Inferring fine scale wild species distribution from spatially aggregated data

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

Inferring fine scale wild species distribution from spatially aggregated data

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

[1] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Nantes, France

[2] Institute for Aquatic Resources, Section for Marine Living Resources, Technical University of Denmark, Kemitorvet, Kongens Lyngby, Denmark

[3] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Brest, France

[4] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Nantes, France

[5] Mathematical Research Institute of Rennes IRMAR, Rennes University, Rennes, France

[*] To whom correspondence should be addressed; E-mail: baptiste.alglave@agrocampus-ouest.fr.

Abstract

In spatial ecology, huge amount of aggregated spatial data such as hunting data or fishermen declarations data offer possibilities to map wild species distribution at fine scale by combining them with high resolution data. However, this requires to properly handle the difference in spatial resolution between the different data sources. Such issue is often referred as the change of support (COS) problem. In ecological applications, accounting for COS can be challenging as the observations data can be complex (e.g. zero-inflated positive continuous data) and this can complicate the way COS is handled. In this paper, we develop a hierarchical approach that allows (1) to handle COS for a mixture of zero-inflated positive continuous data and (2) to combine fine scale data and aggregated data. We develop and apply the approach based on a fishery application where fishermen declarations data are registered at rough scale, but are used in combination with scientific survey data that are exactly geolocalized to infer fine scale species distribution. We compare (1) a rough but standard way to refine the resolution of declarations by proportionally reallocating declarations on fishing locations and (2) a model that handle COS by explicitly modeling the probability distribution of the aggregated declaration conditionally upon the exact locations observation. The rough approach leads to a loss of the species-habitat relationship, to smoothed maps of species distribution and to an overweighted contribution of declarations data into inference in comparison with scientific data. By contrast, the COS

approach allows to provide unbiased estimates of the habitat effect and more accurate spatial predictions. Furthermore, scientific data contributes in a more significant way to inference. This approach is a valuable contribution for a wider use of spatially aggregated data in spatial ecology in order to properly integrate datasets that do not have the same spatial resolution to make fine scale inferences of species distribution.

Introduction

Context

With the progress of new technologies, spatial ecological data are becoming more and more accessible every day thanks to the huge effort of the scientific community to generate and get access to intensive information for ecology, evolution and conservation (Nathan et al. 2022; Hampton et al. 2013; Grémillet, Chevallier, and Guinet 2022). These data are crucial to face the current challenges related to large- and small-scale ecological questions: for instance, following animal movement (Nathan et al. 2022), mapping species distribution (Isaac et al. 2020) or tracking climate change (Maureaud et al. 2020).

These data sources are often highly heterogeneous in size, type and sampling design, making their combination a methodological challenge (Fletcher et al. 2019; Isaac et al. 2020; Miller et al. 2019; Pacifici, Reich, Miller, Gardner, et al. 2017; Renner, Louvrier, and Gimenez 2019). For instance, in species distribution modeling, recent studies have investigated how to combine scientific standardized data with auxiliary data such as citizen science data (Fletcher et al. 2019). Typically, count data from planned surveys on birds communities can be combined with other counts data coming from citizen science programs (e.g. eBird program - Sullivan et al. (2014)). These first ones benefit from a standardized protocol, a controlled sampling plan and they are designed to cover the full range of species distribution. The second ones provide a larger amount of data with lower cost, but they arise from non-standardized sampling and consequently they may not cover the whole area. Integrating these data sources typically allows to benefit from the good coverage of the survey while improving spatial prediction accuracy through the massive amount of data available through citizen science programs.

Another massive source of information are declaration data (we refer to declaration

data as the mandatory data that must be reported by some agent as a legal requirement to proceed with his activity). As they are mandatory, declaration data are usually very large datasets (much larger than scientific or citizen science datasets). They can proove highly valuable to map wildlife species distribution. In fisheries science, a common example of such data sources are commercial catch declaration data. They can be used to map fish distribution and provide valuable information to identify spawning areas or nursery grounds Alglave et al. (2022) and Azevedo and Silva (2020).

Although massive, these data are most often registered at the scale of coarse spatial units while scientific survey and citizen science data are usually reported with their exact locations. Generally, these administrative units do not have a resolution that is relevant for ecological analysis (Pacifici, Reich, Miller, and Pease 2019).

Developing statistical methods that properly handle spatially aggregated data and integrate these with higher resolution data is then a major challenge to make precise and unbiased inference of species distribution at a fine scale.

The change of support issue

Inferring fine-scale spatial processes from coarse data and reconciling spatial scales properly when different set of observations do not have the same resolution is a well known issue in geography, ecology, agriculture, geology and statistics (Gotway and Young 2002). In the statistical literature, *Change of Support* (COS) refers to 'the summary or analysis of spatial data at a scale different from that at which it was originally collected' (Gotway and Young 2002; Gelfand 2010). It is often also reffered as 'downscaling/upscaling' or Modifiable areal unit problem (MAUP) in the literature (Wikle, Zammit-Mangion, and Cressie 2019). This is typically the case where data are aggregated over larger geographical scales, but one would like to infer processes at a different resolution. In such case,

conclusions from a fine-resolution analysis can strongly differ from an analysis at a coarser scale based on the aggregation of the fine-resolution data. Such phenomena is also called the ecological fallacy (Wakefield and Lyons 2010).

Since 2000, several studies have described how COS issues could be overcome; Mugglin, Carlin, and Gelfand (2000), Gelfand, Zhu, and Carlin (2001), Gotway and Young (2007) and Wikle and Berliner (2005) proposed generic approaches (and extensions of these approaches - Kim and Berliner (2016)) for addressing COS in a spatial or spatio-temporal context. In health analysis, Young and Gotway (2007) proposed to compare some rough approach based on centroids of areal units to relate environmental and health outcomes with an approach that honors the spatial support of the data (size, shape, orientation). Berrocal, Gelfand, and Holland (2010a) and Berrocal, Gelfand, and Holland 2010b proposed a spatio-temporal method for fusing several air pollution data: one from coarse resolution but with full spatial coverage and another recorded at point level, with sparse distribution but where records almost corresponds to the true value of the process. In climate science, Reich, Chang, and Foley (2014) and Parker, Reich, and Sain (2015) proposed a spectral statistical approach to downscale information from large-scale model to lower scale. In the field of ecology, some recent studies have tackled such issues: Finley, Banerjee, and Cook 2014 provided a framework for integrating spatially misaligned data, Hefley, Brost, and Hooten (2017) proposed a solution based on COS to account for location error in presence-only data, Pacifici, Reich, Miller, and Pease 2019 introduced a framework for integrating data sources of different resolution to map species distribution. Applying similar ideas, Gilbert et al. (2021) integrated harvest data (aggregated data) and camera trap (precisely geolocalized data) to map several wildlife species in Wisconsin.

Focus of the paper

One of the main challenge limiting the number of application consists in the type of observation data that can be fitted to the existing COS framework. Indeed, the frameworks that were developed so far and their related applications mainly limited their scope to relatively simple observation data: count data were modeled through Poisson processes (Gilbert et al. 2021; Gotway and Young 2007; Mugglin, Carlin, and Gelfand 2000; Pacifici, Reich, Miller, and Pease 2019) and continuous data were modeled through Gaussian or Gamma distributions (Berrocal, Gelfand, and Holland 2010a; Gelfand, Zhu, and Carlin 2001; Wikle and Berliner 2005). However, ecological data do not always consist of observations that can be modeled with standard probability distributions. For instance, in frequent cases data may be zero-inflated and positive-continuous data. Several studies have developed models to handle properly such data in a computationally efficient way Lecomte et al. 2013; Thorson 2018. However, these may complicate a bit the way COS is tackled when dealing with an aggregation of such complex data as their convolution may not be as simple as Poisson or Gaussian ones.

In this paper, we aim at illustrating how to deal with change of support in ecological applications when the observation data are complex (e.g. zero-inflated and positive continuous). We base our approach on an existing framework developed by Alglave et al. (2022) in the field of marine and fisheries ecology. The framework aims at predicting the spatial distribution of fish species based on 2 datasets: scientific survey data and commercial catch declaration data. Commercial catch declarations are declared at the level of ICES rectangles (resolution of 0.5° x 1°) while scientific data benefit from exact location records. Usually in standard processing, declaration data are reallocated uniformly over their GPS fishing positions (available through Vessel Monitoring Satellites - VMS) in order to improve their spatial resolution (Hintzen et al. 2012b). However, the consequence

of this procedure on inference has never been explored. In particular, one can suspect that this could lead to strong homogeneization of the catch, to deleterious bias on model parameters and consequently to a loss of information for inference (Gotway and Young 2007; Pacifici, Reich, Miller, and Pease 2019). There is a need to understand how this procedure negatively affects inference and how the related bias can be corrected through alternative approaches properly handling COS.

In the following, we first describe the original model integrating both data sources and propose a generic statistical solution that allow to properly tackle the change of support issue and adapt it to our specific case (Section 1).

