Stoch Environ Res Risk Assess (2013) 27:1171–1180
DOI 10.1007/s00477-012-0652-3

ORIGINAL PAPER

Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models

Facundo Muñoz · M. Grazia Pennino · David Conesa · Antonio López-Quílez · José M. Bellido

Published online: 20 October 2012

Abstract A methodological approach for modelling the occurrence patterns of species for the purpose of fisheries management is proposed here. The presence/absence of the species is modelled with a hierarchical Bayesian spatial model using the geographical and environmental characteristics of each fishing location. Maps of predicted probabilities of presence are generated using Bayesian kriging. Bayesian inference on the parameters and prediction of presence/absence in new locations (Bayesian kriging) are made by considering the model as a latent Gaussian model, which allows the use of the integrated nested Laplace approximation (`INLA`) software (which has been seen to be quite a bit faster than the well-known MCMC methods). In particular, the spatial effect has been implemented with the stochastic partial differential equation (SPDE) approach. The methodology is evaluated on Mediterranean horse mackerel (*Trachurus mediterraneus*) in the Western Mediterranean. The analysis shows that environmental and geographical factors can play an important role in directing local distribution and variability in the occurrence of species. Although this approach is used to recognize the habitat of mackerel, it could also be for other different species and life stages in order to improve knowledge of fish populations and communities.

F. Muñoz (✉) · D. Conesa · A. López-Quílez
Departament d'Estadística i Investigació Operativa, Universitat de Valéncia, C/Dr. Moliner 50, Burjassot 46100 Valencia, Spain
e-mail: facundo.munoz@uv.es

M. G. Pennino · J. M. Bellido
Instituto Español de Oceanografía, Centro Oceanográfico de Murcia, C/Varadero 1, San Pedro del Pinatar 30740 Murcia, Spain

Keywords Bayesian kriging · Bayesian hierarchical models · Fisheries · Integrated nested Laplace approximation (`INLA`) · Modelling distribution species · Stochastic partial differential equations

1 Introduction

Modelling patterns of the presence/absence of the species using local environmental factors has been a growing problem in Ecology in the last few years (Chakraborty et al. 2010). This kind of modelling has been extensively used to address several issues, including identifying essential fish habitats (EFHs) in order to classify and manage conservation areas (Pressey et al. 2007), and predicting the response of species to environmental features (Midgley and Thuiller 2007; Loarie et al. 2008).

Different approaches and methodologies have been proposed for modelling the distribution of species (Guisan and Thuiller 2005; Hijman and Graham 2006; Wisz et al. 2008). Generalized linear and additive models (GLM and GAM) (Guisan et al. 2002), species envelope models such as BIOCLIM (Busby 1991), neural networks (Zhang 2007; Zhang et al. 2008) and the multivariate adaptive regression splines (MARS) (Leathwick et al. 2005) are some of them.

Most of these applications are only explanatory models that seek to assess the relationship between the presence of species and a suite of one or more explanatory variables (e.g. precipitation, bathymetry, etc.) (Guisan et al. 2002). Moreover, the theory of these methods is based on the fact that the observations are independent, while the fishery data are often inclined to spatial autocorrelation (Kneib et al. 2008). Spatial autocorrelation should be taken into account in the 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 (Underwood 1981; Hurlbert 1984). As a consequence, ignoring spatial correlations in this type of analysis could lead to misleading results (Kneib et al. 2008). Note also that extensive spatiotemporal variability, which characterizes dynamic marine ecosystems, presents inherent difficulties for the development of predictive species-habitat models (Valavanis et al. 2008).

Other complications also arise in the modelling of the occurrence of species due to imperfect survey data such as observer error (Royle et al. 2007; Cressie et al. 2009), gaps in the sampling, missing data, and spatial mobility of the species (Gelfand et al. 2006).

It is also worth mentioning that only a few studies have been developed for predictive models although these models, in addition to offering an estimate of the processes that drive the distribution of species, also provide the probability of the occurrence of species in unsampled areas (Chakraborty et al. 2010).

Our interest here is to propose a hierarchical Bayesian model to predict the occurrence of species by incorporating the environmental and spatial features of each fishing location. The Bayesian approach is appropriate to spatial hierarchical model analysis because it allows both the observed data and model parameters to be random variables (Banerjee et al. 2004), resulting in a more realistic and accurate estimation of uncertainty (see, for instance, Haining et al. 2007, as an example of the advantages over conventional—non-Bayesian—modelling approaches).

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 autocorrelative effects from ordinary non-spatial linear effects (Gaudard et al. 2006). Finally, an important feature of our approach is that maps of predicted probabilities of presence in unsampled areas are generated using Bayesian kriging (Handcock and Stein. 1993; Gaudard et al. 1999).

As usual with this kind of hierarchical model, there is no closed expression for the posterior distribution of all the parameters, and so numerical approximations are needed. In our case, we use the integrated nested Laplace approximations (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 for this choice is the speed of calculation: MCMC simulations require much more time to run, and performing prediction has been practically unfeasible. In contrast, INLA produces almost immediately accurate approximations to posterior distributions even in complex models. Another advantage of this approach is its generality, which makes it possible to perform Bayesian analysis in a straightforward way and to compute model comparison criteria and various predictive measures so that models can be compared easily (Rue et al. 2009). INLA 's performance has been compared with MCMC and has shown a similar reliability (Held et al. 2010).

In particular, we have applied our approach to describing the distribution of Mediterranean horse mackerel (*Trachurus mediterraneus*) in the Western Mediterranean. We have used the geographical characteristics, such as latitude, longitude and bathymetry, of each fishing location. Environmental satellite data, such as the monthly data on precipitation, sea surface temperature and chlorophyll-*a* concentration have also been included in the analysis.

