Going from coarse landings data to fine scale species distribution

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

Going from coarse landings data to fine scale species distribution

To do:

- **email to Kasper**
- **redaction of M&M**
- **run single square simulations**
- **run multiple square simulations**

References for change of support issues in integrated modelling:

https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1002/ecy.2710

https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12793

https://www.sciencedirect.com/science/article/pii/S0169534719302551#b0010

https://www.sciencedirect.com/science/article/pii/S0006320721001993

Material and methods

We base our approach on the model developped by Alglave et al. (2022) and build on these developpments to provide a solution to overcome the issues related to change of support in our specific context.

Defining the problem from a modelling point of view

Let's define the latent field S the punctual observations Y. S is a spatial Gaussian Field (GF) defined in space such as $S(x) = \mu + \beta.\Gamma + \delta(x)$. $\beta.\Gamma$ is the covariate term with Γ the matrix of deisgn of the coavariates and β the effect of the covariates. $\delta(x)$ is GF capturing spatial correlation ($\delta \sim \mathcal{N}(0, \Sigma)$).

Punctual observations Y_i are conditionnal on the latent field values at fishing location x_i for observation i. Observations are supposed to follow some distribution $\mathcal{L}_Y$. In our case, data are positive continuous zero inflated data and they are modelled through a zero-inflated lognormal model described in supplementary material (SM) previously introduced by Thorson (2018) and already used in Alglave et al. (2022).

$$Y_i|S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma)$$

All observations belong to a catch declaration D_j such as:

$$D_j = \sum_{i \in \mathcal{P}_j} Y_i$$

$j \in [\![1, n]\!]$ is the declaration index with n the number of declarations. $\mathcal{P}_j$ is the vector of all fishing positions related to the j^{th} declaration. The punctual observations Y_i are indexed through $i \in [\![1, m_j]\!]$ with m_j the number of fishing positions belonging to the j^{th} declaration.

In standard processing, we know D_j through logbooks data, x_i through VMS data and D_j are reallocated uniformly on related x_i so that derived punctual observations Y_i^* are computed as $Y_i^* = D_j/m_j$. The likelihood is then directly computed on reallocated Y_i^* assuming these are the exact punctual observations. This strongly simplifies the actual process of observation and have most likely repercussion on model performance.

To overcome such limitation, an alternative is to consider that only the catch declarations are observed, but they can inform of the punctual observations Y_i which are not-observed (i.e. they are latent variables).

Such way, we define some distribution $\mathcal{L}_D$ as the distribution of the catch declarations $D_j|S_{\mathcal{P}_j}, \mathcal{P}_j \sim \mathcal{L}_D(S_{\mathcal{P}_j}, \xi, \sigma)$ with $S_{\mathcal{P}_j} = (S(x_1), ..., S(x_i), ..., S(x_{m_j}))$.

Our approach is to match the 2 first moments of D_j (obtained from the moments of $\mathcal{L}_Y$ i.e. Y_i probability distribution) and $\mathcal{L}_D$.

First, as at the punctual level, the declarations can be decomposed in 2 components: (1) the probability to obtain a zero value and, if the declaration is positive, (2) the probability to obtain a certain declaration value.

(1) We derive the probability to obtain a zero-declaration $P(D_j = 0|S_{\mathcal{P}_j}, \mathcal{P}_j)$ by simply multiplying the probability to obtain a zero-punctual observation $P(Y_i = 0|S(x_i), x_i)$ to all fishing points $x_i \in \mathcal{P}_j$ as $P(D_j = 0|S_{\mathcal{P}_j}, \mathcal{P}_j) = \prod_{i \in \mathcal{P}_j} P(Y_i = 0|S(x_i), x_i)$.

(2) If the declarations is positive, based on the moments of $\mathcal{L}_Y$, we derive $E(D_j|D_j > 0)$ and $Var(D_j|D_j > 0)$ (see Table ...). Assuming $D_j|D_j > 0$ follows some probability distribution $\mathcal{L}_{D_j|D_j>0}$, we make the two first moments of the $\mathcal{L}_{D_j|D_j>0}$ coincides with $E(D_j|D_j > 0)$ and $Var(D_j|D_j > 0)$. Considering $\mathcal{L}_{D_j|D_j>0}$ has a lognormal distribution, such link is done through the formulas $\mu_D = E(D|D_j > 0)$ for the mean component of $\mathcal{L}_{D_j|D_j>0}$ and $\sigma_D = ln(\frac{Var(D|D_j>0)}{E(D|D_j>0)^2} + 1)$ for its variance component.

