Spatial Statistics 22 (2017) 434–450

Contents lists available at ScienceDirect

Spatial Statistics

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

Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling

Iosu Paradinas [a,b,*], David Conesa [b], Antonio López-Quílez [b], José María Bellido [c]

[a] *Asociación Ipar Perspective, Karabiondo kalea 17, 48600, Sopela, Basque Country, Spain*
[b] *Departament d'Estadística i Investigació Operativa, Universitat de València, C/ Dr. Moliner 50, Burjassot, 46100, Valencia, Spain*
[c] *Instituto Español de Oceanografía. Centro Oceanográfico de Murcia, C/ Varadero 1, San Pedro del Pinatar, 30740, Murcia, Spain*

ARTICLE INFO

Article history:
Received 30 October 2016
Accepted 2 August 2017
Available online 1 September 2017

Keywords:
Bayesian hierarchical models
Bayesian kriging
Integrated nested Laplace approximation (`INLA`)
Species modelling distribution
Hurdle model
Two-part model

ABSTRACT

Understanding the spatio-temporal dynamism and environmental relationships of species is essential for the conservation of natural resources. Many spatio-temporally sampled processes result in continuous positive $[0, \infty)$ abundance datasets that have many zero values observed in areas that lie outside their optimum niche. In such cases the most common option is to use two-part or hurdle models, which fit independent models and consequently independent environmental effects to occurrence and conditional-to-presence abundance. This may be correct in some cases, but not as much in others where the detection probability is related to the abundance.

The aim of this work is to infer the spatio-temporal dynamism of ecological processes and to fit more robust environmental effects in two-part models. On the one hand we propose different spatio-temporal structures to infer the fundamental spatio-temporal behaviour of the process under study. On the other hand, we propose the use of shared component modelling (SCM) to estimate more robust model effects in related semi-continuous datasets by combining information from occurrence and conditional-to-presence abundance. We use a simulation study to test the application of shared components in two different types of semi-continuous datasets. Lastly, we implement all the proposed model structures

[*] Corresponding author at: Asociación Ipar Perspective, Karabiondo kalea 17, 48600, Sopela, Basque Country, Spain.
E-mail address: paio@uv.es (I. Paradinas).

http://dx.doi.org/10.1016/j.spasta.2017.08.001
2211-6753/© 2017 Elsevier B.V. All rights reserved.

in a case study on the recruitment of hake in the western Mediterranean.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Species distribution modelling links spatially referenced records of species with maps of environmental variables in order to create a statistical model of the relationship between a species and its environment. However, most natural processes also change in time and space driven by a set of unknown factors and interactions that result in spatially and/or temporally autocorrelated model residuals.

Ignoring these dependencies, as most of the popular generalized linear and additive models (GLM and GAM) do, not only restricts their predictive capacity, but may also lead to incorrect results (Fortin and Dale, 2009; Legendre et al., 2002). A spatial autocorrelation term includes such spatial dependency among neighbouring locations based on the principle that close locations have more in common than distant ones (Tobler, 1970), resulting in better predictions at unsampled locations (Krige, 1951).

Sampled locations are often aggregated into previously arranged spatial areas (e.g. grids, regions, etc.), usually referred to as lattice or areal data (Cressie, 2015). This type of data is sensitive to the selected grid size, which can influence our results and conclusions (Dinmore et al., 2003). Geostatistics, on the other hand, incorporates spatial autocorrelation among point-referenced observations (Gelfand et al., 2000), allowing us to predict outcomes across a natural continuous field.

Many spatially sampled studies are repeated periodically for (or over) long periods of time (Aizpurua et al., 2015; Gitzen, 2012) (e.g. fisheries research surveys, plant coverage surveys, air pollution surveys, etc.). A fundamental reason for such a research plan is the interest in the spatial evolution of the system under study, whose intensity is expected to vary in both time and space. Similar to spatial statistics, time series analysis is based on the principle that long runs of repeated measurements may display autocorrelation. Therefore, as with the spatial domain, temporally close observations tend to be more related than temporally distant observations (Cressie and Wikle, 2011). As a consequence, including a temporal correlation term is likely to improve both our models and predictions.

Temporal correlation depends basically on the same principle as spatial correlation but since temporal and spatial scales are different, spatio-temporal analysis is more complicated than the simple addition of an extra dimension to the continuous spatial domain (Cressie and Wikle, 2011). Moreover, most ecological surveys sample the target study area during a fixed time window of the year, which obliges scientists to both assume static systems during each time window and to discretize time.

Another common feature in environmental datasets is the semi-continuous nature of the response variable. Semi-continuous processes are measured in the $[0, \infty)$ interval (rain, plant coverage, chemical concentrations, etc.) for which neither standard probability distribution nor straightforward transformation is available. A popular approach is to model these data as two independent processes, known as two-part models or hurdle models. In these models, one process determines probability of occurrence, while the second determines the intensity when the response is non-zero (Martin et al., 2005).

Nevertheless, although this popular representation (Balderama et al., 2016; Neelon et al., 2013) is statistically convenient, it is not always correct. The reason underneath is that there are two possible ways of producing a semi-continuous dataset. One is when it is produced by two unrelated processes, e.g. there is little probability of rain in dry climates but if it rains, it can rain a lot. In such case an independent two-part model makes sense since the probability of rain is not necessarily related to its abundance. But a semi-continuous dataset can also be originated from a single abundance process where the presence–absence process is a consequence of the underlying abundance process. It seems sensible to assume that the detection probability of a species depends on its relative abundance and the sampling effort (e.g. time observing, size of the net, etc.). This translates into proportional-to-abundance detection probabilities given the unit effort. In such cases, conventional independent two-part models could be prone to overfitting the data because each sub-process ignores valuable information from each other, i.e. the distribution of zeros in the abundance sub-process and the actual abundance in the presences of the occurrence sub-process.

Shared component analysis (SCM) allows fitting common latent fields among related processes (Hogan and Laird, 1997) and in the case of semi-continuous datasets, it also allows to link low abundance intensities to low probabilities of occurrence and vice versa. SCM has already been widely used to characterize the relationship between longitudinal and time-to-event processes (Hogan and Laird, 1997; Tsiatis and Davidian, 2004) and has also been introduced in the context of disease mapping by Knorr-Held and Best (2001) in order to model the relationship between different diseases. In spite of its great value to fit related two-part models, SCM has not been used much to fit semi-continuous datasets. A good example of their application in that context can be found in the work by Quiroz et al. (2015), who applied them in an anchovy fisheries spatial dataset. Although their final model did not consider any shared effects, the comments raised there already showed the possible usefulness of this kind of models.

The purpose of this study is twofold. On the one hand, we propose the comparison of four generic spatio-temporal structures to characterize the behaviour of the process under study in space and time. The idea behind comparing model structures is that by inferring the goodness-of-fit of one structure over the rest, we can provide valuable information on the overall behaviour of the spatial distribution over time, the spatio-temporal interpretation of spatial statistics (Waller, 2014). On the other hand, we test the use of SCM as a more realistic approach to fit semi-continuous datasets where the abundance and the detection probability are related. In these cases SCM allows embracing information from both the abundance and occurrence sub-processes to fit better process-environment relationships.

We illustrate the appropriate scenario where SCM provides a valuable tool to fit semi-continuous datasets using a simulation study. Then, as case study, we analyse an important issue in fisheries management, the identification of important fishery ecosystems (FAO, 2008). Fish distribution maps play an essential role in assessing these areas, and it is therefore important to improve the quality and accuracy of these maps. In particular, we further investigate a previously published study on the recruitment of hake (*Merluccius merluccius*) in the western Mediterranean (Paradinas et al., 2015) by providing more refined hake recruit distribution maps.

