Spatial Statistics 42 (2021) 100446

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

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.

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).

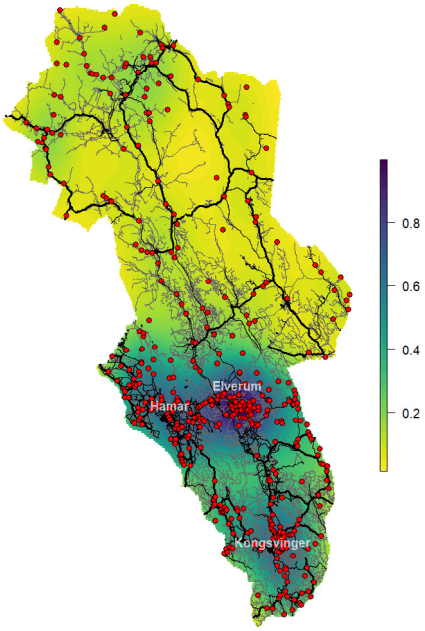


Fig. 1. Moose observations (red points) and road system (lines) in the county of Hedmark, Norway. Bold lines indicate main roads.

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 km $\times$ 1 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 (ht tps://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

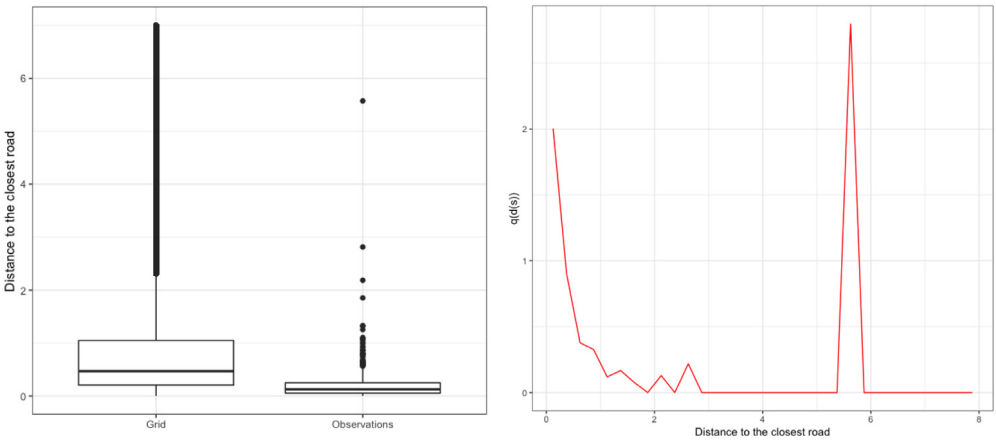


Fig. 2. (a) Boxplots of distance to the road system. Left: Dense grid of about 400 thousand points. Right: 472 reports of moose in Hedmark (b) Relationship between the observed ratio $\hat{q}(\mathbf{s}_d)$ and the distances to closest road, $\mathbf{s}_d$.

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 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.

