneous poisson point process for species distribution modelling: relative performance of methods accounting for sampling bias and imperfect detection', *Modeling Earth Systems and Environment* pp. 1–14.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006), 'Maximum entropy modeling of species geographic distributions', *Ecological Modelling* **190**(3), 231–259.
URL: *http://www.sciencedirect.com/science/article/pii/S030438000500267X*

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009), 'Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data', *Ecological Applications* **19**(1), 181–197.
URL: *https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-2153.1*

Royle, J. A., Kéry, M., Gautier, R. & Schmid, H. (2007), 'Hierarchical spatial models of abundance and occurrence from imperfect survey data', *Ecological Monographs* **77**(3), 465–481.

Santangeli, A., Pakanen, V.-M., Bridgeford, P., Boorman, M., Kolberg, H. & Sanz-Aguilar, A. (2020), 'The relative contribution of camera trap technology and citizen science for estimating survival of an endangered african vulture', *Biological Conservation* **246**, 108593.

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 Statistics* **42**, 100446.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. & 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.

Strickfaden, K. M., Fagre, D. A., Golding, J. D., Harrington, A. H., Reintsma, K. M., Tack, J. D. & Dreitz, V. J. (2020), 'Dependent double-observer method reduces false-positive errors in auditory avian survey data', *Ecological Applications* **30**(2), e02026.

Telenius, A. (2011), 'Biodiversity information goes public: Gbif at your service', *Nordic Journal of Botany* **29**(3), 378–381.

Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C. & Timmer, J. (2021), 'On structural and practical identifiability', *Current Opinion in Systems Biology* .

Wright, W. J., Irvine, K. M., Almberg, E. S. & Litt, A. R. (2020), 'Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity', *Methods in Ecology and Evolution* **11**(1), 71–81.

Yuan, Y., Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., Buckland, S. T., Rue, H., Gerrodette, T. et al. (2017), 'Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales', *The Annals of Applied Statistics* **11**(4), 2270–2297.

Zyphur, M. J. & Oswald, F. L. (2015), 'Bayesian estimation and inference: A user's guide', *Journal of Management* **41**(2), 390–420.