With all this in mind, the remainder of this paper is organized as follows: in Section 2, we briefly describe Gaussian latent models for species distribution modelling and the underpinnings of `INLA` (Rue et al., 2009) as a tool for performing fast Bayesian inference. Section 3 introduces the spatio-temporal extensions proposed to infer the spatio-temporal behaviour of the process under study. Then, in Section 4 we discuss the suitability of SCM to deal with two different types of semi-continuous processes and use a simulation study in Section 5 to illustrate this. In Section 6 we present a case study on hake recruitment, the results of its analysis and a brief discussion on the improvements achieved compared to previous studies. The paper ends with some conclusions in Section 7.

2. Gaussian latent models for species distribution modelling

Spatial distribution modelling involves linking spatially referenced observations to maps of environmental variables in order to understand and predict the distribution of a wide range of processes: diseases, air pollution, natural resources, etc. Unfortunately, the unmeasurable complexity of most ecological spatio-temporal processes is hardly ever completely explained by the collected environmental variables. This often results in spatially and/or temporally autocorrelated model residuals that may compromise both model fit and prediction (Fortin and Dale, 2009; Legendre et al., 2002). Therefore, a good species distribution model often needs to account for these dependencies, specially when management decisions have to be made, as in the case of disease alerts, marine protected areas, climate change, etc.

Most of the available data in the context of species distribution modelling come from designed field-based monitoring programmes. In general, three main characteristics would determine the resulting model, the first of these being the particular type of spatial data (point-referenced data in this case) and the second the inclusion of possible covariates in the modelling. The third and final piece of the puzzle is the variable of interest (response variable). Depending on the process under study and the sampling design, the nature of the response variable frequently differs from the usual Gaussian distribution, so other likelihoods may need to be considered. Discretely measured processes are typical examples where normality cannot be assumed, e.g. occurrence data require a Bernoulli distribution and count data Poisson or negative binomial distributions. Similarly, non-transformed continuous abundance processes, by definition, only take positive values, and thus the use of log-normal or Gamma distributions may be more appropriate. The same occurs when modelling proportions, bounded to the [0,1] interval, where the beta distribution is the natural option of choice.

For the sake of simplicity, we will only present here the most typical case, i.e., that of occurrence data, where the process under study is modelled as a binary presence/absence random variable. It is worth noting that other response variables would produce similar models (with their corresponding link functions). In particular, if Y_s denotes the spatially distributed occurrence process at location s, then:

$$\begin{aligned} Y_s &\sim \text{Ber}(\pi_s),\ s = 1, \ldots, n \\ \text{logit}(\pi_s) &= \boldsymbol{X}_s\boldsymbol{\beta} + W_s \end{aligned} \tag{1}$$

where the probability of presence π_s is linked to the linear predictor through the usual logit link and $X_s\boldsymbol{\beta}$ denotes the fixed effects of the model. The spatial effect $\boldsymbol{W} \sim N(0, \sigma^2\boldsymbol{H}(\phi))$ is a Gaussian latent field with zero mean and covariance matrix $\sigma^2\boldsymbol{H}(\phi)$, that depends on the Euclidean distance between observations, with hyperparameters σ^2 and ϕ representing respectively the variance and the range of the spatial effect.

The spatial model in (1) has been described under the generic term of model-based geostatistics (Diggle and Ribeiro, 2007), although it can also be naturally represented as a hierarchical model (Gelfand, 2012), especially under the Bayesian paradigm which allows for a more realistic estimation of uncertainty (Banerjee et al., 2003; Haining et al., 2007). In line with this, the previous model can also be represented by means of the following hierarchical Bayesian spatial model with three stages

$$\begin{aligned} &1. \quad \boldsymbol{Y}_s|\boldsymbol{\beta}, W_s \sim \text{Ber}(\boldsymbol{X}_s\boldsymbol{\beta} + W_s) \\ &2. \quad \boldsymbol{W}|\sigma^2, \phi \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{H}(\phi)) \\ &3. \quad \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \phi) \sim p(\boldsymbol{\theta}) \end{aligned} \tag{2}$$

where the first level constitutes the conditionally independent likelihood; the second defines the spatial latent field (to which we attribute a Gaussian distribution with zero mean and covariance matrix $\sigma^2\boldsymbol{H}(\phi)$); and lastly, the third is formed by the prior distributions of the parameters and hyperparameters, where $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}, \sigma^2, \phi)$ can be any sensible prior distribution.

As presented by Rue et al. (2009), Bayesian hierarchical Gaussian latent models, such as the model in (2), can be tackled very effectively by using the integrated nested Laplace approximation (INLA). In these cases, INLA allows Bayesian inference and, more importantly prediction in unknown locations to be performed by means of numerical Laplace approximations, which significantly improve the computational costs of the traditional simulation based Markov Chain Monte Carlo method. This method is readily available in the user friendly `INLA` package for R (Martins et al., 2013).

However, geostatistical models such as the one presented in (2) face the well known "big n problem" (Banerjee et al., 2003). This problem is related to the cost of factorizing the dense $n \times n$ covariance matrix of $\boldsymbol{W}$, which requires n^3 operations (Stein et al., 2004). A handful of solutions have been proposed to tackle the "big n problem" (see Lindgren et al., 2011 for a short summary). In `INLA`, this problem is tackled following the work by Lindgren et al. (2011), where continuously indexed Gaussian Fields with Matérn covariance functions are approximated to discretely indexed Gaussian Markov Random Fields (GMRF) using the stochastic partial differential equation approach (SPDE). The SPDE approach basically uses a finite element representation to define the Matérn field as a linear combination of basis functions defined on the so-called *mesh*, i.e. a Delaunay triangulation (Hjelle and Dæhlen, 2006) of the study area. `INLA` uses this triangulation to construct an indexed observation matrix that links our observations to the spatio-temporal random field, allowing to fit spatio-temporally misaligned data without problem. Furthermore, this remarkable approximation enables computational costs to be reduced to around $n^{3/2}$ thanks to the good computational properties of the GMRFs (Rue and Held, 2005).

Using the SPDE approach, (1) can be reparametrized as:

$$\begin{aligned} Y_s &\sim \text{Ber}(\pi_s),\ s = 1, \dots, n_s \\ \text{logit}(\pi_s) &= \boldsymbol{X}_s\boldsymbol{\beta} + W_s \\ p(\beta_j) &\sim N(\mu_{\beta_j}, \sigma^2_{\beta_j}) \\ \boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ (\log(\kappa), \log(\tau)) &\sim MVN(\boldsymbol{\mu}, \boldsymbol{\rho}) \end{aligned} \tag{3}$$

where now the covariance function $\mathbf{Q}$ of the spatial effect depends on another two parameters, κ and τ, which determine the range of the effect and the total variance. By default INLA sets prior distributions to all the parameters, although all of those priors can easily be changed to a list of available distributions or to ones defined by the user (http://www.r-inla.org/models/priors.). In any case, as it can be seen at the last line of (3), INLA uses a joint prior distribution for the vectors κ and τ, in particular a multivariate normal distribution with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\rho}$ (Lindgren and Rue, 2015). Finally, it is worth noting that INLA also provides a number of model selection scores such as DIC (Spiegelhalter et al., 2002), WAIC (Watanabe, 2010), PIT (Dawid, 1984) and CPO (Geisser, 1993).

