Change of support for heavy-tailed zero inflated data: application to spatially aggregated ecological data for fine-scale species distribution inference

Baptiste Alglave, Bastien Mourguiart, Kasper Kristensen, Etienne Rivot, Mathieu Woillez, Youen Vermard, Marie-Pierre Etienne

Change of support for heavy-tailed zero inflated data: application to spatially aggregated ecological data for fine-scale species distribution inference

Baptiste Alglave

Lab-STICC, Univerisé Bretagne Sud, Vannes, France

E-mail: Baptiste.alglave@univ-ubs.fr

Bastien Mourguiart

DYNECO, Ifremer, Plouzané, France

Kasper Kristensen

DTU-Aqua, Kemitorvet, Kongens Lyngby, Denmark

Etienne Rivot

DECOD (Ecosystem Dynamics and Sustainability), L'Institut Agro, INRAE, Ifremer, Nantes, France

Mathieu Woillez

DECOD (Ecosystem Dynamics and Sustainability), L'Institut Agro, INRAE, Ifremer, Brest, France

Youen Vermard

DECOD (Ecosystem Dynamics and Sustainability), L'Institut Agro, INRAE, Ifremer, Nantes, France

Marie-Pierre Etienne

IRMAR, Rennes University, Rennes, France

Summary. In environmental sciences, available data are often available at a coarse resolution. This can result in a mismatch between the data resolution and the resolution at which process inferences should be made. This misalignment, usually named as change of support (COS) issue, can lead to biased inferences if not addressed in the models.
Yet, solutions to the COS issue have only been proposed for a limited number of simple observational processes (e.g. Poisson or Gaussian processes), which narrows their range of application.
Motivated by a fisheries science case study, we introduce a hierarchical method addressing COS issue for zero-inflated data with highly skewed tails. Such data are common in environmental sciences but are not handled by existing COS methods. Our approach requires to know the spatial locations of point-level data and considers that aggregated available data are convolutions of these point-level data.
We assess the accuracy of our method through a simulation study, describing different scenarios of COS. Subsequently, we apply our model to a motivating case study, focusing on the distribution of the common sole in the Bay of Biscay. Our findings illustrate that our approach provides better estimates and predictions than the ad hoc methods used to geoprocess aggregated data and refine their resolution.

1. Introduction

In the field of environmental science, natural processes acts at different spatial scales and available data do not always have the same spatial resolution. For instance, numerous ecological analyses aim at inferring drivers of species distribution to predict suitable habitats over a continuous spatial domain (Carson and Flemming, 2014; Simpson et al., 2012). Yet, observations (e.g. count of individuals) and predictor variables (e.g. climatic conditions) are often available at coarser spatial resolution, which results from some sorts of spatial aggregation (Gelfand et al., 2010). This discrepancy between the resolution of the process and the data are referred to as spatial misalignment (Wakefield and Shaddick, 2006) and is frequent in environmental sciences such as health science (Young and Gotway, 2007), climate science (Parker et al., 2015) or ecology (Gilbert et al., 2021)).

The spatial misalignment between a finely-resolved local process and its coarsely-resolved descriptors is one of the most classical change of support (COS) issue, also

known as point-to-area misalignment, modifiable area unit problem, ecological fallacy, or ecological bias in spatial statistics (Gotway and Young, 2002). This issue arises from the aggregation process that generates the coarsely-resolved data (Chiles and Delfiner, 2012; Rivoirard, 2005; Gotway and Young, 2007; Young and Gotway, 2007). Indeed, spatial data aggregation can alter the signal and introduce an aggregation bias (also known as a scale effect). It can also modify the distribution of confounding variables, leading to a specification bias (also known as a zoning effect).

Addressing COS is relatively straightforward for count data following a Poisson distribution (Gilbert et al., 2021; Gotway and Young, 2007; Mugglin et al., 2000; Pacifici et al., 2019) or continuous data modelled using Gaussian or Gamma distributions (Berrocal et al., 2010; Gelfand et al., 2001; Wikle and Berliner, 2005). However, environmental data rarely follow exact Poisson or Gaussian observations. It is quite common, especially in ecology, to encounter data that exhibit zero-inflation or heavy-tailed distributions (Lecomte et al., 2013; Thorson, 2018).

In marine ecology, COS issue often arises when inferring fish spatial distribution at a relatively fine resolution using commercial catch declaration data, which are typically registered at the resolution of administrative coarse-areal units (Hintzen et al., 2021). Addressing COS issue is crucial because catch declaration data constitute the most extensive source for mapping fish spatial distribution (Alglave et al., 2023) that is central to design management measures (Jansen et al., 2018). However, no methods exist that correctly model commercial declarations while accounting for COS. The most commonly used approach consists in modelling commercial declaration data using a two-step approach that involves (i) uniformly reallocating areal-level declaration data to point-level fishing locations obtained from Vessel Monitoring System (VMS) data (Hintzen et al., 2012; Bastardie et al., 2010), and (ii) fitting a statistical model to the downscaled point-level data considered as observations (Alglave et al., 2022). However, this approach implicitly assumes that catches are uniform and constant across fishing locations within administrative units, which is unlikely for most species. Moreover, declaration data often exhibit zero-inflation or heavy-tailed structure, rendering the existing COS methods unsuitable (Berrocal et al., 2010; Gotway and Young, 2007).

In this paper, after detailing an example of COS in the context of fisheries science (Section 2), we introduce a novel joint hierarchical model designed to address COS for zero-inflated positive continuous data (Section 3). The model assumes that fishing locations are precisely known from GPS data. We rely on a latent spatial field to explicitly represent the spatial distribution of the resource and its aggregation over a coarser spatial area through a convolution of observations realized over the fixed locations. The observation process is then modelled at the scale of the aggregation through a zero-inflated model. We evaluate the performance of this new method against the commonly used two-step approach through a simulation study (Section 4). Finally, we demonstrate the model's application in mapping the spatial distribution of common sole (*Solea solea*) in the Bay of Biscay (Section 5).

2. An example of COS in the context of fisheries science

A key objective in marine ecology is to predict the spatial distributions of fish biomass (denoted $\mathbf{S}$) on a spatial domain (denoted $\mathcal{D}$) using a set of environmental predictors $\mathbf{\Gamma}$. To achieve this, marine ecologists initially rely on data collected by scientific surveys that record the quantity (as weights) of fish caught at precisely known locations (Nielsen, 2015). They are referred to as **point-level data** and henceforth denoted as $Y(x)$ for a catch at location $x \in \mathcal{D}$. The spatial support of this data align with the intended resolution for modelling (Figure 1). However, scientific surveys are often constrained by low sample sizes.

Other valuable and extensive data are commercial fishing declarations also called logbooks (Gerritsen and Lordan, 2011). These data consist of catch weights of fish aggregated at a coarse spatial resolution (hereafter referred to as **areal-level data**). Typically in Europe, the administrative requirement is to declare, for each species of interest, the total retained weight at the resolution of ICES rectangles, i.e., rectangles of $0.5° \times 1°$covering approximately 3000 km^2. Another valuable data source is VMS that precisely locate fishing vessels at regular time interval. Fisheries scientists have developed censors and algorithms to identify fishing locations (here denoted $(x_1, \ldots, x_m)$) from vessel tracks. If fishers VMS locations are precise, allocating the right behavior (fishing,

steaming or stopping at port or at sea) is dependent on the algorithm performance to distinguish between the different states (Vermard et al., 2010). Fishing declarations are then spatialised using VMS data (Figure 2, see section 3.3.1 for more details on notations).

Commercial catch declarations are massive, but raise a COS issue when using these to infer fish spatial distribution at a fine spatial scale. Indeed, even if the precision of the fishing locations is high, the catches realized at each fishing locations is unknown and the only available data is the total weight of fish caught in each administrative unit $a \in \mathcal{A}$, denoted as $W(a)$, where $\mathcal{A}$ represents the set of all administrative units.

As an additional issue, fishing data are often characterized by an excess of zeroes (no catch of a given species in a given fishing operation), which cannot be modelled by simple observational processes (e.g., Poisson or Gaussian). Motivated by this case study, we formulated a COS method tailored for zero-inflated data with highly-skewed tail.

3. Handling COS with zero-inflation

The point-level and areal-level data sources are integrated within an integrated hierarchical statistical framework to infer a unique latent spatial random field. Below we first present the model for the spatial random field, as defined in a previous paper (Alglave et al., 2022). We then define the sampling distribution of point-level data as a zero-inflated distribution. Then, we elaborate on a common ad hoc two-step approach to bypass the COS problem arising from areal-level data. Finally, we present a new approach that directly relates areal-level data to the hidden spatial random field to account for COS.

3.1. Hidden spatial random field

Let's denote the spatial domain as $\mathcal{D} \subset \mathbb{R}^2$ and $\boldsymbol{S} = (S(x), x \in \mathcal{D})$ a spatial random field of interest, such as the biomass of some species in the context of fisheries science. $\boldsymbol{S}$ is assumed to be a spatial log-Gaussian Random Field (GRF) defined as:

$$\log(S(x)) = \mu + \boldsymbol{\beta} \cdot \Gamma(x) + \delta(x), \tag{1}$$

where $\boldsymbol{\delta} = (\delta(x), x \in D)$ is a zero mean isotropic GRF with a Matern covariance function and $\boldsymbol{\Gamma} = (\Gamma(x), x \in \mathcal{D})$ a known field of predictors, $\boldsymbol{\beta}$ are the parameters linking $\boldsymbol{\Gamma}$ and $\boldsymbol{S}$. In marine ecology $\boldsymbol{\Gamma}$ often represents environmental factors, like the bathymetry or sediment type, and $\boldsymbol{\beta}$ is referred to as the species-environment relationship parameters.