Equations of $E(D_j|D_j > 0)$ and $Var(D_j|D_j > 0)$ are given by:

$$E(D_j|D_j > 0) = \frac{\sum_{i \in \mathcal{P}_j} S(x_i)}{1 - \pi_j}$$

$$Var(D_j|D_j > 0) = \frac{\sum_{i \in \mathcal{P}_j} Var(Y_i)}{1 - \pi_j} - \frac{\pi_j}{(1 - \pi_j)^2} E(D_j)^2$$

$$Var(Y_i) = \frac{S(x_i)^2}{1 - p_i}(e^{\sigma^2} - (1 - p_i))$$

All calculations to obtain these formulas are available in SM.

The estimation is realized through Template Model Builder (TMB), an effective tool to build hierarchical models and perform maximum likelihood estimation through automatic differentiation and Laplace Approximation (Kristensen, Nielsen, Berg, Skaug, & Bell, 2016).

Simulation-estimation

To evaluate what bring the performance of the new approach, we assess model performance through simulation-estimation.

First, we limit the study to a single square resolution assuming only commercial data feed the model with a simplified version of the latent field model. This steps aims at exploring the properties of the model.

Then, we extend the simulation-estimation study to several rectangles. We assume both commercial and survey data feed the model and try to fit to the real case application.

Each time the latent field is modelled through the equation introduced in the previous paragraph. The covariate is modelled as a continous GRF that we suppose known at each point of the grid. The covariate effect is fixed to $\beta_S = 2$ and the intercept is also fixed to $\mu = 2$. The random effect (for the several rectangle simulation) is modelled as a Matérn GRF, but is supposed unkwnon over the area. All parameterizations are detailed in the table ... .

Each time we compare several model configurations and evaluate these in regards to 2 metrics:

- the MSPE which quantifies the accuracy of the spatial predictions of the latent field over the spatial domain (x are the locations, n is the number of total locations).

$$MSPE = \frac{\sum_x (S(x) - \hat{S}(x))^2}{n}$$

- the parameter of the species habitat relationship β_S compared with its estimate $\hat{\beta}_S$

Single-square simulations. We study the effect of 2 variables: - the amount of commercial data: the amount of data is progressively increased from 10, 100 and 1000. To fit the real data, the number of fishing pings per sequence is fixed to 10 as it is the average

number of fishing locations for one declaration. This means that the number of declarations increases respectively (1, 10 and 100). - the number of fishing areas within a declaration sequence. The samples belonging to a specific declaration can be sampled in the same area or can mix catches from distinct areas. Here, we model such process by assuming that the fishing declaration was realised either in a single zone, in 3 distinct zones or in 5 distinct zones. The zone size was chosen to be a square of side 6.

Concretely, for each fishing decaration the fishing pings are sampled in 2 steps: first the centroid of the zones are randomly sampled over the full simulation domain and then the fishing positions are sampled within the radius of the centroid of the zones. At each fishing position a catch is realized conditionnally on the value of the latent field at the related position.

We compare three simulation/estimation configurations: - the golden standard configuration: observations are known exactly and the standard model is fitted to the data. - the one corresponding to the current configuration: simulated catch are reallocated over fishing locations, the standard model is fitted to the data (i.e. the one working at the level of Y_i^*) - the one corresponding to the alternative configuration: simulated catch are reallocated over fishing locations, the alternative model is fitted to the data (i.e. the one working at the level of D_j)

In addition to the metrics we introduced to compare model configuration, we also assess the simulated/estimated values for the intercept μ, the observation variance parameter σ^2 and the zero-inflation parameter ξ.

Multiple-squares simulations. In this case, we extend the study to more than one single rectangle, included scientific data in inference and, in addition to the covariate, we also simulate/estimate a spatial random effect in the latent field.

The study area is based on the case study. It includes the whole coast of the Bay of Biscay (Figure 5A). To fit the case study, we simulated 3000 of fishing positions regrouped in

300 declarations (10 pings per declaration). To illustrate the case that commercial data may not cover the full area, we allow the commercial samples to cover only 2/3 ot the area. The simulation of a commercial points of a catch declaration is realized in 3 steps. (1) The declarations is affected to one of the ICES rectangle. (2) A point is sampled within this statistical rectangle. (3) The fishing locations are randomly sampled around the fishing centroid. The radius of the fishing zone is fixed as in the simulations at the level of a single statistical rectangle. Note that we do not explore the effect of exploring several zones within the same declaration as it is already studied in the single-square simulations.

100 scientific samples are simulated following a random stratified plan thus contrarly to commercial data they cover the entire study domain (similarly to the Orhago survey - see the case study).