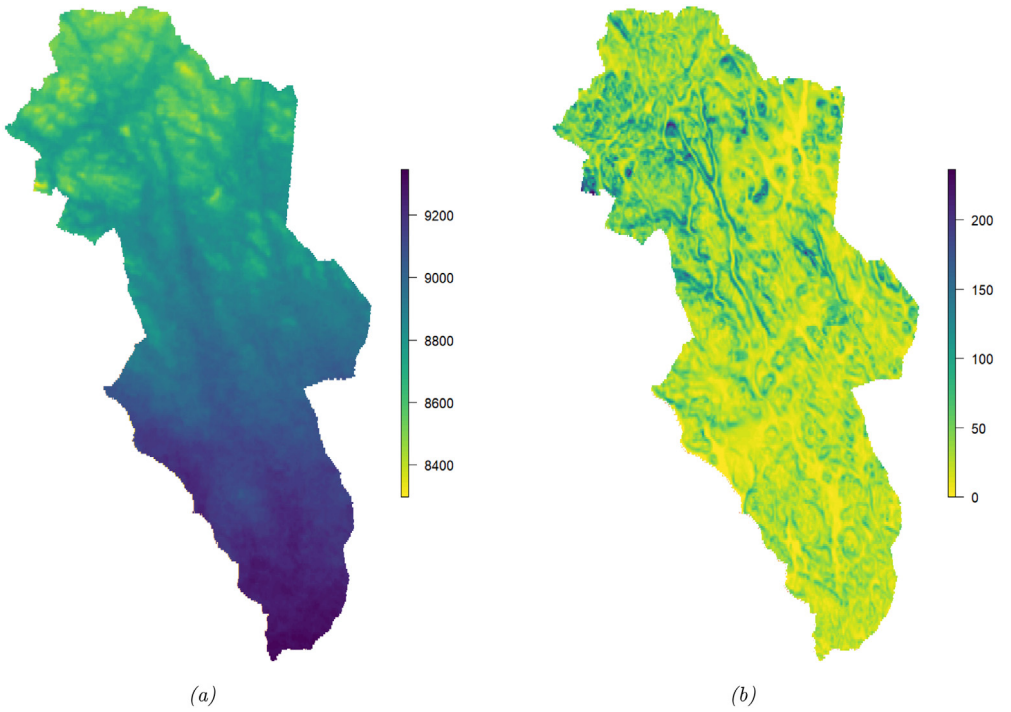


Fig. 3. (a) Solar Radiation (RAD) and (b) Terrain Ruggedness Index (TRI) in the county of Hedmark, Norway.

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:

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

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:

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

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, $B_k(\mathbf{s})$, $k = 1, \ldots, p$ 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 $\boldsymbol{\beta}$ 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 $\zeta_k = \exp(\theta_k)$, $k = 1, \ldots, p$. 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:

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

where $\mathbf{s}$ is a vector of locations in $\mathbb{R}^2$, Δ is the Laplacian. ν, $\kappa > 0$ 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:

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

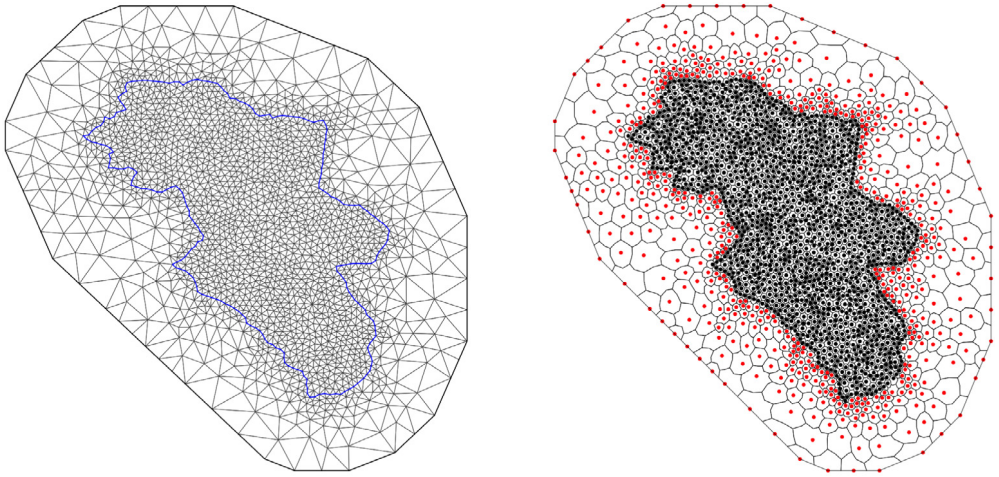
where G is the total number of vertices of the triangulation, $\{\phi_g\}$ is the set of basis functions, and $\{\tilde{\xi}_g\}$ are zero-mean Gaussian distributed weights. This way of representing the Gaussian Random Field has been proven to make more efficient the fitting process. Fig. 4a displays 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 $\boldsymbol{\omega}(\mathbf{s})$ be a finite-dimensional continuously specified random field defined as:

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



(a) (b)

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 $\boldsymbol{\omega}$:

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

can be approximated by :

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

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, $\tilde{\mathbf{s}} = \{\tilde{s}_1, \ldots, \tilde{s}_n\}$ are the nodes of the mesh and $\tilde{\mathbf{a}} = \{\tilde{a}_1, \ldots, \tilde{a}_n\}$ 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

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

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

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

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:

