ICES Journal of Marine Science
May 2022, Volume 79, Issue 4, Pages 1133-1149
https://doi.org/10.1093/icesjms/fsac032
https://archimer.ifremer.fr/doc/00754/86604/

Combining scientific survey and commercial catch data to map fish distribution

Alglave Baptiste [1, 2, *], Rivot Etienne [2], Etienne Marie-Pierre [3], Woillez Mathieu [4], Thorson James T [5], Vermard Youen [1]

[1] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Nantes 44980, France
[2] DECOD (Ecosystem Dynamics and Sustainability), Institut Agro, IFREMER, INRAE, Rennes 35042, France
[3] Mathematical Research Institute of Rennes IRMAR, Rennes University, Rennes 35042, France
[4] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Brest 29280, France
[5] Habitat and Ecological Processes Research Program, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98112, USA

* Corresponding author : Baptiste Alglave, email address : baptiste.alglave@agrocampus-ouest.fr

Abstract :

Developing Species Distribution Models (SDM) for marine exploited species is a major challenge in fisheries ecology. Classical modelling approaches typically rely on fish research survey data. They benefit from a standardized sampling design and a controlled catchability, but they usually occur once or twice a year and they may sample a relatively small number of spatial locations. Spatial monitoring of commercial data (based on logbooks crossed with Vessel Monitoring Systems) can provide an additional extensive data source to inform fish spatial distribution. We propose a spatial hierarchical framework integrating both data sources while accounting for preferential sampling (PS) of commercial data. From simulations, we demonstrate that PS should be accounted for in estimation when PS is actually strong. When commercial data far exceed scientific data, the later bring little information to spatial predictions in the areas sampled by commercial data, but bring information in areas with low fishing intensity and provide a validation dataset to assess the integrated model consistency. We applied the framework to three demersal species (hake, sole, and squids) in the Bay of Biscay that emphasize contrasted PS intensity and we demonstrate that the framework can account for several fleets with varying catchabilities and PS behaviours.

Keywords : hierarchical model, integrated modelling, species distribution model, survey data, Template Model Builder (TMB), VMS and logbook data

1 INTRODUCTION

Developing species distribution models (SDM) is critical in marine and fisheries ecology for assessing the relationship between species and their habitat (Guisan and Zimmermann, 2000), identifying essential habitats (Paradinas *et al.*, 2015), forecasting population and ecosystems response to environmental changes (Cheung *et al.*, 2009). The development of statistical models to predict fishery resources distribution has received considerable attention (Planque *et al.*, 2011; Thorson *et al.*, 2015a, 2015b; Martínez-Minaya *et al.*, 2018; Moriarty *et al.*, 2020). Recent developments have generalized SDM to analyze biological data representing condition, stomach contents, size structure, and other demography and population dynamics features (Thorson, 2015; Grüss *et al.*, 2020). Ongoing research also seek to integrate individual movement, growth, species interactions into SDM (Kristensen *et al.*, 2014; Thorson *et al.*, 2017a, 2019), although these approaches are “data hungry” and therefore require integrating different sources of data within a single model.

Scientific survey and commercial catch data consist in two potentially complementary data sources to estimate harvested fish spatial distribution (Pennino *et al.*, 2016). Scientific surveys are key data sources in fisheries ecology. They most often benefit from a standardized sampling plan and a constant catchability (Hilborn and Walters, 1992; Ocean Studies Board and National Research Council, 2000; ICES, 2005; Nielsen, 2015). They are generally designed to cover the full geographical extent of specific populations including areas of low or null abundance, and are thus adapted to develop unbiased abundance indices and spatial predictions of species distribution (Rivoirard *et al.*, 2008; ICES, 2012). In addition, they often seek to minimize selectivity in order to sample as many species, size groups and life stages as possible. However, the related expansive charges generally comes at the cost of a relatively low sampling density in space and/or time. For instance, trawl survey can sample a limited number of spatial locations, and most often occur once or twice a year. Thus, they may provide poor information regarding intra-annual variability (Pennino *et al.*, 2016; Rufener *et al.*, 2021) and imprecise estimates of species abundance and distribution (ICES, 2005).

Commercial catch declarations (logbooks) data constitute a complementary data source that may benefit of a higher sampling effort than scientific survey. In Europe, catch declarations must be reported in logbooks data for all fishing vessels; besides, geolocation through Vessel Monitoring System (VMS) is mandatory for all fishing boat above 12m long (Hintzen, 2021). Hence, logbook data combined with VMS data can provide high resolution maps of Catch Per Unit Effort (CPUE - Gerritsen and Lordan, 2010; Murray *et al.*, 2013) with a relatively dense spatio-temporal sampling within the range of the commercial fleets. However, inferring SDM with commercial data can be challenging as they generally arise from a preferential sampling (PS) behavior, i.e. a sampling that directly or indirectly depends upon the biomass of the target species. Indeed, fishermen tend to target areas with high biomass or may also favor fishing zones based on other criteria (like bottom substrate or distance to the coast for instance - Hintzen *et al.*, 2021) that are indirectly related to the target species abundance. When not properly considered in statistical models, PS associated with commercial data may lead to biased estimates of fish distribution and biomass (Trenkel *et al.*, 2013; Pennino *et al.*, 2019). In particular, when the biomass is spatially heterogeneous, ignoring PS may overestimate the spatial predictions and the overall biomass estimates.

Recent research has tackled this challenge and developed methods to account for PS in statistical inferences. Model based PS was first introduced by Diggle *et al.* (2010) who proposed a base framework for estimating PS and applied it to led pollution data in Galicia. The authors extended a standard geostatistical approach within a hierarchical framework where the variable of interest is jointly modelled with the spatial intensity of the sampling effort which also contributes to the inference and accounts for PS towards the variable of interest. This approach was extended by Pati *et al.* (2011) who introduced covariates and random effects in the model. Conn *et al.* (2017) followed the same ideas and developed a more generic model for ecological applications, which they applied to aerial seal count data. Pennino *et al.* (2019) applied similar ideas to infer the distribution of shrimps from onboard fishery data.

Provided PS is accounted for, integrated models (IM) appear as an attractive tool to combine fishery-independent and fishery-dependent data to infer harvested fish spatial distribution. IM have received considerable attention in the ecological literature (Schaub and Abadi, 2011; Parent and Rivot, 2012; Gimenez *et al.*, 2014). By sharing the information between different data types, IM may provide more accurate estimates and predictions compared with separate analysis of different data types. Recently, Rufener *et al.* (2021) demonstrated the potential of IM to integrate scientific data and onboard observer count data to improve SDM of fishery resources. However, although onboard observer data provide useful complementary information to scientific survey, they generally only represent a small proportion of all sea trips (1% in average for the French observer programs - Cornou *et al.*, 2021). By contrast, the combination of commercial catch declarations in logbooks with VMS data provides a more extensive data source to map fish spatial distribution. Furthermore, the potential of embedding PS within a hierarchical SDM to integrate catch declaration data and scientific survey is still an open challenge and new methodology are required to handle PS behaviors of commercial fleets while accounting for all the complexity related to fishing locational choice (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017).

In this paper, we develop an IM model to infer fish spatial distribution by combining both scientific and commercial catch declaration data while taking into account the PS induced by fishing targeting behavior.

To assess the challenges, the benefits and also the limits of the approach, we evaluate the performance of our IM based on simulated data. Simulations are primarily designed to assess the respective contribution of each data source to inference for different model configurations. We first evaluate how the balance between the commercial and scientific sample sizes affect the model outputs. Because the commercial data may often only partially cover the distribution area of a targeted species, we assess how this issue may affect the quality of estimation and how scientific data may contribute to reduce the effect of this gap in the commercial data. Introducing PS within an IM framework involves adding new parameters, complexifying the model structure and then increasing the computational cost. We therefore assess how perform a more parsimonious model that would ignore PS. Last, in addition to the PS, the fishing locations can be controlled by other factors independent from the species distribution (e.g. logistical constraints, management regulations – see Girardin *et al.*, 2017; Ducharme-Barth *et al.*, 2022). We therefore assess how such process blurring strict PS may affect the quality of inferences.

We demonstrate the flexibility of the approach by fitting the model to three different important European demersal fishery resources in the Bay of Biscay: common sole (*Solea solea*, Linnaeus, 1758), hake (*Merluccius merluccius*, Linnaeus, 1758) and squids (*Loliginidae* family). With these contrasted examples, we illustrate the capacity of the framework to handle multiple commercial fleets with potentially distinct PS intensities and different fishing behaviors.

2 MATERIAL AND METHODS

2.1 Spatial integrated model

Below we provide the core elements of the modelling approach. Additional details are provided in supplementary material (SM 1). The model is structured in four layers: observations (here commercial and scientific CPUE in weight per unit of effort), sampling process, latent field (here fish biomass relative density) and parameters (Figure 1 - all notations are available in SM 1.1, Table S1). Sampling process is usually ignored in hierarchical models as it is mostly considered independent of the quantity of interest, and then has no consequence on the estimation procedure (Diggle *et al.*, 2010). Here, the spatial distribution of commercial fishing is explicitly modelled as a inhomogenous Poisson point process whose intensity may depend on the biomass field and contributes to the likelihood. The observation processes of scientific and commercial data are conditional upon the biomass latent field and the sampled locations.

All processes are considered to occur in a discrete fine grid (see for instance SM 2.1, Figure S2.1 or SM 3.1, Figure S3.1). We assume the density of the point process is piecewise constant in each cell grid which brings simplification in the expression of the likelihood of the point process (Diggle, 2013 - see SM 1.2). The time component is omitted and both commercial and scientific data are assumed to occur at the same time step.

The IM is designed to assimilate the scientific data of several surveys and/or the commercial data of several fleets. In the following, the subscript j refers to the different data sources either scientific or commercial. For instance, in a model with one scientific survey and two commercial fleets, j will take the values $j = 1,2,3$, with $j = 1$ for the scientific data and $j = 2,3$ for the two commercial fleets.

2.1.1 Latent field of relative biomass

The fish biomass relative density S (eq. (1) – (2)) is modeled through a latent log Gaussian spatial field defined on the same discrete spatial domain as the point process. The mean of the Gaussian field depends on environmental covariates through a log link where the linear predictor combines an intercept α_S, the linear effect of environmental covariates $\Gamma_S(x)$ (effects captured by the corresponding fixed parameters β_S representing the species-habitat relationship). The remaining spatial variation is accounted for through a zero-mean Gaussian random field (GRF) denoted $\delta(x)$ parameterized with a Matérn correlation function $M(x,x';\kappa,\phi)$, characterized by the shape κ and the scale ϕ (Cressie, 1993; Gelfand *et al.*, 2010; Lindgren *et al.*, 2011 and Banerjee *et al.*, (2014)). The shape can be expressed in term of range $\rho = \frac{\sqrt{8}}{\kappa}$ where ρ is the distance for which the correlation between points is near 0.1.

$$\log\big(S(x)\big) = \alpha_S + \Gamma_S(x)^T \cdot \beta_S + \delta(x) \qquad (1)$$

$$\delta(x) \sim GRF(0, M(x,x';\kappa,\phi)) \qquad (2)$$

2.1.2 Sampling process

Recent literature has emphasized the complexity of the targeting behavior processes (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Abbott *et al.*, 2015; Girardin *et al.*, 2017; Hintzen, 2021). In this paper, we did not attempt to model explicitly all those processes (e.g. resource distribution, logistical constraints, tradition, management regulations) and opted for a simplified representation where the spatial targeting directly depends on the biomass field S and on an additional spatially structured random term.

