VNIVERSITAT DE VALÈNCIA

Departamento de Estadística
e Investigación Operativa

Programa de doctorado en Estadística i Optimització

PhD Thesis

Species distribution modelling in fisheries science

Iosu Paradinas Aranjuelo

Supervisors:
David V. Conesa Guillén
Antonio López Quílez
José María Bellido Millán

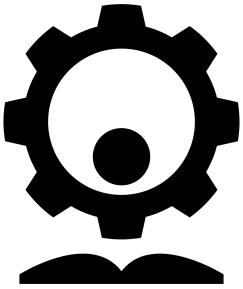
October 2016

Pixkat-asko ta amunt.
Eskerrikasko.

Principio de Bayes social:

El principio ético liberal, es decir, el esfuerzo individual como principio de igualdad, no se sostiene desde el momento en el que las distribuciones a priori para cada individuo depende de su contexto social. Hace años que sabemos de su influencia sobre las distribuciones a posteriori. Le llaman sensibilidad.

Previas iguales para tod@s...



... y dejemos las distribuciones a posteriori en manos de cada un@.

Acknowledgments

Agradezco de todo corazón a toda le gente que se preocupa y lucha por los derechos de la gente. Si, vosotros, esos a los que llaman radicales, esos sin los que negros y mujeres estarían aún sin votar, esos que consiguieron la gran mayoría de derechos fundamentales. Si os llaman antisitema, no os preocupéis porque sin vosotros nunca hubiera existido el sistema de bienestar. Y una sola sugerencia para esos que os llaman antisitema; si ser antisitema es intentar mejorar la sociedad.

DESPERTAD

Porque no he podido estar ahí, en la calle, apoyando causas de moral universal, muchísimas gracias a todos vosotros. Mi alma y mi voz están con vosotros. Esperarme que estaré ahí cuando menos lo esperéis.

Extended summary

Latest fisheries directives propose adopting an ecosystem approach to manage fisheries (FAO, 2003). Such an approach aims to protect important ecosystems based on the principle that healthy ecosystems produce more and thus enhance sustainability. Unfortunately, quantifying the importance of an ecosystem is a difficult task to do due the immense number of interactions involved in marine systems.

This PhD dissertation relies on the fact that good fisheries distribution maps could play a very important role as they allow a visual and intuitive assessment of different marine areas. Unfortunately, the limited amount of data available and the inherent difficulties of modelling fishery data has resulted in relatively low quality maps in the near past (see (Heesen et al., 2015) and http://www.ices.dk/marine-data/maps/Pages/ICES-FishMap.aspx). As a result, the spatial fisheries management framework requires competent statistical approaches to quantify the importance of different marine areas with an appropriate measure of uncertainty associated to the estimates.

The aim of this PhD is to provide competent spatial and spatio-temporal modelling approaches that allow us characterise different fishery processes that are relevant for their sustainable management. More specifically, the objectives of this PhD are:

- To propose a spatial modelling framework that properly assess the fishing-suitability of a fishing ground in terms of fishery discards.

- To propose effective modelling frameworks to map the spatial or spatio-temporal distribution of economically important fisheries. In this regard, different modelling approaches are required to tackle different types of fishery data:
  - On-board or fishery dependent data is sampled preferentially, thus corrections are needed when modelling target species. An objective of this PhD has been, therefore, to test the use of Log-Gaussian Cox Process models to correct the model components of preferentially sampled fish abundance datasets.
  - Survey or fishery independent data provide information to assess changes in the macro-scale of fisheries distribution over the years. Another objective of this PhD has been to propose useful modelling structures to infer the spatio-temporal dynamism of different fishery processes, e.g. spawning and nursery grounds.
- To propose an effective framework to fit appropriate model components in two-part or Hurdle models.
- To assess the performance of point-referenced regression models in fishery transect data, including Euclidean distance-based geostatistical models.

Our baseline statistical approach has been model based geostatistics. In particular we have developed structures upon it to adequate for different fishery processes and fishery data. Bayesian methods allow direct and intuitive quantification of the uncertainty through explicit probabilistic inference. Furthermore the Bayesian hierarchical model formulation allows defining complex statistical models, such as geostatistical models, in a rather easy and intuitive way. However, the computational cost of Bayesian methods can be a problem, specially in big and complex datasets. To tackle the computational burden of the proposed models, we have used the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) method an the Stochastic Partial Differential Equations (Lindgren et al., 2011) (SPDE) approach.

In Chapter (1) we present the main problem in current fisheries management that motivated this PhD, the quantitative spatial assessment of our fisheries. Then we briefly present the main types of spatial data followed by a brief summary of the main species distribution modelling approaches, from linear regression to geostatistical models. Next, we introduce the benefits of Bayesian hierarchical models in spatial statistics and the different types of Bayesian computing approaches. In this chapter, we specially describe the INLA (Rue et al., 2009) method and the SPDE (Lindgren et al., 2011) approach to deal with complex geostatistical structures at assumable computational costs. Finally, we end up summarising the main model selection scores used along this PhD dissertation.

The second Chapter (2) is dedicated to fishery discards, which spatial distribution has most of the times been assessed using biomass based units, e.g. discards per unit effort (DPUE) (Feekings et al., 2012, 2013; Viana et al., 2013a; Cosandey-Godin et al., 2014; Pennino et al., 2014). The fishing suitability of a given area, however, should contrast the actual biomass benefit against biomass loss of a fishing operation. To do so, we propose using spatial beta regression to model discard proportions (discarded biomass divided by the total catch of a fishing operation). Along the chapter, we review the different approaches used in the past to model proportions and end up proposing a Bayesian hierarchical spatio-temporal beta regression model to identify fishing suitable areas.

The third Chapter (3) approaches the modelling of target species using fishery dependent data. The main property of fishery dependent data is that fishermen choose fishing locations based on their knowledge (best locations to catch more target species biomass) and therefore our sample is subject to the preferential sampling problem (Diggle et al., 2010). As a consequence, the sampling process and the process being modelled are not stochastically independent, which violates a basic statistical modelling assumption. To correct for this bias, we make use of joint-modelling techniques between the marks (caught abundances) and the point pattern of the fishery (selected fishing locations). This way we are able to combine information derived from the spa-

tial distribution of the samples (point pattern), as a proxy to the fishermen's knowledge about the underlying fish abundance distribution, and information coming from the fished abundances (marks). As a consequence, we manage to better inform our models, overcoming the preferential sampling problem, thus obtaining a better approximation of the underlying spatial field.

Chapter (4) deals with fishery survey data, which is the most widely used data for fisheries management. Fishery survey data, or fishery independent data, usually cover very wide areas and provide a macroscopic view of the fishery over the years. As most species distribution datasets, fishery data is also prone to zero observations at unfavourable conditions, resulting in spatio-temporal semi-continuous datasets. This chapter is devoted, on the one hand to improve the usual two-part modelling framework to deal with the semi-continuous nature of the data and on the other hand to infer the spatio-temporal behaviour of the fishery process under study. To do so, we compare different spatio-temporal structures and end up using joint-modelling techniques to fit better informed environmental effects in Hurdle models.

In Chapter (5) we investigate on the implications of point-referencing fishery data, which in reality represent a transect between the starting and ending points of the fishing operation (except purse seiners that fish almost static). This could be specially problematic when applying geostatistics, based in Euclidean distances, in small-scale study areas. In this chapter, we also propose an algorithm, that recognize the transect nature of the data, to approximate the underlying spatial field when enough data and enough cross-overs between fishing operations are present.

Finally, Chapter (6) presents some concluding remarks and future lines of research.

Consequently the main contributions of this study to the knowledge in fisheries distribution modelling are:

- The spatial analysis of discard proportions instead of total discard biomass units is a good alternative to assess the fishing-suitability of an area in terms of discards.

- The use of LGCP models to correct the analysis of preferentially sampled data improves significantly the predictive capacity of the abundance models. This allows us use on-board fishery data to model the spatial distribution of targeted fisheries. The use of within-sample and similar model selection scores, e.g. WAIC, DIC, LCPO, etc., can be misleading as they fail to assess the out-of-sample predictive capacity.

- The spatio-temporal distributional behaviour of fisheries can be effectively inferred by comparing a set of spatio-temporal structures.

- Joint modelling techniques can improve fitted effects in two-part or Hurdle models. Visual validation of the models is important in the model selection process.

- The point-referenced representation of fishery transects allows fairly good regression estimates fitting both; process-covariate relationships; and geostatistical fields even in small-scale study areas with respect to the size of the fishery transect.

- The remarkable flexibility of R-INLA in extending common hierarchical models allows fitting complex structures that better resemble natural sciences.

- The spatio-temporal representation of different fish species can effectively improve our understanding of fish ecology. Therefore, extending the hake and cod case studies of this thesis to other species could be very valuable to EAFM policy makers.

Contents

List of Figures

List of Tables

Chapter 1

Spatial statistics for fisheries management

From ancient times, fishing has been a major source of food for humanity and a provider of employment and economic benefits to those engaged in this activity. However, with increased knowledge and the dynamic development of fisheries, it was realized that living aquatic resources, although renewable, are not infinite and need to be properly managed, if their contribution to the nutritional, economic and social well-being of the growing world's population was to be sustained (FAO, 1999).

Nowadays, many of the world's fish populations are overexploited and the ecosystems that sustain them are degraded (FAO, 2002). The unintended consequences of fishing, including habitat destruction, incidental mortality of non-target species, evolutionary shifts in population demographics, and changes in the function and structure of ecosystems are increasingly recognized. Fishery management as it is today has proven to be ineffective in many places (Walters and Maguire, 1996; Pauly et al., 2002). It generally focuses on maximizing the catch of single target species and ignores habitat, predators and prey of target species, as well as other ecosystem components. Today, we

know that the social and economic costs of focusing on single species can be substantial (Pikitch et al., 2004).

In this regard, a variety of advisory panels have recommended the Ecosystem Approach to Fishery Management (EAFM) framework. The founding principles and conceptual goals for EAFM emerge from a decades-long process of elaboration of the foundations for sustainable development, aiming at both human and ecosystem well-being (Garcia, 2003). In summary, EAFM targets the conservation of full ecosystems rather than single species stocks based on the principle that healthy ecosystems produce more.

A key instrument for the effective conservation of ecosystems is the implementation of marine protected areas (MPA) (Hilborn et al., 2004; Claudet et al., 2008). We know that MPAs boost fish productivity in its surrounding areas and that different fishing communities have increased their harvest around the world this way (Mangi et al., 2011; Williams et al., 2009; Claudet et al., 2008). Unfortunately, the creation of marine protected areas is still a very controversial issue among most fishermen and politicians due to its high initial economic cost and the rather unpredictable time required to experience its benefits. It is then understandable the amount of pressure held on the scientific community to quantitatively characterize appropriate, highly productive areas as proposed by FAO (2003).

Assessing the potential productivity of a marine area would be easy if we were able to directly and continuously observe what is happening. Unfortunately, this is infeasible in marine systems, thus we have to rely on data collected through different sea sampling schemes: commercial landings, onboard observers, fishery surveys, ... It is most often the case that the rather scarce amount of data available does not provide enough information to assess the importance of a given area by plain descriptive analysis. Therefore, statistical methods are essential to create an effective marine spatial planning under the EAFM framework.

In this regard, statistical models can allow us to understand the relationships of the process under study with the environment and to quantify the uncertainty related to our estimates. More precisely, marine spatial planning,

as its name suggests, needs to quantify such relationships in a spatial framework, which may require of spatial models to characterise the importance of different marine areas. In this chapter, we discuss the different sorts of spatial models. Lets start by describing the types of spatial data that are available in the fishery world.

1.1 Spatial fishery data

In fisheries we have two broad types of data: fishery dependent data and fishery independent data. The difference between the two types of data falls on the sampling actor: whether we sample the fish caught by the commercial fleet or we sample what the fish caught during scientific surveys. Each of the sampling schemes has its pros and cons.

1.1.1 Fishery dependent data

Fishery dependent data refers to the data that is collected sampling the commercial fleet. Typically, this data is collected by an on-board observer that performs a stratified random sampling (European Comission, 2009) to collect biological data, species composition, discards, etc.

According to European Comission (2009), the number of samples collected each month in each fishery (fishing gear that targets an specific fish stock) is proportional to the total fishing effort (number of fishing days) of that fishery in that particular area. Therefore, in theory the sampling scheme should translate into a good representation of each fishery both in time and space. However, due to logistic and economic reasons the spatial distribution of the data is generally too scarce and patchy to provide a good macro-scale spatial representation of fish distribution.

On the meso-scale however, the spatial resolution of the samples is much better. In this scale we can formulate different hypotheses of the fishery using these data. Nevertheless, for modelling purposes, we should bear in mind that fishing locations are not randomly distributed in space, but located in

those places where fishers expect a better catch. This is often referred as a preferential sampling problem (Diggle et al., 2010). Therefore, depending on the specific process that we are willing to analyse, we need to adjust for the preferential sampling problem or not. For instance, modelling the distribution of cod in a cod fishery is subject to the preferential sampling problem but modelling the quantity of discards (part of the fishing that is thrown back to sea due to its low economic value) of different fisheries is not, because discards are not targeted. We elaborate more about this along next chapter 2.

1.1.2 Fishery independent data

Fishery independent data refers to the data that is collected through scientific surveys. Scientific surveys are not influenced by harvesting activities and provide critical information on the status of fish stocks (Morgan and Burgess, 2005). The spatial sampling design is random and the quality of the data is very high, i.e. number of fish per size, per sex and species measured.

In Europe, there are a good number of Fishery scientific surveys (STECF, 2007) that cover most of the European fishing grounds and stocks. Data collected through scientific surveys are used for stock assessment purposes, yet due to the immense economic cost of these surveys, the temporal resolution is typically very low. For instance, most surveys are repeated only once per year (e.g. Mediterranean Trawl Survey), maximum twice (e.g. International Bottom Trawl Survey). Therefore, we must be aware of its temporal resolution when analysing these data and be very careful with its conclusions.

So, in summary, the properties of fishery independent data are remarkable for analysis, including spatial analysis in a macro-scale. However, they only provide a temporal snapshot of the spatial distribution of fish over the year. In chapter 3 we further elaborate on the modelling opportunities that fishery survey data provide.

1.1.3 Trawl data characteristics

Trawling is the most destructive and most used, fishing gear (Jones, 1992). Therefore, a lot of on-board observer effort is put on commercial trawlers. Many scientific surveys also use trawling gear to sample the ocean because it provides a very representative sample of what there is in the bottom of the sea.

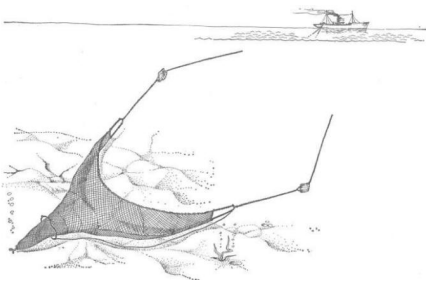


Figure 1.1. An ilustration of a trawl fishing operation

A fishing trawl operation (Figure 1.1) constitutes a three-dimensional transect defined by: the length of the haul, the width of the gear and its height. The essential difference with conventional transects in other disciplines is that, in the fishery world, it is not known what has been observed where, in other words, it is only known the total catch of the transect. As a consequence, it is customary to represent the trawl operation as a point in space, generally the centroid of the fishing operation.

Another important issue when dealing with fishery data is the fact that not all fishing operations are the same (different duration, size of the fishing gear, etc.), thus we have to deal with different so-called efforts. In order to adjust for different efforts in fisheries, we usually work with catch per unit effort (CPUE) units. As we have already mentioned, effort can be measured in several ways (volumes, areas, time, etc.), the most usual one being time. Most trawlers target species associated to the sea floor (so-called demersal species), so per-volume effort should not provide more information than per-

area effort. However, the reasons why per-time effort is the most popular are that; 1) most of the times we do not know the width of the gear, e.g. when sampling commercial vessels, and 2) scientific surveys usually use the same gear in all trawling operations, thus time and area are proportional.

1.2 Types of spatial data

So far, we have summarised the different types of fishery data and the general characteristics of a trawling operation. Now we give an overall overview of the different types of spatial data so that we can then select the spatial treatment that we want to give to our spatial fishery data.

When performing spatial statistics, our observations $\boldsymbol{Y}(\boldsymbol{s})$ are defined over a spatial region $\boldsymbol{s} \in \mathcal{D}$ and specific locations $\boldsymbol{s} = \{s_1, \ldots, s_n\}$. Depending on the nature of the data and the spatial aggregation that we give to it, we can differentiate three types of spatial data: areal, point-referenced and point-pattern data.

1.2.1 Areal data

Areal data, also known as *lattice* data, represent an aggregation of observations over a predefined areal unit. The outcome of such aggregation $\boldsymbol{Y}(\boldsymbol{s})$ is defined over some discrete region $\mathcal{D}$ with fixed number of locations $\boldsymbol{s}$. Therefore, $\mathcal{D}$ is divided into a finite collection of areal units with well defined boundaries.

The hypothesis when modelling areal data is whether adjacent regions share information in the sense that close areas have more in common than distant areas. Modelling areal-data involves borrowing information from adjacent regions. The most usual model structure in these sort of cases is the conditional autoregressive model (Besag et al., 1991), best known as CAR or BYM model after the authors initials. These models induce autoregressive spatial autocorrelation through an adjacency structure of the areal units.

In fisheries for example, the North Sea has often been discretized on a

lattice domain $\mathcal{D}$ (see Figure 1.2). Observations are then aggregated $\boldsymbol{Y}(\boldsymbol{s})$ at each grid cell.

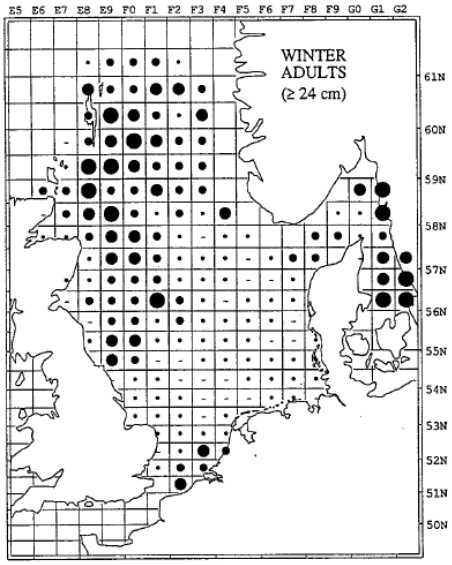


Figure 1.2. Winter adult herring distribution in the North Sea aggregated by ICES statistical rectangles.

1.2.2 Point-referenced data

Point-referenced data, as its name suggests, is constituted by a random variable $\boldsymbol{Y}(\boldsymbol{s})$ collected in a fixed set locations $\boldsymbol{s}$ over a continuous spatial field Λ. Space is typically treated as two dimensional, defined by its longitude and latitude, but it could also include altitude or depth to make it three dimensional. In fisheries, and specially in trawl fisheries, the use two dimensions is sensible because we only fish on the sea floor.

When modelling point-referenced, also known as geostatistical data, we expect our data to be spatially correlated given our explanatory variables. Our main purpose when modelling point-referenced data is to infer the spatial

structure of our data to enhance prediction, using kriging techniques (Cressie, 1990) at unsampled locations.

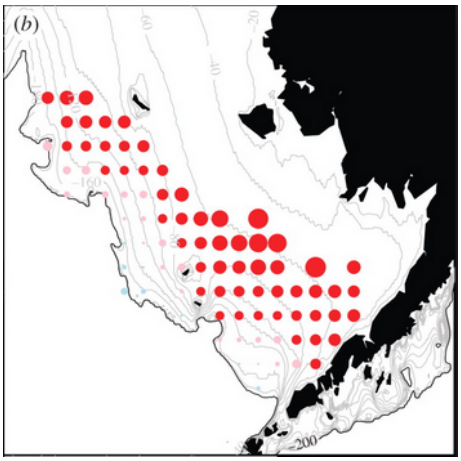


Figure 1.3. Distribution of large fish in the Bering sea.

For example Ciannelli et al. (2012) modelled the distribution of large fish in the Bering Sea (see Figure 1.3). Measurements $\boldsymbol{Y}(\boldsymbol{s})$ are taken in discrete locations $\boldsymbol{s}$ of the continuous domain $\mathcal{D}$. Obviously, the spatial dimension could be extended to the spatio-temporal domain by adding the correlation of fish abundance between time events (i.e. every hour, day, etc).

1.2.3 Point-pattern data

A point-pattern is a process where we observe the exact location at which the subject of interest is, for example the distribution of vessels in the sea. In this case, our interest is not to measure how "many" vessels there are in a location, but to study the spatial arrangement of the vessels in space as a proxy of fishing grounds for example.

In point-patterns, the spatial field Λ itself is random. This random spatial field is what generates the point pattern, whose observations $\boldsymbol{Y}(\boldsymbol{s})$ are equal

to one for all $\boldsymbol{s}$, or maybe also provide some numerical information resulting in a *marked* point-pattern. In point-patterns, the response $\boldsymbol{Y}$ is fixed (presence) and the set of locations $\boldsymbol{s}$ is randomly generated from the spatial field Λ. The underlying question in point-pattern data is often related to the event of clustering, where we usually want to determine whether the spatial distribution of the observed point pattern is homogeneous over space or is a clustered process, and if clustered, what is it that drives the clustering.

Figure 1.4. Aerial view of commercial herring fishing taking place in Sitka Sound, Alaska. Photo by Scott Dickerson.

An example of a point pattern could be the distribution fishing operations in the ocean (see Figure 1.4). Λ represents the distribution of fish in the ocean and $\boldsymbol{s}$ the locations where vessels have been fishing at a given moment. Sometimes, the purely spatial domain should be extended to the spatio-temporal domain. If in the following day we observe a different distribution of vessels in the same area $\boldsymbol{s_{t+1}}$, has the spatial field Λ_{t+1} changed? or is it just a different realization of the same spatial field Λ?

1.3 Species distribution modelling

We have already seen the main problem in contemporary fisheries management, the different types of data available in fisheries and the three different

branches of spatial statistics to deal with different types of spatial data. Now, we find it necessary to slightly relocate ourselves into the scope of this PhD; Species distribution modelling in fisheries science.

Species distribution modelling (SDM) is the framework in which ecologists allocate all the statistical procedures and/or models used to characterise the distribution of species. The starting point for most SDM statistical models is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{Y}$ is the vector of the random sample corresponding to the values of the response variable at some locations, $\boldsymbol{X}$ is a matrix with the values of the covariates and $\boldsymbol{\beta}$ is a vector of parameters that quantifies the fixed effects of our covariates $\boldsymbol{X}$ on our response variable $\boldsymbol{Y}$. The final piece $\boldsymbol{\epsilon}$ corresponds to the vector of errors, each one meant to be normally distributed with mean zero and constant variance $N(0, \sigma^2)$.

Based on this baseline model (1.1), we can elaborate and define more appropriate models to describe the process under study. The progression of the model depends on the nature of the response variable $\boldsymbol{Y}$, its relationship with the covariates $\boldsymbol{X}$ and the structure of the model residuals ϵ.

1.3.1 Different response variables

Depending on the sampling design and the nature of the process under study, our response variable $\boldsymbol{Y}$ can be expressed in terms of different probability distributions (different properties and distributional shapes of the data). The most usual case is the ordinary linear model, where the data is assumed to be normally distributed with a given mean and variance. However, many processes such as counts or proportions do not fit such a continuous and unbounded distribution so we need to extend the linear model to the generalised linear model (GLM) (Nelder and Wedderburn, 1972)

$$\boldsymbol{Y} \sim \pi(y), \tag{1.2}$$

where $\pi(y)$ is any given probability function that suits our data $\boldsymbol{Y}$.

GLMs extend the linear model to other probability distributions of the exponential family, e.g. binomial, Poisson, gamma, etc. This way, we can model diverse types of measurements: animal counts through a Poisson distribution, presence-absence of a given species through a Bernoulli distribution, biomass data using the continuous and always positive gamma distribution, etc. The bounded nature of the response variables in GLMs requires of a link function that allows modelling the expected mean in the whole real line $(-\infty, \infty)$ and transform it to its original domain. More precisely,

$$g(\mu_y) = X\boldsymbol{\beta} \tag{1.3}$$

being $g()$ the link function that relates the mean μ_y with the linear predictor, the most usual being the *log* and *logit* links.

1.3.2 Relationship with the covariates

In the previous subsection 1.3.1, we were modelling $\boldsymbol{Y}$ by applying linear functions over the covariates $\boldsymbol{X}$. However, it is most often the case that process-covariate (environment) relationships show non-linear trends (Guisan et al., 2002).

In this regard, Hastie and Tibshirani (1990) developed the generalised additive models (GAM). GAMs are a semi-parametric extension of GLMs, where we assume that the structure of the linear predictor is additive and that some components are smooth. In GAMs, these smooth components are typically modelled through different types of smoothing-splines (Eilers and Marx, 1996), that allow fitting non-linear effects to the covariates

$$\boldsymbol{Y} = \beta_0 + \sum_{l=1}^{L} \boldsymbol{f_l}(\boldsymbol{x_l}) + \boldsymbol{\epsilon} \tag{1.4}$$

where β_0 is the intercept of the model and $\boldsymbol{f_l}()$ are non-linear functions applied to the covariates.

1.3.3 Structure of model residuals

So far in SDM, we have described how to plug in different likelihood functions to deal with processes of different nature through GLMs. We have also seen the extension from GLMs to GAMs in order to allow fitting non-linear effects to our covariates. With these, we can already apply a notable number of models that typically do inference on the mean of the response variable μ_y.

Every statistical model has model residuals $\boldsymbol{\epsilon}$, i.e. the deviation of the observed values $\boldsymbol{Y}$ from the fitted mean of the model

$$\boldsymbol{Y} = \beta_0 + \sum_{l=1}^{L} \boldsymbol{f_l}(x_l) + \boldsymbol{\epsilon}, \tag{1.5}$$

where model residuals $\boldsymbol{\epsilon}$ should be independent given the model.

However, it is often the case that model residuals display non-independent patterns or structures that our model covariates have not been able to explain. The presence of correlated model residuals compromises the fit of the whole model and its quantification of uncertainty (Næs and Mevik, 2001; Fortin and Dale, 2009; Legendre et al., 2002). Therefore, we should try to get rid of these unobserved structures.

Depending on the process under study and its sampling design, the unobserved structure can take several correlation structures. For instance, if we have repeated measurements of a process at each sampling site, we may expect correlated residuals within site because within site measurements are likely to be more similar among them than with other sites that have similar characteristics. Therefore, assigning a random noise effect (iid) to each site could solve the problem.

Similarly, if we sample this very same process over time, model residuals may be temporally correlated. A usual way of dealing with this temporal structure is by means of time series analysis techniques, which can introduce different kinds of temporal correlation terms by, for example, applying Holt-Winters exponential smoothing trends (Chatfield and Yar, 1988) or auto-regressive integrated moving-average models (Wei, 1994) among others.

In the spatial case, model residuals are also prone to spatial correlation (Kneib et al., 2008; Carl and Kühn, 2007). In order to deal with such spatial correlation, we rely on the principle that "*near things are more related than distant things*" (Tobler, 1970). Again, depending on the nature of the data and its spatial domain $\mathcal{D}$, the spatial structure of the residuals can vary. In the case of areal data, correlation structures are often specified using conditional autoregressive models with a given order of neighbouring regions (Besag et al., 1991). In the case of point-patterns and point-referenced data, the spatial domain $\mathcal{D}$ is continuous, thus correlation functions need also to be continuous over distance. We discuss more about continuous spatial fields and continuous autocorrelation functions in the following section 1.4.

1.4 Continuous spatial autocorrelation and kriging

As already mentioned in section 1.1, fishery data is typically represented as a point in space, so we usually deal with a finite set of point-referenced data $\boldsymbol{Y}(\boldsymbol{s})$ over a continuous, generally two-dimensional, fixed spatial domain $\mathcal{D}$. While, it is sensible to assume that the probability of presence is measurable at all infinite possible sites in the domain, in practice the data are only a partial realisation of the whole spatial process. In other words, we only have measurements at a finite number of locations out of an infinite number of possible locations. For example, $\boldsymbol{Y}(\boldsymbol{s})$ may represent the biomass of a given species at sites $\boldsymbol{s}$. The main problem that we face in these cases is that we have to perform inference about the spatial structure of $\boldsymbol{Y}(\boldsymbol{s})$, i.e. infer a distance based covariance function that best represents the underlying spatial field of our data, and then predict at unsampled locations using kriging (Cressie, 1990) interpolation based on this covariance function.

The underlying spatial process is typically assumed to be a Gaussian field (GF), which means that in a set of locations $\boldsymbol{s}$, the vector of observations $\boldsymbol{Y}(\boldsymbol{s})$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu}$ and a spatially

structured covariance matrix $\mathbf{\Sigma}$. The key for modelling these situations is finding a distance based covariance function $\mathcal{C}()$ that represents the covariance matrix $\mathbf{\Sigma}$

$$\Sigma_{ab} = Cov(y(s_a), y(s_b)) = \mathcal{C}(y(s_a), y(s_b)) \tag{1.6}$$

where $s_a, s_b \in \boldsymbol{s}$.

In order to perform inference on the covariance function of our GF, we usually assume that it fulfils two characteristics:

- The GF is *second order stationary*, which means that the field has constant mean and its covariance function only depends on the distance vector $(s_a, s_b) \in \mathbb{R}^2$, i.e. $Cov(y(s_a), y(s_b)) = \mathcal{C}(s_a - s_b)$.

- The GF is *isotropic*, which means that the covariance function does not depend on the direction of the distance but just the Euclidean distance between observations $||s_a - s_b|| \in \mathbb{R}$.

In the scope of this thesis we assume second-order stationary and isotropic GFs, but obviously, not all GFs fulfil these two characteristics. For example, when modelling the distribution of whales near the coast, we cannot expect the GF to behave equally in all directions because no whales should be expected on land. See pages 31-32 and 63-70 in Banerjee et al. (2014) for a more in depth text on anisotropic and non-stationary processes respectively.

From now on, we will assume that the spatial correlation of the data is a function of distance between points solely, i.e. the spatial correlation is determined by an *isotropic* and *second order stationary* covariance function.

Over the years many covariance functions have been proposed (see Banerjee et al. (2014) for an extended description). Among all, the Matérn class of covariance models, named by Stein (1999) after the Swedish forestry statistician Bertil Matérn (Matérn, 2013), is the most flexible as it embraces a number of covariance functions depending on the value of its smoothing parameter. The Matérn covariance between two points separated by $||s_a - s_b||$ distance units and parametrised as given by Handcock and Wallis (1994) looks

like this

$$\mathcal{C}(s_a, s_b) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}||s_a - s_b||}{r} \right)^{\nu} K_\nu \left(\frac{2\nu^{1/2}||s_a - s_b||}{r} \right), \quad (1.7)$$

where ν is the smoothness parameter ($\nu > 0$), K_ν is a modified Bessel function of the second kind (Abramowitz and Stegun, 1964) and order ν, Γ is the gamma function and r is the range parameter ($r > 0$), which measures the distance at which the covariance is assumed to be zero (there is no autocorrelation). Depending on the smoothness parameter ν and range r, the Matérn covariance function can take various shapes (Figure 1.5).