We compare several model configurations: - to assess what bring our alternative approach, we compare models built on reallocated commercial catch data Y_i^* versus models built on catch declaration data D_j. - to assess the information brought by each data source, we compare models built on scientific data only (scientific-based models), models built on commercial data only (commercial-based models) and models combining both data sources (integrated models).

In addition to the 2 metrics introduced at the begginning of the section (MSPE and species-habitat parameter), we also compare the simulated/estimated values of the range parameter.

Case-study: sole of the Bay of Biscay

We applied the approach to the sole of the Bay of Biscay.

VMS-logbooks data was extracted for the bottom trawlers. The methods to cross VMS-logbooks data and to filter the fleet is already extensively described in the previous papers (Alglave et al., 2022) and is not developped further here.

To ease convergence, we also integrated in the analysis onboard observer data for the same fleet. They are considered as exact data for catches. Then, integrating these data allow to have direct information on Y_i and then to better estimate the underlying parameters (i.e. observation variance and zero-inflation parameter of commercial data).

Scientific data were extracted from the DATRAS database for the Orhago beam trawl survey (Gérard, 2003; ICES, 2018b). To align the commercial and the scientific data, we filtered scientific data based on the minimum size of sole (24 cm for sole - ICES (2018a)).

Furthermore, some parameters were hardly estimated (e.g. the range, the marginal variance parameter) and consequently we adopted a step estimation procedure for these parameters. We first fitted the model on Y_i^* to estimate these parameters. Then, we fitted the model on D_j while fixing the parameters hard to estimate with the estimates of the previously fitted model. Last, we let free the tricky parameters while fixing the initial values of these parameters with the estimates of the 1^{st} fitted model.

We compared the outputs of the scientific based model, of the integrated model based on scientific data and reallocated commercial data (Y_i^* model) and the integrated model based on catch declarations data (D_j model).

Results

Single square analysis

Reallocation has a major effect on predictions and estimates accuracy (Figure 2). Reallocating data conduct to a 10 to 200 times decrease of predictions accuracy when working at punctual observations level (MSPE gold compared to red boxplots) and decreases as the number of visited zones within a declaration increases. It also leads to the loss of the species habitat relationship; β_S is biased an tends towards 0 as the number of fishing zones within a declaration increases. Increasing the number of samples only increases the accuracy of the estimates and spatial predictions but overall spatial predictions remain inaccurate and

β_S remain biased. Regarding other parameters, the zero-inflation parameter is over-estimated (i.e. when reallocating, the quantity of data is under-estimated), the observation variance is underestimated (i.e. the data is estimated to be less noisy than they actually are) and the interecept of the latent field is slightly over-estimated (Figure 3).

Working at the level of catch declarations D_j allows to recover the species-habitat relationship and to improve the accuracy of spatial predictions even though accuracy is not as accurate as the golden-standard (Figure 2). Furthermore, the zero-inflation parameter is unbiased when the likelihood is built on catch declarations. Other parameters (observation variance, intercept) are also better estimated even though they remain slightly biased (Figure 3). This alternative model have some convergence difficulties (Table 1) as 8% of the model did not converged when sample size is medium (100 pings) and only 3% did not when sample size is large (1000 pings).

Multiple square analysis

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```

The scientific-based model provides species-habitat relationship and range estimates that are unbiased. Whether the model is built on Y_i or on D_j, the contribution of either scientific or commercial data can be clearly evidenced from the MSPE plot: the errors related to the integrated model are smaller than the single-data models. Comparing this to the maps obtained from Figure 5, we highlight that the commercial-based models logically miss the hotspot at 2.5°W - 46.75°N as commercial data do not cover this hotspot, while scientific data provide predictions that do not well capture the local hotposts. Integrating scientific and commercial data allows to (1) capture the hotspot missed by commercial data through scientific data and and (2) better capture the local correlation structures through the dense commercial data.

Furtermore, consistently with single-square simulations, working at the level of

punctual observations (Y_i) conduct to a loss in both the species-habitat relationship and predictions accuracy (Figure 4) compared to the model built on commercial declarations (D_j).

Interestingly, reallocating does not affect only the species-habitat relationship but also the spatial autocorrelation terms such as the range parameter. When working at the Y_i level, the range parameter is biased, while it is not when working at the D_j level.

This is a consequence of the loss of the species-habitat relationship: when working at Y_i level part of the variability related to the covariate effect is captured by the random effect and then the range parameter captures both autocorrelation related to the actual random effect and to the covariate.

Working at the level of catch declarations in estimation allows to recover and disentangle the effect of the species-habitat relationship and of the random effect. This is evidenced in Figure 5 where the model based on Y_i provide smoothed maps and do not capture the relatively small scale patterns that are shaped by the covariate effect. On the other hand the model built on D_j (and the scientific-based model too) better capture and disentangle the covariate effect and the spatial random effect and then provide predictions that better emphasize to the small-scale patterns of species distribution.

