

Doctoral theses at NTNU, 2023:85

Jorge Sicacha Parada

Spatial statistical models using citizen science data and professional surveys for biodiversity insight

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences



Norwegian University of
Science and Technology

Jorge Sicacha Parada

Spatial statistical models using citizen science data and professional surveys for biodiversity insight

Thesis for the Degree of Philosophiae Doctor

Trondheim, March 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences



Norwegian University of
Science and Technology

NTNU

Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

© Jorge Sicacha Parada

ISBN 978-82-326-6457-3 (printed ver.)
ISBN 978-82-326-6549-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:85

Printed by NTNU Grafisk senter

Spatial statistical models using citizen science
data and professional surveys for biodiversity
insight

Jorge Sicacha Parada

December 2022

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU). The work was carried out at the Department of Mathematical Sciences during the years 2018-2022 and funded by NTNU digital transformation initiative.

First of all, I want to thank my supervisor Ingelin Steinsland for her guidance, support and advice throughout these four years. I really appreciate her patience and valuable ideas in the moments I got stuck. I would also like to thank my co-supervisors Bob O'Hara and Jan Kjetil Rød for the fruitful discussions and contributions in the meetings of the project Transforming Citizen Science for Biodiversity. This project integrated an interdisciplinary and multicultural group of PhD candidates. I want to thank Kwaku, Ben, Jan, Philip, Caitlin and Wouter for the meetings, conferences, thoughts and beers shared.

Big part of my thesis was made in collaboration with Diego Pavon-Jordan from the Department of Terrestrial Ecology at the Norwegian Institute for Nature Research (NINA). I am very grateful for his valuable contributions, the long and enjoyable discussions as well as the visits to NINA.

My experience these four years was enjoyable in large part thanks to the environment at the department. I would like to thank all my coworkers at the department, in particular my friends Michail Spitieris and Kwaku Peprah Adjei. Our discussions about statistics, football, life and many other subjects made my life as PhD much easier.

Lastly, the biggest thanks goes to my mother Barbara, and my wife Alejandra. You are the motivation of my days and the reason why I will never stop dreaming. Without your support, feedback and unconditional love this thesis would have never become reality. LAS AMO.

Jorge Sicacha Parada
Trondheim, December 2022

Contents

1	Introduction	1
1.1	Spatial statistics	7
1.1.1	Gaussian Random Fields	8
1.1.1.1	Gaussian Markov Random Fields	10
1.1.2	Spatial point patterns	11
1.2	Bayesian Statistics	14
1.2.1	Prior distribution	15
1.2.2	Latent Gaussian Models	17
1.2.3	Computational methods in Bayesian statistics	18
1.2.3.1	The Integrated Nested Laplace Approximation (INLA)	19
1.2.4	The Stochastic Partial Differential Equation (SPDE) approach	21
1.2.4.1	The SPDE approach and Log-Gaussian Cox Processes	23
1.2.5	Nonlinear functions in the linear predictor: the <i>inlabru</i> approach	25
1.2.6	Bayesian data integration	26
1.3	Established methods for ecological and biodiversity data	28
1.3.1	Presence/absence data	29

1.3.2	Count data	29
1.3.3	Citizen Science data	30
2	Scientific papers	35

Chapter 1

Introduction

Citizen Science (CS) is defined as the open engagement of the public in activities formerly exclusive of trained people in scientific projects, (Newman et al., 2012). Information and data have been readily available to the society in general as consequence of the expansion of technology. The convenience offered by technology has encouraged people to contribute in scientific research. Hence, Citizen Science in practice contributes mostly in the stage of data collection of the scientific projects (Strasser et al., 2019) in a handful of fields, such as social sciences (Tauginienė et al., 2020), astronomy (Lingard et al., 2020; Beaumont et al., 2014) or biodiversity (Peter et al., 2021). Citizen scientists also contribute to CS projects by planning, processing and analysing data, or by assessing the final outcome of the projects, (European Commission and Directorate-General for Research and Innovation, 2020). CS has gained a space within the scientific community as an approach to address research questions. In fact, it has been recognized within policy makers as a new framework for research and innovation, and has also been acknowledged as part of the conceptualization of open science (Hecker et al., 2018). In CS projects in biodiversity conservation, citizen scientists play a key role as data collectors. Some CS databases contain data for only one country, e.g. Artobservasjoner (www.artobservasjoner.no), which contains nearly 29 million CS reports made in Norway. Citizen scientists also report their observations to species-specific databases as eBird (ebird.org), managed by the Cornell Lab of Ornithology and with more than 100 million bird sightings reported annually. Ultimately, reports from all the world and from a broader range of species are available in databases such as the Global Bio-

diversity Information Facility, GBIF (*gbif.org*), which contains above 2 billion reports. In these databases the participants of CS projects voluntarily report which species they have observed, the spatial location and time of the occurrence, as well as other specific details.

About 1 million species are threatened with extinction, mostly due to human activities (IPBES, 2019). More affectation on ecosystems would mean a considerable impact for the economy, food security and quality of life around the world. Both UN Sustainability Development Goal (SDG) 15 (United Nations, 2015) and IPBES report on biodiversity and ecosystem services (IPBES, 2019) urge for actions that contribute to restore and protect nature. In particular, SDGs 15.4 and 15.5 highlight the necessity of reducing the degradation of ecosystems as well as the importance of halting the loss of biodiversity that inhabits these ecosystems. Thus, biodiversity conservation has become a relevant research field in ecology for identifying strategies to protect endangered species and their ecosystems (Asaad et al., 2017). One of the necessities of biodiversity conservation is to make the society aware of the importance of biodiversity. Through CS projects, not only the participants contribute to science, but also learn about biodiversity and change their habits so that they can contribute to protect biodiversity in their daily lives (Peter et al., 2021). Biodiversity conservation needs information on species as scientific evidence to target specific research questions (e.g. conservation in residential landscapes (Cooper et al., 2007) or management of invasive species (Clements et al., 2021), see also UN SDG 15.8). CS data provides a massive source of information for overcoming spatial and species information gaps (Amano et al., 2016). Thus, CS data offer tools to assess biodiversity and prevent the extinction of endangered species.

Research questions in ecology and biodiversity are also addressed using many other data types. For example, high quality data are collected by scientists through professional surveys. These data are collected through standardized sampling protocols by skilled observers. Hence, sampling effort is even and a report of presence/absence of species is possible (Gelfand and Shirota, 2019). Despite the quality of the sampling design of professional surveys, these surveys are expensive in terms of money, time and effort. Hence, they are not as massive as CS data, their spatial coverage is low and their temporal resolution is coarse. Trying to solve ecological questions based only on this data type might produce accurate estimates, but with high prediction uncertainty due to its low spatial coverage. Another potential issue using data from professional surveys that requires attention is preferential sampling (Diggle et al., 2010), as the sampling

design of professional surveys is made by experts whose knowledge can make the selection of sites to visit depend on the target ecological process.

Compared to data from professional surveys, CS data offer a cost-efficient alternative as their collection happens on a voluntary basis, they have broader spatial coverage and thus massive observations are reported every day. Therefore, large amounts of information are available in the repositories of CS projects. For example, by August 2022, about 2.2 billion occurrences had been reported in GBIF. Despite being massive and simple to collect, CS data have some drawbacks, namely biases in their collection process. These biases can be classified in four groups: temporal bias, geographical bias, uneven sampling effort per visit and differences in detectability (Isaac et al., 2014). As usual in most applied problems only a sample of the population is observed. There are three types of missing-data mechanisms (Little and Rubin, 2019) that relate the values of the data and their missingness. The data are called missing completely at random (MCAR) if missingness does not depend on the values of the data; missing at random (MAR) when missingness depends only on the values of observed data and missing not at random (MNAR) when the missingness depends on the values of the data. Based on these definitions, it is reasonable to assume that CS data are generated by MNAR mechanisms. Inferring a parameter based on these data requires thorough consideration as the most frequently used statistical methods assume the observations are obtained from a random sample. Using these methods on MNAR data might yield biased inference of the parameters (Diggle et al., 2010; Gelfand and Shirota, 2019). For example, Maxent, which is one of the most used data analysis methods for biodiversity data, fails to account for common features of biodiversity data as spatial autocorrelation. Furthermore, as Maxent is an approach based on a deterministic algorithm, it does not report the uncertainty of the variables predicted (Gelfand and Shirota, 2019). These flaws added to the fact that the biases in the collection process of CS data are not properly accounted for by MaxEnt (Chakraborty et al., 2011) highlight the necessity of using methodological approaches that deal with these problems and hence produce enhanced inference of the ecological state of interest. Other common approach is to create pseudo-absences (Ferrier et al., 2002) in order to increase spatial coverage and fit a logistic regression on the new dataset. This approach implies, however, to incorporate absences in locations where it is not certain if the species is actually absent. Multiple efforts have been made to account for the biases in the collection of CS data, and hence contribute to much more appropriate use of these data. Spatial filtering is performed to standardize the amount of observations across the region of study (Robinson et al.,

2018; Borgelt et al., 2022). For many species this means to discard considerable amount of data. Other researchers have also considered the inclusion of spatial covariates to account for biases (Fithian et al., 2015). Many of these methods do not consider the spatial nature of the problem and do not account explicitly for spatial autocorrelation (Gelfand and Shirota, 2019), or aggregate CS data to fit spatial areal models (Conn et al., 2017). Given the massive amount of CS records and the precision of the spatial location of each report, aggregating these data means loss of information and results that will depend on the size of the aggregations. Therefore, recent efforts have been made to propose spatial point process models for CS data while still trying to account for biases in CS data collection. Chakraborty et al. (2011) presented CS data as the result of the degradation of the species occurrence caused by factors such as land use transformation and sampling effort. Fithian et al. (2015) proposed to explicitly account for sampling biases in CS data by understanding these data as a thinned point pattern. The models are specified as Log-Gaussian Cox Processes (LGCPs) and the biases are assumed to be log-linear. Other approaches have tried to fusion CS data and other types of biodiversity data so that they borrow strength from each other and thus better inferential and predictive performance can be achieved (Pacifi et al., 2017; Miller et al., 2019). However, Simmonds et al. (2020) highlight the necessity of accounting for the biases in CS data even when it is integrated with other data types.

As a response to the necessity of methods that make proper use of CS data, the main aim of this thesis is to contribute with new methods that account for the biases in the collection process of CS data. By accounting for these biases, we expect to produce better inferences of model parameters and prediction of ecological variables. We also aim to take advantage of the potential and quality of data from professional surveys by proposing methods to properly integrate them with CS data and other data types so that prediction accuracy is improved and uncertainty reduced. This thesis relies on three main hypotheses: i) CS data are a realization of a thinned point process, ii) accounting for biases in CS data collection process is necessary to make valid inference of the ecological variables of interest, and iii) integrating CS with other data sources can improve statistical inference and predictive performance. The appropriate use of methods to simultaneously utilize both CS and professional surveys data to answer several questions in ecology and biodiversity is still an open question for statisticians and practitioners. This thesis is composed of five papers (paper 1-5) that propose and develop methodological approaches that are useful for addressing ecological questions with models when only one data type type is

available, or when there are two (or more) data types to be integrated. Each paper offers a new contribution for better use of biodiversity data.

In many applied problems the only available source of information are CS data. Figuring out new methods for and how relevant accounting for the biases in comparison to not accounting for them is of paramount importance. In paper 1, we develop this comparison for both simulated and real data of moose observations with distance to roads as proxy for differences in accessibility, a recurring source of bias in the collection process of CS data (Fithian et al., 2015; Isaac et al., 2014). This bias is accounted for by incorporating appropriate functional forms to the likelihood of a Log-Gaussian Cox Process (LGCP) (Møller et al., 1998). Two functional forms are proposed in this paper, one that follows the half-normal detection function, typical of distance sampling (Yuan et al., 2017) and a more flexible specification that makes use of I-spline basis functions (i.e. monotonic non-increasing functions ,Ramsay (1988)). This is a relevant contribution to the use of CS data as it provides a simple way to account for one of the most fundamental sources of bias for CS data and shows how relevant it is to account for these biases as the posterior distribution of the ecological parameters are considerably less biased when the proposed model is used.

For other research questions multiple professional surveys are available. Gelfand and Shirota (2019); Miller et al. (2019) introduce modeling frameworks that rely on the idea of shared process models (Diggle et al., 2010; Wackernagel, 2003; Banerjee et al., 2008; Wang and Wall, 2003; Knorr-Held and Best, 2001) for integrating multiple data types. It assumes that the observed data types are realizations of one or more shared underlying Gaussian Random Fields (GRF). Based on this assumption, in paper 2 we propose modeling frameworks that fusion multiple bird monitoring surveys that are designed with different sampling protocols and cover disjoint portions of the space. By integrating multiple data types we enhance the predictive power of our models. This framework offers a novel way to use professional surveys to infer any ecological state beyond any defined national border. These ecological states can be used as input for other ecological models.

When data from professional surveys and CS data are available for solving research questions in ecology, fusioning these data types offers the possibility of covering larger portions of space and hence reducing prediction uncertainty and estimating more accurately parameters that drive the ecological process studied. (Simmonds et al., 2020; Wang et al., 2021). Integrating data types with

differences in their collection process requires that these processes are accounted for, in order to avoid biased inferences (Simmonds et al., 2020). In paper 3, we propose a flexible Bayesian spatial modeling framework for integrating CS data and professional surveys while accounting for biases in data collection, such as preferential sampling, differences in sampling effort, accessibility, among others. In addition to the fusion of these two data types, we also provide methods to explicitly account for these factors. Moreover, we address challenges in the implementation of these models such as identifiability issues that may arise as more parameters are introduced to the model and the fact that accounting for some of these biases introduces non-linear terms into the linear predictor of a point process model. We show the utility of these models through both simulation studies and a case study about powerline-induced death of birds. This paper offers both statistical solutions for performing data fusion of different types of spatial data, and also methodological tools based on data fusion to effectively tackle identifiability issues. The methods in this paper are accessible to practitioners by providing open-source code so that CS data can become more widely used by researchers in ecology and biodiversity.

The diversity of data types available in ecology and biodiversity offers the possibility of applying plenty of statistical methods to solve important research questions for the ecological community. For example, in paper 4, we have proposed a simple Bayesian spatial model for finding out if the presence of human populations across Europe represents a factor that determines the spatial distribution of big mammal species. Although simple, these models are of high value for ecologists as they are useful tools for making decisions for the conservation of species of mammals. The proposed model is useful for presence/absence data aggregated at large extensions. Given the size of the aggregated cells, methods for areal data were used in this paper. The results of this paper convey a simple, but powerful message for stakeholders in big mammals conservation: the spatial distribution of big mammals species is not affected by anthropogenic factors such as human disturbance or the existence of protected areas. Hence, it is up to human populations to decide whether or not to allow for coexistence with big mammals.

The volume and spatial coverage of CS data gives them great potential for aiding ongoing and future conservation programs. In paper 5, thousands of citizen science observations around the globe are used to produce species range of plant species. These species are characterized by not being as studied as, for example vertebrate species, despite their relevance for assessing anthropogenic impact

and for defining conservation priorities. This paper makes use of existing native regions for around 47,000 plants species and opportunistic occurrences reported by citizen scientist to predict the spatial range of plant species. The resulting dataset offers a simple tool to access spatial data of understudied species and thus support the conservation of plant species. Moreover, the resulting dataset shows how CS data can support the resolution of relevant research questions through proper management.

At least one Gaussian Random Field (GRF) is part of the models we propose. Performing Bayesian inference for these random effects is computationally expensive. Therefore, we mostly use the Integrated Nested Laplace Approximation (INLA) and the Stochastic Partial Differential Equation (SPDE) approach for fitting our models (Rue et al., 2009; Lindgren et al., 2011). While INLA reaches computational efficiency by producing a numerical approximation of the marginal posterior distribution of the parameters of the model, the SPDE approach is an efficient way of approximating a continuous spatial process by making use of a Gaussian Markov Random Field. The SPDE approach is especially relevant for efficiently approximating the likelihood of a LGCP as pointed out in Simpson et al. (2016). One of the most challenging tasks when fitting some of the proposed models is the fact that the likelihood of the LGCP is not log-linear with respect to the parameters of the model. Bachl et al. (2019) propose to solve this by iteratively linearizing the likelihood until the posterior mode has been found. As we expect the proposed modeling framework to be used by a broad group of practitioners, we use the Penalized Complexity (PC) prior (Simpson et al., 2017) as an intuitive way of specifying the prior distribution of the parameters of our models. PC priors are defined as probabilistic statements about the prior distribution of a parameter.

1.1 Spatial statistics

The models presented along these thesis have a spatial nature as they explicitly account for spatial autocorrelation by including both spatially referenced covariates and spatial random effects. The origins of spatial statistics are linked to the graphical depiction of data on maps and the solution of problems in agriculture during the early twentieth century (Cressie, 2015). Later, advances in

geostatistics (Matheron, 1963), spatial autocorrelation hypothesis testing (Cliff and Ord, 1973) and areal data models (Besag, 1974) set the basis for new methods in spatial statistics. In modern times, the increasing capacity of computers and Geographic Information Systems (GIS) have increased the popularity of spatial statistics among statisticians and practitioners. This has broadened the fields where spatial statistics is applied, including among others oil and mining, ecology and hidrology. A broader historical overview of spatial statistics is presented in Cressie (2015).

Banerjee et al. (2015) classify spatial data types in three big groups depending on how a random variable Y is observed and how the space is assumed:

- Geostatistical data: the space is assumed continuous as observations are collected at the point level in fixed locations defined by a sampling design.
- Lattice data: the space is discretized into areal units and a summary measure of Y is observed at each of them.
- Point pattern data: observations are collected at the point level, so the space is assumed continuous. Unlike geostatistical data, the locations of the events are assumed random.

Despite the differences between these data types, the notion of distance plays a fundamental role in any statistical spatial analysis. Its role is stated in Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things." (Tobler, 1970). Let \mathbf{s} vary over $D \subset \mathbb{R}^d$ to generate the process $Z(\mathbf{s})$, when D is assumed continuous the distance between two locations \mathbf{s}_i and \mathbf{s}_j affects the association between the process $Z(\mathbf{s})$ at these locations. The relation between distance and association is formally defined in the covariance function $Cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j))$.

In this thesis we focus mostly on models for geostatistical and point pattern data. Hence, in this section we introduce Gaussian Random Fields and spatial point patterns.

1.1.1 Gaussian Random Fields

A Random Field (RF), $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is as a collection of random variables that vary over D , a fixed subset of \mathbb{R}^d (Cressie, 2015; Banerjee et al., 2015). Along

this thesis we will assume $d = 2$. A RF is defined through the finite-dimensional distribution

$$F_{s_1, \dots, s_m}(z_1, \dots, z_m) = P(Z(s_1) \leq z_1, \dots, Z(s_m) \leq z_m) \quad m \geq 1 \quad (1.1)$$

If $F_{s_1, \dots, s_m}(z_1, \dots, z_m)$ is an m -variate Gaussian distribution, the process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is a Gaussian Random Field (GRF). In order to determine all distributions of a GRF, only the mean and covariance structure of $Z(\mathbf{s})$ need to be specified (Banerjee et al., 2015). The mean structure of $Z(\mathbf{s})$ is described by $\boldsymbol{\mu}(\mathbf{s}) = E(Z(\mathbf{s}))$ and the covariance structure by the covariance function $C(s_1, s_2) = Cov(Z(s_1), Z(s_2))$.

In the models we will present along this thesis GRFs are included as random effects to account for the dependency between spatial units. A GRF is stationary and isotropic if $\boldsymbol{\mu}(\mathbf{s}) = \boldsymbol{\mu}$ and $C(s_1, s_2) = Cov(\|s_2 - s_1\|)$ (Rue and Held, 2005). That is, a GRF is stationary and isotropic if the mean does not depend on the location and the covariance function depends on the Euclidean distance between a pair of locations, but not on the direction of the vector that defines the distance between these two points.

A covariance function is valid if it induces a positive covariance function. It means:

$$\sum_i \sum_j a_i a_j C(s_i, s_j) > 0 \quad (1.2)$$

for any coefficients $a_i, a_j \in \mathbb{R}$ and locations $s_i, s_j \in \mathbb{R}^2$ (Cressie, 2015). Valid covariance functions include, among others, exponential, Gaussian and Matérn covariance functions. In the models to be presented hereafter, the GRFs are assumed stationary, zero-mean, isotropic and with Matérn covariance function

$$C(s_i, s_j) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa \|s_j - s_i\|)^{\nu} K_{\nu}(\kappa \|s_j - s_i\|) \quad (1.3)$$

with $\nu > 0$ a parameter that defines the smoothness of the field and its mean-square differentiability (Blangiardo and Cameletti, 2015), K_{ν} the Bessel function of the second kind and order ν , σ^2 the marginal variance of the field, $\kappa > 0$ a scaling parameter, Γ the Gamma function and $\|s_j - s_i\|$ the Euclidean distance between the locations s_i and $s_j \in \mathbb{R}^d$. A Matérn covariance function can be characterized by its marginal standard deviation σ and its spatial range ρ . The spatial range for a stationary GRF is defined as the Euclidean distance

at which the correlation between two points is nearly 0.1 (Rue and Held, 2005). Therefore, in the models proposed in this thesis, σ and ρ will be the parameters estimated for the GRFs. The parameter ν is usually difficult to identify, hence it is usually fixed (Lindgren et al., 2011).

1.1.1.1 Gaussian Markov Random Fields

As the number of observations increase, performing inference for GRFs imply high computational burden as they are linked to continuously indexed locations (Rue and Held, 2005). Available solutions in literature include, among others, covariance tapering (Furrer et al., 2006), likelihood approximation (Stein et al., 2004) and low rank approximations (Banerjee et al., 2008; Eidsvik et al., 2012). Gaussian Markov Random Fields (GMRFs) offer a computationally efficient solution to the so called big N problem as they are discretely indexed. They are based on the Markov property which means they consider a neighboring structure, which ensures the sparsity of the precision matrices.

A GMRF can be represented via a graph that represents the nonzero pattern of the precision matrix. Undirected graphs are useful to understand the notion of GMRFs. An undirected graph \mathcal{G} is defined as a tuple $\mathcal{G} = \mathcal{V}, \mathcal{E}$ where \mathcal{V} is the set of nodes in the graph and \mathcal{E} is the set of edges $\{i, j\}$, where $i, j \in \mathcal{V}$ and $i \neq j$. If $\{i, j\} \in \mathcal{E}$ there is an undirected edge from edge i to node j . Let $x = (x_1, \dots, x_n)^T$ be normally distributed and $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, \dots, n\}$ and \mathcal{E} such that there is no edge between node i and j if and only if $x_i \perp x_j | \mathbf{x}_{-ij}$. The vector $x = (x_1, \dots, x_n)^T$ is a GMRF with respect to \mathcal{G} with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$ if and only if its density has the form

$$\pi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (1.4)$$

with $Q_{ij} \neq 0$ if and only if $\{i, j\} \in \mathcal{E}$ for all $i \neq j$. The next Markov properties of the GMRF \mathbf{x} with respect to the graph \mathcal{G} are equivalent:

The pair Markov property:

$$x_i \perp x_j | \mathbf{x}_{-ij} \quad \text{if } \{i, j\} \notin \mathcal{E} \text{ and } i \neq j$$

The local Markov property:

$$x_i \perp \mathbf{x}_{-i, ne(i)} | \mathbf{x}_{ne(i)} \quad \text{for every } i \in \mathcal{V}$$

with $ne(i)$, the nodes in \mathcal{G} that have an edge to node i .
The global Markov property:

$$\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$$

for all disjoint A,B and C where C separates A and B, and A and B are non-empty.

The Markov properties imply that the neighboring structure of a GMRF will only be given by those non-zero entries of the precision matrix \mathbf{Q} , so techniques such as the Stochastic Partial Differential Approach (SPDE) approach (see Sec. 1.2.4.1, Lindgren et al. (2011)) for approximating continuous GRFs can rely on the sparseness of \mathbf{Q} to be efficient computationally.

1.1.2 Spatial point patterns

As previously mentioned, the locations of a spatial point pattern are random. Spatial point processes aim to analyse the geometrical structure of spatially randomly distributed points (Illian et al., 2008). If each of these points contain information, hereafter marks, we say the point pattern is marked. Unless otherwise is mentioned, we will assume our point patterns have no marks. For a bounded region $D \subset \mathbb{R}^2$, both the amount of points in D , $N(D)$, and the location of these points in the point pattern \mathcal{S} are random (Gelfand and Schliep, 2018). As a random variable, an expected value, $E(N(D))$, can be considered for $N(D)$. Following the notation in Gelfand and Schliep (2018):

$$\lambda(D) = E(N(D))$$

with $\lambda(D)$ the intensity measure, which is defined as:

$$\lambda(D) = \int_D \lambda(s) ds \tag{1.5}$$

where $\lambda(s)$ is the intensity function.

To model probabilistically \mathcal{S} , a distribution is needed for $N(D)$ and a multivariate location density over D^n for any n . The Poisson process is the most used model for point patterns. Thus, we assume for any $D \subset \mathbb{R}^2$, $N(D) \sim Poisson(\lambda(D))$. Under a Poisson process, for disjoint sets D_1 and D_2 , $N(D_1)$

and $N(D_2)$ are independent. Hence, under a Poisson process we have a conditionally independent location distribution. A key property of Poisson processes is:

$$P(N(\partial\mathbf{s}) = 1) \approx E(N(\partial\mathbf{s})) = \lambda(\partial\mathbf{s}) \approx \lambda(\mathbf{s})|\partial\mathbf{s}|$$

with $\partial\mathbf{s}$ a small circular neighborhood around \mathbf{s} (Gelfand and Schliep, 2018). Depending on the nature of $\lambda(\mathbf{s})$, a Poisson process can be homogeneous or inhomogeneous. A homogeneous Poisson process (HPP) assumes $\lambda(\mathbf{s}) = \lambda$, while a nonhomogeneous Poisson process (NHPP) assumes $\lambda(\mathbf{s})$ is not constant. When a point process is homogeneous, it exhibits complete spatial randomness. Thus, the distribution of points is expected to be uniform across space. Complete spatial randomness represents a baseline for point process theory, as for example the exploratory tools to determine whether the assumption of complete spatial randomness holds, see Illian et al. (2008) for further details. If a point pattern is nonhomogeneous, i.e. the assumption of complete spatial randomness does not hold, processes of attraction or repulsion between point might happen. These point patterns can be characterized through first- and second-order properties. Most of the interest in the study of nonhomogeneous spatial point patterns is focused on modeling the intensity function $\lambda(\mathbf{s})$. Approaches for modeling $\lambda(\mathbf{s})$ include those that depend on a bivariate density function or on spline surfaces. The most common way of modeling $\lambda(\mathbf{s})$ is by assuming that a group of spatial covariates, $x^T(\mathbf{s})$, drive it. That is, assuming $\log \lambda(\mathbf{s}) = x^T(\mathbf{s})\beta$ (Gelfand and Schliep, 2018). Understanding $\lambda(\mathbf{s})$ as a Gaussian process (see Section 1.1.1), gives rise to the Log Gaussian Cox Process (LGCP) (Møller et al., 1998). It has the typical expression of a Cox process, $\lambda(\mathbf{s}) = g(x^T(\mathbf{s})\beta)\lambda_0(\mathbf{s})$ with $g(\cdot) > 0$ and $\lambda_0(\mathbf{s})$ a local adjustment process with $Z(\mathbf{s}) \equiv \log(\lambda_0(\mathbf{s}))$ a Gaussian process (Gelfand and Schliep, 2018).

We now focus on LGCPs as they provide an interpretable expression for the intensity function. Assume $N(D) = n$, then the location density $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$ of a point pattern is given by (Gelfand and Schliep, 2018):

$$f(\mathbf{s}_1, \dots, \mathbf{s}_n | N(D) = n) = \prod_i \frac{\lambda(\mathbf{s}_i)}{\lambda(D)}$$

As mentioned above, we assume $N(D) \sim \text{Poisson}(\lambda(D))$, thus:

$$f(\mathbf{s}_1, \dots, \mathbf{s}_n, N(D) = n) = \prod_i \frac{\lambda(\mathbf{s}_i)}{\lambda(D)} (\lambda(D))^n \frac{e^{-\lambda(D)}}{n!}$$

Thus, the likelihood of a LGCP is expressed as:

$$L(\{\lambda(\mathbf{s}), \mathbf{s} \in D\}; \mathcal{S}_{obs}) = e^{-\lambda(D)} \prod_i \lambda(\mathbf{s}_i) \quad (1.6)$$

where $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$. Note that part of the likelihood depends on an integral over the whole area D , which makes computations challenging as discussed in upcoming sections. For more details about the theory of point processes see Gelfand and Schliep (2018); Illian et al. (2008).

Even though we assume the spatial distribution of species arguably follows a spatio(-temporal) point process model, one of our working hypotheses is that CS data are a degraded version of the reality (Chakraborty et al., 2011). Therefore, we regard CS data as realizations of a thinned point pattern.

Thinned point pattern

Observed point patterns can be generated after some basic operations are performed on other point patterns. These operations include for example, thinning, clustering and superposition (Illian et al., 2008). A point pattern \mathcal{S}_t is thinned when a specified rule determines which points of the original point pattern \mathcal{S} are deleted. These rules are classified in Illian et al. (2008) as:

p -thinning : Each point is deleted with probability $1 - p$. This rule does not depend on the location of the points, or on the deletion of other points in \mathcal{S}

$p(\mathbf{s})$ -thinning : Now, a deterministic function $p(\mathbf{s})$, which depends on the location \mathbf{s} , determines whether or not a point is deleted. For this rule the deletion of a point in \mathcal{S} is independent of the deletion of other points.

$P(\mathbf{s})$ -thinning : Still a thinning where the deletion of one point in H has no association with the deletion of the other points in the point pattern. However, it generalizes $p(\mathbf{s})$ -thinning as $p(\mathbf{s})$ is not deterministic anymore. Now, it is assumed that $p(\mathbf{s})$ is the realization of a random process $P(\mathbf{s})$, which is independent of \mathcal{S} .

As the process \mathcal{S}_t is the result of a thinning operation on the process \mathcal{S} , its first- and second-order properties depend on the properties of the process \mathcal{S}

(Illian et al., 2008). If $\lambda(\mathbf{s})$ is the intensity of the process \mathcal{S} , then the intensity of \mathcal{S}_t will be:

$$\lambda_t(\mathbf{s}) = p(\mathbf{s})\lambda(\mathbf{s}) \quad (1.7)$$

For a stationary process $P(\mathbf{s})$ with mean p , the intensity of the $P(\mathbf{s})$ -thinned process would be given by

$$\lambda_t(\mathbf{s}) = p\lambda(\mathbf{s})$$

Note that When Poisson processes are thinned, the resulting point pattern remains in the family of Poisson processes. This introduction to thinned point patterns has been completely based on Section 6.2.1 of Illian et al. (2008), where more technical details about these type of point processes are presented.

1.2 Bayesian Statistics

For the research questions we aim to solve in this thesis, quantifying the uncertainty is relevant to communicate the results obtained. Furthermore, in many real-life scenarios experts have prior knowledge about some of the parameters we want to estimate. Bayesian statistics offer an appealing methodological approach to learn about these parameters. The origins of Bayesian statistics date back to 1763 when Bayes' theorem was proposed by Thomas Bayes. During the nineteenth century the lack of clarity about the concept of prior distribution stopped the progress of Bayesian statistics. It was only during the end of the twentieth century that Bayesian methods became popular between statisticians and practitioners as computers with more processing capacity were released and new computational approaches for Bayesian statistics were developed. For more details about the history of Bayesian statistics see Gamerman and Lopes (2006); Carlin and Louis (2008). In this section we will go through the basic elements of Bayesian statistics. Then, we go into more detail about prior distributions, hierarchical modelling and computational methods for Bayesian statistics. In particular, we focus our interest on the class of Latent Gaussian Models and the INLA-SPDE approach.

Whereas in the traditional frequentist setting, parameters $\boldsymbol{\theta}$ are regarded as fixed quantities, in the Bayesian setting the elements of $\boldsymbol{\theta}$ are random variables. Assume we are interested in performing inference on $\boldsymbol{\theta}$ and we have prior knowledge about $\boldsymbol{\theta}$ expressed in a prior distribution $\pi(\boldsymbol{\theta})$ and data denoted as \mathbf{y} . Bayesian analysis aims to update one's prior belief based on Bayes' theorem

(Givens and Hoeting, 2012; Gamerman and Lopes, 2006):

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})}{\pi(\mathbf{y})} \propto \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta}) \quad (1.8)$$

with $\pi(\mathbf{y}|\boldsymbol{\theta})$ the likelihood function and $\pi(\boldsymbol{\theta}|\mathbf{y})$ the posterior distribution, which is used to perform statistical inference on $\boldsymbol{\theta}$. Summary measures as the posterior mean, median and variance are obtained from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ (Gamerman and Lopes, 2006). In addition to the posterior distribution of the parameters in $\boldsymbol{\theta}$, prediction can be performed within the Bayesian setting through the predictive distribution of a new observation \mathbf{y}^* :

$$\pi(\mathbf{y}^*|\mathbf{y}) = \int_{\Theta} \pi(\mathbf{y}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \quad (1.9)$$

1.2.1 Prior distribution

The definition of a prior distribution is a fundamental step when Bayesian analysis is performed. It can reflect some previous knowledge regarding the (hyper)parameters involved in the model, regardless of how certain it is. In some cases the prior distributions reflect information from past studies, subjective opinions from experts, belong to a familiar distributional family, or in cases with too little information, it is defined so that the data become the most influential part in the posterior distribution (Carlin and Louis, 2008).

When there is some information about the parameters to be estimated, prior elicitation (Carlin and Louis, 2008; Wolfson, 1995; Chaloner, 1996) is performed in order to propose a prior distribution that reflect this knowledge (O’Hagan and Kendall, 1994). However, elicitation is not as widely used as it could be due to the lack of methods that fit into the Bayesian workflow and can be performed efficiently (Mikkola et al., 2021). In other cases a computationally convenient choice is to select a prior distribution that is conjugate to the likelihood distribution so that the resulting posterior distribution is part of the same family as the prior distribution. Within the exponential family there are plenty of conjugate priors (Carlin and Louis, 2008; Gamerman and Lopes, 2006). In case a conjugate prior lacks flexibility to reflect any existing prior knowledge, a finite mixture of conjugate priors might reach the desired flexibility while still preserving the computational simplicity of the resulting posterior distribution.

In case little knowledge about the parameters we want to estimate is available, non-informative priors can be defined. This is a controversial topic among Bayesians as some non-informative priors are improper, which in some cases leads to improper posteriors. That is, distributions that do not integrate to 1. Two broadly accepted non-informative prior distributions are Jeffrey's prior (Jeffreys, 1998) and the ones resulting from reference approach proposed by Bernardo (1979). Jeffrey's prior is based on the Fisher information matrix and is invariant to parametric transformations. Reference priors are based on expected discrepancy measures of information. Given the chance of getting an improper posterior, it is recommended to pick carefully non-informative priors (Carlin and Louis, 2008; Gamerman and Lopes, 2006). When little prior information is available, it is also usual to propose priors with known functional forms, but with large prior variance so that this distribution has little or null influence on the construction of the posterior distribution.

Recently, Simpson et al. (2017) proposed the Penalized Complexity (PC) priors. These prior distributions are the result of penalizing models that deviate from a simple base model. The construction of a PC prior distribution relies on four principles. First, the base model is preferred over more complex models. Second, complexity is measured by the Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951), which is a measure of the information lost when a complex model is approximated by the base model and used as a measure of distance between two models. Third, the penalization to more distant models according to KLD measure is made using a constant decay-rate so that deviations from the base model are equally penalized across the whole parameter space. Finally, the user has an idea about values of the parameter of interest. This idea can be expressed through a probabilistic statement of the form:

$$\text{Prob}(Q(\xi) > U) = \alpha$$

with $Q(\xi)$ an interpretable transformation of the parameter of interest, U an user-defined upper bound that defines a tail event and $\alpha \in (0, 1)$ the weight assigned to this event. Say, for example, we want to construct a PC prior for the precision of a Gaussian random effect $x \sim \mathcal{N}(\mathbf{0}, \tau^{-1})$ (Simpson et al., 2017). The PC prior for τ is then given by:

$$\pi(\tau) = \frac{\lambda}{2} \tau^{-3/2} \exp\left(-\lambda \tau^{-1/2}\right), \quad \tau > 0, \lambda > 0 \quad (1.10)$$

where λ determines the magnitude of the penalty for deviating from the base model. λ can be determined by specifying (U, α) so that $\text{Prob}(1/\sqrt{\tau} > U) = \alpha$.

Hence, $\lambda = -\ln(\alpha)/U$. In this way a simple probabilistic statement about the standard deviation $1/\sqrt{\tau}$ in terms of U and α can construct the prior distribution of τ .

Within spatial statistics, Fuglstad et al. (2019) constructed PC priors for both the spatial range, ρ , and the marginal standard deviation, σ , of a Gaussian Random Field with Matérn covariance function (see Eq. (1.3)). As previously presented, the joint prior distribution of ρ and σ can be obtained from simple probabilistic statements about (functions of) the parameters of interest. In this case, the user has to specify the pairs of quantiles and probabilities (ρ_0, α_1) and (σ_0, α_2) in:

$$P(\rho < \rho_0) = \alpha_1 \quad P(\sigma > \sigma_0) = \alpha_2 \quad (1.11)$$

Even though major advances have been made in the definition of prior distributions this is still an open research topic, especially in the definition of the prior distributions of the parameters of the GRFs. Sørbye et al. (2019) point out the relevance of appropriate, careful prior specification of the spatial hyperparameters in Log-Gaussian Cox Processes and suggest the necessity of having clear interpretation and communication of the prior choices of the spatial hyperparameters, which is facilitated by clear probabilistic statements made for PC priors.

PC priors in Fuglstad et al. (2019) will be used for the parameters of the GRFs of the models proposed along this thesis. As little is known about the fixed effects that are part of the model, non-informative prior distributions are defined for them.

1.2.2 Latent Gaussian Models

Now, assume the prior distribution $\pi(\boldsymbol{\theta})$ depends on a parameter $\boldsymbol{\psi}$. Then, the posterior distribution of $\boldsymbol{\theta}$ can be expressed as:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\int \pi(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})d\boldsymbol{\psi}}{\pi(\mathbf{y})} \quad (1.12)$$

This adds a new stage represented by the distribution $\pi(\boldsymbol{\psi})$ which is a prior distribution for the so called hyperparameter $\boldsymbol{\psi}$. Usually $\boldsymbol{\psi}$ is not known and the posterior distribution $\pi(\boldsymbol{\psi}|\mathbf{y})$ becomes interesting for inferential purposes.

An alternative is to estimate $\boldsymbol{\psi}$ as a maximizer, $\hat{\boldsymbol{\psi}}$, for the marginal distribution $\pi(\mathbf{y}|\boldsymbol{\psi}) = \int \pi(\mathbf{y}|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\psi})d\boldsymbol{\theta}$ (Banerjee et al., 2015) and then treat $\boldsymbol{\psi}$ as known, this is known as Empirical Bayes analysis (Carlin and Louis, 2008). Note that plugging $\hat{\boldsymbol{\psi}}$ into Eq. (1.12) implies not accounting for the uncertainty of $\boldsymbol{\psi}$ in the updating process described by Eq. (1.8). This modelling is called hierarchical because the addition of the hyperparameter $\boldsymbol{\psi}$ and its prior distribution $\pi(\boldsymbol{\psi})$ represent a new layer in the hierarchy.

Latent Gaussian Models (LGMs) (Rue et al., 2009) have a hierarchical structure. These models are a subclass of the class of Bayesian additive models with a response variable y_i whose distribution belongs to the exponential family and has a mean μ_i which is linked to a linear predictor η_i as follows:

$$g(\mu_i) = \eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \quad (1.13)$$

with g a link function, α an intercept, $f^{(j)}(\cdot)$ unknown functions of u_{ij} that can be used to model, among others, spatial correlation structures, β_{ki} the linear effects of covariates \mathbf{z} , and ε_i unstructured terms (Rue et al., 2009). LGMs assign Gaussian prior distributions to $\alpha, \{f^{(j)}(\cdot)\}, \{\beta_k\}$ and $\{\varepsilon_i\}$. We define $\mathbf{x} = (\alpha, \{f^{(j)}(\cdot)\}, \{\beta_k\}, \{\varepsilon_i\})$ as the latent Gaussian field. Associated to each element of \mathbf{x} there are hyperparameters $\boldsymbol{\theta}$ which do not have to be Gaussian. In LGMs the density $\pi(\mathbf{x}|\boldsymbol{\theta}_1)$ is Gaussian with zero mean and precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$. The response variables y_i have distribution $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_2)$ and are assumed conditionally independent given \mathbf{x} and $\boldsymbol{\theta}_2$. Based on this hierarchical structure the posterior distribution is computed as

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{i \in \mathbb{I}} \pi(y_i|x_i, \boldsymbol{\theta}) \quad (1.14)$$

$$\propto \pi(\boldsymbol{\theta})|\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp \left[-\frac{1}{2}\mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_{i \in \mathbb{I}} \log\{\pi(y_i|x_i, \boldsymbol{\theta})\} \right] \quad (1.15)$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Our aim is now to estimate $\pi(x_i|\mathbf{y})$, $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(\theta_j|\mathbf{y})$.

1.2.3 Computational methods in Bayesian statistics

Bayesian methods are most frequently used to compute the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ (see Eq. (1.12)). In most of the cases this is a complex task as

conjugacy is difficult to achieve, or computing $\pi(\mathbf{y}) = \int \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ involves an integral over a large parametric space. In such situations, simulation methods offer an appealing alternative to learn about $\pi(\boldsymbol{\theta}|\mathbf{y})$. These methods include, for example, Monte Carlo (MC) techniques, which involve the simulation of independent samples of the posterior distribution (Blangiardo and Cameletti, 2015; Gamerman and Lopes, 2006; Givens and Hoeting, 2012). When the posterior distribution does not have a standard functional form, or the parameter space is large, Markov Chain Monte Carlo (MCMC) methods offer an alternative based on the simulation of a Markov Chain whose stationary distribution is the posterior distribution (Blangiardo and Cameletti, 2015; Gamerman and Lopes, 2006; Givens and Hoeting, 2012). Some of the most used MCMC techniques include among others, the Gibbs sampler (Casella and George, 1992) and the Metropolis-Hastings algorithm (Gamerman and Lopes, 2006). For more details of MCMC methods see (Gamerman and Lopes, 2006; Robert et al., 1999).

In particular, spatial models have proven to demand many computational resources as the dimensions of the parameter space for these models is usually large (Moraga et al., 2021), which results in long computation times. Approximate methods offer an efficient alternative to MCMC. Example of approximate methods include Approximate Bayesian Computation (ABC) (Beaumont et al., 2009; Grazian and Fan, 2020), variational methods (Jaakkola and Jordan, 2000) and empirical likelihood (Owen, 2001). As we deal with LGMs in this thesis, then an approximated method such as the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) offers an efficient, yet accurate, approach to estimate $\pi(\boldsymbol{\theta}|\mathbf{y})$.

1.2.3.1 The Integrated Nested Laplace Approximation (INLA)

As previously mentioned one of the most relevant and demanding tasks when fitting Bayesian models is to estimate the posterior distributions $\pi(\boldsymbol{\theta}|\mathbf{y})$. Traditionally inference for LGMs was made using Markov Chain Monte Carlo (MCMC) methods. As mentioned in Rue et al. (2009), Bayesian inference for LGMs using MCMC has performance issues due to the the high correlation among the elements of the latent Gaussian field \mathbf{x} , and the high correlation between the elements of \mathbf{x} and the hyperparameters $\boldsymbol{\theta}$. A joint proposal based on a Gaussian approximation of the full conditional of \mathbf{x} has been made by

Gamerman (1997); Knorr-Held et al. (2002); Rue et al. (2004) to overcome the problems caused by the strong dependence between the elements of \mathbf{x} . To deal with the strong correlation between \mathbf{x} and $\boldsymbol{\theta}$ a block update of \mathbf{x} and $\boldsymbol{\theta}$ from $\boldsymbol{\theta}$ and $\mathbf{x}|\boldsymbol{\theta}$ is proposed in Knorr-Held et al. (2002). Despite these solutions to the poor performance of MCMC methods for LGMs, these techniques are still computationally slow.

The Integrated Nested Laplace Approximation (INLA), introduced by Rue et al. (2009), is a deterministic algorithm based on Gaussian approximations to the posterior densities, which is reasonable given the prior specification of the latent field for LGMs. We are interested in the posterior marginal distributions:

$$\begin{aligned}\pi(x_i|\mathbf{y}) &= \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ \pi(\boldsymbol{\theta}_j|\mathbf{y}) &= \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}\end{aligned}\tag{1.16}$$

with \mathbf{x} the latent field (see Section 1.2.2) and $\boldsymbol{\theta}$ the hyperparameters of the model. The INLA approach proposes the next approximations to posteriors in Eq. (1.16):

$$\begin{aligned}\tilde{\pi}(x_i|\mathbf{y}) &= \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ \tilde{\pi}(\boldsymbol{\theta}_j|\mathbf{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}\end{aligned}\tag{1.17}$$

These approximations to the posterior distributions rely on the Laplace approximation of $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ which is used to approximate $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. Hence,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \approx \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Bigg|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}\tag{1.18}$$

with $\mathbf{x}^*(\boldsymbol{\theta})$ the mode of $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ for a given $\boldsymbol{\theta}$. $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is a reasonable approximation as $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is almost Gaussian (Rue et al., 2009). The other approximation we need to compute is $\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})$. Computing this approximation is more complex than approximating $\pi(\boldsymbol{\theta}|\mathbf{y})$ because the dimension of \mathbf{x} is larger than the dimension of $\boldsymbol{\theta}$. Three ways of solving this task are proposed by Rue and Held (2005). The first one is to use Gaussian approximations $\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})$ as the marginals of $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ in Eq. (1.18). However, this approach can lead to inaccurate approximations (Blangiardo and Cameletti, 2015). The second

approach is to use Laplace approximations to compute:

$$\pi(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{\pi(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \approx \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})} \quad (1.19)$$

with $\tilde{\pi}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ the Laplace Gaussian approximation to $\pi(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ and $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$ its mode. Despite the good quality of this approach, this strategy can become computationally expensive as $\tilde{\pi}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ needs to be computed for each value of \mathbf{x} and $\boldsymbol{\theta}$ (Blangiardo and Cameletti, 2015). The third approach is the simplified Laplace approximation, which is based on a Taylor's series expansion of the Laplace approximation $\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})$ in Eq. (1.19). This approach is good enough in many applications and computationally is more efficient than the other two approaches (Blangiardo and Cameletti, 2015).

After $\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})$ and $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ have been computed, the marginal posteriors $\pi(x_i|\mathbf{y})$ in Eq. (1.16) are approximated as the finite weighted sum:

$$\tilde{\pi}(x_i|\mathbf{y}) \approx \sum_j \tilde{\pi}(x_i|\boldsymbol{\theta}^{(j)}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}^{(j)}|\mathbf{y}) \Delta_j$$

for integration points $\{\boldsymbol{\theta}^{(j)}\}$ with weights $\{\Delta_j\}$. Finding the integration points $\{\boldsymbol{\theta}^{(j)}\}$ is a demanding task as this requires an exploration of the joint posterior distribution $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. Two exploration schemes were proposed by Rue et al. (2009), the grid strategy and the central composite design. The grid strategy is computationally expensive and its cost increases exponentially with the number of hyperparameters. Hence, it is recommended to have at most 4 hyperparameters. The central composite design uses the mode $\boldsymbol{\theta}^*$ and the Hessian to determine the integration points $\boldsymbol{\theta}$ to perform a second-order approximation to a response variable. For more details about INLA, see Rue et al. (2009); Blangiardo and Cameletti (2015).

1.2.4 The Stochastic Partial Differential Equation (SPDE) approach

The models we propose along this thesis explain spatial autocorrelation through Gaussian Random Fields. Factorizing the $n \times n$ (n =number of observations) dense covariance matrix of a GRF requires performing $\mathcal{O}(n^3)$ operations, this is

known as the big n problem. Solutions to this problem include, among others, approaches that approximate the likelihood through a sequential representation (Vecchia, 1988; Stein et al., 2004), doing exact computations on a simplified Gaussian model of low rank (Banerjee et al., 2008; Eidsvik et al., 2012) or applying covariance tapering to make covariance matrices sparse (Furrer et al., 2006). Lindgren et al. (2011) propose the representation of a continuous spatial process (e.g. a Gaussian Random Field) as a discretely indexed spatial random process by making use of the Stochastic Partial Differential Equation (SPDE) approach. This approach substitutes a GRF with a GMRF in order to achieve a sparse representation of the covariance matrix and thus reach higher computational efficiency. The SPDE approach starts from the definition of the SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau\omega(\mathbf{s})) = \mathcal{W}(\mathbf{s}), \quad \mathbf{s} \in \mathcal{R}^d, \quad \kappa > 0 \quad (1.20)$$

whose stationary solution is the GRF $\omega(\mathbf{s})$ with Matérn covariance function (see Eq. (1.3)). Eq. (1.20) also includes κ the scale parameter of the Matérn covariance function, Δ the Laplacian, α the parameter that controls the smoothness, τ which controls the variance, $\mathcal{W}(\mathbf{s})$ a Gaussian spatial white noise process, and d the dimension of the spatial domain. The parameters in the SPDE in Eq. (1.20) are related to the parameters of the Matérn covariance function (the marginal variance σ^2 and the smoothness parameter ν) through:

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2} \quad (1.21)$$

$$\nu = \alpha - d/2$$

The solution to the SPDE in Eq. (1.20), $\omega(\mathbf{s})$, can be approximated using the finite element method (Ciarlet, 2002; Brenner et al., 2008; Quarteroni and Valli, 2008) through the basis representation:

$$\omega(\mathbf{s}) = \sum_{k=1}^m \phi_k(\mathbf{s})w_k \quad (1.22)$$

In this representation m is the total number of vertices in a triangulation of the spatial domain \mathcal{D} (see Fig. 1.1), $\phi_k(\mathbf{s})$ represents a set of deterministic basis functions and w_k are weights which are normally distributed with zero mean. In order to achieve computational efficiency, the basis functions are chosen to be piecewise linear in each triangle and to have local support. That is, $\phi_k(\mathbf{s})$ is

1 at vertex k and 0 elsewhere. In the case $\alpha = 2$, the precision matrix \mathbf{Q} for the vector $\mathbf{w} = \{w_1, \dots, w_m\}$ is given by:

$$\mathbf{Q} = \tau^2 (\kappa^4 \mathbf{C} + 2\kappa^2 \mathbf{G} + \mathbf{G}\mathbf{C}^{-1}\mathbf{G}) \quad (1.23)$$

with \mathbf{C} a matrix with generic element $C_{ii} = \int \phi_i(\mathbf{s})d\mathbf{s}$ and \mathbf{G} with generic element $G_{ij} = \int \nabla \phi_i(\mathbf{s})\nabla \phi_j(\mathbf{s})d\mathbf{s}$. As the precision matrix \mathbf{Q} is sparse, then \mathbf{w} is a GMRF with Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$. The sparsity of the matrix \mathbf{Q} makes the SPDE approach specially useful for fitting spatial models efficiently. Thus, the SPDE approach and INLA have been used in the models of this thesis to estimate their parameters.

1.2.4.1 The SPDE approach and Log-Gaussian Cox Processes

Log-Gaussian Cox Processes (LGCPs) (see Sec. 1.1.2) are a function of a GRF. Hence, estimating their parameters can be computationally expensive. The likelihood of a LGCP, presented in Eq. (1.6), is intractable as it depends on an integral over the whole study area. Little has been written on model fitting and model comparison of LGCP models (Illian et al., 2012). MCMC methods have been proposed to fit LGCPs, but given their iterative nature they can become computationally expensive (Møller and Waagepetersen, 2003; Møller and Waagepetersen, 2007). As LGCPs are part of the family of LGMs, they can be fitted using the INLA-SPDE approach as introduced by Illian et al. (2012). This approach is based on a discretization of the spatial domain in grid cells and the definition of a Poisson random variable which approximates the true LGCP likelihood. The quality of the inferences produced by this technique depends on the resolution of the grid cells, and can be computationally wasteful if the process intensity is high or the observation window is large or oddly shaped (Simpson et al., 2016).

Simpson et al. (2016) propose a new method for developing computational inference on LGCPs. As mentioned before, the likelihood of a LGCP contains an integral over the whole spatial domain of the point process, which makes this likelihood hard to handle. Instead of defining the GRF over a lattice as in Illian et al. (2012), it is defined as:

$$\omega(\mathbf{s}) = \sum_{j=1}^n z_j \phi_j(\mathbf{s}) \quad (1.24)$$

where $\mathbf{z} = (z_1, \dots, z_n)^T$ is a multivariate Gaussian vector and $\{\phi_i(\mathbf{s})\}_{i=1}^n$ is a set of deterministic basis functions. The expression in Eq. (1.24) resembles Eq.(1.22) with $m = n$ and $\mathbf{z} = \mathbf{w}$. The integral in Eq.(1.6) can be written as:

$$\int_D \lambda(\mathbf{s}) d\mathbf{s} = \int_D \exp\{\omega(\mathbf{s})\} d\mathbf{s} \approx \sum_{i=1}^p \tilde{\alpha}_i \exp\{\omega(\tilde{s}_i)\} \quad (1.25)$$

Based on the Equations (1.24) and (1.25), the log-likelihood in Eq. (1.6) can be expressed as:

$$\log(\pi(y|z)) \approx - \sum_{i=1}^p \tilde{\alpha}_i \exp\left\{ \sum_{j=1}^n z_j \phi_j(\tilde{s}_i) \right\} + \sum_{i=1}^N \sum_{j=1}^n z_j \phi_j(s_i) \quad (1.26)$$

which can be rewritten in matrix form as:

$$\log(\pi(y|z)) \approx -\tilde{\boldsymbol{\alpha}}^T \exp(A_1 \mathbf{z}) + \mathbf{1}^T A_2 \mathbf{z} \quad (1.27)$$

with $[A_1]_{ij} = \phi_j(\tilde{s}_i)$ a matrix that contains the values of the basis functions in Eq. (1.24) at the integration nodes \tilde{s}_i and $[A_2]_{ij} = \phi_j(s_i)$ that contains the basis functions in Eq. (1.24), evaluated at the observation points.

Let $\log \eta = (\mathbf{z}^T A_1^T, \mathbf{z}^T A_2^T)^T$, $\boldsymbol{\alpha} = (\tilde{\boldsymbol{\alpha}}^T, \mathbf{0}^T)^T$ and $\mathbf{y} = (\mathbf{0}^T, \mathbf{1}^T)^T$ a vector of pseudo-observations. Then, the likelihood in (1.27) can be expressed as:

$$\pi(\mathbf{y}|\mathbf{z}) \approx \prod_{i=1}^{N+p} \eta_i^{y_i} \exp(-\alpha_i \eta_i) \quad (1.28)$$

which is similar to a likelihood for $N + p$ conditionally independent Poisson random variables with means $\alpha_i \eta_i$ and observed values y_i .

In practice, the integration points \tilde{s}_i and the weights $\tilde{\alpha}_i$ are based on a dual mesh constructed from the mesh of the SPDE approach (Lindgren et al., 2011) (see Figure 1.1). The centroids of the dual mesh are selected as the integration points and the area of each polygon as the weights $\tilde{\alpha}_i$ (Simpson et al., 2016). This approximation has optimal convergence as the mesh in Figure 1.1 is refined. Furthermore, as this approximation can be framed within the INLA-SPDE approach, the whole fitting process of a LGCP can be done efficiently.

1.2.5 Nonlinear functions in the linear predictor: the *inlabru* approach

As stated in the introduction of this thesis, one of our contributions is to account for biases inherent to Citizen Science data by regarding these data as a thinned point pattern. Each possible source of bias acts as a thinning operation on the actual point pattern. Modelling these biases imply the inclusion of nonlinear functions into the linear predictor of a traditional Log-Gaussian Cox Process. In order to fit such models, Bachl et al. (2019) have proposed an iterative approach for linearizing these terms, which is ready to use in the R-package *inlabru* (Bachl et al., 2019). In this section we present in detail the iterative approach introduced in Bachl et al. (2019) and used, for example, to model thinned points patterns affected by distance sampling (Yuan et al., 2017).