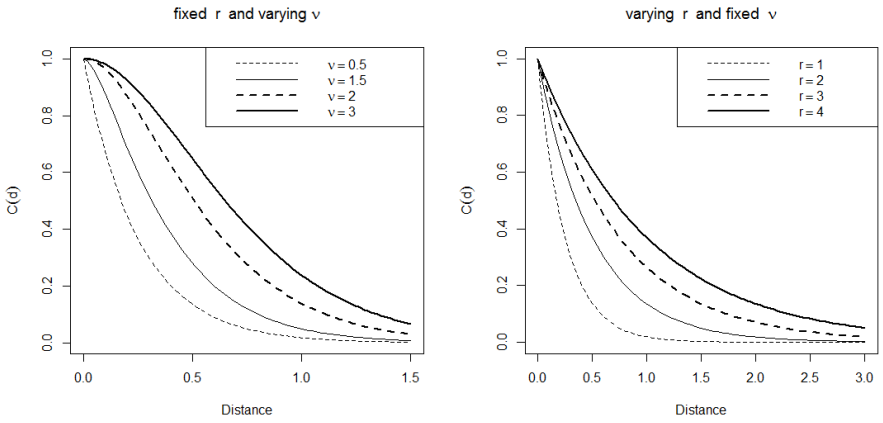


Figure 1.5. Different shapes of the Matérn covariance function depending on its range r and smoothing parameter ν.

1.5 Bayesian hierarchical spatial modelling

Spatial modelling is essentially concerned with three issues: model specification, estimation and inference of parameter estimates, and prediction. It is well known that the Bayesian approach can more easily address model specification, and therefore inference and prediction as well (Banerjee et al., 2014;

Wasserman, 2013). In fact, Bayesian statistics have more attractive features as compared to the frequentist approach as we will elaborate through this section.

Bayesian statistics is based on the Bayes' conditional probability theorem:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{1.8}$$

where A and B are two random variables.

The Bayes' probability theorem (1.8) can also be seen in the statistical modelling framework as:

$$p(hypothesis|data) = \frac{p(data|hypothesis)p(hypothesis)}{p(data)} \tag{1.9}$$

where typically the *hypothesis* is expressed in terms of the parameters of the statistical model ($\boldsymbol{\theta}$). As opposed to the classical or frequentist field, parameters in Bayesian statistics are defined by probability distributions, not just point estimates. $p(data|hypothesis)$ represents the likelihood of the model and $p(hypothesis)$ the prior distribution of the parameters, i.e. the information that we have about the parameters prior to the data that we are analysing in our model.

Lately, Bayesian statistics are becoming increasingly popular among ecologists (Clark, 2005) for at least two reasons: on the one hand, its direct and intuitive quantification of the uncertainty through explicit probabilistic inference is of great help for decision making purposes. On the other hand, as mentioned before, Bayesian statistics allows defining complex statistical models in a rather easy and intuitive way as discussed in the next subsection 1.5.1. However, the computational cost of Bayesian methods can be a problem, specially in big and complex datasets as we discuss in subsection 1.6.

1.5.1 The Bayesian hierarchical approach

Many datasets are organized into a hierarchy of successive levels. For example, students are in classes, classes are in schools, schools are in cities, etc. This

way, we can explain the expected outcome of a student as a sum of hierarchical effects for the class, the school, city, etc.

In a similar way, when modelling, the hierarchical approach decomposes (or accommodates) the complexity of the data into different levels. A good example of hierarchical modelling may be the use of spatial latent fields that model the remaining (unobserved) spatial correlation of the data, given the covariates, by applying distance based functions. In this setting, we have a set of parameters $\boldsymbol{\theta}$ with their respective prior distributions that quantify the fixed effects of our model (intercept, linear effects of the covariates, etc.). However, the spatial latent field follows a distribution $N(0, \Sigma)$ that depends on some hyperparameters Ω (with their own prior distributions) that characterise the structure of the spatial latent field. In this setting, it is evident that we need one more level or stage in our model, a level to specify the distribution of the latent variable.

Let Y be a normally distributed and spatially correlated process given the observed covariates $\boldsymbol{X}$. We can express Y in three stages.

First stage: $Y|\boldsymbol{\theta}, W \sim N(\boldsymbol{X}\boldsymbol{\beta} + W, \rho)$,

where Y is conditionally independent and normally distributed given the parameters ($\boldsymbol{\theta} = \boldsymbol{\beta}$) and a spatial latent field (W).

Second stage: $W|\boldsymbol{\Omega} \sim N(0, \boldsymbol{\Omega})$,

where W is a latent Gaussian spatial model with hyperparameters $\boldsymbol{\Omega}$.

Third stage: priors on $(\boldsymbol{\beta}, \rho, \boldsymbol{\Omega})$.

In other words, the first stage of the Bayesian hierarchical model specifies the likelihood of the model by characterising the data Y given the parameters $\boldsymbol{\theta} = \boldsymbol{\beta}$ of the model and the fitted latent process W. The second stage specifies the latent process W through its hyperparameters $\boldsymbol{\Omega}$ and the third stage specifies the prior distributions of all the parameters and hyperparameters

involved in the model.

A usual problem with this kind of hierarchical models is the fact that, often, there is no closed expression for the marginal posterior distributions of the parameters $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, Y)$, so numerical approximations are needed.

1.6 Bayesian computing

As we have see, performing Bayesian inference means combining likelihood $p(Y|\boldsymbol{\theta})$ and priors $p(\boldsymbol{\theta})$ to get the posterior distributions $p(\boldsymbol{\theta}|Y)$ of our parameters and hyperparameters (here just parameters for the sake of simplicity)

$$p(\boldsymbol{\theta}|Y) \propto p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.10}$$

Note that, following Bayes' theorem as in (1.8), we miss $p(Y)$ in the denominator and in exchange we assign proportionality rather than equality. The proportionality symbol $\propto$ expresses the fact that the product of the likelihood function $p(Y|\boldsymbol{\theta})$ and the prior distributions $p(\boldsymbol{\theta})$ on the right hand side of (1.10) must be scaled to integrate to one over the range of plausible parameter values.

Unless the posterior distribution and the prior distribution belong to the same family, in which case we have a conjugated model (see Gutiérrez-Peña et al. (1997) for a review on conjugate models), the integrals involved in getting the posterior distributions of the models are generally analytically intractable. Consequently, we require of numerical approximations to get these posteriors.

1.6.1 Markov Chain Monte Carlo

The most widely used computing tools in Bayesian statistics today follow Markov chain Monte Carlo (MCMC) simulation methods. MCMC methods are a class of algorithms that allow us to draw samples from some probability distribution without having to know their exact density at any point. By

means of this property, when applying MCMC, we are able to draw independent samples from the posterior distributions in such a way that the number of times visited a given value is proportional to its density in the posterior distribution. Therefore, in MCMC we do not get a closed form of the posterior but a sample of values from it. A key functional property of MCMC is their ability to do inference on highly dimensional problems, by reducing such dimension to low-dimensional (often unidimensional) problems.

There are many different MCMC algorithms available. In this section we briefly describe only two of these (see Gamerman and Lopes (2006) for a more detailed text on MCMC methods) to get an idea of what MCMC implies as compared to the Integrate Nested Laplace Approximation (INLA) method used in the following chapters.

Gibbs sampler

One of the attractive methods for setting up an MCMC algorithm is Gibbs sampling. Suppose that the parameter vector of interest is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. The joint posterior distribution of $\boldsymbol{\theta}$, which we denote by $(\boldsymbol{\theta}|Y)$, may be of high dimension and difficult to summarize. Suppose we define the set of conditional distributions:

$$\begin{aligned} &(\theta_1|\theta_2, \ldots, \theta_l, Y) \\ &(\theta_2|\theta_1, \ldots, \theta_l, Y) \\ &\ldots \\ &(\theta_l|\theta_1, \ldots, \theta_{l-1}, Y) \end{aligned} \tag{1.11}$$

The Basic idea under Gibbs sampling is that we can set up a Markov chain simulation algorithm from the joint posterior distribution by successfully simulating individual parameters from the set of l conditional distributions.

Metropolis-Hastings

Maybe, the most popular way of constructing a Markov chain is by using a Metropolis-Hastings algorithm. A Metropolis-Hastings algorithm begins with

an initial value θ_l^0 and specifies a rule for simulating the t^{th} value in the sequence θ_l^t given its value in $(t-1)$. This rule consists of a proposal density, which simulates a candidate value θ_l, and a computation of an acceptance probability P, which indicates the probability that the candidate value is accepted as the next value in the sequence:

1. Simulate a candidate value θ^* from a proposal density $p(\theta_l^*|\theta_l^{t-1})$
2. Compute the ratio $R = \frac{h(\theta_l^*)p(\theta_l^{t-1}|\theta_l^*)}{h(\theta_l^{t-1})p(\theta_l^*|\theta_l^{t-1})}$
3. Compute the acceptance probability $P = min\{R, 1\}$ Sample a value θ_l^t such that $\theta_l^t = \theta_l^*$ with probability P; otherwise $\theta_l^t = \theta_l^{t-1}$

where $h(\theta_l)$ is proportional to the desired probability distribution $P(x)$.

Among other important MCMC features, such as monitoring convergence of different chains, computational cost is maybe the most limiting factor of Bayesian MCMC methods. When models become complex, specially models with hierarchical structures, MCMC algorithms may be extremely slow or even become computationally unfeasible (Gelfand, 2012; Taylor, 2015). This computational crush occurs particularly in the case of spatial and spatio-temporal models, which is usually known as the "big n problem" (Banerjee et al., 2014; Lasinio et al., 2013). A good alternative to MCMC methods able to reduce the computational costs of Bayesian inference is the Integrated Nested Laplace Approximation (INLA) algorithm (Rue et al., 2009).

1.6.2 Integrated Nested Laplace Approximation

The INLA algorithm, proposed by Rue et al. (2009) and available in the `R-INLA` software package, is a numerical approximation method (rather than simulation as in MCMC) for Bayesian inference. The most remarkable feature of INLA, as opposed to MCMC, is that it allows the posterior distributions to be accurately approximated through Laplace approximations (Laplace, 1986; Tierney and Kadane, 1986), even for complex models without becoming computationally prohibitive (Rue et al., 2009).

INLA is applicable to a very popular subset of structured additive regression models named *latent Gaussian models*. Latent Gaussian models are the subset of all the Bayesian hierarchical models with a structured additive predictor which latent models have Gaussian prior distributions (as in step 2 of 1.5.1), but not necessarily their hyperparameters $\boldsymbol{\Omega}$. As most structured Bayesian models follow such form (Gelman et al., 2004; Rue et al., 2009), we can say that latent Gaussian models embrace a very wide range of statistical models: mixed models, survival models, random walk smoothing models, spatial and spatio-temporal models, etc.

A key feature for the implementation of INLA is that many latent Gaussian models admit conditional independence properties. This means that many latent Gaussian fields can be expressed as a Gaussian Markov Random Field (GMRF) with sparse precision matrix (Rue and Held, 2005), which allows using computationally efficient numerical methods for sparse matrices (Rue and Held, 2005).

How INLA works

As previously mentioned, INLA relies on Laplace approximation methods (Laplace, 1986; Tierney and Kadane, 1986) to numerically approximate posterior distributions. This method performs Gaussian approximations of the parameters by inferring their mode. Although posterior distributions do not necessarily have to be Gaussian, INLA relies on the fact that for most real problems and datasets, the conditional posterior of the latent field looks "almost" Gaussian (Rue et al., 2009). This is clearly assisted by the, non-negligible, impact of the Gaussian priors on the posteriors.

The approximation of model parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\Omega}$ in INLA is computed in three steps:

1. The first step approximates the posterior marginal $p(\boldsymbol{\Omega}|Y)$ by using the

Laplace approximation

$$\begin{aligned} p(\boldsymbol{\Omega}|Y) &= \frac{p(\boldsymbol{\theta},\boldsymbol{\Omega}|Y)}{p(\boldsymbol{\theta}|\boldsymbol{\Omega},Y} \\ &\propto \frac{p(Y|\boldsymbol{\theta},\boldsymbol{\Omega})p(\boldsymbol{\theta}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})}{p(\boldsymbol{\theta}|\boldsymbol{\Omega},Y)} \\ &\approx \frac{p(Y|\boldsymbol{\theta},\boldsymbol{\Omega})p(\boldsymbol{\theta}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})}{\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{\Omega},Y)}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\Omega})} = \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{\Omega},Y) \end{aligned} \tag{1.12}$$

where $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{\Omega},Y)$ is the Gaussian approximation – given by the Laplace method – of $p(\boldsymbol{\theta}|\boldsymbol{\Omega},Y)$ and $\boldsymbol{\theta}^*(\boldsymbol{\Omega})$ is the mode for a given $\boldsymbol{\Omega}$.

2. The second step approximates $p(\theta_i|\boldsymbol{\Omega},Y)$ by using again Laplace approximations. Given that $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$

$$\begin{aligned} p(\theta_i|\boldsymbol{\Omega},Y) &= \frac{p((\theta_i,\boldsymbol{\theta}_{-i})|\boldsymbol{\Omega},Y)}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\Omega},Y)} \\ &\propto \frac{p(\theta_i|\boldsymbol{\Omega},Y)}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\Omega},Y)} \\ &\approx \frac{p(\boldsymbol{\theta}|\boldsymbol{\Omega},Y)}{\widetilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\Omega},Y)}\bigg|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\Omega)} = \widetilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\Omega},Y) \end{aligned} \tag{1.13}$$

where $\widetilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\Omega},Y)$ is the Laplace Gaussian approximation of the probability distribution $p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\Omega},Y)$ and $\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\Omega})$ is its mode. This strategy can be very computationally expensive since $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{\Omega},Y)$ has to be recomputed for each value of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$. A more computationally efficient but slightly less accurate approach is the so-called "simplified Laplace approximation". This method is based on a Taylor's series expansion of the Laplace approximation $\widetilde{p}(\theta_i|\boldsymbol{\Omega},Y)$. The result is then corrected using a spline term to increase the fit of the required distribution.

3. The third step is to compute the marginal posterior distributions of $p(\theta_i|Y)$ by using $\widetilde{p}(\theta_i|\boldsymbol{\Omega},Y)$ and $\widetilde{p}(\boldsymbol{\Omega}|Y)$ from the previous two steps.

$$p(\theta_i|Y) \approx \int \widetilde{p}(\theta_i|\boldsymbol{\Omega},Y)\widetilde{p}(\boldsymbol{\Omega}|Y)d\boldsymbol{\Omega}, \tag{1.14}$$

where the integral can be solved numerically through a finite weighted sum applied in certain integration points and then interpolating in between. For a more detailed text on the selection of integration points see section 3.1(c) in Rue et al. (2009).

1.6.3 Spatial Gaussian fields in INLA: the SPDE approach

Spatial Gaussian Fields (GF) are widely used in geostatistical and point-pattern problems. The biggest problem with spatial GFs is the so called "big n problem" that concerns the computational costs required for performing algebra operations with dense covariance matrices $\mathbf{\Sigma}$ to infer the spatial covariance function $\mathcal{C}$ that best suit $\mathbf{\Sigma}$ as in (1.6).

In the previous section, we have mentioned that INLA exploits the computational properties of GMRFs (with sparse precision matrices) to fit a good number of latent Gaussian models. Unfortunately, spatial GFs are continuous, so do not satisfy the properties of a GMRF. Therefore, in principle, GFs should not be applicable INLA.

In this regard, Lindgren et al. (2011) found an explicit link between GFs and GMRFs through the Stochastic Partial Differential Equation (SPDE) approach. This approach allows representing a spatial GF with Matérn covariance function by a GMRF. It is important to note that, while the computational cost of a two-dimensional GF is approximately n^3, the computational cost of a two-dimensional GMRF is $n^{3/2}$. Furthermore, the GMRF property allows us use the computationally efficient INLA approach.

The Stochastic Partial Differential Equation approach

As we have introduced, Lindgren et al. (2011)'s approximation of a GF requires that its covariance function is of the Matérn family. Following Lindgren et al. (2011)'s notation, we can reparametrise the Matérn covariance function

in 1.5 for an stationary and isotropic GF as

$$\mathcal{C}(d) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||s_i - s_j||)^\nu K_\nu(\kappa||s_i - s_j||), \tag{1.15}$$

where now κ is a scaling parameter that determines the effective range r of the spatial effect by $r = \sqrt{8\nu}/\kappa$, corresponding to the distance where correlations are near 0.1.

The great discovery by Lindgren et al. (2011) falls on the fact that a GF $z(\boldsymbol{s})$ with Matérn covariance function is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} z(\tau \boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0, \tag{1.16}$$

where Δ is the Laplacian, τ controls the variance, d is the dimension of the GF and $\mathcal{W}$ is a Gaussian spatial white noise process.

The link between the SPDE in 1.16 and the parameters of the Matérn covariance function 1.15 is given by the following equations that involve the smoothness parameter ν and the marginal variance σ^2

$$f(n) = \begin{cases} \nu & = \alpha - d/2 \\ \sigma^2 & = \frac{\Gamma(\nu)}{\Gamma(\alpha)()4\pi)^{d/2}\kappa^{2\nu}\tau^2} \end{cases}$$

As we usually work in the two-dimensional framework ($d = 2$), it can be rewritten as

$$f(n) = \begin{cases} \nu & = \alpha - 1 \\ \sigma^2 & = \frac{\Gamma(\nu)}{\Gamma(\alpha)()4\pi)\kappa^{2\nu}\tau^2} \end{cases}$$

In R-INLA, the default value of $\alpha = 2$ which translates into $\nu = 1$ (see Figure 1.5), but $0 \leq \alpha < 2$ are also available, although not fully tested (Lindgren et al., 2011). For this particular case of the Matérn covariance function the range r and the marginal variance of the field σ^2 are approximately:

$$\begin{aligned} r &= \sqrt{8}/\kappa \\ \sigma^2 &= 1/(4\pi\kappa^2\tau^2) \end{aligned} \tag{1.17}$$

In R-INLA, the default parametrisation of the SPDE is in terms of $log(\tau) = \theta_1$ and $log(\kappa) = \theta_2$. Then a joint Gaussian prior distribution is given to θ_1 and θ_2.

As Lindgren et al. (2011) show, the solution to the SPDE can be approximated using the finite element method through a deterministic basis function representation defined on a triangulation of the domain $\mathcal{D}$ (see Figure 1.6). Such triangulation, typically named as `mesh`, of the study area is based on Delaunay triangulations (Delaunay, 1934), which as opposed to a regular grid, it allows a flexible partition of the region into triangles, that can satisfy different types of constraints to better accommodate different characteristics of the study area. See Krainski et al. (2016) for a complete tutorial on how to create a good mesh in R-INLA.

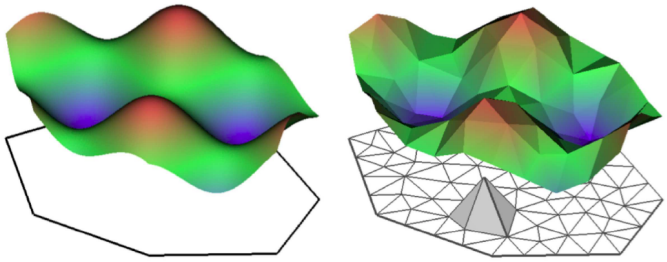


Figure 1.6. Left: example of a Gaussian Field $z(\boldsymbol{s})$. Right: corresponding finite element representation of the Gaussian Field $z(\boldsymbol{s})$. Figure extracted from Cameletti et al. (2013)

For the non-stationary case of the Gaussian field, e.g. cases where unmeasured covariates may affect the behaviour of the spatial correlation in different locations, it is also possible to extend the SPDE approach by specifying spatially varying hyperparameters $\kappa(s)$ and $\tau(s)$ (Lindgren et al., 2011; Bolin and Lindgren, 2011). We do not further describe the approach for non-stationary Gaussian field because it is out of the scope of this PhD dissertation. If in-

terested, Lindgren et al. (2011) discusses this case in section 3.2, while Bolin and Lindgren (2011) shows an application in global ozone and lately Krainski et al. (2016) has included a generic example on how to apply it in INLA.

Fitting an stationary geostatistical model in R-INLA

In what follows we briefly describe how to fit a generic two dimensional geostatistical model in R-INLA. To do so, we first need to define the SPDE model over the triangulation of the study area:

```
# Sampled locations
loc <- cbind(longitude,latitude)

# Construct boundary of sampled locations
boundary <- inla.nonconvex.hull(points=loc)

# Create the mesh
mesh <- inla.mesh.2d(boundary=boundary, loc=null)

# Create the SPDE model over the mesh
spde <- inla.spde2.matern(mesh, constr=TRUE)
```

Here the mesh is created without the sampling locations assigned to any particular node of the mesh. If that was a requirement, we could do so by setting the `loc=loc` statement inside the `inla.mesh.2d()` function. Similarly, when creating the SPDE model an integrate-to-zero constraint can be applied by adding the `constr=TRUE` statement inside the `inla.spde2.matern()` function. The rest of the parameters for the creation of the boundary, mesh and SPDE model are the default implemented in R-INLA. See R-INLA help documents for more details.

Then, it is necessary to create the indexation of the observations and create weight matrices of the estimation locations into the nodes of the mesh

```
# Sampled locations indexation
```

```
idx.mat.est <- inla.spde.make.A(mesh=mesh,loc=loc)

# Prediction locations indexation
idx.mat.prediction <- inla.spde.make.A(mesh=mesh,
                          loc=mesh$loc[,1:2)
```

where again the rest of parameters are set as default. These other parameters are essential to add temporal and/or bloc indexations.

Lastly, before fitting the model using the `inla()` call function, we need to stack the data set and the formula of the model

```
# stack data
est <- inla.stack(data=list(y=y), A=list(idx.mat.est),
                  effects=list(spat=1:spde$n.spde))
pred <- inla.stack(data=list(y=rep(NA,mesh$n),
                  A=list(idx.mat.pred),
                  effects=list(spat=1:spde$n.spde)))
data <- inla.stack(est,pred)

# Set formula
formula <- 1 + f(spat,model=spde)
```

where we fit a simple geostatistical model with an intercept and the geostatistical term.

Starting from this simple ordinary kriging example, along the following chapters we develop more complex geostatistical structures and regressors, with linear and different sorts of non-linear effects, to identify the model that best fit the process under study. For that we need to assess the quality of our models based on model selection criteria.

1.7 Bayesian model selection scores

Bayesian models can be evaluated and compared in several ways (Schwarz, 1978; Akaike, 1998; Geisser, 1993; Berger and Pericchi, 1996; Gelman et al.,

1996; Spiegelhalter et al., 2002; Watanabe, 2010; Vehtari et al., 2012; Gelman et al., 2014). Arguably, the ultimate goal of every statistical model is to have good predictive properties, thus we need to evaluate their predictive accuracy, compare them and select the most appropriate model for our particular data or problem.

Even if all of the models being considered have mismatches with the data, it is informative to evaluate their predictive accuracy, compare them, and consider where to go next. The challenge then is to estimate predictive model accuracy, correcting for the bias inherent in evaluating a model's predictions of the data that were used to fit it.

The natural way of assessing the predictive quality of a model is undoubtedly cross-validation (Geisser and Eddy, 1979) but this requires several repeated model fits and can be computationally prohibitive in complex models or simply not possible due to having little data. Therefore, for practical reasons many other model selection scores have been proposed (Hooten and Hobbs, 2015).

In this regard, R-INLA can compute, as by-product of the main computations, few model selection scores: Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002), marginal likelihoods, conditional predictive ordinate (CPO) (Geisser, 1993), the cross-validated probability integral transform (PIT) (Czado et al., 2009) and the Watanabe Akaike Information Criterion (WAIC)(Watanabe, 2010).

In `R-INLA`, these scores are automatically computed by including the following statement within the `inla()` call:

```
# compute DIC, CPO and WAIC scores
inla(..., control.compute=list(cpo = TRUE , waic = TRUE,
                               dic = TRUE))
```

For the purpose of this PhD dissertation, in what follows we end up this chapter describing DIC, CPO and WAIC.

1.7.1 Deviance Information Criterion

DIC is a within-sample predictive score, defined as

$$DIC = 2 * E(D(\boldsymbol{\theta})) - D(E(\boldsymbol{\theta})), \tag{1.18}$$

where the Deviance is $D(\boldsymbol{\theta}) = -2 * log(p(y|\boldsymbol{\theta}))$ and $E(D(\boldsymbol{\theta})) - D(E(\boldsymbol{\theta}))$ is the effective number of parameters.

While the deviance benefits good fit, the effective number of parameters penalizes the DIC score. However, DIC scores may underpenalize complex models with many random effects (Plummer, 2008).

1.7.2 Cross-validatory Predictive Score (CPO) and and its logarithmic score (LCPO)

CPO is a leave-one-out cross-validation score:

$$CPO_i = p(y_i|y_{-i}), \tag{1.19}$$

where Y_{-1} corresponds to the observations y with the ith observation removed.

Therefore CPO expresses the posterior probability of observing the value of y_i when the model is fitted to all data except y_i. Based on the CPO values of each observation, we can calculate the logarithmic score LCPO (Gneiting and Raftery, 2007) as:

$$LCPO = -\sum_i log(CPO_i). \tag{1.20}$$

Therefore smaller LCPO scores indicate a better predictive quality of the model.

1.7.3 Watanabe-Akaike Information Criterion (WAIC)

WAIC is a within-sample predictive score, defined as:

$$WAIC = \sum_i var_{post}(log(p(y_i|\boldsymbol{\theta}))) \tag{1.21}$$

where we compute the posterior variance of the log predictive density for each data point y_i. Therefore, WAIC is fully Bayesian using posterior densities more effectively than the DIC (Gelman et al., 2014) and CPO.

Chapter 2

Spatial beta regression to identify fishing-suitable areas

Fishery discards refer to the portion of the catch that is not retained on board during commercial fishing operations and is returned to the sea (Catchpole et al., 2005). Discards include non-commercial species, non-marketable commercial material and marketable organisms. Discarding patterns change both in time and space as a consequence of changing economic, sociological, environmental and biological factors (Gillis et al., 1995; Maynou and Sardà, 2001). So far, the only way of quantifying discards has been through on-board sampling programmes, also known as fishery dependent data.

As we introduced at the beginning of the previous chapter, fishery dependent data refers to fishery data collected on board of commercial vessels. These data is usually collected through a stratified sampling scheme that samples proportionally to the number of fishing days performed by a given fishing gear at a given fishing area (European Comission, 2009).

The original goal of these data is to get a spatially and temporally repre-

sentative sample of fishers fishing activity. Unfortunately, due to logistic and budgetary reasons such resolution is generally not enough to cover all national fishing grounds. For example, the data that we will be using in this chapter refers to one of the three locations sampled in the whole southern Spanish Mediterranean. In Figure 2.1 we can see that there is a trade-off between the macro and meso scale sampling, where large areas are not sampled at the macro-scale but the spatial resolution at the meso-scale is pretty good.

Figure 2.1. On-board sampling locations in the souther Spanish Mediterranean. Resolution is good at the meso-scale but the macroscopic spatial coverage not as much.

Fishery dependent data, unlike fishery independent data, allow quantifying fishery discards. Fishery discards have been a matter of debate over the last decades. Unwanted catches and discards constitute a substantial waste of natural resources that negatively affect the sustainable exploitation of marine ecosystems and the financial viability of fisheries (Kelleher, 2005; Viana et al., 2013b). As a consequence, the new EU Common Fisheries Policy plan (European Comission, 2013) proposed for 2014-2020, is controversial in its goal to enforce the landing of fishing discards as a measure to encourage their reduction (Sardà et al., 2015). This measure implies that fishers will have to land all regulated species regardless of their size and that landing quotas will be

replaced by catch quotas. While many potential consequences of this policy are still unclear for different fishing grounds and fisheries (Bellido et al., 2014; Catchpole et al., 2005; Sigurðardóttir et al., 2015), it will undoubtedly have a negative economic impact on the primary fishery sector (Catchpole et al., 2014; Poseidon, 2013). Indeed, fishers will be obliged to land products of little value, which will devalue the economic potential of their catch quotas. Similarly, in those areas where the fishery is not managed by quotas (e.g. Mediterranean Sea), fishers will have to keep the fish on-board which may imply additional costs associated with handling the fish and reduced storage capacity.

Under these circumstances, a possible consequence of the upcoming EU fishing scenario is that it will motivate fishers to maximise revenues by fishing in areas that minimise the catch of unwanted regulated fish species (Vilela and Bellido, 2015). In this context, spatial analysis could help to identify areas where by-catch and/or discards are minimised. By providing spatial or spatio-temporal by-catch/discard predictive maps, both management bodies and fishers could better asses the fishing suitability of a given area.

Several studies have assessed discard concentration areas based on the expected amount of total discards per unit effort (DPUE) (Feekings et al., 2012, 2013; Viana et al., 2013a; Cosandey-Godin et al., 2014; Pennino et al., 2014). However, the use of DPUE units as a criterion to identify these areas can lead to ecologically and economically misleading results for two main reasons. Firstly, such an approach does not include the landed portion of the fishing haul in the analysis, and so it does not identify whether the amount of discard is disproportionate to the catch or not. This is crucial to quantify the economic and ecologic balance between the marketed food biomass and the lost biomass. Secondly, from a technical point of view, modelling discards involves dealing with a wide range of commercial vessels with different characteristics (e.g. length of the vessel, engine-power, etc.), haul duration, and other effort characteristics. Consequently, calculating a standardised DPUE criterion may be difficult or even infeasible in most cases.

A better approach may be to use discard and by-catch ratios, defined

as the discarded or by-caught biomass divided by the total haul biomass. By contrast to the DPUE criterion, discard ratios implicitly include benefit versus loss, which allows us quantify both, the ecological impact in terms of "*food biomass vs. wasted biomass*" and the economic impact by quantifying the percentage of quota loss (if applicable) and percentage of storage room occupied in the vessel by non-marketable fish. Therefore, discard ratios allow a better identification of both, economically and ecologically, fishing-suitable areas. In addition, technically speaking, discard/by-catch ratios are inherently standardised to a wide range of effort variables (vessel size, fishing time, etc.) apart from the most gear specific ones (hook size, mesh size, etc.).

Interestingly, proportions have been widely used in many descriptive studies on fishery discards (Tsagarakis et al., 2013); however, we found no fishery study that applies statistical regression to them. Vilela and Bellido (2015) proposed a random forest based algorithm to assess the fishing suitability of an area. Their algorithm is based on a fishing-suitable/unsuitable (binomial) response variable that is created by manually setting a cut-off discard ratio that classifies hauls as suitable or unsuitable. It is therefore somewhat intuitive that results using this method may be very sensitive to the cut-off percentage set at the beginning of the analysis.

An easier and more straightforward approach is to directly apply regression on discard or by-catch ratios using beta distribution. The beta distribution has historically had a very wide range of applications (Gupta and Nadarajah, 2004), although not until recently has it been used in regression modelling (Ferrari and Cribari-Neto, 2004).

2.1 Modelling proportions

Many ecological processes are spatially or spatio-temporally sampled and measured as proportions; one example is sea-grass coverage in a area. The traditional approach in ecology has been to, first transform proportional data to approximate normality, and then analyse them using Gaussian linear models, such as analysis of variance or linear regression.