Let us denote $X_{com\,j}$ the spatial point process where commercial vessels of fleet j are identified as fishing. In the following, all vessels in the same commercial fleet are assumed to have homogeneous behaviors. Following Diggle *et al.* (2010), the set of fishing locations are modeled conditionally on S, as a inhomogeneous Poisson point process with piecewise constant intensity $\lambda_j(x)$ (eq. (3) - (4)).

$$X_{com\ j} \sim \mathcal{IPP}(\lambda_j(x)) \qquad (3)$$

$$\log\left(\lambda_j(x)\right) = \alpha_{X\ j} + b_j \cdot \log(S(x)) + \eta_j(x) \qquad (4)$$

For any fleet j, intensity $\lambda_j(.)$ of the Poisson point process is modeled as a log-linear combination of the intercept $\alpha_{X\ j}$, the logarithm of the relative biomass $S(.)$ scaled by a parameter b_j, and a residual spatial effect $\eta_j(.)$ with the same structure as $\delta(.)$ but with specific parameters κ and ϕ. All parameters $\alpha_{X\ j}$, b_j and the spatial random effect $\eta_j(x)$ are specific to each fleet.

The parameter b_j quantifies the strength of PS by scaling the relationship between the local value of the resource field and the local fishing intensity.

Fishing locations potentially depend on many other factors than fish distribution such as distance to harbor, logistical constraints, management regulations - spatial closures, quotas – or fishing habits/tradition (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017). The spatial random effect $\eta_j(.)$ is needed to capture any remaining additional effect not captured by the dependence to $S(.)$.

In that sense, a zero value for b_j indicates that the choice of the sampling locations does not depend on the fish biomass relative density but only on the spatial random effect.

In addition to b_j, a dimensionless spatial metric was developed to quantify the strength of PS (SM 1.3).

2.1.3 Observation process

Both scientific and commercial observations are considered as proportional to the underlying biomass through a zero-inflated observation process. In our applications, observations are expressed as CPUE (in weights / unit effort), with high proportion of zeros (zeros represent on average 30% of the commercial data and 10 to 50% of scientific data).

Observations are modelled through a zero-inflated lognormal model conditionally on biomass $S(x)$ in cell x (eq. (5-6)). The model is derived from Thorson *et al.* (2016) or Thorson (2018). We assume that the expected catch $\mu_j(x)$ for any fleet/data source j in the cell x depends on the latent field value $S(x)$ and a catchability coefficient q_j (eq. (5)). A zero catch ($y = 0$) is modeled as a Bernoulli random variable with parameter $exp(-e^{\xi_j} \cdot \mu_j(x))$, where ξ_j is the parameter controlling the intensity of zeros relatively to the expected catch (eq. (6)). Then, $\mu_j(x)$ being fixed, the higher (resp., the lower) ξ_j, the lower (resp. the higher) the probability of obtaining a zero-catch.

The distribution of a positive catch $y > 0$ at a given x is defined as the combination of the probability of obtaining a non-zero catch $(1 - exp(-e^{\xi_j} \cdot \mu_j(x)))$ times a positive continuous distribution L (here a lognormal distribution) with expected value $\frac{\mu_j(x)}{(1-exp(-e^{\xi_j} \cdot \mu_j(x)))}$ and standard deviation σ_j. This formulation allows to represent the zero catch while assuring that the expected catch still equals $\mu_j(x)$.

$$\mu_j(x) = q_j \cdot S(x) \qquad (5)$$

$$\mathrm{P}(Y = \mathrm{y}|x, S(x)) =$$

$$\begin{cases} \exp\left(-e^{\xi_j} \cdot \mu_j(x)\right) & \text{if } y = 0 \\ \left(1 - \exp\left(-e^{\xi_j} \cdot \mu_j(x)\right)\right) \cdot L\left(y, \frac{\mu_j(x)}{\left(1-\exp\left(-e^{\xi_j} \cdot \mu_j(x)\right)\right)}, \sigma_j^2\right) & \text{if } y > 0 \end{cases} \qquad (6)$$

Per se, catchability q_j are not identifiable as there is no information in the model to estimate the absolute scale of S. Commercial catches and/or scientific surveys will be only informative about fish biomass relative density and additional information must be provided to ensure statistical identifiability. If only one data type feeds the model (only scientific or commercial data), relative catchability is fixed to 1 and the spatial random field values is in the same scale as the data. If two data types (or more) are used to feed the model, one of the relative catchability (denoted q_{ref}) has to be fixed, the other ones being estimated relatively to the first one through a scaling factor k_j (eq. (7)).

$$q_j = k_j * q_{ref} \qquad (7)$$

As it is illustrated further in the simulation-estimation study (see section 3.1.1), the choice of the reference level can have important consequences on the precision of estimation.

2.1.4 Maximum likelihood estimation

The estimation of the model is performed with TMB (Template Model Builder - Kristensen *et al.* (2016)) and the spatial random effects are estimated through the SPDE approach (Lindgren *et al.*, 2011) within the R software (R Core Team, 2020). More details on estimation are available in the supplementary material (SM 1.4).

2.1.5 Integrated model validation

A key issue with IM is whether the different data sources provide consistent or conflicting information (Saunders *et al.*, 2019; Zipkin *et al.*, 2019; Peterson *et al.*, 2021). In our framework, the key question is whether integrating commercial data in addition to scientific data will complement or will disrupt the inferences obtained from the scientific data, considered as a reference source of information. To address this issue, we propose a validation procedure based on the consistency check initially developed by Rufener *et al.* (2021) and designed to check whether estimates obtained from the IM are consistent with those obtained from the model fitted to scientific data only. The procedure would reject consistency if the parameters estimates from the IM fall outside the 95% confidence region of parameters estimates from scientific data only (see SM 1.5 for more details on the procedure). This validation step is applied to both simulations and case studies.

2.2 Simulation-estimation experiments

We conducted simulation-estimation experiments to assess the performance of the method for different data/model configurations (Table 1, see also SM 2 for extended details on simulations). For all scenarios, simulations of data, covariates and GRF were parameterized to tailor the case studies described hereafter. All scenarios and configurations are repeated 100 times so as to capture the variability between replicates.

Simulation-estimation experiments were specifically designed to address four questions detailed below. In all cases, commercial data were simulated with various levels of PS ($b = 0$ for uniform sampling, $b = 1$ for moderate PS, $b = 3$ for strong PS) to assess the effect of PS on model's performance (Figure 2).

*(**Q1**) How does each data source contribute to inferences?*

In real case study, commercial data sample size may be far superior to scientific data (specifically when using landings data) which might result in commercial data that dominate inferences. To assess how the balance between the scientific and commercial sample sizes drives the relative contribution of each data source, simulations were conducted with few scientific samples (50 each) with increasing commercial samples (50=small, 400=medium and 3000=large), and with a large commercial sample size (3000) with increasing scientific sample size (50=small, 400=medium, 3000=large). No scenario with more scientific samples than commercial samples is presented here as it is a very unlikely configuration when using logbook catch data.

For each combination of commercial and scientific sample size, we fitted four different models: a model fitted to scientific data only, a model fitted to commercial data only, and two IM fitted to both commercial and scientific data, one with the scientific data used as reference level and another one using the commercial data as reference level (Cf. eq. (7)).

For questions Q2, Q3 and Q4, all simulations were conducted using $n_{scientific} = 50$ and $n_{commercial} = 3000$ to tailor the case studies. Commercial data are used as the reference for catchability in the IM.

*(**Q2**) How does a partial coverage of the study area by the commercial data affect the quality of the estimation?*

While scientific surveys are supposed to cover the full population distribution area, partial coverage of the area by commercial fishing boats may arise from different sources like spatial management closures (e.g. box closure) or too expensive travels from the coast. To assess how a partial coverage by commercial data can affect estimates, we simulated data with the commercial sampling intensity arbitrarily fixed to 0 in a fixed 9x9 box (15% of the domain) while some biomass and some scientific samples are still simulated in this area. We compared estimates of the biomass in the entire area with those obtained with commercial data available on the whole domain.

*(**Q3**) What is the cost of ignoring PS in estimation when sampling is preferential?*

Modelling preferential sampling involves conditioning results upon a specified structural assumption about sampling as well as increased computational cost. Here, we assess how ignoring PS would affect the quality of inferences when sampling is actually preferential. We voluntary introduce misspecification between the model used for simulating the data (with various levels of PS intensity) and the one used in the estimation procedure (b is alternatively estimated or arbitrarily fixed at 0).

*(**Q4**) How does the estimation perform when additional processes other than PS drive the fishing locations?*

Fishing locations potentially depend on many other factors independent from the species distribution (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017). To assess how such process blurring strict PS may affect the quality of inferences, we simulate data with a sampling intensity that depends on both the biomass distribution (PS) and an additional spatial random terms $\eta_f(.)$ independent from the biomass distribution (eq. (4); see Table 1 for more details on $\eta_f(.)$ parameterization), and compare the inferences obtained from a data set simulated with strict PS ($\eta_f(.) = 0$ on the full domain).

Note that for questions Q1, Q2 and Q3, the random effect η was fixed to 0 in simulations (but it is still estimated in the estimation model), so that the sampling process only depends on the distribution of biomass.

2.2.1 Performance metrics

The performance of the estimation method was assessed using different metrics on key model outputs such as the total biomass, the PS parameter b and the spatial biomass predictions.

The quality of the total biomass estimation (the sum over all grid cells, $B = \sum_x S\,(x)$) was explored through the relative bias $\frac{(B-\hat{B})}{B}$, that quantifies how much the total biomass is over or under-estimated.

The quality of the estimation of the parameter b is assessed through the relative bias defined as $\frac{b-\hat{b}}{b}$ (except for $b = 0$, where only the absolute bias is considered). We also assessed the relative bias of the species-habitat relationship estimate $\hat{\beta}_S$ and range parameter ρ as these parameters are meaningful for understanding species distribution.

The precision of the spatial predictions was studied with the mean squared prediction error between the simulated and the estimated latent field values $\frac{1}{n}\sum_x\left(S(x) - \widehat{S(x)}\right)^2$ (MSPE – n stands for the number of grid cells).

2.3 Case studies

We applied the approach on three case studies of demersal fisheries in the Bay of Biscay: the common sole (*Solea solea*, Linnaeus, 1758), the hake (*Merluccius merluccius*, Linnaeus, 1758) and the squids (Loliginidae family). These case studies were selected because they emphasize different intensities of preferential sampling. Further details on case studies and data are provided in SM 3.

To compare models on the same spatial domain for the three species, we limited the analysis to scientific and commercial data available on the Bay of Biscay only (SM 3.1, Figure S3.1 for the spatial grids). Besides, to get some replicates of the analysis, we applied the approach on 2 years for each case study (2017 and 2018 for common sole – 2014 and 2015 for hake and squid). To keep it synthetic, only the data and the results of the models for hake in 2014, sole in 2017 and squids in 2015 are presented in this manuscript as the related IM pass the consistency check and they emphasize contrasted level of PS.

