Going from coarse landings data to fine scale species distribution

Author1[1] & Author2[1,2]

[1] Instit1

[2] Instit2

Author Note

Test

The authors made the following contributions. Author1: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Author2: Writing - Review & Editing.

Correspondence concerning this article should be addressed to Author1, Postal address. E-mail: my@email.com

Abstract

Test

Keywords: keywords

Word count: X

Going from coarse landings data to fine scale species distribution

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). In this framework, all observations were supposed to be realised at the same scale. Here, we extend the framework to potentially provide fine scale predictions from data that have a coarse resolution (landings data) and potentially combine these data with other (more accurate) datasets such as scientific data.

Defining the problem from a modelling point of view

Definition of the variables

Let's define the latent field S and the punctual observations Y. S is a spatial log-Gaussian Field (GF) defined as $\log(S(x)) = \mu + \beta.\Gamma(x) + \delta(x)$ (Figure 1). We consider a discrete study domain where $x \in [\![1, n]\!]$ stands for the discrete locations. $\beta.\Gamma(x)$ is the covariate term with Γ the matrix of design of the covariates and β the effect of the covariates. $\delta(x)$ is a GF capturing spatial correlation ($\delta \sim \mathcal{N}(0, \Sigma)$).

All fishing locations x_i are supposed to be known through VMS data. At each fishing location x_i, fishermen realize a catch Y_i conditionnally on the latent field values (referred hereafter as a punctual catch). Observations are supposed to follow some distribution $\mathcal{L}_Y$.

$$Y_i|S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma^2)$$

In our case, observations are zero-inflated positive continuous data and they are modelled through a zero-inflated lognormal model previously introduced by Thorson (2018) and already used in Alglave et al. (2022). $\mathcal{L}_Y$ is decomposed it 2 parts:

- the probability to obtain a zero-catch which is modelled as a Bernoulli variable with probability $p_i = \exp(-e^{\xi}.S(x_i))$. ξ is a parameter controlling zero-inflation. When ξ increases, the amount of zero in the data decreases.

- if the catch is positive, the probability to obtain some observation value y_i is modelled through a positive continous distribution L (here a lognormal distribution - its parameterization is described in SM)) with mean component $\frac{S(x_i)}{1-p_i}$ and variance term σ^2.

$$\mathrm{P}\left(Y_i = y_i | x_i, S(x_i)\right) = \begin{cases} p_i & \text{if } y_i = 0 \\ (1-p_i) \cdot \mathrm{L}\left(y_i, \frac{S(x_i)}{(1-p_i)}, \sigma^2\right) & \text{if } y_i > 0 \end{cases}$$

Fishermen must declare what they catch at the level of the statistical rectangle at a daily time step. These data are recorded in logbooks data. A declaration (denoted D_j) is a summation of all the Y_i realized on the fishing positions belonging to the catch declaration D_j. They can be expressed as:

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

$\mathcal{P}_j$ is the vector of the fishing observations belonging to the declaration D_j. 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.

Catch reallocation: uniform reallocation or model-based reallocation

In standard processing, D_j (available from logbooks) are reallocated uniformly on related x_i (available from VMS) so that derived punctual observations Y_i^* are computed as $Y_i^* = D_j/m_j$. This is what we call uniform (or proportional) reallocation. In this case, inference of species distribution is directly computed through 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 D_j are observed while Y_i are not (they are latent variables). Such way, we define some distribution $\mathcal{L}_D$ as the distribution of the catch declarations:

$$D_j|\mathcal{P}_j, S_{\mathcal{P}_j} \sim \mathcal{L}_D(S_{\mathcal{P}_j}, \xi, \sigma^2)$$

with $S_{\mathcal{P}_j} = (S(x_1), ..., S(x_i), ..., S(x_{m_j}))$, ξ the zero-inflation parmeter and σ^2 the variance parameter. As $D_j = \sum_{i \in \mathcal{P}_j} Y_i$, it is possible to relate D_j, Y_i and $S_{\mathcal{P}_j}$ through the distribution $\mathcal{L}_Y$ and $\mathcal{L}_D$. Our approach is to match the moments of D_j (obtained from the moments of $\mathcal{L}_Y$) and $\mathcal{L}_D$ and then to infer $S(x)$ and Y_i from D_j. This is what we call model-based reallocation. This model will be referred as the D_j model. In contrast, the model fitted on reallocated observations Y_i^* will be referred as the Y_i^* model.

In the following, Y and D will be assumed conditionnal on the fishing positions and the latent field values, but for notation simplicity they will be simply denoted as Y_i or D_j instead of $Y_i|x_i, S(x_i)$ and $D_j|\mathcal{P}_j, S_{\mathcal{P}_j}$.

Matching $\mathcal{L}_Y$ and $\mathcal{L}_D$

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

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

$$P(D_j = 0) = \prod_{i \in \mathcal{P}_j} P(Y_i = 0),$$
$$= \exp\left\{ - \sum_{i \in \mathcal{P}_j} e^{\xi}.S(x_i) \right\} = \pi_j.$$

(2) If the declaration is positive, we assume the probability to obtain catch d_j follows a continous distribution $\mathcal{L}_{D_j|D_j>0}$ with mean component μ_j and variance component σ_j. From $\mathcal{L}_Y$ we can derive $E(D_j|D_j > 0)$ and $Var(D_j|D_j > 0)$ (given below) and re-express μ_j and σ_j as a function of these two moments.