Following this approach, a handful of species distribution studies have been published in the last few years using different types of data: occurrence of species (Muñoz et al., 2013; Pennino et al., 2013; González-Warleta et al., 2013), number of individuals (Cosandey-Godin et al., 2014; Heegaard et al., 2016), biomass (Pennino et al., 2014; Quiroz et al., 2015; Paradinas et al., 2015) and proportions of species (Paradinas et al., 2016, in press). However, species monitoring programmes run for years or even decades, resulting in spatio-temporal point-referenced datasets. Such processes tend to be not only spatially but also temporally correlated. Similarly, few process–covariate relationships in nature are linear, thus requiring non-linear effects to be fitted in the model.

Our intention in what follows is precisely to extend the spatial model presented in (3) to the spatio-temporal framework by introducing some possible spatio-temporal structures that better reflect species distribution behaviour over time, as well as allowing the use of smoothing effects in the covariates.

3. Gaussian latent spatio-temporal structures

As mentioned above, the distribution of species not only changes in space but also in time. Consequently, extending the spatial model in (3) to a spatio-temporal model can provide further description and/or understanding of the species under study, improving the predictive capacity of our models. Indeed, depending on the nature of the process under study and the available sampling resolution, the spatio-temporal behaviour of the data can vary.

In order to incorporate possible spatio-temporal and smoothing effects, the model introduced in the previous section can be rewritten as:

$$\begin{aligned} Y_{st} &\sim \text{Ber}(\pi_{st}),\ s = 1, \dots, n \\ \text{logit}(\pi_{st}) &= \alpha + \sum_{i=1}^{I} f_i(x_{ist}) + U_{st} \end{aligned} \tag{4}$$

where $t = 1, \dots., T$ is the temporal index and $s = 1, \dots, n_t$ is the spatial location, potentially different at each t. U_{st} represents the spatio-temporal structure of the model, x_{ist} is the value of an explanatory variable i at a given st and f represents any latent model applied to the covariates (linear, non-linear, etc.). Note that smoothing effects in INLA can easily be incorporated using first or second order random walk (RW) latent models that, based on constant increments x, perform as Bayesian smoothing splines (Fahrmeir and Lang, 2001). In fact, RW models are expressed as computationally efficient GMRFs by Rue and Held (2005).

As mentioned above, the spatio-temporal behaviour of the data can vary. As a result, knowing which spatio-temporal behaviour better reflects the data becomes an important issue. (Paradinas et al., 2015 is a good example of how assessing the spatial persistence of nursery areas over time can be highly effective when designing marine protected areas.) A feasible way to do so is by comparing the goodness-of-fit of different spatio-temporal structures.

We here propose the comparison of four basic structures for U_{st} in (4), each one allowing different degrees of flexibility in the temporal domain of the spatio-temporal model:

- **Opportunistic spatial distribution**: a very flexible structure consists of decomposing U_{st} into different spatial realizations of the same spatial field for each time unit. This structure is a good proxy to those processes where the spatial structure varies considerably among different time units and unrelatedly among neighbouring times. In particular,

$$\begin{aligned} U_{st} &= W_{s_t} \\ \boldsymbol{W}_t &\sim N(0, \mathbf{Q}(\kappa, \tau)) \end{aligned} \tag{5}$$

  where U_{st} is decomposed in a different spatial realization $\boldsymbol{W}_t$ at each time t while sharing a common covariance function (same κ and τ) to avoid having too many hyperparameters in the model. This structure is likely to favour the goodness-of-fit of temporally inconsistent spatial processes. The first row of Fig. 1 shows a simulated opportunistic spatio-temporal distribution over four periods. This structure has already been used in Cosandey-Godin et al. (2014) and in Paradinas et al. (2015).
- **Persistent spatial distribution with random intensity changes over time**: another structure treats time as a zero mean Gaussian random noise effect V_t. This structure may perform well in those cases where mean intensities vary unrelatedly among time events but the spatial realization is similar for every time unit, that is,

$$\begin{aligned} U_{st} &= W_s + V_t \\ \boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ V_t &\sim N(0, \sigma_V^2) \end{aligned} \tag{6}$$

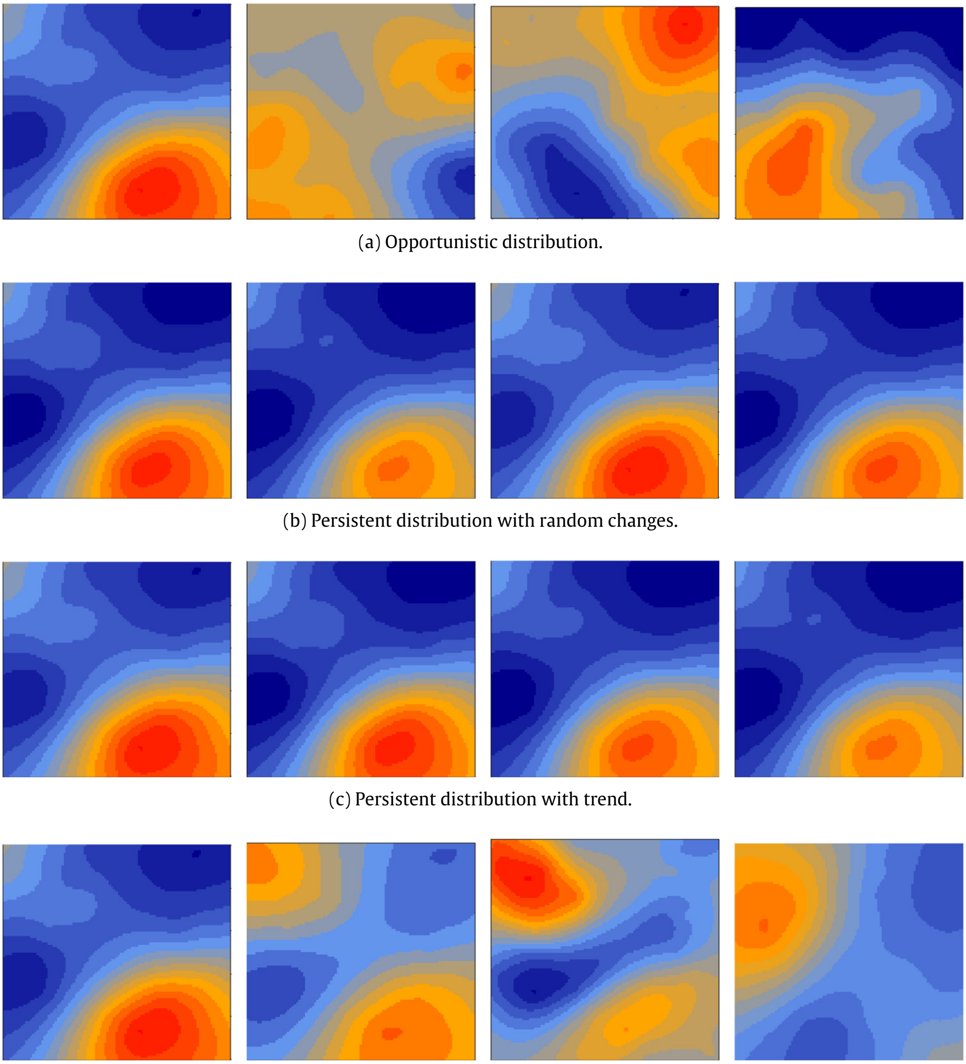
  where U_{st} is decomposed in a common spatial realization W_s along with a random noise effect V_t that absorbs the different mean intensities at each time t. This structure may better accommodate those processes where the spatial structure is somewhat persistent in time. The second row of Fig. 1 shows a simulated persistent spatio-temporal distribution with random intensity changes over four periods. This structure has been used by Pennino et al. (2014) and in Paradinas et al. (2015).
- **Persistent spatial distribution with temporal intensity trend**: alternatively, the process could show a temporal progression in its intensity. Such a case would best fit in our third proposed structure, which includes a temporal trend effect $g(t)$ through a linear or non-linear effect,

