



HAL
open science

Inferring fine scale wild species distribution from spatially aggregated data

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

► To cite this version:

Baptiste Alglave, Kasper Kristensen, Etienne Rivot, Mathieu Woillez, Youen Vermard, et al.. Inferring fine scale wild species distribution from spatially aggregated data. 2023. hal-03878990

HAL Id: hal-03878990

<https://hal.science/hal-03878990v2>

Preprint submitted on 9 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Inferring fine scale wild species distribution from spatially aggregated data

Journal:	<i>Journal of the Royal Statistical Society: Series C</i>
Manuscript ID	JRSSC-Jan-2023-0009
Manuscript Type:	Original Article
Date Submitted by the Author:	09-Jan-2023
Complete List of Authors:	<p>Alglave, Baptiste; University of Washington, School of Aquatic and Fisheries Science Kristensen, Kasper; DTU Aqua Rivot, Etienne; Agrocampus Ouest, UMR DECOD Woillez, Mathieu; Ifremer Vermard, Youen; Ifremer Centre Atlantique, EMH Etienne, Marie-Pierre; IRMAR</p>
Keywords:	spatial statistics, change of support, hierarchical modeling, species distribution model

SCHOLARONE™
Manuscripts

Inferring fine scale wild species distribution from spatially aggregated data

Baptiste Alglave^{1*}, Kasper Kristensen², Etienne Rivot¹,
Mathieu Woillez³, Youen Vernard⁴, Marie-Pierre Etienne⁵

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 and non-aggregated spatial data offer possibilities to map wild species distribution. 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 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 assess the framework through simulations and apply it on real data for the common sole of the Bay of Biscay.

Keywords: spatial statistics, change of support, integrated hierarchical model, species distribution model, fisheries data.

17 Introduction

18 Context

19 With the progress of new technologies, spatial ecological data are becoming more and more
20 accessible every day thanks to the huge effort of the scientific community to generate and
21 get access to intensive information for ecology, evolution and conservation (Nathan et al.
22 2022; Hampton et al. 2013; Grémillet, Chevallier, and Guinet 2022). These data are cru-
23 cial to face the current challenges related to large- and small-scale ecological questions: for
24 instance, following animal movement (Nathan et al. 2022), mapping species distribution
25 (Isaac et al. 2020) or tracking climate change (Maureaud et al. 2020).

26 These data sources are often highly heterogeneous in size, type and sampling design,
27 making their combination a methodological challenge (Fletcher et al. 2019; Isaac et al.
28 2020; Miller et al. 2019; Pacifici, Reich, Miller, Gardner, et al. 2017; Renner, Louvrier,
29 and Gimenez 2019). For instance, in species distribution modeling, recent studies have
30 investigated how to combine scientific standardized data with auxiliary data such as citizen
31 science data (Fletcher et al. 2019). Typically, count data from planned surveys can be
32 combined with other counts data coming from citizen science programs (e.g. see the
33 eBird program for bird ecology - Sullivan et al. (2014)). These first ones benefit from a
34 standardized protocol, a controlled sampling plan and they are designed to cover the full
35 range of species distribution. The second ones provide a larger amount of data with lower
36 cost, but they arise from non-standardized sampling and consequently they may not cover
37 the whole area. Integrating these data sources typically allows to benefit from the good
38 coverage of the survey while improving spatial prediction accuracy through the massive
39 amount of data available through citizen science programs.

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

1
2
3
4
5
6
7 41 data as the mandatory data that must be reported by some agent as a legal requirement to
8
9 42 proceed with his activity). As they are mandatory, declaration data are usually very large
10
11 43 datasets (much larger than scientific or citizen science datasets). They can provide highly
12
13 44 valuable to map wildlife species distribution. In fisheries science, a common example
14
15 45 of such data sources are commercial catch declaration data. They can be used to map
16
17 46 fish distribution and provide valuable information to identify spawning areas or nursery
18
19 47 grounds Alglave et al. (2022) and Azevedo and Silva (2020).

20
21 48 Although massive, these data are most often registered at the scale of coarse spatial
22
23 49 units while scientific survey and citizen science data are usually reported with their exact
24
25 50 locations. Generally, these administrative units do not have a resolution that is relevant
26
27 51 for ecological analysis (Pacifi, Reich, Miller, and Pease 2019).

28
29 52 Developing statistical methods that properly handle spatially aggregated data and
30
31 53 integrate these with higher resolution data is then a major challenge to make precise and
32
33 54 unbiased inference of species distribution at a fine scale.

34 35 55 **The change of support issue**

36
37
38 56 Inferring fine-scale spatial processes from coarse data and reconciling spatial scales prop-
39
40 57 erly when different set of observations do not have the same resolution is a well known
41
42 58 issue in geography, ecology, agriculture, geology and statistics (Gotway and Young 2002).
43
44 59 In the statistical literature, *Change of Support* (COS) refers to ‘the summary or analysis
45
46 60 of spatial data at a scale different from that at which it was originally collected’ (Gotway
47
48 61 and Young 2002; Gelfand 2010). It is often also referred as ‘downscaling/upscaling’ or
49
50 62 Modifiable Areal Unit Problem (MAUP) in the literature (Wikle, Zammit-Mangion, and
51
52 63 Cressie 2019). This is typically the case where data are aggregated over larger geograph-
53
54 64 ical scales, but one would like to infer processes at a different resolution. In such case,
55
56
57
58
59
60

1
2
3
4
5
6
7 65 conclusions from a fine-resolution analysis can strongly differ from an analysis at a coarser
8
9 66 scale based on the aggregation of the fine-resolution data. Such phenomena is also called
10
11 67 the ecological fallacy (Wakefield and Lyons 2010).

12
13 68 Since 2000, several studies have described how COS issues could be overcome; Mugglin,
14
15 69 Carlin, and Gelfand (2000), Gelfand, Zhu, and Carlin (2001), Gotway and Young (2007)
16
17 70 and Wikle and Berliner (2005) proposed generic approaches (and extensions of these
18
19 71 approaches - Kim and Berliner (2016)) for addressing COS in a spatial or spatio-temporal
20
21 72 context. In health analysis, Young and Gotway (2007) proposed to compare some rough
22
23 73 approach based on centroids of areal units to relate environmental and health outcomes
24
25 74 with an approach that honors the spatial support of the data (size, shape, orientation).
26
27 75 Berrocal, Gelfand, and Holland (2010a) and Berrocal, Gelfand, and Holland (2010b)
28
29 76 proposed a spatio-temporal method for fusing several air pollution data: one from coarse
30
31 77 resolution but with full spatial coverage and another recorded at point level, with sparse
32
33 78 distribution but where records almost corresponds to the true value of the process. In
34
35 79 climate science, Reich, Chang, and Foley (2014) and Parker, Reich, and Sain (2015)
36
37 80 proposed a spectral statistical approach to downscale information from large-scale model
38
39 81 to lower scale. In the field of ecology, some recent studies have tackled such issues: Finley,
40
41 82 Banerjee, and Cook (2014) provided a framework for integrating spatially misaligned
42
43 83 data, Hefley, Brost, and Hooten (2017) proposed a solution based on COS to account for
44
45 84 location error in presence-only data, Pacifici, Reich, Miller, and Pease (2019) introduced a
46
47 85 framework for integrating data sources of different resolution to map species distribution.
48
49 86 Applying similar ideas, Gilbert et al. (2021) integrated harvest data (aggregated data)
50
51 87 and camera trap (precisely geolocalized data) to map several wildlife species in Wisconsin.
52
53
54
55
56
57
58
59
60

88 **Focus of the paper**

89 One of the main challenge limiting the number of application consists in the type of ob-
90 servation data that can be fitted to the existing COS framework. Indeed, the frameworks
91 that were developed so far and their related applications mainly limited their scope to
92 relatively simple observation data: count data were modeled through Poisson processes
93 (Gilbert et al. 2021; Gotway and Young 2007; Mugglin, Carlin, and Gelfand 2000; Pacifici,
94 Reich, Miller, and Pease 2019) and continuous data were modeled through Gaussian or
95 Gamma distributions (Berrocal, Gelfand, and Holland 2010a; Gelfand, Zhu, and Carlin
96 2001; Wikle and Berliner 2005). However, ecological data do not always consist of ob-
97 servations that can be modeled with standard probability distributions. For instance, in
98 frequent cases data may be zero-inflated and positive-continuous data. Several studies
99 have developed models to handle properly such data in a computationally efficient way
100 Lecomte et al. 2013; Thorson 2018. However, these may complicate a bit the way COS is
101 tackled when dealing with an aggregation of such complex data as their convolution may
102 not be as simple as Poisson or Gaussian ones.

103 In this paper, we aim at illustrating how to deal with change of support in ecological
104 applications when the observation data are complex (e.g. zero-inflated and positive con-
105 tinuous). We base our approach on an existing framework developed by Alglave et al.
106 (2022) in the field of marine and fisheries ecology. The framework aims at predicting the
107 spatial distribution of fish species based on 2 datasets: scientific survey data and com-
108 mercial catch declaration data. Commercial catch declarations are declared at the level of
109 ICES rectangles (resolution of $0.5^\circ \times 1^\circ$) while scientific data benefit from exact location
110 records. Usually in standard processing, declaration data are reallocated uniformly over
111 their GPS fishing positions (available through Vessel Monitoring Satellites - VMS) in or-
112 der to improve their spatial resolution (Hintzen et al. 2012b). However, the consequence

1
2
3
4
5
6
7 113 of this procedure on inference has never been explored. In particular, one can suspect that
8
9 114 this could lead to strong homogenization of the catch and to bias parameters inference
10
11 115 (Gotway and Young 2007; Pacifici, Reich, Miller, and Pease 2019). There is a need to
12
13 116 understand how this procedure negatively affects inference and how the related bias can
14
15 117 be corrected through alternative approaches properly handling COS.