2.3.1 *Survey data*

Scientific data (CPUE, in kg/hour - Figure 3) were derived from the Orhago survey for common sole and EVHOE survey for hake and squids (ICES, 2020a). The sampling density (number of data points / km^2) of those two surveys revealed representative of the sampling density of the main European trawl surveys from the DATRAS database (see SM 3.2). In comparison, commercial data used in the case studies are denser by 2 orders of magnitude. Scientific data was aligned on commercial data by filtering only individuals above the minimum landing size when available (24 cm for sole, 27 cm for hake - ICES, 2020). The Orhago survey provides 49 samples for 2017 and 2018 and the EVHOE survey provides 86 samples for 2014 and 2015.

2.3.2 *Commercial data*

For each species, we filtered commercial data for ‘bottom trawlers’ as they cover a wide part of the study area (Figure 3) and provide easy to compute and reliable CPUE. Commercial data were standardized by the fishing effort in (kg/hour). For hake and sole, we filtered the métier targeting demersal fish (called OTB_DEF) and for squids, the métier targeting cephalopods (called OTB_CEP).

In comparison with scientific data, the orders of magnitude of commercial sample size is much larger. For hake (i.e. OTB_DEF), there are 6852 commercial samples in 2014 and 5000 in 2015. For squids (i.e. OTB_CEP), there are 7486 commercial samples in 2014 and 9611 in 2015. For sole (i.e. OTB_DEF), there are 2401 samples in 2017 and 3325 in 2018.

2.3.3 Habitat covariates

Two covariates classically used to describe benthic species distribution were selected: depth and sediment type (Le Pape *et al.*, 2003; Witman and Roy, 2009; Rochette *et al.*, 2010). Depth was separated into several categories and was considered (as sediment) as a categorical variable (SM 3.7, 3.8).

2.3.4 Model configurations

As for the simulation-estimation experiments, the models of the case studies were fitted under different configurations. To assess the information brought by each dataset, we compared the model fitted to scientific data only, to commercial data only and to both scientific and commercial data. To assess the effect of PS on model outputs, we compared the IM accounting for PS (b is estimated) with the IM where PS is ignored (b is fixed to 0).

For the sole case study, we compared results obtained from the IM by considering one homogeneous or two distinct fleets with specific catchability and targeting parameters. Note that splitting one fleet in 2 distinct fleets is performed through a PCA coupled with a HCPC analysis on vessels characteristics data derived from both logbooks and VMS data. All the clustering analysis is described in SM 3.9.

2.3.5 Model evaluation

Uncertainty of the predictions are quantified through the coefficient of variation and all estimates (e.g. fixed parameters, total biomass) are represented with related 95 % confidence intervals. We assess the consistency of the IM through the statistical tests described in section 2.1.5 and in SM 1.5. Finally, the different IM are compared through a 5-fold cross validation, and model performance was quantified based on two metrics: the $MSPE_{fit}$ that measures goodness of fit (MSPE – mean squared prediction error), and the PCV that measures predictive capacity (see SM 3.10 for more details on the metrics and guidelines for interpretation). For both metrics, the lower the values, the better the model fits/predicts the data.

3 RESULTS

3.1 Simulations

We summarize the main results of the simulation-estimation experiments below. Additional results are provided in SM 4.

3.1.1 Contribution of each data source in the integrated model

Models fitted on scientific data only provide systematically unbiased estimates of total biomass (the mean bias is close to 0 for all sample size - Figure 4, 1^{st} row), and the variance of estimations decreases with scientific sample size. Note that the species-habitat relationship estimates $\hat{\beta}_S$ are also unbiased (see SM 4.1).

Overall, inferences from the IM revealed consistent with those obtained from scientific data only (SM 4.2.1). Even when the commercial sample size is large and the scientific sample size is small, only 3% of the p-values fall below the 0.05 threshold for the fixed effect test (the test wrongly rejects consistency). For the random effect test, the results are more contrasted as 10% of the p-values fall below the 0.05 threshold when data size are very unbalanced (low scientific sample – high commercial sample).

In almost all configurations, the IM provide unbiased and more precise estimates for total biomass and spatial biomass predictions compared to the model fitted to scientific data only (Figure 4). As expected, the larger the commercial and the scientific sample size, the more accurate the spatial predictions, the PS parameter b and total biomass estimates. Estimates of b are unbiased in most cases except when commercial sample size is small and PS is strong (Figure 4, 2^{nd} row).

As expected, the contribution of each data sources in the IM directly depends on the balance in the sample size. When sample size is balanced between the data sources, then integrating the two data sources in the model systematically improves the inferences with regards to situations where only one data source is analyzed. For instance, for large commercial and scientific sample size (com.L_sci.L) and no PS, the precision is 1.5 higher (i.e. the MSPE is 1.5 lower) for the IM compared to single-data models (either scientific or commercial - Figure 4, 3^{rd} row, 1^{st} column). However, when the sample sizes are unbalanced, the data source with the larger sample size (here commercial data) dominates inference and integrating another data source with a smaller sample size (here scientific data) contributes to a much lesser extent to inference. See for instance the situation where commercial sample size is large and scientific sample size is small (com.L_sci.S - Figure 4, 3^{rd} row, 1^{st} column). In this case, the performances of the model fitted to commercial data alone – with reference level fixed to commercial data - are very close to those of the IM whatever the intensity of PS.

Interestingly, the higher the intensity of PS, the higher the benefits of fitting commercial data in the model (Figure 4, 3^{rd} row); for instance, when both datasets have large sample sizes (com.L_sci.L), increasing PS reduces error predictions (i.e. increases accuracy) by 2 each time (i.e. for $b = 0$, $E(MSPE) = 20$; for $b = 1$, $E(MSPE) = 10$; for $b = 3$, $E(MSPE) = 5$).

Still, the simulations also reveal some limits in the inferences. First, the range parameter might be poorly estimated and slightly biased when the sample size is small while being better estimated when increasing the sample size or integrating additional data in the analysis (see SM 4.3).

Also, in unbalanced cases the accuracy of total biomass estimates from the IM revealed highly sensitive to the choice of the reference level (Figure 4, 1^{st} row). When the commercial sample size far exceeds the scientific sample size, setting the reference level to the commercial data produces more precise estimates than setting the reference level to scientific data. When defining scientific data as reference level, the intercept of the latent field of relative biomass is estimated from the few scientific samples and resulting estimates are less precise than when defining the reference level with a more numerous data source (here commercial data). This is also true - to a lesser extent - for spatial predictions (Figure 4, 3^{rd} row).

In the following, only the case where commercial samples exceed scientific samples and the reference level is fixed with commercial data is explored further as it is the closest to the case studies configuration (Table 1).

3.1.2 Impact of a partial coverage of the study area by the commercial data

When commercial data only partially cover the distribution area, commercial data still provide valuable information to predict biomass spatial distribution whatever the PS intensity is (Figure 5, 2^{nd} column). When sampling is not preferential (data simulated with $b = 0$), a partial coverage of the distribution area produces on average 1.5 less precise spatial predictions but estimates remain unbiased (Figure 5, 3^{rd} row, comparing 1^{st} and 2^{nd} column). When sampling is preferential (either moderate or high), biomass estimates are slightly underestimated. Integrating scientific data in the analysis does not correct this bias.

Finally, all model configurations allow for unbiased and precise estimation of the species-habitat parameters $\hat{\beta}_S$ whether or not there is a partial coverage of the domain (see SM 4.1) and overall almost all IM are consistent with scientific-based model (SM 4.2.2).

3.1.3 How does ignoring PS impact inferences?

As expected, the impact of ignoring PS in the estimation model is negligible when data is simulated with no PS, and becomes more and more detrimental when the intensity of PS increases in the truth (Figure 5, 3^{rd} column). With no surprise, when data are generated with no PS ($b = 0$), ignoring PS in the estimation procedure has no effect on the estimation performance. When PS is moderate, total biomass estimates are 5 % overestimated ($b = 1$). In the case of strong PS ($b = 3$), ignoring PS in the estimation strongly deteriorates the quality of inferences regarding total biomass estimates (Figure 5, 1^{st} row, 3^{rd} column). Total biomass estimates are overestimated by 50% on average. However, the main spatial patterns are well identified with or without consideration of PS, even though more precise when accounting for PS (Figure 5, 3^{rd} row, 1^{st} column). SM 4.4 (Figure S4.4.1) presents maps comparing a simulated biomass field and model predictions obtained by considering or ignoring PS when $b = 3$. The areas with high biomass values (i.e. where commercial sampling is dense) are well predicted by the models accounting for PS or not. The main differences are localized in poorly sampled areas where biomass is low. Accounting for PS in estimation allows to interpret the low sampling intensity areas as low-density areas, and therefore to reduce the bias in those areas (SM 4.4, Figure S4.4.2).

Finally, from a computational point of view, accounting for PS on average multiplies by 4 the computational time (see SM 4.5).

3.1.4 Effect of other spatially structured processes affecting fishing locations

As expected, precision of estimates are deteriorated when fishing locations actually depend upon a combination of biomass distribution (PS) and other mechanisms (here captured by a spatially structured random term - Figure 5, 4^{th} column). In this case, the IM still provides valuable inferences on fish distribution, fish total biomass and estimates of b, although estimations are less accurate than the base case. For instance, MSPE are 5 times lower when nothing else than PS affects sampling locations compared with a case where sampling locations depend on both PS and other independent spatial processes (Figure 5, 3^{rd} row, 1^{st} and 4^{th} column). But interestingly, the weight of scientific data increases when the sampling distribution of commercial data is blurred by spatial processes independent from biomass spatial distribution. MSPE and relative bias provided by the IM are both 1.4 smaller compared to those obtained when the model is fitted to commercial data only.

3.2 Case studies

Below we summarize the main results obtained from the application of the framework to the three case studies. Additional results and maps are provided in SM 5.

3.2.1 Contribution of each dataset to the inferences

Almost all the case studies successfully passed the consistency test between the IM and the model fitted to scientific data only (see SM 5.1).

Models based on scientific data provide different spatial predictions compared with the IM. Predictions for sole and squids from the scientific-based model are mainly shaped by the covariate effects (Figure 6; for further analysis see SM 5.2, SM 5.3 and SM 5.4). On the other hand, predictions from the IM are mainly shaped by the spatial random effect as commercial data allow to better capture the local spatial correlation structures.

Consistently with simulations, inferences from the IM are mainly driven by the commercial data (Figure 6). This logically arise from the much larger sample size of commercial data compared with scientific data, combined with the good coverage of commercial data in high-density areas (Figure 3). As commercial data is denser than scientific data, they will better capture local spatial correlation structures than scientific data. SM 5.5 provide some additional analysis of the information brought by commercial data in the IM.

In this configuration, scientific data bring information to model predictions in areas poorly covered by the commercial data (SM 5.6 - e.g. for squids, the offshore predictions are downscaled by scientific data).

3.2.2 Preferential sampling and other processes affecting fishing locations

In this section and related SM (SM 5.7 to SM 5.10), we focus on results from the IM only.