$$\begin{aligned} U_{st} &= W_s + g(t) \\ \boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \end{aligned} \tag{7}$$

  where U_{st} is decomposed in a common spatial realization W_s and a temporal trend $g(t)$ to absorb the temporal progression of the process. Processes where such a temporal tendency is present will benefit from this structure. The third row of Fig. 1 shows a simulated persistent spatio-temporal distribution with a negative temporal trend over four periods. This structure was proposed by Paradinas et al. (2016) to identify intra-annual trends in fishery discards.
- **Progressive spatio-temporal distribution**: our final proposed structure for U_{st} incorporates both spatial and temporal correlation of the data to accommodate those cases where the spatial realizations change in a related manner over time. In particular,

$$\begin{aligned} U_{st} &= W_{st} + R_{st} \\ \boldsymbol{W}_t &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ R_{st} &= \sum_{k=1}^{K} \rho_k U_{s(t-k)} \end{aligned} \tag{8}$$

  where U_{st} is decomposed in a common spatial realization W_{st} and an autoregressive temporal term R_{st} expressing the correlation among neighbours of order K. This structure may be



(a) Opportunistic distribution.

(b) Persistent distribution with random changes.

(c) Persistent distribution with trend.

(d) Progressive spatio-temporal distribution.

Fig. 1. Simulated types of spatio-temporal scenarios.

favoured when the spatial realization varies between different times *t* but not as much as in (5). Note also that this structure could be applied along with that in (7). The last row of Fig. 1 shows a simulated progressive spatio-temporal distribution over four periods. This structure has been used by Cameletti et al. (2011, 2013) and also by Cosandey-Godin et al. (2014).

It is rather evident that we could have proposed several more complex temporal structures. Unfortunately, the temporal resolution of spatio-temporal datasets is typically too low to fit highly structured models. In this regard, it is worth noting that comparison of the fit of these four basic spatio-temporal structures (5), (6), (7), and (8) allows us to discern the nature of the general spatial behaviour of the process over time.

4. Shared component analysis for two-part models

As previously introduced, many spatio-temporally sampled abundance processes are prone to zero value observations at non-favourable conditions (e.g. species abundance, plant coverage, etc.), and are thus measured continuously in the $[0, \infty)$ interval, resulting in semi-continuous datasets. The absence of distributions capable of dealing with such datasets has persuaded scientists to apply two-part models (Martin et al., 2005) by decomposing the dataset into two independent sub-processes, an occurrence process and a conditional-to-presence continuous process.

Following the notation of the occurrence model in (4), we represent a two-part model: Y_{st} being the occurrence sub-process and Z_{st} the conditional-to-presence abundance sub-process. In particular,

$$\begin{aligned} Y_{st} &\sim \text{Ber}(\pi_{st}) \\ \text{logit}(\pi_{st}) &= \alpha^{(1)} + \sum_{i=1}^{I} f_i^{(1)}(x_{is}) + U_{st}^{(1)} \\ Z_{st} &\sim \text{Ga}(a_{st}, b_{st}) \\ \log(\mu_{st}) &= \alpha^{(2)} + \sum_{i=1}^{I} f_i^{(2)}(x_{is}) + U_{st}^{(2)} \end{aligned} \tag{9}$$

where the probability of occurrence, π_{st}, is modelled through the usual logit link, and the abundance $\mu_{st} = a_{st}/b_{st}$ through its logarithm; $s = 1, \ldots.., n_t$ is the spatial location, and $t = 1, \ldots., T$ is the temporal index; the $\alpha^{(1)}$ and $\alpha^{(2)}$ terms represent the intercepts, $f^{(1)}$ and $f^{(2)}$ represent additive latent models applied to the covariates and both $U_{st}^{(1)}$ and $U_{st}^{(2)}$ refer to any of the spatio-temporal structures already introduced in Section 3. Note that we have chosen to work with a Gamma distribution in order to restrict abundance estimates to the positive real line, although other continuous distributions restricted to the positive real line could have been applied if necessary.

The widely used approach in (9) formulates independent models for each of the sub-processes of the two-part model. However, in some cases, the detection probability and the abundance are related, thus the assumption of independence between the occurrence and abundance sub-processes may be incorrect. The approach in (9) inherently assumes that any abundance has equal weight in the probability of presence and that zero observations have no impact on the abundance model. In this regard, joint modelling techniques could provide a valuable tool to combine information from both sub-processes to fit common effects when the detection probability is suspected to be related to the abundance and vice versa.

Joint modelling has been used to address similar problems, e.g. to characterize the relationship between a longitudinal process and a time-to event process (Hogan and Laird, 1997; Henderson et al., 2000). This approach was also introduced in spatial statistics by Knorr-Held and Best (2001) and further developed by Held et al. (2005) to allow for more than two processes sharing a model component. In the scope of two-part models, SCM may allow us to combine information from related occurrence and abundance sub-processes and so, to fit more robust model components.

In order to introduce SCMs in (9) to fit common model components that share information from both sub-processes, we propose the following modelling:

$$\begin{aligned} \text{logit}(\pi_{st}) &= \alpha^{(1)} + \sum_{i=1}^{I} f_i(x_{is}) + U_{st} \\ log(\mu_{st}) &= \alpha^{(2)} + \sum_{i=1}^{I} \theta_i f_i(x_{is}) + \theta_U U_{st} \end{aligned} \tag{10}$$

where notation is the same as in (9), but the analysed effects, $f_i(x_{is})$ and U_{st}, are now common and have been multiplied in one of the predictors by some unknown parameters, θ_i and θ_U, in order to scale the effects between both sub-processes. Note that it is not necessary for all effects to be shared, there are thus as many models to compare as possible combinations of effects in our linear predictors.

It is also worth noting that we have proposed four different spatio-temporal structures and a case dependent number of shareable effects θ_i as an approach to tackle spatio-temporally sampled semi-continuous datasets. This may imply a high number of comparable model structures (summing approximately $4 * 2^i$, where i is the number of terms in the linear predictor), and thus a large number of relatively complex models to compare. INLA therefore becomes an ideal candidate to deal with these models thanks to its computational efficiency (Cameletti et al., 2013), especially when dealing with some of the proposed spatio-temporal extensions in Section 3.

5. Simulation study

This section is dedicated to test, through simulation, the suitability of joint-modelling techniques to fit shared components in spatial semi-continuous datasets of different nature. To do so, we simulated a hundred Gaussian spatial fields from both types of semi-continuous datasets (those generated from a single abundance process or those generated from two different presence–absence and abundance processes) and fitted both independent and shared models. See supplementary material for further detail and code used in the simulation study.

Model fit was compared based on WAIC scores (Watanabe, 2010), which may perform better than DIC (Gelman et al., 2014), and also on the Conditional Predictive Ordinate (CPO) (Geisser, 1993), through its mean logarithmic score (LCPO) (Gneiting and Raftery, 2007). As INLA does not provide CPO values for the mixture of likelihoods, as in a joint hurdle model, we assessed the predictive capacity of the models using the LCPO scores of each likelihood separately.

