

# THESE DE DOCTORAT DE

L'INSTITUT AGRO RENNES - ANGERS

ECOLE DOCTORALE N° 598  
*Sciences de la Mer et du littoral*  
Spécialité : Halieutique

Par

**Baptiste Alglave**

**Inférer la distribution spatio-temporelle des espèces d'intérêt halieutique et identifier leurs habitats essentiels : modéliser l'échantillonnage préférentiel et le changement de support pour intégrer des sources de données hétérogènes**

Thèse présentée et soutenue à Rennes, le 9 décembre 2022

Unité de recherche : UMR DECOD

Thèse N° : H-117

## Rapporteurs avant soutenance :

Samuel Soubeyrand      Directeur de recherche, BioSP, INRAE, Avignon, France

Benjamin Planque      Directeur de recherche, Institute of Marine Research, Tromsø, Norway

## Composition du Jury :

Président : Samuel Soubeyrand

Clara Ulrich      Directrice scientifique adjointe, Ifremer, Nantes, France

Jan Jaap Poos      Professeur associé, Wageningen University and Research, Pays-Bas

Pierre Gloaguen      Professeur Associé, MIA, Agroparistech, Paris-Saclay, France

Directeur de thèse

Etienne Rivot      ICPEF, Institut Agro, Rennes, France

Co-directeur de thèse

Marie-Pierre Etienne      Professeur associé, Institut Agro, IRMAR, Rennes, France

## Invité(s)

Mathieu Woillez      Chargé de recherche, Ifremer, Brest, France

Youen Vermard      Chargé de recherche, Ifremer, Nantes, France

*"Soyons ce que nous sommes et soyons le bien!"*

# ACKNOWLEDGEMENT

---

Dure, dure de remercier en quelques mots tous ceux qui ont contribué de près ou de loin à ces trois années de thèse. D'abord parce que, mine de rien, on croise du monde en trois ans de thèse ! Et puis, en réalité, ce ne sont pas les trois dernières années qu'il faudrait regarder, mais au moins ces huit longues et (très) belles années d'étude. Enfin parce qu'il faut trouver les mots justes et ça n'est pas si simple... Mais il faut quand même tenter le coup ! J'espère que chacun se retrouvera dans les paragraphes qui suivent. Et dans le cas contraire... et ben vous pouvez toujours m'envoyer un mail pour vous plaindre, ça sera l'occasion de taper la discute ;)

D'abord un immense merci à mes encadrants : Etienne, Marie, Mathieu et Youen. Quatre, c'était le nombre juste et je pense qu'ensemble on a su trouver un bel équilibre tout au long de la thèse. Merci pour votre confiance, pour vos feedbacks toujours pertinents et pour la liberté que vous m'avez laissée. J'ai beaucoup appris sur le plan scientifique et humain et j'espère pouvoir redonner un jour tout ce que vous m'avez donné.

Merci aussi aux membres du jury de thèse (Samuel, Benjamin, Jan Jaap, Pierre et Clara) et aux membres du comité de suivi (Éric, Thomas, Olivier et Sandrine). Les échanges ont été riches et constructifs, j'espère qu'on se recroisera un jour ou l'autre pour discuter science (mais pas que !)

Merci ensuite à tous ceux que j'ai côtoyés au labo (permanents et non-permanents), dans toutes les conférences et autres rencontres scientifiques. J'ai eu la chance de profiter des labos de Rennes et de Nantes, de passer à Brest, de participer aux sessions d'été de 'State Of The R' et de faire le GdR Ecostat à trois reprises. S'il y a bien un point commun entre tous ces lieux, c'est la bonne ambiance qui s'en dégage. Surtout, gardez-là, c'est ça qui donne du goût ! Un merci particulier à ceux qui nous ont formés, moi et les halieutes qui sont passés par la formation. Vous savez transmettre la petite étincelle qui vous anime et c'est une grande richesse.

Il ne faut pas oublier le DTU-Aqua ! Pour l'accueil de Kasper et la bienveillance de Vanessa. Je crois que je peux aussi remercier Youen une deuxième fois. Malgré le COVID et avec un peu de patience, on a réussi à faire cette mobilité et les choses se sont bien goupillées.

Merci à toutes les bonnes bandes de copains que j'ai pu côtoyer pendant ces années de thèse : les maraudeurs du Secours Cath' et de Solidarité Dans la Rue, la coloc' de Nantes, les équipes Magis, les compagnons de bar du dimanche soir. Merci aussi à mes locataires successifs (Philippe Rio et Philippe Hébert), pour leur flexibilité et la générosité de leur accueil.

À tous les amis que j'ai rencontrés pendant les études : à Janson, à l'agro et en spé halieut'. Il y a du monde que j'ai perdu de vue, d'autres avec qui je garde contact, d'autres qui refont surface (concours de la providence !). Je garde des souvenirs très heureux de ces années et j'espère qu'on se recroisera au gré des courants.

Enfin à ma famille qui a été un soutien permanent depuis le début de mon parcours et sans qui rien de tout cela n'aurait été possible.

Le départ à Seattle approche. Je vois cette étape comme la fin d'un chapitre très dense, mais aussi comme le début d'un autre chapitre. Difficile de se projeter, il reste beaucoup d'éléments incertains et impossible de dire comment tout ce qui s'est construit durant ces années va s'agencer.

Ce qui est certain par contre, c'est que tout ce qui a été vécu jusqu'à maintenant a eu beaucoup de saveur et tous ceux que j'ai rencontrés y sont pour beaucoup. Je continue la route et j'espère tous vous recroiser à un moment ou à un autre.

Bon vent à tous et à la prochaine !



## Résumé

**Mots clés :** modélisation spatiale et spatio-temporelle, modèle hiérarchique, intégration de données, échantillonnage préférentiel, changement de support, zone fonctionnelle halieutique.

Des sources de données massives et hétérogènes deviennent disponibles dans différents domaines de l'écologie spatiale. Ces données ouvrent des opportunités pour inférer la distribution des espèces à une résolution spatio-temporelle fine à condition de développer des méthodes pour intégrer ces différentes sources de données.

En halieutique, de nouvelles sources de données sont accessibles via le système de suivi de l'activité de pêche (AIS ou VMS) et via les déclarations des pêcheurs. Dans le cadre de l'aménagement de l'espace marin, ces sources de données sont utilisées pour cartographier l'effort de pêche et la distribution des captures. Plus récemment, leur utilisation a été évoquée pour cartographier la distribution des espèces d'intérêt halieutique et pour identifier leurs habitats essentiels. Une bonne connaissance de ces habitats est cruciale pour assurer le renouvellement des espèces, mais leur localisation (et plus particulièrement celles des frayères) reste mal connue pour de nombreuses espèces exploitées. Ces nouvelles sources de données massives pourraient venir compléter l'information disponible via les sources de données standards utilisées pour cartographier ces espèces.

Les données de référence pour cartographier la distribution des espèces d'intérêt halieutique et identifier leurs habitats essentiels sont issues de campagnes scientifiques. Elles bénéficient d'un plan d'échantillonnage standardisé, elles couvrent un domaine spatial large et elles sont considérées comme des données de bonne qualité. Toutefois, elles nécessitent des moyens humains et matériels lourds et par conséquent elles ont lieu une à deux fois dans l'année pas nécessairement sur la période de reproduction. Par ailleurs, le nombre d'échantillons récoltés reste faible ce qui ne permet pas de prédire avec précision la distribution des espèces.

Les données d'observateurs embarqués permettent de compléter l'information disponible grâce aux campagnes scientifiques en fournissant des observations directes des captures des pêcheurs sur l'ensemble de l'année. Toutefois, le nombre d'échantillons récoltés reste faible et est dans le même ordre de grandeur que les données de campagnes. Par conséquent, les modèles qui sont ajustés à ces données sont construits à une résolution temporelle grossière (e.g. le trimestre) pour augmenter le nombre d'échantillons disponible par pas de temps et les prédictions spatiales obtenues sont peu précises. Les campagnes scientifiques et les données d'observateurs embarqués sont donc limitées pour identifier les frayères des espèces d'intérêt halieutique.

Des sources de données massives deviennent progressivement disponibles et ouvrent des perspectives pour cartographier la distribution des espèces à une résolution spatio-temporelle fine. Les données de déclarations de pêche (logbooks) sont des bases de données où sont répertoriées les déclarations de capture des pêcheurs. Elles sont définies à une résolution spatiale grossière *i.e.* la résolution des rectangles statistiques du CIEM (0.5° de latitude sur 1° de longitude). Pour affiner leur résolution spatiale, les données logbooks sont croisées avec les données de position des navires de pêche disponibles via le système de suivi des navires de pêche (VMS - Vessel Monitoring System). En identifiant les points VMS en pêche, en les reliant à la donnée de déclaration correspondante et en réallouant la déclaration sur les points VMS, il est possible d'obtenir une image très fine de la distribution des débarquements dans l'espace et dans le temps. Le croisement de ces deux jeux de

données produit un nouveau jeu de données (la donnée ‘VMS x logbook’) ayant une densité d’échantillonnage supérieure aux données scientifiques et aux données d’observateurs en mer. Elle pourrait être très informative de la distribution des espèces.

Cette thèse à un **double objectif** :

- Un objectif méthodologique : développer un modèle statistique spatio-temporel qui permet d’inférer la distribution des espèces d’intérêt halieutique en combinant les données ‘VMS x logbook’ et les données scientifiques. Le modèle doit prendre en compte l’échantillonnage préférentiel des pêcheurs et le changement de support (i.e. les différences de résolution spatiale) entre les différents jeux de données.
- Un objectif écologique : produire des cartes précises de la distribution des espèces et, sur la base de ces cartographies, identifier les habitats essentiels des espèces d’intérêts halieutiques.

Une **introduction méthodologique** présente les éléments de modélisation spatio-temporels à la base des modèles développés dans cette thèse ainsi que les outils d’inférence utilisés pour estimer les paramètres de ces modèles. Cette partie développe les bases de la modélisation hiérarchique dans un cadre spatio-temporel et l’approche par maximum de vraisemblance utilisé pour estimer les paramètres des modèles. Elle détaille plus précisément les méthodes d’approximation qui permettent d’estimer les paramètres et les effets aléatoires (l’approximation de Laplace et l’approche SPDE).

La thèse s’articule ensuite autour de 3 articles qui développent progressivement le modèle spatio-temporel et présente ses applications.

Le **premier article** présente les bases méthodologiques d’un modèle spatial combinant données ‘VMS x logbook’ et données scientifiques et prenant en compte l’échantillonnage préférentiel des pêcheurs. Un premier volet du chapitre présente des résultats de simulations/estimations. Il illustre la contribution des différents jeux de données dans l’inférence et l’effet de l’échantillonnage préférentiel sur les sorties du modèle.

Sur la base de ces simulations, nous démontrons que l’échantillonnage préférentiel doit être pris en compte dans la méthode d’inférence lorsque l’échantillonnage préférentiel est fort. Lorsque les données commerciales sont plus volumineuses que les données scientifiques, les données scientifiques apportent peu d’informations aux prédictions spatiales dans les zones échantillonées par les données commerciales, mais apportent de l’information dans les zones faiblement échantillonées par les données commerciales. Elles fournissent également un jeu de données de validation pour évaluer la consistance du modèle intégré avec les données scientifiques.

Le modèle est appliqué à trois espèces démersales du Golfe de Gascogne présentant des configurations d’échantillonnage contrastées (le merlu – échantillonnage préférentiel faible, la sole - échantillonnage préférentiel modéré, les encornets - échantillonnage préférentiel fort). Les applications permettent également d’illustrer comment le modèle peut prendre en compte différentes flottilles avec des capturabilités et des niveaux d’échantillonnage préférentiel différents.

Le **deuxième article** reprend les développements méthodologiques du premier article et ajoute une dimension temporelle au modèle. Le modèle intègre les données scientifiques et les données ‘VMS x logbooks’ et permet d’inférer la distribution spatio-temporelle des espèces sur l’ensemble de l’année à un pas de temps mensuel en prenant en compte la variation temporelle de l’échantillonnage préférentiel. Le modèle est ajusté entre 2010 et 2018 sur 3 espèces du golfe de Gascogne pour

lesquelles les connaissances de leur zone de reproduction sont contrastées (la sole – informations disponibles via les campagnes de suivies d’œufs et de larves, le merlan – informations disponibles via les observations des individus matures au moment de la reproduction, les encornets – connaissance de la saison de reproduction uniquement). Les résultats illustrent l’apport de l’échantillonnage préférentiel dans le modèle ainsi que l’apport de l’intégration des différentes flottilles à l’inférence. Les prédictions spatiales permettent d’identifier les zones d’agrégation persistantes au moment de la saison de reproduction. Les zones d’agrégation coïncident avec les zones de reproduction identifiées pour la sole et le merlan et permettent d’identifier des zones d’agrégation persistantes pour l’encornet au moment de la reproduction.

Le **troisième article** explore différentes approches pour réconcilier les différences de résolution spatiales entre les différents jeux de données. Dans les deux premiers chapitres, le croisement des données de déclarations avec les données VMS fournit une source de donnée massive pour inférer la distribution spatio-temporelle des espèces d’intérêt halieutique (la donnée ‘VMS x logbook’). Cependant, le croisement des deux jeux de données se base sur l’hypothèse de réallocation uniforme des déclarations sur les points de pêche associés. Cette procédure est efficace et pragmatique, mais elle est susceptible de transformer la donnée et d’introduire un biais dans les estimations et les prédictions spatiales du modèle. Ce chapitre introduit un modèle intégré alternatif permettant de relâcher l’hypothèse de réallocation uniforme des captures sur les points VMS en prenant en compte le changement de support dans la méthode d’inférence. Dans ce chapitre, nous comparons par simulation/estimation l’approche par réallocation uniforme et l’approche prenant en compte le changement de support dans la méthode d’inférence.

La réallocation uniforme introduit un biais dans les estimations de la relation espèce-habitat et produit des cartes plus lisses que la distribution réelle de l’espèce. Le fait de prendre en compte le changement de support dans la méthode d’inférence permet d’estimer de façon non biaisée la relation espèce-habitat et de produire des prédictions spatiales plus proches de la distribution réelle de l’espèce. Les deux approches ont été appliquées sur la sole du Golfe de Gascogne et produisent des résultats consistants avec les simulations.

Le **dernier chapitre** de la thèse résume les conclusions principales de la thèse, présente les applications potentielles du cadre de modélisation et développe les extensions futures du modèle.

L’application la plus directe des sorties du modèle pour la gestion concerne l’identification de zones de fermeture spatiales ou spatio-saisonnieres afin d’orienter l’aménagement de l’espace marin. L’approche a déjà pu être appliquée dans le cadre d’un groupe de travail du CSTEP pour identifier des zones de fermetures potentielles pour des espèces démersales du Golfe du Lion.

L’approche pourrait être complexifiée pour modéliser la dynamique spatio-temporelle et/ou le mouvement des espèces à un pas de temps mensuel. Le processus d’échantillonnage pourrait également être complexifié pour prendre en compte des covariables influençant la distribution de l’effort de pêche (*e.g.* la distance à la côte), les processus d’inertie temporelle (*i.e.* la tradition) et la dépendance entre les points de pêche appartenant à la même trajectoire.

Enfin, l’extension du modèle à une approche multi-spécifique permettrait de modéliser le ciblage des pêcheurs vers un ensemble d’espèces plutôt que vers une seule espèce et de se rapprocher d’un cadre réaliste.

**En conclusion**, le modèle spatio-temporel développé dans cette thèse fournit une base solide pour identifier les habitats essentiels des espèces d'intérêt halieutique et pour étudier les patrons spatio-temporels qui structurent la distribution de ces espèces.

Dans l'ensemble, cette thèse illustre les enjeux méthodologiques liés à l'utilisation des données de déclarations de pêche pour inférer la distribution spatio-temporelle des espèces d'intérêt halieutique. En particulier le développement de modèles spatiaux qui prennent en compte la différence de résolution spatiale entre les différents jeux de données et leurs caractéristiques d'échantillonnage est capital pour inférer de façon précise la distribution des espèces.

Enfin, la contribution principale de ce travail va au-delà du domaine de l'halieutique et pourrait trouver une application dans d'autres domaines de la modélisation spatiale dès lors que les données sont agrégées dans l'espace et que l'échantillonnage des données est préférentiel (par exemple, en épidémiologie ou en science du climat). Le développement d'outils opérationnels pour rendre ces méthodes accessibles à la communauté scientifique est une nécessité pour une utilisation plus large des données déclaratives dans la recherche et l'expertise.

# TABLE OF CONTENTS

---

<b>1 General introduction</b>	<b>17</b>
1.1 Methodological challenges when analyzing declarative data to infer species distribution . . . . .	19
1.1.1 Methodological issues related to declarative data . . . . .	19
1.1.2 Strategy to overcome these methodological issues . . . . .	20
1.2 Enhancing the knowledge for Marine Spatial Planning . . . . .	23
1.2.1 Marine spatial planning and essential fish habitats . . . . .	23
1.2.2 The importance of scientific survey data . . . . .	25
1.2.3 Complementing scientific data with commercial data . . . . .	29
1.3 Integrating ‘VMS x logbook’ data with scientific survey data to infer fish spatial distribution . . . . .	33
<b>2 Statistical tools for spatial and spatio-temporal modeling</b>	<b>37</b>
2.1 Base definition of hierarchical models . . . . .	37
2.2 Hierarchical models in a spatial and spatio-temporal context . . . . .	39
2.3 Inference in hierarchical models . . . . .	42
2.3.1 The likelihood: a tricky component of hierarchical models estimation	42
2.3.2 Approximation of the likelihood based on Laplace approximation .	44
2.3.3 Sparse representation of spatial random effects: the SPDE approach	45
<b>3 Combining scientific survey and commercial catch data to map fish distribution</b>	<b>51</b>
3.1 Introduction . . . . .	53
3.2 Material and methods . . . . .	56
3.2.1 Spatial integrated model . . . . .	56
3.2.2 Simulation-estimation experiments . . . . .	61
3.2.3 Case studies . . . . .	66
3.3 Results . . . . .	70
3.3.1 Simulations . . . . .	70

---

**TABLE OF CONTENTS**

---

3.3.2	Case studies . . . . .	77
3.4	Discussion . . . . .	81
3.4.1	Main findings . . . . .	81
3.4.2	Challenges in modeling preferential sampling . . . . .	82
3.4.3	Relative contribution of scientific and commercial data . . . . .	83
3.4.4	The limits of reallocated catch data . . . . .	84
3.4.5	Perspectives . . . . .	85
<b>4</b>	<b>Identifying mature fish aggregation areas during spawning season</b>	<b>87</b>
4.1	Introduction . . . . .	89
4.2	Material and methods . . . . .	91
4.2.1	Case studies . . . . .	91
4.2.2	Data . . . . .	92
4.2.3	Spatio-temporal integrated model . . . . .	94
4.2.4	Evaluating the interest of integrating multiple fleets . . . . .	98
4.2.5	Evaluating the value of modeling preferential sampling . . . . .	99
4.2.6	Investigating spatio-temporal dynamics and identifying reproduction grounds . . . . .	99
4.3	Results . . . . .	101
4.3.1	Assessing the contribution of each data sources to inference . . . . .	101
4.3.2	Interpreting estimates of preferential sampling intensity . . . . .	105
4.3.3	Evaluating the influence of preferential sampling on spatial distribution . . . . .	109
4.3.4	Investigating spatio-temporal dynamics of fish biomass . . . . .	109
4.3.5	Aggregation index and reproduction grounds . . . . .	110
4.4	Discussion . . . . .	114
4.4.1	Main findings . . . . .	114
4.4.2	Combining our results with other data sources to refine the identification of spawning grounds . . . . .	116
4.4.3	Limits and perspectives for the approach . . . . .	117
4.4.4	Future use for Marine Spatial Planning . . . . .	119
<b>5</b>	<b>Inferring fine scale wild species distribution from spatially aggregated data</b>	<b>121</b>
5.1	Introduction . . . . .	123

