



Species distribution modeling: a statistical review with focus in spatio-temporal issues

Joaquín Martínez-Minaya¹ · Michela Cameletti² · David Conesa¹ · Maria Grazia Pennino³

Published online: 19 April 2018

© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

The use of complex statistical models has recently increased substantially in the context of species distribution behavior. This complexity has made the inferential and predictive processes challenging to perform. The Bayesian approach has become a good option to deal with these models due to the ease with which prior information can be incorporated along with the fact that it provides a more realistic and accurate estimation of uncertainty. In this paper, we first review the sources of information and different approaches (frequentist and Bayesian) to model the distribution of a species. We also discuss the Integrated Nested Laplace approximation as a tool with which to obtain marginal posterior distributions of the parameters involved in these models. We finally discuss some important statistical issues that arise when researchers use species data: the presence of a temporal effect (presenting different spatial and spatio-temporal structures), preferential sampling, spatial misalignment, non-stationarity, imperfect detection, and the excess of zeros.

Keywords Geostatistics · Hierarchical Bayesian models · INLA · Point processes · Preferential sampling · SPDE

1 Introduction

Understanding spatio-temporal dynamics of species or diseases is a key issue in many research areas such as ecology or epidemiology. Indeed, the so-called Species Distribution Models (SDMs), which link information on the presence/absence or abundance of a species to environmental variables to predict where (and how much of) a species is likely to be present in unsampled locations or time periods, are important tools in many applied fields.

In the particular case of ecology, SDMs have been implemented in different theoretical and practical cases,

including the identification of critical habitats (Zhang 2007; Zhang et al. 2008; Paradinas et al. 2015; Rufener et al. 2017; Sadykova et al. 2017), the study of the risk associated with invasive species (Fitzpatrick et al. 2007; Luo and Opaluch 2011), the potential effects of climate change (Iverson et al. 2004; Araújo et al. 2005; Brown et al. 2016), the design of protected areas, the protection of threatened species (Parviainen et al. 2008; Roos et al. 2015), the distribution of bioclimatic indices (Barber et al. 2017), the reintroduction of vulnerable species (Danks and Klein 2002; Martínez-Meyer et al. 2006; Hendricks et al. 2016), the delineation of hot spots of biodiversity and species richness (Jiménez-Valverde and Lobo 2007; Gotelli et al. 2009; Goetz et al. 2014), the potential distribution of infectious diseases (Peterson et al. 2002; Fatima et al. 2016; Juan et al. 2017; Martínez-Bello et al. 2017; Martínez-Minaya et al. 2018), among many others.

SDMs have also been used in many other contexts, for instance evolutionary biology, where they have been applied to topics such as speciation or hybrid zones (Kozak et al. 2008); in humans epidemiology, to predict the spread of diseases in humans (Gosoni et al. 2006), in veterinary epidemiology (González-Warleta et al. 2013; Barber et al.

✉ Joaquín Martínez-Minaya
Joaquin.Martinez-Minaya@uv.es

¹ Departamento de Estadística e Investigación Operativa, Universidad de Valencia, C/ Dr. Moliner 50, 46100 Burjassot, Valencia, Spain

² Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy

³ Centro Oceanográfico de Murcia, Instituto Español de Oceanografía, C/ Varadero 1, 30740 San Pedro del Pinatar, Murcia, Spain

2016), in plants epidemiology (Meentemeyer et al. 2011; Václavík and Meentemeyer 2009; Neri et al. 2014; White et al. 2017), etc.

Several review papers on SDMs already exist (see for example, Guisan and Thuiller 2005; Elith and Leathwick 2009), but most of them are focused on the modeling of species data, maintaining a more general overview of the statistical critical issues. Our intention in this review is to describe in more detail some of the statistical issues that arise when dealing with SDMs.

In addition, the quantity and the quality of available datasets has substantially increased over the past ten years, resulting in a higher complexity of the statistical issues that have to be addressed when a SDM is performed. Moreover, a detailed spatial and temporal description of the modeled phenomenon is becoming mandatory in many research fields. As a consequence of this increasing complexity, the performance of the SDM inferential and predictive processes are becoming more challenging, forcing researchers to develop new sophisticated statistical techniques. Accordingly, new modeling approaches continue to be developed because using only geographic information systems (GIS) tools is not totally satisfactory because of the type of spatial data usually available. Indeed, over time model complexity has generally increased over time from the use of simple environmental matching (two good examples are BIOCLIM, Busby 1991, and DOMAIN, Carpenter et al. 1993) to the use of models incorporating more complex non-linear relationships between species presence and the environment, such as generalized additive models (Guisan et al. 2002), neural networks (Park et al. 2003), or multivariate adaptive regression splines (Leathwick et al. 2005).

But more importantly, although most of the methods described in previous reviews (see for example, Guisan and Thuiller 2005; Elith and Leathwick 2009) have increased in their complexity, they are based on the assumption that the observations are conditionally-independent, while species distribution data often depict residual spatial autocorrelation (Kneib et al. 2008; Beale et al. 2010). In this review, we will focus on the fact that the spatial autocorrelation should be taken into account in species distribution models, even if the data were collected in a standardized sampling, since the observations are often close and subject to similar environmental features (Muñoz et al. 2013). Other complications also arise in the modeling of the species due to imperfect survey data such as observer error, gaps in the sampling, missing data, the spatial mobility of the species (Latimer et al. 2006) and the fact that data have been collected over long periods of time. As a consequence, ignoring these issues in this type of analysis could lead to misleading results.

As a consequence, the use of spatial and spatio-temporal models has grown enormously, allowing the incorporation of all these issues into the modeling process (Banerjee et al. 2014). Although there are other types of spatial data that could describe the behavior of a species (see for instance, Gelfand et al. 2010, for a detailed description of the three types of spatial data), we will focus in this review on geostatistical or point-referenced data that derive from those situations where the concern is to analyze spatially continuous phenomena. Bearing in mind that we want to include the effect of possible covariates in the modeling or to apply it to situations in which the stochastic variation in the data is known to be non-Gaussian, we will deal with the model-based geostatistics approach (Diggle and Ribeiro 2007).

This combination of non-Gaussian data, a linear predictor and unobserved latent variables usually makes estimation and prediction computationally difficult. Bayesian inference proves to be a good option to deal with spatial hierarchical models because it allows both the observed data and model parameters to be random variables (Banerjee et al. 2014), resulting in a more realistic and accurate estimation of uncertainty. Another advantage of the Bayesian approach is the ease with which prior information can be incorporated. Note that prior information can usually be very helpful in discriminating spatial autocorrelation effects from ordinary non-spatial linear effects (Gaudard et al. 1999). But, as is usual in Bayesian complex models, inference needs numerical approaches. Among them, in this review we will emphasize on the use of the integrated nested Laplace approximation (INLA) methodology (Rue et al. 2009) and software (<http://www.r-inla.org>) as an alternative to Markov chain Monte Carlo (MCMC) methods, the main reason being the speed of calculation.

To summarize, our intention in this review is to describe in more detail the main statistical issues that arise when dealing with these models. In particular, in Sect. 2 we focus on the statistical aspects of the available data, while Sect. 3 discusses the basic structure of these models and how to perform inference. In particular, we provide a critical review of the Bayesian approach along with a detailed description of INLA. Our review also includes a discussion on some of the particularities appearing when dealing with them, including temporal correlation, preferential sampling, spatial misalignment, non-stationarity, imperfect detection and excess of zeros in Sect. 4. Finally Sect. 5 concludes. To be noted is that we have tried to be simple in the notation so that the paper is readable by a large community of scientists.

2 Sources of information in SDMs

SDMs require basically two types of data input: data describing the observed species' distribution, and data describing the landscape and the environmental characteristics in which the species can be found. In this Section we first present biological data, i.e. the observed species distribution, and then the environmental data and the usual covariates that characterize the species distribution.

2.1 Biological data

The first type of data, which usually represent the response variable, can be either presence-only (i.e. records of localities where the species has been observed), presence/absence (i.e. records of presence and absence of the sampling localities), abundance data (i.e. the quantity of the species at the sampling locations), or proportional data (i.e. the proportion of the species at the sampling locations). Consequently, biological data can be measured at nominal (e.g. presence/absence type), ordinal (e.g. ranked abundance), ratio (e.g. frequency of detection) or continuous (e.g. abundance, richness) levels, which impacts on the selection of the appropriate types of modeling algorithms to use, and subsequently the measurement level of model of this kind (e.g. probability or suitability of occurrence, type, expected mean).

Presence-only data lack absence observations, so that this type of dataset is unsuitable for many of the commonly used species distribution algorithms, unless *pseudo-absences* are assigned to unsampled portions of the study area. Inclusion of *pseudo-absences* records can seriously bias analyses. Indeed, methods used to generate pseudo-absences and their effects on model performance are an open research field in the species distribution context (Barbet-Massin et al. 2012; Iturbide et al. 2015).

With respect to abundance, this could be expressed as a continuous variable (biomass of the species) or as count data (number of individuals). Abundance data reflect the quantitative spatial distribution of the species within the area of interest, while presence/absence information can be used to measure the relative occurrence of species, thereby giving a different approximation. Although abundance data provide greater information for conservation and management purposes, they are less common, because occurrence data are easier and less expensive to be collected. Indeed, abundance estimations are sensitive to detectability, and sampling methods seldom detect all individuals present in an area. Consequently, many research studies rely on approximations of species abundance from species occurrence, although whether abundance can be inferred from

such information has been questioned, because detection is not perfect and occurrence probability may not be linearly related to density (Nielsen et al. 2005; Joseph et al. 2006).

Proportional data are also widely used in many ecological processes. The traditional approach in ecology is based on Gaussian linear models with previous transformation in the proportions. However, model parameters cannot be easily interpreted in terms of the original response, and measures of proportions typically display asymmetry: hence, inferences based on the normality assumption can be misleading (Ferrari and Cribari-Neto 2004). Beta regression has recently appeared as a good alternative to deal with data of this type, allowing bounded estimates and intervals with model parameters that are directly interpretable in terms of the mean of the response (Paradinas et al. 2016, 2017b).

Also to be noted is that different species do not behave independently. There are several species whose abundance (or presence) is constrained by competition: a large increase in one is unavoidably linked to declines in others. In these cases, the response variable should be considered by using a joint distribution. The models used for data of this type are known as joint species distribution models (Clark et al. 2014; Pollock et al. 2014; Hui 2017; Taylor-Rodríguez et al. 2017).

All these types of biological data describing the observed species distribution can be obtained in a variety of ways, such as museum collection, designed field surveys, related activities (i.e. fisheries) or on-line resources.

