Spatial Statistics 70 (2025) 100930

Contents lists available at ScienceDirect

Spatial Statistics

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

Joint model for zero-inflated data combining fishery-dependent and fishery-independent sources☆

Daniela Silva [a,b,*], Raquel Menezes [c], Gonçalo Araújo [d,e,f], Renato Rosa [g], Ana Moreno [a], Alexandra Silva [a], Susana Garrido [a]

[a] Division of Modeling and Management of Fishery Resources, Portuguese Institute for the Sea and Atmosphere (IPMA), Lisbon, Portugal
[b] Centre of Mathematics, Minho University, Braga, Portugal
[c] Centre of Mathematics, Minho University, Guimarães, Portugal
[d] Nova School of Business and Economics, Nova University Lisbon, Lisbon, Portugal
[e] CCMar - Centre of Marine Sciences, University of Algarve, Faro, Portugal
[f] University of Algarve, Faro, Portugal
[g] Centre for Business and Economics Research, Coimbra University, Coimbra, Portugal

ARTICLE INFO

Keywords:
Species distribution model
Integrating data sources
Preferential sampling
Geostatiscal modeling
Fish data

ABSTRACT

Accurately identifying spatial patterns of species distribution is crucial for scientific insight and societal benefit, aiding our understanding of species fluctuations. The increasing quantity and quality of ecological datasets present heightened statistical challenges, complicating spatial species dynamics comprehension. Addressing the complex task of integrating multiple data sources to enhance spatial fish distribution understanding in marine ecology, this study introduces a pioneering five-layer Joint model. The model adeptly integrates fishery-independent and fishery-dependent data, accommodating zero-inflated data and distinct sampling processes. A comprehensive simulation study evaluates the model performance across various preferential sampling scenarios and sample sizes, elucidating its advantages and challenges. Our findings highlight the model's robustness in estimating preferential parameters, emphasizing differentiation between presence–absence and biomass observations. Evaluation of estimation of spatial covariance and prediction performance underscores the model's reliability. Augmenting sample sizes reduces parameter estimation variability, aligning with the principle that increased information enhances certainty. Assessing the contribution of each data source reveals successful integration, providing a comprehensive representation of biomass patterns. Empirical application within a real-world context further solidifies the model's efficacy in capturing species' spatial distribution. This research advances methodologies for integrating diverse datasets with different sampling natures further contributing to a more informed understanding of spatial dynamics of marine species.

☆ This article is part of a Special issue entitled: 'SPASTA_Metma: health and environment' published in Spatial Statistics.

* Corresponding author at: Division of Modeling and Management of Fishery Resources, Portuguese Institute for the Sea and Atmosphere (IPMA), Lisbon, Portugal.

E-mail addresses: daniela.dasilva@ipma.pt (D. Silva), rmenezes@math.uminho.pt (R. Menezes), goncalo.araujo@novasbe.pt (G. Araújo), renato.rosa@novasbe.pt (R. Rosa), amoreno@ipma.pt (A. Moreno), asilva@ipma.pt (A. Silva), susana.garrido@ipma.pt (S. Garrido).

https://doi.org/10.1016/j.spasta.2025.100930
Received 29 November 2024; Received in revised form 1 September 2025; Accepted 4 September 2025
Available online 11 September 2025

1. Introduction

Scientific tools capable of identifying species distribution patterns are important not only for ecological understanding but also for effective resource management. This is particularly true in fisheries science, where precise knowledge of species abundance and distribution supports sustainable exploitation and conservation efforts. For example, understanding the spatio-temporal dynamics of fish populations is crucial for setting quotas, monitoring biodiversity, and anticipating ecological shifts driven by climate or anthropogenic pressures. However, this task is increasingly challenging due to the growing volume, complexity, and heterogeneity of ecological datasets (Martínez-Minaya et al., 2018).

In fisheries science, data collection is typically achieved through two main sources, fishery-independent data (FID) and fishery-dependent data (FDD).

FID refers to information collected through methods that are independent of fishing activities. It involves conducting surveys or research specifically designed to assess fish populations, often using standardized sampling techniques, such as trawl surveys or acoustic monitoring. These data provides a more reliable assessment of fish population size, abundance, and distribution (Ault et al., 1998), as it is not influenced by fishing behavior or economic interests. In contrast, FDD refers to information collected directly from commercial fishing activity —- e.g. logbooks, fishery surveys, or monitoring programs. It reflects the actual catches by fishers, helps estimating fishing mortality rates and monitor changes in fishing effort over time (Rosenberg et al., 2005).

Research surveys are usually limited to specific operational time frames, occurring once or twice a year, covering a larger spatial region. Conversely, FDD is typically available at higher temporal resolution due to the nature of the underlying activity, but it is also subject to preferential sampling (PS) once fishers tend to operate in areas with known or expected high yields, resulting in non-random sampling patterns.

Because FID and FDD capture different facets of the underlying fish population dynamics – one unbiased but sparse, the other dense but biased – their integration into a single statistical frameworks can yield more robust inference. In particular, having access to multiple data types simultaneously improves the identifiability of spatial patterns and model parameters (Steele and Tucker, 2008; Kirk et al., 2012; Ferreras et al., 2021), enabling more accurate estimation of species distributions under realistic sampling conditions (Doser et al., 2021; Tehrani et al., 2022). However, joint modeling is challenging as it requires approaches capable of handling different data structures, sampling designs, and biases. For instance, classical tools handle standardized sampling designs but not the preferential nature of commercial data which affects both the resulting predictive surface (Diggle et al., 2010) and parameters estimation (Gelfand et al., 2012).

The concept of PS was formalized in the statistics literature by Diggle et al. (2010), who introduced a hierarchical modeling framework in which the locations of observations are viewed as a realization of a spatial point process whose intensity depends on an unobserved latent process. This latent process, in turn, is directly linked to the true underlying spatial distribution of the quantity of interest (e.g., species abundance or biomass). This two-part model framework, typically implemented using a log-Gaussian Cox process (LGCP), allows for explicit modeling of the sampling mechanism and its interaction with the latent field.

Subsequent research has extended this work in several directions. Pati et al. (2011) incorporated covariates and hierarchical random effects into the LGCP framework. Despite this progress, challenges remain — particularly in accounting for zero-inflation, a common feature of ecological data due to true absences or non-detection. To address this, zero-inflated (ZI) models (Lambert, 1992) and their hierarchical variants have become widely used in species distribution modeling (MacKenzie et al., 2006; Guillera-Arroita and Lahoz-Monfort, 2012), offering improved model fit and interpretability when excess zeros are present.

More recent models have specifically tackled the joint modeling of FID and FDD. For example, Rufener et al. (2021) introduced a three-layer hierarchical model combining scientific survey data with commercial catch data, incorporating distinct observation models and catchability parameters. Similarly, Alglave et al. (2022) developed a four-layer model using a shared latent biomass field, modeled with spatial random effects, and inhomogeneous Poisson processes (IPPs) to describe the sampling mechanisms. The degree of PS is quantified by the scaling parameter between the process of interest and the sampling process. This approach highlights the growing recognition of PS as a fundamental issue in joint modeling and species distribution inference.

While the aforementioned studies offer valuable contribution, there is still a need for flexible and interpretable joint models that can: (i) distinguish between presence/absence and biomass/abundance data; (ii) account for different sampling processes across FID and FDD; (iii) incorporate PS effects in a structured way; and (iv) handle zero-inflation and spatial autocorrelation jointly.

In this study, we present a novel spatial Joint model designed for ZI data, aiming to infer the spatial distribution of fish by integrating information from both FID and FDD, while effectively addressing the challenges associated with PS. Our model consists of a five-layer structure, separating the presence/absence observations and biomass/abundance processes and explicity modeling the distinct observation mechanisms associated to each data source. PS is incorporated via a spatially structured latent process that influences both the observation and sampling layers.

To evaluate the performance, challenges, and advantages of our Joint model, we conducted a simulation study under multiple PS scenarios and varying sample sizes. We also applied the proposed model to a real-world case study: the spatial distribution of European sardine (*Sardina pilchardus*, Walbaum 1792) along the southern coast of Portugal, using both scientific survey data and commercial fisheries data represented in Fig. 1.

The paper is organized into five sections. Section 2 introduces the proposed Joint modeling framework, including its components, assumptions, and inference methodology. Section 3 describes the case study on European sardine distribution and presents the corresponding results. Section 4 outlines the simulation study, including its configuration and results. Section 5 discusses the findings and implications and explores potential avenues for future research.

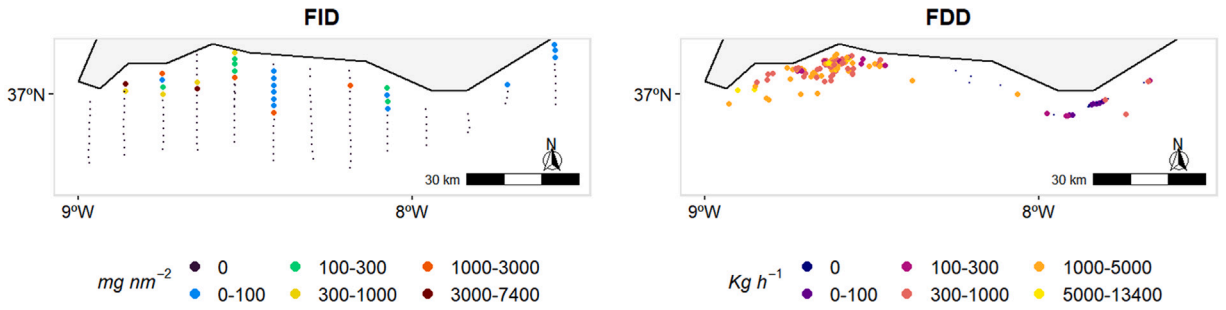


Fig. 1. Observed biomass index of sardine off the southern coast of Portugal during 2017 from FDD source in $\mathrm{mg}^2\,\mathrm{nm}^{-2}$ (first column) and from FDD source in $\mathrm{Kg\,h}^{-1}$ (second column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

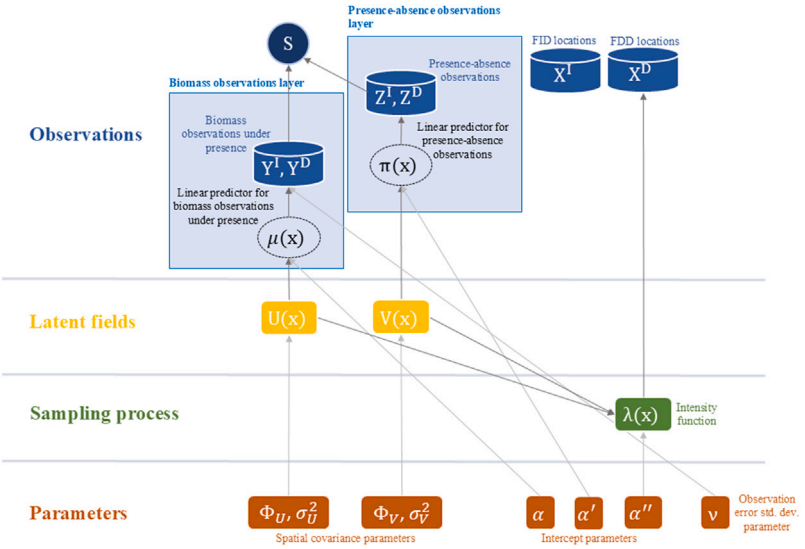


Fig. 2. Diagram of the joint model, including the preferential sampling (PS) for fishery-dependent data (FDD), and distinguishing between presence–absence and biomass observations to address zero-inflated (ZI) data.