Finally, we would like to mention that this approach could also be employed in different settings with other species and life stages in order to improve knowledge of fish populations and communities.

The remainder of this article is organized as follows. After this introduction, in Sect. 2, we present a general Bayesian hierarchical spatial model that accounts for the presence/absence of fish species, allowing both for inference and prediction in unsampled locations. This is commonly known as Bayesian kriging (Banerjee et al. 2004). In Sect. 3, we describe how to implement this model using INLA. In Sect. 4, we apply this methodology in a particular setting with fishery data from Southern Spain in order to provide a realistic view of the methods. Finally, in Sect. 5, we present some concluding remarks and future lines of research.

2 Modelling fish presence

This section will describe Bayesian kriging and its application to presence/absence data in fishing. We also discuss the implementation of this kind of model with INLA and introduce the SPDE approach to modelling the spatial component.

2.1 Bayesian kriging for a binary response

Point-referenced spatial models (Gelfand et al. 2000) are very suitable for situations in which we have observations made at continuous locations occurring within a defined spatial domain. This particular case of spatial models also has the appealing characteristic that the spatial domain is unchanging, even though the precise locations will change over time. In fisheries, this resolves the dimensional control guaranteeing that the inference is realized in relation to the domain instead of the current observed positions, which can change over the years.

In these models, the estimation of the response in unsampled locations can be seen as a statistical prediction problem. When the response is Normal, this is known as

kriging prediction. Using a Bayesian hierarchical model (Banerjee et al. 2004) such as the one we present in this section allows naturally for non-Gaussian responses, and for taking into account uncertainty in the model parameters. This is known as Bayesian kriging, and the rest of this section discusses its application to fishery data.

Basically, when analyzing fish species distribution, we can encounter two different types of observed data: the amount of catch or just presence/absence data. In the first case it is possible to calculate the absolute abundance of species by standardizing the catch with the fishing effort of the studied fleet, and so it is possible to assess the quantitative spatial distribution of the species within the area of interest. In the second case, presence/absence information can be used as a measure of the relative occurrence of species at each precise observed location, thereby giving a different (but very valid and useful) approximation for the spatial distribution of the species.

For most species, especially for those which are not targeted, information about the absolute abundance of the species is not available. In these situations, the spatial distribution can be obtained by using presence/absence as a response variable of interest instead of absolute abundance. Then, assuming that the probability of catching a species is related to its presence, we model presence/absence by using a point-referenced spatial hierarchical model in line with Diggle et al. (1998).

Specifically, if Z_i represents presence (1) or absence (0) at location i $(i = 1, \ldots, n)$ and π_i is the probability of presence, then:

$$\begin{aligned} Z_i &\sim \text{Ber}(\pi_i) \\ \text{logit}(\pi_i) &= \mathbf{X}_i\beta + W_i \end{aligned} \tag{1}$$

where $\mathbf{X}_i\beta$ represents the linear predictor for observation i; W_i represent the spatially structured random effect; and the relation between π_i and the covariates of interest and random effects is the usual logit link. W_i is assumed to be Gaussian with a given covariance matrix $\sigma_W^2 H(\phi)$, depending on the distance between locations, and with hyperparameters σ_W^2 and ϕ representing respectively the variance (partial sill in kriging terminology) and the range of the spatial effect:

$$\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \sigma_W^2 H(\phi)). \tag{2}$$

This modelling could be augmented by incorporating an additional pure error term (usually Gaussian distributed with variance called nugget effect in kriging terminology) describing the "noise" associated with replication of measurement at each location. Nevertheless, as in this case we are dealing with Bernoulli response, sensitivity to prior assumptions on those random effects precision parameters should be dealt carefully (Roos and Held 2011).

Once the model is determined, the next step is to estimate its parameters. As we are using the Bayesian paradigm, we have to specify the prior distributions for each parameter involved in the model $(\boldsymbol{\beta}, \sigma_W^2, \phi)$. In this context, the usual choice (see, for instance, Banerjee et al. 2004) is to deal with independent priors for the parameters, i.e.

$$p(\beta, \sigma_W^2, \phi) = p(\beta)p(\sigma_W^2)p(\phi). \tag{3}$$

When there is an aim of expressing initial vague knowledge about the parameters, useful (but not the only) candidates are non-informative Gaussian prior distributions for $\boldsymbol{\beta}$ and inverse gamma distributions for σ_W^2. Specification of $p(\phi)$ will depend on the choice of the correlation function which determines the covariance matrix H. Note that the final choice for the priors will also depend on the type of modelling and parameterization chosen. We will return to this topic later on.

As mentioned above, expressions from (1, 2, 3) contain all our knowledge about the spatial occurrence but do not yield closed expressions for the posterior distributions of all the parameters. And so in order to make inference about them, numerical approximations are needed. One possible choice for doing this would be using Markov chain Monte Carlo (MCMC) methods. This could be done using WinBUGS (Spiegelhalter et al. 1999), flexible software for performing the Bayesian analysis of complex statistical models (see Banerjee et al. 2004) for examples of how to implement spatial hierarchical Bayesian models with WinBUGS). Nevertheless, this option turns out to be very slow when interest is focused on prediction (as in our case), so we have to resort to another approach.

2.2 Implementing Bayesian kriging with `INLA`

The key idea underlying what follows is to realize that these hierarchical models can be seen as *Structured Additive Regression (STAR) models* (see, for instance Fahrmeir and Tutz (2001) for a detailed description of them and Chien and Bangdiwala (2012) for an applied example of their use). In other words, models in which the mean of the response variable Z_i is linked to a structured predictor that accounts for the effects of various covariates in an additive way. But, more specifically, point referenced spatial hierarchical Bayesian models can also be seen as a particular case of STAR models called *Latent Gaussian models* (Rue et al. 2009), namely those assigning Gaussian priors to all the components of the additive predictor. In this framework, all the latent Gaussian variables can be seen as components of a vector which is the latent Gaussian field.