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

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,

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

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

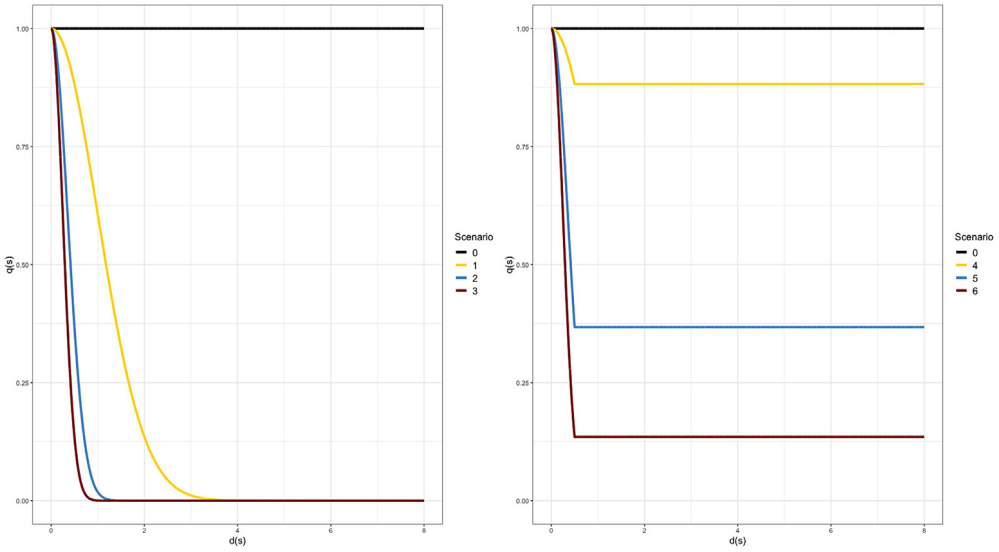


Fig. 5. (a) Relationship between $d(\mathbf{s})$ and $q(\mathbf{s})$ for the simulation scenarios 0,1,2 and 3 (b) Relationship between $d(\mathbf{s})$ and $q(\mathbf{s})$ for the simulation scenarios 0,4,5 and 6.

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 10 000 realizations $\{\theta^p_{jkl}\}, j = , 1 \ldots, 10\,000$, from the posterior distribution of each parameter θ for point pattern $k = 1, \ldots, 100$ 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:

$$bias_{kl} = \frac{1}{10000} \sum_{j=1}^{10000} \left(\theta^p_{jkl} - \tilde{\theta}\right)$$

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

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

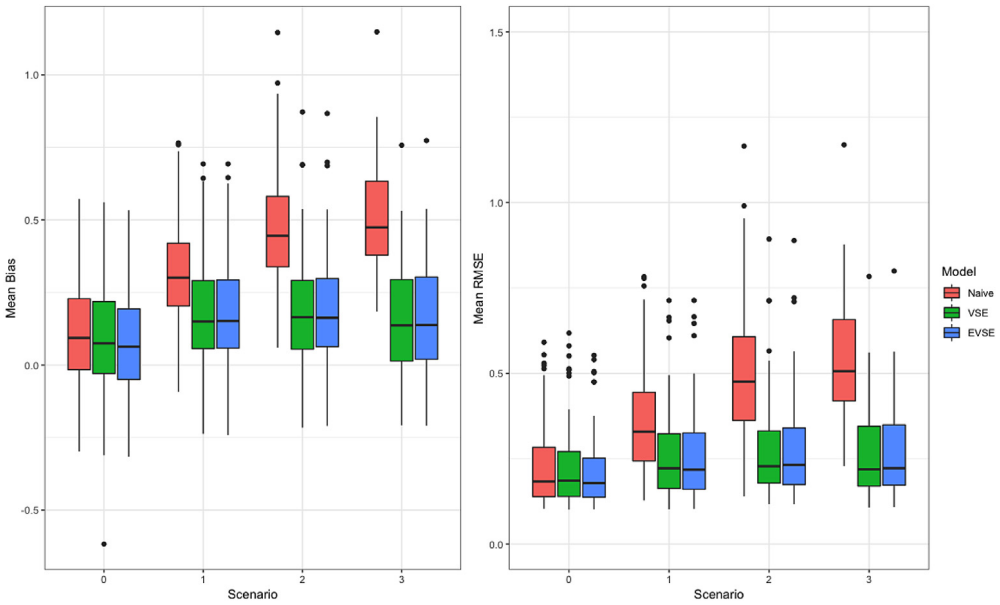


Fig. 6. Boxplots of mean bias (left) and mean RMSE (right) of β_1 for all datasets in each scenario (scenarios 0,1,2,3) and for each model.