16
17 118 In the following, we first describe the original model integrating both data sources and
18
19 119 propose a generic statistical solution that allow to properly tackle the change of support
20
21 120 issue and adapt it to our specific case (Section 1).

22
23 121 Then, we assess the effect of reallocation through a first set of simulations (Section 2.1).
24
25 122 In these simulations, the framework is simplified as much as possible to investigate the
26
27 123 impact of reallocation on inference alone: are the estimates biased when reallocating catch
28
29 124 declarations data? What is the gain of our alternative approach? For these simulations,
30
31 125 the domain is reduced to a single statistical rectangle, only commercial declarations feed
32
33 126 the model and the fish distribution is simply considered to arise from a known covariate.

34
35 127 In a second set of simulations, we get closer to a real application (Section 2.2). In
36
37 128 particular, the integrated dimension of the problem is added to the simulation configura-
38
39 129 tion and both scientific and commercial data feed the model. This allows to investigate
40
41 130 the contribution of both data sources to inference in addition to the effect of reallocation.
42
43 131 The study domain is enlarged to several statistical rectangles. The model is complexified
44
45 132 and species distribution is supposed to arise from a known covariate and a spatial random
46
47 133 effect.

48
49 134 Finally, we compare the 2 methods on a real case study (common sole in the Bay of
50
51 135 Biscay - Section 3) and we outline the consistency between simulation results and the real
52
53 136 case of application.
54
55
56
57
58
59
60

1 A spatialized catch model for aggregated data

Alglave et al. (2022) have proposed a hierarchical spatial model to combine scientific survey data obtained through a standardized sampling protocol and catch data recorded by fishers. The proposed model will be denoted by GeoCatch in the following and assumes that the catch are precisely geolocalized.

A short presentation of the GeoCatch model

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 isotropic GRF with a Matern covariance function and $(\Gamma) = (\Gamma(x), x \in \mathcal{D})$ a field of covariate.

Following Thorson (2018), the zero-inflated positive continuous data (Catch Per Unit of Effort - CPUE), $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ at sampled sites $(x_1, \dots, x_n)^\top$ are assumed to be independent conditionally on (S) and for any i in $\llbracket 1..n \rrbracket$,

$$Y_i | S(x_i) = p_i \delta_0 + (1 - p_i) \mu_i e^{\sigma N_i - \frac{\sigma^2}{2}}, \quad N_i \sim \mathcal{N}(0, 1), \quad (1)$$

where δ_0 stands for the Dirac mass in 0 and N_i a standard Normal random variable. In the following, this mixture distribution will be denoted by

$$Y_i | S(x_i) \stackrel{ind}{\sim} \mathcal{M}_Y(p_i, \mu_i, \sigma^2), \quad (2)$$

with $p_i := \exp(-e^\xi S(x_i))$ the proportion of the mixture, e^ξ a parameter controlling zero-inflation, $\mu_i := \frac{S(x_i)}{1-p_i}$ the expected catch when positive (on the natural-scale) and

1
2
3
4
5
6
7 156 σ^2 the variance parameter on the log-scale. While accounting for the zero inflation in
8
9 157 the data, this choice allows to represent continuous positive data and ensures that the
10
11 158 expected catch at site x equals the local biomass $S(x)$. A more detailed presentation
12
13 159 is available in the Supplementary Material. The three main quantities of interest are
14
15 160 summed up by the following equations for any $i = 1, \dots, n$:

$$\begin{aligned} \mathbb{P}(Y_i = 0 | S(x_i)) &= p_i = \exp(-e^\xi S(x_i)) \\ \mathbb{E}(Y_i | Y_i > 0, S(x_i)) &= \mu_i = \frac{S(x_i)}{1 - p_i}, \\ \text{Var}(Y_i | Y_i > 0, S(x_i)) &= \mu_i^2 (e^{\sigma^2} - 1), \end{aligned} \quad (3)$$

161 Aggregated observation layer for commercial data

31
32 162 In the approach of Alglave et al. (2022) all fishing locations x_i and the corresponding
33
34 163 individual catch Y_i are supposed to be recorded. However, fishers do not declare the
35
36 164 individual catch but only the total daily catch aggregated at a given administrative spatial
37
38 165 unit named statistical rectangles in the fisheries management vocabulary. Those units are
39
40 166 represented for the Bay of Biscay map in Figure 6. A given vessel fishing with a given
41
42 167 gear on a given day declares the total catch realized in a statistical rectangle, this will be
43
44 168 referred to as a declaration and denoted by D . D_{vgda} is therefore the sum of all individual
45
46 169 catch Y_i realized by vessel v with gear g on day d in administrative unit a . For the sake
47
48 170 of simplicity, we fix the vessel, gear and day in the presentation and we omit them, so
49
50 171 that D_a is defined by:

$$D_a = \sum_{i|x_i \in \mathcal{R}_a} Y_i, \quad (4)$$

172 \mathcal{R}_a being the geographical area corresponding to the administrative unit a .

173 Hence, in Alglave et al. (2022), declarations have been preprocessed and reallocated
 174 on fishing locations previously identified from VMS data as it is classically done when
 175 defining spatialized CPUE (Hintzen et al. 2012a; Murray et al. 2013).

176 This process consists in identifying the locations $x_i \in \mathcal{R}_a$, associated with declaration
 177 D_a and defining for each the associated reallocated individual observation Y_i^r :

$$Y_i^r := \frac{D_a}{m_a} \mathbb{1}_{\{\mathcal{R}_a\}}(x_i), \quad \forall i = 1, \dots, n, \quad (5)$$

178 where $\mathbb{1}_{\{\mathcal{R}_a\}}(x)$ stands for the characteristic function which equals 1 when x belongs to
 179 the geographical area corresponding to the administrative unit and m_a the cardinal of
 180 set $\{x_i \in \mathcal{R}_a\}$. As noted by Alglave et al. (2022), this process has several drawbacks.
 181 First, as a consequence of the reallocation process, the reconstructed observations tend
 182 to exhibit smoother patterns than the original observations. Second, the actual sample
 183 size is the total number of declarations while the new sample size after the reallocation
 184 process is the number of fishing locations, which is approximately 10 times the number
 185 of declarations. From a statistical point of view, this artificial data augmentation tend to
 186 overestimate the information brought by the data and for example to produce excessively
 187 narrow confidence intervals.

188 To circumvent such limitations, we propose an alternative approach that models the
 189 declarations \mathbf{D} instead of the reconstructed individual catch \mathbf{Y}^r . By defining the observa-
 190 tion process at the declaration level, we expect to avoid some of the drawbacks previously
 191 mentioned. To specify an aggregated version of the GeoCatch model, we need to specify
 192 the probability distribution of a sum of a fixed number of random variables, each following
 193 the mixture defined in Equation 1. Although this distribution has no known analytical
 194 form, it also exhibits some zero-inflation and a long tail repartition of the values and thus

1
2
3
4
5
6
7 195 a mixture model is a good candidate to model \mathbf{D} and we propose to model \mathbf{D} using the
8
9 196 same mixture form as in Equation 1 but with adequate parameters:

$$D_a | \mathbf{S}, (x_{a1}, \dots, x_{am_a}) \sim \mathcal{M}_D (p_a^D, \mu_a^D, \sigma_a^{D^2}) \quad (6)$$

10
11
12
13
14
15 197 with $(x_{a1}, \dots, x_i, \dots, x_{am_a})$ is the list of all the fishing positions associated to the decla-
16
17 198 ration D_a in area \mathcal{R}_a , μ_a^D the expected positive biomass, p_a^D the proportion of the mixture
18
19 199 and $\sigma_a^{D^2}$ the variance parameter.
20
21
22

23 201 In order to relate the individual observation level \mathbf{Y} and the declaration level \mathbf{D} ,
24
25 202 we choose to match the key quantities of the two distributions. In the following, both
26
27 203 \mathbf{Y} and \mathbf{D} are defined conditionally on the latent field (S) and on the related fishing
28
29 204 positions. With no loss of generality, we can rename the sequence $\mathbf{Y} = (Y_i)_{i=1, \dots, n}$ in
30
31 205 $(Y_{ai})_{a=1, \dots, A, i=1, \dots, m_a}$, A being the total number of administrative units.
32
33

34 206 1. As the Y_{a1}, \dots, Y_{am_a} are independent conditionally on \mathbf{S} , we have:
35
36
37

$$p_a^D = \mathbb{P}(D_a = 0) = \prod_{i=1}^{m_a} \mathbb{P}(Y_{ai} = 0) = \exp \left\{ - \sum_{i=1}^{m_a} e^{\xi} \cdot S(x_{ai}) \right\} \quad (7)$$

38
39
40
41 207 2. The continuous component of the mixture is defined by the expected mean of a
42
43 208 positive declaration and a transformation of its variance (see Equations 8 and SM).
44

45 209 It is straightforward to prove that
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

$$\begin{aligned}\mu_a^D &= \mathbb{E}(D_a | D_a > 0) = \frac{\sum_{i=1}^{m_a} S(x_{ai})}{1 - p_a^D} \\ \text{Var}(D_a | D_a > 0) &= \frac{\sum_{i=1}^{m_a} \text{Var}(Y_{ai})}{1 - p_a^D} - \frac{p_a^D}{(1 - p_a^D)^2} \mathbb{E}(D_a)^2\end{aligned}\quad (8)$$

$$\text{with } \text{Var}(Y_{ai}) = \frac{S(x_{ai})^2}{1 - p_{ai}} (e^{\sigma^2} - (1 - p_{ai})) \quad \text{and } p_{ai} = \mathbb{P}(Y_{ai} = 0).$$

We suggest to approximate the distribution of $\mathbb{P}(D_a | D_a > 0)$ as a Lognormal distribution. This is an approximation that we discuss later.