Semi-continuous data generated from two different processes

Sometimes semi-continuous datasets are generated from the combination of two clearly different occurrence and abundance processes as it is the case with rain data (for example, it rarely rains in dry climates but if it rains, it can rain as much as in wet climates). To generate this type of semi-continuous datasets, we simulated different fields for the presence–absence process and for the abundance process. We created the presence–absence datasets from one of the fields and extracted the abundances from the other one at those locations where we got a presence.

Semi-continuous data generated from a single process

Other times the semi-continuous dataset is produced from the abundance process alone. For example, if we survey the number of worms in a grass field by randomly collecting small samples of soil in space, it is more likely to observe worms when the overall abundance is high and vice versa. To generate this type of semi-continuous datasets, we first simulated the underlying abundance fields. Then, we created the presence–absence datasets based on proportional-to-abundance detection probabilities. Finally, abundances were extracted from the simulated field at those locations where we got a presence.

Simulation study results

Fig. 2 summarizes the results obtained in the simulation study by comparing the improvement of SCM over independent models in a hundred different semi-continuous datasets of both types. Results show that, SCM performs well when the presence–absence and conditional-to-presence abundance sub-processes are related (positive improvement in both LCPO and WAIC scores, and recovery rate of the Gaussian fields standard deviation). On the contrary, when both processes are unrelated SCM does not fit well as denoted by the negative improvement in both LCPOs and WAIC scores, as well as a worse recovery of the original Gaussian fields standard deviation (see Fig. 2).

6. Modelling hake recruitment in the western mediterranean

In order to exemplify the importance of the models proposed in Sections 3 and 4, we apply our proposal to fisheries, a natural resource of great economic importance. The latest management

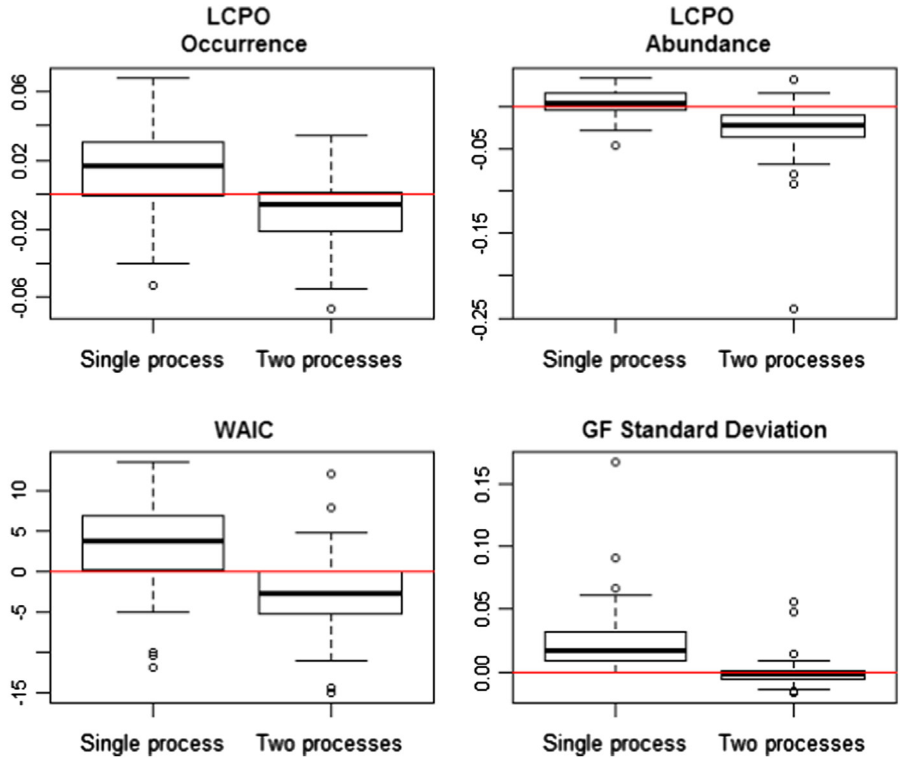


Fig. 2. Improvement of model fit scores (LCPO and WAIC) and recovery rate of the Gaussian fields standard deviation using SCM. Comparison is based on 100 semi-continuous datasets originated from two different processes (occurrence and abundance) and another 100 semi-continuous datasets originated from a single abundance process. Note that positive values represent an improvement on model fit and vice versa.

directives in the context of fisheries (e.g. the Ecosystem Approach to Fisheries Management) require the understanding of marine biological processes on a spatial scale (FAO, 2008). However, quantifying the ecosystemic importance of a marine area is a challenging task due to the inherent constraints of sampling at sea and the fact that fish distributions can vary with time. Information on the distribution of fish populations is generally obtained through research surveys at sea, which provide high quality data. In our case, we will focus on the abundance of European hake (*Merluccius merluccius*) recruitment in the Mediterranean Sea.

The data for this study come from the EU-funded *MEDIterranean Trawl Survey* (MEDITS), and comprise yearly spring hake recruit abundance data from 2000 to 2012 in the Mediterranean GSA06, where the number of observations and their location is different every year. In total the dataset contains information on 1048 hauls that have been georeferenced in the centroid of each fishing operation (see Paradinas et al., 2015 for more details). The calculated Catch Per Unit Effort data (kg per 30 min tow) contain 38% of zero observations while if present, hake recruit abundance shows a right skewed distribution ranging from 0.01 to 26.4 with its mean at 1.5.

A previous study modelled this western Mediterranean hake recruitment using the aforementioned two-part independent modelling approach (Paradinas et al., 2015). The resulting models of that study included bathymetric effects, spatial effects and temporal unstructured effects, as in Eq. (6), for both the occurrence and abundance models, using Bernoulli and Gamma likelihoods respectively. However, as mentioned in the previous section, those models can be improved by including the shared component analysis and also by comparing other different spatio-temporal structures, thereby allowing for a better understanding of recruitment behaviour during the whole of the period analysed.

Table 1
Model fit scores of the most representative fitted models (see full table in supplementary material).

	Model	WAIC	Abundance LCPO	Occurrence LCPO
1	Persistent with iid (independent)	1916.9	0.99	0.23
2	Persistent with trend (independent)	1925.4	0.99	0.24
3	Opportunistic (independent)	1954.9	1.12	0.35
17	Progressive (independent)	**1836.2**	1.02	0.26
18	Progressive with shared bathymetric effect	**1839.9**	0.96	0.23

6.1. Modelling hake recruitment

Following the two-part modelling structure in Paradinas et al. (2015), we have analysed the recruitment dataset comparing all the resulting models obtained by implementing the four spatio-temporal structures and the shared component analysis described in Sections 3 and 4 respectively. We used a fine-scale mesh with almost 2500 nodes, of which more than 1700 nodes were distributed along the 57 915 square kilometres of the study area to allow capturing a good resolution of the spatial autocorrelation. The rest of mesh nodes created a wide outer area to avoid having a boundary effect in our estimates. With respect to the bathymetric and temporal trend effects, we fitted them by means of smooth second order random walk (RW2) latent models (Rue and Held, 2005). In the case of the fourth temporal structure in Eq. (8), we only considered first order autoregressive (AR1) models due to the rather short time series of thirteen years available.