A very common transformation is the arcsine square root transformation. This transformation can be useful to stabilise variances and normalise the data but there are several reasons why it should be avoided. Firstly, model parameters cannot be easily interpreted in terms of the original response (Warton and Hui, 2011; Ferrari and Cribari-Neto, 2004). Secondly, the efficacy of the arcsine transformation in normalising proportional data is heavily dependent on the sample size, and does not perform well at extreme ends of the distribution (Warton and Hui, 2011; Wilson and Hardy, 2002). Thirdly, measures of proportions typically display asymmetry, and hence inference based on the normality assumption can be misleading (Ferrari and Cribari-Neto, 2004).

An alternative that is becoming more prevalent in ecological analyses is the logistic regression, an analytical method designed to deal with binomial proportional data (Steel et al., 1997; Wilson and Hardy, 2002; Warton and Hui, 2011), i.e. proportions measured as *x out of n*. The logistic regression provides a more biologically and ecologically interpretative analysis and is not sensitive to sample size. Nonetheless, such binomial data is prone to overdispersion, resulting in an incorrect quantification of the uncertainty when applying the proposed binomial Generalised Linear Regression (GLM). In these cases, the inclusion of a random intercept term using Generalised Linear Mixed Models (GLMMs) may improve the assessment of uncertainty (Wilson and Hardy, 2002).

When data are non-binomial, that is, observations do not follow the *x out of n* pattern, the logistic regression is no longer applicable. As an alternative approach, Warton and Hui (2011) suggested the logit transformation of the data, which overcomes the problems of interpretability and shape of the posterior estimates using the arcsine square root transformation is used. However, any transformation of the data (y_t) implies that regression parameters are only interpretable in terms of the mean of y_t and not the mean of the original data. In this regard, beta regression (Ferrari and Cribari-Neto, 2004) provides a natural approach to deal with non-binomial proportional data.

The beta distribution is a well known distribution that satisfies the characteristics of proportions, bounded to the [0, 1] interval with asymmetric shapes.

It has long been used in a wide range of applications involving proportions and probabilities (Gupta and Nadarajah, 2004). However, only recently has it been applied to regression modelling (Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006; Liu and Kong, 2015), allowing bounded posterior distributions and model parameters that are directly interpretable in terms of the mean of the response.

Aside from the likelihood function, it is well known that changes in ecological processes in time and space are driven by a set of factors and interactions. Understanding these drivers is very often the ultimate goal among scientists seeking to manage natural resources effectively. However, the immeasurable complexity of ecological spatial processes often means that the spatial variability of the data exceed the variability explained by the explanatory variables. This phenomenon usually results in spatially autocorrelated model residuals that can yield incorrect results and a restricted predictive capacity of the models (Fortin and Dale, 2009; Legendre et al., 2002).

2.2 Beta regression

Traditionally the beta distribution is denoted by two scaling parameters $Be(a, b)$. However, in order to apply regression, it is necessary to reparametrize its density distribution in terms of its mean $\mu = \frac{a}{a+b}$ and dispersion $\rho = a + b$, so that:

$$\pi(y) = \frac{\Gamma(\rho)}{\gamma(\mu\rho)\gamma(\rho(1-\mu))} y^{\mu\rho-1}(1-y)^{(1-\mu)\rho-1},\ 0 < y < 1 \qquad (2.1)$$

where Γ is the gamma function, $E(y) = \mu$ and $Var(y) = \frac{\mu(1-\mu)}{1+\rho}$. It is important to note that, as opposed to the Gaussian distribution, the variance of the beta distribution depends on the mean, which translates into maximum variance at the centre of the distribution and minimum at the edges, to support the truncated nature of the beta distribution.

It is also important to note that the probability density (equation 2.1) does not provide a satisfactory description of the data at both ends of the

distribution, zero and one. An ad hoc solution may be to add a small error value to the observations to satisfy this criterion (Warton and Hui, 2011); otherwise zero and one inflated models are required (Liu and Kong, 2015).

Following the $Be(\mu, \rho)$ reparametrisation and the notation used in chapter 1, a given set of observations $y_1, ..., y_n$ that represent proportions can be related to a set of covariates and functions using a similar approach to the generalised linear model:

$$\begin{aligned} logit(\mu_i) &= \eta_i \\ \eta_i &= \beta_0 + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{l=1}^{L} f_l(x_l) + v_i \end{aligned} \tag{2.2}$$

where, η_i enters the likelihood through a logit link, β_0 is the intercept of the model, β_j are the fixed effects of the model, $f_k()$ denote any smooth effects (including spatial dependence effects) and v_i are unstructured error terms (random variables).

At the time of writing, a handful of R packages allow beta regression: `betareg` (Grün et al., 2011), `mgcv` (Wood, 2011) and `gamlss` (Stasinopoulos and Rigby, 2007) in the frequentist field and `Bayesianbetareg` (Marin et al., 2014), `zoib` Liu and Kong (2015) and `R-INLA` (Martins et al., 2013) in the Bayesian counterpart. `zoib` allows zero/one inflated beta regression but only `R-INLA` allows a wide range of flexible hierarchical models to be fitted at a user-friendly and computationally efficient environment.

2.3 Bayesian hierarchical spatial beta regression

Bayesian hierarchical methods are becoming very popular among ecologists due to the complexity of the relationships involved in natural systems (Clark, 2005). Modelling these relationships often requires specifying sub-models inside the additive predictor that allow inferring a suspected hidden or latent effect that characterise these relationships.

A good example, as we saw in Chapter 1, may be the use of spatial latent fields that apply distance-based functions to model the spatial dependence of the data. In these cases, the main intensity of the process is driven by a set of covariates $X\beta$, also called large-scale variation, to which a spatial term is added based on a correlation function that helps us describe the unobserved small-scale variation. Consequently, we end up with a spatial correlation model, which depends on its own hyperparameters, as part of a broader model that characterises the intensity of the process; in other words, we have a hierarchical model with a spatial latent variable.

The most popular point-referenced spatial model, the geostatistical model, has the characteristic that the spatial covariance function is continuous over the range of the spatial effect. Based on this function, it is customary to assume a Gaussian latent field $W \sim N(0, Q(\kappa, \tau))$ with covariance matrix Q that depends on two hyperparameters, in the case of `R-INLA`, κ and τ. These hyperparameters determine the range and the variance of the spatial latent field. When we include this in the additive predictor of a beta distributed process Y, we obtain a hierarchical model with at least three stages:

- First stage: $Y|\beta, W \sim Be(X\beta + W, \rho)$

  where Y are conditionally independent given W.

- Second stage: $W|\kappa, \tau \sim N(0, Q(\kappa, \tau))$

  where W is a Gaussian latent spatial model.

- Third stage: priors on $(\beta, \rho, \kappa, \tau)$

As commented in the previosu chapter, a common problem with this kind of hierarchical model is that there is no closed expression for the marginal posterior distributions of the parameters $p(\beta, \rho, \kappa, \tau|Y)$, so numerical approximations are needed. In this thesis we use the INLA methodology (Rue et al., 2009) instead of the more usual Markov Chain Monte Carlo (MCMC) methods due to its computational efficiency (Simpson et al., 2011). Furthermore, inference and prediction under a geostatistical Gaussian field W entail the

so-called "big n problem" (Banerjee et al., 2014). This problem is related to the dense covariance matrix Q, which traduces into very high computational costs.

In this vein, and as we saw in Section 1.6.3 of Chapter 1, the work by Lindgren et al. (2011) provides a clever approximation of continuously indexed Gaussian Fields with Matérn covariance function to Gaussian Markov Random Fields (GMRF) using the stochastic partial differential equation approach. In other words, it allows the approximation of a continuous covariance function and dense covariance matrix by, respectively, a computationally efficient neighbourhood structure and a sparse covariance matrix. This approximation reduces the required number of computations from $O(n^3)$ (Stein et al., 2004) to $O(n^{3/2})$ (Cameletti et al., 2013) in the two dimensional spatial domain.

2.4 Case study

2.4.1 Discard data

Trawl discard data were collected according to European Comission (2009), which establishes a métier-based sampling programme of discards. Specifically, this study was based on bottom trawl data collected in the central Spanish Mediterranean Sea (Figure 2.2). Bottom trawlers in this area are segregated into two different métiers due to the difference in catch composition at different depths: the bottom otter trawl for demersal species métier (OTB-DES) and the bottom otter trawl for deep-water species métier (OTB-DWS) (see Pennino et al., 2014, for a more detailed description of the métiers).

The database, provided by the *Instituto Español de Oceanografía* (IEO, Spanish Oceanographic Institute), contains a total of 391 hauls collected between March 2009 and December 2012, including catch and discard data disaggregated by species. Two by-catch/discard proportion (henceforward simply discard) response variables were created. A total discard proportion variable was created to assess the global ecological impact of the fishery by substract-

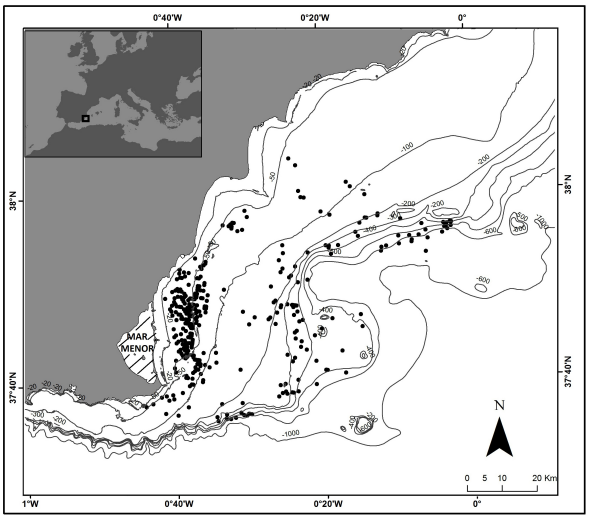


Figure 2.2. Map of the study area, located in the south-eastern part of the Spanish Mediterranean Sea. Black dots represent the centroids of the 391 sampled hauls.

ing the total catch biomass to the discarded biomass. In addition, a discard proportion of regulated species variable was created to account for the non-profitable but also non-discardable fraction of the haul by substracting the total catch to the regulated biomass that had been discarded.

$$\text{Total discard proportion:} \quad \frac{\text{Discard biomass}}{\text{Total biomass}}$$

$$\text{Regulated discard proportion:} \quad \frac{\text{Regulated species discard biomass}}{\text{Total biomass}}$$

Fishing haul characteristics, such as date, time, geolocation and depth were extracted directly from the onboard observer database. Fishing geolocation and depth were computed using an average point between the start and end point of each fishery operation. The total catch of each fishing haul, in

kilograms, was also included as a potential predictor.

It is well known that fisheries are prone to seasonal patterns. However, the spatio-temporal resolution of the data was rather small (see Table 2.1) for the identification of spatio-temporal (interaction) changes of discards distribution. Yet, the purely temporal resolution of the data along the year was good however, so different temporal trends were included in the analysis:

- Similarly a 'Ordinal day' variable was created using the *date* package (Therneau et al., 2014), which assign an integer value to each fishing haul based on the date of the haul, starting from 1 (1st January) to 365 (31st December).

- The moon is another factor that is known to affect fish distribution. For that, a 'Moon phase' variable was created using the *phenology* package (Marc, 2015). This variable can take any continuous value between 0 and 100, where 0 and 100 represent full moon and 50 new moon.

		Month											
		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Year	2009	0	0	11	11	10	11	11	8	1	4	4	8
	2010	10	9	2	8	10	0	12	9	11	1	9	12
	2011	12	8	8	7	7	11	12	10	1	14	14	11
	2012	6	3	9	10	10	12	0	13	3	13	14	11

Table 2.1. Contingency table quantifying the monthly sampling resolution across the different years.

The six most discarded fish species were: bogue (*Boops boops*) representing about 12% of the total discards, followed by the axillary seabream (*Pagellus acarne*) with 6%, the small-spotted catshark (*Scyliorhinus canicula*) with 4.5%, horse-mackerel (*Trachurus sp.*) with a 3.8% and the common pandora (*Pagellus erythrinus*) with 2.3%. Of these, the axillary seabream, the horse mackerel and the common pandora must be landed under the new EU discard ban.

2.4.2 Modelling trawl fishery discards

The analysis of trawl discards included the total catch of each fishing haul, the mean bathymetry of the haul, two temporal variables, a geostatistical term and a vessel effect as possible predictors (Table 2.2). Based on the work by Rochet and Trenkel (2005), who found that discard proportions are not fully proportional to the catch, the total catch of each haul $C = (c_1, \ldots, c_n)$ was introduced as a linear effect with vague normal prior distributions as implemented by default in `R-INLA`. The vessel effect was assigned a random noise effect as in (Pennino et al., 2014) to absorb the variability on discard ratios due to different skipper's decisions and similars.

The exploratory analysis revealed non-linear relationships between depth and the discard proportion, so a second order random walk (RW2) latent model was applied based on constant depth increments d_j. These RW2 models, which perform as Bayesian smoothing splines (Fahrmeir and Lang, 2001), can be expressed as a computationally efficient GMRF (Rue and Held, 2005), and are therefore applicable in INLA. The smoothing of the bathymetric effect was selected visually by subsequently changing its prior distribution while models were scaled to have a generalized variance equal to one (Sørbye and Rue, 2014). The temporal effects also had RW2 smooth effects fitted but with cyclic indexations where, for example, in the ordinary day variable 1st of January comes after the 31st of December and so on. Finally, a geostatistical latent model $\boldsymbol{W}$ was introduced to identify fine-scale hot spots.

Therefore, assuming that the discard proportion Y_{st} at location s and time t follows a beta distribution and including all the effects summarised in

Table 2.2, the final model can be expressed as

$$\begin{aligned}
Y_{st} &\sim \text{Be}(\mu_{ij}, \phi_{ij}),\ s = 1, \ldots, S \quad t = 1, \ldots, T \\
\text{logit}(\mu_{st}) &= \beta_c c_s + d_j + e_t + W_s \\
\beta_c &\sim N(0, 0.001) \\
\Delta^2 d_j &= d_j - 2d_{j+1} + d_{j+2} \sim N(0, \rho_D), \quad j = 1, \ldots, J \quad (2.3) \\
\log\rho_D &\sim LogGamma(0.5, 0.00005) \\
\Delta^2 y_t &= y_t - 2y_{t+1} + y_{t+2} \sim N(0, \rho_D) \\
\log\rho_k &\sim LogGamma(1, 0.00005) \\
\boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\
(2\log\kappa, \log\tau) &\sim MN(\boldsymbol{\mu}_w, \boldsymbol{\rho}_w)
\end{aligned}$$

where the mean of discard proportions enters the model through the logit link, i indexes the location of each haul, j indexes different depths (d_j, representing the different values of bathimetry starting at $d_1 = 40$ metres till $d_{m=30} = 720$ metres) and t has a cyclic indexation ($T + 1 = t$) for either the moon phase or the ordinal day (e) of the haul.

This code allows us fit a second order random walk effects in `R-INLA`

```
# cluster covariate (haul.depth) into "n" groups
d <- inla.group(haul.depth,n=n)

# Bayesian spline in formula environment
Y ~ ... + f(d,model="rw2",prior=prior.depth,cyclic=FALSE/TRUE)
```

where the `inla.group()` function allows us to cluster the covariate of interest in n number of groups. These clustering values will perform as the second order transition nodes, denoted as d_j in equation (2.3). The terms of the form `f(...)` in the `R-INLA` formula environment account for the $f_j()$ terms in expression (2.2)and the `cyclic=FALSE/TRUE` statement will allow us assign cyclic or non-cyclic indexations to the fitted splines.

As we saw in Chapter 1, the two dimensional geostatistical latent model $\boldsymbol{W}$, introduced to identify fine-scale hot-spots, depends on two hyperparam-

eters κ and τ that define the variance and the range of the spatial effect. Specifically, and with the smoothing parameter of the Matérn (1.15) fixed ($\nu = 1$), the range of the spatial terms is approximately $\sqrt{8}/\kappa$ and the variance $1/(4\pi\kappa^2\tau^2)$. The priors for κ and τ are specified over the $\log\tau$ and $2\log\kappa$. In this case, default `R-INLA` prior distributions were used, where μ_κ is specified so that the range of the field is 20% of the longest distance in the field and μ_τ is chosen so that the mean variance of the field is one. The rest of the prior distributions in use are described in (2.3).

Variable	Description	Unit	Effect
Total catch	Total catch of the haul	Kilograms	Linear
Location	Geolocation	UTM	Geostatistical
Depth	Mean depth of the haul	Meters	Non-linear effect
Moon phase	Moon phase of the haul	Cyclic ordinal [0,100]	Non-linear cyclic effect
Ordinal day	Day of the year	Cyclic ordinal [1,365]	Non-linear cyclic effect
Vessel	Sampled vessel ID	-	Random noise effect

Table 2.2. List of covariates included in the analysis and the effect assigned to them. In the moon phase variable, values of 0/100 represent full moon and 50 new moon. Similarly, 1 represents 1 January and 365 represents 31 December in the ordinal day variable.

2.4.3 Results

Final models for both response variables included a non-liner bathymetric effect and the total catch of the haul as explanatory variables (Table 2.3). Specifically, the total catch of the haul had a positive effect on the expected ratios of both the total discard (posterior mean = 0.0023; 95% CI = [0.0018,0.0029]) and the regulated discard ratio (posterior mean = 0.038; 95% CI = [0.0027, 0.0049]).

Predicted total discard ratios showed a marked relationship with respect to bathymetry (Figure 2.3). Highest total discard ratios were observed in

		Total discard		Regulated discard	
		WAIC	LCPO	WAIC	LCPO
Temporal trend	I + M	-66.86	-0.085	-1109.2	-1.418
	I + OD	-67.25	-0.086	-1109.3	-1.419
	I	-67.32	-0.086	-1110.4	-1.420
Final model selection	I + D + V + C + S	-	-	-1201.0	-1.534
	I + D + C + S	-	-	**-1201.6**	**-1.535**
	I + D + V + C	-273.9	-0.354	-1170.9	1.504
	I + D + C	**-274.1**	**-0.355**	-1165.3	-1.493
	I + D	-201.3	-0.268	-1117.3	-1.422
	I + C	-189.2	-0.243	-1145.7	-1.470
	I + S	-183.1	-0.23	-1164.9	-1.479

Table 2.3. Model comparison for the total discard and regulated discard proportions. Missing values represent a bad fit of the spatial latent models, whose variance converged to nearly zero. Lower WAIC and LCPO scores represent a better compromise between fit, parsimony and predictive quality of the models. I = intercept, D = depth, V = vessel, M = moon phase, OD = ordinal day, C = total catch and S = spatial effect.

shallow waters, between 40 and 200 m. Regulated species discards, however, showed a maximum expected discard ratio in the 75-175 depth strata, while remaining relatively low in shallower and deeper waters.

No relevant temporal patterns were found in the study area. Indeed, all models with temporal effects, showed higher WAIC and LCPO scores than those without them for both response variables (Table 2.3). Similarly, the

model selection process dismissed the vessel random effect from both models, suggesting that discard ratios were fairly homogeneous across the different commercial vessels (Table 2.3).

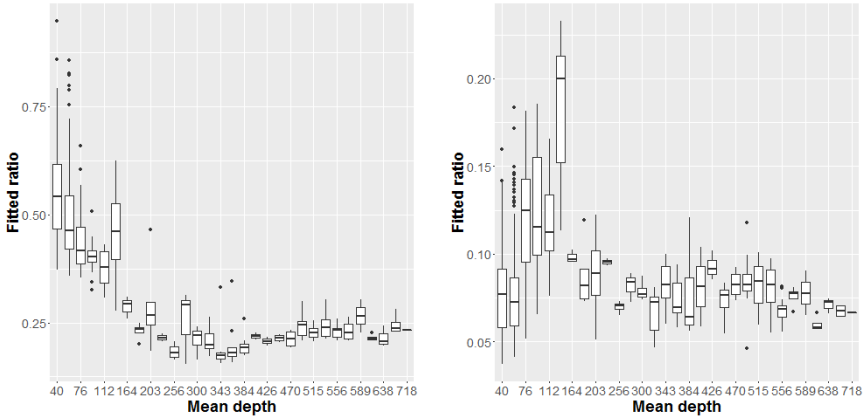
The spatial effect was only included in the regulated discard ratios model (Table 2.3). The estimated mean range was 0.53 (CI = [0.51,0.58]) and mean variance of 1.35 (CI =[1.14,1.77]). Figure 2.4 displays the posterior mean and standard deviation of the spatial component. This component showed two main low discard areas (negative values in Figure 2.4), which translates into lower expected discard ratios than those expected by the rest of variables. Specifically, one low discard area is located in the shallow waters in front of the Mar Menor lagoon and another along the central part of the 0.3 west meridian (Figure 2.4). These two low discard areas could constitute two fishing-suitable areas where expected levels of unwanted regulated species are lower than in other zones of the study area with similar bathymetric conditions. Similarly, a high discard hot-spot (positive values in Figure 2.4) was identified around the latitude 37.7 north and longitude 0.7 west coordinate area.

The total discard ratio predictive map (Figure 2.5) confirmed the key role of depth in the distribution of discard ratios. The posterior predictive map of regulated discard ratios (Figure 2.5) showed a similar pattern but with the added small-scale spatial variability provided by the spatial effect.

The predicted hot-spot of regulated discard ratios in the northern coastal zone of the study area (Figure 2.5) is driven by the marginal bathymetric effect due to the absence of observations in the area. Therefore, this discard hot-spot should not be considered while new observations suggest the contrary. Such uncertainty is displayed by the standard deviation map associated to these predictions (Figure 2.5).

2.4.4 Discussion

The present study proposes a new framework to characterise fishing-suitable areas under the upcoming EU discard ban. This study proposes using spatial beta regression models applied to discard or by-catch ratios. Specifically, we



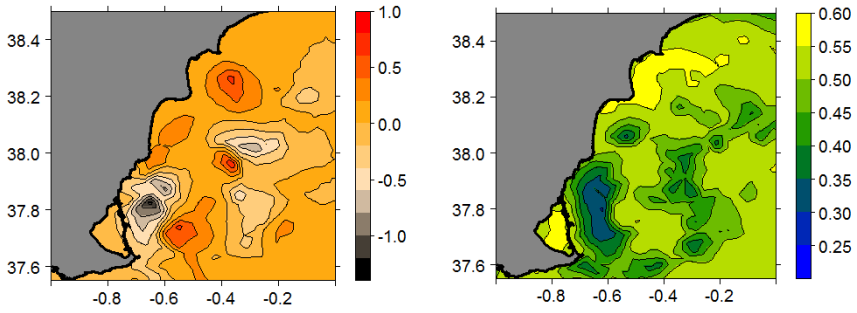
(a) Total discard ratios (b) Regulated discard ratios

Figure 2.3. Fitted discard ratios with respect to the mean depth of the observed hauls.

use total discard ratios and discard ratios of regulated species as a proxy to assess the global ecological impact and economic impact on fishers respectively.

The use of discard ratios is also a good alternative to the widely used discards per unit effort (DPUE) criterion. In contrast to DPUEs, discard ratios represent benefit versus loss, and thus allow researchers to assess whether the amount of discards is disproportionate to the catch or not. Discard ratios allow assessing the economic impact by quantifying the percentage of quota loss (if applicable) and percentage of storage room occupied in the vessel by non-marketable fish when fishing in a given sub-area. Similarly, the ecological impact can be quantified in terms of gained food biomass against wasted biomass either by looking at total discard ratios or regulated species discard ratios as a proxy to the amount of biomass removed from the system that before the landing obligation would be returned to the system.

From a methodological perspective, results showed that discard ratios have



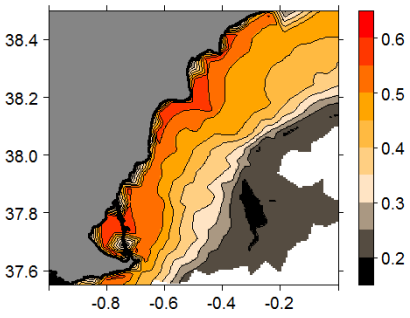
(a) Mean (b) Standard deviation

Figure 2.4. Posterior mean and standard deviation of the spatial component of the regulated species discard ratio.

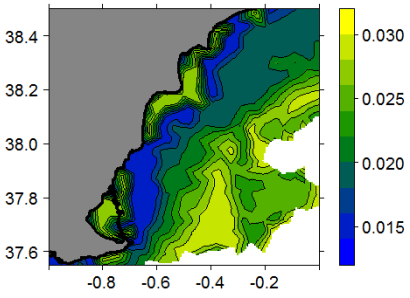
a good standardising capability across different vessels. The random effect assigned to absorb extra variability among vessels was dismissed during the model selection process. Conversely, the study by Pennino et al. (2014), using DPUEs in the same study area, found that this component was relevant in the analysis. Our results using discard ratios compared to the results in Pennino et al. (2014) could provide initial evidence of the good standardizing capacity of discard ratios compared to the more usual DPUE units.

The resulting discard ratio predictive maps (Figure 2.5) provide intuitive tools to assess the fishing suitability of a sub-area. Fishers and policy makers, could combine information on the proportion of total and regulated species discards to select economically and ecologically fishing-suitable areas. In this regard, the Bayesian approach provides an added value, which is the straightforward quantification of the uncertainty in our predictions, visualised here with the standard deviation maps.

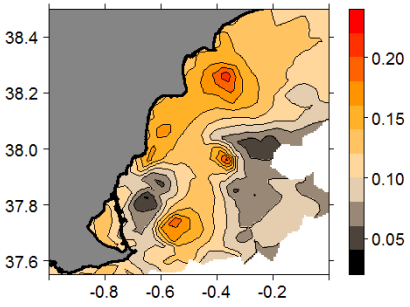
The marginal spatial effect also provides a very informative tool for de-



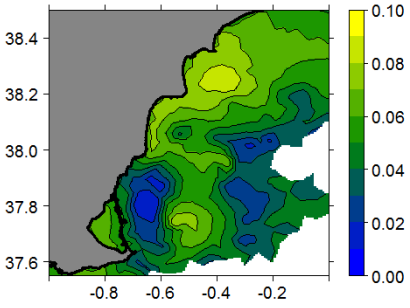
(a) Mean of the total discard ratio



(b) Standard deviation of the total discard ratio



(c) Mean of the regulated discard ratio



(d) Standard deviation of the regulated discard ratio

Figure 2.5. Posterior predictive mean and standard deviation of the total discard ratios (top) and the regulated discard ratios (bottom).

cision making since it represents the spatial fine-scale variability of discard ratios given the effect of the covariates. In other words, the hot-spots of the spatial effect identify areas where more discards are expected as compared to other areas with similar environmental conditions. As a consequence, a marine spatial planning framework could consider these areas for protection so that discarded/wasted biomass is minimised. Inversely, and following the same reasoning, cold-spots in the spatial effect characterise fishing-suitable areas due to low expected discard biomass. Consequently, the map of the spatial effect is particularly useful for policy makers to design an effective marine spatial plan.

This study identified two main fishing-suitable sub-areas based on the proportion of discarded regulated fish (Figure 2.4). Fishing in these sub-areas could reduce fishers' economic loss due to quota reduction (if applicable) or the minimization of ship hold occupied by non-marketable species. Furthermore, fishing in these sub-areas may minimise the ecological biomass loss generated by the landing obligation. Under the landing obligation it is mandatory to land some of the previously discarded biomass, which results in higher energy removals from the system than before. Regarding the total proportion of unwanted fish, results showed a clear longitudinal gradient related to the bathymetry. Indeed discard ratios were higher in shallow waters (Figure2.3) along the coastline (Figure 2.5) and may reflect the distribution of target species of these métiers. As also highlighted by Pennino et al. (2014), the depth-related variations of discard ratios are linked to differences in species composition of fish communities and in the length-frequency distribution of some species. Species replace each other according to their bathymetric and geographical preferences. Thus, the bogue, the most discarded species, is particularly abundant between 50 and 200 m, which may explain the increase of total discard ratios in shallow waters.

Interestingly, and although fish distribution is known to vary seasonally, none of the models found any temporal trend on the discard ratios of the study area. Discard ratios, as well as the DPUE criterion, constitute the aggregation of many different species and thus mixed species-specific distribution patterns.

In this respect, a detailed species-specific study of discard could better identify masked temporal and/or spatial patterns in this study.

Lastly, results confirmed that discard ratios consistently increase with the amount of total catch, as shown previously by Pennino et al. (2014) and Rochet and Trenkel (2005). In this regard, Rochet and Trenkel (2005) proposed limited hold capacity of the vessels as a possible explanation for the increased discard ratios when the catch is high. This could not be the case in this area since the local fleet operates on the basis of day-trips and the total catch seldom exceeds hold capacity. An alternative reason may be related to high grading, where mid-priced fish species could be landed when the catch is low to make the trip profitable but thrown away when the catch is good enough. A more detailed study of these mid-priced fish species combined with sales notes information could confirm this hypothesis.

2.5 Conclusions

In this chapter, we used a Bayesian hierarchical spatial beta model to analyse spatio-temporally sampled proportion data. To this end, we used a simple reparametrisation of the beta distribution to apply regression on the mean of the process. The Bayesian approach allows a straightforward quantification of uncertainty, which is important for decision making, while the hierarchical structure allows a more natural model specification, especially when including complex latent models such as geostatistical terms.

Beta regression overcomes all the drawbacks of the traditional data transformations (Warton and Hui, 2011; Ferrari and Cribari-Neto, 2004). First, it allows a direct interpretation of model parameters in terms of the original data; second, the analysis is not sensitive to the sample size; and lastly, posterior distributions are expected to concentrate well within the bounded range of proportions. It is only when observations on the extremes of the distribution are present, i.e. 0 and 1, that the beta distribution does not provide a satisfactory description of the data. A possible solution to this problem is to add some small value to the proportion, which introduces minimal bias

while still satisfying the above criteria (Warton and Hui, 2011); otherwise, zero and/or one inflated models may be required (Ospina and Ferrari, 2012), now available in the `zoib` package (Liu and Kong, 2015) for R.

The incorporation of spatial random effects in beta regression models can be very useful in a wide range of disciplines. For example mapping plant coverage in ecology; mapping budget allocation in econometrics; mapping the percentage of retirees in sociology, mapping sex-ratios in species, etc. Furthermore, combining the Bayesian spatial hierarchical modelling approach (Banerjee et al., 2014) and the temporal extension of Da-Silva and Migon (2016), the beta regression framework can be extended to the spatio-temporal domain. Consequently, it is possible to tackle problems such as the evolution of plant epidemics (Stein et al., 1994), the spatio-temporal evolution of temperature (Hengl et al., 2012) or the understanding of the spatial dynamism of species over time, as in Paradinas et al. (2015). It must be taken into account that the computational burden of these models can be even more demanding than in the purely spatial domain, making `R-INLA` and its SPDE module two almost necessary tools to deal with them.

The Bayesian hierarchical modelling approach is used due to its flexibility in model formulation, which makes it suitable to deal with complex ecological problems (Clark, 2005). Furthermore, the Bayesian formulation of posterior parameters is particularly straightforward to implement as it relies on quite direct probability rules (Clark, 2005). Similarly, the Bayesian approach is especially appealing for management purposes since it improves quantification of uncertainty as compared to the classical approach. Under the Bayesian approach, the quantification of uncertainty in model predictions is incorporated though the uncertainty associated to the estimated parameters, whereas in the classical application, parameters are considered to be known.