---

TABLE OF CONTENTS

---

5.1.1	Context . . . . .	123
5.1.2	The change of support issue . . . . .	124
5.1.3	Focus of the paper . . . . .	126
5.2	A spatialized catch model for aggregated data . . . . .	127
5.3	Simulation studies . . . . .	132
5.3.1	Single-square simulations . . . . .	133
5.3.2	Multiple-square simulations . . . . .	138
5.4	Case-study: sole of the Bay of Biscay . . . . .	143
5.5	Discussion . . . . .	147
5.5.1	The benefit of a statistical approach for change of support . . . . .	147
5.5.2	The hierarchical structure of the approach and the punctual observation layer . . . . .	148
5.5.3	Future perspectives for the framework . . . . .	149
<b>6</b>	<b>Discussion</b>	<b>151</b>
6.1	Challenge of data integration . . . . .	152
6.1.1	Integrating highly unbalanced datasets . . . . .	152
6.1.2	Challenge of change of support . . . . .	156
6.2	Enhancing integrated ecosystem assessment . . . . .	157
6.2.1	Identifying essential fish habitats . . . . .	157
6.2.2	Towards seasonal spatio-temporal population dynamics . . . . .	160
6.2.3	Perspectives for an implementation in Marine Spatial Planning . . . . .	161
6.3	Improving the realism of fishermen targeting behavior . . . . .	165
6.3.1	Modeling targeting behavior as multifactorial process . . . . .	165
6.3.2	Modeling fishing locations as a non-random trajectory . . . . .	166
6.3.3	Adapting the framework to mixed fisheries . . . . .	168
6.4	Conclusion . . . . .	170
<b>Bibliography</b>		<b>171</b>
<b>Supplementary material</b>		<b>194</b>
<b>A Statistical tools for spatial and spatio-temporal modeling</b>		<b>195</b>
A.1	Properties of spatio-temporal covariance functions . . . . .	195
A.2	Automatic differentiation . . . . .	196
A.3	Laplace approximation . . . . .	198

---

**TABLE OF CONTENTS**

---

A.4 Some remarks on the matrix of the SPDE approach . . . . .	199
<b>B Combining scientific survey and commercial catch data to map fish distribution</b>	<b>201</b>
B.1 Modeling framework . . . . .	201
B.1.1 Notations . . . . .	201
B.1.2 Simplification of the density function of an inhomogeneous Poisson point process on a discrete domain . . . . .	204
B.1.3 Targeting metric . . . . .	205
B.1.4 Using TMB for maximum likelihood estimation . . . . .	206
B.1.5 Consistency check . . . . .	207
B.2 Simulations material and methods . . . . .	209
B.2.1 Description of simulations . . . . .	209
B.3 Case studies material and methods . . . . .	212
B.3.1 Spatial grids . . . . .	212
B.3.2 Sampling intensity of European bottom trawl surveys . . . . .	213
B.3.3 Species and stock description . . . . .	214
B.3.4 Scientific data . . . . .	215
B.3.5 Commercial data . . . . .	216
B.3.6 Building commercial data . . . . .	217
B.3.7 Habitat covariates . . . . .	218
B.3.8 Effect of substrate and depth on scientific and commercial observations . . . . .	220
B.3.9 Fleet structure analysis: results of the PCA and the HCPC conducted on commercial data . . . . .	222
B.3.10 Goodness-of-fit and predictive capacity metrics . . . . .	226
B.4 Simulation results . . . . .	227
B.4.1 Relative bias of covariates effect estimates . . . . .	227
B.4.2 Consistency check for simulations-estimations . . . . .	228
B.4.3 Relative bias of range estimates when increasing sample size . . . . .	231
B.4.4 Comparison between simulations and predictions for strong preferential sampling ( $b = 3$ ) . . . . .	232
B.4.5 Fitting time for simulation-estimation . . . . .	235
B.5 Case studies results . . . . .	237

---

TABLE OF CONTENTS

B.5.1	Consistency check for each case study . . . . .	237
B.5.2	Species-habitat relationship . . . . .	238
B.5.3	Proportion of variance of the latent field ( $\log(S(x))$ ) explained by the random effect $\delta(x)$ and the covariates ( $\Gamma_S(x)^T \cdot \beta_S$ ) . . . . .	239
B.5.4	Spatial correlation parameter . . . . .	240
B.5.5	Information brought by commercial data . . . . .	242
B.5.6	Contribution of scientific data to the integrated model spatial predictions . . . . .	245
B.5.7	Targeting metric $T_j(x)$ maps . . . . .	246
B.5.8	Random effect $\eta_j(x)$ of the sampling process . . . . .	247
B.5.9	Proportion of variance of fishing intensity ( $\log(\lambda_j(x))$ ) explained by the random effect $\eta_j(x)$ and the preferential sampling term $b_j \cdot \log(S(x))$ . . . . .	248
B.5.10	Comparison of goodness-of-fit and predictive capacity metrics for models accounting or not for preferential sampling . . . . .	249
B.5.11	Comparison of goodness-of-fit and predictive capacity metrics for models considering 1 fleet or 2 distinct fleets . . . . .	251
B.5.12	Comparison of spatial prediction for models considering 1 fleet or 2 distinct fleets . . . . .	252
B.5.13	Maps and related uncertainty . . . . .	253
<b>C</b>	<b>Identifying mature fish aggregation areas during spawning season by combining catch declarations and scientific survey data</b>	<b>256</b>
C.1	Filtering the mature fraction from landings . . . . .	256
C.2	Discretization grid . . . . .	258
C.3	Survey sampling locations . . . . .	259
C.4	The SPDE approach . . . . .	260
C.5	Estimating the point process . . . . .	262
C.6	Maximum likelihood estimation . . . . .	262
C.7	Biomass predictions and related coefficient of variation for November 2018	263
C.8	Spatial predictions with and without PS (November 2018) . . . . .	264
C.9	Monthly average biomass predictions . . . . .	267
C.10	Persistence index maps . . . . .	269

---

**TABLE OF CONTENTS**

---

<b>D Inferring fine scale wild species distribution from spatially aggregated data</b>	<b>272</b>
D.1 Notations . . . . .	272
D.2 Reparameterization of the Lognormal distribution . . . . .	272
D.3 $D_k$ probability distribution and moments . . . . .	273
D.4 Probability of obtaining a zero declaration . . . . .	273
D.5 Expectation of a positive declaration . . . . .	275
D.6 Variance of a positive declaration . . . . .	275
D.7 Sum up of the main formulas . . . . .	277
<b>E Discussion</b>	<b>278</b>
E.1 The value of ‘VMS x logbook’ data to explore fish spatio-seasonal patterns and species phenology at fine spatio-temporal scale . . . . .	278
E.1.1 Material and methods . . . . .	278
E.1.2 Results . . . . .	283
E.1.3 Discussion . . . . .	292
E.2 Modeling the spatial distribution of the sardine ( <i>Sardina pilchardus</i> ) in the Bay of Biscay by integrating commercial and scientific data: challenges and limits . . . . .	295
E.3 Report of the ad-hoc contract for the preparation of STECF EWG 22-01 concerning closure areas to protect juveniles and spawners of all demersal stocks in western Mediterranean Sea . . . . .	387

# LIST OF ABBREVIATIONS

---

- AD: Automatic Differentiation  
COS: Change of Support  
CPUE: Catch Per Unit Effort  
CS: Citizen Science  
EOF: Empirical Orthogonal Functions  
GM(R)F: Gauss-Markov (Random) Field  
G(R)F: Gaussian (Random) Field  
HM: Hierarchical Model  
IM: Integrated Model  
INLA: Integrated Nested Laplace Approximation  
iSDM: integrated Species Distribution Model  
LPUE: Landing Per Unit Effort  
MCMC: Markov-Chain Monte Carlo  
MLE: Maximum Likelihood Estimation  
MPA: Marine Protected Areas  
MSP: Marine Spatial Planning  
MSPE: Mean Squared Prediction Error  
PCV: Predictive Cross Validation  
PS: Preferential Sampling  
pdf: probability density function  
SDM: Species Distribution Model  
SPDE: Stochastic Partial Differential Equation  
STECF: Scientific, Technical and Economic Committee for Fisheries  
TMB: Template Model Builder  
VMS: Vessel Monitoring System



# GENERAL INTRODUCTION

---

**New massive and highly diverse data** sources are becoming available in all fields of spatial ecology to infer spatial ecological processes and specifically to map species distribution. This is a common consensus that these data sources open new gates to unravel ecological processes that were not accessible to observation so far and to study ecological processes at finer spatial and temporal scale (Isaac et al., 2020).

For instance, the progress in wildlife tracking techniques through acoustic or satellite transmitters are progressively transforming animal movement ecology from a data-poor science to a data-rich one. This offers opportunities to investigate animal habitats preferences at large spatial scales with great precision (Hussey et al., 2015; Kays et al., 2015; Nathan et al., 2022). The development of new mobile phone applications for citizen science programs allows to collect data from volunteer observers for a scientific purpose at low cost to map birds, plants or any other species of interest (Sullivan et al., 2014; Dobson et al., 2020; Botella et al., 2021). Online survey databases allow to map fish distribution at an extent that was not possible before and offer opportunities to track the effect of climate change on biodiversity (Maureaud et al., 2020; Moriarty et al., 2020). Data from digital sensors or remote imagery enables to follow land uses, habitat suitability, ecological interactions, impact of climate change, response to anthropogenic disturbances, effect of conservation policies in near real time (Garrigues, Allard, and Baret, 2008; Constantin, Fauvel, and Girard, 2021; Poggi et al., 2021).

Among these databases, some are recorded for a clear scientific purpose with standardized protocol, in controlled experimental conditions through scientific recording systems (Nielsen, 2015; Farley et al., 2018; Isaac et al., 2020). These data are well suited for scientific analysis and they are so built that standard approaches allow to derive unbiased statistical estimators to describe the underlying process of interest (Cochran, 1977) e.g. abundance indices in the case of scientific survey data. These will be referred as **scientific data** in the following of this thesis. However, in many other cases, data arise from a more opportunistic observation scheme outside of a pre-established statistical observation

protocol and with a poor control over the sampling effort.

This is typically the case for **citizen science (CS) data** where volunteer agents record data for a scientific purpose – see for instance Sullivan et al. (2014) and Botella et al. (2021) for ecological applications. Usually, CS programs are deployed for a scientific use. Sampling agents – or at least the organizers of the CS programs – have a clear intention to use these data for scientific analysis. However, the sampling protocol may not be standardized and some bias may arise in the analysis from species misidentification or uncontrolled repartition of sampling effort (e.g., agents that may preferentially sample areas where they will find the specific outcome they are searching for). Huge literature on CS data lays basis to overcome the methodological issues inherent to these data (Conrad and Hilchey, 2011; MacPhail and Colla, 2020; Feldman et al., 2021).

A second classical example concerns data collected for routine, administrative or regulatory purposes. By contrast with CS, those data are generally not recorded per se for a scientific analysis. They will be referred as **declarative data** in the following. Typical examples include hunting records (Gilbert et al., 2021), administrative healthcare data (Morel et al., 2020) or fishermen catch declarations data and Vessel Monitoring System (VMS) (Bastardie et al., 2010; Gerritsen and Lordan, 2011; Hintzen et al., 2012).

Only few (almost none) publications in the statistical literature specifically address the challenges raised by the analysis of declarative data apart from very specific applications (e.g. in fisheries science as Bastardie et al. (2010) and Hintzen et al. (2012), in health science as Young and Gotway (2007), in terrestrial ecology as Gilbert et al. (2021)) or very specific issues (e.g. data aggregation - Wakefield and Lyons (2010)). Declarative data are poorly identified in systematic reviews and they are easily mixed up with other types of data such as CS data (see for instance Dobson et al. (2020)). This is probably a consequence of both:

1. the fact that they are not first dedicated to scientific analysis per se and they are characterized by huge heterogeneity regarding their type and structure which make them hard to ‘grip’.
2. the methodological difficulties when using these data for scientific analysis. From this point of view, they face similar methodological challenges as CS data (e.g. non-standardized sampling design), but they also face their own specific methodological challenges (e.g. data aggregation over rough spatial units).

Still, they are usually massive data and they are now ubiquitous in the field of spatial ecology but also in many other fields – e.g. economics, health science (Cressie and Wikle,

2015; Lokers et al., 2016). Provided that appropriate methods are developed to mitigate their methodological issues, declarative data could provide highly valuable information to infer spatial ecological processes and specifically species distribution (Dobson et al., 2020).

## 1.1 Methodological challenges when analyzing declarative data to infer species distribution

Inferring species distribution from non-standardized data such as declarative data raises strong methodological issues and requires to correctly integrate these in species distribution models (SDM) in order to provide unbiased estimates of the processes under study.

### 1.1.1 Methodological issues related to declarative data

Those data are often the reported results of standard action of the actors of a system. As so, they do not derive from a standardized well established sampling plan. Typically, they may only partially cover the area under study during specific periods (Fithian et al., 2015). Then, by **combining several data sources** – and specifically by combining non-standardized data sources with standardized ones -, one can expect to provide a wider and denser coverage of the study area and provide better inference of species distribution. Combining several data sources in inference is commonly referred as integrated statistical modeling.

As the result of standard searching behavior, the **sampling can be preferential** towards areas where the outcome under study is higher (Diggle, Menezes, and Su, 2010). This is typically the case of volunteers that target areas of higher species density (Warton, Renner, and Ramp, 2013) or fishermen that target areas of higher biomass (Pennino et al., 2019). This spatial targeting behavior can be challenging to handle in inference as it goes against a standard assumption made in spatial statistics where sampled locations  $\mathbf{X}$  is assumed independent from the process under study (denoted  $\mathbf{S}$  - we follow Diggle, Menezes, and Su (2010)). In such case, the joint probability  $\mathbb{P}[\mathbf{S}, \mathbf{X}]$  can be expressed as  $\mathbb{P}[\mathbf{S}, \mathbf{X}] = \mathbb{P}[\mathbf{S}] \cdot \mathbb{P}[\mathbf{X}]$  and conditioning on  $\mathbf{X}$  allows to use standard geostatistical approach to infer  $\mathbf{S}$  from observations (denoted  $\mathbf{Y}$ ). When spatial targeting towards  $\mathbf{S}$  arise,  $\mathbf{S}$  and  $\mathbf{X}$  cannot be assumed independent and  $\mathbb{P}[\mathbf{S}, \mathbf{X}] = \mathbb{P}[\mathbf{S}|\mathbf{X}] \cdot P[\mathbf{X}] \neq \mathbb{P}[\mathbf{S}] \cdot \mathbb{P}[\mathbf{X}]$ . Such process

is well known in the statistical literature and is often referred as preferential sampling (PS). If not properly accounted for, PS can lead to biased spatial predictions (Diggle, Menezes, and Su, 2010; Pati, Reich, and Dunson, 2011; Gelfand, Sahu, and Holland, 2012). Specifically, if areas of highest density are preferentially targeted, low-density areas may be overestimated if PS is not explicitly considered because the information from low-density areas is ignored in inference.

Declarative data are also often **aggregated over large administrative scales**. Typically, healthcare data (Wakefield and Lyons, 2010) or hunting harvest data are aggregated at the county level (Bauder et al., 2021). Fish catch declarations data are aggregated over rough statistical rectangles of size  $0.5^\circ \times 1^\circ$  (Hintzen et al., 2012). Such aggregation of the data may mask the processes that occur at fine scale and can potentially lead to what is called ‘ecological fallacy’ when used in inference to infer fine-scale processes (Wakefield and Lyons, 2010). Downscaling such information and reconciling data recorded at exact locations with data recorded at an aggregated level is a challenge often referred as the Change of Support (COS) problem (Cressie and Wikle, 2015; Gelfand, 2010).

Other issues can arise from this type of data, but they will not be treated in this manuscript. For instance, locations of the samples can be affected by location errors which can affect the species-habitat relationship (Hefley, Brost, and Hooten, 2017). Detection bias can arise too: detectability of a species may vary in space, time or following habitats leading to potential negative bias if the conditions imply poor detectability for the outcome under study (Coggins Jr, Bacheler, and Gwinn, 2014).

### **1.1.2 Strategy to overcome these methodological issues**

#### **Integrated spatial modeling**

Extensive literature exists regarding the combination of several data sources to infer the same ecological process (for instance in population dynamics modeling or spatial modeling – Zipkin and Saunders (2018)). These models are often referred as integrated models. In the context of SDM, these are often referred as **integrated Species Distribution Models (iSDM)**. iSDM integrate several data sources of different types (e.g. presence-only with presence-absence data - Gelfand and Shirota (2019)) with distinct sampling designs (Fletcher et al., 2019; Isaac et al., 2020; Lauret et al., 2021) to infer a single spatial field of species distribution and get more accurate predictions of species distribution than with single datasets. In some cases, these also integrate datasets that

allow to infer supplementary mechanistic processes in addition to species distribution. See for instance, Thorson et al. (2021b) who integrate mark-recapture data to model movement within a complex SDM or Cao et al. (2020) who model population dynamics of an harvested species (the Snow Crab of the Bering Sea) by integrating survey data and commercial catch data.

## Preferential sampling

**To handle PS**, Diggle, Menezes, and Su (2010) introduced a hierarchical framework where the sampling locations are defined conditionally on the latent field values and therefore contribute to inference through this dependence (see box below). In their seminal formulation, the authors consider a Gaussian Random Field (GRF) latent process  $\mathbf{S}$  with Matérn covariance  $M(x, x'; \kappa, \varphi)$  where  $M(\cdot, \cdot)$  controls the covariance of  $\mathbf{S}$  between the locations  $x$  and  $x'$  through the parameters  $\kappa, \varphi$ . Observations  $\mathbf{Y}$  are defined conditionally on  $\mathbf{S}$  and assumed to be Normally distributed and depends on a variance parameter  $\sigma^2$ . The originality comes from the sampling process  $\mathbf{X}$  which is modeled explicitly through an inhomogeneous Poisson point process where the intensity of the point process  $\lambda(x)$  is related to the latent field  $\mathbf{S}$ . That way, the sampling location  $\mathbf{X}$  contribute to the total likelihood of the model.

Several other developments and applications followed the publication of Diggle, Menezes, and Su (2010). For instance, we can mention the literature regarding PS in air pollution science (Shaddick and Zidek, 2014; Zidek, Shaddick, and Taylor, 2014; Shaddick, Zidek, and Liu, 2016; Watson, Zidek, and Shaddick, 2019). Other applications exist in econometrics (Paci et al., 2020) and phylodynamics (Karcher et al., 2016). In ecology, Gelfand and Shirota (2019) proposed a framework to fuse presence only data and presence/absence data while accounting for PS in the presence/absence dataset. Pennino et al. (2019) and Rufener et al. (2021) took up the same model construction as the one of Diggle, Menezes, and Su (2010) and applied it to fisheries data assuming fishermen locations were recorded following PS. Finally, Conn, Thorson, and Johnson (2017) proposed a generalized version of these models for ecological applications and fitted the model to aerial seal count data in the Eastern Bering Sea. Their framework allows (1) to handle discrete counts data while previous ones only suited for Gaussian responses and (2) to account for spatially varying PS parameter – although authors recognize this last extension leads to over parameterization of the model.

In their applications, most of the previous frameworks assume a relatively simplistic

representation of sampling effort where sampling intensity only depends on the process of interest and eventually on additional covariates. However, sampling locations (in particular in the case of fisheries dynamics) may depend on highly complex processes that are not all known or observable. Then, building a framework that is both generic and parsimonious while being able to capture and disentangle the effect of PS from the effect of other processes is a major challenge.

### Hierarchical model from Diggle, Menezes, and Su (2010)

#### Latent field equation

$$\mathbf{S} \sim GF(0, M(x, x'; \kappa, \varphi))$$

#### Sampling location equation

$$\mathbf{X} \sim \mathcal{IPP}(\lambda(x))$$

$$\log(\lambda(x)) = \alpha + \beta \cdot S(x)$$

#### Observation equation

$$\mathbf{Y} \sim \mathcal{N}(\mu + \mathbf{S}, \sigma^2)$$

**S:** spatial latent field,  $GF(, )$ : Gaussian Field,  $M(x, x'; \kappa, \varphi)$ : Matérn covariance function, **X**: fishing positions modeled as an inhomogeneous Poisson point process with intensity  $\lambda(x)$ .  $\alpha$ : intercept of the point process,  $\beta$ : PS parameter, **Y**: observations,  $\mu$ : intercept of the latent field,  $\sigma^2$ : observation variance.

## Change of support

Reconciling the spatial scales of several datasets and providing fine-scale predictions based on spatially aggregated data with various level of spatial aggregation is typically a problem that is referred as the ‘**change of support**’ (**COS**) issue in the statistical literature. From a generic point of view, COS refers to ‘the summary or analysis of spatial data at a scale different from that at which it was originally collected’ (Gelfand et al., 2010). It is also often referred as ‘upsampling/downscaling’ or ‘Modifiable Areal Unit Problem’ (Cressie and Wikle, 2015). This is typically the case where variables are

aggregated at specific scales and one would like to represent the distribution of these variables at a new level of aggregation (this can be either at larger or finer scale, the different blocks of aggregation can be nested or non-nested).