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

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

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

If we consider $\mathcal{L}_{D_j|D_j>0}$ is lognormal (with parameterization described in SM) then, $\mu_j = E(D|D_j > 0)$ and $\sigma_j = ln(\frac{Var(D|D_j>0)}{E(D|D_j>0)^2} + 1)$. Such way, by fitting the model on declaration D_j rather than reallocated catch Y_i and linking $\mathcal{L}_D$ to $\mathcal{L}_Y$, we intend to better follow the observation process of the commercial data and potentially improve inference compared with the approximation realized through uniform reallocation. All calculations to obtain these formulas are available in SM.

The inference 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).

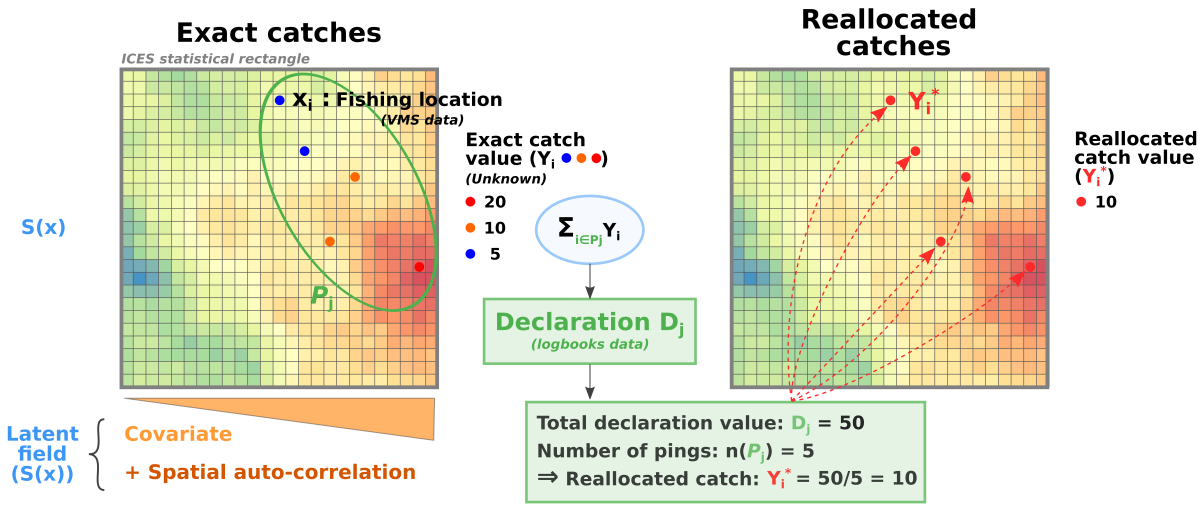

Figure 1. Schematic representation of the reallocation process.

Simulation-estimation

To evaluate the impact of reallocation on model outputs and how our alternative approach can improve model accuracy, we conduct simulation-estimation.

First, we conduct basic simulations in order to explore the effect of some key variables on model performance and also the base properties of the alternative models. We assume the study domain is a single statistical rectangle, only commercial data feed the model and the expression of the latent field in both simulation and estimation is simplified (the variability only depends on one covariate, not on a spatial random term). These simulations will be referred as "single-square simulations".

Then, we extend the simulation-estimation study to several rectangles to get closer to a real case study configuration. We add a spatial random effect in the latent field (for both simulation and estimation) and we also simulate scientific data (in addition to commercial data) to explore the contribution of both datasets in inference. These simulations will be referred as multiple-square simulations.

In these two sets of simulation-estimation, 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 $\mu = 2$. Regarding commercial data, the number of fishing pings per declaration is fixed to 10 as it is the average number of fishing locations for a single declaration in real data. All parameterizations are detailed in the Table 1.

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 (n is the number of locations over the grid).

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

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

Single-square simulations

Two important variables may affect the accuracy of model outputs: the amount of commercial data and the number of fishing zones explored within one single declaration. In these simulations-estimations, we intend to explore the effect of these 2 variables.

First, the amount of commercial data; it is expected that when increasing the amount of data the accuracy of the predictions will be improved. In simulations, we progressively increase the number of fishing positions (10, 100 and 1000 fishing pings). As there are 10 pings per declaration, this means that the number of declarations increases respectively (1, 10 and 100 declarations).

Furthermore, the samples belonging to a single declaration can be sampled in a single restricted zone or can mix catches from distinct zones (within a unique statistical rectangle). Note that we make a distinction between fishing zones and statistical rectangles: fishing

Table 1

Parameter values for the simulations

Parameters	Single-square simulations	Multiple-square simulations
μ	2	2
β_S	2	2
Range	–	1.5
Marginal variance	–	0.5
ξ_{com}	-1	-1
σ_{com}	1	1
k_{com}	–	1
ξ_{sci}	–	0
σ_{sci}	–	0.8

zones are included in a statistical rectangle and there can be several fishing zones within a statistical rectangles. Consequently, a declaration can mix catches that have been realized in zones of high biomass and in zones of low biomass. For instance, a declaration can aggregate data from several fishing operations that occured the same day within the same statistical rectangle, but that were realized on 2 distinct types of fishing grounds (and then on two distinct habitats). Reallocating uniformly the declaration will strongly homogeneise the actual catch, will mix up information from to different grounds and then may lead to a strong loss of accuracy in model outputs. It is expected that when the number of fishing zones within a declaration increases, the accuracy of the model outputs decreases. To assess the effect of such process, we simulated the pings of a fishing declaration assuming they were either realized in a single zone, in 3 distinct zones or in 5 distinct zones.