2.2 Environmental data

With respect to the explanatory variables that could help to describe the species behavior, a wide range of environmental variables have been usually incorporated in SDMs. These variables are commonly related to climate (e.g. temperature, precipitation), topography (e.g., elevation, aspect, bathymetry, slope of the seabed), land cover type or seabed type in marine ecosystems. Variables tend to describe primarily the abiotic environment, although there is potential to include biotic interactions within the modeling.

These variables can be collected in situ, but they are usually derived from remoted sensing data. CRU (New et al. 2002), WorldClim (Hijmans et al. 2005), and MARSPEC (Sbrocco and Barber 2013) are all examples of spatially explicit datasets of climatic remote sensing conditions. These datasets encompass climatic information based on interpolations from global weather stations. However, interpolations are only as good as the underlying data, and uneven geographical coverage leads to high model uncertainty, especially in developing countries where few weather stations are in place (Daly 2006; He

et al. 2015). When uncertainty in spatial climate variables is not accounted for, coefficient estimates tend to be biased, and this leads to poor performances of the SDMs, as recently shown with simulations by Stoklosa et al. (2015). This problem, also known as misalignment, is treated in this review in Sect. 4.3.

3 Inference

In what follows, after presenting the traditional methods that have been used to perform inference in SDMs, we first discuss the hierarchical modeling as one of the most flexible and encompassing approaches to deal with them. The second subsection presents the Bayesian framework as a good option for dealing with hierarchical models. The final subsection deals with the INLA approach to approximating the marginal posterior distributions of the parameters involved in the SDMs.

3.1 Gaussian fields and hierarchical modeling

A number of alternative modeling algorithms have been applied to classify species distribution as a function of a set of environmental variables. A first group of methods developed to deal with presence-only datasets includes maximum entropy algorithm, environmental distance, similarity, and envelope methods such as MAXENT (Phillips et al. 2006), Gower metric, Mahalanobis distance, and ecological niche factor analysis, all of which describe some measure of habitat suitability.

A second group involves machine-learning algorithms that are iterative in nature, such as artificial neural networks. These ensemble methods (e.g. Boosting Regression Trees, Classification Trees and Random Forests) generally involve developing multiple models on different subsets of the data, the results of which are averaged (Franklin 2010).

A third group of methods relates to traditional regression and includes generalized linear models (GLM) and their non-parametric extension, generalized additive models (GAM), both of which can handle several measurement levels of the response variable by using a different link function (e.g. logistic for presence/absence or log for counts). GAM and a related method, multivariate adaptive regression splines (MARS), are more flexible than GLM as they are fitted using smoothing and piecewise linear splines, respectively, and are particularly useful for identifying the shape of species responses (Leathwick et al. 2005). MARS is computationally faster than GAM and the results are more easily converted to map predictions in a GIS; however, the currently used algorithms require normally distributed error terms. This makes MARS unsuitable for use with presence/absence data unless the basis

functions are extracted and used to parameterize a GLM (Leathwick et al. 2005). Rodríguez de Rivera and López-Quílez (2017) present a comparison of these three groups of methodologies stating that GAM models gave the best results.

However, most of the above mentioned methods are based on the assumption that the observations are conditionally-independent. But this is not always the case because data of species distribution usually present residual spatial autocorrelation (Kneib et al. 2008). GAMs and MARS can model spatial and temporal autocorrelation using smoothing splines. A very powerful and flexible alternative is to incorporate this spatial relationship by considering the species distribution data as point-referenced or geostatistical data. Data of this type appear in those situations where the interest is to analyze spatially continuous phenomena. The most basic format for data of this kind is a pair composed by the spatial location coordinates defined throughout a continuous study region and the measurement value observed in the location. Geostatistical data require methods that make it possible to relate the species data with potential related covariates by quantifying the spatial dependence. However, one of the main interests in geostatistics concerns predicting the underlying process on those non-observed locations (Cressie and Wikle 2011; Banerjee et al. 2014).

Geostatistical or point-referenced data can be seen as realizations of a spatial process (random field) $\{y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ characterized by a spatial index \mathbf{s} which varies continuously in the fixed domain \mathcal{D} . This process is called a Gaussian field (GF) if for any $n \geq 1$ and for each set of locations $(\mathbf{s}_1, \dots, \mathbf{s}_n)$, the vector $(y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_n))$ and with covariance matrix $\boldsymbol{\Sigma}$ defined by a covariance function $\mathcal{C}(\cdot, \cdot)$, such that $\Sigma_{ij} = Cov(y(\mathbf{s}_i), y(\mathbf{s}_j)) = \mathcal{C}(y(\mathbf{s}_i), y(\mathbf{s}_j))$. If the mean is constant in space, i.e. $\mu(\mathbf{s}_i) = \mu$ for each i , and the generic spatial covariance matrix element depends only on the difference vector $(\mathbf{s}_i - \mathbf{s}_j) \in \mathbb{R}^2$, the spatial process is second-order stationary. In addition, if the covariance function only depends on the Euclidean distance $\|\mathbf{s}_i - \mathbf{s}_j\|$, the process is said to be isotropic.

In a hierarchical framework, the first step in defining a model for a random field is to identify a probability distribution for the observations available at n spatial locations and represented by the vector $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n)) = (y_1, \dots, y_n)$ (the notation is simplified and the index i is used for denoting the generic spatial points \mathbf{s}_i). At the first level of the hierarchy, we usually select a distribution from the exponential family, characterized by a set of parameters. These parameters are linked with a linear predictor which also includes a latent

GF denoted by $\zeta(s)$ whose covariance function Σ depends on two parameters: σ^2 which represents the variance (partial sill in kriging terminology) and the range ϕ of the spatial effect.

Computational costs required to estimate these parameters are high when we deal with the spatial covariance function because the generated matrices are dense. This problem is known as the “big n problem” (Banerjee et al. 2014; Jona Lasinio et al. 2013) and despite computational power today, it is still a computational bottleneck in many situations. A computationally effective alternative is given by the stochastic partial differential equation (SPDE) approach proposed by Lindgren et al. (2011) (see Sect. 3.3).

In addition to the spatial pattern, the temporal variation could be equally important because the phenomenon can vary not only in space but also in time (see Hefley and Hooten 2016, for a comprehensive overview of modeling species distribution with a spatio-temporal perspective). Then, extending the spatial case to the spatio-temporal case including a time dimension, the process indexed by space and time can be defined as $\{y(s, t), (s, t) \in \mathcal{D} \subset \mathbb{R} \times \mathbb{R}\}$, and is observed at n spatial locations and at T time points.

The general structure for modeling the spatial distribution of species is given by the following formulation and notation. If $\mathbf{y} = (y_1, \dots, y_n)$ represents the observed values of the corresponding response variable Y with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$, each μ_i can be easily linked to a structured additive predictor η_i through a link function $g(\cdot)$, so that $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. The structured additive predictor $\boldsymbol{\eta}$ accounts for the effect of various covariates in an additive way:

$$\eta_i = \beta_0 + \sum_{m=1}^M \beta_m x_{mi} + \sum_{l=1}^L f_l(z_{li}), \tag{3.1}$$

where β_0 corresponds to the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_M\}$ quantify the (linear) effect of some covariates $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_M)$ on the response; and $\mathbf{f} = \{f_1(\cdot), \dots, f_L(\cdot)\}$ are unknown functions of the covariates $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_L)$, and can assume different forms such as smooth nonlinear effects of covariates, time trends and seasonal effects, random intercept and slopes as well as temporal or spatial random effects. Note that this general structure can also be seen as a Generalized Additive Mixed Model (GAMM). Also to be noted is that here it is assumed that covariates are observed at the same locations of the response variable. The situation where covariates are observed in locations different from those of the response variable (misalignment) will be discussed in Sect. 4.3.

In many statistical applications, in particular, in SDMs, the model involves multiple parameters that can be regarded as related or connected in some way by the structure of the problem, implying that a joint probability model for

these parameters should reflect their dependence (Gelman et al. 2014). It is common to model such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which in turn are given a probabilistic specification in terms of further parameters, adding various levels of the modeling and thus defining a hierarchical model (HM). Note that Hierarchical models provide a generalization of all the models presented here; and moreover that they are able to deal with all the types of the data that we can be found when dealing with SDMs. Table 1 describes all the models mentioned in this subsection along with a diagram emphasizing their nested nature.

Although other approaches can be used such as maximum likelihood (MLE; Le Cam 1990), restricted maximum likelihood (RMLE; Bartlett 1937), quasi-maximum likelihood (QMLE; Cox and Reid 2004), the method of moments (Bowman and Shenton 2006), the generalized method of moments (GMM; Hansen 1982), M-estimators (Shapiro 2000), the maximum spacing estimation (MSE; Anatolyev and Kosenok 2005), etc., here we will focus on the Bayesian approach to making inference for hierarchical models with a linear predictor of the form (3.1).

3.2 Bayesian approach

The use of the Bayesian framework as a way to make inference has increased in the past 50 years and it has been applied in different areas, such as social sciences (Jackman 2009), medicine and public health (Berry and Stangl 1999), finance (Rachev et al. 2008), ecology (McCarthy 2007), bioinformatics (Mallick et al. 2009), health economics (Baio 2012), physical sciences (Andreon and Weaver 2015) and econometrics (Gómez-Rubio et al. 2014). Bayesian reasoning is based on the assumption that parameters are random variables, and prior knowledge has to be incorporated via the corresponding prior distributions of the said parameters. Bayes’ theorem is the tool that combines prior information with the likelihood yielding the posterior distributions. To be noted is that the Bayesian approach is perfectly suited for complex spatial models such as SDMs because it allows model parameters to be random variables, resulting in a more realistic and accurate estimation of uncertainty.

SDMs are a very good example of a hierarchical structure that can be expressed as a hierarchical Bayesian model (Wikle and Hooten 2010; Hefley and Hooten 2016). They can be structured in three levels: the first one refers to the data and is conditioned on the process and parameters in whatever aspects of the process are appropriate. The second level contains the latent components, which can be spatial and/or dynamic and the stochastic form can be univariate or multivariate. Finally, the third stage defines

Table 1 Matching of models presented and data types

Explanatory variable (s)	Response variable distribution	
	NORMAL	OTHER DIST. EXP. FAMILY
LP	LM	GLM
R. effects	LMM	GLMM
Non-Lin. effects	AM	GAM
R. effects + Non-Lin. effects	AMM	GAMM

LM linear models, *LMM* linear mixed models, *GLM* generalized linear models, *GLMM* generalized linear mixed models, *AM* additive models, *AMM* additive mixed models, *GAM* generalized additive models, *GAMM* generalized additive mixed models, *HM* hierarchical models

By construction, these models are nested: LM < GLM < GAM < GAMM < HM

the priors for the parameters on which the latent processes depend. The parameters in this level are also known as hyperparameters.