The great bonus here is that for latent Gaussian models, we can directly compute very accurate approximations of the posterior marginals using `INLA` (Rue et al. 2009). In

spite of its wide acceptance and its good behaviour in many Latent Gaussian models (see for instance, Schrodle and Held (2011) for a description of how to use INLA in spatio-temporal disease mapping), until now it has not been feasible to fit the particular case of continuously indexed Gaussian Fields with INLA, as is the case with our spatial component $\mathbf{W}$. The underlying reason is that a parametric covariance function needs to be specified and fitted based on the data, which determines the covariance matrix H and enables prediction in unsampled locations. But from the computational perspective, the cost of factorizing these (dense) matrices is cubic in their dimension. Despite computational power today, this problem is still a computational bottleneck in many situations.

Lindgren et al. (2011) have proposed an alternative approach by using an approximate stochastic weak solution to a stochastic partial differential equation (SPDE) as a Gaussian Markov random field (GMRF, Rue and Held 2005; GMRF, Rue et al. 2009) approximation to continuous Gaussian Fields with Matern covariance structure. Specifically, they use the fact that a Gaussian Field $x(\mathbf{u})$ with Matern covariance is a solution to the linear fractional SPDE

$$\begin{aligned} &(\kappa^2 - \Delta)^{\alpha/2} x(\mathbf{u}) = \mathcal{W}(\mathbf{u}), \\ &\mathbf{u} \in \mathbb{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0, \end{aligned} \tag{4}$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo-differential operator defined in terms of its spectral properties (see Lindgren et al. 2011). They then use a finite-elements method on a triangulation of the region (see Figure 1) to construct an approximate GMRF representation of the Matern Field with parameters κ and $\nu = 1$. They fix ν to 1 for identifiability reasons. An additional parameter τ is used to adjust the scale of the field.

Some important features arise here. Firstly, a GMRF is a discretely indexed Gaussian field $\mathbf{x} = (x_1, \ldots, x_n)$, where the full conditionals $\pi(x_i|\mathbf{x}_{-i}), i = 1, \ldots, n$ depend only on a set of neighbours of each site i. This Markov property makes their precision matrix sparse, enabling the use of efficient (and faster) numerical algorithms.

Secondly, the Matern covariance function is a really flexible and general family of functions generalizing many of the most-used covariance models in spatial statistics. Its expression, giving the covariance between the values of a random field at locations separated by a distance $d > 0$, can be parameterized as

$$C(d) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa d)^{\nu} K_{\nu}(\kappa d),$$

where K_{ν} is the modified Bessel function of the second kind and order $\nu > 0$ (Abramowitz and Stegun 1972, § 9.6), $\kappa > 0$ is a scaling parameter and σ^2 is the marginal variance. The parameter ν is a *smoothness* parameter determining the mean-square differentiability of the underlying process, although it is fixed in the SPDE approach since it is poorly identified in typical applications. For more information on the Matern covariance model see Handcock and Stein (1993); Stein (1999); Guttorp and Gneiting (2006). Finally, GMRFs fit seamlessly with the INLA approach, which requires the latent field to be a GMRF.

Under this perspective, for each vertex $i = 1, \ldots, n$, the full model can be stated as follows:

$$\begin{aligned} Z_i|\pi_i &\overset{\text{iid}}{\sim} \text{Ber}(\pi_i) \\ \text{logit}\, \pi_i &= \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{W_i} \\ \pi(\beta_0) &\propto 1 \\ \beta_j &\overset{\text{iid}}{\sim} \mathcal{N}(0, 1\text{e-}05) \\ \mathbf{W} &\sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\kappa, \tau)) \\ 2\log\kappa &\sim \mathcal{N}(m_\kappa, q_\kappa^2) \\ \log\tau &\sim \mathcal{N}(m_\tau, q_\tau^2). \end{aligned} \tag{5}$$

In contrast with the previous specification, when using the SPDE approach the correlation function is not modelled directly. Instead, the Gaussian field $\mathbf{W}$ is found numerically as a (weak) solution of the SPDE (4), depending now on two different parameters κ and τ which determine the range of the effect and the total variance, respectively. More precisely, the range is approximately $\phi \approx \sqrt{8}/\kappa$ while the variance is $\sigma_W^2 = 1/(4\pi\kappa^2\tau^2)$.

Consequently, we have to specify the prior distributions for the parameters involved in this approach $(\beta_0, \boldsymbol{\beta}, \kappa, \tau)$. We set the intercept apart because INLA by default specifies a flat improper prior on the intercept, and independent zero-mean Gaussian priors with a fixed vague precision (1e–05) a priori on the fixed effects in $\boldsymbol{\beta}$. The priors for κ and τ are specified over the reparameterizations $\log\tau$ and $2\log\kappa$ as independent Gaussian distributions. We also used the default values for their parameters. Specifically, m_κ is chosen automatically such that the range of the field is about 20 % of the diameter of the region, while m_τ is chosen so that the corresponding variance of the field is 1. For instance, in the dataset described in Sect. 4, this gives $m_\kappa = -16.8$ and $m_\tau = 7.16$. Finally, the default a priori precisions for $\log\tau$ and $2\log\kappa$ distributions are $q_\kappa^2 = q_\tau^2 = 0.1$.

The INLA program can be used through the R (R Development Core Team 2010) package of the same name. It is worth noting that the SPDE module of INLA is still under development and enhancement, but a fully-functional version is readily available by upgrading INLA from R with the command `inla.upgrade(testing=TRUE)`. As there is still a lack of documentation, there is a downloadable

worked-out case study in http://www.r-inla.org/examples/case-studies/lindgren-rue-and-lindstrom-rss-paper-2011 that demonstrates the functionality of the module.