3.2. Observation process of point-level zero-inflated data

Ideally, the observations needed to infer the latent field $S(x)$ would be available at the point level, let's say $\mathbf{Y} = (Y(x_1), \ldots, Y(x_n))$ sampled at a given set of fishing locations $\mathbf{x} = (x_1, \ldots, x_n)$.

Conditionally on the latent field, the observation process for point-level data is modelled through the zero-inflated distribution as proposed by Thorson (2018). Positive continuous data are assumed to be independent conditionally on $\boldsymbol{S}$. The catch at the sampled location x, denoted $Y(x)$, is defined through its conditional density $f_{S(x)}$ defined through a mixture of a Dirac mass at 0 and a lognormal distribution:

$$f_{S(x)}(y) = p(x)\mathbb{I}_{\{y=0\}}(y) + (1 - p(x))\,\Psi\left(\frac{S(x)}{1-p(x)}, \sigma^2\right)\mathbb{I}_{\{y>0\}}(y), \tag{2}$$

where

- $p(x) := \exp(-e^{\xi}S(x))$ is the proportion of the mixture representing the probability to get a zero observation,
- ξ is a parameter controlling the zero-inflation,
- Ψ stands for the density of a lognormal distribution,
- $\mathbb{I}_B(y)$ stands for the indicator function which equals 1 when y belongs to any ensemble B,
- σ^2 is the variance parameter on the log-scale.

The distribution $f_{S(x)}$ is completely specified with three parameters: p the weight of the Dirac mass (i.e., the probability of obtaining a zero), μ the mean of the lognormal

distribution, and its variance σ^2, denoted as $\mathcal{M}_y(p, \mu, \sigma^2)$. Note that at location x,

$$\begin{aligned}
\mathbb{P}\left(Y(x) = 0 \,|\, S(x)\right) &= \exp(-e^{\xi} S(x)) = p(x), \\
\mathbb{E}\left(Y(x) \,|\, Y(x) > 0, S(x)\right) &= \frac{S(x)}{1 - p(x)} = \mu(x), \\
\mathrm{Var}(Y(x) \,|\, Y(x) > 0, S(x)) &= \mu(x)^2 (e^{\sigma^2} - 1),
\end{aligned} \tag{3}$$

While accounting for the zero inflation in the data, this choice allows to represent continuous positive data and ensures that the expected catch at site x equals the local biomass $S(x)$. A more detailed presentation is available in the Supplementary Material (Section S-I).

3.3. Modelling observations from areal-aggregated data

The COS issue arises as we rely on data available only at the aggregated areal-level denoted $W(a)$, obtained as the sum of all point-level data $Y(x)$ in the unit a:

$$W(a) = \sum_{i | x_i \in \mathcal{R}_a} Y(x_i), \tag{4}$$

where $\mathcal{R}_a$ denotes the geographical area corresponding to the administrative unit a. Below we provide two different methods to integrate those areal-aggregated data. Note that both methods work by considering the number and the position of the sampling points within each zone $\mathcal{R}_a$ are known.

3.3.1. The ad hoc two-step COS approach

The two-step approach consists of (1) considering the spatial locations of sampling point as known and generate a pseudo data set by reallocating the aggregated data to those sampling point, (2) integrating those pseudo point-wise data through the sampling distribution defined previously. In the context of fisheries science, the reallocation is typically uniform on each fishing location within a declaration area. It consists of defining the reallocated data $\mathbf{Y}_{a,i}$ associated with declaration/areal observation $W(a)$ as:

$$\mathbf{Y}_{a,i} := \frac{W(a)}{m_a} \mathbb{I}_{\{\mathcal{R}_a\}}(x_i), \quad \forall i = 1, \ldots, n, \tag{5}$$

where m_a is the cardinal of the set $\{x_i \in \mathcal{R}_a\}$. A schematic visualization of the reallocation (or imputation) process is proposed in Figure 2.

As noted by Alglave et al. (2022), the uniform reallocation along the fishing route likely does not match with the true spatial distribution of the catches along the fishing routes which is most often highly heterogeneous. Hence, using those pseudo data is likely to underestimate the spatial heterogeneity of the spatial field within each declaration area.

Moreover, imputing point-level values prior to model fitting artificially increases the sample size by a factor corresponding to the ratio between the number of fishing locations and the number of areal units. This artificial data augmentation, akin to pseudo-replication, can lead to an underestimation of uncertainty associated with estimates (Alston et al., 2023).

Finally, note that this process is not specific to fishery applications. In many other fields of environmental science, aggregated data are often geo-processed to refine their resolution through ad hoc arithmetic methods, such as proportional allocation and zonal addition (Young and Gotway, 2007; Gotway and Young, 2007).

3.3.2. Joint COS model

To overcome the limitations of the two-step approach, we developed a joint COS model that addresses the COS issue. Instead of considering the imputed point-level catches as observed data, we directly relate areal-level data to the underlying point-level spatial process of interest. This avoids the strong hypothesis needed for the reallocation of the total catches to the fishing locations.

Let's denote by $\mathcal{M}_W$ the conditional distribution of areal-level observations $W(a)$ associated with area a, given the sampling points $\boldsymbol{x_a} = (x_{a,1}, ..., x_{a,i}, ..., x_{a,m_a})$ and the spatial random field $\boldsymbol{S}$.

$\mathcal{M}_W$ results from the convolution of zero-inflated lognormal distribution. Its mathematical form is unknown as there is no analytical form for a convolution of zero inflated lognormal distributions. However, given that the individual observations are zero-inflated with heavy tails, $\mathcal{M}_W$ takes the form of a mixture distribution with zero inflation

and long tail. Our approach rely on the hypothesis that the distribution of the variable $W(a)$ has the same zero inflated lognormal form as the distribution of point-level distribution defined in Equation (2). We will discuss this assumption later.

So, let's denote $\mathcal{M}_W$ the corresponding zero inflated lognormal mixture distribution:

$$W(a) \,|\, \boldsymbol{S}, \boldsymbol{x_a} \sim \mathcal{M}_W(p_a^W, \mu_a^W, \sigma_a^{W\,2}) \tag{6}$$

with

- $\boldsymbol{x_a} = (x_{a,1}, ..., x_{a,i}, ..., x_{a,m_a})$ the fishing positions associated with the declaration W_a in area $\mathcal{R}_a$,
- μ_a^W the expected positive biomass,
- p_a^W the proportion of zeros in the mixture,
- $\sigma_a^{W\,2}$ the variance parameter.

Conditionally on the random field $\boldsymbol{S}$ and on the sampling locations, the quantities $p_a^W, \mu_a^W, (\sigma_a^W)^2$ are defined as follow :

(a) Defining $\boldsymbol{Y_a} = (Y(x_{a,1}), \ldots, Y(x_{a,m_a}))$ and using conditional independence with respect to $\boldsymbol{S}$, we have:

$$p_a^W = \mathbb{P}(W_a = 0) = \prod_{i=1}^{m_a} \mathbb{P}(Y(x_{a,i}) = 0) = \exp\left\{ -\sum_{i=1}^{m_a} e^{\xi} \cdot S(x_{a,i}) \right\} \tag{7}$$

(b) The continuous component of the mixture is defined by the expected mean of a positive declaration and a transformation of its variance, which correspond to:

$$\begin{aligned}
&\mu_a^W = \mathbb{E}(W(a) \,|\, W(a) > 0) = \frac{\sum_{i=1}^{m_a} S(x_{a,i})}{1 - p_a^W} \\
&\mathrm{Var}(W(a) \,|\, W(a) > 0) = \frac{\sum_{i=1}^{m_a} \mathrm{Var}\left(Y(x_{a,i})\right)}{1 - p_a^W} - \frac{p_a^W}{(1 - p_a^W)^2} \mathbb{E}(W(a))^2 \\
&\text{with } \mathrm{Var}\left(Y(x_{a,i})\right) = \frac{S(x_{a,i})^2}{1 - p_{a,i}} (e^{\sigma^2} - (1 - p_{a,i})) \quad \text{and } p_{a,i} = \mathbb{P}(Y(x_{a,i}) = 0).
\end{aligned} \tag{8}$$

The details of the computation are given in the Supplementary Material (from section S-II to S-VI).

4. Simulation study

We perform a simulation study to investigate the effect of COS on models' performance, and to explore the influence of the level of aggregation as well as the intensity of zero-inflation in the sampling process.

4.1. Data simulation

We simulate a spatial random field that represents the species distribution (*i.e.*, the spatial distribution of fish biomass) across a virtual spatial domain with dimensions similar to the Bay of Biscay (Figure S1). The spatial distribution of the virtual species is simulated following equation 1 with a single continuous predictor and a spatial random effect, which were both simulated by GRFs with respectively ranges of 1.5 and 0.6 (approximately 50 km) and marginal variances of 0.5 and 1 respectively. The species-habitat relationship (β) is fixed to 2 (the table 1 shows all the parameter values used for the simulation.).

Conditionally on this random field, point-level and areal-level data were simulated to mimic different scenarios of data clustering and zero-inflation. Each simulated scenario was replicated 100 times.

Baseline scenario

Point-level data were generated by virtually sampling 100 points following a stratified sampling scheme over the virtual Bay of Biscay (Figure S1). For each point, biomass observations are simulated following the observation equation of $\mathcal{M}_Y$ (with specific parameters, see Table 1). This corresponds to the standardized scientific data.

We simulate 300 area-level data in square cells across a spatial domain covering 2/3 of the virtual Bay of Biscay (Figure S1). For each areal-level data, we simulate 10 point-level locations spatially clustered to mimic real fishing zones (*i.e.*, small areas targeted by fishermen). This spatial clustering is simulated using a Neymann-Scott process (Waagepetersen, 2007) in two steps: (1) within the areal unit where the areal-level data is simulated, we sample a spatial point representing the center of a single fishing zone, and (2) the 10 sampled locations are uniformly sampled around the fishing zone center

within a squared area approximating the distance of a trawl haul (Figure S2). Finally, at each fishing location, a point-level catch is sampled conditionally on the value of the spatial random field following the distribution $\mathcal{M}_Y$. Last, the point-level data are aggregated at the areal-level.