Concretely, for each fishing declaration the fishing pings are sampled in 2 steps: first the centroid of the zones are randomly sampled over the statistical rectangle (here the

simulation domain) and then the fishing positions are randomly sampled within the radius of the centroid of the zones. The zone size was chosen to be a square of side 7. This is a realistic order of magnitude if we consider each zone is one fishing operation and that for each fishing operation the distance of the operation will not exceed 30km.

At each fishing position, a catch is realized conditionnally on the value of the latent field and follow the probability distribution described above ($\mathcal{L}_y$).

We compare three simulation/estimation configurations:

- a golden standard configuration: punctual observations (Y_i) are known exactly and the punctual-data model is fitted to the exact data.
- a configuration corresponding to the actual situation (uniform reallocation): simulated catches are summed into declarations and reallocated over the related fishing locations. We then fit the Y_i^* model (the model fitted to punctual reallocated data).
- a configuration corresponding to our alternative approach: simulated catch are reallocated over fishing locations and we fit the D_j model to the data (i.e. the one fitted at the level of the declarations data).

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

Multiple-square simulations

In this case, we extend the study to more than one single rectangle, we simulate scientific data and include these in inference. Finally, in addition to the covariate, we also simulate/estimate a spatial random effect in the latent field as a Matérn GRF.

The study area is based on the case study; it includes the whole coast of the Bay of Biscay and cover several statistical rectangles (Figure 5A). To tailor the case study, we

simulate 3000 of fishing positions regrouped in 300 declarations (10 pings per declaration). Commercial data may not cover the full area, and consequently we allow the commercial samples to cover only 2/3 ot the area. Similarly to the single-square simulations, the sampling of the commercial points of a declaration are realized in 3 steps. (1) The declaration is randomly affected to one of the ICES rectangle. (2) The centroid of the fishing zone is randomly sampled within this statistical rectangle. (3) The 10 fishing punctual observations are randomly sampled within the fishing zone. The radius of the fishing zone is set so as the extent of a fishing operation does not exceed 30 km. Note that we do not explore the effect of exploring several zones within the same declaration as it is already done in the single-square simulations.

100 scientific samples are simulated following a random stratified plan; contrarly to commercial data they cover the entire study domain (Figure 5A). Scientific observations are simulated following the observation equations of $\mathcal{L}_Y$ (with specific parameters for scientific data - Table 1).

We compare several model configurations:

- to assess what brings our alternative approach, we compare the Y_i^* model and the D_j model.

- 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 beginning of the section ($MSPE$ and species-habitat parameter β_S), we also compare the simulated/estimated values of the range parameter.

Note that when including 2 datasets in the model, a difference in catchability should be accounted for and one of the two dataset will be considered as reference (see Alglave et al. (2022) for more details on parameterization). Here the reference level is set with scientific data and the parameter k_{com} accounts for the relative ratio between scientific and commercial observation values. In simulations, k_{com} is set to 1 so that both datasources are assumed to have the same catchability.

Case-study: sole of the Bay of Biscay

To illustrate our method on a real case study, we applied the approach to the common sole of the Bay of Biscay. VMS-logbooks data were extracted for the bottom trawlers fleet (OTB). 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. 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)). To illustrate the method, we compare the outputs of (1) the scientific-based model, (2) the integrated model fitted to scientific data and reallocated commercial data (Y_i^* model) and (3) the integrated model fitted to scientific and declarations data (D_j model).

The D_j model faced convergence issues (some of the parameters were hardly estimated e.g. the range parameter). To ease convergence, we integrated in the analysis onboard observer data for the same fleet. They are considered as exact commercial data of catches. Integrating these data allow to have direct information on Y_i and to better estimate the observation equations parameters (i.e. observation variance and zero-inflation parameter of commercial data).

Furthermore, we adopt a step estimation procedure to initialize the optimization algorithm for the D_j model. We first fit the Y_i^* model and use the estimates of the model to

initialize the optimization of the D_j model. We eventually fix the parameters that are hard to estimate in first steps of optimization (intercept μ, covariate effect β_S, range and marginal variance) and let them free iteratively.

Results

Single-square simulations

Reallocation has a major effect on predictions and estimates accuracy (Figure 2). Reallocating data conduct to a 10 to 200 times decrease in accuracy for spatial predictions when fitting the model to reallocated catch (MSPE gold compared to red boxplots). Accuracy decreases as the number of visited zones within a declaration increases. Besides, β_S estimates are biased and reallocation leads to the loss of the species-habitat relationship as the number of fishing zones (within a declaration) increases (β_S estimates get closer to 0). Increasing the number of samples does not improve inference. Regarding other parameters estimates, the zero-inflation parameter (ξ) is over-estimated (i.e. when uniformly reallocating commercial data, the quantity of positive observations is under-estimated), the observation variance (σ) is underestimated (i.e. the data is estimated to be less noisy than they actually are) and the intercept of the latent field (μ) is slightly over-estimated (Figure 3).