Then, we assess the effect of reallocation through a first set of simulations (section 2.1). In these simulations, the framework is simplified as much as possible to investigate the impact of reallocation on inference alone: are the estimates biased when reallocating catch declarations data? What is the gain of our alternative approach? For these simulations, the domain is reduced to a single statistical rectangle, only commercial declarations feed the model and the fish distribution is simply considered to arise from a known covariate.

In a second set of simulations, we get closer to a real application 2.2. In particular, the integrated dimension of the problem is added to the simulation configuration and both scientific and commercial data feed the model. This allows to investigate the contribution of both data sources to inference in addition to the effect of reallocation. The study domain is enlarged to several statistical rectangles. The model is complexified and species distribution is supposed to arise from a known covariate and a spatial random effect.

Finally, we compare the 2 methods on a real case study (common sole in the Bay of Biscay - Section 3) and we outline the consistency between simulation results and the real case of application.

1 A spatialized catch model for aggregated data

Alglave et al. (2022) propose a hierarchical spatial model to combine scientific survey data obtained through a standardized sampling protocol and catch data recorded by fishermen. This section provides a brief overview of the key ingredient of this model and raises the main concerns on the spatialization of the commercial catch data.

Latent field

Let $D \subset \mathbb{R}^2$ be a spatial domain and $(S) = (S(x), x \in D)$ a spatial random field which represents the biomass for a species of interest. (S) is assumed to be a spatial log-Gaussian Random Field (GRF) defined as $\log(S(x)) = \mu + \beta \cdot \Gamma(x) + \delta(x)$ (Figure 1) where $(\delta) = (\delta(x), x \in D)$ is a zero mean GRF with covariance matrix $\mathbf{\Sigma}$ and $(\Gamma) = (\Gamma(x), x \in \mathcal{D})$ a field of covariate.

Punctual observation layer for commercial data

As the observations are zero-inflated positive continuous data (Catch Per Unit of Effort - CPUE), following Thorson (2018), the authors model an observation at a sampled site x_i with a mixture of a Dirac mass at 0 and a Log Normal distribution (Equation 7), the proportion of zeroes in the mixture being defined conditionally on the random field $S(x_i)$.

$$Y_i | S(x_i), x_i \overset{ind}{\sim} \mathcal{M}_Y \left(p_i, \mu_i, \sigma^2\right), \tag{1}$$

with $p_i := \exp(-e^{\xi} S(x_i))$ the proportion of the mixture, e^{ξ} a parameter controlling zero-inflation, $\mu_i := k_{com} \frac{S(x_i)}{1-p_i}$ the expected catch when positive (on the natural-scale) and σ^2 a transformation of its variance (on the log-scale). When only commercial data feed the model, k_{com} is fixed to 1 and (S) is in the same unit as the commercial observations. When

other data sources feed the model, k_{com} is a scaling factor between commercial observations and the other data sources e.g. scientific data (see section 'Integrating scientific data in the model' for more details).

This parametrization of the mixture model corresponds to equations 2 (see the supplementary material for more details):

$$\begin{aligned}
&\mathbb{P}\left(Y_i = 0 | S(x_i), x_i\right) = p_i = \exp(-e^{\xi} S(x_i)) \\
&\mathbb{E}\left(Y_i | Y_i > 0, S(x_i), x_i\right) = \mu_i = k_{com} \frac{S(x_i)}{1 - p_i}, \\
&\mathbb{V}ar(Y_i | Y_i > 0, S(x_i), x_i) = \mu_i^2 (e^{\sigma^2} - 1),
\end{aligned} \tag{2}$$

Aggregated observation layer for commercial data

In the approach of Alglave et al. (2022) all fishing locations x_i and the corresponding individual observations Y_i are supposed to be known. However, fishermen do not declare this precise information in logbooks, they only declare the total daily catch aggregated at a given administrative spatial unit named statistical rectangles. For a given vessel fishing with a given gear on a given day, a declaration (denoted D) is therefore the sum of all individual observations Y_i realized in the administrative unit $\mathcal{A}_D$ associated with the declaration D (Equation 3), more formally:

$$D = \sum_{i | x_i \in \mathcal{A}_D} Y_i \tag{3}$$

Hence, in Alglave et al. (2022), declarations have been preprocessed and reallocated on fishing locations previously identified from VMS data (Hintzen et al. 2012a; Murray et al. 2013).

This process consists in counting, for a given vessel, on a given day, the number m_k of fishing points in $\mathcal{A}_{D_k}$ associated with declaration D_k and define for each $x_i \in \mathcal{A}_{D_k}$, the associated reallocated individual observation $Y_i^r := D_k/m_k$. As noted by Alglave et al. (2022), this process has several drawbacks. First, as a consequence of the reallocation process, the reconstructed observations tend to exhibit smoother patterns than the original observations. Second, the actual sample size is the total number of declarations while the new sample size after the reallocation process is the number of fishing locations, which is approximately 10 times the number of declarations. From a statistical point of view, this overestimation of the number of informative data tends to produce excessively narrow confidence intervals.

To circumvent such limitations, we propose an alternative approach that models the declarations D_k instead of the reconstructed individual observation Y_i^r. By defining the observation process at the declaration level D, we expect to avoid some of the drawbacks of the estimation based on reallocated individual observations Y^r. As we aim to propose a model compatible with the original one, we specify the distribution of D_k which consists of a sum of m_k random variables following a mixture distribution. However, this distribution has no known analytical form. The declaration data also exhibit some zero-inflation and a long tail repartition of the values, thus as for Y_i a mixture model is a good candidate to model D_k. We define then the distribution of D_k through the different key quantities, i.e. the mixture proportion, the expected positive declaration and its variance (Equation 4):

$$D_k | \boldsymbol{S}_{\mathcal{P}_k}, \mathcal{P}_k \sim \mathcal{M}_D \left(p_k^D, \mu_k^D, {\sigma_k^D}^2 \right) \tag{4}$$

with $\mathcal{P}_k = (1, ..., i, ..., m_k)$ the index of the individual observations belonging to the k^{th} declaration $D_k = \sum_{i \in \mathcal{P}_k} Y_i$, $(x_1, ..., x_i, ..., x_{m_k})$ is the list of all the fishing positions of

the k^{th} declaration, $\boldsymbol{S}_{\mathcal{P}_k}$ the latent field values at fishing positions $\mathcal{P}_k$, μ_k^D the expected positive biomass, p_k^D the proportion of the mixture and $\sigma_k^{D\,2}$ the variance parameter.

In order to relate the individual observation level Y and the declaration level D, we choose to match the key quantities of the two distributions. In the following, both Y_i and D_k are supposed to be conditional on the latent field (S) and on the related fishing positions (either x_i or $\mathcal{P}_k$).

1. As the $Y_1, \ldots, Y_{m_k}$ are independent conditionally on $\boldsymbol{S}_{\mathcal{P}_k}$, the probability of a zero-declaration $\mathbb{P}(D_k = 0)$ is obtained by simply multiplying the probability to obtain a zero-punctual observation $\mathbb{P}(Y_i = 0)$ at all fishing points $i \in \mathcal{P}_k$ (Equation 5).

$$\begin{aligned} \mathbb{P}(D_k = 0) &= \Pi_{i \in \mathcal{P}_k} \mathbb{P}(Y_i = 0) \\ &= \exp\left\{ - \sum_{i \in \mathcal{P}_k} e^{\xi}.S(x_i) \right\} = p_k^D \end{aligned} \tag{5}$$

2. The continuous component of the mixture is defined by the expected mean of a positive declaration and a transformation of its variance (see Equations 6 and SM). It is straightforward to prove that

$$\begin{aligned} &\mathbb{E}(D_k | D_k > 0) = \frac{\sum_{i \in \mathcal{P}_k} \mathbb{E}(Y_i)}{1 - p_k^D} = \frac{\sum_{i \in \mathcal{P}_k} S(x_i)}{1 - p_k^D} \\ &\mathbb{V}ar(D_k | D_k > 0) = \frac{\sum_{i \in \mathcal{P}_k} \mathbb{V}ar(Y_i)}{1 - p_k^D} - \frac{\pi_k}{(1 - p_k^D)^2} \mathbb{E}(D_k)^2 \\ &\text{with } \mathbb{V}ar(Y_i) = \frac{S(x_i)^2}{1 - p_i}(e^{\sigma^2} - (1 - p_i)) \\ &\text{and } p_i = \mathbb{P}(Y_i = 0) \end{aligned} \tag{6}$$

Knowing the moment of D_k (that only depends on the distribution of Y_i), it is required then to define a probability distribution for $D_k|D_k > 0$ and express $\mathbb{P}(D_k|D_k > 0)$ as a function of $\mathbb{E}(D_k|D_k > 0)$ and $\mathbb{V}ar(D_k|D_k > 0)$. As a good proxy for this distribution, we propose to use the same family of distribution for $D_k|D_k > 0$ as the one used for the individual observation $Y_i|Y_i > 0$ i.e. a Lognormal distribution. This is an approximation that we discuss later.