Effect of spatial clustering

Alternative scenarios were simulated to investigate the effects of the spatial clustering of the point-level fishing locations within the areal units. We simulate three types of areal-level data with one, three, or five visited zones, respectively, within each administrative areal unit (see Figure S2 for illustration).

Proportion of zeros

We modify the observation process for areal-level data to generate four levels of zero-inflation with 0%, 7%, 37%, or 70% of zeroes in the areal-level data.

4.2. Model fitting and comparison

We consider three types of models described in table 2:

- a model with point-level data only (point-level model).
- a model fitted to scientific data and to the reallocated data (two-step approach)
- the model accounting for COS (joint COS approach).

All three models were fitted to the simulated datasets representing the different scenarios (*i.e.*, the baseline, the three scenarios regarding the number of fishing zones, and the four levels of zero-inflation), each replicated 100 times, resulting in 800 fits for each model. These fitted models were then used to predict the species distribution across the entire simulated area.

Inference was conducted using maximum likelihood methods with the package Template Model Builder (Kristensen et al., 2016). We leverage the SPDE approach to efficiently estimate the spatial random effect δ (Lindgren et al., 2011).

We compared the explanatory and predictive performance of the three models by assessing the discrepancy between the estimated parameter $\hat{\beta}$ and the simulated one β, and by computing the mean squared prediction errors (MSPE), respectively. The MSPE quantifies the accuracy of the predictions $\hat{\mathbf{S}}$ over the spatial domain by comparing the simulated latent field $\mathbf{S}$ with the estimated ones $\hat{\mathbf{S}}$ using the formula:

$$MSPE = \frac{\sum_{i=1}^{n}(S(x_i) - (\hat{S}(x_i)))^2}{n}$$

where n is the number of points in the spatial domain.

4.3. Results

The species-environment relationship is unbiased for both the point-level and the joint COS models (Figure 3, right). Conversely, the two-step approach produces biased estimates tending towards zero. The ad hoc two-step approach also tends to produce overly smoothed species distribution predictions (Figure 4). These are direct consequences of the uniform reallocation of areal-level data that smooth the distribution and blur the perception of the effects of predictors in the species-environment relationship. This can also be seen in the variance and zero-inflation parameters that are respectively strongly under- and over-estimated in the two-step approach compared with the joint COS model (Figure S3).

By comparison, in addition to accurately estimating the species-environment relationship, the joint COS model demonstrates the most accurate predictions, with mean squared prediction errors (MSPE) being 1.5 and 2 times lower than those of the two-step and point-level models, respectively (Figure 3, left). The higher predictive performance of the joint COS model may be attributed to the COS part of the model that improves the capacity to capture fine-scale predictor effects. However, it is important to note that the joint COS model is the only model that face convergence issues, with only 63% rate of convergence (Table 3).

The ad hoc two-step approach demonstrates high sensitivity to the number of fishing zones within the administrative areal units (*i.e.*, the clustering resolution of the fishing locations), with the MSPE increasing from 0.3 to 0.75 on average and a bias in species-environment relationship coefficients increasing with the number of fishing zones (Figure

5, left). By contrast, the predictive and explanatory performances of the joint COS model remains almost constant and better than the point-level model when increasing numbers of fishing zones.

The proportion of zeroes in observations negatively impacts the performance of the two-step approach and the joint COS model (Figure 5, right). Predictive performances of the two-step and joint COS models decrease as the amount of zeroes in the data increases but they still have better predictive performance than the point-level model (Figure 5, top-right). Furthermore, the joint COS model outperforms the two-step approach in predictive accuracy. The benefit of choosing the joint COS model over the two-step approach increases with higher levels of zero-inflation.

The prevalence of zeroes also increases convergence issues for the joint COS model, with only 17% of the fitted models that converged in the worst-case scenario (70% zeroes).

5. Application: Fine-scale distribution of common sole in the Bay of Biscay (NE Atlantic)

The three models were applied using a real case study to map the distribution of common sole (*Solea solea*, Linnaeus, 1758) in the Bay of Biscay (NE Atlantic).

5.1. Data

Point-level data are scientific data sourced from the DATRAS database for the Orhago beam trawl survey (Coupeau and Biais, 2019), which employs similar fishing methodologies as commercial fishermen. The scientific dataset were filtered to retain specimens exceeding the minimum catch size threshold *i.e.* $>$ 24 cm (ICES), ensuring alignment with the size structure observed in commercial data.

Area-aggregated data are commercial catch declaration data, including logbook entries detailing catch weights, and data from the VMS providing geolocations of fishing activities. Commercial data were gathered from "bottom trawlers" operating in the Bay of Biscay, which target *Solea solea*. For a comprehensive understanding of the data preprocessing procedures, including the integration of information from both logbooks and

VMS data, as well as the filtration of targeted fishing activities, refer to Alglave et al. (2022).

Biomass data from both sources were standardized by fishing effort to yield Catch Per Unit of Effort (CPUE), measured in kg.h^{-1}.

Sediment type can strongly influence sole spatial distribution and is used as an environmental predictor in the model (Holzhauer et al., 2019; Künitzer et al., 1992). Sediment type data were extracted from the EMODNET platform with resolution 0.05°, categorized into two main types: 'sand and coarse substrate' and 'mud'. Sediment type was coded as a binary variable taking value 0 for 'sand and coarse substrate' and 1 for 'mud' sediments. Therefore, in equation 1, μ represents the expected biomass in sand and coarse substrate, while β is the effect of "mud" on sole distribution.

Finally, note the different data sources may not share the same scale of response, meaning they may not exhibit the same probability and efficiency of catch. To address such differences, we introduce an offset parameter $k = \frac{\mathbb{E}(Y^{(1)}(x_i) \mid S(x_i), x_i)}{\mathbb{E}(Y^{(2)}(x_i) \mid S(x_i), x_i)}$, serving as a scaling factor between some data sources 1 and 2 (*i.e.* here $Y^{(1)}(x_i)$ and $Y^{(2)}(x_i)$ are the commercial and the scientific datasets at the point-level). In the context of fisheries, such scaling factor typically denotes relative differences in catchability between scientific survey and commercial catch data.

5.2. Model fitting

The joint COS model encountered convergence issues, particularly in estimating some parameters such as the range parameter. To facilitate convergence, we incorporated onboard observer data from the same fleet into the analysis. These data can be regarded as point-level commercial catch data, comprising 86 samples for the corresponding time step. The integration of these data provides direct point-level observations of the declarations data and helps estimating the observation parameters of the area-level data (including observation variance and zero-inflation parameter of the declaration data).

Moreover, as commonly done in fisheries modelling with automatic differentiation methods (Fournier et al., 2012), we applied a phased optimization approach to initialize the optimization algorithm for the joint COS model. The estimates derived from the

two-step approach serve as the starting point for the optimization algorithm used in estimating the joint COS model. Subsequently, parameters that are hard to estimate during the initial optimization phases (such as intercept μ, covariate effect β, range, and marginal variance) were fixed, and left free one by one in the next phases of estimation.

5.3. Results

Predictor effects and spatial predictions of species distribution differ between the joint COS model and the ad hoc two-step approach (Figures 6, 7). Notably, the joint COS model estimates a larger substrate effect compared to the ad hoc two-step approach, aligning more closely with estimates derived from the point-level model (Figure 6). Consistently with simulations, the zero-inflation parameter ξ estimated by the joint COS model for areal data is notably smaller than the one estimated by the ad hoc two-step approach (Figure S3). Also, the observation variance of areal-data estimated by the joint COS model is higher, suggesting that the this model estimates more noise in the areal data.

Furthermore, the joint COS model produces wider confidence interval than the ad hoc approach (Figure 6). Compared with the two-step approach, the joint COS model yields notably wider confidence intervals for the species-habitat relationship (β), the marginal variance, the range, the zero-inflation parameter (ξ_{areal}), and the variance of areal-data (σ_{areal}). This divergence indicates an underestimation of uncertainty by the two-step approach.

Regarding the spatial predictions of species distribution (Figure 7), the spatial patterns are overall consistent for the two-step approach and the joint COS model. Areas of high densities are consistent in the North of the Bay of Biscay (4°W - 47.5°N), offshore the estuary of Gironde (1.5°W, 45.5°N) and along the Vendée coast (2°W - 46.5°N). The point-level data is shaped by the sediment effect and produces a smooth pattern due to low sampling density.

Consistently with simulations, the two-step approach generates smoother spatial predictions relative to the joint COS model. The coefficient of variation of the latent field in the two-step approach equals 0.47 while it is 1.07 for the joint COS approach. Also,

some low biomass areas are evidenced by the joint COS model for instance in the North of the Bay of Biscay (3°W - 47°).

As for the point-level model, the joint COS model predictions are shaped by the sediment effect while the two-step approach predictions are not.

6. Discussion

6.1. The benefit of a statistical approach for COS

Dealing with COS is a key issue in spatial statistics. An extensive literature has been dedicated to develop statistical methods that predict fine-scale processes from coarse resolution data (Wikle et al., 2019; Wakefield and Lyons, 2010). However, in many cases, data resolution refinement is often achieved through ad hoc arithmetic methods such as proportional allocation or zonal addition. These methods can alter the data and result in information loss (Young and Gotway, 2007; Gotway and Young, 2002) or artificially inflate the amount of data (Alglave et al., 2022). In this paper, we propose a model capable of handling complex environmental data, including zero-inflated or highly skewed data distributions. We demonstrate that failure to properly account for COS can bias the link with linear predictors and substantially overestimates the precision of estimates.