The case study presented here applies spatial beta regression to identify fishery discard hot-spots based on discard proportions, which, as opposed to total discard units, assess the biomass benefit against the amount of wasted biomass that constitute discards. Our results have identified at least one high discard proportion hot-spot in the study area. Under a marine spatial

planning framework that seeks to minimise the ecological impact of the fishing activity, the characterisation of hot-spots could be specially useful for policy makers, as it would allow them to protect those hot-spots as areas of special interest.

To conclude, we would like to mention that the geostatistical beta regression approach proposed here to analyse proportions is not only applicable to non-binomial proportional data but also to binomial proportional data, i.e. proportions measured as *x out of n*. In fact, applying beta regression in these cases may be an easier and more natural approach to avoid the usual problem of overdispersion in logistic regression than that proposed in Wilson and Hardy (2002) using GLMMs.

The case study for identifying fishing-suitable areas with regards to fishery discards has been accepted for publication in the ICES journal of marine science and selected as Editor's Choice (http://www.oxfordjournals.org/our_journals/icesjms/editorial_board.html***). Similarly, the spatial beta regression approach for modelling ecological proportions used in this section has been accepted for publication in Revstat Statistical Journal.***

Chapter 3

Modelling preferentially sampled fish distribution

Another key reason for the collection of fishery dependent data is to assess the state of our stocks, i.e. targeted species. In the spatial framework, using fishery dependent data to model the distribution of target species is problematic because sampling locations are deliberately chosen by the skipper to maximize the catch of target species. As a consequence, fishery dependent data lack observations in areas out of the optimum ecological niche of the target species, resulting in preferentially biased samples for modelling their spatial distribution. This property of the data is known in the literature as the preferential sampling problem (Diggle et al., 2010).

The preferential sampling issue was not relevant in the previous Chapter (2), where we modelled fishery discards using fishery dependent data, because discards are a residual process of the fishery, which means that discards are not targeted by the fishers. However, when our interest falls on the prediction of targeted stocks distribution using fishery dependent data, then, the preferential sampling is indeed a problem. Fishers go fishing where they expect to catch highest volumes of economically valuable fish, therefore our sample has

no observations at environmentally unfavourable locations.

Typical geostatistical models, as that of the previous chapter, assume that sampling locations are non-informative, which means that the sampling process and the process being modelled are stochastically independent. When using preferentially sampled data however, this assumption is no longer correct and standard geostatistical methods will potentially lead to biased results.

3.1 Modelling fish distribution under preferential sampling

As we already mentioned, predicting target species distribution using preferentially sampled data requires full awareness of the scientist. Basically, under these circumstances, our sampling distribution is not random any more and thus basic model assumptions are violated. In order to overcome this problem we need to correct for the fact that we have sampled where we expect to catch more.

A sensible approach to do so may be to make use of fishers knowledge on the distribution of target species. The basis of this approach is to assume that fishers know the preferred habitats of these species and fish in those locations. In other words, we believe that fishers have a good idea of the underlying spatial field that we want to predict (distribution of target species) and choose their best fishing locations accordingly.

As Rue et al. (2010) indicate in the discussion on Diggle et al. (2010)'s paper, preferential sampling may be accounted for in a marked point process model, in particular a log-Gaussian Cox process (LGCPs), which is an example of a latent Gaussian model. As a result, it is plausible to perform Bayesian inference based on INLA.

Among all spatial point processes, the class of Cox processes embrace non-homogeneous Poisson processes that arise from a random spatial field Λ. These models provide a statistically tractable class of models for aggregated point patterns in which its spatial distribution can be associated to unknown

conditions such as environmental variables, etc.

LGCPs are a particular class of Cox processes in which the logarithm of the intensity surface is a Gaussian field. More formally:

$$\log(\Lambda(s)) = V(s), \tag{3.1}$$

where $V(s)$ is a Gaussian random field that is conditional to the unknown conditions. In other words, given the random field, our observations are independent and hence they form a non-homogeneous Poisson process.

In the case of fisheries preferential sampling, the amount of fish collected in each location $\mathbf{Y} = (y_1, \ldots, y_n)$ represents the intensity of the underlying spatial field in the sampled locations. However, as fishers aim to fish as much as possible, very few samples, if any, are available in those locations where the intensity is low, thus kriging interpolation will be positively biased. In this regard, a LGCP model applied over the sampling locations $\mathbf{s} = (s_1, \ldots, s_n)$ allows us to identify low fishing pressure areas, potentially due to the amount of fish in those locations being low according to fishers knowledge. These low fishing pressure areas, assumed to be low fish abundance areas according to fishers knowledge, could be used to correct the absence of low abundance observations in the kriging interpolation commented before. To incorporate such information in the abundance model it is necessary to use joint modelling techniques, which allow fitting shared model components between the two linear predictors, the one for the observed abundances and the one for the point process distribution.

So, in summary, sharing information between the fishers knowledge (selection of fishing locations) and the actual abundances of the spatial field could help us improve posterior abundance estimates, specially in low abundance areas. Therefore, on the one hand of a preferentially sampled model, we fit a spatial point pattern $\mathbf{s}$ representing observers approximated knowledge about the underlying Gaussian random field $V(s)$ that forms a LGCP model with an intensity function $\Lambda(s)$ of the form:

$$\Lambda(s) = \exp\{\beta_{0_\Lambda} + W_s\}, \tag{3.2}$$

where β_{0_Λ} is the intercept of the LGCP and W_s is the spatial effect of the model. Note that covariates and other types of correlations, not included here for the sake of simplicity, could also be included in the additive predictor.

On the other hand, caught fish biomass $\mathbf{Y}$ represents the intensity of the underlying species abundance at sampling locations, which are assumed to follow an exponential family distribution such as a gamma distribution with parameters μ and σ^2.

$$log(\mu_s) = \beta_{0_m} + \alpha W_s \tag{3.3}$$

where the mean abundance is entered using a logarithmic link. Again β_{0_m} is the intercept of the model and W_s is the spatial correlation term of the model that is shared with the LGCP but scaled by α to allow for the differences in scale between the abundances and the LGCP intensities. As well as in the LGCP model, other covariates and correlation terms could be included as well.

This way, the resulting spatial field of the preferentially sampled abundances is corrected by incorporating information about the distribution of the point pattern (that indirectly quantify fishermen's knowledge).

3.2 Simulation study

To illustrate the preferential sampling method proposed by Rue et al. (2010) based on Diggle et al. (2010)'s paper, we first apply this approach in a simulated scenario to prove the effectiveness of the approach.

A simulated Gaussian random field with Mátern covariance function was created over a 100 by 100 grid using the `RandomFields` package (Schlather et al., 2015). Over this spatial field 100 locations were selected as sampling locations. In order to mimic the preferential sampling scenario, these locations were selected based on the sum to one standardisation of exponential abundances in each location using the function *sample()* in R (R Core Team, 2016):

As a result, we got a simulated Gaussian random field (Figure 3.1) which values range between 0 and 10.3. Locations were selected by applying an exponential transformation over the intensities of the simulated field to create the preferentially selected sample, i.e. the point-pattern process in (3.2). Then, the abundances of the process were extracted from the simulated Gaussian field (Figure 3.1) based on the locations of the point-pattern.

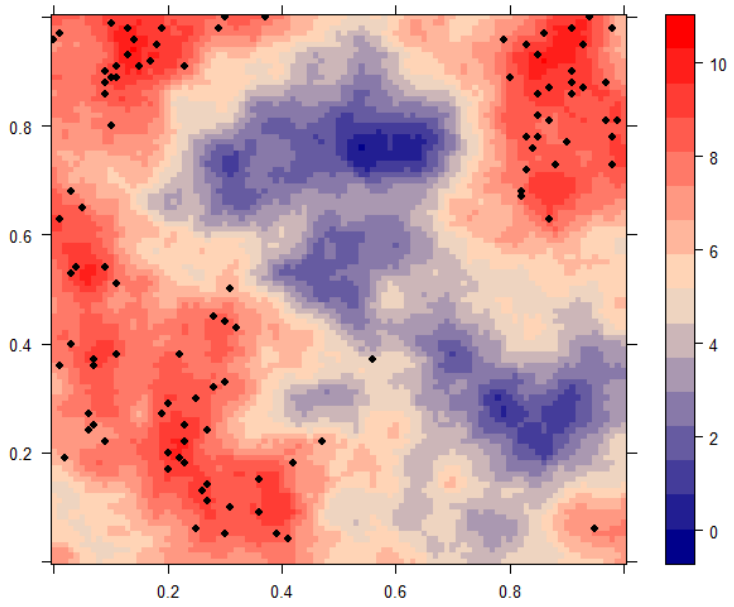


Figure 3.1. The simulated Gaussian field and point pattern that represent the sampling locations of the process under study.

In order to test for the applicability of the preferential sampling approach, we compared the following two models:

1. The non-preferential model of the abundances

$$\begin{aligned} Y_s &\sim Ga(\mu_s, \rho), \quad s = 1, \ldots, S \\ log(\mu_s) &= \beta_0 + W_s \\ \boldsymbol{W} &\sim N(0, \boldsymbol{Q}(\kappa, \tau)) \\ (2\log\kappa, \log\tau) &\sim MN(\boldsymbol{\mu}_w, \boldsymbol{\rho}_w) \\ \beta_0 &\sim N(0, 0) \end{aligned} \tag{3.4}$$

   where abundances are modelled through a geostatistical term (W_s).

2. The preferential sampling model

$$\begin{aligned} \Lambda(s) &\sim Po(\lambda_s), \quad s = 1, \ldots, S \\ Y_s &\sim Ga(\mu_s, \rho) \\ log(\lambda_s) &= \exp\{\beta_{0_\Lambda} + W_s\}, \\ log(\mu_s) &= \beta_{0_m} + \alpha W_s \\ \boldsymbol{W} &\sim N(0, \boldsymbol{Q}(\kappa, \tau)) \\ (2\log\kappa, \log\tau) &\sim MN(\boldsymbol{\mu}_w, \boldsymbol{\rho}_w) \\ \boldsymbol{\beta_0} &\sim N(0, 0) \end{aligned} \tag{3.5}$$

   where the common geostatistical term (W_s) is corrected by combining information from both the LGCP and the abundance process scaled by α in the second predictor to allow for differences in scale.

Within `R-INLA` the formula environment syntax required to share information from both processes to fit a common spatial field in both predictors is the following:

```
# Fitting a shared spatial effect
formula <- ... + f(point_pattern_spat, model=spde) +
        f(abund_spatial, copy="point_pattern_spat",
        fixed = FALSE)
```

where, and despite the `copy` syntax, geostatistical terms are fitted jointly and scaled in the abundance model by an scalar (α in equation 3.5). If the scale of

the predictors were the same, we could set this by passing the `fixed = TRUE` statement instead.

The models discussed in equations (3.4 and 3.5) were fitted and compared in terms of the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) as a measure of goodness-of-fit, the Log-Conditional Predictive Ordinates (LCPO)(Roos and Held, 2011) as a leave-one-out predictive score and the predictive Mean Absolute Error (MAE) (Willmott and Matsuura, 2005) as a measure of the overall out-of-sample predictive score.

	Model	MAE	DIC	LCPO
1	Non-Preferential	1.92	90.58	0.66
2	Preferential sampling	0.96	78.50	0.56

Table 3.1. Model comparison based on the Deviance Information Criterion (DIC), the Log-Conditional Predictive Ordinate (LCPO) and the predictive Mean Absolute Error (MAE). In all cases, smaller scores represent better fit.

The best predictive model was the model with the preferential sampling correction, which had better DIC and LCPO scores (Table 3.1), but most importantly the out-of-sample MAE score was very much improved in the preferential model (Table 3.1) as compared to the model without preferential correction. Similarly, correlation between observed and predicted values was about 0.93 in the preferential model and 0.72 in the model without the preferential correction. Essentially, neither of the models was able to make good predictions at low abundance locations (Figure 3.2) because there were no observations made there but the non-preferential model performed significantly worse.

Figure 3.3 shows the posterior predictive mean of the simulated abundance process without (a) and with (b) the preferential sampling correction. As already mentioned before, and although both models show a similar predictive spatial patterns, the preferentially corrected model predicted better at low

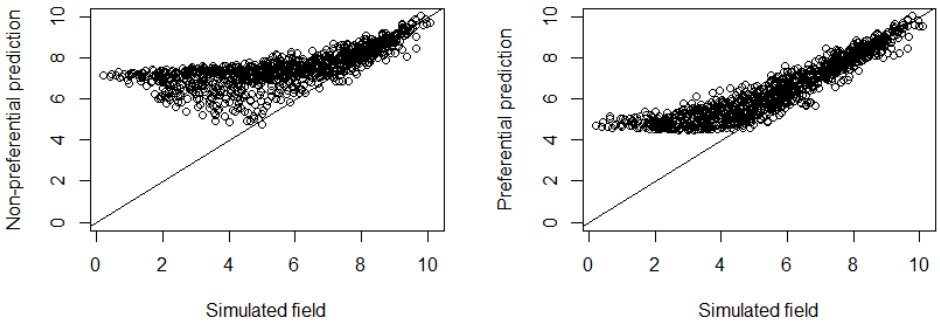
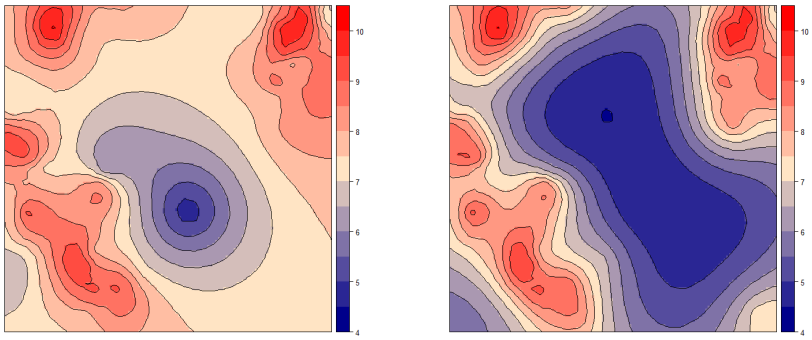


Figure 3.2. Simulated abundance against predicted abundance in the non-preferential model (left) and in the model with the preferential correction (right). The non-preferential model predicts worse than the preferential model at low abundance areas.

abundance areas.

3.3 Case study: modelling red shrimp abundance using red shrimp fishery data

In this section we present a practical application of the preferential approach in a real case scenario. Specifically, within the fishery context, a red shrimp fishery (*Aristeus antennatus*) was used. The red shrimp is a very important stock that is fished exclusively by trawlers and is distributed in between the 300 and 900 metres of depth (Gorelli et al., 2014). The price of the red shrimp at the market reach peaks of $200/kg$ during particular periods such as Christmas or summer holidays. Therefore this is a very important fishery in the Spanish Mediterranean that is named as "bottom otter trawl for deep-water species" métier (OTB-DWS) under EC rules (European Comission,



(a) Non-preferential sampling model (b) Preferential sampling model

Figure 3.3. Posterior predictive mean maps of the simulated abundance process without (left) and with (right) the preferential sampling correction.

2009).

3.3.1 Red shrimp fishery data

Specifically, we analyzed data collected by onboard observers from 2009 to 2012 in the Gulf of Alicante (Spain). The data set includes 77 OTB-DWS hauls collected in 9 different trawlers (Figure 3.4) and has been provided by the *Instituto Español de Oceanografía* (IEO, Spanish Oceanographic Institute). The database provided contained information on the amount of red shrimp hauled in kilograms, the location of the haul and its bathymetry.

3.3.2 Modelling red shrimp distribution

Following the ideas from Diggle et al. (2010) and Rue et al. (2010) we applied the preferential sampling model to a red shrimp stock in the Western Mediterranean. The fitted effects for the abundance of red shrimp (Y) were corrected by jointly modelling the abundances model and the LGCP model, which represents fishers knowledge about the distribution of red shrimp (be-

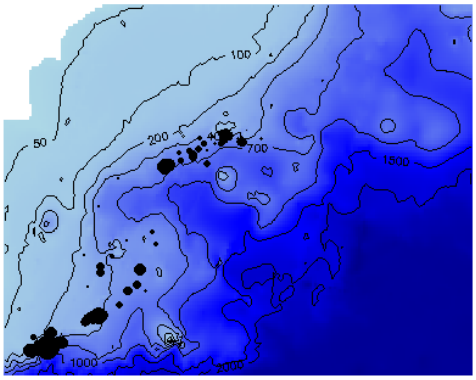


Figure 3.4. OTB-DWS on-board sampling locations and red shrimp (*Aristeus antennatus*) abundance in the Gulf of Alicante (Spanish Mediterranean).

cause fishermen fish where they expect to fish most). Therefore, assuming that the LGCP intensity follows a Poisson distribution and shrimp abundance a gamma distribution at locations s, the model looks like this:

$$
\begin{aligned}
\Lambda(s) &\sim Po(\lambda_s), \quad s = 1, \ldots, S \\
Y_s &\sim Ga(\mu_s, \rho) \\
log(\lambda_s) &= \beta_{0_\Lambda} + f(d) + W_s, \\
log(\mu_s) &= \beta_{0_Y} + \alpha_d f(d) + \alpha_w W_s \\
\boldsymbol{W} &\sim N(0, \boldsymbol{Q}(\kappa, \tau)) \\
\Delta d_j &= d_j - d_{j+1} \sim N(0, \rho_D), \quad j = 1, \ldots, m \\
\boldsymbol{\beta} &\sim N(0, 0) \\
(2 \log \kappa, \log \tau) &\sim MN(\mu, \rho) \\
\rho_D &\sim LogGam(2, 0.00001)
\end{aligned}
\tag{3.6}
$$

where s indexes the location of each haul and j indexes different depths (d_j, representing the different values of bathimetry observed in the study area starting at $d_1 = 90$ metres till $d_{m=40} = 920$ metres). Each predictor has its own intercept ($\beta_{0_\Lambda}, \beta_{0_Y}$) but bathymetric $f(d)$ and spatial effects W_s are shared in both predictors. The bathymetric effect was assigned a first order random walk (RW1) latent model based on constant depth increments, which perform as Bayesian smoothing splines (Fahrmeir and Lang, 2001), and W_s is a geostatistical term based on a Matérn covariance function, as described in section 1.6.3, that depends on the hyperparameters κ and τ. These hyperparameters are linked to the range and the total variance of the effect (see Section 1.6.3). Both, the bathymetric and the spatial effects were scaled by α_d and α_w respectively to allow for the differences in scale between red shrimp abundances and the LGCP intensities.

Every Bayesian model precise of prior distributions and as usual in Bayesian notation we work with precisions, not variances. In this case, we used an informative prior for the range of the spatial effect. Specifically we set the multivariate normal prior of the spatial field so that the median of the prior for the range is approximately half the maximum distance in the study area. This was selected because the study area is rather small, and thus the default prior mean range in R-INLA (20% of the study area) would have been, a priori, too small. To do so, we used the following code to assign the desired prior to the SPDE model:

```
# Desired median over prior of the range
range0 = max.size / 2
kappa0 = sqrt(8) / range0
tau0 = 1 / (sqrt(4 * pi) * kappa0 * sigma0)
spde = inla.spde2.matern(mesh,
                         B.tau = cbind(log(tau0),-1,+1),
                         B.kappa = cbind(log(kappa0),0,-1),
                         theta.prior.mean = c(0,0),
                         theta.prior.prec = c(0.1,1) )
```

where `max.size` stand for the maximum distance in our study area.

It is also important to note that in equation (3.6) we specify the full model, where both spatial and bathymetric effects are relevant to the distribution of red shrimp and are shared between the abundance and LGCP models. However, this may not necessarily be true, thus we fitted and compared all the possible combinations of independent and shared effects.

3.3.3 Results

We run all the possible models derived from (3.6), the most relevant results are presented in Table 3.2. While analysing the data we observed that both the bathymetric and the spatial terms of the LGCP accounted for approximately the same information. As a consequence, full models did not converge in the point-pattern process, which restricted the model comparison in Table 3.2 to correcting only one of the effects, either the bathymetric or the spatial effect.

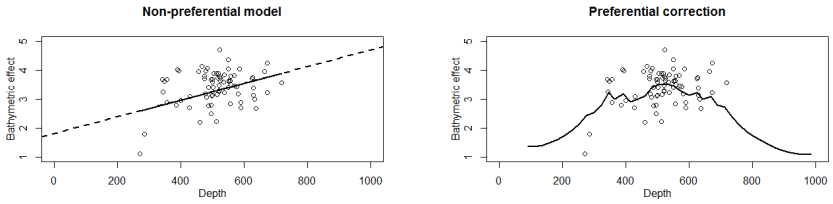
	Model	DIC	LCPO
1	Intc + Depth	673.49	4.38
2	Intc + Spatial	665.52	4.35
3	Intc + Depth + Spatial	657.76	4.29
4	Intc + **Depth**	674.77	4.40
5	Intc + **Spatial**	671.43	4.37
6	Intc + **Depth** + Spatial	661.06	4.34

Table 3.2. Model comparison for the abundance of the red shrimp (*Aristeus antennus*) based on DIC and LCPO scores. Intc = Intercept and **Bold** terms = shared components

Finally, we selected model 6 in Table 3.2 which included a shared bathymetric effect and an independent spatial effect for the abundances that absorb the spatial variability of the data given red shrimp's bathymetric preference.

As shown in Table 3.2, DIC and LCPO scores are slightly better for the non-preferential full model (model 3). However, as Figure 3.5 shows, and despite the spiky relationship of the preferential bathymetric effect (slightly overfitted by the RW1 latent model), the estimated effect for the bathymetry is far more natural than the linear bathymetric effect fitted in the non-preferential model. However, even if the preferential model improves the bathymetric effect, new observations at deeper waters could further improve this relationship as other studies on the distribution of red shrimp suggest (Gorelli et al., 2014). The non-preferentially corrected spatial effect (right panel in Figure 3.6) accounted for the residual spatial heterogeneity derived from the shared bathymetric effect in the abundance process.

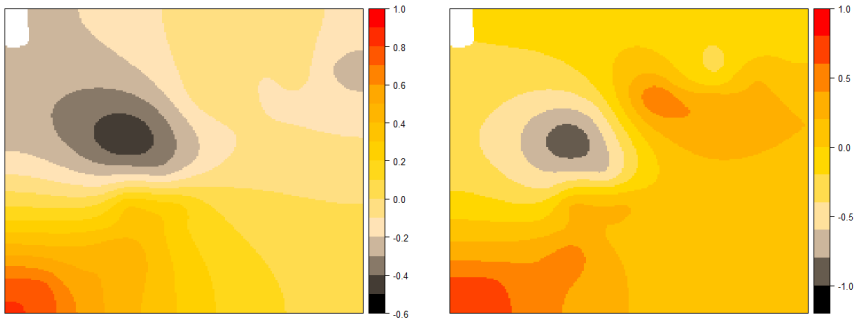
Observing a more natural relationship in the corrected fitted effects (preferential model) despite the worse model selection scores was not surprising. The reason for this is that within-sample (DIC) or similar (LCPO) scores are not able to properly measure the out-of-sample predictive capacity of the model. Therefore, selecting the best predictive model based on these model selection scores alone may be problematic because we may end up selecting overfitted models.



(a) Non-preferential model

(b) Preferential model

Figure 3.5. Bathymetric effect in the models without and with the preferential sampling correction. Dashed lines represent the extrapolated effect in the non-preferential linear effect.



(a) Non-preferential model (b) Preferential model

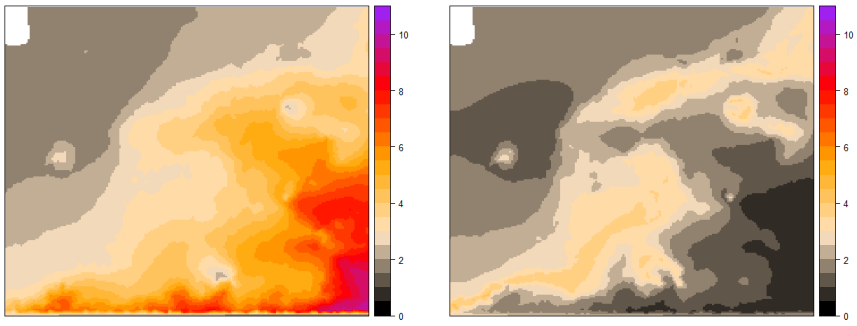
Figure 3.6. Maps of the mean of the posterior distribution of the spatial effect in the model without (left) and with (right) preferential sampling.

Figure 3.7 shows the posterior predictive mean of the red shrimp distribution without and with the preferential bathymetric correction. Each map shows a very different pattern, the non-preferential model is driven by the linear positive bathymetric effect (posterior mean = 0.003; 95% CI = [0.0009, 0.0049]) that extrapolates to very high abundances at high depths. The preferential model however, is able to correct this linearity of the data and provide a more natural bathymetric distribution of red shrimp.

3.4 Conclusions

In this section we have presented a modelling approach that could be very useful to assess the spatial distribution of fishery stocks using fishery dependent data.

The simulated case study demonstrates the extent at which a preferentially sampled processes can be corrected following Diggle et al. (2010)'s proposal. In general, high-abundance areas are predicted fairly well but low abundance



(a) Non-preferential model (b) Preferential model

Figure 3.7. Posterior predictive mean maps of the red shrimp (*Aristeus antennatus*) species, without and with the preferential sampling correction.

areas tend to be overestimated due to the low number of samples (if any) at unfavourable conditions.

The practical application on red shrimp, included here as a real world scenario, shows that the preferential sampling phenomena can be driven either by space or any other important covariate of the process, e.g. bathymetry in this case. Resulting predictive maps significantly improve the prediction of the target species when the model accounts for preferential sampling. Consequently, we conclude that this approach could suppose a major step forward in the understanding of target species meso-scale ecology given that most of the available data today are fisheries dependent data.

The real case application also raised a very important issue regarding the model selection process of preferentially sampled data. Conventional model selection scores, e.g. DIC and LCPO, are likely to prefer non-corrected abundance models as the fit of within-sample observations is not restricted to the LGCP model. This situation may result in better within-sample predictive capacity of the non-preferentially corrected models while the out-of-sample

predictive capacity of the model is compromised due to overfit.

In fisheries, the application of these approaches to other fishery dependent data collected under the EC regulation (European Comission, 2009) or any other on-board sampling scheme around the world may help us implement the EAFM framework at the meso-scale. The following step in modelling fishery dependent data should assess the spatio-temporal distributional patterns of the different fisheries. To do so however, the sparse spatio-temporal resolution of the data to date should be improved.

Chapter 4

Spatio-temporal structures with shared components for species distribution modelling

In the previous two chapters we have seen a couple of applications using fishery dependent data. Now, in this chapter our goal is to squeeze the capacities of fishery independent data to respond few ecological questions: which is the spatial distribution of fish? Is it persistent or does it change over time? If so, how? To test such hypothesises, along this chapter, we present a set of spatio-temporal extensions of the usual geostatistical model.

Fishery research surveys play a very important role in the management of our fisheries. Fishery survey data, or fishery independent data, usually cover very wide areas and provide a macroscopic view of the fishery. This is a very important feature for spatial management purposes because it allows us to quantify the importance of areas in the macro-scale, which is the scale at

which marine protected areas should be designed (Fenberg et al., 2012). An important drawback of fishery surveys is that they are typically repeated only once per year (e.g. MEDITS survey, Bertrand et al. (2002)) or twice per year at most (e.g. IBTS, The International Bottom Trawl Survey Working Group (2012)), so we must be very careful when drawing conclusions in this respect.

4.1 Assessing the temporal persistence of a spatial process.

As introduced in the first chapter, the ecosystem approach to fisheries management (EAFM) aims the protection of productive ecosystems based on the principle that healthy ecosystems produce more and will secure a sustainable exploitation of fishery resources. Therefore, understanding the spatial pattern of different life stages has attracted the main focus of attention. In fact, one of the fundamental objectives of an Ecosystem Approach to Fisheries Management (EAFM) is to reduce any adverse impact on recruitment habitats, primarily from fishing (Garofalo et al., 2011).

There is some controversy on how to define a nursery ground. Some definitions were proposed in the last decade (Beck et al., 2001; Dahlgren et al., 2006). These definitions generally rely on direct measurements of juvenile movement from nursery habitats to the adult population (Beck et al., 2001; Gillanders et al., 2003). Unfortunately, direct measurements become infeasible for deeper water species, whose nursery grounds tend to be located in deeper waters as well. European hake (*Merluccius merluccius*) is one of such species which recruits tend to inhabit the continental shelf and the upper slope at 80-250 m depths (Maynou et al., 2003; Recasens et al., 1998). As a result, when dealing with these kind of species an alternative is needed. A good one was introduced by Colloca et al. (2009), who suggested assessing the temporal persistence of abundance hot-spots as a proxy to identify nursery areas.

But, how can we assess the persistence of a spatial process? Over the years, visual assessment has been the way to go. However, and even if eye-

sight seems like a sensitive approach, the seek of statistical support to assess persistence motivated the work by Colloca et al. (2009). Their study fitted a different Bayesian geostatistical model for each time event and then applied aggregation curves using the posterior mean estimates of the models at the different prediction points to assess the persistence of the process. Unfortunately, their methodology is rather tedious to implement, does not consider the uncertainty associated to the estimates and does little emphasis on the quality of the model used to get the posterior predictive estimates, which play a crucial role in the results of this methodology.

Therefore, the purpose of this study here presented is to propose a method that assess the persistence of a spatial process by means of statistical inference rather than using post-analysis algorithms (Colloca et al., 2009). To do so, we propose comparing the goodness-of-fit of two different Bayesian hierarchical spatio-temporal models.

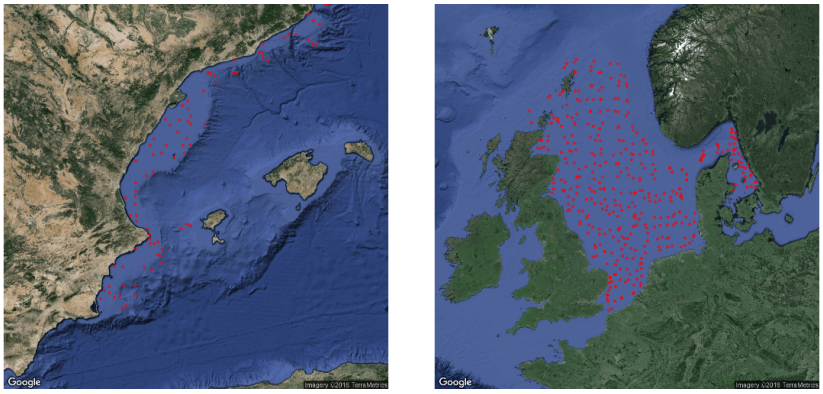
4.1.1 Data

In this study, two different datasets are used to test our approach for assessing the persistence of fish distribution (or any other spatial process).