Overall, there is a wide statistical literature on COS (Matheron, 1985; Gelfand, Zhu, and Carlin, 2001; Gelfand et al., 2010; Wikle and Berliner, 2005) and several applications already exist in many fields of research e.g. health analysis – Young and Gotway (2007) –, air pollution – Berrocal, Gelfand, and Holland (2010a) –, climate science – Reich, Chang, and Foley (2014) – and ecology – Pacifici et al. (2019). Still, poor literature exists regarding COS in the case of complex data such as zero-inflated positive continuous data. Most publications do not address the issue as they assume simple observation processes for the aggregated data (Gamma distribution or Normal distribution for continuous data and Poisson distribution for count data). As declarative data become more accessible, there will be a growing need to handle COS for various types of data and possibly data that are more complex than Gaussian or Poisson observations e.g. mixture distribution such as zero-inflated lognormal observations.

*N.b.* Some more details on the mathematical background regarding hierarchical modeling, spatial models and inference methods are given in the chapter 2.

## 1.2 Enhancing the knowledge for Marine Spatial Planning

### 1.2.1 Marine spatial planning and essential fish habitats

In marine ecology and fisheries science, accurate spatial information is crucial to manage the marine space and ensure resource renewal. Sea ecosystems face strong pressures from multiple anthropogenic activities and there is a competition between marine sectors to exploit the marine space e.g. fishery, energy, shipping, conservation, gravel extraction, recreation (Bastardie et al., 2015; Campbell et al., 2014). **Marine spatial planning (MSP)** is precisely about finding a balance between all these activities and the ecosystems that face these anthropogenic pressures. MSP is often seen as a practical strategy to implement the ecosystem-based fisheries management framework (Qiu and Jones, 2013). Thoroughly selecting the areas that may be either open to fishing, closed to fishing, or protected to ensure the renewal of the resources requires an accurate knowledge of fish spatio-temporal dynamics (Fock, 2008; Janßen et al., 2018).

More specifically, improving our knowledge of **essential fish habitats** and of their functional role for individuals, populations and ecosystems is critical. Essential fish habitats are zones where fish spend at least one stage of their life cycle, these areas ensure the realization of either reproduction, growth, feeding or migration functions (Figure 1.1 - Brown et al. (2018)). To determine whether an essential habitat is important in the species life-cycle, Delage and Le Pape (2016) point out that the area should be restricted in space, should represent a high concentration of individuals for a specific life stage and should contribute in a consistent manner to the next stage. In temperate waters, the most important essential habitats are classically defined as nurseries, spawning areas, and also migration corridors for amphihaline species.

Specifically, **spawning areas** are crucial habitats for species renewal as the reproductive capacity of a population relies on the good functioning of the reproduction grounds (Seitz et al., 2014). Furthermore, reproduction is typically a period where the population is particularly vulnerable because large mature (and hence often high commercial value) individuals aggregate and strong overexploitation can arise from intensive targeting of fish spawning aggregation (Sadovy and Domeier, 2005; Biggs et al., 2021).

Still, spawning areas are often poorly known relative to other essential fish habitats (Di Stefano et al., 2022). Typically, ecology of coastal nursery grounds is relatively well known and their economic and ecological value for fisheries are unquestionable (Seitz et al., 2014; Brown et al., 2018). By contrast, spawning areas of marine fishes may be more subject to uncertainty regarding their spatial location and the temporal timing of the reproduction. For instance, Regimbart, Guitton, and Le Pape (2018) clearly outlines the lack of knowledge that remains regarding the spawning grounds in the French territorial waters; besides, many recent examples illustrate how spawning grounds still remained relatively uncertain until recently (Laptikhovsky et al., 2022; Di Stefano et al., 2022).

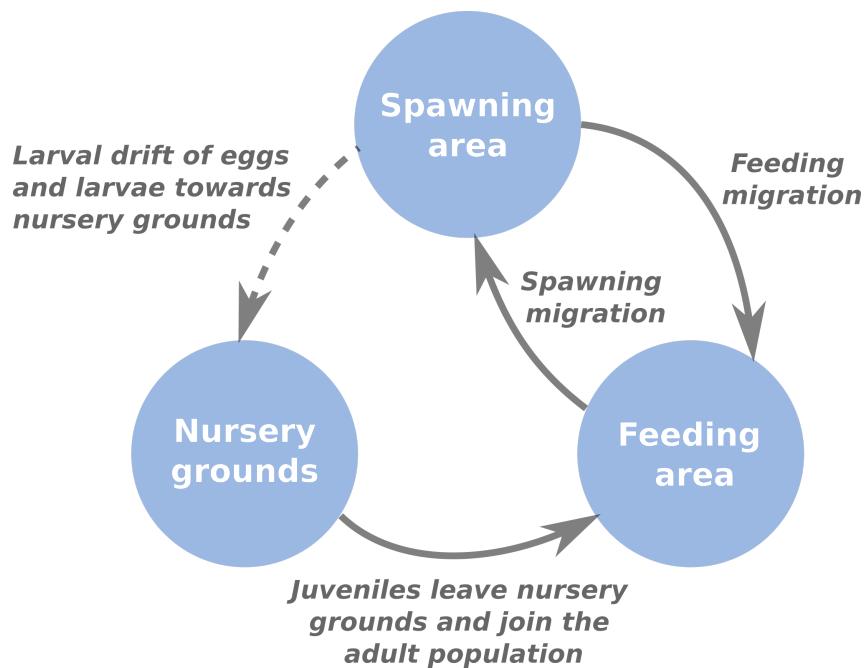


Figure 1.1 – Conceptual diagram of common life-history stages of fish in coastal habitats based on Brown et al. (2018). During reproduction, mature individuals aggregate and reproduce in spawning areas. After spawning, eggs and larvae are transported to nursery grounds through hydrodynamic circulation and active tidal migration. When reaching nurseries, larvae settle and metamorphose into juveniles. Juveniles usually stay one to several years in a nursery. When getting older, juveniles leave the nursery, mix with the adult population and become mature. Mature individuals alternate reproduction season and feeding season; if the reproduction and the feeding areas are distinct, the transition between these two periods are characterized by migrations between spawning areas and feeding areas.

### 1.2.2 The importance of scientific survey data

The reference data for mapping species distribution and identifying spawning grounds are mainly **scientific surveys**. Surveys typically benefit from a standardized sampling plan, a controlled protocol and an equipment that is held constant over the years justifying that the probability to catch a species conditionally on its presence (also called ‘catchability’ in fisheries science) does not change for technical reasons (Board and Council, 2000; Hilborn and Walters, 1992; Nielsen, 2015). Furthermore, they provide the exact locations of the record and they are typically designed to cover the full geographical extent

of the populations under study and are designed to compute unbiased abundance indices and spatial predictions (Rivoirard et al., 2008; ICES, 2015). Their selectivity is often minimized to sample as many species, size groups, and life stages as possible, they are independent from fishery data and thus they have a central role for fisheries ecology and fish stock assessment (both single and multispecies). Typically, they provide exhaustive and reference data on marine populations and allow to compute unbiased estimates on the variation of fish abundance in space and time (Nielsen, 2015). Classically, bottom trawl and beam trawls are dedicated to demersal species, and acoustic methods coupled with pelagic trawls are dedicated to pelagic fish populations. Surveys allow to map fish distribution at multiple scales (global scale - Maureaud et al. (2020); continental scale - Moriarty et al. (2020); local scale - Cariou et al. (2021)) and they are also keystone for studying marine communities and ecosystem food webs (Mérillet et al., 2022).

Regarding the identification of spawning grounds, systematic or punctual **surveys occurring during fish reproduction season** (e.g. see the PELGAS survey for a systematic survey - Doray et al. (2018) - or Arbault, Camus, and Bec (1986) for a punctual survey) typically gives access to spawning areas in both space and time. Observation can be either direct (egg and larvae) or indirect (observation of mature individuals - Doray et al. (2018) and Fox et al. (2008)). Lot of information can be gathered to infer fish reproduction biology and ecology through these surveys and they typically bring information to:

- study the timing of the reproduction and confront the outputs of the model to the available data on reproducing individual condition available through these surveys (Pecquerie, Petitgas, and Kooijman, 2009);
- understand the relationship between environmental variables and spawning habitats to identify spawning habitats (habitats where the environmental conditions are suitable for spawning) and realized spawning habitat (habitat where spawning actually occurs - Planque et al. (2011));
- assess the main spatio-temporal patterns that structure reproduction areas (Petitgas et al., 2020) when relatively long time series are available.

However, **scientific surveys also exhibit strong limits**. First, they often provide a relatively small number of samples with only a restricted temporal coverage due to high costs and heavy technical constraints (Nielsen, 2015). Typically, European bottom trawl surveys occur once or twice a year and provide a small number of samples for large areas (see for instance the Orhago survey on Figure 1.2 that occurs each November

along the Bay of Biscay). Because of the relatively low spatial density of sampling, the spatial predictions that can be derived from the surveys may be imprecise (ICES, 2005). Besides, in many cases these surveys mismatch the key life cycle stages and thus they are of limited use to identify fish essential habitats such as spawning grounds (Pennino et al., 2016; Hilborn and Walters, 1992).

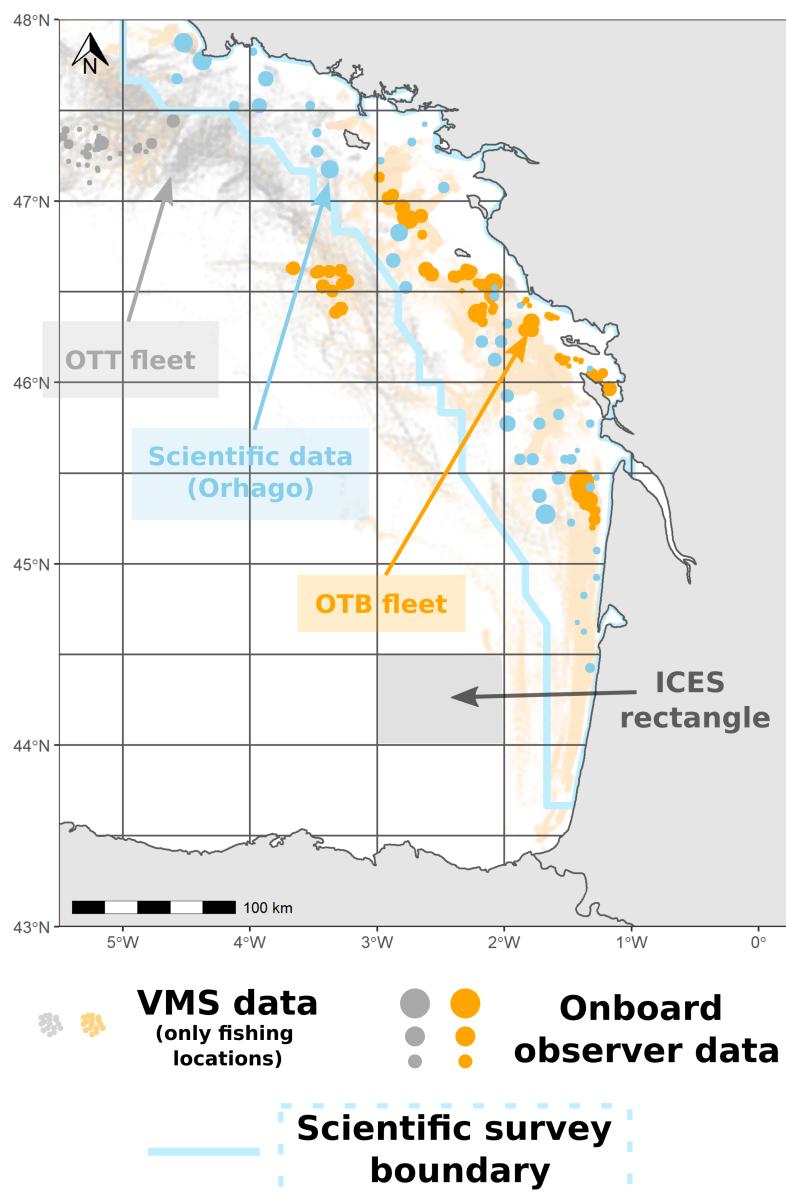


Figure 1.2 – Illustration of the multiple data sources available to map fish distribution in the Bay of Biscay. Data come from the last quarter of 2018. Survey data is the Orhago survey, a yearly beam trawl survey occurring in November that is specifically designed for sole (Coupeau and Biais, 2019). The point size of onboard observer data and scientific survey data corresponds to the catch value for common sole. Two commercial fleets are presented on the figure, the OTT fleet (otter trawls - grey) and the OTB fleet (bottom trawls - orange). Both have VMS records and onboard observer observations.

### 1.2.3 Complementing scientific data with commercial data

Recent analysis have focused on the use of **commercial data** (also called fishery-dependent data) to fill the gaps inherent to scientific data. While scientific data mainly provide information on the month of the survey, commercial fleet benefit from a continuous activity throughout the year. Using fishermen as samplers of fish distribution can provide information outside the time span of the survey and can bridge the gaps between the months of the surveys (Rufener et al., 2021).

#### Onboard observer data

The most common commercial data used to complement scientific data are **onboard observer data** (see Figure 1.2). These data are recorded on fishing vessels by onboard observers that record data through direct recording of the catches (i.e. counts/weights of the species caught and discarded) together with the exact location of the record. Then, they can be assumed to be as accurate as scientific data. Observer data can provide information over the full year for all caught species to study intra-annual spatial variability of both fish (Kai et al., 2017; Pinto et al., 2019) and discards (Stock et al., 2019; Yan et al., 2022), as well as to identify fish essential habitats (Pennino et al., 2013; Pennino et al., 2016).

Onboard observer data are complementary with scientific survey data; recently, Rufener et al. (2021) proposed a framework to combine scientific and on-board observer data to infer fish spatio-temporal distribution. Thanks to the fine demographic resolution of both data sources, the framework separately models the distribution of each age class while accounting for spatial and temporal correlations.

However, onboard observer data are also limited in sample size and in spatial and temporal coverage. They usually only represent a small proportion of all fishing trips. In France for instance, only 1% of all sea trips are covered by the French observer programs (Cornou et al., 2021). Then, the sampling intensity is limited and the number of observation are in the same order of magnitude as the survey data (Figure 1.2). When using these data in modeling, this often constrains to consider rough temporal scales such as semesters or quarters to improve the spatial coverage of the whole area (Kai et al., 2017; Pinto et al., 2019; Rufener et al., 2021). However, some lifecycle events such as the reproduction peak may be tighter than this temporal resolution (month or week resolution - Biggs et al. (2021)) which prevents to identify accurately spawning grounds from these

data.

Furthermore, they are fishery-dependent data that may be preferentially sampled in zones of higher biomass and PS needs to be accounted for in inference when using these data to map fish distribution (Pennino et al., 2019; Rufener et al., 2021).

### **'VMS x logbook' data as valuable massive data to map fish distribution**

Declarative data are progressively becoming available for fisheries research through **catch declarations data (logbook) and fishing geolocation available from VMS** (Hintzen, 2021).

Commercial catch declarations are intensive datasets regrouping the daily declarations from the fishermen at the scale of statistical rectangles (Hintzen, 2021). Besides, VMS provides information on the fishing locations of the fishermen (see Figure 1.2 - 1.3). By combining VMS data with logbook data, it is possible to refine the spatial resolution of the catches and produce fine scale maps of catch distribution (Hintzen et al., 2012).

These extensive datasets record data on fishing activity (both effort and catch) with a much better fleet coverage with high spatio-temporal resolution. This opens perspectives for mapping fish distribution at much finer spatio-temporal resolution than what is possible through either onboard or scientific surveys provided the methodological challenges inherent to the data are tackled.

#### **Logbook declaration data**

**Logbooks are the declared catch weights for all harvested species.** Logbook data are daily declared by fishermen at the resolution of ICES statistical rectangle (one longitudinal degree by 0.5 latitudinal degree, see Figure 1.2 - Hintzen et al. (2021)). First records of logbook data date back to the early 1960's with progressive implementation in Europe until their mandatory implementation in 1987 as a consequence of low fishermen compliance and important illegal practices (EC, 1987; Berg, 1999). Today, logbooks are typically used to check for quotas consumption and compliance to technical restrictions (e.g. mesh size). They also allow to map fishing effort and catch declarations at rough scale and they are used as inputs in stock assessment models.

Note that logbook are mainly representative of landings and not of the actual catch as discards are generally not reported in logbooks (even though they should since the Landings Obligation implementation - Lehuta and Vermaud (2022) and Ulrich (2021)). In the following, they will be referred either as catch declarations or landings, but we fully

acknowledge that a missing part of the catch are missing from landings data.

### **Vessel Monitoring System data**

**VMS data provide information on longitude, latitude and speed.** Each fishing location recorded is called a fishing ping. In Europe, ping intervals have to be less than 2 hours and the interval is set to one hour in France. VMS data were first introduced in the 1990s (EC, 1993). They have been mandatory for vessels superior to 24m in length from 2000 to 2004 before being extended to vessels superior to 15m from 2005 to 2011 (Hintzen, 2021). In Europe, they are now mandatory for all vessels over 12m. Today, VMS information are used for security to monitor illegal fishing activity and to cross-check the ICES rectangles recorded in logbook data. There are confidentiality constraints on these data; these constraints are progressively being relaxed as VMS data are progressively being used routinely.

Access to VMS data opened new gates to **explore the location of fishing** (Murawski et al., 2005; Fock, 2008; Stelzenmüller, Rogers, and Mills, 2008), the factors that underlies the spatial distribution of fishing (Tidd et al., 2015; Hintzen, 2021), the effect of trawling on the seabed (Hiddink et al., 2006) and the interactions between vessels and other activities (Poos and Rijnsdorp, 2007; Campbell et al., 2014; Bastardie et al., 2015).

They also offer the possibility to **study fishermen trajectories and behavior** by considering fishers are foragers that either forage (i.e. fishing activity) or search for food (i.e. steaming) (Vermaud et al., 2010; Walker and Bez, 2010; Gloaguen, Etienne, and Le Corff, 2018; Gloaguen, 2015). Consequently, these trajectories may reflect to some extent the spatial distribution of the resource (Bertrand et al., 2005) and the fishing locations can be assumed to arise from PS towards areas of higher biomass (Pennino et al., 2019). In addition, fishers' locational choices can arise from several other factors (Salas and Gaertner, 2004; Girardin et al., 2017). For instance, habitat strongly determines fishing locations: trawlers do not fish on rocky habitats and depth affects fishing locations (Hintzen, 2021). Fishermen can select fishing areas in such way to maximize catches or revenues (Eales and Wilen, 1986), fishing locations may depend on management closures as bycatch avoidance leading to large-scale shift in fishing effort and to potential modifications in catch composition (Abbott, Haynie, and Reimer, 2015). Logistical constraints (transit costs, spatial management, and sea/ice state, distance to harbor), tradition, information sharing and cooperation also affect the spatial distribution of fishing (Haynie, Hicks, and Schnier, 2009; Girardin et al., 2017). Disentangling the relationship between

targeting towards some harvested species and the other factors that structure fishing locations is a major challenge that the access to VMS data could unravel.

### **Combining logbooks and VMS data**

When combined, **VMS crossed with logbook data** (denoted ‘VMS x logbook’ data hereafter - Figure 1.3) **open perspectives for mapping catch and fishing effort** (Gerritsen and Lordan, 2011; Bastardie et al., 2010) **and potentially computing CPUE at fine spatial resolution to produce maps of species distribution**. Murray et al. (2013) demonstrated that CPUE obtained from ‘VMS x logbook’ data were consistent with abundance estimates from scientific survey data. Recently, Azevedo and Silva (2020) used similar data to produce high-resolution maps of species annual landings by commercial size category and age group. They highlighted that the different life stages had different distribution patterns and interpreted these shifts in regards to ontogenetic migrations of the species. Other technical reports used similar data to map species distribution in the Mediterranean Sea in order to identify potential closure areas for overexploited species (Billet et al., 2021). However, the temporal coverage and resolution of the studies remained relatively limited (only one time step was considered – either a month or a year) and the potential of the data was not fully exploited. Besides, only the raw discretized data were used to produce maps and no modelisation exercise was realized to (1) properly fit the raw data, (2) reconstruct species distribution (eventually on a continuous domain) and (3) quantify the related uncertainty.