For the three case studies, estimates of b are positive, suggesting sampling by fishermen is preferential towards high biomass density areas. The hake case study has the lowest PS parameter ($\hat{b} = 0.88,\ sd(\hat{b}) = 0.107$), followed by sole ($\hat{b} = 2.4,\ sd(\hat{b}) = 0.046$) and squids ($\hat{b} = 3.5,\ sd(\hat{b}) = 0.025$). For more intuition concerning the strength of PS and how it varies in space, refer to SM 5.7. In all case studies, the spatial random term η in the sampling process turned out to be spatially structured (SM 5.8) and captures 25% to 97% of the spatial variability of fishing locations (SM 5.9). This highlights the importance of other spatial mechanisms in the choice of fishing locations compared to strict PS towards biomass distribution.

Consistently with simulations, the higher the PS intensity, the higher the differences between inferences obtained with and without considering PS. When comparing biomass field values (Figure 7, left column), ignoring PS increases predictions in poorly sampled areas (all red areas – compare with Figure 3). This effect is particularly marked for the squid case study where the relative difference is the strongest in the offshore areas. However, considering PS or not has relatively little effect in areas where sampling is spatially denser (all white areas). Ignoring PS affects total biomass indices estimates and the relative difference between biomass estimates with or without PS increases with the value of *b* estimates (Figure 7, right column).

When the estimated PS intensity is high (i.e. in the case of squids) accounting for PS can improve model goodness-of-fit and predictive capacity (SM 5.10).

3.2.3 Benefits of considering different fleets in the estimation model

Based on the sole case study, we demonstrate the capacity of the model to integrate multiple commercial fishing fleets, each with specific parameters (catchability and targeting). In the sole case studies, considering two different fleets in the IM (instead of one homogeneous) improves goodness-of-fit towards scientific data (SM 5.11, y-axis) and modifies spatial predictions (SM 5.12).

4 DISCUSSION

Main findings

Combining multiple sources of data to build more informative spatio-temporal models for fish distribution is a major challenge in fishery ecology. Commercial catch per unit effort data have long been recognized as a valuable source of information eventually highly complementary to scientific survey data. But the complexity of the mechanisms driving the way fishermen sample in space and time make the combination of scientific and commercial data challenging.

In this paper, we provide a hierarchical framework to integrate scientific surveys and commercial catch declaration data to infer species distribution while considering the effect of PS on fishing points distribution. The new model allows for exploring and questioning the challenges raised by such integration. The benefit but also the limits of the new approach were evaluated using simulations and through the application of the model to three contrasted demersal case studies (sole, hake and squids) of the Biscay Bay fishery.

Both simulations and case studies demonstrate that ignoring PS in the inference may be highly detrimental when the intensity of PS is strong. The present framework can serve as a tool to assess the benefit of including PS in analysis, depending on the intensity of PS but also on the modelling objectives. As already shown in previous studies (Conn *et al.*, 2017; Pennino *et al.*, 2019), when PS actually occurs in commercial catches, ignoring this process may bias inferences on total biomass estimates. Even if ignoring PS may not hamper the capacity to detect areas of high biomass, the biomass in low-density areas may be overestimated. Therefore, if the objective is to compute biomass indices integrated over a large area, then it might be worth accounting for PS to avoid biased results. By contrast, if the objective is to identify hotspots, the benefits of considering PS may be small with regard to the additional computational time it requires.

The three case studies illustrated the potential of the model to handle the variability of PS behavior among species and fleets. Low PS was revealed for hake, while a moderate and strong PS was revealed for sole and squids, respectively, which is consistent with the expert knowledge on the behavior of those bottom trawls fleets (Y. Vermard, *com. pers.*).

Results also demonstrate the capacity of the framework to integrate commercial catch data from multiple fleets, and the benefits for the quality of inferences when those fleets have different features such as distinct catchabilities or targeting behaviors. For the sole case study, this approach proves useful to distinguish two segments in the bottom trawl fleet, which improved model outputs. This framework could be extended to more than two fleets and combined with other studies analyzing fleets structure (Pelletier and Ferraris, 2000; Ferraris, 2002; Stephens and MacCall, 2004; Deporte *et al.*, 2012; Winker *et al.*, 2013; Okamura *et al.*, 2018).

Challenges in modelling PS

Still, modelling the spatial distribution of commercial fishing locations remains highly challenging (Hintzen, 2021; Hintzen *et al.*, 2021). Our framework is shaped to integrate data from homogeneous fishing fleets supposed to share the same fishing behavior, which simplifies the modelling of the non-uniform spatial intensity of fishing for each fleet. We propose a parsimonious model where the dependence of the sampling intensity to the biomass is supposed to be linear in the log scale. This is a strong hypothesis and departure from this hypothesis may obviously exist in the truth. For instance, the intensity of PS could vary in space such as in Conn *et al.* (2017) who considered that the degree of PS could change across the landscape. On the other hand, however, the log-log linear assumption is easy to implement in other software including the VAST R package used for operational assessments in some management regions (Thorson *et al.*, 2019).

Of course, many other factors may drive the spatial intensity of fishing, and those were simply captured in our model through an additional spatial random term. For instance, fishers' behavior may depend on prior knowledge of fish spatial distribution, on information sharing within fishing cooperatives, on expected distribution of bycatch species, or logistical constraints (e.g., transit costs) (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017). Targeting behavior may also be directed toward an assemblage of species rather than toward a single species (Bourdaud *et al.*, 2019).

The random effect should be able to capture additional variations whenever the departure from a continuous Gaussian random field is not too high. If not, for instance in the case of fishery closures where fishing activity suddenly drops to very low levels (as explored in simulation-estimation), the model may produce biased estimates due to model misspecification. We did not detect such misspecification in our case study, but we recommend that future analyses based on fishery-dependent data present a log-log plot between sampling intensity and predicted biomass density to diagnose strong departure from model hypothesis.

Still, some non-spatial targeting has been reported from multi-species catch records (Stephens and MacCall, 2004; Okamura *et al.*, 2018). Efforts to integrate these methods into spatio-temporal models are underway (Thorson *et al.*, 2016), although these methods have not previously been extended to jointly analyzing multi-species fishery and survey data.

Relative contribution of scientific and commercial data

Our analysis exemplifies that a key issue in such integrated modelling exercise is to get a sensible evaluation of the relative contribution of the different sources of data in estimation. In particular, critical issues with the IM are whether the different data sources provide eventually highly unbalanced quantity of information (then the inferences are fully dominated by one of the data sources; Fletcher *et al.*, 2019) and whether they provide complementary or conflicting information to the final inferences (Saunders *et al.*, 2019; Zipkin *et al.*, 2019; Peterson *et al.*, 2021).

We implemented a likelihood ratio-test (Rufener *et al.*, 2021) to check for model consistency between the IM and the scientific-based model. In most cases, models passed the consistency check successfully, although it was rejected in some cases. Some further analysis should investigate in detail the reasons of these inconsistencies as they could probably shed light on some new research avenues for model improvement. For instance, some neglected vessel effect (*e.g.*, difference in catchability among vessels - Thorson and Ward, 2014) or some too simplistic representation of the sampling and/or the observation process of commercial data might partly explain these inconsistencies.

Simulations revealed that when scientific data and commercial data have balanced sample size, they both contribute to inference and the IM will provide better biomass predictions than models based on single-data set. As expected, when the sample size of commercial data far exceeds scientific data, inference about spatial patterns is mainly driven by the commercial data. In the three case studies, we used commercial data with sample sizes that far exceed the scientific one. In that case, scientific data have relatively limited weight in the final inference. Still, they bring valuable information in areas that are not sampled by the commercial fishery. Also, scientific data remain a critical component in the analysis as they provide some reference data through a standardized sampling plan and a controlled protocol allowing then to assess for the IM consistency. It would be worth applying our framework to other case study that may consist in more balanced data sets, such as models seeking to combine scientific with onboard observer data (Rufener *et al.* 2021), or in pelagic fisheries where acoustic surveys can provide continuous observations over the full domain.

Our results also point out the importance of setting the reference level for the catchability coefficient with either the scientific or the commercial data. In particular, when the sample size of the commercial data far exceeds the scientific survey, fixing the reference level with scientific surveys generally results in higher imprecision, due to the smaller sample size. But still, in certain cases, the scientific data may provide absolute information on biomass and fixing the catchability factor associated with the survey data can result in an interpretable measure of index scale (Thorson *et al.*, 2021). Hence, the choice of the reference level could be a matter of tradeoffs between precision of inferences and interpretation of the results in terms of scale.

The limits of reallocated catch data

Probably one of the major limits of our approach is that the actual framework ignores the uncertainty that arises from the procedure used to reallocate the catch declarations in space. Obtaining the spatialized CPUE inputs used in the model requires pre-treatment of the commercial catch declaration data to allocate declaration data to VMS positions (Hintzen et al., 2012). Raw data corresponds to fishing operations that are daily aggregated and reported at coarse administrative spatial units (0.5° latitude by 1° longitude rectangles). These declarations are then reallocated uniformly on all GPS locations previously identified as fishing in the vessel path. This procedure has been demonstrated to be robust while being a fast and a pragmatic approach for reallocating landings to VMS pings (Gerritsen and Lordan, 2010; Murray *et al.*, 2013). However, it implies strong hypotheses that may artificially increase or transform the information provided by the data. Typically, the uniform reallocation of catch declarations on all GPS positions identified as fishing may smooth the spatial signal, which could potentially explain the lack of species-habitat relationship obtained from the IM. The effect of such reallocation should be explored in further study to better understand its consequences on model predictions/estimates and further model development should investigate how to mitigate its consequences.

Perspectives

Our work raises some major challenges which all constitutes exciting tracks for future research.

Data-weighting approaches could be explored further to better control the contribution of the two sources of data and eventually assess if increasing scientific data weight could improve model predictive capacity. Data-weighting methods intend to modify the relative influence of the data sources by assigning or estimating a weight for each data source (Francis, 2017; Punt, 2017; Wang and Maunder 2017; Punt *et al.*, 2020). Only very few studies have already explored the potential for data weighting in the SDM context (Fletcher *et al.*, 2019). Still, several questions regarding the weight specification remain open or largely debated. For instance, how to rigorously fix/estimate/interpret the weight? Also, when can we consider that a data-weighting approach is relevant or is it only a matter of model misspecification? Some theoretical and modelling development could be highly valuable to provide a generic and rigorous formalization for either data weighting or model correction in the context of SDM (but see for instance the approach provided by Thorson *et al.* (2017b) for composition data in the context of stock assessment models).

Another option would consist in developing an alternative observation model for the commercial CPUE in order to better capture the uncertainty associated with the reallocation procedure. As a general idea, an observation model could be developed to explicitly represent that CPUE are available at the scale of the daily fishing activity (the scale that corresponds to the catch declaration), rather than artificially reallocating uniformly catch declarations on related VMS pings. Doing so, the quantity of information provided by commercial data would be more representative of the information they really contain.

Future work should also seek to better integrate the discrete-choice and econometric analyses emphasizing the complexity of the processes related to the choice of fishing locations. For instance, the sampling process could account for the pluri-specific nature of fisheries (Bourdaud *et al.*, 2019) and additional factors other than fish distribution could be included to explain the variability of sampling intensity in space and time (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017).