2. Proposed model and inference

2.1. Joint model

To infer species distribution taking advantage of the information provided by both FID and FDD sources, we propose a joint hierarchical model with five layers: presence–absence observations, biomass observations under presence, the latent, the sampling process, and the parameters (Fig. 2).

2.1.1. Observations

Let us denote the spatial biomass process by $\mathbf{S} = \{s(\mathbf{x}_1), \dots, s(\mathbf{x}_n)\}$ at location $\mathbf{x}_i \in \mathcal{A} \subset \mathbb{R}^2$, where $\mathcal{A}$ is the study region and n represents dimension of the data. The presence–absence process (PAP) $\mathbf{Z} = \{z(\mathbf{x}_1), \dots, z(\mathbf{x}_n)\}$, with presence probability $\pi(\mathbf{x}_i)$, takes the binary value 0 if no species was observed at location $\mathbf{x}_i$, and 1 otherwise. The biomass process under the presence $\mathbf{Y} = \mathbf{S}|(\mathbf{Z} = 1) = \{y(\mathbf{x}_1), \dots, y(\mathbf{x}_n)\}$ takes the strictly positive values of the biomass process $\mathbf{S}$.

Consequently, the distribution of the process of interest $\mathbf{S}$ is given by the product of the distribution of the PAP $\mathbf{Z}$ and the distribution of the biomass process under the presence $\mathbf{Y}$ such that

$$P(S_i = s_i) = P(S(\mathbf{x}_i) = s(\mathbf{x}_i)) = \begin{cases} 1 - \pi(\mathbf{x}_i) & \text{if} \quad s(\mathbf{x}_i) = 0 \\ \pi(\mathbf{x}_i)\ p(s(\mathbf{x}_i)|\mu(\mathbf{x}_i)) & \text{if} \quad s(\mathbf{x}_i) > 0 \end{cases} \tag{1}$$

where $p(s(\mathbf{x}_i)|\mu(\mathbf{x}_i))$ represents a probability mass function for $\mathbf{Y}$, the biomass process under presence (e.g., Gamma and Log-normal distributions). The same is observed for the main statistics of the interest process, mean $E[\mathbf{S}] = E[\mathbf{Z}]E[\mathbf{Y}]$ and median $F_{\mathbf{S}}(0.5) = E[\mathbf{Z}]F_{\mathbf{Y}}(0.5)$ (Silva et al., 2024).

We propose a two-part model ((2) and (3)) designed for accommodate ZI data, taking into account the distinct conditions influencing both the PAP (2) and the biomass process under presence of the species (3). PAP $\mathbf{Z}$ is assumed to come from a Bernoulli distribution with probability π such that $Z(\mathbf{x}_i) \sim Bernoulli(\pi(\mathbf{x}_i))$. The biomass process under the presence $\mathbf{Y}$ requires a continuous distribution such as Gamma distribution with shape parameter $a(\mathbf{x}_i) = \mu(\mathbf{x}_i)^2/\upsilon^2$ and scale parameter $b(\mathbf{x}_i) = \upsilon^2/\mu(\mathbf{x}_i)$, that is, $Y(\mathbf{x}_i) \sim Gamma(a(\mathbf{x}_i), b(\mathbf{x}_i))$. $\mu(\mathbf{x}_i)$ and υ represent the mean for location $\mathbf{x}_i$ and the standard deviation of biomass under presence, respectively.

$$logit(\pi(\mathbf{x}_i)) = \alpha' + V(\mathbf{x}_i) \tag{2}$$

$$log(\mu(\mathbf{x}_i)) = \alpha + U(\mathbf{x}_i). \tag{3}$$

α and α' denote the intercepts of the linear predictors for the respective processes. $\mathbf{U}(\mathbf{X})$ and $\mathbf{V}(\mathbf{X})$ are spatial latent fields as described below.

2.1.2. Latent fields

The spatial latent fields $\mathbf{U}(\mathbf{X})$ and $\mathbf{V}(\mathbf{X})$ are modeled as independent to reflect the distinct spatial structures and ecological drivers underlying the PAP $\mathbf{Z}$ and biomass process $\mathbf{Y}$, respectively. This assumption is supported by empirical findings in Silva et al. (2024), which demonstrate that the spatial patterns associated with PAP and biomass exhibit different degrees of spatial variability and are influenced by ecosystem conditions in non-overlapping ways. Each latent field denotes the spatial dependency and variation that is accounted for through a zero-mean Gaussian Markov Random Field (GMRF) with a Matérn covariance function $M(\mathbf{x}, \mathbf{x}'; \phi_j, \sigma_j, \nu)$, $j = \{\mathbf{U}, \mathbf{V}\}$ with spatial range ϕ_j, marginal variance σ_j^2 and smoothing parameter ν such that $\mathbf{U}(\mathbf{X}), \mathbf{V}(\mathbf{X}) \sim GMRF(\mathbf{0}, M(\mathbf{x}, \mathbf{x}'; \phi_j, \sigma_j, \nu))$.

2.1.3. Sampling process

Let us denote the spatial point processes underlying FID and FDD by $\mathbf{X}^I$ and $\mathbf{X}^D$, respectively.[1] The intensity of a point process can exhibit either spatial constancy, yielding a homogeneous or stationary pattern, or spatial variability with a discernible spatial trend, resulting in an inhomogeneous pattern.

FID typically provides a more objective characterization of fish distribution, as its sampling locations $\mathbf{x}^I$ are selected independently of the underlying ecological process $\mathbf{S}$ - often through randomized or systematic designs. In this context, if the sampling is randomized and spatially uniform, $\mathbf{X}^I$ can be represented as a realization of a Homogeneous Poisson Process (HPP) with constant intensity λ^{HPP}. However, since the location process is assumed to be independent of the ecological processes of interest, its contribution to the overall likelihood is constant and thus does not influence parameter estimation. Conversely, if the sampling follows a systematic scheme, the locations $\mathbf{X}^I$ are deterministic and treated as fixed in the analysis. In both cases, the sampling design does not convey information about the latent ecological fields, and so $\mathbf{X}^I$ is either omitted from the likelihood or conditioned upon during model fitting. This treatment preserves inferential validity while accommodating a range of FID survey strategies.

In contrast, the assumption of independence between sampling design and ecological processes is often violated in the case of FDD. Fishery-dependent locations $\mathbf{x}^D$ are typically influenced by fisher behavior, regulatory constraints, and prior knowledge of high-yield areas — introducing a PS mechanism. As a result, the distribution of $\mathbf{X}^D$ is stochastically dependent on the ecological processes of interest, particularly abundance or presence patterns.

To account for this dependence, we follow the framework of Diggle et al. (2010) and model $\mathbf{X}^D$ as a realization of an IPP. Here, the spatial intensity function $\lambda(\mathbf{x}^D)$ is explicitly linked to the latent fields $\mathbf{U}(\mathbf{X})$ and $\mathbf{V}(\mathbf{X})$ such that $\mathbf{X}^D \sim IPP(\lambda(\mathbf{x}^D))$. This formulation enables the model to capture the non-random structure of the sampling process, correct for biases in inference and prediction, and quantify the strength of PS via scaling parameters.

The logarithm of the intensity function for location $\mathbf{x}_i$ is expressed as:

$$log(\lambda(\mathbf{x}_i^D)) = \alpha'' + \beta' V(\mathbf{x}_i^D) + \beta U(\mathbf{x}_i^D). \tag{4}$$

Therefore, the intensity function of the IPP is described by the logarithmic link function of the linear combination of the intercept α'' and the latent effects $U(\mathbf{x}_i^D)$ and $V(\mathbf{x}_i^D)$. β' and β quantify the degree of PS by scaling the relationship between the local fishing intensity and the local value of each process of interest $\mathbf{Z}$ and $\mathbf{Y}$, respectively.

Fig. 3 illustrates a realization of each process contributing to the determination of a single realization of the ecological process $\mathbf{S}$.

2.2. Inference

Model estimation and parameter inference are carried out by maximizing the Laplace approximation to the marginal likelihood (Skaug and Fournier, 2006). The full joint distribution and likelihood formulation are provided in Section 2.2.1. Standard errors for the parameter estimates are obtained from the inverse of the observed Hessian matrix. The model is implemented in C++ and fitted using the TMB (Template Model Builder) package in R (Kristensen et al., 2016), which facilitates efficient computation through automatic differentiation and sparse matrix operations.

To derive each spatial latent field, we employ an approximation method based on stochastic partial differential equations (SPDEs), as introduced by Lindgren et al. (2011). SPDE approach enables the approximation of a spatial continuous field, represented by

[1] I is the acronym for fishery-Independent data in $\mathbf{X}^I$ and D for fishery-Dependent data in $\mathbf{X}^D$.

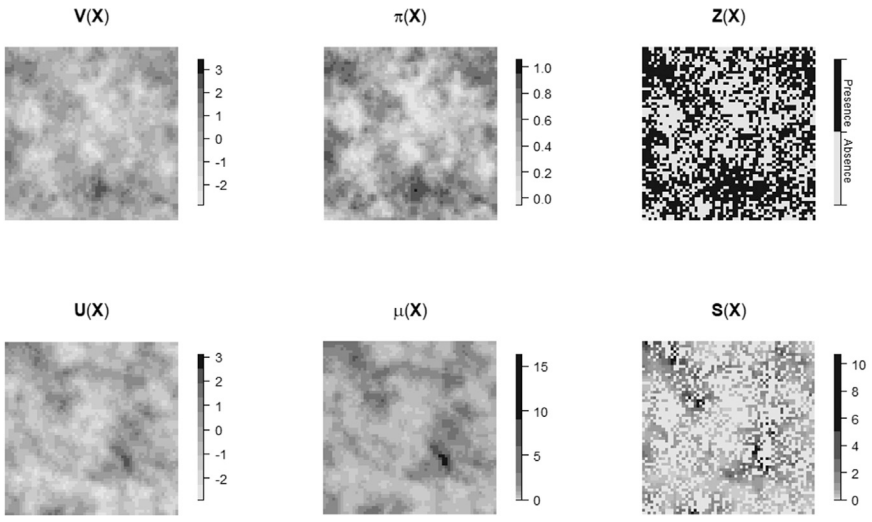


Fig. 3. Example of simulated latent fields. $\mathbf{V}(\mathbf{X})$ and $\mathbf{U}(\mathbf{X})$ represent realizations of the simulated GMRFs. $\pi(\mathbf{X})$ and $\mu(\mathbf{X})$ are the probability field of species presence and mean field of species biomass under presence given $\mathbf{V}(\mathbf{X})$ and $\mathbf{U}(\mathbf{X})$, respectively. $\mathbf{Z}(\mathbf{X})$ represents realizations of the PAP determined by $\pi(\mathbf{X})$. The biomass process $\mathbf{S}(\mathbf{X})$ is derived as $\mathbf{S}(\mathbf{X}) = \mathbf{Z}(\mathbf{X}) \cdot \mathbf{Y}(\mathbf{X})$, where $\mathbf{Y}(\mathbf{X})$ represents realizations of the biomass process under presence determined by $\mu(\mathbf{X})$.

a Matérn covariance function, to a GMRF which is discretly indexed. The adoption of this approximation is motivated by its computational efficiency.