Our lack of prior information about the model parameters led us to adopt an objective Bayesian approach (Bayarri and Berger, 2004) and to assign vague prior distributions as implemented by default, in `INLA`. Only the prior of the bathymetric RW2 precision was changed to a $LogGa(2, 0.00005)$ to restrict its smoothing capacity and avoid overfit. This prior was selected visually to allow a sensible process–covariate relationship after scaling the RW2 model to obtain a generalized variance equal to 1 (Sørbye and Rue, 2014). A sensitivity analysis was performed to verify that the posterior distributions concentrated well within the support of all the priors.

6.2. Results and discussion

All the resulting model structures were fitted and compared on the basis of WAIC scores (Watanabe, 2010) and the Conditional Predictive Ordinate (CPO) (Geisser, 1993), through its mean logarithmic score (LCPO) (Gneiting and Raftery, 2007). As mentioned in the simulation study, `INLA` provides CPO scores for each likelihood, thus model selection was performed based on the LCPO scores obtained for each likelihood. In both model selection scores, the smaller the score the better the model. As highlighted in the WAIC scores of Table 1, two structures performed reasonably better than the rest (see supplementary materials for the full model selection table). Both models include a first order autoregressive temporal term, with independent bathymetric effects in the occurrence and the abundance sub-processes in model 17, while model 18 fits a shared bathymetric effect to both.

We finally selected model 18 over model 17 for a number of reasons. Firstly, model 18 fits a biologically more natural bathymetric effect (see Fig. 3), where the abundance of hake recruits decreases gradually after the optimum 150–200 metre strata. Secondly, model 17 clearly overfits the bathymetric effect of the abundance sub-process due to the lack of zero observations in it (see highlighted box in Fig. 3). Interestingly, even if WAIC scores slightly prefer model 17 over model 18, the predictive LCPO scores clearly benefit model 18. Lastly, model 17 is unable to predict hake recruit abundance in the whole sampled depth range without extrapolation.

The selection of an autoregressive temporal term in the model suggests the presence of a certain degree of temporal continuity in the spatial distribution of hake recruits in the study area. Moreover, such a temporal correlation term allows for a better informed interpolation and thus a better representation of the distribution of hake recruitment in the western Mediterranean. Indeed, as the posterior mean estimate maps in Fig. 4 show, although the recruitment of hake is mainly concentrated in the

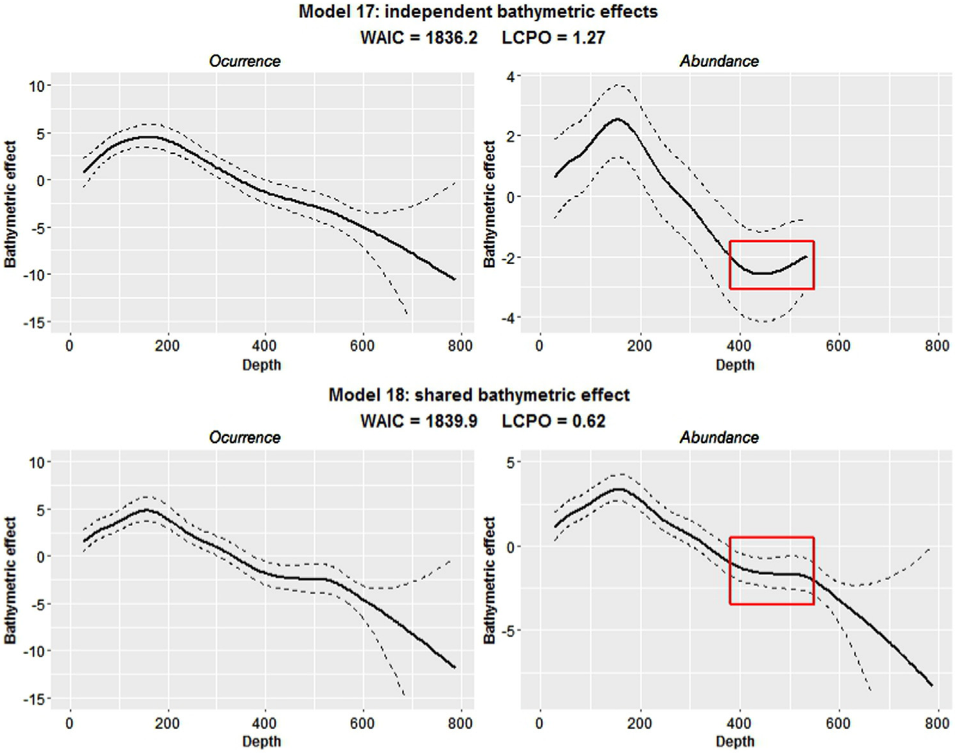
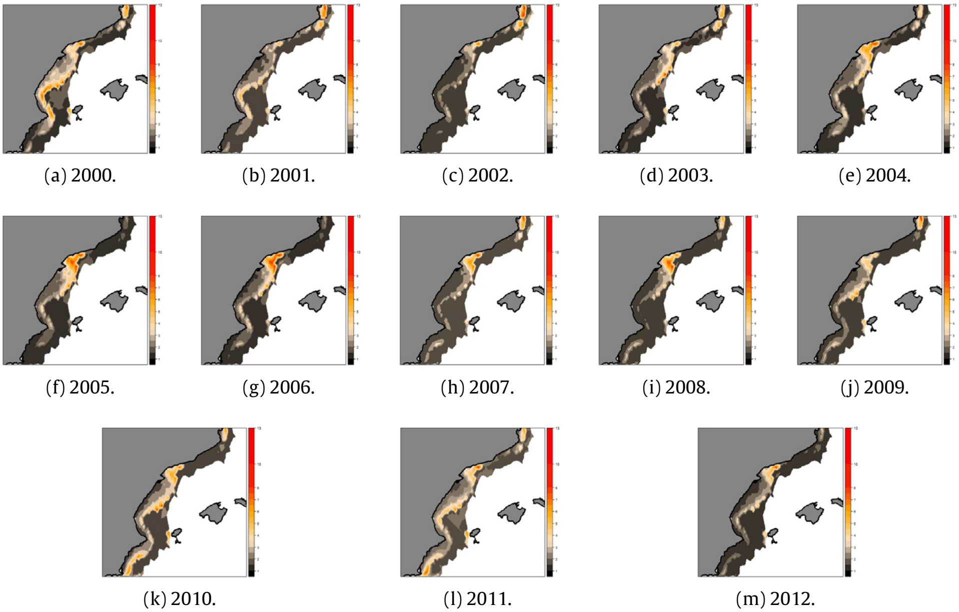


Fig. 3. Fitted smooth bathymetric effects in models 17 and 18 (Table 1). The solid line represents the mean of the effect and the dashed lines its 95% credibility interval. The marked box highlights the importance of SCM to fit a biologically more natural bathymetric effect for hake recruit abundance.

central and northern parts of the study area (as already discussed in Paradinas et al. (2015)) there have been smooth changes in the relative abundance and the spatial location of hake recruitment hot-spots from year to year (see Fig. 5), which may provide important insight for management purposes.