Integrating scientific data in the model

Scientific survey observations are availabale at their exact location and they can provide punctual observations to feed the model. They are integrated in inference through an observation process that has the same parameterization as the model of the punctual observation layer for commercial data.

$$Y_i^{(sci)}|S(x_i), x_i \overset{ind}{\sim} \mathcal{M}_Y\left(p_i^{(sci)}, \mu_i^{(sci)}, \sigma^2_{sci}\right) \tag{7}$$

with $p_i^{(sci)} := \exp(-e^{\xi_{sci}} S(x_i))$, $\mu_i^{(sci)} := k_{sci} \frac{S(x_i)}{1-p_i^{(sci)}}$. The parameters ξ_{sci}, σ^2_{sci} and k_{sci} are specific to the scientific observation model.

When scientific and commercial data feed the model, either k_{sci} or k_{com} need to be fixed so that the other parameter is estimated. In our case, we choose to fix the scientific parameter k_{sci} so that the latent field (S) is in the same unit as the scientific data and k_{com} is estimated and serves as scaling factor between scientific and commercial data.

Inference method

The inference is based on maximum likelihood approach with two approximations. We use the Stochastic Partial Differential Equations (SPDE) approach to represent the spatial Gaussian random field as a Gauss-Markov random field (Lindgren, Rue, and Lindström

2011) and we use the Laplace approximation to approximate the marginal likelihood of the model. The stochastic random field is also approximated by a piecewise constant process defined on a fine grid. The optimization of the likelihood relies on Template Model Builder (TMB), an effective tool to build hierarchical models and perform maximum likelihood estimation through automatic differentiation and Laplace approximation (Kristensen et al. 2016).

Alternative model configuration

In the following, we will first use simulations to compare three alternatives to estimate the spatial field of biomass from declaration data: a spatial model where individual observations are supposed to be known exactly, a 'reallocated model' where the model is fitted to reallocated observations, a 'declaration Model' where the model is fitted to spatially aggregated observations. The different model configurations are summarised in Table 1.

The latter two models are then tested on a real case study.

2 Simulation

To assess the drawbacks and the advantages of the different approaches, we conduct two different simulation studies.

First, we assess the effect of reallocation alone based on a simplistic statistical model. To do so, we conduct the simulation at the level of a single statistical rectangle on estimates, based on commercial data alone and with a very simple spatial latent field which only depends on one covariate (with no spatial random effect). This allows to clearly identify and illustrate the effect of reallocation on model estimates without confounding the effect of reallocation with other factors (e.g. the configuration of the study domain, artefacts that could arise from a more complex model). These simulations will be referred

as **single-square simulations**.

Then, we extend the analysis to get closer to a real case study and we investigate how integrating several data sources into inference while accounting for change of support improve model predictions. We simulate precisely geolocalized scientific data (in addition to commercial data), we shape the simulation domain to fit the case study domain (i.e. the Bay of Biscay area) which covers several statistical rectangles and we add a spatial random effect in the latent field. These simulations will be referred as **multiple-square simulations**.

In these two sets of simulation studies, there is a unique covariate that we suppose known at each point of the grid. Parameters values are detailed in the Table 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.

The locations of the individual commercial observation are generally organized in spatial clusters (they are named fishing zones in the following). The simulation process mimics this property by sampling the fishing points using a Neymann Scott process: the centers of the fishing zones are sampled according to a Poisson process and the fishing points are then uniformly sampled within a squared area that approximates the distance of a trawl haul. At each fishing position, an observation is sampled conditionally on the value of the latent field according to the model $\mathcal{M}_Y$.

We compare the performance of the Spatial Model (the gold standard), the Reallocated Model and the Declaration Model configurations in regards to several metrics/estimates.

The MSPE (Equation 8) quantifies the accuracy of the spatial predictions of the latent field over the spatial domain (n is the number of locations over the grid).

$$MSPE = \frac{\sum_i^n (S(x_i) - \hat{S}(x_i))^2}{n} \tag{8}$$

The estimates of the parameter β_S is also a key parameter of species distribution models as it quantifies the species habitat relationship.

In addition to the $MSPE$ and the species-habitat parameter $\hat{\beta}_S$, we look at the quality of the estimation for the intercept of the latent field $\hat{\mu}$, the observation variance parameter $\hat{\sigma}^2$ and the zero-inflation parameter $\hat{\xi}$. When a spatial random effect is simulated/estimated in the latent field (i.e. the mutliple square simulation), we also investigate the range estimates.

To get enough replicates, we run the simulations 100 times for both single-square and multiple square simulations.

2.1 Analysing the effect of data reallocation alone: single-square simulations

Two important variables may affect the accuracy of model outputs: the sample size of commercial data and the number of fishing zones explored and aggregated within a declaration. The single-square simulations intend to explore the effect of these two variables.

First, increasing the amount of data is expected to improve the estimates and the spatial prediction accuracy. We explore the potential improvement of the spatial predictions brought by an increasing amount of fishing points (10, 100 and 1000) which correspond respectively to 1, 10 and 100 declarations, the number of fishing locations within a declaration being fixed to 10.

Furthermore, the number of fishing zones within the statistical rectangle associated with a declaration might also affect the performance of the different approaches. We expect that the reallocation process will be less problematic when all the individual observations are spatially close as this situation is likely to correspond to a more homogeneous

underlying density than a situation with distant fishing zones. The accuracy of the Reallocated Model outputs is expected to decrease when the number of fishing zones increases. To assess the effect of such process, we simulated the fishing locations associated with a declaration assuming they were either realized in a single zone, in 3 distinct zones or in 5 distinct zones (Figure 2).

The results are presented in Figures 3 and 4.

The reallocation process has a major effect on predictions and estimates accuracy (Figure 3). As expected, the reallocation process conducts to a 10 to 200 times decrease in accuracy for spatial predictions. Accuracy decreases as the number of visited zones related to a declaration increases. Besides, the estimation of $\hat{\beta}_S$ is biased and reallocation leads to the loss of the species-habitat relationship as the number of fishing zones (related to a declaration) increases ($\hat{\beta}_S$ estimates get closer to 0). Increasing the number of samples does not improve inference.

The zero-inflation parameter (ξ) is also overestimated when using the Reallocated Model (Figure 4). When ξ increases, the amount of zero in the data decreases. Then, an overestimation of the ξ parameter means the model estimates that the amount of zero is smaller than what is actually simulated. This is not surprising: as soon as at least one of the individual observations Y_i associated with the same declaration D_k is non-zero, uniform reallocation will lead to a positive observation for each reallocated individual observation Y^r, hence to an underestimation of the proportion of zero. Consequently, this will tend to decrease the proportion of zero and will lead to the over-estimation of the ξ parameter.

The observation variance (σ) is underestimated - i.e. the data are estimated to be less noisy than they actually are - which is also a direct effect of uniform reallocation of declarations. The intercept of the latent field (μ) is slightly over-estimated (Figure 4).

Fitting the model to aggregated declaration allows to recover the species-habitat relationship and to improve the accuracy of the spatial predictions (Figure 3) even so the model outputs are not as accurate as the ones of the Spatial Model. Furthermore, the zero-inflation parameter is unbiased when the model is fitted to aggregated declarations. Other parameters (observation variance, intercept) are also better estimated than with the Reallocated Model even though they remain slightly biased (Figure 4). This alternative model has some convergence issues (Table 3) as 8% of the model runs did not converge when sample size is medium (100 pings) and only 3% did not when sample size is large (1000 pings).

2.2 Integrating several data sources with different spatial resolution: multiple-square simulations

In these simulations, the latent biomass process is modeled as the sum of a covariate effect and a random spatial field which represents the spatial structure not captured by the covariate. We also simulate precisely located scientific data as another source of information used to infer the spatial hidden biomass field and assess the constribution of scientific data in inference.

The study area is based on the case study; it includes the whole coast of the Bay of Biscay and covers several statistical rectangles (Figure 6A). To tailor the case study, we simulate 3000 of fishing positions grouped in 300 declarations (10 individual observations per declaration). Commercial data may not cover the full area and consequently we allow the commercial samples to cover 2/3 of the area similarly as in the case study. Similarly to the single-square simulations, the sampling of the commercial fishing points associated with a declaration is realized in three steps. (1) The declaration is randomly affected to one of the ICES rectangles. (2) The centroid of a fishing zone is uniformly sampled

within this statistical rectangle. (3) The 10 fishing punctual observations are randomly sampled within the fishing zone. The side of the squared 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 precisely localized scientific fishing points are simulated following a random stratified plan; contrary to commercial data they cover the entire study domain (Figure 6A). Scientific observations are simulated following the observation equation of $\mathcal{M}_Y$ (with specific parameters for scientific data - Table 2).

We compare several model configurations:

- to assess what brings our alternative approach, we compare the Reallocated Model to the Declaration 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).

Note that as in the single square simulation, the Declaration Model face some difficulty in convergence as only 75% of the model built on aggregated declarations converge (Table 4).

In addition to the 2 metrics introduced at the beginning of the section ($MSPE$ and species-habitat parameter β_S), we also compare the precision of the estimates for the range parameter.