Finally, including a temporal dimension in the model and fitting a longer time series looks a fruitful research avenue. Moving to spatio-temporal modelling that would consider temporal autocorrelation in the spatial distribution may be methodologically challenging (Cameletti et al., 2013), but represents an exciting step towards a better understanding of the seasonal spatial distribution of fish resources. Indeed, commercial data are often available all along the year, when scientific surveys most often occur once or twice a year. Combining scientific and catch declarations data within an integrated spatio-temporal framework built at an infra-annual time step (e.g., season, month) would allow to complement the gap of information to investigate fish spatio-temporal distribution at a finer temporal scale than what is possible using scientific data only (Bourdaud *et al.*, 2017; Pinto *et al.*, 2019; Rufener *et al.*, 2021). It would offer new opportunities to interpret seasonal patterns of distribution (Kai et al., 2017), identify fish functional habitats such as spawning areas (Paradinas et al., 2015; Delage and Le Pape, 2016), and provide the required knowledge for protecting those habitats (Schmitten, 1999; Erisman et al., 2020).

SUPPLEMENTARY MATERIAL

All the supplementary material documents are available at the ICESJMS online version of the manuscript. They provide additional information on the modelling framework (SM1), material and methods for simulations (SM2) and case studies (SM3), results for simulations (SM4) and case studies (SM5).

ACKNOWLEDGMENTS

The authors are grateful to the Direction des pêches maritimes et de l’aquaculture (DPMA) and Ifremer (Système d’Informations Halieutiques - SIH) who provided the aggregated VMS data. The findings and conclusions of the present paper are those of the authors.

The authors thank David Eme who provided tidy environmental covariates data as well as Kasper Kristensen, Jean-Baptiste Lecomte, Louise Day and Pierre-Yves Hernvann for their feedbacks and their highly valuable advice. The authors thanks also Maxime Olmos, John Best, and two anonymous reviewers whose feedbacks greatly improved the manuscript.

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’ (https://cran.r-project.org/web/packages/icesDatras/index.html). Logbooks and VMS data are confidential data and they are available on specific request to DPMA. Codes that support the findings of this study are on gitlab and can be given access on request at the address: baptiste.alglave@agrocampus-ouest.fr.

FIGURES AND TABLES

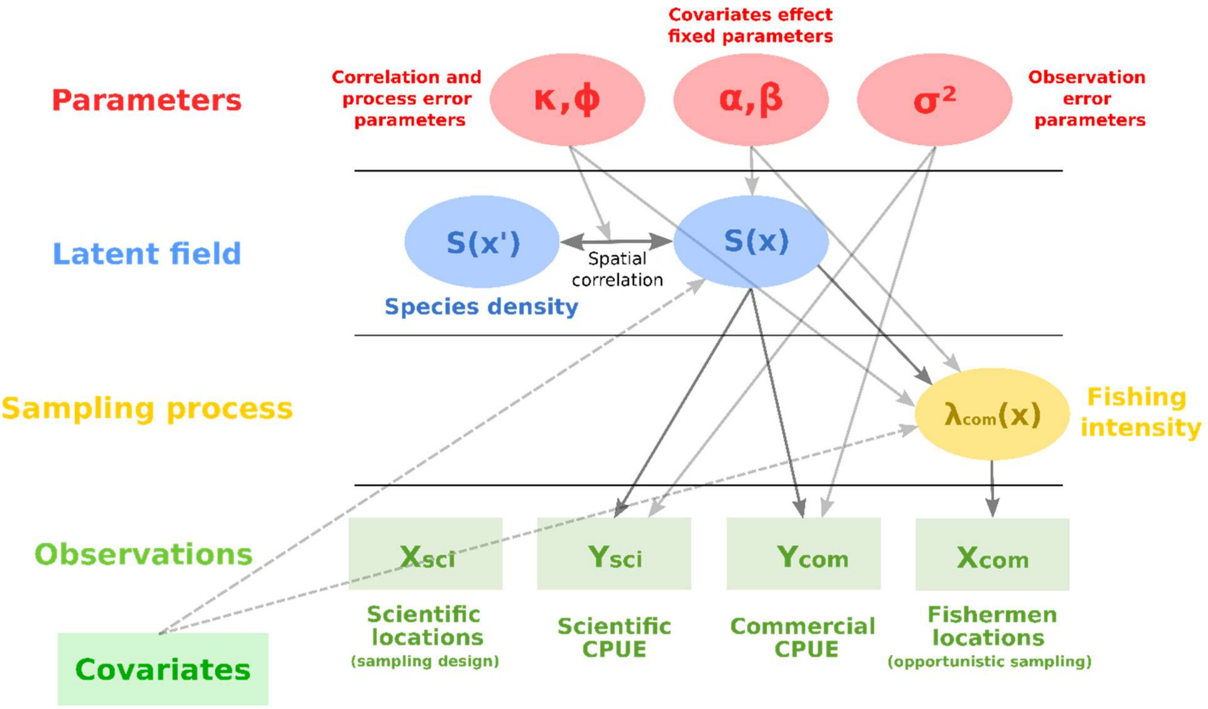


Figure 1. Diagram of the spatial integrated model including preferential sampling for commercial data. Locations of scientific trawls do not contribute directly to the likelihood.

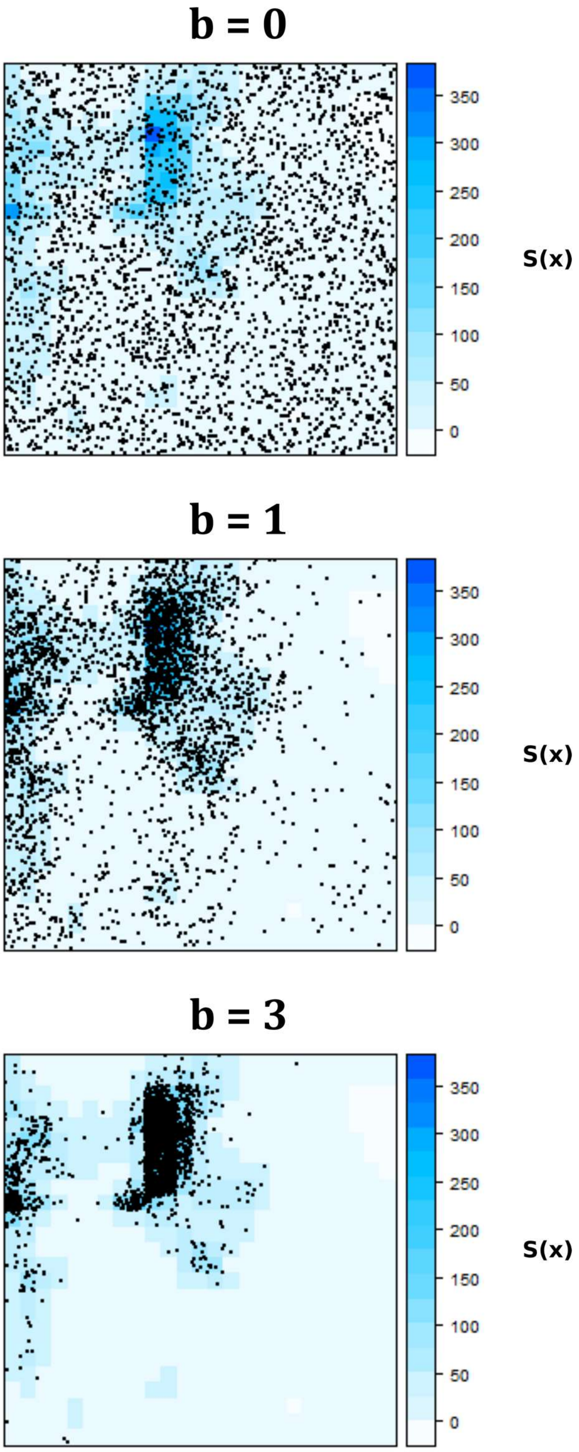


Figure 2. Maps of simulated commercial sampling points obtained for three values of preferential sampling (b=0, b=1, b=3). Blue scale: values of the simulated biomass field. Dots: fishing points. For $b = 0$,the targeting metric $T_j(x) = 1$. For $b = 1$,

$\underset{x}{arg\,max}\, T_j(x) = 12,\ q_{50\%}\{T_j(x)\} = 0.4$ *. For* $b = 3,\ \underset{x}{arg\,max}\, T_j(x) = 80,\ q_{50\%}\{T(x)\} = 0.002$ *(SM 1.3).*

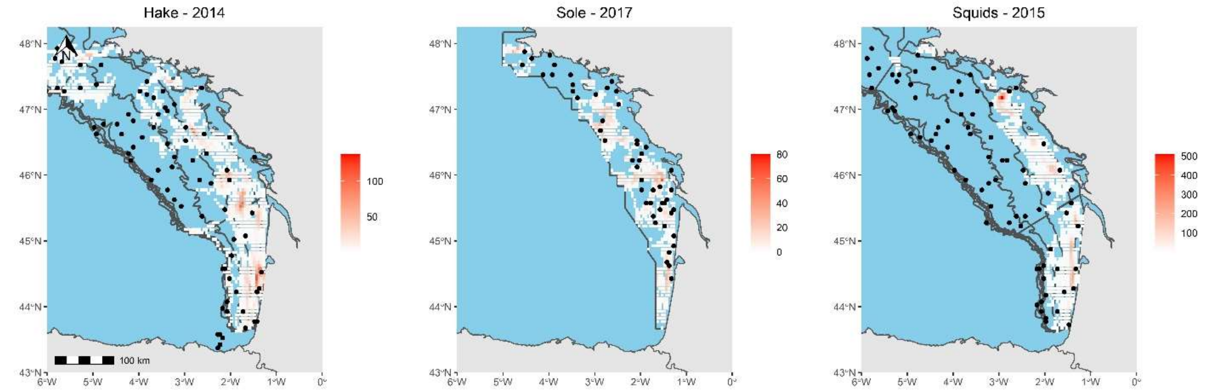


Figure 3. Map of scientific samples (black dot) and commercial sampling distribution (red color scale – unit: fishing hours). Note that all scientific hauls last around 30 minutes. Black lines - limits of the spatial domains covered by the scientific survey (Orhago and EVHOE) that delineate the study area. Left – Hake, November 2014 (EVHOE; commercial data from otter bottom trawls targeting demersal species OTB_DEF). Middle – Sole, November 2017 (Orhago; commercial data from otter bottom trawls targeting demersal species OTB_DEF). Right – squid, year 2015 (EVHOE; commercial data from otter bottom trawls targeting cephalopods OTB_CEP).