6.2. The hierarchical structure of the approach and the point-level sampling distribution

Our strategy for handling COS follows the conventional hierarchical framework structure. The hierarchical structure allows us to establish the link between the hidden spatial random field and areal-level data. Given the observation distribution of a zero-inflated point-level observation, we calculated the probability to get a zero at the areal-level, and the sampling distribution of a positive value was approximated by a lognormal distribution by identifying the first two moments.

Our approach stands on the hypothesis that the convolution of zero-inflated lognormal distribution is well approximated by a zero-inflated lognormal distribution. Previously, some studies have intended to identify approximations for convolutions of lognormal ob-

servations. They concluded that in practice this convolution could be modelled through a lognormal distribution by matching the moment of the single variables and the resulting convoluted variable, though there is no explicit relationship (Furman et al., 2020; Beaulieu et al., 1995). Here, we demonstrated through the simulations that this approximation holds reasonably well for our model. However, it is worth noting that this hypothesis can be violated depending on the context. Exploring alternative observation models that satisfy additive properties, such as the Gamma distribution, could be an interesting avenue for future research.

Another common approach in the COS literature is 'Block kriging' (Gelfand et al., 2001; Pacifici et al., 2017). In this approach, the aggregation process is modelled within the latent field. By defining a spatial block B (e.g., a statistical rectangle), the average latent field over the spatial block is considered as $S(B) = |B|^{-1} \int_B S(x)dx$. Here, observations are assumed to stem from a distribution $\mathcal{M}_B$ conditionally on $S(B)$, following $D_j|S(B) \sim \mathcal{M}_B(S(B), \sigma^2)$. While this method considers areal-level data arising from the averaged biomass over the spatial block, it may encounter similar challenges as reallocated data and could lead to smoothed estimates of the linear predictor. Indeed, the observations and the latent field are still linked at a coarse resolution and the observations are not downscaled in the model. Furthermore, our approach maintains sparsity in the Hessian of the likelihood and improves computation time, whereas Block kriging would entail losing sparsity by integrating over block areas B.

6.3. Future perspectives for the framework

Aggregated declarative data are important source of information in numerous fields of environmental science such as ecology, epidemiology, and environmental sciences. These datasets typically include hunting records (Gilbert et al., 2021), administrative healthcare data (Morel et al., 2020), and teledetection data (Garrigues et al., 2008). While these data are not specifically tailored for scientific analysis, they hold immense potential for research and expertise once related methodological challenges are addressed. Several drawbacks may hinder the use of these data, including issues with data aggregation, sampling bias as seen in citizen science programs (Botella et al., 2021), and species mis-

specification (Botella et al., 2018). Also, in our specific case, vessels may cross several administrative rectangles for one single declaration and there might be uncertainty on the administrative unit the observation is assigned to (Gábor et al., 2022). Tackling these sources of bias all together with COS is a major challenge, and our model could be extended to account for these other bias.

Acknowledgments

The authors are grateful to the Direction générale des affaires maritimes, de la pêche et de l'aquaculture (DGAMPA) and Ifremer (Système d'Informations Halieutiques - SIH) who provided the aggregated VMS and logbooks data. The findings and conclusions of the present paper are those of the authors.

Fundings

The authors declare no specific funding for this work.

Competing interests

The authors declare there are no competing interests

Data availability statement

Survey data are available through the DATRAS portal (https://www.ices.dk/data/data-portals/Pages/DATRAS.aspx) with the package 'icesDatras'. Logbooks and VMS data are confidential data and they are available on specific request to DGAMPA.

References

Alglave, B., Rivot, E., Etienne, M.-P., Woillez, M., Thorson, J. T. and Vermard, Y. (2022) Combining scientific survey and commercial catch data to map fish distribution. *ICES Journal of Marine Science*, fsac032. URL: `https://doi.org/10.1093/icesjms/fsac032`.

Alglave, B., Vermard, Y., Rivot, E., Etienne, M.-P. and Woillez, M. (2023) Identifying mature fish aggregation areas during spawning season by combining catch declarations and scientific survey data. *Canadian Journal of Fisheries and Aquatic Sciences*, **80**, 808–824.

Alston, J. M., Fleming, C. H., Kays, R., Streicher, J. P., Downs, C. T., Ramesh, T., Reineking, B. and Calabrese, J. M. (2023) Mitigating pseudoreplication and bias in resource selection functions with autocorrelation-informed weighting. *Methods in Ecology and Evolution*, **14**, 643–654.

Bastardie, F., Nielsen, J. R., Ulrich, C., Egekvist, J. and Degel, H. (2010) Detailed mapping of fishing effort and landings by coupling fishing logbooks with satellite-recorded vessel geo-location. *Fisheries Research*, **106**, 41–53.

Beaulieu, N. C., Abu-Dayya, A. A. and McLane, P. J. (1995) Estimating the distribution of a sum of independent lognormal random variables. *IEEE Transactions on Communications*, **43**, 2869.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010) A bivariate space-time downscaler under space and time misalignment. *The annals of applied statistics*, **4**, 1942.

Botella, C., Joly, A., Bonnet, P., Monestiez, P. and Munoz, F. (2018) Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences*, **6**, e1029.

Botella, C., Joly, A., Bonnet, P., Munoz, F. and Monestiez, P. (2021) Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, **12**, 933–945.

Carson, S. and Flemming, J. M. (2014) Seal encounters at sea: A contemporary spatial approach using r-inla. *Ecological Modelling*, **291**, 175–181.

Chiles, J.-P. and Delfiner, P. (2012) *Geostatistics: modeling spatial uncertainty*, vol. 713. John Wiley & Sons.

Coupeau, Y. and Biais, G. (2019) ORHAGO 19. URL: `https://campagnes.flotteoceanographique.fr/campagnes/18001044/`. Publisher: Sismer.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A. and Sibert, J. (2012) Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, **27**, 233–249.

Furman, E., Hackmann, D. and Kuznetsov, A. (2020) On log-normal convolutions: An analytical–numerical method with applications to economic capital determination. *Insurance: Mathematics and Economics*, **90**, 120–134.

Gábor, L., Jetz, W., Lu, M., Rocchini, D., Cord, A., Malavasi, M., Zarzo-Arias, A., Barták, V. and Moudrỳ, V. (2022) Positional errors in species distribution modelling are not overcome by the coarser grains of analysis. *Methods in Ecology and Evolution*, **13**, 2289–2302.

Garrigues, S., Allard, D. and Baret, F. (2008) Modeling temporal changes in surface spatial heterogeneity over an agricultural site. *Remote Sensing of Environment*, **112**, 588–602.

Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010) *Handbook of spatial statistics*. CRC press.

Gelfand, A. E., Zhu, L. and Carlin, B. P. (2001) On the change of support problem for spatio-temporal data. *Biostatistics*, **2**, 31–45.

Gerritsen, H. and Lordan, C. (2011) Integrating vessel monitoring systems (VMS) data with daily catch data from logbooks to explore the spatial distribution of catch and effort at high resolution. *ICES Journal of Marine Science*, **68**, 245–252.

Gilbert, N. A., Pease, B. S., Anhalt-Depies, C. M., Clare, J. D., Stenglein, J. L., Townsend, P. A., Van Deelen, T. R. and Zuckerberg, B. (2021) Integrating harvest and camera trap data in species distribution models. *Biological Conservation*, **258**, 109147.

Gotway, C. A. and Young, L. J. (2002) Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**, 632–648.

— (2007) A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, **16**, 115–135.

Hintzen, N. T., Aarts, G., Poos, J. J., Van der Reijden, K. J. and Rijnsdorp, A. D. (2021) Quantifying habitat preference of bottom trawling gear. *ICES Journal of Marine Science*, **78**, 172–184.

Hintzen, N. T., Bastardie, F., Beare, D., Piet, G. J., Ulrich, C., Deporte, N., Egekvist, J. and Degel, H. (2012) VMStools: Open-source software for the processing, analysis and visualisation of fisheries logbook and VMS data. *Fisheries Research*, **115**, 31–43. Publisher: Elsevier.

Holzhauer, H., Borsje, B. W., Van Dalfsen, J. A., Wijnberg, K. M., Hulscher, S. J. and Herman, P. M. (2019) Benthic species distribution linked to morphological features of a barred coast. *Journal of marine science and engineering*, **8**, 16.

ICES () Working group for the bay of biscay and the iberian waters ecoregion (WGBIE). URL: `http://www.ices.dk/sites/pub/PublicationReports/Forms/DispForm.aspx?ID=36841`. Publisher: ICES.

Jansen, H., Bastardie, F., Eero, M., Hamon, K. G., Hinrichsen, H.-H., Marchal, P., Nielsen, J. R., Le Pape, O., Schulze, T. and Simons, S. (2018) Integration of fisheries into marine spatial planning: Quo vadis? *Estuarine, Coastal and Shelf Science*, **201**, 105–113.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. and Bell, B. M. (2016) TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, **70**, 1–21. URL: `https://www.jstatsoft.org/index.php/jss/article/view/v070i05`.

Künitzer, A., Basford, D., Craeymeersch, J., Dewarumez, J., Dörjes, J., Duineveld, G., Eleftheriou, A., Heip, C., Herman, P., Kingston, P. et al. (1992) The benthic infauna of the north sea: species distribution and assemblages. *ICES Journal of Marine Science*, **49**, 127–143.

Lecomte, J.-B., Benoıt, H. P., Ancelet, S., Etienne, M.-P., Bel, L. and Parent, E. (2013) Compound poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling. *L'Institut des Sciences et Industries du Vivant et de l'Environnement (AgroParisTech)*, 37.