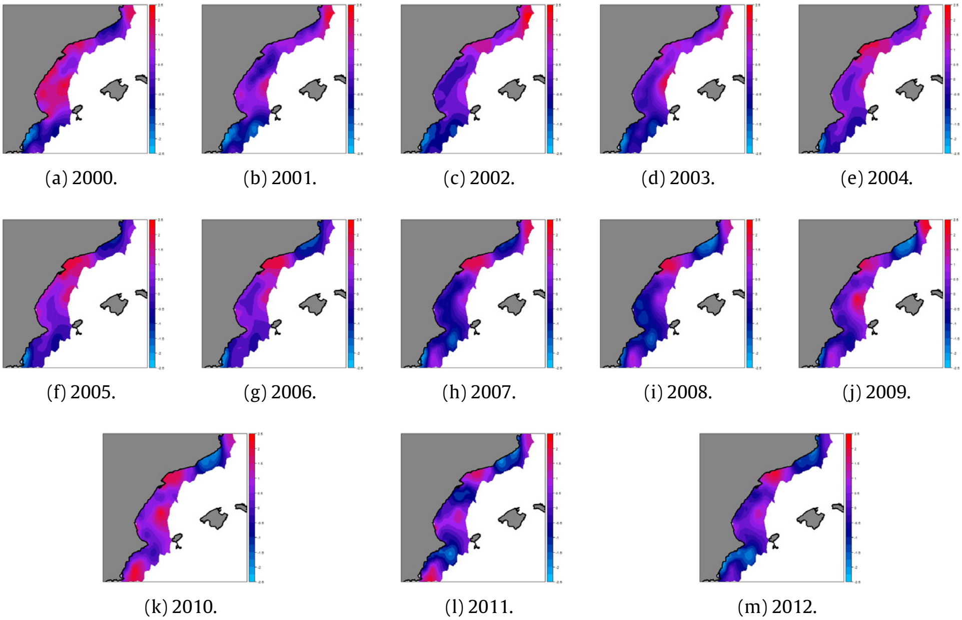
Concerning the spatial or spatio-temporal fields, shared components did not improve fitted models, as also occurred in Quiroz et al. (2015). This can be seen in Fig. 6, where the fitted spatial covariance functions of the occurrence and the abundance sub-processes are very different. The same occurred in the case of the autoregressive term, where the independent occurrence (posterior median = 0.98; 95% CI = [0.95,0.99]) and abundance estimates (posterior median = 0.87; 95% CI = [0.67,0.95]) also differed widely, and thus the shared component (posterior median = 0.95; 95% CI = [0.87,0.98]) fitted neither of the two, especially the abundance sub-process. This significant difference between the spatio-temporal patterns of both sub-processes may, at first sight, suggest that the semi-continuous data of hake recruitment is generated through two completely different processes; the probability of observing hake recruits and, if present, their abundance. However, the nature of the process under study induces to believe that this apparent independence is a consequence of the high sampling effort of the survey relative to the abundance of hake recruits, rather than being two different processes. The MEDITS fishery survey trawls relatively big areas (over one mile), therefore the probability of observing at least one individual of an abundant fish species, such as hake, is quite high at environmentally not-too-challenging areas. Similarly, if effort was diminished, the detection probability would decrease proportionally and thus record a lot more zeros in our dataset.


(a) 2000. (b) 2001. (c) 2002. (d) 2003. (e) 2004.

(f) 2005. (g) 2006. (h) 2007. (i) 2008. (j) 2009.

(k) 2010. (l) 2011. (m) 2012.

Fig. 4. Yearly hake recruitment estimated posterior mean abundance maps.


(a) 2000. (b) 2001. (c) 2002. (d) 2003. (e) 2004.

(f) 2005. (g) 2006. (h) 2007. (i) 2008. (j) 2009.

(k) 2010. (l) 2011. (m) 2012.

Fig. 5. Yearly estimated mean spatial effects.

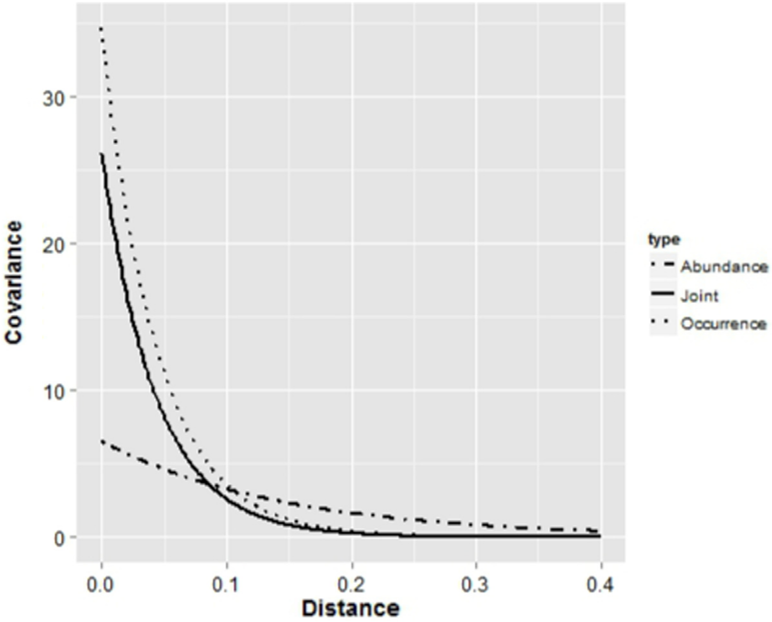


Fig. 6. Fitted Matérn covariance functions in the unit scale. The solid line represents the joint covariance function, the dotted line represents the covariance function for independent occurrence model and the dot-dashed line that for the independent abundance model.

7. Concluding remarks

In this paper we have presented a model structure comparison for spatio-temporally sampled datasets as an approach to infer further information on the distributional behaviour of a process over time. Furthermore, we have proposed the use of SCM to deal with non-independent semi-continuous data. By using the proposed approaches, we have significantly improved the information that was available for the spatial management of hake recruitment in the western Mediterranean (until now, the results were in line with the model in Paradinas et al. (2015) that corresponds to Model 1 in Table 1). In this regard, the INLA package for R (Rue et al., 2009) not only provides a computationally efficient tool to fit complex models but also a wide range of modelling possibilities in a reasonably user-friendly environment.

Acknowledging the spatio-temporal behaviour of a natural process is crucial for management purposes and decision making (Smit et al., 2013). The spatio temporal structures proposed in Section 3 characterize four basic, yet informative, spatio-temporal behaviours. Basically, these structures allow us to distinguish the extent to which the spatial distribution of the process under study varies along the sampled time intervals. For instance, if the spatial distribution of the process varies unrelatedly from time to time, different spatial realizations for each time will be necessary to fit our data. On the contrary, if the spatial structure is reasonably persistent, a unique spatial realization may be sufficient, to which either a zero mean random effect or a temporal trend (Paradinas et al., 2015) could be fitted to absorb the different mean intensities of the process over time. Lastly, if the spatial realization varies over time but in a structured manner, as in our example, a temporal correlation structure will suit best our data.

Regarding the use of SCM in semi-continuous datasets, this study has shown that fitted environment-process effects can be improved by combining information on occurrence and conditional-to-presence abundance. However, its application over fisheries spatial random fields may require further research as shown by this study and that by Quiroz et al. (2015), where both sub-processes seem to have non-proportional spatial variabilities. This apparent independence may be derived from the high sampling effort of the fishery surveys relative to the abundance of the process under study. Generally speaking, and specially at low sample sizes and high sampling effort, the variability of a presence–absence sub-process as a function of distance may not be comparable to that of the abundance sub-process. This might be the reason why, as opposed to the simulation study on the single originating process, the shared geostatistical term did not improve the hake recruitment two-part independent modelling, as also occurred in Quiroz et al. (2015).