The D_j model allows to recover the species-habitat relationship and to improve the accuracy of the spatial predictions (Figure 2). Still, the D_j model outputs are not as accurate as the ones of the golden-standard. Furthermore, the zero-inflation parameter is unbiased when the model is fitted to catch declarations D_j. 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 2) 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).

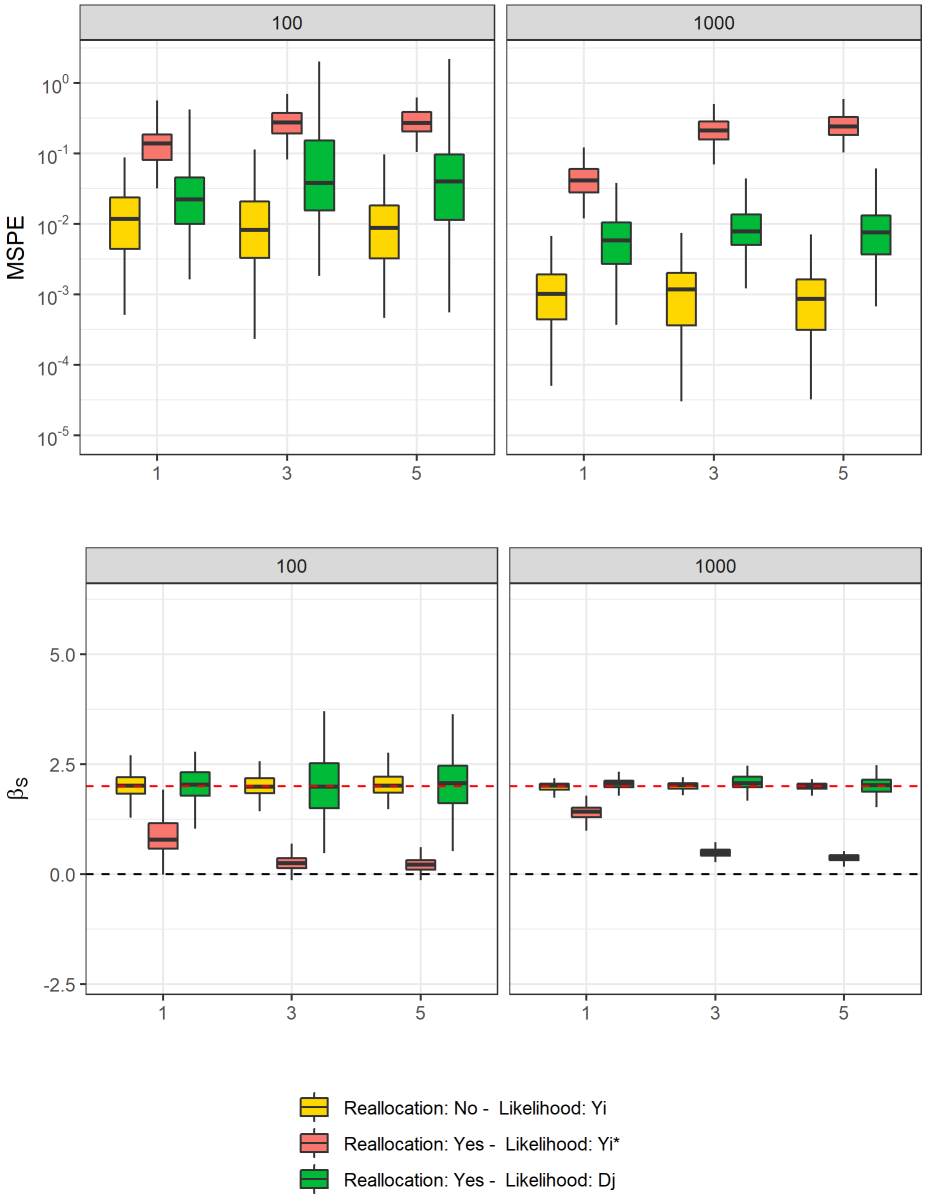

Figure 2. Performance metric for single-square simulations. Columns: number of commercial fishing positions. x-axis: number of zones visited within each declaration. "Reallocation:", data are or are not reallocated in simulations. "Likelihood:", the likelihood is computed on exact punctual observations Y_i, on reallocated observations Y_i^* or on catch declarations D_j. Gold: golden standard. Red: uniform reallocation (Y_i^* model). Green: model-based reallocation (D_j model). Simulations conducted on 10 fishing positions are not represented as they mostly did not converged for the D_j model (see Table 2).