Lindgren, F., Rue, H. and Lindstrom, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498. Publisher: Wiley Online Library.

Morel, M., Bacry, E., Gaïffas, S., Guilloux, A. and Leroy, F. (2020) Convsccs: convolutional self-controlled case series model for lagged adverse event detection. *Biostatistics*, **21**, 758–774.

Mugglin, A. S., Carlin, B. P. and Gelfand, A. E. (2000) Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, **95**, 877–887.

Nielsen, J. R. (2015) *Methods for integrated use of fisheries research survey information in understanding marine fish population ecology and better management advice: improving methods for evaluation of research survey information under consideration of survey fish detection and catch efficiency*. Wageningen University.

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. and Collazo, J. A. (2017) Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, **98**, 840–850.

Pacifici, K., Reich, B. J., Miller, D. A. and Pease, B. S. (2019) Resolving misaligned spatial data with integrated species distribution models. *Ecology*, **100**, e02709.

Parker, R. J., Reich, B. J. and Sain, S. R. (2015) A multiresolution approach to estimating the value added by regional climate models. *Journal of Climate*, **28**, 8873–8887.

Rivoirard, J. (2005) Concepts and methods of geostatistics. In *Space, Structure and Randomness: Contributions in Honor of Georges Matheron in the Field of Geostatistics, Random Sets and Mathematical Morphology*, 17–37. Springer.

Simpson, D., Lindgren, F. and Rue, H. (2012) Think continuous: Markovian gaussian models in spatial statistics. *Spatial Statistics*, **1**, 16–29.

Thorson, J. T. (2018) Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative. *Canadian Journal of Fisheries and Aquatic Sciences*, **75**, 1369–1382. Publisher: NRC Research Press.

Vermard, Y., Rivot, E., Mahévas, S., Marchal, P. and Gascuel, D. (2010) Identifying fishing trip behaviour and estimating fishing effort from vms data using bayesian hidden markov models. *Ecological Modelling*, **221**, 1757–1769.

Waagepetersen, R. P. (2007) An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics*, **63**, 252–258.

Wakefield, J. and Lyons, H. (2010) Spatial aggregation and the ecological fallacy. *Handbook of spatial statistics*, **541**, 558.

Wakefield, J. and Shaddick, G. (2006) Health-exposure modeling and the ecological fallacy. *Biostatistics*, **7**, 438–455.

Wikle, C. K. and Berliner, L. M. (2005) Combining information across spatial scales. *Technometrics*, **47**, 80–91.

Wikle, C. K., Zammit-Mangion, A. and Cressie, N. (2019) *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.

Young, L. J. and Gotway, C. A. (2007) Linking spatial data from different sources: the effects of change of support. *Stochastic Environmental Research and Risk Assessment*, **21**, 589–600.

Table 1. Parameter values for the simulations. 'areal' refer to the areal-level data. 'point' refer to the point-level data.

Parameters	Simulation values
μ	2
β	2
Range of δ	0.6 ($\approx$ 50 km)
Marginal variance of δ	1
ξ_{areal}	-1
σ_{areal}	1
k_{areal}	1
ξ_{point}	0
σ_{point}	0.8

Table 2. Model configurations.

Model name	Configuration
Point-level model	The model fitted to point-level data only.
Two-step approach	The original model fitted with imputed point-level data in Alglave et al. (2022).
Joint COS model	The alternative approach introduced in this paper where the biomass model is fitted using areal-level data and few precisely point-level data.

Table 3. (Simulations) Percentage of convergence for the alternative models (point-level only, two-step and joint approach).

Model	Convergence (%)
Two-step approach	100
Joint approach	63
Point-level model	100

Table 4. (Simulations) Percentage of convergence for various level of zero-inflation. Two-step approach: integrated model fitted to reallocated observations. Joint approach: integrated model accounting for change of support.

Model	Proportion of zero values	Convergence (%)
Two-step approach	0 %	100
Two-step approach	7 %	100
Two-step approach	37 %	100
Two-step approach	70 %	99
Joint approach	0 %	78
Joint approach	7 %	63
Joint approach	37 %	49
Joint approach	70 %	17

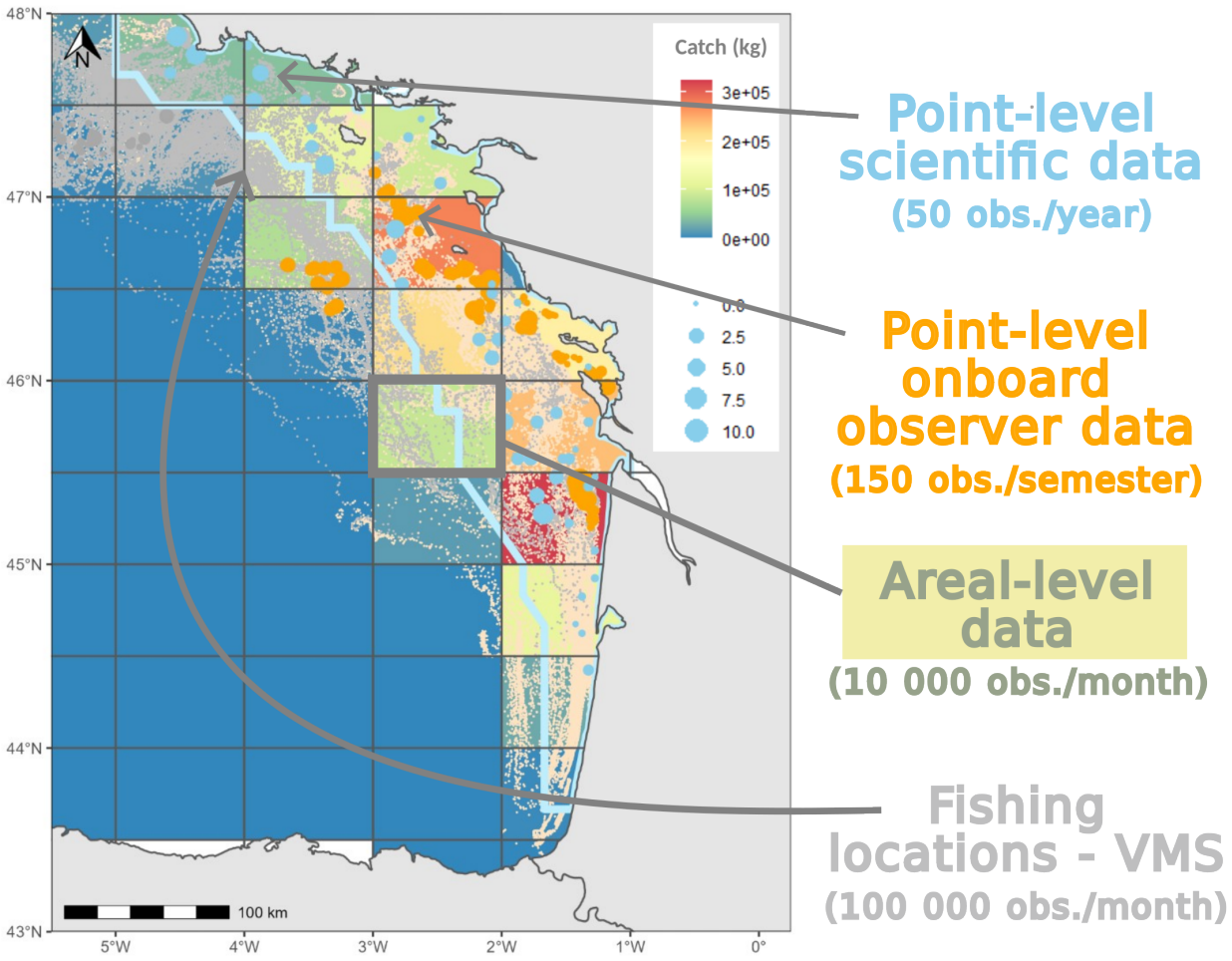

Fig. 1. Maps of the different data sources with the corresponding spatio-temporal sampling densities. Grey rectangles represent the level of aggregation of the catch declarations (*i.e.* the grid resolution).

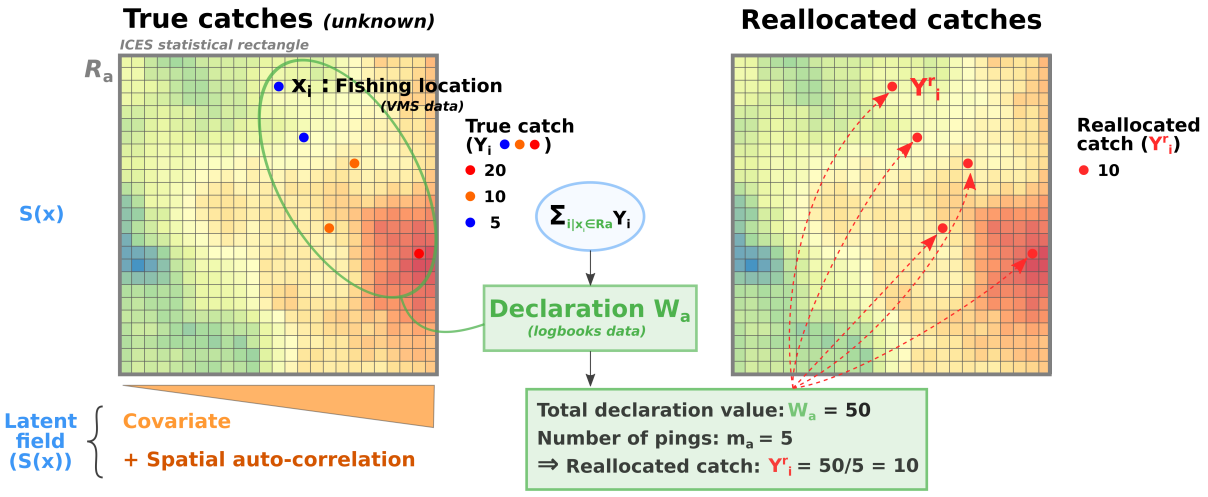

Fig. 2. Schematic representation of the reallocation process. The biomass field (the spatial field, $S(x)$) depends on a covariate ($\Gamma(x)$) and a GRF ($\delta(x)$). The covariate is the x-axis. It has a positive effect on biomass values (*i.e.* biomass is higher on the right of the grid than on the left). The spatial random effect creates an area of high density on the bottom-right of the latent field. The study domain is considered as a rectangle (grey square). Points represent catches made by fishermen and their colors are related to the weights of the catches. These punctual catches (Y_i) belong to the same rectangle R_a and are summed to constitute the declaration $W_a = \sum_{i|x_i \in \mathcal{R}_a} Y(x_i) = 50$ that is recorded in catch declaration data (logbook) at the resolution of the statistical rectangle. Based on VMS data, we know the fishing positions x_i. In standard processing, to refine the spatial resolution of the declaration, W_a is uniformly reallocated over the related fishing positions x_i. This strongly homogenizes the catch and the effect of the environment disappears from the reallocated catch Y^r_i.