The approach most commonly used to perform Bayesian inference for spatial species distribution models is based on MCMC methods (Gelfand et al. 2006); they are flexible computational tools which can be easily adapted to any kind of inferential problem. The software most frequently used to implement MCMC algorithms are WinBUGS (Lunn et al. 2000; Brooks et al. 2011), OpenBUGS (Lunn et al. 2009) and JAGS (Plummer 2016), which can also be run within other programs like R (through the R2OpenBUGS, R2WinBUGS, BRugs and rjags packages), Stata and SAS. Alternatively other R packages are BayesX (Brezger et al. 2003), CARBayes (Lee 2013), stocc (for binary data only), spatcounts (for count data only), CAR-ramps (for Gaussian data only), and spdep (for Gaussian data only). Several hierarchical models including ecological processes (habitat suitability, spatial dependence and anthropogenic disturbance) and observation processes (species detectability) can also be performed using the hSDM package of R developed by Vieilledent et al. (2014). Functions in this R package use an adaptive Metropolis algorithm (Robert and Casella 2011) in a Gibbs sampler (Gelfand and Smith 1990) to obtain the posterior distribution of model parameters. The Gibbs sampler is written in C code and compiled to optimize computation efficiency. Thus, the hSDM package can be used for very large data-sets while drastically reducing the computation time. However, with hSDM it is not possible at present to model spatio-temporal or proportion response variables.

Despite their generalized use, to be noted is that MCMC methods still have many challenges to deal with (like the so-called “big n problem” mentioned above; see Banerjee et al. 2014; Jona Lasinio et al. 2013). Indeed, they can be extremely slow and even computationally unfeasible especially when the models are extremely complex (with many random effects or hierarchical levels) or when big datasets are considered in the space-time setting.

As a result, other options have appeared to make inference in SDMs. Taking advantage of the hierarchical structure of SDMs, Golding and Purse (2016) propose the use of an empirical Bayesian approach. In particular, they maximize the marginal posterior density of the model, which, in their words, enables the incorporation of prior knowledge over hyperparameters whilst being much less computationally intensive than fully Bayesian inference.

Here, we will focus on the integrated nested Laplace approximation (INLA) methodology (Rue et al. 2009), as a computational effective alternative to MCMC. Our choice is due to two considerations: the speed of calculation, and the ease with which model comparison can be performed.

3.3 INLA and SPDE framework

The INLA methodology is now a well-established tool for Bayesian inference in several research fields, including ecology, epidemiology, econometrics and environmental science (Rue et al. 2017). It can be used through R with the R-INLA package. For more details on INLA for spatial and spatio-temporal models we refer the reader to Blangiardo et al. (2013) and Blangiardo and Cameletti (2015), where practical examples and code guidelines are also provided.

The reason why INLA can be used is that SDMs can be seen as latent Gaussian models (Rue and Held 2005), for which the class of models INLA is designed. After identifying the distribution for the observed data, we can link its corresponding mean to the linear predictor as in Eq. (3.1). If conditional independence is assumed, the distribution of the n observations is given by the likelihood

$$p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^n p(y_i | \theta_i, \boldsymbol{\psi}), \quad (3.2)$$

where $\boldsymbol{\theta}$ represents the set of latent (nonobservable) components of interest $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \mathbf{f}\}$, also known as the latent field, and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)$ denotes the vector of K hyperparameters. As we can observe in Eq. (3.2), each data

point y_i is connected to one element θ_i in the latent field. This assumption can be relaxed, and each observation can be connected with a linear combination of elements in θ (Martins et al. 2013). In addition, the multiple likelihood case can also be taken into account.

In the context of latent Gaussian models, assumed is a multivariate Normal prior distribution on θ with mean θ and precision matrix $\mathbf{Q}(\psi)$, i.e., $\theta \sim N(\theta, \mathbf{Q}^{-1}(\psi))$ with density function given by

$$p(\theta | \psi) = (2\pi)^{-n/2} |\mathbf{Q}(\psi)|^{1/2} \exp\left(-\frac{1}{2} \theta' \mathbf{Q}(\psi) \theta\right), \quad (3.3)$$

being $|\cdot|$ the matrix determinant and $'$ the transpose operation. When the precision matrix $\mathbf{Q}(\psi)$ is sparse a GF becomes a Gaussian Markov random field (GMRF, Rue and Held 2005). Interestingly, when making inference with GMRFs, linear algebra operations are performed using numerical methods for sparse matrices, and this yields computational benefits.

In spite of the wide acceptance of INLA, its precision and its computational efficiency in many latent Gaussian models (see for instance, Martino et al. 2011; Schrödle et al. 2011; Ruiz-Cárdenas et al. 2012, for a description of how to use INLA in spatio-temporal disease mapping, in state-space models and in survival models, respectively), INLA cannot be directly applied when dealing with models that incorporate geostatistical data (that is, continuously indexed Gaussian Fields). The underlying reason is that a parametric covariance function needs to be specified and fitted based on the data, which determines the covariance matrix Σ and enables prediction in unsampled locations. But from the computational perspective, the cost of factorizing the dense covariance matrix Σ is cubic in its dimension. Despite current computational power, in many situations it is still challenging to factorize it for computing the inverse and the determinant.

Lindgren et al. (2011) proposed an alternative approach by using an approximate stochastic weak solution to a Stochastic Partial Differential Equation (SPDE) as a GMRF approximation to a continuous Gaussian Field (GF) with Matérn covariance structure. Specifically, they used the fact that a Gaussian Field $\zeta(\mathbf{s})$ with Matérn covariance is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} (\tau \zeta(\mathbf{s})) = \mathcal{W}(\mathbf{s}), \quad (3.4)$$

$$\mathbf{s} \in \mathbb{R}^d, \alpha = \nu + \delta/2, \kappa > 0, \nu > 0,$$

where Δ is the Laplacian, α controls the smoothness, κ is the scale parameter, τ controls the variance, and $\mathcal{W}(\mathbf{s})$ is a Gaussian spatial white noise process. The exact and stationary solution to this SPDE is the stationary GF $\zeta(\mathbf{s})$ with Matérn covariance function given by:

$$\begin{aligned} \text{Cov}(\zeta(\mathbf{s}_i), \zeta(\mathbf{s}_j)) &= \mathcal{C}(\xi_i, \xi_j) \\ &= \frac{\sigma^2}{2^{v-1} \Gamma(v)} (\kappa \|\mathbf{s}_i - \mathbf{s}_j\|)^v K_v(\kappa \|\mathbf{s}_i - \mathbf{s}_j\|), \end{aligned} \quad (3.5)$$

being $\|\mathbf{s}_i - \mathbf{s}_j\|$ the Euclidean distance between two locations $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^d$, and σ^2 the marginal variance. Moreover, K_v is the modified Bessel function of the second kind and order $v > 0$, which measures the degree of smoothness of the process. This parameter is usually kept fixed due to its poor identifiability. Conversely, $\kappa > 0$ is a scaling parameter related to the distance at which the spatial correlation becomes almost null, i.e., the range (for more information on the Matérn covariance model see Handcock and Stein 1993; Stein 1999). Typically, as pointed out in Lindgren et al. (2011), the empirically derived definition for the range is $r = \frac{\sqrt{8\nu}}{\kappa}$, with r corresponding to the distance at which the spatial correlation is close to 0.1, for each $\nu \geq \frac{1}{2}$.

The link between Eqs. (3.4) and (3.5) is given by the expressions $\nu = \alpha - \frac{\delta}{2}$, and $\sigma^2 = \frac{\Gamma(v)}{\Gamma(\alpha)(4\pi)^{\delta/2} \kappa^{2\nu} \tau^2}$. In the particular case where the dimension is 2, i.e., $\delta = 2$, it follows that $\nu = \alpha - 1$ and $\sigma^2 = \frac{\Gamma(v)}{\Gamma(\alpha)(4\pi)\kappa^{2\nu}\tau^2}$.

Finally, in R-INLA, the Gaussian field $\zeta(\mathbf{s})$ is found numerically as a weak solution to the SPDE in (3.4), and by default the smoothness parameter α is fixed to 2, corresponding with $\nu = 1$. With this assumption, the range is given by $\phi \approx r = \sqrt{8}/\kappa$, while the variance is given by $\sigma^2 = 1/(4\pi\kappa^2\tau^2)$.

Bayesian geostatistical analysis using R-INLA has already been applied in various contexts. Along with introducing the `geostatsinla` package for performing geostatistics with INLA in an easy way, Brown (2015) applies it in the context of mapping the *Loa loa* filiarasis disease (a dataset previously cited in Diggle and Ribeiro 2007). Moreover, Karagiannis-Voules et al. (2013) have used Bayesian geostatistical negative binomial models to analyze reported incidence data of cutaneous and visceral leishmaniasis in Brazil covering a 10-year period, while González-Warleta et al. (2013) have used Bayesian geostatistical binomial models to predict the probability of infection of paramphistomosis in Galicia (NW Spain). In the context of fisheries, Bayesian geostatistical analysis using R-INLA has also been used to predict the presence/absence, the abundance, or the proportion of fish species (Muñoz et al. 2013; Pennino et al. 2013, 2014, 2016, 2017; Paradinas et al. 2015, 2016; Cosandey-Godin et al. 2015; Quiroz et al. 2015; Roos et al. 2015; Rufener et al. 2017).

4 Extending statistical modeling of species distribution

There are a number of additional potential sources of bias and error that should be carefully considered when analyzing and modeling species distribution data. Errors may arise through the incorrect identification of species, or inaccurate spatial referencing of samples. Biases can also be introduced because collectors tend to sample in easily accessible locations. Here we discuss some of these issues.

4.1 Temporal autocorrelation

As mentioned above, in addition to the spatial pattern, the temporal variation could be equally important because the phenomenon may vary not only in space but also in time. This happens in problems such as the evolution of epidemics (Stein et al. 1994; Hefley et al. 2017b), the spatio-temporal evolution of temperature (Hengl et al. 2012) or the understanding of the spatial dynamism of species over time (Wikle 2003; Hooten et al. 2007; Hooten and Wikle 2008; Paradinas et al. 2015, 2017a; Williams et al. 2017).

As pointed out by Cressie and Wikle (2011), temporal correlation depends on the same principle as spatial correlation: temporally close observations tend to be more related than temporally distant ones. Consequently, model fitting and predictions improve when a temporal term is added. However, temporal and spatial scales are different and the spatio-temporal analysis is more complicated than the simple addition of an extra dimension to the continuous spatial domain.