4.1.1.1 Hake recruitment in the western Mediterranean Sea

On the one hand, we chose the European hake *(Merluccius merluccius)* because it constitutes one of the most important commercial species in the Mediterranean Sea, suffering from high fishing pressure and currently overexploited (Lleonart, 2005). In fact, in many Mediterranean countries there is still an important illegal market of small hake (Bellido et al., 2014). As a result, the juvenile fraction is particularly exposed, especially to trawl fishery after the bottom settlement stage, when they aggregate over nursery grounds.

Data on hake recruits were collected during the EU-funded *MEDIterranean Trawl Survey* (MEDITS) (Bertrand et al., 2002) project, carried out from spring to early summer (April to June) from 2000 to 2012. The MEDITS



(a) Western Mediterranean GSA 06 (b) North Sea

Figure 4.1. Sampling locations of MEDITS (left) and IBTS (right) surveys in the GSA 06 and North Sea respectively.

project uses a stratified sampling design based on depth (5 bathymetric strata: $10-50$, $51-100$, $101-200$, $201-500$ and $501-700$ m) and Geographical Sub-Area (GSA). Sampling stations were placed randomly within each stratum at the beginning of the project. In all subsequent years sampling was performed in similar locations. This study concerns the trawl-able grounds of GSA 06 (see left panel in Figure 4.1) which borders the northern Iberian Mediterranean coast. In total the dataset contains information on 1048 hauls that have been georeferenced in the centroid of each fishing operation.

Only hake recruits were considered, defined as those individuals less than 15 *cm* in total length. This length limit was selected using the *slicing* method (Lassen and Medley, 2001). A catch per unit effort (CPUE) response variable (Kg per 30 min tow) was created. As it is usual when dealing with biomass and analogous terms in other disciplines (e.g. rain volume), CPUEs showed a semi-continuous behaviour: 38% of zero observations, while if present, hake recruit abundance showed a right skewed distribution that ranges from 0.01 to 26.4 with its mean at 1.5 (Figure 4.2).

Potentially relevant environmental variables were included in this study. As mentioned before, bathymetry is a very important explanatory variable in the distribution of hake (Maynou et al., 2003; Recasens et al., 1998). We also included the type of substratum as a potentially relevant variable. Both variables were obtained as shapefiles from the IEO geoportal, accessible through the website of the Spanish Institute of Oceanography (http://www.ieo.es). The type of substratum shapefile include three levels: sand, mud and rock. As for the bathymetry, we also included a quadratic term in order to account for the bathymetric preference of hake recruits.

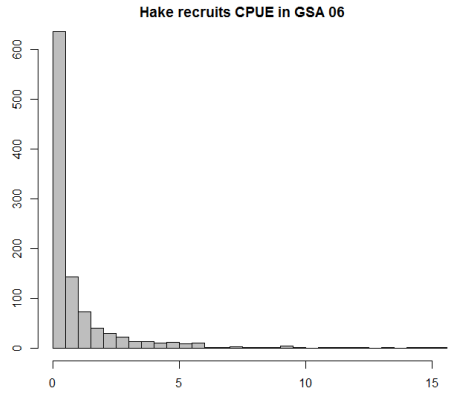


Figure 4.2. Histograms of observed CPUEs in hake recruitment between 2000 and 2012. Note that there is a 38% of zeros in the dataset.

4.1.1.2 Cod in the Northern Sea

On the other hand, we chose North Sea (NS) cod *(Gadus morhua)* because it constitutes the most important commercial species in Northern latitudes and it has suffered severe fishing pressure (Hutchings, 2000). Cod is known to be homogeneously distributed in the NS during winter, while it migrates north avoiding the warmer waters of the southern NS in summer (ICES, 2014). Therefore, we would expect our method to identify a non-persistent

distribution between quarters of the year.

Data on cod biomass were collected through the EU-funded The International Bottom Trawl Survey Working Group (2012) (IBTS) project. The IBTS in the North Sea is carried out two times per year, first in winter an then in summer (1st and 3rd Quarters). Sampling stations are more or less repeated every year and distributed among different countries that converge in the North Sea (Figure 4.1). The complete database is available on-line in `http://www.ices.dk/marine-data/data-portals/Pages/DATRAS.aspx`.

The cod dataset used here contains a total of 10865 hauls, which were collected twice a year (1st and 3rd Quarters) in between 2000 and 2014. This study considered cod biomass through a catch per unit effort (CPUE) response variable (Kg per 30 min tow). As with the hake, CPUEs showed a semi-continuous behaviour; 29% of the observations were zero, while if present, hake recruit abundance showed a right skewed distribution with mean 40.2 and median 11.2 (Figure 4.3). The dataset also contained information on depth and the location of the starting and finishing points of the haul. Hauls were finally georeferenced in the centroid of each fishing operation.

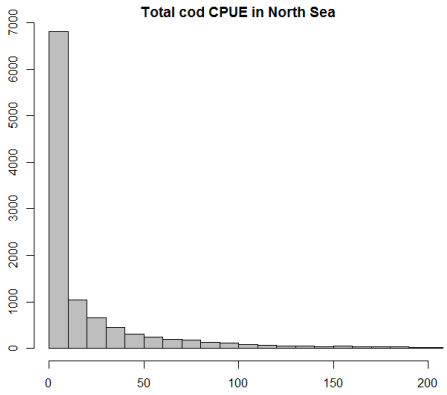


Figure 4.3. Histograms of observed CPUEs in NS cod between 2000 and 2014. Note that there is 29% of zeros in the dataset.

4.1.2 Modelling semi-continuous data

The most popular approach is to model semi-continuous data as two independent sub-processes, known as two-part models or Hurdle models. In these models, one sub-process determines whether the response is zero (using presence-absence data), while the second determines the intensity when the response is non-zero (Martin et al., 2005). Using this approach, one is able to estimate the probability of presence and if present, estimate the abundance (CPUE).

Let Y be the occurrence and Z the conditional-to-presence abundance (biomass) process at locations $\boldsymbol{s} = s_1, \quad , s_n$. Then Y_s and Z_s can be modelled as

$$\begin{aligned} Y_s &\sim \text{Ber}(\pi_s), \quad s = 1, ..., n \\ \text{logit}(\pi_s) &= \boldsymbol{X\beta^{(y)}} + W^{(y)} \\ Z_s|Y_s &\sim \text{Ga}(\mu_s, \rho), \quad s = 1, ..., n \\ \log(\mu_s) &= \boldsymbol{X\beta^{(z)}} + W^{(z)} \end{aligned} \tag{4.1}$$

where the probability of occurrence, π_s, is modelled through the usual logit link, and the mean abundance μ_s through its logarithm in location s. $\boldsymbol{X\beta}$ represents the fixed effects of the linear predictor and W represents the spatio-temporal structure of the data. Note that y and z supra-indices are used to point out that within the usual Hurdle model, both $\boldsymbol{\beta}$ and W are independent between sub-processes (e.g. bathymetric effect). Also note that we have chosen to work with a Gamma distribution in order to restrict abundance estimates to the positive real line, although the use of other distributions could be discussed.

4.1.3 Method to assess persistence of a spatial process

As we already introduced, the purpose of this study is to propose a method that allows us to assess the persistence of a given spatial process by means of statistical inference. For that, we propose two spatio-temporal decompositions of W, from now on W_{st} with locations denoted by s and time by t:

- The first structure consists of decomposing W_{st} into different spatial realizations at each time unit. This structure is a good proxy to those processes where the spatial structure vary considerably among different time units and unrelatedly among neighbouring times. In particular,

$$\begin{aligned} W_{st} &= W_{s_t} \\ \boldsymbol{W}_{s_t} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \end{aligned} \tag{4.2}$$

  where W_{st} is decomposed in a different spatial realization $\boldsymbol{W}_{s_t}$ at each time t. In this case, all $\boldsymbol{W}_{s_t}$s shared a common covariance function (same κ and τ, as in 1.15) to avoid having too many hyperparameters in the model. This structure is likely to favour the goodness-of-fit of temporally inconsistent spatial processes, i.e. when the distribution of fish vary substantially between time events.

- The other 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,

$$\begin{aligned} W_{st} &= W_s + V_t \\ \boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ V_t &\sim N(0, \sigma^2) \end{aligned} \tag{4.3}$$

  where W_{st} is decomposed in a common spatial realization W_s along with a random noise effect V_t that absorb different mean intensities at each time t. This structure may accommodate better those processes where the spatial structure is somewhat persistent in time.

4.1.4 Results

All models obtained by combining environmental variables with the different decompositions of the spatio-temporal structure were fitted and compared. In this case, model selection was based on the Deviance Information Criterion

(DIC) and the Log-Conditional Predictive Ordinates (LCPO) (see Section 1.7 for more information).

4.1.4.1 Hake recruitment results

All models including a quadratic term for bathymetry had better DIC and LCPO values than those including only a linear relationship. The type of substratum was discarded from the model because estimates of all level categories were centred on zero and had very high standard deviations. Table 4.1 shows a selection of the most representative models based on the DIC goodness-of-fit and LCPO predictive quality measures for the occurrence and conditional-to-presence abundance sub-processes.

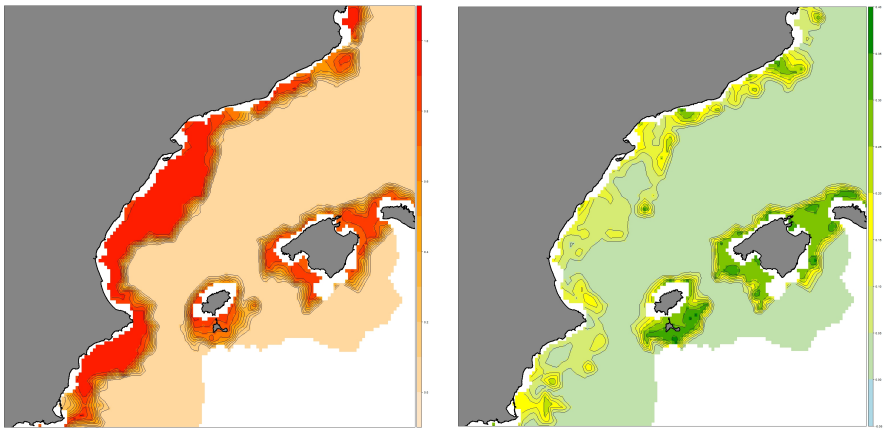
	Occurrence		Abundance	
Model	DIC	LCPO	DIC	LCPO
Only covariates	638.4	0.31	2854.7	1.88
Common spatial effect without covariates	518.4	0.40	2631.7	1.70
Covariates + common spatial + iid for year	493.9	0.23	2594.9	1.69
Covariates + yearly spatial effect	627.1	0.30	2707.3	1.80

Table 4.1. Model comparison for the hake occurrence and conditional-to-presence abundance models.

Following the principle of parsimony, the selected models for both occurrence and abundance were the models with the spatio-temporal decomposition in equation (4.6), which share a common spatial effect for all observations and a random noise effect for year in addition to the bathymetric effect. In other words, the selected models are those suggesting temporal persistence hake recruits spatial distribution.

Hake recruit occurrence

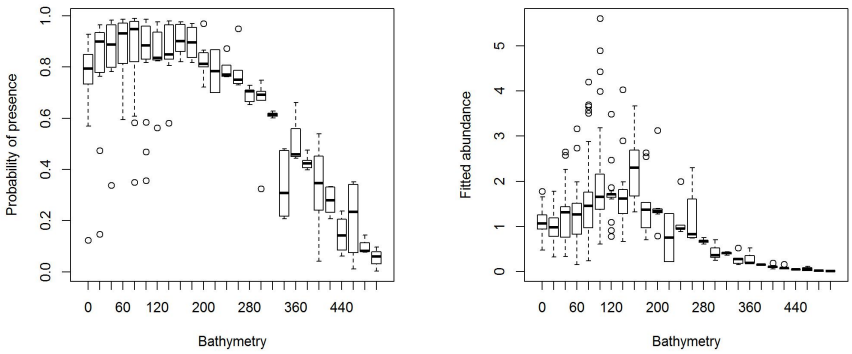
The selected model for the occurrence of recruits revealed the highest probability of presence along the continental shelf and the upper slope (Figure 4.4). Accordingly, hake recruitment showed an occurrence peak at between 40 and 180 m depth (Figure 4.5 left panel). However, the model also identified some low probability patterns along the continental shelf, especially off the Mar Menor, in the waters off Barcelona and the Palamós Canyon, in the northern corner of the study area 4.1.



(a) Mean (b) Standard deviation

Figure 4.4. Posterior mean (left) and standard deviation (right) for the hake occurrence probability.

The range of the spatial effect was estimated to be around 50 kilometres in the conditional-to-presence abundance model. The median variance of the unstructured temporal effect for year was three orders of magnitude smaller than that of the spatial variance (Figure 4.6). Yearly mean estimates of the unstructured random effect for year showed a possible pattern that



(a) occurrence

(b) abundance

Figure 4.5. Mean predicted values at different depths of the hake occurrence model (left) and the conditional-to-presence abundance (right). Each boxplot corresponds to a 20 meter interval.

may represent non-independence among neighbouring years mean intensities (Figure 4.7, right panel).

Hake recruit abundance

The highest abundance areas were also located along the continental shelf and upper slope (Figure 4.8), coinciding with the estimated effect of the bathymetry. The bathymetric peak abundance was around the 80 to 180 m strata derived from the predicted abundance estimates in Figure 4.5 (right panel). As opposed to the occurrence probabilities, abundance hotspots were much more localised. In fact, the sizes of these areas were around 10 km in diameter (very appropriate for protection purposes).

The range of the spatial effect was estimated to be around 35 kilometres in the conditional-to-presence abundance model. The median variance of the

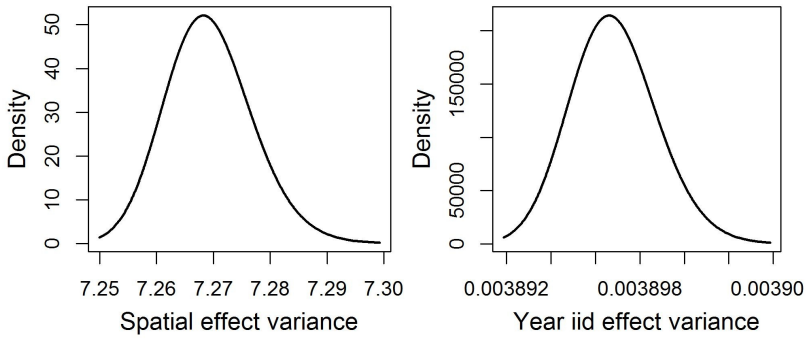
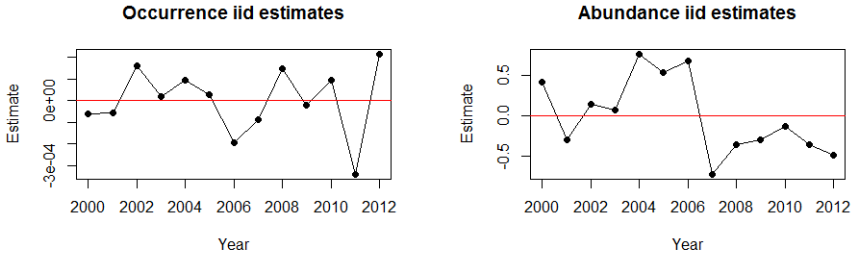


Figure 4.6. Estimated distribution of the variance for the spatial effect (left) and independent random effect for year (right) in the hake occurrence model.

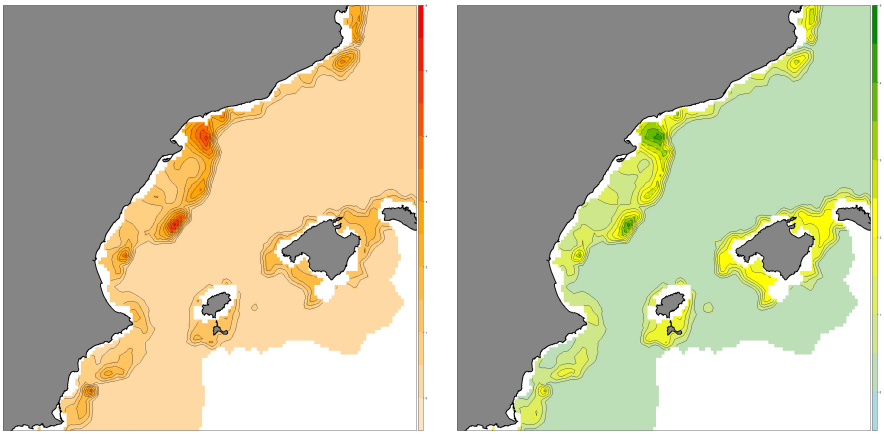


(a) Mean

(b) Standard deviation

Figure 4.7. Yearly mean estimates of the unstructured random effect for year in the hake occurrence model (left) and contitional-to-presence abundance model (right).

unstructured temporal effect for year was three times smaller than that of the spatial variance (Figure 4.9) and the estimates of the marginal unstructured temporal effect showed no apparent correlation (Figure 4.7, left panel).



(a) Mean (b) Standard deviation

Figure 4.8. Posterior mean (left) and standard deviation (right) for the hake contitional-to-presence abundance.

Hake nursery grounds

As we have seen in the results, both the occurrence and conditional-to-presence abundances preferred a persistent spatial realisation as proposed in equation 4.6. The models identified at least 3 high abundance and occurrence areas. A small hotspot was located a few kilometres off the city of Valencia, while the highest abundance hotspot was located some kilometres to the north-east, around the Columbretes Islands. This hotspot extended transversally to the bathymetric slope and connected through a moderate density region to an other high density area north of the Ebro delta. These 2 highest abundance hotspots encompass around 650 km2 of the total 18000 km2 area of the 50 to 200 m depth strata in the GSA 06. The areas close to the Palamós Canyon and Mar Menor showed relatively high abundance estimates, while the estimated occurrences were not that high. This behaviour

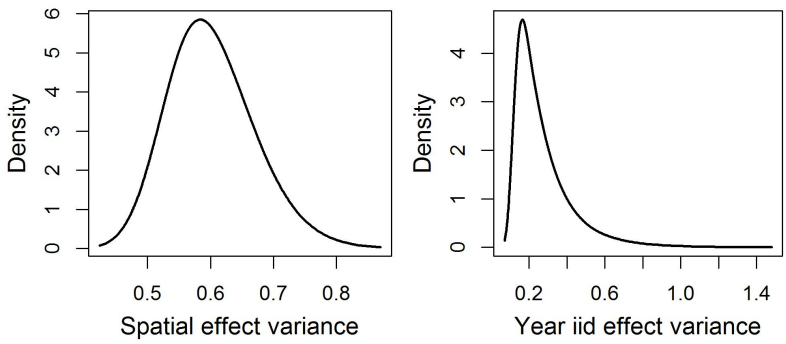


Figure 4.9. Estimated distribution of the variance for the spatial effect (left) and independent random effect for year (right) in the hake conditional-to-presence abundance model.

suggests that the aggregation patterns are diffuse, and hence these areas were not considered to be important nursery grounds.

4.1.4.2 Cod results

In the case of cod, the structure selection process was performed in two steps according to the temporal resolution of the data. First, we assessed the persistence of cod distribution along the year by applying the model comparison in 4.1.3 to assess whether the distribution of cod in the first and third quarters of the year were consistent. Secondly, we assessed the persistence of the model in between years as with the hake recruitment.

The bathymetric effect in this case was modelled applying a second order random walk model (RW2), which performs like Bayesian smoothing splines (Fahrmeir and Lang, 2001) to allow for non-linear effects to be fitted. Table 4.2 shows the WAIC goodness-of-fit and LCPO predictive quality measures of the models.

Following the principle of parsimony, the selected model for the occurrence has different spatial realisations each quarter and a consistent distribution

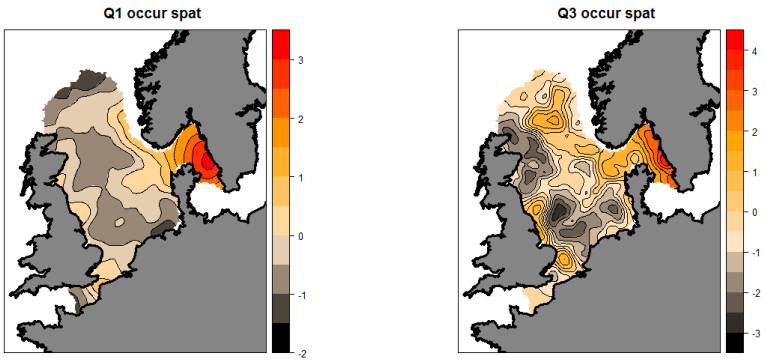
	Occurrence		Abundance	
Model	WAIC	LCPO	WAIC	LCPO
Persistent between quarters	10879	0.502	63020	4.106
Inconsistent between quarters	10342	0.477	61989	4.048
Persistent among years	10324	0.476	61910	4.07
Inconsistent among years	10392	0.479	61794	4.42

Table 4.2. Model comparison for the cod occurrence and conditional-to-presence abundance models.

between years. Regarding the conditional-to-presence abundance, results also show different spatial patterns in winter and summer but the assessment of the abundance spatial distribution over the years is less clear. While WAIC scores clearly benefit the temporarily inconsistent structure in the abundance model, LCPO scores prefer the annually persistent structure. This might occur due to the presence of some very influential observations (e.g. accidental big school catches). Modelling cod may require some extensions from the usual Hurdle model to accommodate large aggregation observations (schooling effect) as we will mention in the conclusions of this PhD dissertation as future study lines in fisheries distribution modelling.

Cod occurrence

The selected model for cod occurrence revealed highest presence probabilities in the north-east of Denmark both in winter and summer (Figure 4.10). However, in summer expected probabilities are higher in the northern part of the study area and the south-west (English channel). Cod showed a occurrence peak in between the 100 and 150 metres (Figure 4.12) in both quarters of the year.



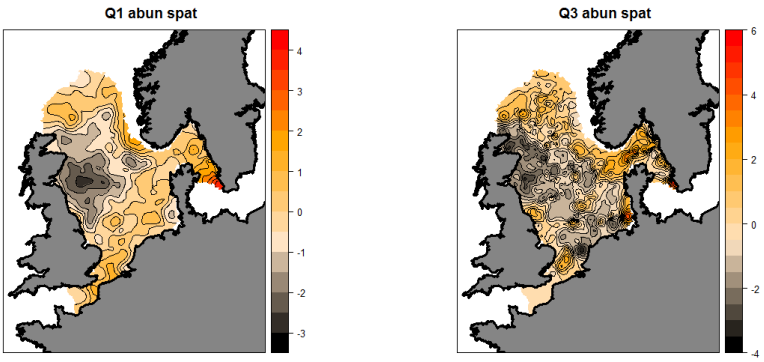
(a) Q1 (b) Q3

Figure 4.10. Posterior mean of the spatial effect in winter (left) and summer (right) for cod occurrence.

Cod abundance

Figure 4.11 show the mean spatial effect for each quarter assuming persistent distributions along the years. Note, especially in the third quarter, the patchy effect of the spatial field. This is likely to happen due to the presence of a more complex spatio-temporal pattern in cod abundances, as suggested by the WAIC scores.

Cod abundance revealed again very high values in the north-east of Denmark in both quarters. In summer, abundances are generally higher around Denmark. In both quarters a significant cold-spot is observed in the north-east of the United Kingdom. Cod showed an abundance peak in between the 80 and 150 metres (Figure 4.12) in both quarters of the year, while less cod is expected in deeper waters over winter than in summer. This could be due to cod's preference to cold waters.



(a) Q1 (b) Q3

Figure 4.11. Posterior mean of the spatial effect in winter (left) and summer (right) for cod conditional-to-presence abundance.

4.1.5 Discussion

The methodology proposed in Section 4.1.3 assesses the persistence of a spatial process by comparing 2 spatio-temporal structures, while density hotspots are identified by combining information from independent occurrence and abundance sub-models. Consequently, compared to the methodology proposed by Colloca et al. (2009), this approach not only reduces the number of steps needed to assess the persistence of the spatial process but also includes information on absence observations through the occurrence sub-model, so as to better characterise the spatial presence of hake recruits. In fact, areas where high abundance estimates concur with low occurrence estimates have not been highlighted as important nursery grounds in the hake recruitment scenario.

Results suggest a persistent spatial distribution of hake recruit occurrence and abundance in the western Mediterranean while in the case of cod, different distributions are inferred for winter and summer (inconsistent pattern), confirming an already know phenomenon (ICES, 2014). Results are slightly more complicated to interpret when assessing the temporal persistence of cod dis-

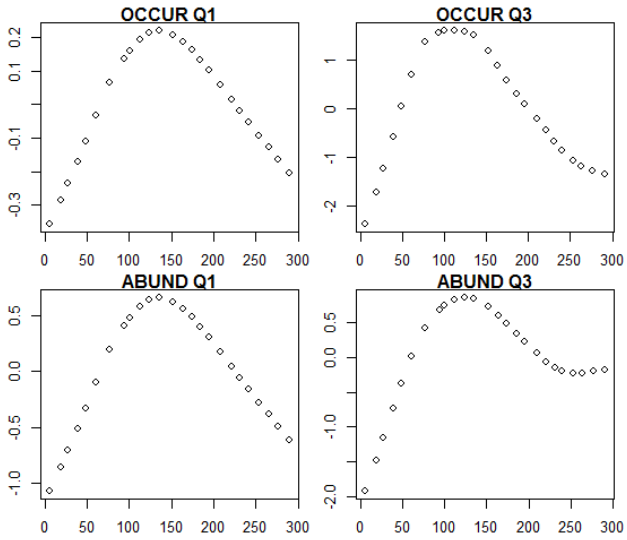
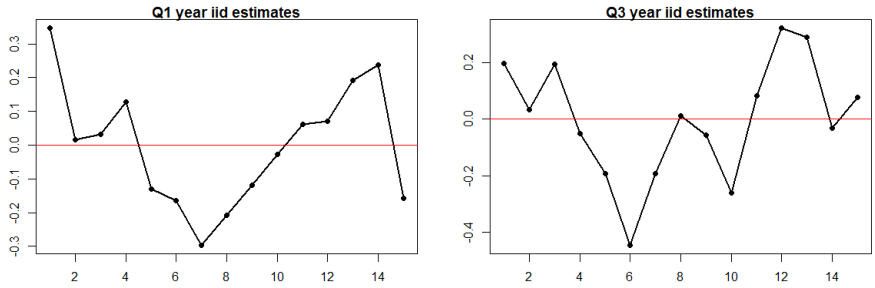


Figure 4.12. Fitted marginal bathymetric effects for cod. Winter occurrence (top-left), summer occurrence (top-right), winter abundance (bottom-left) and summer abundance (bottom-right).



(a) Mean (b) Standard deviation

Figure 4.13. Yearly mean estimates of the cod unstructured random effect for year in winter (left) and summer (right).

tribution at each quarter. While the spatial distribution of the occurrence of cod seem to be persistent over the years for each quarter, WAIC and LCPO scores suggest different models in the case of the abundance process. This phenomena may be related to the presence of a number of high abundance hauls (schooling effect outliers) or, as suggest by the patchy spatial effect observed in summer (Figure 4.11 right panel), because the distribution is not completely persistent over time. These results could be suggesting the fact that more complex spatio-temporal models are necessary.

Moreover, a quick look at the fitted yearly mean estimates of the unstructured random effects for year (Figures 4.7 and 4.13) may show temporal correlation patterns. In the following Section (4.2), we propose another two spatio-temporal structures to further diagnose the spatio-temporal behaviour of the process under study. In addition, we investigate on an approach to deal with the independence/non-independence of the occurrence and conditional-to-presence abundance sub-processes.

4.2 Comparing different spatio-temporal structures and shared components fore spatio-temporally sampled semi-continuous data.

In the previous section, we have proposed the comparison of two spatio-temporal structures to assess the temporal persistence of a spatial process. However, it is rather obvious that the spatio-temporal correlation structure of a process can be much more complex. Unfortunately, in the case of fish, fishery surveys do not generally allow intra-annual temporal analysis of the spatial distribution since they are usually performed only once per year (with the exception of the IBTS survey in the NS). However, every process in nature evolves in time, and therefore if fishery surveys are carried out during a reasonably long period of time, we might be able to see certain patterns. Likewise, the basic principle of time series analysis is that long runs of repeated measurements over time can display temporal tendencies and, with regards to the analysis performed in this chapter, fitted temporal random effects in Section 4.1 suggest that hake recruitment and cod abundances may have inferable temporal patterns too (see Figures 4.7 and 4.13).

In this regard, one of the pillars of this section is to propose a handful of spatio-temporal model structures to approach different types of spatio-temporal data/process scenarios. Specifically, we propose four generic spatio-temporal structures, including both structures proposed in the previous section. The idea behind these structures comparison is to, by means of goodness-of-fit criteria, characterise the overall spatio-temporal beahviour of the process under study: opportunistic, persistent or progressive (in the sense that it evolves with time) patterns. Such an spatio-temporal understanding of fisheries distribution can be essential to fisheries spatial management policy makers.

The second pillar of this chapter is to tackle the fact that the assumption

of independence between the two sub-processes of a Hurdle model is rather unnatural. In fact, in nature, low intensities are expected to be linked to low probabilities of occurrence and vice versa. Acknowledging this is fundamental to fit robust process-environment relationships and thus to species distribution modelling. In this vein, this section proposes the use of shared component modelling (SCM) techniques (Tsiatis and Davidian, 2004; Knorr-Held and Best, 2001) to fit common process-environment effects by embracing information from both the occurrence and conditional-to-presence abundance sub-processes.

4.2.1 Gaussian latent spatio-temporal structures for species distribution modelling

The distribution of species not only changes in space but also in time. Depending on the nature of the process under study and the available sampling resolution, the spatio-temporal behaviour of the data can vary. Consequently, and as we have done in the previous section (4.1), comparing different spatio-temporal model extensions provides further description and/or understanding of the species under study. This will result in an improved predictive capacity of our models, which, in cases like the EAFM, may be crucial for management purposes.

In order to incorporate other spatio-temporal and smoothing effects, the sub-models (4.1) introduced in the previous section can be rewritten as:

$$\begin{aligned} Y_{st} &\sim \text{Ber}(\pi_{st}),\ s = 1, \ldots, n \\ \text{logit}(\pi_{st}) &= \alpha^{(y)} + \sum_{i=1}^{I} f_i^{(y)}(x_{ist}) + U_{st}^{(y)} \\ Z_{st}|Y_{st} &\sim \text{Ga}(\mu_{st}, \rho),\ s = 1, \ldots, n \\ \log(\mu_{st}) &= \alpha^{(z)} + \sum_{i=1}^{I} f_i^{(z)}(x_{ist}) + U_{st}^{(z)} \end{aligned} \tag{4.4}$$

where $t = 1, ..., T$ is the temporal index and $s = 1, ..., n_t$ is the spatial location of each sub-process and potentially different at each t. U_{st} represents the

spatio-temporal structure of the models, 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.).