The idea of *inlabru* is to add a linearization step to the traditional INLA-SPDE approach for fitting LGMs. Let $\tilde{\boldsymbol{\eta}}(\mathbf{u})$ be a non-linear predictor. In the *inlabru* approach it is approximated by the first order Taylor approximation at \mathbf{u}_0 , $\bar{\boldsymbol{\eta}}(\mathbf{u})$, given by:

$$\bar{\boldsymbol{\eta}}(\mathbf{u}) = \tilde{\boldsymbol{\eta}}(\mathbf{u}_0) + B(\mathbf{u} - \mathbf{u}_0) = [\tilde{\boldsymbol{\eta}}(\mathbf{u}_0) - B(\mathbf{u}_0)] + B(\mathbf{u}) \quad (1.29)$$

with B the derivative matrix evaluated at \mathbf{u}_0 . Once the approximation $\bar{\boldsymbol{\eta}}(\mathbf{u})$ has been computed, the non-linear observation model $p(\mathbf{y}|g^{-1}[\tilde{\boldsymbol{\eta}}(\mathbf{u}, \boldsymbol{\theta})])$ is approximated by $p(\mathbf{y}|g^{-1}[\bar{\boldsymbol{\eta}}(\mathbf{u}, \boldsymbol{\theta})])$. Furthermore, as the non-linear model posterior is factorized as:

$$\tilde{p}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) = \tilde{p}(\boldsymbol{\theta}|\mathbf{y})\tilde{p}(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}) \quad (1.30)$$

the linear approximation is analogously factorized as:

$$\bar{p}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) = \bar{p}(\boldsymbol{\theta}|\mathbf{y})\bar{p}(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}) \quad (1.31)$$

A relevant part of this approach is determining the linearization point \mathbf{u}_0 . This is done through the fixed point iteration method (Burden et al., 2015). Let $f(\bar{p}_v)$ be a function of the posterior distribution linearized at v , this method seeks a point \mathbf{u}_0 , such that $\mathbf{u}_0 = f(\bar{p}_{\mathbf{u}_0})$. The function f can be either the posterior expectation $\bar{E}(\mathbf{u}|\mathbf{y})$ or the joint conditional mode

$$f(\bar{p}_v) = \arg \max_{\mathbf{u}} \bar{p}_v(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\theta}}) \quad (1.32)$$

with $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \bar{p}_v(\boldsymbol{\theta}|\mathbf{y})$. The next steps summarize the algorithm to find the point \mathbf{u}_0 :

1. Let \mathbf{u}_0 be an initial linearization point
2. Compute the predictor linearization at \mathbf{u}_0
3. Compute the linearized INLA posterior $\bar{p}(\boldsymbol{\theta}|\mathbf{y})$
4. Let $\mathbf{u}_1 = f(\bar{p}_{\mathbf{u}_0})$ be the initial candidate for the linearization point
5. Let $\mathbf{u}_\alpha = (1 - \alpha)\mathbf{u}_0 + \alpha\mathbf{u}_1$, and find α that minimizes $\|\tilde{\boldsymbol{\eta}}(\mathbf{u}_\alpha) - \tilde{\boldsymbol{\eta}}(\mathbf{u}_1)\|$
6. Set the linearization point as \mathbf{u}_α and repeat from step 1 until convergence is reached

How fast the point \mathbf{u}_0 is found and how accurate the approximation is depends on the function that is linearized. Further information on this method is available on <https://inlabru-org.github.io/inlabru/articles/>.

1.2.6 Bayesian data integration

Ecological and biodiversity phenomena are a constant source of information as they occur on a permanent basis. However, observing these processes requires effort in terms of time and economic resources. Depending on the phenomenon or the research question to be addressed, different data types, sampling protocols and observers are used to collect information. The increase in access to technology has made it easier to observe ecological processes. Hence, one ecological question can be addressed by using one of the many available datasets. Nevertheless, as ecological processes occur over vast extensions of space and on very fine time resolution, combining multiple datasets that inform on the same process is a sensible, though technically complex, idea. Combination of more than one data type has been used in research to model variables such as annual runoff in hidrology (Roksvåg et al., 2021), abundance of birds (Sicacha-Parada et al., 2022), concentration of fine particular matter (Moraga et al., 2017), spatial distribution of species (Gelfand and Shirota, 2019), among others. In all these cases improved predictive performance was achieved when multiple datasets were combined, with respect to models that used only one data type at a time.

The most known models for combining multiple datasets are the Linear Models of Coregionalization (LMC) (Wackernagel, 2003; Banerjee et al., 2008), the spatial factor model (Wang and Wall, 2003) and the shared component model

(Knorr-Held and Best, 2001). LMCs are a class of multivariate spatial models that intend to model measurements that covary jointly. These models are suited for combining data collected at point level. These models can be specified as:

$$\mathbf{Z}(\mathbf{s}) = \mathbf{A}\mathbf{Y}(\mathbf{s}) \quad (1.33)$$

with $Z(\mathbf{s})$ a $p \times 1$ random vector, $Y(\mathbf{s})$ a $m \times 1$ random vector of orthogonal latent spatial effects and A a matrix of coefficients, which have to be estimated. These models can be either fitted using MCMC techniques (Banerjee et al., 2015), or using the INLA-SPDE approach (Blangiardo and Cameletti, 2015; Krainski et al., 2018), which makes it convenient to jointly model observations collected at different spatial locations.

Spatial factor model (Wang and Wall, 2003) assume the existence of a random $p \times 1$ vector Z observed at n spatial units (either points or areas). Each Z_{ij} is assumed to have a distribution F in the exponential family with mean parameter θ_{ij} and variance parameter σ_j^2 . This model assumes Z_{ij} are conditionally independent given θ_{ij} and σ_j^2 . The parameter θ_{ij} is specified as:

$$g(\theta_{ij}) = \boldsymbol{\alpha}_j + \lambda_j f_i, \quad i = 1, \dots, n, \quad j = 1, \dots, p \quad (1.34)$$

with g an appropriate link function, $\boldsymbol{\alpha}_j$ the intercept, λ_j the slope parameter and \mathbf{f} a common Gaussian Random Field that affects all the p random variables. Depending on the structure of the covariance matrix of the spatial random effect, \mathbf{f} , the model can be either a geostatistical model or an areal data model.

Knorr-Held and Best (2001) proposed the shared component model for combining data observed at the areal level. This model is motivated by the existence of common risk factors that affect multiple diseases. This model assumes two Poisson random variables y_{i1} and y_{i2} for $i = \{1, \dots, n\}$ areal units. These random variables are specified as follows:

$$y_{i1} \sim \text{Poisson}(e_{1i}\theta_i^\delta \phi_{1i}) \quad (1.35)$$

$$y_{i2} \sim \text{Poisson}(e_{2i}\theta_i^{1/\delta} \phi_{2i}) \quad (1.36)$$

with expected counts e_{1i} and e_{2i} , a shared component θ_i and disease specific components ϕ_{1i} and ϕ_{2i} . The term δ acts as a scaling parameter to quantify the contribution of the shared component to the overall risk. This model assumes: i) that y_{i1} and y_{i2} are conditionally independent given θ_i , ϕ_{1i} and ϕ_{2i} , and ii) that θ_i, ϕ_{1i} and ϕ_{2i} are independent. This model becomes especially relevant

for the literature of data integration models as it defines a common random effect to each data type and two dataset-specific effects that can be thought of as “residuals” of the common structure (Knorr-Held and Best, 2001), or as “idiosyncratic” components for each data type (Müller et al., 2004).

Integrating multiple data types implies a trade-off between model flexibility and computational complexity (Wang et al., 2021). A considerable part of the Bayesian models for data integration are implemented through MCMC techniques such as Gibbs sampler (Wang and Wall, 2003), Metropolis-Hastings (Banerjee et al., 2015), reversible jump MCMC (Knorr-Held and Best, 2001), among others. The INLA-SPDE approach offers a more computationally-efficient alternative when these models are specified as Latent Gaussian Models (Rue et al., 2009).

The models we propose along this thesis are based on the existing literature on Bayesian data integration, and integrate multiple data types to serve two purposes: i) borrow strength from each data type by integrating data types with higher spatial coverage as CS data, and high-quality data obtained through standardized sampling protocols, and ii) tackle eventual identifiability issues that could appear when biases in the collection of each data type are accounted for as more and more parameters become part of the model.

1.3 Established methods for ecological and biodiversity data

Multiple research questions are constantly formulated by ecologists as there is an increasing concern about topics such as conservation of species and assessment of impacts of human activity on biodiversity. These questions have been addressed in literature by making use of Species Distribution Models (SDMs). As the amount of research questions increase, the amount of data collected in biodiversity also does. The constant evolution of technological solutions for observing ecological variables and reporting the occurrence of events in real time have contributed to collect more data in a cost-efficient way. Hence, more variety of data types is available to fit SDMs and knowing how to handle these data types has become a considerable area of research. The available types of

data for constructing SDMs include, but are not limited to presence/absence data, presence-only data and count data. In this subsection we will go through some of the most frequently used statistical methods in ecology and biodiversity conservation.

1.3.1 Presence/absence data

Site-occupancy models were first proposed by MacKenzie and Kendall (2002) and Tyre et al. (2003) to model available presence/absence data \mathbf{y} . These models rely on the concept of hierarchical models (see Section 1.2.2) as they model simultaneously the actual process, so called state process \mathbf{z} , and the observation process $\mathbf{y}|\mathbf{z}$. The state process, which generates the presences/absences at n sites is specified as:

$$z_i \sim \text{Bernoulli}(\psi) \quad (1.37)$$

with ψ the occupancy probability. The observation process is specified as:

$$y_i|z_i \sim \text{Bernoulli}(z_i p) \quad (1.38)$$

with p the detection probability. These models have a simple structure and are suitable for data collected through repeated surveys that report whether or not a given species was present at a site so that detection error can be estimated. For a wider perspective of methods available for modeling presence/absence data, see (Kéry and Royle, 2015).

1.3.2 Count data

Abundance is defined as the number of individuals at some place and time. The concept of abundance is very relevant to ecologists as it provides better information about regions with high density of individuals (Johnston et al., 2015), which is of paramount importance in conservation (Massimino et al., 2017). For example, estimating abundance hotspots can inform and help authorities to select sites that may qualify to be included in the network of protected areas. Despite its relevance, this data are hard to collect. Some common measurement errors that are made in the collection of abundance data are detection errors and counting multiple times the same individual. N-mixture models (Royle, 2004) are proposed as a natural way to account for detection error to produce improved estimates of abundance. These models utilize counts of unmarked

individuals (so double-counting is a possibility) as the collection of this data is much more convenient and less expensive in terms of effort compared to, for example, capture-recapture data, which in some cases require the actual capture of the marked individuals. As site-occupancy models, N-mixture models are hierarchical models, with one layer explaining the underlying process for abundance, and a second layer built to explain the observation process of the counts. The first layer of the N-mixture model is then, the underlying process of counts N_i at m sites, specified as:

$$N_i \sim \text{Poisson}(\lambda), \quad i = \{1, \dots, m\} \quad (1.39)$$

and an observation process, with counts C_{ij} , product of p repeated surveys, which are specified as:

$$C_{ij}|N_i \sim \text{Binomial}(N_i, p) \quad (1.40)$$

with p the per-individual detection probability. For a more detailed review on N-mixture models see (Kéry and Royle, 2015).

These are very established methods among ecologists, and have proven to work efficiently to solve a handful of research questions. However, as mentioned before, the evolution of technology and the increasing access to technology for millions of people have fostered the existence of more data types as Citizen Science data.

1.3.3 Citizen Science data

Citizen Science (CS) is defined as the open engagement of the public in scientific tasks. This is a convenient way of collecting data as the participants of CS projects are voluntary. Specifically in biodiversity, CS data is captured through mobile applications or websites such as iNaturalist, GBIF and Art-sobsevasjoner. Hence, these data contain large amount of records, cover vast areas and represent a considerable amount of species. Despite the advantages of these data, they are the result of non-standardized sampling designs as people collect observation in accessible places, or areas where they expect to see more occurrences, citizen scientists have different skills for detecting species, and have differences in sampling effort. In addition to these factors, CS data are affected by differences in activity of citizen scientists in different moments of the year as well as preference for reporting certain group of species.

Given these characteristics of the sampling process of CS data, modeling these data should account for these characteristics. Otherwise, the inference produced without accounting for these flaws could be heavily biased towards characteristics of the sampled areas (Isaac et al., 2014; Sicacha-Parada et al., 2021). An open question between the users of CS data is how to properly use them. In fact, among the ecological community, there is still skepticism with respect to these data because their poor quality could affect their research (Fischer et al., 2021). We now introduce two of the most frequently used methods for modeling CS data, creating pseudo-absences (Ferrier et al., 2002) and MaxEnt (Phillips et al., 2006, 2009).

Pseudo-absences for modeling Citizen Science data

Citizen Science data in biodiversity are typically collected in the form of presence-only data as they often are the product of opportunistic detections. That is, an observer reports only the occurrences she/he detected, but do not provide further information about the locations visited and the effort made there to observe occurrences. Hence, a typical CS dataset is a large collection of points in space (i.e. a spatial point pattern of presences) without any additional information about these points. As an attempt to provide a methodological framework to utilize these data, Ferrier et al. (2002) proposed to create background zeroes so that these data could be handled afterwards as typical presence/absence data. The question of how, where and how many zeroes should be added to the original dataset has been broadly approached in ecology literature (Barbet-Massin et al., 2012). Some of the strategies proposed include, for example adding pseudo-absences beyond a minimum distance from the observed presences (Zaniewski et al., 2002; Lobo et al., 2010) or placing zeroes where other species have been reported as present (Phillips et al., 2009). Barbet-Massin et al. (2012) make recommendations about how many and how to select the pseudo-absences depending on the modeling technique used, while Pearce and Boyce (2006) and Ward et al. (2009) have made efforts to adjust the logistic regression for the resulting presence/pseudo-absence data to account for the bias induced by the selection of the background zeroes. Even though this approach has enabled many researchers to use CS data, this is still based on an arbitrary selection of background locations and do not recognize the random nature of the amount of reports and their location (Gelfand and Shirota, 2019).

Maximum entropy approach

The maximum entropy (MaxEnt) approach (Phillips et al., 2006, 2009) is arguably one of the most popular methodological tools to model CS data among ecologists as this offers a simple way of analyzing large volumes of data through an open source software (Merow et al., 2013). This method takes a gridded version of the space as starting point. The first input of MaxEnt is a list of presence location (i.e. presence-only data), then background locations are extracted to be contrasted against the presence locations. MaxEnt takes the explanatory variables \mathbf{Z} at spatial locations x_i , i.e. $Z(x_i)$ as inputs to maximize a gain function defined as:

$$gain = \frac{1}{M} \sum_{i=1}^M z(x_i)\lambda - \log \sum_{i=1}^N Q(x_i) \exp\{z(x_i)\lambda\} - \sum_{j=1}^J |\lambda_j| \beta \sqrt{s^2 [z_j] / M} \quad (1.41)$$

where the first term represents the likelihood of the presence data, and depends on coefficients λ which act as weights for the sum of the covariates at the M presence locations. The second term is a sum, in log-scale, of $\exp\{z(x_i)\lambda\}$ weighted by prior information in $Q(x_i)$, and finally the last term represents a regularization penalty on the coefficients λ , so that the most relevant covariates are retained. The strength of the regularization is determined by the parameters β and $s^2 [z_j]$ are the variances of the covariates. The definition of this gain has a positive relation with $z(x_i)$ at the presence locations, while it seeks to downweight the relevance of places where the species is expected to occur through the weights $Q(x_i)$ and avoid overfitting through the last term. The optimization of this gain function is constrained so that the moments of the prediction match the empirical moments of the data. Hence, MaxEnt through the optimization of the gain function subject to the constraints in equation aims to find the species density that matches more closely the prior belief, usually a uniform distribution in geographic space (i.e. $Q(x_i) = 1/N$ meaning all the cells are equally likely to contain an individual). MaxEnt, thus produces a surface of relative occurrence rate, which describes the relative probability that a cell is contained in a collection of presence samples. Gelfand and Shirota (2019) point out that the algorithmic nature of MaxEnt impedes quantifying the uncertainty of the estimates produced. For further details about MaxEnt see Phillips et al. (2006, 2009); Gelfand and Shirota (2019)

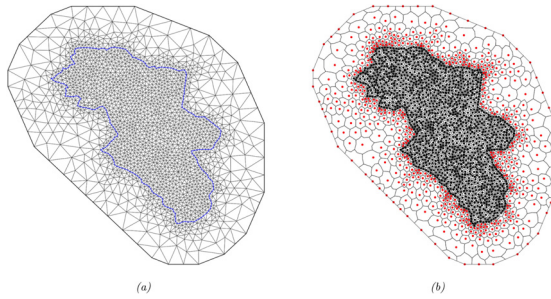


Figure 1.1: (a) Typical discretization of the space using the SPDE approach, and (b) Dual mesh for approximating the likelihood of a LGCP. The red points represent the locations \tilde{s}_i and the areas of the polygons the weights $\tilde{\alpha}_i$ in Equation (1.28)

Chapter 2

Scientific papers

In this section we present each of the papers that are part of this thesis. The topics presented in the introduction are used to develop the statistical methods proposed in each paper. Hence, in addition to the presentation of each paper, we link them with the topics in the previous section, making a brief description of the innovations and methods proposed, and identifying the fundamental aspects that make them relevant as methodological tools for modeling biodiversity data.

The papers

Paper I 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.

Paper II Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B., Øien, I. J. (2022). A Spatial Modeling Framework for Monitoring Surveys with Different Sampling Protocols with a Case Study for Bird Abundance in Mid-Scandinavia. *Journal of Agricultural, Biological and Environmental Statistics*, 27(3), 562-591.

Paper III Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B. (2022) New spatial models for integrating standardized detection-nondetection

and opportunistic presence-only data: application to estimating bird mortality hotspots linked to powerlines. In preparation.

Paper IV Cretois, B., Linnell, J. D., Van Moorter, B., Kaczensky, P., Nilsen, E. B., Parada, J., Rød, J. K. (2021). Coexistence of large mammals and humans is possible in Europe's anthropogenic landscapes. *Iscience*, 24(9), 103083.

Paper V Borgelt, J., Sicacha-Parada, J., Skarpaas, O., Verones, F. (2022). Native range estimates for red-listed vascular plants. *Scientific Data*, 9(1), 1-12.

The overall goal of thesis is to propose novel modeling methods that contribute to make better use of biodiversity data, in particular of Citizen Science data. This data type is continuously growing since citizen scientist have more technical tools to report what they observe. It means these data cover large portions of space and are reported at very fine temporal resolution. With the papers we demonstrate how Citizen Science data can be efficiently used to address multiple research questions in ecology by providing methods to efficiently account for biases in their collection process when only CS data are available, or when multiple datasets are available.

Documentation

Paper I: “Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway”

Some research questions in ecology can be addressed by making use of CS data as these are open, massive data. However, many scientists are still reluctant to the idea of using data collected through non-standardized sampling protocols and reported by observers with diverse levels of knowledge of biodiversity. This paper serves as a way to raise awareness on the potential CS data have by making proper use of them.

Several factors affect the collection of CS data in reality and represent a risk of producing biased statistical inference (Isaac et al., 2014). To address the demand of methods that make proper use of CS data, we proposed a simple modeling framework to account for differences in accessibility, as this is a factor that yields varying sampling effort in CS data. This model is built upon the conceptualization of CS data as thinned point pattern (Chakraborty et al., 2011; Fithian et al., 2015). In this particular case, accessibility was the factor that determined the thinning of the true point pattern (see Section 1.1.2). This factor has a particular relation with the activity of scientists as it is well-known that the activity of citizen scientists is focused in areas of easy accessibility (Fithian et al., 2015; Monsarrat et al., 2019), hence chances of having locations nearby roads visited are larger (i.e. lower thinning probabilities) than for the most inaccessible places. With this information in mind, two functional forms were utilized to account for accessibility as a factor that produces the observed point pattern: i) a parametric form that made use of the half-normal detection in distance sampling (Yuan et al., 2017), and ii) a semi-parametric functional form based on I-spline basis functions (Ramsay, 1988). These models were specified as Log-Gaussian Cox Processes (LGCPs; Møller and Waagepetersen (2007)) and fitted through the INLA-SPDE approach (Rue et al., 2009; Lindgren et al., 2011), using the ideas in Simpson et al. (2016).

This paper shows through simulation studies and a real-data application using data from CS projects that it is possible to make efficient use of CS data, but some extra work needs to be done. A first step is to figure out which factors are behind the degeneration of the true unobserved point pattern and then trying to account for them based on existing covariates (e.g. distance to roads), or through functional forms (e.g. I-splines), which may not provide interpretable estimates of the thinning of the point pattern, but contribute to obtain more accurate estimates of the parameters involved in the ecological process that is being investigated.

This paper is an useful tool for both statisticians and practitioners to use CS data because is based on a simple Bayesian specification and is reproducible as code for fitting the proposed model is readily available.

Paper II: “A Spatial Modeling Framework for Monitoring Surveys with Different Sampling Protocols with a Case Study for Bird Abundance in Mid-Scandinavia”

One of the most relevant questions for ecologists is about the amount of individuals that inhabit an area. Abundance (see Section 1.3.2) has proven to be beneficial to support the resolution of several research questions. However, the collection of these data requires repeated sampling, expertise and usually structured sampling designs (Kéry and Royle, 2015). The observed abundance data are usually the result of long periods of observations where lack of detection or double counting are influential factors. The effort necessary to count individuals makes the collection of these data expensive, and thus their spatial coverage is large, but their spatial and temporal resolution is coarse as a site is not visited every year and the sampling sites are placed in a cost-efficient way so that a considerable portion of the space can be covered.

In many countries data on abundance are collected through professional surveys as part of monitoring programs. These surveys hardly ever follow the same sampling protocols, so the resulting count data can be measured in different units and be the product of different sampling efforts. For this reason, producing abundance maps for more than one country is a challenging task. This paper contributes to the existing literature by building Bayesian spatial models that jointly model data from multiple professional surveys which cover disjoint portions of the space and have different sampling protocols. The models proposed in this paper take elements from existing spatial fusion models (see Section 1.2.6) by assuming that the observed counts have a common ecological process underlying. This ecological process is determined by a group of random effects and a Gaussian Random Field. Conditional on these effects, the observed counts are assumed independent. Through these models, it is also possible to quantify the effect of the design of each sampling protocol through the inclusion of covariates or random effects that explain particular characteristics of each sampling protocol.

The proposed models were fitted for both simulation studies and real-data application for monitoring programs of Norway and Sweden. The results show improvement in the predictive performance of the models when data of all the protocols are integrated. In the real-data application, a model that used the idea of an ‘idiosyncratic’ GRF (i.e. specific to one of the sampling protocols, see

Knorr-Held and Best (2001)) produced improved predictions both in terms of accuracy and precision. The methods proposed in this paper open the possibility for ecologists to construct abundance maps that cover larger spatial extents since the differences in the observed counts are properly accounted for.

Paper III: “New spatial models for integrating standardized detection-nondetection and opportunistic presence-only data: application to estimating bird mortality hotspots linked to powerlines”

As mentioned in section 1.3, more datasets are available for addressing research questions in ecology. In particular, the applied research question we address in this paper is where the riskier locations for powerline-induced deaths of birds are located, and which factors drive this process. To solve this research question two datasets are available, professional surveys data from the Norwegian Institute for Nature Research (NINA) and Citizen Science data. Both data types are regarded as thinned point patterns.

In this paper we proposed Bayesian spatial models to fusion both data types. As done in spatial fusion literature (Wackernagel, 2003; Knorr-Held and Best, 2001; Wang and Wall, 2003; Wang et al., 2021), we assume both observed point patterns are generated from a common underlying point process and what differs from dataset to dataset is how the observation process is performed. The observed data from professional surveys is the result of preferential sampling since the experts use their prior knowledge to determine which powerlines to visit. To account for this, we have assumed the selection of sites to sample by experts begins by selecting which powerline should be visited. That is, the modeling of the preferential sampling occurs at the areal level. The thinning of the true point pattern that generates the observed CS data is explained as a multi-stage thinning process that is associated to factors such as accessibility, detectability and willingness to report an occurrence. Unlike the process that generates professional survey data, we assume that the thinning process that produces CS data occur at the point level.

The contribution of this paper is to propose flexible models to fusion two or more data types that report the same ecological process, but following different sampling process, while accounting for the factors that bias the collection of

the data types fused. Fitting the proposed models implies in this case the integration of point-level data and areal data as done in, for example, Roksvåg et al. (2021) and Wang et al. (2021). Producing fusion models with improved predictive performance requires to account for the factors linked to biases in the collection of each data set (Simmonds et al., 2020). In this paper we have also proposed techniques for explicitly accounting for preferential sampling and biases in CS data collection with aid of the linearization technique in the *inlabru* R-package (Bachl et al., 2019).

This flexible framework offers a methodology for efficiently fusing data collected by professionals and citizen scientists. The benefits of these models were analyzed in a simulation study and real data of powerline-induced deaths. The simulation studies showed the importance of integrating both data sets as accuracy in parameter estimates improved. The results of the case study highlighted the relevance of the proposed models as the effect of the amount of exposed birds on the risk of powerline-induced deaths had considerable differences between models.

Paper IV: “Coexistence of large mammals and humans is possible in Europe’s anthropogenic landscapes”

Spatial distribution of species is a major concern for the ecological community because knowing which factors determine the spatial distribution and size of species’ niche is of paramount importance for supporting conservation policies. This paper aims to determine the influence of human populations on the spatial distribution of large mammal species across Europe as a way to support existing conservation programs.

The data for solving this question was collected from existing literature about mammal populations and was aggregated into $10\text{km} \times 10\text{km}$ presence/absence cells. Hence, the proposed model was a typical areal data model with binary response. A spatial random effect with intrinsic Conditional Autoregressive (iCAR) structure was specified as a way to account for the spatial autocorrelation not accounted for by the fixed effects included in the model. The fixed effects of the model included anthropogenic factors considered as the Human Footprint Index, and the spatial coverage of protected areas within each $10\text{km} \times 10\text{km}$ (i.e. the number of $1\text{km} \times 1\text{km}$ cells within each $10\text{km} \times 10\text{km}$ cell that corresponded to protected areas). The anthropogenic factors were con-

trasted against biophysical drivers of big mammals distribution such as Terrain Ruggedness Index, Potential Evapotranspiration of the Warmest Quarter and Snow Cover Duration. Given that the proposed model was specified as a Latent Gaussian Model (LGM, Rue et al. (2009)), it was fitted using the INLA-SPDE approach (Rue et al., 2009; Lindgren et al., 2011).

The results of the model showed that for most of the studied species, the presence of human populations does not represent the main driver of their spatial distribution. The spatial location of these species was instead determined by the biophysical factors. Hence, coexistence between human populations and big mammals seems a possibility that does not harm populations of big mammals. This paper showed how simple statistical techniques can support large-scale studies in ecology and have impact on decision making for species conservation.

Paper V: “Native range estimates for red-listed vascular plants”

Conservation policies for biodiversity around the globe rely on information about the spatial distribution of species. However, comprehensive ready-to-use datasets are only available for few vertebrate groups. The necessity of more information regarding the spatial distribution of fundamental species for assessing anthropogenic impact and define conservation policies is not available. CS data offers the opportunity of producing range maps for thousands of species around the globe, given the massive amount of georeferenced occurrences that are available in CS databases.

This paper contributes to the development of conservation policies for terrestrial vascular plants listed at the global IUCN red list by providing accessible datasets for about 47,675 species. These data include pre-defined native regions for all the aforementioned species, the spatial density of CS reports in the Global Biodiversity Information Facility (GBIF) for above 30,000 species and predicted suitable areas within species’ native regions for about 27,000 species. These predictions were made using MaxEnt (see Section 1.3.3) while using the available CS data and a group of environmental covariates. As known, CS data is the result, in many cases, of biased sampling protocols. As a way to deal with the potential biases, the data from CS databases, available in 30 arc min. cells, were filtered out in three different levels: no filter, presence cells (i.e. removing duplicated cells) and thinned presence cells (i.e. cells within two-cell

from a presence cell were removed). The predicted maps for 4,257 species were validated by comparing them against expert-drawn range maps from IUCN.

This paper and the resulting datasets and maps offer an innovative tool to access information of thousands of plant species while making efficient use of millions of CS reports. Not only supports this tool ongoing and new conservation initiatives, but also encourages citizen scientists to keep contributing as their work has actual value for science and biodiversity.

Bibliography

- Amano, T., Lamming, J. D., and Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, 66(5):393–400.
- Asaad, I., Lundquist, C. J., Erdmann, M. V., and Costello, M. J. (2017). Ecological criteria to identify areas for biodiversity conservation. *Biological conservation*, 213:309–316.
- Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an r package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2ed. edition.
- Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.
- Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in ecology and evolution*, 3(2):327–338.
- Beaumont, C. N., Goodman, A. A., Kendrew, S., Williams, J. P., and Simpson, R. (2014). The milky way project: leveraging citizen science and machine learning to detect interstellar bubbles. *The Astrophysical Journal Supplement Series*, 214(1):3.

- Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990.
- Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.
- Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.
- Borgelt, J., Sicacha-Parada, J., Skarpaas, O., and Verones, F. (2022). Native range estimates for red-listed vascular plants. *Scientific Data*, 9(1):1–12.
- Brenner, S. C., Scott, L. R., and Scott, L. R. (2008). *The mathematical theory of finite element methods*, volume 3. Springer.
- Burden, R. L., Faires, J. D., and Burden, A. M. (2015). *Numerical analysis*. Cengage learning.
- Carlin, B. P. and Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.
- Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.
- Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and 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.
- Chaloner, K. (1996). Elicitation of prior distributions. *Bayesian biostatistics*, 141:156.
- Ciarlet, P. G. (2002). *The finite element method for elliptic problems*. SIAM.
- Clements, K. R., Karp, P., Harris, H. E., Ali, F., Candelmo, A., Rodríguez, S. J., Balcázar-Escalera, C., Fogg, A. Q., Green, S. J., and Solomon, J. N. (2021). The role of citizen science in the research and management of invasive lionfish across the western atlantic. *Diversity*, 13(12):673.

- Cliff, A. and Ord, J. (1973). *Spatial Autocorrelation*. Monographs in spatial and environmental systems analysis. Pion.
- Conn, P. B., Thorson, J. T., and Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8(11):1535–1546.
- Cooper, C. B., Dickinson, J., Phillips, T., and Bonney, R. (2007). Citizen science as a tool for conservation in residential ecosystems. *Ecology and society*, 12(2).
- Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.
- Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.
- Eidsvik, J., Finley, A. O., Banerjee, S., and Rue, H. (2012). Approximate bayesian inference for large spatial datasets using predictive process models. *Computational Statistics & Data Analysis*, 56(6):1362–1380.
- European Commission and Directorate-General for Research and Innovation (2020). *Citizen Science : elevating research and innovation through societal engagement*. Publications Office of the European Union.
- Ferrier, S., Drielsma, M., Manion, G., and 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.
- Fischer, H. A., Gerber, L. R., and Wentz, E. A. (2021). Evaluating the fitness for use of citizen science data for wildlife monitoring. *Frontiers in Ecology and Evolution*, page 705.
- Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.
- Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. *Journal of the American Statistical Association*, 114(525):445–452.

- Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.
- Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68.
- Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press.
- Gelfand, A. E. and Schliep, E. M. (2018). Bayesian inference and computing for spatial point patterns. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, volume 10, pages i–125. JSTOR.
- Gelfand, A. E. and 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.
- Givens, G. H. and Hoeting, J. A. (2012). *Computational statistics*, volume 703. John Wiley & Sons.
- Grazian, C. and Fan, Y. (2020). A review of approximate bayesian computation methods via density estimation: Inference for simulator-models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(4):e1486.
- Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., and Bonn, A. (2018). Innovation in open science, society and policy—setting the agenda for citizen science. *Citizen Science: Innovation in open science, society and policy*, pages 1–23.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.
- Illian, J. B., Sørbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). *The annals of applied statistics*, 6(4):1499–1530.
- IPBES (2019). Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. Technical report, IPBES secretariat.

- Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., and 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.
- Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.
- Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.
- Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., Hallstein, E., Merrifield, M. S., Matsumoto, S., and Kelling, S. (2015). Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications*, 25(7):1749–1756.
- Kéry, M. and Royle, J. A. (2015). Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in r and bugs.
- Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85.
- Knorr-Held, L., Raßer, G., and Becker, N. (2002). Disease mapping of stage-specific cancer incidence data. *Biometrics*, 58(3):492–501.
- Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.
- Lingard, T. K., Masters, K. L., Krawczyk, C., Lintott, C., Kruk, S., Simmons, B., Simpson, R., Bamford, S., Nichol, R. C., and Baeten, E. (2020). Galaxy zoo builder: four-component photometric decomposition of spiral galaxies guided by citizen science. *The Astrophysical Journal*, 900(2):178.

- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- Lobo, J. M., Jiménez-Valverde, A., and Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114.
- MacKenzie, D. I. and Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance? *Ecology*, 83(9):2387–2393.
- Massimino, D., Johnston, A., Gillings, S., Jiguet, F., and Pearce-Higgins, J. W. (2017). Projected reductions in climatic suitability for vulnerable british birds. *Climatic Change*, 145(1):117–130.
- Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.
- Merow, C., Smith, M. J., and Silander Jr, J. A. (2013). A practical guide to maxent for modeling species’ distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069.
- Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., et al. (2021). Prior knowledge elicitation: The past, present, and future. *arXiv preprint arXiv:2112.01380*.
- Miller, D. A., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species’ distributions. *Methods in Ecology and Evolution*, 10(1):22–37.
- Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.
- Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC press.
- Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684.
- Monsarrat, S., Boshoff, A. F., and Kerley, G. I. H. (2019). Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. *Ecography*, 42(1):125–136.

- Moraga, P., Cramb, S. M., Mengersen, K. L., and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41.
- Moraga, P., Dean, C., Inoue, J., Morawiecki, P., Noureen, S. R., and Wang, F. (2021). Bayesian spatial modelling of geostatistical data using inla and spde methods: A case study predicting malaria risk in mozambique. *Spatial and Spatio-temporal Epidemiology*, 39:100440.
- Müller, P., Quintana, F., and Rosner, G. (2004). Hierarchical meta-analysis over related non-parametric bayesian models. *Journal of Royal Statistical Society, Series B*, 66:735–749.
- Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., and Crowston, K. (2012). The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304.
- O’Hagan, A. and Kendall, M. (1994). *Kendall’s advanced theory of statistics: Vol. IIB*. Edward Arnold.
- Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.
- Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., and Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, 98(3):840–850.
- Pearce, J. L. and Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of applied ecology*, 43(3):405–412.
- Peter, M., Diekötter, T., Höffler, T., and Kremer, K. (2021). Biodiversity citizen science: Outcomes for the participating citizens. *People and Nature*, 3(2):294–311.
- Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259.
- Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197.

- Quarteroni, A. and Valli, A. (2008). *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media.
- Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sci.*, 3(4):425–441.
- Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.
- Robinson, O. J., Ruiz-Gutierrez, V., and Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, 24(4):460–472.
- Roksvåg, T., Steinsland, I., and Engeland, K. (2021). A two-field geostatistical model combining point and areal observations—a case study of annual runoff predictions in the voss area. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(4):934–960.
- Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.
- Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Monographs on statistics and applied probability 104. Chapman & Hall CRC, 1 edition.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.
- Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):877–892.
- Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B., and Øien, I. J. (2022). A spatial modeling framework for monitoring surveys with different sampling protocols with a case study for bird abundance in mid-scandinavia. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–30.
- Sicacha-Parada, J., Steinsland, I., Cretois, B., and Borgelt, J. (2021). Accounting for spatial varying sampling effort due to accessibility in citizen science data: A case study of moose in norway. *Spatial Statistics*, 42:100446.

- Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., and O’Hara, R. B. (2020). Is more data always better? a simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43(10):1413–1422.
- Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., and Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.
- Sørbye, S. H., Illian, J. B., Simpson, D. P., Burslem, D., and Rue, H. (2019). Careful prior specification avoids incautious inference for log-gaussian cox point processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):543–564.
- Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.
- Strasser, B., Baudry, J., Mahr, D., Sanchez, G., and Tancoigne, E. (2019). ” Citizen Science”? Rethinking Science and Public Participation. *Science & Technology Studies*, 32(ARTICLE):52–76.
- Tauginienė, L., Butkevičienė, E., Vohland, K., Heinisch, B., Daskolia, M., Suškevičs, M., Portela, M., Balázs, B., and Průse, B. (2020). Citizen science in the social sciences and humanities: the power of interdisciplinarity. *Palgrave Communications*, 6(1):1–11.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.
- Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., and Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, 13(6):1790–1801.
- United Nations (2015). Transforming our world: The 2030 agenda for sustainable development.

- Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.
- Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.
- Wang, C., Furrer, R., Group, S. S., et al. (2021). Combining heterogeneous spatial datasets with process-based spatial fusion models: A unifying framework. *Computational Statistics & Data Analysis*, 161:107240.
- Wang, F. and Wall, M. M. (2003). Generalized common spatial factor model. *Biostatistics*, 4(4):569–582.
- Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. R. (2009). Presence-only data and the em algorithm. *Biometrics*, 65(2):554–563.
- Wolfson, L. J. (1995). *Elicitation of priors and utilities for Bayesian analysis*. PhD thesis, Carnegie Mellon University.
- Yuan, Y., Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., Buckland, S. T., Rue, H., Gerrodette, T., and Others (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.
- Zaniewski, A. E., Lehmann, A., and Overton, J. M. (2002). Predicting species spatial distributions using presence-only data: a case study of native new zealand ferns. *Ecological modelling*, 157(2-3):261–280.

Paper I

**Accounting for spatial varying sampling effort due to accessibility
in Citizen Science data: A case study of moose in Norway.**

Sicacha-Parada, J., Steinsland, I., Cretois, B., & Borgelt, J.(2021) published in *Spatial
Statistics*



ELSEVIER

Contents lists available at ScienceDirect

Spatial Statistics

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

Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway



Jorge Sicacha-Parada ^{a,*}, Ingelin Steinsland ^a,
Benjamin Cretois ^b, Jan Borgelt ^c

^a Department of Mathematical Sciences, NTNU (Norwegian University of Science and Technology), Norway

^b Department of Geography, NTNU, Norway

^c Department of Energy and Process Engineering, NTNU, Norway

ARTICLE INFO

Article history:

Received 1 November 2019

Received in revised form 5 March 2020

Accepted 8 April 2020

Available online 18 April 2020

Keywords:

Bayesian modeling

Citizen Science

Variation in sampling effort

Thinned point process models

Integrated Nested Laplace Approximation (INLA)

Log-Gaussian Cox process (LGCP)

ABSTRACT

Citizen Scientists together with an increasing access to technology provide large datasets that can be used to study e.g. ecology and biodiversity. Unknown and varying sampling effort is a major issue when making inference based on citizen science data. In this paper we propose a modeling approach for accounting for variation in sampling effort due to accessibility. The paper is based on an illustrative case study using citizen science data of moose occurrence in Hedmark, Norway. The aim is to make inference about the importance of two geographical properties known to influence moose occurrence; terrain ruggedness index and solar radiation. Explanatory analysis shows that moose occurrences are overrepresented close to roads, and we use distance to roads as a proxy for accessibility. We propose a model based on a Bayesian Log-Gaussian Cox Process specification for occurrence. The model accounts for accessibility through two functional forms. This approach can be seen as a thinning process where probability of thinning, i.e. not observing, increases with increasing distances. For the moose case study distance to roads are used. Computationally efficient full Bayesian inference is performed using the Integrated Nested Laplace Approximation and the Stochastic Partial Differential Equation approach for spatial modeling. The proposed model as well as the consequences of

* Corresponding author.

E-mail addresses: jorge.sicacha@ntnu.no (J. Sicacha-Parada), ingelin.steinsland@ntnu.no (I. Steinsland), bernjamin.cretois@ntnu.no (B. Cretois), jan.borgelt@ntnu.no (J. Borgelt).

<https://doi.org/10.1016/j.spasta.2020.100446>

2211-6753/© 2020 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

not accounting for varying sampling effort due to accessibility are studied through a simulation study based on the case study. Considerable biases are found in estimates for the effect of radiation on moose occurrence when accessibility is not considered in the model.

© 2020 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the expansion of technology, information and data have become readily available not only for the scientific community, but also for society in general. Citizen Science (CS), i.e. the engagement of the public in activities formerly exclusive of trained people in scientific projects, has emerged as a consequence, (Newman et al., 2012). The convenience offered by technology has encouraged people to contribute to different fields of scientific research ranging from social sciences (www.ancientlives.org, www.oldweather.org) or astronomy (www.galaxyzoo.org) to biodiversity (e.g. www.artsobservasjoner.no, www.eBird.org and www.iNaturalist.org).

According to the typology of Citizen Science introduced in Strasser et al. (2019), CS projects in biodiversity are regarded as “sensing” projects. It means that the role of volunteers is to collect information and submit it to a large database. These projects take advantage of the participants local knowledge on their environment and reach high spatial coverage. The impact of these projects can be measured in the amount of observations that are stored in their databases. For example, by September 2019, about 1.3 billion of occurrences had been reported in the global biodiversity information facility (GBIF). The Norwegian biodiversity information centre (Artsdatabanken) has about 21 million of occurrences reported. Despite being cost-efficient, easy to retrieve and its massive amount, CS data have some drawbacks. Given their “open” nature, there is no systematic sampling design to collect data, meaning citizens record observations at convenient sampling locations and times. Additionally, no scientific background is required to be part of a CS project, which implies that some species may get misidentified, (Kelling et al., 2015).

The differences in knowledge and expertise of participants in CS projects is only one of the potential sources of bias. As described in Isaac et al. (2014), the biases in the sampling processes can be classified in four groups: temporal bias, understood as varying activity of observation and reporting across time; geographical bias, meaning more reports in more convenient locations, (Mair and Ruete, 2016); uneven sampling effort per visit and differences in detectability. Preference for reporting a specific type of species constitutes another typical bias in CS sampling designs. All these biases yield in uneven sampling effort across space and time. Moreover the sampling process is not always independent of the variable intended to be measured or observed, known as preferential sampling, (Diggle et al., 2010). An issue that is not exclusive to CS records and that needs to be considered when uncertain about the independence between observation and sampling design.

Furthermore, ideally citizens record both locations where species have been observed and locations where species have been absent. This type of data is known as presence–absence data. In this case the locations are fixed and presence or absence of a species is recorded. However, CS databases in biodiversity contain mostly presence-only data. Hence, the only information given is the presence of a species in random locations whereas the rest of the landscape remains unknown. They can be actual absences or locations that have not been sampled yet. Then, there is an evident necessity of modeling CS data in a way that acknowledges the randomness of the number and the location of the observations and that accounts for different biases in the underlying sampling process.

The focus of this paper is on presence-only data and geographical bias due to accessibility. A common approach to model this data is turning some of the unobserved locations into pseudo-absences, then the available observations could be modeled as presence–absence data, (Ferrier et al.,

2002) and (Barbet-Massin et al., 2012) However, it does not account for the spatial autocorrelation for presences and absences across space, (Gelfand and Shirota, 2019). Arguably the most common approach for modeling presence-only data is Maxent, Phillips et al. (2009, 2006). This is an algorithmic strategy that aims to find an optimal species density subject to some constraint. Given its nature, Maxent does not account for the uncertainty of the predictions. Furthermore, it provides the relative chance of finding a species in comparison to other locations rather than a probability of presence or absence at each location. In Chakraborty et al. (2011) presence-only data is regarded as a realization of a spatial point process which, for the particular case of CS data, is subject to degradation. This approach was proven to perform better than Maxent in terms of goodness-of-fit statistics in a scenario with biased sampling.

The source of variation in sampling effort targeted in this paper is spatial bias due to differences in accessibility. It has been discussed in Gelfand and Shirota (2019) and addressed in Monsarrat et al. (2019) that studies historical large mammal records in South Africa where accessibility depends on proximity to freshwater and European settlements. There, an accessibility index is computed as the average of two functions defined as the half-normal function, characteristic of distance sampling. This functional form is also mentioned in Yuan et al. (2017) as an approach to model the probability of detection as a function of the perpendicular distance to a transect line segment.

In this paper we aim to emphasize the importance of accounting for differences in accessibility when CS data is modeled. We do it by making use of the Bayesian spatial approach proposed in Chakraborty et al. (2011) and Gelfand and Shirota (2019) to model the intensity of the point process associated to the distribution of a species. It means the observed point process is understood as the resulting process after the potential point process has been degraded by the probability of having access to each location. Our working hypothesis is that the distance to the road system is a good indicator of accessibility. Thus, we account for accessibility by making use of two functional forms introduced in Yuan et al. (2017): (a) the half-normal function that assumes an exponential decay of the probability of accessing a location as the distance to the closest road increases and (b) a semi-parametric approach that explains the decay of this probability as a function of a linear combination of I-spline basis functions, (Ramsay, 1988). These functional forms are then included as part of the models that explain the observed intensity. We refer to these models as the Varying Sampling Effort (VSE) model and the Extended Varying Sampling Effort (EVSE) model. A common goal of ecological studies is to explore the importance of geographical, climatic or biological quantities that drive the distribution of a species. Hence, we also aim to see how accounting for accessibility impacts the parameters estimates in a Bayesian spatial model, changing then the way the dynamics of a species is understood. Gelfand and Shirota (2019) uses a Markov chain Monte Carlo (MCMC) sampling for inference, which is computationally expensive. The Integrated Nested Laplace Approximation (INLA), (Rue et al., 2009) is a non-sampling approach to full Bayesian inference. INLA can also be used for spatial models based on Gaussian Matern Processes using the stochastic partial differential equation (SPDE) approach, (Lindgren et al., 2011), also in point process modeling, (Simpson et al., 2016). We use INLA for inference, and its computational efficiency enable us to do a simulation study.

We consider an illustrative case study of CS presence data of moose (*Alces alces*) in the county of Hedmark, Norway. Moose is a large ungulate distributed across most of the Norwegian landscape. It utilizes a wide variety of environments, including forests, wetlands and farmland, (Hundertmark, 2016). The species contributes to ecosystem health parameters by providing key ecological processes such as browsing on both broad-leaved and needle-leaved trees as well as shrubs (for a review see Shipley (2010)). Moose survival and fitness are highly determined by competition for food, e.g. Messier (1991). Hence, moose tend to avoid areas dominated by steep slopes, deep and enduring snow cover as well as poor food availability. In order to proxy this knowledge, we use two explanatory variables: solar radiation (RAD) and terrain ruggedness index (TRI). Solar radiation has been shown to influence fine scale movement of moose due to its effects on air temperature, snow cover and plant phenology, (Pomeroy et al., 1998). Moose are more likely to select areas receiving higher levels of solar energy as snow cover is shallow and plant productivity higher. Ruggedness, or terrain heterogeneity also has a major role in moose distribution as a high ruggedness increase their energy expenditure, (Leblond et al., 2010).

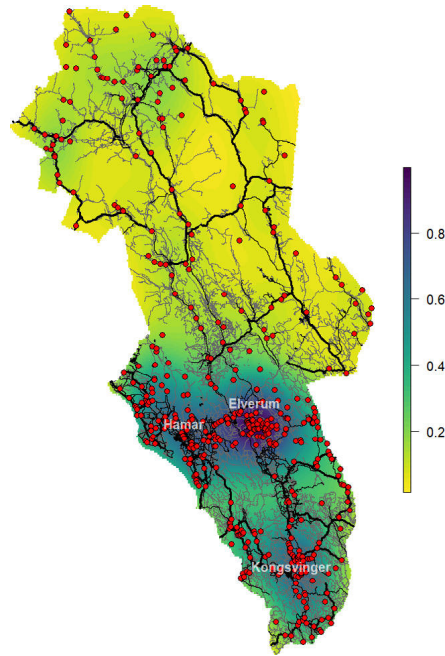


Fig. 1. Moose observations (red points) and road system (lines) in the county of Hedmark, Norway. Bold lines indicate main roads.

This paper is organized as follows: In Section 2, the dataset of the case study is introduced and explored. In Section 3, models are presented, as well as the inference method and measures for evaluating and comparing them. In Section 4, we perform a simulation study comparing the models that account for variation in sampling effort and a model not accounting for it. In Section 5 results of both the simulation study and the moose case study are shown. The paper finishes in Section 6 with the discussion of the results and concluding remarks.

2. Case study: Moose in Hedmark and exploratory analysis

In this paper we study moose distribution using locations recorded by citizen scientists and retrieved from GBIF (<https://gbif.org>). It corresponds to 472 observations product of human observation from 2000 to 2019, NBIC (2019b,a), Blindheim (2019) and iNaturalist.org (2019). These observations correspond to locations of moose in the county of Hedmark, Norway, see Fig. 1. Further, we have two explanatory variables available: RAD and TRI. RAD is computed as the yearly average of the monthly solar radiation retrieved from WorldClim (<http://worldclim.org/version2>), Fick and Hijmans (2017). TRI was obtained from the ENVIREM dataset (<https://envirem.github.io>). Both variables are available at approximately $1 \text{ km} \times 1 \text{ km}$ resolution, Title and Bemmels (2018).

Our working hypothesis is that spatial variation in sampling effort can be partly explained by accessibility due to distance to roads. In order to determine whether or not it happens, we used the road system of Hedmark retrieved from the spatial crowd-sourcing project OpenStreetMap (<https://www.openstreetmap.org>). This dataset includes a detailed network of roads that ranges from highways to footways. Fig. 1 shows the roads as well as reported moose presences in Hedmark. Most of the observations are made in southern Hedmark and near populated zones of the region, such as Hamar, Elverum and Kongsvinger, or in zones with many roads.

To explore if the observed locations are more accessible than the mass of locations in the region, we compare the citizen science dataset that contains the 472 observed points with a grid of about 400 thousand evenly distributed points. We computed the closest distance to the road network for

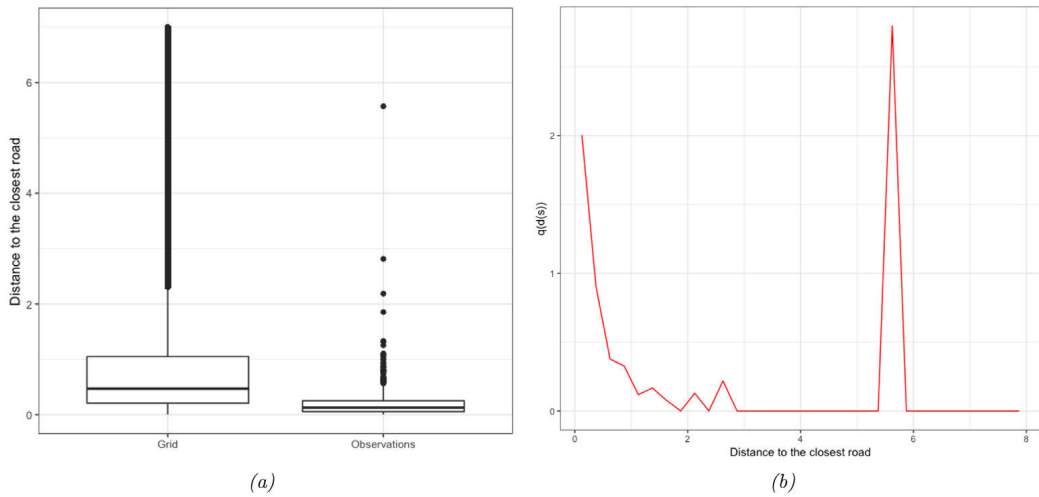


Fig. 2. (a) Boxplots of distance to the road system. Left: Dense grid of about 400 thousand points. Right: 472 reports of moose in Hedmark (b) Relationship between the observed ratio $\hat{q}(\mathbf{s}_d)$ and the distances to closest road, \mathbf{s}_d .

both datasets. The boxplots of these distances for each set of points are displayed in Fig. 2a. 91% of the observations reported are located less than 500 m away from a road. On the other hand, the grid has points that are more distant from the road system. The boxplots show that locations further away than 1 km are not represented in the observed point pattern. A Kolmogorov–Smirnov test was performed on the two sets of distances in order to determine if these two sets of distances follow the same distribution or not. The result ($p - value < 2.2e - 16$) let us conclude that, as suspected, the sets of distances do not follow the same distribution. This is an indication of a non-random sampling process. Following our working hypothesis we explore the relationship between the distance to the closest road, $d(\mathbf{s})$ and $q(\mathbf{s})$, the probability of retaining a point located at distance $d(\mathbf{s})$ (i.e. not thinning) in the observed pattern. To proxy $q(\mathbf{s})$, we grouped both sets of distances into bins, \mathbf{s}_d , of width 0.25 and for each of them we computed:

$$\hat{q}(\mathbf{s}_d) = \frac{\hat{p}_{obs}(\mathbf{s}_d)}{\hat{p}_{grid}(\mathbf{s}_d)}$$

with $\hat{p}_{obs}(\mathbf{s}_d)$ and $\hat{p}_{grid}(\mathbf{s}_d)$, the proportion of points that are part of the bin \mathbf{s}_d in the observed pattern and the dense grid, respectively. In Fig. 2b we observe a considerable decrease of $\hat{q}(\mathbf{s}_d)$ from $\mathbf{s}_d = [0, 0.25]$ to $\mathbf{s}_d = (1.5, 1.75]$. After this distance, $\hat{q}(\mathbf{s}_d)$ becomes 0, except for $\mathbf{s}_d = \{(2, 2.25]; (2.5, 2.75]; (5.5, 5.75]\}$ where few observations were reported.

According to the shape of $\hat{q}(\mathbf{s}_d)$ obtained from our sample, an exponential decay function as the one introduced in Yuan et al. (2017) arguably describes well the relationship between $d(\mathbf{s})$ and $q(\mathbf{s}_d)$. In addition to it, a semi-parametric approach also presented in Yuan et al. (2017) could be used. Both approaches are explained in more detail in Sections 3.1.2 and 3.1.3.

3. Modeling and inference approach

In this section we introduce three models that will be fitted and compared. They are based on the specification of a Log-Gaussian Cox Process. The first of them, the naive model, does not account for any difference in accessibility, while the second and third model account for accessibility as a potential source of variation in sampling effort. Then, we briefly describe the inference methods we will use. Finally, we introduce the criteria to assess and compare these models.

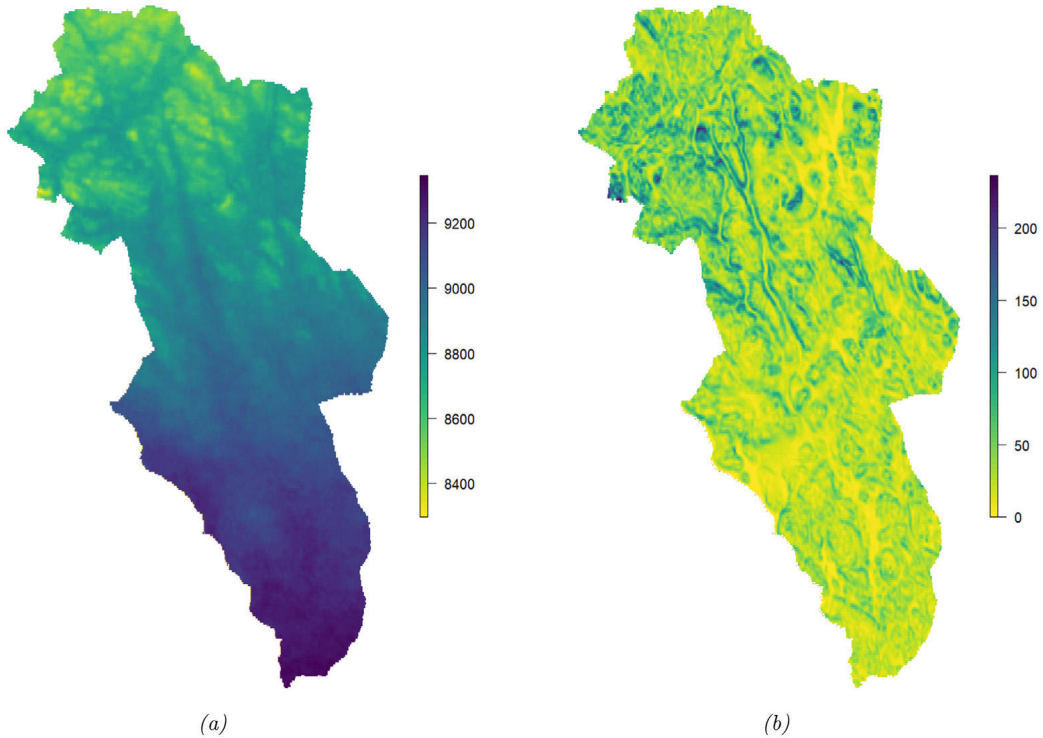


Fig. 3. (a) Solar Radiation (RAD) and (b) Terrain Ruggedness Index (TRI) in the county of Hedmark, Norway.