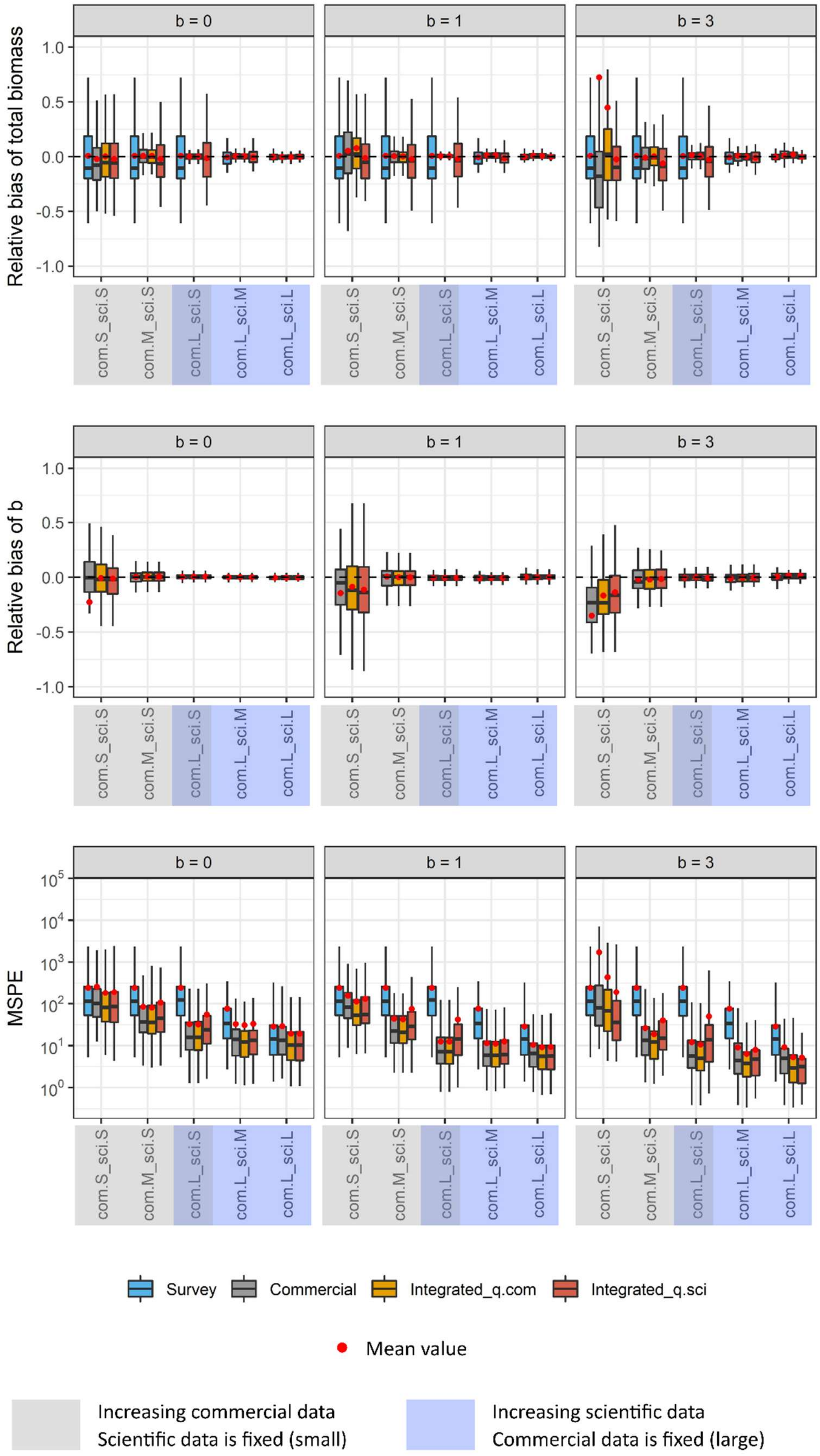

b = 0
b = 1
b = 3
Relative bias of total biomass
1.0
0.5
0.0
-0.5
-1.0
Relative bias of b
MSPE
10^5
10^4
10^3
10^2
10^1
10^0
com.S_sci.S
com.M_sci.S
com.L_sci.S
com.L_sci.M
com.L_sci.L
Survey
Commercial
Integrated_q.com
Integrated_q.sci
Mean value
Increasing commercial data
Scientific data is fixed (small)
Increasing scientific data
Commercial data is fixed (large)

Figure 4. Performance metrics obtained for various commercial and scientific data sample size. Column: intensity of the preferential sampling in simulated data. x-axis: 5 combinations of commercial and scientific sample size. 'com' *stands for commercial,* 'sci' *stands for scientific,* S *stands for small sample size (50),* M *stands for medium sample size (400),* L *stands for large sample size (3000). Colors: model configurations. Integrated_q.com: integrated model with catchability fixed to 1 for commercial data; Integrated_q.sci: integrated model with catchability fixed to 1 for scientific data. Boxplots represent the variability among the 100 replicates.*

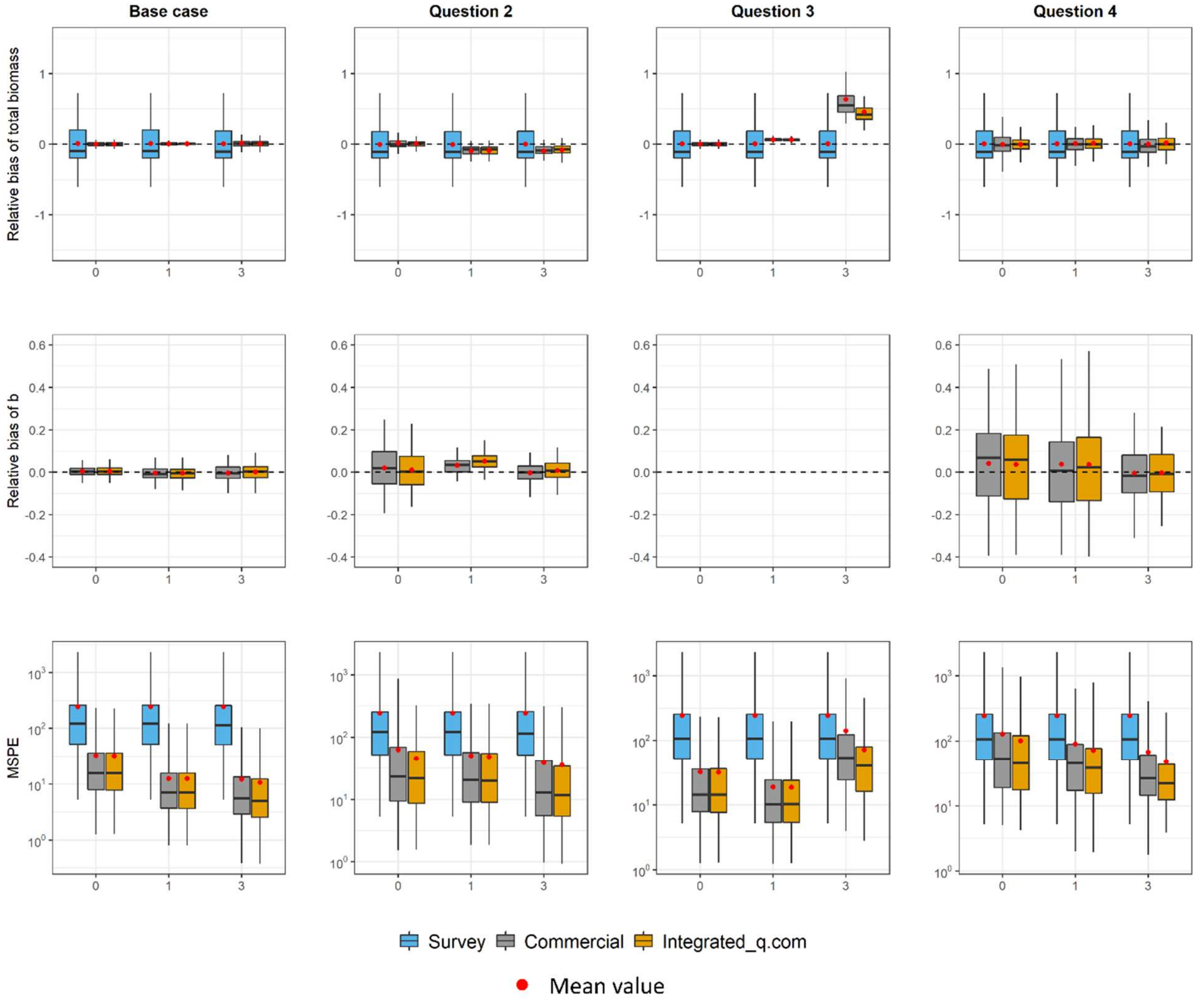


Figure 5. Performance metrics obtained in different data and model configurations. Red points: mean value. 1^{st} column: no discrepancy between simulation and estimation. 2^{nd} column: commercial data do not cover a 9 x 9 zone of the grid. 3^{rd} column: b is arbitrarily fixed to 0 in the estimation models. 4^{th} column: data simulated with a random effect η in the sampling intensity process. In all configurations, simulations are conducted for three levels of preferential sampling (x-axis: b = 0, b = 1, b = 3). Colors: data sources used in the integrated model for inferences. Integrated_q.com: integrated model with catchability fixed with commercial data. Boxplots represent the variability among the 100 replicates.