However, this goes with some difficulty in convergence as only 75% of the model built on catch declarations converge.

Real case study

Passing from punctual observations to catch declarations modifies the overall distribution patterns and some of the parameters estimates and related confidence intervals. Consistently with simulations, model built on Y_i tend to provide biased estimates and under-estimate the uncertainty related to other estimates compared with the scientific model and the model buit on D_j. Mainly, the effect of substrate is recovered in the integrated

model built on D_j (Figure 7). The zero-inflation parameter ξ is revised downwards (i.e. there are actually more zero-values than in the reallocated data) while the observation variance of commercial data is revised upwards (i.e. the commercial data are more noisy than expected when building likelihood on Y_i). In the model built on Y_i, the confidence interval of some parameters (specifically β_S, the marginal variance, the range, ξ_{com}, σ_{com}) are very narrow while in the model built on D_j confidence intervals are larger.

Furthermore, when comparing the scientific-based model estimates and the integrated model built on D_j, some parameters are better estimated. For instance, while in the scientific-based model the substrate effect was not significant, in the integrated model built on D_j substrate is significant and the confidence interval is smaller. Other parameters are more precisely estimated such as the marginal variance, the range, ξ_{sci} and σ_{sci}.

On the contrary, other parameters do not seems well estimated in either Y_i or D_j models. For instance, compared to the scientific-based model, the intercept is revised upwards when building the likelihood on Y_i and revised downards when working on D_j. This is consistent with simulations results, see Figure 3.

Regarding the maps of the spatial predictions, working at D_j level strongly modify the model biomass field compared with the Y_i model. Mainly, the effect of the covariate have a sharper effect on species distribution. Overall, the strength of the hotspots are revised and the map of the model fitted on D_j is less homogeneous than the map of the model built on Y_i.

Finally, the D_j model built on commercial data only does not converge (while the one built on Y_i does) emphasizing the model build on catch declarations face difficulties to converge and require punctual observations (here survey data and on-board observer data) to converge on real data. Interestingly, apart from the species-habitat relationship that are similar between the scientific-based model and the integrated model built on D_j, maps are not similar emphasizing commercial data brings information to the spatial predictions and better allow to capture local spatial patches of biomass.

Discussion

References

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

Gérard, B. (2003). ORHAGO. https://doi.org/10.18142/23

ICES. (2018a). *Report of the Working Group for the Bay of Biscay and the Iberian Waters Ecoregion (WGBIE)* (p. 642). Copenhagen, Denmark.

ICES. (2018b). *Report of the Working Group on Beam Trawl Surveys (WGBEAM)* (p. 121). Galway, Ireland.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, *70*(1), 1–21. https://doi.org/10.18637/jss.v070.i05

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*(9), 1369–1382.

Table 1

Percentage of convergence per simulation/model configuration at the level of a single rectangle

Nb samples	Fishing sequence	Reallocation	Likelihood level	Convergence (%)
10	1	No	Yi	99.668
10	1	Yes	Yi	0.333
10	1	Yes	Dj	0.000
100	10	No	Yi	100.000
100	10	Yes	Yi	100.000
100	10	Yes	Dj	92.000
1000	100	No	Yi	100.000
1000	100	Yes	Yi	100.000
1000	100	Yes	Dj	97.333

Table 2

Percentage of convergence per simulation/model configuration at the level of several ICES rectangles

Model	Likelihood level	Convergence (%)
Commercial model	Yi	100.000
Commercial model	Dj	75.377
Integrated model	Yi	100.000
Integrated model	Dj	76.382
Scientific model		100.000