In the context of species distribution modeling, most studies (surveys, plant coverage surveys, air pollution surveys, etc.) have been repeated periodically for long periods of time (Gitzen 2012; Aizpurua et al. 2015). Although the main interest is the spatial evolution of the system under study, it must be considered that it varies not only in space but also in time. Here we focus on this most common situation of discrete and regular time observations. For situations in which data are collected in irregular time-lags—that is, when the issue is handling continuous-time data—a good option is to consider 1D SPDE models with a second order B-Spline basis representation (Lindgren and Rue 2015a, b).

The spatio-temporal behavior of the data can vary depending on the nature of the process under study and the available sampling resolution. In particular, the basic model in (3.1) can be rewritten by splitting the f term into two terms, one indicating different possible spatio-temporal structures, and the other indicating any other latent model or non-linear effect. If y_{it} represents the response variable analyzed at location s_i ($i = 1, \dots, n$) at time t

($t = 1, \dots, T$), then the mean of the response variable μ_{it} is linked to the linear predictor with a link function $g(\cdot)$, as

$$\eta_{it} = g(\mu_{it}) = \beta_0 + \sum_{m=1}^M \beta_m x_{mit} + \sum_{k=1}^K f_k(z_{kit}) + u_{it}, \quad (4.1)$$

where β_0 corresponds to the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_M\}$ quantify the linear effect of some covariates on the response; u_{it} represents the spatio-temporal structure of the model; z_{kit} is the value of the k -th explanatory variable at a given location s_i and time t ; and f represents any latent model applied to the covariates.

Among other structures, and following Paradinas et al. (2017a), we comment here on four basic structures for u_{it} , each one allowing for different degrees of flexibility in the temporal domain of the spatio-temporal model. Paradinas et al. (2017a) provide a figure that schematically illustrates all these structures:

- **Opportunistic spatial distribution** this flexible structure consists in expressing u_{it} as different spatial realizations $\mathbf{w}_t = \{w_{1t}, \dots, w_{it}, \dots, w_{nt}\}$ of the same spatial field for each time unit t , while sharing a common covariance function (same κ and τ) to avoid overfitting:

$$\begin{aligned} u_{it} &= w_{it}, \\ \mathbf{w}_t &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\kappa, \tau)). \end{aligned} \quad (4.2)$$

This structure is a good approximation for processes where the spatial distribution varies considerably among different time units and unrelatedly among neighboring times. This structure has been used in Cosandey-Godin et al. (2015) and in Paradinas et al. (2015).

- **Persistent spatial distribution with random intensity changes over time** when the pattern of spatial variation persists over time, but with possibly varying scales of intensity, a time structure is introduced into the model using a zero mean Gaussian random noise effect v_t . In this case, u_{it} is decomposed in a common spatial realization w_{it} along with an independent random noise effect v_t that absorbs the different mean intensities at each time t :

$$\begin{aligned} u_{it} &= w_{it} + v_t, \\ \mathbf{w}_t &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\kappa, \tau)), \\ v_t &\sim N(0, \tau_v^{-1}). \end{aligned} \quad (4.3)$$

For processes where the spatial component persists in time, this structure may be the most suitable. It has been used by Pennino et al. (2014) and in Paradinas et al. (2015).

- **Persistent spatial distribution with temporal intensity trend** the process could show a temporal progression in its mean. To model that, a temporal trend effect $h(t)$ can be added to the linear predictor. In this case, u_{it} is decomposed into a common spatial realization w_i and an independent temporal structured trend $h(t)$ to absorb the temporal progression of the process:

$$\begin{aligned} u_{it} &= w_i + h(t), \\ \mathbf{w} &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\kappa, \tau)). \end{aligned} \tag{4.4}$$

This structure is highly recommended in situations where a temporal tendency is present. It was proposed by Paradinas et al. (2016) to identify intra-annual trends in fishery discards.

- **Progressive spatio-temporal distribution** this structure incorporates both spatial and temporal correlation of the data to accommodate those cases where the spatial realizations change in a related manner over time. Here, u_{it} is decomposed into a common spatial realization w_{it} and an autoregressive temporal term r_{it} expressing the correlation among temporal neighbors of order K :

$$\begin{aligned} u_{it} &= w_{it} + r_{it}, \\ \mathbf{w}_t &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\kappa, \tau)), \\ r_{it} &\sim N\left(\sum_{k=1}^K \rho_k r_{i(t-k)}, \tau_r^{-1}\right). \end{aligned} \tag{4.5}$$

This structure is preferred when the spatial realization varies between different times but not as much as in (4.3). Indeed, the structure has been used by Cameletti et al. (2011, 2013) and also by Cosandey-Godin et al. (2015).

Note that this list is only an overview of the different spatio-temporal structures which allow us to discern the nature of the general spatial behavior of the process over time. Unfortunately, the temporal resolution of spatio-temporal datasets is typically too low to fit most of the highly structured models.

4.2 Preferential sampling

In studies on species distributions, collecting data on the species of interest is not a trivial problem. With the exception of a few studies, species distribution models rely on opportunistic data collection due to the high cost and time-consuming nature of collecting data in the field, especially on a large spatial scale. As an example, studies on bird monitoring data are often collected by volunteers who concentrate the sampling process on areas where they expect to find species of interest. These types of

opportunistically collected data tend to suffer from a specific complication: the sampling process that determines the data locations and the species observations are not independent (Diggle et al. 2010). Statistical models used for species distribution usually assume, if only implicitly, that sampling is non-preferential and that the selection of the sampling locations does not depend on the values of the spatial variable. However, opportunistic data are a clear example of preferential sampling, that occurs because sampling locations are deliberately chosen in areas where the values of the species of interest are thought likely to be particularly high or low (Diggle et al. 2010).

Hence, applying standard geostatistical methods to preferentially sampled data potentially yields biased results if the choice of monitoring locations is not accounted for in the modeling process. A possible approach to correct this issue is to interpret the data as a marked point pattern (Fortin and Dale 2005; Diggle 2013) where the sampling locations form a point pattern and the observations taken in those locations are the marks. By assuming that the intensity of the point process depends on the amount of species of interest, the marks and the pattern become not independent.

A preferential sampling model can be considered as a two-part model that share information. Firstly, it is supposed that the observed locations (s_1, \dots, s_n) come from a non-homogeneous Poisson process with intensity $A_i = \exp\{\alpha_1 + w_i\}$, i.e., a log-Gaussian Cox process (LGCP; Fortin and Dale 2005; Diggle 2013) is assumed, being α_1 the intercept of the LGCP and w_i the spatial effect of the model and $i = 1, \dots, n$ the index corresponding to the s_i location. Secondly, the species characteristic (usually the abundance) y_i is assumed to follow an exponential family distribution (such as a Normal or a Gamma distribution when dealing with abundances, although other options such as exponential, lognormal, etc., could clearly be possible), whose mean is related with the spatial term using a link function $g(\cdot)$, $g(\mu_i) = \alpha_2 + \beta w_i$, being α_2 the intercept of the model and w_i the spatial term shared with the LGCP, but scaled by β to allow for the differences in scale between the abundances and the LGCP. More formally, the model can be expressed as follows:

$$\begin{aligned} y_i &\sim F(\mu_i, \gamma^2) \\ g(\mu_i) &= \alpha_2 + \beta w_i \\ \mathbf{w} &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\kappa, \tau)) \end{aligned} \tag{4.6}$$

where $\mathbf{w} = \{w_1, \dots, w_n\}$, the precision matrix $\mathbf{Q}(\kappa, \tau)$ is computed internally by the SPDE approach and represents the GMRF approximation to the continuous GF (see Illian et al. 2012; Krainski et al. 2017; Pennino et al. 2018, for details about how to implement these models within

INLA), and $F(\mu, \gamma)$ represents a distribution coming from the Exponential family with mean μ and variance γ^2 .

4.3 Spatial misalignment

A crucial issue in studying the effect of environmental physical factors on species distribution concerns spatial misalignment (Clark and Gelfand 2006; Gelfand et al. 2010; Foster et al. 2012; Miller 2012).

This occurs when the response biological variable (e.g. presence/absence of the species) is observed in locations which are different from the spatial points where covariate data are available. Additionally, it can happen that covariates have a different spatial scale if they are defined at the area or cell grid level (as in the case of remote sensing data).

The naïve solution for spatial misalignment is a two-stage approach: the first step consists in the prediction of the covariate in the spatial locations where the response variable is observed (through a geostatistical model by means of kriging or inverse-distance weighting) or in the downscaling of the gridded covariate to the point-level resolution (usually considered is the value of the cell where the spatial point is located). Then, at the second stage, these predicted values are plugged into the linear predictor (3.1) as known constants. The problem with this approach is that it does not take account of the uncertainty related to the covariate spatial estimation of the first stage, with the consequence of erroneous inference of the statistical model and a potential biased estimate of the environmental variable effect on the response variable (Foster et al. 2012).

A solution to incorporate the spatial prediction uncertainty in SDMs consists in implementing one of the so-called *errors-in-variables models* (Carroll et al. 2016) which can be estimated in a frequentist (by means of the EM-algorithm) or Bayesian framework (with MCMC or INLA). If we assume for example that the predicted covariate is a noisy version of the true one, a classic measurement error model can be adopted (Stoklosa et al. 2015). Otherwise, a Berkson-error model can be considered if the predicted covariate is a smoothed (i.e. less variable) version of the true variable (Foster et al. 2012). As reported in Stoklosa et al. (2015) “Which of these two types of error models to consider will depend on what the analyst believes to be the true underlying explanatory variable, and how the data were collected/measured. The analyst must take into account: how and whether the species responds to a particular climate observation (Berkson); or that it might respond to an average, such that relatively minor deviations from this are immaterial (classical)”.

Another alternative to the two-stage approach is the joint modeling strategy implemented in Barber et al. (2016) to

evaluate the presence of the *Fasciola hepatica* in Galicia (Spain) using the annual mean temperature as covariate. In this case a spatial geostatistical model is specified for the covariate and is estimated jointly with the species distribution models in a Bayesian context. The joint model is specified as follows

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 \phi_i + w_i \\ w &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\kappa, \tau)) \\ x_i &\stackrel{iid}{\sim} N(\phi_i, \sigma_x^2) \\ \phi &\sim N(\mathbf{0}, \mathbf{Q}^{-1}(\gamma, \delta)) \end{aligned} \quad (4.7)$$

where π_i is the probability of occurrence at site s_i , x_i is the covariate of interest whose spatial distribution is specified through its mean (a realization of the Matérn Gaussian process ϕ depending on the parameters γ and δ), and through its variance σ_x^2 , which is introduced to express any possible measurement error. The model also includes another spatial process for the response represented by w . This kind of model pertains to the latent Gaussian model family and can be estimated using the SPDE-INLA approach (see Blangiardo and Cameletti, 2015, Chap. 8 and Muff et al. 2015). The advantage is that this joint model allows to properly propagate all the uncertainty related to the covariate prediction; on the other it can be extremely computationally expensive especially when there is more than one explanatory variable.