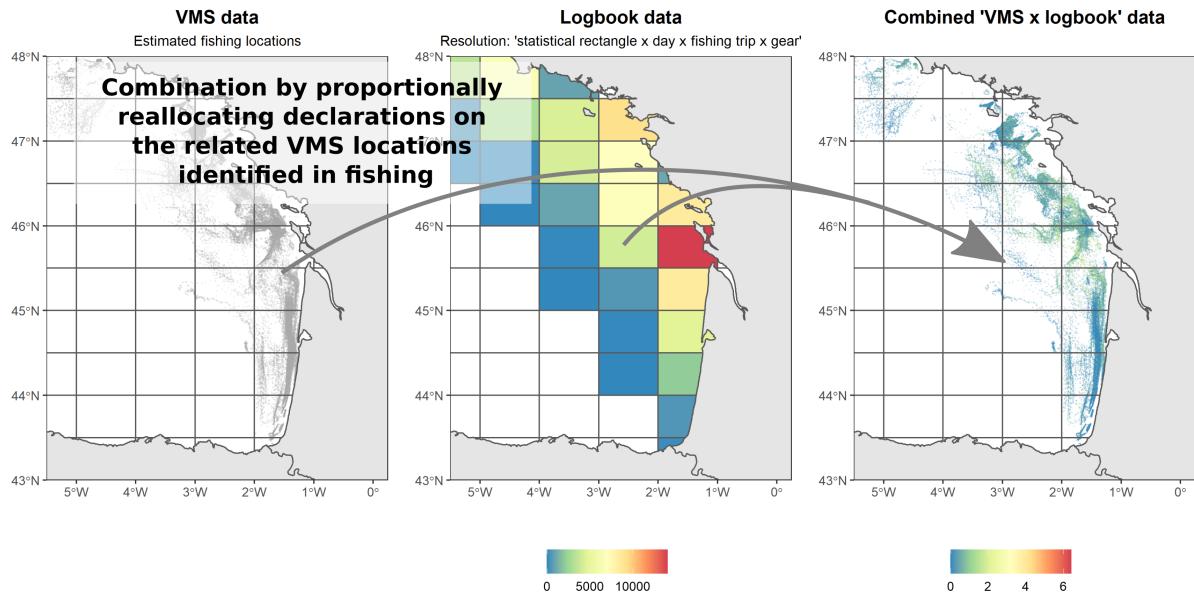


Figure 1.3 – VMS data, logbook data (sole catches) and combined ‘VMS x logbook’ data in the last quarter of 2018. logbook data are in kg. ‘VMS x logbook’ data are log-scaled ( $\log(\text{catch} + 1)$ ). Extensive details on the combination of both data sources are given in the first chapter (Section 3) and in the last chapter (Section 5 - Figure 5.1).

## 1.3 Integrating ‘VMS x logbook’ data with scientific survey data to infer fish spatial distribution

‘VMS x logbook’ data provide massive data to map fish distribution and complement scientific survey data. By **building a spatio-temporal integrated framework combining both ‘VMS x logbook’ and survey data**, it is possible to provide inferences that benefit from the information of the survey data on the period of the survey and from the information available in ‘VMS x logbook’ on the remaining part of the year.

As fishermen tend to target areas of higher biomass, **the framework should account for PS** of the distinct fleets and the temporal variation of PS as fleets can have varying PS behavior through time.

Finally, catch declarations are declared at the level of statistical rectangles while scientific data are recorded at their exact locations. The scale of statistical rectangles are not relevant for an ecological analysis. Then, **the approach needs to refine the spa-**

tial resolution of logbook data to infer species distribution at a fine scale and combine these with scientific data. Two alternative approaches are suitable:

1. a pragmatic approach of COS: upstream to the fitting procedure, reallocate proportionally logbooks declarations on fishing locations to refine their spatial resolution.
2. a statistical approach of COS: account for COS within the integrated framework and consider that logbook data have been observed at rough scale as a summation of exactly located catches.

### **Objective of the PhD**

In this manuscript, I aim at presenting the methodological basis of a statistical spatio-temporal model that integrates ‘VMS x logbook’ data with standard data sources (scientific data and onboard data) while accounting for the potential bias that are inherent to the declaration data. The objective of the manuscript is then twofold:

**Methodological objective** Develop the spatio-temporal statistical methods to properly integrate the different data sources and produce monthly maps of species distribution.

**Ecological objective** Identify fish essential habitats (specifically spawning grounds) based on the integration of the available data sources (catch declarations data and scientific survey data).

Both issues are strongly connected, as identifying essential fish habitats requires reliable integration of the data sources to predict fish spatio-temporal distribution. Integrating properly catch declarations data with scientific data raises several challenges: integrating several potentially unbalanced data sources, modeling PS while disentangling PS from other processes affecting fishing locations, accounting for differences in spatial scales between the data sources.

**Methodological basis** of the spatial and spatio-temporal models that were built in this PhD are described in detail as a foreword part of the manuscript in chapter 2.

**Chapter 3** of the manuscript provides the basis of a spatial model which combines ‘VMS x logbook’ data and scientific survey data while accounting for PS of commercial data. It addresses the question of the contribution of the several data sources in inference, the importance of modeling PS and the effect of a potential misspecification of the

sampling process on model outputs. The model is illustrated on three demersal species of the Bay of Biscay with distinct PS targeting behaviors. This chapter is now published in ICES Journal of Marine Science.

**Chapter 4** extends the framework in time and emphasizes how the model outputs can be used to identify recurrent aggregation areas during the reproduction season. These are confronted with literature knowledge of spawning areas for three species of the Bay of Biscay that face contrasts regarding the available knowledge on spawning grounds.

Finally, **Chapter 5** handle the COS issue. We explore (1) how the standard way to roughly cross logbook data with VMS data may induce a bias in model estimates and (2) how the several data sources can be combined properly to reconcile the spatial scales of the data.

The manuscript ends with a **Discussion** that outlines the possible applications of the framework developed in this thesis for fisheries science and for other fields of research. It also details the limits of the approach and finally details some potential future extension of the model (e.g. complexifying the sampling process of the data, including population dynamics in the framework).



# STATISTICAL TOOLS FOR SPATIAL AND SPATIO-TEMPORAL MODELING

---

The main methodological notions that underly the models and the fitting procedure that are used in the thesis are described hereafter. I **first** introduce few definitions that I adopt regarding hierarchical models. **Second**, I describe the structure of the spatio-temporal models that are used in this thesis. **Last**, I develop the main ideas behind the estimation method that was used to infer the model estimates i.e. Maximum Likelihood Estimation (MLE) through Template Model Builder (TMB - Kristensen et al. (2016)). To integrate the likelihood over the latent effects, we use (1) the Laplace approximation (which is implemented in TMB) and (2) the SPDE approach, an approximation method that is specific to spatial modeling and that allows to keep sparsity in the precision matrix of the spatial random effect and enables fast estimation of spatial fields.

## 2.1 Base definition of hierarchical models

Some ecological variables are not observable *per se*, but are only accessible through random indirect observations of these non-observed variables (Peyrard, Robin, and Gimenez, 2022).

For example, systematic surveys record count or biomass data for several species of interest that can be used to describe their spatial distribution (e.g. fish research surveys). Even though these counts are supposed to be representative of species distribution, they are also sampled in rough natural conditions, by distinct agents, in a varying environment and consequently they are noisy data. In such configuration, one would like to disentangle the observation stochasticity from the underlying ecological process variability to infer species distribution.

Many similar examples exist in the statistical literature: trajectories and behav-

ior of individuals could be drawn from GPS locations (Etienne and Gloaguen, 2022), species demography and population dynamics from capture-recapture and abundance data (Gimenez et al., 2022; Rivot et al., 2004; Aeberhard, Mills Flemming, and Nielsen, 2018), species colonization mechanisms from detection/non-detection data (Papaïx et al., 2022). Those non-observed variables are often referred as hidden variables in ecology or latent variables in the statistical literature (Peyrard, Robin, and Gimenez, 2022).

Developing statistical tools to disentangle the observation stochasticity from the underlying ecological process to infer this latent ecological process is then a major challenge for ecological analysis.

**Hierarchical models** formalize this idea by expressing the variability of an ecological process through one or several layers of latent variables while observations are modeled conditionally on these latent variables (Auger-Méthé et al., 2021). Hierarchical models are often presented through a graphical representation of the process under study where the graph represents the conditional relationship between the different layers (Figure 2.1 - Parent and Rivot (2012)).

More formally, in a parametric context the distribution of the observations  $\mathbf{Y}$  is defined conditionally on a latent variable  $\mathbf{S}$  and a set of parameters  $\boldsymbol{\theta}_{obs}$  and follow some probability distribution  $\mathcal{L}$ .

$$\mathbf{Y}|\mathbf{S}, \boldsymbol{\theta}_{obs} \sim \mathcal{L}(\mathbf{S}, \boldsymbol{\theta}_{obs})$$

The latent process  $\mathbf{S}$  is assumed to follow a distribution  $\mathcal{F}$  depending on a set of parameters  $\boldsymbol{\theta}_{process}$ .

$$\mathbf{S}|\boldsymbol{\theta}_{process} \sim \mathcal{F}(\boldsymbol{\theta}_{process})$$

By explicitly separating out the ecological process of interest  $\mathbf{S}$  from the observation process, the hierarchical approach is typically appropriate to model jointly several data sources. In the ecological community, this type of approach is often referred as an integrated model (Schaub and Abadi, 2011; Zipkin and Saunders, 2018). The data sources can be of different types with distinct observation processes. For instance, in species distribution modeling one can combine presence-absence data with presence-only data (Gelfand and Shirota, 2019). In this case, presence-absence data are modeled through a Bernoulli distribution, presence-only data through a Poisson process and both data sources are integrated to infer the same hidden process (species distribution).

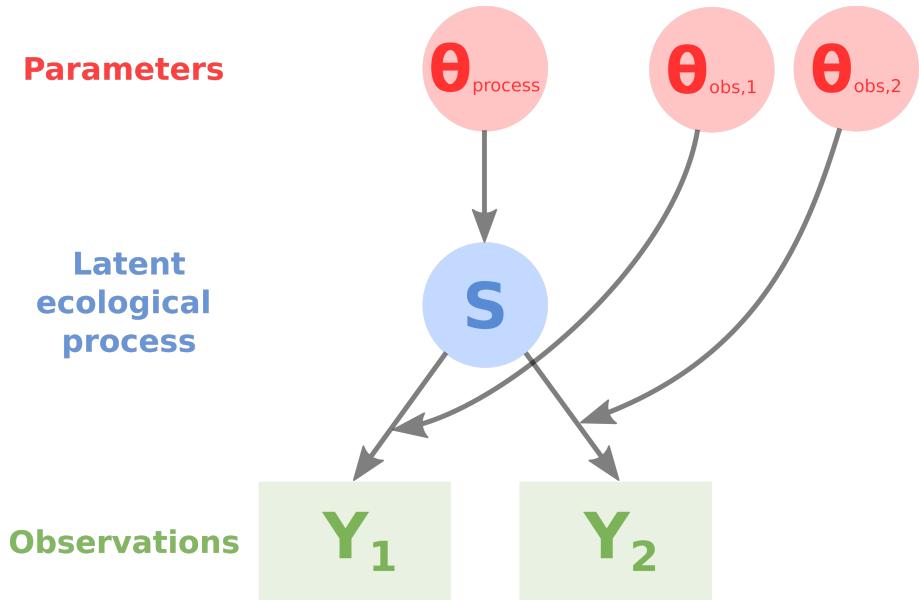


Figure 2.1 – Graph of a hierarchical model.  $\mathbf{Y}_1$  and  $\mathbf{Y}_2$  are two distinct data sources with their own observation process and observation parameters  $\theta_{obs,1}$  and  $\theta_{obs,2}$ .

## 2.2 Hierarchical models in a spatial and spatio-temporal context

As stated previously, there is a huge diversity of hierarchical models according to the process under study (e.g. animal movement, population demography, dispersal mechanisms, community structure). In the following, we will focus on **simple spatio-temporal models** describing the distribution of an outcome as a function of covariates and random terms that capture spatio-temporal correlation processes. In this case, the latent field can be formulated as a combination of (1) a trend term representing the effect of the covariates on the spatial distribution and (2) a spatial or spatio-temporal component. This last term represents that values are more similar in nearby locations and time steps than those that are distant in space and time i.e. latent field values are correlated and the degree of correlation decreases with distance and duration between observations. This term captures all the spatial and/or temporal correlation processes that are not captured by the covariates included in the model (Krainski et al., 2018). This spatio-temporal term can result either from the effect of some covariates that are not included in the model or some aggregation patterns that cannot be captured by any covariates.

**N.b.** In the following, if we denote a random process ( $U$ ),  $\mathbf{U}$  refers to a random vector

(or matrix) of  $(U)$ .  $U_i$  is the  $i^{th}$  term of  $\mathbf{U}$ .  $U(x, t)$  is the random variable  $(U)$  at location  $x$  and time  $t$ . Also, all matrices and vectors are denoted in **bold**. When convenient, realizations of the random variable  $U_i$  are denoted in lower case  $u_i$  (even though they are not used in the formulas of this section, but rather in the following chapters).

Let's assume the process  $(S) = (S(x, t), x \in \mathbb{R}^2, t \in \mathbb{R}^+)$  depends on both space and time so that  $S(x, t)$  is the value of the latent field at location  $x$  and time  $t$ . Following classical generalized linear modeling approach, the decomposition of the latent field can be reformulated as:

$$g(S(x, t)) = \mu + \boldsymbol{\Gamma}(x, t) \cdot \boldsymbol{\beta} + \delta(x, t)$$

where  $g$  is a link function,  $\mu$  is the intercept of the latent field,  $\boldsymbol{\Gamma}(x, t)$  is the vector of the covariates measured at time  $t$  and location  $x$ ,  $\boldsymbol{\beta}$  is the vector of the parameters related to the covariates,  $\cdot$  stands for the scalar product.  $\delta(x, t)$  stands for the random term that captures spatial/spatio-temporal correlation.

In spatial statistics, the random effect  $(\delta) = (\delta(x, t), x \in \mathbb{R}^2, t \in \mathbb{R}^+)$  is classically represented as a Gaussian Field (GF) with a 0 mean and variance-covariance function  $\mathcal{C}(x, y; t, r)$ ,  $(x, y) \in \mathbb{R}^2$  and  $(t, r) \in \mathbb{R}$  which controls the spatial dependence between all  $\delta(x, t)$ . Some base properties are required to define a covariance function. These are described in appendix A.1 (e.g. non-negative and positive-definiteness, stationarity, isotropy, separability).

$$(\delta) \sim \mathcal{N}(0, \mathcal{C}(x, y; t, r))$$

Many kinds of **spatial covariance functions** exist e.g. the spherical, the exponential, the powered exponential, the Gaussian model (Banerjee, Carlin, and Gelfand, 2014). In this manuscript, we only defined the models based on the Matérn covariance function for the spatial component and through a simple first-order autoregressive form for the time component (Cameletti et al., 2013). This process is second-order stationnary, isotropic and separable. This choice has been made to ease the computational burden (see section 2.3.3).

In this kind of model, locations  $x$  are defined continuously in space  $x \in \mathbb{R}^2$  and time steps  $t$  are defined on a discrete and regular domain so that  $t \in \llbracket 1, T \rrbracket^1$ . The equation of

---

1. On the continuous time domain, the similar specification would consist in the Ornstein–Uhlenbeck process  $\delta(x, t) = e^{\varphi|t-s|}\delta(x, s) + \omega(x, t)$

$(\delta)$  is given by:

$$\delta(x, t) = \varphi \cdot \delta(x, t - 1) + \omega(x, t) \text{ for } t = 2, \dots, T$$

with  $\varphi \in ]-1; 1[$  the autoregressive temporal term.  $\omega(x, 1)$  derives from the stationary distribution  $\mathcal{N}(0, \sigma^2/(1 - \varphi^2))$ .

$\omega(x, t)$  is an innovation term between each time step that is modeled as a zero mean spatial GF. It is assumed to be temporally independent and is parameterized through the Matérn covariance function:

$$\mathcal{C}(\omega(x, t), \omega(y, r)) = \begin{cases} 0 & \text{if } t \neq r \\ \sigma_\omega^2 \cdot \text{Cor}(h) & \text{if } t = r \end{cases} \quad \text{for } x \neq y$$

where  $h = \|x - y\| \in \mathbb{R}$  and  $\text{Cor}(h)$  is the Matérn correlation function. Furthermore,  $\text{Var}(\omega(x, t)) = \sigma_\omega^2$ , where  $\sigma_\omega^2$  is the marginal variance of  $(\omega)$ .

Then, following Cameletti, Iagnaccolo, and Bande (2011), the covariance of  $(\delta)$  can be written as:

$$\mathcal{C}(\delta(x, t), \delta(y, r)) = \frac{\varphi^{|t-r|}}{1 - \varphi^2} \sigma_\omega^2 \text{Cor}(h)$$

with  $|t - r|$  the time lag between time step  $t$  and  $r$ .

The separability of  $\mathcal{C}(\delta(x, t), \delta(y, r))$  clearly appears in a separable multiplicative form where  $\sigma_\omega^2 \text{Cor}(h)$  is the spatial component of the spatio-temporal covariance function and  $\frac{\varphi^{|t-r|}}{1 - \varphi^2}$  is the temporal component of the spatio-temporal covariance.

**The Matérn correlation function** takes the form:

$$\text{Cor}(h) = \frac{(\kappa \cdot h)^\nu K_\nu(\kappa \cdot h)}{\Gamma(\nu) \cdot 2^{\nu-1}}$$

$K_\nu$  is the Bessel function of the second kind and of order  $\nu$ .  $\nu$  controls the smoothness of the process, in 2 dimensions it is usually fixed to 1.  $\kappa$  is a scaling parameter that controls the degree of spatial auto-correlation. It is commonly re-expressed in terms of range  $\rho = \sqrt{8\nu}/\kappa$  which is the distance where spatial correlation decreases below 0.1.

The Matérn covariance is commonly used in spatial statistics literature. It has nice physical interpretation for the range parameters  $\rho$  and the smoothness  $\nu$ . Besides, it is quite convenient as GF defined through Matérn covariance can be represented as Gauss-Markov Random Field (GMRF) which can strongly fasten computation (Cf. next section

on the SPDE approach - Lindgren, Rue, and Lindström (2011)).

For a given set of locations and time  $((x_1, t_1), \dots, (x_n, t_n))$  of  $(\delta)$ , the corresponding random vector  $\boldsymbol{\delta}$  follows a Normal distribution  $\delta \sim \mathcal{N}(0, \Sigma)$  with **probability density function (pdf)**:

$$\pi(\boldsymbol{\delta}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{\delta}^T \Sigma^{-1} \boldsymbol{\delta}\right)$$

so that  $\Sigma_{ij} = \mathcal{C}((x_i, t_i), (x_j, t_j))$ .  $\Sigma$  is a square matrix whose size equals the number of terms contained in  $\boldsymbol{\delta}$ , whose diagonal terms equals the variance of each  $\delta$  that we denote  $\sigma_\delta^2$  and whose off-diagonal elements equal  $\mathcal{C}(x, y; t, r)$ .

## 2.3 Inference in hierarchical models

Estimating the parameters and the latent effects in hierarchical models can be highly challenging and raises several issues. In the following, we shortly highlight these issues. Each one is emphasized by a box where we synthetize the issue and the related solution. These are then developed more extensively in the following.

### 2.3.1 The likelihood: a tricky component of hierarchical models estimation

The **likelihood** (i.e. the pdf of the data knowing the parameters  $\mathbb{P}(\mathbf{Y}|\boldsymbol{\theta})$ ) is the backbone of most inference methods. Maximum likelihood methods consist in looking for the parameter values that maximize the likelihood  $\mathbb{P}(\mathbf{Y}|\boldsymbol{\theta})$  while Bayesian inference consists in computing the posterior distribution of the parameters  $\boldsymbol{\theta}$  knowing the data  $\mathbf{Y}$  through the Bayes rule  $\mathbb{P}(\boldsymbol{\theta}|\mathbf{Y}) = \frac{\mathbb{P}(\mathbf{Y}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathbf{Y})}$ , where  $\mathbb{P}(\boldsymbol{\theta})$  is the prior distribution and  $\mathbb{P}(\mathbf{Y})$  is the probability of the data.