3.1. Models

3.1.1. Naive model

The observed data are regarded as a realization of a point process. It means both the number of points and their locations are random. The intensity measure, understood as the mean number of points per area unit, is the variable we are interested in modeling. In what follows, we will assume the observed point pattern is a realization of an inhomogeneous Poisson Process (NHPP), Illian et al. (2008), over the region $D \subset \mathbb{R}^2$. Thus, the number of points in D is assumed to be random and to have a Poisson distribution with mean $\int_D \lambda(x) dx$. We assume the point process is a Log-Gaussian Cox Process (LGCP). Hence, $\lambda(\mathbf{s})$, $\mathbf{s} \in D$ can be expressed as:

$$\log(\lambda(\mathbf{s})) = \mathbf{x}^T(\mathbf{s})\beta + \omega(\mathbf{s}) \quad (1)$$

with $\mathbf{x}(\mathbf{s})$ a set of spatially-referenced covariates and $\omega(\mathbf{s})$ a zero-mean Gaussian process that accounts for residual spatial autocorrelation between locations in D . For our case study the set of spatial covariates $\mathbf{x}(\mathbf{s})$ are: TRI and RAD, displayed in Fig. 3. A flexible family of covariance functions is the Matérn class:

$$\frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} (\kappa \|s_i - s_j\|)^\nu K_\nu(\kappa \|s_i - s_j\|) \quad (2)$$

with $\|s_i - s_j\|$ the Euclidean distance between two locations $s_i, s_j \in D$. σ^2 stands for the marginal variance, and K_ν represents the modified Bessel function of the second kind and order $\nu > 0$. ν is the parameter that determines the degree of smoothness of the process, while $\kappa > 0$ is a scaling parameter.

3.1.2. Variation in sampling effort (VSE) model

Degeneration of the point process has to be considered in the model. We associate it to a thinned intensity. That is, we now assume that the intensity of the observed point process is $\lambda(\mathbf{s})q(\mathbf{s})$ with $\lambda(\mathbf{s})$ the intensity modeled in the naive model, named in Chakraborty et al. (2011) as the potential intensity and $q(\mathbf{s})$ the thinning factor which ranges between 0 and 1, with 0 representing total degradation and 1 no degradation. In our application, the degradation is associated to accessibility based on distances to a road network. Thus, as $d(\mathbf{s})$ approaches 0, $q(d(\mathbf{s}))$ approaches 1.

The way $q(\mathbf{s})$ can be specified is still an open question, and several alternatives are available, depending on the sources of variation in sampling effort that are considered in the model. For example, in the case of moose distribution in Hedmark, $q(\mathbf{s})$ could be associated to accessibility to the road system, (Gelfand and Shirota, 2019), to populated areas and freshwater, (Monsarrat et al., 2019), or land transformation, (Chakraborty et al., 2011). As pointed out in Yuan et al. (2017), in case $q(\mathbf{s})$ is not log-linear, the estimation of the parameters is not part of the latent Gaussian model framework of INLA. Thus, following the half normal detection function in distance sampling, (Yuan et al., 2017), we aim to account for differences in accessibility by making use of the functional form:

$$q(\mathbf{s}) = \exp(-\zeta \cdot d(\mathbf{s})^2/2); \quad \zeta > 0 \tag{3}$$

where ζ is a scale parameter and $d(\mathbf{s})$ is the closest distance from location \mathbf{s} to the road system. Thus, the model we propose, which accounts for differences in accessibility is:

$$\log(\lambda(\mathbf{s})q(\mathbf{s})) = \mathbf{x}^T(\mathbf{s})\beta + \omega(\mathbf{s}) + \log(q(\mathbf{s})) \tag{4}$$

This model requires that the variables that are used to explain $q(\mathbf{s})$, in our application distance to the road system, are available at every $\mathbf{s} \in D$.

3.1.3. Extended variation in sampling effort model (EVSE)

Even if the VSE model accounts for variation in sampling effort, the functional form of $q(\mathbf{s})$ does not offer enough flexibility in situations with thinning processes that do not follow an exponential functional form. A natural, convenient way of overcoming this issue and still keeping a log-linear relationship between $d(\mathbf{s})$ and $q(\mathbf{s})$, is by means of a non-parametric approach. We can specify $-\log(q(\mathbf{s}))$ as a linear combination of basis functions as proposed in Yuan et al. (2017). In order to guarantee the monotonicity of $-\log(q(\mathbf{s}))$, we should use a basis of monotone functions, $B_k(\mathbf{s})$, $k = 1, \dots, p$ in the linear combination:

$$-\log(q(\mathbf{s})) = \sum_{k=1}^p \zeta_k B_k(\mathbf{s}) \tag{5}$$

with ζ_k a set of parameters constrained to be positive, (Yuan et al., 2017) and (Ramsay, 1988). Since this specification of $q(\mathbf{s})$ is only implemented in INLA for independent ζ_k , p should not be more than 2 or 3. Otherwise the resulting $q(\mathbf{s})$ would not be smooth, (Yuan et al., 2017). A graphical overview of the relationship between the basis function $B_k(\mathbf{s})$ and $q(\mathbf{s})$ is available in Appendix A.

3.1.4. Prior specification

The parameter ν in the Matérn covariance function (2) is fixed to be 1. On the other hand, the interest is put on the spatial range ρ and on σ , with ρ related to κ in (2) through $\rho = \sqrt{8}/\kappa$. These two parameters are specified by making use of PC priors, (Fuglstad et al., 2019). In this case we set $P(\rho < 15) = 0.05$ and $P(\sigma > 1) = 0.05$. It means that under this prior specification a standard deviation greater than 1 is regarded as large, while a spatial range less than 15 is considered unlikely. The parameters in β have Normal prior with mean 0 and precision 0.01. Finally, let $\zeta = \exp(\theta)$. For the hyperparameter θ a Normal prior distribution with mean 1 and precision 0.05 is specified. In (5), let $\zeta_k = \exp(\theta_k)$, $k = 1, \dots, p$. Each θ_k has a normal prior with mean 1 and precision 0.05.

3.2. Inference and computational approach

The models introduced in Section 3.1 will be fitted making use of the Integrated Nested Laplace Approximation (INLA), (Rue et al., 2009), the SPDE approach, (Lindgren et al., 2011), and the approach introduced in Simpson et al. (2016) for fitting spatial point processes.

3.2.1. The Integrated Nested Laplace Approximation (INLA)

The traditional approach for performing Bayesian inference for latent Gaussian models is Monte Carlo Markov Chains (MCMC). However, the Integrated Nested Laplace Approximation (INLA), (Rue et al., 2009), has emerged as a reliable alternative, (Illian et al., 2013; Humphreys et al., 2017) and (Sadykova et al., 2017). While MCMC requires considerable time to perform Bayesian inference for complex structures such as those inherent to spatial models, INLA requires less time to do the same task since, unlike MCMC which is simulation based, INLA is a deterministic algorithm, (Blangiardo and Cameletti, 2015). The aim of INLA is to produce a numerical approximation of the marginal posterior distribution of the parameters and hyperparameters of the model. In addition to its computational benefits, implementing INLA is simple by making use of the R-INLA library.

3.2.2. The SPDE approach

A useful and efficient way to represent a continuous spatial process based on a discretely indexed spatial random process is the Stochastic Partial Differential Equation (SPDE) approach, (Lindgren et al., 2011). This is based on the solution to the SPDE:

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}(\tau\xi(\mathbf{s})) = \mathbf{W}(\mathbf{s}) \quad (6)$$

where \mathbf{s} is a vector of locations in \mathbb{R}^2 , Δ is the Laplacian. $\nu, \kappa > 0$ and $\tau > 0$ are parameters that represent a control for the smoothness, scale and variance, respectively. $\mathbf{W}(\mathbf{s})$ is a Gaussian spatial white noise process. The solution for this equation, $\xi(\mathbf{s})$, is a stationary Gaussian Field with Matérn covariance function (2). This solution can be approximated through a basis function representation defined on a triangulation of the spatial domain D :

$$\xi(\mathbf{s}) = \sum_{g=1}^G \phi_g(\mathbf{s})\tilde{\xi}_g \quad (7)$$

where G is the total number of vertices of the triangulation, $\{\phi_g\}$ is the set of basis functions, and $\{\tilde{\xi}_g\}$ are zero-mean Gaussian distributed weights. This way of representing the Gaussian Random Field has been proven to make more efficient the fitting process. Fig. 4a displays the triangulation for the moose distribution example.

3.2.3. Approach for modeling LGCPs

The traditional way of fitting point process models is by gridding the space and then modeling the intensity on a discrete number of cells. However, this approach becomes unfeasible and computationally expensive as the number of grids increases. Given that gridding the space also implies approximating the location of the observations, it also represents a waste in information in contexts such as Citizen Science where the locations of the observations are collected with considerable precision. Since a better approximation of the continuous random field is achieved by making the size of the cells as small as possible, lattice-based methods become unfeasible as stressed in Simpson et al. (2016). The approach there introduced is especially useful in situations with uneven sampling effort since the resolution of the approximation can be locally adapted in those regions with low sampling. Some additional details of this approach are now presented.

Let $\omega(\mathbf{s})$ be a finite-dimensional continuously specified random field defined as:

$$\omega(\mathbf{s}) = \sum_{i=1}^n \omega_i \phi_i(\mathbf{s}) \quad (8)$$

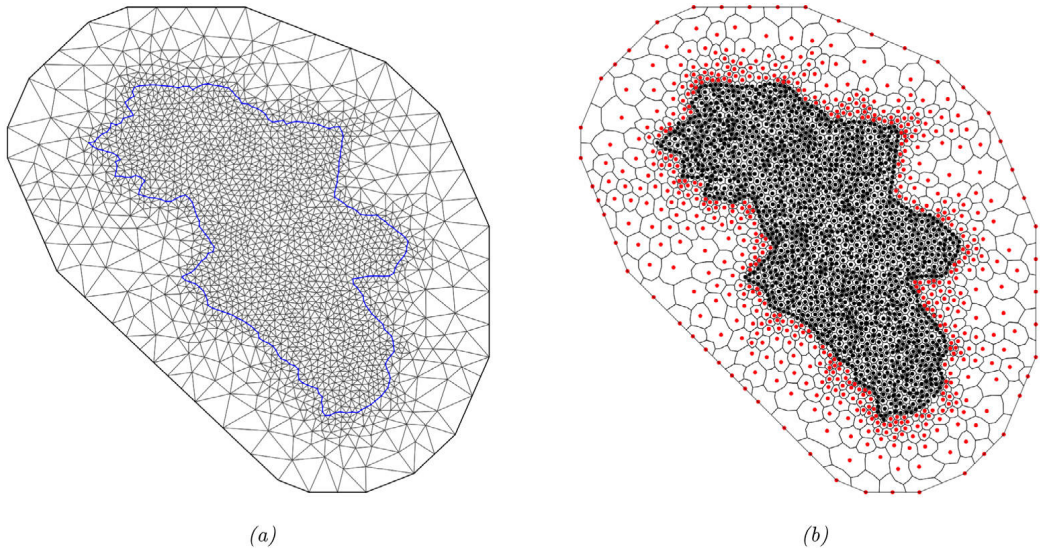


Fig. 4. (a) Triangulation of Hedmark according to the SPDE approach (b) Dual mesh for approximating the likelihood of the LGCP associated to moose distribution in Hedmark. The points are the locations \tilde{s}_i in Eq. (10) and the areas of the polygons are the weights \tilde{a}_i in Eq. (10).

Based on this specification, the likelihood of a LGCP conditional on a realization of ω :

$$\log(\pi(\lambda(\cdot)|\omega)) = |\omega| - \int_{\omega} \exp(\omega(\mathbf{s}))d\mathbf{s} + \sum_{i=1}^N \omega(s_i) \tag{9}$$

can be approximated by :

$$\log(\pi(\lambda(\cdot)|\omega)) \approx C - \sum_{i=1}^p \tilde{\alpha}_i \exp\left\{ \sum_{j=1}^n \omega_j \phi_j(\tilde{s}_i) \right\} + \sum_{i=1}^N \sum_{j=1}^n \omega_j \phi_j(s_i) \tag{10}$$

with \tilde{a}_i and \tilde{s}_i a set of deterministic weights and locations that can be obtained from a dual mesh with polygons centered at each node of the mesh. Then, $\tilde{\mathbf{s}} = \{\tilde{s}_1, \dots, \tilde{s}_n\}$ are the nodes of the mesh and $\tilde{\mathbf{a}} = \{\tilde{a}_1, \dots, \tilde{a}_n\}$ the areas of the polygons linked to each centroid. These polygons are constructed by making use of the midpoint rule, (Simpson et al., 2016). The dual mesh for our application is shown in Fig. 4b.

3.3. Model assessment

In order to assess and compare competing models such as the ones we are fitting in upcoming sections, we employ the Deviance Information Criterion (DIC), (Spiegelhalter et al., 2002), the Watanabe–Akaike Information Criterion (WAIC), Watanabe (2010), and the logarithm of the pseudo marginal likelihood (LPML). DIC makes use of the deviance of the model

$$D(\theta) = -2 \log(p(\mathbf{y}|\theta))$$

to compute the posterior mean deviance $\bar{D} = E_{\theta|\mathbf{y}}(D(\theta))$. In order to penalize the complexity of the model, the effective number of parameters,

$$p_D = E_{\theta|\mathbf{y}}(D(\theta)) - D(E_{\theta|\mathbf{y}}(\theta)) = \bar{D} - D(\bar{\theta})$$

is added to \bar{D} . Thus,

$$DIC = \bar{D} + p_D.$$

The Watanabe–Akaike Information Criterion is based on the posterior predictive density, which makes it preferable to the Akaike and the deviance information criteria, since according to [Gelman et al. \(2014\)](#) it averages over the posterior distribution rather than conditioning on a point estimate. It is empirically computed as

$$-2 \left[\sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S p(y_i | \theta^s) \right) + \sum_{i=1}^n V_{s=1}^S (\log p(y_i | \theta^s)) \right]$$

with θ^s a sample of the posterior distribution and $V_{s=1}^S$ the sample variance
Another criterion to compare the models is LMPL, defined as:

$$LPML = \sum_{i=1}^n \log(CPO_i)$$

It depends on CPO_i , the Conditional Predictive Ordinate at location i , ([Pettit, 1990](#)), a measure that assesses the model performance by means of leave-one-out cross validation. It is defined as:

$$CPO_i = p(y_i^* | y_f)$$

with y_i^* the prediction of y at location i and $y_f = y_{-i}$.

4. Simulation studies

Our simulation studies aim to show: (i) the implications of not accounting for variations on sampling effort when CS data is modeled, (ii) how accounting for at least one source of variation in sampling effort can contribute to improve the inference made about the point process underlying the spatial distribution of a species and (iii) see how misspecification of $q(\mathbf{s})$ in the VSE model can affect the quality of the inference. In order to do it, we make use of the same region map, the road system in the application, the covariate Solar Radiation (RAD), given its association with the sampling process (82% of the reports are made in locations whose solar radiation is above the median solar radiation of the entire region) and its negative correlation, (-0.43), with the distance to the road system. Then a zero-mean Gaussian random field with Matérn covariance function is simulated.

A point pattern whose intensity depends on RAD is simulated. This is specified as a Log-Gaussian Cox Process, $Y(\mathbf{s})$, with log-intensity given by:

$$\log(\lambda(\mathbf{s})) = \beta_0 + \beta_1 \text{RAD}(\mathbf{s}) + \omega(\mathbf{s}) \quad (11)$$

It is simulated with $\beta_0 = -4.25$ and $\beta_1 = 0.82$. The parameters of the Matérn covariance associated to the zero-mean Gaussian field, $\omega(\mathbf{s})$, are assumed to be $\nu = 1$, $\kappa \approx \sqrt{8}/\rho = \sqrt{8}/34$, ([Lindgren et al., 2011](#)), with ρ the practical range, and $\sigma^2 = 0.7$.

After simulating the LGCP, we thin the point pattern using two functional forms. For the first of them a point located at a distance $d(\mathbf{s})$ from the nearest road is retained with probability given by the half-normal function in (3). We create 4 scenarios based on the value of ζ : scenario 0, when $\zeta = 0$; scenario 1, when $\zeta = 1$; scenario 2, when $\zeta = 8$ and scenario 3, when $\zeta = 16$. $\zeta = 0$ corresponds to the case with no thinning. The other three values of ζ represent increasing levels of thinning that result in about 13%, 39% and 50% of observations removed, respectively.

The second functional form is a mix between the half-normal function and a constant probability of retention. In this case the probability of retaining a point follows the same functional form as in (3) until a distance d_1 . After this, the probability becomes constant. That is,

$$q(\mathbf{s}, d_1) = \exp\left(-\frac{\zeta}{2} d^2(\mathbf{s})\right) \mathbb{1}_{[0, d_1)}(d(\mathbf{s})) + \exp\left(-\frac{\zeta}{2} d_1^2\right) \mathbb{1}_{[d_1, \infty)}(d(\mathbf{s})) \quad (12)$$

With $d_1 = 0.5$, three simulation scenarios were created: scenario 4, when $\zeta = 1$; scenario 5, when $\zeta = 8$ and scenario 6, when $\zeta = 16$.

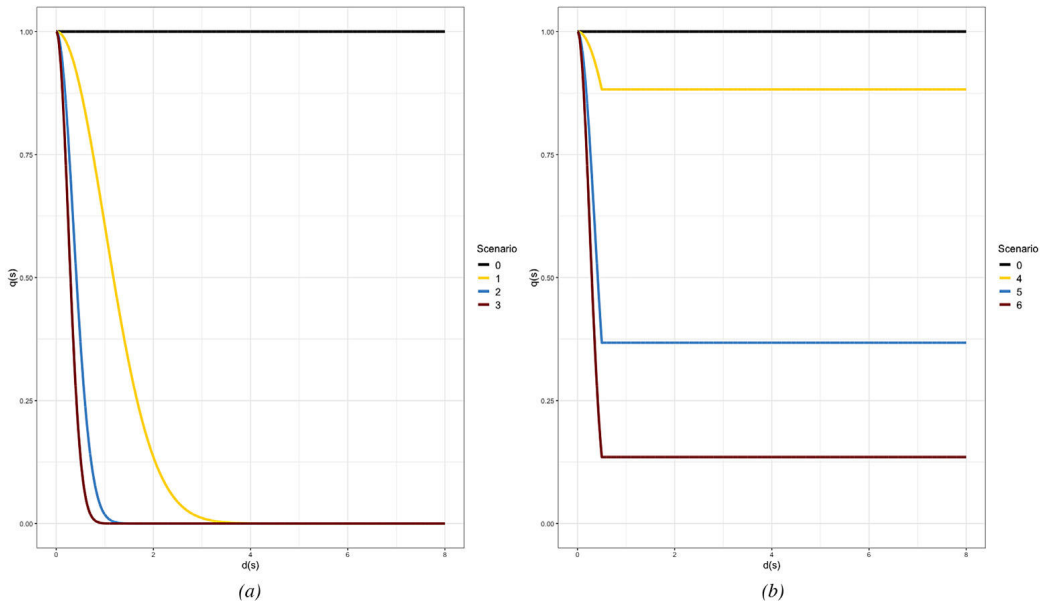


Fig. 5. (a) Relationship between $d(\mathbf{s})$ and $q(\mathbf{s})$ for the simulation scenarios 0,1,2 and 3 (b) Relationship between $d(\mathbf{s})$ and $q(\mathbf{s})$ for the simulation scenarios 0,4,5 and 6.

Table 1
Simulation scenarios.

Scenario	Thinning	ζ	d_1
0	No thinning	0	–
1	Half-normal	1	–
2	Half-normal	8	–
3	Half-normal	16	–
4	Mixed	1	0,5
5	Mixed	8	0,5
6	Mixed	16	0,5

Fig. 5a displays how the functional form of $q(\mathbf{s})$ in Eq. (3) varies as ζ increases, while Fig. 5b shows $q(\mathbf{s})$ as a function of $d(\mathbf{s})$ in each of the proposed scenarios when the functional form associated to the thinning is (12). The process of simulating a LGCP and thinning it according to ζ and d_1 was made for 100 different simulated point patterns. All the simulation scenarios are summarized in Table 1.

To assess the performance of each model for each scenario, we simulate 10000 realizations $\{\theta_{jkl}^p\}, j = 1, \dots, 10000$, from the posterior distribution of each parameter θ for point pattern $k = 1, \dots, 100$ in scenario $l = 0, 1, 2, 3, 4, 5, 6$. Then, the bias and the Root Mean Square Error (RMSE) for point pattern k in scenario l are computed as:

$$bias_{kl} = \frac{1}{10000} \sum_{j=1}^{10000} (\theta_{jkl}^p - \tilde{\theta})$$

$$RMSE_{kl} = \sqrt{\frac{1}{10000} \sum_{j=1}^{10000} (\theta_{jkl}^p - \tilde{\theta})^2}$$

with $\tilde{\theta}$ the actual value of parameter θ .

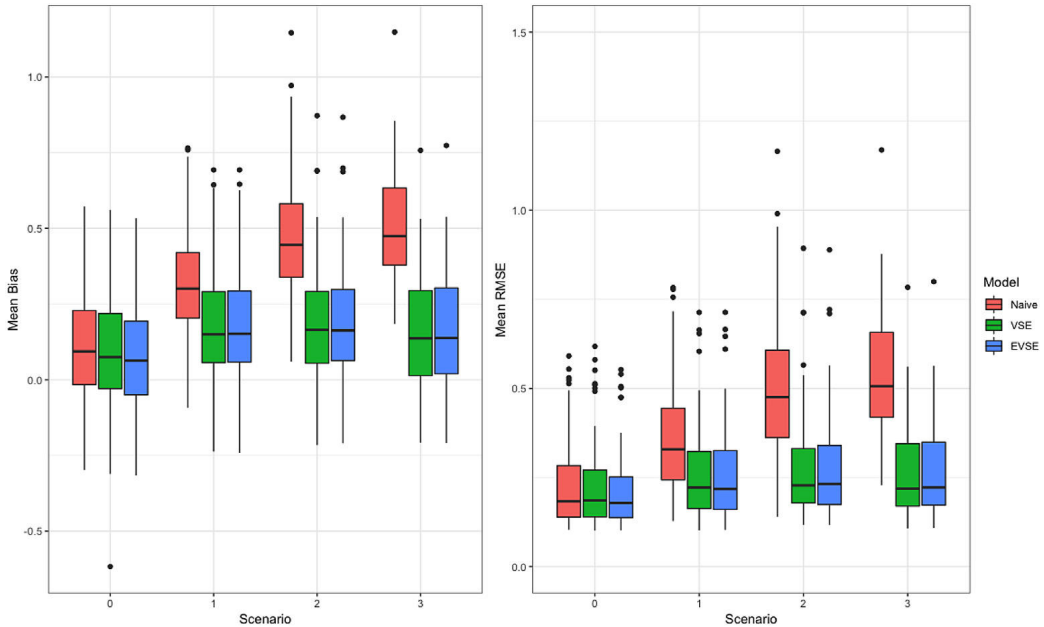


Fig. 6. Boxplots of mean bias (left) and mean RMSE (right) of β_1 for all datasets in each scenario (scenarios 0,1,2,3) and for each model.

5. Results

5.1. Simulation study

5.1.1. Results for half-normal form of $q(\mathbf{s})$

The point patterns obtained for each of the 100 simulations in each scenario described in Section 4 were fitted using the naive, the VSE and the EVSE model with $p = 3$ as suggested in Section 3. The chosen basis functions are plotted in Appendix B. The results are summarized by using measures of performance such as bias, RMSE, already introduced in Section 4, and frequentist coverage.

The parameter β_1 is the parameter of our interest. Fig. 6 presents both the mean bias and the mean RMSE at all simulated datasets for each scenario and model for this parameter. We first notice that when there is no thinning (scenario 0) the models perform similarly according to their mean bias and RMSE. However, as the original process becomes thinned (scenarios 1,2 and 3), the naive model shows poorer results than the models that account for variation in sampling effort. In scenario 3, for example, for 50% of the simulated datasets the mean RMSE for the naive model exceeds 0.5, while for less than 10% of the simulated datasets the mean RMSE is greater than 0.5 for the VSE and the EVSE models.

Table 2 introduces the mean bias and RMSE of parameters β_0 , β_1 , ρ and σ for the three models. The only parameters for which the bias and RMSE are not considerably different between the naive and the other two models are ρ and σ . However, ρ is clearly overestimated by all the models. The spatial variance and the range are the most difficult parameters to estimate and prior distributions that provide more information about these parameters may be useful to improve the accuracy of their estimates, (Cameletti et al., 2019) and (Bakar et al., 2015).

As an additional comparison measure we used the frequentist coverage of the equal-tailed $100(1-\alpha)\%$ Bayesian credible intervals for each parameter. Table 3 presents the frequentist coverage of the parameter β_1 for the three models, the results for the other parameters are available in Appendix B. The coverage of the spatial parameters does not differ between models and scenarios.

Table 2

Mean bias and RMSE for the parameters of the naive, VSE and EVSE models in the 4 scenarios simulated. In parenthesis the standard deviation of each measure.

Scenario	Approach	β_0		β_1		ρ		σ	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0	Naive	0,132 (0,150)	0,265 (0,083)	0,109 (0,186)	0,223 (0,112)	13,698 (7,994)	17,437 (7,698)	-0,055 (0,113)	0,160 (0,054)
	VSE	0,187 (0,357)	0,309 (0,322)	0,089 (0,199)	0,223 (0,114)	13,507 (8,272)	17,410 (7,672)	0,349 (4,016)	0,559 (3,994)
	EVSE	0,192 (0,158)	0,300 (0,099)	0,082 (0,184)	0,213 (0,103)	13,849 (8,019)	17,590 (7,705)	-0,051 (0,112)	0,159 (0,053)
1	Naive	-0,157 (0,154)	0,285 (0,096)	0,310 (0,179)	0,352 (0,154)	14,480 (8,754)	18,594 (8,475)	-0,033 (0,127)	0,170 (0,058)
	VSE	0,125 (0,165)	0,277 (0,084)	0,168 (0,188)	0,258 (0,128)	14,420 (8,190)	18,641 (7,701)	-0,049 (0,121)	0,169 (0,054)
	EVSE	0,121 (0,168)	0,277 (0,081)	0,169 (0,187)	0,258 (0,128)	14,302 (8,394)	18,401 (7,914)	-0,047 (0,121)	0,167 (0,056)
2	Naive	-0,648 (0,166)	0,685 (0,162)	0,463 (0,187)	0,494 (0,177)	15,187 (9,211)	20,263 (9,121)	-0,022 (0,129)	0,179 (0,057)
	VSE	0,025 (0,163)	0,253 (0,075)	0,179 (0,196)	0,276 (0,137)	15,593 (10,625)	21,463 (12,145)	-0,075 (0,131)	0,190 (0,058)
	EVSE	-0,007 (0,166)	0,254 (0,076)	0,182 (0,196)	0,278 (0,138)	14,803 (12,403)	20,063 (13,296)	-0,074 (0,149)	0,193 (0,069)
3	Naive	-0,890 (0,168)	0,918 (0,167)	0,503 (0,181)	0,534 (0,174)	14,856 (10,104)	20,600 (10,055)	-0,016 (0,129)	0,183 (0,055)
	VSE	-0,025 (0,158)	0,252 (0,067)	0,161 (0,193)	0,271 (0,130)	15,371 (10,847)	22,397 (11,051)	-0,094 (0,135)	0,203 (0,064)
	EVSE	-0,068 (0,160)	0,259 (0,079)	0,164 (0,194)	0,272 (0,131)	15,174 (14,158)	20,900 (13,846)	-0,087 (0,151)	0,195 (0,077)

Table 3

Frequentist coverage of the equal-tailed 95% Bayesian credible interval for β_1 . In parenthesis, mean length of the intervals.

Scenario	Model		
	Naive	VSE	EVSE
0	0,76 (0,49)	0,76 (0,48)	0,79 (0,49)
1	0,43 (0,55)	0,73 (0,54)	0,72 (0,54)
2	0,19 (0,63)	0,81 (0,61)	0,79 (0,61)
3	0,16 (0,67)	0,81 (0,64)	0,81 (0,64)

It is worth noting that smaller coverages are obtained for β_0 for the naive model in comparison to the other two models as the parameter ζ increases.

The model comparison methods based on the deviance and on the predictive distribution as the ones introduced in Section 3 are used to compare the results of the three models. In the scenario with $\zeta = 0$ the naive model is the true model and, as expected, it performed better than the other two models in about 40% of the simulated point patterns. This situation changes as the thinning parameter increases, the models that account for variation in sampling effort perform better than the naive one for all the simulated datasets.

5.1.2. Results for mixed functional form of $q(\mathbf{s})$

As explained in Section 4, we now thin differently the simulated point processes. The function $q(\mathbf{s})$ is now half-normal up to a distance d_1 , where it becomes constant. We fit the resulting observations using the same three models. Fig. 7 displays the mean bias and RMSE for the three models in each scenario.

In scenarios with low values of the thinning parameter ($\zeta = 0, 1$), there are not large differences in terms of bias and RMSE for the posterior median of β_1 for the three approaches. On the other

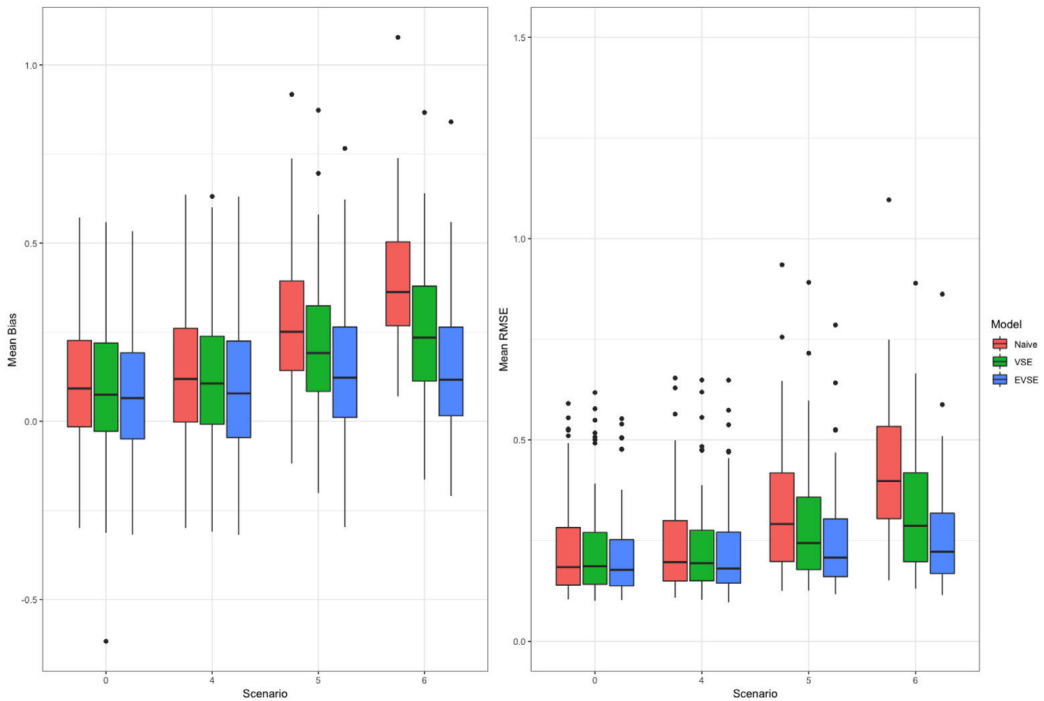


Fig. 7. Boxplots of mean bias (left) and mean RMSE (right) of β_1 for all datasets in each scenario (scenarios 0,4,5,6) and for each model.

hand, as ζ increases the differences between the three models become evident. While the EVSE model produces mean bias and RMSE consistent with scenarios of low thinning, the mean bias and RMSE of the VSE model increase, but not as much as for the naive model. Table 4 has the mean bias and RMSE of the parameters β_0 , β_1 , ρ and σ .

The same pattern described for the bias and RMSE of the parameter β_1 occurs for the intercept β_0 . In contrast, for the spatial hyperparameters, ρ and σ , there are not considerable differences in mean bias or RMSE between the three models. As made for scenarios 0,1,2 and 3, the frequentist coverage of each parameter in each scenario was computed. In Table 5, the frequentist coverage of β_1 is reported. The frequentist coverage for the other parameters is available in Appendix B.

The frequentist coverage of β_1 is very similar between the three models when the thinning is moderate, i.e. scenarios 0 and 1. However, as more observations are removed from the original point pattern, the differences between the models become larger, with the EVSE model having about 80% of coverage, while the VSE model has less than 70% and the naive model less than 60%. Finally, in terms of DIC, WAIC and CPO, the EVSE model outperforms the other two models when the thinning of the model is high.

5.2. Results for moose distribution in Hedmark application

The models introduced in Section 3 are fitted for the dataset introduced in Section 2. Table 6 reports the posterior mean and standard deviation of the parameters for each of these models. Terrain Ruggedness Index (TRI) is negatively related to the intensity, while Solar Radiation (RAD) has positive association with it for all the models. This suggests, as expected, that moose occurrences are more likely found in locations with higher solar radiation and where the terrain is less rough. The variability and range of the Gaussian field have right skewed posterior distributions based on their posterior medians and means. There is a difference in the posterior mean of RAD coefficient

Table 4

Mean bias and RMSE for the parameters of the naive and the VSE model under the 3 scenarios simulated with mixed thinning. In parenthesis the standard deviation of each measure.

Scenario	Approach	β_0		β_1		ρ		σ	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
4	Naive	0,059 (0,151)	0,244 (0,068)	0,132 (0,188)	0,235 (0,117)	13,749 (8,344)	17,640 (8,035)	-0,056 (0,115)	0,163 (0,054)
	VSE	0,088 (0,157)	0,255 (0,073)	0,117 (0,188)	0,229 (0,112)	13,793 (8,311)	17,696 (7,952)	-0,057 (0,114)	0,163 (0,054)
	EVSE	0,142 (0,162)	0,277 (0,086)	0,097 (0,187)	0,221 (0,107)	14,045 (8,377)	17,923 (8,050)	-0,051 (0,115)	0,161 (0,054)
5	Naive	-0,369 (0,158)	0,427 (0,137)	0,265 (0,191)	0,321 (0,155)	14,110 (8,378)	18,823 (8,040)	-0,062 (0,118)	0,172 (0,055)
	VSE	-0,264 (0,167)	0,348 (0,128)	0,207 (0,190)	0,284 (0,139)	14,344 (8,121)	19,049 (7,772)	-0,067 (0,116)	0,172 (0,053)
	EVSE	-0,047 (0,162)	0,248 (0,077)	0,132 (0,191)	0,245 (0,119)	14,403 (8,134)	19,027 (7,758)	-0,068 (0,117)	0,172 (0,055)
6	Naive	-0,733 (0,169)	0,763 (0,167)	0,391 (0,184)	0,428 (0,171)	13,587 (9,549)	19,145 (9,338)	-0,044 (0,124)	0,177 (0,055)
	VSE	-0,465 (0,184)	0,514 (0,167)	0,247 (0,184)	0,315 (0,145)	14,155 (9,629)	19,905 (9,432)	-0,071 (0,120)	0,181 (0,056)
	EVSE	-0,147 (0,164)	0,281 (0,103)	0,145 (0,191)	0,258 (0,126)	14,607 (9,920)	20,545 (9,546)	-0,085 (0,126)	0,189 (0,063)

Table 5

Frequentist coverage of the equal-tailed 95% Bayesian credible interval for β_1 . In parenthesis, mean length of the intervals.

Scenario	Model		
	Naive	VSE	EVSE
0	0,76 (0,49)	0,76 (0,48)	0,79 (0,49)
4	0,72 (0,50)	0,76 (0,50)	0,77 (0,50)
5	0,53 (0,56)	0,66 (0,56)	0,81 (0,56)
6	0,36 (0,62)	0,63 (0,62)	0,82 (0,61)

between the models. It is larger when differences in accessibility are not considered in the model. In addition to it, both parameters associated to the Matérn Gaussian field have lower posterior medians for the models that account for variation in sampling effort.

RAD is the most influential parameter for the three models. We see from Fig. 8 and Table 6 that the posteriors of this parameter shift considerably between the models. While the naive model has the largest posterior mean for RAD, the EVSE model has the smallest posterior mean.

The parameter ζ in the VSE model with posterior median 0.87 indicates that the observed point pattern is a thinned version of the real one, while the posterior medians of ζ_1 , ζ_2 and ζ_3 seem to give more weight to the first basis function. The basis functions used for modeling $q(\mathbf{s})$ are presented in Appendix C. Fig. 9 shows the estimated relationship between distance (in kilometers) to the road system and $q(\mathbf{s})$ for the VSE and the EVSE models. According to the results of the VSE model a point located more than 3 km away from the road system can be regarded as inaccessible for citizen scientists. On the other hand, the EVSE model does not consider any location as inaccessible for citizen scientists. Instead, it assigns constant $q(\mathbf{s}) \approx 0.05$ for locations more than 1.5 km away from the nearest road.

Fig. 10 displays the map of differences in posterior median and standard error of the logarithm of the intensity between the EVSE and the naive model. The maps with the differences in posterior median and standard error between all the models are available in Appendix C. The largest differences occur in zones that are distant to the nearest road and that have no occurrences of moose recorded. These places have lower solar radiation than the rest of the region and have considerable elevation in some locations. For the zones that are more observed, accounting for differences in

Table 6 Posterior summaries of the parameters of the naive and the VSE model for the moose presence data in Hedmark, Norway.

Parameter	Model														
	Naive					VSE					EVSE				
	Mean	Sd	0.025q	0.50q	0.975q	Mean	Sd	0.025q	0.50q	0.975q	Mean	Sd	0.025q	0.50q	0.975q
Intercept	-4.87	0.23	-5.32	-4.87	-4.41	-4.56	0.21	-4.97	-4.56	-4.14	-4.17	0.20	-4.57	-4.17	-3.77
TRI	-0.20	0.08	-0.35	-0.20	-0.04	-0.20	0.08	-0.35	-0.20	-0.04	-0.16	0.08	-0.32	-0.16	-0.01
RAD	1.01	0.19	0.64	1.01	1.38	0.73	0.18	0.37	0.73	1.10	0.57	0.17	0.23	0.57	0.91
ζ	-	-	-	-	-	0.88	0.21	0.52	0.87	1.33	-	-	-	-	-
ρ	39.78	9.06	26.14	38.32	61.37	37.73	7.68	25.88	36.59	55.74	36.45	7.04	25.26	35.52	52.78
σ	1.12	0.16	0.85	1.10	1.48	1.04	0.13	0.81	1.03	1.34	0.99	0.12	0.78	0.98	1.26
ζ_1	-	-	-	-	-	-	-	-	-	-	2.79	0.35	2.20	2.75	3.57
ζ_2	-	-	-	-	-	-	-	-	-	-	0.09	0.14	0.00	0.04	0.45
ζ_3	-	-	-	-	-	-	-	-	-	-	0.09	4.32	0.00	0.03	0.51

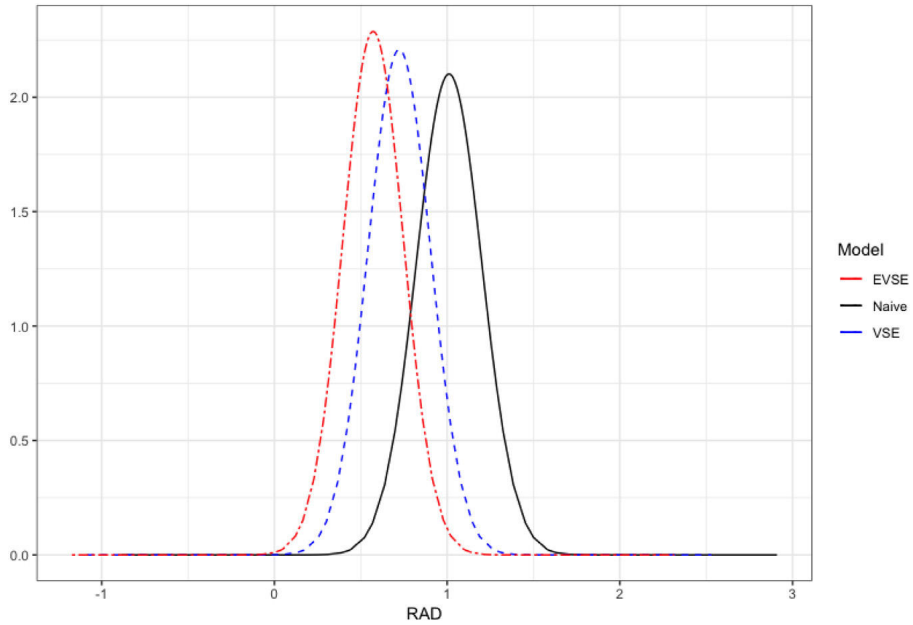


Fig. 8. Posterior density of RAD for the three models.

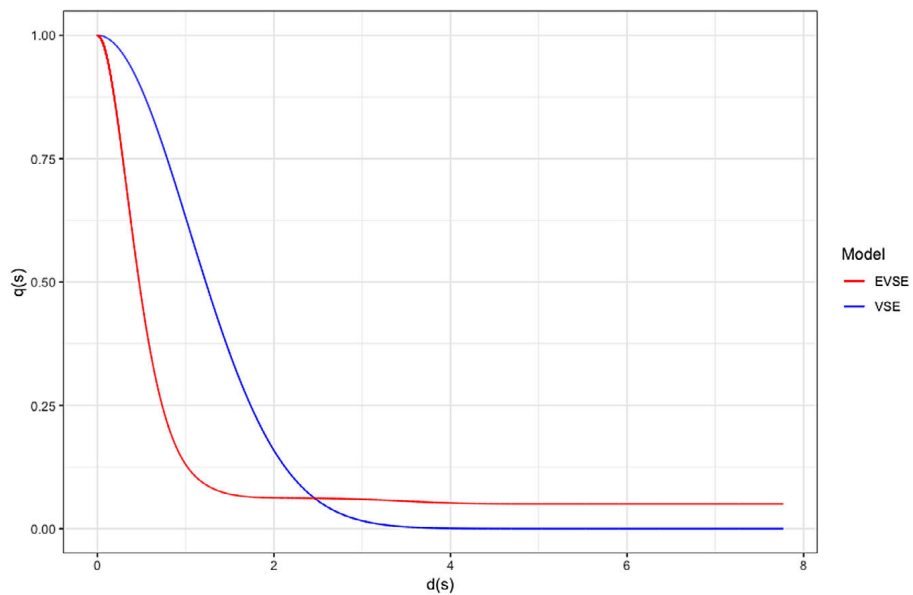


Fig. 9. Estimated relation between distance to the road system, in kilometers, and the probability of having access to location s .

accessibility does not affect the posterior median intensity and the uncertainty. The uncertainty is smaller for the EVSE model in most of the locations, except for some that include bodies of water such as lakes Mjøsa and Femun and national parks like Forollhogna national park.

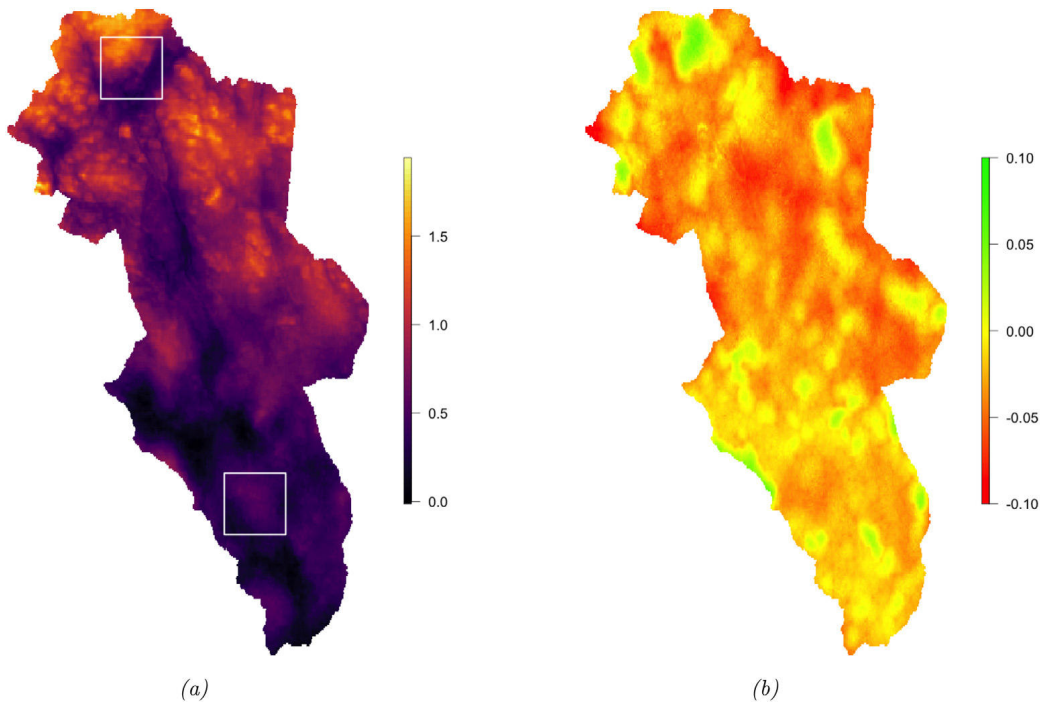


Fig. 10. (a) Differences in posterior median intensity and (b) differences in standard error of the posterior median intensity obtained through the VSE model and the naive model. In (a) the two squares represent the zones that are focused in Fig. 11 (south of Hedmark) and 12 (north of Hedmark).

Table 7
Comparison criteria for the naive and VSE model fitted to moose location reports.

	Model		
	Naive	VSE	EVSE
DIC	4377,51	4344,90	4265,77
WAIC	4505,36	4471,39	4400,91
LPML	-2467,61	-2446,98	-2428,182

The magnitude of the differences in the posterior median intensity between the VSE and the naive model is lower than between the EVSE and the naive model. The places with the highest differences in intensity and uncertainty are the same as between the EVSE and the naive model. The differences between the VSE and the EVSE model are considerably small. The three models are compared by making use of the DIC, the WAIC and the LPML. Table 7 introduces the value of each criterion for each model.

For the case of moose in Hedmark the results in Table 7 indicate that accounting for variation in sampling effort represents an improvement in terms of goodness of fit since both the DIC and WAIC are smaller, and the LPML is larger for the VSE and the EVSE model, with the latter showing better results in this sense than the former model.

Now we will focus on two specific zones of Hedmark to see with more detail how the posterior median and its associated uncertainty vary between the models. The two zones are bounded by a $30 \text{ km} \times 30 \text{ km}$ square and are highlighted in Fig. 10. The first zone is located on the southern half of Hedmark between Kongsvinger and Hamar. It is accessible only through service roads, which are

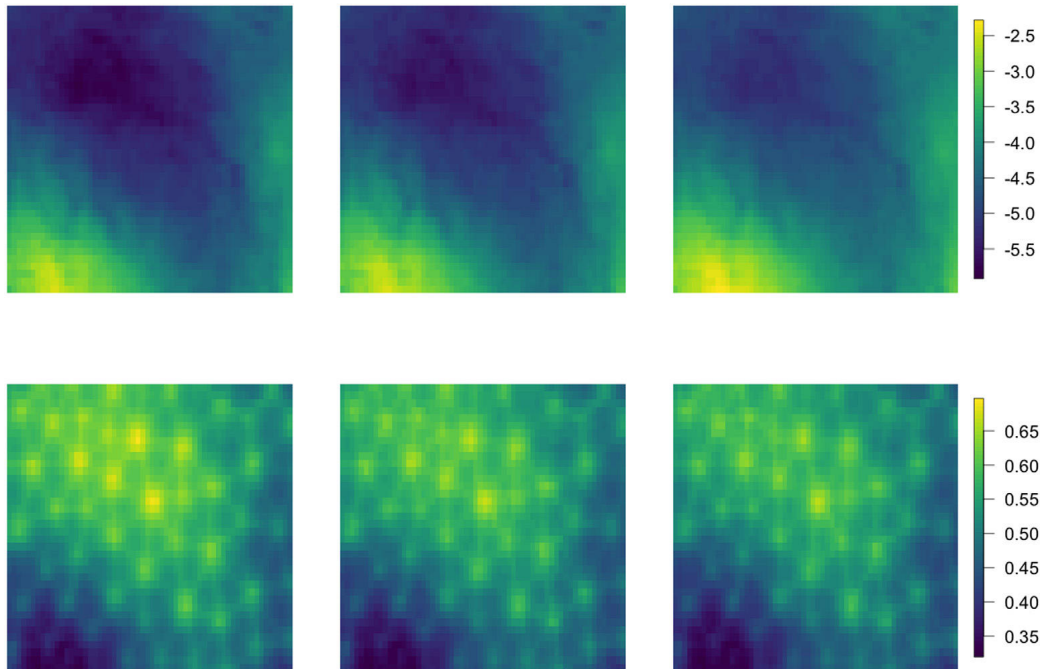


Fig. 11. Posterior median intensity (top) and associated standard error (bottom) for the naive model (left), the VSE model (middle) and the EVSE model (right) in zone 1.

not as visited as the main roads of the region, while the second square corresponds to one of the most distant zones of the region, which is located on the northern border of Hedmark. For zone 1 the posterior median intensity and its associated standard error for all the models are displayed in Fig. 11. The posterior median intensity is similar for the three models as well as the associated uncertainties. Given that the zone is regarded as highly accessible, considerable differences are not expected. In contrast, for zone 2 the EVSE model increases the intensity in most locations compared to the other two models. In terms of uncertainty the three models produce similar results. However, it becomes larger in some few zones under the VSE model, see Fig. 12.

6. Discussion and conclusions

The main goal of this paper was to highlight the importance of accounting for sources of variation in sampling effort for CS data. Bayesian spatial models that account for variation in sampling effort by including proxies for external processes that degrade the intensity of the point process have been introduced.

This paper focused on differences in accessibility across space. In the simulation studies performed in Section 4, we created scenarios where the only source of degradation for the actual point pattern was the distance to the nearest road. Two of the functional forms presented in Yuan et al. (2017) were used to link it to the intensity of the point pattern. The first of them is the half-normal function, characteristic of distance sampling. The second one is a function of a linear combination of a set of monotone functions with strictly positive coefficients. The aim of ecological studies is often to learn about the effect of covariates. The results of both the simulation study and the real data application suggest that in situations with some evidence of uneven sampling effort accounting for differences in accessibility improves performance indices, such as bias and RMSE, and model selection indices, such as DIC, WAIC and LMPL. In the scenario with no thinning on the point pattern

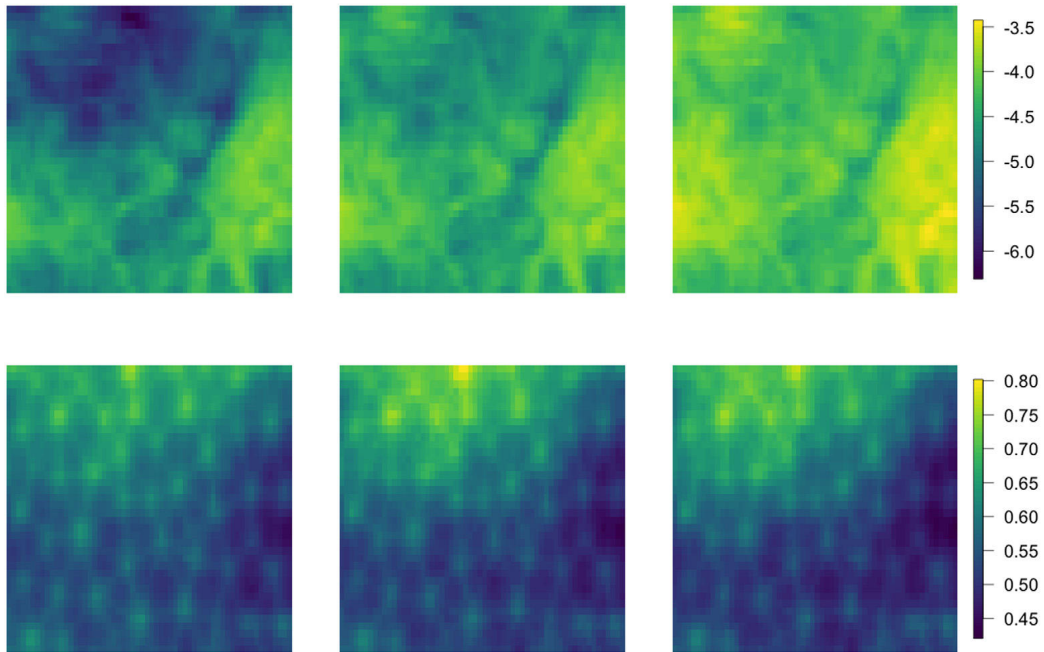


Fig. 12. Posterior median intensity (top) and associated standard error (bottom) for the naive model (left), the VSE model (middle) and the EVSE model (right) in zone 2.

due to variation in sampling effort, we found that including a term that accounts for it does not affect the quality of the inference. Furthermore, differences in the covariates posterior summaries in the simulation study showed that in cases with sampling biases the effect of an explanatory variable may be incorrectly estimated if they are not considered in the model. It is also important to note that the VSE model was proved not robust to misspecification of the relationship between $d(\mathbf{s})$ and $q(\mathbf{s})$ in scenarios with considerable thinning.

In our case study we focused on two zones of Hedmark. The large difference in intensity between the naive and the other two models in Zone 2 shows how the models that account for variation in sampling effort regard some locations on the west of this zone as possibly thinned given that they are located above 2 km away from a road and their geographical characteristics make them suitable for moose presences. The differences and the uncertainty on the north side indicate a need for increased sampling effort in this region, marking the area around Forollhogna national park. This area is one of the few mountainous areas in Norway with relatively gentle slopes and is therefore called the “friendly mountains”. Moose occasionally passes through this area, however, only few CS observations have been made so far which might partly be due to a low accessibility and therefore low CS activity. In contrast, the road network in zone 1 is rather dense. Therefore, the values of $q(\mathbf{s})$ are estimated to be relatively high and the model assumes high CS activity in this area. However, the road network here is mainly composed of service roads and small tracks. Therefore, no CS observations of moose in this area might be a result of a low visiting rate of people rather than moose being absent. However, we only accounted for differences in accessibility of sampling locations in space, therefore, the habitat is predicted to be not suitable, which seems to be wrong from an ecological perspective. Accounting for differences in visits of sampling locations in time, for instance by using spatially refined information on type of road or population data could further increase modeling performance. The results highlight, that not only accessibility (e.g. roads) are important features for quantifying preferential sampling in CS data, but also how frequent sampling sites are being visited. Small service roads and hiking tracks are likely to have a lower turnover of

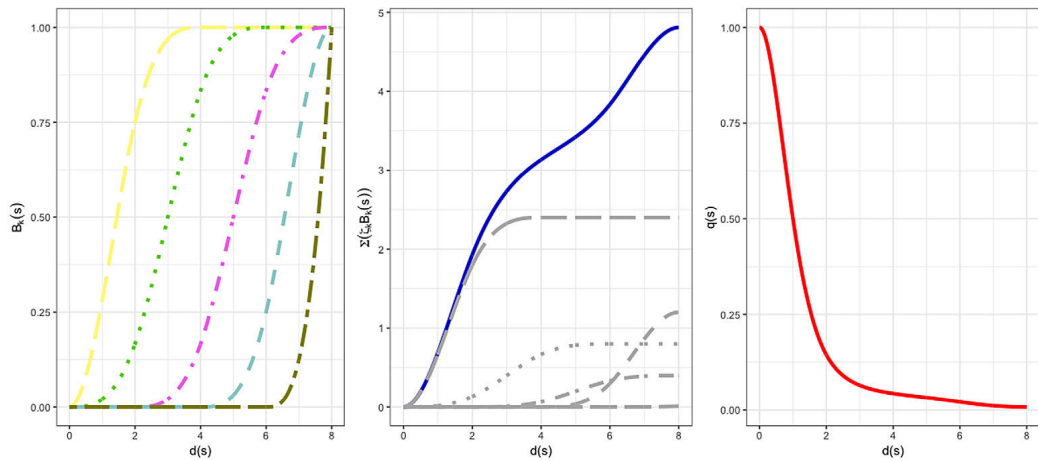


Fig. A.13. Illustration of the relationship between the basis functions and $q(\mathbf{s})$ in the EVSE model. Left, basis functions $B_k(\mathbf{s}), k = 1, \dots, 5$. Middle, weighted basis functions by the coefficients $\zeta_k, k = 1, \dots, 5$ (gray); linear combination of the weighted basis functions (solid, blue). Right, estimated $q(\mathbf{s})$ computed as Eq. (A.1).

visiting people than larger roads, and hence, CS more frequently register observations close to larger roads than close to small and remote roads.