The contribution of either scientific or commercial data can be clearly evidenced from the MSPE plot: the errors related to the integrated model at the declaration level or

at the individual reallocated observation level are always smaller than those obtained from models based on scientific data only or commercial data only. This can be well illustrated from Figure 6. 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, the Reallocated Model conducts to a loss in both the predictions accuracy and the species-habitat relationship (Figure 5) compared to the Declaration Model.

Interestingly, in addition to the species-habitat relationship, uniform reallocation also affects the range parameter. The Reallocated model provides biased range estimates while the Declaration Model provides unbiased estimates. Then, the Declaration Model (as the scientific-based model) better captures and disentangles the covariate effect and the spatial random effect and provides predictions that better fit to the small-scale patterns of the species distribution.

3 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-logbook data were extracted for the bottom trawlers fleet (OTB). The methods to cross VMS-logbook data and to filter the fleet is already extensively described in the previous paper (Alglave et al. 2022) and is not developed 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 Spatial Model fitted with scientific data only, (2) the Integrated Reallocated Model fitted to both scientific data

with known fishing location and declaration data uniformly reallocated on fishing locations and (3) the Integrated Declaration Model fitted to both scientific and declaration data aggregated at the scale of statistical squares.

The Integrated Declaration Model faced convergence issues (some of the parameters were hardly estimated e.g. the range parameter). To favor convergence, we integrated in the analysis onboard observer data from the same fleet. They can be considered as precisely geolocalized commercial catch data (86 samples are available for the related time step). Integrating these data allows to have direct information on Y_i and to better estimate the observation equation parameters (i.e. observation variance and zero-inflation parameter of commercial data).

Furthermore, as commonly done in complex fisheries model using automatic differentiation method (Fournier et al. 2012), we adopt a phase optimization procedure to initialize the optimization algorithm for the Declaration Model. We first fit the Reallocated model and use the estimates of this model as starting point of the optimization algorithm used for the Declaration Model estimation. We eventually fix the parameters that are hard to estimate in the initial optimization phases (intercept μ, covariate effect β_S, range and marginal variance) and finally let them free in the following phases of estimation.

Consistently with simulations, the Declaration Model shows differences with the Reallocated Model in both parameters estimates and spatial pattern of the species distribution (Figures 7, 8). In particular, the substrate effect is recovered in the Declaration Model and fall in the same range as estimates obtained from the scientific-based model (Figures 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 estimated with the Reallocated Model).

In addition, uncertainty is also revised when fitting the model at the declaration level. For instance, the confidence intervals of β_S, the marginal variance, the range, ξ_{com}, σ_{com} obtained from the Declaration Model are much wider than those obtained from the Reallocated Model. This emphasizes that uncertainty is probably underestimated in the Reallocated Model compared with the Declaration Model.

On the contrary, other parameters do not seem well estimated in either the Reallocated or the Declaration Models. For instance, compared to the scientific-based model, the intercept μ is revised upwards when building the likelihood on the individual precisely geolocalized observations and revised downwards when estimated with the Reallocated Model. This is consistent with the simulations results, see Figure 4.

Regarding the maps of the species distribution, fitting the model at the declaration level strongly modifies the model biomass field compared with the Reallocated Model. In particular, the substrate covariate have a sharper effect on species distribution and the intensity of the hotspots are revised when fitting the Declaration Model.

4 Discussion

The benefit of a statistical approach for COS

Handling change of support is a key issue in spatial statistics and extensive literature has intended to provide statistical methods to infer fine spatial processes based on data aggregated over rough scales (Wikle, Zammit-Mangion, and Cressie 2019; Wakefield and Lyons 2010). Such methods are key to integrate data that have different spatial resolution to make fine-scale inference on spatial processes (Pacifici, Reich, Miller, and Pease 2019). Still, in many cases, one often refines data resolution through ad-hoc arithmetic methods (proportional allocation, zonal addition) that can transform the data and lead to a loss of information (Young and Gotway 2007; Gotway and Young 2007) or artificially increase

the weight of such data when integrating several data sources (Alglave et al. 2022).

In this paper, we assessed how the well established method of proportional reallocation of declaration on fishing locations biases the parameter estimation and tend to produce overly smooth species distribution maps. Based on the framework of Alglave et al. (2022), we proposed an alternative integrated spatial framework that combines the two datasets to provide fine resolution maps of species distribution.

The base study explored in this paper highlights that even though prediction maps based on uniform reallocation allows to capture the main patterns of species distribution through the spatial random effect, uniform reallocation leads to the loss of the species-habitat relationship (parameters estimates are close to 0). Furthermore, results emphasize that uncertainty estimation is also strongly under estimated by uniform reallocation.

This is particularly problematic as one of the main objective of species distribution modeling lies in understanding the effect of habitat on species distribution (Guisan and Zimmermann 2000). Reallocated declarations data can provide information on the overall pattern of species distribution through the autocorrelation structures captured by the spatial random effect; however, they will not provide any information on species habitat preferences as the parameters of the species-habitat relationship will be biased.

The model that accounts for COS allows to recover the species-habitat relationship and provides more accurate spatial predictions of species distribution. Then, such method COS is key to estimate properly the species-habitat relationship from declarations data. More generally, COS approaches should be preferred when dealing with aggregated data because they allow (1) to properly reconcile the spatial scale of several data sources within the inference procedure, (2) to provide unbiased estimates of model parameters and (3) to better quantify model uncertainty.

The hierarchical structure of the approach and the punctual observation layer

The overall approach that we adopted to handle COS follows the standard structure of hierarchical frameworks. We assumed that both data sources (scientific data and commercial declarations data) arise from a shared latent process (species distribution) and that, while scientific data are recorded at their exact locations, commmercial declarations are recorded at a rough scale and are a convolution of exact location observations. Linking fine scale with rough scale for commercial data is made possible by relating the moments of the fine-scale observation probability distribution to the rough scale observation probability distribution.

The general approach that we propose (i.e. considering that aggregated data are convolutions of exact locations data) is relatively generic. To adapt the model to another application, only the moment equations and the probability distribution of the aggregated level would require to be adapted to the distribution of the underlying punctual observation level. However, considering that a convolution of zero-inflated lognormal distribution follows a zero-inflated lognormal is an approximation that can be questionned. We showed that this approximation is reasonnably good in our context (Alglave et al. 2022). However, exploring alternative observation models that verify additive property as the Gamma distribution would be an interesting perspectives for the future.

Finally, another approach that is common in the COS literature is 'Block krigeing' (Gelfand, Zhu, and Carlin 2001; Gelfand 2010; Pacifici, Reich, Miller, Gardner, et al. 2017). In such approach, the aggregation process is modeled in the latent field. By denoting a block B (i.e. a statistical rectangle), one can consider the latent field average over the block as $S(B) = |B|^{-1} \int_B S(x)dx$. In this case, the observations are supposed to arise from a distribution $\mathcal{M}_B$ conditionally on $S(B)$ following $D_j|S(B) \sim \mathcal{M}_B(S(B), \sigma^2)$.

This approach considers declarations arise from the averaged biomass over the statistical rectangle. This may suffer from the same difficulty as reallocated data and could tend to smoothed species-habitat relationship. By contrast, our approach considers that all observations are realized at given fishing locations and are then aggregated to constitute the declarations. It valorizes the information on fishing locations available through VMS data and then considers the catch has been realized over these locations conditionally on the related latent field values. In this case, COS is modeled in the observation layer, not in the latent field layer. This allows to remain closer to the actual process occuring during data aggregation (data are first observed and then aggregated). Furthermore, our approach allows to keep sparsity in the hessian of the likelihood and improve computation time, while Block krigeing would imply to loose sparsity by integrating over block areas B.

Future perspectives for the framework

More and more declarative data are now becoming available in the field of ecology, epidemiology and environnemental science. Typically, these are hunting records (Gilbert et al. 2021), administrative healthcare data (Morel et al. 2020), teledetection data (Garrigues, Allard, and Baret 2008). They are not specifically designed for a scientific analysis, but they can provide huge information for research and expertise provided the methodological challenges related or these data are overcome. Many drawbacks may impede the use of these data. Data aggregation is one of these issues, but as in citizen science programs sampling bias (Botella, Joly, Bonnet, Munoz, et al. 2021) as well as species misspecification can arise (Botella, Joly, Bonnet, Monestiez, et al. 2018). The approach that we propose is a step forward for a wider use of declarative data for scientific analysis and should be combined with other methods that have been developped to correct for the

several potential deleterious bias that can arise in non-standardized data (Dobson et al. 2020).

Acknowledgments

Include acknowledgments of funding, any patents pending, where raw data for the paper are deposited, etc.

References

[1] Baptiste Alglave et al. "Combining scientific survey and commercial catch data to map fish distribution". In: *ICES Journal of Marine Science* (Mar. 2022), fsac032. ISSN: 1054-3139. DOI: 10.1093/icesjms/fsac032. URL: https://doi.org/10.1093/icesjms/fsac032 (visited on 03/09/2022).