Parameterization of the latent fields is performed in terms of marginal variance σ_j^2 and range of influence ϕ_j, enhancing the model interpretability and computational advantages. Under certain circumstances, a reparametrization of these parameters (as defined in Section 2.1.2) proves to be advantageous. In this study, mainly dictated by the functionality of the `TMB` `R` package, our proposed model is implemented using the parameters κ_j and τ_j. Subsequently, assuming the fixed value of $\nu = 1$, a reparametrization is undertaken to enhance the interpretability of the results according to $\phi_j = \frac{\sqrt{8\nu}}{\kappa_j}$ and $\sigma_j = \frac{\sqrt{\Gamma(\nu)}}{\sqrt{\Gamma(\nu+1)}\cdot\kappa_j^{\nu}\cdot\tau_j\cdot\sqrt{4\pi}}$, $j = \{\mathbf{U}, \mathbf{V}\}$.

2.2.1. Likelihood of the model

The joint distribution (5) is determined by the distribution of the biomass process under the presence (conditioned on the GMRF $\mathbf{U}$ and the sampling processes $\mathbf{X}^I$ and $\mathbf{X}^D$), the distribution of the PAP (conditioned on the GMRF $\mathbf{V}$ and the sampling processes $\mathbf{X}^I$ and $\mathbf{X}^D$), the distributions of the sampling processes $\mathbf{X}^D$ (conditioned on the GMRFs $\mathbf{U}$ and $\mathbf{V}$) and $\mathbf{X}^I$, and the distributions of both GFs $\mathbf{U}$ and $\mathbf{V}$. Each of these distributions is characterized by a specific expression, with the distribution of $\mathbf{X}^I$ remaining constant as it is assumed as a HPP.

$$\begin{aligned}\left[\mathbf{Y}, \mathbf{Z}, \mathbf{X}^D, \mathbf{X}^I, \mathbf{U}, \mathbf{V}\right] &= \left[\mathbf{Y}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{U}\right]\left[\mathbf{Z}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{V}\right]\left[\mathbf{X}^D, \mathbf{X}^I, \mathbf{U}, \mathbf{V}\right] \\ &= \left[\mathbf{Y}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{U}\right]\left[\mathbf{Z}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{V}\right]\left[\mathbf{X}^D|\mathbf{U}, \mathbf{V}\right][\mathbf{U}][\mathbf{V}]\left[\mathbf{X}^I\right]. \end{aligned} \tag{5}$$

Given the result presented in (5) and denoting the space of parameters as Θ, the likelihood of the model is

$$\mathcal{L}(\Theta) = \mathcal{L}(\mu, \sigma; \mathbf{y}) \times \mathcal{L}(\pi; \mathbf{z}) \times \mathcal{L}(\lambda; \mathbf{x}) \times \mathcal{L}(\sigma_U, \phi_U) \times \mathcal{L}(\sigma_V, \phi_V), \tag{6}$$

where $\mathcal{L}(\mu, \sigma; \mathbf{y})$ represents the likelihood for $\mathbf{Y}|\mathbf{X}^D$, $\mathbf{X}^I, \mathbf{U}$ (A.1), the likelihood for $\mathbf{Z}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{V}$ is denoted by $\mathcal{L}(\pi; \mathbf{z})$ (A.3), $\mathcal{L}(\lambda; \mathbf{x})$ identifies the likelihood for $\mathbf{X}^D|\mathbf{U}, \mathbf{V}$ (A.5), and $\mathcal{L}(\sigma_U, \phi_U)$ and $\mathcal{L}(\sigma_V, \phi_V)$ represent the likelihoods for $\mathbf{U}$ and $\mathbf{V}$ (A.7), respectively. Hence, the joint log-likelihood is given by

$$\ell(\Theta) = \ell(\mu, \sigma; \mathbf{y}) + \ell(\pi; \mathbf{z}) + \ell(\lambda; \mathbf{x}) + \ell(\sigma_U, \phi_U) + \ell(\sigma_V, \phi_V). \tag{7}$$

Details on each component composing the likelihood of the model are provided in Appendix A.

3. Application

Given the socioeconomic significance of the European sardine for Portugal and Spain, and the abundance of available data pertaining to this species, we undertake the task of predicting its spatial distribution within the Portuguese shelf. For illustrative purposes, the application focuses specifically on the southern region of the Portuguese coast, an important area for sardine fisheries. In our predictive modeling approach, we incorporate two data sources – FID and FDD – to comprehensively represent the two main sources of fishery data described in Section 3.1.

Table 1
Summary of recorded locations of sardine off the southern coast of Portugal, along with the count and percentage of locations exhibiting strictly positive values (i.e., presences), for both data sources FID (PELAGO survey series data) and FDD (commercial data obtained through the AIS) during 2017. The table includes the overall estimate of the nautical area-scattering coefficient (NASC) derived from FID and the total Catch Per Unit Effort (CPUE) from FDD for sardine.

Data source	Number of locations	Sardine positive locations	Total estimated/captured
FID	144	29 (20%)	26 142.06 mg nm^{-2}
FDD	151	127 (84%)	182 434 kg h^{-1}

3.1. Data

3.1.1. Fishery-independent data

The spatial distribution of sardine biomass is assessed using data from the Portuguese spring acoustic (PELAGO) series (first row of Fig. 1), conducted by the Portuguese Institute for the Sea and Atmosphere (IPMA) in Continental Portuguese waters during 2017.

The primary objective of the PELAGO series is to monitor the spatial distribution of abundance, biomass, and various biological parameters of sardines and other small pelagic fish. The survey design involves continuous daytime acoustic measurements along parallel transects, facilitated by a calibrated 38-kHz echosounder. Data processing includes integrating and averaging the resulting backscatter from the water column over 1 nm intervals, expressed as nautical area-scattering coefficients [NASC; S_A (in m^2 nm^{-2})]. The inter-transect distance is consistently 6 nm. The detailed methodology underpinning the PELAGO series is outlined in Doray et al. (2021).

Each NASC value, representing as a proportion of fish density, is utilized as a biomass proxy for each pair of coordinates (longitude and latitude). The FID source incorporates 144 sardine NASC values recorded in 2017, where the majority (about 80%) represent species absence (Table 1).

3.1.2. Fishery-dependent data

For the same area of interest, the FDD source consists of recent output data by Araújo and Rosa (2023) generated from Automatic Identification System (AIS) data obtained under a commercial license for the Portuguese mainland purse seiners. Importantly, this commercial data aligns with the period when the scientific survey was conducted, ensuring consistency in the temporal scope to avoid variations in species distribution patterns that may occur throughout the year, specifically between April 24th and June 6th of 2017. The dataset from commercial source is standardized by fishing effort (total duration in hours of the fishing event), quantified in kilograms per hour (Kg h^{-1}), enhancing comparability across samples. The FDD dataset comprises 151 commercial samples, providing valuable insights into the spatial distribution of sardine biomass in the studied region (second row of Fig. 1). Conversely, the majority (about 84%) of the FDD observations indicate species presence (Table 1).

3.2. Catchability effect

Given the distinct biomass indices for sardines derived from the FID (measured in NASC units) and the FDD (expressed as catch in Kg h^{-1}), the proposed model is applied to estimate the relative biomass index, denoted as the underlying process of interest $\mathbf{S}$. In this framework, we assume that the expected biomass index value, $\zeta(\mathbf{x}_i, j)$, for each vessel j (whether associated with FID or FDD) and spatial location $\mathbf{x}_i$, is a function of the expected relative biomass $\mu(\mathbf{x}_i)$ and a catchability parameter k_j, defined as

$$\zeta(\mathbf{x}_i, j) = k_j \times \mu(\mathbf{x}_i) \tag{8}$$

The catchability parameter allows to adjust for measurement differences between the two data sources, to capture the vessel-specific differences in catch efficiency, and to ensure that the relative biomass estimates are comparable across sources allowing each index to be proportional to the underlying biomass $\zeta(\mathbf{x}_i, j)$. In this setting, we assume that $Y(\mathbf{x}_i, j) \sim Gamma(\zeta(\mathbf{x}_i, j)^2/v^2, v^2/\zeta(\mathbf{x}_i, j))$ instead of $Y(\mathbf{x}_i) \sim Gamma(\mu(\mathbf{x}_i)^2/v^2, v^2/\mu(\mathbf{x}_i))$, which does not account for variation across vessels.

In the present case study, the survey data were collected using a single research vessel, whereas the commercial data were obtained from observations across fifteen distinct fishing vessels. Accordingly, the index j takes values $j = \ldots, 16$ where $j = 1$ corresponds to the survey vessel, and $j = 2, \ldots, 16$ represent the fifteen commercial vessels.

3.3. Results

The proposed Joint model, integrating both FDD and FID, provided insights into the sardine distribution. The three models under consideration exhibited slight differences in their parameter estimates, as outlined in Table 2, revealing the influence of each data source.

Additionally, the findings shed light on the pronounced positive correlation of the sampling process from the FDD data source on both sardine presence and biomass. This is prominently illustrated in Fig. 1, where the majority of FDD samples align with higher FID observations. Compared to the FDD model, the Joint model produced more robust estimates for the preferential degrees, especially for β. The discrepancy in the estimation of this parameter can be attributed to the fact that the scientific survey detected

Table 2
Parameter estimates (and standard errors) involved in FID, FDD, and Joint models. Standard errors for $\sigma_{\cdot}$ and $\phi_{\cdot}$ are not provided since they resulted from the reparametrization of $\kappa_{\cdot}$ and $\tau_{\cdot}$ (see Section 2.2).

Parameter	FID	FDD	Joint
α'	−10.36 (0.15)	−2.77 (0.51)	−5.52 (2.84)
α	5.61 (0.10)	4.81 (0.20)	6.25 (1.04)
α''		1.28 (1.46)	1.43 (1.64)
β'		0.96 (0.03)	0.96 (0.32)
β		9.06 (0.03)	0.85 (0.23)
$\phi_{\mathbf{V}}$ (km)	5.19	40.91	25.48
$\sigma_{\mathbf{V}}$	20.39	3.31	3.60
$log(\kappa_{\mathbf{V}})$	−0.61 (0.61)	−2.67 (0.11)	−2.20 (0.29)
$log(\tau_{\mathbf{V}})$	−3.67 (1.24)	0.21 (0.21)	−0.35 (0.31)
$\phi_{\mathbf{U}}$ (km)	6.86	36.91	15.05
$\sigma_{\mathbf{U}}$	2.56	0.27	2.05
$log(\kappa_{\mathbf{U}})$	−0.89 (0.11)	−2.57 (0.10)	−1.67 (0.29)
$log(\tau_{\mathbf{U}})$	−1.32 (0.19)	2.63 (0.18)	−0.31 (0.23)

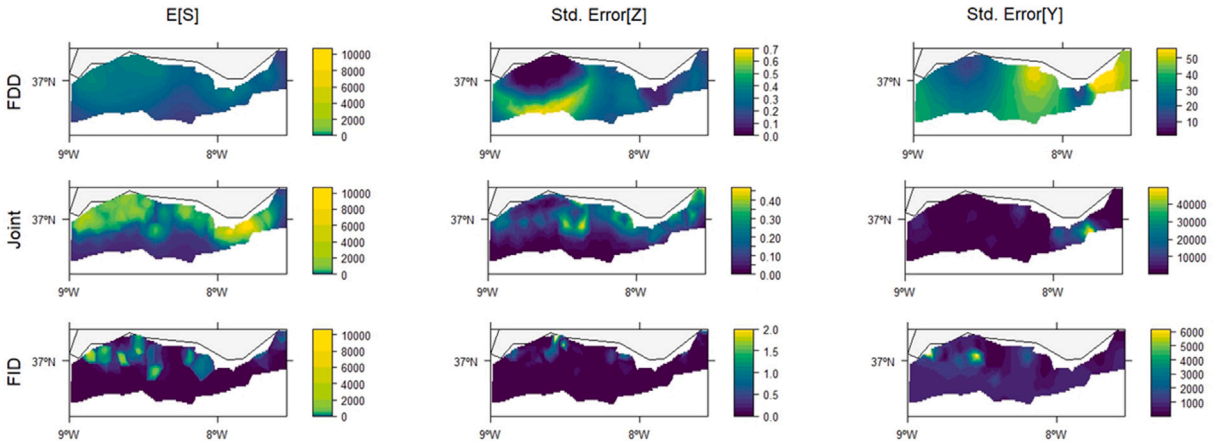


Fig. 4. Predicted relative biomass index of sardine in Portuguese south coast for 2017 (first column) and standard errors associated to the predicted processes **Z** (second column) and **Y** (third column). First row: FDD model prediction - results obtained from the model fitted exclusively to the FDD (commercial) source. Second row: Joint model prediction - results from modeling both FID and FDD (survey) sources. Third row: FID model prediction - results derived from the model fitted exclusively to the FID source. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sardine in areas that were not explored by the commercial fishing vessels. This suggests that incorporating FID data also contributes in defining the preferential effect associated with the FDD, as it provides additional spatial information beyond the regions covered by the fishermen.