An important part of the VSE and the EVSE models are the parameters ζ and $\zeta_k, k = 1, \dots, 3$, which are necessary to determine to what extent the differences in accessibility affect the observed process. Interpreting and including them in the model is more difficult for the EVSE model given that the basis functions need to be chosen. The prior specification of the parameters that are part of the spatial Gaussian field $\omega(\mathbf{s})$ is a complex task in spatial statistics. In this paper PC priors were used as a way to incorporate prior knowledge about these parameters in a straightforward way. Alternative prior specifications using PC priors are introduced in Sørbye et al. (2019).

The VSE and EVSE models are a first step for modeling CS data in a way that accounts for its inherent sources of bias. More effort is required for e.g. extending the sampling effort model to more quantities (e.g. cell phone coverage or geographical parameters). Extending the VSE and the EVSE to more splices would be an interesting approach for learning more about citizen science sampling effort in general.

Acknowledgments

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

Appendix A. Illustration of the EVSE model

In the EVSE model we assume

$$q(\mathbf{s}) = \exp\left(-\sum_{k=1}^p \zeta_k B_k(\mathbf{s})\right) \tag{A.1}$$

That is, $q(\mathbf{s})$ is assumed as a function of a linear combination of p basis functions $B_k(\mathbf{s}), k = 1, \dots, p$. As mentioned in Section 3, $B_k(\mathbf{s}), k = 1, \dots, p$ are a set of monotone nondecreasing functions . In addition to it, the coefficients $\zeta_k, k = 1, \dots, p$ are constrained to be positive in order to guarantee monotonicity, (Yuan et al., 2017) and (Ramsay, 1988). Fig. A.13 illustrates, similarly as made in Yuan et al. (2017), how the relationship between these basis functions and $q(\mathbf{s})$ works .

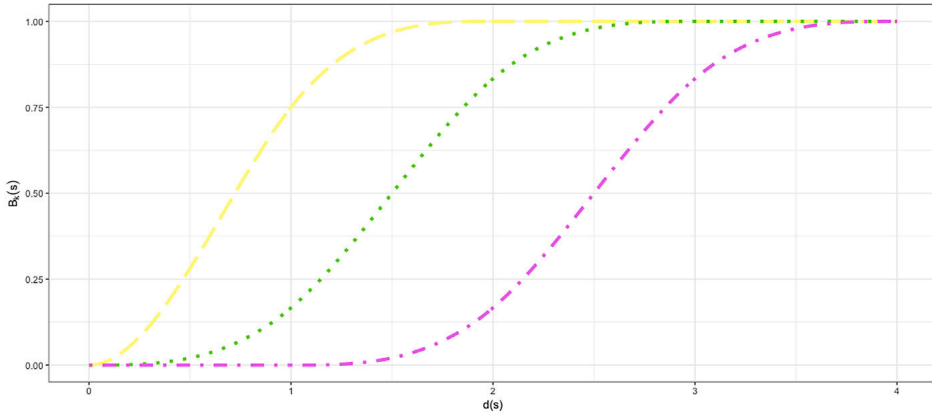


Fig. B.14. Basis functions used to fit the EVSE model in the simulation study.

Table B.8

Frequentist coverage of the equal-tailed 95% Bayesian credible interval for all the parameters in the simulations. In parenthesis, mean length of the intervals.

Parameter	Scenario	Model		
		Naive	VSE	EVSE
β_0	0	0,93 (0,74)	0,91 (0,74)	0,85 (0,75)
	1	0,92 (0,79)	0,92 (0,78)	0,94 (0,77)
	2	0,09 (0,83)	0,99 (0,8)	0,99 (0,8)
	3	0 (0,85)	0,99 (0,8)	0,99 (0,8)
	4	0,97 (0,75)	0,95 (0,75)	0,92 (0,75)
	5	0,53 (0,77)	0,73 (0,77)	0,98 (0,76)
β_1	0	0,76 (0,49)	0,76 (0,49)	0,79 (0,49)
	1	0,43 (0,55)	0,73 (0,54)	0,72 (0,54)
	2	0,19 (0,63)	0,79 (0,61)	0,79 (0,61)
	3	0,16 (0,67)	0,81 (0,64)	0,81 (0,64)
	4	0,72 (0,5)	0,76 (0,5)	0,77 (0,5)
	5	0,53 (0,56)	0,66 (0,57)	0,81 (0,56)
ρ	0	0,75 (39,88)	0,73 (39,56)	0,7 (39,96)
	1	0,72 (42,94)	0,75 (43,23)	0,72 (42,21)
	2	0,79 (49,35)	0,74 (47,08)	0,74 (47,08)
	3	0,87 (52,19)	0,73 (47,15)	0,73 (47,15)
	4	0,75 (40,67)	0,75 (40,75)	0,72 (40,88)
	5	0,8 (45,67)	0,79 (46,15)	0,7 (45,57)
σ	0	0,87 (0,43)	0,86 (0,43)	0,88 (0,43)
	1	0,92 (0,47)	0,9 (0,46)	0,89 (0,45)
	2	0,92 (0,51)	0,73 (0,44)	0,73 (0,44)
	3	0,94 (0,54)	0,74 (0,43)	0,74 (0,43)
	4	0,87 (0,44)	0,87 (0,44)	0,87 (0,44)
	5	0,88 (0,46)	0,88 (0,46)	0,82 (0,46)
6	0,91 (0,5)	0,88 (0,49)	0,76 (0,5)	

Appendix B. Simulation study: Extra tables and figures

See Fig. B.14 and Table B.8.

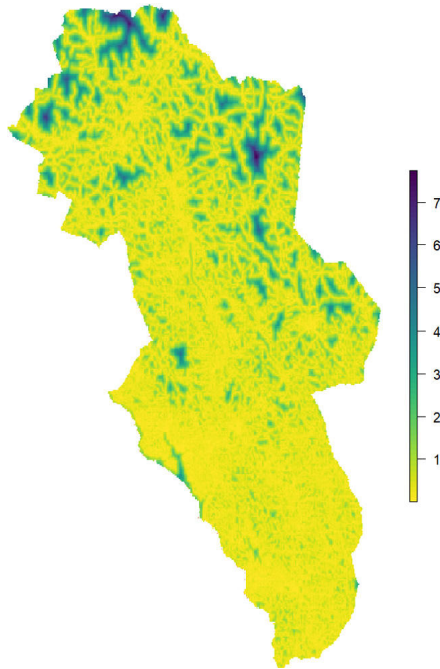


Fig. C.15. Distance to the nearest road for all locations in Hedmark.

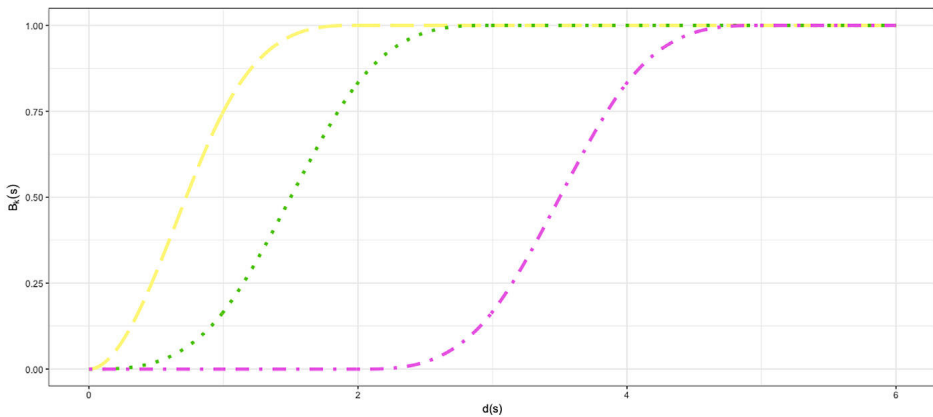


Fig. C.16. Basis functions used to fit the EVSE model for the real dataset application.

Appendix C. Moose in Hedmark application: Extra figures

See [Figs. C.15–C.17](#).

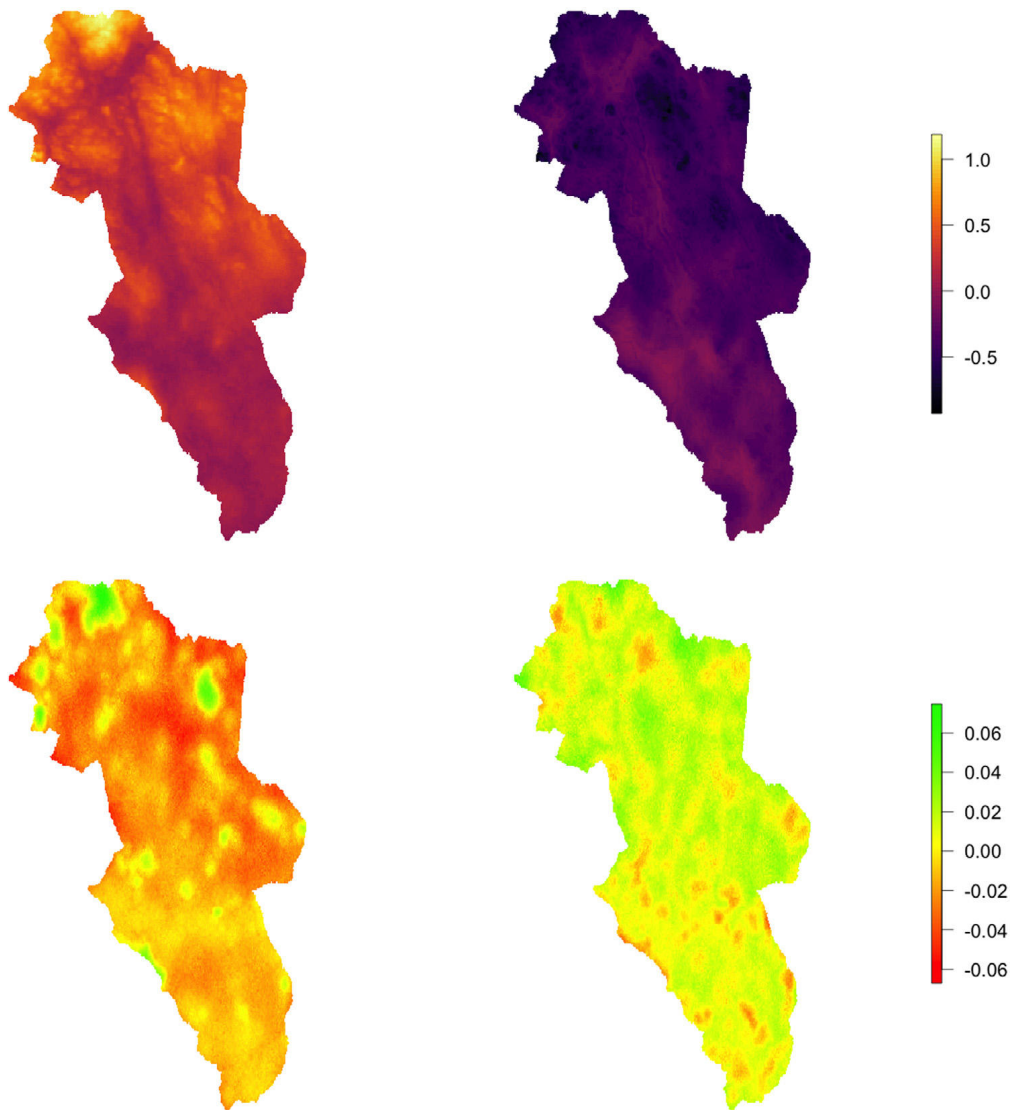


Fig. C.17. Differences in posterior median (top) and standard deviation (bottom), in log-scale, between the naive and the VSE model (left) and between the VSE and the EVSE model (right).

References

- Bakar, K.S., Sahu, S.K., et al., 2015. Sptimer: Spatio-temporal bayesian modelling using R. *J. Stat. Softw.* 63 (15), 1–32.
- Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* 3 (2), 327–338, URL <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2011.00172.x>.
- Blangiardo, M., Cameletti, M., 2015. *Spatial and Spatio-Temporal Bayesian Models With R-INLA*. John Wiley & Sons.
- Blindheim, T., 2019. Biofokus. <http://dx.doi.org/10.15468/jxbhqx>, Accessed via GBIF.org on 2019-12-02.
- Cameletti, M., Gómez-Rubio, V., Blangiardo, M., 2019. Bayesian modelling for spatially misaligned health and air pollution data through the INLA-spde approach. *Spat. Statist.* 31, 100353, URL <http://www.sciencedirect.com/science/article/pii/S2211675318301799>.

- 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 (5), 757–776, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2011.00769.x>.
- Diggle, P.J., Menezes, R., Su, T.-L., 2010. Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 59 (2), 191–232, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2009.00701.x>.
- 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. *Biodivers. Conserv.* 11 (12), 2309–2338, <http://dx.doi.org/10.1023/A:1021374009951>.
- Fick, S.E., Hijmans, R.J., 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37 (12), 4302–4315, URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086>.
- Fuglstad, G.-A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of Gaussian random fields. *J. Amer. Statist. Assoc.* 114 (525), 445–452, <http://dx.doi.org/10.1080/01621459.2017.1415907>.
- Gelfand, A.E., Shirota, S., 2019. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecol. Monograph* 89 (3), e01372, URL <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1372>.
- Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* 24 (6), 997–1016, <http://dx.doi.org/10.1007/s11222-013-9416-2>.
- Humphreys, J.M., Elsner, J.B., Jagger, T.H., Pau, S., 2017. A Bayesian geostatistical approach to modeling global distributions of lygodium microphyllum under projected climate warming. *Ecol. Model.* 363, 192–206, URL <http://www.sciencedirect.com/science/article/pii/S0304380017304064>.
- Hundertmark, K., 2016. Alces alces. the iucn red list of threatened species 2016: e.T56003281a22157381.. Downloaded on 29 October 2019. URL <http://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T56003281A22157381.en>.
- Illian, J.B., Martino, S., Sørbye, S.H., Gallego-Fernández, J.B., Zunzunegui, M., Esquivias, M.P., Travis, J.M.J., 2013. Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods Ecol. Evol.* 4 (4), 305–315, URL <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210x.12017>.
- Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. *Statistical Analysis and Modelling of Spatial Point Patterns*, Vol. 70. John Wiley & Sons.
- iNaturalist.org, 2019. iNaturalist research-grade observations. <http://dx.doi.org/10.15468/ab35x>, Accessed via GBIF.org on 2019-12-02.
- Isaac, N.J.B., 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 Ecol. Evol.* 5 (10), 1052–1060, URL <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210x.12254>.
- 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.
- Leblond, M., Dussault, C., Ouellet, J.-P., 2010. What drives fine-scale movements of large herbivores? A case study using moose. *Ecography* 33 (6), 1102–1112, <http://dx.doi.org/10.1111/j.1600-0587.2009.06104.x>.
- Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73 (4), 423–498, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x>.
- Mair, L., Ruete, A., 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLOS ONE* 11 (1), 1–13, <http://dx.doi.org/10.1371/journal.pone.0147796>.
- Messier, F., 1991. The significance of limiting and regulating factors on the demography of moose and white-tailed deer. *J. Anim. Ecol.* 377–393.
- Monsarrat, S., Boshoff, A.F., Kerley, G.I., 2019. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. *Ecography* 42 (1), 125–136, <http://dx.doi.org/10.1111/ecog.03944>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.03944>.
- NBIC, 2019a. Norwegian biodiversity information centre - other datasets. <http://dx.doi.org/10.15468/tm56sc>, Accessed via GBIF.org on 2019-12-02.
- NBIC, 2019b. Norwegian species observation service. <http://dx.doi.org/10.15468/zjbzel>, Accessed via GBIF.org on 2019-12-02.
- Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., Crowston, K., 2012. The future of citizen science: emerging technologies and shifting paradigms. *Front. Ecol. Environ.* 10 (6), 298–304, URL <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/110294>.
- Pettit, L., 1990. The conditional predictive ordinate for the normal distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 52 (1), 175–184, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1990.tb01780.x>.
- Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* 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. *Ecol. Appl.* 19 (1), 181–197, URL <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-2153.1>.
- Pomeroy, J.W., Gray, D.M., Shook, K.R., Toth, B., Essery, R.L.H., Pietroniro, A., Hedstrom, N., 1998. An evaluation of snow accumulation and ablation processes for land surface modelling. *Hydrol. Process.* 12 (15), 2339–2367, [http://dx.doi.org/10.1002/\(SICI\)1099-1085\(199812\)12:15<2339::AID-HYP800>3.0.CO;2-L](http://dx.doi.org/10.1002/(SICI)1099-1085(199812)12:15<2339::AID-HYP800>3.0.CO;2-L).
- Ramsay, J.O., 1988. Monotone regression splines in action. *Statist. Sci.* 3 (4), 425–441, <http://dx.doi.org/10.1214/ss/1177012761>.

- Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71 (2), 319–392, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00700.x>.
- Sadykova, D., Scott, B.E., De Dominicis, M., Wakelin, S.L., Sadykov, A., Wolf, J., 2017. Bayesian joint models with INLA exploring marine mobile predator–prey and competitor species habitat overlap. *Ecol. Evol.* 7 (14), 5212–5226, URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.3081>.
- Shipley, L., 2010. Fifty years of food and foraging in moose: lessons in ecology from a model herbivore. *Alces: J. Devot. Biol. Manage. Moose* 46, 1–13.
- Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H., 2016. Going off grid: computationally efficient inference for log-Gaussian cox processes. *Biometrika* 103 (1), 49–70. <http://dx.doi.org/10.1093/biomet/asv064>.
- Sørbye, S.H., Illian, J.B., Simpson, D.P., Burslem, D., Rue, H., 2019. Careful prior specification avoids incautious inference for log-Gaussian cox point processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 68 (3), 543–564, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12321>.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64 (4), 583–639, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00353>.
- Strasser, B., Baudry, J., Mahr, D., Sanchez, G., Tancoigne, E., 2019. “Citizen science”? Rethinking science and public participation. *Sci. Technol. Stud.* 32 (ARTICLE), 52–76.
- Title, P.O., Bemmels, J.B., 2018. Envirem: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* 41 (2), 291–307, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02880>.
- Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11, 3571–3594, URL <http://dl.acm.org/citation.cfm?id=1756006.1953045>.
- 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. *Ann. Appl. Stat.* 11 (4), 2270–2297.


Paper II

A Spatial Modeling Framework for Monitoring Surveys with Different Sampling Protocols with a Case Study for Bird Abundance in Mid-Scandinavia.

Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B., & Øien, I. J.
(2022) published in *Journal of Agricultural, Biological and Environmental Statistics*



A Spatial Modeling Framework for Monitoring Surveys with Different Sampling Protocols with a Case Study for Bird Abundance in Mid-Scandinavia

Jorge SICACHA-PARADA , Diego PAVON-JORDAN, Ingelin STEINSLAND, Roel MAY, Bård STOKKE, and Ingar Jostein ØIEN

Quantifying the total number of individuals (abundance) of species is the basis for spatial ecology and biodiversity conservation. Abundance data are mostly collected through professional surveys as part of monitoring programs, often at a national level. These surveys rarely follow exactly the same sampling protocol in different countries, which represents a challenge for producing biogeographical abundance maps based on the transboundary information available covering more than one country. Moreover, not all species are properly covered by a single monitoring scheme, and countries typically collect abundance data for target species through different monitoring schemes. We present a new methodology to model total abundance by merging count data information from surveys with different sampling protocols. The proposed methods are used for data from national breeding bird monitoring programs in Norway and Sweden. Each census collects abundance data following two different sampling protocols in each country, i.e., these protocols provide data from four different sampling processes. The modeling framework assumes a common Gaussian Random Field shared by both the observed and true abundance with either a linear or a relaxed linear association between them. The models account for particularities of each sampling protocol by including terms that affect each observation process, i.e., accounting for differences in observation units and detectability. Bayesian inference is performed using the Integrated Nested Laplace Approximation (INLA) and the Stochastic Partial Differential Equation (SPDE) approach for spatial modeling. We also present the results of a simulation study based on the empirical census data from mid-Scandinavia to assess the performance of the models under model misspecification. Finally, maps of the expected abundance of birds in our study region in

J. Sicacha-Parada (✉) · I. Steinsland, Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
(E-mail: jorge.sicacha@ntnu.no; ingelin.steinsland@ntnu.no)

D. Pavon-Jordan · R. May · B. Stokke, Department of Terrestrial Ecology, Norwegian Institute for Nature Research (NINA), P.O. Box 5685, Torgarden, 7485 Trondheim, Norway

I. J. Øien
Norwegian Ornithological Society-BirdLife Norway, Sandgata 30 B, 7012 Trondheim, Norway.

© 2022 The Author(s)

Journal of Agricultural, Biological, and Environmental Statistics, Volume 27, Number 3, Pages 562–591
<https://doi.org/10.1007/s13253-022-00498-y>

mid-Scandinavia are presented with uncertainty estimates. We found that the framework allows for consistent integration of data from surveys with different sampling protocols. Further, the simulation study showed that models with a relaxed linear specification are less sensitive to misspecification, compared to the model that assumes linear association between counts. Relaxed linear specifications of total bird abundance in mid-Scandinavia improved both goodness of fit and the predictive performance of the models.

Supplementary materials accompanying this paper appear on-line.

Key Words: Data integration; Joint species distribution models; Bayesian statistics; Latent Gaussian Models; Gaussian Random Fields.

1. INTRODUCTION

Understanding why organisms are where they are and what drives changes in their abundances is one of the main pillars of spatial ecology (Brodie et al. 2020) and is critical to propose effective measures to preserve biodiversity. In this regard, species distribution models (SDMs) have typically been used to gain a better understanding of species–habitat relationships (Brodie et al. 2020; Bradter et al. 2021) and to guide conservation practitioners and policy makers (Araujo et al. 2019). Previous SDMs using abundance data have revealed higher predictive performance in comparison with those using occurrence data (Howard et al. 2014; Johnston et al. 2015). Yet, the majority of SDMs published to date used presence/absence (i.e., occurrence) data (Araujo et al. 2019; Yu et al. 2020), rather than abundance data (count of individuals), especially in large-scale studies (Miller et al. 2019). This limits our ability to robustly infer, for example, regions with high density of individuals (Johnston et al. 2015), which is of paramount importance in conservation (Massimino et al. 2017). For example, estimating abundance hotspots can inform and help authorities to select sites that may qualify to be included in the network of protected areas. Indeed, one of the main criteria to identify important areas for conservation under the European Union’s Bird Directive (i.e., Special Protection Areas; SPA) is that a site accommodates regularly 1% of the total biogeographical population of a species of conservation concern or more than 20,000 individuals of wetland birds (EU’s Birds Directive, 2009/147/EC 2009). Moreover, this Directive states that “*The measures to be taken must apply to various factors which may affect the numbers of birds, namely the repercussions of man’s activities and in particular the destruction and pollution of their habitats[...]*”. Abundance data can also be useful to detect and predict areas where human-wildlife conflicts may arise (e.g., May et al. 2020), informing the corresponding authorities that infrastructure and further human development such as siting of powerlines and wind farms must be planned carefully (e.g., De Lucas et al. 2008; May et al. 2020). Information about abundance is ultimately requested by national (e.g., Directorates, Environmental Agencies) and international (e.g. European Commission) authorities as basis to propose biodiversity conservation policies at different scales. This information should be based on all available count data.

Most countries have monitoring programs following national law and as signatories of international biodiversity conservation Directives and Conventions. These different national monitoring schemes may cover the same taxon (e.g., most countries have a national monitoring scheme for breeding birds) but can differ in the species recorded (different set

of species may occur in different countries and at different densities) and, most importantly, they usually follow different sampling protocols, which makes the information obtained by different schemes not directly comparable. Furthermore, not all species are well represented in the data gathered within a single ‘general’ protocol. For this reason, many countries have, for example, additional targeted monitoring schemes that complement the information for species that are considered poorly represented in the more general monitoring scheme, for example, colonial birds such as herons in Greece, raptors and waterbirds in Finland, nocturnal birds in Spain; see also [Buckland and Johnston \(2017\)](#). National common bird monitoring schemes and those targeting particular (groups of) species provide together the largest datasets known on species abundance in time and space. However, at the (sub)national level, these datasets have mainly been used independently ([Kålås 2010](#); [Bevanger et al. 2014](#); [Kéry and Royle 2009](#); [Soykan et al. 2016](#)) and multi-country studies have mostly analyzed these data either independently for each country to later draw common conclusions from the country-specific estimates ([Lehikoinen et al. 2019](#)) or combining the raw data with limited account for sampling differences (e.g., total abundance of waders; [Lindström et al. 2019](#)). Thus, overlooking the potential of integrating such a large amount of standardized data seems like an under usage of the effort and resources spent in collecting these data, especially when the taxa included in such monitoring schemes are very diverse, allowing not only to carry out species-specific analyses but also, potentially, community-level studies. This study was motivated by the need for estimates of the total abundance of birds in mid-Scandinavia based on high quality (i.e., standardized surveys) localized data on bird abundances from the common breeding bird monitoring programs in Norway (TOV-E) and Sweden (BBS). An estimate of the total abundance of birds can be used as an input for models that inform on the risk of infrastructure development (e.g., new powerlines and wind farms) for birds. The TOV-E and the BBS both provide standardized count data, but they differ in their sampling protocols. Both countries collect observations in point counts and transect surveys. In Norway, the main point counts (all species recorded) are complemented with line transects (only a subset of ‘rare’ species also included in point counts are recorded—see further details in Sect. 2). However, in Sweden, the line transects and the point counts can be regarded as two different censuses (i.e., all species are counted in both census methods). These differences present the challenge of integrating the four sources of spatial information (points and transects in both Norway and Sweden) with different sampling protocols into one estimate for the spatial distribution of bird abundance for the entire region of interest ([Brodie et al. 2020](#); [Gruss and Thorson 2019](#)).

The scarcity of studies applying large-scale abundance SDMs is likely related to (i) the generally lower availability of abundance data compared to occurrence data for most species ([Miller et al. 2019](#); [Buckland and Johnston 2017](#) and references therein) and (ii) statistical and computational challenges of modeling abundance data. Great methodological advancements to overcome some of these problems have been developed in the past decade, especially for integrating different data types, see [Miller et al. \(2019\)](#) and references therein. Most of these efforts have focused on enabling the use of casually collected (non-standardized) presence-only data to increase spatial coverage and data points of certain species (see also [Buckland and Johnston 2017](#)). The possibility of improving SDMs by integrating abundance (count) data collected under different standardized monitoring schemes

is most often neglected. Thus, in addition to the integration of data from different countries, merging data from different schemes (from one or several countries) can thus improve the estimates of abundance obtained from all available count data.

Given the existing gap in methodology for proper integration of standardized count data, we here propose a generic modeling framework that integrates standardized count data from various monitoring schemes (i.e., designed surveys) with different sampling protocols. The models can ultimately produce one single estimate of abundance (total abundance of birds in our case study) and its uncertainty based in data from different sampling protocols. In addition, it also gives interpretable estimates of the ecological parameters driving this abundance. Our methodology, thus, analyzes these data in a unique, single framework to produce models that account for different sampling processes, and describe and predict the spatial distribution of abundance.

Spatial modeling of multiple data sources has been approached for example in the context of coregionalization models (Banerjee et al. 2015; Blangiardo and Cameletti 2015; Krainski 2019) and recently reviewed in Miller et al. (2019). These are multivariate models for measurements that vary jointly over a region and have been defined through a hierarchical structure and fitted using Markov Chain Monte Carlo (MCMC) techniques (Banerjee et al. 2015). For the family of Spatial Latent Gaussian Models (Rue and Held 2005), the INLA-SPDE approach (Rue et al. 2009; Lindgren et al. 2011) and its easy implementation in the INLA library of R have emerged as a faster alternative to jointly model multiple sources of information. Such method has been applied to multivariate models related with, for example, air pollution data (Cameletti et al. 2019), and hydrology (Roksvåg et al. 2020). The proposed framework assumes the existence of a latent process, underlying all the observed abundances, that represents the true expected abundances. The true expected abundance varies in space through spatial covariates as well as a spatial random effect. Given the true expected abundance, we assume that the observed abundances follow Poisson distributions. For each observation process a linear relation between the expected counts and the true expected abundances is assumed. Further, we assume the existence of a common spatial random effect that drives the observed counts (cf. Miller et al. 2019) for all the observation processes. Given that the linear assumption may not depict the true relationship between the expected counts and the true expected abundances, we also propose models that allow deviations from this assumption. The proposed models are suitable doing computational fast inference using the INLA-SPDE approach, which approximates the posterior densities of parameters and hyperparameters.

To the best of our knowledge, methodologies for jointly modeling spatial abundance using data from multi-country standardized biodiversity monitoring programs with different sampling protocols have not been published before. By properly integrating data from different monitoring schemes, our method can be part of solving some of the issues inherent to monitoring data raised in Buckland and Johnston (2017), such as the scarcity of data, low representability, and small geographical scale. This opens new possibilities for more robust international assessments of species distributions and abundance using count data from diverse national monitoring programs, which is of paramount importance for understanding global change impacts on biodiversity (Buckland and Johnston 2017; Massimino et al. 2017). We validate this framework with a case study aiming at estimating total bird

abundance in mid-Scandinavia and a simulation study that explores the effects of misspecification on the proposed models.

This paper is organized as follows: In Sect. 2, we describe the data from the Norwegian and Swedish monitoring programs in detail. Moreover, we explain how we preprocessed these census data, present an exploratory analysis and introduce the set of candidate explanatory variables for our models. In Sect. 3, models as well as inference methodology and measures for evaluating and comparing models are presented. In Sect. 4, we set up a simulation study to explore how the proposed models perform in scenarios with different relation between the observed and the true abundances. In Sect. 5, results of both the simulation study and the case study using bird counts in mid-Scandinavia are presented. The paper finishes in Sect. 6 with the discussion of the results and concluding remarks.

2. BIRD MONITORING SURVEYS DATA

2.1. TOV-E AND BBS DATA

The Norwegian common bird monitoring scheme (TOV-E), coordinated by the Norwegian Institute for Nature Research (NINA) and the Norwegian Ornithological Society (NOF) since 2006, was established to monitor population variation for common breeding terrestrial birds on a national scale in a representative way. Surveys (i.e., count of pairs of birds of all observed species) are carried out by experienced ornithologists that follow a standardized protocol (Kålås 2002). Each census route ($n = 492$) contains between 12 and 20 (average = 18.8) point counts 300 m apart describing a square (see Fig. 1) with side = 1.5 km (deviation of this shape are allowed and recorded when the geographic/topographic conditions do not allow the observer to walk, e.g., sea/lakes, glaciers, rough mountainous terrain). A total of 229 species are heard or seen at the entirety of the point counts of TOV-E during 5 minutes. Approximately 121 of the species are less abundant and/or difficult to detect, so observers are asked to record these species during a line transect between point counts (see Fig. 1—figure with the configuration of a census site with the twenty points). A random selection of 370 census routes (out of a total of 492 routes across Norway) is visited once a year during the period 20th May to 10th June. TOV-E is designed to cover all relevant habitats throughout the altitudinal and latitudinal gradient in Norway and reports ‘pairs of individuals’ as sampling unit. The Swedish breeding bird survey (hereafter BBS) has been coordinated by Lund University since 1996 and consists of 716 fixed sites across Sweden within a 25-km grid (one route per grid cell, see Lindstrom et al. 2013). These sites are surveyed once a year between mid-May and mid-June (the breeding period for most bird species in Sweden) though not all sites are surveyed every year (mean = 353 sites per year). The 25-km grid makes sure that the habitats of Sweden are monitored in proportion to their abundance in the country as well as the entire altitudinal and latitudinal gradient where birds are present. At each site, the observer walks an 8-km transect describing a 2×2 km square and records all bird species heard and/or seen within 8 h. In addition, the observer has eight 5-min point counts where all birds seen or heard must also be recorded. The point counts take place at each of the corners of the square and at the middle point of the transect (see Fig. 1). Of the circa 250 species breeding in Sweden, 244 are reported in BBS, thus

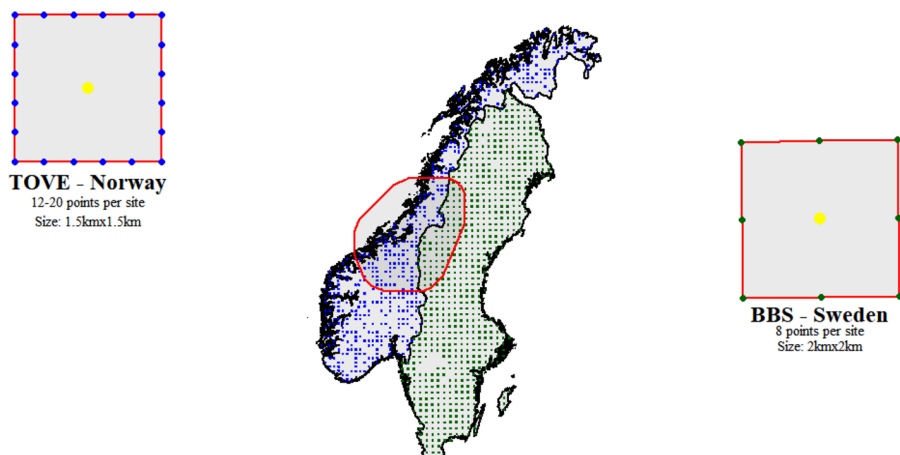


Figure 1. Spatial location of census sites and sampling points and line transects according to each sampling protocol. Left: graphical display of sampling protocol of TOV-E census. Blue points: 20 locations for point counts (the number of points vary between 12 and 20 in different sites). Red lines: line transects. Yellow point: centroid associated with each census site (see Sect. 2.2). Center: spatial distribution of census sites across Norway (blue sites) and Sweden (green sites). The red polygon represents the study area described in Sect. 2.1. Right: graphical display of sampling protocol of BBS census. Green points: 8 locations for point counts. Red lines: line transects. Yellow point: centroid associated with each census site (see Sect. 2.2).

ensuring a good coverage of the breeding birds (Lindstrom et al. 2013). The BBS reports ‘individuals’ as sampling unit, which differs from TOV-E’s reporting unit (pairs; see above).

Although these monitoring programs are designed to cover a large part of both countries (Fig. 1), for our case study, we only selected census sites that lie within a polygon defined to produce an approximation of a Gaussian Random Field and make inference about a point pattern in Trøndelag Country, central Norway (see red polygon in Fig. 1, (Lindgren et al. 2011; Simpson et al. 2016)). This polygon covers a total area of 173.634 km² and contains 113 census sites in Norway and 70 in Sweden. The main motivation to reduce the study region from the entire country to a smaller area (defined by the polygon) was strictly computational and for an easier compilation of covariate information. In addition, this region, which is basically within Trøndelag County in central Norway, is largely representative of habitat types, topography and biodiversity found elsewhere in Norway.

2.2. EXPLORATORY ANALYSIS

Our main goal was to develop and validate a new modeling framework to integrate abundance data from standardized monitoring schemes with different sampling protocols. Such a framework can ultimately be used, for example, to detect hotspots of abundance of birds, as in the case we illustrate here (note: we are not interested in the distribution of particular species, but in the distribution of total abundance of birds regardless of the species). In other words, we apply our modeling framework to produce maps of total abundance of birds based on count data from multiple sources—information gathered as part of standardized national bird monitoring schemes in Norway and Sweden that differ in the sampling protocols. The

data preparation consisted in averaging across all years (2006–2019) the total count of all individuals (regardless of the species) found at each survey site. That is, we first added up the counts of all individual birds recorded in the points or lines of a given census site and assigned this total count of individuals (regardless of the species present) to the site's centroid (see Fig. 1) so that each census site will have one single value of total abundance of birds per year. Next, for each site, we averaged the yearly total abundance of birds across all years that the site was sampled (note: not all sites are censused every year) in the period between 2006 and 2019, so that we ended up with one single value of total abundance of birds per site (temporal average). Although estimating single-species abundance and distribution maps are commonly used to inform about species of conservation concern, here we wanted to report the total abundance of birds across the region (note: our methodology can also be used to estimate single-species abundances). Estimating total abundance of individuals across a region (as opposed to single-species abundance) has clear implications in spatial conservation planning and prioritization (Lehtomäki and Moilanen 2013). For example, De Lucas et al. (2008) estimated the total abundance of raptors in a region to assess the impacts of wind farms on this group of birds. Lindström et al. (2019) attempted to estimate total density of wading birds across Fennoscandia by combining count data from Norway, Sweden and Finland. However, they did not account for many differences in the sampling protocols. Our modeling framework thus can be applied to account for such differences. Another example of potential use of our method is to get more robust estimates of total abundance of birds to inform authorities and stakeholders where powerlines (Bevanger et al. 2014) or wind farms (De Lucas et al. 2008) may cause large mortality rates. Although here we present a simplified and more generic analysis (all species have weight = 1, and thus their abundance has the same influence in the resulting map), each species abundance can be multiplied (weighted) by a factor relative to their sensitivity to e.g., powerlines (D'Amico et al. 2019) so that the resulting map will highlight total abundance hotspots in relation to their sensitivity to the particular issue. Since we include data from both Norway and Swe-

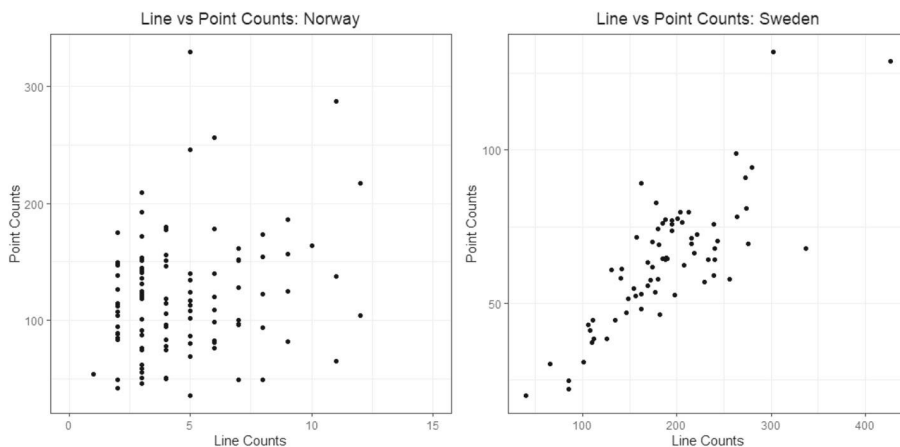


Figure 2. Scatterplots of line vs point counts in Norway (number of pairs, left) and Sweden (number of individuals, right).

den, we explore how the relation of point and line counts differ between surveys from both countries. In Fig. 2, we display a scatterplot with the points and line counts at each of the TOV-E ($n=113$) and BBS ($n=70$) sites.

These scatterplots show a linear relation between point and line counts in Sweden, whereas in Norway there is no clear linear association between the counts in points and lines. This is somehow expected due to the census design in Norway, where the line counts are meant to record a reduced subset of species compared to the point counts. This is a common issue highlighted by [Buckland and Johnston \(2017\)](#) and is often found in many countries when certain species are monitored with special censuses in addition to the general monitoring scheme. Therefore, this is not only an issue when integrating between-country datasets (e.g., to increase the geographical extent), but also within-country datasets (to increase the representability and number of data points).

2.3. EXPLANATORY VARIABLES

In our case study, we want to apply our new methodology not only to estimate total abundance of birds, but also to produce interpretable estimates of ecological factors associated with it across the region. We have selected three candidate ecological factors that are commonly used in SDMs to explain distribution of birds (e.g. [Bradter et al. 2021](#); [Lissovsky et al. 2021](#); [Soultan et al. 2022](#)): (i) climatic variables—temperature (average daily temperature from April to July over 2006–2019, downloaded from [seNorge.no](#)) and precipitation (average daily precipitation from April to July over 2006–2019, downloaded from [seNorge.no](#)), (ii) topography - elevation (Digital Elevation Model at a 10m resolution, DEM10, downloaded from <https://kartkatalog.geonorge.no/>), and (iii) the land cover surrounding each location expressed as the percentage of each of the following six land covers (urban, mountains, rocky area, water body, forest, and open area) in a square neighborhood of $2\text{km} \times 2\text{km}$. Land cover information was depicted from the N50 layer (downloaded from <https://kartkatalog.geonorge.no/>). All rasters files have resolution of $1\text{km} \times 1\text{km}$ (the elevation data from DEM10 was aggregated to this resolution prior analysis) and are shown in the Supplementary Information. As a first stage of model selection, we computed the correlation coefficient between all the candidate covariates on a fine grid of about 600,000 points. Only one variable in those pairs with $|\rho| > 0.7$ was left as a candidate. Those pairs with high correlation were: 1) elevation and temperature ($\rho = -0.81$). Temperature was discarded; 2) % of open area and % of forest ($\rho = -0.83$). % of open area was discarded.

3. MODELING AND INFERENCE APPROACH

The specification of our models relies on the assumption that our four sources of observations are obtained from a common underlying ecological process ([Miller et al. 2019](#)). This assumption arguably makes sense if we consider the fact that national borders of neighboring countries are not, in general, a key factor for natural changes in biodiversity, although there might be slight differences in conservation policies and governance. Hence, we can assume that a common nonzero mean Gaussian Random Field (GRF) is involved in the

generation of the number of individuals at each census site. However, the two different sampling protocols (points and lines), which also differ between the two countries (complementary surveys in Norway and independent surveys in Sweden), result in four groups of counts observed. Moreover, TOV-E counts (Norway) are reported as ‘number of pairs’ of each species, whereas BBS counts (Sweden) are reported as ‘number of individuals’ of each species. Therefore, direct inference and comparisons between these four response variables should be made with caution. The true total bird counts random variable, $Y_{\text{true}}(\mathbf{s})$ with $\mathbf{s} \in D \subset \mathbb{R}^2$, is assumed to follow a Poisson distribution with expected value $\lambda_{\text{true}}(\mathbf{s})$, expressed as

$$\log(\lambda_{\text{true}}(\mathbf{s})) = X^T(\mathbf{s})\beta + \omega_1(\mathbf{s}) \quad (1)$$

with $X^T(\mathbf{s})$ a set of spatial covariates and $\omega_1(\mathbf{s})$ a zero-mean GRF that aims at accounting for residual spatial dependency. Both $X^T(\mathbf{s})$ and $\omega_1(\mathbf{s})$ can include well-established factors that influence variation in the total abundance of birds; in our case study these factors are precipitation and elevation. We assume a Matérn covariance function for $\omega_1(\mathbf{s})$

$$\frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} (\kappa \|s_i - s_j\|)^\nu K_\nu(\kappa \|s_i - s_j\|) \quad (2)$$

with $\|s_i - s_j\|$ the Euclidean distance between two locations $s_i, s_j \in D$. σ^2 stands for the marginal variance, and K_ν represents the modified Bessel function of the second kind and order $\nu > 0$. ν is the parameter that determines the degree of smoothness of the process, while $\kappa > 0$ is a scaling parameter. For $\omega_1(\mathbf{s})$, let $\kappa = \kappa_{1,\nu} = \nu_1$ and $\sigma^2 = \sigma_1^2$. We assume that the observed counts for each sampling protocol are realizations of four random variables conditionally independent given the true abundance, $\lambda_{\text{true}}(\mathbf{s})$. That is, we assume the four groups of observed counts are realizations of the Poisson random variables:

$$\begin{aligned} Y_1(\mathbf{s}) &\sim \text{Poisson}(\lambda_1(\mathbf{s})) && \text{(Point counts in Norway)} \\ Y_2(\mathbf{s}) &\sim \text{Poisson}(\lambda_2(\mathbf{s})) && \text{(Line counts in Norway)} \\ Y_3(\mathbf{s}) &\sim \text{Poisson}(\lambda_3(\mathbf{s})) && \text{(Point counts in Sweden)} \\ Y_4(\mathbf{s}) &\sim \text{Poisson}(\lambda_4(\mathbf{s})) && \text{(Line counts in Sweden)} \end{aligned}$$

where $\lambda_j(\mathbf{s})$, $j = \{1, 2, 3, 4\}$ are the expected values of the random variables $Y_j(\mathbf{s})$. Additionally, we assume $Y_1(\mathbf{s}) + Y_2(\mathbf{s}) \approx Y_{\text{NO}}(\mathbf{s})$ as a proxy for total abundance since the line transects are complementary to the point counts in Norway. This assumption does not hold for Sweden since, as mentioned in Sect. 1, line transects and point counts are regarded as two different independent censuses. In case we wanted to suggest a proxy for the total abundance in Sweden using $Y_3(\mathbf{s})$ and $Y_4(\mathbf{s})$, we would need to account for a potential overlap (double counting) between the counts observed in points and line transects. Given that we assume a common latent process underlying all the observed abundances, $Y_1(\mathbf{s}) + Y_2(\mathbf{s})$ works also as a proxy for total abundance of birds in Sweden. This variable is used to produce the predicted total abundance of birds in Sect. 3. Our final assumption is that there

are no differences in observer skills between countries since the census are performed by experienced ornithologists.

3.1. MODELS

In this section, we introduce three model specifications for integrating data from the four sampling protocols introduced in Sect. 2. Model 1 (see Sect. 3.1.1) is a model that assumes a linear relation between the expected counts of the four sampling protocols. This is achieved by the introduction of a unique intercept for each sampling scheme. In Sect. 3.1.2, model 2 is presented. This model allows for a relaxation of the assumption of linear relation between expected counts by incorporating terms that allow to explain any deviation from this assumption through the GRF $\omega_1(\mathbf{s})$. Finally, model 3 (see Sect. 3.1.3) is introduced. This model adds a second GRF, $\omega_2(\mathbf{s})$, which aims to account for spatial sources of variation not accounted for in the other parts of the model and not explained by known covariates, (Simmonds et al. 2020; Selle et al. 2020). It is worth noting that as each of the models proposed depend on $\lambda_{\text{true}}(\mathbf{s})$, they explicitly account for the factors that influence the variation in abundance.

3.1.1. Model 1

Based on our exploratory analysis and the four sampling processes present in our dataset, in model 1 we assumed a linear relation between the expected values of the four random variables representing each sampling protocol and $\lambda_{\text{true}}(\mathbf{s})$. That is,

$$\begin{aligned}\lambda_1(\mathbf{s}) &= \zeta_1^* \cdot \lambda_{\text{true}}(\mathbf{s}); & \log(\zeta_1^*) &\sim N(0, \tau_1^*) \\ \lambda_2(\mathbf{s}) &= \zeta_2^* \cdot \lambda_{\text{true}}(\mathbf{s}); & \log(\zeta_2^*) &\sim N(0, \tau_2^*) \\ \lambda_3(\mathbf{s}) &= \zeta_3^* \cdot \lambda_{\text{true}}(\mathbf{s}); & \log(\zeta_3^*) &\sim N(0, \tau_3^*) \\ \lambda_4(\mathbf{s}) &= \zeta_4^* \cdot \lambda_{\text{true}}(\mathbf{s}); & \log(\zeta_4^*) &\sim N(0, \tau_4^*)\end{aligned}\quad (3)$$

with $\zeta_j^* \geq 0$, $j = 1, \dots, 4$ the factors that determine the association between the observed and the true counts for each protocol. In real-life problems, ζ_j^* can explain multiple sources of variation that are common to sampling of bird species such as observer differences, observed units, differences in detection probability, among others. The inclusion of this term is also useful to deal with overdispersion (Gomez-Rubio 2020), a common issue when working with count data. In order to avoid identifiability issues, we restate the model in (3) in terms of $\lambda_1(\mathbf{s})$. That is,

$$\begin{aligned}\lambda_2(\mathbf{s}) &= \zeta_2 \cdot \lambda_1(\mathbf{s}); & \log(\zeta_2) &\sim N(0, \tau_2) \\ \lambda_3(\mathbf{s}) &= \zeta_3 \cdot \lambda_1(\mathbf{s}); & \log(\zeta_3) &\sim N(0, \tau_3) \\ \lambda_4(\mathbf{s}) &= \zeta_4 \cdot \lambda_1(\mathbf{s}); & \log(\zeta_4) &\sim N(0, \tau_4)\end{aligned}\quad (4)$$

where $\zeta_j \geq 0$ and $\zeta_j = \frac{\zeta_j^*}{\zeta_1^*}$, $j = \{2, 3, 4\}$.

3.1.2. Model 2

In model 2, we relax the assumption of linear relation between the expected value of the number of observed individuals with protocol j , $\lambda_j(\mathbf{s})$, and the true intensity, $\lambda_{\text{true}}(\mathbf{s})$, by including spatial varying terms $(\psi_j^* - 1) \cdot \omega_1(\mathbf{s})$, $j = \{1, 2, 3, 4\}$. These terms aim to explain any deviation from a linear relation between expected values as a function of a GRF $\omega_1(\mathbf{s})$. It is worth noting that model 1 (see above) is a special case of model 2 with $\psi_j^* = 1$. We define model 2 as:

$$\begin{aligned}\lambda_1(\mathbf{s}) &= \zeta_1^* \cdot \lambda_{\text{true}}(\mathbf{s}) \cdot \exp\{(\psi_1^* - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_1^*) &\sim N(0, \tau_1^*) \\ \lambda_2(\mathbf{s}) &= \zeta_2^* \cdot \lambda_{\text{true}}(\mathbf{s}) \cdot \exp\{(\psi_2^* - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_2^*) &\sim N(0, \tau_2^*) \\ \lambda_3(\mathbf{s}) &= \zeta_3^* \cdot \lambda_{\text{true}}(\mathbf{s}) \cdot \exp\{(\psi_3^* - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_3^*) &\sim N(0, \tau_3^*) \\ \lambda_4(\mathbf{s}) &= \zeta_4^* \cdot \lambda_{\text{true}}(\mathbf{s}) \cdot \exp\{(\psi_4^* - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_4^*) &\sim N(0, \tau_4^*)\end{aligned}\quad (5)$$

Again, to avoid identifiability issues, we restate the model in (5) in terms of $\lambda_1(\mathbf{s})$ as:

$$\begin{aligned}\lambda_2(\mathbf{s}) &= \zeta_2 \cdot \lambda_1(\mathbf{s}) \cdot \exp\{(\psi_2 - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_2) &\sim N(0, \tau_2) \\ \lambda_3(\mathbf{s}) &= \zeta_3 \cdot \lambda_1(\mathbf{s}) \cdot \exp\{(\psi_3 - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_3) &\sim N(0, \tau_3) \\ \lambda_4(\mathbf{s}) &= \zeta_4 \cdot \lambda_1(\mathbf{s}) \cdot \exp\{(\psi_4 - 1) \cdot \omega_1(\mathbf{s})\}; & \log(\zeta_4) &\sim N(0, \tau_4)\end{aligned}\quad (6)$$

In the scales of the linear predictors in (5), $\psi_j = \psi_j^* - \psi_1^* + 1$, $j = \{2, 3, 4\}$ are scaling coefficients for the common GRF, $\omega_1(\mathbf{s})$, in each likelihood. They quantify to what extent the departure of the assumption of linearity is explained by $(\psi_j^* - 1) \cdot \omega_1(\mathbf{s})$. In real-life scenarios, this departure can be related with sources of variation with spatial structure such as differences in detectability, among others. Therefore, we would expect posterior densities for ψ_3 and ψ_4 to be around 1 in our case study, while for ψ_2 we expect different results because line and point counts in Norway do not seem to follow a linear relation (see Sect. 2; Fig. 2). Due to different characteristics of line transect surveys in Norway, we propose model 3.

3.1.3. Model 3

In addition to causing departure from a linear relation between true and observed counts, species detectability may also change with the census technique used (i.e., one of our data sources, the line transects in TOV-E, targeted only a subset of species as it is regarded as a complementary survey to the point counts). Hence, in model 3 we included a second GRF, $\omega_2(\mathbf{s})$ to try to account for the characteristics of this observation process. In case that no explanatory variable that explains the particular characteristics of the sampling protocol is available, a second GRF can be added as a way to account for them, (Simmonds et al. 2020). This is included as an additive term in the linear predictor, as follows:

$$\begin{aligned}\lambda_2(\mathbf{s}) &= \zeta_2 \cdot \lambda_1(\mathbf{s}) \cdot \exp\{(\psi_2 - 1)\omega_1(\mathbf{s})\} \cdot \exp\{\omega_2(\mathbf{s})\} \\ \lambda_3(\mathbf{s}) &= \zeta_3 \cdot \lambda_1(\mathbf{s}) \cdot \exp\{(\psi_3 - 1)\omega_1(\mathbf{s})\} \\ \lambda_4(\mathbf{s}) &= \zeta_4 \cdot \lambda_1(\mathbf{s}) \cdot \exp\{(\psi_4 - 1)\omega_1(\mathbf{s})\}\end{aligned}\quad (7)$$

We assume a Matérn covariance function as in (2) for $\omega_2(\mathbf{s})$, with parameters $\kappa = \kappa_2$, $\nu = \nu_2$ and $\sigma^2 = \sigma_2^2$

3.1.4. Prior Specification

For the GRFs $\omega_k(\mathbf{s})$, $k = \{1, 2\}$, the parameters ν_k in the Matérn covariance function are fixed to be 1. The interest is put on the spatial ranges ρ_k , and on the standard deviation of the GRFs, σ_k . ρ_k are related to κ_k through $\rho_k = \sqrt{8}/\kappa_k$. The prior distributions of these two parameters are specified by making use of Penalized Complexity (PC) priors, (Fuglstad et al. 2019). In this case, we set $P(\rho_1 < 20000) = 0.1$ and $P(\sigma_1 > 1) = 0.1$ for $\omega_1(\mathbf{s})$, while $P(\rho_2 < 2000) = 0.1$ and $P(\sigma_2 > 3) = 0.1$ for $\omega_2(\mathbf{s})$. This means, for example, that under this prior specification, a standard deviation greater than 1 is regarded as large, while a spatial range below 20 kilometers is considered unlikely for $\omega_1(\mathbf{s})$. The parameters in β have Normal prior with mean 0 and precision 0.01. Let $\log(\zeta_j) \sim N(0, \tau_j)$, $j = \{2, 3, 4\}$, where the logarithm of each τ_j has a log-Gamma prior with parameters 1 and 0.00005. For the parameters ψ_j , $j = 2, 3, 4$ in models 2 and 3, we set a normal prior with mean 1 and precision 0.1. We have now defined a group of three candidate models. In the upcoming subsections, we introduce the methodological approach for fitting them and for selecting a model that suits best for our problem.

3.2. INFERENCE AND COMPUTATIONAL APPROACH

The models introduced in Sect. 3.1 were fitted making use of the Integrated Nested Laplace Approximation (INLA), (Rue et al. 2009) and the Stochastic Partial Differential Equation (SPDE) approach (Lindgren et al. 2011). INLA is a faster alternative to Monte Carlo Markov Chains (MCMC) for performing Bayesian inference for latent Gaussian models. INLA aims at producing a numerical approximation of the marginal posterior distribution of the parameters and hyperparameters of the model. Further details can be found in Rue et al. (2009) and Blangiardo and Cameletti (2015). Since we deal with continuous spatial processes in our models, the SPDE approach emerges as an efficient representation of $\omega_1(\mathbf{s})$ and $\omega_2(\mathbf{s})$. It is based on the solution of a SPDE which can be approximated through a basis function representation defined on a triangulation of the spatial domain. More details are available in Lindgren et al. (2011) and Blangiardo and Cameletti (2015).