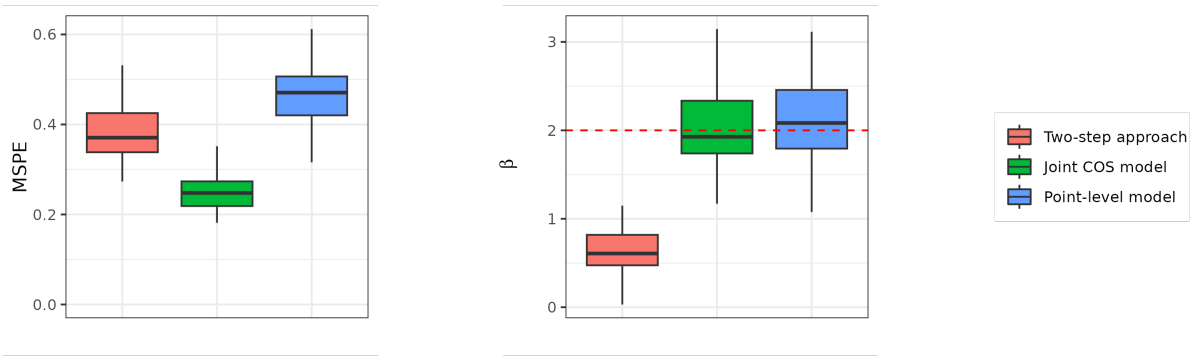

Fig. 3. (Simulations) Boxplot of the performance metrics for the alternative models. Blue boxplots: model fitted to point-level data only. Red boxplots: integrated model fitted to reallocated data. Green boxplots: integrated model accounting for change of support, joint approach. MSPE : mean squared prediction error. β: species-environment parameter. Red line: true value for the species-environment parameter.

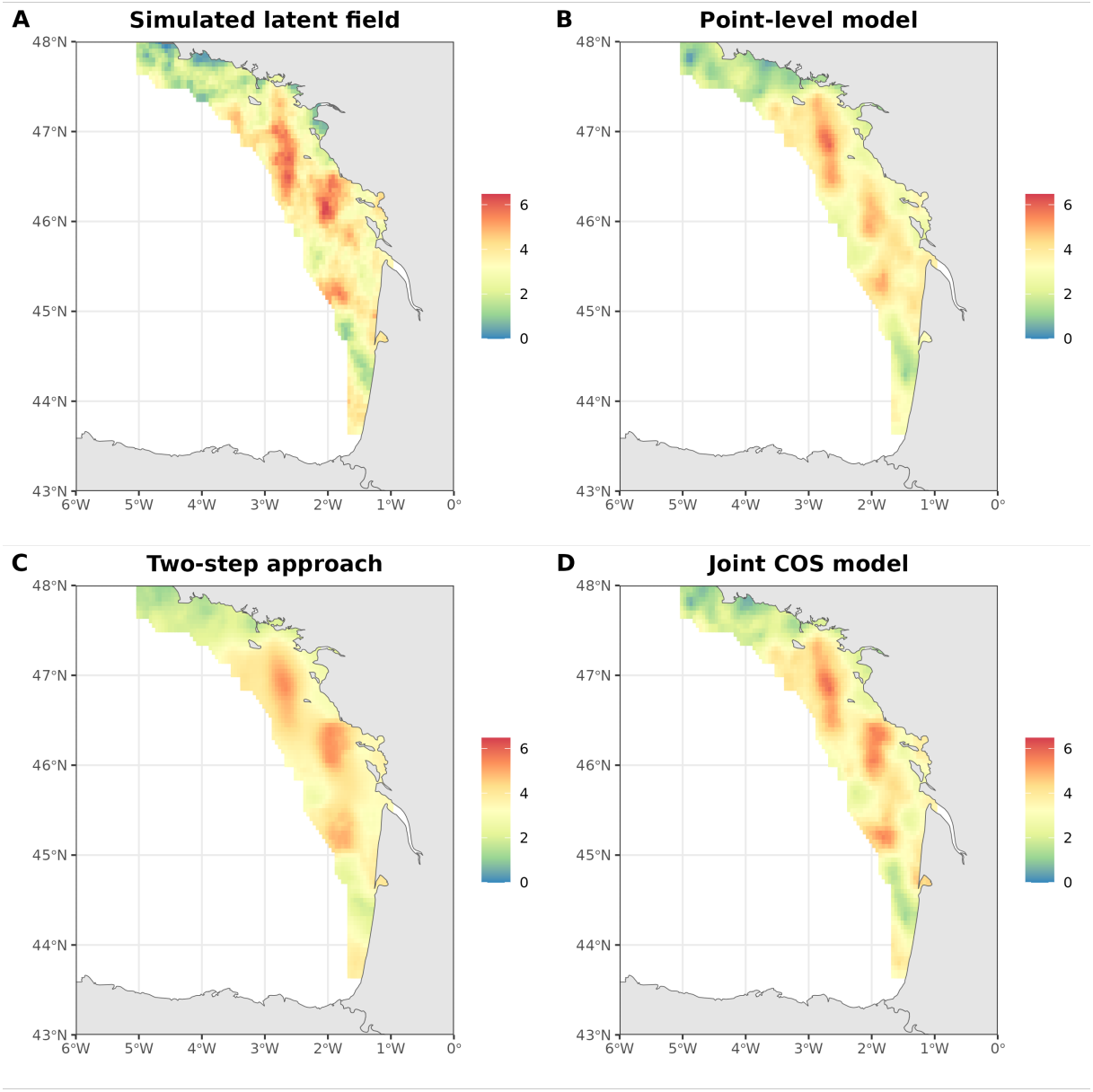

Fig. 4. (Simulations) Distribution of simulated/estimated biomass field in the log scale for alternative model configurations.

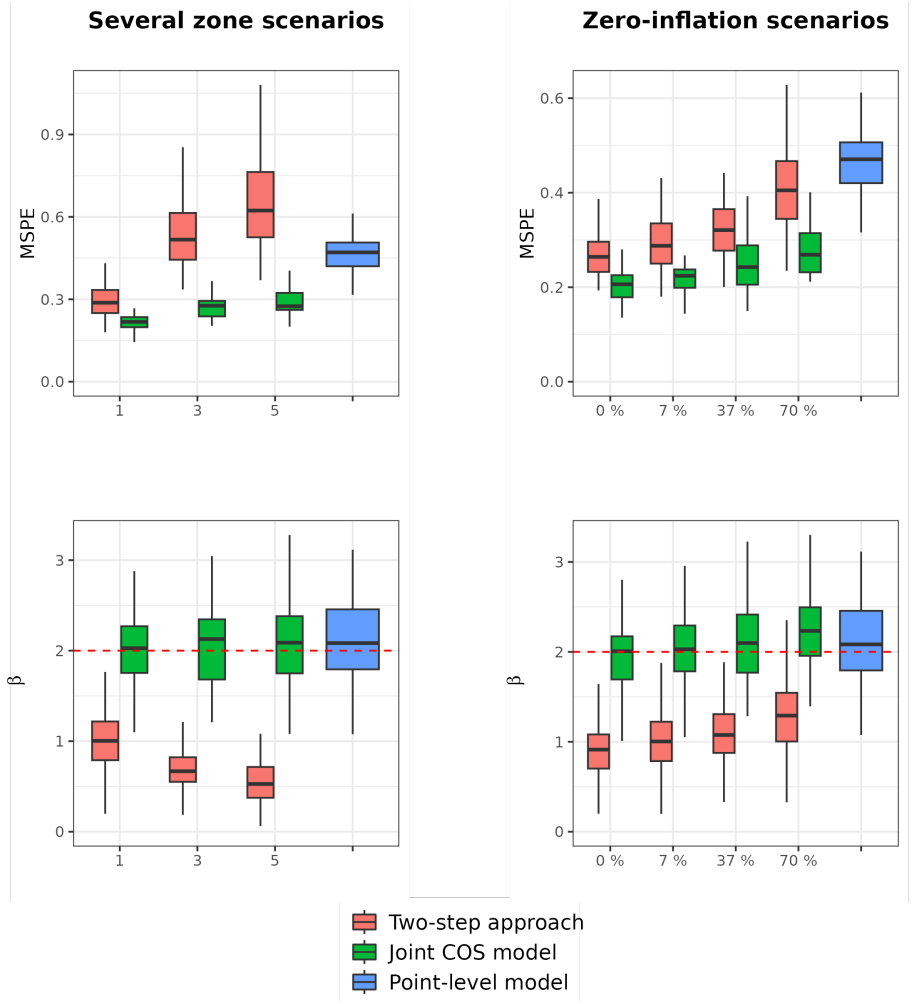

Fig. 5. (Simulations) Boxplots of the performance metrics of alternative scenarios for the number of fishing zones within a single declaration (left) and the level of zero-inflation of the data (right). x-axis, left: number of fishing zones for a single declaration. x-axis, right: amount of zero in the data (at the point-level $Y(x_i)$). Blue boxplots: model fitted to point-level data only. Red boxplots: integrated model fitted to reallocated observations. Green boxplots: integrated model accounting for change of support. MSPE: mean squared prediction error. β: species-habitat parameter. Red line: true value for the species-habitat parameter.