Based on this structure we propose the comparison of four basic decompositions for U_{st} in (4.4), each one allowing different degrees of flexibility in the temporal domain of the spatio-temporal model. Please note that two of these structures were already proposed in the previous section, but for the reader's ease are written again:

- The most flexible structure consists of decomposing U_{st} into different spatial realizations of the same spatial field for each time unit. This structure may be a good proxy to those processes where the spatial structure vary considerably among different time units and unrelatedly among neighbouring times. In particular,

$$\begin{aligned} U_{st} &= W_{s_t} \\ \boldsymbol{W}_t &\sim N(0, \mathbf{Q}(\kappa, \tau)) \end{aligned} \tag{4.5}$$

  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, as mentioned in the previous section (4.1).

  In `R-INLA`, these flexible structure that fits different spatial realizations at each time and shares hyperaparameters can be fitted by including the following syntax in the formula environment:

```
## Fit different spatial realizations at each time
# that share the hyperparameters of the covariance
formula <- Y ~ ... + f(spat, model=spde,
                       replicate=s.replicate)
```

  where `s.replicate` is the temporal indexation that has previously been created using the `inla.spde.make.A()` function (as in Chapter 1).

- 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,

$$\begin{aligned} U_{st} &= W_{st} + V_t \\ \boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ V_t &\sim N(0, \sigma^2) \end{aligned} \tag{4.6}$$

  where U_{st} is decomposed in a common spatial realization W_{st} 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 but intensities vary unrelatedly through time.

- Alternatively, the mean intensities at each time t could show a temporal progression or tendency. Such a case would best fit in our third proposed structure, which includes a mean temporal trend effect $g(t)$ through a linear or non-linear effect,

$$\begin{aligned} U_{st} &= W_{st} + g(t) \\ \boldsymbol{W} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \end{aligned} \tag{4.7}$$

  where U_{st} is decomposed in a common spatial realization W_{st} and a temporal trend *g(t)* to absorb the temporal progression of the process. Processes where the spatial distribution is persistent but mean intensities show a temporal tendency will benefit from this structure.

  In `R-INLA`, the structures in the previous equations (4.6) and (4.7) are fitted using the following syntax in the formula environment:

```
## Fit different spatial realizations at each time
# that share the hyperparameters of the covariance
formula <- Y ~ ... + f(spat, model=spde) +
                 f(time, model="XXX", prior = prior)
```

where `XXX` can be any one dimensional model available in `R-INLA` (see http://www.r-inla.org/models/latent-models for all the available latent models). For example `"iid"` stands for the unstructured random effect in equation (4.6).

- 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,

$$\begin{aligned} U_{st} &= W_{st} + R_{st} \\ \boldsymbol{W}_t &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ R_{st} &= \sum_{k=1}^{K} \rho_k U_{s(t-k)} \end{aligned} \tag{4.8}$$

  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 favoured when the spatial realization varies between different times t but not as much as in (4.5). Note also that this structure could be applied along with that in (4.7).

  This spatio-temporally correlated model can be fitted using the following syntax in the formula environment of `R-INLA`:

```
## Fit a spatio-temporal (spde + ar(p)) field
formula <- Y ~ ... + f(spat, model=spde,
                       group=s.group,control.group = list(
                       model="ar",order = p))
```

  where `s.group` is the temporal indexation that has previously been created using the `inla.spde.make.A()` function and `p` is the order of the auto-regressive temporal correlation term.

It is rather evident that we could have proposed several more complex temporal structures. Unfortunately, as we previously mentioned, the temporal resolution of spatio-temporal fisheries datasets is typically too low to fit highly structured models. Nevertheless, comparing the goodness-of-fit of these four basic spatio-temporal structures (4.5), (4.6), (4.7), and (4.8) allows us infer the general spatial behaviour of the process over time, which can per se provide very useful information for decision making.

4.2.2 Shared component analysis for Hurdle models

As we saw in the previous section 4.1, many spatio-temporally sampled abundance processes are prone to zero value observations at non-favourable conditions (e.g. rain, species abundance, plant coverage, chemical concentrations, etc.), and are thus measured continuously in the $[0, \infty)$ interval, resulting in semi-continuous datasets. The absence of distributions capable of plugging into such datasets has persuaded scientists to apply two-part or Hurdle models (Martin et al., 2005) by decomposing the dataset into two independent sub-processes, an occurrence process and a conditional-to-presence continuous process. However, in nature both sub-processes are often related: low intensities are linked to low probabilities of occurrence and vice versa. Fitted effects may then be incomplete due to substantial information being ignored in each sub-process, such as zero observations in the abundance model and observed abundances in the occurrence model.

The widely used approach in (4.4) formulates independent models for each of the sub-processes of the two-part model. However, ecologically speaking, the assumption of independence between the occurrence and abundance sub-processes may not be a natural approach, and therefore fitted effects in each independent sub-process could be biased due to the lack of substantial information when fitting such effects. For instance, this approach inherently assumes that any abundance has equal weight in the probability of presence and that zero observations have no impact on the abundance model. However, from an ecological point of view, such assumptions are likely to be erroneous

and should be tackled in some way to ensure that fitted process-environment effects share information from both sub-processes. A good approach may be the use of shared components in both linear predictors by means of joint modelling.

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 the occurrence and abundance sub-processes and therefore fit more robust model components.

In order to introduce SCMs in (4.4) and fit common model components that share information from both sub-processes, we propose modelling both sub-processes together:

$$\begin{aligned} \text{logit}(\pi_{st}) &= \alpha^{(y)} + \sum_{i=1}^{I} f_i(x_{is}) + U_{st} \\ log(\mu_{st}) &= \alpha^{(z)} + \sum_{i=1}^{I} \theta_i f_i(x_{is}) + \theta_U U_{st} \end{aligned} \tag{4.9}$$

where notation is the same as in (4.4), but fitted 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.

In summary, in this section we have proposed four different spatio-temporal structures and a case dependent number of shareable effects θ_i, θ_U as an approach to tackle spatio-temporally sampled semi-continuous processes. 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. R-INLA again (as it

has been along all this thesis) becomes an ideal candidate to deal with these models thanks to its computational efficiency as discussed in chapter 1.

4.2.3 Case study: hake recruitment

In this section we have used the recruitment dataset to compare all the resulting models obtained by implementing the four spatio-temporal structures and the shared component analysis described in sections 4.2.1 and 4.2.2 respectively. 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) that resemble Bayesian smoothing splines (Fahrmeir and Lang, 2001). In the case of the fourth temporal structure in equation (4.8), 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 most 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 R-INLA. Only the prior of the bathymetric RW2 precision was changed to a $Loggamma(2, 0.00005)$ 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.

4.2.4 Results

All the resulting model structures were fitted and compared on the basis of WAIC scores and LCPO scores. As highlighted in the WAIC scores of Table 4.3, two structures performed reasonably better than the rest. Both models include a first order autoregressive temporal term, with independent bathymetric effects in the occurrence and the abundance sub-processes in model 14, while model 15 fits a shared bathymetric effect to both.

	Model	WAIC	LCPO		Model	WAIC	LCPO
1	X + W + V_t	1916.9	1.22	10	**X** + W + $\boldsymbol{g(t)}$	1931.6	0.66
2	X + W + $g(t)$	1925.4	1.23	11	**X** + W + $\boldsymbol{V_t}$	1931.6	0.65
3	X + W_t	1954.9	1.46	12	**X** + **W** + $\boldsymbol{V_t}$	1975.6	0.72
4	X + **W** + V_t	1969.9	0.63	13	**X** + **W** + $\boldsymbol{V_t}$	1979.6	0.69
5	**X** + W + V_t	1922.3	0.65	14	X + W + R_{st}	**1836.2**	1.27
6	X + W + $\boldsymbol{V_t}$	1913.6	0.53	15	**X** + W + R_{st}	**1839.9**	0.62
7	X + W + $\boldsymbol{g(t)}$	1917.1	0.54	16	X + **W** + R_{st}	2097.5	0.65
8	X + **W** + $\boldsymbol{g(t)}$	1977.7	0.62	17	**X** + **W** + $\boldsymbol{R_{st}}$	2098.1	0.80
9	**X** + **W** + V_t	1971.6	0.70				

Table 4.3. Model fit scores for the most representative model structures. X = bathymetry, W = spatial effect, W_t = yearly spatial realisations , V_t = unstructured random effect for time, R_{st} = first order autoregressive structure for time, $g(t)$ = smooth temporal trend for time. **Bold** terms = shared components. The highlighted WAIC scores represent the models that perform best.

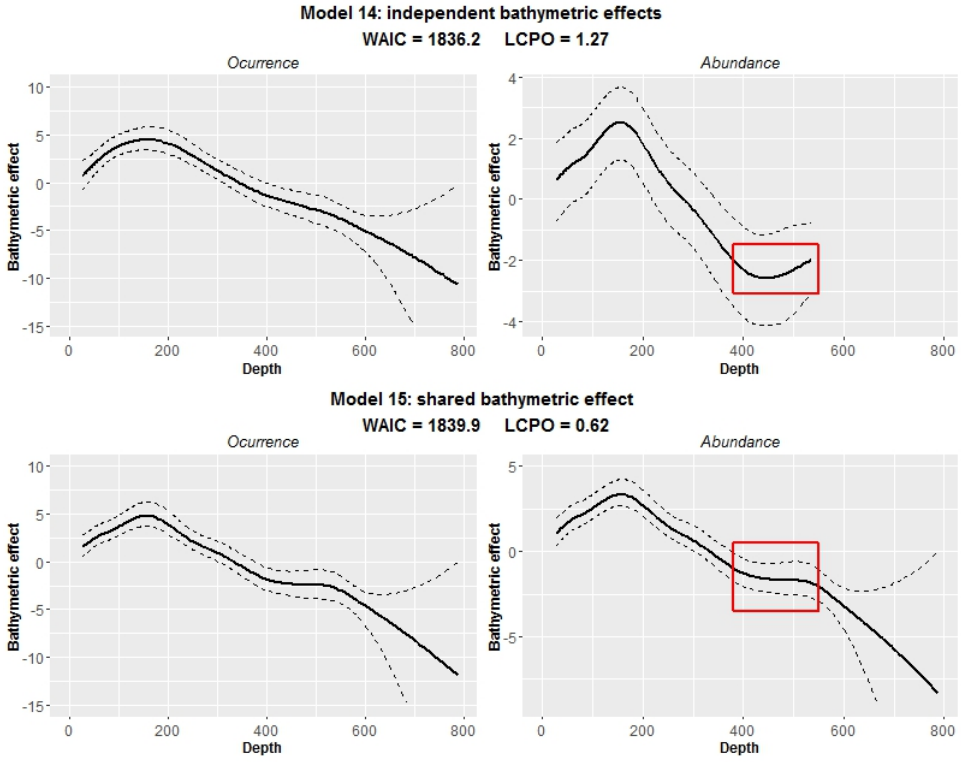
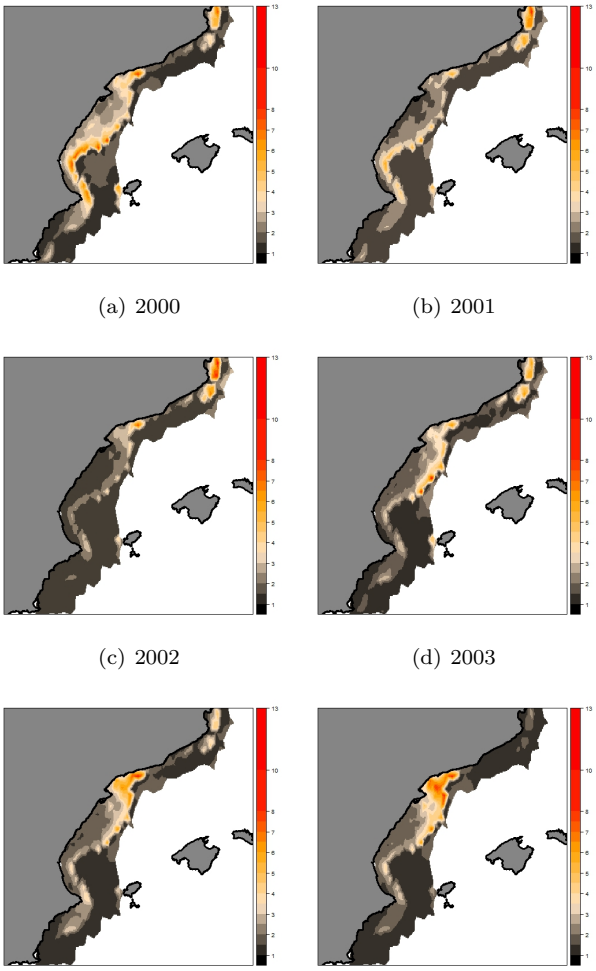


Figure 4.14. Fitted smooth bathymetric effects in models 14 and 15 (Table 4.3). The solid line represents the mean of the effect and the dashed lines its 95% credibility interval. The marked box highlights the importance of SCM to fit a biologically more natural bathymetric effect for hake recruit abundance.

We finally selected model 15 over model 14 for a number of reasons. Firstly, model 15 fits a biologically more natural bathymetric effect (see Figure 4.14), where the abundance of hake recruits decreases gradually after the optimum 150-200 meter strata. Secondly, model 14 clearly overfits the bathymetric effect of the abundance sub-process due to the lack of zero observations in it (see highlighted box in Figure 4.14). Interestingly, even if WAIC scores slightly prefer model 14 over model 15, the predictive LCPO scores clearly benefit model 15. Lastly, model 14 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 that there is certain relation between temporally neighbouring points in space. Moreover, such temporal correlation term allows a better informed interpolation and thus a better representation of the distribution of hake recruitment in the western Mediterranean. Indeed, as the posterior predictive maps in Figures 4.15 and 4.16 show, the recruitment of hake is mainly concentrated in the central and northern parts of the study area (as already discussed in the previous section). We can observe smooth changes in abundance and the distribution of hake recruitment hot-spots from year to year (see Figures 4.17 and 4.17), 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). In our case, the variability of the occurrence sub-process as a function of distance differed too much from that of the abundance sub-process. Consequently, the fitted joint spatial field failed to satisfy either sub-model, particularly so in the case of the abundance sub-process. This can be seen in Figure 4.19, 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.

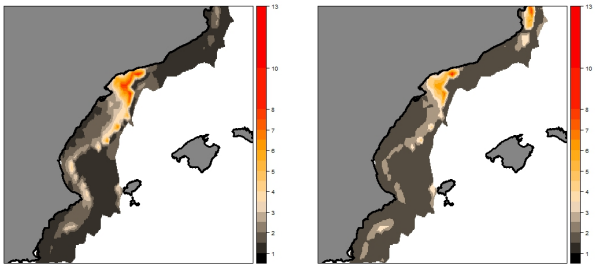


(a) 2000 (b) 2001

(c) 2002 (d) 2003

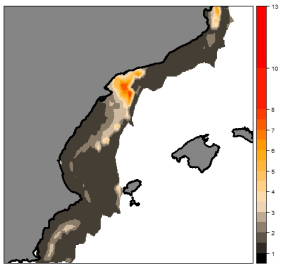
(e) 2004 (f) 2005

Figure 4.15. Yearly hake recruitment posterior predictive mean abundance maps (2000 to 2005).

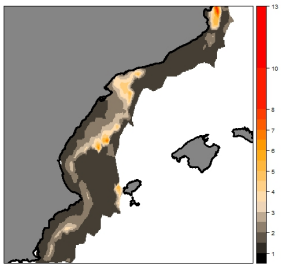


(a) 2006

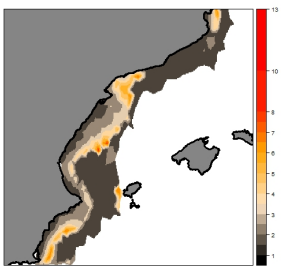
(b) 2007



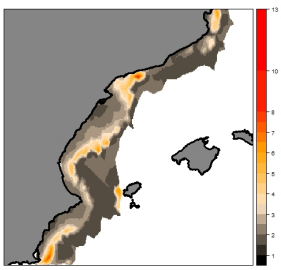
(c) 2008



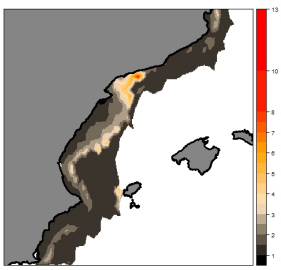
(d) 2009



(e) 2010

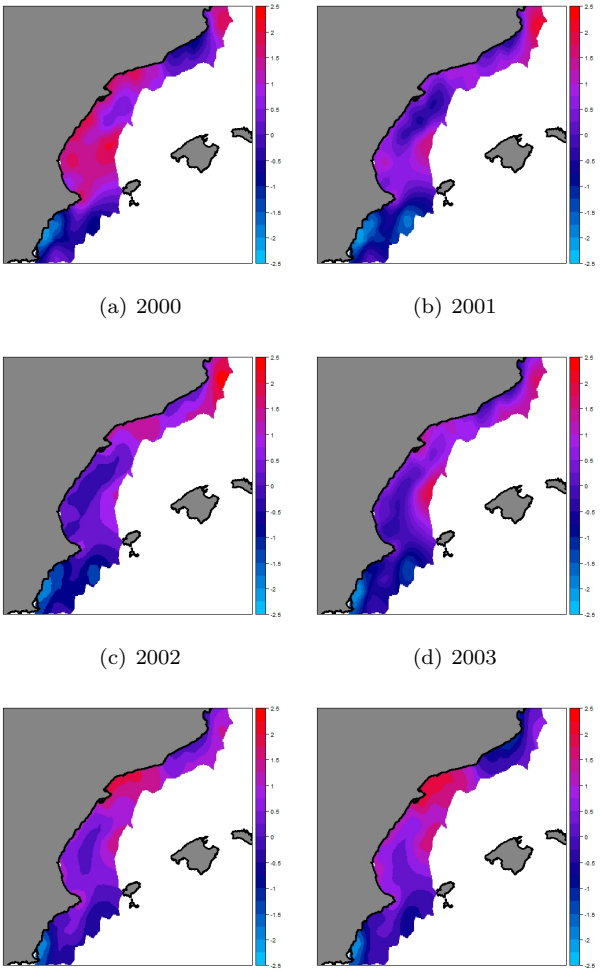


(f) 2011



(g) 2012

Figure 4.16. Yearly hake recruitment posterior predictive mean abundance maps (2006 to 2012).

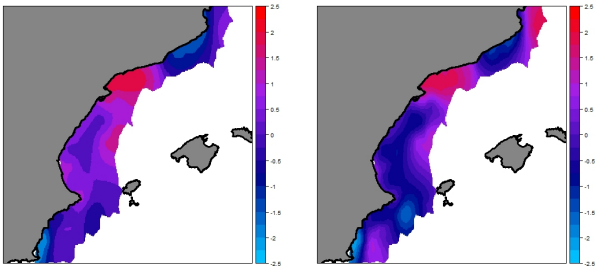


(a) 2000 (b) 2001

(c) 2002 (d) 2003

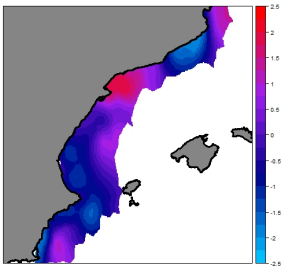
(e) 2004 (f) 2005

Figure 4.17. Yearly hake recruitment posterior spatial effect (2000 to 2005).

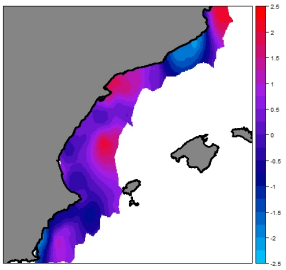


(a) 2006

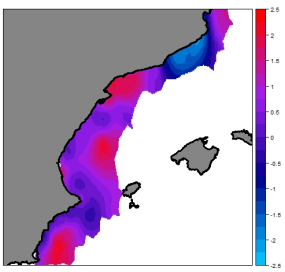
(b) 2007



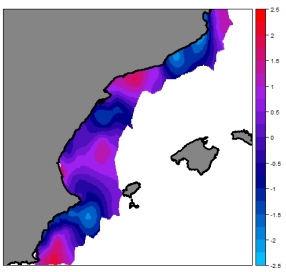
(c) 2008



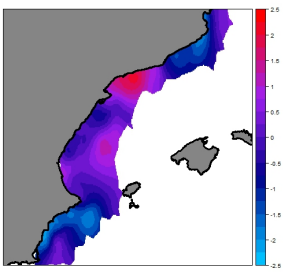
(d) 2009



(e) 2010



(f) 2011



(g) 2012

Figure 4.18. Yearly hake recruitment posterior spatial effect (2006 to 2012).

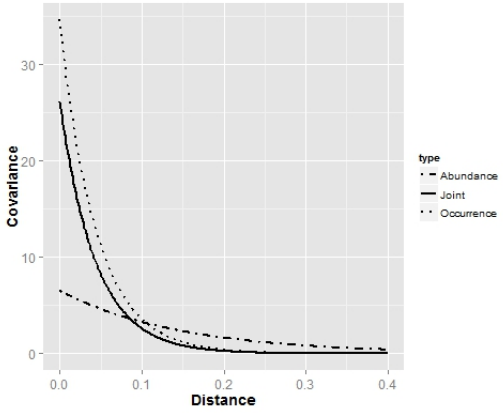


Figure 4.19. Fitted Matérn covariance functions in the unit scale. The solid line represents the joint covariance function, the dotted line represents the covariance function for independent occurrence model and the dot-dashed line that for the independent abundance model.

4.2.5 Discussion

In this section 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 as an effective approach to dealing with fitted effects in two-part models for semi-continuous data. By using the proposed approaches, we have significantly improved the information available for the management of hake recruitment in the western Mediterranean. The same approach could be used to improve the fit and prediction of other spatio-temporally sampled semi-continuous processes. 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 resource is cru-

cial for management purposes and decision making. For this reason, the spatio temporal structures proposed in section 4.2.1 make it possible to identify 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 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 the hake recruitment example, a correlation structure will suit best our data.

Regarding the use of SCM in semi-continuous processes, this study has proved that fitted environment-process effects can be improved by combining information on occurrence and conditional-to-presence abundance. However, common model selection scores such as WAIC may benefit independent two-part models over the use of shared components due to overfit effects in independent two-part models. In such cases, cross-validation scores such as LCPO may help us select the best model.

However, in the case of the spatial field, fitting a shared component in semi-continuous processes may not always perform well. Generally speaking, the variability of a presence-absence sub-process as a function of distance, may not be comparable to that of the abundance sub-process, and hence SCM may not improve two-part models 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 here 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.

4.3 Conclusions

Fishery research surveys play a very important role in the management of our fisheries. Fishery survey data, or fishery independent data, cover very wide areas and allow us understand the macroscopic view of the fisheries. This is specially relevant under a global fisheries spatial management framework, i.e. the ecosystem approach to fisheries management (EAFM) (FAO, 2003), because it allows us quantify the importance of marine areas in the macro-scale, which is the scale at which marine protected areas should be designed (Fenberg et al., 2012).

As we have seen along this chapter, spatial statistical models can play a key role in the assessment of the EAFM. For instance, assessing the spatial persistence of nursery areas and spawning areas of different fish species is of great value (Beck et al., 2001; Gillanders et al., 2003; Colloca et al., 2009). In this regard, the approach proposed in Section 4.1 to assess such persistence has proven to work fairly well in both the hake recruitment and the cod processes. In fact, in the case of the cod fishery, results confirmed what fishery experts already knew, that cod has different spatial distributions in summer and winter (ICES, 2014). Nevertheless, it is important to note that the model comparison in Section 4.1 assesses a relative spatial persistence of the process under study. An example of this may be the confusing model selection scores obtained in the yearly spatial distribution of the cod fishery and its resulting patchy distribution in the persistent model.

In Section 4.2, we have further developed the spatio-temporal in Section 4.1 structure comparison by incorporating both a spatially persistent model with a temporal mean trend effect and a spatio-temporally correlated structure. This way, we have been able to infer further information on the hake recruitment process of the Spanish Mediterranean, where even though hake recruit hot-spots are located in similar places every year (relatively persistent), some distributional changes are inferred. Similarly, and following the slightly confusing results obtained with cod in Section 4.1, it is likely that the spatio-temporally correlated model proposed in equation (4.8) will produce

more meaningful results.

This chapter has also investigated on the fitted effects of Hurdle models in SCM. From a biological point of view fitting independent occurrence and conditional-to-presence abundance sub-processes in a Hurdle model is unnatural at the very least because we expect that low intensities are linked to low probabilities of occurrence and vice versa. In this vein, this chapter has also proved that fitting shared components in the occurrence and conditional-to-presence abundance processes significantly improve fitted process-covariate relationships. The reason for this improvement relies on the fact that this way zero abundance observations do influence the abundance model and likewise, different abundance observations affect the occurrence probability.

Regarding model selection in shared spatial component models, usual within-sample model selection scores such as WAIC may not perform well (as also occurred in the case of preferential sampling problem of Section 3). These scores may benefit independent two-part models over the use of shared components due to overfitted effects in independent two-part models. In this case, leave-one-out cross-validation scores such as LCPO performed better in this study. Nevertheless, this topic requires further investigation as some influential observations may have helped the leave-on-out predictive score to identify the overfitting issue of independent Hurdle models.

In the case of the spatial field, fitting a shared component in semi-continuous processes may not perform that well. Generally speaking, the variability of a presence-absence sub-process as a function of distance, may not be comparable to that of the abundance sub-process, and hence SCM may not improve two-part models as also occurred in Quiroz et al. (2015).

The method to asssess the persistence of a spatial process over time presented in Section 4.1 has been published in the Marine Ecology Progress Series (MEPS) journal (http://www.int-res.com/journals/meps/meps-home/) (Paradinas et al., 2015).

Similarly, the comparison of different spatio-temporal structures

with shared components for species distribution modelling presented in Section 4.2 has been sent for peer reviewed publication.

Chapter 5

Point-referenced vs transect data

Previous chapters faced fishery dependent and independent data by means of model based geostatistics (Diggle et al., 1998), which rely in point-referenced observations to calculate Euclidean distances among them. However, as all fishery spatial studies have done to date, we obviated the fact that a trawling operation represents a transect in space, not a point. Until now, we have used the typical point-referenced representation at the centroid of the fishing operation. A priori, we assume that the error in the point-referenced representation is negligible when the study area is big with respect to the transect size. But, what if the study area is a small-scale fishing ground? Then the representation of the fishing operation by its centroid point could be problematic as can be seen in Figure 5.1.

In this setup, we face two problems. On the one hand, a fishing transect is likely to catch fish at different habitats (characterised by different depths and types of substratum for example) while we typically do inference based on the value of these covariates in the centroid of the fishing operation alone. On the other hand, when applying geostatistics, the spatial random effect

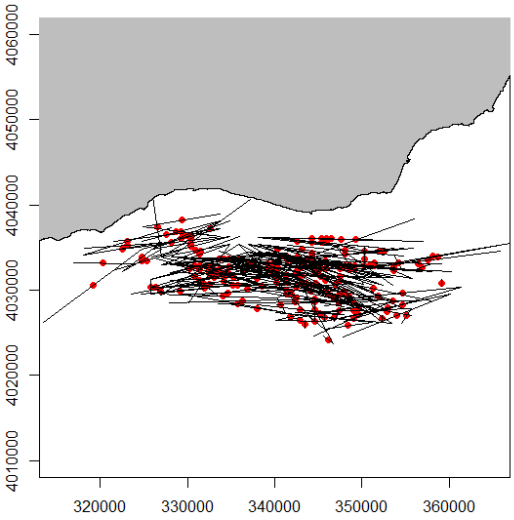


Figure 5.1. Transect and centroid representation of onboard sampling data in the southern Spanish Mediterranean. Lines represent the transect performed by the fishery operation. Red dots represent the centroid of each fishing operation.

is characterised by a covariance matrix that is typically based on Euclidean distances. However, as we can see in Figure 5.1, it is rather obvious that the distance between two fishing hauls is not necessarily the distance between the centroid points of the transects.

This chapter will investigate on the sensitivity of point-referenced representation of fishery transects to approximate the underlying spatial fields and examine the applicability of a new approach to estimate the underlying spatial field when enough data is available.

5.1 An algorithm to approximate the spatial field by overlaying fishery transects

In this section we describe the steps of an algorithm that we propose to approximate the underlying spatial field when enough data is available and fishery transects cross over enough times as in Figure 5.1. The proposed algorithm goes as follows:

1. Create a grid of the study area with the resolution of interest ($n \times m$).
2. Draw a transect line between the starting and finishing points of the fishing operation.
3. Set a d distance and divide the transect line into equidistant points of d distance.
4. Assign proportional abundance from the catch to each point of the transect. For example, 1000 kilograms divided into 200 points results in 5 kilograms assigned to each of them.
5. Aggregate at each cell the number of points that fall in it.
6. Do the same with every haul by going back to step 1.
7. Finally, compute the mean abundance at each cell of the grid.

The principle behind this approach is that if enough data were available and transects cross over enough, this algorithm could be able to approximate the distribution of fish in the study area, i.e. the underlying spatial field.

To test the performance of this algorithm we have applied it into two types of simulated spatial fields. Simulation study 1 (Section 5.2) tests the problem derived from the application of geostatistics & kriging based on Euclidean distances, while the real distance between two transects is way more complex (see Figure 5.1). Simulation study 2 (Section 5.3) tests the fact that a fishing transect can trawl at different habitats but we typically do inference

based only on the centroid value of the covariates. Lastly, the method has been tested in a real dataset (Section 5.4). In the simulation studies, results were compared with point-referenced regression models to assess both, the estimation error of the algorithm and the predictive error of point-referenced regression models.

5.2 Simulation study 1

The first simulation study aims to investigate the fact that the Euclidean distance between two transect centroids may not properly represent the distance between the two transects. As a consequence, applying geostatistics over these point-referenced representations could lead to biased results. In order to test for this, we have contrasted the results obtained by using the algorithm proposed in Section 5.1 and those obtained by means of conventional geostatistical methods in a set of simulated fields.

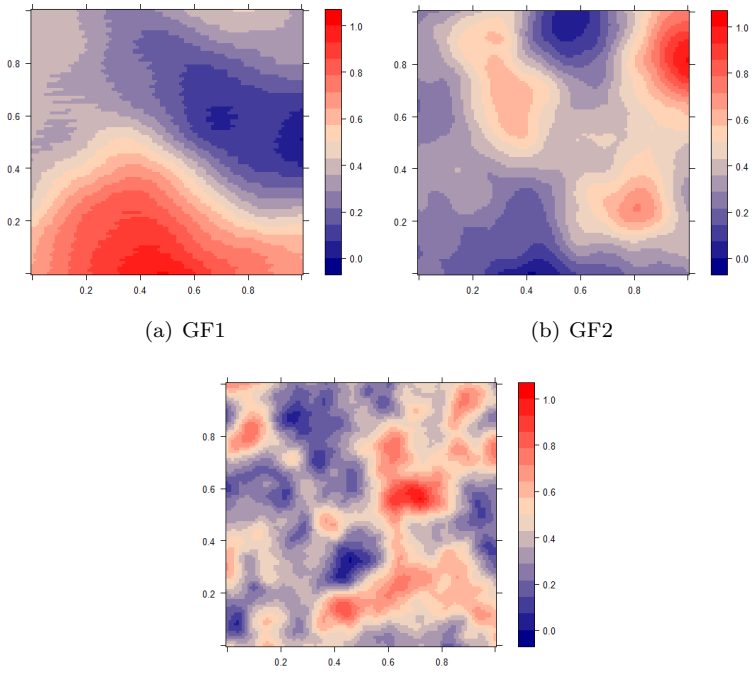
We have created three simulated Gaussian fields (GF) of different complexities over a grid. These simulated fields have been created using the `RandomFields` package for R and a Matérn covariance function. In all three simulated GFs the smoothness parameter (ν) has been fixed to 2 and different variance and scales have been used to allow for different types of fields to be fitted. After the simulation, the intensities of each GF have been scaled to be in between 0 and 1 for the sake of comparability.