[2] Manuela Azevedo and Cristina Silva. "A framework to investigate fishery dynamics and species size and age spatio-temporal distribution patterns based on daily resolution data: a case study using Northeast Atlantic horse mackerel". In: *ICES Journal of Marine Science* 77.7 (Dec. 1, 2020), pp. 2933–2944. ISSN: 1054-3139. DOI: 10.1093/icesjms/fsaa170. URL: https://doi.org/10.1093/icesjms/fsaa170 (visited on 04/15/2022).

[3] Veronica J Berrocal, Alan E Gelfand, and David M Holland. "A bivariate space-time downscaler under space and time misalignment". In: *The annals of applied statistics* 4.4 (2010), p. 1942.

[4] Veronica J Berrocal, Alan E Gelfand, and David M Holland. "A spatio-temporal downscaler for output from numerical models". In: *Journal of agricultural, biological, and environmental statistics* 15.2 (2010), pp. 176–197.

[5] Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, et al. "Species distribution modeling based on the automated identification of citizen observations". In: *Applications in Plant Sciences* 6.2 (2018), e1029.

[6] Christophe Botella, Alexis Joly, Pierre Bonnet, François Munoz, et al. "Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data". In: *Methods in Ecology and Evolution* 12.5 (2021), pp. 933–945.

[7] Andrew DM Dobson et al. "Making messy data work for conservation". In: *One Earth* 2.5 (2020), pp. 455–465.

[8] Andrew O Finley, Sudipto Banerjee, and Bruce D Cook. "Bayesian hierarchical models for spatially misaligned data in R". In: *Methods in Ecology and Evolution* 5.6 (2014), pp. 514–523.

[9] Robert J. Fletcher et al. "A practical guide for combining data to model species distributions". en. In: *Ecology* 100.6 (2019), e02710. ISSN: 1939-9170. DOI: 10.1002/ecy.2710. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2710 (visited on 06/17/2021).

[10] David A Fournier et al. "AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models". In: *Optimization Methods and Software* 27.2 (2012), pp. 233–249.

[11] Sébastien Garrigues, Denis Allard, and Frédéric Baret. "Modeling temporal changes in surface spatial heterogeneity over an agricultural site". In: *Remote Sensing of Environment* 112.2 (2008), pp. 588–602.

[12] Alan E Gelfand. "Misaligned Spatial Data; The Change of Suport Problem". In: *Handbook of spatial statistics* 29 (2010), pp. 495–515.

[13] Alan E Gelfand, Li Zhu, and Bradley P Carlin. "On the change of support problem for spatio-temporal data". In: *Biostatistics* 2.1 (2001), pp. 31–45.

[14] BIAIS Gérard. "ORHAGO". In: (2003). Publisher: Sismer. DOI: 10.18142/23.

[15] Hans Gerritsen and Colm Lordan. "Integrating vessel monitoring systems (VMS) data with daily catch data from logbooks to explore the spatial distribution of catch and effort at high resolution". In: *ICES Journal of Marine Science* 68.1 (2010), pp. 245–252.

[16] Neil A Gilbert et al. "Integrating harvest and camera trap data in species distribution models". In: *Biological Conservation* 258 (2021), p. 109147.

[17] Carol A Gotway and Linda J Young. "A geostatistical approach to linking geographically aggregated data from different sources". In: *Journal of Computational and Graphical Statistics* 16.1 (2007), pp. 115–135.

[18] Carol A Gotway and Linda J Young. "Combining incompatible spatial data". In: *Journal of the American Statistical Association* 97.458 (2002), pp. 632–648.

[19] David Grémillet, Damien Chevallier, and Christophe Guinet. "Big data approaches to the spatial ecology and conservation of marine megafauna". In: *ICES Journal of Marine Science* (2022).

[20] Antoine Guisan and Niklaus E. Zimmermann. "Predictive habitat distribution models in ecology". In: *Ecological modelling* 135.2-3 (2000). Publisher: Elsevier, pp. 147–186.

[21] Stephanie E Hampton et al. "Big data and the future of ecology". In: *Frontiers in Ecology and the Environment* 11.3 (2013), pp. 156–162.

[22] Trevor J Hefley, Brian M Brost, and Mevin B Hooten. "Bias correction of bounded location errors in presence-only data". In: *Methods in Ecology and Evolution* 8.11 (2017), pp. 1566–1573.

[23] Niels T. Hintzen et al. "VMStools: Open-Source software for the processing, analysis and visualisation of fisheries logbook and VMS data". In: *Fisheries Research* 115 (2012). Publisher: Elsevier, pp. 31–43.

[24] Niels T. Hintzen et al. "VMStools: Open-source software for the processing, analysis and visualisation of fisheries logbook and VMS data". In: *Fisheries Research* 115 (2012). Publisher: Elsevier, pp. 31–43.

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

[26] ICES. *Report of the Working Group on Beam Trawl Surveys (WGBEAM)*. en. Tech. rep. Galway, Ireland, 2018, p. 121.

[27] Nick JB Isaac et al. "Data integration for large-scale models of species distributions". In: *Trends in ecology & evolution* 35.1 (2020), pp. 56–67.

[28] Yongku Kim and L Mark Berliner. "Change of spatiotemporal scale in dynamic models". In: *Computational Statistics & Data Analysis* 101 (2016), pp. 80–92.

[29] Kasper Kristensen et al. "TMB: Automatic Differentiation and Laplace Approximation". English. In: *Journal of Statistical Software* 70.1 (Apr. 2016), pp. 1–21. ISSN: 1548-7660. DOI: 10.18637/jss.v070.i05. URL: https://www.jstatsoft.org/index.php/jss/article/view/v070i05 (visited on 01/24/2020).

[30] Jean-Baptiste Lecomte et al. "Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling". In: *L'Institut des Sciences et Industries du Vivant et de l'Environnement (AgroParisTech)* (2013), p. 37.

[31] Finn Lindgren, H\aavard Rue, and Johan Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011). Publisher: Wiley Online Library, pp. 423–498.

[32] Aurore Maureaud et al. "Are we ready to track Climate-driven shifts in marine species across international boundaries? - A global survey of scientific bottom trawl data". English. In: *Global Change Biology* (Oct. 2020). tex.options: useprefix=true, gcb.15404. ISSN: 1354-1013, 1365-2486. DOI: 10.1111/gcb.15404. URL: https://onlinelibrary.wiley.com/doi/10.1111/gcb.15404 (visited on 10/18/2020).

[33] David AW Miller et al. "The recent past and promising future for data integration methods to estimate species' distributions". In: *Methods in Ecology and Evolution* 10.1 (2019), pp. 22–37.

[34] Maryan Morel et al. "ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection". In: *Biostatistics* 21.4 (2020), pp. 758–774.

[35] Andrew S Mugglin, Bradley P Carlin, and Alan E Gelfand. "Fully model-based approaches for spatially misaligned data". In: *Journal of the American Statistical Association* 95.451 (2000), pp. 877–887.

[36] Lee G. Murray et al. “The effectiveness of using CPUE data derived from Vessel Monitoring Systems and fisheries logbooks to estimate scallop biomass”. In: *ICES Journal of Marine Science* 70.7 (2013), pp. 1330–1340.

[37] Ran Nathan et al. “Big-data approaches lead to an increased understanding of the ecology of animal movement”. In: *Science* 375.6582 (2022), eabg1780.

[38] Krishna Pacifici, Brian J Reich, David AW Miller, Beth Gardner, et al. “Integrating multiple data sources in species distribution modeling: a framework for data fusion”. In: *Ecology* 98.3 (2017), pp. 840–850.

[39] Krishna Pacifici, Brian J Reich, David AW Miller, and Brent S Pease. “Resolving misaligned spatial data with integrated species distribution models”. In: *Ecology* 100.6 (2019), e02709.

[40] Ryan J Parker, Brian J Reich, and Stephan R Sain. “A multiresolution approach to estimating the value added by regional climate models”. In: *Journal of Climate* 28.22 (2015), pp. 8873–8887.

[41] Benjamin Planque et al. “Understanding what controls the spatial distribution of fish populations using a multi-model approach”. English. In: *Fisheries Oceanography* 20.1 (2011), pp. 1–17. ISSN: 1365-2419. DOI: 10.1111/j.1365-2419.2010.00546.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2419.2010.00546.x (visited on 08/31/2020).

[42] Brian J Reich, Howard H Chang, and Kristen M Foley. “A spectral method for spatial downscaling”. In: *Biometrics* 70.4 (2014), pp. 932–942.

[43] Ian W Renner, Julie Louvrier, and Olivier Gimenez. “Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization”. In: *Methods in Ecology and Evolution* 10.12 (2019), pp. 2118–2128.

[44] Brian L. Sullivan et al. “The eBird enterprise: An integrated approach to development and application of citizen science”. In: *Biological Conservation* 169 (Jan. 1, 2014), pp. 31–40. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2013.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0006320713003820 (visited on 04/19/2022).