3 Estimation and prediction using INLA

In what follows we present the basis of how to perform the fitting and prediction in unobserved locations for the Latent Gaussian model in (5) using INLA's SPDE module and a brief guide to its syntaxis. It is worth saying that both model fitting and prediction are done simultaneously. Moreover, the fact that INLA can be used through R provides a familiar interface with the model specification, which is accomplished through the R's *formula* approach. However, INLA provides some additional syntaxis for the definition of random effects, namely the f() terms.

Using this syntaxis, the latent field in model (5) can be specified as

$$\text{formula} = \text{Y} \sim 1 + \text{X} + \text{f}(\text{W}, \text{model} = \text{spde})$$

where 1 stands for the intercept term, X is a fixed linear effect and W represents a smooth spatial effect. More terms could be added in the same way if additional covariates were available (for instance, + X2 + X3) or if a noise term were required (+ f(U, model = 'iid')). It is worth noting that while X is a variable containing the covariate values at each observation, W is only a numeric vector linking each observation with a spatial location.

INLA provides different approximation strategies for the posterior marginal distributions. In this study we have used the default ones: the simplified Laplace approximation for the marginalization, and the Central Composite Design for the numerical integration of the hyperparameters. These are the default and recommended settings providing reasonable accuracy with maximum computational efficiency (Held et al. 2010).

The standard output of a run returns the marginal posterior distributions for all the parameters in the model as well as summary statistics, together with several model selection and predictive measures. Specifically, the Deviance Information Criterion (DIC) is a well-known Bayesian model-choice criterion for comparing complex hierarchical models (Spiegelhalter et al. 2002). Additionally, the Conditional Predictive Ordinate (CPO, Geisser 1993) is defined as the cross-validated predictive density at a given observation, and can be used to compute predictive measures such as the logarithmic score (Gneiting and Raftery 2007) or the cross-validated mean Brier Score (Schmid and Griffith 2005). The latter is more adequate for a binary response, measuring the degree to which the fitted probabilities of fish presence at location i coincide with the observed binary outcomes Z_i (Roos and Held 2011).

As mentioned above, along with the inferential results about the parameters in (5), INLA's SPDE module can be used simultaneously to perform prediction in unobserved locations, which constitutes the real interest in this problem. The basic idea is to deal with the species' occurrence at a new location as a random variable with a certain probability of *success* and to calculate a point estimation of this probability, and even its full predictive density.

The SPDE module has a handful of functions to create prediction locations. It allows the construction of a Delaunay triangulation (Hjelle and Daehlen 2006) covering the region. As opposed to a regular grid, a triangulation is a partition of the region into triangles, satisfying constraints on their size and shape in order to ensure smooth transitions between large and small triangles. Initially, observations are treated as initial vertices for the triangulation, and extra vertices are added heuristically to minimize the number of triangles needed to cover the region subject to the triangulation constraints. These extra vertices are used as prediction locations. This has at least two advantages over a regular grid. First, the triangulation is denser in regions where there are more observations and consequently there is more information, and more detail is needed. Second, it saves computing time, because prediction locations are typically much lower in number than those in a regular grid. This partition is usually called *mesh* and an example (the one obtained using the data introduced in the following section) can be appreciated in Fig. 1.

Once the prediction is performed in the selected location, there are additional functions that linearly interpolate the results within each triangle into a finer regular grid. As a result of the process, a faceted surface prediction is obtained which approximates to the true predictive surface.

The prediction in INLA is performed simultaneously with the inference, considering the prediction locations as points where the response is missing.

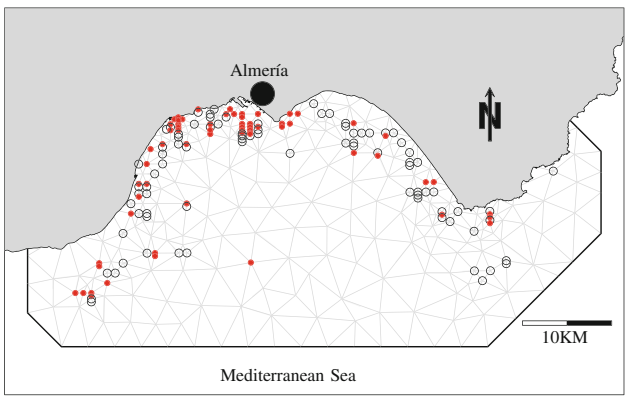


Fig. 1 Sampling locations for the presence (*filled circle*) and the absence (*open circle*) of the Mediterranean horse Mackerel in the bay of Almeria; each mesh vertex is either an observed point or a prediction point

Please refer to http://www.r-inla.org and references therein for directions on how to use INLA for inference and prediction.