Integrating scientific data in the model

Scientific survey observations are available 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.

$$Y_{ai}^{(sci)} | S(x_{ai}), x_{ai} \stackrel{ind}{\sim} \mathcal{M}_Y \left(p_{ai}^{(sci)}, \mu_{ai}^{(sci)}, \sigma_{sci}^2 \right) \quad (9)$$

$$\text{with } p_{ai}^{(sci)} := \exp(-e^{\xi_{sci}} S(x_{ai})), \mu_{ai}^{(sci)} := k_{sci} \frac{S(x_{ai})}{1 - p_{ai}^{(sci)}}.$$

k_{sci} is a scaling parameter named catchability in the fisheries science literature to account for a proportionality coefficient between expected commercial catch and scientific catch i.e. $k_{sci} = \mathbb{E}(\mathbf{Y}^{(sci)}) / \mathbb{E}(\mathbf{Y}^{(com)})$. The parameters ξ_{sci} , σ_{sci}^2 are specific to scientific data.

When combining scientific and commercial data, we can either estimate k_{sci} and express the latent field in the same scale than the commercial data (which is better if the amount of commercial data is larger than the scientific one as mentioned in Alglave et al. (2022) or on the opposite express the latent field in the same unit than the scientific data and add the corresponding k_{com} parameter to define the commercial catch distribution. In

1
2
3
4
5
6
7 227 our case, we choose to express the latent field (S) is in the same unit as the scientific data
8
9 228 and estimate a parameter k_{com} to match the classical choice in fisheries science.

11 229 **Inference method**

12
13
14 230 The inference is based on maximum likelihood approach with two approximations. We use
15
16 231 the Stochastic Partial Differential Equations (SPDE) approach to represent the spatial
17
18 232 Gaussian random field as a Gauss-Markov random field (Lindgren, Rue, and Lindström
19
20 233 2011) and we use the Laplace approximation to approximate the marginal likelihood of the
21
22 234 model. The stochastic random field is also approximated by a piecewise constant process
23
24 235 defined on a fine grid. The optimization of the likelihood relies on Template Model Builder
25
26 236 (TMB), an effective tool to build hierarchical models and perform maximum likelihood
27
28 237 estimation through automatic differentiation and Laplace approximation (Kristensen et
29
30 238 al. 2016).

31 32 239 **Alternative model configuration**

33
34
35 240 In the following, we will first use simulations to compare three alternatives to estimate
36
37 241 the spatial field of biomass from declaration data: a first configuration called ‘Spatial
38
39 242 Model’ refers to the GeoCatch model fitted to the true individual observations (as if
40
41 243 individual observations were known), a second configuration called ‘Reallocated Model’
42
43 244 refers to the GeoCatch model fitted to the reallocated observations, a last configuration
44
45 245 called ‘Declaration Model’ refers to the model that accounts for COS and that is fitted to
46
47 246 the aggregated data.

48
49 247 where individual observations are supposed to be known exactly, a ‘reallocated model’
50
51 248 where the model is fitted to reallocated observations, a ‘declaration Model’ where the
52
53 249 model is fitted to spatially aggregated observations. The different model configurations
54
55
56
57
58
59
60

1
2
3
4
5
6
7 are summarised in Table 1.

8
9 The latter two models are then tested on a real case study.

10 11 2 Simulation

12
13
14
15 To assess the drawbacks and the advantages of the different approaches, we conduct two
16
17 different simulation studies.

18
19 First, we assess the effect of reallocation alone based on a simplistic statistical model.
20
21 To do so, we conduct the simulation at the level of a single statistical rectangle on esti-
22
23 mates, based on commercial data alone and with a very simple spatial latent field which
24
25 only depends on one covariate (with no spatial random effect). This allows to clearly
26
27 identify and illustrate the effect of reallocation on model estimates without confounding
28
29 the effect of reallocation with other factors (e.g. the configuration of the study domain,
30
31 artefacts that could arise from a more complex model). These simulations will be referred
32
33 as **single-square simulations**.

34
35 Then, we extend the analysis to get closer to a real case study and we investigate
36
37 how integrating several data sources into inference while accounting for change of support
38
39 improve model predictions. We simulate precisely geolocalized scientific data (in addition
40
41 to commercial data), we shape the simulation domain to fit the case study domain (i.e.
42
43 the Bay of Biscay area) which covers several statistical rectangles and we add a spatial
44
45 random effect in the latent field. These simulations will be referred as **multiple-square**
46
47 **simulations**.

48
49 In these two sets of simulation studies, there is a unique covariate that we suppose
50
51 known at each point of the grid. Parameters values are detailed in the Table 2.

52
53 Regarding commercial data, the number of fishing pings per declaration is fixed to 10
54
55 as it is the average number of fishing locations for a single declaration in real data.

1
2
3
4
5
6
7 274 The locations of the individual commercial observation are generally organized in
8
9 275 spatial clusters (they are named fishing zones in the following). The simulation process
10
11 276 mimics this property by sampling the fishing points using a Neymann Scott process: the
12
13 277 centers of the fishing zones are sampled according to a Poisson process and the fishing
14
15 278 points are then uniformly sampled within a squared area that approximates the distance
16
17 279 of a trawl haul. At each fishing position, an observation is sampled conditionally on the
18
19 280 value of the latent field according to the model \mathcal{M}_Y .

20
21 281 We compare the performance of the Spatial Model (the gold standard), the Reallocated
22
23 282 Model and the Declaration Model configurations in regards to several metrics/estimates.

24
25 283 The MSPE (Equation 10) quantifies the accuracy of the spatial predictions of the
26
27 284 latent field over the spatial domain (n_{cells} is the number of locations over the grid).

$$MSPE = \frac{\sum_j^{n_{cells}} (S(x_j) - \hat{S}(x_j))^2}{n} \quad (10)$$

28
29
30
31
32
33 285 The estimates of the parameter β_S is also a key parameter of species distribution
34
35 286 models as it quantifies the species habitat relationship.

36
37 287 In addition to the $MSPE$ and the species-habitat parameter $\hat{\beta}_S$, we look at the qual-
38
39 288 ity of the estimation for the intercept of the latent field $\hat{\mu}$, the observation variance
40
41 289 parameter $\hat{\sigma}^2$ and the zero-inflation parameter $\hat{\xi}$. When a spatial random effect is simu-
42
43 290 lated/estimated in the latent field (i.e. the multiple square simulation), we also investigate
44
45 291 the range estimates.

46
47 292 To get enough replicates, we run the simulations 100 times for both single-square and
48
49 293 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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

319 leads to the loss of the species-habitat relationship as the number of fishing zones (related
320 to a declaration) increases (i.e. $\hat{\beta}_S$ estimates get closer to 0). Increasing the number of
321 samples does not improve inference.

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

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

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

- 368 • to assess what brings our alternative approach, we compare the Reallocated Model
369 to the Declaration Model.
- 370 • to assess the information brought by each data source, we compare models built on
371 scientific data only (scientific-based models), models built on commercial data only
372 (commercial-based models) and models combining both data sources (integrated
373 models).

374 Note that as in the single square simulation, the Declaration Model face some difficulty
375 in convergence as only 75% of the model built on aggregated declarations converge (Table
376 4).

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

380 The contribution of either scientific or commercial data can be clearly evidenced from
381 the MSPE plot: the errors related to the integrated model at the declaration level or
382 at the individual reallocated observation level are always smaller than those obtained
383 from models based on scientific data only or commercial data only. This can be well
384 illustrated from Figure 6. Integrating scientific and commercial data allows to (1) capture
385 the hotspot missed by commercial data through scientific data and (2) better capture the
386 local correlation structures through the dense commercial data.

387 Furthermore, consistently with single-square simulations, the Reallocated Model con-
388 ducts to a loss in both the predictions accuracy and the species-habitat relationship (Fig-
389 ure 5) compared to the Declaration Model.

390 Interestingly, in addition to the species-habitat relationship, uniform reallocation also
391 affects the range parameter. The Reallocated model provides biased range estimates while

1
2
3
4
5
6
7 392 the Declaration Model provides unbiased estimates. Then, the Declaration Model (as the
8
9 393 scientific-based model) better captures and disentangles the covariate effect and the spatial
10
11 394 random effect and provides predictions that better fit to the small-scale patterns of the
12
13 395 species distribution.

14 15 396 **3 Case-study: sole of the Bay of Biscay**

16
17
18 397 To illustrate our method on a real case study, we applied the approach to the common
19
20 398 sole of the Bay of Biscay. VMS-logbook data were extracted for the bottom trawlers
21
22 399 fleet (OTB). The methods to cross VMS-logbook data and to filter the fleet is already
23
24 400 extensively described in the previous paper (Alglave et al. 2022) and is not developed
25
26 401 further here. Scientific data were extracted from the DATRAS database for the Orhago
27
28 402 beam trawl survey (Gérard 2003; ICES 2018b). To align the commercial and the scientific
29
30 403 data, we filtered scientific data based on the minimum size of sole (24 cm for sole - ICES
31
32 404 (2018a)). To illustrate the method, we compare the outputs of (1) the Spatial Model fitted
33
34 405 with scientific data only, (2) the Integrated Reallocated Model fitted to both scientific data
35
36 406 with known fishing location and declaration data uniformly reallocated on fishing locations
37
38 407 and (3) the Integrated Declaration Model fitted to both scientific and declaration data
39
40 408 aggregated at the scale of statistical squares.