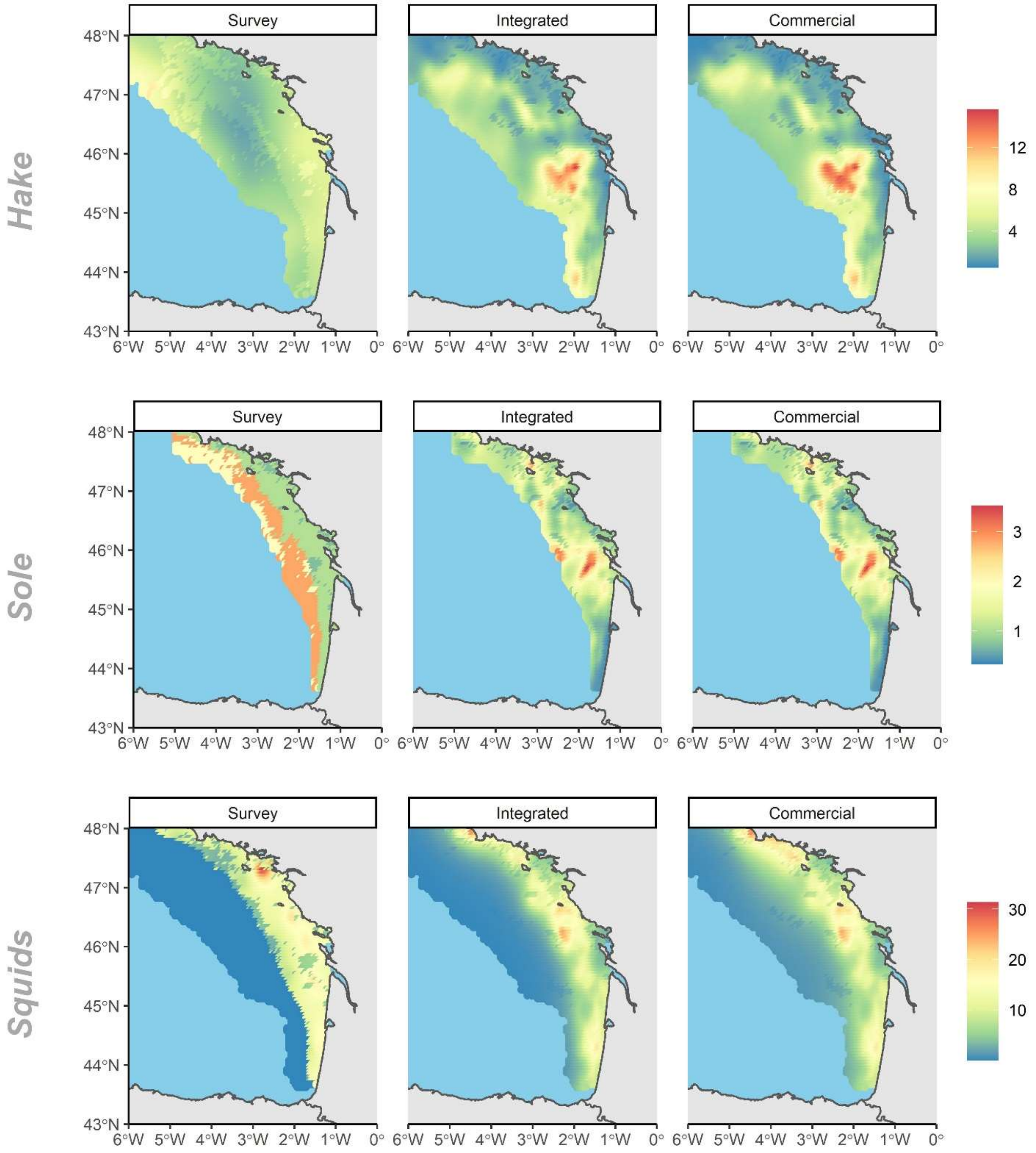
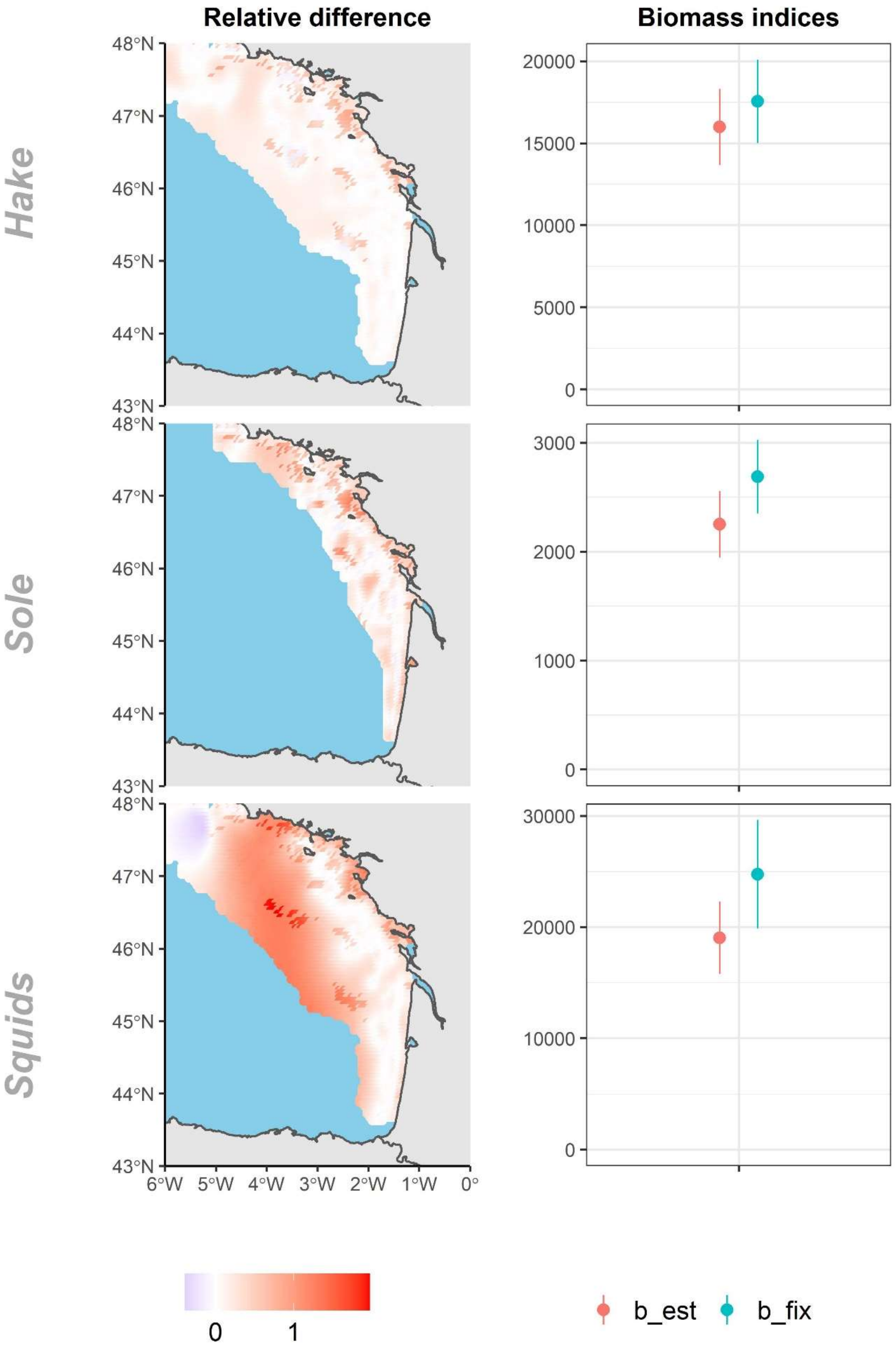


Figure 6. Prediction of the relative biomass for each case study. 1^{st} column: model fitted to scientific data only; 2^{nd} column: integrated model accounting for PS; 3^{rd} column: commercial-based model accounting for PS. When the model is fitted to scientific data only, relative biomass is rescaled with the relative catchability parameter estimated within the integrated model so that all maps are in the same scale.

772



Relative difference
Biomass indices
Hake
Sole
Squids
48°N
47°N
46°N
45°N
44°N
43°N
6°W
5°W
4°W
3°W
2°W
1°W
0°
20000
15000
10000
5000
0
3000
2000
1000
0
30000
20000
10000
0
0
1
b_est
b_fix

Figure 7. Comparison of relative difference in biomass spatial predictions (calculated as ($S_{b\_fix}(x) - S_{b\_est}(x))/S_{b\_est}(x)$) in space (left) and of total biomass (sum on the spatial domain; right) obtained with the integrated models from the 3 case studies when accounting or not for preferential sampling. b_est: PS is estimated. b_fix: PS is not accounted for.

Table 1: Simulations

General simulations description	
Biomass field	Depends on one continuous covariate (Γ_S) and one random spatial effect (δ). Simulated within a 25 x 25 grid. Both are simulated independently through a GRF with Matérn covariance function. Their range (ρ) and marginal variance are fixed respectively to 10 and 1. n.b. the marginal variance quantifies the variability of the spatial process. For more details on marginal variance parameterization, see Lindgren *et al.* (2011).
Scientific data	Random stratified plan within 4 strata (see Figure S2.1) Catchability fixed to 1 Simulated with 10% of zeroes ($\xi_j = 0$)
Commercial data	Simulated according to three PS levels (i.e. three values for b - see Figure 2). - $b = 0$: commercial sampling is not preferential; - $b = 1$: preferential sampling is moderate, commercial vessels mainly target areas where fish biomass is high; - $b = 3$: commercial sampling is highly preferential and vessels strongly target zones where biomass is high. η is set to 0 for Q1, Q2, Q3. For Q4, η is set to tailor the sole case study. The range of η is set to 40 (4 times the range of δ), the marginal variance is set to 5 (5 times the marginal variance of δ). Catchability fixed to 1 Simulated with 30 % of zero when PS is null ($\xi_j = -1$)

	Simulation scenarios					**Model configurations**		
	b	**Scientific sample size**	**Commercial samples size**	**Coverage of the study area**	**Additional random effect in sampling intensity (η)**	**Data sources considered in the model**	**PS estimated**	**Fixed catchability**
Question 1: How do each data source contribute to inferences?	0,1,3	50	50, 400, 3000	Full	No	Scientific only, commercial only, both	yes	Scientific or Commercial
	0,1,3	50, 400, 3000	3000	Full	No	Scientific only, commercial only, both	yes	Scientific or Commercial
Question 2: How does a partial coverage of the study area by the commercial data affect the quality of the estimation?	0,1,3	50	3000	No fishing in a 9x9 cells box	No	Scientific only, commercial only, both	yes	Commercial
Question 3: What is the cost of ignoring PS in estimation when sampling is preferential?	0,1,3	50	3000	Full	No	Scientific only, commercial only, both	no (b fixed at 0)	Commercial
Question 4: How does the estimation perform when additional processes other than PS drive the fishing locations?	0,1,3	50	3000	Full	Yes	Scientific only, commercial only, both	yes	Commercial

REFERENCES

Abbott, J., Haynie, A., and Reimer, M. 2015. Hidden Flexibility: Institutions, Incentives, and the Margins of Selectivity in Fishing. Land Economics, 91: 169–195.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. 2014. Hierarchical modeling and analysis for spatial data. CRC press.

Bourdaud, P., Travers-Trolet, M., Vermard, Y., Cormon, X., and Marchal, P. 2017. Inferring the annual, seasonal, and spatial distributions of marine species from complementary research and commercial vessels' catch rates. ICES Journal of Marine Science, 74: 2415–2426.

Bourdaud, P., Travers-Trolet, M., Vermard, Y., and Marchal, P. 2019. Improving the interpretation of fishing effort and pressures in mixed fisheries using spatial overlap metrics. Canadian Journal of Fisheries and Aquatic Sciences, 76: 586–596.

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

Cheung, W. W., Lam, V. W., Sarmiento, J. L., Kearney, K., Watson, R., and Pauly, D. 2009. Projecting global marine biodiversity impacts under climate change scenarios. Fish and fisheries, 10: 235–251. Wiley Online Library.

Conn, P. B., Thorson, J. T., and Johnson, D. S. 2017. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. Methods in Ecology and Evolution, 8: 1535–1546.

Cornou, A.-S., Quinio-Scavinner, M., Sagan, J., Cloâtre, T., Dubroca, L., and Billet, N. 2021. Captures et rejets des métiers de pêche francais - Résultats des observations à bord des navires de pêche professionnelle en 2019. Ifremer.

Cressie, N. A. 1993. Statistics for spatial data. John Willy and Sons. Inc., New York.

Delage, N., and Le Pape, O. 2016. Inventaire des zones fonctionnelles pour les ressources halieutiques dans les eaux sous souveraineté française. Première partie: définitions, critères d'importance et méthode pour déterminer des zones d'importance à protéger en priorité. Rapport de recherche. Pôle halieutique AGROCAMPUS OUEST, Rennes.

Deporte, N., Ulrich, C., Mahévas, S., Demanèche, S., and Bastardie, F. 2012. Regional métier definition: a comparative investigation of statistical methods using a workflow applied to international otter trawl fisheries in the North Sea. ICES Journal of Marine Science, 69: 331–342. Oxford University Press.

Diggle, P. J., Menezes, R., and Su, T. 2010. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59: 191–232.

Diggle, P. J. 2013. Statistical analysis of spatial and spatio-temporal point patterns. CRC press.

Ducharme-Barth, N. D., Grüss, A., Vincent, M. T., Kiyofuji, H., Aoki, Y., Pilling, G., Hampton, J., *et al.* 2022. Impacts of fisheries-dependent spatial sampling patterns on catch-per-unit-effort standardization: A simulation study and fishery application. Fisheries Research, 246: 106169.