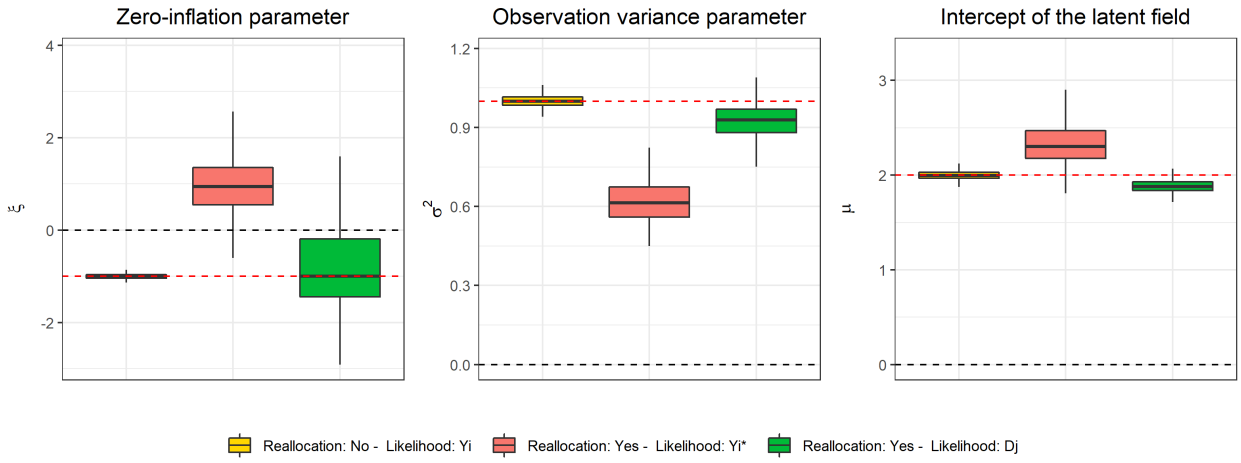

Figure 3. Parameters relative bias for single-square simulations. "Reallocation:", data are or are not reallocated in simulations. "Likelihood:", the likelihood is computed on exact punctual observations Y_i, on reallocated observations Y_i^* or on catch declarations D_j. Gold: golden standard. Red: uniform reallocation (Y_i^* model). Green: model-based reallocation (D_j model). Only the simulations with 1000 fishing positionss are represented. Black line: zero value. Red line: parameter true value.

Multiple-square simulations

Whether the model is fitted at the punctual level or at the declaration level, 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. This can be well illustrated from Figure 5. Integrating scientific and commercial data allows to (1) capture the hotspot missed by commercial data through scientific data and (2) better capture the local correlation structures through the dense commercial data.

Furthermore, consistently with single-square simulations, uniform reallocation conducts to a loss in both the predictions accuracy and the species-habitat relationship (Figure 4) compared to the model built on commercial declarations (D_j).

Interestingly, in addition to the species-habitat relationship, uniform reallocation also

Table 2

Single-square simulations - Percentage of convergence per simulation-estimation configuration.

Fishing positions	Declarations	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

affects the spatial autocorrelation terms such as the range parameter. The model fitted on reallocated data (Y_i^*) provides biased range estimates while the D_j model provides unbiased estimates. This is a consequence of the loss of the species-habitat relationship: when uniformly reallocating declarations, part of the variability related to the covariate effect is captured by the random effect. Consequently, the range parameter captures both the autocorrelation related to the actual random effect and to the covariate. The D_j model 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 fitted to reallocated catch Y_i^* provides smoothed maps and does not capture the relatively small scale patterns that are shaped by the covariate. On the other hand, the D_j model (as the scientific-based model) better captures and disentangles the covariate effect and the spatial random effect and then provides predictions that better fit to the small-scale patterns of the species distribution.

However, this goes with some difficulty in convergence as only 75% of the model built on catch declarations converge (Table 3).

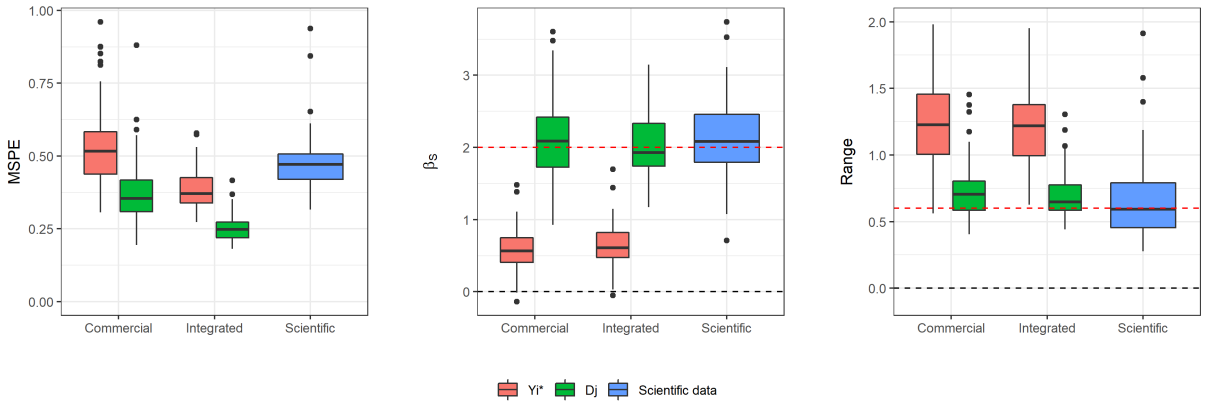

Figure 4. Performance metric for the multiple-square simulations. Red line: true value for the range and the species-habitat parameter (β_S). Red: uniform reallocation (Y_i^* model). Green: model-based reallocation (D_j model). Blue: scientific-based model.

Real case study

Consistently with simulations, the D_j model modifies some of the parameters estimates and revises the spatial pattern of the species distribution compared with the Y_i^* model. In particular, the substrate effect is recovered in the D_j model and fall in the same range as the scientific-based estimate (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 noisier than expected when fitting the model on Y_i^*).