41
42 409 The Integrated Declaration Model faced convergence issues (some of the parameters
43
44 410 were hardly estimated e.g. the range parameter). To favor convergence, we integrated
45
46 411 in the analysis onboard observer data from the same fleet. They can be considered as
47
48 412 precisely geolocalized commercial catch data (86 samples are available for the related
49
50 413 time step). Integrating these data allows to have direct information on Y_{ai} and to better
51
52 414 estimate the observation equation parameters (i.e. observation variance and zero-inflation
53
54 415 parameter of commercial data).

1
2
3
4
5
6
7 416 Furthermore, as commonly done in complex fisheries model using automatic differenti-
8
9 417 ation method (Fournier et al. 2012), we adopt a phase optimization procedure to initialize
10
11 418 the optimization algorithm for the Declaration Model. We first fit the Reallocated model
12
13 419 and use the estimates of this model as starting point of the optimization algorithm used
14
15 420 for the Declaration Model estimation. We eventually fix the parameters that are hard
16
17 421 to estimate in the initial optimization phases (intercept μ , covariate effect β_S , range and
18
19 422 marginal variance) and finally let them free in the following phases of estimation.

20
21 423 Consistently with simulations, the Declaration Model shows differences with the Real-
22
23 424 located Model in both parameters estimates and spatial pattern of the species distribution
24
25 425 (Figures 7, 8). In particular, the substrate effect is recovered in the Declaration Model
26
27 426 and fall in the same range as estimates obtained from the scientific-based model (Figures
28
29 427 7). The zero-inflation parameter ξ is revised downwards (i.e. there are actually more zero-
30
31 428 values than in the reallocated data) while the observation variance of commercial data is
32
33 429 revised upwards (i.e. the commercial data are noisier than estimated with the Reallocated
34
35 430 Model).

36
37 431 In addition, uncertainty is also revised when fitting the model at the declaration
38
39 432 level. For instance, the confidence intervals of β_S , the marginal variance, the range, ξ_{com} ,
40
41 433 σ_{com} obtained from the Declaration Model are much wider than those obtained from the
42
43 434 Reallocated Model. This emphasizes that uncertainty is probably underestimated in the
44
45 435 Reallocated Model compared with the Declaration Model.

46
47 436 On the contrary, other parameters do not seem well estimated in either the Reallocated
48
49 437 or the Declaration Models. For instance, compared to the scientific-based model, the
50
51 438 intercept μ is revised upwards when building the likelihood on the individual precisely
52
53 439 geolocalized observations and revised downwards when estimated with the Reallocated
54
55 440 Model. This is consistent with the simulations results, see Figure 4.

1
2
3
4
5
6
7 441 Regarding the maps of the species distribution, fitting the model at the declaration
8
9 442 level strongly modifies the model biomass field compared with the Reallocated Model. In
10
11 443 particular, the substrate covariate have a sharper effect on species distribution and the
12
13 444 intensity of the hotspots are revised when fitting the Declaration Model.

15 445 **4 Discussion**

18 446 **The benefit of a statistical approach for COS**

20
21 447 Handling change of support is a key issue in spatial statistics and extensive literature
22
23 448 has intended to provide statistical methods to infer fine spatial processes based on data
24
25 449 aggregated over rough scales (Wikle, Zammit-Mangion, and Cressie 2019; Wakefield and
26
27 450 Lyons 2010). Such methods are key to integrate data that have different spatial resolution
28
29 451 to make fine-scale inference on spatial processes (Pacifi, Reich, Miller, and Pease 2019).
30
31 452 Still, in many cases, one often refines data resolution through ad-hoc arithmetic methods
32
33 453 (proportional allocation, zonal addition) that can transform the data and lead to a loss
34
35 454 of information (Young and Gotway 2007; Gotway and Young 2007) or artificially increase
36
37 455 the weight of such data when integrating several data sources (Alglave et al. 2022).

38
39 456 In this paper, we assessed how the well established method of proportional reallocation
40
41 457 of declaration on fishing locations biases the parameter estimation and tend to produce
42
43 458 overly smoothed species distribution maps. Based on the framework of Alglave et al.
44
45 459 (2022), we proposed an alternative integrated spatial framework that combines the two
46
47 460 datasets to provide fine resolution maps of species distribution.

48
49 461 The base study explored in this paper highlights that even though prediction maps
50
51 462 based on uniform reallocation allows to capture the main patterns of species distribution
52
53 463 through the spatial random effect, uniform reallocation leads to the loss of the species-
54
55 464 habitat relationship (parameters estimates are close to 0). Furthermore, results emphasize

1
2
3
4
5
6
7 465 that uncertainty estimation is also strongly under estimated by uniform reallocation.

8 466 This is particularly problematic as one of the main objective of species distribution
9
10 467 modeling lies in understanding the effect of habitat on species distribution (Guisan and
11
12 468 Zimmermann 2000). Reallocated declarations data can provide information on the overall
13
14 469 pattern of species distribution through the autocorrelation structures captured by the
15
16 470 spatial random effect; however, they will not provide any information on species habitat
17
18 471 preferences as the parameters of the species-habitat relationship will be biased.

20 472 The model that accounts for COS allows to recover the species-habitat relationship
21
22 473 and provides more accurate spatial predictions of species distribution. Then, such method
23
24 474 accounting for COS is key to estimate properly the species-habitat relationship from
25
26 475 declarations data. More generally, COS approaches should be preferred when dealing
27
28 476 with aggregated data because they allow (1) to properly reconcile the spatial scale of
29
30 477 several data sources within the inference procedure, (2) to provide unbiased estimates of
31
32 478 model parameters and (3) to better quantify model uncertainty.

35 479 **The hierarchical structure of the approach and the punctual ob-** 36 37 480 **servaion layer**

38
39 481 The overall approach that we adopted to handle COS follows the standard structure of
40
41 482 hierarchical frameworks. We assumed that both data sources (scientific data and commer-
42
43 483 cial declarations data) arise from a shared latent process (species distribution) and that,
44
45 484 while scientific data are recorded at their exact locations, commercial declarations are
46
47 485 recorded at a rough scale and are a convolution of exact location observations. Linking
48
49 486 fine scale with rough scale for commercial data is made possible by relating the moments
50
51 487 of the fine-scale observation probability distribution to the rough scale observation prob-
52
53 488 ability distribution.

1
2
3
4
5
6
7 489 The general approach that we propose (i.e. considering that aggregated data are con-
8
9 490 volutions of exact locations data) is relatively generic. To adapt the model to another
10
11 491 application, only the moment equations and the probability distribution of the aggre-
12
13 492 gated level would require to be adapted to the distribution of the underlying punctual
14
15 493 observation level. However, considering that a convolution of zero-inflated lognormal dis-
16
17 494 tribution follows a zero-inflated lognormal is an approximation that can be questioned.
18
19 495 We showed that this approximation is reasonably good in our context (Alglave et al.
20
21 496 2022). However, exploring alternative observation models that verify additive property
22
23 497 as the Gamma distribution would be an interesting perspectives for the future.

24
25 498 Finally, another approach that is common in the COS literature is ‘Block krige-
26
27 499 ing’ (Gelfand, Zhu, and Carlin 2001; Gelfand 2010; Pacifici, Reich, Miller, Gardner,
28
29 500 et al. 2017). In such approach, the aggregation process is modeled in the latent field
30
31 501 and one usually consider the latent field average over the statistical rectangle (or block)
32
33 502 $S(\mathcal{R}_a) = |\mathcal{R}_a|^{-1} \int_{\mathcal{R}_a} S(x) dx$. In this case, the observations are supposed to arise from a
34
35 503 distribution $\mathcal{M}_{\mathcal{R}}$ conditionally on $S(\mathcal{R}_a)$ following $D_a | S(\mathcal{R}_a) \sim \mathcal{M}_{\mathcal{R}}(S(\mathcal{R}_a), \sigma^2)$. This
36
37 504 approach considers declarations arise from the averaged biomass over the statistical rect-
38
39 505 angle. This may suffer from the same difficulty as the reallocated data and could tend to
40
41 506 smoothed the species-habitat relationship. By contrast, our approach considers that all
42
43 507 observations are realized at given fishing locations and are then aggregated to constitute
44
45 508 the declarations. It valorizes the information on fishing locations available through VMS
46
47 509 data and then considers the catch has been realized over these locations conditionally
48
49 510 on the related latent field values. In this case, COS is modeled in the observation layer,
50
51 511 not in the latent field layer. This allows to remain closer to the actual process occurring
52
53 512 during data aggregation (data are first observed and then aggregated). Furthermore, our
54
55 513 approach allows to keep sparsity in the hessian of the likelihood and improve computation

1
2
3
4
5
6
7 514 time, while Block kriging would imply to loose sparsity by integrating over block areas
8 515 \mathcal{R}_a .

11 516 **Future perspectives for the framework**

13
14 517 More and more declarative data are now becoming available in the field of ecology, epi-
15
16 518 demiology and environmental science. Typically, these are hunting records (Gilbert
17
18 519 et al. 2021), administrative healthcare data (Morel et al. 2020), teledetection data (Gar-
19
20 520 rrigues, Allard, and Baret 2008). They are not specifically designed for a scientific analysis,
21
22 521 but they can provide huge information for research and expertise provided the method-
23
24 522 ological challenges related or these data are overcome. Many drawbacks may impede
25
26 523 the use of these data. Data aggregation is one of these issues, but as in citizen science
27
28 524 programs sampling bias (Botella, Joly, Bonnet, Munoz, et al. 2021) as well as species
29
30 525 misspecification can arise (Botella, Joly, Bonnet, Monestiez, et al. 2018). The approach
31
32 526 that we propose is a step forward for a wider use of declarative data for scientific analysis
33
34 527 and should be combined with other methods that have been developed to correct for the
35
36 528 several potential deleterious bias that can arise in non-standardized data (Dobson et al.
37
38 529 2020).