3.3. ASSUMPTIONS AND POSSIBLE EXTENSIONS

This new modeling framework is developed to integrate count data collected in designed surveys that follow different standardized protocols. Particularly, in the case study presented here, the bird surveys introduced in Sect. 2 are designed to minimize biases due to variation in the time of sampling or observer expertise. For this reason, the models presented in our case study assume, in principle, that these external sources of variation that could affect the observation process are constant across sites or negligible. However, these models are flexible enough to explicitly account for factors that may affect the observation process of each sampling protocol, and can thus be accounted for. There may be, however, other potential sources of variation when working with monitoring data, which also depend on the

taxon being surveyed. Hence, as mentioned in Sect. 3, our method includes relevant terms for quantifying the effect of potential sources of noise in the observation process. Our models incorporate the terms ζ to explain what proportion of the true abundance is explained by each of the observation processes. That is, ζ_j quantifies the effect of each sampling protocol on the observed abundances. This effect comprises sources of variation such as differences in the observed units, differences in detectability, and potential differences in the expertise of the observers. In many real-life scenarios, these terms do not provide enough quantification of the effect of the sampling protocols as there are sources of variation in the sampling process that have spatial variation that cannot be summarized in one term. Therefore, the Gaussian Random Field that drives the true abundance (in our case study, the total abundance of birds) or a second GRF is also used to account for sources of variation that have a spatial behavior. This modeling framework also allows to explicitly account for factors that affect the observation process of each sampling protocol. To show how this can be done, we take model 2 as our reference to explicitly account for a factor that influences the observed number of individuals. We now assume that unlike our case study, there are several factors affecting the observed total abundance of birds. As seen in equation (6) in Sect. 3.1.2, the term $\zeta_j \cdot \exp\left\{(\psi_j - 1)\omega_1(\mathbf{s})\right\}$ accounts for the effect the sampling protocol j has on the observed abundance. In addition to the spatial effect driven by $\omega_1(\mathbf{s})$, the term ζ_j can be further explained, for example, by a fixed effect z as follows:

$$\zeta_j = \alpha_{0j} + \alpha_{1j}z \quad (8)$$

This is a straightforward way to explicitly account for multiple factors that may influence the observation process of the sampling protocol j . Factors with a spatial or temporal structure can be accounted for through random effects with these structures. Given the additional parameters to be estimated and the increased complexity of the model when the effect of these factors is accounted for explicitly, structural identifiability issues may arise. Therefore, in order to overcome these issues, it is recommended to constrain the parameters in (8). This can be achieved by either having additional data that inform on these factors or informative prior information of the parameters involved in (8). Acquiring additional data to account for factors that affect the observation process of each sampling protocol might be possible by integrating data, for example, from schemes with sampling protocols designed to gather information on species detection probabilities through repeated visits to the sites or distance sampling (Järvinen and Väisänen 1983; Miller et al. 2019). In our case study, the temporal variation in birds is not considered to compute the total abundance of birds across the study region. Rather, this temporal variation is removed by averaging the total count of birds at each site over the 14 years (2006–2019). This is also a convenient assumption as we do not have information (counts) at every census site every year (i.e., not all sites are surveyed every year). Furthermore, we believe that the overall state of important sites for birds has remained similar in the past 14 years (i.e. bird-rich areas in 2006, at the beginning of the monitoring scheme are still bird-rich areas in 2019, even if the species composition might have changed slightly).

3.4. MODEL ASSESSMENT

In order to assess and compare competing models such as the ones we are fitting in upcoming sections, we employed the Deviance Information Criterion (DIC), (Spiegelhalter et al. 2002), the Watanabe–Akaike Information Criterion (WAIC), (Watanabe 2010), the logarithm of the pseudo marginal likelihood (LPML) (Blangiardo and Cameletti 2015) and the Continuous Rank Probability Score (CRPS) (Gneiting and Raftery 2007).

DIC makes use of the deviance of the model

$$D(\theta) = -2 \log(p(\mathbf{y}|\theta))$$

to compute the posterior mean deviance $\bar{D} = E_{\theta|\mathbf{y}}(D(\theta))$. In order to penalize the complexity of the model, the effective number of parameters

$$p_D = E_{\theta|\mathbf{y}}(D(\theta)) - D(E_{\theta|\mathbf{y}}(\theta)) = \bar{D} - D(\bar{\theta})$$

is added to \bar{D} . Thus,

$$\text{DIC} = \bar{D} + p_D.$$

The Watanabe–Akaike Information Criterion is based on the posterior predictive density, which makes it preferable to the Akaike and the deviance information criteria, since according to Gelman et al. (2014) it averages over the posterior distribution rather than conditioning on a point estimate. It is empirically computed as

$$-2 \left[\sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S p(y_i | \theta^s) \right) + \sum_{i=1}^n V_{s=1}^S (\log p(y_i | \theta^s)) \right]$$

with θ^s a sample of the posterior distribution and $V_{s=1}^S$ the sample variance. Another criterion to compare the models is LMPL, defined as:

$$\text{LPML} = \sum_{i=1}^n \log(\text{CPO}_i)$$

It depends on CPO_i , the Conditional Predictive Ordinate at location \mathbf{s}_i , (Pettit 1990), a measure that assesses the model performance by means of leave-one-out cross validation. It is defined as:

$$\text{CPO}_i = p(y_i^* | y_f)$$

with y_i^* the prediction of y at location \mathbf{s}_i and $y_f = y_{-i}$. Lastly, we will compare the predictive performance of our models using the Continuous Rank Probability Score (CRPS). It makes possible to compare the estimated posterior mean and our observed values while accounting for the uncertainty of the estimation, (Gneiting and Raftery 2007; Selle et al. 2019). It is defined as:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(u) - 1\{y \leq u\})^2 du$$

with F , the cumulative distribution of the estimated posterior mean, and y is the observed value. The smaller CRPS is, the closer the estimated value is to the observed one.

4. SIMULATION STUDIES

We set up three simulation studies based on the case study of total abundance of birds in mid-Scandinavia that allow us to assess the performance of the models proposed in Sect. 3, when the true data generating model either assume linear relation between the counts (Scenario 1), deviate from this assumption due to some spatial factor explained by a GRF (Scenario 2) or when one group of observed counts is considerably affected by additional spatial sources of variation (Scenario 3). We used the same sites as the observations in the TOV-E and BBS surveys (Fig. 1). To start, we simulated the true intensity, $\lambda_{true}(\mathbf{s})$ as:

$$\log(\lambda_{true}(\mathbf{s})) = \beta_0 + \beta_1 \text{PREC}(\mathbf{s}) + \omega_1(\mathbf{s})$$

with $\text{PREC}(\mathbf{s})$, the precipitation at location \mathbf{s} in the study region (see Figure S.1.), and $\omega_1(\mathbf{s})$ a GRF with range $\rho = 15\text{km}$ and $\sigma^2 = 0.14$. Further, we specified $\beta_0 = 4.70$ and $\beta_1 = -0.20$. These values were chosen based on the posterior marginal distribution of these parameters in the real-data application. Next, we simulated observations representing the surveys, i.e., using four different Poisson models with parameters $\lambda_j(\mathbf{s})$, $j = \{1, \dots, 4\}$. Table 1 summarizes the two simulation scenarios proposed for $\lambda_j(\mathbf{s})$

For each scenario, we simulated 100 datasets with $\zeta_1^* = 0.91$, $\zeta_2^* = 0.04$, $\zeta_3^* = 0.57$ and $\zeta_4^* = 1.72$. While we assume a linear relation between $\lambda_j(\mathbf{s})$ and $\lambda_{true}(\mathbf{s})$ in Scenario 1, in Scenarios 2 and 3 the relation between $\lambda_j(\mathbf{s})$ and $\lambda_{true}(\mathbf{s})$ is assumed to follow (5) with $\psi_1^* = 1$, $\psi_2^* = 1.57$, $\psi_3^* = 1.09$ and $\psi_4^* = 1.21$. These settings are based on the posterior marginal distribution of the parameters in the real data case study (presented in Sect. 5.2). The three simulation scenarios closely mimicked real data application by making two of the simulated counts only observed in Norway and the other two only observed in Sweden. For each simulated dataset, we fitted the three models proposed in Sect. 3. A second group of simulation scenarios was proposed by taking more extreme values of the posterior marginal distributions. The results and more details on this simulation scenario are discussed in Sect. 5.1 and the supplementary information.

Table 1. Simulation scenarios

Scenario	Simulated $\lambda_j(\mathbf{s})$
1	$\lambda_j(s) = \zeta_j^* \cdot \lambda_{true}(\mathbf{s})$
2	$\lambda_j(s) = \zeta_j^* \cdot \lambda_{true}(\mathbf{s}) \cdot \exp((\psi_j^* - 1) \cdot \omega_1(\mathbf{s}))$
3	$\lambda_j(s) = \zeta_j^* \cdot \lambda_{true}(\mathbf{s}) \cdot \exp((\psi_j^* - 1) \cdot \omega_1(\mathbf{s})); j = \{1, 3, 4\}$ $\lambda_2(s) = \zeta_2^* \cdot \lambda_{true}(\mathbf{s}) \cdot \exp((\psi_2^* - 1) \cdot \omega_1(\mathbf{s}) + \omega_2(\mathbf{s}))$

To assess the performance of each model in each scenario, we simulated 10000 realizations $\{\theta_{jkl}^p\}$, $j = 1, \dots, 10000$, from the posterior distribution of each parameter θ for dataset $k = 1, \dots, 100$ in scenario $l = 1, 2, 3$. Thus, the mean bias and the Root Mean Square Error (RMSE) for dataset k in scenario l are computed as:

$$\text{bias}_{kl} = \frac{1}{10000} \sum_{j=1}^{10000} (\theta_{jkl}^p - \tilde{\theta})$$

$$\text{RMSE}_{kl} = \sqrt{\frac{1}{10000} \sum_{j=1}^{10000} (\theta_{jkl}^p - \tilde{\theta})^2}$$

with $\tilde{\theta}$ the true value of the parameter θ .

5. RESULTS

5.1. SIMULATION STUDIES

The 100 datasets generated in each of the proposed scenarios were fitted using the three proposed models in Sect. 3 and the results summarized here using the measures of performance introduced in Sect. 4. We only show the mean bias and RMSE for the parameters ζ_2^* , ζ_3^* and ζ_4^* as they are key to understand how different response variables interact with each other (Fig. 3).

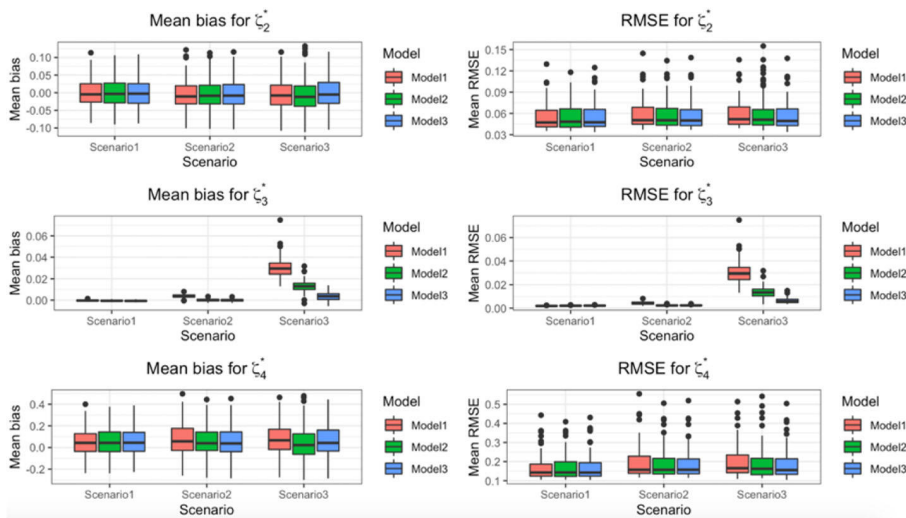


Figure 3. Mean bias (left) and RMSE (right) for parameters ζ_2^* (upper panels), ζ_3^* (central panels) and ζ_4^* (lower panels) for each model in simulation scenario 1 (assumption of linear relationship between expected abundances), scenario 2 (non-linear relation between expected abundances explained by $\omega_1(s)$) and scenario 3 (an extra spatial source of variation affecting only one of the groups of observed counts).

Table 2. Mean bias and RMSE for parameters β_0 , β_1 , ρ and σ in simulation scenario 1 (assumption of linear relationship between expected abundances), scenario 2 (non-linear relation between expected abundances explained by $\omega_1(\mathbf{s})$) and scenario 3 (an extra spatial source of variation affecting only one of the groups of observed counts)

Scenario	Model	β_0		β_1		ρ (km)		σ	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	1	-0.112	0.120	$1.62 \cdot 10^{-3}$	0.040	-1.567	4.729	0.076	0.097
		(0.037)	(0.035)	(0.028)	(0.011)	(3.982)	(1.946)	(0.080)	(0.076)
	2	-0.116	0.125	$2.74 \cdot 10^{-3}$	0.043	-1.380	4.812	0.115	0.129
		(0.036)	(0.034)	(0.028)	(0.011)	(4.330)	(2.267)	(0.096)	(0.099)
	3	-0.120	0.129	$1.48 \cdot 10^{-4}$	0.044	-1.497	4.762	0.122	0.135
		(0.038)	(0.035)	(0.030)	(0.011)	(3.978)	(2.021)	(0.088)	(0.093)
2	1	-0.112	0.119	$1.37 \cdot 10^{-3}$	0.040	-1.132	4.681	0.066	0.089
		(0.037)	(0.035)	(0.028)	(0.011)	(4.152)	(2.057)	(0.070)	(0.065)
	2	-0.111	0.120	$1.05 \cdot 10^{-4}$	0.038	-0.880	4.704	0.069	0.091
		(0.036)	(0.034)	(0.024)	(0.009)	(4.160)	(2.148)	(0.070)	(0.065)
	3	-0.104	0.113	$-1.92 \cdot 10^{-3}$	0.048	-0.952	4.629	0.058	0.082
		(0.038)	(0.035)	(0.049)	(0.025)	(4.013)	(1.981)	(0.067)	(0.060)
3	1	-0.112	0.120	$1.42 \cdot 10^{-3}$	0.040	-1.056	4.652	0.063	0.087
		(0.037)	(0.035)	(0.028)	(0.011)	(3.927)	(1.972)	(0.069)	(0.065)
	2	-0.111	0.120	$5.06 \cdot 10^{-5}$	0.038	-0.596	4.783	0.069	0.089
		(0.036)	(0.034)	(0.024)	(0.009)	(4.288)	(2.139)	(0.070)	(0.065)
	3	-0.112	0.119	$-2.05 \cdot 10^{-4}$	0.040	-1.837	5.024	0.088	0.107
		(0.037)	(0.035)	(0.028)	(0.011)	(4.240)	(2.064)	(0.079)	(0.076)

In parentheses, the standard error of each performance measurement

Figure 3 shows that the estimation of the proportional relation between the four likelihoods performed similarly for the three models when the truth is that the four likelihoods are linearly related (Scenario 1). Model 1 (which assumes linear relationship between the expected counts) performed, as expected, slightly better than the other two models as this is the model that generated the datasets. However, when we introduced some deviation from the assumption of linearity in our data generating process (Scenario 2), model 1 underperformed relative to the other two models. This is true for the three parameters of interest (Fig. 3). Models 2 and 3 performed better in terms of bias and RMSE, whereas the estimates produced by model 1 were biased and showed higher variability. Lastly, when an additional source of variation affected only one of the likelihoods (Scenario 3), the three models performed similarly as in Scenario 2, except for the hyperparameter ζ_3^* , which is part of the likelihood affected by the extra source of variation. For this hyperparameter, the differences in performance between the three models increased considerably as model 3 produced less biased and variable estimates of this hyperparameter.

Our results show that there are only marginal differences in the fixed effects β_0 and β_1 between the three models in all the scenarios. However, larger differences are observed for the hyperparameters of $\omega_1(\mathbf{s})$. For example, in the three scenarios the bias of ρ was smaller for model 2 compared to the other two models, but at the same time it produced estimates of ρ with larger RMSE than the other two models. In this simulation study, we have also explored the selection of the best model according to the comparison criteria DIC, WAIC



Figure 4. Differences in DIC, WAIC and LMPL between the model that generated the observed counts in each simulation scenario (Scenario 1, generated according to model 1; scenario 2, generated according to model 2 and scenario 3, generated according to model 3) and the other two models proposed in Sect. 3.

and LMPL (See Sect. 3.4). For each scenario, we computed the differences in each criterion between the model that generated the 100 datasets of the scenario and the other two models. The summaries of these differences are displayed in Fig. 4.

Figure 4 shows small differences in DIC and WAIC between the three models when model 1 generates the observed counts (Scenario 1). In Scenario 1, the predictive performance, measured by LMPL, was similar for model 1 (the one that generated that data) and model 2, while model 3 underperformed. In Scenario 2, model 2 (generating model) and model 3 performed similarly based on all performance comparisons, but model 1 underperformed considerably. In scenario 3, where the observed counts are generated according to a more complex specification (i.e., one sampling protocol is affected by an additional source of variation), model 3 had better goodness of fit and predictive performance with large differences in DIC, WAIC and LMPL with respect to the other two models. The difference in performance between models increases as the complexity of the data generating process increases (Fig. 4).

The results for the second group of simulations can be found in the Supplementary Information. Results match those obtained with the first group of simulations above. In Scenario 1, all three models perform similarly. As in the first simulation study, when the complexity of the model that generates the data increases, models 2 and 3 outperform model 1. Nevertheless, unlike in the first simulation study, model 2 outperformed model 3 in Scenario 3 (generated by model 3) for ζ_3 as it produced less biased estimates.

5.2. RESULTS OF THE CASE STUDY ON TOTAL ABUNDANCE OF BIRDS IN MID-SCANDINAVIA

We fitted our three models (see Sect. 3) to count data from the common bird monitoring schemes in Norway and Sweden (see Sect. 2) to estimate total abundance of birds across mid-Scandinavia with precipitation and elevation as explanatory variables. These two were selected from all the variables considered *a priori*, as it was the subset of candidate variables that produced the best results in terms of goodness of fit (see Supplementary Information for an overview of the performance of other competing models). The most demanding model in terms of computation time was model 3, which run in 60 seconds. In Table 3, we report the posterior mean, standard deviation and quartiles of the most relevant parameters from the three models.

Table 3 shows the associations between precipitation (PREC) and elevation (ELEV) with the expected counts are negative for all the models. The posterior means of the parameters of these two variables have small differences, model 2 estimated stronger association of the explanatory variables (precipitation and elevation) and the response variable (total abundance of birds). The posterior summarizes of PREC and ELEV suggest that those locations with higher levels of precipitation and high elevation are expected to have lower total bird counts. The variability and range of the Gaussian field have right skewed posterior distributions based on their posterior medians and means.

Figure 5 and Table 3 show that the posterior densities of ζ_2 are different between models, with higher posterior mean for model 1 compared to the other models. This result agrees with the exploratory analysis of Sect. 2, which suggested the necessity of specifying a relaxed linear relationship between the line and point counts in Norway (linearity was met in Sweden, but not in Norway, see Fig. 2). However, the posterior densities of ζ_3 and ζ_4 are almost identical for models 1 and 3, whereas model 2 estimated posterior distributions for ζ_3 and ζ_4 that are shifted toward lower values (Fig. 5). Large differences in the posterior mean of ψ_2 in models 2 and 3 are observed when $\omega_2(\mathbf{s})$ is introduced to account for the particularities of the sampling protocol of the line counts in Norway (i.e., in general terms, to account for added complexity due to one of the data collecting protocols considered). While model 2 gives high prevalence to $\omega_1(\mathbf{s})$ (posterior mean of $\psi_2 = 1.90$) as determinant of the departure from linear association, model 3 reduces this prevalence (posterior mean of $\psi_2 = 0.63$). It arguably means that $\omega_2(\mathbf{s})$ accounts for what is particular of this sampling protocol (the added complexity) and what at the same time reduces the leverage of what is shared between this sampling protocol (the line transect in Norway in this case study) and the other protocols. We expect these differences in contribution of $\omega_1(\mathbf{s})$ across models to impact their predictive performance. In Figure S.2, we show the posterior mean of $Y_1(\mathbf{s}) + Y_2(\mathbf{s})$, understood as a proxy for the total abundance of birds in our study region (see Sect. 3). Given the high similarity across mid-Scandinavia, hereafter, we explore the differences in the predicted mean of $Y_1(\mathbf{s}) + Y_2(\mathbf{s})$ between the three models in a smaller sub-region (highlighted with a red square in Fig. 6), which encompasses the locations surrounding Trondheimsfjorden and the Norwegian Sea.

Our three models predicted high total bird counts along the eastern coast of Trondheimsfjorden and on the islands of Hitra and Frøya (Fig. S.9) and low counts at higher elevations

Table 3. Posterior mean, standard deviation and quartiles of the most relevant parameters of the models proposed in Sect. 3

Model Model 1					
Parameter	Mean	SD	0.025q	0.50q	0.975q
Intercept	4.69	0.04	4.61	4.69	4.77
PREC	-0.12	0.04	-0.19	-0.12	-0.04
ELEV	-0.29	0.04	-0.38	-0.29	-0.21
ζ_2	0.05	0.00	0.04	0.05	0.05
ζ_3	0.51	0.03	0.45	0.51	0.57
ζ_4	1.50	0.09	1.33	1.50	1.68
ψ_2					
ψ_3					
ψ_4					
$\rho(m)$	$1.80 \cdot 10^4$	$4.00 \cdot 10^3$	$1.11 \cdot 10^4$	$1.77 \cdot 10^4$	$2.68 \cdot 10^4$
σ	0.36	0.02	0.32	0.36	0.41
Model Model 2					
Parameter	Mean	SD	0.025q	0.50q	0.975q
Intercept	4.68	0.03	4.62	4.68	4.75
PREC	-0.20	0.03	-0.26	-0.20	-0.14
ELEV	-0.39	0.04	-0.46	-0.39	-0.32
ζ_2	0.04	0.00	0.04	0.04	0.05
ζ_3	0.48	0.03	0.43	0.48	0.54
ζ_4	1.42	0.08	1.27	1.42	1.58
ψ_2	1.86	0.14	1.59	1.86	2.13
ψ_3	1.26	0.13	1.00	1.26	1.52
ψ_4	1.30	0.12	1.07	1.30	1.54
$\rho(m)$	$1.80 \cdot 10^4$	$3.88 \cdot 10^3$	$1.17 \cdot 10^4$	$1.75 \cdot 10^4$	$2.69 \cdot 10^4$
σ	0.31	0.02	0.27	0.31	0.36
Model Model 3					
Parameter	Mean	SD	0.025q	0.50q	0.975q
Intercept	4.69	0.04	4.61	4.69	4.77
PREC	-0.11	0.04	-0.18	-0.11	-0.04
ELEV	-0.27	0.04	-0.35	-0.27	-0.19
ζ_2	0.04	0.00	0.04	0.04	0.05
ζ_3	0.51	0.03	0.45	0.51	0.57
ζ_4	1.50	0.09	1.32	1.49	1.69
ψ_2	0.61	0.16	0.30	0.61	0.91
ψ_3	1.09	0.12	0.86	1.09	1.34
ψ_4	1.18	0.12	0.96	1.18	1.42
$\rho(m)$	$2.01 \cdot 10^4$	$4.12 \cdot 10^3$	$1.29 \cdot 10^4$	$1.98 \cdot 10^4$	$2.90 \cdot 10^4$
σ	0.34	0.03	0.29	0.34	0.39

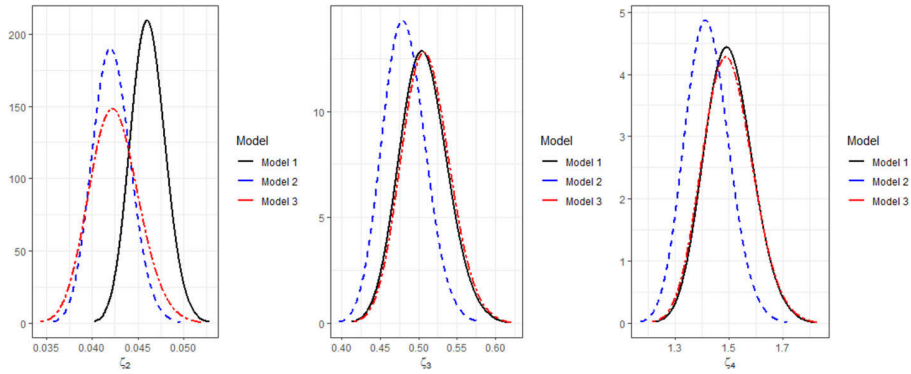


Figure 5. Posterior densities of ζ_2 (left), ζ_3 (center) and ζ_4 (right) for each model.

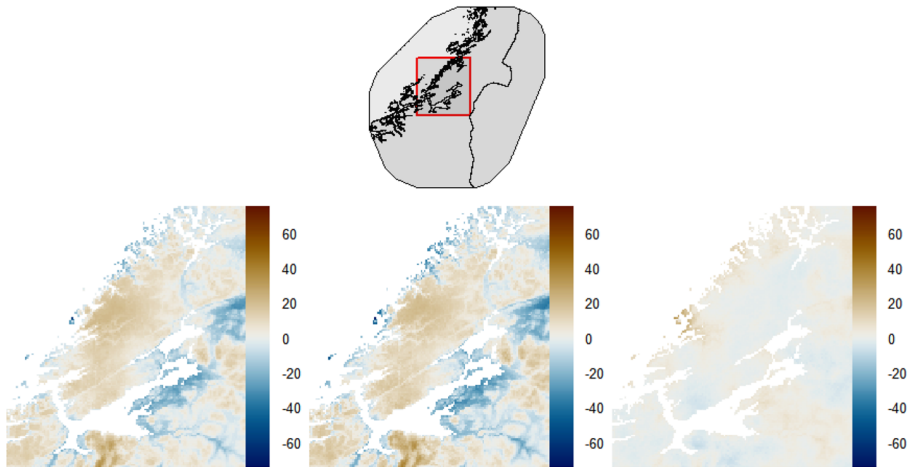


Figure 6. Top(small): Study region with the red square that encloses the zone chosen for analyzing differences between models. Bottom: differences in the posterior mean of $Y_1(s) + Y_2(s)$ (i.e. total abundance of birds) between: model 1 - model 2 (left), model 3 - model 2 (center) and model 1 - model 3 (right).

such as in the mountainous in the southwest and the north of the study region (Fig S.9.). Model 2 estimates higher counts compared to the other two models along the fjord's coast (dark blue) and lower abundance inland (mainly in the mountains; light brown; Fig. 6). The differences in predicted counts between model 1 and model 3 are smaller (Fig. 6, right panel) compared to those with model 2. However, larger predicted counts are produced by model 3 around the island of Linesøya. Our modeling framework allows for computing the uncertainty of our predictions. Here, we assess this by computing the standard error of $Y_1(s) + Y_2(s)$ (see Fig. 7 for the standard error of the sub-region highlighted in Fig. 6, and see Fig. S.10 for the standard errors across the entire study region).

The standard error of model 1 is larger than the other two models in most regions (see brown colors, left and right panels in Fig. 7). In the zones with higher predicted counts (the coast on the Norwegian Sea and Trondheimsfjorden), model 2 produced predictions with higher uncertainty (dark blue in the central panel), while on the mountains the uncer-

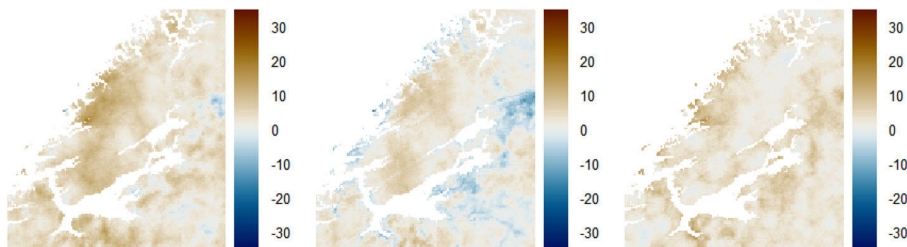


Figure 7. Differences in the posterior standard error of $Y_1(\mathbf{s}) + Y_2(\mathbf{s})$ for: model 1 and model 2 (left), model 3 and model 2 (center) and model 1 and model 3 (right).

tainty produced by model 3 was larger (light brown in the central panel). As a way to better appreciate the numerical differences between models, we explored the total predicted counts at the 113 sampling sites in Norway by comparing them against the observed counts (Fig. 8).

Figure 8 shows the comparison between the predicted and the observed values of total abundance of birds ($Y_1(\mathbf{s}) + Y_2(\mathbf{s})$). Model 1 and model 3 predict very similar values, and thus, we also compared the observed and predicted values of the counts gathered via point counts $Y_1(\mathbf{s})$ and line transects $Y_2(\mathbf{s})$ separately. Although model 1 and model 3 produce very similar predictions of total abundance of birds, model 3 predicted Y_1 and Y_2 separately more accurately. This is due to the inclusion of the GRF $\omega_2(\mathbf{s})$, as it makes it possible to better distribute the abundance between likelihoods and is flexible enough to capture more complex relationships between the census processes. We have highlighted the predicted and observed counts of the site located in the island of Linesøya (in red in Fig. 8) as this is a site where big discrepancies are observed between all the models. Model 1 and model 2 are not

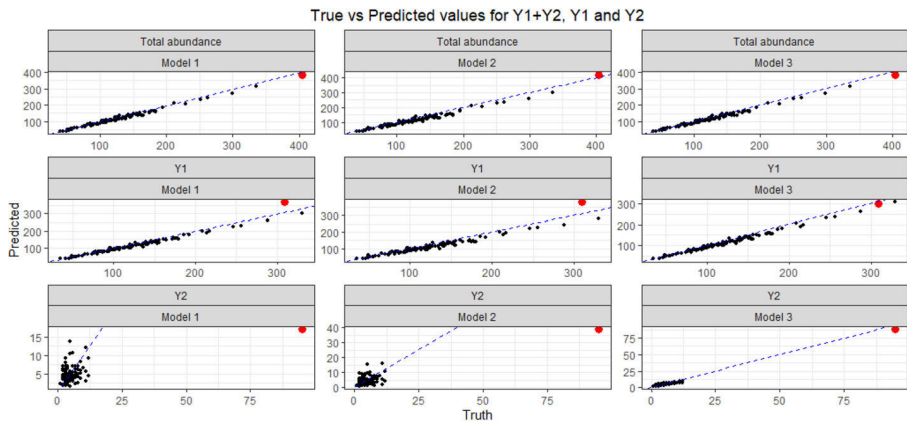


Figure 8. Comparison of observed vs predicted counts for: total abundance ($Y_1(\mathbf{s}) + Y_2(\mathbf{s})$; top row), counts produced via point counts $Y_1(\mathbf{s})$ (middle row) and counts produced via line transect counts $Y_2(\mathbf{s})$ (bottom row). The performance of model 1, model 2 and model 3 is displayed in the first, second and third columns, respectively. A particular site with high total abundance of birds due to presence of gregarious species (in this case) that is only captured by one (the line transect in Norway) of the four census protocols is highlighted in red to allow for a quick assessment of discrepancies between the three models.

able to accurately predict the counts reported in this site by the line transect survey in Norway. This site is a special location where gregarious geese belonging to several species aggregate and form large gaggles (similar examples elsewhere might be sites with (multi-species) colonies, roosting sites or wetlands hosting thousands of waterbirds). Such information is only available if data from several census protocols are combined and properly analyzed—our new modeling framework can account for these differences, as our model 3 does in comparison with model 1 that assumes a linear relation.

Figure 9 shows the posterior mean of $\omega_1(\mathbf{s})$ for the three models, as well as $\omega_2(\mathbf{s})$ for model 3. $\omega_1(\mathbf{s})$ is, in general, similar for the three models. The largest difference occurs in $\omega_1(\mathbf{s})$ for model 2, which has a shorter spatial range in comparison to the other two models. In addition, the highest contribution of $\omega_2(\mathbf{s})$ occurs in Linesøya, an island where high total abundance of birds can be recorded during the line transects, due to high concentrations of geese from several species (see above). Such species form large groups of individuals (so called, gaggles) in some of the islands along the Norwegian coast. Lastly, we compared our three models in terms of goodness of fit and predictive performance (Table 4) using the measures of performance introduced in Sect. 3.4 and out-of-sample predictive performance measures such as RMSE after brute-force Leave-One-Out Cross Validation (CV), (Vehtari et al. 2016) and Leave-One-Site-Out CV. In the former CV scheme, we removed one data point at a time, while for the other we removed both the point and line transect counts. This procedures were computationally demanding, but feasible for our problem as it took 1.76 hours for model 1, 4.2 hours for model 2 and 4.1 hours for model 3.

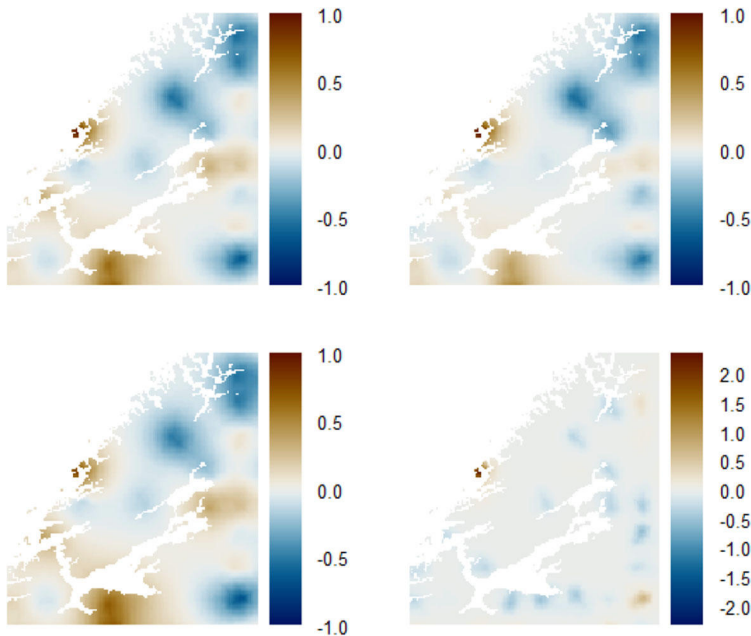


Figure 9. Posterior mean of $\omega_1(\mathbf{s})$ for model 1 (upper left), model 2 (upper right) and model 3 (bottom left). Posterior mean of $\omega_2(\mathbf{s})$ for model 3 (bottom right).

Table 4. Measures of performance (see Sect. 3.3) for models 1, 2 and 3

Measure of performance	Model 1	Model 2	Model 3
DIC	2728.79	2751.04	2603.82
WAIC	2876.19	2876.96	2593.69
RMSE	165.60	145.00	38.63
LMPL	-1673.50	-1656.35	-1425.04
Mean CRPS	27.21	25.93	20.95
RMSE (Leave-one-site-out CV)	45.70	46.39	45.88

In bold, the model with the best performance

The results show a considerable improvement in the goodness of fit when a second GRF to account for the particular characteristics in one of the observed data sources (line transects in Norway) is added. Moreover, the improvement in predictive performance of model 3 is exemplified by its low values of RMSE for the point count surveys in Norway, its high value of LMPL and its low CRPS for the point transect counts in Norway. The result of the leave-on-site-out CV shows small differences, but model 1 outperformed the other two models.

6. DISCUSSION AND CONCLUSIONS

The main goal of this paper was to introduce a modeling framework that allows us to model jointly multiple sources of information (count data) that are collected under different sampling protocols. We also presented a simple case study where we used this new methodology to estimate the total abundance of birds in mid-Scandinavia using bird counts in Norway and Sweden. These two countries have well-established bird monitoring programs, but differ in the sampling protocols. Therefore, we proposed a set of models that assumed the same coefficients for the fixed effects in each likelihood and a common GRF. The only difference between the different likelihoods is random intercepts in the linear predictor that aim at accounting for differences in the sampling protocols. For example, while the observed point counts in Norway have pairs of birds as the unit reported, Sweden reports individuals. Having different random intercepts makes possible to establish a proportional relation between the observed counts in the data sources. This is arguably a sensible choice since the biological processes that determine the abundance of species do not generally depend on national borders. Although the assumption of linear relation is reasonable for this case, it is also true that when working with real data allowing for some flexibility with respect to this assumption may correspond better to reality in most cases. This is why, we proposed a model that has a common GRF, but with a coefficient that explains how far we are from a linear relation. As seen in the exploratory analysis (Sect. 2), one of our data sources did not seem to follow the assumption of linear association with the other likelihoods. Hence, we suggested the inclusion of a second GRF to account for the differences of this likelihood. The inclusion of the second GRF, $\omega_2(\mathbf{s})$, was especially useful in our case as we do not have variables at the spatial point level that explicitly inform on the differences of the line count surveys in

Norway with respect to the other likelihoods. [Simmonds et al. \(2020\)](#) show the benefits of including an extra GRF to account for sources of bias in the sampling process of Citizen Science data. We assessed the performance of the three models when the key assumptions in the specification of each of them were not met in two simulation studies. The results of these simulations showed that a flexible specification performed similarly to the model that assumed a linear relation (model 1) when the latter model was used to generate the data. On the other hand, when the linear assumption was not met by the data generating model, the gap in performance between models became more evident. This suggests that using the models with flexible specification is always advised, regardless of the nature of the data. The estimates of the parameters in model 1 (the model assuming a linear relation between the observed counts) were biased and more uncertain than the estimates of the same parameters in the other two models. When a more complex scenario was proposed, model 3 (the model with two GRFs) clearly outperformed model 1 and model 2 in every comparison criteria. From the two simulation studies, we can conclude that model 3 is more robust than the other two models to misspecification of the functional form of the model. The parameters that showed higher differences in terms of bias and mean RMSE in the simulation study were the hyperparameters ζ_j . This might be caused by the fact that these parameters are the only ones that are not constrained to be the same for all the likelihoods, and therefore, they are more sensitive to misspecification. A biased estimate of these hyperparameters might have an impact on the predictions of our models (total abundance of birds, in our case study) as these coefficients can be used as weighting of different likelihoods when computing the total abundance. The data of the simulation studies were also used to show why integrating the four sources of information is better for predicting the total counts of birds in more than one country (See Section S.1.2. of the Supplementary Information). We compared the predictive performance of a set of models that include (i) only one of the four sources of information, (ii) two sources of information (from the same country to predict abundance in a given location within the corresponding country—e.g., points and lines from Norway to predict within Norway), and (iii) the four sources of information (points and lines from both countries) (see Table S.2.). The results show that if the goal of the study is to produce predictions in more than one country, then integrating sources of information from both countries is recommended. If the goal of the study is to only produce within-country predictions, then integrating information for more than one country would not provide any additional benefit as the models with two sources of information performs as well as the models with the four sampling protocols. When we applied this methodology to the case study of estimating total bird abundance in mid-Scandinavia, we found some very high counts on the island of Linesøya (compared to elsewhere in the region). This count was recorded during a line transect sampling, which model 1 and model 2 failed to explicitly account for. This is arguably why the differences in goodness of fit between model 1 and model 2 were negligible. The inclusion of a second GRF in model 3 to explain extra complexity (in this case, the line counts in Norway that may produce large number of birds) made sense for our research problem since it was able to explain the large counts in Linesøya, when a large number of geese congregate around these islands. Adding GRFs to the likelihoods in order to account for particularities of each observed response seemed useful and practical in other cases when researchers need to account for complexity that

can not be explained with available covariate information. However, this addition should have a clear justification and be applied with caution since giving an ecological interpretation to this random effect may not be a trivial task. Our modeling framework offers, thus, advantages to integrate data from surveys with different sampling protocols and disjoint spatial locations. In its most simple parametrization, it does not explicitly account for any factor that affects the observed total abundance (i.e., detection). For example, in our case study, we have assumed these factors are negligible. However, this modeling framework is flexible enough to explicitly account for factors that influence the observed abundance. As shown in Sect. 3, these factors can be accounted for by explaining each of the terms ζ_j in the models proposed as a function of fixed and random effects that affect the observation process. Given the complexity of the models, identifiability issues may arise if the parameters that explain the effect of the factors related to the observation process are not constrained. This issue can be overcome by integrating data that inform on these parameters, or informative prior knowledge about them. The proposed framework does not explicitly accommodate species-specific characteristics. In our case study, it was not necessary as we assumed all the species have the same weight on the estimated total abundance. However, this modeling framework can work for a broader range of goals. For example, if one or a group of species are of interest when studying anthropogenic impacts on birds (e.g., total raptor counts (De Lucas et al. 2008)), the raw data can be preprocessed according to the purpose of the study. If the goal is to model one species of concern, then getting the subset of the raw data that belong to this species would suffice to apply our methodology and obtain satisfactory results. If, in another case, the question we want to solve is linked to the risk of collision of birds with powerlines (e.g., D'Amico et al. 2019) or rotor blades in wind farms (see De Lucas et al. 2008), we can account for the differences in sensitivity between species (for example soaring raptors, which are proportionally scarce in common bird monitoring schemes, are more sensitive than other bird species). Thus, one would multiply (apply weights) the count of each species in the dataset by a 'species-specific sensitivity factor' to that particular human impact (in this case, counts of raptor species would have a larger weight than other species). Then, one would proceed by summing up the new weighted counts to obtain a 'total weighted abundance of birds' at each census site. Our methodology, thus, can provide estimates of such a total weighted abundance across the entire region of interest and maps of 'sensitivity-adjusted hotspots.' An open question would be then, how to decide the values of these weights, which might be decided based on, for example, expert opinion, traits databases (Tobias et al. 2022) and published literature (D'Amico et al. 2019). A limitation of this modeling framework is that it lies in the category of purely spatial SDMs and thus it is not possible to explicitly account for any potential temporal variation at small (e.g., within a day) or large (e.g., across years) scale. In our case study, this was not a major concern as the temporal span of our data (14 years) is not considered a period in which the distribution of the total abundance of birds has varied a lot in the study region. The ultimate goal of developing this methodology is to integrate the different sources of bird count data to predict total abundance of birds across Norway, information that will be used in further studies of human impact on biodiversity, including predicting bird mortality hotspots due to powerlines and wind farms (Bernardino et al. 2018; Bevanger 1995, 2001; Serrano et al. 2020). Therefore, achieving a good predictive performance of our models is

of paramount importance to properly assess the vulnerability of different regions to human development based on the total local abundance of birds. Although we found differences in goodness of fit between the three models, the differences in predictive performance were small. However, a flexible model specification seemed the best choice for ensuring good predictions. For example, model 3 (which included $\omega_2(\mathbf{s})$ to account for particularities of the line counts in Norway) yields the most accurate predictions at the observed locations in Norway. This is associated with the extra complexity found between line transects and point counts in Norway, which unlike the two sampling protocols in Sweden did not have a clear linear relation, as they are only complementary to one another. In conclusion, in this paper we propose models to integrate multiple professional surveys with differences in their sampling protocols. These differences are usually determined by the country of origin of the data (sampling protocol) or by the specific targets of each monitoring scheme. The INLA-SPDE approach implemented in the R-INLA package makes it straightforward to perform full Bayesian inference for models that integrate multiple sources of information, even if they are not standardized or report the observed counts in different units. A natural extension of this work is the application of the proposed modeling framework to solve a broader range of ecological questions at larger geographical scales or for species with poor data (**Buckland and Johnston**) that incorporate more sources of information given its convenience and simple implementation.

ACKNOWLEDGEMENTS

The Norwegian terrestrial bird monitoring (TOV-E) is coordinated by BirdLife Norway and Norwegian Institute for Nature Research and is financed by the Ministry of Climate and Environment and the Norwegian Environment Agency. The Swedish Bird Survey is supported by grants from the Swedish Environmental Protection Agency, with additional financial and logistic support from the Regional County Boards (Länsstyrelsen).

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital)

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

[Received March 2021. Revised March 2022. Accepted March 2022. Published Online May 2022.]

REFERENCES

- Araújo MB, Anderson RP, Márcia Barbosa A, Beale CM, Dormann CF, Early R, Garcia RA, Guisan A, Maiorano L, Naimi B, O'Hara RB, Zimmermann NE, and Rahbek C (2019). Standards for distribution models in biodiversity assessments. *Sci Adv*, 5(1)
- Banerjee S, Carlin BP, Gelfand AE (2015) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2ed. edition
- Bernardino J, Bevanger K, Barrientos R, Dwyer J, Marques A, Martins R, Shaw J, Silva J, Moreira F (2018) Bird collisions with power lines: state of the art and priority areas for research. *Biol Cons* 222:1–13
- Bevanger K (1995) Estimates and population consequences of tetraonid mortality caused by collisions with high tension power lines in Norway. *J Appl Ecol* 32(4):745–753
- Bevanger K, Bartzke G, Brøseth H, Dahl E, Gjershaug J, Hanssen F, Jacobsen K-O, Kleven O, Kvaløy P, May R, Meås R, Nygård T, Refsnæs S, Stokke S, Thomassen J (2014) Optimal design and routing of power lines: ecological, technical and economic perspectives (optipol). final report: findings 2009–2014
- Bevanger K (2001) Bird collisions with power lines - an experiment with ptarmigan (*lagopus* spp.). *Biol Cons* 99(3):341–346
- Blangiardo M, Cameletti M (2015) Spatial and spatio-temporal Bayesian models with R-INLA. Wiley, New York
- Bradter U, Ozgul A, Griesser M, Layton-Matthews K, Eggers J, Singer A, Sandercock BK, Haverkamp PJ, Snäll T (2021) Habitat suitability models based on opportunistic citizen science data: Evaluating forecasts from alternative methods versus an individual-based model. *Divers Distrib* 27(12):2397–2411
- Brodie SJ, Thorson JT, Carroll G, Hazen EL, Bograd S, Haltuch MA, Holsman KK, Kotwicki S, Samhoury JF, Willis-Norton E, Selden RL (2020) Trade-offs in covariate selection for species distribution models: a methodological comparison. *Ecography* 43(1):11–24
- Buckland S, Johnston A (2017) Monitoring the biodiversity of regions: key principles and possible pitfalls. *Biol Cons* 214:23–34
- Cameletti M, Gómez-Rubio V, Blangiardo M (2019) Bayesian modelling for spatially misaligned health and air pollution data through the INLA-SPDE approach. *Spat Stat* 31:100353
- De Lucas M, Janss GFE, Whitfield DP, Ferrer M (2008) Collision fatality of raptors in wind farms does not depend on raptor abundance. *J Appl Ecol* 45(6):1695–1703
- D'Amico M, Martins RC, Álvarez Martínez JM, Porto M, Barrientos R, Moreira F (2019) Bird collisions with power lines: prioritizing species and areas by estimating potential population-level impacts. *Divers Distrib* 25(6):975–982
- Fuglstad G-A, Simpson D, Lindgren F, Rue H (2019) Constructing priors that penalize the complexity of gaussian random fields. *J Am Stat Assoc* 114(525):445–452
- Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. *Stat Comput* 24(6):997–1016
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102(477):359–378
- Gomez-Rubio V (2020) Bayesian inference with INLA. CRC Press, Boca Raton
- Gruss A, Thorson JT (2019) Developing spatio-temporal models using multiple data types for evaluating population trends and habitat usage. *ICES J Mar Sci* 76(6):1748–1761
- Howard C, Stephens PA, Pearce-Higgins JW, Gregory RD, Willis SG (2014) Improving species distribution models: the value of data on abundance. *Methods Ecol Evol* 5(6):506–513
- Järvinen O, Väisänen RA (1983) Correction coefficients for line transect censuses of breeding birds. *Ornis Fennica* 60(4):97–104
- Johnston A, Fink D, Reynolds MD, Hochachka WM, Sullivan BL, Bruns NE, Hallstein E, Merrifield MS, Matsumoto S, Kelling S (2015) Abundance models improve spatial and temporal prioritization of conservation resources. *Ecol Appl* 25(7):1749–1756

- Kéry M, Royle JA (2009). Inference about species richness and community structure using species-specific occupancy models in the national swiss breeding bird survey MHB, pp 639–656. Springer US, Boston, MA
- Krański ET (2019) Advanced spatial modeling with stochastic partial differential equations using R and INLA. CRC Press, Boca Raton
- Kålås J (2010). The 2010 norwegian red list for species
- Kålås J, Husby M(2002). Ekstensiv overvaking av terrestrre fugl i norge
- Lehikoinen A, Brotons L, Calladine J, Escandell CT, Flousek V, Grueneberg J, Haas C, Harris F, Herrando S, Jiguet HM, Kålås F, Lindström JA, Lorrilliere A, Molina R, Pladevall B, Calvi C, Sattler G, Schmid T, Trautmann H (2019) Declining population trends of european mountain birds. *Glob Change Biol* 25(2):577–588
- Lehtomäki J, Moilanen A (2013) Methods and workflow for spatial conservation prioritization using zonation. *Environ Modell Softw* 47:128–137
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J R Stat Soc Ser B (Stat Methodol)* 73(4):423–498
- Lindström Å, Green M, Husby M, Kålås JA, Lehikoinen A, Stjernman M, et al. (2019). Population trends of waders on their boreal and arctic breeding grounds in northern europe. *Wader Study*
- Lindstrom A, Green M, Paulson G, Smith HG, Devictor V (2013) Rapid changes in bird community composition at multiple temporal and spatial scales in response to recent climate change. *Ecography* 36(3):313–322
- Lisovsky AA, Dudov SV, Obolenskaya EV (2021) Species-distribution modeling: advantages and limitations of its application. 1. General approaches. *Biol Bull Rev* 11(3):254–264
- Massimino D, Johnston A, Gillings S, Jiguet F, Pearce-Higgins JW (2017) Projected reductions in climatic suitability for vulnerable British birds. *Clim Change* 145(1):117–130
- May R, Middel H, Stokke BG, Jackson C, Verones F (2020) Global life-cycle impacts of onshore wind-power plants on bird richness. *Environ Sustain Indicators* 8:100080
- Miller DAW, Pacifici K, Sanderlin JS, Reich BJ (2019) The recent past and promising future for data integration methods to estimate species' distributions. *Methods Ecol Evol* 10(1):22–37
- Pettit LI (1990) The conditional predictive ordinate for the normal distribution. *J R Stat Soc Ser B (Methodol)* 52(1):175–184
- Roksvåg T, Steinsland I, Engeland K (2020). A geostatistical two field model that combines point observations and nested areal observations, and quantifies long-term spatial variability—a case study of annual runoff predictions in the voss area
- Rue H, Held L (2005). Gaussian Markov random fields: theory and applications. Monographs on statistics and applied probability 104. Chapman & Hall CRC, 1 edition
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Ser B (Stat Methodol)* 71(2):319–392
- Selle ML, Steinsland I, Hickey JM, Gorjanc G (2019) Flexible modelling of spatial variation in agricultural field trials with the r package Inla. *Theor Appl Genet* 132(12):3277–3293
- Selle ML, Steinsland I, Powell O, Hickey JM, Gorjanc G (2020) Spatial modelling improves genetic evaluation in smallholder breeding programs. *Genet Sel Evol* 52(1):1–17
- Serrano D, Margalida A, Pérez-García JM, Juste J, Traba J, Valera F, Carrete M, Aihartza J, Real J, Mañosa S, Flaquer C, Garin I, Morales MB, Alcalde JT, Arroyo B, Sánchez-Zapata JA, Blanco G, Negro JJ, Tella JL, Ibañez C, Tellería JL, Hiraldo F, Donazar JA (2020) Renewables in Spain threaten biodiversity. *Science* 370(6522):1282–1283
- Simmonds EG, Jarvis SG, Henrys PA, Isaac NJB, O'Hara RB (2020) Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography* 43(10):1413–1422
- Simpson D, Illian JB, Lindgren F, Sørbye SH, Rue H (2016) Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika* 103(1):49–70
- Soultan A, Pavón-Jordán D, Bradter U, Sandercock BK, Hochachka WM, Johnston A, Brommer J, Gaget E, Keller V, Knaus P, Aghababayan K, Maxhuni Q, Vintchevski A, Nagy K, Raudonikis L, Balmer D, Noble D, Leitão D, Øien IJ, Shimmings P, Sultanov E, Caffrey B, Boyla K, Radišić D, Lindström Å, Veleviski M, Pladevall

- C, Brotons L, Karel Š, Rajković DZ, Chodkiewicz T, Wilk T, Szép T, van Turnhout C, Foppen R, Burfield I, Vikstrøm T, Mazal VD, Eaton M, Vorisek P, Lehtikoinen A, Herrando S, Kuzmenko T, Bauer H-G, Kalyakin MV, Voltzit OV, Sjeničić J, Pärt T (2022) The future distribution of wetland birds breeding in Europe validated against observed changes in distribution. *Environ Res Lett* 17(2):024025
- Soykan CU, Sauer J, Schuetz JG, LeBaron GS, Dale K, Langham GM (2016) Population trends for north American winter birds based on hierarchical models. *Ecosphere* 7(5):e01351
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. *J R Stat Soc Ser B (Stat Methodol)* 64(4):583–639
- Tobias JA, Sheard C, Pigot AL, Devenish AJM, Yang J, Sayol F, Neate-Clegg MHC, Alioravainen N, Weeks TL, Barber RA, Walkden PA, MacGregor HEA, Jones SEI, Vincent C, Phillips AG, Marples NM, Montaña-Centellas FA, Leandro-Silva V, Claramunt S, Darski B, Freeman BG, Bregman TP, Cooney CR, Hughes EC, Capp EJR, Varley ZK, Friedman NR, Korntheuer H, Corrales-Vargas A, Trisos CH, Weeks BC, Hanz DM, Töpfer T, Bravo GA, Remeš V, Nowak L, Carneiro LS, Moncada R (2022) Avonet: morphological, ecological and geographical data for all birds. *Ecol Lett* 25(3):581–597
- Vehtari A, Mononen T, Tolvanen V, Sivula T, Winther O (2016) Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *J Mach Learn Res* 17:103:1-103:38
- Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 11:3571–3594
- Yu H, Cooper AR, Infante DM (2020) Improving species distribution model predictive accuracy using species abundance: application with boosted regression trees. *Ecol Modell* 432:109202

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper III

**New spatial models for integrating standardized
detection-nondetection and opportunistic presence-only data:
application to estimating bird mortality hotspots linked to
powerlines.**

Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B. (2022)

New spatial models for integrating standardized detection-nondetection and opportunistic presence-only data: application to estimating risk factors associated to powerline-induced death of birds

Jorge Sicacha-Parada ^{*1}, Diego Pavon-Jordan ^{†2}, Ingelin Steinsland ^{‡1}, Roel May ^{§2}, and Bård Stokke ^{¶2}

¹Department of Mathematical Sciences. Norwegian University of Science and Technology (NTNU). Trondheim. Norway

²Department of Terrestrial Ecology. Norwegian Institute for Nature Research (NINA). P.O. Box 5685 Torgarden. N-7485. Trondheim. Norway

Abstract

The constant increase in energy consumption has created the necessity of extending the energy transmission and distribution network. Placement of powerlines represent a risk for bird population. Hence, better understanding of deaths induced by powerlines, and the factors behind them are of paramount importance to reduce the impact of powerlines. To address this concern, professional surveys and citizen science data are available. While the former data type is observed in small portions of the space by experts through expensive standardized sampling protocols, the latter is opportunistically collected over large extensions by citizen scientists.

In this paper we set up full Bayesian spatial models that 1) fusion both professional surveys and citizen science data and 2) explicitly account for preferential sampling that affects professional surveys data and for factors that affect the quality of citizen science data. The proposed models are part of the family of latent Gaussian models as both data types are interpreted as thinned spatial point patterns and modeled as log-Gaussian Cox processes. The specification of these models assume the existence of a common latent spatial process underlying the observation of both data types.

The proposed models are used both on simulated data and on real-data of powerline-induced death of birds in the Trøndelag county in Norway. The simulation studies clearly show increased accuracy in parameter estimates when both data types are fusioned and factors that bias their collection processes are properly accounted for. The study of powerline-induced deaths shows a clear association between the density of the powerline network and the risk that powerlines represent for bird populations. The choice of model is relevant for the conclusions that could be drawn from this case study as different models estimated the association between risk of powerline-induced deaths and the amount of exposed birds differently.