Erisman, B. E., Grüss, A., Mascareñas-Osorio, I., Lícon-González, H., Johnson, A. F., and López-Sagástegui, C. 2020. Balancing conservation and utilization in spawning aggregation fisheries: a trade-off analysis of an overexploited marine fish. ICES Journal of Marine Science, 77: 148–161. Oxford University Press.

Ferraris, J. 2002. Fishing fleet profiling methodology. Food & Agriculture Org.

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M. 2019. A practical guide for combining data to model species distributions. Ecology, 100: e02710.

Francis, R. C. 2017. Revisiting data weighting in fisheries stock assessment models. Fisheries Research, 192: 5–15.

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

Gerritsen, H., and Lordan, C. 2010. 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.

Gimenez, O., Buckland, S. T., Morgan, B. J. T., Bez, N., Bertrand, S., Choquet, R., Dray, S., *et al.* 2014. Statistical ecology comes of age. Biology Letters, 10: 20140698. Royal Society.

Girardin, R., Hamon, K. G., Pinnegar, J., Poos, J. J., Thébaud, O., Tidd, A., Vermard, Y., *et al.* 2017. Thirty years of fleet dynamics modelling using discrete-choice models: What have we learned? Fish and Fisheries, 18: 638–655.

Grüss, A., Thorson, J. T., Carroll, G., Ng, E. L., Holsman, K. K., Aydin, K., Kotwicki, S., *et al.* 2020. Spatio-temporal analyses of marine predator diets from data-rich and data-limited systems. Fish and Fisheries, 21: 718–739.

Guisan, A., and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. Ecological modelling, 135: 147–186. Elsevier.

Haynie, A. C., Hicks, R. L., and Schnier, K. E. 2009. Common property, information, and cooperation: Commercial fishing in the Bering Sea. Ecological Economics, 69: 406–413.

Hilborn, R., and Walters, C. J. (Eds). 1992. Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty. Springer US. https://www.springer.com/gp/book/9780412022715 (Accessed 14 June 2021).

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

Hintzen, N. T. 2021. Zooming into small-scale fishing patterns: The use of vessel monitoring by satellite in fisheries science. Wageningen University.

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.

ICES. 2005. Report of the Workshop on Survey Design and Data Analysis (WKSAD). Sète, France.

ICES. 2012. Manual for the international bottom trawl surveys. SISP 1-IBTS Copenhagen, Denmark.

ICES. 2020a. International Bottom Trawl Survey Working Group (IBTSWG). ICES Scientific Reports. ICES. http://www.ices.dk/sites/pub/Publication Reports/Forms/DispForm.aspx?ID=37066 (Accessed 28 May 2021).

ICES. 2020b. Working Group for the Bay of Biscay and the Iberian Waters Ecoregion (WGBIE). ICES Scientific Reports. ICES. http://www.ices.dk/sites/pub/Publication Reports/Forms/DispForm.aspx?ID=36841 (Accessed 28 May 2021).

Kai, M., Thorson, J. T., Piner, K. R., and Maunder, M. N. 2017. Spatiotemporal variation in size-structured populations using fishery data: an application to shortfin mako (Isurus oxyrinchus) in the Pacific Ocean. Canadian Journal of Fisheries and Aquatic Sciences, 74: 1765–1780.

Kristensen, K., Thygesen, U. H., Andersen, K. H., and Beyer, J. E. 2014. Estimating spatio-temporal dynamics of size-structured populations. Canadian Journal of Fisheries and Aquatic Sciences, 71: 326–336.

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.

Le Pape, O., Chauvet, F., Mahévas, S., Lazure, P., Guérault, D., and Désaunay, Y. 2003. Quantitative description of habitat suitability for the juvenile common sole (Solea solea, L.) in the Bay of Biscay (France) and the contribution of different habitats to the adult population. Journal of Sea Research, 50: 139–149.

Lindgren, F., Rue, H., and Lindström, 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. Wiley Online Library.

Martínez-Minaya, J., Cameletti, M., Conesa, D., and Pennino, M. G. 2018. Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stochastic environmental research and risk assessment, 32: 3227–3244. Springer.

Moriarty, M., Sethi, S. A., Pedreschi, D., Smeltz, T. S., McGonigle, C., Harris, B. P., Wolf, N., *et al.* 2020. Combining fisheries surveys to inform marine species distribution modelling. ICES Journal of Marine Science, 77: 539–552. Oxford University Press.

Murray, L. G., Hinz, H., Hold, N., and Kaiser, M. J. 2013. The effectiveness of using CPUE data derived from Vessel Monitoring Systems and fisheries logbooks to estimate scallop biomass. ICES Journal of Marine Science, 70: 1330–1340.

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.

Ocean Studies Board, and National Research Council. 2000. Improving the collection, management, and use of marine fisheries data. National Academies Press.

Okamura, H., Morita, S. H., Funamoto, T., Ichinokawa, M., and Eguchi, S. 2018. Target-based catch-per-unit-effort standardization in multispecies fisheries. Canadian Journal of Fisheries and Aquatic Sciences, 75: 452–463.

Paradinas, I., Conesa, D., Pennino, M., Muñoz, F., Fernández, A., López-Quílez, A., and Bellido, J. 2015. Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. Marine Ecology Progress Series, 528: 245–255.

Parent, E., and Rivot, E. 2012. Introduction to hierarchical Bayesian modeling for ecological data. CRC Press.

Pati, D., Reich, B. J., and Dunson, D. B. 2011. Bayesian geostatistical modelling with informative sampling locations. Biometrika, 98: 35–48.

Pelletier, D., and Ferraris, J. 2000. A multivariate approach for defining fishing tactics from commercial catch and effort data. Canadian Journal of Fisheries and Aquatic Sciences, 57: 51–65. NRC Research Press.

Pennino, M. G., Conesa, D., Lopez-Quilez, A., Munoz, F., Fernández, A., and Bellido, J. M. 2016. Fishery-dependent and-independent data lead to consistent estimations of essential habitats. ICES Journal of Marine Science, 73: 2302–2310. Oxford University Press.

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa, D. 2019. Accounting for preferential sampling in species distribution models. Ecology and evolution, 9: 653–663.

Peterson, C. D., Courtney, D. L., Cortés, E., and Latour, R. J. 2021. Reconciling conflicting survey indices of abundance prior to stock assessment. ICES Journal of Marine Science, 78: 3101–3120.

Pinto, C., Travers-Trolet, M., Macdonald, J. I., Rivot, E., and Vermard, Y. 2019. Combining multiple data sets to unravel the spatiotemporal dynamics of a data-limited fish stock. Canadian Journal of Fisheries and Aquatic Sciences, 76: 1338–1349. NRC Research Press.

Planque, B., Loots, C., Petitgas, P., LINDSTRøM, U. L. F., and Vaz, S. 2011. Understanding what controls the spatial distribution of fish populations using a multi-model approach. Fisheries Oceanography, 20: 1–17.

Punt, A. E. 2017. Some insights into data weighting in integrated stock assessments. Fisheries Research, 192: 52–65.

Punt, A. E., Dunn, A., Elvarsson, B. Þ., Hampton, J., Hoyle, S. D., Maunder, M. N., Methot, R. D., *et al.* 2020. Essential features of the next-generation integrated fisheries stock assessment package: A perspective. Fisheries Research, 229: 105617.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rivoirard, J., Simmonds, J., Foote, K. G., Fernandes, P., and Bez, N. 2008. Geostatistics for estimating fish abundance. John Wiley & Sons.

Rochette, S., Rivot, E., Morin, J., Mackinson, S., Riou, P., and Le Pape, O. 2010. Effect of nursery habitat degradation on flatfish population: Application to Solea solea in the Eastern Channel (Western Europe). Journal of sea Research, 64: 34–44. Elsevier.

Rufener, M.-C., Kristensen, K., Nielsen, J. R., and Bastardie, F. 2021. Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. Ecological Applications: e02453.

Salas, S., and Gaertner, D. 2004. The behavioural dynamics of fishers: management implications. Fish and Fisheries, 5: 153–167.

Saunders, S. P., Farr, M. T., Wright, A. D., Bahlai, C. A., Ribeiro Jr., J. W., Rossman, S., Sussman, A. L., *et al.* 2019. Disentangling data discrepancies with integrated population models. Ecology, 100: e02714.

Schaub, M., and Abadi, F. 2011. Integrated population models: a novel analysis framework for deeper insights into population dynamics. Journal of Ornithology, 152: 227–237. Springer.

Schmitten, R. A. 1999. Essential fish habitat: opportunities and challenges for the next millennium. *In* American Fisheries Society Symposium, p. 10.

Stephens, A., and MacCall, A. 2004. A multispecies approach to subsetting logbook data for purposes of estimating CPUE. Fisheries Research, 70: 299–310.

Thorson, J. T., and Ward, E. J. 2014. Accounting for vessel effects when standardizing catch rates from cooperative surveys. Fisheries Research, 155: 168–176.

Thorson, J. T., Ianelli, J. N., Munch, S. B., Ono, K., and Spencer, P. D. 2015a. Spatial delay-difference models for estimating spatiotemporal variation in juvenile production and population abundance. Canadian journal of fisheries and aquatic sciences, 72: 1897–1915.

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K. 2015b. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution, 6: 627–637. Wiley Online Library.

Thorson, J. T. 2015. Spatio-temporal variation in fish condition is not consistently explained by density, temperature, or season for California Current groundfishes. Marine Ecology Progress Series, 526: 101–112.

Thorson, J. T., Fonner, R., Haltuch, M. A., Ono, K., and Winker, H. 2016. Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data. Canadian Journal of Fisheries and Aquatic Sciences, 74: 1794–1807.

Thorson, J. T., Jannot, J., and Somers, K. 2017a. Using spatio-temporal models of population growth and movement to monitor overlap between human impacts and fish populations. Journal of Applied Ecology, 54: 577–587.

Thorson, J. T., Johnson, K. F., Methot, R. D., and Taylor, I. G. 2017b. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research, 192: 84–93.

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. NRC Research Press.

Thorson, J. T., Adams, G., and Holsman, K. 2019. Spatio-temporal models of intermediate complexity for ecosystem assessments: A new tool for spatial fisheries management. Fish and Fisheries, 20: 1083–1099.

Thorson, J. T., Cunningham, C. J., Jorgensen, E., Havron, A., Hulson, P.-J. F., Monnahan, C. C., and von Szalay, P. 2021. The surprising sensitivity of index scale to delta-model assumptions: Recommendations for model-based index standardization. Fisheries Research, 233: 105745.

Trenkel, V. M., Beecham, J. A., Blanchard, J. L., Edwards, C. T., and Lorance, P. 2013. Testing CPUE-derived spatial occupancy as an indicator for stock abundance: application to deep-sea stocks. Aquatic living resources, 26: 319–332. EDP Sciences.

Winker, H., Kerwath, S. E., and Attwood, C. G. 2013. Comparison of two approaches to standardize catch-per-unit-effort for targeting behaviour in a multispecies hand-line fishery. Fisheries Research, 139: 118–131.

Witman, J. D., and Roy, K. 2009. Marine Macroecology. University of Chicago Press. 442 pp.

Zipkin, E. F., Inouye, B. D., and Beissinger, S. R. 2019. Innovations in data integration for modeling populations. Ecology, 100: e02713.