4 Presence of Mediterranean horse mackerel in the bay of Almería

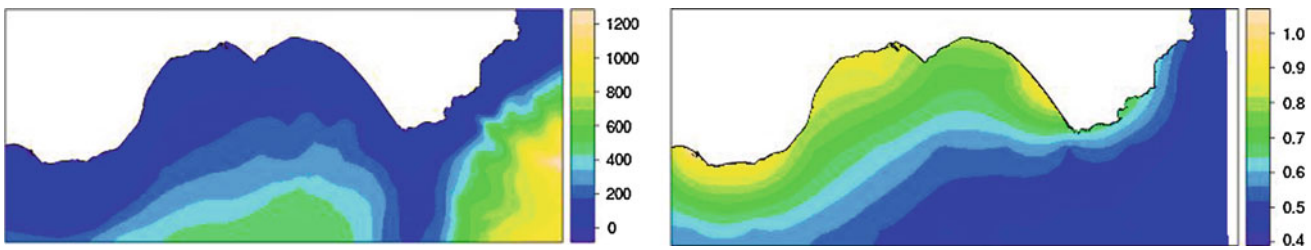
In order to show the usefulness of the approach presented, we have applied it to analyze the distribution of Mediterranean horse mackerel (*Trachurus mediterraneus*) in the bay of Almería, Spain (see Fig. 1 for a map of the region). The observed area is a transition zone between the Mediterranean and Atlantic sea, containing a mix of fish species and characteristics (Tintore et al. 1991). It is worth noting that, in spite of its low commercial value, this species plays an important role in the ecosystem being a food source for other commercially important predators (Froese and Pauly 2011). But more importantly, this is not a targeted species for commercial fishing, so its occurrence is an unbiased indicator of its presence/absence pattern. Moreover, it also means that the selection of the sampling locations does not depend on the values of the spatial variable and so these are stochastically independent of the field process. This is an important issue as it allows us to predict in all the locations of the bay, including those in which there is no information about the presence/absence of the species.

The reference fleet for this study was the purseine fleet with landings in the southwestern Spanish ports. This fleet operates in waters on the continental shelf around 200 m. isobaths. The fishing time for each haul lasts around one hour. The data set includes 147 hauls of 15 different purseine vessels and has been provided by the *Instituto Espanol de Oceanografia* (IEO, Spanish Oceanographic Institute). The IEO provides the national input of the European Plan for collecting fishery data. In particular, they collect samples from the commercial fleet with observers on board. This sampling has been carried out for six years, usually involving about 2-3 observations every month. From this database we have used the geographical location and occurence of the mackerel for each haul.

With respect to the environmental covariates used in this analysis, we have included those we had information about and those we thought were potentially relevant for a pelagic species like Mediterranean horse mackerel. In particular, the two covariates used were chlorophyll-*a* (an environmental covariate that usually provides great spatial and temporal coverage Valavanis et al. 2004) and bathymetry (see Fig. 2 for two maps of both covariates in the region analyzed). The chlorophyll-*a* data were obtained from satellite data provided by the IEO, while the bathymetry data were obtained from the WFS service of the Spatial Data Infrastructure of the Junta de Andalucia (Andalucian Local Government). It is worth noting that if we had had information about other factors such as precipitation, sea surface temperature, etc., they could have been included in the analysis via the linear predictor.

All the resulting models obtained from combining those two covariates and the logarithm of the bathymetry were fitted and compared. DIC was used as a measure for goodness-of-fit, while the logarithmic score (LCPO) and the cross-validated mean Brier score (BS) measure the predictive quality of the models. As shown in Table 1, all measures agree on the same model, with a reasonable predictive quality. In particular, the model comparison indicates that (apart from the spatial effect) the logarithm of the bathymetry and the chlorophyll-*a* concentration play a determining role in Mediterranean horse mackerel distribution.

As can be seen in Table 2 and Fig. 3, both covariates have a significant influence on driving the mackerel distribution. Table 2 shows a numerical summary of the posterior distribution of the effects, shown in Fig. 3. In both cases, they show that depth affects the distribution of the species studied negatively, while the chlorophyll-*a* concentration has a positive relationship. Results therefore indicate that the occurrence of Mediterranean horse mackerel is greater in shallow waters (near the coast)



(a) Bathymetry **(b)** Chlorophyll-*a*

Fig. 2 Maps of the covariates in the bay of Almeria. The bathymetry map is presented as it was obtained via the Andalucian Government, while the Chlorophyll-*a* is the result of the IEO processing of satellite data

Table 1 Model comparison

	Model	LCPO	BS	DIC
1	1	0.69	0.25	202.77
2	1 + Depth	0.69	0.24	200.87
3	1 + logDepth	0.67	0.24	197.03
4	1 + Chlorophill-*a*	0.67	0.24	197.60
5	1 + θ	0.66	0.23	195.59
6	1 + Depth + Chlorophill-*a*	0.67	0.23	196.19
7	1 + Depth + θ	0.67	0.23	195.13
8	1 + logDepth + Chlorophill-*a*	0.66	0.23	192.18
9	1 + logDepth + θ	0.65	0.23	191.21
10	1 + Chlorophill-a + θ	0.65	0.23	192.18
11	1 + Depth + Chlorophill-*a* + θ	0.66	0.23	191.48
12	1 + logDepth + Chlorophill-*a* + θ	0.64	0.22	187.83

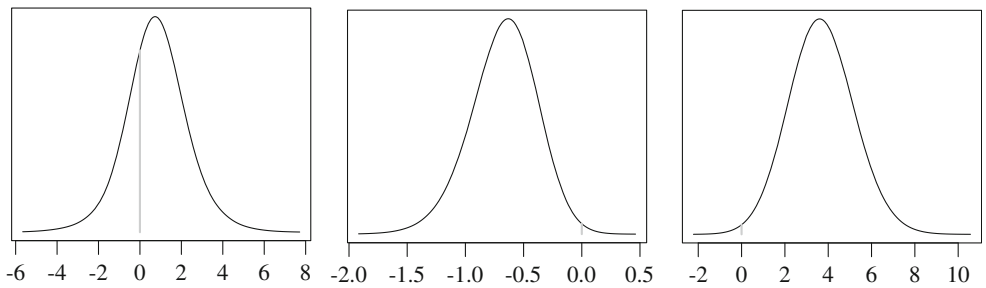
Table 2 Numerical summary of the posterior distributions of the fixed effects

	Mean	s.d.	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$
(Intercept)	0.80	1.56	−2.31	0.78	3.99
log depth	−0.67	0.29	−1.29	−0.66	−0.14
Chlorophyll-*a*	3.69	1.52	0.79	3.66	6.75

where the concentration of the chlorophyll-*a* is higher with to respect to deeper waters. The underlying reason may be that Mediterranean horse mackerel is a pelagic migratory fish occurring at a depth of between 40 and 500 m., usually in surface waters, but at times near the bottom (Ragonese et al. 2003).