For both the classical and the Bayesian paradigms, inference on hierarchical models including latent effects requires integrating the joint likelihood  $\mathbb{P}(\mathbf{Y}, \boldsymbol{\delta}|\boldsymbol{\theta})$  over the latent effects  $\boldsymbol{\delta}$  to compute the likelihood  $P(\mathbf{Y}|\boldsymbol{\theta})$  (that we denote also  $L_M(\boldsymbol{\theta})$  with  $q$  the size of the latent effect).

$$L_M(\boldsymbol{\theta}) = \mathbb{P}(\mathbf{Y}|\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \mathbb{P}(\mathbf{Y}, \boldsymbol{\delta}|\boldsymbol{\theta}) d\boldsymbol{\delta}$$

This is a major challenge as this integral cannot be written in a closed form. Then,

numerical technics (either sampling or approximation technics) are required to compute  $L_M(\boldsymbol{\theta})$ .

During twenty years, **Bayesian methods** have been very popular to infer hierarchical models through sampling methods such as **MCMC** algorithms<sup>2</sup> (Parent and Rivot, 2012).

MCMC methods are sampling technics that consist in sampling into the joint posterior distribution  $\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{\delta} | \mathbf{Y})$  to infer the probability distribution of parameters and latent effects. The most common algorithm to perform MCMC is the **Metropolis-Hastings** algorithm. If we denote the vector of parameters and latent effects  $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\delta})$ , this algorithm consists in sampling a new set of parameters and latent effects  $\boldsymbol{v}'$  in a proposal distribution  $p(\cdot | \boldsymbol{v})$  conditionally on the parameters of a previous iteration  $\boldsymbol{v}$ . When a proposal parameter is drawn, it is accepted with the probability:

$$\min \left\{ 1, \frac{p(\boldsymbol{v} | \boldsymbol{v}') \mathbb{P}(\boldsymbol{v}' | \mathbf{Y})}{p(\boldsymbol{v}' | \boldsymbol{v}) \mathbb{P}(\boldsymbol{v} | \mathbf{Y})} \right\} = \min \left\{ 1, \frac{p(\boldsymbol{v} | \boldsymbol{v}') \mathbb{P}(\mathbf{Y} | \boldsymbol{v}') \mathbb{P}(\boldsymbol{v}')}{p(\boldsymbol{v}' | \boldsymbol{v}) \mathbb{P}(\mathbf{Y} | \boldsymbol{v}) \mathbb{P}(\boldsymbol{v})} \right\}$$

$p(\boldsymbol{v} | \boldsymbol{v}')$  is the density of the proposal distribution  $p(\cdot | \boldsymbol{v}')$  evaluated at  $\boldsymbol{v}$ .

Such procedure does not require any integration over the latent effects, only  $\mathbb{P}(\mathbf{Y} | \boldsymbol{v})$  need to be evaluated at  $\boldsymbol{v}$  and  $\boldsymbol{v}'$ .

In basic Metropolis-Hastings algorithm, the terms of  $\boldsymbol{v}$  are updated one by one. Convergence to the posterior distribution can take a number of iterations before reaching a stationary distribution and usually the first batch of simulations (called burn-in) are discarded for inference. The convergence time can get very long as the size of  $\boldsymbol{v}$  increases (which is the case for spatio-temporal models - this is often referred as the big  $n$  problem). Furthermore, the component of  $\boldsymbol{\delta}$  and  $\boldsymbol{\theta}$  are very interdependent which makes that Markov chains will move around very slowly towards the target posterior distribution. Some solutions to overcome this interdependence exist e.g. building blocks of parameters that are updated together (Rue, Martino, and Chopin, 2009; Rue and Held, 2005). However even so, convergence for complex spatio-temporal models remains inefficient.

---

2. Note that sampling technics also exist in the frequentist paradigm. They consist in integrating over latent effects through technics such as Monte Carlo methods - see in Cressie and Wikle (2015) for instance. They will not be detailed here as MCMC methods remains the most widespread sampling methods to make inference on hierarchical models.

### 2.3.2 Approximation of the likelihood based on Laplace approximation

In the Bayesian paradigm, the **Integrated Nested Laplace Approximation** is gaining ground as an efficient alternative to MCMC methods for spatio-temporal models as it bypasses the integration issue by simplifying the marginal posterior distribution of the parameters  $\mathbb{P}(\boldsymbol{\theta}|\mathbf{Y})$  and the marginal posterior distribution of the latent effects  $\mathbb{P}(\boldsymbol{\delta}|\mathbf{Y})$  by several nested Laplace approximations (Rue, Martino, and Chopin, 2009).

The **frequentist paradigm** is also progressively coming back as a suitable alternative to Bayesian inference. More specifically, the Laplace approximation is gaining ground as it allows to approximate the marginal likelihood and to bypass the burden of numerical integration issue for spatio-temporal models (Skaug and Fournier, 2006; Kristensen et al., 2016).

The **Laplace approximation** simplifies the expression of the marginal likelihood through Taylor expansion. Overall, this approximation replaces the integration of  $\int L(\boldsymbol{\theta}, \boldsymbol{\delta})d\boldsymbol{\delta}$  by a maximization step (here searching for  $\hat{\boldsymbol{\delta}}_\theta$ ). It suits for any kind of hierarchical models with random effect in the hidden layer (not only spatial or spatio-temporal ones) provided that the random effect is Gaussian (or nearly Gaussian) in addition to some regularity conditions (e.g. the maximum of  $\mathbb{P}(\mathbf{Y}, \boldsymbol{\delta}|\boldsymbol{\theta})$  relatively to  $\boldsymbol{\delta}$  is unique).

For numerical reasons, one often work on the joint log-likelihood  $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$  or on the negative joint log-likelihood  $f_{nll}$  to maximise the Marginal likelihood.

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \log \mathbb{P}(\mathbf{Y}, \boldsymbol{\delta}|\boldsymbol{\theta}) = -f_{nll}(\boldsymbol{\theta}, \boldsymbol{\delta})$$

By approximating  $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$  through Taylor series around the maximum  $\hat{\boldsymbol{\delta}}_\theta = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}(\ell(\boldsymbol{\theta}, \boldsymbol{\delta}))$ , we obtain:

$$L_M(\boldsymbol{\theta}) \approx L_M^*(\boldsymbol{\theta}) = (2\pi)^{q/2} |\mathbf{H}(\boldsymbol{\theta})|^{-1/2} \exp[-f_{nll}(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta)]$$

with  $\mathbf{H}(\boldsymbol{\theta})$  the Hessian of the negative joint log-likelihood  $f_{nll}(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta)$  and  $q$  the size of the latent effect. The  $(j, k)^{th}$  element of the Hessian are written as  $\frac{\partial^2}{\partial \delta_j \partial \delta_k} f_{nll}(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta)$ . Some additional details on the calculus are provided in appendix A.3.

Then, the marginal negative log-likelihood (the function being optimized in TMB) can be rewritten as:

$$-\log L_M^*(\boldsymbol{\theta}) = -q \log \sqrt{(2\pi)} + \frac{1}{2} \log |\mathbf{H}(\boldsymbol{\theta})| + f_{nll}(\hat{\boldsymbol{\delta}}_\theta, \boldsymbol{\theta})$$

Such approximation allows huge computational gains compared with standard MCMC algorithms. However, it only suits for continuous and unimodal distribution such as Gaussian (or nearly Gaussian) ones.

To derivate  $f_{nll}$ , TMB performs automatic differentiation. Some details on automatic differentiation are given in appendix A.2 and more details are available in Kristensen et al. (2016).

A tricky and time consuming term in the expression of  $-\log L_M^*(\boldsymbol{\theta})$  is the computation of the Hessian matrix  $\mathbf{H}(\boldsymbol{\theta})$  and its determinant. A way to gain computation time is to specify latent effects that benefit from Markovian properties (i.e. conditional independence properties) which ensures the sparsity of the Hessian matrix; this enables to use sparse computation technics for fast estimation under MLE (Skaug and Fournier, 2006).

### 2.3.3 Sparse representation of spatial random effects: the SPDE approach

The major breakthrough brought by Lindgren, Rue, and Lindström (2011) through the **SPDE approach** is to provide a sparse representation of a continuous GF by:

1. approximating the GF by a GMRF to benefit from the Markovian properties of GMRF for fast computation.
2. extending this first set of results (valid on a regular grid only) to a continuous spatial domain through the Finite Element Method.

Note that a deep description of all the theory behind the SPDE approach would require much more development. We only limit the description of the theory to the main conclusions of the paper of Lindgren, Rue, and Lindström (2011). More details are developed in Lindgren, Rue, and Lindström (2011), Rue and Held (2005), Krainski et al. (2018), Bakka et al. (2018), and Bakka (2018).

#### Defining a Gaussian Markov Random Field

**A GMRF is a spatial process defined over a discrete domain that admits conditional independence relations.** For instance, let's introduce a 0 mean GMRF model denoted  $\boldsymbol{\delta}^* \sim GMRF(0, \mathbf{Q}^{-1})$  defined over a set of locations  $x = (x_1, \dots, x_n)'$ . A

GMRF is modeled through its precision matrix  $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$  where 0 coefficients indicate conditional independence between random variables. The pdf of  $\boldsymbol{\delta}^*$  can be written as:

$$\pi(\boldsymbol{\delta}^*) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} \boldsymbol{\delta}^{*T} \mathbf{Q} \boldsymbol{\delta}^*\right)$$

The Markovian property of the GMRF suppose to define a neighborhood structure that breaks the dependencies: the value of  $\delta^*(x)$  depends on the values of  $\boldsymbol{\delta}^*$  in the neighborhood of  $x$  (denoted  $o(x)$ ), but is independent from the other values of  $\boldsymbol{\delta}^*$  conditionally on the values within the neighborhood  $o(x)$  (conditional independence property).

$$\delta^*(x) \perp \boldsymbol{\delta}_{-(x,o(x))}^* | \boldsymbol{\delta}_{o(x)}^*$$

Such conditional independence relationship (which is controlled by the neighborhood structure) leads to 0 values in the precision matrix between conditionally independent locations. This results in a sparse precision matrix i.e. a matrix filled with zero values outside the neighborhood structure of each location and with non-zero patterns within the neighborhood structure of each location.

$$Q_{xy} \neq 0 \text{ if } y \in \{x, o(x)\} \text{ and } Q_{xy} = 0 \text{ otherwise}$$

Using sparse matrix computation technics can lead to huge computational gains. From a matrix factorization that requires  $\mathcal{O}(n^3)$  flops (floating point operations per second) for a dense matrix as is the case for GF, one pass to a factorization that requires  $\mathcal{O}(n)$  flops for temporal GMRF models,  $\mathcal{O}(n^{3/2})$  flops for spatial GMRF models and  $\mathcal{O}(n^2)$  flops for spatio-temporal GMRF models.

## Approximating a Gaussian Field as a Gauss-Markov Random Field

Lindgren, Rue, and Lindström (2011) **link GF to GMRF** through the Stochastic Partial Differential Equation (SPDE):

$$(\kappa^2 - \Delta)^{\alpha/2} \cdot \delta(x) = \mathcal{W}(x)$$

with  $x \in \mathbb{R}^d$ ,  $\alpha = \nu + d/2$ ,  $\kappa > 0$ ,  $\nu > 0$ ,  $\Delta$  is the Laplacian,  $\mathcal{W}(x)$  is a spatial white noise Gaussian stochastic process with unit variance,  $d$  is the number of dimensions (here there are 2 dimensions).

A GF  $\boldsymbol{\delta}$  defined through the Matérn covariance is solution of the SPDE defined above.

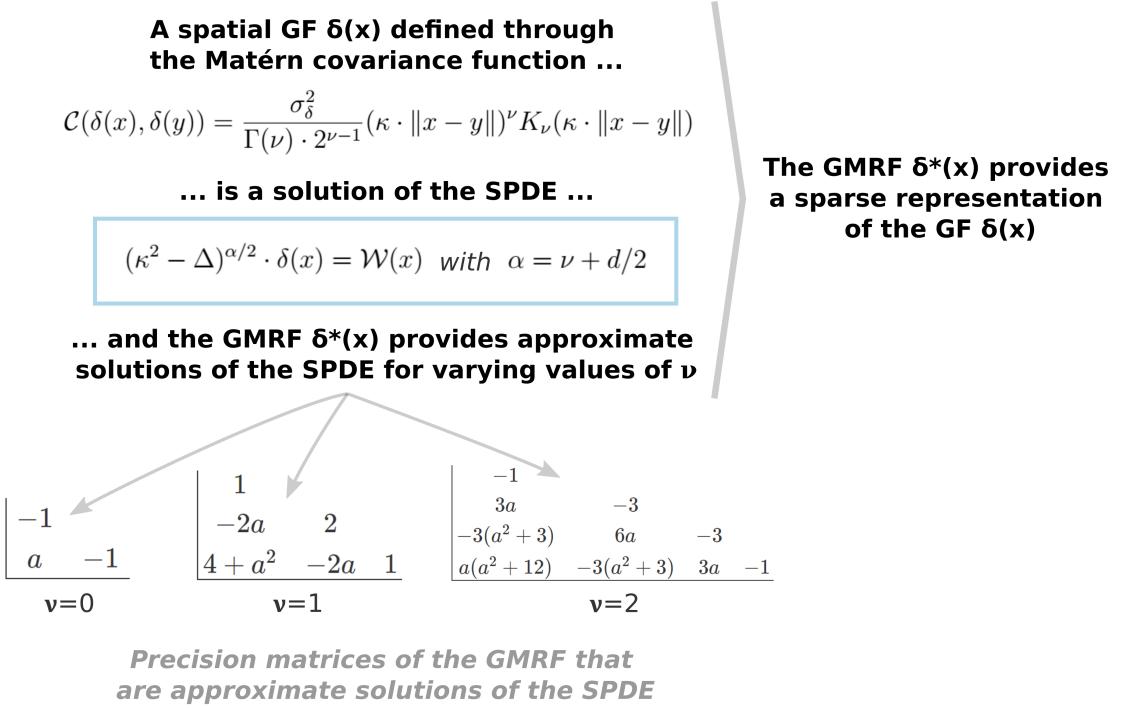


Figure 2.2 – Graphical representation of the main theoretical link between a GMRF model and a GF model through the SPDE approach. For the precision matrices, only the upper right quadrant are shown with the corner value being the central element of the matrix (e.g.  $a$  for the bottom left matrix). All other values of the matrices are 0 values.  $a$  is the equivalent of the autoregression parameter in standard conditional autoregressive model - see Krainski et al. (2018).

Lindgren, Rue, and Lindström (2011) demonstrated that for increasing values of  $\nu$  ( $\nu = 1, 2, \dots$ ), a GMRF  $\delta^*$  with increasing neighborhood (the related precision matrices are given in Figure 2.2) are approximate solutions of the SPDE. Then these GMRF provide sparse representation of GF defined through Matérn covariance for varying values of  $\nu$ .

By making a theoretical link between GF and GMRF, the SPDE approach allows to approximate a GF by a GMRF. This enables to benefit from the nice properties of GF (interpretable parameters of the covariance functions such as the range  $\rho$ ) while benefiting from the sparse precision matrix of GMRF and its computational properties which derives from the Markovian property of the GMRF.

Still, these results are valid on regular grids only. To go one step further, Lindgren, Rue, and Lindström (2011) extended their approach to a continuous spatial domain.

## Moving to a continuous domain

Similarly as in the finite element method literature, Lindgren, Rue, and Lindström (2011) pass on a continuous domain by considering that **the solution of the SPDE  $\delta$  can be approximated through piecewise linear approximation by introducing a weak formulation of the SPDE** (see Bakka (2018) for more details).

This requires to build a triangulated mesh covering the whole domain (Figure 2.3) and to express the random field values  $\delta(x)$  at position  $x$  as a linear combination of basis functions  $\psi_k(x)$  and Gaussian weights  $w_k$  such as:

$$\delta(x) = \sum_{k=1}^n \psi_k(x) \cdot w_k$$

with  $n$  the number of nodes of the triangulated mesh.

At each node of the mesh,  $\delta(x_k) = w_k$  and the basis function related to the  $k^{th}$  node equals 1 while all other basis functions are null. At all other locations (for instance, the red points inside the triangle on the top left of the Figure 2.3), the value of  $\delta(x)$  is a combination of the Gaussian weights  $w_k$  and the (linear) basis functions  $\psi_k(x)$  related to each node of the triangle to which belongs the location. The value of the basis functions depends on the distance to the related node and the 3 basis functions always sum to 1.

A critical (and often ad hoc) point in the SPDE approach is the construction of the mesh: one has to find a compromise between computation time and density/geometry of the mesh to obtain the most accurate inference in a reasonable amount of time.

Some further comments on the expression of the precision matrix when defined over the mesh are provided in supplementary material A.4.

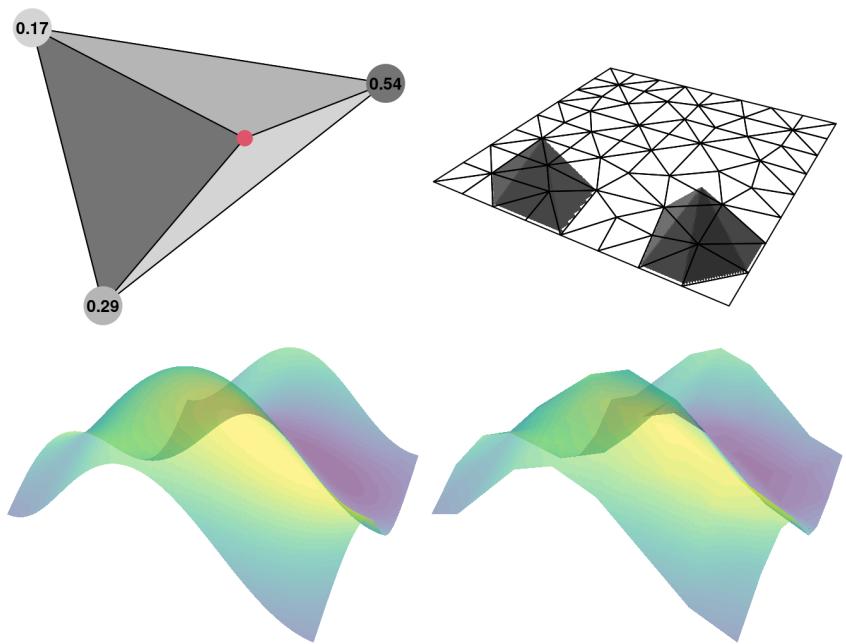


Figure 2.3 – Illustration of the piecewise linear approximation (Krainski et al., 2018). Top left: One triangle of the mesh. The red point is a data point. Top right: Mesh and basis functions for 2 nodes of the mesh. Bottom left: True random field. Bottom right: Approximation of the random field.



# **COMBINING SCIENTIFIC SURVEY AND COMMERCIAL CATCH DATA TO MAP FISH DISTRIBUTION**

---

This first chapter lays the basis of the integrated framework in a purely spatial context. We aim at illustrating how scientific survey and ‘VMS x logbook’ data can be combined on the time span of the scientific survey and how to account for preferential sampling of commercial data in inference. We assess the model through simulations and apply the approach on three contrasted case studies regarding preferential sampling behavior. We last discuss the limitations of this study, the potential applications of the framework and the next extensions of this spatial framework.

This chapter led to a publication in the ICES Journal of Marine Science:

See the link: <https://doi.org/10.1093/icesjms/fsac032>

## **Abstract**

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

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

### **3.1 Introduction**

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

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

Commercial catch declarations data (logbooks) constitute a complementary data source that may benefit of a higher sampling effort than scientific survey. In Europe, catch declarations must be reported in logbooks data for all fishing vessels; besides, geolocation through Vessel Monitoring System (VMS) is mandatory for all fishing boats above 12 m

long (Hintzen, 2021). Hence, logbook data combined with VMS data can provide high resolution maps of Catch Per Unit Effort i.e. CPUE (Gerritsen and Lordan, 2011; Murray et al., 2013) with a relatively dense spatio-temporal sampling within the range of the commercial fleets. However, inferring SDM with commercial data can be challenging as they generally arise from a preferential sampling (PS) behavior, i.e. a sampling that directly or indirectly depends upon the biomass of the target species. Indeed, fishermen tend to target areas with high biomass and may also favour fishing zones based on other criteria (like bottom substrate or distance to the coast for instance — Hintzen (2021)) that are indirectly related to the target species abundance. When not properly considered in statistical models, PS associated with commercial data may lead to biased estimates of fish distribution and biomass (Trenkel et al., 2013; Pennino et al., 2019). In particular, when the biomass is spatially heterogeneous, ignoring PS may overestimate the spatial predictions and the overall biomass estimates.