An important disparity in parameter estimates emerges in the spatial covariance parameters $\sigma_{\cdot}$ and $\phi_{\cdot}$ (refer to rows 6–7 and 10–11 of Table 2). Given that 80% of FID observations indicate species absence, while 84% of FDD observations are positive, an anticipated dissimilarity in variability emerges in the PAP for FID (rows 7 and 11 of Table 2). These findings underscore the Joint model's efficacy in assimilating the variability from individual data sources, enhancing our understanding of the interplay between different datasets in the modeling process.

The FDD and FID models yield distinct spatial prediction patterns (first column of Fig. 4). The FID model reveals seven well-defined (colored) regions that distinctly indicate species presence (first column, third row of Fig. 4). In contrast, the FDD model presents a smoother pattern of relative biomass, characterized by stronger spatial dependence (first column, first row of Fig. 4). Finally, predictions from the Joint model integrate features from both the FID and FDD models, underscoring the complementary contributions of each data source. Specifically, the FID data aids in delineating a clear presence–absence pattern and identifying certain hotspots, while the FDD data contributes to capturing the spatial dependency associated with both the PAP and the relative biomass process.

These distinctive contributions can be attributed to the prevalence of zero values in the FID data and a substantial number of positive values in the FDD data. Thus, the Joint model can effectively synthesize these unique contributions, offering a comprehensive spatial prediction that captures the nuances of both data sources.

4. Simulation study

To evaluate the performance and robustness of the proposed Joint model under controlled conditions, we conducted a simulation study that mirrors key characteristics of the fisheries data.

4.1. Scenarios of sampling

Various combinations of the preferential parameters β' and β give rise to distinct intensity functions of the point process, and consequently to diverse sampling scenarios. These scenarios may range from extremes, where sampling is solely contingent on either interest process $\mathbf{Z}$ or $\mathbf{Y}$, to situations where it is dependent on both processes as demonstrated by the results of the application. This array of scenarios enables the projection of real-world situations in fishery science, as fishermen often concentrate their efforts on sampling based on their prior knowledge of the species.

The case study results support this interpretation, with both preferential parameters estimated as positive, indicating joint dependence on species presence and biomass. Consequently, the simulation scenarios were carefully designed to reflect this empirical configuration under *Scenario 3* ($\beta' = 1$ and $\beta = 1$) as well as to explore variations in the relative influence of biomass (*Scenarios 2* and *4*), allowing for the evaluation of model robustness under realistic variations.

Additionally, we include scenarios representing extremes cases: one where sampling is driven purely by species presence regardless of biomass (*Scenario 5*), which may arise due to regulatory constraints such as quotas or minimum landing sizes; and another where sampling is based solely on biomass (*Scenario 1*), modeling situations where abundance information is prioritized over presence certainty. This simulation design ensures that model performance is evaluated under conditions that are both empirically grounded and relevant to applied fisheries science.

Below, we enumerate some of these scenarios whose representation of the set of sample locations is available in Fig. B.1.

- *Scenario 1:* Strong PS dependent on $\mathbf{Y}$
  The sampling process for simulated FDD is entirely and strongly contingent on the biomass process under presence. Hence, $\beta' = 0$ and $\beta = 2$.
- *Scenario 2:* Moderate PS dependent on $\mathbf{Z}$ and weak PS dependent on $\mathbf{Y}$
  The sampling process depends on both the PAP and biomass under the presence process, with parameters set at $\beta' = 1$ and $\beta = 0.5$.
- *Scenario 3:* Moderate PS dependent on $\mathbf{Z}$ and $\mathbf{Y}$
  The FDD was simulated under the combination $\beta' = 1$ and $\beta = 1$, reflecting the preferential pattern observed in the case study (Table 2).
- *Scenario 4:* Moderate PS dependent on $\mathbf{Z}$ and strong PS dependent on $\mathbf{Y}$
  The sampling process for FDD is dependent on both processes of interest, with a higher weight assigned to the biomass process under presence $\mathbf{Y}$. In this setting, $\beta' = 1$ and $\beta = 2$.
- *Scenario 5:* Strong PS dependent on $\mathbf{Z}$
  The sampling locations for simulated FDD are contingent on the PAP of the species. The preferentiality parameters are set such that $\beta' = 2$ and $\beta = 0$.

4.2. Simulation–estimation experiments

Simulation–estimation experiments are carried out to evaluate the performance of the proposed model across various data and model configurations. Each scenario is simulated on a regular 60×60 grid within the domain $[0, 1] \times [0, 1]$. Range and marginal variance parameters are individually set for each GMRF, denoted as $\mathbf{U}(\mathbf{X})$ and $\mathbf{V}(\mathbf{X})$, to assess the model performance concerning distinct spatial dependencies of both responses $\mathbf{Z}$ and $\mathbf{Y}$. That is, assuming that both responses are governed by different processes. Specifically, $(\phi_{\mathbf{V}}, \sigma^2_{\mathbf{V}}) = c(0.15, 0.80)$ and $(\phi_{\mathbf{U}}, \sigma^2_{\mathbf{U}}) = c(0.20, 1.00)$. Additionally, the intercept parameters, namely α, α', and α'', were assumed to be zero across all scenarios.

To assess how the sample sizes of both FID (n^I) and FDD (n^D) influence the relative contribution of each data source, simulations are conducted with various combinations of sample sizes $C_n(n^I, n^D)$. These configurations are chosen to reflect realistic sampling conditions commonly encountered in fisheries studies. Therefore, the selected combinations are centered around the sample sizes observed in the empirical application – approximately similar magnitudes for both FID and FDD – while also exploring scenarios in which one source substantially outweighs the other.

The selected combinations include the configuration $C_n(100, 100)$, representing a balanced scenario in which FID and FDD have equal sample sizes — consistent with the sampling structure observed in the application case study. Recognizing that in practical situations, FDD often exhibits larger dimensions compared to FID due to factors such as financial constraints and time-intensive surveying (also stated by Alglave et al., 2022), additional combinations are explored. These include $C_n(100, 200)$ and a more asymmetric scenario $C_n(100, 500)$. Conversely, to account for scenarios where a greater emphasis on FID may arise due to fishery restrictions or limited interest in specific species by fishermen, a combination with larger FID dimensions is considered, denoted as $C_n(200, 100)$. This scenario reflects a less common, yet plausible.

Furthermore, we conduct a comprehensive assessment of the contribution of each data source in *Scenario 3* ($\beta' = 1$ and $\beta = 1$) through a comparative analysis of three models: the FDD model, the FID model, and the Joint model. The FDD model encompasses

our proposed model but exclusively utilizes FDD data. In contrast, the FID model represents a simplified version of our proposed model, neglecting the modeling of the sampling process and relying solely on FID data. The selection of *Scenario 3* ($\beta' = 1$ and $\beta = 1$) for this evaluation was deliberate, aiming to capture a scenario that closely mirrors real-world application presented in Section 3 and thus provides insights into the distinctive contributions of each data source.

To ensure robustness, each scenario and configuration is repeated 100 times, allowing for the capture of variability among replicates.

4.3. Performance metrics

The assessment of the estimation performance of the proposed model involves a comprehensive analysis of various model outputs. The evaluation encompasses all estimated parameters, including the intercept (results in Appendix C.1), preferential (Section 4.4.1), and spatial covariance (Section 4.4.2) parameters, as well as the spatial predictions (Sections 4.4.3 and 4.4.4 and Appendices C.2 and C.3).

To gauge the accuracy of intercept parameters estimation, the distribution of estimates is scrutinized across 100 replicas. Given the potential for asymmetric distributions in spatial covariance parameters across replicas, their estimation quality is performed through the identification of the median and the interquartile interval.

In addition to assessing parameters estimation, the predictive performance of the proposed method is thoroughly evaluated using three distinct metrics: RMSE, MAE, and the Hellinger distance (Le Cam, 1986). These metrics provide a robust evaluation of the model's ability to generate spatial predictions that align closely with observed data. In this context, the Hellinger distance measures the similarity between the observed data and the predicted data distributions, ranging from 0, indicating equality, and 1, indicating "total difference".

4.4. Results

4.4.1. Evaluation of the estimation of preferential parameters

The proposed model demonstrates a capacity to yield valuable estimates for preferential parameters (Fig. 5). When assuming β' or β as zero, the model provides accurate estimates for these parameters. Moreover, under scenarios of moderate or weak effects of PS ($\beta' = 1$ and/or $\beta = \{0.5, 1\}$), the model reliably estimates preferential parameters for FDD dimension up to 200, albeit slight underestimation is observed for FDD dimensions of 500. In instances of a strong PS effect, while there is a tendency for parameter underestimation, the estimates remain statistically significant.

Furthermore, increasing the dimensions of the FDD enables the selection of locations with lower corresponding values of the intensity function, thereby capturing samples with reduced values (see Fig. D.1 of Appendix D). This expansion results in a decreased mean for the process of interest. The augmentation of FDD sample size also emerges as a pivotal factor in reducing variability in parameter estimation.

Conversely, the increase in sample size of FID does not exert a discernible impact on the estimation of preferentiality parameters. This outcome aligns with expectations, as β and β' parameters are only utilized to describe the spatial arrangement of FDD. The lack of influence from FID on preferentiality parameter estimation underscores the distinct roles played by FID and FDD in shaping the precision and reliability of the parameter estimates.

4.4.2. Evaluation of the estimation of spatial covariance parameters

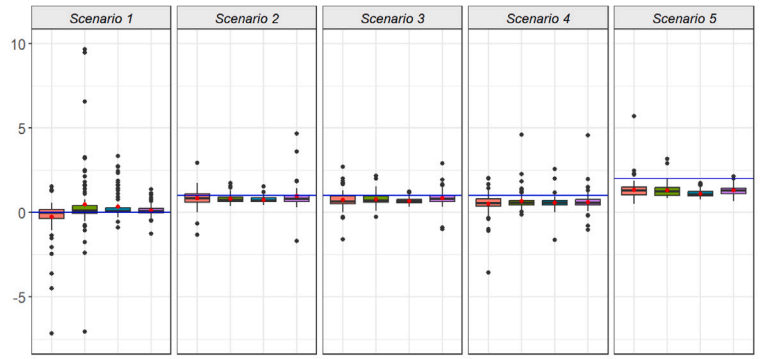
In most considered model configurations, encompassing various combinations of sample sizes and scenarios, we observe a pattern of slight overestimation in the range parameter, denoted as $\phi_{\mathbf{V}}$ (refer to Table 3). Generally, as sample sizes increase, the estimated $\phi_{\mathbf{V}}$ distantiates to its true value. *Scenario 1*, characterized by a robust stochastic dependence influenced by the biomass process, yields more accurate $\phi_{\mathbf{V}}$ estimates irrespective of sample sizes. Conversely, *Scenario 5*, representing a scenario where the PS is only influenced by the PAP, produces more biased estimates of the range parameter within the GMRF $\mathbf{V}(\mathbf{X})$. The precision of $\phi_{\mathbf{V}}$ estimation is found to be sensitive to the sample sizes, with increased variability as both sample sizes grow.