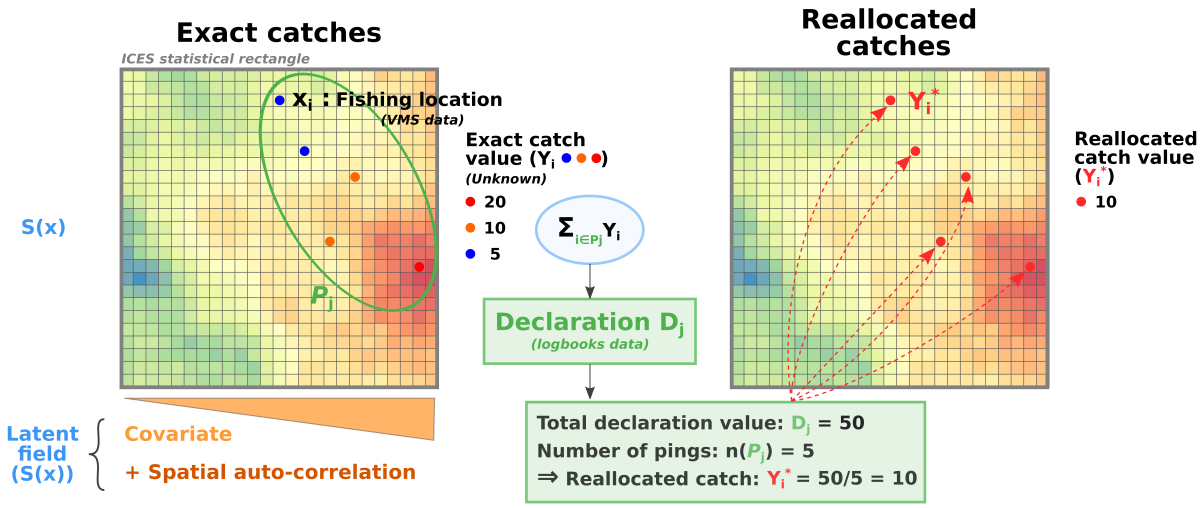

Figure 1. Reallocation process.