Finally, another alternative is the one proposed by Gómez-Rubio and Rue (2017) that, using a more general approach, deals with missing values in the covariates, based on fitting conditional latent Gaussian models where covariates are imputed using a Metropolis-Hastings algorithm.

4.4 Non-stationarity

The Matérn spatial covariance function $\mathcal{C}(\cdot, \cdot)$ specified by Eq. (3.5) enjoys the second-order stationarity and isotropy property, i.e. it depends only on the distance between the spatial locations and not on the direction or the coordinates. In some situations, this stationarity assumption, which is very convenient to simplify the inferential procedures, may not be suitable. For example, for some applications it is not realistic to assume that the spatial dependence structure is the same throughout the domain considered, especially when geographical elements or physical barriers (river, lakes, islands, etc.) exist. In such situations characterized by spatial heterogeneity and barriers, it may be more reasonable to adopt a non-stationary Gaussian field (see Gelfand et al. 2010, Chapter 9 and Risser 2016 for a review).

In ecological applications, heterogeneity in space (i.e. non-stationarity) occurs when a latent *global* process is also affected by some underlying local processes (Miller 2012). A local modeling technique to include this heterogeneity in SDMs is given by the geographically weighted regression (GWR) characterized by covariate coefficients which vary spatially and are specific for each spatial location; a spatial kernel function is used to define spatial neighborhoods (see e.g. Brunsdon et al. 1998; Windle et al. 2010; Holloway and Miller 2015; Liu et al. 2017). Some authors do not completely agree with the use of these models due to the large degree of multicollinearity that their coefficients tend to exhibit, as well as strong positive spatial autocorrelation. As an alternative, spatial filtering provides a methodology for dealing better with multicollinearity, while accounting for spatial autocorrelation (see e.g. Griffith 2008). The Bayesian counterpart of GWR models, which are usually estimated by weighted least squares, is given by spatially-varying coefficients models (Gelfand et al. 2003; Finley 2011).

In the SPDE framework non-stationarity is achieved by allowing the Matérn covariance function parameters to vary smoothly over space according to a log-linear function: thus, we will have $\sigma^2(s)$ for the marginal variance in (3.5) and $r(s)$ for the spatial range (Ingebrigtsen et al. 2014; Lindgren and Rue 2015b). Bakka et al. (2016) extend this approach to solve specifically the barrier problem for SDMs. In particular, they force the spatial correlation to go around the barriers (and not through them) by means of a partition of the considered spatial field—in a normal and in a barrier area—and in the specification of two corresponding non-stationary processes with different range parameters (in particular for the barrier region the range parameter is almost zero). The application considered in Bakka et al. (2016) regards fish larvae data in the Finnish archipelago.

4.5 Imperfect detection

Studies on species abundance and distribution are often imperfect due to observer error (Nichols et al. 2000), species rarity (Dettmers et al. 1999) or because detection varies with confounding variables such as environmental conditions (Gu and Swihart 2004; Pennino et al. 2017). When detection is imperfect, additional steps are usually needed to improve inference. Indeed, failure to do so could result in biased estimation and erroneous conclusions.

In recent years, new models called site-occupancy (Hoeting et al. 2000; MacKenzie et al. 2002) for presence-absence data and N-mixture models (Royle 2004) for abundance data have been developed to solve this problem. These models combine two processes: an ecological

process to describe habitat suitability and an observation process to take imperfect detection into account. To estimate detectability, these models use information from repeated observations at several sites. Detectability may vary with site characteristics such as habitat variables, or survey characteristics such as weather conditions, since suitability relates only to site characteristics. Various studies showing the advantages of site occupancy and N-mixture models over classical models that do not consider the problem of detectability can be found in the literature: Royle (2004), Dorazio et al. (2006) for birds, MacKenzie et al. (2002) for amphibians or Pennino et al. (2017) for cetaceans. In addition to the detectability problem, a variety of methods have been developed to correct for the effects of spatial autocorrelation (Latimer et al. 2006; Johnson et al. 2013; Hefley et al. 2017a).

A Bayesian version for site-occupancy spatial models and N-mixture spatial models could also be implemented to take simultaneously account of both imperfect detection and spatial autocorrelation. To describe Bayesian site-occupancy spatial models, let z_i be a random variable describing habitat suitability at site s_i . It can take the value 1 or 0 depending on the habitat suitability, i.e. $z_i = 1$ or $z_i = 0$, thus a Bernoulli distribution is assumed with parameter π_i . Several visits at time $t = 1, \dots, T$ can happen at site i . Let y_{it} be a random variable representing the presence of the species at site i and time t . The species is observed at site i ($\sum_t y_{it} \geq 1$) only if the habitat is suitable ($z_i = 1$). The species is unobserved at site i ($\sum_t y_{it} = 0$) if the habitat is not suitable ($z_i = 0$), or if the habitat is suitable ($z_i = 1$) but the probability α_{it} of detecting the species at site s_i and time t is lower than 1. Then, y_{it} follows a Bernoulli distribution of parameter $z_i \alpha_{it}$, and the model is expressed as follows

Ecological process

$$z_i \sim \text{Bernoulli}(\pi_i),$$

$$\text{logit}(\pi_i) = \beta_0 + \sum_{m=1}^{M_1} \beta_m x_{mi}^{(1)} + w_i, \tag{4.8}$$

Detection process

$$y_{it} \sim \text{Bernoulli}(z_i \alpha_{it}),$$

$$\text{logit}(\alpha_{it}) = \gamma_0 + \sum_{m=1}^{M_2} \gamma_m x_{mit}^{(2)}, \tag{4.9}$$

where $\{\beta_0, \dots, \beta_{M_1}\}$ and $\{\gamma_0, \dots, \gamma_{M_2}\}$ are the parameters that quantify the linear effects of some covariates $(x_1^{(1)}, \dots, x_{M_1}^{(1)})$ and $(x_1^{(2)}, \dots, x_{M_2}^{(2)})$ in the ecological and observation process respectively. These covariates are usually variables referred to site characteristics such as habitat variables or survey characteristics such as weather

conditions. $\mathbf{w} = (w_1, \dots, w_n)$ represents the spatial effect in the ecological process. Normally, this spatial effect is a Gaussian process that can be incorporated as geostatistical terms (in the way already introduced in Sect. 3), but other options are possible (such as CAR Normal distributions, as in Pennino et al. (2017)). The R-package hSDM, which make inference using MCMC, can be used easily to fit some of these models. In addition, the inlabru package also handle the problem of detectability (Yuan et al. 2016).

With respect to N-mixture models, which are used for count data with imperfect detection, they implement a Poisson distribution for the ecological process, while using a Binomial distribution for the observability process (Royle and Nichols 2003; Dodd and Dorazio 2004; Royle 2004). The structure of the model is similar to the site-occupancy model, in particular:

Ecological process

$$N_i \sim \text{Poisson}(\lambda_i),$$

$$\log(\lambda_i) = \beta_0 + \sum_{m=1}^{M_1} \beta_m x_{mi}^{(1)} + w_i, \tag{4.10}$$

Detection process

$$y_{it} \sim \text{Bernoulli}(N_i \alpha_{it}),$$

$$\text{logit}(\alpha_{it}) = \gamma_0 + \sum_{m=1}^{M_2} \gamma_m x_{mit}^{(2)}. \tag{4.11}$$

The R-package hSDM allow us to fit some of these models. In addition, the INLA group is developing some methods to fit N-mixture models (Meehan et al. 2017).

4.6 Excess of zeros

The study of datasets with zero excess has an important role in the literature, particularly, in species distribution modeling (Agarwal et al. 2002; Ver Hoef and Jansen 2007; Neelon et al. 2013), becoming highly relevant in recent years especially. Bayesian softwares like INLA already contain different functions to handle situations with zero excess. Generally, these situations are a source of overdispersion caused by a disagreement between the data and the distribution assumed: there are more zeros in the dataset than the proposed distribution could reasonably explain.

Zero-inflated models are a widely known tool for dealing with this problem. These models assume that the data follow a finite mixture of a degenerate distribution with all its mass at zero with a discrete distribution with support in $\mathbb{Z}^+ \cup \{0\}$ (Yau et al. 2003). If $1 - \pi_i$ represents the probability of species presence, π_i the probability of the species absence, i.e., $p(y_i|\pi_i) = \pi_i$ and $p(y_i > 0) = 1 - \pi_i$, and h a

probability mass function (pmf) of some parametric discrete distribution with support on $\mathbb{Z}^+ \cup \{0\}$, the distribution of y_i has the following mixture density:

$$p(y_i|\pi_i, \mu_i, \boldsymbol{\psi}_1) = \pi_i \delta_0 + (1 - \pi_i)h(y_i|\mu_i, \boldsymbol{\psi}_1), \tag{4.12}$$

being δ_0 the Dirac delta function, μ_i and $\boldsymbol{\psi}_1$ hyperparameters depending on h , and h is a pmf coming from a Poisson, binomial or negative-binomial (note that this latter distribution is one of those considered to account for overdispersion). The model is completed when linking π_i and μ_i with the linear predictors by means of:

$$\begin{aligned} \text{logit}(\pi_i) = \eta_i^{(1)} &= \alpha^{(1)} + \sum_{m=1}^{M^{(1)}} \beta_m^{(1)} x_{mi}^{(1)} + \sum_{l=1}^{L^{(1)}} f_l^{(1)}(z_{li}^{(1)}), \\ g(\mu_i) = \eta_i^{(2)} &= \alpha^{(2)} + \sum_{m=1}^{M^{(2)}} \beta_m^{(2)} x_{mi}^{(2)} + \sum_{l=1}^{L^{(2)}} f_l^{(2)}(z_{li}^{(2)}), \end{aligned} \tag{4.13}$$

where logit denotes the link function between the linear predictor $\eta_i^{(1)}$ and the probability of absence π_i , and $g(\cdot)$ is an appropriate link for the mean of h .