*jorge.sicacha@ntnu.no

†diego.pavon-jordan@nina.no

‡ingelin.steinsland@ntnu.no

§roel.may@nina.no

¶bard.stokke@nina.no

1 Introduction

Energy consumption is anticipated to rise by ca. 50% by 2050 (Conti et al., 2016) and hence the global network for energy transmission must also be extended, particularly to meet the UN’s Sustainable Development Goal (SDG) 7 “*universal access to affordable, reliable, and modern energy services*” (United Nations, 2018). In addition to the high expenditure that this implies (Bernardino et al., 2018), such a development in infrastructure and land sparing will have an enormous environmental cost (Biasotto and Kindel, 2018). Accumulating evidence shows that powerlines are an important threat for many avian species (Martin, 2011), with overhead wires fragmenting the airspace used by birds, increasing mortality risk by collision (Bernardino et al., 2018) and with masts used by many species as perching structures, causing increased death rates by electrocution (Hernández-Lambrano et al., 2018). Davis (2002) estimated that the deaths caused by powerlines easily reach a billion birds per year. Since then, the network of powerlines has increased by 5% annually (Jenkins et al., 2010), and the number of deaths has most likely increased despite the success of effective local risk mitigation actions (Pavón-Jordán et al., 2020; Barrientos et al., 2012).

From the conservation point of view and in line to the SDG 9 “*build resilient and sustainable infrastructure*”, SDG 11 “*make cities and human settlements sustainable*”, SDG 12, “*ensure sustainable consumption and production patterns*”, SDG 15 “*protect, restore and promote sustainable use of terrestrial ecosystems and halt biodiversity loss*” (United Nations, 2018), it is of paramount importance to gain a better understanding of under which circumstances (i.e. how, where and when) powerlines suppose a high death hazard. This can target mitigation actions and plan new power lines such that their ecological impact is reduced.

Until now studies of mortality impact of powerlines have been based on standardised detection-nondetection data (sensu Miller et al. (2019)), are often carried out at specific locations and relatively small spatial scale (e.g. Bevinger (1995); Bevinger and Brøseth (2001)) and/or focus on target species of concern (e.g. JANSS and FERRER (2001); López-López et al. (2011)). This hinders our ability to draw broader conclusions about the factors involved in this recognised human-wildlife conflict. One way to increase the range of species and habitats as well as the geographical extent represented in such datasets is to use the vast information contained in Citizen Science (CS) portals (e.g. <https://ebird.org>, www.artsobservasjoner.no). Some of these CS platforms allow citizen scientists to report additional information on their observations, including whether the finding was a dead animal and even the potential cause of death (e.g. electrocution, collision with powerline wires, collision with fences, roadkill). These two sources of information (standardised detection-nondetection and opportunistic presence-only data), however, are not directly comparable as they come with different inherent biases, especially regarding survey effort (Botella et al., 2021).

Aware of the potential benefits the integration of multiple data types has to offer, much research has in the past years attempted to overcome the challenges of integrating more than one data source (Koshkina et al., 2017; Pacifici et al., 2017; Miller et al., 2019; Gelfand and Shirota, 2019; Zipkin et al., 2021; Wang et al., 2021). Evidence of the benefits in inferential and predictive performance when multiple data types are integrated is accumulating (Simmonds et al., 2020; Gelfand and Shirota, 2019). These benefits include, inter alia, reduction in the uncertainty of the predicted variable in comparison to the results when each data type is modelled separately and increased accuracy in parameter estimation. Nevertheless, accounting for potential biases in the collection process is of paramount importance for fusion models to perform as expected (Simmonds et al., 2020).

This paper is motivated for a case study whose aim is to determine which factors are associated to high risk of powerline-induced death of birds, and hence highlight riskier areas within the powerline network of Trøndelag, Norway. Two data types are available to address this question, data collected through professional surveys performed by the Norwegian Institute for Nature Research (NINA) and opportunistic records collected from two sources: i) Artsobservasjoner, a database of the Norwegian Biodiversity Information Center (NIBC), and ii) the Norwegian Bird Ringing Centre. Extensive research in this field show that factors such as visibility, land use, the density of the powerlines and the amount of bird that are exposed to collide with the powerlines are associated to the risk of powerline-induced deaths (Bevinger et al., 2014; Drewitt and Langston, 2008; Martin and Shaw, 2010).

Detection-nondetection data collected in professional surveys for ecological research are often analysed using site-occupancy models (MacKenzie and Kendall, 2002) and geostatistical spatial models (Banerjee et al., 2015) as the data locations are fixed and defined during the sampling design. For our case study, the data collected by NINA contains exact locations of where powerline-induced deaths have occurred. As a census is performed in the area around a powerline once it has been selected to be sampled, we regard these data as a thinned point pattern (Illian et al., 2008) with thinning probability depending on the area to which the points belong. If the removal of points from the original point pattern is driven by Missing Not At Random (MNAR) mechanisms (Little and Rubin, 2019) that depend on our ecological process, we have thinning caused by a preferential sampling design (Diggle et al., 2010). Otherwise, the removal of points is assumed to occur randomly.

CS data are also regarded as a thinned point pattern, but due to a MNAR mechanism (Little and Rubin, 2019) that depends on factors such as differences in the sampling effort, detectability, reporting effort and/or misclassification. A general flexible framework for generating CS data has been proposed by **Peprah, Sicacha-Parada, 2022**. This framework links thinning operations for point patterns (Illian et al., 2008) with the biases in CS data and provides a novel perspective for modeling CS data. This framework relies on the idea of a shared process model for modeling data generated through MNAR mechanisms (Little and Rubin, 2019). Hence, this model assumes a common latent effect that drives multiple observed data. In our case, we assume this common latent effect affects both the observed data and the missingness process. For our case study CS data is obtained from two sources, both of them share biases, such as uneven sampling effort, which can be affected by factors such as accessibility and/or land use (Monsarrat et al., 2019; Sicacha-Parada et al., 2021), differences in detectability, which can be explained by land use, habitat type, the size of the dead bird and/or moment of the observation (Domínguez del Valle et al., 2020), and uneven reporting effort (August et al., 2020).

Hence, the aim of this paper is to introduce a modeling framework that integrates professional surveys and CS data while accounting for the biases in the collection of each of the data types as suggested in Simmonds et al. (2020). As a natural result of the modeling framework proposed, we also expect to highlight riskier areas for powerline-induced deaths. This framework extends the state of the art of data integration models as it simultaneously models CS data and their biases and professional surveys data and their sampling process, which might be preferential (Diggle et al., 2010). This modeling framework is specified as a group of Bayesian models that depend on shared spatial random effects and lies within the class Latent Gaussian Models (LGMs, Rue et al. (2009)). As these models belong to the family of LGMs, they are suitable for being fitted using both the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) and the Stochastic Partial Differential Equation (SPDE) (Lindgren et al., 2011) which offer efficient approximation of both the posterior distribution of the parameters of the models and the Gaussian Random Fields (GRF) involved in the specification of the models (Simpson et al., 2016). This is a flexible framework for integrating the two available data types and accounting for multiple sources of bias in both professional surveys and CS data. Hence, the models we present can be applied directly in assessments not only of the impacts of powerlines on animal mortality (our case study), but also in those of other human infrastructures such as roads and windfarms. Such assessments are critical in times where the rapid increase of new projects linked to renewable energy development are adding mortality to that of the traditional roadkills (Barrientos et al., 2021) and powerlines (Bernardino et al., 2018) and are facing strong public rejection (Serrano et al., 2020).

We show the performance of the models we propose for both simulated data and our case study. Through the simulation study we show the relevance, potential benefits for parameter estimation and challenges of integrating both data types. In the case study, despite not knowing the ground truth, we expect to find riskier areas for powerline-induced deaths, which factors determine this risk, as well as comparing each of the models proposed.

This paper is organized as follows. In Section 2 we present the two available data types for our case study as well as details of their collection process that are relevant for the specification of our modeling framework. This framework is introduced in Section 3, technical details regarding the models for integrating the two data

types and accounting for their biases are also presented in this section. In Section 4, the simulation studies to assess the properties of our models as well as the necessity of accounting for biases and of integrating both data types are presented. Section 5 contains the analysis of powerline-induced death of birds and the results. Finally, in Section 6 we discuss the proposed framework and propose future extensions of it.

2 Data and case study: Death of birds caused by powerlines in Trøndelag, Norway

Here we use two different datasets on bird casualties due to powerlines. First, we use standardised data from professional surveys conducted by the Norwegian Institute for Nature Research (NINA) aiming at finding all bird carcasses under a specific section of a powerline (Bevanger et al., 2014). Second, we use two opportunistic presence-only CS data: 1) records found at the Norwegian CS portal www.artsobservasjoner.no and reported as “dead bird by a powerline” since 2016 and 2) Records of dead birds reported to the Norwegian Bird Ringing Centre. Figure 1 displays the occurrence of both types of data. Since the detection of birds can occur up to about 100m away from a powerline, and for computational convenience, the spatial domain of the case study is a buffered version (100m on each side) of the networks of powerlines. In this study we use data from Trøndelag, in central Norway. The large variation in environmental conditions of this county and the existence of areas with high abundance of birds (Sicacha-Parada et al., 2022) makes this region suitable for our case study.

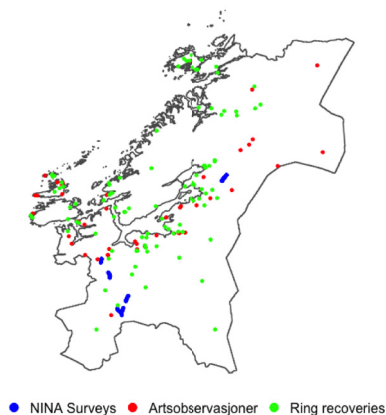


Figure 1: *Data sources considered for studying the risk of powerline-induced bird mortality and their spatial distribution. In blue: Spatial distribution of observations collected by NINA. In red: Opportunistic records reported by citizen scientists in Artsobservasjoner. In green: Opportunistic records reported through the Norwegian Bird Ringing Centre.*

2.1 Professional surveys of bird mortality

The data of professional surveys on bird mortality caused by powerlines were collected within the OPTIPOL project (Bevanger et al., 2014). Carcasses were searched for using a trained dog (wachtel) under a 7.1 km-long section of a high-voltage powerline (300 kV transmission line) during 2011-2014. Every year, the same dog patrolled the same section following the same protocol - crisscrossing under the powerline in the clearcut area (see Bevanger et al. (2014) for further details and study design). Found carcasses were removed to avoid double-counting. For this study we used 147 observations of carcasses collected at point level as the exact geographic location is available for each observation.

2.2 Opportunistic records of bird mortality

We retrieved opportunistic (presence-only) records from two sources. First, Artsobservasjoner, a database of the Norwegian Biodiversity Information Center (NBIC; www.artsobservasjoner.no) where everyone, volunteers and professionals, can report the occurrence of any species alongside the location, date/time and additional information that the observer deems important to be linked to the observation. For example, observers can report if the observed occurrence was dead and the cause of death if it is straightforward (e.g. electrocuted bird, broken wing due to collision). The second source of opportunistic records is the Norwegian Bird Ringing Centre, run by the Natural History Museum in Stavanger (<https://www.museumstavanger.no/en/forskning/dennorske-ringmerkingssentralen-1>), whose database registers more than 8 million entries, including ringing data (i.e. tagging birds with metal rings with a unique identifier) and ring-recovery data. The recovery data (i.e. when a bird with a metal ring is found dead and reported to the national ringing office) allows us to, inter alia, gather information on location and causes of death (provided that the observer reports the cause of death, which may not always be the case). As precise geographic information is available for each report of this data type are available, these data are handled as point-level data.

In total, for our case-study in Trøndelag County, we used 98 observations from ring recoveries (dead birds found with a metal ring and were reported to the ringing center) and 46 observations of dead birds killed by a powerline from the CS portal Arts Observasjoner ($n = 144$).

2.3 Explanatory variables

Two groups of explanatory variables are required to fit the models proposed in the upcoming sections. The first group of candidate variables aim to explain the ecological process underlying the powerline-induced deaths and the second group explain the sampling process of both the professional surveys and the opportunistic records.

In the group of candidate variables for explaining the ecological process, we include elevation (Digital Elevation Model; DEM), mean temperature and precipitation (Norwegian Meteorological Institute), bird abundance estimated from the Norwegian common bird monitoring scheme (<https://tov-e.nina.no>; see also Sicacha-Parada et al. (2022)), powerline density (The Norwegian Water Resources and Energy Directorate; NVE) cloud cover (<https://www.earthenv.org/cloud>) and land cover (AR50, <https://www.nibio.no/tema/jord/arealressurser/>). All covariate information was rasterized to a scale of 1 x 1 kilometers. The second group of variables explain the sampling process of the professional surveys and the CS data. The candidate covariates to explain the sampling process include distance to the closest (tertiary) road (road network obtained from OpenStreetMap <https://www.openstreetmap.org>), as well as the distance to the nearest water body (where citizen scientists frequently go for birdwatching due to high bird abundance). To explain the sampling process that yield the observed professional surveys data, we use elevation gradient.

2.4 Exploratory analysis

Despite the standard sampling effort made in NINA's professional surveys, these data have some limitations. First, their spatial coverage is small as carrying out these surveys is time-consuming and expensive. Second, the selection of the survey sites (powerlines) is not completely random (Bevanger et al., 2014) as expert knowledge is used to determine which powerlines should be visited. Now we explore the sampled locations during the CS projects and determine whether or not there is indication of preferential sampling in these data. To do it, we make two datasets, one with a 100mx100m grid of points along a 100-meter buffer of the network of powerlines in Trøndelag and the other one with a 100mx100m grid of points defined over the powerlines that have been visited by NINA. Our comparison focuses on two covariates, powerline density and cloud cover. The results are presented in Figure 2. Note that both covariates are standardized based on their values over Trøndelag.

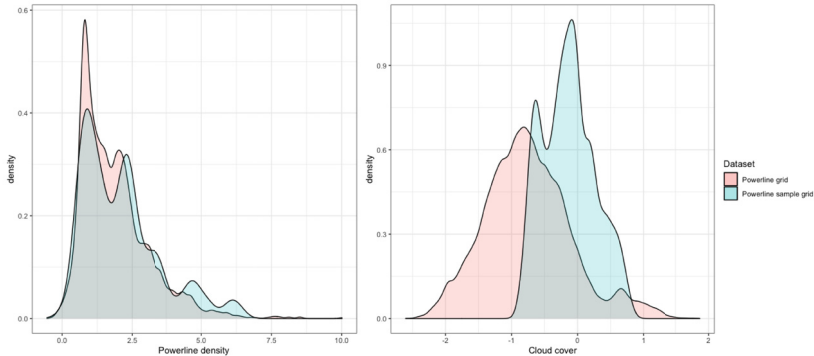


Figure 2: Density plot for powerline density and cloud cover for the two datasets previously defined. In red, the density plot for the grid defined over the whole network of powerlines. In blue, the density plot defined for the grid of points defined for the sampled powerlines.

Even though both densities seem similar for the powerline density, we notice that denser powerlines are more frequent among the sampling locations chosen by the experts. On the other hand, the right panel of Figure 2 shows that those powerlines selected by the experts are located where cloudcover is higher than usual along the network of powerlines. It is known that both powerline density and visibility (proxied by cloud cover) are factors associated with powerline-induced mortality (Drewitt and Langston, 2008). Hence, there is apparently indication of preferential sampling in the sampling design performed by experts at NINA.

Both Artsobservasjoner data and ring recoveries can be regarded as CS data. Both of them share biases, such as uneven sampling effort, differences in detectability and uneven reporting effort. The latter might be regarded as an important source of bias as it is more common for rare occurrences, as in the case of dead birds, which might not be as convenient to report as the occurrence of alive individuals. Now, we explore the available CS data and determine if there is indication of a sampling design affected by factors such as accessibility or land use. We define again two datasets in this case. The first one, a 100mx100m grid of points defined over Trøndelag and the second one, the locations where powerline-induced bird death have occurred. The results are displayed in Figure 3.

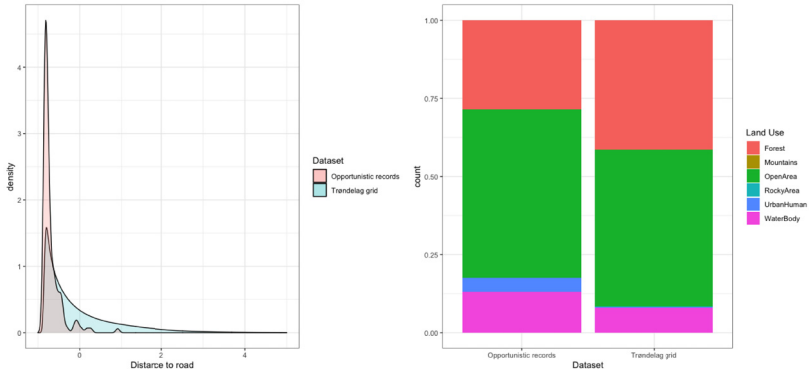


Figure 3: *Density plot for distance to (tertiary) roads and partition of land uses for the two datasets previously defined. On the left panel: in blue, the density plot for the grid defined over the whole county of Trøndelag; in red, the density plot defined for the locations with opportunistic records. On the right panel: left bar: partition of land use in the locations with opportunistic records; right bar: partition of land use over the whole county of Trøndelag*

On the left panel of Figure 3 we notice that the locations where citizen scientists have reported occurrences of dead birds are considerably closer to tertiary roads than all the other locations in the study area. Regarding the land use, we see from the right panel of Figure 3 that the percentage of reports of dead birds that occur in open areas is higher than the percentage of open areas in Trøndelag. On the other hand, the percentage of reports made in forest areas is much smaller than the actual portion of Trøndelag that is covered by forest. As it is known that accessibility has no association with the risk of powerline-induced mortality, there seems to be an indication of bias due to accessibility in the sampling process of CS data, as previously mentioned in Fithian et al. (2015); Monsarrat et al. (2019); Sicacha-Parada et al. (2021). Unlike accessibility, land use might be associated with differences in the risk of death due to powerlines. However, it is a factor that is also associated with higher or lower detectability. Therefore, in the upcoming sections we will examine whether or not land use should be considered as a factor linked to differences in risk of powerline-induced death and differences in detectability.

3 Models

The generating process of occurrence of dead birds related to the network of powerlines in Trøndelag is the target model of this work. We model occurrence of dead birds along the network of powerlines $P \subset \mathbb{R}^2$ as a point pattern with intensity $\lambda_{true}(\mathbf{s})$. This intensity is further modeled as a sum of a linear combination of known environmental variables $X(\mathbf{s})$ as, for example, powerline density and abundance of birds, and a spatial Gaussian Random Field, $\omega_1(\mathbf{s})$:

$$\log(\lambda_{true}(\mathbf{s})) = X(\mathbf{s})\beta + \omega_1(\mathbf{s}) \quad (1)$$

This true intensity of birds dead due to powerlines and its components are the quantities of interest. In our case study we have two data sources available, the citizen science (CS) data and the professional data as described in Section 2. Each data type has two different observation processes which we propose models for below.

3.1 Modelling data process of professional surveys

There are two aspects we want to include in our model of the professional surveys: 1) sampling locations might be chosen preferentially, and 2) the sampling effort is very high and we assume that the process is perfectly sampled. Therefore, we interpret the observed data by experts as a realization of the preferential

sampling process described by Diggle et al. (2010). Each of the m buffered powerlines (see Section 2) has a probability $\phi_i(\mathbf{s}); i = \{1, \dots, m\}$ of being sampled. If the selection of the lines that are visited is completely random we assume $\phi_1(\mathbf{s}) = \dots = \phi_m(\mathbf{s})$. However, as in our case study, experts usually have prior knowledge on where they can find what they want to investigate, we assume the probabilities in $\Phi = [\phi_1(\mathbf{s}), \dots, \phi_m(\mathbf{s})]^T$ are stochastically dependent on the latent process $\omega_1(\mathbf{s})$ (see Eq. 1) and eventually on some of the covariates in $X(\mathbf{s})$. In addition to $\omega_1(\mathbf{s})$, the probabilities in Φ are explained by known variables $Z(\mathbf{s})$, which might include some of the covariates in $X(\mathbf{s})$, as follows:

$$\text{logit}(\phi_i(P_i)) = Z(P_i)\gamma + \zeta\omega_1(P_i) \quad (2)$$

The coefficient ζ is relevant for this specification as it determines the extent of the dependence between the sampling selection of power lines to be sampled and the true ecological process that defines hotspots for bird collision or electrocution (Gelfand and Shirota, 2019).

Once the powerlines to be sampled have been selected, we assume, as stated at the beginning of this subsection, that within each selected area, a complete census with perfect detection of the dead birds is performed. Hence, we observe a point pattern \mathbf{P}_{PA} with intensity $\lambda_{PA}(\mathbf{s}) = \lambda_{true}(\mathbf{s})\phi_k(\mathbf{s})$, with k the powerline where location \mathbf{s} belongs. Figure S.1. explains graphically the observation process performed in professional surveys.

It is worth noting that as we work with a buffered version of the network P , the model in Eq. (2) is an areal data model (Banerjee et al., 2015) and accounting for preferential sampling while avoiding identifiability issues implies modelling jointly $\lambda_{PA}(\mathbf{s})$ and the selection probabilities $\phi_i(\mathbf{s})$. More details about this part of this modelling framework are provided in Section 3.1.1.

3.1.1 Modelling preferential sampling

As previously stated, modelling preferential sampling for our case study implies the integration of a point process and areal data. Both data types depend on the Gaussian Random Field $\omega_1(\mathbf{s})$. However, while a continuous version of the GRF generates the point process, a discretized version of it produces the areal data. Integration of two or more types of spatial data has been previously approached by Roksvåg et al. (2020) and Wang et al. (2021) with applications to hydrology and epidemiology, respectively. In general, a GRF at the area \mathbf{A} is expressed and approximated as:

$$\omega_1(\mathbf{A}) = \frac{1}{|\mathbf{A}|} \int_{u \in \mathbf{A}} \omega_1(u) du \approx \frac{1}{H} \sum_{\mathbf{s} \in \mathbf{A}} \omega_1(\mathbf{s}) \quad (3)$$

with $\mathbf{s} = \{s_1, \dots, s_H\}$ a set of sampling points in \mathbf{A} . However, as $\omega_1(P_i)$ is a nonlinear function of $\phi_i(P_i)$ in Eq. (2), we express and approximate $\omega_1(\mathbf{A})$ for our case study as:

$$\omega_1(\mathbf{A}) = \text{logit} \left(\frac{1}{|\mathbf{A}|} \int_{u \in \mathbf{A}} \frac{\exp(\omega_1(u))}{1 + \exp(\omega_1(u))} du \right) \approx \text{logit} \left(\frac{1}{H} \sum_{\mathbf{s} \in \mathbf{A}} \frac{\exp(\omega_1(\mathbf{s}))}{1 + \exp(\omega_1(\mathbf{s}))} \right) \quad (4)$$

as presented in Wang et al. (2021) in order to avoid ecological bias (Greenland, 1992).

3.2 CS data process model

We assume CS data is the result of a thinning process of the true point pattern of dead birds with three components: sampling effort, detectability and reporting effort. They can be sources of bias due to uneven sampling effort understood as differences in the probability of location \mathbf{s} being sampled by citizen scientists, due to differences in detectability according to the location of the occurrence of the event and due to differences in reporting probability amongst the observers. Based on this, we assume CS data follow a point process model, to be specific a log Gaussian Cox process (LGCP), \mathbf{P}_{CS} with intensity $\lambda_{CS}(\mathbf{s})$ given by:

$$\lambda_{CS}(\mathbf{s}) = \lambda_{true}(\mathbf{s}) \cdot \tau(\mathbf{s}) \cdot \psi(\mathbf{s}) \cdot \delta(\mathbf{s}) \quad (5)$$

where $\lambda_{true}(\mathbf{s})$ is the intensity of the true occurrences of dead birds due to powerlines, $\tau(\mathbf{s})$ is the probability of location \mathbf{s} being part of the locations sampled by citizen scientists, $\psi(\mathbf{s})$ is the probability of detecting an

occurrence at location \mathbf{s} given that it has been visited and $\delta(\mathbf{s})$ is the probability of reporting an occurrence at \mathbf{s} given that the location has been visited by a citizen scientist and an occurrence has been detected.

We focus on accounting for three sources of bias: uneven sampling effort, detectability and reporting effort. However, this modeling framework is flexible enough as to account for any other source of bias that might need to be accounted for.

3.2.1 Sampling effort

The sampling process of CS data is not standardized (Isaac et al., 2014) and is often biased towards locations with higher accessibility, or locations where observers expect to find more occurrences, i.e. locations that are preferentially sampled, (Fithian et al., 2015; Monsarrat et al., 2019). Hence, we propose modeling the term $\tau(\mathbf{s})$ as:

$$\text{logit}(\tau(\mathbf{s})) = Z(\mathbf{s})\alpha + \omega_2(\mathbf{s}) \quad (6)$$

where $Z(\mathbf{s})$ are covariates that aim to explain which locations are most likely visited by citizen scientists and $\omega_2(\mathbf{s})$ is a spatial random effect.

3.2.2 Differences in detectability

In many real-life scenarios the detection of an event of interest, (a dead bird in our case) can not be assumed constant. For our case-study, a relevant factor that affects whether or not a dead bird can be detected is the land use. As pointed out in previous studies (Domínguez del Valle et al., 2020), habitats such as open areas allow for easier detection of such events. Hence, the differences in detectability are regarded as a second thinning factor for the actual point pattern of dead birds. We define the probability of detecting the occurrence of a dead bird at location \mathbf{s} given that it has been visited as:

$$\text{logit}(\psi(\mathbf{s})) = W(\mathbf{s})\nu \quad (7)$$

with $W(\mathbf{s})$ the covariates that explain the factors that affect detectability, in our case land use or habitat type.

3.2.3 Reporting effort

Estimating the reporting effort represented in $\delta(\mathbf{s})$ requires information that is almost always unavailable. However, we here propose two ways of accounting for the noise that the uneven reporting effort generates on the observed point pattern of reports of dead birds.

Simple report effort model

This proposal assumes no structure in the random effect that drives the differences in reporting error. Hence, $\delta(\mathbf{s})$ is assumed to depend on a hyperparameter θ . The relation between $\delta(\mathbf{s})$ and θ is given by:

$$\delta(\mathbf{s}) = \frac{\exp(\theta)}{1 + \exp(\theta)} \quad (8)$$

where θ has a normal prior $\pi(\theta) = N(0, 1)$.

Observer-specific report effort model

This is a more complex approach as we assume the probability of reporting associated to each occurrence depends on its location and the observers whose citizen science activity occurs around it (i.e. how active an observer is around the location of the occurrence based on her/his other reported observations). In this case we express $\delta(\mathbf{s})$, the probability of reporting an occurrence at location s given that it was visited by a citizen scientist and an occurrence was detected as:

$$\delta(\mathbf{s}) = \sum_{j \in \{obs\}} \psi_j(\mathbf{s})\kappa_j(\mathbf{s}) \quad (9)$$

where $\psi_j(\mathbf{s})$ is the probability that the observer of an occurrence that was sampled is observer j and $\kappa_j(\mathbf{s})$ is the probability that this occurrence that observer j has detected becomes finally reported. The expression in (9) can be interpreted as a weighted average of the probabilities that each observer reported the occurrence of a dead bird once they saw one. The weights $\psi_j(\mathbf{s})$ depend on characteristics of the observers such as the distance to \mathbf{s} , or how active the observer (i.e. number of observations in the CS portal) is. In this modeling framework, $\psi_j(\mathbf{s})$ is taken as a deterministic input while the aim of the model is to estimate $\kappa_j(\mathbf{s})$. A broader explanation of how ψ_j could be estimated is available in the Supplementary Information.

3.3 Prior specification

The spatial GRFs $\omega_1(\mathbf{s})$ and $\omega_2(\mathbf{s})$ in Eq (1) and Eq (6) are assumed to follow a Matérn covariance function given by:

$$\frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\kappa\|s_i - s_j\|)^\nu K_\nu(\kappa\|s_i - s_j\|) \quad (10)$$

with $\|s_i - s_j\|$ the Euclidean distance between two locations $s_i, s_j \in D$. σ^2 stands for the marginal variance, and K_ν represents the modified Bessel function of the second kind and order $\nu > 0$. ν is the parameter that determines the degree of smoothness of the process, while $\kappa > 0$ is a scaling parameter. The parameter ν is fixed to be 1. The spatial range ρ is expressed as $\rho = \sqrt{8}/\kappa$. The prior distribution of the parameters ρ and σ are specified by making use of Penalized Complexity (PC) priors, (Fuglstad et al., 2019). The parameter vectors β, γ, α and ν have Normal prior distribution with mean 0 and precision 0.01.

3.4 Fitting the models: inlabru R-package

As our modeling framework lies within the framework of the Latent Gaussian Models (LGMs; Rue et al. (2009)), our models can be conveniently fitted using the INLA-SPDE approach (Rue et al., 2009; Lindgren et al., 2011). INLA produces fast, reliable inference as it aims to produce a numerical approximation of the marginal posterior distribution of the parameters and hyperparameters of the model. The SPDE approach is based on the solution of a SPDE which can be approximated through a basis function representation defined on a discretization of the spatial domain. Simpson et al. (2016) propose using the SPDE approach based on a tesselation of the space to efficiently approximate the likelihood of a LGCP.

This modeling framework incorporates nonlinear terms in the linear predictor in order to account for the biases in CS data. An alternative to deal with these non-linearities is to iteratively linearize them. This is done using the inlabru R-package (Bachl et al., 2019), which makes an approximation by linearizing the non-linear function at a so-called linearization point and then approximating the posterior distribution through a Taylor series approximation of second order at this point. The fixed point iteration method (Burden et al., 2015) is used to find the linearization point. Additional details and examples are available at <https://inlabru-org.github.io/inlabru/articles/method.html>.

4 Simulation Studies

In order to explore the importance of accounting for the sampling process of CS data as well as the uneven reporting effort between citizen scientists, we conduct a simulation study based on the Trøndelag case study with the powerlines of the region and some of the explanatory variables defined in Section 2.3. We set up for simulation scenarios representing different selection schemes of the powerlines visited by the experts (random or preferential sampling) and different willingness (low or high) to report dead birds once they have been detected by citizen scientists. For each of these scenarios 100 datasets are generated and fitted to a group of models that range from models that use each of the data sources separately to models that integrate professional surveys and opportunistic records while account for some biases in the collection of both data types.

4.1 Datasets simulation models

4.1.1 Simulation model of the true occurrence

To start our simulation study, we generated 100 spatial point patterns with intensity $\lambda_{true}(\mathbf{s})$ that represents the true risk intensity as:

$$\log(\lambda_{true}(\mathbf{s})) = -2 + 0.75 \text{ CLOUDCOVER}(\mathbf{s}) + \omega_1(\mathbf{s}) \quad (11)$$

where represents each of the 100 realizations of the GRF $\omega_1(\mathbf{s})$ with range $\rho_1 = 1$ and variance $\sigma_1^2 = 0.3$.

4.1.2 Citizen science simulation models

The points generated following the specification in Equation (11) were thinned with a thinning probability that depends on the sampling process of citizen scientists (see Sec. 3). In this case the sampling process of citizen scientists is represented through a point process with intensity $\lambda_{CS}(\mathbf{s})$ specified as:

$$\log(\lambda_{CS}(\mathbf{s})) = -4 - 2 \text{ DISTANCE}(\mathbf{s}) + \omega_2(\mathbf{s}) \quad (12)$$

where *DISTANCE* represents the distance to the closest tertiary road and $\omega_2(\mathbf{s})$ a GRF with range $\rho_2 = 100$ and variance $\sigma_2^2 = 1.3$. As the coefficient of the covariate *DISTANCE* is negative, those occurrences located at more accessible locations for citizen scientists have higher probability of being retained. The next stage of thinning depended on the detection probability of an occurrence of a dead bird based on factors such as land use, which affects how likely a dead bird is detected by an observer that has visited a location with an occurrence. The probability of retaining an occurrence, $\psi(\mathbf{s})$, was expressed as:

$$\text{logit}(\psi(\mathbf{s})) = 1 - 2\text{MOUNTAIN}(\mathbf{s}) + 1.2\text{OPEN}(\mathbf{s}) + 1.4\text{ROCKY}(\mathbf{s}) + 1.8\text{URBAN}(\mathbf{s}) - 3\text{WATER}(\mathbf{s}) \quad (13)$$

Each of the covariates in this model are indicator variables for each land use as defined in Section 2.3. The last stage of thinning depended on the probability of reporting the occurrence of a dead bird given that it has been detected and an observer has reached the place where the dead body was located. 10 observers were assumed as the population of citizen scientists and $\phi_j(\mathbf{s})$, as defined in Section 3.2.3, was modeled as:

$$\text{logit}(\phi_j(\mathbf{s})) = 10 - 0.3\text{DISTTOOBS}_j(\mathbf{s}) + \zeta\text{ACTLEVEL}_j \quad (14)$$

with $\text{DISTTOOBS}_j(\mathbf{s})$ the distance from the centroid of the citizen science activity of observer $j = \{1, \dots, 10\}$ to location \mathbf{s} and ACTLEVEL_j a categorical ordinal variable with the activity level (in scale 1-5) of each observer. The vector $\zeta = [0, 0.5, 1, 1.5, 2]$ determines the magnitude of the relation between each activity level and $\phi_j(\mathbf{s})$. Two definitions were given for $\kappa_j(\mathbf{s})$ depending of the willingness to report of the population of citizen scientists. Low willingness to report was assumed when the values of $\kappa_j(\mathbf{s})$ were generated using a *Beta*(2, 5) distribution (i.e. mean ≈ 0.29) and high willingness when a *Beta*(5, 1.5) distribution (i.e. mean ≈ 0.77) was used. A complete summary of the generation of each dataset is displayed in Figure S.1. in the supplementary information.

4.1.3 Professional surveys simulation models

These data were simulated to resemble the professional surveys carried out by NINA. For each simulated dataset we assume that the buffer around a powerline, P_i , is sampled with probability $\phi_i(P_i)$. This probability is assumed the same ($\phi_i(P_i) = p$) for all the buffers when random sampling is assumed and when preferential sampling is assumed, it is given by:

$$\text{logit}(\phi_i(P_i)) = -2.5 - 1.5\text{ELEVGRADIENT}(\mathbf{s}) + 3.5\text{CLOUDCOVER}(\mathbf{s}) \quad (15)$$

In this case we assume $\omega_1(\mathbf{s})$ is not part of the preferential sampling model in order to reduce the computational complexity of the simulations. Once a segment has been selected, all the occurrences within the buffer are assumed to be observed, detected and reported as the sampling effort in these surveys is high since the sampling is performed with trained dogs. A graphical summary of this sampling is presented in Figure S1

4.2 Simulation scenarios

For each simulation scenario 100 observed datasets were generated by thinning the 100 point patterns generated in Section according to Equation (11). The way these point patterns are thinned define the four simulation scenarios displayed in Figure 4.

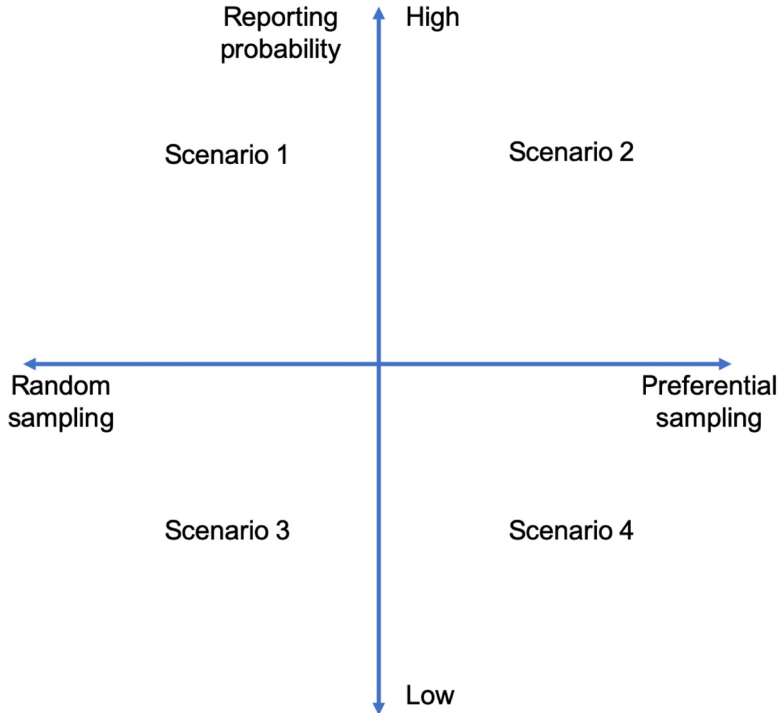


Figure 4: Scenarios for the simulation studies of Section 4.

Two factors define the four simulation scenarios: i) whether the probability of reporting a detected occurrence is high or low (see Section 4.1.2), and ii) if the selection of the powerlines in the professional surveys is completely random or preferential (see Section 4.1.3). In Scenarios 1 and 2 we assume the probabilities of reporting occurrences that were detected by the observers are high, but in Scenario 1 the sampling of the powerlines is assumed completely random whereas in Scenario 2 the sampling of the powerlines is preferential as in Equation (15). A similar relation holds between Scenarios 3 and 4. However, the probability of reporting for these scenarios are assumed low.

4.3 Model comparison

We want to compare the results of fitting different models (i.e. models that use only one data type and models that fusion both data types) for the datasets generated in the different simulation scenarios previously described. We defined the eight models in Table 1 and fitted them for each of the 400 generated datasets (100 true occurrence point patterns \times 4 simulation scenarios). Since models 7 and 8 are computationally expensive and specifically devised to for scenarios with low reporting probability (scenarios 3 and 4), they were not fitted in Scenarios 1 and 2.

Table 1: Description of the models fitted in the simulation study. In the data sources column: PS stands for Professional Surveys and CS for Citizen Science.

Model	Data sources	Description
1	PS	Only data from professional surveys (Equation (1))
2	PS	Only data from professional surveys accounting for preferential sampling (Equations (1) and (2))
3	CS	Only data with locations from CS reports (Equation (1))
4	CS	Same as model 3, but accounting for sampling process of CS (Equation (5) with $\delta(\mathbf{s})=1$)
5	CS	Same as model 4, but also accounting for the detection process (Equation (5) with $\delta(\mathbf{s})=1$)
6	PS + CS	Integration of models 2 and 5 accounting for sampling and detection process of CS and preferential sampling. (Equations (1), (2), and (5) with $\delta(\mathbf{s})=1$)
7	PS + CS	Integration of models 2 and 5 accounting for sampling, detection and reporting process of CS and preferential sampling. (Equations (1), (2), and (5) with $\delta(\mathbf{s})$ as in Equation (8))
8	PS + CS	Integration of models 2 and 5 accounting for sampling, detection and reporting process of CS and preferential sampling. (Equations (1), (2), and (5) with $\delta(\mathbf{s})$ as in Equation (9))

4.4 Results

The results of this simulation study are summarized for each of the parameters of the model using bias and RMSE. The effect of the covariates on the ecological state is of paramount importance for use as decision support. Hence, in Figure 5 we present the performance measures for the parameter β_1 , which in our simulation study represents the effect of the covariate *cloud cover* on the risk of powerline-induced deaths.

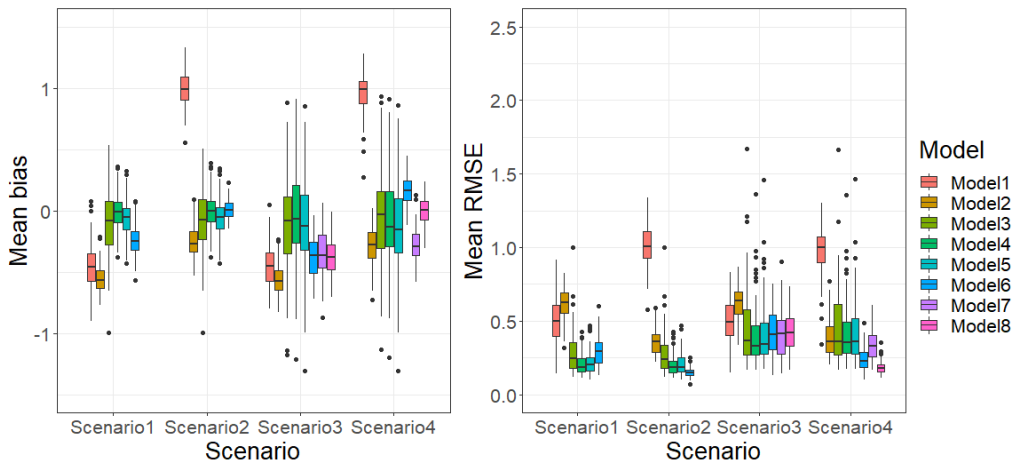


Figure 5: Mean bias and RMSE for β_1 for each of the eight models proposed.

We first observe that model 1 (based only on professional surveys and not accounting for preferential

sampling) performs poorly in scenarios with preferential sampling (scenarios 2 and 4), but outperforms model 2 (based only on professional surveys and accounting for preferential sampling) in scenarios with random sampling (Scenarios 1 and 3). Nevertheless, these two models were outperformed by all the models that are based on CS data (models 3-8) in all the simulation scenarios. Models based only on CS data (models 3, 4 and 5) performed better in scenarios with random selection of the sampled powerlines (scenarios 1 and 3). On the other hand, in the simulation scenario that assumed preferential selection of the powerlines and high willingness to report (Scenario 2), model 6 (both data types while accounting for preferential sampling and biases in CS data collection except reporting effort) performs much better than the other competing models showing, in addition, consistently less variability in the posterior means of the parameter β_1 . Model 8 (both data types while accounting for preferential sampling and all biases in CS data collection) performed better than all the competing models in both bias and RMSE for the scenario with preferential sampling and low willingness to report (scenario 4). In this scenario model 6 outperformed model 7 (both data types and simple method to account for reporting effort) arguably due to the lack of structure of model 7 in the explanation of the differences in reporting effort. Finally, note that the variability in the posterior means of β_1 is much smaller in scenarios with high willingness to report (scenarios 1 and 2) for all the models, compared to scenarios with low willingness to report (scenarios 3 and 4) as much more information is available in CS data.

The summaries for both the fixed effects and the spatial hyperparameters are presented in Table S1 of the Supplementary Information. For β_0 we find out that models that do not account for any bias in the collection process (models 1 and 3) underestimate its value for all the simulation scenarios, while model 5 (based only on CS data and accounting for accessibility and detectability) is the one that performs the best for scenarios 1 and 2. Besides, in scenarios with high willingness to report (scenarios 3 and 4), models 6, 7 and 8 (models that fusion both data types) outperformed all the other competing models. Note that the estimates of the spatial hyperparameters of the GRF $\omega_1(\mathbf{s})$ are inaccurate for the most complex models (models 2, 6, 7 and 8). To explore in more detail the effect of this bias, we have fixed the spatial hyperparameters for one of the datasets of the simulation studies and have fitted models 6, 7 and 8. The marginal posterior distributions of the fixed effects are displayed in the Supplementary Information. We arrive to similar results regarding the accuracy of the marginal posterior distributions for parameters β_0 and β_1 . However, we noticed that these posterior distributions are more precise than when the spatial hyperparameters were not fixed. Given the sensitivity of the posterior distributions of the spatial hyperparameters to the prior specification of the spatial hyperparameters in models 6, 7 and 8, more informative prior information could contribute to obtain more accurate posterior distributions for both the fixed effects and the spatial hyperparameters.

Our proposed framework does not only allow us to infer the posterior distribution of the parameters that drive the ecological process, but also to learn about the processes that drive the biases in CS data and professional surveys. In Figures S.2.-S.15. and tables S.1.-S.4. in the supplementary information, we present the comparison measures for the parameters involved in the thinning of the true point pattern to produce both CS and professional surveys data. In these figures we find out that the parameters of the sampling process model are similarly estimated by models 4 to 8, with small biases for both α_0 and α_1 . On the other hand, the parameters associated with the detectability exhibit biases and large RMSE. This might be linked to poor identifiability of these parameters, which could be remediated with informative prior information or more information about the relation between land use and detectability. Finally, the parameters of the preferential sampling model are accurately estimated in Scenarios 2 and 4 as expected. A particularity of model 8 is that it allows for posterior inference about the willingness to report occurrences of powerline-induced deaths for each observer. Figure S.16. shows the marginal posteriors for each of the observers in the simulation study. Though inaccurate due to lack of information, the information in Figure S.16. has potential for CS projects to target a group of observers.

Finally, we compare the proposed models in terms of predictive performance. We focus our comparison on two aspects: the accuracy of the predictions and how uncertain these predictions are. These predictions are only based on the fixed effects of the linear predictor. To compare the accuracy of the predictions produced, we computed the Root Mean Squared Error for the predicted probabilities by each model in each simulation scenario on a dense grid along the powerline network of Trøndelag. The RMSE maps for each model in each scenario are presented in Figures S.17., S.22, S.27 and S.32. The results show that in scenarios

with high willingness to report (scenarios 1 and 2), model 5 (based only on CS data and accounting for accessibility and detectability) outperforms all the other models. In scenarios with low willingness to report (scenarios 3 and 4), model 5 was outperformed by models that fusion both data types (models 6-8) in vast portions of the study region.

Regarding the uncertainty of the predictions produced in the different simulation scenarios, we computed the average width of the 95% prediction intervals for each model over the dense grid mentioned above. The maps with the length of the prediction intervals are displayed in Figures S.19., S.24, S.29 and S.34. Model 2 (using only professional surveys while accounting for preferential sampling) produces the larger uncertainties, while the most complex models (models 6 to 8) have larger uncertainties than the models based only on CS data (models 3-5).

5 Case study of bird mortality and power lines in Trøndelag, Norway

Our real data application study aims to gain a better understanding of the role of the landscape (environmental covariates) in creating riskier regions for powerline-induced deaths for birds. Based on the effect of these on the risk of powerline-induced deaths, risk maps of powerline-induced deaths can be made to inform electricity companies and the Water and Energy Directorate (NVE). To achieve this goal we have two sources of information (see Section 2): 1) professional surveys performed by NINA and 2) opportunistic records collected by citizen scientists. In this section we fit models 1 to 7, presented in Section 4, then show model predictions on the risk of powerline-induced death of birds, and conclude by comparing the predictions obtained with the proposed models.

The selection of the candidate variables to explain the processes behind both observed data types was based on expert knowledge and the exploratory analysis performed in Section 2. As higher occurrences of powerline-induced deaths are expected as more birds are exposed and the powerline network is dense, we have considered the covariates *powerline density* (The Norwegian Water Resources and Energy Directorate; NVE) and *bird abundance* (Sicacha-Parada et al., 2022) as the first candidate variables. *Land use* (AR50, <https://www.nibio.no/tema/jord/arealressurser/ar50>), which explains the type of land around a powerline, and *cloud cover* (<https://www.earthenv.org/cloud>), which proxies the visibility the birds have as they fly, were also considered as candidate variables, but were discarded as land use was correlated with the two chosen covariates and adding cloud cover did not improved the existing models. As argued previously in Fithian et al. (2015) and Monsarrat et al. (2019), accessibility is one of the main factors that determine where citizen scientists are more prone to collect information on biodiversity. For this reason, the covariates *distance to (tertiary) roads* and *distance to water bodies* (sea, lakes and rivers) were considered to explain where citizen scientists collect their observations. Land use was not considered to explain where citizen scientists go to collect observations as this covariate is also correlated with the distance variables. Another factor that needs consideration is the detectability of a dead bird. This can be affected by the size of the bird (Borner et al., 2017; Ponce et al., 2010), the time of the year the observation is made (Bevanger, 1995), the land use of the place where the carcass is located (Philibert et al., 1993; Schutgens et al., 2014; Domínguez del Valle et al., 2020) among other factors. In our case, we consider the different land uses as proxies for explaining differences in detectability. Finally, whether or not citizen scientists are willing to report a dead bird is an unsolved question. Hence, we have opted for considering this as another relevant factor that affects what is observed in CS databases and account for it using Model 7.

Preferential sampling was also considered as a possible flaw in the collection process of the data from professional surveys performed by NINA since the experts leading these projects have prior knowledge they might use when determining where to collect observations. The results of Section 2 suggest powerline density and cloud cover as possible drivers for preferential sampling. According to information provided by NINA, elevation gradient was also relevant for choosing which lines to visit. Given its correlation with cloud cover (Pearson correlation coefficient, $\rho = 0.76$), the variables chosen to explain the preferential sampling were the elevation gradient and powerline density.

In Section 3, we pointed out the potential identifiability issues that may arise when trying to account for biases in CS data as the number of parameters to estimate increases considerably. Therefore, in addition to the two observed data types, we have also included a point pattern with the locations of CS reports so that the parameters of the sampling process can be identified. For the parameters that link the different land uses and the detection of dead birds, we have proposed informative prior distributions based on expert knowledge (Domínguez del Valle et al., 2020) as no studies that explain the relation between land use and detectability for Norway are available. We fitted the seven proposed models while integrating the available data types for each source of bias in the collection process. The posterior summaries for each of the fixed effects involved in the ecological process are graphically presented in Figure 6.

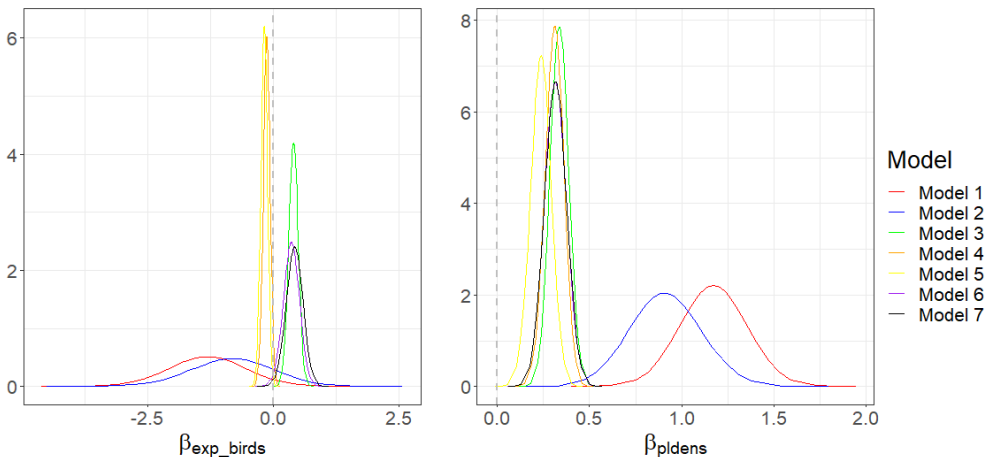


Figure 6: Posterior distributions for the fixed effects of the seven models fitted

For the effect of the covariate *Exposed birds* we observe considerable disagreement between the models proposed. Models 1,2, 4 and 5 have marginal posterior distributions centered below 0, whereas for models 3,6 and 7 this distribution is located mostly above 0. The effect of the covariate *powerline density* is more consistent for all the models proposed, with marginal posterior distributions centered around the same values, except for models 1 and 2. It is worth noting that the posterior distributions of models 1 and 2 show much higher uncertainty than for the other models. The posterior summaries for these and the other parameters in the model are presented in the supplementary information. Note that, as in the simulation study, the spatial range parameter and the marginal variance took much higher values for models 2, 6 and 7 than for models 3, 4 and 5.

The risk of powerline-induced deaths was predicted along the powerline network in Trøndelag using each of the seven models proposed. The posterior medians of the risk are presented in Figure 7

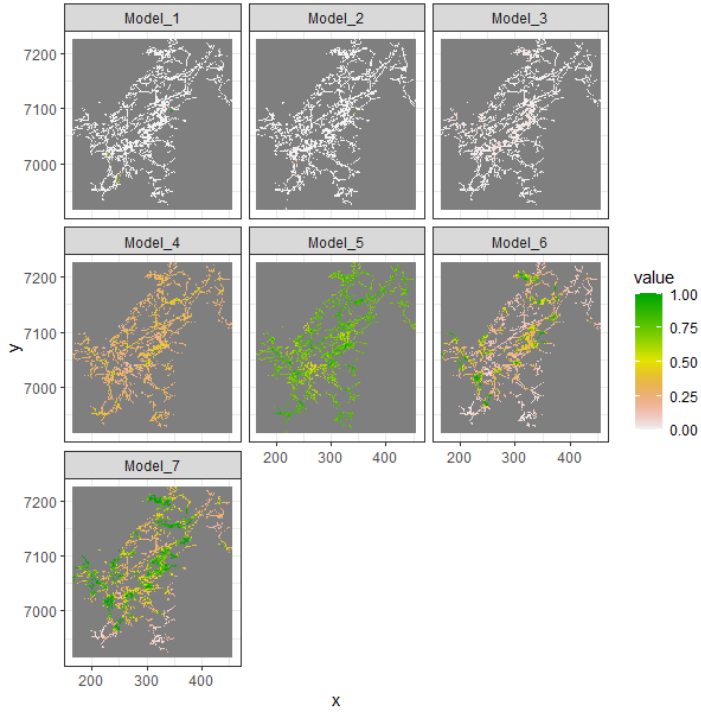


Figure 7: Posterior median of the risk of powerline-induced deaths

As seen in Table S.5., the posterior means of the intercept in models 1, 2 and 3 make the posterior median of the predicted risk to take low values, except for locations where sampling has been performed. More realistic predictions are produced by models 4-7. Model 4 highlights very few spots in the study region, while model 5 predicts larger risk all over the region after accounting for differences in detectability. Once both data types have been integrated (models 6 and 7) fewer areas are highlighted as riskier for powerline-induced deaths. Although similar, the probabilities predicted using model 7 are higher than using model 6 as the former accounts for differences in reporting effort. The uncertainty of the predictions was computed through the standard deviation of the predicted risks and is presented in Figure 8.

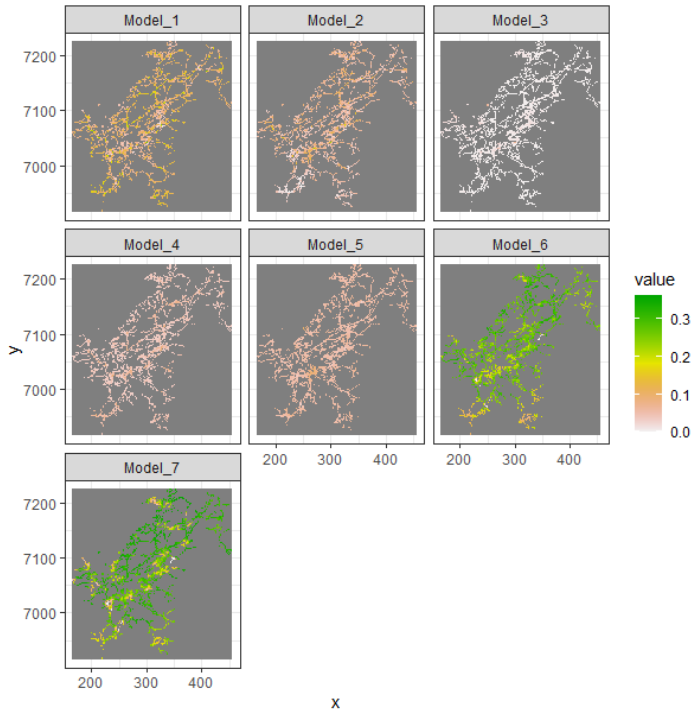


Figure 8: Standard deviation of the predicted risk of powerline-induced deaths

As expected, the predictions based only on professional surveys, which cover a small portion of the study area, have higher uncertainty than those obtained using only opportunistic records in CS data. As seen in Figure S.40., the uncertainty of the term $\omega_1(\mathbf{s})$ is considerable for models 6 and 7 across all the study area, while the uncertainty of the fixed effects is larger for models 1 and 2. After transforming to the scale of the risk of powerline-induced death, models 6 and 7 are the most uncertain.

These results are consistent with the predictions made for the simulated datasets in Section 4 as models that use only CS data produce the less uncertain predictions. However, models 6 and 7, which use only data from professional surveys, produce predictions with higher uncertainty than all the other models.