[45] James T. Thorson. “Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative”. In: *Canadian Journal of Fisheries and Aquatic Sciences* 75.9 (2018). Publisher: NRC Research Press, pp. 1369–1382.

[46] Jonathan Wakefield and Hilary Lyons. “Spatial aggregation and the ecological fallacy”. In: *Handbook of spatial statistics* 541 (2010), p. 558.

[47] Christopher K Wikle and L Mark Berliner. “Combining information across spatial scales”. In: *Technometrics* 47.1 (2005), pp. 80–91.

[48] Christopher K Wikle, Andrew Zammit-Mangion, and Noel Cressie. *Spatio-temporal Statistics with R*. Chapman and Hall/CRC, 2019.

[49] Linda J Young and Carol A Gotway. "Linking spatial data from different sources: the effects of change of support". In: *Stochastic Environmental Research and Risk Assessment* 21.5 (2007), pp. 589–600.

Tables

Table 1: Model configurations.

Model name	Configuration
Spatial Model	Baseline configuration (or gold standard). the commercial observations are known at there exact locations. This is an ideal situation with no actual application and it is used as a reference for the comparison between the two alternatives.
Reallocated Model	The original model fitted with commercial reallocated individual catch (and potentially few precisely geolocalized scientific data) as done in Alglave et al. (2022).
Declaration Model	The alternative approach introduced in this paper where the biomass model is fitted using commercial catch declaration at a coarse spatial level and potentially few precisely geolocalized scientific data.

Table 2: Parameter values for the simulations

Parameters	Single-square simulations	Multiple-square simulations
μ	2	2
β_S	2	2
Range of δ	–	0.6 ($\approx$ 50 km)
Marginal variance of δ	–	1
ξ_{com}	-1	-1
σ_{com}	1	1
k_{com}	–	1
ξ_{sci}	–	0
σ_{sci}	–	0.8

Table 3: Single-square simulations - Percentage of convergence per simulation-estimation configuration.

Fishing positions	Declarations	Reallocation	Likelihood level	Convergence (%)
10	1	No	Y_i	99.668
10	1	Yes	Y_i^r	0.333
10	1	Yes	D_j	0.000
100	10	No	Y_i	100.000
100	10	Yes	Y_i^r	100.000
100	10	Yes	D_j	92.000
1000	100	No	Y_i	100.000
1000	100	Yes	Y_i^r	100.000
1000	100	Yes	D_j	97.333

Table 4: Multiple-square simulations - Percentage of convergence per simulation-estimation configuration.

Model	Likelihood level	Convergence (%)
Commercial model	Y_i^r	100.000
Commercial model	D_j	75.377
Integrated model	Y_i^r	100.000
Integrated model	D_j	76.382
Scientific model		100.000

692

Figures

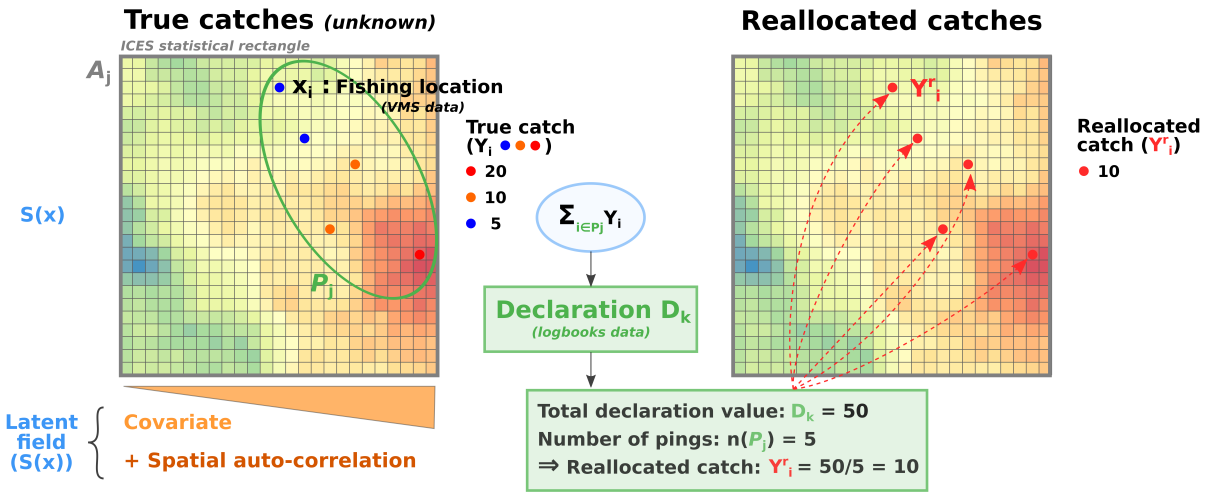

Figure 1: Schematic representation of the reallocation process. The biomass field (the background field) depends on a covariate and a spatial random effect. The covariate is the x axis. It has a positive effect on biomass values (i.e. biomass is higher on the right of the grid than on the left). The spatial random effect conduct to a hotspot on the bottom-right of the latent field. The study domain is considered as a statistical rectangle (grey square). Fishermen sample observations in areas of poor biomass where the covariate is relatively low (blue points) and in areas of higher biomass where the covariate is higher and eventually in the hotspot of biomass (orange and red points). These catches belong to the same declaration k and are summed to constitute the declaration $D_k = 50$. The declaration is declared at the level of the statistical rectangle. From VMS data, we know the fishing positions x_i. In standard processing, D_k are then uniformly reallocated over the fishing positions x_i. This strongly homogenizes the catch. In particular, the effect of the habitat is no more evidenced in the reallocated catch Y_i^r.

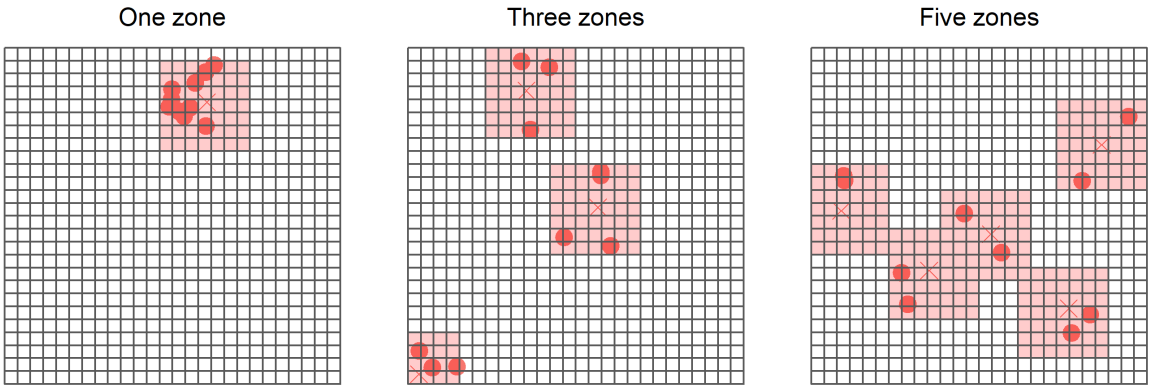

Figure 2: Simulations of 10 fishing points within 1, 3 and 5 fishing zones. The full grid corresponds to a statistical rectangle. Cross are the centroid of the fishing zones. A declaration declared at the level of the statistical rectangle would be uniformly reallocated over these fishing points.