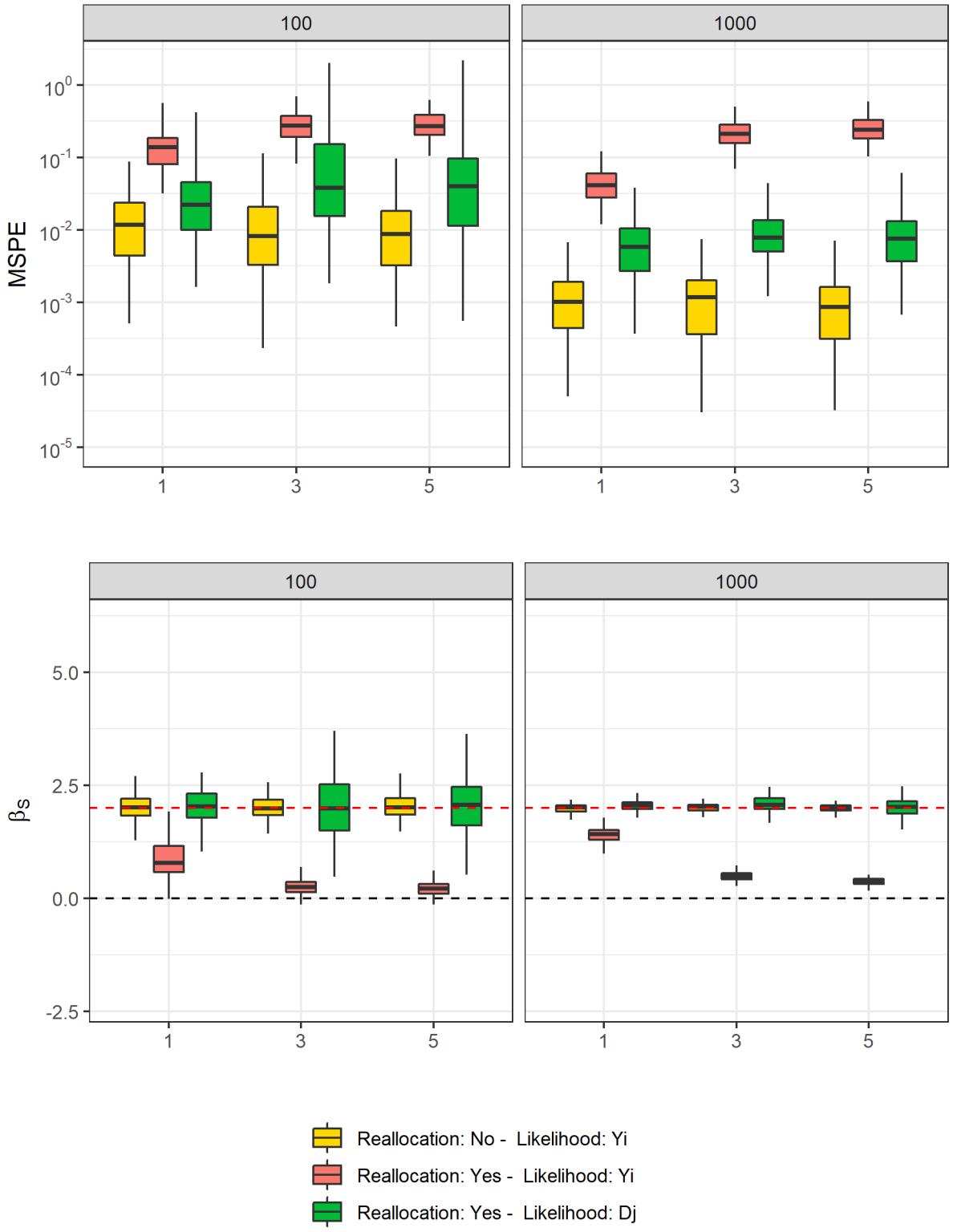

Figure 2. Performance metric for single-square simulations. columns: number of commercial pings. x-axis: number of zones visited within each declaration. "Reallocation:", data are or are not reallocated in simulations. "Likelihood:", the likelihood is computed on reallocated observations Y_i or on catch declarations D_j. Gold: golden standard. Red: actual situation (Y_i level). Green: alternative model (D_j level).

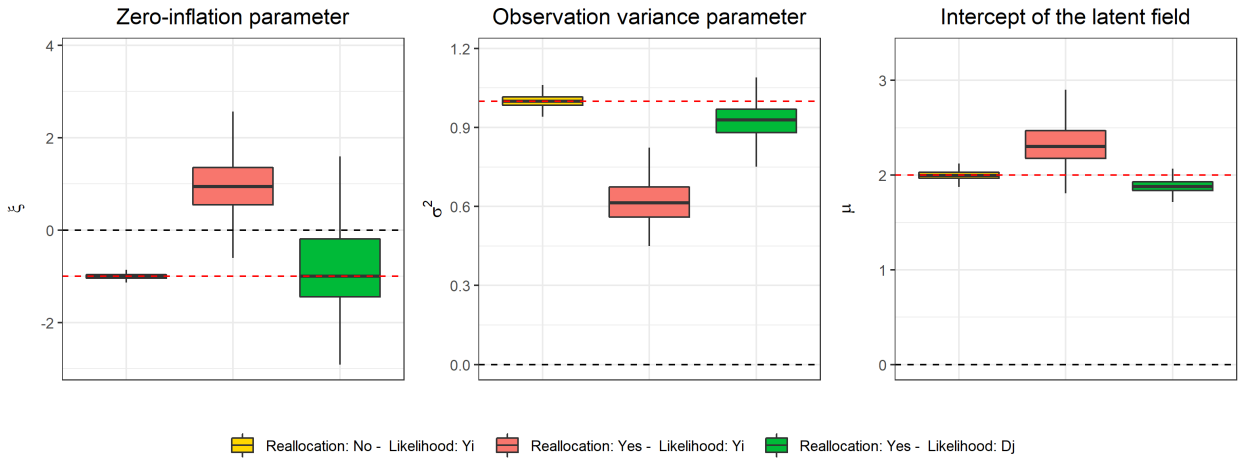

Figure 3. Parameters relative bias for single-square simulations. "Reallocation:", data are or are not reallocated in simulations. "Likelihood:", the likelihood is computed on reallocated observations Y_i or on catch declarations D_j. Gold: golden standard. Red: actual situation (Y_i level). Green: alternative model (D_j level). Only the simulations with 1000 observations are represented. Black line: zero value. Red line: parameter true value.

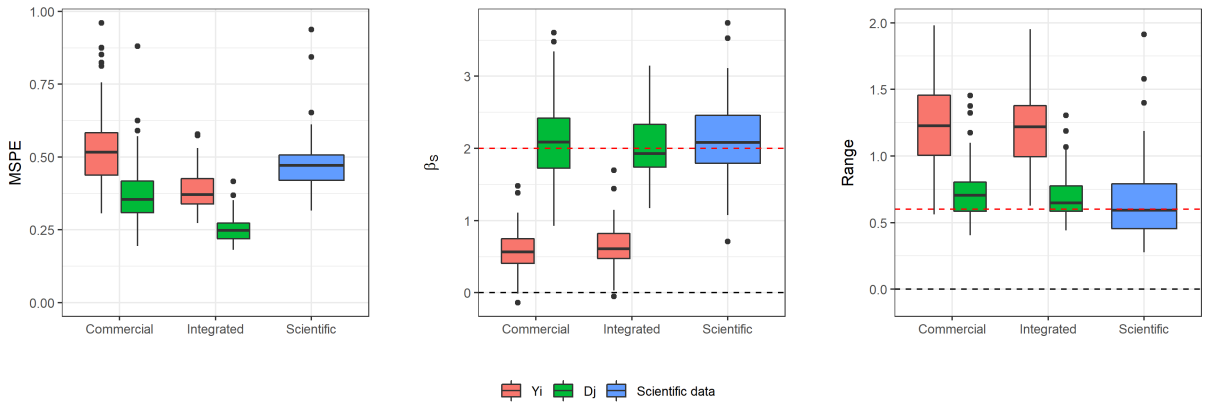

Figure 4. Performance metric for several rectangles simulations. columns: commercial data coverage. x-axis: likelihood level. 1^{st} row, red line: true value of β_S.