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

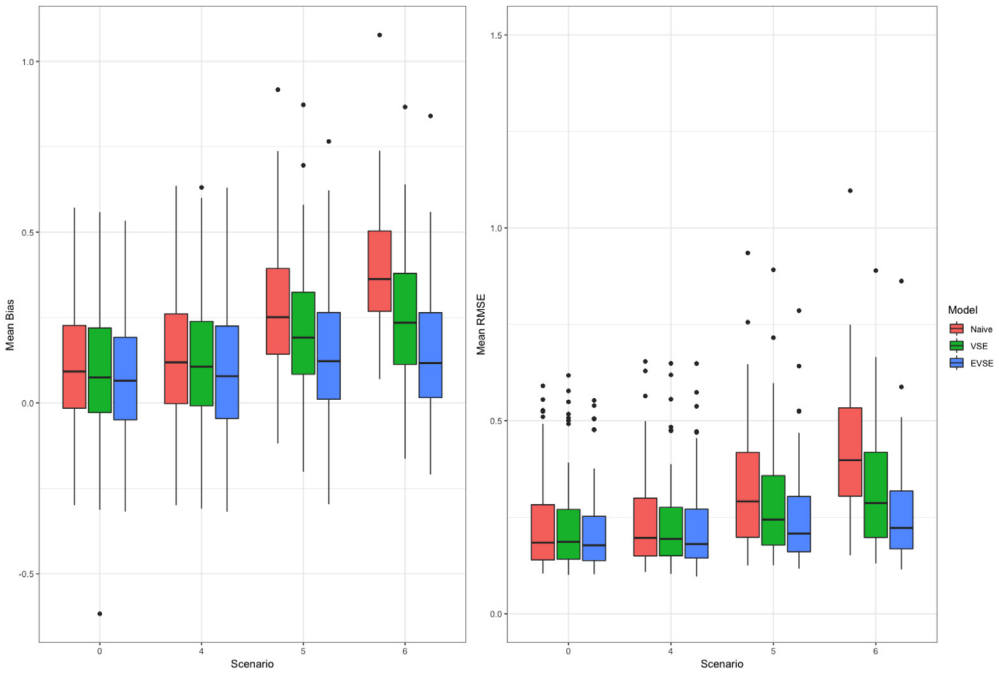


Fig. 7. Boxplots of mean bias (left) and mean RMSE (right) of β_1 for all datasets in each scenario (scenarios 0,4,5,6) and for each model.

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 $q(s) \approx 0.05$ 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

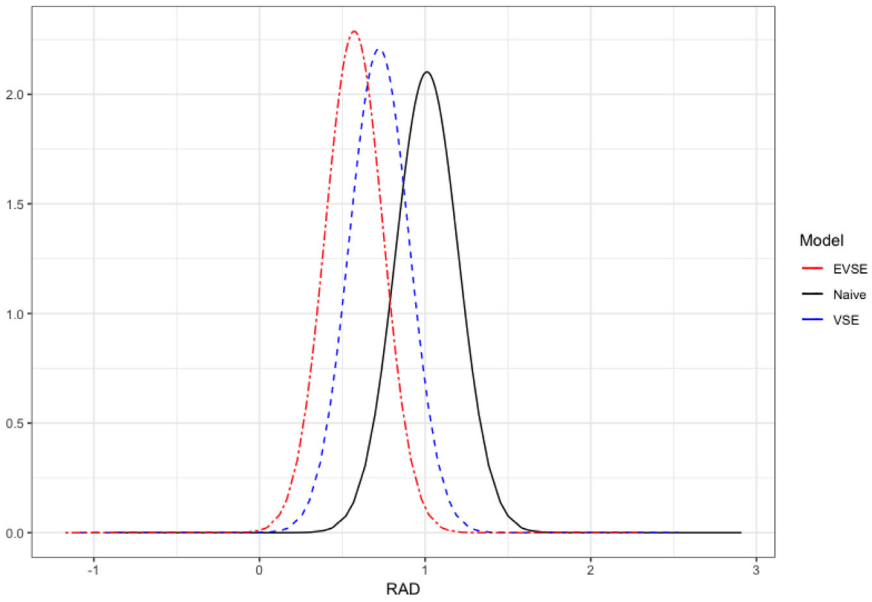


Fig. 8. Posterior density of RAD for the three models.