530 **Acknowledgment**

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

535 **Fundings**

536 The authors declare no specific funding for this work.

537 **Competing interests**

538 The authors declare there are no competing interests.

539 **Data availability statement**

540 Survey data are available through the DATRAS portal (<https://www.ices.dk/data/>
541 [data-portals/Pages/DATRAS.aspx](https://www.ices.dk/data/data-portals/Pages/DATRAS.aspx)) with the package 'icesDatras' ([https://cran.r-project.](https://cran.r-project.org/web/packages/icesDatras/index.html)
542 [org/web/packages/icesDatras/index.html](https://cran.r-project.org/web/packages/icesDatras/index.html)). Logbooks and VMS data are confidential
543 data and they are available on specific request to DPMA.

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.

- 1
2
3
4
5
6
7 580 [11] Sébastien Garrigues, Denis Allard, and Frédéric Baret. “Modeling temporal changes
8 581 in surface spatial heterogeneity over an agricultural site”. In: *Remote Sensing of*
9 582 *Environment* 112.2 (2008), pp. 588–602.
- 10
11 583 [12] Alan E Gelfand. “Misaligned Spatial Data; The Change of Support Problem”. In:
12 584 *Handbook of spatial statistics* 29 (2010), pp. 495–515.
- 13
14 585 [13] Alan E Gelfand, Li Zhu, and Bradley P Carlin. “On the change of support problem
15 586 for spatio-temporal data”. In: *Biostatistics* 2.1 (2001), pp. 31–45.
- 16
17 587 [14] BIAIS Gérard. “ORHAGO”. In: (2003). Publisher: Sismser. DOI: 10.18142/23.
- 18
19 588 [15] Hans Gerritsen and Colm Jordan. “Integrating vessel monitoring systems (VMS)
20 589 data with daily catch data from logbooks to explore the spatial distribution of catch
21 590 and effort at high resolution”. In: *ICES Journal of Marine Science* 68.1 (2010),
22 591 pp. 245–252.
- 23
24 592 [16] Neil A Gilbert et al. “Integrating harvest and camera trap data in species distribu-
25 593 tion models”. In: *Biological Conservation* 258 (2021), p. 109147.
- 26
27 594 [17] Carol A Gotway and Linda J Young. “A geostatistical approach to linking geo-
28 595 graphically aggregated data from different sources”. In: *Journal of Computational*
29 596 *and Graphical Statistics* 16.1 (2007), pp. 115–135.
- 30
31 597 [18] Carol A Gotway and Linda J Young. “Combining incompatible spatial data”. In:
32 598 *Journal of the American Statistical Association* 97.458 (2002), pp. 632–648.
- 33
34 599 [19] David Grémillet, Damien Chevallier, and Christophe Guinet. “Big data approaches
35 600 to the spatial ecology and conservation of marine megafauna”. In: *ICES Journal of*
36 601 *Marine Science* (2022).
- 37
38 602 [20] Antoine Guisan and Niklaus E. Zimmermann. “Predictive habitat distribution mod-
39 603 els in ecology”. In: *Ecological modelling* 135.2-3 (2000). Publisher: Elsevier, pp. 147–
40 604 186.
- 41
42 605 [21] Stephanie E Hampton et al. “Big data and the future of ecology”. In: *Frontiers in*
43 606 *Ecology and the Environment* 11.3 (2013), pp. 156–162.
- 44
45 607 [22] Trevor J Hefley, Brian M Brost, and Mevin B Hooten. “Bias correction of bounded
46 608 location errors in presence-only data”. In: *Methods in Ecology and Evolution* 8.11
47 609 (2017), pp. 1566–1573.
- 48
49 610 [23] Niels T. Hintzen et al. “VMStools: Open-Source software for the processing, analysis
50 611 and visualisation of fisheries logbook and VMS data”. In: *Fisheries Research* 115
51 612 (2012). Publisher: Elsevier, pp. 31–43.
- 52
53 613 [24] Niels T. Hintzen et al. “VMStools: Open-source software for the processing, analysis
54 614 and visualisation of fisheries logbook and VMS data”. In: *Fisheries Research* 115
55 615 (2012). Publisher: Elsevier, pp. 31–43.

- 1
2
3
4
5
6
7 616 [25] ICES. *Report of the Working Group for the Bay of Biscay and the Iberian Waters*
8 617 *Ecoregion (WGBIE)*. Tech. rep. Copenhagen, Denmark, 2018, p. 642.
- 9 618 [26] ICES. *Report of the Working Group on Beam Trawl Surveys (WGBEAM)*. en. Tech.
10 619 rep. Galway, Ireland, 2018, p. 121.
- 12 620 [27] Nick JB Isaac et al. “Data integration for large-scale models of species distributions”.
13 621 In: *Trends in ecology & evolution* 35.1 (2020), pp. 56–67.
- 15 622 [28] Yongku Kim and L Mark Berliner. “Change of spatiotemporal scale in dynamic
16 623 models”. In: *Computational Statistics & Data Analysis* 101 (2016), pp. 80–92.
- 18 624 [29] Kasper Kristensen et al. “TMB: Automatic Differentiation and Laplace Approxima-
19 625 tion”. English. In: *Journal of Statistical Software* 70.1 (Apr. 2016), pp. 1–21. ISSN:
20 626 1548-7660. DOI: 10.18637/jss.v070.i05. URL: [https://www.jstatsoft.org/](https://www.jstatsoft.org/index.php/jss/article/view/v070i05)
21 627 [index.php/jss/article/view/v070i05](https://www.jstatsoft.org/index.php/jss/article/view/v070i05) (visited on 01/24/2020).
- 23 628 [30] Jean-Baptiste Lecomte et al. “Compound Poisson-gamma vs. delta-gamma to han-
24 629 dle zero-inflated continuous data under a variable sampling”. In: *L’Institut des Sci-*
25 630 *ences et Industries du Vivant et de l’Environnement (AgroParisTech)* (2013), p. 37.
- 27 631 [31] Finn Lindgren, H\aaavard Rue, and Johan Lindström. “An explicit link between
28 632 Gaussian fields and Gaussian Markov random fields: The stochastic partial dif-
29 633 ferential equation approach”. In: *Journal of the Royal Statistical Society: Series B*
30 634 *(Statistical Methodology)* 73.4 (2011). Publisher: Wiley Online Library, pp. 423–498.
- 32 635 [32] Aurore Maureaud et al. “Are we ready to track Climate-driven shifts in marine
33 636 species across international boundaries? - A global survey of scientific bottom trawl
34 637 data”. English. In: *Global Change Biology* (Oct. 2020). tex.options: useprefix=true,
35 638 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).
- 38 640 [33] David AW Miller et al. “The recent past and promising future for data integration
39 641 methods to estimate species’ distributions”. In: *Methods in Ecology and Evolution*
40 642 10.1 (2019), pp. 22–37.
- 42 643 [34] Maryan Morel et al. “ConvSCCS: convolutional self-controlled case series model for
43 644 lagged adverse event detection”. In: *Biostatistics* 21.4 (2020), pp. 758–774.
- 45 645 [35] Andrew S Mugglin, Bradley P Carlin, and Alan E Gelfand. “Fully model-based
46 646 approaches for spatially misaligned data”. In: *Journal of the American Statistical*
47 647 *Association* 95.451 (2000), pp. 877–887.
- 49 648 [36] Lee G. Murray et al. “The effectiveness of using CPUE data derived from Vessel
50 649 Monitoring Systems and fisheries logbooks to estimate scallop biomass”. In: *ICES*
51 650 *Journal of Marine Science* 70.7 (2013), pp. 1330–1340.
- 53 651 [37] Ran Nathan et al. “Big-data approaches lead to an increased understanding of the
54 652 ecology of animal movement”. In: *Science* 375.6582 (2022), eabg1780.