An alternative to these models is given by hurdle models (Mullahy 1986; Cameron and Trivedi 1998), where data are assumed to follow a finite mixture of a degenerate distribution with all its mass at zero and a zero truncated discrete distribution. That is, unlike the zero inflated models, in hurdle models, all observed zeros come from the zero-degenerate distribution. Following the same notation of Eq. (4.12), a hurdle model can be expressed as follows:

$$p(y_i|\pi_i, \mu_i, \boldsymbol{\psi}_1) = \pi_i \delta_0 + (1 - \pi_i)h(y_i|\mu_i, \boldsymbol{\psi}_1)I_{[y_i > 0]}. \tag{4.14}$$

As in (4.13), the hurdle model is completed when linking π_i and μ_i with their corresponding linear predictors.

However, the response variable is not always a discrete variable. Semi-continuous processes like rain, plant coverage, chemical concentrations, etc., are measured in the $[0, \infty)$ interval having high proportions of zero values, and there are neither an appropriate probability distribution nor a transformation available to fit them adequately. To model processes of this type, an extension of hurdle models for continuous data is required (Aitchison 1955; Quiroz et al. 2015). Again, data are modeled as two independent sub-processes: one determines whether the response is zero, and the other determines the intensity when the response is non-zero using a continuous well known distribution like the log-Normal or the Gamma (Stefánsson 1996; Brynjarsdóttir and Stefánsson 2004; Paradinas et al. 2017b). In this case, hurdle models are defined as a finite mixture of a degenerate distribution with point mass at zero and a distribution with support on \mathbb{R}^+ . If h is a pdf of some

parametric continuous distribution with support on \mathbb{R}^+ (e.g. Gamma, log-Normal or log-logistic), the hurdle model for y_i (now assumed to be a continuous distribution) has the same mixture density as in (4.14). Although there exist an extensive list of zero-inflated or hurdle models dealing with correlated discrete data in many fields (Agarwal et al. 2002; Ver Hoef and Jansen 2007), this approach has not been widely used with continuous responses.

It is worth noting that all the models commented upon in this section are a mixture of two processes, and in almost all cases, they are modeled independently (Neelon et al. 2013; Balderama et al. 2016). However, generally both sub-processes are related: low intensities are linked to low probabilities of presence and vice versa. Shared component modeling (SCM) is a good tool to deal with it by combining information both from the two subprocesses (Paradinas et al. 2017b).

5 Discussion

This paper has reviewed some of the statistical challenges that can arise when the distribution of the species is modeled using geostatistical or point-referenced data. In particular, after describing in detail data and methods commonly used to model species distribution, we have focused on complex issues and we have discussed how they can be solved using Bayesian hierarchical spatio-temporal models. Specifically, in this review we have focused on the Bayesian approach and the INLA methodology (Rue et al. 2009) because they have several benefits with respect to the classic geostatistical methods. INLA makes it possible to perform complex models with a minimum computational effort while obtaining accurate estimates. Its importance in the context of SDMs can be even more appreciated with the appearance of the recent project *inlabru* which has been created to develop and implement innovative methods to model spatial distribution and change from ecological survey data (<https://sites.google.com/inlabru3.org/inlabru>). In addition, classic geostatistical methods typically overestimate their predictive accuracy by using plug-in estimations of parameters in their predictive equations. (Diggle and Ribeiro 2007). On the contrary, inference about uncertainty, based on the observations and models, is a byproduct of the model predictions when the Bayesian framework is employed.

However, some limitations can arise when the INLA approach is used. For example, INLA can not handle missing values in spatially structured covariates. This issue can be framed in the misalignment problem discussed in Sect. 4.3; this means that it could be overcome by applying a two-stage or joint modeling approach that allows

prediction of the covariate values in the locations where they were not measured. As mentioned above, an alternative is the one proposed by Gómez-Rubio and Rue (2017) that, using a more general approach, deals with missing values in the covariates, based on fitting conditional latent Gaussian models where covariates are imputed using a Metropolis-Hastings algorithm.

We would like to remark that, due to space limitations, we have not fully reviewed the several complications that can derive from the sampling process. Indeed, we have only focused on the preferential sampling problem (Diggle et al. 2010), which, as previously mentioned, refers to the possibility that the sample design is stochastically dependent on the studied process. Nevertheless, other types of sampling procedures could produce different issues that should be taken into account in the statistical analysis. For example, one of the most popular methods used in ecology to estimate an animal population's size is the capture-recapture method that involves capturing, marking and releasing an initial sample of individuals (Otis et al. 1978; McInerney and Purves 2011). Subsequently, a second sample of animal individuals is obtained independently and it is noted how many of them in that sample were marked. To model data of this type, a feasible solution could be the implementation of Bayesian hierarchical N-mixture models described in Sect. 4.5, which are currently being developed in INLA (Meehan et al. 2017).

Finally, an important point to consider is that INLA is not the only computational approach to making inference for Bayesian spatio-temporal models. In recent years, other approaches that also make it possible to achieve accurate species distribution models results, such as *stan* (Stan Development Team 2017; Monnahan et al. 2017), have been widely used.

Acknowledgements JM-M would like to thank Generalitat Valenciana for support via VALi+d grant ACIF/2016/455, while DC would like to thank the Ministerio de Educación y Ciencia (Spain) for financial support (jointly financed by the European Regional Development Fund) via Research Grant MTM2016-77501-P. MC has been supported by the PRIN EphaStat Project (Project No. 20154X8K23, <https://sites.google.com/site/ephastat/>) provided by the Italian Ministry for Education, University and Research. We would also like to thank Facundo Muñoz for his detailed and careful reading of our paper. All his comments have helped us in identifying areas where clarification/changes/additional details were needed.

References

- Agarwal DK, Gelfand AE, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. *Environ Ecol Stat* 9(4):341–355
- Aitchison J (1955) On the distribution of a positive random variable having a discrete probability mass at the origin. *J Am Stat Assoc* 50(271):901–908

- Aizpurua O, Paquet JY, Brotons L, Titeux N (2015) Optimising long-term monitoring projects for species distribution modelling: how atlas data may help. *Ecography* 38(1):29–40
- Anatolyev S, Kosenok G (2005) An alternative to maximum likelihood based on spacings. *Econom Theory* 21(2):472–476
- Andreon S, Weaver B (2015) Bayesian methods for the physical sciences: learning from examples in astronomy and physics. Springer series in astrostatistics, vol 4. Springer, Berlin
- Araújo MB, Pearson RG, Thuiller W, Erhard M (2005) Validation of species-climate impact models under climate change. *Glob Change Biol* 11(9):1504–1513
- Baio G (2012) Bayesian methods in health economics. CRC Chapman and Hall, Boca Raton
- Bakka H, Vanhatalo J, Illian J, Simpson D, Rue H (2016) Accounting for physical barriers in species distribution modeling with non-stationary spatial random effects. [arXiv:1608.03787](https://arxiv.org/abs/1608.03787)
- Balderama E, Gardner B, Reich BJ (2016) A spatial-temporal double-hurdle model for extremely over-dispersed avian count data. *Spat Stat* 18:263–275
- Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data. CRC, Boca Raton
- Barber X, Conesa D, Lladosa S, López-Quílez A (2016) Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. *Geospat Health* 11:415
- Barber X, Conesa D, López-Quílez A, Mayoral A, Morales J, Barber A (2017) Bayesian hierarchical models for analysing the spatial distribution of bioclimatic indices. *SORT-Stat Oper Res Trans* 1(2):277–296
- Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol Evol* 3(2):327–338
- Bartlett MS (1937) Properties of sufficiency and statistical tests. *Proc R Soc Lond A* 160(901):268–282
- Beale CM, Lennon JJ, Yearsley JM, Brewer MJ, Elston DA (2010) Regression analysis of spatial data. *Ecol Lett* 13(2):246–264
- Berry DA, Stangl D (1999) Bayesian biostatistics. Marcel Dekker, New York City
- Blangiardo M, Cameletti M (2015) Spatial and spatio-temporal Bayesian models with R-INLA. Wiley, Hoboken
- Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with R-INLA. *Spat Spatio-temporal Epidemiol* 7:39–55
- Bowman K, Shenton L (2006) Estimation: method of moments. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, Johnson NL (eds) Encyclopedia of statistical sciences
- Brezger A, Kneib T, Lang S (2003) BayesX: Analysing Bayesian structured additive regression models. Tech. rep., Discussion paper//Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München
- Brooks S, Gelman A, Jones GL, Meng XL (2011) Handbook of markov chain monte carlo. CRC Press, Boca Raton
- Brown P (2015) Model-based geostatistics the easy way. *J Stat Softw* 63:1–24
- Brown CJ, O’connor MI, Poloczanska ES, Schoeman DS, Buckley LB, Burrows MT, Duarte CM, Halpern BS, Pandolfi JM, Parmesan C, Richardson AJ (2016) Ecological and methodological drivers of species distribution and phenology responses to climate change. *Glob Change Biol* 22:1548–1560
- Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression. *J R Stat Soc Ser D (The Statistician)* 47(3):431–443
- Brynjarsdóttir J, Stefánsson G (2004) Analysis of cod catch data from Icelandic groundfish surveys using generalized linear models. *Fish Res* 70(2):195–208
- Busby JR (1991) BIOCLIM – A bioclimate analysis and prediction system. In: Margules CR, Austin MP (eds) Nature conservation: cost effective biological surveys and data analysis. CSIRO, Melbourne, pp 64–68
- Cameletti M, Ignaccolo R, Bande S (2011) Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics* 22(8):985–996
- Cameletti M, Lindgren F, Simpson D, Rue H (2013) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *ASTA Adv Stat Anal* 97(2):109–131
- Cameron CA, Trivedi PK (1998) Regression analysis count data. Cambridge University Press, New York
- Carpenter G, Gillison A, Winter J (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodivers Conserv* 2(6):667–680
- Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu LA (2016) Measurement error in nonlinear models: a modern perspective, 2nd edn. Chapman and Hall/CRC, Boca Raton
- Clark J, Gelfand A (2006) Hierarchical modeling for the environmental sciences. Statistical methods and applications. Oxford University Press, New York
- Clark JS, Gelfand AE, Woodall CW, Zhu K (2014) More than the sum of the parts: forest climate response from joint species distribution models. *Ecol Appl* 24(5):990–999
- Cosandey-Godin A, Krainski ET, Worm B, Flemming JM (2015) Applying Bayesian spatio-temporal models to fisheries bycatch in the Canadian Arctic. *Can J Fish Aquat Sci* 72(2):186–197
- Cox DR, Reid N (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91(3):729–737
- Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, Hoboken
- Daly C (2006) Guidelines for assessing the suitability of spatial climate data sets. *Int J Climatol* 26(6):707–721
- Danks F, Klein D (2002) Using GIS to predict potential wildlife habitat: a case study of muskoxen in northern Alaska. *Int J Remote Sens* 23(21):4611–4632
- Dettmers R, Buehler DA, Bartlett JG, Klaus NA (1999) Influence of point count length and repeated visits on habitat model performance. *J Wildl Manag* 63:815–823
- Diggle PJ (2013) Statistical analysis of spatial and spatio-temporal point patterns. CRC, Boca Raton
- Diggle PJ, Ribeiro PJ (2007) Model-based geostatistics. Springer, Berlin
- Diggle PJ, Menezes R, Su TL (2010) Geostatistical inference under preferential sampling. *J R Stat Soc Ser C (Appl Stat)* 59(2):191–232
- Dodd CK Jr, Dorazio RM (2004) Using counts to simultaneously estimate abundance and detection probabilities in a salamander community. *Herpetologica* 60(4):468–478
- Dorazio RM, Royle JA, Söderström B, Glimskär A (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* 87(4):842–854
- Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* 40:677–697
- Fatima SH, Atif S, Rasheed SB, Zaidi F, Hussain E (2016) Species distribution modelling of *Aedes aegypti* in two dengue-endemic regions of Pakistan. *Trop Med Int Health* 21:427–436
- Ferrari SLP, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. *J Appl Stat* 31(7):799–815
- Finley AO (2011) Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods Ecol Evol* 2(2):143–154
- Fitzpatrick MC, Weltzin JF, Sanders NJ, Dunn RR (2007) The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Glob Ecol Biogeogr* 16(1):24–33