6 Discussion

In this paper we have proposed and evaluated a methodological framework for integrating multiple data types to address questions in ecology and biodiversity. We focus on two types of data to integrate: professional surveys and opportunistic records collected by citizen scientists. A fundamental assumption of this work is that the observations in both data types have the same underlying ecological process as their origin, but they have different observation and reporting processes, which generates observations with different spatial coverages and different observation efforts. We approached the challenge of integrating these data types in first place by modeling the factors that determine the generation of each data type. In particular, the data collected by experts is affected by prior knowledge of the professionals that conduct these surveys, thus a preferential sampling design that gives some powerlines higher chances of being visited than others is assumed. On the other hand, CS data is affected by factors accessibility, detectability and uneven reporting effort among citizen scientists. Both data types were managed as realizations of thinned point processes, but

with different processes behind the thinning of the actual ecological process. While CS data was assumed affected by processes that occur at point level, we assumed that data from professional surveys was affected by selection probabilities occurring at areal level.

To account for the factors that affected the collection process of both data types and avoid identifiability issues (Fithian et al., 2015), we used additional sources of information that provided knowledge of the processes involved in the thinning. As no previous studies that inform about the link between land use/habitat and detectability of dead birds was available for our study region, priors based on expert knowledge were proposed. We proposed Bayesian spatial models that resemble the generation of each data type and lie within the class of Latent Gaussian Models, hence they were fitted using the INLA-SPDE approach as this is a computationally efficient approach for fitting this class of spatial Bayesian models. However, as seen in Section 3, the terms of the linear predictor that explain the effect of biases in the collection process on the generation of the observed point patterns do not follow the log-linear requirement underlying the INLA methodology for log-Gaussian Cox processes. For this reason, as a complement to the traditional INLA-SPDE approach, we used the approach in the *inlabru* R-package (Bachl et al., 2019), which is based on an iterative linearization of the non-linear terms in the linear predictor. This extra step might represent additional computational burden when a GRF is part of the nonlinear part of the linear predictor since the sparsity induced by the SPDE approach and the approximation of the Gaussian Random Fields as Gaussian Markov Random Fields might be reduced. For the most complex models (models 7 and 8 defined in Section 4), we experienced some numerical issues that might be related to this approximation as convergence was not reached.

A simulation study and a case study were performed in order to study characteristics and properties of the models proposed in Section 3. Models that used one or two data types and that accounted for biases in the collection process to different extents were fitted for 100 different datasets in scenarios with professional surveys that select powerlines preferentially and randomly and with high and low willingness to report powerline-induced deaths by citizen scientists. The results of the simulation studies show the relevance of integrating both data types in scenarios with preferential sampling as model 6-8 produce more accurate estimates for the fixed effects in Scenarios 2 and 4 (scenarios with preferential sampling), which are the scenarios that more closely resemble what occurs in the case study. In Scenarios 1 and 3 (scenarios without preferential sampling), models that account for preferential sampling do not perform as well as in the other scenarios, probably due to lack of identifiability of the large amount of parameters these models introduce.

The assessment of the predictive performance of the proposed models show that models that integrate both data types (models 6-8) performed better than those based on only one data type (models 1-5). Regarding the uncertainty of the predictions, those models based only on professional surveys (models 1 and 2) predicted with larger uncertainty than those models based on only CS data (models 3-5), likely due to the larger area covered by CS data. Also the models that integrated both data types (models 6-8) produced predictions with larger uncertainty than model based only on CS data (models 3-5). A similar pattern was observed in our case study, where predictions using models 1, 2, 6 and 7 had higher uncertainty than the predictions of models 3-5.

In the simulation studies we found that the posterior estimates of the spatial hyperparameters of the most complex models (models 2 and 6-8) were considerably more biased compared to the estimates obtained using only CS data. In particular, in both the simulation studies and the case study the spatial range of $\omega_1(\mathbf{s})$ was much larger for these models. Arguably the choice of prior distribution is more influential on the models based only on professional surveys (models 1 and 2) as the spatial coverage of the professional surveys is small.

Through the simulation studies we noticed that the proposed framework contributes to account for both the fixed effects behind an ecological process and the spatial autocorrelation that determines it. Besides, this framework has been proven useful as well to account for and quantify the factors that affect the collection process of both CS data and professional surveys. Hence, the estimates of our models can be used to devise more informed sampling of CS data by focusing sampling efforts in areas with higher uncertainty or with low sampling effort. The simulation studies also showed the importance of knowing more about the preferences

of citizen scientists. If more is known about these preferences, more targeted activities for citizen scientists can be launched in order to target specific research questions and the differences in reporting effort could be accounted for better inference.

The case study of deaths caused by powerlines in Trøndelag, Norway motivated this paper. Through this case study we showed the importance of proposing methods that combine more than one data type as we showed how the effect of a factor like the amount of exposed birds varies amongst models. Moreover, the prediction maps produced highlighted zones with higher risk of powerline-induced deaths. These maps could be used by conservation programs to target mitigation measures for powerlines with higher risks.

The proposed methods in this paper are made available also through code such that practitioners in ecology and biodiversity can use them to address many other questions using multiple sources of information while accounting for other sources of bias given the flexible specification of the models for both professional surveys and CS data.

Bibliography

- August, T., Fox, R., Roy, D. B., and Pocock, M. J. (2020). Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. *Scientific reports*, 10(1):1–12.
- Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an r package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2ed. edition.
- Barrientos, R., Ascensão, F., D’Amico, M., Grilo, C., and Pereira, H. M. (2021). The lost road: Do transportation networks imperil wildlife population persistence? *Perspectives in Ecology and Conservation*, 19(4):411–416.
- Barrientos, R., Ponce, C., Palacín, C., Martín, C. A., Martín, B., and Alonso, J. C. (2012). Wire marking results in a small but significant reduction in avian mortality at power lines: a baci designed study. *PLoS One*, 7(3):e32569.
- Bernardino, J., Bevanger, K., Barrientos, R., Dwyer, J., Marques, A., Martins, R., Shaw, J., Silva, J., and Moreira, F. (2018). Bird collisions with power lines: State of the art and priority areas for research. *Biological Conservation*, 222:1–13.
- Bevanger, K. (1995). Estimates and population consequences of tetraonid mortality caused by collisions with high tension power lines in norway. *Journal of Applied Ecology*, 32(4):745–753.
- Bevanger, K., Bartzke, G., Brøseth, H., Dahl, E., Gjershaug, J., Hanssen, F., Jacobsen, K.-O., Kleven, O., Kvaløy, P., May, R., Meås, R., Nygård, T., Refsnæs, S., Stokke, S., and Thomassen, J. (2014). Optimal design and routing of power lines: ecological, technical and economic perspectives (optipol). final report: findings 2009-2014.
- Bevanger, K. and Brøseth, H. (2001). Bird collisions with power lines—an experiment with ptarmigan (*lagopus* spp.). *Biological Conservation*, 99(3):341–346.
- Biasotto, L. D. and Kindel, A. (2018). Power lines and impacts on biodiversity: A systematic review. *Environmental Impact Assessment Review*, 71:110–119.
- Borner, L., Duriez, O., Besnard, A., Robert, A., Carrere, V., and Jiguet, F. (2017). Bird collision with power lines: estimating carcass persistence and detection associated with ground search surveys. *Ecosphere*, 8(11):e01966.

- Botella, C., Joly, A., Bonnet, P., Munoz, F., and Monestiez, P. (2021). Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, 12(5):933–945.
- Burden, R. L., Faires, J. D., and Burden, A. M. (2015). *Numerical analysis*. Cengage learning.
- Conti, J., Holtberg, P., Diefenderfer, J., LaRose, A., Turnure, J. T., and Westfall, L. (2016). International energy outlook 2016 with projections to 2040. Technical report, USDOE Energy Information Administration (EIA), Washington, DC (United States
- Davis, G. (2002). A roadmap for pier research on avian collisions with power lines in california.
- Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.
- Domínguez del Valle, J., Cervantes Peralta, F., and Jaquero Arjona, M. I. (2020). Factors affecting carcass detection at wind farms using dogs and human searchers. *Journal of Applied Ecology*, 57(10):1926–1935.
- Drewitt, A. L. and Langston, R. H. (2008). Collision effects of wind-power generators and other obstacles on birds. *Annals of the New York Academy of Sciences*, 1134(1):233–266.
- Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.
- Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. *Journal of the American Statistical Association*, 114(525):445–452.
- Gelfand, A. E. and 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.
- Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in medicine*, 11(9):1209–1223.
- Hernández-Lambráño, R. E., Sánchez-Agudo, J. Á., and Carbonell, R. (2018). Where to start? development of a spatial tool to prioritise retrofitting of power line poles that are dangerous to raptors. *Journal of Applied Ecology*, 55(6):2685–2697.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.
- Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., and 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.
- JANSS, G. F. and FERRER, M. (2001). Avian electrocution mortality in relation to pole design and adjacent habitat in spain. *Bird Conservation International*, 11(1):3–12.
- Jenkins, A. R., Smallie, J. J., and Diamond, M. (2010). Avian collisions with power lines: a global review of causes and mitigation with a south african perspective. *Bird Conservation International*, 20(3):263–278.
- Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., and Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4):420–430.
- Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.
- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

- López-López, P., Ferrer, M., Madero, A., Casado, E., and McGrady, M. (2011). Solving man-induced large-scale conservation problems: the spanish imperial eagle and power lines. *PLoS one*, 6(3):e17196.
- MacKenzie, D. I. and Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance? *Ecology*, 83(9):2387–2393.
- Martin, G. and Shaw, J. (2010). Bird collisions with power lines: failing to see the way ahead? *Biological Conservation*, 143(11):2695–2702.
- Martin, G. R. (2011). Understanding bird collisions with man-made objects: a sensory ecology approach. *Ibis*, 153(2):239–254.
- Miller, D. A. W., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species’ distributions. *Methods in Ecology and Evolution*, 10(1):22–37.
- Monsarrat, S., Boshoff, A. F., and Kerley, G. I. H. (2019). Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. *Ecography*, 42(1):125–136.
- Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., and Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: a framework for data fusion*. *Ecology*, 98(3):840–850.
- Pavón-Jordán, D., Stokke, B. G., Åström, J., Bevanger, K., Hamre, Ø., Torsæter, E., and May, R. (2020). Do birds respond to spiral markers on overhead wires of a high-voltage power line? insights from a dedicated avian radar. *Global Ecology and Conservation*, 24:e01363.
- Philibert, H., Wobeser, G., and Clark, R. (1993). Counting dead birds: examination of methods. *Journal of Wildlife Diseases*, 29(2):284–289.
- Ponce, C., Alonso, J. C., Argandoña, G., García Fernández, A., and Carrasco, M. (2010). Carcass removal by scavengers and search accuracy affect bird mortality estimates at power lines. *Animal Conservation*, 13(6):603–612.
- Roksvåg, T., Steinsland, I., and Engeland, K. (2020). Estimation of annual runoff by exploiting long-term spatial patterns and short records within a geostatistical framework. *Hydrology and Earth System Sciences*, 24(8):4109–4133.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.
- Schutgens, M., Shaw, J. M., and Ryan, P. G. (2014). Estimating scavenger and search bias for collision fatality surveys of large birds on power lines in the karoo, south africa. *Ostrich*, 85(1):39–45.
- Serrano, D., Margalida, A., Pérez-García, J. M., Juste, J., Traba, J., Valera, F., Carrete, M., Aihartza, J., Real, J., Mañosa, S., et al. (2020). Renewables in spain threaten biodiversity. *Science*, 370(6522):1282–1283.
- Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B., and Øien, I. J. (2022). A spatial modeling framework for monitoring surveys with different sampling protocols with a case study for bird abundance in mid-scandinavia. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–30.
- Sicacha-Parada, J., Steinsland, I., Cretois, B., and Borgelt, J. (2021). Accounting for spatial varying sampling effort due to accessibility in citizen science data: A case study of moose in norway. *Spatial Statistics*, 42:100446.

- Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., and O’Hara, R. B. (2020). Is more data always better? a simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43(10):1413–1422.
- Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., and Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.
- United Nations (2018). The sustainable development goals report 2018. Technical report, United Nations.
- Wang, C., Furrer, R., Group, S. S., et al. (2021). Combining heterogeneous spatial datasets with process-based spatial fusion models: A unifying framework. *Computational Statistics & Data Analysis*, 161:107240.
- Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Itter, M. S., and Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, 19(1):30–38.

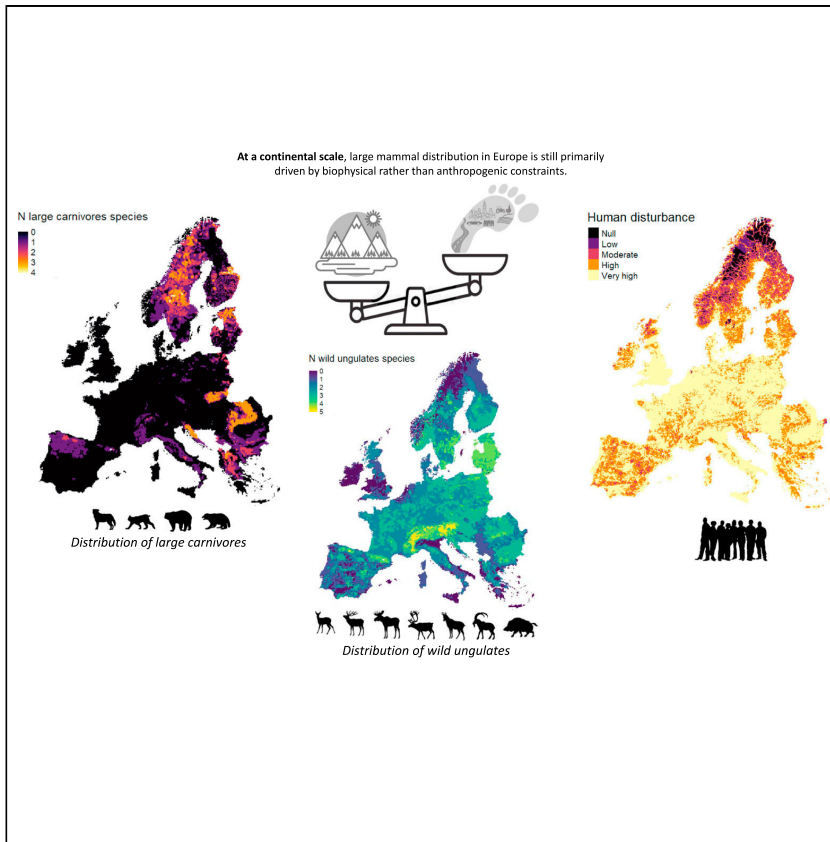
Paper IV

Coexistence of large mammals and humans is possible in Europe's anthropogenic landscapes

Cretois, B., Linnell, J. D., Van Moorter, B., Kaczensky, P., Nilsen, E. B., Parada, J., & Rød, J. K. (2021) published in *iScience*

Article

Coexistence of large mammals and humans is possible in Europe's anthropogenic landscapes



Benjamin Cretois,
John D.C. Linnell,
Bram Van
Moorter, Petra
Kaczensky, Erlend
B. Nilsen, Jorge
Parada, Jan Ketil
Rød

benjamin.cretois@nina.no,
bencretois@gmail.com

Highlights

Biophysical factors had a far greater impact on species distribution than human factors

This indicates that most species have a broad tolerance of human land use at coarse scales

We provide grounds for cautious optimism for wildlife conservation in the Anthropocene

Cretois et al., iScience 24,
103083
September 24, 2021 © 2021
The Author(s).
<https://doi.org/10.1016/j.isci.2021.103083>



Article

Coexistence of large mammals and humans is possible in Europe's anthropogenic landscapes

Benjamin Cretois,^{1,2,5,*} John D.C. Linnell,^{2,3} Bram Van Moorter,² Petra Kaczensky,^{2,3} Erlend B. Nilsen,² Jorge Parada,⁴ and Jan Ketil Rød¹

SUMMARY

A critical question in the conservation of large mammals in the Anthropocene is to know the extent to which they can tolerate human disturbance. Surprisingly, little quantitative data is available about large-scale effects of human activity and land use on their broad scale distribution in Europe. In this study, we quantify the relative importance of human land use and protected areas as opposed to biophysical constraints on large mammal distribution. We analyze data on large mammal distribution to quantify the relative effect of anthropogenic variables on species' distribution as opposed to biophysical constraints. We finally assess the effect of anthropogenic variables on the size of the species' niche by simulating a scenario where we assumed no anthropogenic pressure on the landscape. Results show that large mammal distribution is primarily constrained by biophysical constraints rather than anthropogenic variables. This finding offers grounds for cautious optimism concerning wildlife conservation in the Anthropocene.

INTRODUCTION

Even though most conservation actions have the primary objective of safeguarding the long-term persistence of wildlife, there is substantial disagreement about the most effective strategies to achieve these goals (e.g., land sparing vs land sharing, [Phalan et al., 2011](#)). Some conservationists advocate for implementing a spatial dichotomy, where "wild areas" would be subject to minimal human intervention (land sparing) acting as refugia for wildlife against human disturbance. Another paradigm consists of a diversity of coexistence strategies (land sharing), which envisions the possibility of shared landscapes where human and wildlife interactions are allowed, managed and sustained by effective institutions ([Carter and Linnell, 2016](#); [Linnell and Kaltenborn, 2019](#)).

Adopting a land sharing strategy requires a mutual adaptation in behavior from both humans and wildlife ([Carter and Linnell, 2016](#)). This may seem especially challenging for large animals as they are more likely to be negatively impacted directly (e.g., through persecution and exploitation) and indirectly (loss and fragmentation of habitats) by human activities owing to their larger spatial and resource requirements and the potential for human-wildlife conflicts ([Redpath et al., 2013](#)). Because of their size, large animals with wide-ranging behavior and slow reproductive rates are frequently viewed as being at a disproportionately high risk of extinction ([Ripple et al., 2014, 2015](#)).

Coexistence with large mammals has been a historical challenge in Europe. Large carnivores were extensively persecuted in retaliation for killing livestock while large ungulates were overexploited for sport and meat hunting and to minimize damage to crops and forests ([Ripple et al., 2014, 2015](#)). This resulted in populations of both taxa being driven to the edge of a near continent-wide extinction in the 19th and early 20th centuries ([Chapron et al., 2014](#); [Apollonio et al., 2010](#)). Even though European landscapes are among the most affected by humans ([Venter et al., 2016](#)), strict regulations, reintroduction programs, effective wildlife management institutions, reforestation and agricultural abandonment have allowed most large mammal species to recover. Nowadays, these species are again found across very large areas of the European landscape ([Chapron et al., 2014](#); [Linnell and Zachos, 2010](#); [Linnell et al., 2020](#)).

Another factor which could potentially have contributed to the re-establishment of these species and their widespread distribution is the widespread protected area network created throughout Europe. However, because of the diverse legislative framework and multiple goals (i.e., encouraging tourism and allowing

¹Department of Geography, Norwegian University of Science and Technology, 7491 Trondheim, Norway

²Norwegian Institute for Nature Research, PO Box 5685, Torgard, 7485 Trondheim, Norway

³Department of Forestry and Wildlife Management, Inland Norway University of Applied Sciences, 2480 Koppang, Norway

⁴Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway

⁵Lead contact

*Correspondence :: benjamin.cretois@nina.no, bencretois@gmail.com

<https://doi.org/10.1016/j.isci.2021.103083>



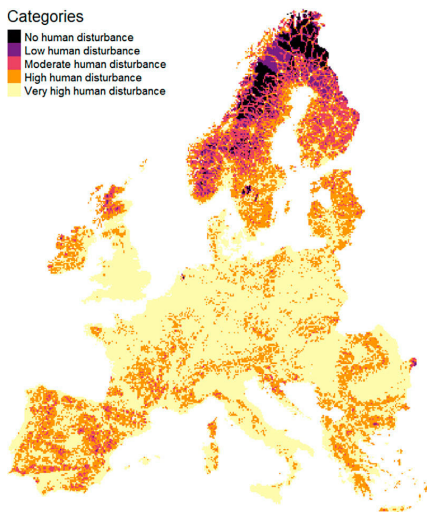


Figure 1. Map of the distribution of human disturbance levels in Europe

hunting and different forms of traditional land use) and their small sizes, the conservation effectiveness of protected areas in Europe has been widely disputed for highly mobile, large mammals (Linnell et al., 2015; Gaston et al., 2008).

Although there is an increasing body of literature addressing the influence humans have on large mammals (Tucker et al., 2018; Carter et al., 2012; Alexander et al., 2016), we are not aware of any attempts to quantify the extent to which the contemporary recovering distributions of large predators and their prey in Europe are constrained by the presence of humans' modification in their habitat as opposed to underlying biophysical constraints. The issue is important to understand the factors limiting the potential for large-scale land-sharing in a crowded and human-modified continent.

In this study we evaluate the relative effects of both the human footprint, a proxy for human disturbance levels widely used in large-scale ecological studies (Belote et al., 2020; Tucker et al., 2018, 2021) and protected areas (i.e., to which extent human footprint and protected areas explain species distribution) after accounting for natural heterogeneity. We compare the effect of these two human variables with the effects of biophysical environmental variables such as climate and terrain on large mammal distribution at a continental scale. We use Bayesian hierarchical models to estimate the importance of these variables on species' distributions and compare the environmental niche of these species with and without accounting for human variables by simulating a scenario where the European landscape is free of human influence.

RESULTS

For ease of interpretation, we consider five disturbance levels (Venter et al., 2016). A 'no human disturbance' area has a human footprint of 0; a 'low disturbance' area with a human footprint of 1–2; a 'moderate disturbance' area a human footprint of 3–5; a 'high disturbance' area; a human footprint of 6–11; and 'very high disturbance' area with a human footprint of 12–50, following the definition by Venter et al. (2016).

With a median human footprint of 12.2, summary statistics show that more than 50% of Europe's area is in an area of very high human disturbance, whereas less than 8% of Europe has no to low human footprints (Figure 1). Protected areas are spread throughout Europe with the median area of protected areas per 100 km² (i.e. per 10 km × 10 km grid cell) being 9 km² (Q1 = 0 km², Q3 = 41 km²). Grid cells containing at least 50 km² of protected areas tended to have on average a slightly lower human footprint than grid cells containing less than 50 km² of protected areas (median = 10.04 and 12.98 respectively).

The seven large ungulates and four large carnivores demonstrate great variability in their presence across the human footprint gradient (Figure 2). Roe deer (median of 12.8, Q1 = 8.2, Q3 = 18.2) and wild boar

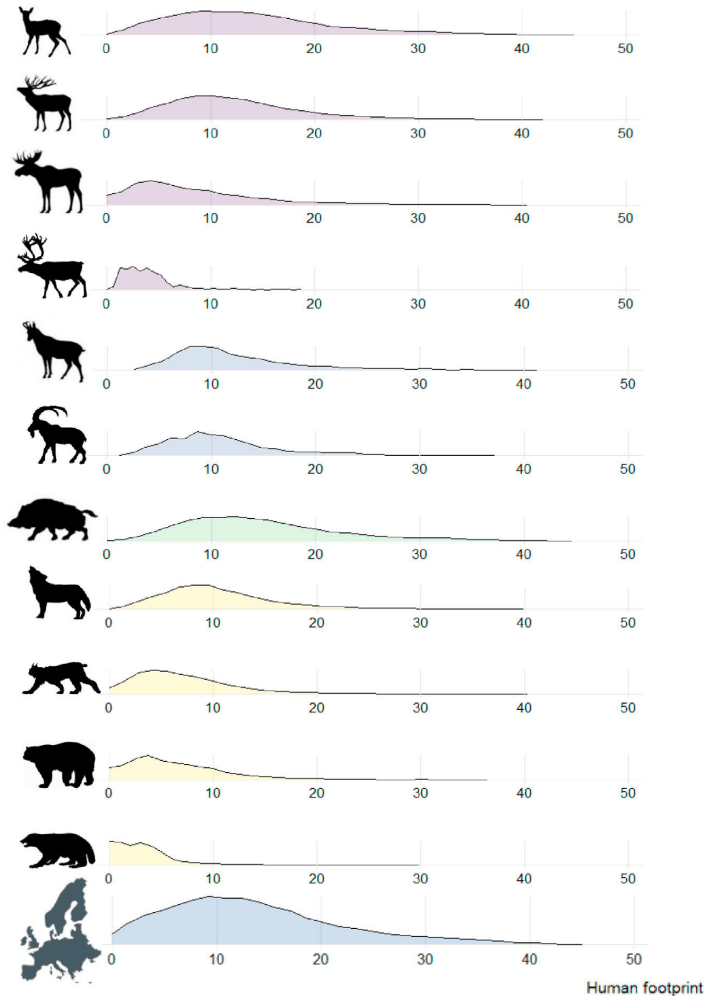


Figure 2. Ridge plot displaying the species' distributions data across the human footprint gradient

From top to bottom: roe deer, red deer, moose, wild reindeer, chamois, ibex, wolf, lynx, bear, wolverine, and the European human footprint distribution.

(median = 13.5, Q1 = 9.2, Q3 = 18.7) are the species present at the highest human footprints. These statistics show that more than 50% of the roe deer and wild boar distribution occurs in areas of very high human footprint. Wild reindeer (median = 3.9, Q1 = 2.1, Q3 = 4.8) and wolverines (median = 2.7, Q1 = 1.1, Q3 = 4.4) are at the other end of the spectrum with distributions in places that are least impacted by human disturbance. Our data also shows that wolves are not restricted to “wild” remote places but live in areas where human disturbance is high (median = 9.6, Q1 = 6.8, Q3 = 13). More than 25% of their distribution is in areas where human disturbance is very high.

Results from the dominance analysis show that the distributions of all 11 species are largely explained by the biophysical variables (Figure 3). In fact, biophysical variables consistently dominate the models (with a relative importance close to 100%) and the influence of anthropogenic variables in our models is shown to

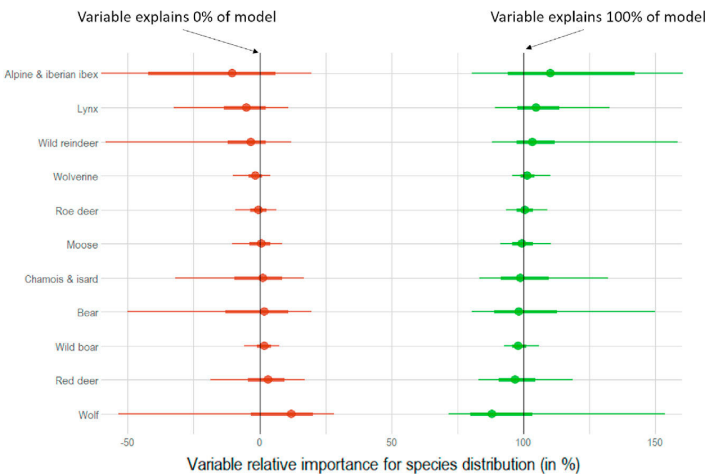


Figure 3. Relative importance for model fit (in percentage) of anthropogenic variables (human footprint and protected area coverage; in red) and biophysical variables (winter and summer severity and terrain ruggedness; in green) to species distribution

Negative importance indicates a drop in the R^2 when the variable is included in the model. Points represent the median value, thick lines represent the 50% credible interval, and thin lines represent the 95% credible interval.

be close to 0% or even negative (i.e., the R^2 of the model gets worse as we include these variables). Only for red deer and wolf do anthropogenic variables increase the models' R^2 values (median = 3.3% and median = 12%, respectively), although their effects were still considerably lower than those of the biophysical variables.

Finally, in [Figure 4](#) we show that human modifications on the landscape hardly influence the area of species' potential distribution. The suitable area for most studied mammals (i.e., ibex, wild reindeer, bears, wolverines, red deer, and moose) is weakly influenced by setting both human footprint and protected areas to zero. Only in the case of chamois and roe deer, we did observe a strong decrease in predicted suitable area when setting the anthropogenic variables to zero (median = $-13,900$ and $-284,400$ km^2 , respectively). We also observed a decrease of the predicted suitable area for wolverine, wild reindeer, and ibex when removing anthropogenic effects (median = $-12,900$ and $-6,200$ km^2 , respectively), because of the removal of protected areas (see [Figures S1](#) and [S2](#) in the Annexes). In contrast, the total predicted suitable area available for wolf, lynx, and wild boar increases when anthropogenic effects are set to zero (median = $50,700$, $133,400$ and $131,200$ km^2 respectively). These predicted gains represent 17%, 6%, and 4% of the actual lynx, wolf, and wild boar distributions, respectively.

DISCUSSION

In this study we have demonstrated that the large-scale distributions of Europe's main large mammalian species include large areas of high to very high human disturbance. Even though there is a wide distribution of high human disturbance combined with a rarity of wild places in the European landscape ([Venter et al., 2016](#)) these results show that large mammals can maintain a presence in these heavily modified multi-use landscapes. We have further shown that human disturbance and protected area coverage are only minor drivers of large mammal distributions at the continental scale. Overall, for all large mammals, our results show that the anthropogenic variables are poor predictors of species distribution compared to the other biophysical environmental variables.

Large-scale studies (e.g., with a continental scope) and finer scale studies (e.g., with a sub-national scope) do not answer the same questions, and their results can apparently be in contradiction. Failure to consider scale can lead to misinterpretation of results ([Johnson, 1980](#)) and conservation scientists should be careful

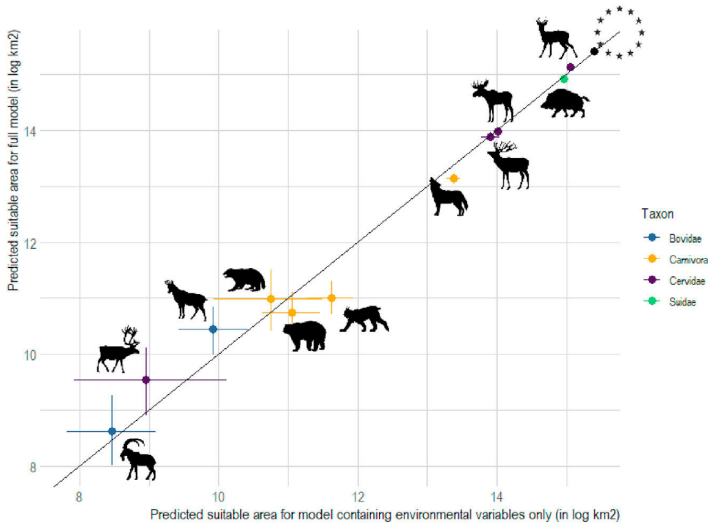


Figure 4. Predicted environmental niche (in log in km²) of European large mammals in the presence (y axis) and absence (x axis) of anthropogenic variables

A value below the iso-line indicates an increase in potentially suitable area when removing anthropogenic variables. Thin lines represent the 95% credible intervals. The stars represent the area of our study area, Europe. Species symbols, from bottom left, are ibex, wild reindeer, chamois, wolverine, brown bear, Eurasian lynx, wolf, red deer, moose, wild boar, roe deer.

about the scale used to answer their research questions. Although our models of first order habitat selection (distribution range) suggest that anthropogenic factors such as protected area coverage and human disturbance are minor drivers of large ungulate and large carnivore distribution in Europe, results should not be generalized to higher order habitat selection at finer spatial scales (*sensu Johnson 1980*). Indeed, many fine scale studies find that the presence or habitat use of large mammals is mainly negatively affected by their proximity to human infrastructure such as trails, roads, or cities (for red and roe deer see *D'Amico et al., 2016*; *Polfus et al., 2011* for moose, *Lesmerises et al., 2013* for wolves, *Gundersen et al., 2019* for wild reindeer, *May et al., 2006* for wolverines, *Pełksa and Ciach, 2018* for chamois, *Basille et al., 2013* for lynx and *Støen et al., 2015* for bear). Furthermore, studies demonstrate that species are often forced to adapt to the proximity of humans through temporal segregation (e.g., animals become primarily night active, *Gaynor et al., 2018*). As different ecological processes drive distributions at different scales, it is therefore not surprising that results will vary across studies at different scales. For instance, although mountain ungulates forage on steep slopes, human settlements are usually located in the valley bottoms, allowing a vertical coexistence in close proximity. Thus, topographic complexity can provide refuge areas that facilitate human-wildlife proximity (*Richard and Côté, 2016*). The Human Footprint Index is an aggregated metric of human pressure appropriate for analysis of coarse scale data like ours. Finer scale analyses of other datasets would benefit from breaking down its component layers to explore mechanistic relationships between the different aspects of human activity and land use.

The low effect of anthropogenic variables in our models also implies a weak effect of protected areas on large mammal distributions in Europe. (for ungulates see *Linnell et al., 2020*, for carnivores *Chapron et al., 2014*). A main reason is the small size of most European protected areas relative to the spatial requirements of large mammals (for ungulates see *Linnell et al., 2020*, for carnivores *Chapron et al., 2014*). Moreover, although European protected areas have on average a lower human footprint, they are not free of human disturbance. In fact, most European protected areas permit harvesting or culling of large herbivores as well as livestock grazing, extensive agriculture, and forestry (*van Beeck Calkoen et al., 2020*; *Linnell et al., 2015*), and they encourage tourism. It should be noted that these disturbances are not captured by the Human Footprint Index which focuses on infrastructure, implying that the actual disturbance level of protected areas might be higher than the ones used in this analysis. Only in the case of the wolverine and the wild reindeer does

protected area coverage increase the suitable area available because their actual distribution is largely located within protected areas. The mechanistic relationship between the presence of these species and protected area management is however unclear, although for both species human activity and infrastructure has been shown to have negative effects (Nellemann et al., 2000).

This demonstration of the weak effect of human footprint on species distribution compared to the effect of biophysical covariates indicates that most of the large mammals included in our study are flexible enough to adapt to the dramatic anthropogenic impacts which have occurred within their bioclimatic envelope in the European landscape during recent centuries. This is reflected by the overall generalist behavior of these species. For instance, moose seem to adapt to road presence and associated forage in their proximity (Eldegard et al., 2012), whereas agricultural landscapes help roe deer to supplement their diet (Abbas et al., 2011).

Limitation of the study

Similar to other large-scale studies such as Belote et al. (2020) or Pacifici et al. (2020), our analysis is also limited to distributional data whose quality is highly variable and coarse, and we do not analyze effects on density, behavior or demography. Therefore, while our results document the ability of populations of ungulates and carnivores to persist and use areas in the general proximity to areas of high human footprint, this does not mean these species are not influenced by humans in other ways and at finer spatiotemporal scales. Another challenge is the lack of historical distribution data which makes inferences about causal relationships between human activities and land uses with changes in distributions and population of ungulates populations. Although some attempts to reconstruct large mammals' historical distribution are made, they generally rely on current distribution (Belote et al., 2020).

Conclusion

Our results contribute to advancing the science of human-wildlife coexistence in the heavily modified landscapes that are typical of the Anthropocene. Although several papers rightly point out that large mammals are threatened by human impacts in many parts of the world (Ripple et al., 2014, 2015) we argue that the European experience demonstrates that coexistence between humans and wild large mammals at broad scales, and continental scale recovery, are both possible. We suggest that it is impossible for nature conservation authorities to rely on a land-sparing policy for large mammals because protected areas large enough to support viable populations of these spaces demanding species don't exist. Ultimately, the challenge of coexistence may not be about whether species are able to cope with human modification to the landscape but whether humans are willing to share their landscape and host wildlife in their backyards (Title and Bemmels, 2018). Europe has multiple layers of formal and informal institutions at continental, national and local scales that effectively manage wildlife and human-wildlife interactions and which appear to have an instrumental role in facilitating this coexistence (Linnell and Kaltenborn, 2019). Overall, the results permit cautious optimism concerning the possibility for wildlife conservation in the Anthropocene.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Material availability
 - Data and code availability
- METHOD DETAILS
 - Distribution data
 - Explanatory variables
- QUANTIFICATION AND STATISTICAL ANALYSIS
 - Model specification
 - Evaluation of variables' importance for species' distribution
 - Quantifying the effect of anthropogenic variables on the size of the species' suitable habitat

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2021.103083>.

ACKNOWLEDGMENTS

This study was funded by the Research Council of Norway (grant number 251112) and a PhD scholarship to BC from the Norwegian University of Science and Technology. We thank the two anonymous reviewers for their useful comments and suggestions that have improved this manuscript.

AUTHOR CONTRIBUTIONS

BC and JDCL conceptualized the idea of the manuscript. BC analyzed the data, drafted and revised the manuscript. BM provided technical support. All authors contributed to the writing and the improvement of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 4, 2021

Revised: August 11, 2021

Accepted: August 30, 2021

Published: September 24, 2021

REFERENCES

- Abbas, F., Morellet, N., Hewison, A.M., Merlet, J., Cargnelutti, B., Lourtet, B., Angibault, J.M., Daufresne, T., Aulagnier, S., and Verheyden, H. (2011). Landscape fragmentation generates spatial variation of diet composition and quality in a generalist herbivore. *Oecologia* 167, 401–411.
- Alexander, J.S., Gopalaswamy, A.M., Shi, K., Hughes, J., and Riordan, P. (2016). Patterns of snow leopard site use in an increasingly human-dominated landscape. *PLoS One* 11 (5), e0155309.
- M. Apollonio, R. Andersen, and R. Putman, eds. (2010). *European Ungulates and Their Management in the 21st Century* (Cambridge University Press).
- Araújo, M.B., and Peterson, A.T. (2012). Uses and misuses of bioclimatic envelope modeling. *Ecology* 93, 1527–1539.
- Azen, R., and Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychol. Methods* 8, 129.
- Basille, M., Van Moorter, B., Herfindal, I., Martin, J., Linnell, J.D., Odden, J., Andersen, R., and Gaillard, J.M. (2013). Selecting habitat to survive: the impact of road density on survival in a large carnivore. *PLoS ONE* 8, e65493.
- Beguín, J., Martino, S., Rue, H., and Cumming, S.G. (2012). Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. *Methods Ecol. Evol.* 3, 921–929.
- Belote, R.T., Faurby, S., Brennan, A., Carter, N.H., Dietz, M.S., Hahn, B., McShea, W.J., and Gage, J. (2020). Mammal species composition reveals new insights into Earth's remaining wilderness. *Front. Ecol. Environ.* 18, 376–383.
- Carter, N.H., and Linnell, J.D. (2016). Co-adaptation is key to coexisting with large carnivores. *Trends Ecol. Evol.* 31, 575–578.
- Carter, N.H., Shrestha, B.K., Karki, J.B., Pradhan, N.M.B., and Liu, J. (2012). Coexistence between wildlife and humans at fine spatial scales. *Proc. Natl. Acad. Sci. U S A* 109, 15360–15365.
- Chapron, G., Kaczensky, P., Linnell, J.D., Von Arx, M., Huber, D., Andrén, H., López-Bao, J.V., Adamec, M., Álvares, F., Anders, O., and Balciuskas, L. (2014). Recovery of large carnivores in Europe's modern human-dominated landscapes. *Science* 346, 1517–1519.
- D'Amico, M., Périquet, S., Román, J., and Revilla, E. (2016). Road avoidance responses determine the impact of heterogeneous road networks at a regional scale. *J. Appl. Ecol.* 53, 181–190.
- Dietz, A.J., Kuenzer, C., and Dech, S. (2015). Global SnowPack: a new set of snow cover parameters for studying status and dynamics of the planetary snow cover extent. *Remote Sens. Lett.* 6, 844–853.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., and Münkemüller, T. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46.
- Eldegard, K., Lyngved, J.T., and Hjeljord, O. (2012). Coping in a human-dominated landscape: trade-off between foraging and keeping away from roads by moose (*Alces alces*). *Eur. J. Wildl. Res.* 58, 969–979.
- Gaston, K.J., Jackson, S.F., Nagy, A., Cantú-Salazar, L., and Johnson, M. (2008). Protected areas in Europe: principle and practice. *Ann. NY Acad. Sci.* 1134, 97–119.
- Gaynor, K.M., Hojnowski, C.E., Carter, N.H., and Brashares, J.S. (2018). The influence of human disturbance on wildlife nocturnality. *Science* 360, 1232–1235.
- Guisan, A., and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009.
- Gundersen, V., Vistad, O.I., Panzacchi, M., Strand, O., and van Moorter, B. (2019). Large-scale segregation of tourists and wild reindeer in three Norwegian national parks: management implications. *Tourism Manag.* 75, 22–33.
- Johnson, D.H. (1980). The comparison of usage and availability measurements for evaluating resource preference. *Ecology* 61, 65–71.
- Leblond, M., Dussault, C., and Ouellet, J.P. (2010). What drives fine-scale movements of large herbivores? A case study using moose. *Ecography* 33, 1102–1112.
- Leroux, S.J., Krawchuk, M.A., Schmiegelow, F., Cumming, S.G., Lisgo, K., Anderson, L.G., and Petkova, M. (2010). Global protected areas and IUCN designations: do the categories match the conditions? *Biol. Conserv.* 143, 609–616.
- Lesmerises, F., Dussault, C., and St-Laurent, M.H. (2013). Major roadwork impacts the space use behaviour of gray wolf. *Landscape Urban Plan* 112, 18–25.
- Lindgren, F., and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* 63, 1–25.
- Linnell, J.D.C., and Kaltenborn, B. (2019). Institutions for achieving human-wildlife coexistence. In *The Case of Large Herbivores and Large Carnivores in Europe. Human Wildlife Interactions: Turning Conflict into Coexistence*, B. Frank, J.A. Glikman, and S. Marchini, eds. (Cambridge University Press), pp. 288–310.
- Linnell, J.D.C., and Zachos, F.E. (2010). Status and Distribution Patterns of European Ungulates: Genetics, Population History and Conservation. *Ungulate Management in Europe: Problems and Practices*, pp. 12–53.
- Linnell, J.D., Kaczensky, P., Wotschikowsky, U., Lescureux, N., and Boitani, L. (2015). Framing the relationship between people and nature in the context of European conservation. *Conserv. Biol.* 29, 978–985.

- Linnell, J.D., Cretois, B., Nilsen, E.B., Rolandsen, C.M., Solberg, E.J., Veiberg, V., Kaczensky, P., Van Moorter, B., Panzacchi, M., Rauset, G.R., and Kaltenborn, B. (2020). The challenges and opportunities of coexisting with wild ungulates in the human-dominated landscapes of Europe's Anthropocene. *Biol. Conserv.* *244*, 108500.
- May, R., Landa, A., van Dijk, J., Linnell, J.D., and Andersen, R. (2006). Impact of infrastructure on habitat selection of wolverines *Gulo gulo*. *Wildl. Biol.* *12*, 285–295.
- Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods Ecol. Evol.* *4*, 133–142.
- Nellemann, C., Jordhøy, P., Støen, O.G., and Strand, O. (2000). Cumulative impacts of tourist resorts on wild reindeer (*Rangifer tarandus tarandus*) during winter. *Arctic*, 9–17.
- Nellemann, C., Støen, O.G., Kindberg, J., Swenson, J.E., Vistnes, I., Ericsson, G., Katajisto, J., Kaltenborn, B.P., Martin, J., and Ordiz, A. (2007). Terrain use by an expanding brown bear population in relation to age, recreational resorts and human settlements. *Biol. Conserv.* *138*, 157–165.
- Pacifici, M., Rondinini, C., Rhodes, J.R., Burbidge, A.A., Cristiano, A., Watson, J.E., Woinarski, J.C., and Di Marco, M. (2020). Global correlates of range contractions and expansions in terrestrial mammals. *Nat. Commun.* *11*, 1–9.
- Pęksa, Ł., and Giach, M. (2018). Daytime activity budget of an alpine ungulate (*Tatra chamois Rupicapra rupicapra tatraica*): influence of herd size, sex, weather and human disturbance. *Mammal Res.* *63*, 443–453.
- Phalan, B., Onial, M., Balmford, A., and Green, R.E. (2011). Reconciling food production and biodiversity conservation: land sharing and land sparing compared. *Science* *333*, 1289–1291.
- Polfus, J.L., Hebblewhite, M., and Heinemeyer, K. (2011). Identifying indirect habitat loss and avoidance of human infrastructure by northern mountain woodland caribou. *Biol. Conserv.* *144*, 2637–2646.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria2020. <https://www.R-project.org/>.
- Redpath, S.M., Young, J., Evelyn, A., Adams, W.M., Sutherland, W.J., Whitehouse, A., Amar, A., Lambert, R.A., Linnell, J.D., Watt, A., and Gutierrez, R.J. (2013). Understanding and managing conservation conflicts. *Trends Ecol. Evol.* *28*, 100–109.
- Richard, J.H., and Côté, S.D. (2016). Space use analyses suggest avoidance of a ski area by mountain goats. *J. Wildl. Manag.* *80*, 387–395.
- Ripple, W.J., Estes, J.A., Beschta, R.L., Wilmers, C.C., Ritchie, E.G., Hebblewhite, M., Berger, J., Elmhagen, B., Letnic, M., Nelson, M.P., and Schmitz, O.J. (2014). Status and ecological effects of the world's largest carnivores. *Science* *343*, 1241484.
- Ripple, W.J., Newsome, T.M., Wolf, C., Dirzo, R., Everatt, K.T., Galetti, M., Hayward, M.W., Kerley, G.I., Levi, T., Lindsey, P.A., and Macdonald, D.W. (2015). Collapse of the world's largest herbivores. *Sci. Adv.* *1*, e1400103.
- Støen, O.G., Ordiz, A., Evans, A.L., Laske, T.G., Kindberg, J., Frøbert, O., Swenson, J.E., and Arnemo, J.M. (2015). Physiological evidence for a human-induced landscape of fear in brown bears (*Ursus arctos*). *Physiol. Behav.* *152*, 244–248.
- Svenning, J.C., Fløjgaard, C., and Baselga, A. (2011). Climate, history and neutrality as drivers of mammal beta diversity in Europe: insights from multiscale deconstruction. *J. Anim. Ecol.* *80*, 393–402.
- Tattersall, G.J., Sinclair, B.J., Withers, P.C., Fields, P.A., Seebacher, F., Cooper, C.E., and Maloney, S.K. (2012). Coping with thermal challenges: physiological adaptations to environmental temperatures. *Comprehensive physiology* *2*, 2151–2202.
- Title, P.O., and Bemmels, J.B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* *41*, 291–307.
- Tucker, M.A., Böhning-Gaese, K., Fagan, W., Fryxell, J., Van Moorter, B., Albers, S., Ali, A., Allen, A., Attias, N., Avgar, T., et al. (2018). Moving in the anthropocene: global reductions in terrestrial mammalian movements. *Science* *359*, 466–469.
- Tucker, M.A., Santini, L., Carbone, C., and Mueller, T. (2021). Mammal population densities at a global scale are higher in human-modified areas. *Ecography* *44*, 1–13.
- van Beeck Calkoen, S.T.S., Mühlbauer, L., Andrén, H., Apollonio, M., Balciuskas, L., Belotti, E., Carranza, J., Cottam, J., Filli, F., Gatiso, T.T., et al. (2020). Ungulate management in European national parks: Why a more integrated European policy is needed. *J. Environ. Manag.* *260*.
- Venter, O., Sanderson, E.W., Magrath, A., Allan, J.R., Beher, J., Jones, K.R., Possingham, H.P., Laurance, W.F., Wood, P., Fekete, B.M., and Levy, M.A. (2016). Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nat. Commun.* *7*, 1–11.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Wild ungulates distribution data	Linnell et al. 2020	https://doi.org/10.1016/j.biocon.2020.108500
Large carnivores distribution data	Chapron et al. 2014	https://doi.org/10.1126/science.1257553
Terrain Ruggedness Index	Title & Bemmels 2018	https://doi.org/10.1111/ecog.02880
Potential Evapotranspiration	Title & Bemmels 2018	https://doi.org/10.1111/ecog.02880
Snow Cover Duration	Dietz et al. 2015	https://doi.org/10.1080/2150704X.2015.1084551
Human Footprint Index	Venter et al. 2016	https://doi.org/10.1038/ncomms12558
Protected Area	World Database on Protected Areas	https://protectedplanet.net/
Software and algorithms		
R Statistical Software	R Core Team, 2020	https://www.r-project.org/
ArcGIS Pro	ESRI	https://www.esri.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Benjamin Cretois (benjamin.cretois@nina.no).

Material availability

This study did not generate new materials.

Data and code availability

- The dataset and scripts used to conduct all analyses presented in this manuscript are fully available and has been deposited on Open Science Framework (<https://doi.org/10.17605/OSF.IO/XV8NH>).
- Data concerning wild ungulates distribution are fully available and has been extracted from Linnell et al. (2020) (<https://doi.org/10.1016/j.biocon.2020.108500>).
- Data on large carnivores' distribution are fully available and has been extracted from Chapron et al. (2014) (<https://doi.org/10.1126/science.1257553>).

METHOD DETAILS

Distribution data

In this paper we focus on wild large mammals which are native to Europe and whose distribution is not intensively managed (i.e. doesn't depends on intensive interventions such as the European bison *Bison bonasus*, Linnell et al., 2020). This includes nine large ungulates: roe deer (*Capreolus capreolus*), red deer (*Cervus elaphus*), moose (*Alces alces*), wild reindeer (*Rangifer tarandus*), Alpine chamois (*Rupicapra rupicapra*), Pyrenean chamois (*Rupicapra pyrenaica*), Alpine ibex (*Capra ibex*), Iberian ibex (*Capra pyrenaica*) and wild boar (*Sus scrofa*). We extracted the distribution data provided in Linnell et al. (2020) for all these species. Because the distribution of the mountain ungulates is restricted and because several species belong to the same genus and have similar ecological requirements, we merged the distribution of the Iberian and Alpine ibex, and the distribution of the Alpine and Pyrenean chamois creating *Capra* spp. and *Rupicapra* spp. distributions, respectively. Data come from many sources spread across a period from c. 1990 to 2019. Distribution data for the four species of large carnivore present in Europe; wolves (*Canis lupus*), Eurasian lynx (*Lynx lynx*), brown bears (*Ursus arctos*) and wolverines (*Gulo gulo*) were derived from published data (Chapron et al., 2014), and are derived from the period 2008–2011. Distribution data for all species had a spatial resolution of 10 km × 10 km and take the value 0 if the species is absent and 1 if

the species is present. As the underlying distribution data is of widely varying quality and resolution, 10 km × 10 km is the finest resolution we would advocate for large-scale studies as it erases uncertainty related to the location of a species observation and is computationally manageable. In addition, the 10 km × 10 km resolution allows the results of our analysis to be comparable to other large-scale studies such as [Tucker et al. \(2018\)](#) or [Chapron et al. \(2014\)](#). We included data on both herbivore and carnivore distribution from 31 countries, consisting of all EU countries (excluding Cyprus and Malta), plus Norway, Switzerland, Serbia, Albania, Northern Macedonia and the United Kingdom.

Explanatory variables

We collected three abiotic covariates, two related to climate and one to terrain relief that are thought to be influential biophysical drivers of species distribution ([Araújo and Peterson, 2012](#)). In addition, we included the two anthropogenic covariates human footprint (HF) and protected area. The biophysical drivers represent potential large-scale and long-term constraints on species' potential distributions (i.e. bioclimatic envelopes) operating through physiological tolerance, rather than fine-scaled and temporally variable environmental factors that typically represent vegetation or habitat patch quality.

Terrain Ruggedness Index and the Potential Evapotranspiration for the Warmest Quarter (PETWQ) were acquired from the ENVIREM dataset ([Title & Bemmels 2018](#)) at a spatial resolution of 2.5 arc minutes (i.e. about 3 km × 3 km at 50°N). The mean snow cover duration (SCD) was derived from the Global SnowPack, a 14-year average available at a 0.25 km × 0.25 km resolution (from 2000 to 2014, [Dietz et al., 2015](#)). We used PETWQ and SCD as proxies for summer and winter severity respectively. Snow cover is widely viewed as being a major limiting factor for species latitudinal and altitude distributions as it correlates with cold winter temperatures, and the physical inhibition of animal movement and access to forage ([Leblond et al., 2010](#)). Evapotranspiration serves as a proxy for hot, dry, unproductive summer conditions that also limit species through thermal stress, and poor forage conditions ([Tattersall et al., 2012](#)). Terrain ruggedness is widely viewed as being an important escape terrain for species (especially ibex and chamois, and potentially wild reindeer) to avoid disturbance and predation ([Nellemann et al., 2007](#)). These three biophysical variables were all obtained as raster data.

As a measure of human disturbance, we chose the Human Footprint Index (HFI version 2009, [Venter et al., 2016](#)). Ranging from 0 to 50, the HFI is a composite raster built from multiple variables related to human disturbance (e.g. the extent of built environment, cropland, pasture lands, human population density, nighttime lights, railways, roads and navigable waterways; [Venter et al., 2016](#)). The HFI has been recently used in multiple continent-wide comparisons of mammal movement rates (e.g. [Tucker et al., 2018, 2021](#)).

Finally, we obtained the protected area coverage from the World Database on Protected Areas: <https://protectedplanet.net/>. We included all protected areas whose status was listed as either "designated", "not reported", "not applicable" or "assigned". Data was available as vector data and was rasterized at a resolution of 1km² using ArcGIS Pro for ease of computation. We finally used aggregation to sum the total number of 1 km × 1 km pixels of protected area within each 10 km × 10 km grid cell (i.e. the grid cell value for protected area varied from 0 for a grid cell containing no protected area to 100 for a grid cell entirely covered by a protected area). Although European protected areas are almost never wilderness areas ([van Beeck Calkoen et al., 2020](#); [Linnell et al., 2015](#)) they are expected to be associated with greater restrictions on human activities that could potentially better limit human impacts on wildlife, and less intensive forms of land use. However, we did not separate the different IUCN categories as previous studies show that there is little difference in human footprint between categories ([Leroux et al., 2010](#)) and there is a high degree of variation between European countries in how they manage protected areas of different IUCN categories ([Gaston et al., 2008](#)).

We assessed the extent of collinearity between the covariates. Winter and summer severity were negatively related ($r = -0.71$), as both display strong coastal-inland and north-south gradients. However, we opted to include both as they reflect different mechanisms for species' ecology. Following [Dormann et al., 2013](#) we made sure to carefully interpret the results of these two variables by interpreting the combined effects of all environmental variables (More detailed explanations in Annexes). Other covariates were not significantly correlated with each other ($r < 0.70$; [Table S1](#) in Annexes). We aggregated all the explanatory variables to the same 10 km × 10 km grid cell resolution.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model specification

Because the residuals of the non-spatial models were strongly spatially correlated, we fitted an intrinsic conditional autoregression (iCAR) model using hierarchical Bayesian models for each of the 13 species. The probability of presence (π) of a given species in a given grid cell was calculated using a Bernoulli distribution and the following model:

$$y_i \sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = \alpha_i + x_i\beta + u_i$$

where x_i is the vector of covariates for cell i , β the vector of parameters to be estimated and u_i the spatially correlated random effect whose prior is defined as:

$$u_i | u_k \sim \text{normal}\left(\frac{\sum_{i \neq k} w_{i,k} u_k}{n_i}, \frac{\sigma_u^2}{n_i}\right)$$

where $w_{i,k} = 1$ if grid cells i and k are neighbors and 0 otherwise. n_i is the total number of neighbors of grid cell i . We define two cells as being neighbors if they directly share a single boundary point. All models assume a vague prior for the regression parameters $\beta \sim \text{normal}(\text{Mean} = 0, \text{SD} = 1000)$ and we used a penalized complexity prior on the spatial effect to avoid risks of overfitting.

As we expect species to have an optimal niche for environmental variables, we included linear and quadratic terms for winter and summer severity and ruggedness (Svenning et al., 2011). We also included linear and quadratic terms for human footprint as we suspected certain species to have an optimal niche in the moderate human disturbance level. We only included a linear effect for protected area coverage as we only expected a linear response.

To fit the spatial models, we used the Integrated Nested Laplace Approximation (INLA) approach with the package R-INLA (Lindgren and Rue, 2015). INLA is a faster alternative to Markov Chain Monte Carlo approaches and yields similar, if not identical, results (Beguin et al., 2012). We standardized the covariates to enable direct comparison between the regression coefficients. All analyses were conducted in R 3.6.1.

We validated the models by plotting residual values against covariates for each model. We also plotted the leave-one out cross validation scores (conditional predictive ordinate CPO in our case) to estimate model fit.

Evaluation of variables' importance for species' distribution

We estimated the relative importance of both environmental and anthropogenic variables using dominance analysis (Azen and Budescu, 2003), which is a procedure to quantify the importance of a random variable through examination of the R^2 values (or similar metrics) for all possible subset models of a predefined full model. In a dominance analysis, the higher the dominance score the more useful is the random variable in predicting the response variable. Because the number of models required to estimate the importance of a single random variable grows exponentially with the total number of random variables, we did not quantify the importance of each single variable, but rather the importance of the combined effect of summer and winter severity and ruggedness ("environmental variables") and human footprint and protected area coverage ("anthropogenic variables"). Thus, we fitted 3 models for each of the 11 species: a full model containing all variables, a model containing only the environmental variables and a model containing only the anthropogenic variables. For all models we computed the R^2_{glm} , a modified version of the classic R^2 which is suitable for mixed models (Nakagawa and Schielzeth, 2013). We sampled 1,000 values from the posterior distribution of the model parameters and bootstrapped the R^2_{glm} 1,000 times. We finally rescaled the dominance score for it to range from 0 to 100%.