- 1
2
3
4
5
6
7 653 [38] Krishna Pacifici, Brian J Reich, David AW Miller, Beth Gardner, et al. “Integrating
8 654 multiple data sources in species distribution modeling: a framework for data fusion”.
9 655 In: *Ecology* 98.3 (2017), pp. 840–850.
- 10 656 [39] Krishna Pacifici, Brian J Reich, David AW Miller, and Brent S Pease. “Resolving
11 657 misaligned spatial data with integrated species distribution models”. In: *Ecology*
12 658 100.6 (2019), e02709.
- 13
14 659 [40] Ryan J Parker, Brian J Reich, and Stephan R Sain. “A multiresolution approach
15 660 to estimating the value added by regional climate models”. In: *Journal of Climate*
16 661 28.22 (2015), pp. 8873–8887.
- 17
18 662 [41] Benjamin Planque et al. “Understanding what controls the spatial distribution of
19 663 fish populations using a multi-model approach”. English. In: *Fisheries Oceanography*
20 664 20.1 (2011), pp. 1–17. ISSN: 1365-2419. DOI: 10.1111/j.1365-2419.2010.00546.x.
21 665 URL: [https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2419.](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2419.2010.00546.x)
22 666 2010.00546.x (visited on 08/31/2020).
- 23
24 667 [42] Brian J Reich, Howard H Chang, and Kristen M Foley. “A spectral method for
25 668 spatial downscaling”. In: *Biometrics* 70.4 (2014), pp. 932–942.
- 26
27 669 [43] Ian W Renner, Julie Louvrier, and Olivier Gimenez. “Combining multiple data
28 670 sources in species distribution models while accounting for spatial dependence and
29 671 overfitting with combined penalized likelihood maximization”. In: *Methods in Ecology and Evolution* 10.12 (2019), pp. 2118–2128.
- 30
31 672
32 673 [44] Brian L. Sullivan et al. “The eBird enterprise: An integrated approach to develop-
33 674 ment and application of citizen science”. In: *Biological Conservation* 169 (Jan. 1,
34 675 2014), pp. 31–40. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2013.11.003. URL:
35 676 <https://www.sciencedirect.com/science/article/pii/S0006320713003820>
36 677 (visited on 04/19/2022).
- 37
38 678 [45] James T. Thorson. “Three problems with the conventional delta-model for biomass
39 679 sampling data, and a computationally efficient alternative”. In: *Canadian Journal*
40 680 *of Fisheries and Aquatic Sciences* 75.9 (2018). Publisher: NRC Research Press,
41 681 pp. 1369–1382.
- 42
43 682 [46] Jonathan Wakefield and Hilary Lyons. “Spatial aggregation and the ecological fal-
44 683 lacy”. In: *Handbook of spatial statistics* 541 (2010), p. 558.
- 45
46 684 [47] Christopher K Wikle and L Mark Berliner. “Combining information across spatial
47 685 scales”. In: *Technometrics* 47.1 (2005), pp. 80–91.
- 48
49 686 [48] Christopher K Wikle, Andrew Zammit-Mangion, and Noel Cressie. *Spatio-temporal*
50 687 *Statistics with R*. Chapman and Hall/CRC, 2019.
- 51
52 688 [49] Linda J Young and Carol A Gotway. “Linking spatial data from different sources:
53 689 the effects of change of support”. In: *Stochastic Environmental Research and Risk*
54 690 *Assessment* 21.5 (2007), pp. 589–600.

691 **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.
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_{ai}	99.668
10	1	Yes	Y_{ai}^r	0.333
10	1	Yes	D_a	0.000
100	10	No	Y_{ai}	100.000
100	10	Yes	Y_{ai}^r	100.000
100	10	Yes	D_a	92.000
1000	100	No	Y_{ai}	100.000
1000	100	Yes	Y_{ai}^r	100.000
1000	100	Yes	D_a	97.333

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

Model	Likelihood level	Convergence (%)
Commercial model	Y_{ai}^r	100.000
Commercial model	D_a	75.377
Integrated model	Y_{ai}^r	100.000
Integrated model	D_a	76.382
Scientific model		100.000

692

For Review Only

693 **Figures**

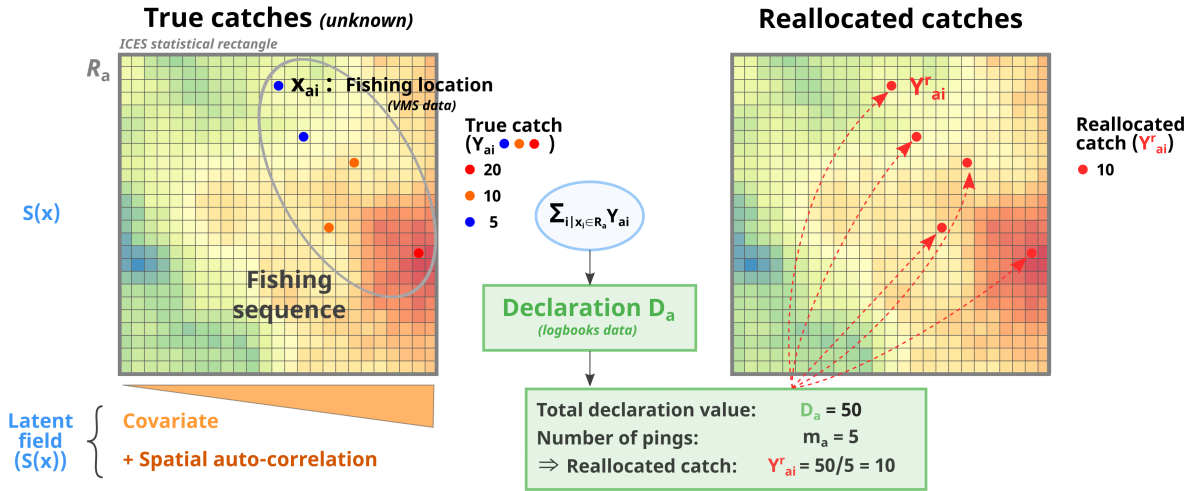


Figure 1: Schematic representation of the reallocation process. The biomass field (the background field) depends on a covariate and a spatial random effect. The covariate is the x axis. It has a positive effect on biomass values (i.e. biomass is higher on the right of the grid than on the left). The 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 a and are summed to constitute the declaration $D_a = 50$. The declaration is declared at the level of the statistical rectangle. From VMS data, we know the fishing positions x_{ai} . In standard processing, D_a are then uniformly reallocated over the fishing positions x_{ai} . This strongly homogenizes the catch. In particular, the effect of the habitat is no more evidenced in the reallocated catch Y_{ai}^r .

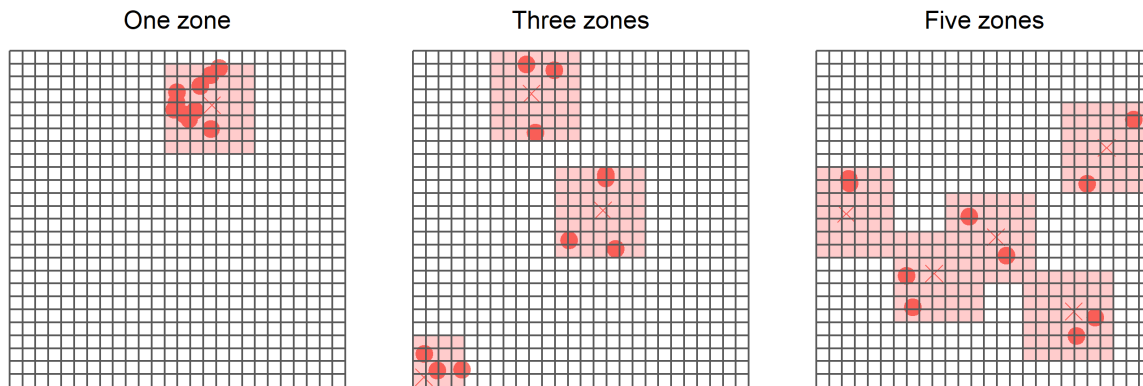


Figure 2: Simulations of 10 fishing points within 1, 3 and 5 fishing zones. The full grid corresponds to a statistical rectangle. Cross are the centroid of the fishing zones. A declaration declared at the level of the statistical rectangle would be uniformly reallocated over these fishing points.

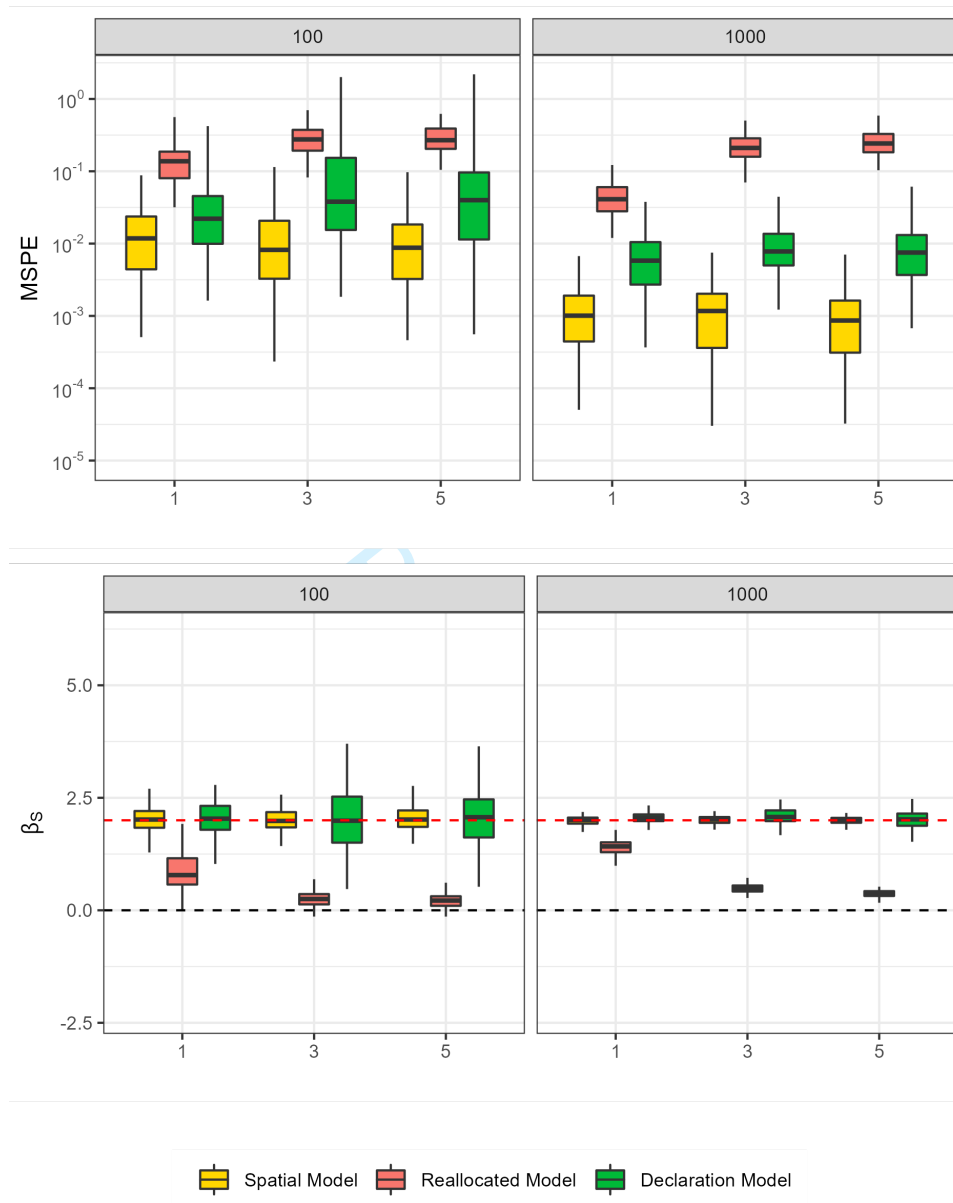


Figure 3: Performance metrics for single-square simulations with a total of 100 or 1000 fishing positions in columns. $MSPE = \frac{\sum_j^{n_{cells}} (S(x_j) - \hat{S}(x_j))^2}{n}$ is the mean squared prediction error and $\hat{\beta}_S$ is the species-habitat relationship parameter. The number of fishing 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.