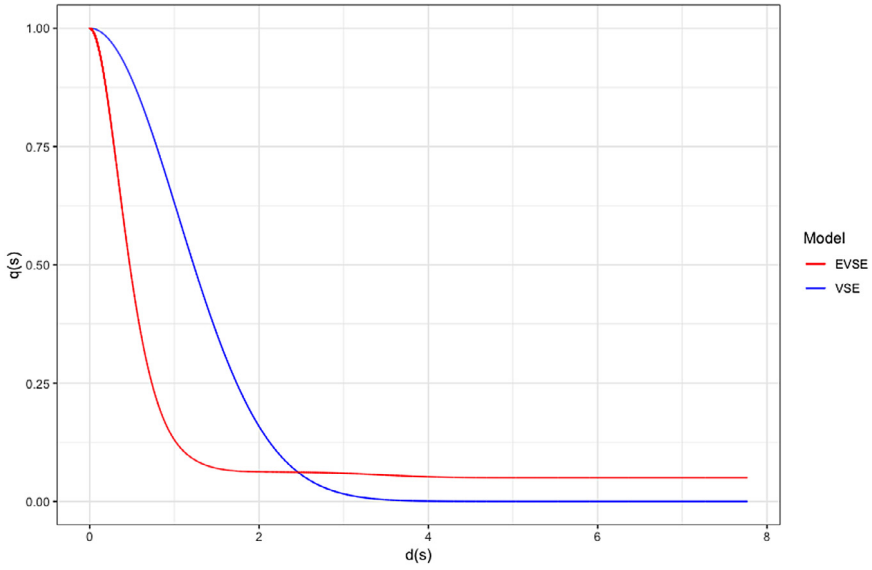


Fig. 9. Estimated relation between distance to the road system, in kilometers, and the probability of having access to location **s**.

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.

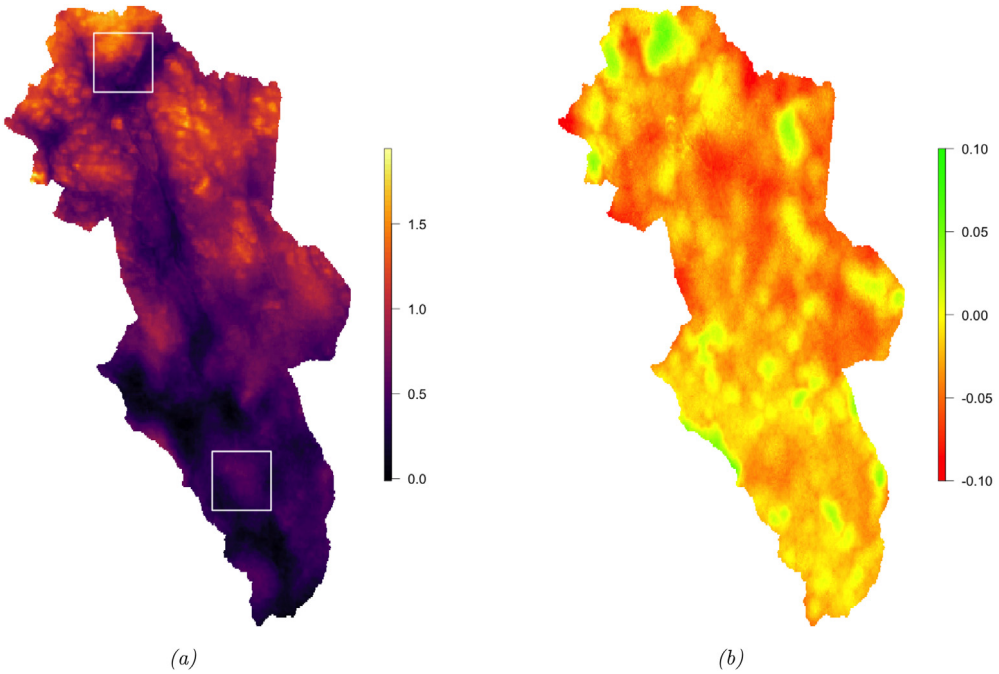


Fig. 10. (a) Differences in posterior median intensity and (b) differences in standard error of the posterior median intensity obtained through the VSE model and the naive model. In (a) the two squares represent the zones that are focused in Fig. 11 (south of Hedmark) and 12 (north of Hedmark).

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 km $\times$ 30 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