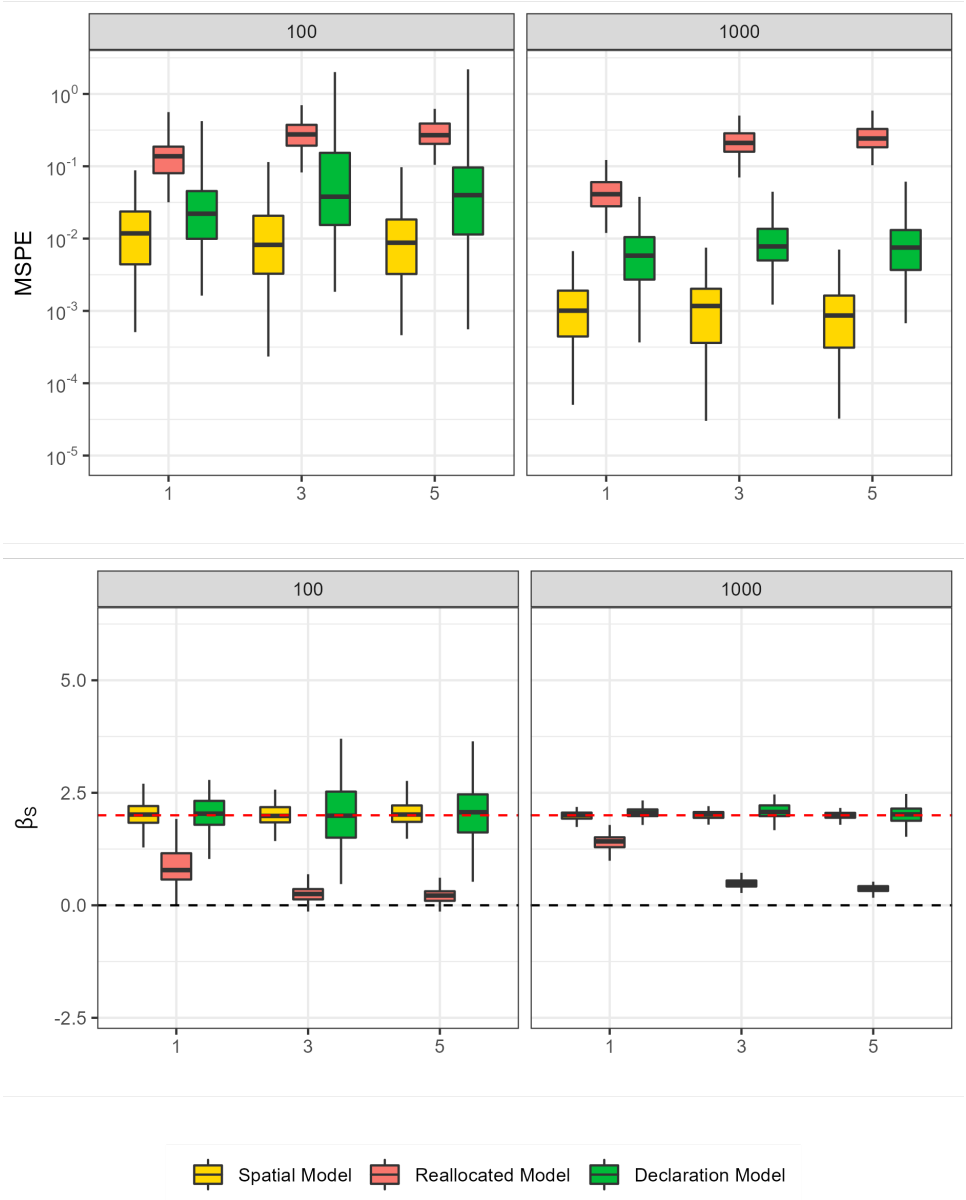

Figure 3: Performance metrics for single-square simulations with a total of 100 or 1000 fishing positions in columns. $MSPE = \frac{\sum_i^n (S(x_i) - \hat{S}(x_i))^2}{n}$ is the mean squared prediction error and $\hat{\beta}_S$ is the species-habitat relationship parameter. The number of fishing zones visited within each declaration is represented on the x-axis. The results of the Spatial model are in yellow, in red the results of the Reallocated Model and in green the Declaration Model. Simulations conducted with 10 fishing positions are not represented as they encounter convergence issues as stated in Table 3.

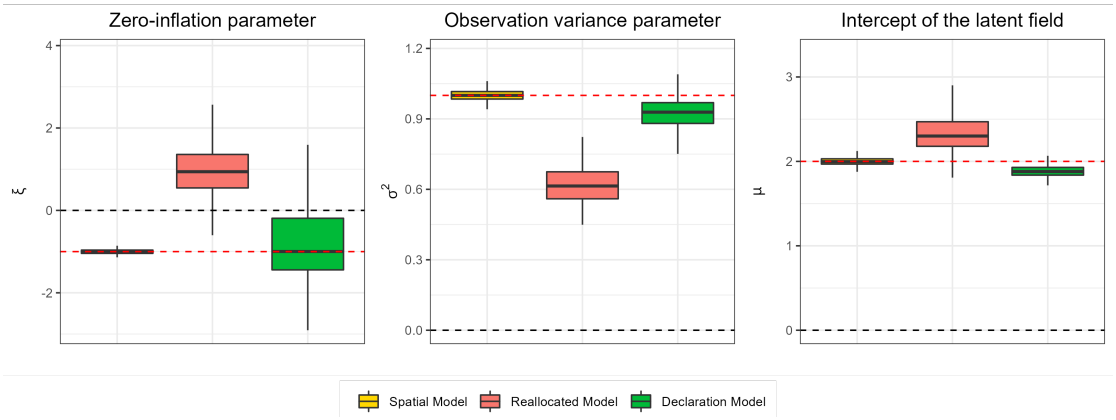

Figure 4: Parameters relative bias for single-square simulations. Only the simulations with 1000 fishing positions are represented. Black line: zero value. Red line: parameter true value.

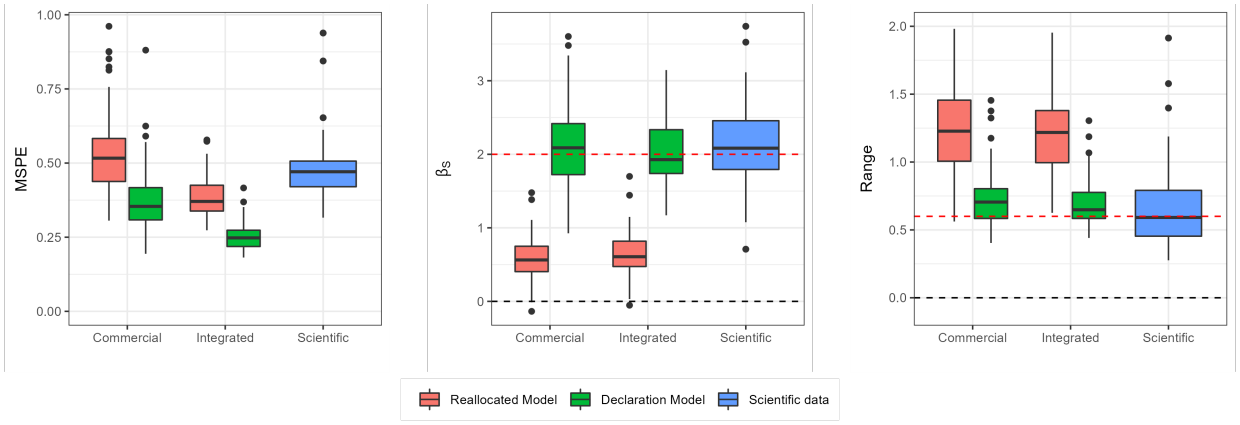

Figure 5: Performance metric for the multiple-square simulations. Red line: true value for the range and the species-habitat parameter (β_S). Red: uniform reallocation (Y_i^r model). Green: model-based reallocation (D_k model). Blue: scientific-based model.