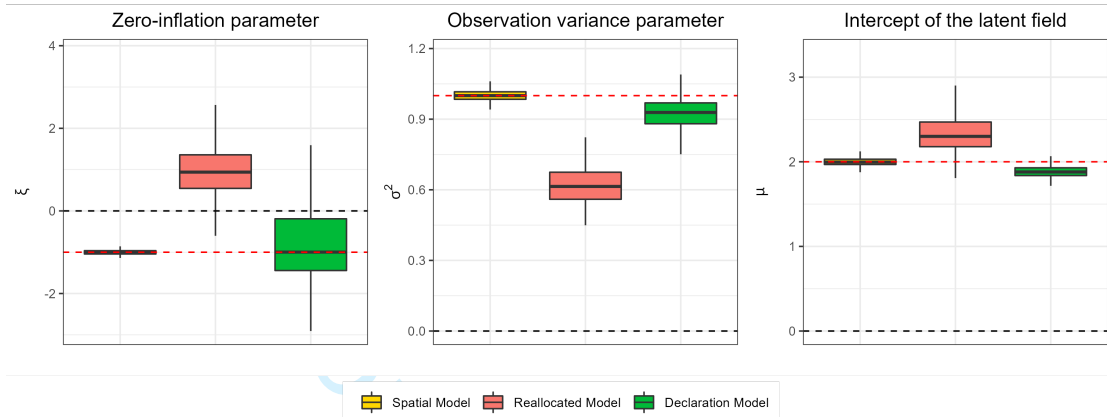


Figure 4: Parameters relative bias for single-square simulations. Only the simulations with 1000 fishing positions are represented. Black line: zero value. Red line: parameter true value.

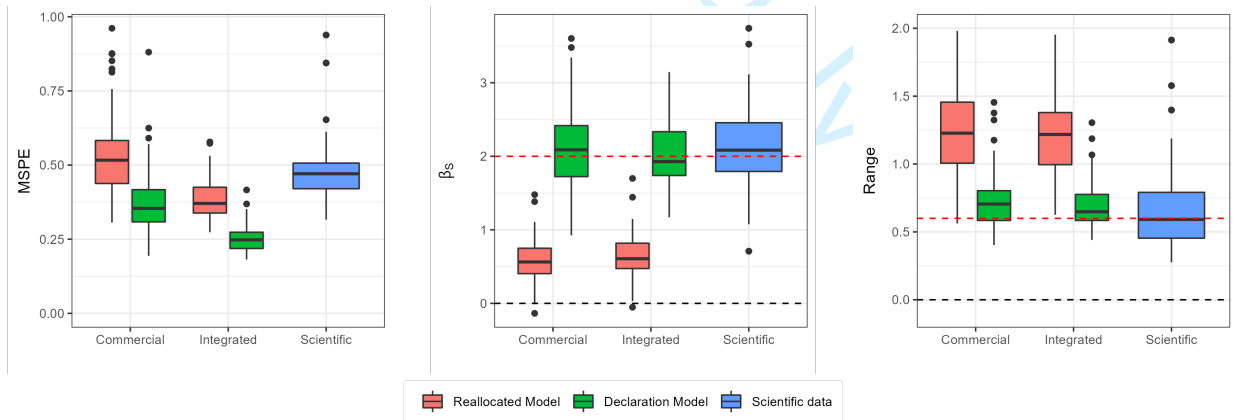


Figure 5: Performance metric for the multiple-square simulations. Red line: true value for the range and the species-habitat parameter (β_S). Blue: scientific-based model.

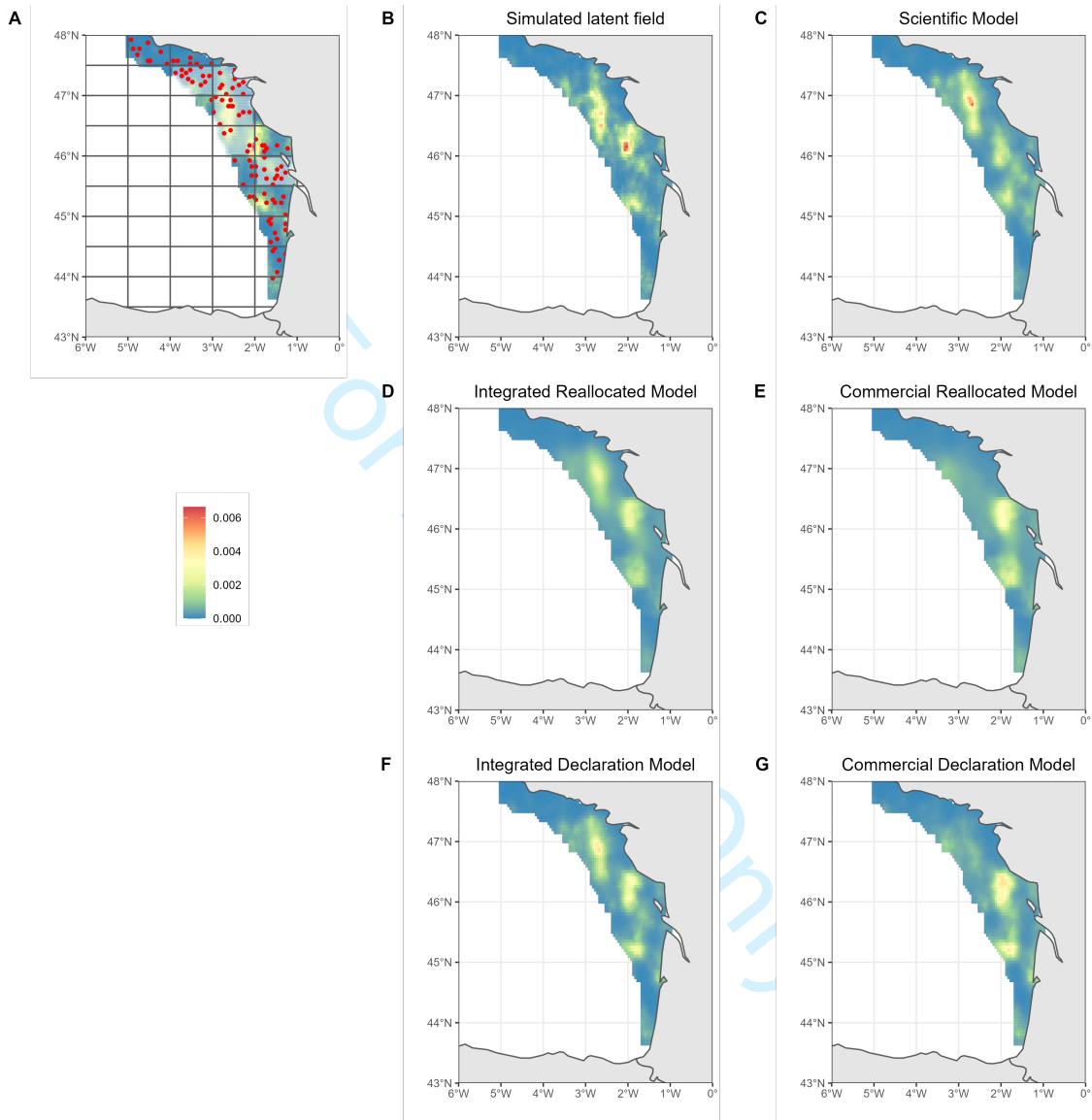


Figure 6: Distribution of simulated/estimated biomass field. A: Simulated biomass field with scientific samples (red) and statistical rectangles. The rectangles that have not been sampled by commercial data are the transparent rectangles. They represent $1/3$ of the full area. B: simulated biomass field.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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

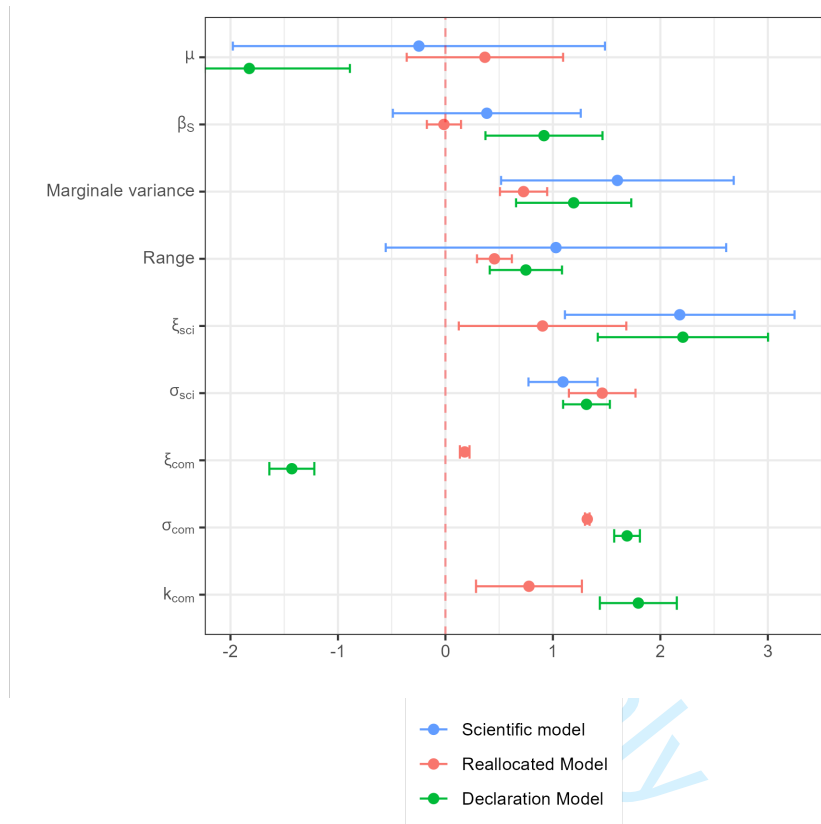


Figure 7: Parameters obtained with the model fitted on scientific data only, the integrated model fitted on reallocated catch Y_{ai}^r and the integrated model fitted on catch declarations D_a .