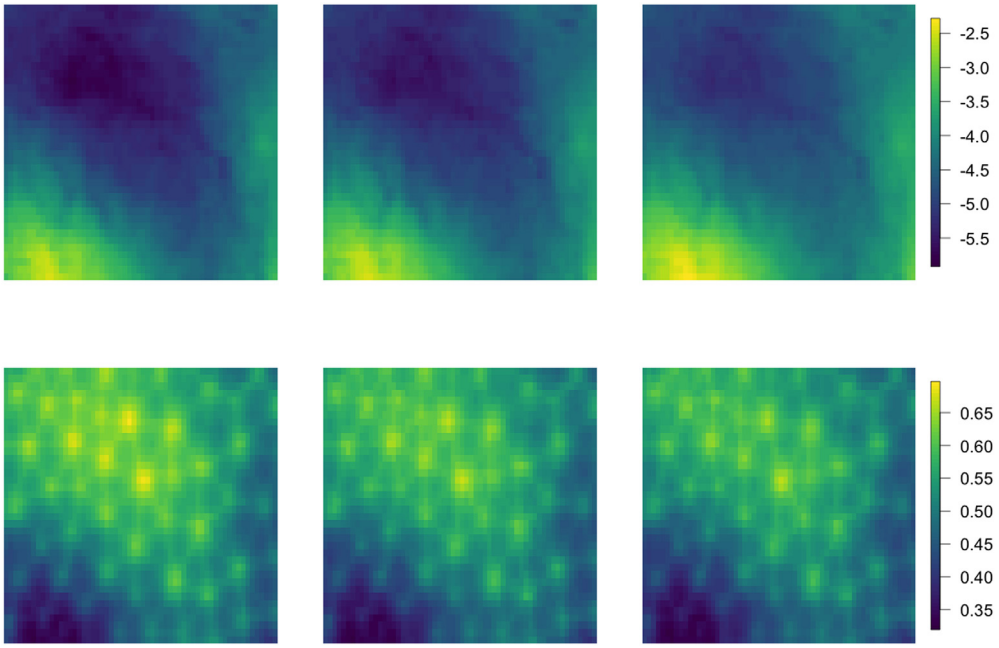


Fig. 11. Posterior median intensity (top) and associated standard error (bottom) for the naive model (left), the VSE model (middle) and the EVSE model (right) in zone 1.

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

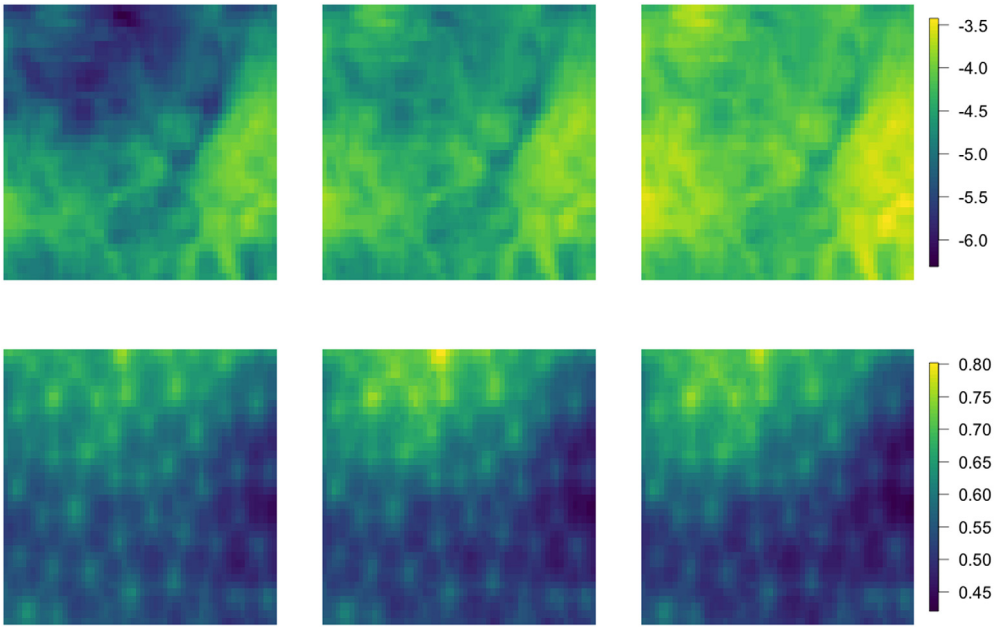


Fig. 12. Posterior median intensity (top) and associated standard error (bottom) for the naive model (left), the VSE model (middle) and the EVSE model (right) in zone 2.