By means of the different parametrisations of the Matérn, our aim has

Simulated field	ν	Var	Scale
GF1	2	8	.5
GF2	2	12	.3
GF3	2	18	.1

Table 5.1. Different parametrisations of the 3 simulated Gaussian fields.

been to create GFs with different levels of complexity and heterogeneity. The smoothest simulated field has been the first simulated field, labelled as GF1 in the top-left panel of Figure 5.2. Then the second field, labelled as GF2 in the top-right panel of Figure 5.2 and lastly GF3 in the bottom panel of Figure 5.2, has been the most complex of all. In summary, the level of heterogeneity/complexity of the GFs increased from GF1 to GF3 (Figure 5.2).



(a) GF1 (b) GF2

(c) GF3

Figure 5.2. Simulated Gaussian fields using a Matérn covariance function with different parameters summarised in Table 5.1

5.2.1 Simulating fishing operations

The next step was to simulate fishing operations in the study area. This was done by randomly choosing a set of starting and ending points spaced by a minimum and maximum distance in between. In this case the minimum and maximum haul lengths were approximately 10% and 50% of the longest distance in the study area. The catch of each transect has been computed by summing up the proportional catch at each cell using the same approach as in step 3 of 5.1 but inversely.

5.2.2 Performance testing

The performance of both, the proposed algorithm (Section 5.1) and the usual geostatistical approach (ordinary kriging in this case) applied over the centroid of the fishing operation have been tested for the three simulated Gaussian fields. The performance testing of both methods has been performed at two levels; comparing the representation error of point-referenced data and transect representation with the real values of the field; and comparing the results obtained with both methods agaithe real field. To assess these, we have used Mean Absolute Error (MAE) (Willmott and Matsuura, 2005) as a measure of the overall out-of-sample predictive score..

Data representation errors

As Figure 5.3 shows, in all three cases the representation error of the point-referenced geostatistical approach was smaller than that obtained by applying the algorithm proposed in Section 5.1. However, as the complexity of the underlying field increased, errors tended to get closer (see bottom panel in Figure 5.3).

Prediction/estimation errors

Similarly, the mean absolute predictive errors of the kriging interpolation approach was smaller than the estimations obtained by applying the algorithm

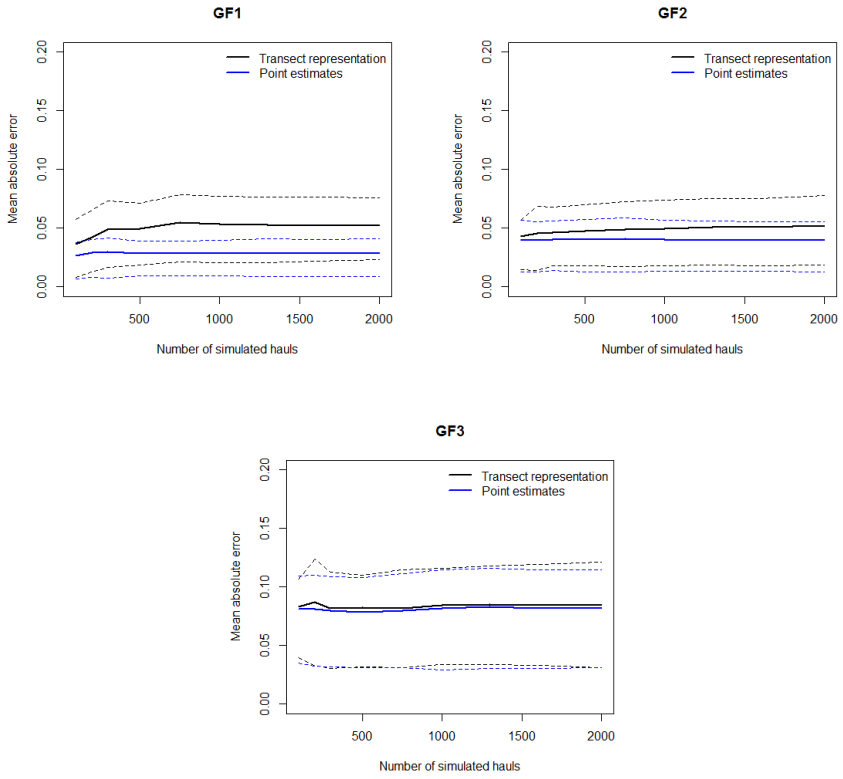


Figure 5.3. Mean absolute errors in the centroid data representation of transects using the conventional point-referenced approach (in blue) and the algorithm proposed in Section 5.1 (in black). Solid lines represent the mean and dashed lines the 95% confidence intervals of the mean absolute errors.

(Figure 5.4). Nevertheless, as with the representation errors, when the complexity of the underlying field increased errors tended to get closer (see bottom panel in Figure 5.3).

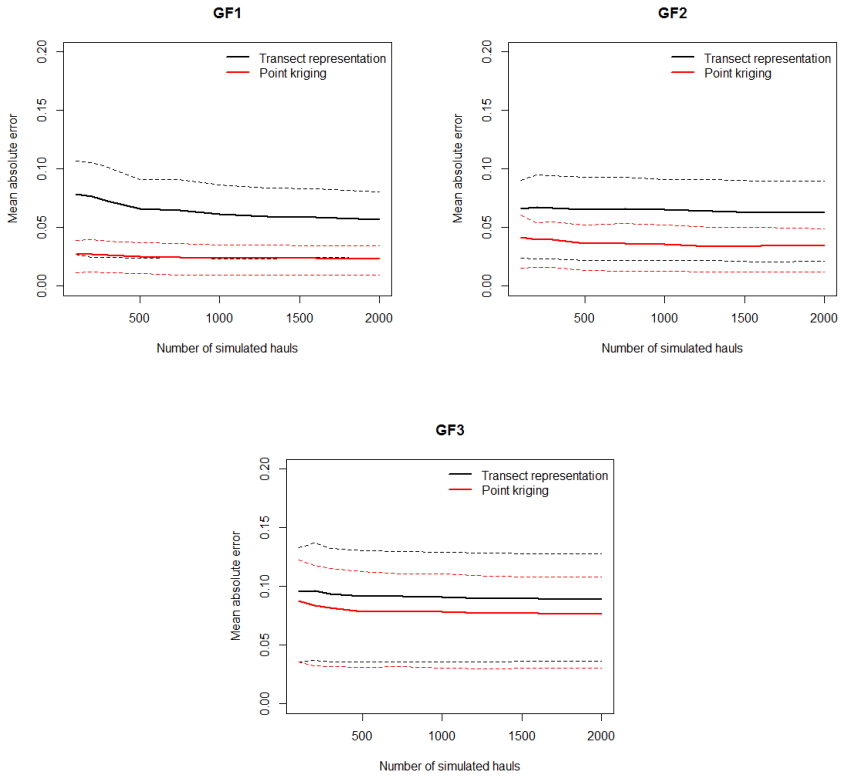


Figure 5.4. Mean absolute errors of the results obtained using the conventional point-referenced geostatistical approach (in red) and the algorithm proposed in Section 5.1 (in black). These errors were computed only in the cells were the algorithm had estimates. Solid lines represent the mean and dashed lines the 95% confidence intervals of the mean absolute errors.

5.2.3 Resulting maps

Although mean absolute errors proved the better performance of geostatistical & kriging methods over the transect superposition algorithm (5.1), at a reasonable number of samples the resulting maps of the algorithm were not that different from the real/simulated field. Figures (5.6, 5.8, 5.10) show the maps obtained by applying Bayesian geostatistical models and Figures (5.5, 5.7, 5.9) the maps obtained by applying the algorithm proposed in this chapter at $N = 100, 300, 500, 750, 1300, 2000$ number of simulated sampling hauls.

5.2.4 Discussion: simulated study 1

Results suggest that the use of geostatistics & kriging in small-scale fishing grounds produce good predictive estimates despite the fact that the use of Euclidean distances between fishery operations is not the most appropriate measure. Furthermore, we have seen that point-referenced geostatistical regression methods perform reasonably better than the algorithm (5.1) in most cases. Only when the heterogeneity of the underlying spatial field is big (in the scale of the transects), the estimation errors obtained through the proposed algorithm are similar to the predictive errors of kriging.

It is important to note, however, that when enough data is available, the proposed algorithm can approximate fairly well the underlying spatial field. This is specially relevant due to the simplicity and almost null computational requirement of the algorithm.

5.3 Simulation study 2

The second simulation study aims to investigate the fact that a fishery transect is likely to fish in different habitats (e.g. different bathymetries and types of substratum) while, when setting up point-referenced models, only the values extracted in the centroid of the fishing operations are used. In order to test for this issue, we have created a new simulated field and have contrasted the results obtained by means of the algorithm proposed in Section 5.1 and

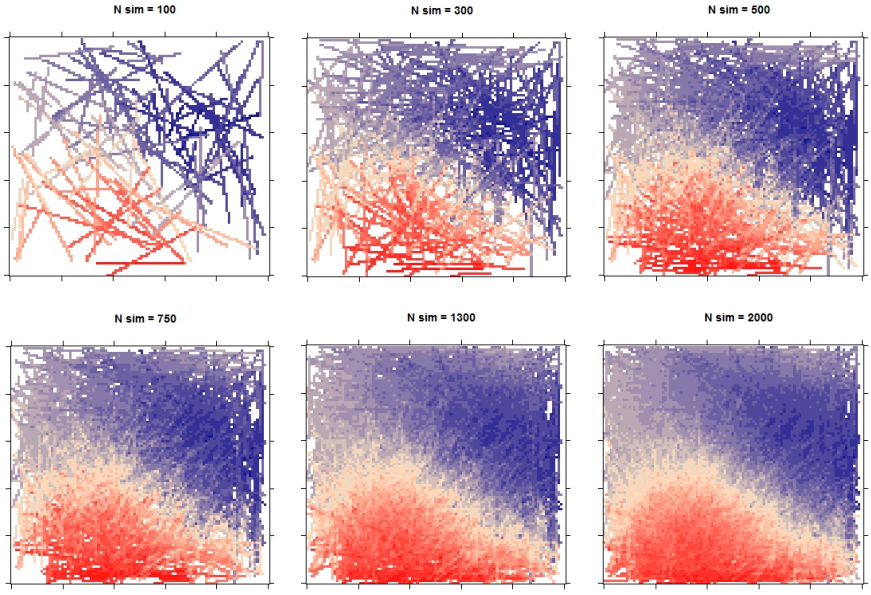


Figure 5.5. Results obtained by applying the proposed algorithm in the first simulated Gaussian field (top-left panel in Figure 5.2) at a different number of simulated sampling hauls.

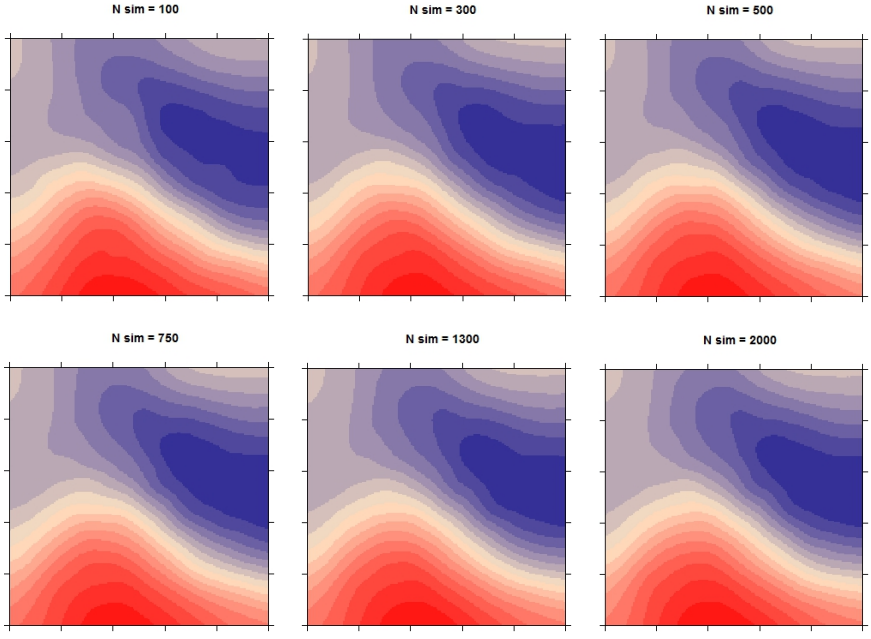


Figure 5.6. Results obtained by applying ordinary kriging in the first simulated Gaussian field (top-left panel in Figure 5.2) at a different number of simulated sampling hauls.

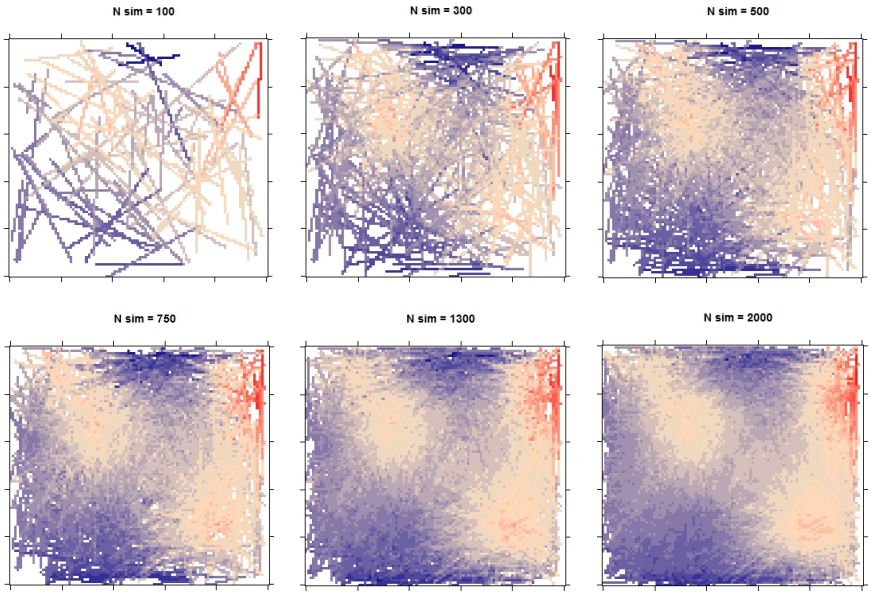


Figure 5.7. Results obtained by applying the proposed algorithm in the second simulated Gaussian field (GF2) (top-right panel in Figure 5.2) at a different number of simulated sampling hauls.

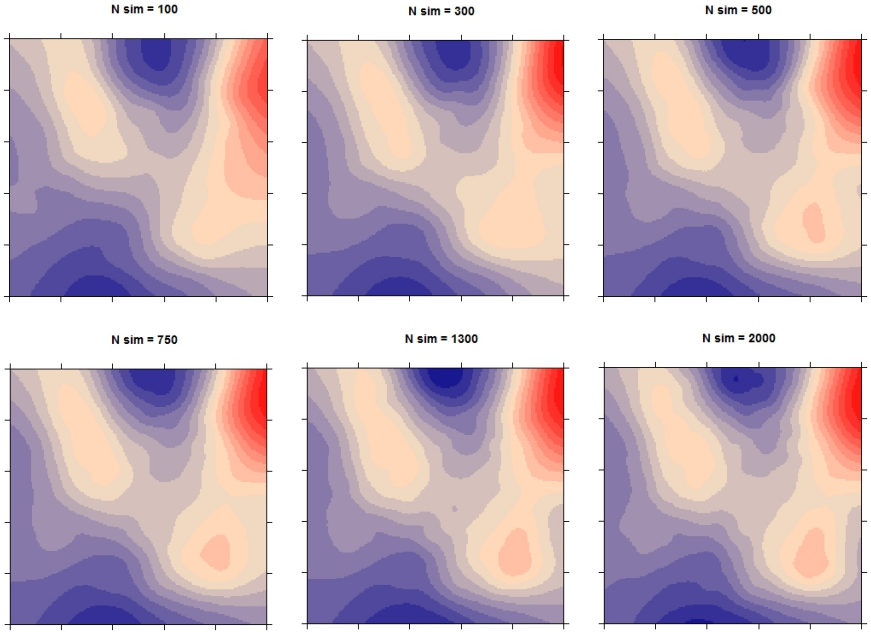


Figure 5.8. Results obtained by applying ordinary kriging in the second simulated Gaussian field (GF2) (top-right panel in Figure 5.2) at a different number of simulated sampling hauls.

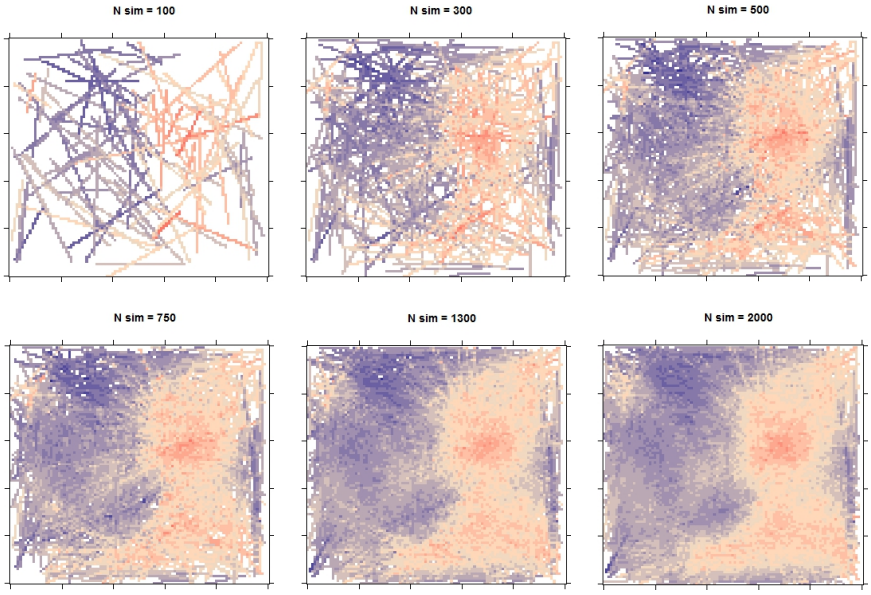


Figure 5.9. Results obtained by applying the proposed algorithm in the third simulated Gaussian field (GF3) (bottom panel in Figure 5.2) at a different number of simulated sampling hauls.

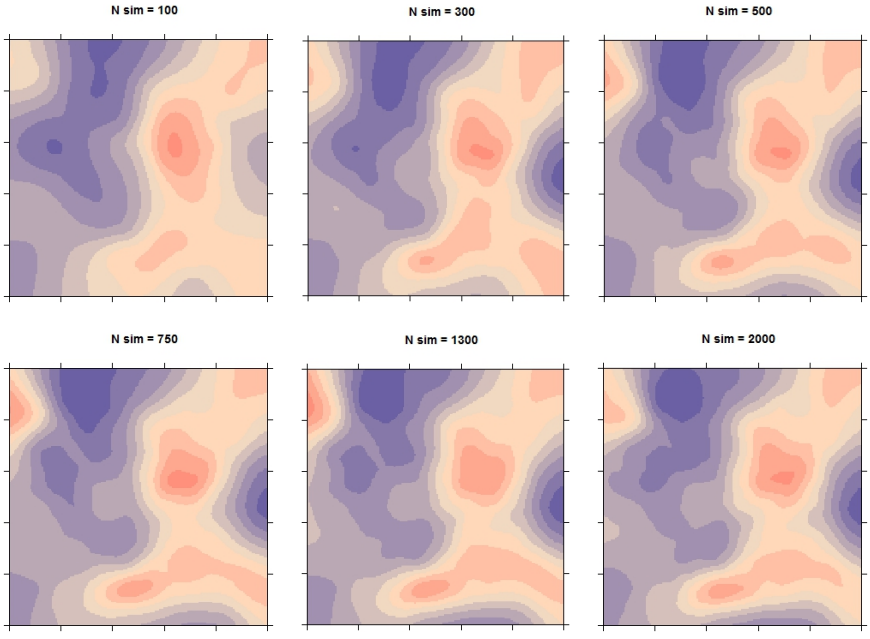


Figure 5.10. Results obtained by applying ordinary kriging in the third simulated Gaussian field (GF3) (bottom panel in Figure 5.2) at a different number of simulated sampling hauls.

those obtained by means of conventional geostatistical methods in a deterministically simulated field.

The spatial field has been simulated using a three-level categorical variable (category A = +4, category B = +10, category C = +6) to resemble the type of substrate, and a non-linear continuous variable to resemble the bathymetric effect on a marine species (Figure 5.11). Let Y be the simulated abundance

$$Y = \beta_i + f(D) + \epsilon, \tag{5.1}$$

where i stands for each of the levels of the categorical variable (see right-panel in Figure 5.11), $f(D)$ stands for the smooth bathymetric effect (see left-panel in Figure 5.11) and ϵ represents an added $N(0, 1)$ error term in the simulation process.

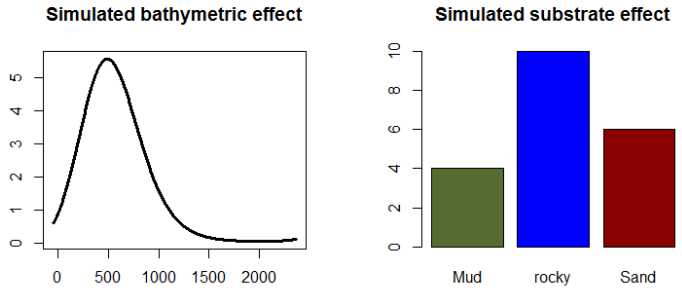


Figure 5.11. Simulated bathymetric (left panel) and substrate (right panel) effects.

The resulting spatial field (bottom panel in Figure 5.12) has been created over a bathymetric map (top-left panel in Figure 5.12) and a type of substrate map (top-right panel in Figure 5.12) of two unidentified areas of the Spanish Mediterranean.

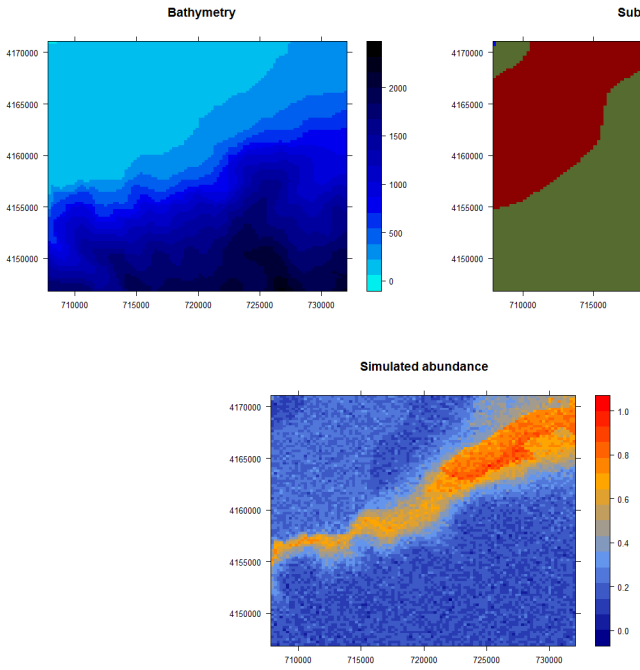


Figure 5.12. Simulated case study maps. Bathymetry map in the top-left panel, type of substrate map in the top-right panel and the resulting map by applying the equation (5.3) in the bottom panel.

5.3.1 Simulating fishing operations

Fishing operations have been simulated using the same approach as in the previous simulation study (Section 5.2). We randomly chose a set of starting and ending points spaced by a minimum and maximum distance in between. As before, the minimum and maximum haul lengths have been approximately 10% and 50% of the longest distance in the study area. The catch of each transect has been computed by summing the proportional catch at each cell using the same approach as in step 3 of 5.1 but the other way around.

5.3.2 Performance testing

In this case, we have only tested for the estimation/predictive capacity of the transect superposition algorithm and point-referenced regression. The point-referenced data representation errors were not tested because, as the previous simulated study showed, it is clear that it performs better than the proposed algorithm (Section 5.1). For that, we have modelled point-referenced representations using Generalised Additive regression Models (GAM) using the `mgcv` package and applied the algorithm in Section 5.1 to the data. Results were once again compared with the simulated spatial field using mean absolute errors.

As Figure 5.13 shows, the representation error of the point-referenced geostatistical approach was smaller than that obtained by applying the algorithm proposed in Section 5.1.

5.3.3 Resulting maps

Although mean absolute errors proved once again the better performance of point-referenced regression methods over the transect superposition algorithm (5.1), at a reasonable number of samples the resulting maps of the algorithm were not that different from the real/simulated field. Figure 5.15 shows the maps obtained by applying the usual point-referenced model while Figures 5.14 shows the maps obtained through the algorithm proposed in this

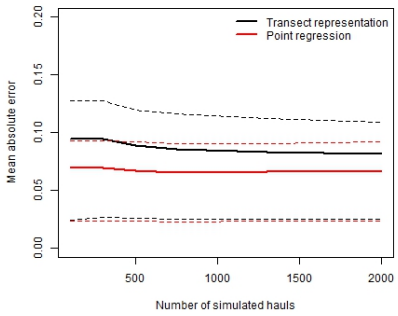


Figure 5.13. Mean absolute error of the results obtained using point-referenced GAMs (in red) and the algorithm proposed in Section 5.1 (in black). These errors were computed only in the cells were the algorithm had estimates. Solid lines represent the mean and dashed lines the 95% confidence intervals of the mean absolute errors.

chapter at $N = 100, 300, 500, 750, 1300, 2000$ number of simulated sampling hauls.

5.3.4 Discussion: simulated study 2

Results suggest that applying point-referenced regression methods produce good predictive estimates, despite the fact that in reality a fishery operation can fish at different habitats, not only in that of the centroid of the operation. Moreover, as in the simulation study 1, we have seen that point-referenced regression methods perform reasonably better than the algorithm in most cases.

Nevertheless, this study confirms that if enough data is available and cross-over enough times, the proposed algorithm can approximate fairly well the underlying spatial field. This is specially relevant due to the simplicity and almost null computational requirement of the algorithm.

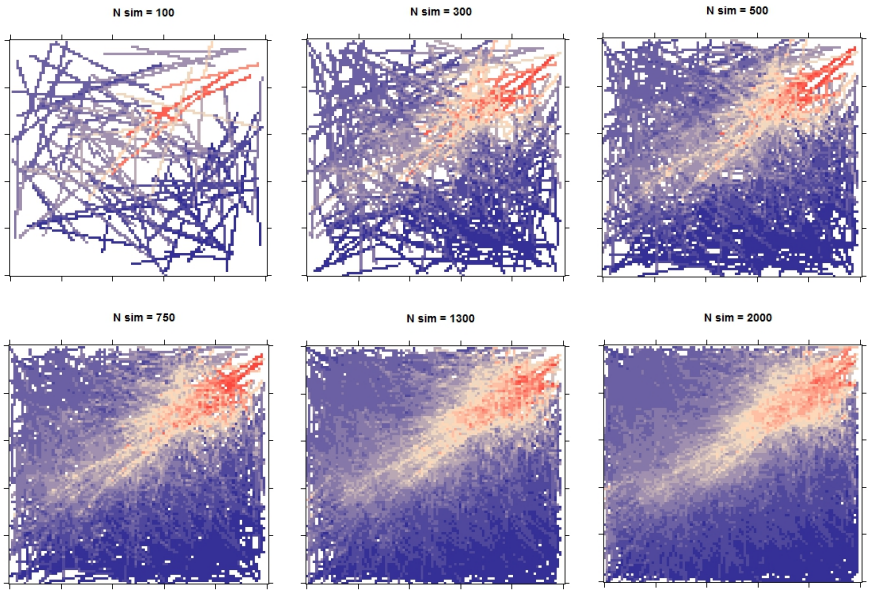


Figure 5.14. Results obtained by applying the proposed algorithm in the simulated field (bottom panel in Figure 5.12) at a different number of simulated sampling hauls.

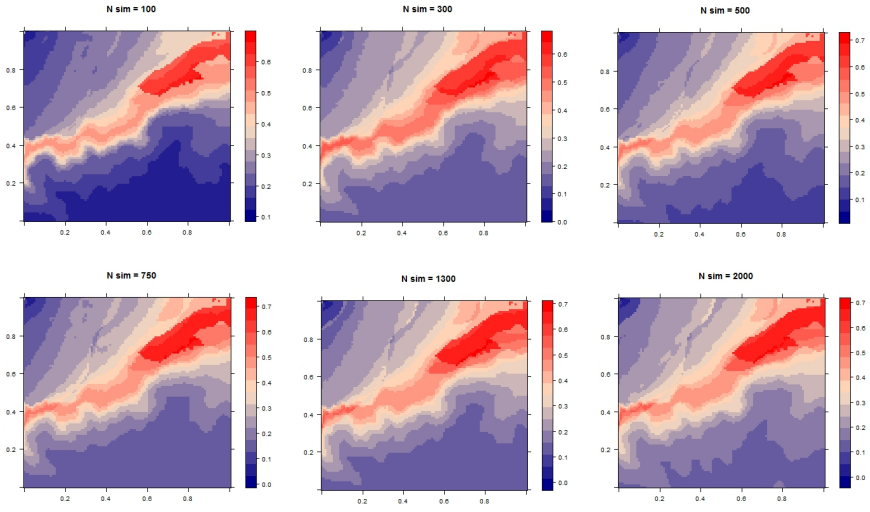


Figure 5.15. Results obtained by applying generalized additive models in the simulated spatial field (bottom panel in Figure 5.12) at a different number of simulated sampling hauls.

5.4 Case study

In this section we have presented a practical application of the proposed algorithm (5.1) against point-referenced regression models in a real case scenario. To do so we have used on-board sampling data in a small-scale fishing ground. Trawl discard data were collected according to the European Comission (2009) decision as commented in chapter 2. More, specifically, this study has been based on bottom trawl data collected in the southern Spanish Mediterranean Sea. The database was comprised by the starting and ending points of the fishing operations and the caught fish kilograms segregated by species.

The database contained a total of 218 observations and more than 100 species. For the purpose of this study we chose three economically important fish species; the blackbellied angler *Lophius budegassa*; the surmullet *Mullus surmuletus*; and the red mullet *Mullus barbatus*. Each species specific subset had a different distribution of zero observations, i.e. semi-continuous nature. For simplification reasons, only those hauls with non-zero abundance have been used in each of the subsets acknowledging that resulting modelling maps are not fully correct. As a consequence the size of the final subsets for each species were: 91 samples for the blackbellied angler, 172 samples for the surmullet and 174 samples for the red mullet.