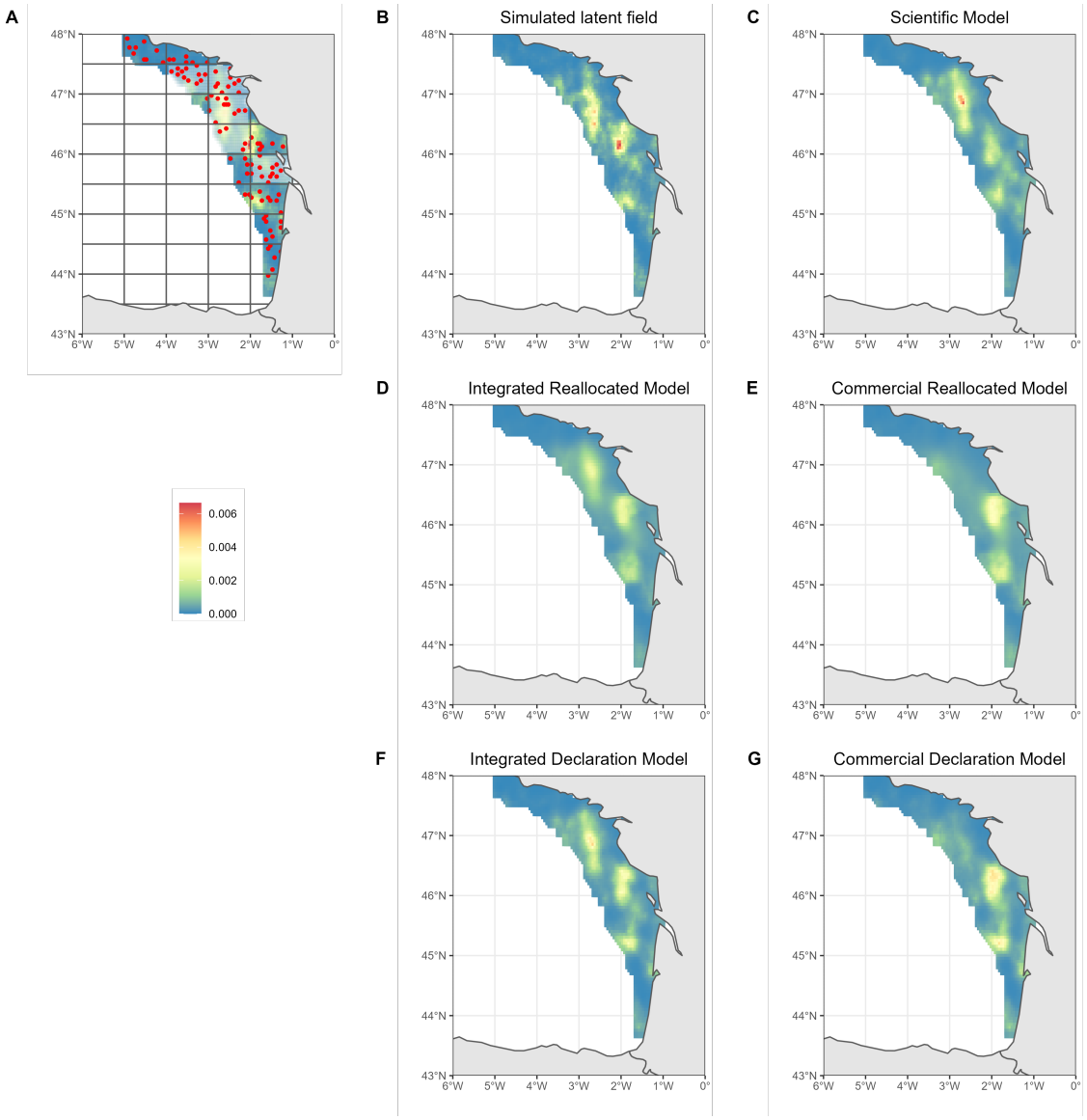

Figure 6: Distribution of simulated/estimated biomass field. A: Simulated biomass field with scientific samples (red) and statistical rectangles. The rectangles that have not been sampled by commercial data are the transparent rectangles. They represent 1/3 of the full area. B: simulated biomass field. C: biomass field from the scientific-based model. Y_i^r: Reallocated Model (D, E). D_k: Declaration Model (F, G). Scientific model: model fitted to scientific data only. Commercial model: model fitted to commercial data only. Integrated model: model fitted to both data sources.

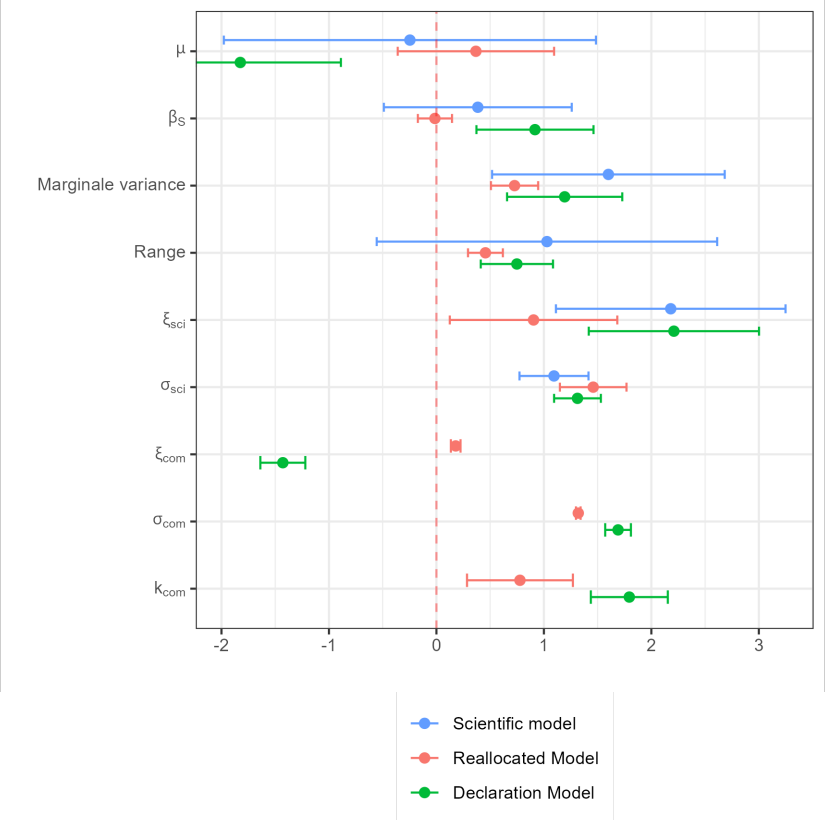

Figure 7: Parameters obtained with the model fitted on scientific data only, the integrated model fitted on reallocated catch Y_i^r and the integrated model fitted on catch declarations D_k.

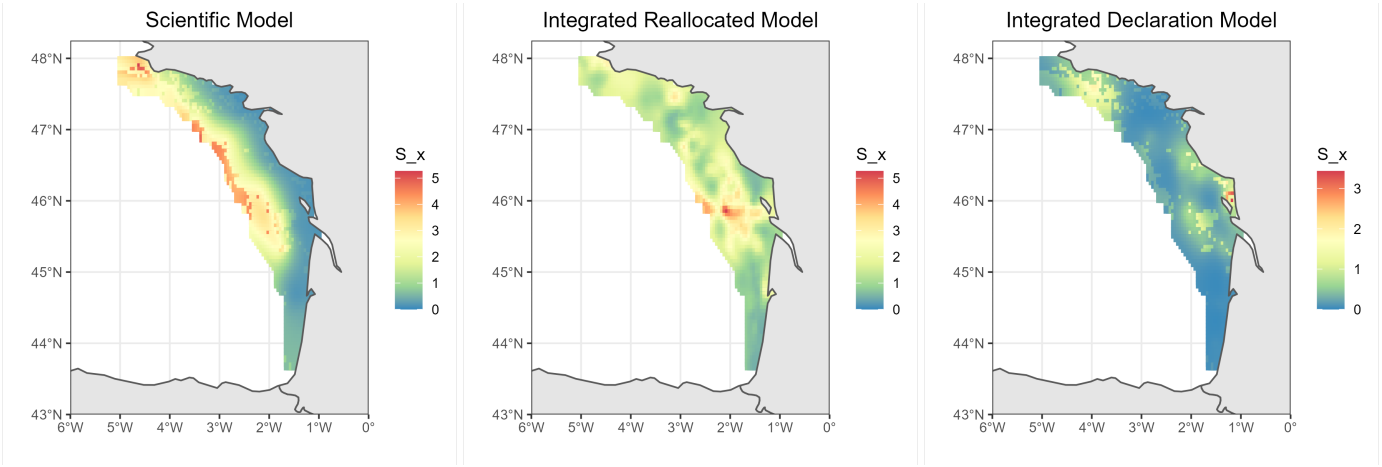

Figure 8: Maps obtained from the scientific-based model (left), the integrated model fitted on reallocated catch Y_i^r (center), the integrated model fitted on catch declarations D_k (right).

Supplementary materials

Notations

We model catch declarations D_k (available at coarse resolution through logbook data) as a sum of Y_i individual observations (which are considered as latent variables) each one realised at one fishing position x_i (known through VMS data).

We note:

- $\mathcal{P}_k = (1, ..., i, ..., m_k)$: the vector of all the individual catches belonging to the k^{th} declaration.
- $k \in \{1, ..., l\}$: the declaration index with l the number of declarations.
- $(x_1, ..., x_i, ..., x_{m_k})$: the vector of all the fishing positions of the k^{th} declaration.

$$D_k = \sum_{i \in \mathcal{P}_k} Y_i$$

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, $\mathbb{E}(Z) = e^{\rho + \frac{\sigma^2}{2}}$ and $\mathbb{V}ar(Z) = (e^{\sigma^2} - 1)e^{2\rho + \sigma^2}$.

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

- $Z = \mu e^{\sigma N - \frac{sigma^2}{2}}$

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

$\mathbf{D_k}$ probability distribution and moments

We have to express the probability distribution of D_k and its moments as a function of Y_i and its related moments. Let's assume $Y_i = C_i \cdot Z_i$ is a zero-inflated Lognormal distribution with C_i and Z_i the two components of the mixture. C_i is a binary random variable and Z_i a Lognormal random variable.

$$C_i|S(x_i), x_i \sim \mathcal{B}(1 - p_i)$$

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

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

Probability of obtaining a zero declaration

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

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

Expectation of a positive declaration

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

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

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

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

Variance of a positive declaration

The variance then can be expressed as:

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

with,

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

and

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

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

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

As the $(Y_i)_{i\in\mathcal{P}_k}$ are independent, $\mathbb{V}ar(D_k) = \sum_{i\in\mathcal{P}_k}\mathbb{V}ar(Y_i) = \sum_{i\in\mathcal{P}_k}\mathbb{V}ar(C_i.Z_i)$.

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

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

Sum up of the main formulas

The main formulas can be summarised as follows:

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

- The probability to obtain a zero declaration

$$\mathbb{P}(D_k = 0) = \exp\left\{ -\sum_{i \in \mathcal{P}_k} e^{\xi}.S(x_i) \right\} = \pi_k$$

- The expectancy of a positive declaration

$$\mathbb{E}(D_k | D_k > 0) = \frac{\sum_{i \in \mathcal{P}_k} S(x_i)}{1 - \pi_k}$$

- The variance of a positive declaration

$$\mathbb{V}ar(D_k | D_k > 0) = \frac{\sum_{i \in \mathcal{P}_k} \mathbb{V}ar(Y_i)}{1 - \pi_k} - \frac{\pi_k}{(1 - \pi_k)^2} \mathbb{E}(D_k)^2$$

- The variance of an individual observation

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

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

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