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

Provided PS is accounted for, integrated models (IM) appear as an attractive tool to combine fishery-independent and fishery-dependent data to infer the spatial distribution of harvested fish. IM have received considerable attention in the ecological literature (Schaub and Abadi, 2011; Parent and Rivot, 2012; Gimenez et al., 2014). By sharing the information between different data types, IM may provide more accurate estimates and predictions compared with separate analysis of different data types. Recently, Rufener et al. (2021) demonstrated the potential of IM to integrate scientific data and onboard observer count data to improve SDM of fishery resources. However, although onboard observer data provide useful complementary information to scientific survey, they generally only represent a small proportion of all sea trips — 1 % in average for the French observer

programs (Cornou et al., 2021). In contrast, the combination of commercial catch declarations in logbooks with VMS data provides a more extensive data source to map fish spatial distribution. Furthermore, the potential of embedding PS within a hierarchical SDM to integrate catch declaration data and scientific survey is still an open challenge and new methodology are required to handle PS behaviors of commercial fleets while accounting for all the complexity related to fishing locational choice (Salas and Gaertner, 2004; Haynie, Hicks, and Schnier, 2009; Girardin et al., 2017).

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

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

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

## 3.2 Material and methods

### 3.2.1 Spatial integrated model

Below we provide the core elements of the modeling approach. Additional details are provided in the Supplementary material (SM B.1). The model is structured in four layers: observations (here commercial and scientific CPUE in weight per unit of effort), the sampling process, the latent field (here fish biomass relative density), and the parameters (Figure 3.1 — all notations are available in SM B.1.1, Supplementary Table B.1). Sampling process is usually ignored in hierarchical models as it is mostly considered independent of the quantity of interest, and then has no consequence on inference (Diggle, Menezes, and Su, 2010). Here, the spatial distribution of commercial fishing is explicitly modelled as a inhomogenous Poisson point process whose intensity may depend on the biomass field and contributes to the likelihood. The observation processes of scientific and commercial data are conditional upon the biomass latent field and the sampled locations.

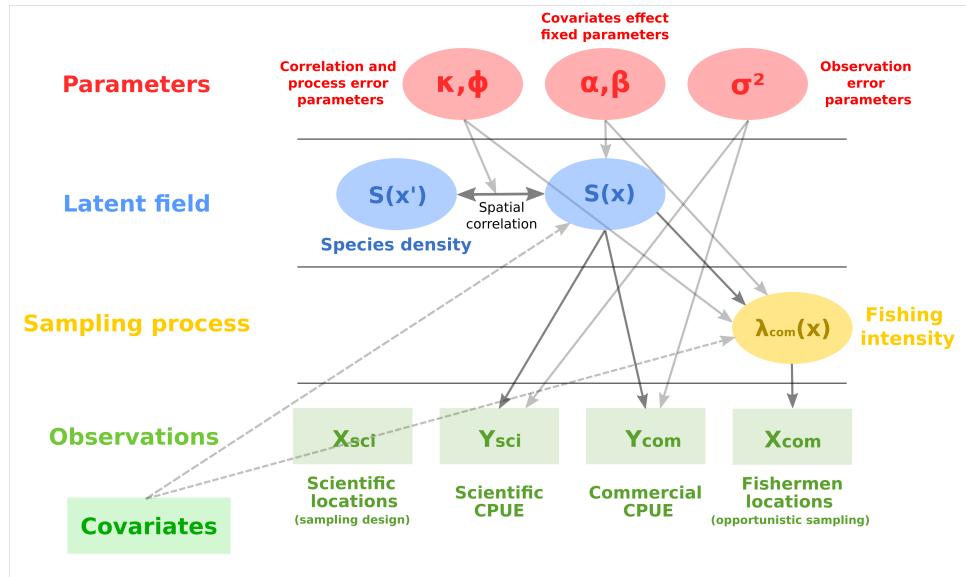


Figure 3.1 – Diagram of the spatial IM including PS for commercial data. Locations of scientific trawls do not contribute directly to the likelihood.

All processes are considered to occur in a discrete fine grid (see for instance SM B.2.1, Supplementary Figure B.1 or SM B.3.1, Supplementary Figure B.2). We assume the density of the point process is piecewise constant in each cell grid, which brings simplification in the expression of the likelihood of the point process (Diggle (2013) — see SM B.1.2). The time component is omitted and both commercial and scientific data are assumed to occur at the same time step.

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

### Latent field of relative biomass

The fish biomass relative density ( $S$ ) (eq. 3.1 - 3.2) is modelled through a latent log Gaussian spatial field defined on the same discrete spatial domain as the point process.

The mean of the Gaussian field depends on environmental covariates through a log link where the linear predictor combines an intercept  $\alpha_S$ , the linear effect of environmental

covariates  $S(x)$  (effects captured by the corresponding fixed parameters  $\beta_S$  representing the species-habitat relationship). The remaining spatial variation is accounted for through a zero-mean Gaussian random field (GRF) denoted  $(\delta(x))$  (eq. 3.2) parameterized with a Matérn correlation function  $M(x, x'; \kappa, \phi)$ , characterized by the shape  $\kappa$  and the scale  $\phi$  (Cressie, 1993; Gelfand, 2010; Lindgren, Rue, and Lindström, 2011; Banerjee, Carlin, and Gelfand, 2014). The shape can be expressed in term of range  $\rho = \frac{\sqrt{8}}{\kappa}$  where  $\rho$  is the distance for which the correlation between points is near 0.1.

$$\log(S(x)) = \alpha_S + \boldsymbol{\Gamma}_S(x)^T \cdot \boldsymbol{\beta}_S + \delta(x). \quad (3.1)$$

$$(\delta) \sim \text{GRF}(0, M(x, x'; \kappa, \phi)). \quad (3.2)$$

## Sampling process

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

Let us denote  $(X_{com_j})$ , the spatial point process, where commercial vessels of fleet  $j$  are identified as fishing. In the following, all vessels in the same commercial fleet are assumed to have homogeneous behaviors. Following Diggle, Menezes, and Su (2010), the set of fishing locations are modelled conditionally on  $(S)$ , as an inhomogeneous Poisson point process with piecewise constant intensity  $\lambda_j(x)$  (eq. 3.3 - 3.4).

$$(X_{com_j}) \sim \text{IPP}(\lambda_j(x)) \quad (3.3)$$

$$\log(\lambda_j(x)) = \alpha_{Xj} + b_j \cdot \log(S(x)) + \eta_j(x) \quad (3.4)$$

For any fleet  $j$ , intensity  $\lambda_j(x)$  of the Poisson point process (eq. 3.3) is modelled as a log-linear combination of the intercept  $\alpha_{Xj}$ , the logarithm of the relative biomass  $S(x)$  scaled by a parameter  $b_j$ , and a residual spatial effect  $\eta_j(x)$  with the same structure

as  $\delta(x)$  but with specific parameters  $\kappa$  and  $\phi$ . All parameters  $\alpha_{Xj}$ ,  $b_j$ , and the spatial random effect  $\eta_j(x)$  are specific to each fleet.

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

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

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

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

## Observation process

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

Observations  $Y_i$  are modelled through a zero-inflated lognormal model conditionally on biomass  $S(x_i)$  in the sampled cell  $x_i$  (eq. 3.5) - (3.6). The model is derived from Thorson et al. (2016a) or Thorson (2018). We assume that the expected catch  $\mu_j(x_i)$  for any fleet/data source  $j$  in the cell  $x_i$  depends on the latent field value  $S(x_i)$  and a catchability coefficient  $q_j$  (eq. 3.5). A zero catch ( $y_i = 0$ ) is modelled as a Bernoulli random variable with parameter  $\exp(-e^{\xi_j} \cdot \mu_j(x_i))$ , where  $\xi_j$  is the parameter controlling the intensity of zeros relatively to the expected catch (eq. 3.6). Then,  $\mu_j(x_i)$  being fixed, the higher (resp., the lower)  $\xi_j$ , the lower (resp. the higher) the probability of obtaining a zero-catch.

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

$$\mu_j(x_i) = q_j \cdot S(x_i) \quad (3.5)$$

$$\begin{aligned} & P(Y_i = y_i | x_i, S(x_i)) \\ &= \begin{cases} \exp(-e^{\xi_j} \cdot \mu_j(x_i)) & \text{if } y_i = 0 \\ (1 - \exp(-e^{\xi_j} \cdot \mu_j(x_i))) \cdot L\left(y_i, \frac{\mu_j(x_i)}{(1 - \exp(-e^{\xi_j} \cdot \mu_j(x_i)))}, \sigma_j^2\right) & \text{if } y_i > 0 \end{cases} \quad (3.6) \end{aligned}$$

Per se, catchability  $q_j$  are not identifiable as there is no information in the model to estimate the absolute scale of  $S(x)$ . Commercial catches and/or scientific surveys will only be informative about fish biomass relative density and additional information must be provided to ensure statistical identifiability. If only one data type feeds the model (only scientific or commercial data), relative catchability is fixed to 1 and the spatial random field values is in the same scale as the data. If two data types (or more) are used to feed the model, one of the relative catchability (denoted  $q_{ref}$ ) has to be fixed, the other ones being estimated relatively to the first one through a scaling factor  $k_j$  (eq. 3.7).

$$q_j = k_j \cdot q_{ref} \quad (3.7)$$

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

## Maximum likelihood estimation

The estimation of the model is performed with TMB (Template Model Builder - Kristensen et al. (2016)) and the spatial random effects are estimated through the SPDE approach (Lindgren and Rue, 2015) within the R software. More details on estimation are available in the Supplementary material (SM B.1.4).

## IM validation

A key issue with IM is whether the different data sources provide consistent or conflicting information (Saunders et al., 2019; Zipkin, Inouye, and Beissinger, 2019; Peterson et al., 2021). In our framework, the key question is whether integrating commercial data in addition to scientific data will complement or will disrupt the inferences obtained from

the scientific data, considered as a reference source of information. To address this issue, we propose a validation procedure based on the consistency check initially developed by Rufener et al. (2021) and designed to check whether estimates obtained from the IM are consistent with those obtained from the model fitted to scientific data only. The procedure would reject consistency if the parameters estimates from the IM fall outside the 95% confidence region of parameters estimates from scientific data only (see SM B.1.5 for more details on the procedure). This validation step is applied to both simulations and case studies

### **3.2.2 Simulation-estimation experiments**

We conducted simulation-estimation experiments to assess the performance of the method for different data/model configurations (Table 1.1, see also SM B.2 for extended details on simulations). For all scenarios, simulations of data, covariates, and GRF were parameterized to tailor the case studies described hereafter. All scenarios and configurations are repeated 100 times so as to capture the variability between replicates.

Simulation-estimation experiments were specifically designed to address four questions detailed below. In all cases, commercial data were simulated with various levels of PS ( $b = 0$  for uniform sampling,  $b = 1$  for moderate PS, and  $b = 3$  for strong PS) to assess the effect of PS on model's performance (Figure 3.2).

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

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

For each combination of commercial and scientific sample size, we fitted four different models: a model fitted to scientific data only, a model fitted to commercial data only,

and two IM fitted to both commercial and scientific data, one with the scientific data used as reference level and another one using the commercial data as reference level (Cf. Equation 3.7).

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

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

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

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

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

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

Fishing locations potentially depend on many other factors independent from the species distribution (Salas and Gaertner, 2004; Haynie, Hicks, and Schnier, 2009; Gibrardin et al., 2017). To assess how such process blurring strict PS may affect the quality of inferences, we simulate data with a sampling intensity that depends on both the biomass distribution (PS) and an additional spatial random term  $\eta_j(x)$  independent from the biomass distribution (Equation 3.4); see Table 1.1 for more details on  $\eta_j(x)$  parameterization), and compare the inferences obtained from a data set simulated with strict PS ( $\eta_j(x) = 0$  on the full domain).

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

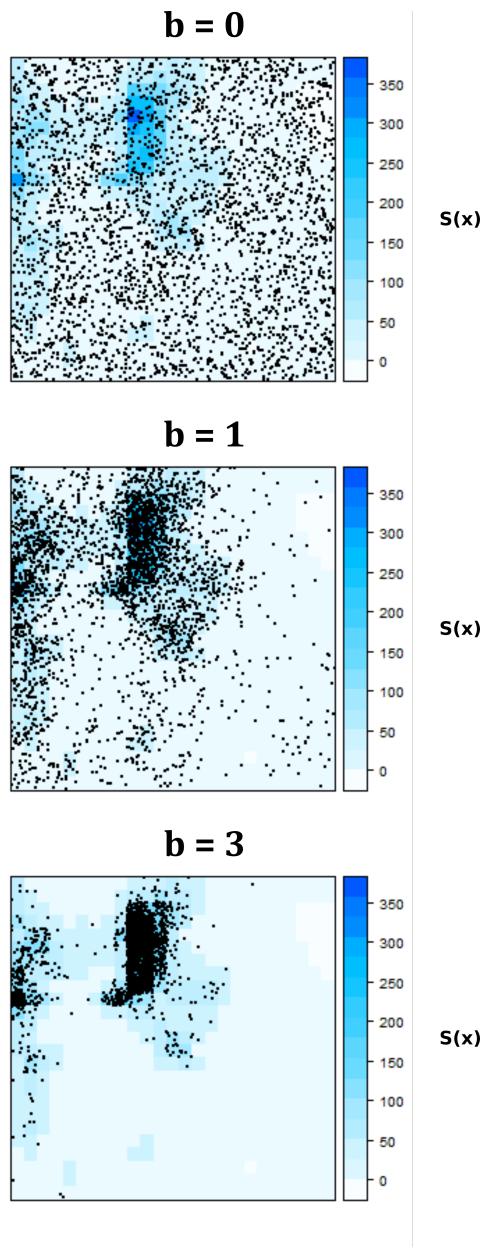


Figure 3.2 – Maps of simulated commercial sampling points obtained for three values of PS ( $b = 0$ ,  $b = 1$ , and  $b = 3$ ). Blue scale: values of the simulated biomass field. Dots: fishing points. For  $b = 0$ , the targeting metric  $T_j(x) = 1$ . For  $b = 1$ ,  $\arg \max(T_j(x)) = 12$ ,  $q_{50\%}(T_j(x)) = 0.4$ . For  $b = 3$ ,  $\arg \max(T_j(x)) = 80$ ,  $q_{50\%}(T_j(x)) = 0.002$  (SM B.1.3).

**Table 1.1** Simulation description.

General simulations description						
Biomass field	Depends on one continuous covariate $\Gamma_S$ and one random spatial effect $\delta$ . Both are simulated independently through a GRF with Matérn covariance function. Their range ( $\rho$ ) and marginal variance are fixed respectively to 10 and 1. n.b. the marginal variance quantifies the variability of the spatial process. For more details on marginal variance parameterization, see Lindgren <i>et al.</i> (2011).					
Scientific data	Random stratified plan within four strata (see Supplementary Figure SM B.2.1)					
Commercial data	Simulated according to three PS levels (i.e. three values for $b$ – see Figure 1.2).  $-b = 0$ : commercial sampling is not preferential; $-b = 1$ : PS is moderate, commercial vessels mainly target areas where fish biomass is high; $-b = 3$ : commercial sampling is highly preferential and vessels strongly target zones where biomass is high. $\eta$ is set to 0 for Q1, Q2, and Q3. For Q4, $\eta$ is set to tailor the sole case study. The range of $\eta$ is set to 40 (four times the range of $\delta$ ), the marginal variance is set to 5 (five times the marginal variance of $\delta$ ). Catchability fixed to 1 Simulated with 30% of zero when PS is null ( $\xi_j = -1$ ). Catchability fixed to 1					
Model configurations						
	<i>b</i>		Scientific sample size	Commercial samples size	Coverage of the study area	Additional random effect in sampling intensity $\eta$
					Data sources considered in the model	PS estimated
					Yes	Fixed catchability
<b>Question 1:</b> How does each data source contribute to inferences?	0,1,3		50	50,400,3 000	Full	No
	0,1,3		50,400,3 000	3 000	Full	No
	0,1,3		50	3 000	No fishing in a 9 × 9 cells box	No
<b>Question 2:</b> How does a partial coverage of the study area by the commercial data affect the quality of the estimation?	0,1,3		50	3 000	Full	No
	0,1,3		50	3 000	No	Scientific only, commercial only, both
<b>Question 3:</b> What is the cost of ignoring PS in estimation when sampling is preferential?	0,1,3		50	3 000	Full	No ( <i>b</i> fixed to 0)
<b>Question 4:</b> How does the estimation perform when additional processes other than PS drive the fishing locations?	0,1,3		50	3 000	Yes	Commercial

## Performance metrics

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

The quality of the total biomass estimation (the sum over all grid cells,  $B = \sum_x S(x)$ ) was explored through the relative bias  $\frac{B - \hat{B}}{B}$ , that quantifies how much the total biomass is over or under-estimated. The quality of the estimation of the parameter  $b$  is assessed through the relative bias defined as  $\frac{b - \hat{b}}{b}$  (except for  $b = 0$ , where only the absolute bias is considered). We also assessed the relative bias of the species–habitat relationship estimate  $\hat{\beta}_S$  and range parameter  $\rho$  as these parameters are meaningful for understanding species distribution.

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

### 3.2.3 Case studies

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

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

## **Survey data**

Scientific data (CPUE, in  $\text{kg} \cdot \text{hr}^{-1}$  - Figure 3.3) were derived from the Orhago survey for common sole and EVHOE survey for hake and squids (ICES, 2020a; ICES, 2020b). The sampling density (number of data points per  $\text{km}^{-2}$ ) of those two surveys revealed representative of the sampling density of the main European trawl surveys from the DATRAS database (see SM B.3.2). In comparison, commercial data used in the case studies are denser by 2 orders of magnitude. Scientific data was aligned on commercial data by filtering only individuals above the minimum landing size when available (24 cm for sole and 27 cm for hake — ICES (2020a)). The Orhago survey provides 49 samples for 2017 and 2018 and the EVHOE survey provides 86 samples for 2014 and 2015.

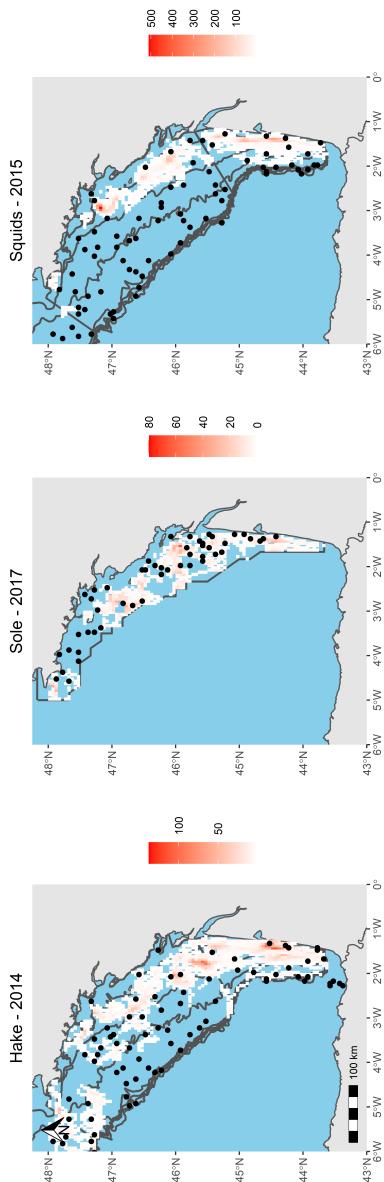


Figure 3.3 – Map of scientific samples (black dot) and commercial sampling distribution (red colour scale-unit: fishing hours). Note that all scientific hauls last around 30 min. Black lines - limits of the spatial domains covered by the scientific survey (Orhago and EVHOE) that delineate the study area. Left - hake, November 2014 (EVHOE; commercial data from otter bottom trawls targeting demersal species OTB\_DEF). Middle - sole, November 2017 (Orhago; commercial data from otter bottom trawls targeting demersal species OTB\_DEF). Right - squids, year 2015 (EVHOE; commercial data from otter bottom trawls targeting cephalopods OTB\_CEP).