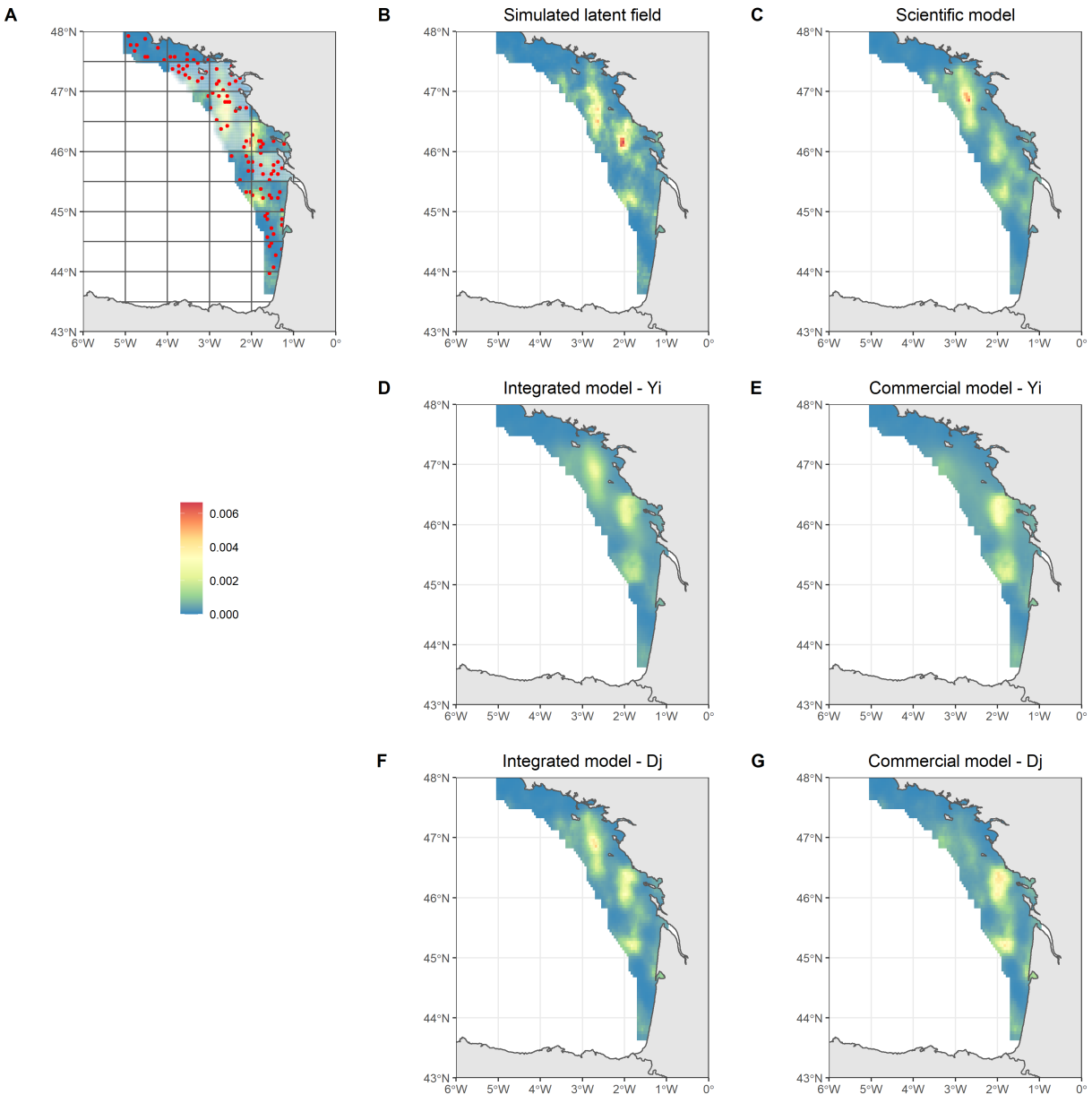

Figure 5. Relative distribution of simulated/estimated biomass field. A: Simulated biomass field with scientific samples (red) and rectangles that have not been sampled by commercial data (transparent rectangles). B: simulated biomass field. C: biomass field from the scientific-based model. Y_i: model fitted at the punctual observation level (D, E). D_j: model fitted at the declaration level (F, G).