In addition, uncertainty is also revised when fitting the model at the declaration level. For instance, when comparing the Y_i^* model to the D_j model, the confidence intervals of β_S, the marginal variance, the range, ξ_{com}, σ_{com} are much wider. This emphasizes that uncertainty is probably underestimated in the Y_i^* model compared with the D_j model. Now, when comparing the scientific-based model and the integrated D_j model, some parameters

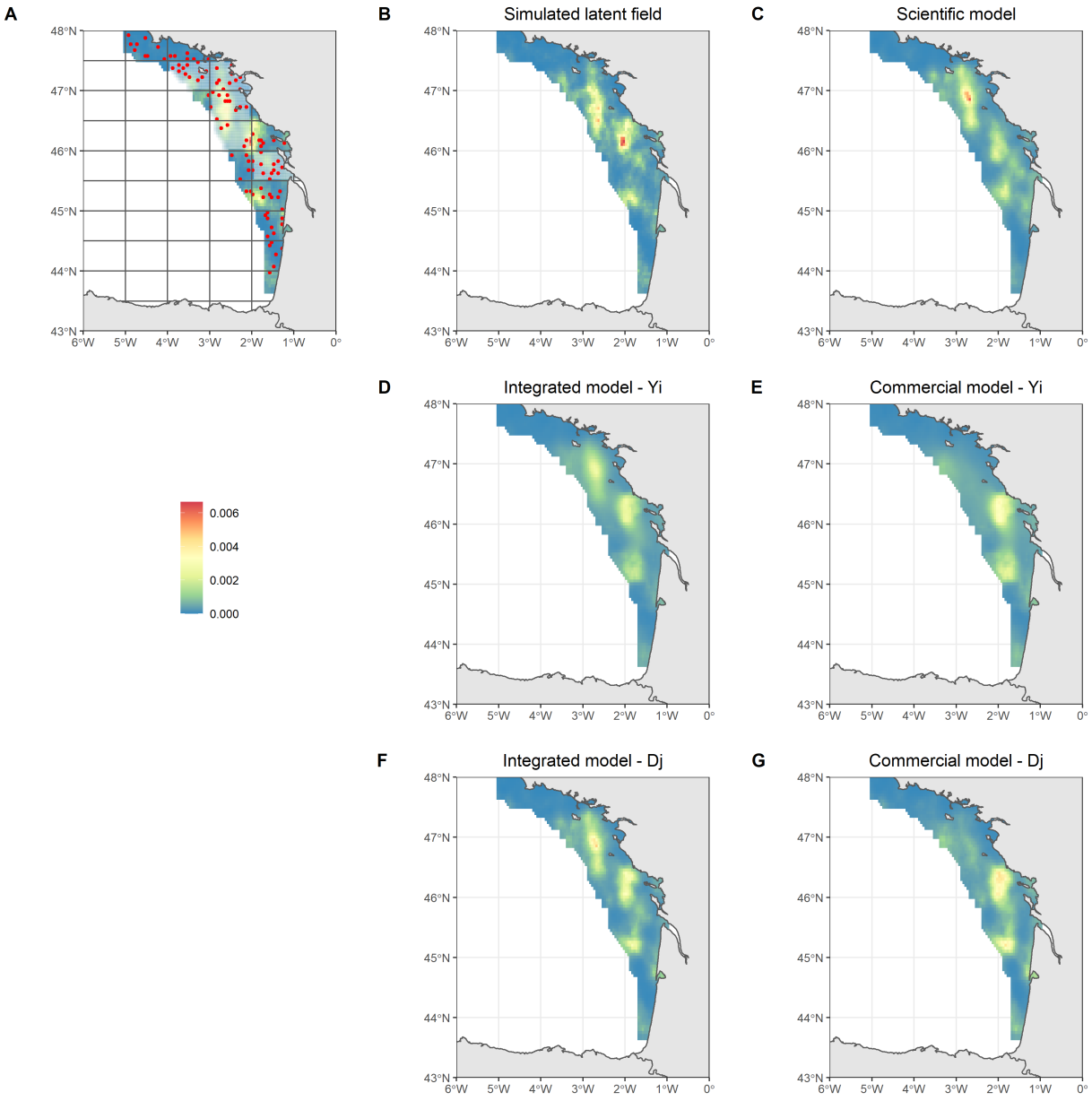

Figure 5. Relative distribution of simulated/estimated biomass field. A: Simulated biomass field with scientific samples (red) and statistical rectangles. The rectangles that have not been sampled by commercial data are the transparent rectangles. They represent 1/3 of the full area. B: simulated biomass field. C: biomass field from the scientific-based model. Y_i^*: Y_i^* model (D, E). D_j: D_j model (F, G). Scientific model: model fitted to scientific data only. Commercial model: model fitted to commercial data only. Integratedd model: model fitted to both data sources.

Table 3

Multiple-square simulations - Percentage of convergence per simulation-estimation configuration.

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

are better estimated than when only scientific data feed the model. 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.

On the contrary, other parameters do not seems be well estimated in either the Y_i^* or the 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 species distribution, fitting the model at the D_j level strongly modifies the model biomass field compared with the Y_i^* model. In particular, the substrate covariate have a sharper effect on species distribution and the intensity of the hotspots are revised when fitting the model on D_j.