The point-referenced modelling of fish distribution was performed using ordinary kriging, i.e. a constant mean (intercept) and a geostatistical term:

$$\begin{aligned} Y_{j_i} &\sim N(\mu_{j_i}, \rho_j) \\ \mu_{j_i} &= \alpha_j + W_i \\ \boldsymbol{W_{j_i}} &\sim N(0, Q(\kappa_j, \tau_j)) \\ \boldsymbol{\alpha} &\sim N(0, 0) \\ (2\log\boldsymbol{\kappa}, \log\boldsymbol{\tau}) &\sim MN(\mu, \rho) \end{aligned} \tag{5.2}$$

where j represents the species under study, i are the species specific observation locations, α represents the intercept of each of the models and $\boldsymbol{W}$ represent the geostatistical terms of each of the models. Finally, the prior

distributions of the models were the default implemented in R-INLA.

5.4.1 Results

The maps that result from the geostatistical models show obvious unnatural behaviours due to; the simplicity of the models proposed in equation (5.2); the stochastic reality of natural systems; the border effect; and the fact that we dismissed all hauls with zero value observation in each species specific subset of the data. However, acknowledging these factors, resulting maps give a good enough perspective of fish distribution patterns to compare with the maps generated by the algorithm.

Interestingly, results show quite similar patterns in the case of the black-bellied angler (maps on the top of Figure 5.16) and the surmullet (maps on the middle of Figure 5.16) for both methodologies. In the case of the red mullet (maps on the bottom of Figure 5.16), both methods show a marked hot-spot in the center of the study area that extends towards the west. This western semi-high abundance area cannot be very well identified in the map produced by the algorithm because there are not many fishing transects in the area (see the non-trawled area in dark-blue) and the algorithm does smooth its results, yet.

5.4.2 Discussion: real case scenario

This part of the study aimed to test the application of the proposed algorithm in a real case scenario. For this purpose we have used an on-board dataset located in a small fishing ground of the southern Spanish Mediterranean.

As expected, due to the low number of samples and their non-random distribution, results were not as satisfactory as in the simulation studies (Sections 5.2,5.3). However, the general distributional pattern of both approaches was rather similar.

As compared to the geostatistical approach, it is notorious that the algorithm does not smooth the estimated abundances. A post smoothing treatment, e.g. linear interpolation, of the resulting estimates could help us predict

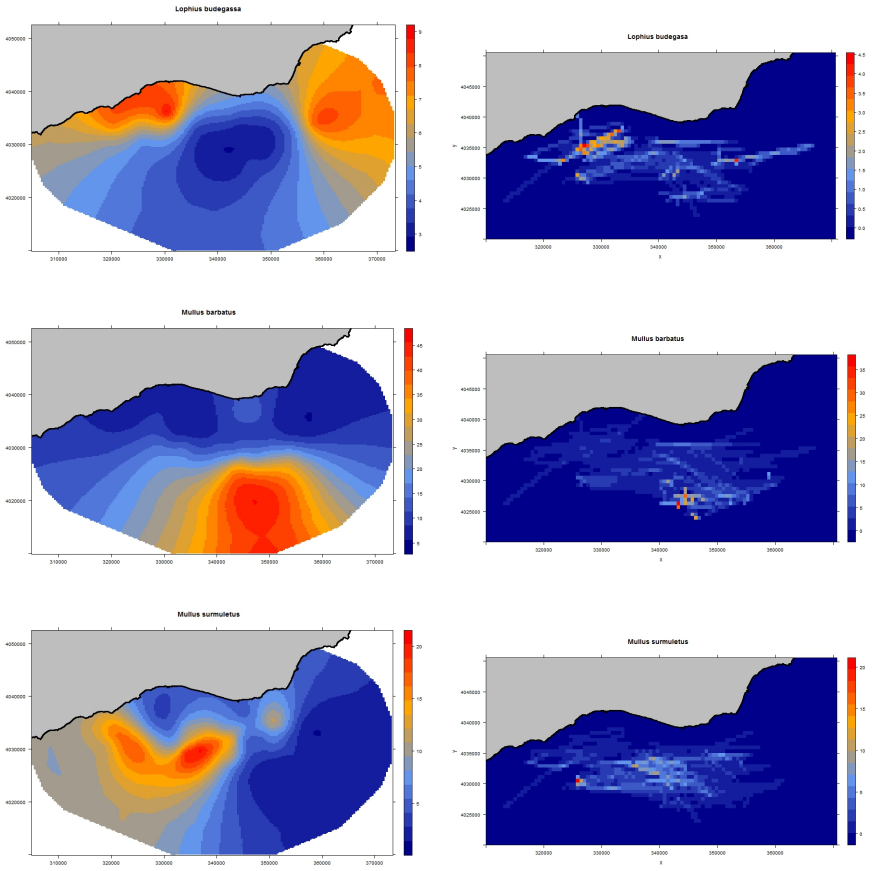


Figure 5.16. Results obtained by applying ordinary kriging (left) and the proposed algorithm (right) in the blackbellied angler *Lophius budegassa*, the surmullet *Mullus surmuletus* and the red mullet *Mullus barbatus* from top to bottom respectively.

at unsampled locations and smooth the discontinuous nature of its results. Further research should assess the performance of different smoothing techniques.

5.5 Conclusions

The aim of this chapter has been to test the performance of the point-referenced representation of fishery transects and the predictive capacity of usual kriging interpolation techniques based on Euclidean distances in small-scale fishery data. In this vein, we have also proposed an alternative algorithm to approximate the underlying spatial field by superposition of fishery transects in the study area.

To test the performance of these approaches we have used two simulation studies. In all cases, the point-referenced representation and regression have performed better than the transect algorithm. However as the complexity/heterogeneity of the simulated spatial field increased (see GF3 in Figure 5.2), mean absolute errors tended to resemble more. The performance of the algorithm and the geostatistical approach have also been tested and compared in a real dataset of the southern Spanish Mediterranean. Again, results have shown similar distributional patterns in both cases, although the results obtained by the algorithm were not as smooth as in the simulation studies. Further research on post spatial smoothing of the estimates could improve the actual estimations and allow us predict at unsampled locations.

Regarding the usual point-referenced representation of fishery transects, this chapter has allowed us to conclude that it performs rather well even if the spatial scale of the study area is small compared to the size of the transects. This conclusion applies to both; the characterization of process-covariate relationships based on their value in the centroid of the haul; and the use of geostatistical techniques based in Euclidean distances between the centroids of the fishery transects.

Concerning the performance of the algorithm in Section 5.1, this chapter showed that, even if the usual point-referenced regression method approximate

better the underlying spatial field, the algorithm performs fairly well too. In fact when the complexity of the underlying spatial field is high and enough data is available, it can perform almost as well as the geostatistical approach. Unfortunately, the proposed algorithm requires a good amount of data and a good crossover rate of the fishery transects.

As a final remark, we would like to mention that, taking into account the simplicity of the proposed algorithm, its overall performance is fairly good. The simplicity of the method is specially relevant when self-updating tools want to be created. In the commercial fishery world two sources of information are available to geolocate fishing operations, the Vessel Monitoring System (VMS) and the Automatic Identification System (AIS). These data, matched with log-book or sales notes data of the vessels as a proxy of the catch, could provide an immense database that could automatically provide updated fish distribution maps using the algorithm.

Chapter 6

Conclusions and future work

In this thesis, we have sought to explore different geostatistical model structures capable of answering the main questions raised by policy makers involved in the spatial management of fisheries:

- Identify economically and ecologically fishing-suitable areas with regards to fishery discards.
- Characterise spawning/nursery grounds in big marine spatial areas to assess the design of marine protected areas.
- Integrate fishery dependent (on-board) data in the assessment of marine spatial planning for target species.

To tackle these issues:

- We have proposed to assess the fishery discards spatial planning based on fishery discard proportions instead of the usual discard per unit effort units. To do so, we have used a Bayesian hierarchical spatial beta regression model.

- We have proposed the comparison of different spatio-temporal model structures to assess the distribution behaviour of fish.

- We have proposed the use of shared components to fit more appropriate process-covariate relationships in the usual species distribution Hurdle models.

- We have tested the use of Log-Gaussian Cox Process models to correct the model components of preferentially sampled fishery datasets (fishery dependent data) as an approach to fit appropriate small/meso-scale fish distribution maps.

- We have tested the performance of point-referenced regression models in fishery transect data, including Euclidean distance-based geostatistical models. Additionally, we have proposed an algorithm that approximate the underlying spatial field when enough data are available.

And we conclude that:

- Fishery discard proportions perform better than usual discard per unit effort units. Analytically, the across-vessel standardisation capacity of discard proportions is better, which may improve the predictive capacity of our models. Ecologically and economically, discard proportions allow a better assessment of fishing suitable areas because they asses the balance between marketed food biomass and biomass loss due to discards. Furthermore, Bayesian hierarchical spatial beta regression has proved to be an effective approach to deal with spatially sampled proportion data.

- The comparison of different spatio-temporal structures allow us to effectively infer the generic spatio-temporal behaviour of the species under study. This is specially relevant to design effective marine protected areas, where assessing the spatial persistence of spawning/nursery areas is particularly important.

- The assumption of independent processes in two-part or Hurdle models to deal with semi-continuous data is prone to overfit the data and therefore to produce incorrect predictions. In this regard, fitting shared components in the occurrence and the conditional-to-presence abundance processes can effectively improve fitted process-covariate relationships.
- The use of log-Gaussian Cox process models to incorporate fishermen's knowledge can effectively improve the predictive performance of preferentially sampled fish distribution models.
- The use of standard model selection scores, e.g. DIC, WAIC, CPO, to assess the predictive capacity of a model can be misleading when applied over semi-continuous and/or preferentially sampled datasets. Therefore, fishery experts knowledge is important to the model selection process.
- The point-referenced representation of fishery transects in the centroid of the fishing operation performs well to both; represent the sampling habitat of the transect; and apply Euclidean distance based geostatistics.

Future work

In overall, this PhD dissertation has proposed a number of model structures that have quite effectively tackled some of the main challenges in fisheries distribution modelling. However, the scope of research in the fishery field is still extensive. Here is a list of topics that we consider of special interest to fisheries science:

- Investigate new model structures that accommodate sporadic high catches into the models, i.e. the schooling effect.
- Apply the models proposed in this thesis to all the marine species available in order to visually assess the ecosystemic importance of different sub-areas.
- Propose multivariate models to investigate the relationship between two or more species in space.

- Investigate the performance of simple smoothing techniques to improve the estimates and prediction of the proposed algorithm in Chapter 5.
- Create an automatic application that map the small/meso-scale distribution of marine species using commercial fleet data. To do so we could use the proposed algorithm and its hypothetic smoothing improvement using; vessel monitoring system (VMS) or (AIS) data to locate the starting and ending points of each fishing operation; and log-books or sales notes data to approximate the catch of each fishing operation.

Bibliography

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964. 15

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998. 27

S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2014. 14, 15, 20, 39, 52

M. J. Bayarri and J. O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, pages 58–80, 2004. 97

M. W. Beck, K. L. Heck, K. W. Able, D. L. Childers, D. B. Eggleston, B. M. Gillanders, B. Halpern, C. G Hays, K. Hoshino, T. J. Minello, et al. The Identification, Conservation, and Management of Estuarine and Marine Nurseries for Fish and Invertebrates A better understanding of the habitats that serve as nurseries for marine species and the factors that create site-specific variability in nursery quality will improve conservation and management of these areas. *Bioscience*, 51(8):633–641, 2001. 72, 107

J. M. Bellido, A. Carbonell, M. Garcia, T. Garcia, and M. González. *The obligation to land all catches – consequences for the Mediterranean.* European

Parliament, Policy Department B: Structural and Cohesion Policies, 2014. ISBN 9789282356043. 33, 73

J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433): 109–122, 1996. 27

J. A. Bertrand, L. Gil De Sola, C. Papaconstantinou, G. Relini, and A. Souplet. The general specifications of the medits surveys. *Scientia marina*, 66 (S2):9–17, 2002. 72, 73

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991. 6, 13

D. Bolin and F. Lindgren. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, pages 523–550, 2011. 25, 26

M. Cameletti, F. Lindgren, D. Simpson, and H. Rue. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97(2):109–131, 2013. xiii, 25, 39

G. Carl and I. Kühn. Analyzing spatial autocorrelation in species distributions using gaussian and logit models. *ecological modelling*, 207(2):159–170, 2007. 13

S. Catchpole, T.and Elliot, D. Peach, and S. Mangi. Final Report: The English Discard Ban Trial. Technical report, Cefas, 2014. 33

T. L. Catchpole, C. L. J. Frid, and T. S. Gray. Discards in North sea fisheries: causes, consequences and solutions. *Marine Policy*, 29(5):421–430, 2005. 31, 33

C. Chatfield and M. Yar. Holt-Winters forecasting: some practical issues. *The Statistician*, pages 129–140, 1988. 12

L. Ciannelli, V. Bartolino, and K.S. Chan. Non-additive and non-stationary properties in the spatial distribution of a large marine fish population. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1743): 3635–3642, 2012. 8

J. S. Clark. Why environmental scientists are becoming Bayesians. *Ecology letters*, 8(1):2–14, 2005. 16, 37, 52

J. Claudet, C. W. Osenberg, L. Benedetti-Cecchi, P. Domenici, J. A. García-Charton, A. Pérez-Ruzafa, F. Badalamenti, J. Bayle-Sempere, A. Brito, F. Bulleri, et al. Marine reserves: size and age do matter. *Ecology letters*, 11(5):481–489, 2008. 2

F. Colloca, V. Bartolino, G. J. Lasinio, L. Maiorano, P. Sartor, G. Ardizzone, et al. Identifying fish nurseries using density and persistence measures. *Marine Ecology, Progress Series*, 381:287–296, 2009. 72, 73, 87, 107

A. Cosandey-Godin, E. T. Krainski, B. Worm, and J. M. Flemming. Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. *Canadian Journal of Fisheries and Aquatic Sciences*, 72(2):186–197, 2014. v, 33

N. Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990. 8, 13

C. Czado, T. Gneiting, and L. Held. Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261, 2009. 28

C. Q. Da-Silva and H. S. Migon. Hierarchical dynamic beta model. *Revstat Statistical Journal*, pages 1–17, 2016. 52

C. P. Dahlgren, G. T. Kellison, A. Adams, B. M. Gillanders, M. S. Kendall, C. A. Layman, J. A. Ley, I. Nagelkerken, and J. E. Serafy. Marine nurseries and effective juvenile habitats: concepts and applications. *Marine Ecology Progress Series*, 312:291–295, 2006. 72

B. Delaunay. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800):1–2, 1934. 25

P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47 (3):299–350, 1998. 111

P. J. Diggle, R. Menezes, and T. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C*, 52:191–232, 2010. v, 4, 55, 56, 58, 63, 68

P. H. C. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996. 11

European Comission. Comission Decision 2010/93/EU, adopting a multiannual Community programme for the collection, management and use of data in the fisheries sector for the period 2011-2013. Technical report, European Comission, 2009. 3, 31, 39, 62, 70, 128

European Comission. Regulation (EU) No 1380/2013 of the European Parliament and the council, on the Common Fisheries Policy, amending Council Regulations (EC) No 1954/2003 and (EC) No 1224/2009 and repealing Council Regulations (EC) No 2371/2002 and (EC) No 639/2004 and Council Decision 2004/585/EC. Technical report, EC (European Comission), 2013. 32

L. Fahrmeir and S. Lang. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied statistics*, pages 201–220, 2001. 42, 65, 84, 97

FAO. *Indicators for sustainable development of marine capture fisheries*, volume 8. FAO, 1999. 1

FAO. The state of world fisheries and aquaculture. Fisheries Department, Rome, 2002. 1

FAO. *The ecosystem approach to fisheries*. 2003. iii, 2, 107

J. Feekings, V. Bartolino, N. Madsen, and T. Catchpole. Fishery discards: factors affecting their variability within a demersal trawl fishery. *PLoS One*, 7(4):e36409, 2012. v, 33

J. Feekings, P. Lewy, N. Madsen, and C. T. Marshall. The effect of regulation changes and influential factors on Atlantic cod discards in the Baltic Sea demersal trawl fishery. *Canadian Journal of Fisheries and Aquatic Sciences*, 70(4):534–542, 2013. v, 33

P. B. Fenberg, J. E. Caselle, J. Claudet, M. Clemence, S. D. Gaines, J. A. García-Charton, E. J. Gonçalves, K. Grorud-Colvert, P. Guidetti, S. R. Jenkins, et al. The science of European marine reserves: Status, efficacy, and future needs. *Marine Policy*, 36(5):1012–1021, 2012. 72, 107

S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004. 34, 35, 36, 51

M. J. Fortin and M. R. T. Dale. Spatial autocorrelation in ecological studies: a legacy of solutions and myths. *Geographical Analysis*, 41(4):392–397, 2009. 12, 36

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006. 19

S. M. Garcia. *The ecosystem approach to fisheries: issues, terminology, principles, institutional foundations, implementation and outlook*. Number 443. Food & Agriculture Org., 2003. 2

G. Garofalo, T. Fortibuoni, M. Gristina, M. Sinopoli, and F. Fiorentino. Persistence and co-occurrence of demersal nurseries in the strait of sicily (central mediterranean): Implications for fishery management. *Journal of Sea Research*, 66(1):29–38, 2011. 72

S. Geisser. *Predictive inference*, volume 55. CRC Press, 1993. 27, 28

S. Geisser and W. F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979. 28

A. E. Gelfand. Hierarchical modeling for spatial data problems. *Spatial statistics*, 1:30–39, 2012. 20

A. Gelman, X. L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996. 27

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 2004. 21

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. 28, 30

B. Gillanders, K. Able, J. Brown, D. Eggleston, and P. Sheridan. Evidence of connectivity between juvenile and adult habitats for mobile marine fauna: an important component of nurseries. *Marine Ecology-Progress Series*, 247: 281–295, 2003. 72, 107

D. M. Gillis, E. K. Pikitch, and R. M. Peterman. Dynamic discarding decisions: foraging theory for high-grading in a trawl fishery. *Behavioral Ecology*, 6(2):146–154, 1995. 31

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. 29

G. Gorelli, F. Sardà, et al. Management strategies for the fishery of the red shrimp Aristeus antennatus in Catalonia (NE Spain). 2014. 62, 67

B. Grün, I. Kosmidis, and A. Zeileis. Extended beta regression in R: Shaken, Stirred, Mixed, and Partitioned. Technical report, Working Papers in Economics and Statistics, 2011. 37

A. Guisan, T. C. Edwards, and T. Hastie. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2):89–100, 2002. 11

A. K. Gupta and S. Nadarajah. *Handbook of beta distribution and its applications*. CRC Press, 2004. 34, 36

E. Gutiérrez-Peña, A. F. M. Smith, J. M. Bernardo, G. Consonni, P. Veronese, E. I. George, F. J. Girón, M. L. Martínez, G. Letac, and C. N. Morris. Exponential and Bayesian conjugate families: review and extensions. *Test*, 6(1):1–90, 1997. 18

M. S. Handcock and J. R. Wallis. An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89(426):368–378, 1994. 14

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990. 11

J. L. H. Heesen, N. Daan, and J. R. Ellis. *Fish Atlas of the Celtic Sea, North Sea and Baltic Sea: Based on International Research Vessel Data*. KNNV Publishing, 2015. iii

L. Held, I. Natário, S. E. Fenton, H. Rue, and N. Becker. Towards joint disease mapping. *Statistical methods in medical research*, 14(1):61–82, 2005. 96

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000. 96

T. Hengl, G. B. M. Heuvelink, M. P. Tadić, and E. J. Pebesma. Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theoretical and Applied Climatology*, 107(1-2):265–277, 2012. 52

R. Hilborn, K. Stokes, J. J. Maguire, T. Smith, L. W. Botsford, M. Mangel, J. Orensanz, A. Parma, J. Rice, J. Bell, et al. When can marine reserves

improve fisheries management? *Ocean & Coastal Management*, 47(3):197–205, 2004. 2

J. W. Hogan and N. M. Laird. Mixture models for the joint distribution of repeated measures and event times. *Statistics in medicine*, 16(3):239–257, 1997. 96

M. B. Hooten and N. T. Hobbs. A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28, 2015. 28

J. A. Hutchings. Collapse and recovery of marine fishes. *Nature*, 406(6798): 882–885, 2000. 75

ICES. ICES FishMap species factsheet: cod. Technical report, ICES, 2014. 75, 87, 107

J. B. Jones. Environmental impact of trawling on the seabed: a review. *New Zealand Journal of Marine and Freshwater Research*, 26(1):59–67, 1992. 5

K. Kelleher. *Discards in the world's marine fisheries: an update*, volume 470. FAO, 2005. 32

T. Kneib, J. Muller, and T. Hothorn. Spatial smoothing techniques for the assessment of habitat suitability. *Enviromental and Ecological Statistics*, 15 (3):343–364, 2008. 13

L. Knorr-Held and N. G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85, 2001. 91, 96

E. T. Krainski, F. Lindgren, D. Simpson, and H. Rue. *The R-INLA tutorial: SPDE models*, 2016. 25, 26

P. S. Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1(3):364–378, 1986. 20, 21

G. J. Lasinio, G. Mastrantonio, and A. Pollice. Discussing the "big n problem". *Statistical Methods & Applications*, 22(1):97–112, 2013. 20

H. Lassen and P. Medley. *Virtual population analysis: a practical manual for stock assessment*. Number 400. Food & Agriculture Org., 2001. 74

P. Legendre, M. R. T. Dale, M. J. Fortin, J. Gurevitch, M. Hohn, and D. Myers. The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, 25(5):601–615, 2002. 12, 36

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73:423–498, 2011. iv, v, 23, 24, 25, 26, 39

F. Liu and Y. Kong. *zoib: an R package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression*, 2015. 36, 37, 52

J. Lleonart. Review of the state of the world fishery resources. *FAO Fish Technical Paper*, 457:49–64, 2005. 73

S. C. Mangi, L. D. Rodwell, and C. Hattam. Assessing the impacts of establishing MPAs on fishermen and fish merchants: the case of lyme bay, uk. *Ambio*, 40(5):457–468, 2011. 2

G. Marc. *phenology: Tools to Manage a Parametric Function that Describes Phenology*, 2015. URL `http://CRAN.R-project.org/package=phenology`. R package version 4.1. 41

M. Marin, J. Rojas, D. Jaimes, H. A. G. Rojas, M. Corrales, M. F. Zarate, R. Duplat, . F. Villarraga, and E. Cepeda-Cuervo. *Bayesianbetareg: Bayesian Beta regression: joint mean and precision modeling*, 2014. URL `http://CRAN.R-project.org/package=Bayesianbetareg`. R package version 1.2. 37

T. G. Martin, B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters*, 8(11):1235–1246, 2005. 77, 95

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67: 68–83, 2013. 37

B. Matérn. *Spatial variation*, volume 36. Springer Science & Business Media, 2013. 14

F. Maynou and F. Sardà. Influence of environmental factors on commercial trawl catches of Nephrops norvegicus (L.). *ICES Journal of Marine Science: Journal du Conseil*, 58(6):1318–1325, 2001. 31

F. Maynou, J. Lleonart, and J. E. Cartes. Seasonal and spatial variability of hake (Merluccius merluccius l.) recruitment in the NW Mediterranean. *Fisheries Research*, 60(1):65–78, 2003. 72, 75

A. C. Morgan and G. H. Burgess. 11. fishery-dependent sampling: total catch, effort and catch composition. *Management techniques for elasmobranch fisheries*, page 182, 2005. 4

F. Muñoz, M. G. Pennino, D. Conesa, A. López-Quílez, and J. M. Bellido. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. *Stochastic Environmental Research and Risk Assessment*, 27:1171–1180, 2013.

T. Næs and B. H. Mevik. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4):413–426, 2001. 12

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. 10

R. Ospina and S. L. P. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, 2012. 52

I. Paradinas, D. Conesa, M. G. Pennino, F. Muñoz, A. M. Fernández, A. López-Quílez, and J. M. Bellido. Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. *Marine Ecology Progress Series*, 528:245, 2015. 52, 108

I. Paradinas, M. Marín, M. G. Pennino, A. López-Quílez, D. Conesa, D. Barreda, M. Gonzalez, and J. M. Bellido. Identifying the best fishing-suitable areas under the new European discard ban. *ICES Journal of Marine Science: Journal du Conseil*, 2016.

I. Paradinas, M. G. Pennino, M. Marín, A. López-Quílez, J. M. Bellido, and D. Conesa. Modelling spatially sampled proportion processes. *Revstat Statistical Journal*, In press.

D. Pauly, V. Christensen, S. Guénette, T. J. Pitcher, U. R. Sumaila, C. J. Walters, R. Watson, and D. Zeller. Towards sustainability in world fisheries. *Nature*, 418(6898):689–695, 2002. 1

M. G. Pennino, F. Muñoz, D. Conesa, A. López-Quílez, and J. M. Bellido. Modeling sensitive elasmobranch habitats. *Journal of sea research*, 83:209–218, 2013.

M. G. Pennino, F. Muñoz, D. Conesa, A. López-Quílez, and J. M. Bellido. Bayesian spatio-temporal discard model in a demersal trawl fishery. *Journal of Sea Research*, 90:44–53, 2014. v, 33, 39, 42, 48, 50, 51

E. K. Pikitch, C. Santora, E. A. Babcock, A. Bakun, R. Bonfil, D. O. Conover, P. Dayton, P. Doukakis, D. Fluharty, B. Heneman, E. D. Houde, J. Link, P. A. Livingston, M. Mangel, M. K. McAllister, J. Pope, and K. J. Sainsbury. Ecosystem-based fishery management. *Science*, 305(Weekly):346–347, 2004. 2

M. Plummer. Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008. 29

Poseidon. A case study review of the potential economic implications of the proposed CFP landings obligation. Technical report, Poseidon, Aquatic Resource Management, 2013. 33

Z. C. Quiroz, M. O. Prates, and H. Rue. A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics*, 71(1):208–217, 2015. ISSN 1541-0420. 100, 106, 108

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`. 58

L. Recasens, A. Lombarte, B. Morales-Nin, and G.J. Tores. Spatiotemporal variation in the population structure of the European hake in the NW Mediterranean. *Journal of fish biology*, 53(2):387–401, 1998. 72, 75

M. J. Rochet and V. M. Trenkel. Factors for the variability of discards: assumptions and field evidence. *Canadian Journal of Fisheries and Aquatic Sciences*, 62(1):224–235, 2005. 42, 51

M. Roos and L. Held. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278, 2011. 61

H. Rue and L. Held. *Gaussian Markov Random Fields. Theory and Applications*. Chapman and Hall/CRC, 2005. 21, 42, 97

H. Rue, S. Martino, and N. Chopin. 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, 2009. iv, v, 20, 21, 23, 38, 105

H. Rue, S. Martino, D. Mondal, and N. Chopin. Discussion on the paper Geostatistical inference under preferential sampling by Diggle, Menezes and Su. *Journal of the Royal Statistical Society, Series C*, 52:221–223, 2010. 56, 58, 63

F. Sardà, M. Coll, J. J. Heymans, and K. I. Stergiou. Overlooked impacts and challenges of the new European discard ban. *Fish and Fisheries*, 16(1): 175–180, 2015. 32

M. Schlather, A. Malinowski, P. J. Menck, M. Oesting, K. Strokorb, et al. Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, 63(8):1–25, 2015. 58

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978. 27

S. Sigurðardóttir, E. K. Stefánsdóttir, H. Condie, S. Margeirsson, T. L. Catchpole, J. M. Bellido, S. Q. Eliasen, R. Goñi, N. Madsen, A. Palialexis, S. S. Uhlman, V. Vassilopoulou, and Rochet M. J. How can discards in European fisheries be mitigated? Strengths, weaknesses, opportunities and threats of potential mitigation methods. *Marine Policy*, 51:366–374, 2015. 33

D. Simpson, J. Illian, F. Lindgren, S. H. Sørbye, and H. Rue. Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 1–19, 2011. 38

M. Smithson and J. Verkuilen. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54, 2006. 36

S. H. Sørbye and H. Rue. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014. 42, 97

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–616, 2002. 28, 61

D. M. Stasinopoulos and R. A. Rigby. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 2007. 37

STECF. Review of list of surveys at sea (Appendix XIV of EU Commission Regulation N°1581/2004) with their priorities (SGRN 07-01). Technical report, cientific, Technical and Economic Committee for Fisheries (STECF), 2007. 4

R. G. D. Steel, J. H. Torrie, and D. A. Dickey. *Principles and procedures of statistics: A biological approach*. McGraw-Hill, 1997. 35

A. Stein, C. G. Kocks, J. C. Zadoks, H. D. Frinking, M. A. Ruissen, and D. E. Myers. A geostatistical analysis of the spatio-temporal development of downy mildew epidemics in cabbage. *Phytopathology*, 84(10):1227–1238, 1994. 52

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 1999. 14

M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296, 2004. 39

B. M. Taylor. Auxiliary Variable Markov Chain Monte Carlo for Spatial Survival and Geostatistical Models. *arXiv preprint arXiv:1501.01665*, 2015. 20

The International Bottom Trawl Survey Working Group. *Manual for the International Bottom Trawl Surveys*. ICES, 2012. 72, 76

T. Therneau, T. Lumley, K. Halvorsen, and K. Hornik. *date: Functions for handling dates*, 2014. URL `http://CRAN.R-project.org/package=date`. R package version 1.2-34. S original by Terry Therneau, R port by Thomas Lumley, Kjetil Halvorsen, and Kurt Hornik. 41

L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81 (393):82–86, 1986. 20, 21

W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, pages 234–240, 1970. 13

K. Tsagarakis, A. Palialexis, and V. Vassilopoulou. Mediterranean fishery discards: review of the existing knowledge. *ICES Journal of Marine Science: Journal du Conseil*, page fst074, 2013. 34

A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834, 2004. 91

A. Vehtari, J. Ojanen, et al. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012. 28

M. Viana, A. L. Jackson, N. Graham, and A. C. Parnell. Disentangling spatio-temporal processes in a hierarchical system: a case study in fisheries discards. *Ecography*, 36(5):569–578, 2013a. v, 33

M. Viana, L. McNally, N. Graham, D. G. Reid, and A. L. Jackson. Ignoring discards biases the assessment of fisheries' ecological fingerprint. *Biology letters*, 9(6):20130812, 2013b. 32

R. Vilela and J. M. Bellido. Fishing suitability maps: helping fishermen reduce discards. *Canadian Journal of Fisheries and Aquatic Sciences*, (ja), 2015. 33, 34

C. Walters and J. J. Maguire. Lessons for stock assessment from the northern cod collapse. *Reviews in fish biology and fisheries*, 6(2):125–137, 1996. 1

D. I. Warton and F. K. C. Hui. The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10, 2011. 35, 37, 51, 52

L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013. 16

S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594, 2010. 28

W. W. S. Wei. *Time series analysis*. Addison-Wesley publ Reading, 1994. 12

I. D. Williams, W. J. Walsh, J. T. Claisse, B. N. Tissot, and K. A. Stamoulis. Impacts of a Hawaiian marine protected area network on the abundance and fishery sustainability of the yellow tang, Zebrasoma flavescens. *Biological Conservation*, 142(5):1066–1073, 2009. 2

C. J. Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1):79–82, 2005. 61, 116

K. Wilson and I. C. W. Hardy. Statistical analysis of sex ratios: an introduction. 2002. 35, 53

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011. 37