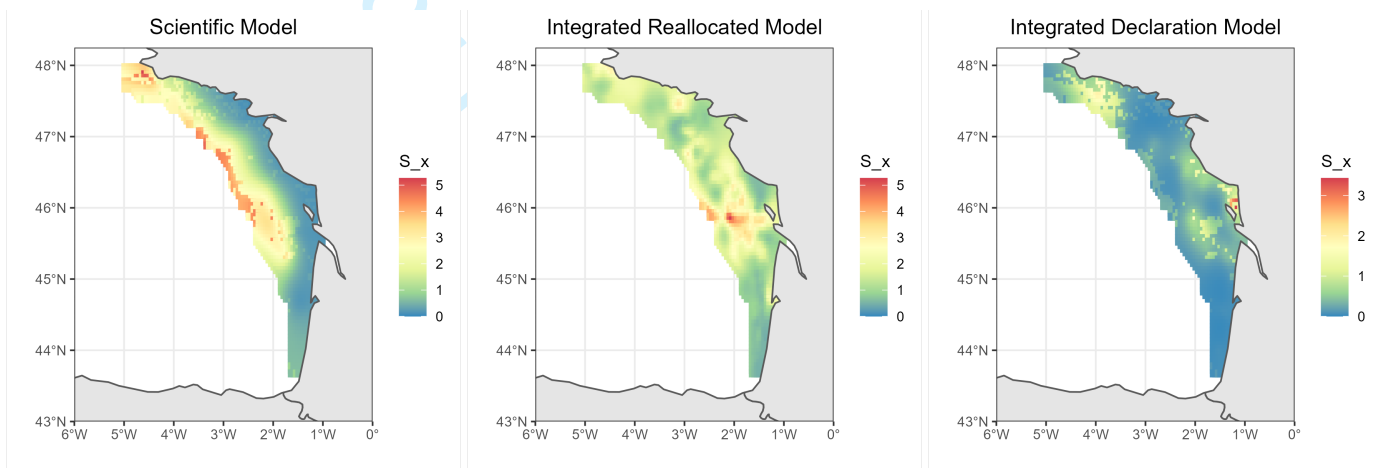


Figure 8: Maps obtained from the scientific-based model (left), the integrated model fitted on reallocated catch Y_{ai}^r (center), the integrated model fitted on catch declarations D_a (right).

Supplementary material

Reparameterization of the Lognormal distribution

The Lognormal distribution can be written as $Z \sim L(\rho; \sigma^2)$ 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 $\text{Var}(Z) = (e^{\sigma^2} - 1)e^{2\rho + \sigma^2}$.

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

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

Probability distribution and moments of declarations D_a

Probability distribution of individual observations Y_{ai}

We have to express the probability distribution of D_a and its moments as a function of Y_{ai} and its related moments. Let's assume $Y_{ai} = C_{ai} \cdot Z_{ai}$ is a zero-inflated Lognormal distribution with C_{ai} and Z_{ai} the two components of the mixture. C_{ai} is a binary random variable and Z_{ai} a Lognormal random variable.

$$C_{ai} | S(x_{ai}), x_{ai} \sim \mathcal{B}(1 - p_{ai})$$

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

$$Z_{ai} | S(x_{ai}), x_{ai} \sim L\left(\frac{S(x_{ai})}{1 - p_{ai}}, \sigma^2\right)$$

1
2
3
4
5
6
712 **Probability of obtaining a zero declaration**

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

$$\begin{aligned}\mathbb{P}(D_a = 0) &= \prod_{i=1}^{m_a} \mathbb{P}(Y_{ai} = 0 | S(x_{ai}), x_{ai}), \\ &= \exp \left\{ - \sum_{i=1}^{m_a} e^{\xi} \cdot S(x_{ai}) \right\} = \pi_a.\end{aligned}$$

715
716
717 **Expectation of a positive declaration**

718 Conditionally on \mathbf{S} and fishing positions x_{ia} .

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

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

$$\begin{aligned}\mathbb{E}(D_a | D_a > 0) &= (1 - \pi_a)^{-1} \mathbb{E}(D_a), \\ &= (1 - \pi_a)^{-1} \sum_{i=1}^{m_a} \mathbb{E}(C_{ai} Z_{ai}), \\ &= (1 - \pi_a)^{-1} \sum_{i=1}^{m_a} (1 - p_{ai}) \frac{S(x_{ai})}{1 - p_{ai}}, \\ &= (1 - \pi_a)^{-1} \sum_{i=1}^{m_a} S(x_{ai}).\end{aligned}$$

1
2
3
4
5
6 **Variance of a positive declaration**
7

8
9 721 The variance then can be expressed as:
10
11
12

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

13
14
15 722 with,
16
17
18
19

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

20
21
22
23
24
25 723 and
26
27
28
29

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

30
31
32
33
34
35 724 Then, using these two expressions in the variance formula gives:
36
37
38
39
40

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

41
42
43
44
45 725 As the $(Y_{ai})_{x_{ai} \in \mathcal{R}_a}$ are independent, $\text{Var}(D_a) = \sum_{i=1}^{m_a} \text{Var}(Y_{ai}) = \sum_{i=1}^{m_a} \text{Var}(C_{ai} \cdot Z_{ai})$.
46
47

48 726 Obtaining $\text{Var}(C_{ai} Z_{ai})$ is then straightforward due to conditional independence prop-
49
50 727 erties:
51
52
53
54
55
56
57
58
59
60

$$\begin{aligned}
\text{Var}(C_{ai}Z_{ai}) &= \mathbb{E}(C_{ai}^2Z_{ai}^2) - \mathbb{E}(C_{ai}Z_{ai})^2, \\
&= \mathbb{E}(C_{ai}^2)\mathbb{E}(Z_{ai}^2) - \mathbb{E}(C_{ai})^2\mathbb{E}(Z_{ai})^2, \\
&= (1 - p_{ai})\mathbb{E}(Z_{ai}^2) - (1 - p_{ai})^2\mathbb{E}(Z_{ai})^2, \\
&= (1 - p_{ai})(\text{Var}(Z_{ai}) + \mathbb{E}(Z_{ai})^2) - (1 - p_{ai})^2\mathbb{E}(Z_{ai})^2, \\
&= \frac{S(x_{ai})^2}{1 - p_{ai}}(e^{\sigma^2} - 1) + \frac{S(x_{ai})^2}{1 - p_{ai}} - S(x_{ai})^2, \\
&= \frac{S(x_{ai})^2}{1 - p_{ai}}(e^{\sigma^2} - (1 - p_{ai}))
\end{aligned}$$

1
2
3
4
5
6
7 **Sum up of the main formulas**

8
9 The main formulas can be summarised as follows:

10 n.b. all the formulas are conditioned on \mathbf{S} and on the fishing positions x_{ai} .

- 11
12
13 • The probability to obtain a zero declaration

14
15
16
17
$$\mathbb{P}(D_a = 0) = \exp \left\{ - \sum_{i=1}^{m_a} e^{\xi} \cdot S(x_{ai}) \right\} = \pi_a$$

- 18
19
20 • The expectancy of a positive declaration

21
22
23
$$\mathbb{E}(D_a | D_a > 0) = \frac{\sum_{i=1}^{m_a} S(x_{ai})}{1 - \pi_a}$$

- 24
25
26 • The variance of a positive declaration

27
28
29
$$\mathbb{V}ar(D_a | D_a > 0) = \frac{\sum_{i=1}^{m_a} \mathbb{V}ar(Y_{ai})}{1 - \pi_a} - \frac{\pi_a}{(1 - \pi_a)^2} \mathbb{E}(D_a)^2$$

- 30
31
32
33 • The variance of an individual observation

34
35
36
$$\mathbb{V}ar(Y_{ai}) = \frac{S(x_{ai})^2}{1 - p_{ai}} (e^{\sigma^2} - (1 - p_{ai}))$$

37
38
39 Then, assuming $D_a | D_a > 0$ also follows a Lognormal distribution we can write:

40
41
42
$$D_a | D_a > 0 \sim L(\mu_a = \mathbb{E}(D_a | D_a > 0), \sigma_a^2 = \ln(\frac{\mathbb{V}ar(D_a | D_a > 0)}{\mathbb{E}(D_a | D_a > 0)^2} + 1))$$