Finally, the D_j model fitted only on commercial data does not converge (while the one fitted on Y_i^* does) emphasizing the model fitted on catch declarations face difficulties to converge and require punctual observations (here survey data and on-board observer data) to converge on real data.

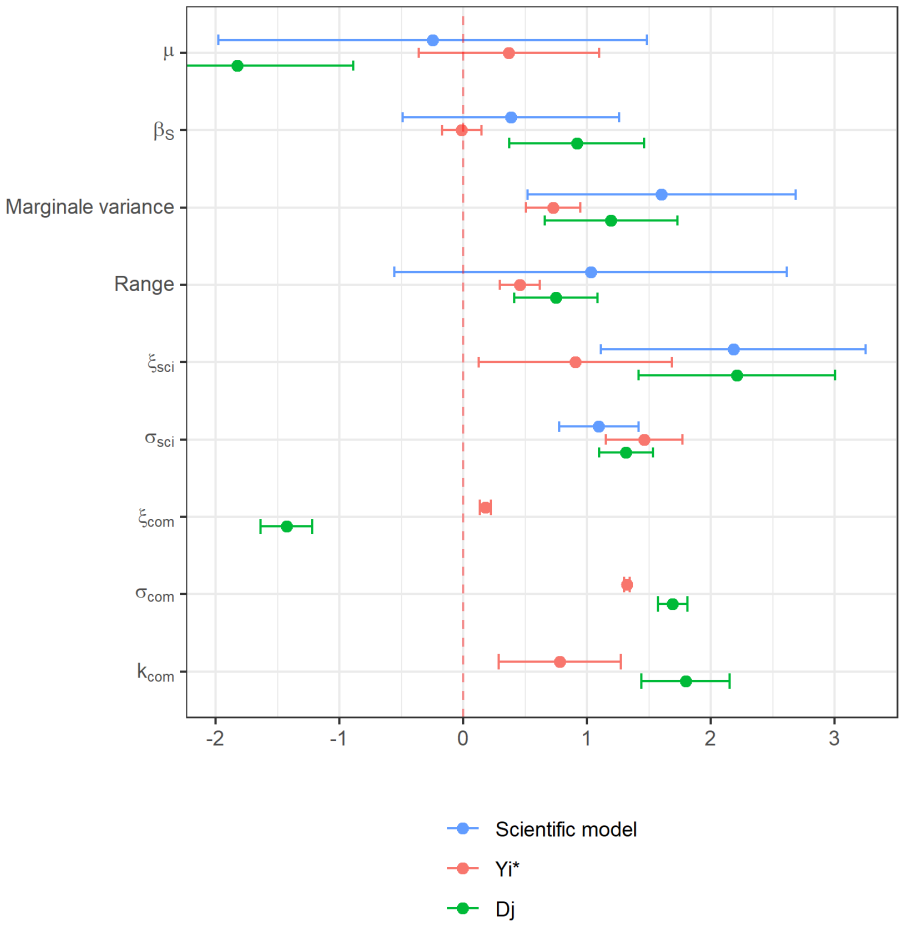

Figure 6. Parameters obtained with scientific-based model, the integrated model fitted on reallocated catch Y_i^* and the integrated model fitted on catch declarations D_j.

Discussion

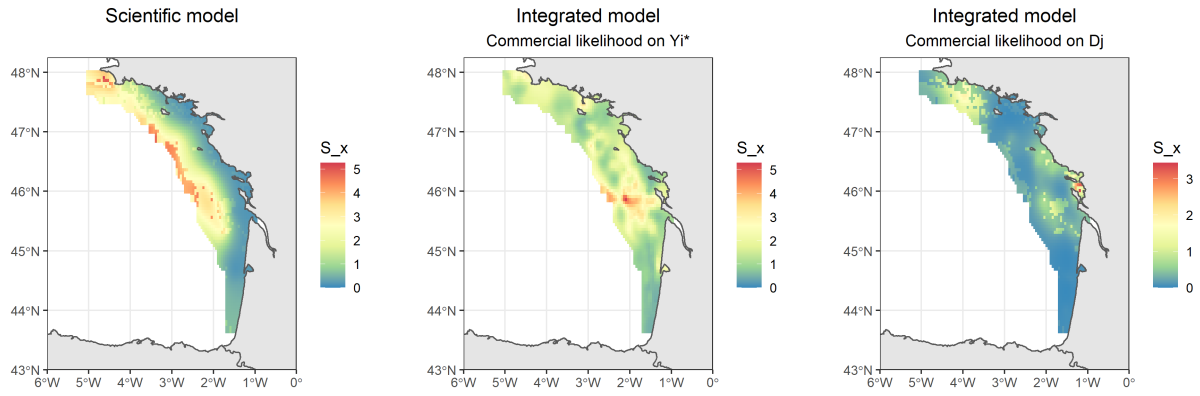

Figure 7. Maps obtained from (left) the scientific-based model, (center) the integrated model fitted on punctual reallocated catch Y_i^*, (right) the integrated model fitted on catch declarations D_j.

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.

Supplementary material

Notations

We model catch declarations D_j (available at coarse resolution through logbooks data) as a sum of Y_i punctual observations (which are unknown i.e. latent variables) each one realised at one fishing position x_i (known through VMS data).