## Commercial data

For each species, we filtered commercial data for ‘bottom trawlers’ as they cover a wide part of the study area (Figure 3.3) and provide easy to compute and reliable CPUE. Commercial data were standardized by the fishing effort in  $\text{kg} \cdot \text{h}^{-1}$ . For hake and sole, we filtered the métier targeting demersal fish (called OTB\_DEF) and for squids, the métier targeting cephalopods (called OTB\_CEP).

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

## Habitat covariates

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

## Model configurations

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

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

## Model evaluation

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

## 3.3 Results

### 3.3.1 Simulations

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

#### Contribution of each data source in the IM

Models fitted on scientific data only provide systematically unbiased estimates of total biomass (the mean bias is close to 0 for all sample size - Figure 3.4, 1st row), and the variance of estimations decreases with scientific sample size. Note that the species-habitat relationship estimates  $\hat{\beta}_S$  are also unbiased (see SM B.4.1).

Overall, inferences from the IM revealed consistent with those obtained from scientific data only (SM B.4.2). Even when the commercial sample size is large and the scientific sample size is small, only 3% of the p-values fall below the 0.05 threshold for the fixed effect test (the test wrongly rejects consistency). For the random effect test, the results are more contrasted as 10% of the p-values fall below the 0.05 threshold when data size are very unbalanced (low scientific sample-high commercial sample).

In almost all configurations, the IM provide unbiased and more precise estimates for total biomass and spatial biomass predictions compared to the model fitted to scientific data only (Figure 3.4). As expected, the larger the commercial and the scientific sample size, the more accurate the spatial predictions, the PS parameter  $b$ , and total biomass

estimates. Estimates of  $b$  are unbiased in most cases except when commercial sample size is small and PS is strong (Figure 3.4, 2nd row).

As expected, the contribution of each data sources in the IM directly depends on the balance in the sample size. When sample size is balanced between the data sources, integrating the two data sources in the model systematically improves the inferences with regards to situations where only one data source is analyzed. For instance, for large commercial and scientific sample size (com.L\_sci.L) and no PS, the precision is 1.5 higher (i.e. the MSPE is 1.5 lower) for the IM compared to single-data models (either scientific or commercial - Figure 3.4, 3rd row, 1st column). However, when the sample sizes are unbalanced, the data source with the larger sample size (here commercial data) dominates inference and integrating another data source with a smaller sample size (here scientific data) contributes to a much lesser extent to inference. See, for instance, the situation where commercial sample size is large and scientific sample size is small (com.L\_sci.S - Figure 3.4, 3rd row, 1st column). In this case, the performances of the model fitted to commercial data alone - with reference level fixed to commercial data - are very close to those of the IM whatever the intensity of PS.

Interestingly, the higher the intensity of PS, the higher the benefits of fitting commercial data in the model (Figure 3.4, 3rd row); for instance, when both datasets have large sample sizes (com.L\_sci.L), increasing PS reduces error predictions (i.e. increases accuracy) by 2 each time (i.e. for  $b = 0$ ,  $\mathbb{E}(MSPE) = 20$ ; for  $b = 1$ ,  $\mathbb{E}(MSPE) = 10$ ; and for  $b = 3$ ,  $\mathbb{E}(MSPE) = 5$ ).

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

Also, in unbalanced cases the accuracy of total biomass estimates from the IM revealed highly sensitive to the choice of the reference level (Figure 3.4, 1st row). When the commercial sample size far exceeds the scientific sample size, setting the reference level to the commercial data produces more precise estimates than setting the reference level to scientific data. When defining scientific data as reference level, the intercept of the latent field of relative biomass is estimated from the few scientific samples and resulting estimates are less precise than when defining the reference level with a more numerous data source (here commercial data). This is also true — to a lesser extent—for spatial predictions (Figure 3.4, 3rd row).

In the following, only the case where commercial samples exceed scientific samples and the reference level is fixed with commercial data is explored further as it is the closest to the case studies configuration (Table 1.1).

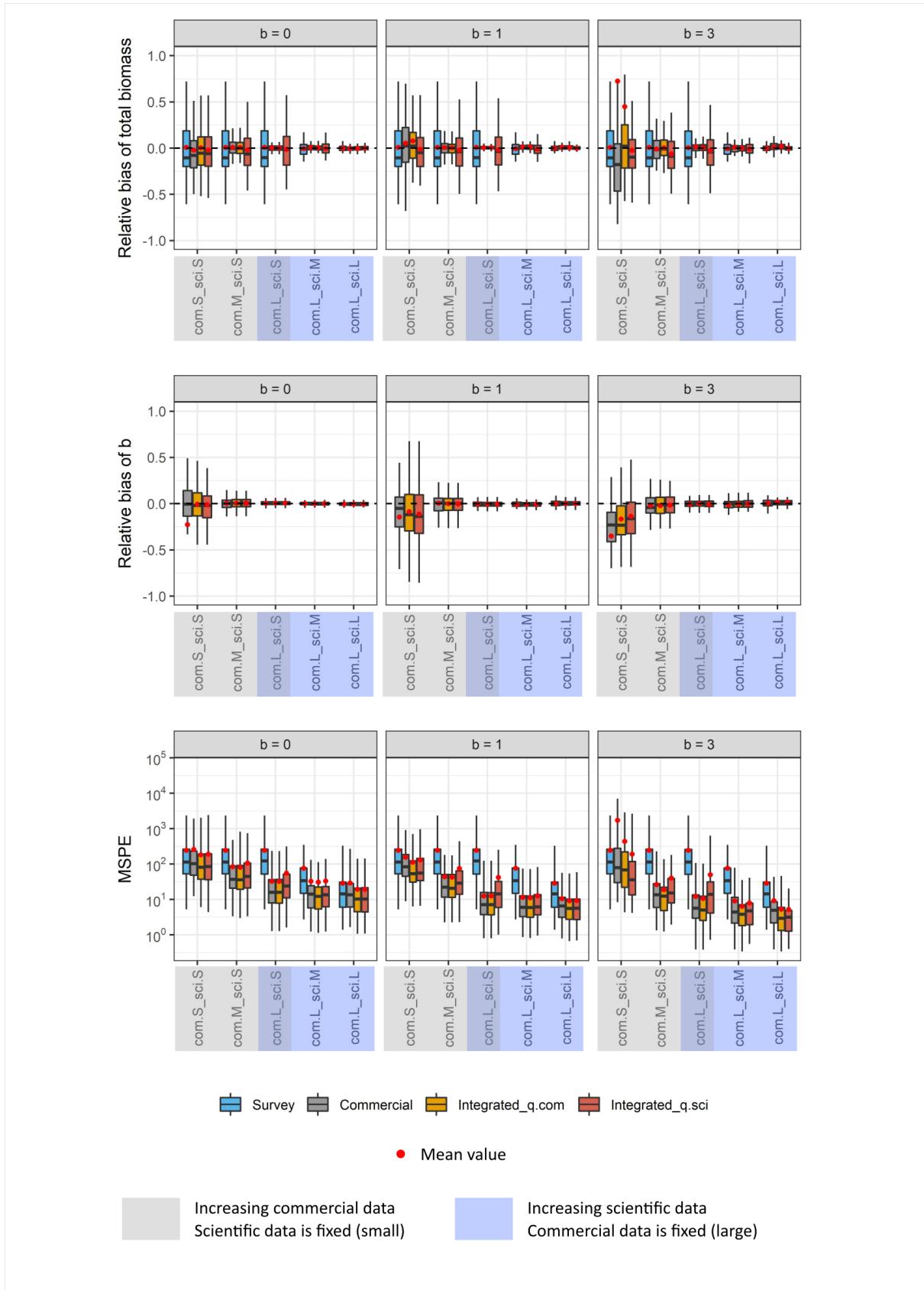


Figure 3.4 – Performance metrics obtained for various commercial and scientific data sample size. Column: intensity of the PS in simulated data. x-axis: five combinations of commercial and scientific sample size. ‘com’ stands for commercial, ‘sci’ stands for scientific, S stands for small sample size (50), M stands for medium sample size (400), and L stands for large sample size (3000). Colours: model configurations. Integrated\_-q.com: IM with catchability fixed to 1 for commercial data; Integrated\_q.sci: IM with catchability fixed to 1 for scientific data. Boxplots represent the variability among the 100 replicates

### **Impact of a partial coverage of the study area by the commercial data**

When commercial data only partially cover the distribution area, commercial data still provide valuable information to predict biomass spatial distribution whatever the PS intensity is (Figure 3.5, 2<sup>nd</sup> column). When sampling is not preferential (data simulated with  $b = 0$ ), a partial coverage of the distribution area produces on average 1.5 less precise spatial predictions but estimates remain unbiased (Figure 3.5, 3<sup>rd</sup> row, comparing 1<sup>st</sup> and 2<sup>nd</sup> column). When sampling is preferential (either moderate or high), biomass estimates are slightly underestimated. Integrating scientific data in the analysis does not correct this bias.

Finally, all model configurations allow for unbiased and precise estimation of the species-habitat parameters  $\hat{\beta}_S$ , whether or not there is a partial coverage of the domain (see SM B.4.1) and overall almost all IM are consistent with scientific-based model (SM B.4.2).

### **How does ignoring PS impact inferences?**

As expected, the impact of ignoring PS in the estimation model is negligible when data is simulated with no PS, and becomes more and more detrimental when the intensity of PS increases in the truth (Figure 3.5, 3<sup>rd</sup> column). With no surprise, when data are generated with no PS ( $b = 0$ ), ignoring PS in the estimation procedure has no effect on the estimation performance. When PS is moderate, total biomass estimates are 5% overestimated ( $b = 1$ ). In the case of strong PS ( $b = 3$ ), ignoring PS in the estimation strongly deteriorates the quality of inferences regarding total biomass estimates (Figure 3.5, 1<sup>st</sup> row, 3<sup>rd</sup> column). Total biomass estimates are overestimated by 50% on average. However, the main spatial patterns are well identified with or without consideration of PS,

even though more precise when accounting for PS (Figure 3.5, 3<sup>rd</sup> row, 1<sup>st</sup> column). SM B.4.4 (Supplementary Figure B.13) presents maps comparing a simulated biomass field and model predictions obtained by considering or ignoring PS when  $b = 3$ . The areas with high biomass values (i.e. where commercial sampling is dense) are well-predicted by the models accounting for PS or not. The main differences are localized in poorly sampled areas where biomass is low. Accounting for PS in estimation allows to interpret the low sampling intensity areas as low-density areas, and therefore, to reduce the bias in those areas (SM B.4.4, Supplementary Figure B.14).

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

### **Effect of other spatially structured processes affecting fishing locations**

As expected, precision of estimates are deteriorated when fishing locations actually depend upon a combination of biomass distribution (PS) and other mechanisms (here captured by a spatially structured random term - Figure 3.5, 4<sup>th</sup> column). In this case, the IM still provides valuable inferences on fish distribution, fish total biomass and estimates of  $b$ , although estimations are less accurate than the base case. For instance, MSPE are five times lower when nothing else than PS affects sampling locations compared with a case where sampling locations depend on both PS and other independent spatial processes (Figure 3.5, 3<sup>rd</sup> row, 1<sup>st</sup> and 4<sup>th</sup> column). But interestingly, the weight of scientific data increases when the sampling distribution of commercial data is blurred by spatial processes independent from biomass spatial distribution. MSPE and relative bias provided by the IM are both 1.4 smaller compared to those obtained when the model is fitted to commercial data only.

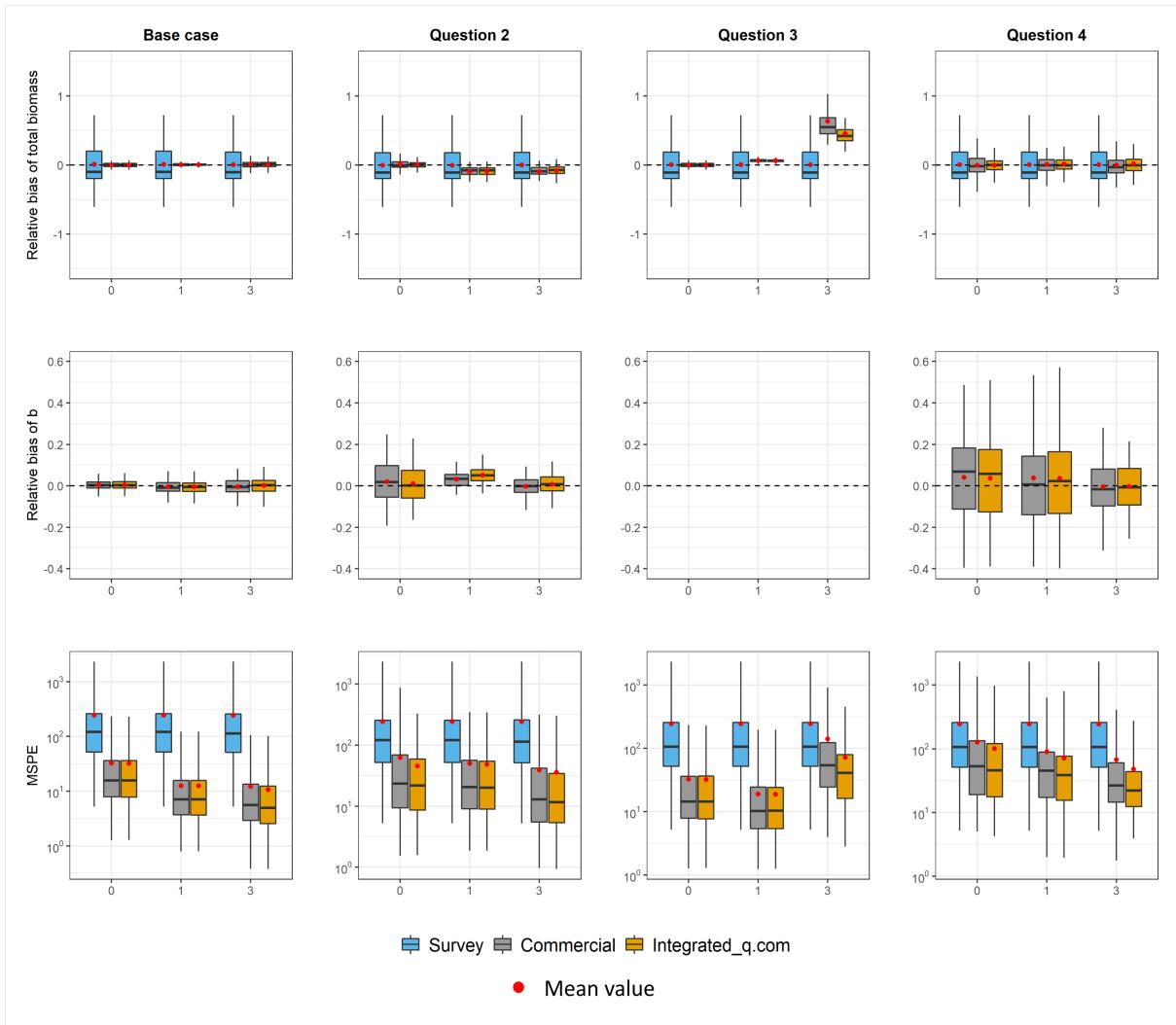


Figure 3.5 – Performance metrics obtained in different data and model configurations. Red points: mean value. 1<sup>st</sup> column: no discrepancy between simulation and estimation. 2<sup>nd</sup> column: commercial data do not cover a  $9 \times 9$  zone of the grid. 3<sup>rd</sup> column:  $b$  is arbitrarily fixed to 0 in the estimation models. 4<sup>th</sup> column: data simulated with a random effect  $\eta$  in the sampling intensity process. In all configurations, simulations are conducted for three levels of PS (x-axis:  $b = 0$ ,  $b = 1$ , and  $b = 3$ ). Colours: data sources used in the IM for inferences. Integrated\_q.com: IM with catchability fixed with commercial data. Boxplots represent the variability among the 100 replicates.

### 3.3.2 Case studies

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

#### Contribution of each dataset to the inferences

Almost all the case studies successfully passed the consistency test between the IM and the model fitted to scientific data only (see SM B.5.1). Models based on scientific data provide different spatial predictions compared with the IM. Predictions for sole and squids from the scientific-based model are mainly shaped by the covariate effects (Figure 3.6; for further analysis see SM B.5.2, SM B.5.3, and SM B.5.4). On the other hand, predictions from the IM are mainly shaped by the spatial random effect as commercial data allow to better capture the local spatial correlation structures. Consistently with simulations, inferences from the IM are mainly driven by the commercial data (Figure 3.6). This logically arise from the much larger sample size of commercial data compared with scientific data, combined with the good coverage of commercial data in high-density areas (Figure 3.3).

On the other hand, predictions from the IM are mainly shaped by the spatial random effect as commercial data allow to better capture the local spatial correlation structures.

Consistently with simulations, inferences from the IM are mainly driven by the commercial data (Figure 3.6). This logically arise from the much larger sample size of commercial data compared with scientific data, combined with the good coverage of commercial data in high-density areas (Figure 3.3).

As commercial data is denser than scientific data, they will better capture local spatial correlation structures than scientific data. SM B.5.5 provides some additional analysis of the information brought by commercial data in the IM.

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

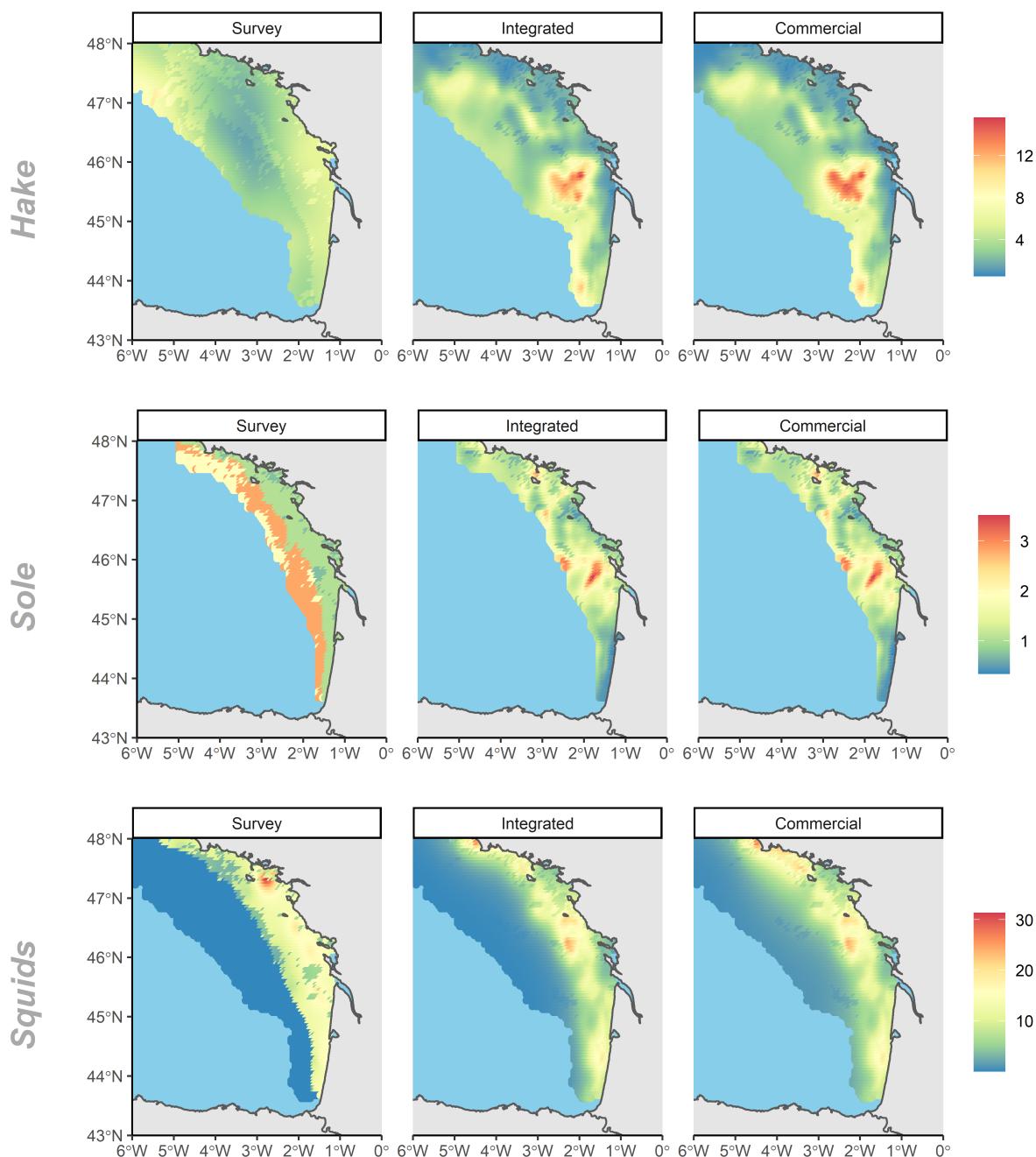


Figure 3.6 – Prediction of the relative biomass for each case study. 1<sup>st</sup> column: model fitted to scientific data only; 2<sup>nd</sup> column: IM accounting for PS; and 3<sup>rd</sup> column: commercial-based model accounting for PS. When the model is fitted to scientific data only, relative biomass is rescaled with the relative catchability parameter estimated within the IM so

that all maps are in the same scale.