In all scenarios and across various combinations of sample sizes, accurate estimation of the $\sigma_{\mathbf{V}}$ parameter is observed. That is, the interquartile intervals consistently encompass the true value, affirming the method adequacy for estimating this parameter. Additionally, as the sample size of the FDD increases, there is a decrease in bias. Upon comparing all scenarios, it is evident that *Scenario 2*, under the moderate dependence of the sampling process on the PAP and weak dependence on the biomass process, stands out by providing more accurate estimates for $\sigma_{\mathbf{V}}$.

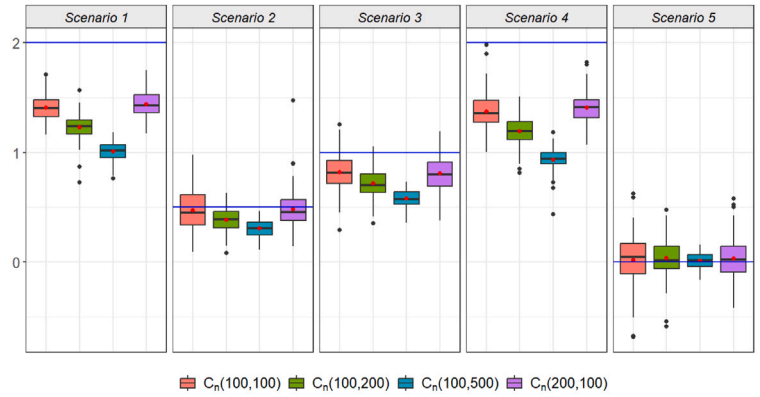
Contrasting with the findings concerning the estimation of $\phi_{\mathbf{V}}$, the estimation of the range parameter $\phi_{\mathbf{U}}$ is reliable across all model configurations. Moreover, the augmentation of sample sizes generally induces higher estimates of the range parameter for biomass under the presence process. *Scenario 1*, characterized by the absence of PS dependent on PAP, yields more biased estimates for the $\phi_{\mathbf{U}}$ parameter. This stands in contrast to *Scenarios 3* and *5* which result in more accurate estimates.

A prevalent accurate estimation of the $\sigma_{\mathbf{U}}$ parameter is observed for all of model configurations, despite the approximation to the true value with the augmentation of data dimension.

In summary, the assessment of spatial covariance estimation reveals consistent patterns, including the tendency for accurate estimation of both range marginal variance parameters across various model configurations, except the observed overestimation of $\phi_{\mathbf{V}}$. Additionally, the reliability of parameter estimation improves when the sample size of FID increases and the effect of PS on PAP is null or moderate.



(a) β'



(b) β

Fig. 5. Estimates of preferential parameters (β' and β) across sampling scenarios (*Scenario 1*: $\beta' = 0$ and $\beta = 2$; *Scenario 2*: $\beta' = 1$ and $\beta = 0.5$; *Scenario 3*: $\beta' = 1$ and $\beta = 1$; *Scenario 4*: $\beta' = 1$ and $\beta = 2$; *Scenario 5*: $\beta' = 2$ and $\beta = 0$) and combination of samples' dimensions $C_n(n^I, n^D)$. Red points represent the mean values of all 100 replicas and blue lines identify the true values assumed for the preferential parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4.3. Evaluation of the prediction performance

In terms of prediction performance (Figs. 6 and C.2), no substantial differences are discernible across various model configurations, although a modest improvement in prediction accuracy is noted with larger datasets.

Analysis of the Hellinger distance estimates across diverse scenarios (Fig. 6) reveals its capability to generate predicted data distributions closely resembling the observed ones.

4.4.4. Evaluation of the contribution of each data source

Comparing RMSE (Fig. C.3(a)) and MAE (Fig. C.3(b)) results, the Joint model consistently outperformed models that use one of both data sources. These discrepancy is more pronounced when analyzing the results of the Hellinger distance (Fig. 7). Larger datasets are required for accurate predictions independently of the model. Consequently, the proposed Joint model demonstrates prediction efficiency, presenting a balanced performance across various dimensions and providing robust predictions.

Fig. 8 illustrates the distinctions in predictions obtained from the three models (FDD, FID, and Joint) for one replica of *Scenario 3*. Independently on the configuration of sample sizes, Joint model yield a predicted pattern more akin to the true one. Moreover, FDD and Joint models demonstrated enhanced capability in capturing biomass hotspots, possibly attributed to the dependence of the sampling process on the biomass under presence. Indeed, the influence of a PS dependent on the biomass process leads to higher sample values of biomass.

The predicted patterns from FDD and Joint models are quite similar, with slight variations reflecting the influence of FID. FID provides a more well-defined pattern of the response process when FID surpasses FDD in dimensionality, $C_n(200, 100)$. In this specific combination, the Joint model showcases the combined contributions of both data sources, highlighting the clear impact of incorporating both FID and FDD dimensions in the Joint model for a comprehensive and nuanced representation of biomass patterns. However, it is important to note that Fig. 8 resulted from one simulated experiment and discrepancies between the models in the prediction fields can be further observed for other simulation experiments.

Table 3
Median values (and interquartile intervals) for $\phi_{\mathbf{V}} = 0.15$, $\sigma_{\mathbf{V}} = \sqrt{0.80}$, $\phi_{\mathbf{U}} = 0.20$ and $\sigma_{\mathbf{U}} = 1.00$ across sampling scenarios (*Scenario 1*: $\beta' = 0$ and $\beta = 2$; *Scenario 2*: $\beta' = 1$ and $\beta = 0.5$; *Scenario 3*: $\beta' = 1$ and $\beta = 1$; *Scenario 4*: $\beta' = 1$ and $\beta = 2$; *Scenario 5*: $\beta' = 2$ and $\beta = 0$) and combinations of samples' dimensions $C_n(n^I, n^D)$.

Param.	*Scen.*	$C_n(100,100)$	$C_n(100,200)$	$C_n(100,500)$	$C_n(200,100)$
$\phi_{\mathbf{V}}$	*1*	0.25 (0.07,0.45)	0.18 (0.07,0.57)	0.17 (0.13,0.62)	0.18 (0.10,0.45)
	2	0.22 (0.15,0.42)	0.26 (0.19,0.45)	0.39 (0.26,0.43)	0.22 (0.17,0.34)
	3	0.23 (0.16,0.37)	0.32 (0.23,0.39)	0.39 (0.29,0.46)	0.22 (0.14,0.35)
	4	0.20 (0.15,0.43)	0.22 (0.16,0.38)	0.28 (0.19,0.41)	0.19 (0.12,0.34)
	5	0.22 (0.17,0.36)	0.29 (0.25,0.42)	0.37 (0.33,0.47)	0.26 (0.17,0.33)
$\sigma_{\mathbf{V}}$	*1*	0.03 (1.05×10^{-4},0.97)	0.01 (5.80×10^{-5},1.02)	0.88 (3.62×10^{-5},1.07)	0.80 (4.74×10^{-4},1.07)
	2	0.91 (0.31,1.17)	0.89 (0.45,1.17)	0.88 (0.70,1.18)	0.79 (0.51,1.07)
	3	0.79 (0.46,1.14)	0.83 (0.64,1.24)	0.93 (0.72,1.32)	0.89 (0.48,1.09)
	4	0.79 (1.93×10^{-3},1.16)	0.75 (0.47,1.10)	0.81 (0.57,1.06)	0.83 (0.35,1.15)
	5	0.85 (0.66,1.22)	0.96 (0.66,1.26)	1.08 (0.84,1.28)	0.90 (0.68,1.16)
$\phi_{\mathbf{U}}$	*1*	0.25 (0.20,0.32)	0.31 (0.22,0.34)	0.29 (0.27,0.38)	0.23 (0.19,0.32)
	2	0.23 (0.19,0.30)	0.24 (0.20,0.32)	0.25 (0.21,0.34)	0.22 (0.19,0.29)
	3	0.23 (0.18,0.31)	0.21 (0.19,0.32)	0.25 (0.19,0.31)	0.22 (0.18,0.31)
	4	0.24 (0.18,0.30)	0.25 (0.22,0.32)	0.30 (0.24,0.35)	0.22 (0.19,0.29)
	5	0.23 (0.18,0.32)	0.23 (0.19,0.30)	0.25 (0.18,0.29)	0.21 (0.17,0.32)
$\sigma_{\mathbf{U}}$	*1*	0.97 (0.86,1.27)	1.10 (0.89,1.26)	1.13 (1.05,1.45)	0.96 (0.86,1.20)
	2	0.92 (0.81,1.16)	0.98 (0.85,1.22)	0.99 (0.91,1.33)	0.93 (0.83,1.18)
	3	0.86 (0.80,1.19)	0.92 (0.84,1.17)	0.99 (0.89,1.23)	0.90 (0.82,1.18)
	4	0.95 (0.87,1.25)	1.00 (0.87,1.22)	1.11 (0.97,1.40)	0.94 (0.88,1.16)
	5	0.89 (0.78,1.11)	0.94 (0.81,1.09)	0.91 (0.85,1.14)	0.88 (0.82,1.12)

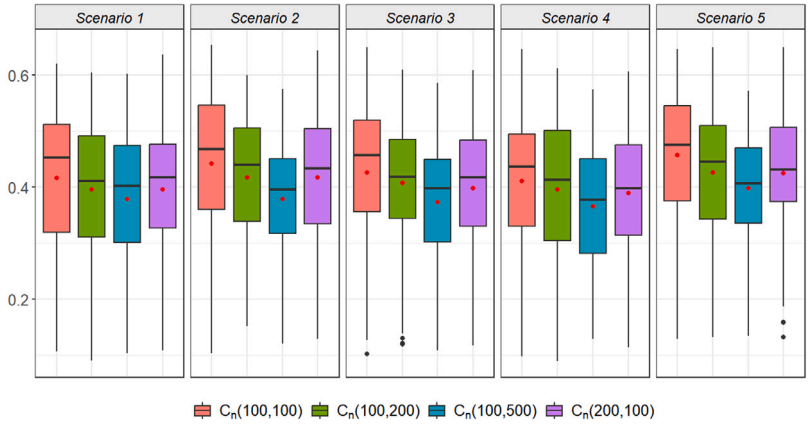


Fig. 6. Evaluation of predictive performance, using the Hellinger distance, across sampling scenarios (*Scenario 1*: $\beta' = 0$ and $\beta = 2$; *Scenario 2*: $\beta' = 1$ and $\beta = 0.5$; *Scenario 3*: $\beta' = 1$ and $\beta = 1$; *Scenario 4*: $\beta' = 1$ and $\beta = 2$; *Scenario 5*: $\beta' = 2$ and $\beta = 0$) and combination of samples' dimensions $C_n(n^I, n^D)$. Red points represent the mean values of all 100 replicas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Discussion

Integrating data from diverse sources to gain more comprehensive insights into spatial fish distribution poses a significant challenge in marine ecology. Both FID and FDD sources have proven their capability to offer valuable information about fish distribution (Pennino et al., 2016; Izquierdo et al., 2022; Silva et al., 2024). However, the distinct sampling processes in these datasets pose challenges when attempting joint modeling.