- Fortin MJ, Dale MR (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge
- Foster SD, Shimadzu H, Darnell R (2012) Uncertainty in spatially predicted covariates: is it ignorable? *J Roy Stat Soc Ser C (Appl Stat)* 61(4):637–652
- Franklin J (2010) Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge
- Gaudard M, Karson M, Linder E, Sinha D (1999) Bayesian spatial prediction. *Environ Ecol Stat* 6(2):147–171
- Gelfand AE, Smith AF (1990) Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85(410):398–409
- Gelfand AE, Kim HJ, Sirmans CF, Banerjee S (2003) Spatial modeling with spatially varying coefficient processes. *J Am Stat Assoc* 98(462):387–396
- Gelfand AE, Silander JA, Wu S, Latimer A, Lewis PO, Rebelo AG, Holder M (2006) Explaining species distribution patterns through hierarchical modeling. *Bayesian Anal* 1(1):41–92
- Gelfand AE, Diggle PJ, Fuentes M, Guttorp P (2010) Handbook of spatial statistics. Chapman & Hall, Boca Raton
- Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Chapman & Hall/CRC, Boca Raton
- Gitzen RA (2012) Design and analysis of long-term ecological monitoring studies. Cambridge University Press, Cambridge
- Goetz SJ, Sun M, Zolkos S, Hansen A, Dubayah R (2014) The relative importance of climate and vegetation properties on patterns of North American breeding bird species richness. *Environ Res Lett* 9(3):034013
- Golding N, Purse BV (2016) Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods Ecol Evol* 7:598–608
- Gómez-Rubio V, Rue H (2017) Markov chain monte carlo with the integrated nested Laplace approximation. [arXiv:1702.07007](https://arxiv.org/abs/1702.07007)
- Gómez-Rubio V, Bivand RS, Rue H (2014) Spatial models using Laplace approximation methods. In: Fischer MM, Nijkamp P (eds) Handbook of regional science. Springer, Berlin, pp 1401–1417
- González-Warleta M, Lladosa S, Castro-Hermida JA, Martínez-Ibeas AM, Conesa D, Muñoz F, López-Quílez A, Manga-González Y, Mezo M (2013) Bovine paramphistomosis in Galicia (Spain): prevalence, intensity, aetiology and geospatial distribution of the infection. *Vet Parasitol* 191(3):252–263
- Gosoni L, Vouunatsou P, Sogoba N, Smith T (2006) Bayesian modelling of geostatistical malaria risk data. *Geospat Health* 1(1):127–139
- Gotelli NJ, Anderson MJ, Arita HT, Chao A, Colwell RK, Connolly SR, Currie DJ, Dunn RR, Graves GR, Green JL (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecol Lett* 12(9):873–886
- Griffith DA (2008) Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environ Plan A* 40(11):2751–2769
- Gu W, Swihart RK (2004) Absent or undetected? effects of non-detection of species occurrence on wildlife-habitat models. *Biol Conserv* 116(2):195–203
- Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecol Lett* 8(9):993–1009
- Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol Model* 157(2):89–100
- Handcock MS, Stein ML (1993) A Bayesian analysis of kriging. *Technometrics* 35(4):403–410
- Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econom J Econom Soc* 50(4):1029–1054
- He KS, Bradley BA, Cord AF, Rocchini D, Tuanmu MN, Schmidtlein S, Turner W, Wegmann M, Pettorelli N (2015) Will remote sensing shape the next generation of species distribution models? *Remote Sens Ecol Conserv* 1(1):4–18
- Hefley TJ, Hooten MB (2016) Hierarchical species distribution models. *Curr Landsc Ecol Rep* 1(2):87–97
- Hefley TJ, Broms KM, Brost BM, Buderman FE, Kay SL, Scharf HR, Tipton JR, Williams PJ, Hooten MB (2017a) The basis function approach for modeling autocorrelation in ecological data. *Ecology* 98(3):632–646
- Hefley TJ, Hooten MB, Hanks EM, Russell RE, Walsh DP (2017b) Dynamic spatio-temporal models for spatial data. *Spat Stat* 20:206–220
- Hendricks SA, Clee PRS, Harrigan RJ, Pollinger JP, Freedman AH, Callas R, Figura PJ, Wayne RK (2016) Re-defining historical geographic range in species with sparse records: implications for the Mexican wolf reintroduction program. *Biol Conserv* 194:48–57
- Hengl T, Heuvelink GB, Tadić MP, Pebesma EJ (2012) Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theoret Appl Climatol* 107(1–2):265–277
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25(15):1965–1978
- Hoeting JA, Leecaster M, Bowden D (2000) An improved model for spatially correlated binary responses. *J Agric Biol Environ Stat* 5:102–114
- Holloway P, Miller JA (2015) Exploring spatial scale, autocorrelation and nonstationarity of bird species richness patterns. *ISPRS Int J Geo-Inf* 4(2):783–798
- Hooten MB, Wikle CK (2008) A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environ Ecol Stat* 15(1):59–70
- Hooten MB, Wikle CK, Dorazio RM, Royle JA (2007) Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics* 63(2):558–567
- Hui FK (2017) Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Comput Stat Data Anal* 105:1–10
- Illian JB, Sørbye SH, Rue H (2012) A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Ann Appl Stat* 6(4):1499–1530
- Ingebrigtsen R, Lindgren F, Steinsland I (2014) Spatial models with explanatory variables in the dependence structure. *Spat Stat* 8:20–38
- Iturbide M, Bedia J, Herrera S, del Hierro O, Pinto M, Gutiérrez JM (2015) A framework for species distribution modelling with improved pseudo-absence generation. *Ecol Model* 312:166–174
- Iverson LR, Schwartz MW, Prasad AM (2004) How fast and far might tree species migrate in the eastern united states due to climate change? *Glob Ecol Biogeogr* 13(3):209–219
- Jackman S (2009) Bayesian analysis for the social sciences. Wiley, Hoboken
- Jiménez-Valverde A, Lobo JM (2007) Determinants of local spider (*Araneidae* and *Thomisidae*) species richness on a regional scale: climate and altitude vs. habitat structure. *Ecol Entomol* 32(1):113–122
- Johnson DS, Conn PB, Hooten MB, Ray JC, Pond BA (2013) Spatial occupancy models for large data sets. *Ecology* 94(4):801–808
- Jona Lasinio G, Mastrantonio G, Pollice A (2013) Discussing the “big n problem”. *Stat Methods Appl* 22(1):97–112
- Joseph LN, Field SA, Wilcox C, Possingham HP (2006) Presence-absence versus abundance data for monitoring threatened species. *Conserv Biol* 20(6):1679–1687
- Juan P, Díaz-Avalos C, Mejía-Domínguez NR, Mateu J (2017) Hierarchical spatial modeling of the presence of Chagas disease