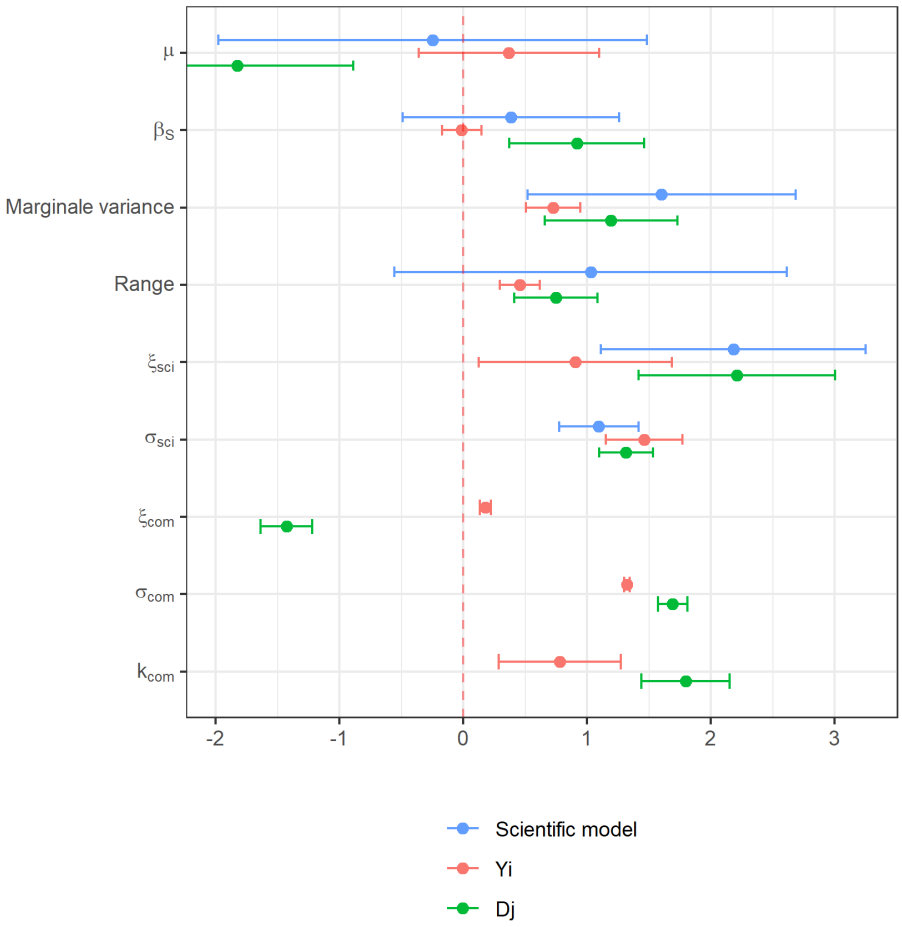

Figure 6. Parameters obtained with scientific-based model, the integrated model built on punctual observations Y_i and the integrated model built on catch declarations D_j.

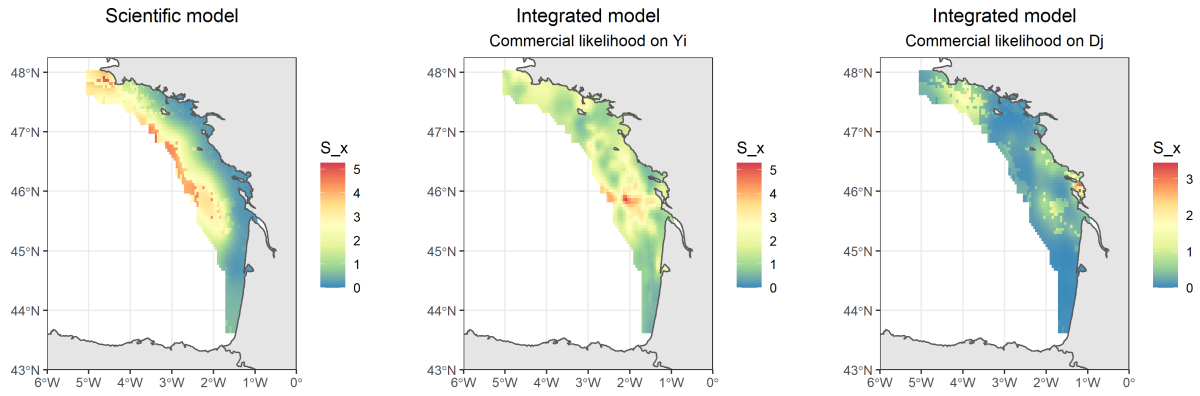

Figure 7. Maps obtained from (left) the scientific-based model, (center) the integrated model built on punctual observations Y_i, (right) the integrated model built on catch declarations D_j.