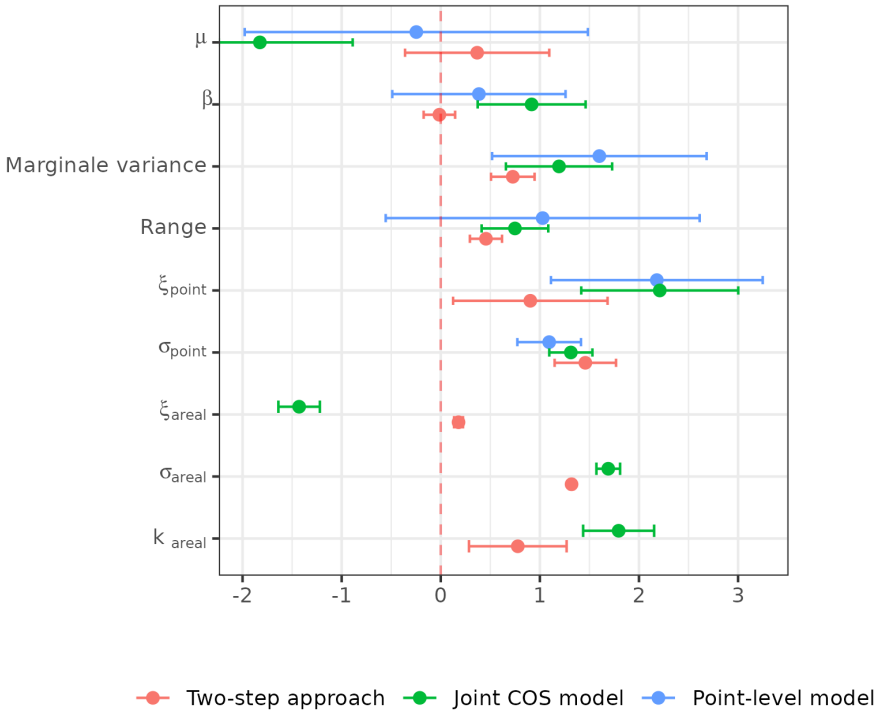

Fig. 6. (Real data) Parameters estimated by alternative models. 'areal' refers to the areal-level data. 'point' refers to the point-level data.

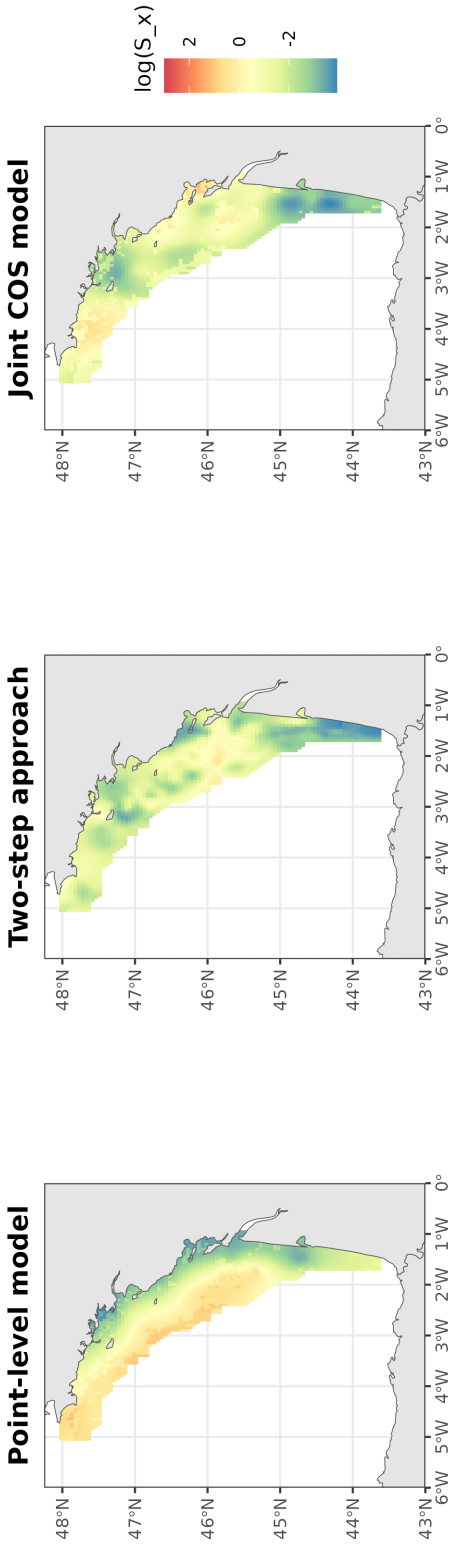

Fig. 7. (Real data) Maps of common sole distribution in the log-scale predicted by the different models.

Supplementary Material

S-I. Reparameterization of the Lognormal distribution

The Lognormal distribution is usually written as $Z \sim \mathrm{L}(\rho; \sigma^2)$ where ρ is the mean component in the log scale and σ is the variance component. Also, $Z = e^{\rho+\sigma N}$ and $N \sim \mathcal{N}(0,1)$. In this case, $\mathbb{E}(Z) = e^{\rho+\frac{\sigma^2}{2}}$ and $\mathbb{V}ar(Z) = (e^{\sigma^2}-1)e^{2\rho+\sigma^2}$.

We choose to reparameterize the Lognormal distribution so that $\rho = ln(\mu) - \frac{\sigma^2}{2}$. Then:

- $Z = \mu e^{\sigma N - \frac{\sigma^2}{2}}$
- $\mathbb{E}(Z) = \mu$
- $\mathbb{V}ar(Z) = \mu^2(e^{\sigma^2}-1) \Leftrightarrow \sigma^2 = ln(\frac{\mathbb{V}ar(Z)}{\mathbb{E}(Z)^2}+1)$

S-II. Mixture probability of the individual observation layer

We have to express the probability distribution of D_a and its moments as a function of $Y(x_i)$ and its related moments. Let's assume $Y(x) = C(x) \cdot Z(x)$ is a zero-inflated Lognormal distribution with $C(x)$ and $Z(x)$ the two components of the mixture. $C(x)$ is a binary random variable and $Z(x)$ a Lognormal random variable.

$$C(x)|S(x), x \sim \mathcal{B}(1-p(x))$$

with $p(x) = exp(-e^{\xi} \cdot S(x))$ the probability to obtain a zero value.

$$Z(x)|S(x), x \sim \mathrm{L}(\frac{S(x)}{1-p(x)}, \sigma^2)$$

S-III. Probability of obtaining a zero declaration

As mentioned in the core text, the probability to obtain a zero declaration is the probability that all individual observations within this declaration are null. This gives:

$$\begin{aligned}\mathbb{P}(D_a = 0) &= \prod_{i|x_i \in \mathcal{R}_a} \mathbb{P}(Y_i = 0|S(x_i), x_i), \\ &= \exp\left\{ - \sum_{i|x_i \in \mathcal{R}_a} e^{\xi}.S(x_i) \right\} = \pi_a.\end{aligned}$$

S-IV. Expectation of a positive declaration

Conditionally on $\boldsymbol{S}$ and $\mathcal{P}_j$.

$$\begin{aligned}\mathbb{E}(D_a|D_a > 0) &= \mathbb{E}(D_a 1_{\{D_a>0\}})/\mathbb{P}\left(D_a > 0\right), \\ &= \mathbb{E}(D_a 1_{\{D_a>0\}})/\left(1 - \pi_a\right).\end{aligned}$$

As $\mathbb{E}(D_a 1_{\{D_a>0\}}) = \mathbb{E}(D_a)$, we can write $\mathbb{E}(D_a|D_a > 0)$ as:

$$\begin{aligned}\mathbb{E}(D_a|D_a > 0) &= (1 - \pi_a)^{-1} \mathbb{E}(D_a), \\ &= (1 - \pi_a)^{-1} \sum_{i|x_i \in \mathcal{R}_a} \mathbb{E}(C_i Z_i), \\ &= (1 - \pi_a)^{-1} \sum_{i|x_i \in \mathcal{R}_a} (1 - p_i) \frac{S(x_i)}{1 - p_i}, \\ &= (1 - \pi_a)^{-1} \sum_{i|x_i \in \mathcal{R}_a} S(x_i).\end{aligned}$$

S-V. Variance of a positive declaration

The variance then can be expressed as:

$$\mathbb{V}ar(D_a|D_a > 0) = \mathbb{E}(D_a^2|D_a > 0) - \mathbb{E}(D_a|D_a > 0)^2.$$

with,

$$\begin{aligned}\mathbb{E}(D_a^2|D_a > 0) &= (1-\pi_a)^{-1}\mathbb{E}(D_a^2 1_{\{D_a>0\}}) \\ &= (1-\pi_a)^{-1}\mathbb{E}(D_a^2)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(D_a|D_a > 0)^2 &= ((1-\pi_a)^{-1}\mathbb{E}(D_a 1_{\{D_a>0\}}))^2 \\ &= (1-\pi_a)^{-2}\mathbb{E}(D_a)^2\end{aligned}$$

Then, using these two expressions in the variance formula gives:

$$\begin{aligned}\mathbb{V}ar(D_a|D_a > 0) &= (1-\pi_a)^{-1}\mathbb{E}(D_a^2) - (1-\pi_a)^{-2}\mathbb{E}(D_a)^2 \\ &= (1-\pi_a)^{-1}(\mathbb{V}ar(D_a) + \mathbb{E}(D_a)^2) - (1-\pi_a)^{-2}\mathbb{E}(D_a)^2. \\ &= (1-\pi_a)^{-1}\mathbb{V}ar(D_a) - \frac{\pi_a}{(1-\pi_a)^2}\mathbb{E}(D_a)^2.\end{aligned}$$

As the $(Y_i)_{i|x_i\in\mathcal{R}_a}$ are independent, $\mathbb{V}ar(D_a) = \sum_{i|x_i\in\mathcal{R}_a}\mathbb{V}ar(Y(x_i)) = \sum_{i|x_i\in\mathcal{R}_a}\mathbb{V}ar(C(x_i)Z(x_i))$.