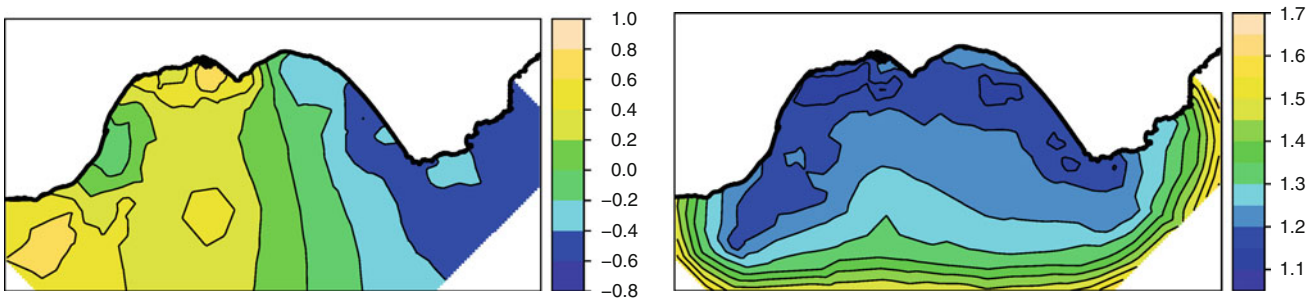
In Ecology, chlorophyll-*a* can be used as an indicator of the primary production of an ecosystem. The spatial variability of the primary production modifies trophic conditions (Katara et al. 2008) of the examined area and thus the distribution of the marine species. Coastal waters are usually zones of high productivity while in surface waters away from coastlines, there is generally plenty of sun but insufficient nutrients. In our case, although captures were scanty in the upper part of the slope (down to 300 m. depth, see Fig. 1), mackerel was caught on the shelf over practically all the area investigated.

Figure 4 displays the posterior mean and standard deviation of the spatial component. This component shows a strong effect with positive values in the western part of the bay of Almeria, with values around zero in the middle and with negative values in the eastern part of the area. This results in a clear dependence with respect to longitude in Mediterranean horse mackerel distribution. The western area of the bay of Almeria is a protected coastline, the Punta Entinas-Sabinar Natural Park, made up of sand dunes



(a) Intercept **(b)** log(Depth) **(c)** Chlorophill-a

Fig. 3 Posterior distributions of the fixed effects



(a) Mean **(b)** Standard deviation

Fig. 4 The posterior mean (*left*) and standard deviation (*right*) of the spatial effect

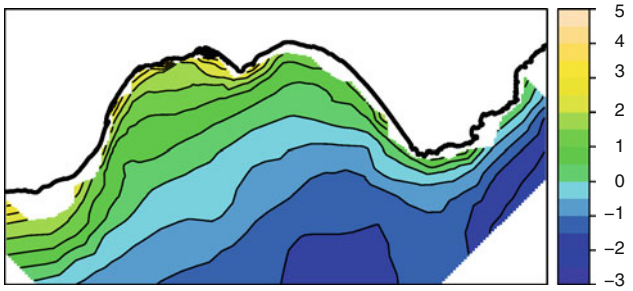


Fig. 5 Posterior mean of the lineal predictor

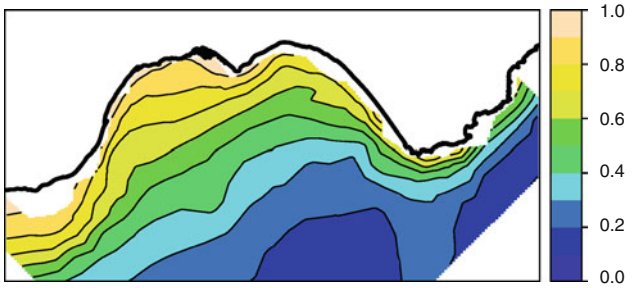


Fig. 6 Median for $\pi_i|\mathbf{Z}$

interspersed with a series of freshwater and saline lakes. Its size and development are directly associated with groundwater flows that, jointly with strong hydrochemical variability and an anthropogenic influence due to intensive agriculture, produce a significant concentration of nutrients in the coastal waters. All these factors make this a highly productive area that is the ideal habitat for Mediterranean horse mackerel.

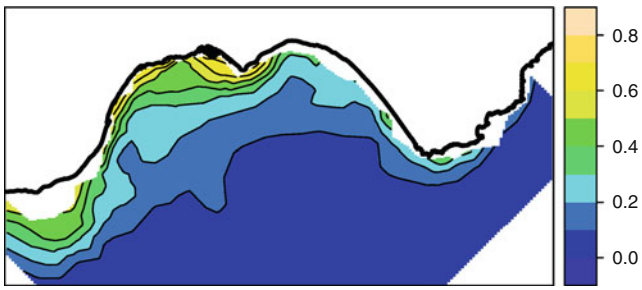
We can also obtain a precise estimation of the complete linear predictor by calculating the corresponding combination of the means of the different effects, as shown in Fig. 5. The posterior mean of the linear predictor confirms that depth plays a key role in the distribution of Mediterranean horse mackerel, along with the concentration of chlorophyll-*a*. Along the coast, mean values of the linear predictor show positive values, where the concentration of chlorophyll-*a* is higher, and as we move away from the coast to the offshore area the mean values become negative.