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

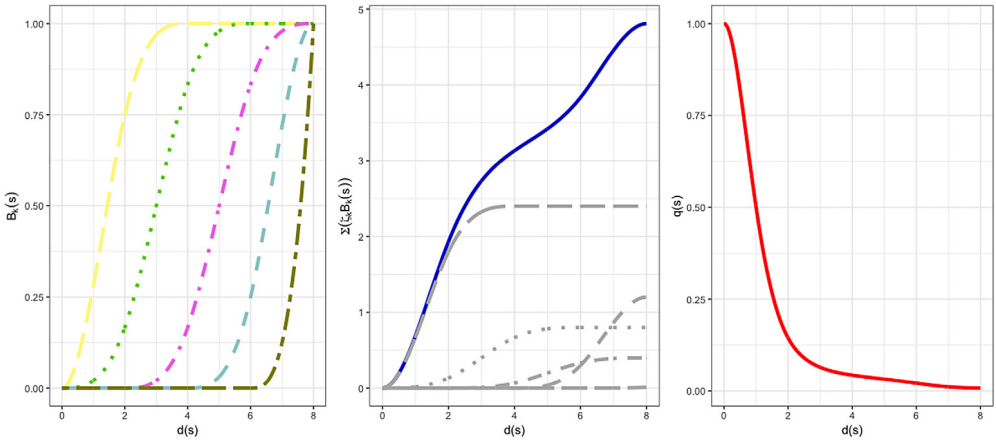


Fig. A.13. Illustration of the relationship between the basis functions and $q(\mathbf{s})$ in the EVSE model. Left, basis functions $B_k(\mathbf{s}), k = 1, \ldots, 5$. Middle, weighted basis functions by the coefficients $\zeta_k, k = 1, \ldots, 5$ (gray); linear combination of the weighted basis functions (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 $\zeta_k, k = 1, \ldots, 3$, 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 spices 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 $B_k(\mathbf{s}), k = 1, \ldots, p$. As mentioned in Section 3, $B_k(\mathbf{s}), k = 1, \ldots, p$ are a set of monotone nondecreasing functions . In addition to it, the coefficients $\zeta_k, k = 1, \ldots, p$ 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 .

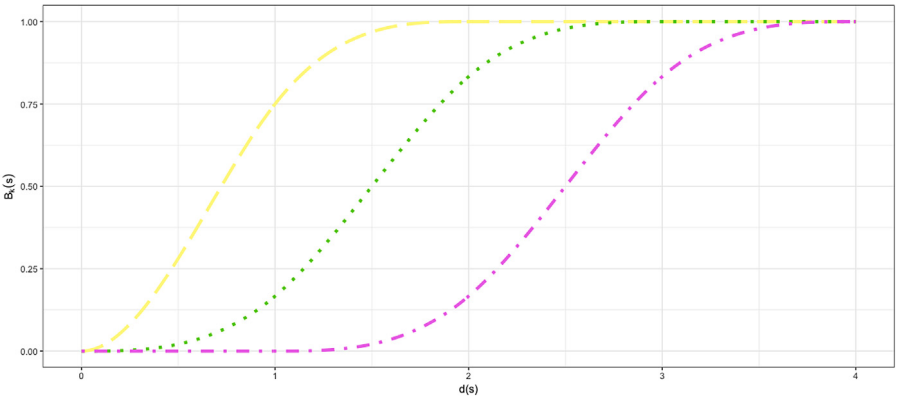


Fig. B.14. Basis functions used to fit the EVSE model in the simulation study.

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)
	6	0,01 (0,8)	0,35 (0,79)	0,94 (0,78)
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)
	6	0,36 (0,62)	0,63 (0,62)	0,82 (0,61)
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)
	6	0,88 (49,07)	0,87 (50,96)	0,68 (52,26)
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.

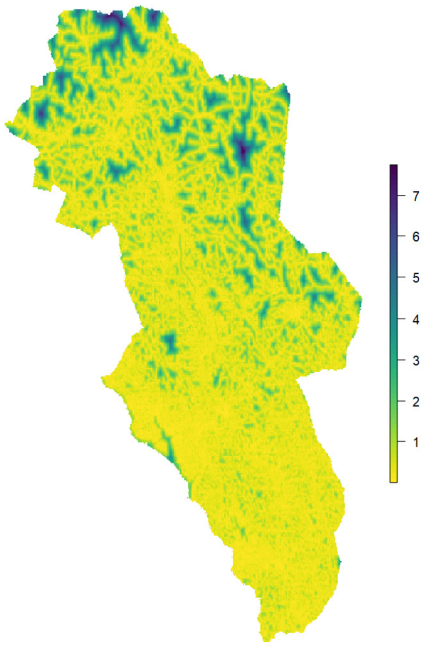


Fig. C.15. Distance to the nearest road for all locations in Hedmark.