To address these challenges, our study introduces a two-part model that integrates FID and FDD to infer fish distribution patterns. The model is designed to accommodate ZI data and account for the unique sampling processes inherent to each dataset.

Beyond estimating fish spatial distribution, the model provides quantitative insights into fishing behavior through the PS parameters. These coefficients characterize how fishing effort responds to species presence and biomass, enabling the assessment of spatial selectivity and effort allocation. This information is critical for understanding patterns of fishing pressure and supports spatially explicit management strategies, such as identifying areas of high exploitation or evaluating the effectiveness of spatial regulations.

We evaluated the model's strengths and limitations through a comprehensive simulation study, complemented by a case study focused on the spatial distribution of European sardine. In this context, the survey data represent FID, while commercial catch

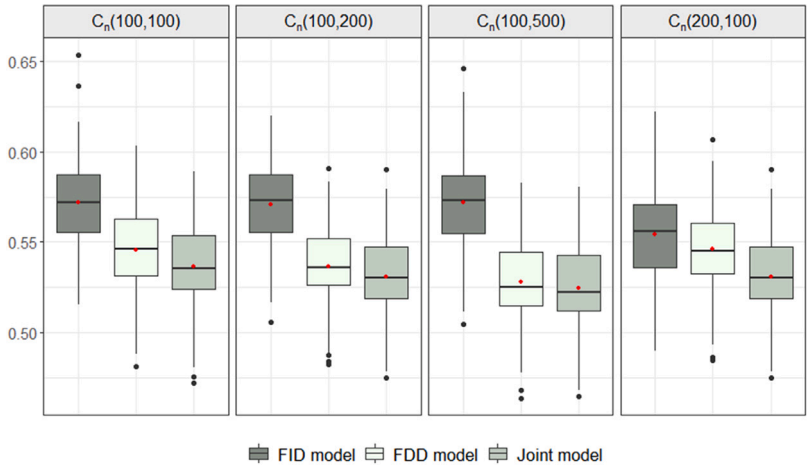


Fig. 7. Evaluation of the contribution of each data source, using the Hellinger distance, across model configurations and combinations of samples' dimensions $C_n(n^I, n^D)$ in *Scenario 3*. Red points depict the mean values of all 100 replicas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

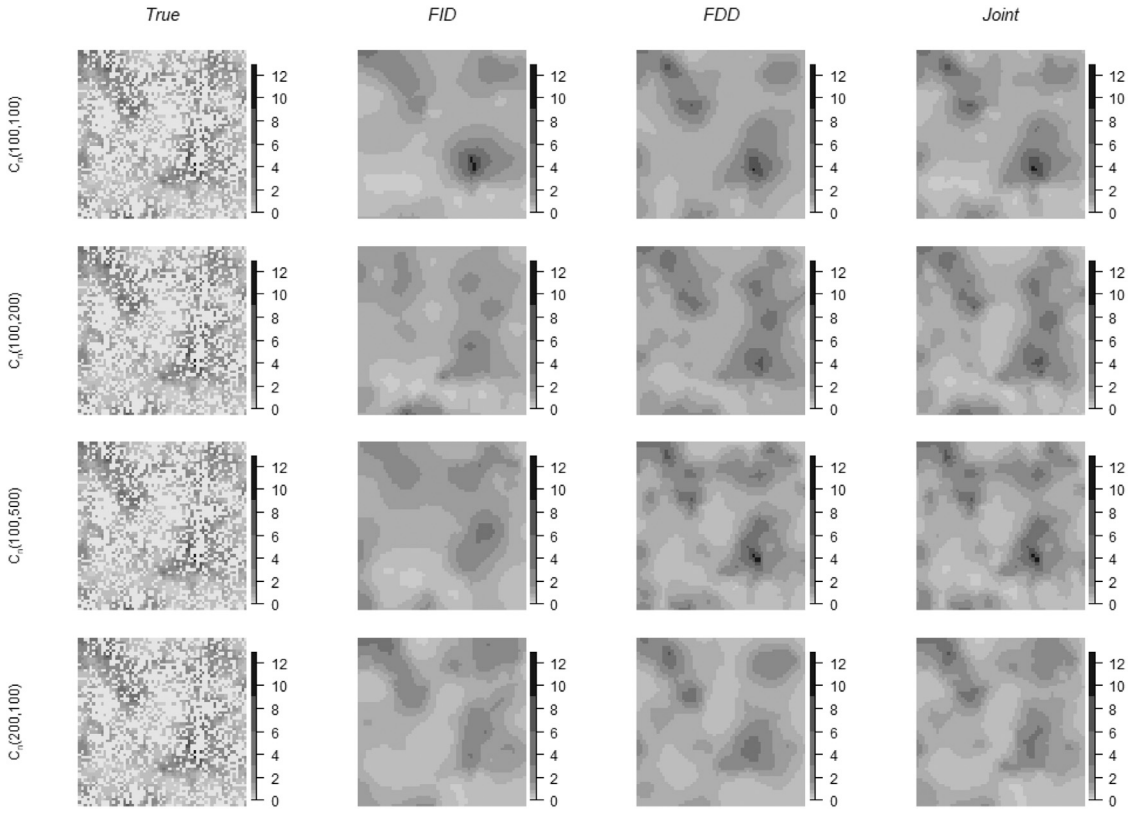


Fig. 8. Results for biomass prediction based on a simulation experiment of *Scenario 3* ($\beta' = 1$ and $\beta = 1$) across various samples' dimensions. First column: observed biomass — the true biomass values $\mathbf{S}$ observed during the simulation experiment. Second column: FDD model prediction $\hat{\mathbf{S}}$ — results obtained from the model fitted exclusively to the FDD source. Third column: Joint model prediction $\hat{\mathbf{S}}$ — predicted biomass values resulting from modeling both FID and FDD sources. Fourth column: FID model prediction $\hat{\mathbf{S}}$ — predictions derived from the model fitted exclusively to the FID source.

records represent FDD. The datasets differ markedly – 80% of FID observations show absence, while 84% of FDD show presence – reflecting their distinct sampling designs. Specifically, fishermen typically rely on prior knowledge and real-time observation of fish

distribution, leading to a higher proportion of presence observations in FDD. Such discrepancies further highlight the necessity of a joint modeling approach to capture a more comprehensive picture of sardine distribution.

Our proposed model exhibits a strong capacity to produce accurate estimates for preferential parameters, attesting to its reliability across various scenarios. Parameter estimation becomes less stable under strong preferential effects simultaneously influencing both PAP and biomass processes, as also noted in Silva and Menezes (2024). Examining spatial covariance estimation reveals a consistent pattern of accurate estimation, except the overestimation of $\phi_{\mathbf{V}}$. Additionally, the model consistently exhibit accuracy across all model configurations for estimating the intercept parameters, excepting the intercept parameter α'' for the intensity function that tends to be overestimated due to its role of representing the intensity mean of the point process (more details in Appendix C.1).

Prediction performance consistently favors models that incorporate both FID and FDD over those utilizing a single data source across all examined scenarios, demonstrating the model's reliability in accurately capturing underlying distribution patterns. When FID and FDD sources share identical dimensions, the single source models yield similar patterns, suggesting that both preferential and non-PS approaches can provide accurate representations of the process of interest. Results from both the simulation study and the empirical application indicate that the FDD and Joint models present a higher capability to capture the spatial correlation of the response process, while FID model may bring information about marine regions not explored by fishermen. Thus, the Joint model integrates contributions from both data sources, providing a more comprehensive representation of biomass patterns.

Although the model introduced by Rufener et al. (2021) demonstrated satisfactory results in predicting fish distribution, it is important to note that the authors did not account for the PS nature of commercial data. However, it is widely recognized that a substantial portion of opportunistic data, such as derived from commercial sources, exhibits relevant PS. Ignoring PS in spatial prediction may introduce significant bias, as emphasized by Diggle et al. (2010). A limitation of this model arises in its suitability for handling ZI data since it is tailored for count data, using the Negative Binomial distribution to accommodate abundance data.

Alglave et al. (2022) demonstrated the robustness and consistency of their proposed model across a spectrum of scenarios, effectively addressing the ZI issue in the data. However, an important aspect not explicitly considered in their model is the potential variability in conditions governing PAP and biomass process, as observed in Silva et al. (2024). In scenarios where local fishing intensity is contingent on fish presence and fish biomass in distinct ways, the model assumption of a unique relationship between the sampling process and the relative biomass field may not capture such variations. In our study, we address this limitation by proposing a two-part model that allows for the differentiation between PAP and biomass process. This consideration enhances the comprehensiveness of our model and contributes to a more detailed understanding of the complex spatial dynamics governing fish distribution in marine environments.

The principal strength of our model lies in its ability to integrate heterogeneous data sources while accounting for PS, leading to improved spatial predictions. Numerous comparative studies (e.g., Diggle et al., 2010; Conn et al., 2017; Alglave et al., 2022) have shown that ignoring PS can lead to biased estimates of spatial distribution. In particular, when high-density areas are preferentially targeted, models that fail to account for PS may overestimate abundance in under-sampled or less-exploited regions. Nonetheless, limitations remain — particularly when strong preferential effects simultaneously influence both PAP and biomass components, where estimation becomes less reliable.

The significance of environmental conditions in influencing species spatial distribution is well-established (Austin, 2007). Incorporating explanatory covariates representing these conditions in our model is not only important for achieving more precise predictions but also offers valuable insights into the complex relationship between the species and the ecosystem (Hefley and Hooten, 2016). As such, a key avenue for future development involves the inclusion of covariates in the model since this consideration is crucial as both the PAP and biomass process can be elucidated by a set of environmental conditions (Pennino et al., 2020; Silva et al., 2024). Moreover, the integration of additional environmental and external variables may be relevant to comprehensively describe the preferentiality for certain locations (Manceur and Kühn, 2014; Pennino et al., 2019). Factors such as distance to the coast and bathymetry can significantly influence the spatial arrangement of fishing locations, since the fishermen tend to stay closer to the port/coast as possible due to fuel costs, contributing to a more detailed understanding of the PS dynamics. This enhancement will not only refine the precision of our predictions but also contribute to a more comprehensive and ecologically informed model.

Beyond the spatial dimension, the temporal scale is a critical component in species distribution modeling, as the temporal evolution holds significant ecological relevance (Paradinas et al., 2017; Martínez-Minaya et al., 2018). Consequently, another point for future investigation involves extending our proposed model to a spatio-temporal framework, enabling the prediction of temporal trends, seasonal variations, and long-term ecological patterns that are integral to a thorough understanding of species distribution dynamics.

In terms of practical biological applications, a potential future direction is to separately model juvenile and adult abundance indexes. Juvenile fish might be avoided by fishermen due to low fishing value and restrictions, and these areas likely correspond to higher biomass, as indicated in the FID, which could impact parameter estimates.

CRediT authorship contribution statement

Daniela Silva: Conceptualization, Methodology, Software, Writing – original draft. **Raquel Menezes:** Conceptualization, Methodology, Validation, Writing – original draft. **Gonçalo Araújo:** Data curation, Formal analysis, Writing – original draft. **Renato Rosa:** Investigation, Validation. **Ana Moreno:** Data curation, Writing – original draft. **Alexandra Silva:** Investigation, Validation, Writing – original draft. **Susana Garrido:** Conceptualization, Validation, Writing – original draft.