Obtaining $\mathbb{V}ar(C(x_i)Z(x_i))$ is then straightforward due to conditional independence properties:

$$
\begin{aligned}
\mathbb{V}ar(C(x_i)Z(x_i)) &= \mathbb{E}(C(x_i)^2 Z(x_i)^2) - \mathbb{E}((x_i)Z(x_i))^2, \\
&= \mathbb{E}(C(x_i)^2)\mathbb{E}(Z(x_i)^2) - \mathbb{E}(C(x_i))^2\mathbb{E}(Z(x_i))^2, \\
&= (1-p(x_i))\mathbb{E}(Z(x_i)^2) - (1-p(x_i))^2\mathbb{E}(Z(x_i))^2, \\
&= (1-p(x_i))(\mathbb{V}ar(Z(x_i)) + \mathbb{E}(Z(x_i))^2) - (1-p(x_i))^2\mathbb{E}(Z(x_i))^2, \\
&= \frac{S(x_i)^2}{1-p(x_i)}(e^{\sigma^2}-1) + \frac{S(x_i)^2}{1-p(x_i)} - S(x_i)^2, \\
&= \frac{S(x_i)^2}{1-p(x_i)}(e^{\sigma^2} - (1-p(x_i)))
\end{aligned}
$$

S-VI. Sum up of the main formulas

The main formulas can be summarised as follows:

n.b. all the formulas are conditionned on $\boldsymbol{S}$ and on the fishing positions (x_i or $\mathcal{P}_j$).

- The probability to obtain a zero areal-level data

$$\mathbb{P}(D_a = 0) = \exp\left\{ - \sum_{i|x_i \in \mathcal{R}_a} e^{\xi} S(x_i) \right\} = \pi_a$$

- The expectancy of a positive declaration

$$\mathbb{E}(D_a | D_a > 0) = \frac{\sum_{i|x_i \in \mathcal{R}_a} S(x_i)}{1 - \pi_a}$$

- The variance of a positive declaration

$$\mathbb{V}ar(D_a | D_a > 0) = \frac{\sum_{i|x_i \in \mathcal{R}_a} \mathbb{V}ar(Y(x_i))}{1 - \pi_a} - \frac{\pi_a}{(1 - \pi_a)^2} \mathbb{E}(D_a)^2$$

- The variance of an individual observation

$$\mathbb{V}ar(Y(x_i)) = \frac{S(x_i)^2}{1 - p_j} (e^{\sigma^2} - (1 - p(x_i)))$$

Then, assuming $D_a | D_a > 0$ also follows a Lognormal distribution we can write:

$$D_a | D_a > 0 \sim \mathrm{L}(\mu_a = \mathbb{E}(D_a | D_a > 0), \sigma_a^2 = ln(\frac{\mathbb{V}ar(D_a | D_a > 0)}{\mathbb{E}(D_a | D_a > 0)^2} + 1))$$

S-VII. Simulation domain and data

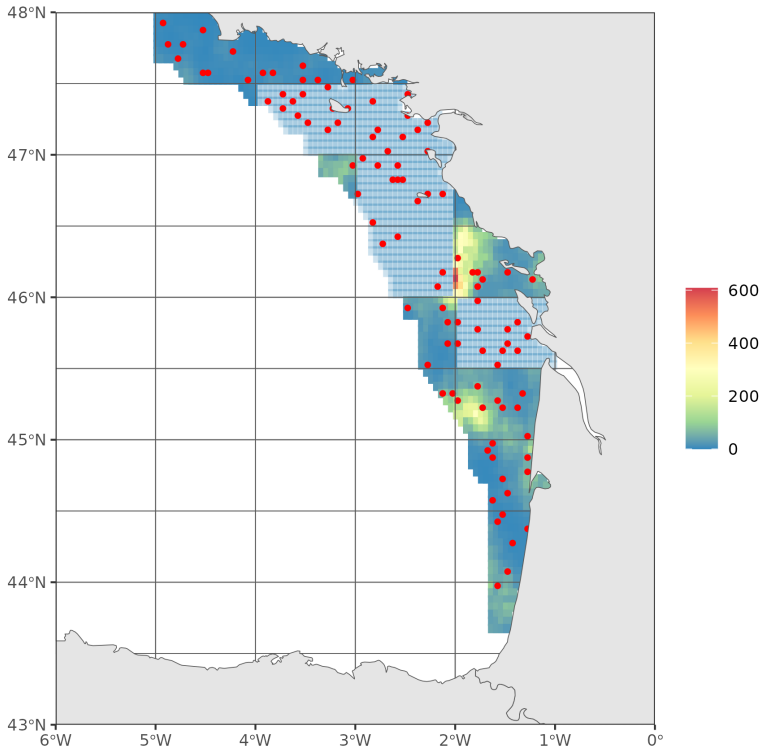

Fig. S1. Simulated biomass field with point-level samples (scientific data are the red dotes) and statistical rectangles. The rectangles that have not been sampled by areal data (commercial declaration) are the transparent rectangles. They represent 1/3 of the full area.

S-VIII. Simulation of several fishing zones

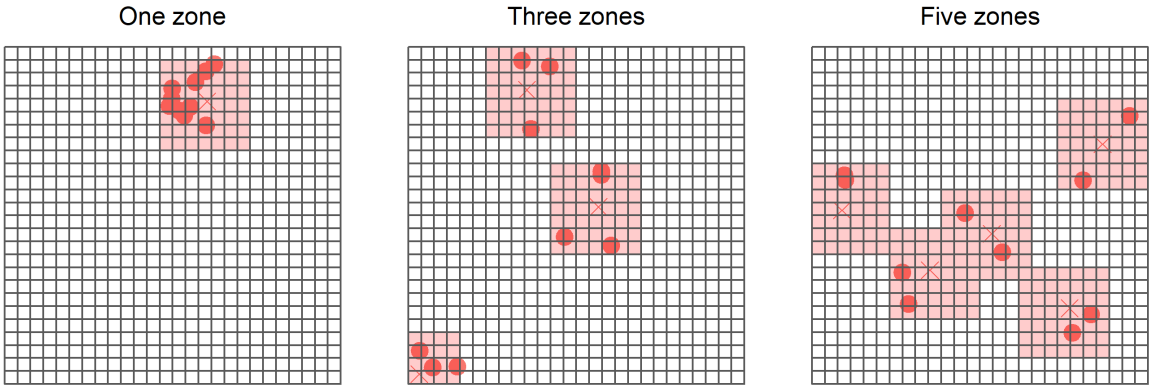

Fig. S2. Simulations of 10 fishing locations within 1, 3 and 5 fishing zones. The full grid corresponds to a statistical rectangle. Crosses are the centroid of the fishing zones. A declaration declared at the resolution of the statistical rectangle would be uniformly reallocated over these fishing locations.

S-IX. Additional parameters of the simulations

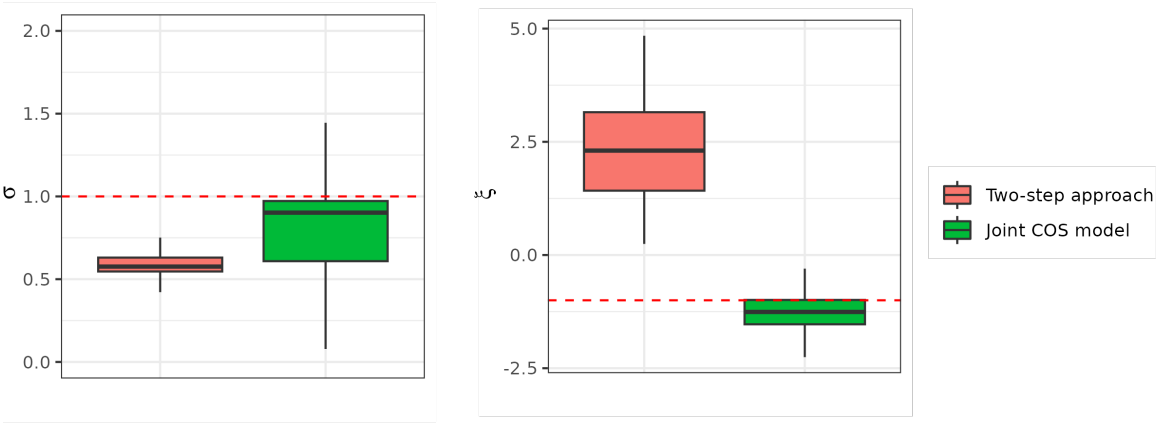

Fig. S3. (Simulations) Boxplot of the variance (σ) and zero-inflation parameter (ξ) of the observation model for the areal-level data. Red boxplots: integrated model fitted to reallocated data. Green boxplots: integrated model accounting for change of support, joint COS model. Red line: true value for the parameters.