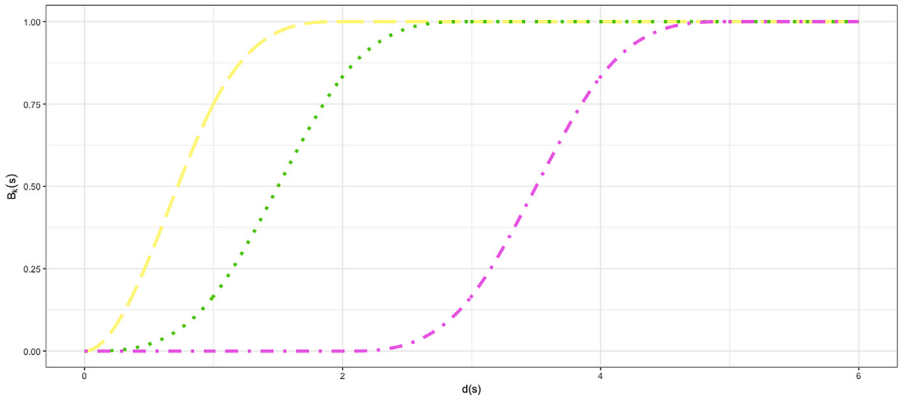


Fig. C.16. Basis functions used to fit the EVSE model for the real dataset application.

Appendix C. Moose in Hedmark application: Extra figures

See Figs. C.15–C.17.

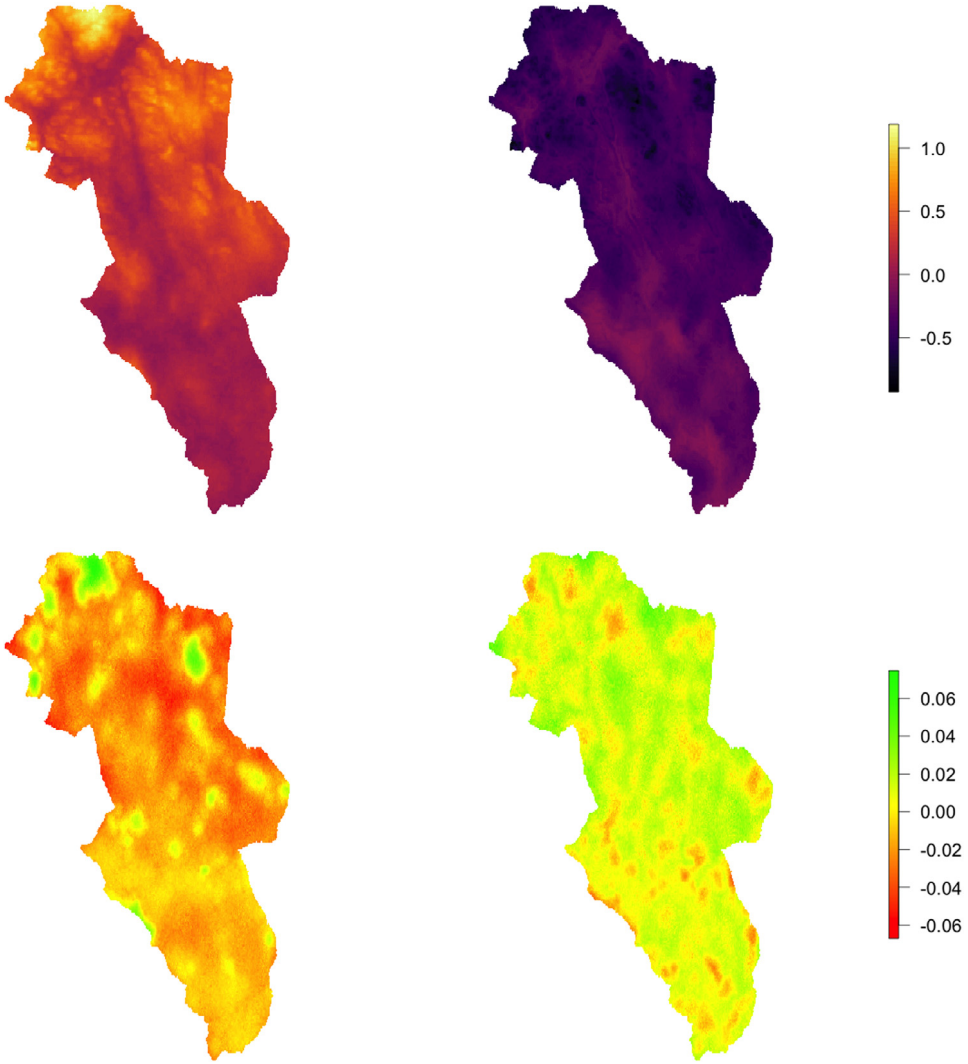


Fig. C.17. Differences in posterior median (top) and standard deviation (bottom), in log-scale, between the naive and the VSE model (left) and between the VSE and the EVSE model (right).

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/ab3s5x, 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.

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.