Funding

This study received support from Portuguese funding provided through the Centre for Mathematics via the following projects: DOI 10.54499/UIDP/00013/2020, DOI 10.54499/UIDB/00013/2020, and the Portuguese Foundation for Science and Technology (FCT), Portugal through the Individual PhD Scholarship PD/BD/150535/2019, the research grant UIDB/04292/2020, and the project PTDC/MAT-STA/28243/2017. Additionally, support was provided by the SARDINHA2030 project (MAR-111.4.1-FEAMPA-00001).

Acknowledgments

The authors would like thanks the teams that collected the data. The biological survey datasets generated during and/or analyzed during the current study are available from the Portuguese Data Collection Framework on reasonable request. It is not ethically feasible to share any AIS data, as it would publicly reveal vessel information indicating where the activity takes place, while disclosing sensitive information. Additionally, the AIS data outputs are ruled by a confidentiality agreement between the different authors, preventing the share of the provided AIS data outputs, any private information regarding the fishing vessel and other related information.

Appendix A. Components of the likelihood

A.1. Likelihood for $\mathbf{Y}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{U}$

$$\mathcal{L}(\mu,\sigma;\mathbf{y}) = \prod_{i=1}^{n} \frac{y_i^{\left(\frac{\mu^2}{\sigma^2}-1\right)} e^{-\left(\frac{\mu}{\sigma^2}\right)y_i} \left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{\sigma^2}}}{\Gamma(\frac{\mu^2}{\sigma^2})}. \tag{A.1}$$

The corresponding log-likelihood is given by

$$\begin{aligned}\ell(\mu,\sigma;\mathbf{y}) &= \sum_{i=1}^{n}\left(\left(\frac{\mu^2}{\sigma^2}-1\right)log(y_i) - \frac{\mu}{\sigma^2}y_i + \frac{\mu^2}{\sigma^2}log\left(\frac{\mu}{\sigma^2}\right) - log\left(\Gamma\left(\frac{\mu^2}{\sigma^2}\right)\right)\right)\\ &= n\frac{\mu^2}{\sigma^2}log\left(\frac{\mu}{\sigma^2}\right) - nlog\left(\Gamma\left(\frac{\mu^2}{\sigma^2}\right)\right) + \sum_{i=1}^{n}\left(\left(\frac{\mu^2}{\sigma^2}-1\right)log(y_i) - \frac{\mu}{\sigma^2}y_i\right)\end{aligned} \tag{A.2}$$

where n represents the dimension of all data ($n = n^I + n^D$).

A.2. Likelihood for $\mathbf{Z}|\mathbf{X}^D, \mathbf{X}^I, \mathbf{V}$

$$\mathcal{L}(\pi;\mathbf{z}) = \prod_{i=1}^{n}\pi^{z_i}(1-\pi)^{1-z_i} = \pi^{\sum_{i=1}^{n} z_i}(1-\pi)^{n-\sum_{i=1}^{n} z_i} \tag{A.3}$$

and, hence, the log-likelihood is given by

$$\ell(\pi;\mathbf{z}) = log(\pi)\sum_{i=1}^{n} z_i + log(1-\pi)\left(n - \sum_{i=1}^{n} z_i\right). \tag{A.4}$$

A.3. Likelihood for $\mathbf{X}^D|\mathbf{U}, \mathbf{V}$

Following Diggle (2013), the likelihood for IPP comes from

$$\mathcal{L}(\lambda;\mathbf{x}) = \prod_{i=1}^{n^D}\frac{e^{-\omega}\omega^{n^D}}{n^D!} \times \frac{\lambda(\mathbf{x}_i^D)}{\omega_t}. \tag{A.5}$$

The log-likelihood is expressed as

$$\begin{aligned}\ell(\lambda;\mathbf{x}) &= log\left(\frac{e^{-\omega}\omega^{n^D}}{n^D!}\right) + \sum_{i=1}^{n^D} log\left(\frac{\lambda(\mathbf{x}_i)}{\omega}\right) = -\omega + n^D \times log(\omega) - log(n^D!) + \sum_{i=1}^{n^D}\left(log(\lambda(\mathbf{x}_i)) - log(\omega)\right)\\ &\simeq \sum_{i=1}^{n^D} log(\lambda(\mathbf{x}_i)) - \omega = \sum_{i=1}^{n^D} log(\lambda(\mathbf{x}_i)) - \int_{\mathcal{A}}\lambda(\mathbf{s})\partial\mathbf{s}.\end{aligned} \tag{A.6}$$

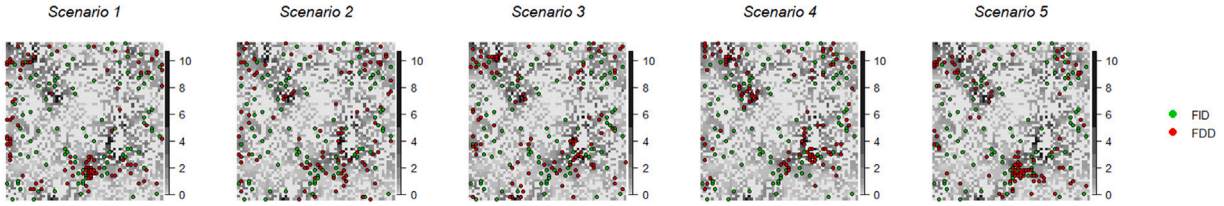


Fig. B.1. Simulated FID and FDD locations across sampling scenarios (*Scenario 1*: $\beta' = 0$ and $\beta = 2$; *Scenario 2*: $\beta' = 1$ and $\beta = 0.5$; *Scenario 3*: $\beta' = 1$ and $\beta = 1$; *Scenario 4*: $\beta' = 1$ and $\beta = 2$; *Scenario 5*: $\beta' = 2$ and $\beta = 0$). Green points represent sample locations for simulated FID (Fishery-independent data), and red points identify the sample locations for simulated FDD (Fishery-dependent data). The depicted latent field is **S**, and each data source has a dimension of 100.

A.4. Likelihoods for **U** *and* **V**

$$\mathcal{L}(\sigma_j, \phi_j) = \frac{1}{(\sqrt{2\pi})^N |\Sigma_j|^{1/2}} exp\{-\frac{1}{2}\mathbf{j}_{\cdot}^{\mathrm{T}} \Sigma_j^{-1} \mathbf{j}_{\cdot}\}, \tag{A.7}$$

and, hence, the corresponding log-likelihoods are given by

$$\ell(\sigma_j, \phi_j) = -\frac{N}{2} log(\pi) - \frac{log(|\Sigma_j|)}{2} - \frac{1}{2}\mathbf{j}_{\cdot}^{\mathrm{T}} \Sigma_j^{-1} \mathbf{j}_{\cdot}, \tag{A.8}$$

where N denotes the dimension of the prediction grid (or mesh), $j = \{\mathbf{U}, \mathbf{V}\}$ and $\mathbf{j} = \{\mathbf{u}, \mathbf{v}\}$.

Appendix B. Scenarios of sampling

See Fig. B.1.

Appendix C. Supplementary results of the simulation study

C.1. Evaluation of the estimation of intercept parameters

The method accurately estimates α and α', as evidenced by the inclusion of the true value (zero) within the interquartile interval across all considered model configurations (Figs. C.1(a) and C.1(b)). Conversely, the intercept parameter associated with the point process, denoted as α'', exhibits a tendency toward overestimation (Fig. C.1(c)), becoming more biased as the dimension of the FDD increases. Referring to the definition of the intensity function (4), α'' represents the expected value of the point process intensity linked to the FDD, given the zero-mean GFs **V** and **U**. Since the intensity function is positive and increases with the sample size, this behavior explains why α'' is consistently estimated above its true value and why the bias grows as the FDD sample dimension increases.

C.2. Evaluation of the prediction performance

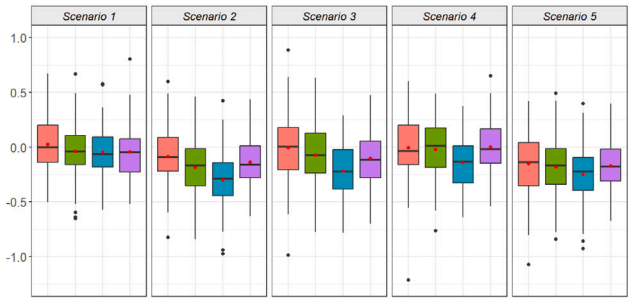
See Fig. C.2.

C.3. Evaluation of the contribution of each data source

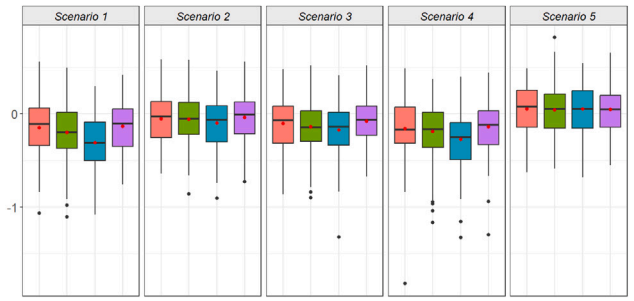
See Fig. C.3.

Appendix D. Impact of sample dimension on observation process

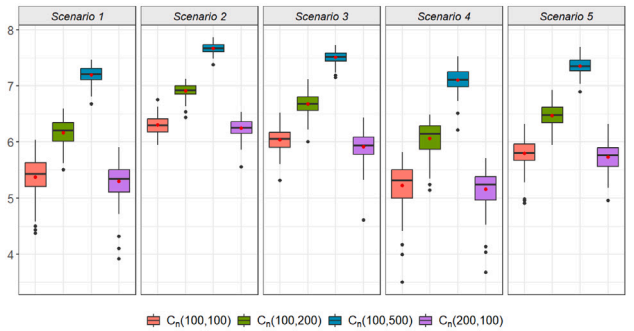
See Fig. D.1.



(a) α'

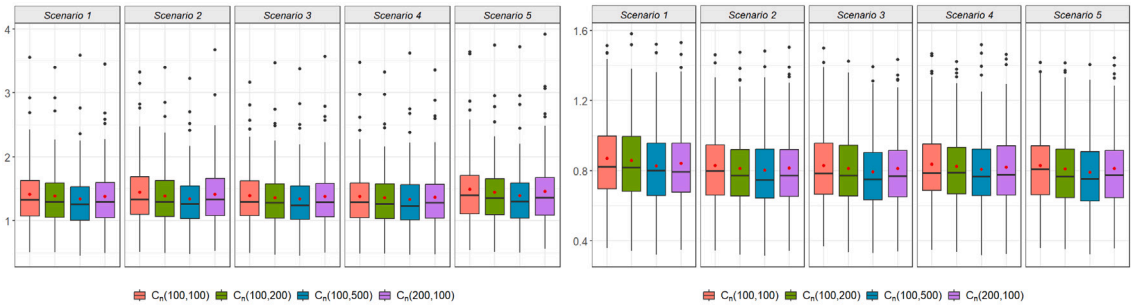


(b) α



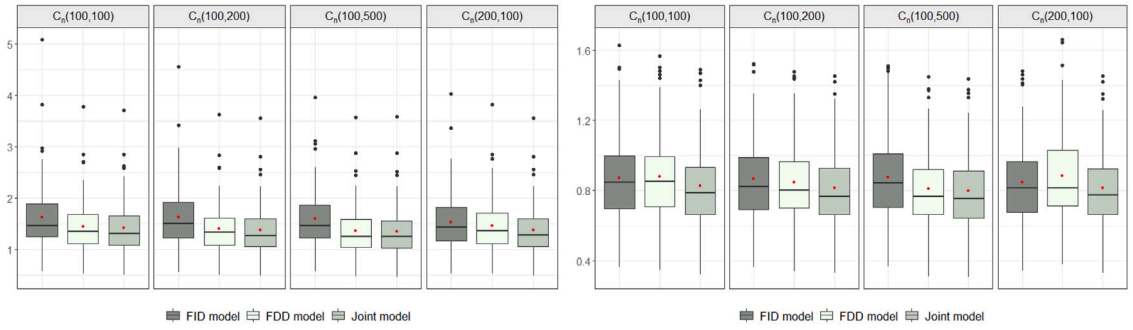
(c) α''