Quantifying the effect of anthropogenic variables on the size of the species' suitable habitat

To further assess the results of the dominance analysis we assessed the relative extent to which anthropogenic variables influence the realized distribution of the studied large mammals we quantified the geographic representation of the suitable habitat for each species (i.e. the potential suitable area available due to environmental predictors only, Guisan and Thuiller, 2005). We predicted the probability of a species' occurrence within a grid cell both when anthropogenic variables were set at their minimum

value (i.e. we simulated a landscape free of all human influence: no human footprint and no protected areas) and when anthropogenic variables are set to their observed values. We summed these predicted occurrences across Europe to estimate the expected number of occupied cells (i.e. the size of a species' suitable area in Europe). A sum of predictions in a human-free landscape higher than a sum of prediction for the full model implies that the species increase its range in absence of human influence in the landscape. We sampled 1,000 values from the posterior distribution of the model parameters and bootstrapped the niche area 10,000 times.

Paper V

Native range estimates for red-listed vascular plants

Borgelt, J., Sicacha-Parada, J., Skarpaas, O., & Verones, F. (2022) published in *Scientific Data*



OPEN

DATA DESCRIPTOR

Native range estimates for red-listed vascular plants

Jan Borgelt¹✉, Jorge Sicacha-Parada², Olav Skarpaas³ & Francesca Veronesi¹

Besides being central for understanding both global biodiversity patterns and associated anthropogenic impacts, species range maps are currently only available for a small subset of global biodiversity. Here, we provide a set of assembled spatial data for terrestrial vascular plants listed at the global IUCN red list. The dataset consists of pre-defined native regions for 47,675 species, density of available native occurrence records for 30,906 species, and standardized, large-scale Maxent predictions for 27,208 species, highlighting environmentally suitable areas within species' native regions. The data was generated in an automated approach consisting of data scraping and filtering, variable selection, model calibration and model selection. Generated Maxent predictions were validated by comparing a subset to available expert-drawn range maps from IUCN ($n = 4,257$), as well as by qualitatively inspecting predictions for randomly selected species. We expect this data to serve as a substitute whenever expert-drawn species range maps are not available for conducting large-scale analyses on biodiversity patterns and associated anthropogenic impacts.

Background & Summary

Life on Earth is essential to human society as it forms the foundation of present welfare¹. The growing human population, modern lifestyles and associated pressures on the planet have already resulted in a significant loss of natural habitat and are threatening biodiversity^{2–6}. Different initiatives promote the protection of biodiversity and aim to halt its loss, such as the UN Sustainable Development Goals⁷, the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services⁸ and the International Union for the Conservation of Nature (IUCN). Different decision-support tools can contribute to this by assessing environmental performances of products, strategies and policies^{2,9–11}. For the development of such tools, but also for the implementation of global conservation strategies and policies itself, spatial data, e.g. in the form of distribution maps of individual species¹², are crucial. However, besides many species remaining undiscovered or undescribed, we still lack spatial information for most of the ones we know¹³. Consequently, comprehensive and ready-to-use datasets for large-scale analyses are only available for a few vertebrate groups^{14–16}. This is concerning, as global conservation strategies and biodiversity impact assessments are limited to these groups, while some hyperdiverse species groups, such as plants, are often not considered^{17,18}.

Here, we provide spatial distribution data for a large fraction of red-listed terrestrial vascular plant species at different levels of spatial detail (Fig. 1), i.e. native regions ($n = 47,675$), occurrence records ($n = 30,906$) and modelled range estimates (i.e. a predicted relative environmental suitability¹⁹ within native regions; $n = 27,208$). The workflow included data scraping and filtering, as well as variable selection, model calibration and model selection, aiming for best practice^{20–22} but within the constraints of data limitations and computational feasibility at this scale. Species-specific native regions were retrieved from a scheme specifically developed to challenge the lack of distributional knowledge for plant species²³. Available native occurrence records were retrieved from the Global Biodiversity Information Facility (GBIF)²⁴ and subsequently filtered. Range estimates were generated using maximum entropy modelling^{19,25–27}, and show where environmentally suitable conditions exist within each species' native regions (Fig. 2a–d).

The underlying occurrence data is known to be highly spatiotemporally aggregated and variable across administrative borders for some species^{28–31}. We aimed at counteracting a potential sampling bias by using three differently treated occurrence data types (i.e. different degree of spatial filtering: no filter, presence cells, thinned presence cells), and by dividing occurrence data in equally-sized bins during model calibration³². Up

¹Industrial Ecology Programme, Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. ²Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. ³Natural History Museum, University of Oslo, Oslo, Norway. ✉e-mail: jan.borgelt@ntnu.no

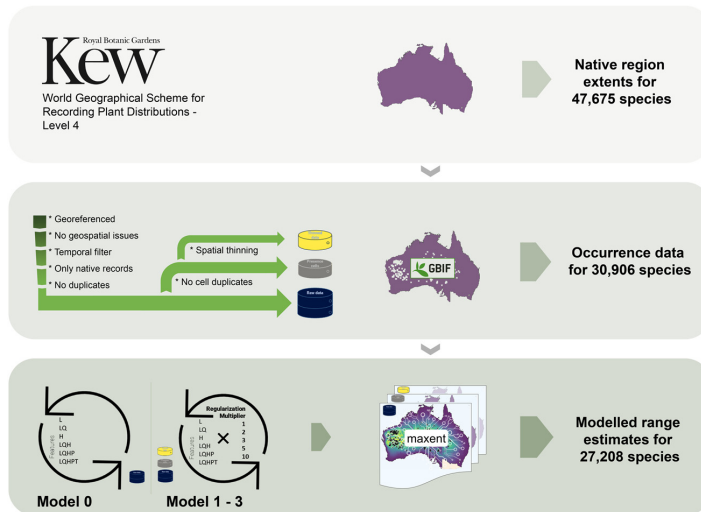


Fig. 1 Schematic summary of the dataset. Top: Native region extents were retrieved from Kew’s Plants of the World online. Middle: Occurrence data was retrieved from the Global Biodiversity Information Facility (GBIF)²⁴ and filtered into three different occurrence data types: raw data (blue), presence cells (grey) and thinned data (yellow). Bottom: The different occurrence data types were used in Maxent models to predict relative environmental suitability indices within native regions (i.e. range estimates). Differences between Model 0 and Model 1 to 3. Model 0 was trained to support variable selection using raw data in k-fold cross validated Maxent models (one model for each combination of feature classes, i.e. linear (L), quadratic (Q), hinge (H), product (P) and threshold (T)). The selected variables and each of the three occurrence data types were used to train a set of separate k-fold cross validated Maxent models (one model for each possible combination of feature classes, regularization multipliers and occurrence data type). The overall best performing model was selected for each species based on performance metrics.

to 96 different models were fitted per species to find optimal variables, model settings and data type. The best prediction was selected for each species based on common performance metrics (i.e. AUC and AUC_{PR}).

However, some predictions will undoubtedly remain flawed by underlying biases. Based on comparisons to expert-drawn range maps available from IUCN ($n = 4,257$) and qualitative inspection of predictions for randomly selected species, we expect this to mainly influence widespread and common species, and hence, only affect the smallest proportion of global biodiversity³³. In addition, the species most vital for assessing anthropogenic impacts or for defining conservation priorities, are more likely to be small-ranged and endemic. Although validating each prediction was not feasible, we found most individually inspected predictions to either offer an improvement compared to elsewhere available data or an acceptable substitute, although at a coarser spatial resolution and less detailed.

We want to stress that the presented dataset is generated for the purpose of global spatial screening studies and for building a basis for future, global biodiversity impact assessment models. In concert with powerful, species-specific trait and conservation-related databases, the provided data can benefit future work, such as assessing global extinction probabilities³⁴, effects of terrestrial acidification³⁵, drivers of invasion success³⁶, progress towards reaching global conservation goals³⁷ and act as pre-assessment prior to expert-based range map generation and red list assessments^{38–41}. With a continuously increasing availability of species occurrence records, the presented dataset can be updated frequently to illustrate the state of knowledge at any time. With more data becoming available, precision is likely to increase in the future.

Methods

Taxonomic scope. A species list containing all terrestrial vascular plants ($n = 52,372$) of the global IUCN red list was retrieved from IUCN in April 2021, IUCN version 2021-1¹⁶. We retrieved each species’ accepted name from Plants of the World Online (POWO)⁴² to facilitate communication to various data portals using the package *taxize*⁴³ in R⁴⁴. Plant family, order and class were retrieved from the Integrated Taxonomic Information System⁴⁵ using the package *taxize*⁴³ in R. Only species outside the IUCN threat categories “Extinct” and “Extinct in the Wild” were kept, and all species considered as subspecies or varieties according to POWO removed. We attempted to assemble spatial data for each of the remaining 48,144 species.

Native regions. Species-specific native regions (Fig. 1) were retrieved from POWO using a customized web-scraper function (see section *Code Availability*) and the packages *taxize*⁴³ and *rvest*⁴⁶ in R. The data follows the World Geographical Scheme for Recording Plant Distributions (WGSRPD)²³ and includes a continental,

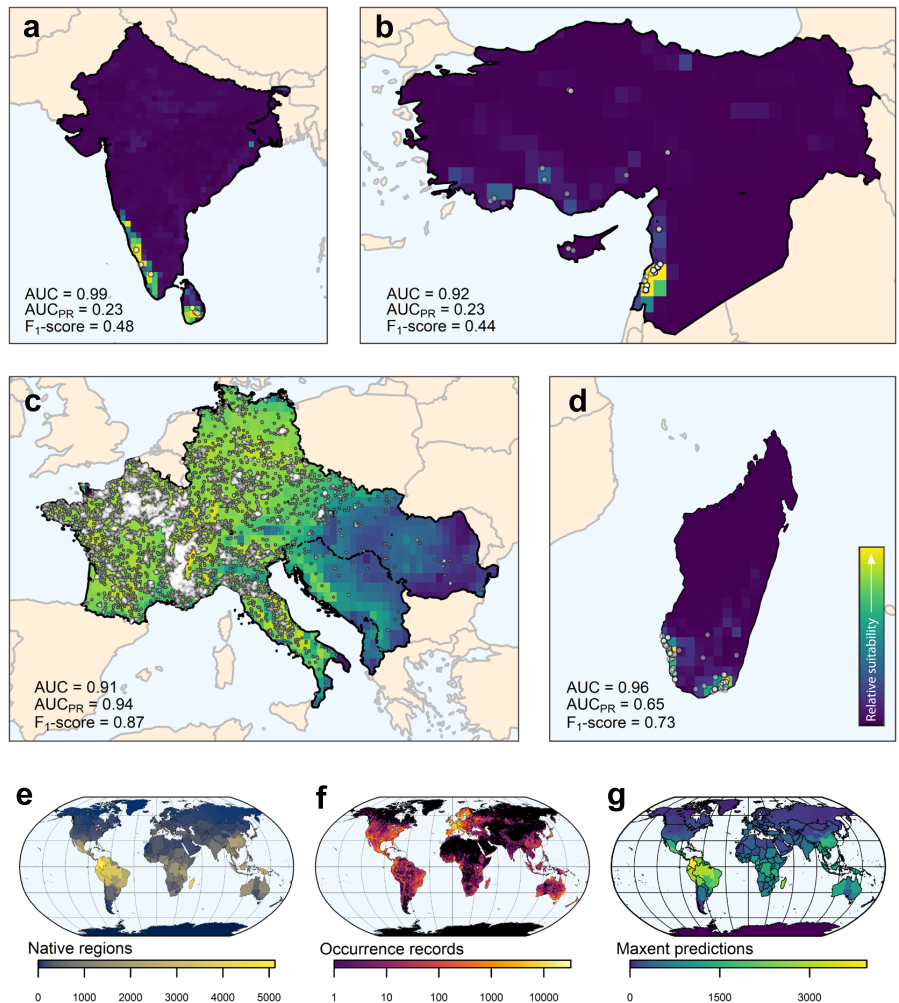


Fig. 2 Data examples for randomly selected species and spatial coverage of the dataset. Best performing Maxent prediction, highlighting environmentally suitable conditions within the species native regions (i.e. modelling extent) along retrieved occurrence records (white points) for (a) *Amomum pterocarpum*, (b) *Cedrus libani*, (c) *Laburnum anagyroides*, (d) *Megistostegium nodulosum*. Performance of the shown predictions indicated by maximum F₁-score and the area under the receiver operating characteristics curve for true vs. false positive rate (AUC) and recall vs. precision (AUC_{PR}). Bottom: number of (e) retrieved native regions, (f) retrieved occurrence records, and (g) generated Maxent predictions across the globe.

country and regional level. Retrieved WGSRPD-*regions* were matched to its corresponding shapefile at level 4, available from the Biodiversity Information Standards GitHub repository⁴⁷ and rasterized at 30 arc minutes spatial resolution (approximately 56 km at the equator).

Occurrence records. For species with given native extents in POWO, the maximum number of most recent occurrence points (i.e. 100,000) per native WGSRPD-*country* was retrieved from the GBIF application programming interface (API) using the package *rgbif*⁴⁸ in R (the equivalent full dataset⁴⁹ is available at <https://doi.org/10.15468/dl.uvd56q>). The considered environmental variables have changed tremendously in the past decades^{50,51} and only cover a limited period of time, i.e. the years 1979–2013 and 2015 respectively (see section *Environmental data*). Therefore, only records between the years 2000 and 2020 were considered to temporally align occurrence data to both sets of environmental variables as best as possible. If less than 25 records were available for a given species after the year 2000, no temporal filter was set to maximize data retrieval. GBIF records without specified coordinates and with flagged geospatial issues⁴⁸ were not considered. As such, we expect

inaccurate coordinate notations as well as records of specimens preserved in museums or other biodiversity facilities to be typically detected. Only points inside reported native WGSRPD-*regions* were kept and duplicated records were removed (hereafter: raw data). The number of raw data records was counted per cell (30 arc min.) using the package *raster*⁵² in R.

Maxent predictions. We generated spatial predictions within species' native WGSRPD-*regions* at 30 arc min. resolution (approximately 56 km at the equator) using maximum entropy modelling (Maxent)^{19,26,27}, for all species with at least 5 raw data records^{53,54} that were distributed across at least 3 cells, and a native region extent of at least 9 cells. Although an arbitrary threshold, we attempted to allocate computational resources to more meaningful predictions, modelled across larger extents. Maxent is a probability density estimation approach widely used for predicting species distributions based on presence-only data⁵⁵. Background information, required to fit response curves⁵⁶, was collected from each cell within each species' native regions⁵⁷. For generating models we utilized a high-performance computing infrastructure⁵⁸ allowing for parallel computations using the Maxent software²⁵ via R packages *dismo*⁵⁹ and *ENMeval*⁶⁰.

Environmental data. We downloaded all CHELSA bioclimatic variables^{61,62} ($n = 19$, see Table 1 for full list) in 30 arc seconds resolution and aggregated, for computational efficiency, to the chosen modelling resolution (30 arc min.) by averaging. CHELSA bioclimatic variables are a set of modelled, biologically relevant, climatic variables based on data collected during the years 1979–2013⁶¹. In addition, fractions for different natural land cover types, including different types and mosaics of forest, shrubland, grassland and sparse vegetation, ($n = 17$, see Table 1 for full list) were calculated based on the European Space Agency's land cover product for the year 2015 in 300 m resolution⁶³. Each land cover class was transformed into a binary raster depicting presence (=1) and absence (=0) of the land cover type. The binary raster was then aggregated to modelling resolution by averaging, resulting in one raster for each land cover class, representing the proportion of land covered by that class per pixel.

Occurrence data types. For some species, several raw data records can be in the same cell at the given spatial resolution (30 arc min.). Although pseudo-replication can inflate model performance (here: during model calibration) and, hence, increases the risk of overfitting, we argue that these occurrence points still contain valid information if they are discrete observations and therefore kept this data. However, we henceforth applied two filters to counteract potential spatial biases, as well as pseudo-replication (Fig. 1). We removed all cell-duplicates from the raw data (hereafter: presence cells), and we applied spatial thinning with a minimum distance of two cells on the presence cells (hereafter: thinned data). Occurrence data was spatially filtered using the R package *spThin*⁶⁴.

Model training. A set of Maxent models was fitted for each species using the differently treated occurrence data types. All models were calibrated using k-fold cross validation. The employed occurrence data was partitioned into training and testing bins. For species with only few data points ($n < 25$), we used k - 1 Jackknife partitioning ($k = n$)⁵⁴. For species with more data points ($n \geq 25$) we used block partitioning ($k = 4$) to account for spatial autocorrelation of occurrence points in larger datasets³². This partitioning splits the occurrence data at a longitudinal and latitudinal line, resulting in approximately equally sized bins⁶⁰.

An initial model (Fig. 1; Model 0) was trained to support the selection of uncorrelated environmental variables using the raw data and all environmental variables ($n = 36$) for each species. Separate models, one for each possible combination out of all included feature classes (i.e. environmental variables and transformations thereof), were trained. We included linear (l), quadratic (q), product (p), hinge (h) and threshold (t) transformations, resulting in 6 possible combinations (i.e. l, lq, h, lqh, lqhp, and lqhpt). The best performing model was selected based on the corrected Akaike information criterion (AICc)^{65–67}. However, if no model performed best in terms of AICc, or if this metric was unavailable for 50% of fitted models, the average testing area under the receiver operating characteristics curve (AUC; see section *Technical Validation*) during model calibration was used instead. Permutation importance was retrieved for all variables in Model 0. Correlated variables were identified using Spearman's rank correlation coefficient (ρ) and defined as $\rho \geq |\pm 0.7|$. In any set of correlated variables, only the variable with the greatest permutation importance was kept.

The selected environmental variables were used to train separate models for each of the three differently treated occurrence data types: raw data (Model 1), presence cells (Model 2), and thinned data (Model 3). Model 1 was trained if at least 5 raw data records were available, distributed across at least 3 cells (see above). Model 2 and Model 3 were trained if at least 3 records of the corresponding data type were available to avoid computational failure. Although a smaller sample size, we argue that if those models performed better than Model 1, the threshold of 5 records becomes arbitrary and the assessed performance indicators (see section *Technical Validation*) more valuable. The same model architecture as in Model 0 was utilized, including model calibration and selection of the best performing model. However, this time, we added five different regularization multipliers (RM; i.e. 1, 2, 3, 5 and 10; based on previous studies^{68–70}) to counteract overfitting^{20,56} and for building simpler, ecologically more relevant, models⁶⁰. Hence, separate models for each possible combination out of feature classes and RMs were trained (Fig. 1; Model 1–3), resulting in 30 trained models for each data type and up to 90 models per species.

Metadata. Metadata was assembled for all data and includes general information about species (taxonomy and red list status), provided data type (native regions, occurrence records or Maxent prediction), bounding box of native regions, and if relevant, information about the occurrence data (number of raw data records, Moran's

Variable	Code
Annual Mean Temperature	CHELSA_BIO1
Mean Diurnal Range	CHELSA_BIO2
Isothermality	CHELSA_BIO3
Temperature Seasonality	CHELSA_BIO4
Max Temperature of Warmest Month	CHELSA_BIO5
Min Temperature of Coldest Month	CHELSA_BIO6
Temperature Annual Range	CHELSA_BIO7
Mean Temperature of Wettest Quarter	CHELSA_BIO8
Mean Temperature of Driest Quarter	CHELSA_BIO9
Mean Temperature of Warmest Quarter	CHELSA_BIO10
Mean Temperature of Coldest Quarter	CHELSA_BIO11
Annual Precipitation	CHELSA_BIO12
Precipitation of Wettest Month	CHELSA_BIO13
Precipitation of Driest Month	CHELSA_BIO14
Precipitation Seasonality	CHELSA_BIO15
Precipitation of Wettest Quarter	CHELSA_BIO16
Precipitation of Driest Quarter	CHELSA_BIO17
Precipitation of Warmest Quarter	CHELSA_BIO18
Precipitation of Coldest Quarter	CHELSA_BIO19
Fraction of mosaic cropland/natural vegetation	X30_ESA_CCI
Fraction of mosaic natural vegetation/cropland	X40_ESA_CCI
Fraction of broadleaved evergreen, closed to open, tree cover	X50_ESA_CCI
Fraction of broadleaved deciduous, closed to open, tree cover	X60_ESA_CCI
Fraction of needleleaved evergreen, closed to open, tree cover	X70_ESA_CCI
Fraction of needleleaved deciduous, closed to open, tree cover	X80_ESA_CCI
Fraction of mixed leaf type tree cover	X90_ESA_CCI
Fraction of mosaic tree and shrub/herbaceous cover	X100_ESA_CCI
Fraction of mosaic herbaceous cover/tree and shrub	X110_ESA_CCI
Fraction of shrubland	X120_ESA_CCI
Fraction of grassland	X130_ESA_CCI
Fraction of lichens and mosses	X140_ESA_CCI
Fraction of sparse vegetation	X150_ESA_CCI
Fraction of tree cover, flooded, fresh or brakish water	X160_ESA_CCI
Fraction of tree cover, flooded, saline water	X170_ESA_CCI
Fraction of shrub or herbaceous cover, flooded, fresh/saline/brakish water	X180_ESA_CCI
Fraction of bare areas	X200_ESA_CCI

Table 1. Environmental data used in this study. The layers ($n = 36$) are based on Karger *et al.*⁶² and the European space agency's land cover product⁶³.

Index⁷¹, calculated as a measure of spatial autocorrelation and based on the number of raw occurrence points obtained per cell), and Maxent metadata: training data (filter treatment, number of training data points), thresholds for converting the prediction into binary range maps⁵⁹, model settings (features, parameters, transformations, regularization multiplier, variables) and out of the box⁶⁰ model performance, including degree of overfit (DOO) quantified as the difference between calibration and testing AUC during k-fold cross validation⁷⁰, as well as self-assessed model performance metrics as described in the section *Technical Validation*.

Data Records

Dataset. The presented dataset is stored in a stable Dryad Digital Repository⁷² and can be explored at <https://plant-ranges.indacol.no>. The dataset includes spatial information for 47,675 species at different levels of detail. In total, range estimates (i.e. relative environmental suitability within native regions) have been predicted for 27,208 species using Maxent, for 30,906 species native occurrence records are provided, and for 47,675 species the spatial extent of its native WGSRPD-*regions* is provided.

All gathered and generated data are stored in netCDF files and can be called by specifying a *varname*. Spatial predictions are provided in Maxent's raw as well as default output (i.e. complementary log-log (cloglog) transformed, but see section *Usage Notes*)^{27,59,60}. The suggested data is stored in folder *basic*. These netCDF files (default output and raw output) assemble the best performing Maxent prediction (*varname*: Maxent prediction) for each species selected based on the highest harmonic mean between AUC and AUC_{PR} (see *Technical Validation*), along with number of occurrence records per cell (*varname*: Presence cells) and rasterized native WGSRPD-*regions* (*varname*: Native region).

	Reference		Red list category						
			DD	LC	NT	VU	EN	CR	Total
AUC	Presence - background	Mean	0.939	0.937	0.95	0.96	0.971	0.957	0.945
		Median	0.961	0.951	0.977	0.985	0.994	0.989	0.964
	Reference range	Mean	0.817	0.89	0.927	0.931	0.929	0.915	0.902
		Median	0.852	0.925	0.972	0.974	0.98	0.987	0.943
AUC _{PR}	Presence - background	Mean	0.576	0.529	0.656	0.69	0.749	0.7	0.589
		Median	0.603	0.535	0.717	0.755	0.833	0.797	0.617
	Reference range	Mean	0.516	0.664	0.686	0.653	0.655	0.592	0.658
		Median	0.527	0.702	0.737	0.712	0.699	0.626	0.702

Table 2. Performance of Maxent predictions in the suggested dataset. Mean and median values of area under the receiver operating characteristics curve for true vs. false positive rate (AUC) and recall vs. precision (AUC_{PR}) for all species and across different IUCN threat categories (i.e. data-deficient (DD), least concern (LC), near-threatened (NT), vulnerable (VU), endangered (EN) and critically endangered (CR)). Calculations are based on presence-background data (n = 27,208) and on comparison to expert-based range maps retrieved from IUCN (i.e. reference range, n = 4,257).

The netCDF files in folder *advanced* contain one Maxent prediction for each occurrence data type (*varname*: Model 1, Model 2 or Model 3), instead of best performing Maxent prediction (i.e. *varname* Maxent prediction is not applicable). Number of occurrence records per cell (*varname*: Presence cells) and rasterized native WGSRPD-regions (*varname*: Native region) are identical in all netCDF files.

Each band in the netCDF files assembles the mentioned variables for one species. The corresponding bands can be looked up in the metadata (i.e. *speciesID*). Furthermore, the metadata can be used to select appropriate cut-off thresholds for generating binary range maps, filter models based on species, performance, or desired datatypes, and to lookup the relevant study extent for masking individual predictions (see *Usage Notes*).

Technical Validation

Maxent predictions. We calculated performance metrics for model 1 to 3 for each species using its corresponding presence cells to validate the Maxent predictions. Receiver operating characteristic curves and the corresponding area under the curve for *recall* (i.e. *true positive rate*, *sensitivity*) versus *false positive rate* (AUC) as well as *precision* versus *recall* (AUC_{PR}) were generated using the packages *ROCR*⁷³ and *PRROC*⁷⁴ in R. *Recall* was calculated as the fraction of correctly predicted presence cells compared to all presence cells of the reference (Eq. 1), the *false positive rate* as the fraction of falsely assigned presence cells compared to all true absence cells (Eq. 2), and *precision* as the fraction of correctly assigned presence cells compared to all predicted presence cells (Eq. 3). In addition, *F*₁-scores (Eq. 4) were calculated as harmonic mean between *recall* and *precision* at all possible cut-off thresholds to transform the Maxent prediction into a binary range map. The maximum obtained *F*₁-score indicates how well a potential binary range map performs at equal importance of *recall* and *precision*.

$$\text{Recall} = \frac{\text{True Presence}}{\text{True Presence} + \text{False Absence}} \quad (1)$$

$$\text{False positive rate} = \frac{\text{False Presence}}{\text{False Presence} + \text{True Absence}} \quad (2)$$

$$\text{Precision} = \frac{\text{True Presence}}{\text{True Presence} + \text{False Presence}} \quad (3)$$

$$F_1 = 2 \left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \right) \quad (4)$$

AUC and AUC_{PR} are threshold-independent performance measures for binary classifiers. An AUC value of 1 indicates a perfect model, an acceptable AUC value (>0.7)⁷⁵ indicates the ability to predict many true presences at a low false positive rate, and an AUC value of 0.5 indicates the model performing as good as a random guess. The average AUC obtained across the suggested dataset was 0.95 when comparing predictions to its corresponding presence cells (Table 2), indicating well-performing models for the majority of species. For 26,977 species (99%), at least one Maxent prediction had an AUC value above 0.7⁷⁵.

AUC_{PR} is not affected by true negatives (i.e. true absence) which often dominated our dataset. A higher AUC_{PR} value indicates a relatively higher ability to correctly predict a high proportion of presumably true range while maintaining a high precision compared to a lower AUC_{PR}. However, the AUC and AUC_{PR} values, as well as max. *F*₁-score, described here were calculated based on presence-background data and are highly influenced by class balances. Strictly speaking, both false presences and true absences cannot be determined

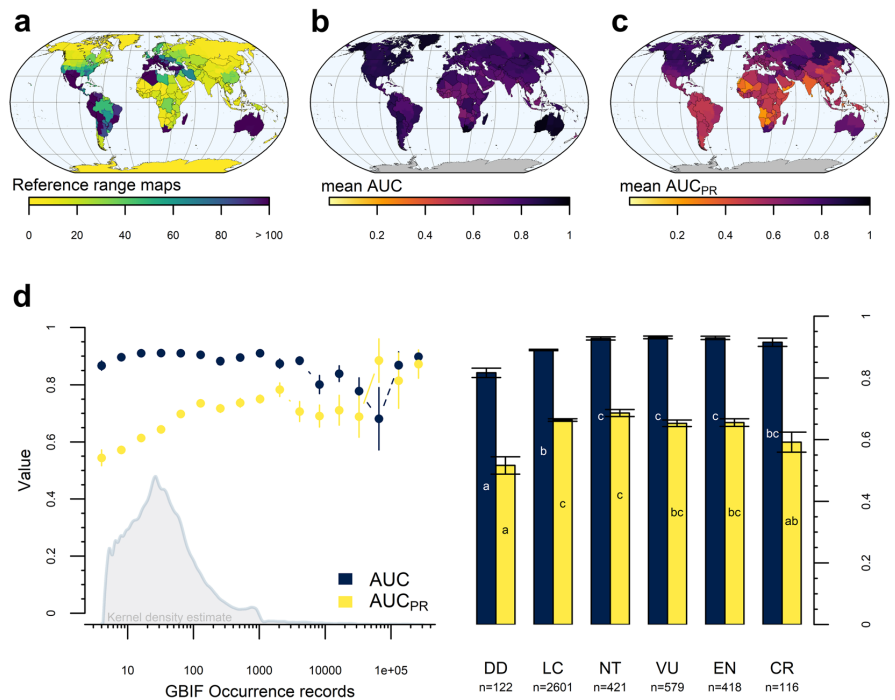


Fig. 3 Performance metrics for the suggested Maxent predictions. **(a)** Number of reference range maps available used for calculating performance metrics. Average values for species native to the corresponding regions of area under the receiver operating characteristics curve for **(b)** true vs. false positive rate (AUC) and **(c)** recall vs. precision (AUC_{PR}). **(d)** Mean and standard deviation of AUC (blue) and AUC_{PR} (yellow) per rounded log-transformed number of raw occurrence data points (left) and for species in different IUCN red list categories (right), i.e. data-deficient (DD), least concern (LC), near-threatened (NT), vulnerable (VU), endangered (EN) and critically endangered (CR). Significant differences across IUCN categories in **d** are indicated by different letters in bars for AUC (white text) and AUC_{PR} (black text).

with presence-only data. Hence, the performance metrics described here can only be used to compare different models for a given species, but not across different species^{76,77}.

Therefore, we evaluated the Maxent predictions by comparison to available expert-based range maps, as an additional evaluation dataset³². Expert-based range maps were retrieved from IUCN, if available (hereafter: reference ranges). Only reference ranges that were labelled as “native” and “extant (resident)” or “probably extant (resident)” were considered. For 4,257 species of our Maxent predictions, range maps were available at IUCN. These species were unevenly distributed in space (Fig. 3a), across IUCN red list categories (Fig. 3d) as well as the plant classes dicots (Magnoliopsida, $n = 3,480$), monocots (Liliopsida, $n = 731$), ferns (Polypodiopsida, $n = 27$), conifers (Pinopsida, $n = 17$), and lycopods (Lycopodiopsida, $n = 2$). Reference ranges were used to calculate the above described performance measures (i.e. max. F_1 -score, AUC and AUC_{PR}). However, this time we dealt, presumably, with actual presences and absences of the given species, making the performance metrics comparable across species⁷⁶. Maxent predictions for species classified as “data-deficient” (DD) obtained the lowest, and predictions for species classified as “near-threatened” (NT), “vulnerable” (VU) and “endangered” (EN) the highest AUC values (Fig. 3d). However, these differences were marginal and all average values consistently high across different IUCN categories (mean AUC: 0.9; Table 2) and across the globe (Fig. 3b). Although AUC is a strong indication of model performance⁷⁵, the predictions seem to rarely accommodate both a high *recall* and a high *precision* (represented in either max. F_1 -score or AUC_{PR} value) when compared to reference ranges. However, we found a large variation and no clear trend in AUC_{PR} values for species across different threat-level categories (Fig. 3d), and although the average AUC_{PR} was lowest for species native to parts of central Africa, India and south-eastern Asia (Fig. 3c), we expect these values to be of little explanatory power due to the limited sample sizes in these regions (Fig. 3a). Moreover, AUC_{PR} seems to increase with increasing data availability (Fig. 3d). We assume that low data coverage in sparsely populated areas influenced modelling performance for some, primarily widespread, species, highlighting that sometimes more spatially distributed occurrence data is required for making expert-alike range maps⁷⁸.

Furthermore, based on a qualitative assessment of predictions for twelve randomly selected species, we expect uncertainties due to differences in data availability across administrative borders as well as for highly naturalized species. For instance, the clustered occurrence records for *Cedrus libani* in Lebanon (Fig. 2b)

resulted in less precise data than elsewhere available for this species⁷⁹, while the prediction for *Laburnum anagyroides* (Fig. 2c) was affected by naturalized occurrence records outside its native origin⁸⁰ but still within its native WGSRPD-*regions*. However, this will be most problematic for abundant, widespread, and naturalized species, and hence only relevant for the smallest fraction of global biodiversity³³. In addition, the predictions for more vulnerable species, presumably small-ranged or endemic, seem to perform better than species in the lowest red list category (i.e. least concern (LC)) in terms of AUC when compared to reference ranges (Fig. 3d).

In fact, the remaining randomly selected predictions were either consistent with point data (e.g. *Terminalia macrostachya*⁸¹), reflected the current knowledge of elsewhere available data, although at a coarser spatial resolution and less detailed (e.g. *Mammillaria grahamii*⁸²), or offered an improvement compared to previously unavailable spatial data (e.g. *Eucalyptus elliptica*⁸³, *Megistostegium nodulosum*⁸⁴ (Fig. 2d), *Memecylon elegantulum*⁸⁵, *Psidium salutare*^{86,87}, *Siparuna conica*^{88,89}, *Trisetaria dufourei*⁹⁰). However, the prediction of *Pyracantha angustifolia* was difficult to evaluate due to poorly understood range dynamics⁹¹, highlighting the need for more data for vascular plant species.

We want to stress that our predictions indicate environmentally suitable conditions even if isolated from known species occurrence locations. For instance, *Amomum pterocarpum* seems to be restricted to southern India and Sri Lanka^{92,93} while our prediction indicates environmentally suitable conditions in north-eastern India (Fig. 2a), which in fact, supports a possible observation nearby⁹⁴. We further detected several expert-based range maps with a substantial mismatch to our data, confirming that some of the expert-based data may be too conservative⁹⁵ (e.g. *Magnolia pugana*)⁹⁶. However, we also found expert-based ranges being smaller (e.g. *Vallesia glabra* or *Tetraclinis articulata*)^{97,98} than predicted environmental suitability indicates, or being incorrectly georeferenced (e.g. *Corylus cornuta*)⁹⁹. Hence, besides highlighting mismatches to expert-based range maps, we expect this dataset to be of sufficient quality to serve as time- and cost-efficient range map substitutes and pre-assessed range estimates for currently unmapped species.

External data. The retrieved native WGSRPD-*regions* are provided by POWO under a CC BY 3.0 license (<https://creativecommons.org/licenses/by/3.0/>) and have been checked for consistency to assure proper workflow of data retrieval from POWO and feature matching to the WGSRPD level 4 shapefile. However, the data provider, POWO, cannot warrant the quality or accuracy of the WGSRPD data⁴². In addition, other data (e.g. ecoregions¹⁰⁰) may ecologically be more relevant than administrative boundaries. However, WGSRPD offers the most detailed data on species' native origins available on a large-scale, to the best of our knowledge. An attempt in matching native WGSRPD-*regions* to ecoregions was discontinued after loss of information due to incompatible geographical boundaries. Hence, we consider the utilized WGSRPD-*regions*, currently, as the best compromise between level of detail and availability of data on species' native origins. Furthermore, spatial inaccuracies and biases in the occurrence data retrieved from GBIF were counteracted by the implemented filtering steps, the coarse spatial resolution, by avoiding non-native occurrence records and the model calibration techniques. However, any unforeseen misclassified or misreported records may flaw predictions for individual species. In addition, data retrieval via GBIF's API was limited to 100,000 occurrence records per request. We extended this limit by sending one request per native country for each species, and hence, expect this issue to be irrelevant for our study. We further want to stress that most of the generated predictions have not been validated individually, and that some predictions may be erroneous either due to data limitations or simply because digitally stored data can contain minor but crucial blunders. For instance, in terms of nomenclature, the red-listed species *Cotoneaster cambricus* is endemic to Wales¹⁰¹, but also seems to be a synonym for a widespread species according to POWO⁴². Consequently, either our spatial prediction or the expert-based range for this species is incorrect.

Usage Notes

All data handling, modelling and visualization was done using R version 4.0.3⁴⁴ in RStudio version 1.4.1103¹⁰². Handling of all spatial data was done using the R packages *raster*, *rgdal*, *maptools*, *rgeos* and *sp*^{52,103–106}. A showcase for opening the different data types for individual species, is available at https://github.com/jannebor/plant_range_estimates. Although functionality of the code may be given at newer, or older, versions, we expect the best user-experience using the versions specified in this descriptor.

Maxent predictions are given as raw and cloglog transformed output. These outputs are related monotonically, meaning that the performance metrics described in this study, as well as a potential binary range map (excluding prevalence dependent thresholds), will be identical for both raw and cloglog output⁵⁶. For users mostly interested in qualitative analyses, both predictions can simply be interpreted as indices of environmental suitability²⁰. However, due to rescaling, the exact interpretation and appearance of each output differs. In general, Maxent's output interpretation depends on the underlying data, and differs, in our case between Model 1 (raw data including pseudo-replicates = abundance) compared to Model 2 and 3 (presence), but gives an estimate of the abundance, or presence, of the species in relation to the true modelled quantity (either abundance or presence). Maxent's raw output reflects the exponential Maxent model itself, and can be interpreted as a relative occurrence (or presence) rate summing up to 1²⁰. The raw output does not rely on any assumptions²⁰, however, it may not perform well in visualizing actual differences in suitability¹⁰⁷. Being rescaled on a more common range from 0 to 1, the cloglog transformation compresses extreme values, and hence facilitates visualization and comparison amongst predictions²⁷. It can, arguably, be interpreted as a relative probability of presence under certain assumptions²⁷. However, as these assumptions are rarely met, we strongly discourage users from this interpretation and suggest interpreting the cloglog output values as an estimate of relative environmental suitability²⁰ instead.

We further suggest using Maxent predictions with an AUC below 0.7 only in exceptions, and in large-scale studies. In general, our predictions may overestimate true range extents of endemic species and underestimate

ranges of widespread species. However, in worst case, the entire native WGSPRD-*regions* are outlined as being environmentally suitable, which may be acceptable in some cases, but not in others.

In addition, Model 1 has been fitted with the suggested minimum number of records for generating meaningful distributions models^{53,54}, but Model 2 and 3 were in some cases trained with less records. Whether this low sample size as well as its implied uncertainty is acceptable or not will differ between users and applications and needs to be considered.

The full data, including Maxent predictions (cloglog transformed), underlying occurrence records, native regions and corresponding metadata, can be explored at <https://plant-ranges.indecol.no>. Here, the predictions based on individual models (Model 1 to 3) as well as a suggested (i.e. best performing) prediction highlight environmentally suitable conditions, if available for the selected species. Predictions can potentially be transformed into a map indicating where the species is most certainly found, as required for local management and conservation actions⁹⁵, or into a conservative range map, best suited for analysing global patterns¹⁰⁸ and highlighting where a species is certainly absent¹⁰⁹. However, the choice of an appropriate cut-off threshold is highly application specific. We outlined “potential range maps” in the data explorer for illustrational purposes only and based on the best performing prediction. We applied different cut-off thresholds to represent different levels of confidence using the R package *dismo*⁵⁹. The threshold at which there was no omission (possibly suitable), the threshold at which the F_1 -score is highest (probably suitable) and presence cells (presence).

Code availability

All data and code is available without restrictions under the terms of a Creative Commons Zero (CC0) waiver (<https://creativecommons.org/share-your-work/public-domain/cc0/>). R code for retrieving and filtering data from POWO and GBIF, and for generating and evaluating Maxent models is available on GitHub (https://github.com/jannebor/plant_range_estimates). Any further requests can be directed to the corresponding author.

Received: 2 June 2021; Accepted: 24 February 2022;

Published online: 29 March 2022

References

1. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis*. (World Resources Institute, 2005).
2. Moran, D. & Kanemoto, K. Identifying species threat hotspots from global supply chains. *Nat. Ecol. Evol.* **1**, 0023 (2017).
3. Newbold, T. Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proc. R. Soc. B Biol. Sci.* **285**, 20180792 (2018).
4. Newbold, T. *et al.* Global effects of land use on local terrestrial biodiversity. *Nature* **520**, 45–50 (2015).
5. Newbold, T. *et al.* Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* (80-). **353**, 288–291 (2016).
6. Veronesi, F., Moran, D., Stadler, K., Kanemoto, K. & Wood, R. Resource footprints and their ecosystem consequences. *Sci. Rep.* **7**, 40743 (2017).
7. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development*. A/RES/70/1 (United Nations, 2015).
8. Díaz, S. *et al.* Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science* (80-). **366**, eaax3100 (2019).
9. Lenzen, M. *et al.* International trade drives biodiversity threats in developing nations. *Nature* **486**, 109–112 (2012).
10. Hellweg, S. & Milà i Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science* (80-). **344**, 1109–1113 (2014).
11. Chaudhary, A. & Brooks, T. M. National Consumption and Global Trade Impacts on Biodiversity. *World Dev.* **121**, 178–187 (2019).
12. Pereira, H. M., Ziv, G. & Miranda, M. Countryside Species-Area Relationship as a Valid Alternative to the Matrix-Calibrated Species-Area Model. *Conserv. Biol.* **28**, 874–876 (2014).
13. Lomolino, M. V. & Heaney, L. R. *Frontiers of Biogeography: New Directions in the Geography of Nature*. (Sinauer Associates Inc. Publishers, 2004).
14. World Wildlife Fund. *WildFinder: Online database of species distributions*. <http://www.worldwildlife.org/WildFinder> (2006).
15. BirdLife International. *IUCN Red List for birds*. <http://www.birdlife.org> (2019).
16. IUCN. *The IUCN Red List of Threatened Species. Version 2021-1* <https://www.iucnredlist.org> (2021).
17. Curran, M. *et al.* Toward Meaningful End Points of Biodiversity in Life Cycle Assessment. *Environ. Sci. Technol.* **45**, 70–79 (2011).
18. Woods, J. S. *et al.* Ecosystem quality in LCIA: status quo, harmonization, and suggestions for the way forward. *Int. J. Life Cycle Assess.* **23**, 1995–2006 (2018).
19. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190**, 231–259 (2006).
20. Merow, C., Smith, M. J. & Silander, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* (Cop.). **36**, 1058–1069 (2013).
21. Araújo, M. B. *et al.* Standards for distribution models in biodiversity assessments. *Sci. Adv.* **5**, eaat4858 (2019).
22. Zurell, D. *et al.* A standard protocol for reporting species distribution models. *Ecography* (Cop.). **43**, 1261–1277 (2020).
23. Brummitt, R. K., Pando, F., Hollis, S. & Brummitt, N. A. World Geographical Scheme for Recording Plant Distributions. *International Working Group on Taxonomic Databases (TDWG)* <https://www.tdwg.org/standards/wgspnd/> (2001).
24. GBIF. The Global Biodiversity Information Facility: What is GBIF? <https://www.gbif.org/what-is-gbif> (2021).
25. Phillips, S. J., Dudík, M. & Schapire, R. E. Maxent software for modeling species niches and distributions (Version 3.4.0). http://biodiversityinformatics.amnh.org/open_source/maxent/ (2016).
26. Phillips, S. J., Dudík, M. & Schapire, R. E. A maximum entropy approach to species distribution modeling. *Proc. Twenty-first Int. Conf. Mach. Learn.* 655–662 (2004).
27. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: an open-source release of Maxent. *Ecography* (Cop.). **40**, 887–893 (2017).
28. Reddy, S. & Dávalos, L. M. Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30**, 1719–1727 (2003).
29. Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M. & Baselga, A. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847–858 (2008).
30. Isaac, N. J. B. & Poock, M. J. O. Bias and information in biological records. *Biol. J. Linn. Soc.* **115**, 522–531 (2015).

31. Feeley, K. J. & Silman, M. R. Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Divers. Distrib.* **17**, 1132–1140 (2011).
32. Radosavljevic, A. & Anderson, R. P. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J. Biogeogr.* **41**, 629–643 (2014).
33. ter Steege, H. *et al.* Hyperdominance in the Amazonian Tree Flora. *Science* (80-). **342**, 1243092 (2013).
34. Kuipers, K. J. J., Hellweg, S. & Verones, F. Potential Consequences of Regional Species Loss for Global Species Richness: A Quantitative Approach for Estimating Global Extinction Probabilities. *Environ. Sci. Technol.* **53**, 4728–4738 (2019).
35. Gade, A. L., Hauschild, M. Z. & Laurent, A. Globally differentiated effect factors for characterising terrestrial acidification in life cycle impact assessment. *Sci. Total Environ.* **761**, 143280 (2021).
36. Géron, C. *et al.* Urban alien plants in temperate oceanic regions of Europe originate from warmer native ranges. *Biol. Invasions* **23**, 1765–1779 (2021).
37. Mair, L. *et al.* A metric for spatially explicit contributions to science-based species targets. *Nat. Ecol. Evol.* **5**, 836–844 (2021).
38. Bachman, S., Moat, J., Hill, A., de la Torre, J. & Scott, B. Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. *Zookeys* **150**, 117–126 (2011).
39. Cardoso, P. red - an R package to facilitate species red list assessments according to the IUCN criteria. *Biodivers. Data J.* **5**, e20530 (2017).
40. Lee, C. K. F., Keith, D. A., Nicholson, E. & Murray, N. J. Redlistr: tools for the IUCN Red Lists of ecosystems and threatened species in R. *Ecography* (Cop.) **42**, 1050–1055 (2019).
41. Bachman, S., Walker, B., Barrios, S., Copeland, A. & Moat, J. Rapid Least Concern: towards automating Red List assessments. *Biodivers. Data J.* **8** (2020).
42. POWO. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. <http://www.plantsoftheworldonline.org/> (2021).
43. Chamberlain, S. *et al.* taxize: Taxonomic information from around the web. R package version 0.9.98. <https://github.com/ropensci/taxize> (2020).
44. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria <https://www.r-project.org/> (2021).
45. ITIS. Integrated Taxonomic Information System. <https://www.itis.gov/> (2021).
46. Wickham, H. rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5. <https://cran.r-project.org/package=rvest> (2019).
47. Desmet, P. & Page, R. WGSRPD. GitHub repository <https://github.com/tdwg/wgsrpd> (2018).
48. Chamberlain, S. *et al.* rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.6.0. <https://cran.r-project.org/package=rgbif> (2021).
49. GBIF. GBIF Occurrence Download. <https://doi.org/10.15468/dl.uvd56q> (2021).
50. Winkler, K., Fuchs, R., Rounsevell, M. & Herold, M. Global land use changes are four times greater than previously estimated. *Nat. Commun.* **12**, 2501 (2021).
51. Sippel, S., Meinschausen, N., Fischer, E. M., Székely, E. & Knutti, R. Climate change now detectable from any single day of weather at global scale. *Nat. Clim. Chang.* **10**, 35–41 (2020).
52. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. R package version 3.0-7. <https://cran.r-project.org/package=raster> (2019).
53. Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* (Cop.) **29**, 773–785 (2006).
54. Pearson, R. G., Raxworthy, C. J., Nakamura, M. & Townsend Peterson, A. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *J. Biogeogr.* **34**, 102–117 (2006).
55. Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* (Cop.) **31**, 161–175 (2008).
56. Elith, J. *et al.* A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
57. Anderson, R. P. & Raza, A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *J. Biogeogr.* **37**, 1378–1393 (2010).
58. Sjalander, M., Jahre, M., Tufté, G. & Reissmann, N. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure. *arXiv* 1–4 (2019).
59. Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. dismo: Species Distribution Modeling. R package version 1.1-4. <https://cran.r-project.org/package=dismo> (2017).
60. Muscarella, R. *et al.* ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods Ecol. Evol.* **5**, 1198–1205 (2014).
61. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).
62. Karger, D. N. *et al.* Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad, Dataset* <https://doi.org/10.5061/dryad.kd1d4> (2018).
63. ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. <http://maps.elie.ucl.ac.be/CCI/viewer/download.php> (2017).
64. Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* (Cop.) **38**, 541–545 (2015).
65. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in *2nd International Symposium on Information Theory* (eds Petrov, B. N. & Csaki, F.) 267–281 (Akademia Kiado, 1973).
66. Hurvich, C. M. & Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika* **76**, 297–307 (1989).
67. Sugiura, N. Further analysts of the data by akaike's information criterion and the finite corrections. *Commun. Stat. - Theory Methods* **7**, 13–26 (1978).
68. Morales, N. S., Fernández, I. C. & Baca-González, V. MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. *PeerJ* **5**, e3093 (2017).
69. Shcheglovitova, M. & Anderson, R. P. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecol. Modell.* **269**, 9–17 (2013).
70. Warren, D. L. & Seifert, S. N. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol. Appl.* **21**, 335–342 (2011).
71. Moran, P. A. P. Notes on Continuous Stochastic Phenomena. *Biometrika* **37**, 17 (1950).
72. Borgelt, J., Sicacha-Parada, J., Skarpaas, O. & Verones, F. Native range estimates for red-listed vascular plants. *Dryad, Dataset* <https://doi.org/10.5061/dryad.qbzkh18h9> (2022).
73. Sing, T., Sander, O., Beerwinkel, N. & Lengauer, T. ROCr: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
74. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
75. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression*. *The Statistician* **45** (Wiley, 2013).
76. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2008).

77. Sofaer, H. R., Hoeting, J. A. & Jarnevich, C. S. The area under the precision-recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* **10**, 565–577 (2019).
78. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
79. Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. Chorological maps for the main European woody species. *Data Br.* **12**, 662–666 (2017).
80. Rivers, M. C. Laburnum anagyroides. *The IUCN Red List of Threatened Species 2017*: e.T79919483A79919650 <https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T79919483A79919650.en> (2017).
81. Botanic Gardens Conservation International Group & IUCN SSC Global Tree Specialist. Terminalia macrostachya. *The IUCN Red List of Threatened Species 2019*: e.T150118895A150118897 <https://doi.org/10.2305/IUCN.UK.2019-3.RLTS.T150118895A150118897.en> (2019).
82. Heil, K., Terry, M. & Corral-Díaz, R. Mammillaria grahamii (amended version of 2013 assessment). *The IUCN Red List of Threatened Species 2017*: e.T152723A121546147 <https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T152723A121546147.en> (2017).
83. Brooker, M. & Kleinig, D. *Field Guide to Eucalypts*. (Blooming Books, 2006).
84. Koopman, M. M. A synopsis of the Malagasy endemic genus Megistostegium Hochr. (Hibiscaceae, Malvaceae). *Adansonia* **33**, 101–113 (2011).
85. World Conservation Monitoring Centre. Memecylon elegantulum. *The IUCN Red List of Threatened Species 1998*: e.T32597A9713234 <https://doi.org/10.2305/IUCN.UK.1998.RLTS.T32597A9713234.en> (1998).
86. Landrum, L. R. A revision of the Psidium salutare complex (Myrtaceae). *SIDA, Contrib. to Bot.* **20**, 1449–1469 (2003).
87. Tropical Plants Database. Ken Fern. *tropical.theferns.info* <https://tropical.theferns.info/viewtropical.php?id=Psidium+salutare> (2021).
88. Bernal, R., Gradstein, S. R. & Celis, M. Siparuna conica S.S.Renner & Hausner. *Catálogo de plantas y líquenes de Colombia* <http://catalogoplantasdecolombia.unal.edu.co> (2015).
89. Renner, S. S. & Hausner, G. New Species of Siparuna (Monimiaceae) II. Seven New Species from Ecuador and Colombia. *Missouri Bot. Gard. Press* **6**, 103–116 (1996).
90. Melendo, M., Giménez, E., Cano, E., Mercado, F. G. & Valle, F. The endemic flora in the south of the Iberian Peninsula: taxonomic composition, biological spectrum, pollination, reproductive mode and dispersal. *Flora - Morphol. Distrib. Funct. Ecol. Plants* **198**, 260–276 (2003).
91. Chari, L. D., Martin, G. D., Steenhuisen, S.-L., Adams, L. D. & Clark, V. R. Biology of Invasive Plants 1. *Pyracantha angustifolia* (Franch.) C.K. Schneid. *Invasive Plant Sci. Manag.* **13**, 120–142 (2020).
92. Sasidharan, N. Amomum pterocarpum Thwaites. *India Biodiversity Portal* <https://indiabiodiversity.org/species/show/258864#habitat-and-distribution> (2013).
93. Contu, S. Amomum pterocarpum. *The IUCN Red List of Threatened Species 2013*: e.T44393013A44450020 <https://doi.org/10.2305/IUCN.UK.2013-1.RLTS.T44393013A44450020.en> (2013).
94. Babyrose Devi, N., Das, A. & Singh, P. Amomum Pterocarpum (Zingiberaceae): a new record in the flora of Manipur. *Int. J. Adv. Res.* **6**, 546–549 (2018).
95. Jetz, W., Sekercioglu, C. H. & Watson, J. E. M. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110–9 (2008).
96. Gibbs, D. & Khela, S. Magnolia pugana. *The IUCN Red List of Threatened Species 2014*: e.T194806A2363344 <https://doi.org/10.2305/IUCN.UK.2014-1.RLTS.T194806A2363344.en> (2014).
97. Sayer, C. Vallesia glabra. *The IUCN Red List of Threatened Species 2015*: e.T62543A72668627 <https://doi.org/10.2305/IUCN.UK.2015-2.RLTS.T62543A72668627.en> (2015).
98. Sánchez Gómez, P., Stevens, D., Fennane, M., Gardner, M. & Thomas, P. Tetraclinis articulata. *The IUCN Red List of Threatened Species 2011*: e.T30318A9534227 <https://doi.org/10.2305/IUCN.UK.2011-2.RLTS.T30318A9534227.en> (2011).
99. Stritch, L., Roy, S., Shaw, K. & Wilson, B. Corylus cornuta (errata version published in 2017). *The IUCN Red List of Threatened Species 2016*: e.T194448A115337731 <https://doi.org/10.2305/IUCN.UK.2016-1.RLTS.T194448A115337731.en> (2016).
100. Olson, D. M. *et al.* Terrestrial ecoregions of the world: A new map of life on Earth. *Bioscience* **51**, 933–938 (2001).
101. Rivers, M. C. Cotonaster cambricus. *The IUCN Red List of Threatened Species 2017*: e.T102827479A102827485 <https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T102827479A102827485.en> (2017).
102. RStudio Team. RStudio: Integrated Development Environment for R. *RStudio, PBC, Boston, MA* <http://www.rstudio.com/> (2021).
103. Bivand, R., Keitt, T. & Rowlingson, B. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. <https://cran.r-project.org/package=rgdal> (2019).
104. Bivand, R. & Lewin-Koh, N. *mapproj: Tools for Handling Spatial Objects. R package version 0.9-5*. <https://cran.r-project.org/package=mapproj> (2019).
105. Bivand, R. & Rundel, C. *rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.5-1*. <https://cran.r-project.org/package=rgeos> (2019).
106. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer New York, 2013).
107. Phillips, S. J. & Elith, J. POC plots: calibrating species distribution models with presence-only data. *Ecology* **91**, 2476–2484 (2010).
108. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci.* **104**, 13384–13389 (2007).
109. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).

Acknowledgements

We want to thank Radek Lonka and the IndEcol Digital Lab for facilitating the use of the high-performance computing infrastructure and hosting the online application. This study is part of the Transforming Citizen Science for Biodiversity project hosted by the Digital Transformation initiative of the Norwegian University of Science and Technology.

Author contributions

J.B. was responsible for study design, methodologies, code writing, code execution, and writing the manuscript. J.S.P. contributed to methods for technical validation of the data and writing the manuscript. O.S. contributed to methodologies, interpretation of the data, and writing the manuscript. F.V. contributed to study design, interpreting the results, and writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

ISBN 978-82-326-6457-3 (printed ver.)
ISBN 978-82-326-6549-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)



NTNU

Norwegian University of
Science and Technology