In order to make the results more understandable, we have also generated maps of predicted probabilities of occurrence using the distribution of the parameter π_i. In this specific case, it is not a linear transformation from the linear predictor, so it is not possible to compute the posterior distribution of the parameter π_i. However, we can obtain any quantile using the corresponding quantiles of the linear predictor.

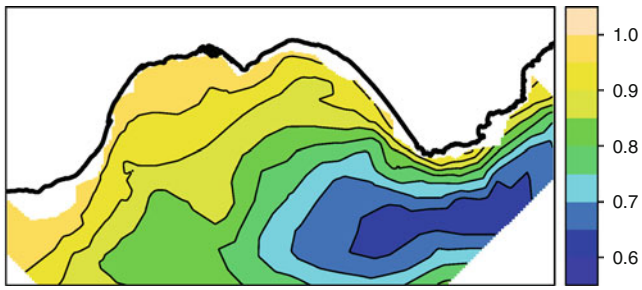
Figure 6 shows the median posterior probability of occurrence, while Fig. 7 shows the first (a) and third (b) quartiles for this probability. In this way we get not only a point estimate for the probability of occurrence, but also an assessment of the uncertainty of this estimation. Figure 6 confirms that the probability of finding this species is greater in areas near the coast at a shallow depth and where the chlorophyll-*a* concentration is higher. In deeper waters the occurence probability is lower where the nutrient concentration is less. Also, the western part of the bay of Almeria shows a higher probability of occurence with respect to the eastern zone due to the presence of the Natural Park and the intensive agriculture that releases a high concentration of organic material into the sea.

5 Concluding remarks

The main advantage of the Bayesian model formulation is the computational ease in model fit and prediction compared to classical geostatistical methods. Both the stationary and especially the non-stationary models have a large number of parameters. Also, in classical geostatistical applications, the full range of uncertainties that are always associated with species distribution models is not correctly measured, as many parameters that are considered to be "known" are actually estimated through the statistical model (Diggle and Ribeiro 2007), a potential cause of optimistic assessments of predictive accuracy. Using Bayesian kriging, we have incorporated parameter uncertainty into the prediction process.



(a) First quartile



(b) Third quartile

Fig. 7 Variability for $\pi_i|\mathbf{Z}$

The main goal of this study has been to predict the occurrence of the species in unsampled areas. To do so, instead of MCMC we have used the novel integrated nested Laplace approximation approach. More precisely, we have applied the work of Lindgren et al. (2011), which provides an explicit link between Gaussian Fields and Gaussian Markov Random Fields through the Stochastic Partial Differential Equation approach. Thanks to the RINLA library, the SPDE approach can be easily implemented providing results in reasonable computing time (in contrast to MCMC algorithms). The simplicity of the SPDE parameter specifications provides a new modelling approach that allows an easy construction of non-stationary models that provides a good, computationally very efficient, local interpretation. For these reasons, the SPDE approach has proved to be a powerful strategy for modelling and mapping complex spatial occurence phenomena.

This modelling could be expanded to the spatiotemporal domain by incorporating an extra term for the temporal effect, using parametric or semiparametric constructions to reflect linear, nonlinear, autoregressive or more complex behaviours. Although the inclusion of independent effects on spatial and temporal domains would be straightforward (for instance using a non-structured random effect), it must be taken into account that introducing non-separable spatiotemporal structures could be much more difficult. A first analysis in this line can be seen in Andrianakis and Challenor (2012). Nevertheless, in our case, the information available did not include a reasonable enough number of years for performing any temporal analyses.

To conclude, we would like to mention that we have described an application to a set of fishery data of Mediterranean horse mackerel from the bay of Almeria, western Spain, to illustrate this approach. The results have shown that the distribution of Mediterranean horse mackerel is influenced by a spatial effect, as well as the depth and the concentration of chlorophyll-*a*.

Finally, we would also like to mention that the analytical approach we used here to document the spatial patterns in the distribution of Mediterranean horse mackerel can be extended to different species and life stages to improve knowledge of the role of habitat for populations and communities.

Acknowledgements This paper was mainly written while Maria Grazia Pennino was visiting the department of Statistics and Operations Research at the University of Valencia. David Conesa, Antonio López-Quílez and Facundo Muñoz would like to thank the Ministerio de Educación y Ciencia for financial support (jointly financed with European Regional Development Fund) via the research Grant MTM2010-19528 and of the Generalitat Valenciana via the research Grant ACOMP11/218. We would also like to thank Havard Rue and Finn Lindgren for their prompt support with technical aspects in the usage of INLA.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions: with formulas, graphs, and mathematical tables, ninth dover printing, tenth gpo printing edn. Dover books on mathematics, Dover

Andrianakis I, Challenor PG (2012) A bayesian hierarchical model for the reconstruction of the sea level in the mediterranean basin for the late 20th century. In: Poster presentation at the ISBA 2012 conference

Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. CRC, London

Busby JR (1991) BIOCLIM: A bioclimatic analysis and predictive system. In: Margules C, Austin M (eds) Nature conservation: cost effective biological surveys and data analysis, CSIRO, Canberra , pp 64–68

Chakraborty A, Gelfand AE, Wilson AM, Latimer AM, Silander JA (2010) Modeling large scale species abundance with latent spatial processes. Ann Appl Stat 4(3):1403–1429

Chien LC, Bangdiwala SI (2012) The implementation of bayesian structural additive regression models in multi-city time series air pollution and human health studies. Stochastic environmental research and risk assessment doi:10.1007/s00477-012-0562-4, in press

Cressie N, Calder CA, Clark JS, Hoef JMV, Wikle CK (2009) Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Ecol Appl 19:553–5701

Diggle PJ, Ribeiro PJ (2007) Model-based geostatistics. Springer series in statistics, Springer, New York

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. J R Stat Soc Ser C Appl Stat 47(3):299–350