Fig. C.1. Estimates of intercept parameters ($\alpha' = 0$, $\alpha = 0$ and $\alpha'' = 0$) across sampling scenarios (*Scenario 1*: $\beta' = 0$ and $\beta = 2$; *Scenario 2*: $\beta' = 1$ and $\beta = 0.5$; *Scenario 3*: $\beta' = 1$ and $\beta = 1$; *Scenario 4*: $\beta' = 1$ and $\beta = 2$; *Scenario 5*: $\beta' = 2$ and $\beta = 0$) and combination of samples' dimensions $C_n(n^I, n^D)$. Red points represent the mean values of all 100 replicas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) RMSE (Root Mean Square Error) (b) MAE (Mean Absolute Error)

Fig. C.2. Evaluation of predictive performance. Performance metrics (RMSE and MAE) across sampling scenarios (*Scenario 1*: $\beta' = 0$ and $\beta = 2$; *Scenario 2*: $\beta' = 1$ and $\beta = 0.5$; *Scenario 3*: $\beta' = 1$ and $\beta = 1$; *Scenario 4*: $\beta' = 1$ and $\beta = 2$; *Scenario 5*: $\beta' = 2$ and $\beta = 0$) and combination of samples' dimensions $C_n(n^I, n^D)$. Red points represent the mean values of all 100 replicas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) RMSE (Root Mean Square Error) (b) MAE (Mean Absolute Error)

Fig. C.3. Evaluation of the contribution of each data source. Performance metrics (RMSE and MAE) across model configurations and combinations of samples' dimensions $C_n(n^I, n^D)$ in *Scenario 3*. Red points depict the mean values of all 100 replicas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

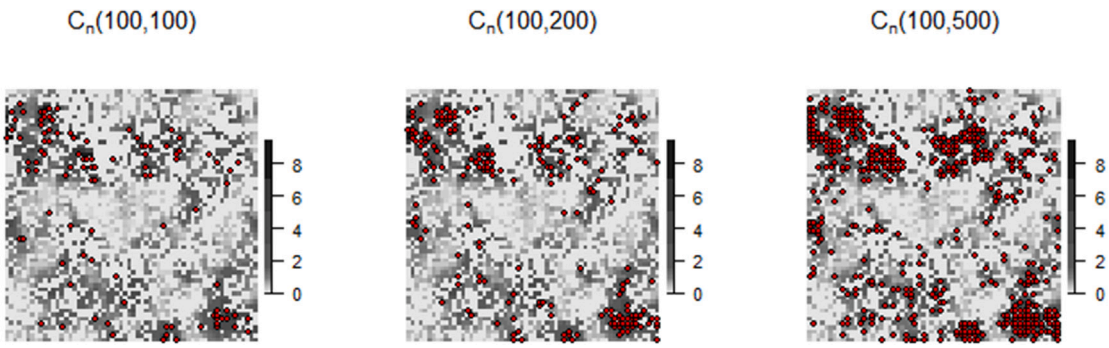


Fig. D.1. Examples of simulated intensity fields and FDD locations for Scenario 4 across combinations of samples' dimensions $C_n(100, 100)$, $C_n(100, 200)$ and $C_n(100, 500)$. Red points identify sample locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Alglave, B., Rivot, E., Etienne, M.-P., Woillez, M., Thorson, J.T., Vermard, Y., 2022. Combining scientific survey and commercial catch data to map fish distribution. ICES J. Mar. Sci. 79 (4), 1133–1149. http://dx.doi.org/10.1093/icesjms/fsac032.

Araújo, G., Rosa, R., 2023. [Unpublished data]. Nova School of Business and Economics, Nova University Lisbon.

Ault, J.S., Bohnsack, J.A., Meester, G.A., 1998. A retrospective (1979–1996) multispecies assessment of coral reef fish stocks in the Florida Keys. Fish. Bulletion 98 (3), 395–414.

Austin, M., 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecol. Model. 200 (1), 1–19. http://dx.doi.org/10.1016/j.ecolmodel.2006.07.005.

Conn, P.B., Thorson, J.T., Johnson, D.S., 2017. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. Methods Ecol. Evol. 8 (11), 1535–1546. http://dx.doi.org/10.1111/2041-210X.12803.

Diggle, P.J., 2013. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, third ed. Chapman and Hall/CRC, http://dx.doi.org/10.1201/b15326.

Diggle, P., Menezes, R., Su, T.-l., 2010. Geostatistical inference under preferential sampling. J. R. Stat. Soc. Ser. C 59 (2), 191–232. http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x.

Doray, M., Boyra, G., van der Kooij, J., 2021. ICES Survey Protocols – Manual for acoustic surveys coordinated under ICES Working Group on Acoustic and Egg Surveys for Small Pelagic Fish (WGACEGG) (Version 1). ICES Tech. Mar. Environ. Sci. (TIMES) 64, 100pp. http://dx.doi.org/10.17895/ices.pub.7462.

Doser, J.W., Leuenberger, W., Sillett, T.S., Hallworth, M.T., Zipkin, E.F., 2021. Integrated community occupancy models: A framework to assess occurrence and biodiversity dynamics using multiple data sources. Methods Ecol. Evol. 13, 919–932. http://dx.doi.org/10.1111/2041-210X.13811.

Ferreras, P., Jiménez, J., Díaz-Ruiz, F., Tobajas, J., Alves, P.C., Monterroso, P., 2021. Integrating multiple datasets into spatially-explicit capture-recapture models to estimate the abundance of a locally scarce felid. Biodivers Conserv. 30, 4317–4335. http://dx.doi.org/10.1007/s10531-021-02309-1.

Gelfand, A.E., Sahu, S.K., Holland, D.M., 2012. On the effect of preferential sampling in spatial prediction. Environmetrics 23 (7), 565–578. http://dx.doi.org/10.1002/env.2169.

Guillera-Arroita, G., Lahoz-Monfort, J.J., 2012. Designing studies to detect differences in species occupancy: power analysis under imperfect detection. Methods Ecol. Evol. 3 (5), 860–869. http://dx.doi.org/10.1111/j.2041-210X.2012.00225.x.

Hefley, T.J., Hooten, M.B., 2016. Hierarchical species distribution models. Curr. Landsc. Ecol. Rep. 1, 87–97. http://dx.doi.org/10.1007/s40823-016-0008-7.

Izquierdo, F., Menezes, R., Wise, L., Teles-Machado, A., Garrido, S., 2022. Bayesian spatio-temporal CPUE standardization: Case study of European sardine (*Sardina pilchardus*) along the western coast of Portugal. Fish. Manag. Ecol. 29 (5), 670–680. http://dx.doi.org/10.1111/fme.12556.

Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., Wild, D.L., 2012. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28 (24), 3290–3297. http://dx.doi.org/10.1093/bioinformatics/bts595.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M., 2016. TMB: Automatic differentiation and Laplace approximation. J. Stat. Softw. 70 (5), 1–21. http://dx.doi.org/10.18637/jss.v070.i05, URL: https://www.jstatsoft.org/index.php/jss/article/view/v070i05.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14. http://dx.doi.org/10.1080/00401706.1992.10485228.

Le Cam, L., 1986. Asymptotic Methods in Statistical Decision Theory. SpringerVerlag.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 423–498. http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x.

MacKenzie, D., Nichols, J.D., Royle, J.A., Pollock, K., Bailey, L., Hines, J.E., 2006. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence, second ed. Academic Press, Burlington, MA.

Manceur, A., Kühn, I., 2014. Inferring model based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. Methods Ecol. Evol. 5, http://dx.doi.org/10.1111/2041-210X.12224.

Martínez-Minaya, J., Cameletti, M., Conesa, D.V., Pennino, M.G., 2018. Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stoch. Environ. Res. Risk Assess. 32, 3227–3244. http://dx.doi.org/10.1007/s00477-018-1548-7.

Paradinas, I., Conesa, D.V., López-Quílez, A., Bellido, J.M., 2017. Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling. Spat. Stat. 22, 434–450. http://dx.doi.org/10.1016/J.SPASTA.2017.08.001.

Pati, D., Reich, B.J., Dunson, D.B., 2011. Bayesian geostatistical modelling with informative sampling locations. Biometrika 98 (1), 35–48. http://dx.doi.org/10.1093/biomet/asq067.

Pennino, M.G., Coll, M., Albo-Puigserver, M., Fernández-Corredor, E., Steenbeek, J., ans Maria González, A.G., Esteban, A., Bellido, J.M., 2020. Current and future influence of environmental factors on small pelagic fish distributions in the Northwestern Mediterranean Sea. Front. Mar. Sci. http://dx.doi.org/10.3389/fmars.2020.00622.

Pennino, M.G., Conesa, D., López-Quílez, A., Muñoz, F., Fernández, A., Bellido, J.M., 2016. Fishery-dependent and -independent data lead to consistent estimations of essential habitats. ICES J. Mar. Sci. 73 (9), 2302–2310. http://dx.doi.org/10.1093/icesjms/fsw062.

Pennino, M.G., Paradinas, I., Illian, J.B., Muñoz, F., Bellido, J.M., López-Quílez, A., Conesa, D., 2019. Accounting for preferential sampling in species distribution models. Ecol. Evol. 9 (1), http://dx.doi.org/10.1002/ece3.4789.

Rosenberg, A.A., Bolster, W.J., Alexander, K.E., Leavenworth, W.B., Cooper, A.B., McKenzie, M.G., 2005. The history of ocean resources: modeling cod biomass using historical records. Front. Ecol. Environ. 3 (2), 84–90. http://dx.doi.org/10.1890/1540-9295(2005)003[0078:THOORM]2.0.CO;2.

Rufener, M.-C., Kristensen, K., Nielsen, J.R., Bastardie, F., 2021. Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. Ecol. Appl. 31 (8), http://dx.doi.org/10.1002/eap.2453.

Silva, D., Menezes, R., 2024. A simulation comparison of spatial models for preferential sampling. In: New Frontiers in Statistics and Data Science. Vol. 398, Springer Nature Switzerland, pp. 259–272. http://dx.doi.org/10.1007/978-3-031-68949-9_19.

Silva, D., Menezes, R., Moreno, A., Teles-Machado, A., Garrido, S., 2024. Environmental effects on the spatiotemporal variability of sardine distribution along the Portuguese Continental coast. JABES http://dx.doi.org/10.1007/s13253-023-00577-8.

Skaug, H.J., Fournier, D.A., 2006. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Comput. Statist. Data Anal. 51 (2), 699–709. http://dx.doi.org/10.1016/j.csda.2006.03.005.

Steele, E., Tucker, A., 2008. Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets. J. Biomed. Inform. 41 (6), 914–926. http://dx.doi.org/10.1016/j.jbi.2008.01.011.

Tehrani, N.A., Naimi, B., Jaboyedoff, M., 2022. A data-integration approach to correct sampling bias in species distribution models using multiple datasets of breeding birds in the Swiss Alps. Ecol. Inform. 69, 101501. http://dx.doi.org/10.1016/j.ecoinf.2021.101501.