- insect vectors in Argentina. A comparative approach. *Stoch Env Res Risk Assess* 31(2):461–479
- Karagiannis-Voules DA, Scholte RG, Guimarães LH, Utzinger J, Voumatsou P (2013) Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. *PLOS Negl Trop Dis* 7(5):e2213
- Kneib T, Müller J, Hothorn T (2008) Spatial smoothing techniques for the assessment of habitat suitability. *Environ Ecol Stat* 15(3):343–364
- Kozak KH, Graham CH, Wiens JJ (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol Evol* 23(3):141–148
- Krainski ET, Lindgren F, Simpson D, Rue H (2017) The R-INLA tutorial: SPDE models. <http://www.math.ntnu.no/inla/r-inla.org/tutorials/spde/spde-tutorial.pdf>. Accessed 18 Apr 2018
- Latimer AM, Wu S, Gelfand AE, Silander JA (2006) Building statistical models to analyze species distributions. *Ecol Appl* 16(1):33–50
- Le Cam L (1990) Maximum likelihood: an introduction. *Int Stat Rev/Rev Int Stat* 58(2):153–171
- Leathwick J, Rowe D, Richardson J, Elith J, Hastie T (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshw Biol* 50(12):2034–2052
- Lee D (2013) CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *J Stat Softw* 55(13):1–24
- Lindgren F, Rue H (2015a) On the second order random walk model for irregular locations. *Scand J Stat* 35(4):691–700
- Lindgren F, Rue H (2015b) Bayesian spatial modelling with R-INLA. *J Stat Softw* 63(19):1–25
- 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
- Liu C, Wan R, Jiao Y, Reid KB (2017) Exploring non-stationary and scale-dependent relationships between walleye (*Sander vitreus*) distribution and habitat variables in lake Erie. *Mar Freshw Res* 68(2):270–281
- Luo M, Opaluch JJ (2011) Analyze the risks of biological invasion. *Stoch Env Res Risk Assess* 25(3):377–388
- Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 10(4):325–337
- Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions. *Stat Med* 28(25):3049–3067
- MacKenzie DI, Nichols JD, Lachman GB, Droege S, Andrew Royle J, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83(8):2248–2255
- Mallick BK, Gold D, Baladandayuthapani V (2009) Bayesian analysis of gene expression data. Wiley, Hoboken
- Martinez-Meyer E, Peterson AT, Servín JI, Kiff LF (2006) Ecological niche modelling and prioritizing areas for species reintroductions. *Oryx* 40(4):411–418
- Martínez-Bello D, López-Quílez A, Prieto AT (2017) Spatiotemporal modeling of relative risk of dengue disease in Colombia. *Stoch Environ Res Risk Assess*. <https://doi.org/10.1007/s00477-017-1461-5>
- Martínez-Minaya J, Conesa D, López-Quílez A, Vicent A (2018) Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in South Africa. A Bayesian latent Gaussian model approach. *Eur J Plant Pathol* (in press)
- Martino S, Akerkar R, Rue H (2011) Approximate Bayesian inference for survival models. *Scand J Stat* 38(3):514–528
- Martins TG, Simpson D, Lindgren F, Rue H (2013) Bayesian computing with inla: new features. *Comput Stat Data Anal* 67:68–83
- McCarthy MA (2007) Bayesian methods for ecology. Wiley, Hoboken
- McInerny GJ, Purves DW (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods Ecol Evol* 2(3):248–257
- Meehan TD, Michel NL, Rue H (2017) Estimating animal abundance with N-mixture models using the R-INLA package for R. [arXiv:1705.01581](https://arxiv.org/abs/1705.01581)
- Meentemeyer RK, Cunniffe NJ, Cook AR, Filipe JA, Hunter RD, Rizzo DM, Gilligan CA (2011) Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990–2030). *Ecosphere* 2(2):1–24
- Miller JA (2012) Species distribution models. *Prog Phys Geogr* 36(5):681–692
- Monnahan CC, Thorson JT, Branch TA (2017) Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods Ecol Evol* 8(3):339–348
- Muñoz F, Pennino MG, Conesa D, López-Quílez A, Bellido JM (2013) Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. *Stoch Environ Res Risk Assess* 27(5):1171–1180
- Muff S, Riebler A, Held L, Rue H, Saner P (2015) Bayesian analysis of measurement error models using integrated nested Laplace approximations. *J Roy Stat Soc Ser C (Appl Stat)* 64(2):231–252
- Mullahy J (1986) Specification and testing of some modified count data models. *J Econom* 33(3):341–365
- Neelon B, Ghosh P, Loebs PF (2013) A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *J R Stat Soc Ser A* 176(2):389–413
- Neri FM, Cook AR, Gibson GJ, Gottwald TR, Gilligan CA (2014) Bayesian analysis for inference of an emerging epidemic: citrus canker in urban landscapes. *PLOS Comput Biol* 10(4):e1003587
- New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. *Clim Res* 21(1):1–25
- Nichols JD, Hines JE, Sauer JR, Fallon FW, Fallon JE, Heglund PJ (2000) A double-observer approach for estimating detection probability and abundance from point counts. *Auk* 117(2):393–408
- Nielsen SE, Johnson CJ, Heard DC, Boyce MS (2005) Can models of presence-absence be used to scale abundance? Two case studies considering extremes in life history. *Ecography* 28(2):197–208
- Otis DL, Burnham KP, White GC, Anderson DR (1978) Statistical inference from capture data on closed animal populations. *Wildl Monogr* 62:3–135
- Paradinas I, Conesa D, Pennino MG, Muñoz F, Fernández AM, López-Quílez A, Bellido JM (2015) Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. *Mar Ecol Prog Ser* 528:245–255
- Paradinas I, Marín M, Pennino MG, López-Quílez A, Conesa D, Barreda D, Gonzalez M, Bellido JM (2016) Identifying the best fishing-suitable areas under the new European discard ban. *ICES J Mar Sci J Cons* 73(10):2479–2487
- Paradinas I, Conesa D, López-Quílez A, Bellido JM (2017a) Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling. *Spat Stat* 22:434–450
- Paradinas I, Pennino MG, López-Quílez A, Marín M, Bellido JM, Conesa D (2017b) Modelling spatially sampled proportion processes. *REVSTATStat J* 16(1):71–86
- Park YS, Céréghino R, Compin A, Lek S (2003) Applications of artificial neural networks for patterning and predicting aquatic

- insect species richness in running waters. *Ecol Model* 160(3):265–280
- Parviainen M, Luoto M, Rytteri T, Heikkinen RK (2008) Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. *J Biogeogr* 35(10):1888–1905
- Pennino MG, Muñoz F, Conesa D, López-Quílez A, Bellido JM (2013) Modeling sensitive elasmobranch habitats. *J Sea Res* 83:209–218
- Pennino MG, Muñoz F, Conesa D, López-Quílez A, Bellido JM (2014) Bayesian spatio-temporal discard model in a demersal trawl fishery. *J Sea Res* 90:44–53
- Pennino MG, Conesa D, López-Quílez A, Muñoz F, Fernández A, Bellido JM (2016) Fishery-dependent and-independent data lead to consistent estimations of essential habitats. *ICES J Mar Sci J Cons* 73(9):2302–2310
- Pennino MG, Mérigot B, Fonseca VP, Monni V, Rotta A (2017) Habitat modeling for cetacean management: spatial distribution in the southern Pelagos Sanctuary (Mediterranean sea). *Deep Sea Res Part II Top Stud Oceanogr* 141:203–211
- Pennino MG, Paradinas I, Illian JB, Muñoz F, Bellido JM, López-Quílez A, Conesa D (2018) Accounting for preferential sampling in species distribution models (**submitted**)
- Peterson AT, Sánchez-Cordero V, Beard CB, Ramsey JM (2002) Ecologic niche modeling and potential reservoirs for chagas disease, Mexico. *Emerg Infect Dis* 8(7):662–667
- Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecol Model* 190(3):231–259
- Plummer M (2016) *rjags*: Bayesian graphical models using MCMC. R Software Package for Graphical Models. URL <https://cran.r-project.org/web/packages/rjags/index.html>
- Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, Vesik PA, McCarthy MA (2014) Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods Ecol Evol* 5(5):397–406
- Quiroz ZC, Prates MO, Rue H (2015) A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics* 71(1):208–217
- Rachev ST, Hsu JS, Bagasheva BS, Fabozzi FJ (2008) *Bayesian methods in finance*, vol 153. Wiley, Hoboken
- Risser MD (2016) Review: nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. [arXiv:1610.02447](https://arxiv.org/abs/1610.02447)
- Robert C, Casella G (2011) A short history of Markov Chain Monte Carlo: subjective recollections from incomplete data. *Stat Sci* 26(1):102–115
- Rodríguez de Rivera O, López-Quílez A (2017) Development and comparison of species distribution models for forest inventories. *ISPRS Int J Geo-Inf* 6(6):176
- Roos NC, Carvalho AR, Lopes PF, Pennino MG (2015) Modeling sensitive parrotfish (Labridae: Scarini) habitats along the Brazilian coast. *Mar Environ Res* 110:92–100
- Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60(1):108–115
- Royle JA, Nichols JD (2003) Estimating abundance from repeated presence-absence data or point counts. *Ecology* 84(3):777–790
- Rue H, Held L (2005) *Gaussian Markov random fields: theory and applications*. Chapman & Hall, Boca Raton
- 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
- Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK (2017) Bayesian computing with INLA: a review. *Annu Rev Stat Appl* 4:395–421
- Rufener MC, Kinas PG, Nóbrega MF, EL Oliveira J (2017) Bayesian spatial predictive models for data-poor fisheries. *Ecol Model* 348:125–134
- Ruiz-Cárdenas R, Krainski ET, Rue H (2012) Direct fitting of dynamic models using integrated nested Laplace approximations-INLA. *Comput Stat Data Anal* 56(6):1808–1828
- Sbrocco EJ, Barber PH (2013) MARSPEC: ocean climate layers for marine spatial ecology. *Ecology* 94(4):979–979
- Schrödle B, Held L, Riebler A, Danuser J (2011) Using integrated nested Laplace approximations for the evaluation of veterinary surveillance data from Switzerland: a case-study. *J Roy Stat Soc Ser C (Appl Stat)* 60(2):261–279
- Sadykova D, Scott BE, De Dominicis M, Wakelin SL, 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
- Shapiro A (2000) On the asymptotics of constrained local M-estimators. *Ann Stat* 28(3):948–960
- Stan Development Team (2017) *Stan Modeling Language Users Guide and Reference Manual*, Version 2.17.0. <http://mc-stan.org>. Accessed 18 Apr 2018
- Stefánsson G (1996) Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES J Mar Sci* 53(3):577–588
- Stein M (1999) *Interpolation of spatial data. Some theory for kriging*. Springer, Berlin
- Stein A, Kocks C, Zadoks J, Frinking H, Ruissen M, Myers D (1994) A geostatistical analysis of the spatio-temporal development of downy mildew epidemics in cabbage. *Phytopathology* 84(10):1227–1238
- Stoklosa J, Daly C, Foster SD, Ashcroft MB, Warton DI (2015) A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods Ecol Evol* 6(4):412–423
- Taylor-Rodríguez D, Kaufeld K, Schliep EM, Clark JS, Gelfand AE (2017) Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Anal* 12(4):939–967
- Václavík T, Meentemeyer RK (2009) Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? *Ecol Model* 220(23):3248–3258
- Ver Hoef JM, Jansen JK (2007) Space-time zero-inflated count models of Harbor seals. *Environmetrics* 18(7):697–712
- Vieilledet G, Latimer A, Gelfand A, Merow C, Wilson A, Mortier F, Silander Jr J (2014) *hSDM*: hierarchical Bayesian species distribution models. R package version 1
- White SM, Bullock JM, Hooftman DA, Chapman DS (2017) Modelling the spread and control of *Xylella fastidiosa* in the early stages of invasion in Apulia, Italy. *Biol Invasions* 19:1825–1837
- Wikle CK (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84(6):1382–1394
- Wikle CK, Hooten MB (2010) A general science-based framework for dynamical spatio-temporal models. *Test* 19(3):417–451
- Williams PJ, Hooten MB, Womble JN, Esslinger GG, Bower MR, Hefley TJ (2017) An integrated data model to estimate spatiotemporal occupancy, abundance, and colonization dynamics. *Ecology* 98(2):328–336
- Windle MJS, Rose GA, Devillers R, Fortin MJ (2010) Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. *ICES J Mar Sci* 67(1):145
- Yau KK, Wang K, Lee AH (2003) Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biom J* 45(4):437–452

- Yuan Y, Bachl F, Lindgren F, Brochers D, Illian J, Buckland S, Rue H, Gerrodette T (2016) Point process models for spatio-temporal distance sampling data. [arXiv:1604.06013](https://arxiv.org/abs/1604.06013)
- Zhang W (2007) Supervised neural network recognition of habitat zones of rice invertebrates. *Stoch Environ Res Risk Assess* 21(6):729–735
- Zhang W, Zhong X, Liu G (2008) Recognizing spatial distribution patterns of grassland insects: neural network approaches. *Stoch Environ Res Risk Assess* 22(2):207–216