Lastly, we would like to mention the possibility of extending the spatio-temporal structure comparison for modelling the distribution of species proposed in this paper to other spatio-temporally sampled processes. Similarly, higher order temporal structures could be proposed to infer more informative behaviours of the process under study when the temporal resolution of the data allows.

Acknowledgements

Iosu Paradinas would like to thank the Fundación Biodiversidad (Spanish Ministry of Agriculture and the Environment) for the funding provided. David Conesa and Antonio López-Quílez would like to thank the Ministerio de Economía y Competitividad (the Spanish Ministry of Economy and Competitiveness) for the funding provided via research grants MTM2013-42323-P and MTM2016-77501-P (jointly financed with the European Regional Development Fund). The authors would also like to thank the INLA-project group, especially Elias T. Krainski, for their support with technical aspects of INLA and its SPDE module. They also express their gratitude to all the people who worked in the MEDITS surveys. Data collection has been co-funded by the EU through the European Maritime and Fisheries Fund (EMFF) within the National Program of collection, management and use of data in the fisheries sector and support for scientific advice regarding the Common Fisheries Policy.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spasta.2017.08.001.

References

Aizpurua, O., Paquet, J., Brotons, L., Titeux, N., 2015. Optimising long-term monitoring projects for species distribution modelling: how atlas data may help. Ecography 38 (1), 29–40.

Balderama, E., Gardner, B., Reich, B.J., 2016. A spatial–temporal double-hurdle model for extremely over-dispersed avian count data. Spat. Stat. 18, 263–275.

Banerjee, S., Gelfand, A.E., Carlin, B.P., 2003. Hierarchical Modeling and Analysis for Spatial Data. Crc Press.

Bayarri, M.J., Berger, J.O., 2004. The interplay of Bayesian and frequentist analysis. Statist. Sci. 58–80.

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.

Cosandey-Godin, A., Krainski, E.T., Worm, B., Flemming, J.M., 2014. Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. Can. J. Fish. Aquat. Sci. 72 (2), 186–197.

Cressie, N., 2015. Statistics for Spatial Data. John Wiley & Sons.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. John Wiley & Sons.

Dawid, A.P., 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. J. Roy. Statist. Soc. Ser. A (Gen.) 278–292.

Diggle, P., Ribeiro, P.J., 2007. Model-Based Geostatistics. Springer Science & Business Media.

Dinmore, T.A., Duplisea, D.E., Rackham, B.D., Maxwell, D.L., Jennings, S., 2003. Impact of a large-scale area closure on patterns of fishing disturbance and the consequences for benthic communities. ICES J. Mar. Sci. J. Cons. 60 (2), 371–380.

Fahrmeir, L., Lang, S., 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. Appl. Stat. 201–220.

FAO, 2008. Fisheries Management. 2. The Ecosystem Approach to Fisheries. FAO.

Fortin, M.J., Dale, M.R.T., 2009. Spatial autocorrelation in ecological studies: a legacy of solutions and myths. Geogr. Anal. 41 (4), 392–397.

Geisser, S., 1993. Predictive Inference, Vol. 55. CRC press.

Gelfand, A.E., 2012. Hierarchical modeling for spatial data problems. Spat. Stat. 1, 30–39.

Gelfand, A.E., Ravishanker, N., Ecker, M.D., 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., pp. 381–394.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24 (6), 997–1016.

Gitzen, R.A., 2012. Design and Analysis of Long-Term Ecological Monitoring Studies. Cambridge University Press.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378.

González-Warleta, M., Lladosa, S., Castro-Hermida, J.A., Martínez-Ibeas, A.M., 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.

Haining, R., Law, J., Maheswaran, R., Pearson, T., Brindley, P., 2007. Bayesian modelling of environmental risk: example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels. Stoch. Environ. Res. Risk Assess. 21 (5), 501–509.

Heegaard, E., Boddy, L., Diez, J.M., Halvorsen, R., Kauserud, H., Kuyper, T.W., Bässler, C., Buntgen, U., Gange, A.C., Krisai-Greilhuber, I., et al., 2017. Fine-scale spatiotemporal dynamics of fungal fruiting: prevalence, amplitude, range and continuity. Ecography 40 (8), 947–959.

Held, L., Natáio, I., Fenton, S.E., Rue, H., Becker, N., 2005. Towards joint a disease mapping. Stat. Methods Med. Res. 14 (1), 61–82.

Henderson, R., Diggle, P., Dobson, A., 2000. Joint modelling of longitudinal measurements and event time data. Biostatistics 1 (4), 465–480.

Hjelle, Ø., Dæhlen, M., 2006. Triangulations and Applications. Springer Science & Business Media.

Hogan, J.W., Laird, N.M., 1997. Mixture models for the joint distribution of repeated measures and event times. Stat. Med. 16 (3), 239–257.

Knorr-Held, L., Best, N.G., 2001. A shared component model for detecting joint and selective clustering of two diseases. J. R. Stat. Soc. Ser. A Stat. Soc. 164 (1), 73–85.

Krige, D.G., 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. J. Chem. Metal. Min. Soc. S. Afr..

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25 (5), 601–615.

Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63 (19).

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.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. 8 (11), 1235–1246.

Martins, T.G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: new features. Comput. Statist. Data Anal. 67, 68–83.

Muñoz, F., Pennino, M.G., Conesa, D., López-Quílez, A., Bellido, J.M., 2013. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. Stoch. Environ. Res. Risk Assess. 27, 1171–1180.

Neelon, B., Ghosh, P., Loebs, P.F., 2013. A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. J. Roy. Statist. Soc. Ser. A 176 (2), 389–413.

Paradinas, I., Conesa, D.V., Pennino, M.G., Muñoz, F., Fernández, A., López-Quílez, A., Bellido, J.M., 2015. A Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. Mar. Ecol. Prog. Ser. 528, 245–255.

Paradinas, I., Marín, M., Pennino, M.G., López-Quílez, A., Conesa, D., Barreda, D., Gonzalez, M., Bellido, J.M., 2016. Identifying the best fishing-suitable areas under the new European discard ban. ICES J. Mar. Sci. J. Cons. 73 (10), 2479.

Paradinas, I., Pennino, M.G., Marín, M., López-Quílez, A., Bellido, J.M., Conesa, D., 2017. Modelling spatially sampled proportion processes. REVSTAT (in press).

Pennino, M.G., Muñoz, F., Conesa, D., López-Quílez, A., Bellido, J.M., 2013. Modeling sensitive elasmobranch habitats. J. Sea Res. 83, 209–218.

Pennino, M.G., Muñoz, F., Conesa, D., López-Quílez, A., Bellido, J.M., 2014. Bayesian spatio-temporal discard model in a demersal trawl fishery. J. Sea Res. 90, 44–53.

Quiroz, Z.C., Prates, M.O., Rue, H., 2015. A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. Biometrics 71 (1), 208–217.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. CRC Press.

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.

Smit, I.P.J., Smit, C.F., Govender, N., Linde, M., MacFadyen, S., 2013. Rainfall, geology and landscape position generate large-scale spatiotemporal fire pattern heterogeneity in an African savanna. Ecography 36 (4), 447–459.

Sørbye, S.H., Rue, H., 2014. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spat. Stat. 8, 39–51.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (4), 583–639.

Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 66 (2), 275–296.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 234–240.

Tsiatis, A.A., Davidian, M., 2004. Joint modeling of longitudinal and time-to-event data: an overview. Statist. Sinica 14 (3), 809–834.

Waller, L.A., 2014. Putting spatial statistics (back) on the map. Spat. Stat. 9, 4–19.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.