Fahrmeir L, Tutz G (2001) Multivariate statistical modelling based on generalized linear models, 2nd edn. springer series in statistics, Springer, New Yorkê

Froese R, Pauly D (2011) Fishbase. World Wide Web electronic publication http://www.fishbase.org, version (10/2011)

Gaudard M, Karson M, Linder E, Sinha D (1999) Bayesian spatial prediction. Environ Ecol Stat 35(6):147–171

Gaudard M, Karson M, Linder E, Sinha D (2006) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecol Biogeography 16(2):129–138

Geisser S (1993) Predictive inference. monographs on statistics & applied probability, CRC, London

Gelfand AE, Ravishanker N, Ecker MD (2000) Modeling and inference for point-referenced binary spatial data. In: Dey D, Ghosh S, Mallick B (eds) Generalized linear models: a Bayesian perspective, Marcel Dekker Inc., New York, pp 381–394

Gelfand AE, Silander JA, Wu SJ, Latimer AM, Rebelo PLAG, Holder M (2006) Explaining species distribution patterns through hierarchical modeling. Bayesian Anal 1(1):41–92

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009

Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol model 157:89–100

Guttorp P, Gneiting T (2006) Studies in the history of probability and statistics XLIX: On the matérn correlation family. Biometrika 93(4):989–995, http://dx.doi.org/10.1093/biomet/93.4.989

Haining R, Law J, Maheswaran R, Pearson T, Brindley P (2007) Bayesian modelling of environmental risk: a small area ecological study of coronary heart disease mortality in relation to

modelled outdoor nitrogen oxide levels. Stochastic Environ Res Risk Assess 21(5):501–509
Handcock MS, Stein ML (1993) A Bayesian analysis of kriging. Technometrics 35(4):403–410
Held L, Schrödle B, Rue H (2010) Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA, Springer, New York
Hijman R, Graham C (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biol 12(12):2272–2281
Hjelle Ø, Dæhlen M (2006) Triangulations and applications. mathematics and visualization, Springer, New York
Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211
Katara I, Illian J, Graham J, Scott B, Wang J (2008) Atmospheric forcing on chlorophyll concentration in the Mediterranean. Hydrobiologia 612:33–48
Kneib T, Muller J, Hothorn T (2008) Spatial smoothing techniques for the assessment of habitat suitability. Environ Ecol Stat 15(3):343–364
Leathwick JR, 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. Freshwater Biol 50:2034–2052
Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). J R Stat Soc Ser B 73:423–498
Loarie S.R., Carter B.E., Hayhoe K., McMahon S., Moe R., Knight C.A., Ackerly D.D. (2008) Climate change and the future of Californias endemic flora. PLoS ONE 3(6):e2502
Midgley G.F., Thuiller W. (2007) Potential vulnerability of Namaqualand plant diversity to anthropogenic climate change. Journal of Arid Environments 70:615–628
Pressey R.L., Cabeza M., Watts E.M., Cowling R.M., Wilson K.A. (2007) Conservation planning in a changing world. Trends in Ecology and Evolution 22:583–592
R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna http://www.R-project.org/
Ragonese R, Fiorentino F, Garofalo G, Gristina M, Levi D, Gancitano S, Giusto GB, Rizzo P, Sinacori G (2003) Distribution, abundance and biological features of picarel (*Spicara flexuosa*), Mediterranean (*Trachurus mediterraneus*) and Atlantic (*T. trachurus*) horse mackerel based on experimental bottom-trawl data (MEDITS, 1994–2002) in the Strait of Sicily. MedSudMed Technical Documents 5:100–114
Roos M, Held L (2011) Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Anal 6(2): 259–278
Royle J.A., Kery M., Gautier R., Schmidt H. (2007) Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecol Monogr 77:465–481
Rue H, Held L (2005) Gaussian Markov random fields. Theory and applications. CRC, New York
Rue H., Martino S., Chopin N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society, Series B 71(2):319–392
Schmid CH, Griffith JL (2005) Multivariate classification rules: calibration and discrimination, 2nd edn, Wiley, pp 3491–3497
Schrödle B., Held L. (2011) Spatio-temporal disease mapping using INLA. Environmetrics 22:725–734
Spiegelhalter DJ, Thomas A, Best NG (1999) Winbugs version 1.2 user manual. MRC Biostatistics Unit
Spiegelhalter D.J., Best N., Carlin B., vanderLinde A. (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64:583–616
Stein ML (1999) Interpolation of spatial data: some theory for Kriging, 1st edn. Springer, New York
Tintore J., Gomis D., Alonso S., Parilla G. (1991) Mesoscale dynamics and vertical motion in the Alboran Sea. Journal of Physical Oceanography 21:811–823
Underwood A.J. (1981) Techniques of analysis of variance in marine Biology and Ecology. Oceanography and Marine Biology Annual Review 19:513–605
Valavanis DV, Katara I, Palialexis A (2004) A GIS-based modelling approach for the mapping of marine productivity hotspots. Aquat Sci 36:234–243
Valavanis D.V., Pierce G.J., Zuur A.F., Palialexis A., Saveliev A., Katara I., Wang J. (2008) Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS. Hydrobiologia 612:5–20
Wisz M.S., Hijmans R.J., Li J., Peterson A.T., Graham C.H., Guisan A. (2008) Effects of sample size on the performance of species distribution models. Divers Distrib 14:763–773
Zhang W. (2007) Supervised neural network recognition of habitat zones of rice invertebrates. Stochastic Environmental Research and Risk Assessment 21(6):729–735
Zhang W., Zhong X., Liu G. (2008) Recognizing spatial distribution patterns of grassland insects: neural network approaches. Stochastic Environmental Research and Risk Assessment 22(2):207–216