We note:

- $\mathcal{P}_j$: the vector of all fishing positions related to the j^{th} declaration.
- $j \in \{1, ..., l\}$ with l the number of declarations.
- $i \in \{1, ..., m_j\}$ with m_j the number of fishing positions belonging to the j^{th} declaration.

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

Reparameterization of the lognormal distribution

The lognormal distribution is usually written as $Z \sim \mathrm{L}(\rho; \sigma^2)$ with $Z = e^{\rho + \sigma N}$ and $N \sim \mathcal{N}(0, 1)$. In this case, $E(Z) = e^{\rho + \frac{\sigma^2}{2}}$ and $Var(Z) = (e^{\sigma^2} - 1)e^{2\rho + \sigma^2}$.

We choose to slightly reparameterize the lognormal distribution. Let's define $\rho = ln(\mu) - \frac{sigma^2}{2}$, then:

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

D_j probability distribution and moments

We have to express the probability distribution of D_j and its moments as a function of Y_i and its related moments. Let's assume $Y_i = C_i.Z_i$ is a zero-inflated lognormal distribution with C_i a binary random variable and Z_i a lognormal random variable.

$$C_i \sim \mathcal{B}(1 - p_i)$$

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

$$Z_i \sim \mathrm{L}(\frac{S(x_i)}{1 - p_i}, \sigma^2)$$

Here, Y_i, C_i and Z_i are observations of a latent field $S(x_i)$ at a sampled point x_i.

Probability of obtaining a zero declaration

$$\begin{aligned} P(D_j = 0|S, X) &= \prod_{i \in \mathcal{P}_j} P(Y_i = 0|S, X), \\ &= \exp\left\{ - \sum_{i \in \mathcal{P}_j} e^{\xi}.S(x_i) \right\} = \pi_j. \end{aligned}$$

Expectancy of a positive declaration

Following calculations are supposed to be conditionnal on S and X.

$$E(D_j|D_j > 0) = \sum_{i \in \mathcal{P}_j} E(C_i Z_i | \exists i \in \mathcal{P}_j, C_i = 1)$$

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

As $E(D_j 1_{\{D_j > 0\}}) = E(D_j)$,

$$\begin{aligned} E(D_j|D_j>0) &= (1-\pi_j)^{-1}\, E(D_j), \\ &= (1-\pi_j)^{-1} \sum_{i\in\mathcal{P}_j} E(C_i Z_i), \\ &= (1-\pi_j)^{-1} \sum_{i\in\mathcal{P}_j} (1-p_i)\frac{S(x_i)}{1-p_i}, \\ &= (1-\pi_j)^{-1} \sum_{i\in\mathcal{P}_j} S(x_i). \end{aligned}$$

338 **Variance of a positive declaration**

$$Var(D_j|D_j>0) = E(D_j^2|D_j>0) - E(D_j|D_j>0)^2.$$

$$E(D_j^2|D_j>0) = (1-\pi_j)^{-1}E(D_j^2 1_{\{D_j>0\}}) = (1-\pi_j)^{-1}E(D_j^2)$$

$$E(D_j|D_j>0)^2 = ((1-\pi_j)^{-1}E(D_j 1_{\{D_j>0\}}))^2 = (1-\pi_j)^{-2}E(D_j)^2$$

339 Then,

$$Var(D_j|D_j>0) = (1-\pi_j)^{-1}E(D_j^2) - (1-\pi_j)^{-2}E(D_j)^2 = (1-\pi_j)^{-1}Var(D_j) - \frac{\pi_j}{(1-\pi_j)^2}E(D_j)^2.$$

340 As the $(Y_i)_{i\in\mathcal{P}_j}$ are independent, $Var(D_j) = \sum_{i\in\mathcal{P}_j} Var(Y_i) = \sum_{i\in\mathcal{P}_j} Var(C_i.Z_i)$.

$$
\begin{aligned}
Var(C_i Z_i) &= E(C_i^2 Z_i^2) - E(C_i Z_i)^2, \\
&= E(C_i^2) E(Z_i^2) - E(C_i)^2 E(Z_i)^2, \\
&= (1-p_i) E(Z_i^2) - (1-p_i)^2 E(Z_i)^2, \\
&= (1-p_i)(Var(Z_i) + E(Z_i)^2) - (1-p_i)^2 E(Z_i)^2, \\
&= \frac{S(x_i)^2}{1-p_i}(e^{\sigma^2} - 1) + \frac{S(x_i)^2}{1-p_i} - S(x_i)^2, \\
&= \frac{S(x_i)^2}{1-p_i}(e^{\sigma^2} - (1-p_i))
\end{aligned}
$$

Sum up of the main formulas

$$P(D_j = 0|S, X) = \exp\left\{ - \sum_{i \in \mathcal{P}_j} e^{\xi}.S(x_i) \right\} = \pi_j$$

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

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

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

Assuming $D_j|D_j > 0$ also follows a lognormal distribution we can write:

$$D_j|D_j > 0 \sim \mathrm{L}(\mu_j = E(D_j|D_j > 0), \sigma_j^2 = ln(\frac{Var(D_j|D_j > 0)}{E(D_j|D_j > 0)^2} + 1))$$