### **PS and other processes affecting fishing locations**

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

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

Consistently with simulations, the higher the PS intensity, the higher the differences between inferences obtained with and without considering PS. When comparing biomass field values (Figure 3.7, left column), ignoring PS increases predictions in poorly sampled areas (all red areas - compare with Figure 3.3). This effect is particularly marked for the squid case study where the relative difference is the strongest in the offshore areas. However, considering PS or not has relatively little effect in areas where sampling is spatially denser (all white areas). Ignoring PS affects total biomass indices estimates and the relative difference between biomass estimates with or without PS increases with the value of  $b$  estimates (Figure 3.7, right column).

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

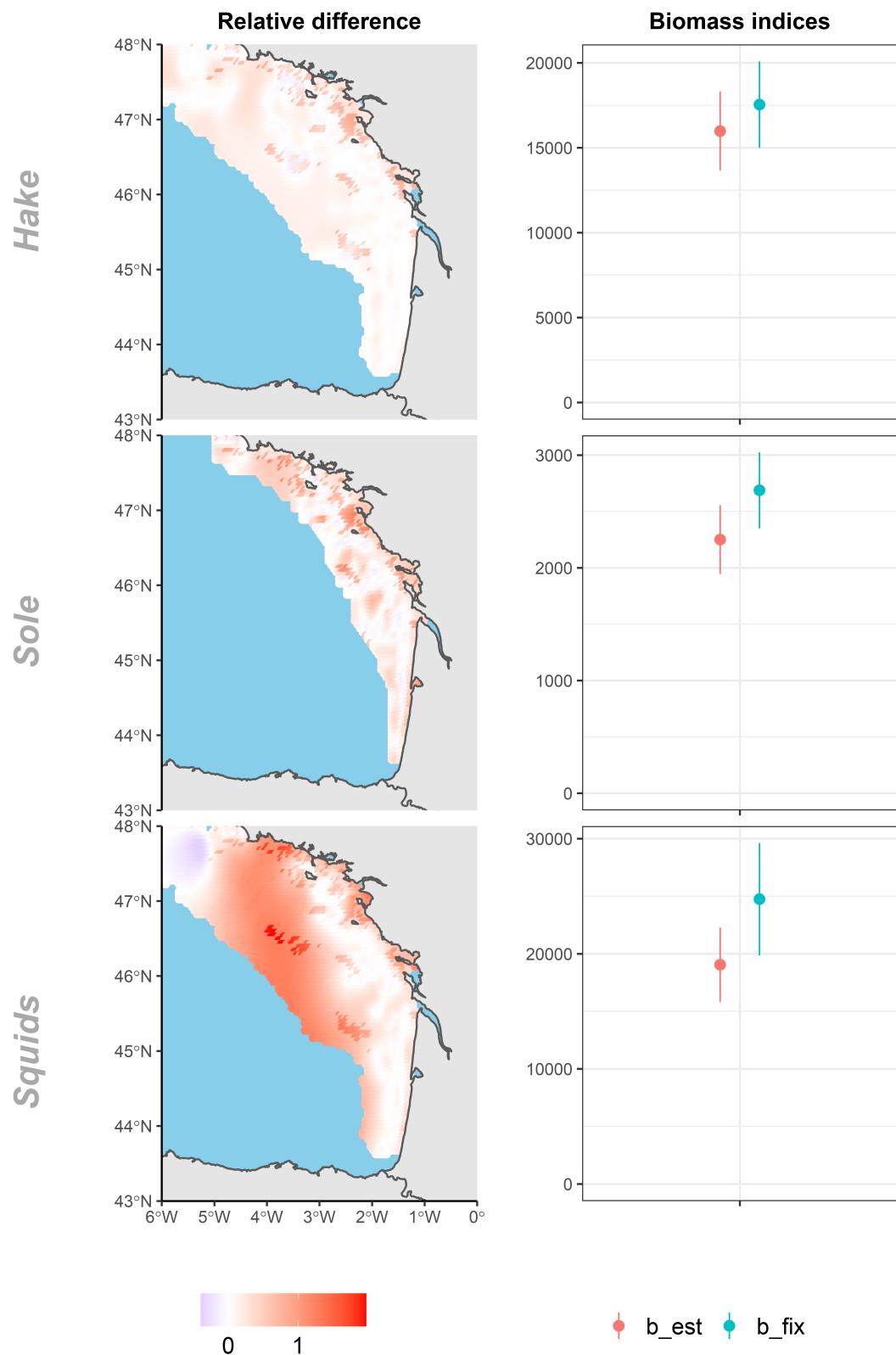


Figure 3.7 – Relative difference in biomass spatial predictions between IM accounting or not for PS for the three case studies (left). Comparison of the total biomass estimates obtained from the IM when accounting or not for PS (right).  $b_{est}$ : PS is estimated.  $b_{fix}$ : PS is not accounted for. The relative bias is calculated as  $(S_{bfix}(x) - S_{best}(x))/S_{best}(x)$ . The total biomass is computed as the sum of the latent field values on the spatial domain.

### **Benefits of considering different fleets in the estimation model**

Based on the sole case study, we demonstrate the capacity of the model to integrate multiple commercial fishing fleets, each with specific parameters (catchability and PS behavior). In the sole case studies, considering two different fleets in the IM (instead of one homogeneous) improves goodness-of-fit towards scientific data (SM B.5.10, y-axis) and modifies spatial predictions (SM B.5.12).

## **3.4 Discussion**

### **3.4.1 Main findings**

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

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

Both simulations and case studies demonstrate that ignoring PS in the inference may be highly detrimental when the intensity of PS is strong. The present framework can serve as a tool to assess the benefit of including PS in analysis, depending on the intensity of PS but also on the modeling objectives. As already shown in previous studies (Conn, 2010; Pennino et al., 2019), when PS actually occurs in commercial catches, ignoring this

process may bias inferences on total biomass estimates. Even if ignoring PS may not hamper the capacity to detect areas of high biomass, the biomass in low-density areas may be overestimated. Therefore, if the objective is to compute biomass indices integrated over a large area, then it might be worth accounting for PS to avoid biased results. In contrast, if the objective is to identify hotspots, the benefits of considering PS may be small with regard to the additional computational time it requires.

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

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

### **3.4.2 Challenges in modeling preferential sampling**

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

Of course, many other factors may drive the spatial intensity of fishing, and those were simply captured in our model through an additional spatial random term. For

instance, fishers' behavior may depend on prior knowledge of fish spatial distribution, on information sharing within fishing cooperatives, on expected distribution of bycatch species, or logistical constraints (e.g. transit costs) (Salas and Gaertner, 2004; Haynie, Hicks, and Schnier, 2009; Girardin et al., 2017). Targeting behavior may also be directed towards an assemblage of species rather than toward a single species (Bourdaud et al., 2019).

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

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

### **3.4.3 Relative contribution of scientific and commercial data**

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

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

process of commercial data might partly explain these inconsistencies.

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

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

### **3.4.4 The limits of reallocated catch data**

Probably one of the major limits of our approach is that the actual framework ignores the uncertainty that arises from the procedure used to reallocate the catch declarations on fishing locations. Obtaining the spatialized CPUE inputs used in the model requires pre-treatment of the commercial catch declaration data to allocate declaration data to VMS positions (Hintzen et al., 2012). Raw data corresponds to fishing operations that are daily aggregated and reported at coarse administrative spatial units ( $0.5^\circ$  latitude by  $1^\circ$  longitude rectangles). These declarations are then reallocated uniformly on all GPS locations previously identified as fishing in the vessel path. This procedure has been

demonstrated to be robust while being a fast and a pragmatic approach for reallocating landings to VMS pings (Gerritsen and Lordan, 2011; Murray et al., 2013). However, it implies strong hypothesis that may artificially increase or transform the information provided by the data. Typically, the uniform reallocation of catch declarations on all GPS positions identified as fishing may smooth the spatial signal, which could potentially explain the lack of species–habitat relationship obtained from the IM. The effect of such reallocation should be explored in further study to better understand its consequences on model predictions/estimates and further model development should investigate how to mitigate its consequences.

### **3.4.5 Perspectives**

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

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

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

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

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

# **IDENTIFYING MATURE FISH AGGREGATION AREAS DURING SPAWNING SEASON BY COMBINING CATCH DECLARATIONS AND SCIENTIFIC SURVEY DATA**

---

In the first chapter, we developed a purely spatial species distribution model integrates scientific data with commercial catch declarations data while accounting for PS of the commercial data. We assessed the model through simulations and applied the approach on three contrasted case studies in the Bay of Biscay on the time span of the survey.

In the following chapter, we extend the model in time and moved to a spatio-temporal analysis of the data. This enables to infer species distribution outside the period of the survey and then to identify essential habitats (here spawning grounds) that potentially mismatch the scientific survey period. Moving to the time dimension also allows to quantify the temporal variation of PS, that is interpreted with regards to the seasonality of fishing behavior. We applied the approach on three contrasted case studies regarding the available literature knowledge on fish spawning grounds. We illustrate how the model outputs provide consistent aggregation areas with known spawning grounds and discuss alternative uses of the model outputs as well as further extensions of the actual spatio-temporal model.

This chapter led to a submission in the Canadian Journal of Fisheries and Aquatic Sciences and is currently under review.

## **Abstract**

Identifying and protecting essential fish habitats like spawning grounds requires an accurate knowledge of fish spatio-temporal distribution. Commercial declarations coupled with Vessel Monitoring System provide fine scale information on the full year to map fish distribution and identify essential habitats. We developed an integrated framework to infer fish spatial distribution on a monthly time step by combining scientific and commercial data while explicitly considering the preferential sampling of fishermen towards areas of higher biomass. We developed a method to identify areas of persistent aggregation of biomass during the spawning season and interpret these as spawning areas. The model is applied to infer maps of relative biomass for three species (sole, whiting, squids) in the Bay of Biscay on a monthly time step over a 9-year period. Integrating several fleets in inference provide a good coverage of the area and improves model predictions. The preferential sampling parameters give insights into the temporal dynamics of the targeting behavior of the different fleets. Last, persistent aggregation areas reveal consistent with the available literature on spawning grounds, highlighting the potential of our approach to identify reproduction areas.

*Keywords:* Species distribution model, Spatio-temporal model, Hierarchical model, VMS and logbook data, Fish reproduction areas, Template Model Builder (TMB)

## 4.1 Introduction

Integrating fisheries into Marine Spatial Planning (MSP) to preserve ecosystem functions and ensure a sustainable exploitation requires an accurate knowledge of fish spatio-temporal distribution and more specifically of fish essential habitats such as reproduction and nursery grounds (Janßen et al., 2018). However, such knowledge is still missing for many species due to a lack of data with sufficient spatial, temporal or demographic resolution (Delage and Le Pape, 2016; Regimbart, Guitton, and Le Pape, 2018).

The available data to map fish distribution and identify essential habitats mainly rely on either scientific survey data (fishery-independent data) or commercial data available through on-board observer programs (fishery-dependent data) (Pennino et al., 2016). Both data sources benefit of direct on-board recording of catches and are usually considered as high quality data. Furthermore, both data sources were proved to be complementary (Rufener et al., 2021). Scientific data benefit from a standardized sampling plan, a standardized catchability and occur each year at the same period. Consequently, they provide standardized data on a large spatial extent for most species and size classes (Hilborn and Walters, 1992; Nielsen, 2015). Observer data potentially provide data over the full year for all caught species, even though they do not follow a standardized protocol as survey data. However, both scientific survey and onboard observer data are characterized by a relatively low sampling intensity in space and time. Because of material limitations, surveys occur only once or twice a year and provide a limited number of sample each time (ICES, 2005) and observer programs only cover a limited fraction of the entire fleet (e.g. only 1% of all sea trips are covered by the French observer programs - Cornou et al. (2021)). The low sampling density of both data sources may lead to imprecise predictions (ICES, 2005; Alglave et al., 2022) and constrains to consider only rough temporal resolution (e.g. semesters, quarters or seasons – see for instance Kai et al. (2017), Pinto et al. (2019), and Rufener et al. (2021)) to ensure a satisfying spatial coverage of the data at each time step. However, the temporality of key biological events, such as the reproduction peak, may be much tighter than the temporal resolution of data (Biggs et al., 2021). Hence, those data alone are likely not sufficient to provide accurate inferences on essential fish habitats such as spawning grounds.

Commercial catch declarations combined with their fishing locations available from VMS (Vessel Monitoring System) were proven to be an interesting alternative to obtain landing per unit effort (LPUE) data with fine spatial and temporal resolution (Pedersen,

Fock, and Sell, 2009; Bastardie et al., 2010; Gerritsen and Lordan, 2011; Hintzen et al., 2012; Murray et al., 2013; Azevedo and Silva, 2020). However, considering commercial fisheries data to infer fish spatial distribution remains highly challenging. Among other challenges, this implies accounting for fishermen sampling behavior. Fishermen typically tend to preferentially sample areas of higher biomass (a process referred to as preferential sampling, PS - Diggle, Menezes, and Su (2010)). Hence, because data preferentially represent areas of highest biomass, ignoring PS in the distribution of fishing effort when estimating spatial distribution on larger areas can lead to overestimated biomass predictions (Conn, Thorson, and Johnson, 2017; Pennino et al., 2019; Alglave et al., 2022).

In a recent paper, Alglave et al. (2022) developed an integrated modeling framework to infer spatial distribution of fish abundance by combining scientific survey CPUE and commercial LPUE data while accounting for PS in the distribution of fishing effort. They applied their framework to commercial data of a single month to match with the scientific survey and did not consider any temporal dimension.

In this paper, we extend the modeling framework from Alglave et al. (2022) by adding a temporal dimension to estimate fish spatio-temporal distribution at a monthly time step. Our new model accounts for the variation over time (monthly time step) in the biomass field as well as in the intensity of PS for distinct fishing fleets. To demonstrate the value of the method, we selected and applied the model to 3 demersal species in the Bay of Biscay (common sole, whiting and squids) characterized by contrasted configurations regarding the available knowledge of their spawning grounds. We used those applications to reinforce results obtained in Alglave et al. (2022) demonstrating how the integrated framework combines the information from several fleets in order to produce accurate maps of spatio-temporal biomass. To illustrate the capacity of the framework to identify areas of aggregation during the spawning season, we processed model outputs to identify areas of recurrent aggregation occurring during the reproduction season and compared these to the information available in the literature.

## 4.2 Material and methods

In this section, we first present the different species, the datasets and how we process and combine them to produce LPUE data in space and time. Second, we extend the model proposed by Alglave et al. (2022) to introduce a temporal dimension on a discrete monthly time step. In our applications, the models were fitted to data from 2010 to 2018 on a monthly time step (108 time steps). Then, we illustrate how the PS component modifies model predictions and can be interpreted, and how integrating several fleets in the analysis further improves model predictions. Last, we detail the method used to investigate spatio-temporal dynamics from model outputs and identify reproduction grounds based on the aggregation patterns of each of three species.

### 4.2.1 Case studies

Sole is a data-rich case. Direct information about reproduction grounds is available through egg and larvae surveys (Arbault, Camus, and Bec, 1986). Reproduction period fall between January and April, but the peak of the reproduction fall in February (Figure 4.7). Discard rates is also very low, which makes the landings data a good proxy of the catch (ICES, 2019a).

By contrast, Whiting is a data-poor case study where only indirect information of reproduction period exists through spring trawl surveys (Houise and Forest, 1993). Reproduction period fall between March and May. Discard rates can be high (about 30%) and thus landings data may provide a biased picture of the real catches (ICES, 2019b).

Our third case study is Squids that represent a mixture of several species declared under a common denomination in the catch (Loliginidae here referred as squids): *Loligo vulgaris* (Lamarck, 1798), *Loligo forbesii* (Steenstrup, 1856) and *Alloteuthis sp* (Lamarck, 1798). Overall scientific survey suggest that the predominant species in the Bay of Biscay is *Loligo vulgaris* (ICES, 2020b). All 3 species are data-poor: no information exist regarding their reproduction grounds but some information of the reproduction period exist for *Loligo Vulgaris* (Moreno et al., 2002). For this species, the reproduction period fall between January and April.

## 4.2.2 Data

### High spatial resolution catch per unit effort data for the mature component of the populations

We pre-processed the VMS and catch declaration (logbook) data to obtain high spatial resolution LPUE data for the mature component of those three stocks, for three different fishing fleet, and for each month of the 2010-2018 time series. In the text, we used the term Landing Per Unit of Effort (LPUE) to refer to commercial observations expressed in kg of (mature) biomass per hour fished. Discards are neglected in our approach, hence LPUE are considered as biomass indices.

Our model can integrate data of different fishing fleets. For the purpose of our application, we selected three different métiers that belong to the same fleet of trawlers (in this case, the ‘métier’ term refers to a combination of gear and of a set of species that are targeted by the vessels): OTB\_DEF (bottom otter trawl targeting demersal species), OTB\_CEP (bottom otter trawl targeting cephalopods) and OTT\_DEF (multi-rig otter trawl targeting demersal species). In the following those three métiers are referred as three fleets. These three fleets were selected as they offer three main advantages. First, their targeting behaviors and technical characteristics are similar. Second, catch per unit effort of trawlers are generally good indicator of fish relative abundance while other gears (longline, gillnet) may face saturation effects leading to non-linear relationship between catches and fishing time (Hovgêrd and Lassen, 2008). Third, the combination of the three fleets cover the full spatial domain (Figure 4.1).

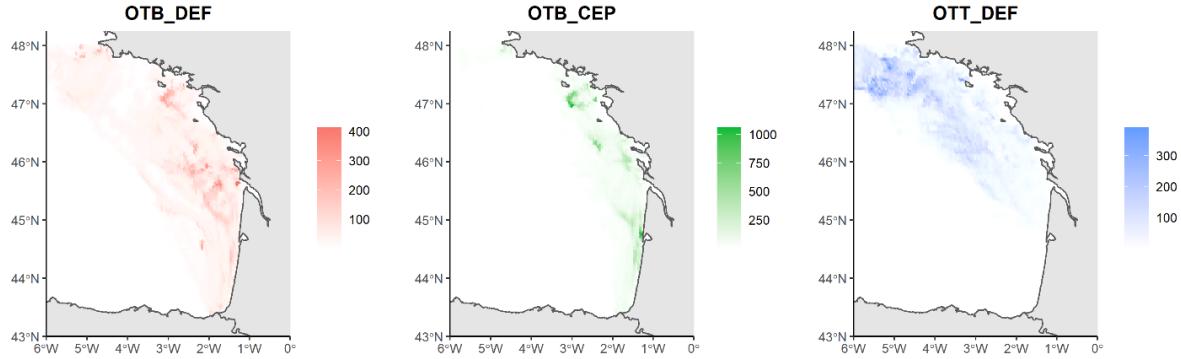


Figure 4.1 – Spatial distribution of each fleet on the whole period (2010-2018). Unit: fishing effort in fishing hour.

Because one of our primary goal is to identify spawning grounds, we filtered only the mature fraction of the landings (i.e. the fraction of the individuals that can potentially reproduce, not per se the fraction of the population that are spawning – this is detailed further in the discussion). This was done by crossing the landings data with length class and maturity data. For most of the landings, information on the commercial size categories is available from the sales notes. These commercial categories are regularly sampled to derive length structure of each commercial category. This enables to estimate the proportion of potentially mature fish in each commercial category by applying maturity ogives and in turn estimate the proportion of mature fish for each landing declaration. See SM C.1 for more detail. Note that this procedure was not possible for squids, as there are no data on maturity and size classes for this species group.

Landing data were then combined with VMS data to finally obtain high spatial resolution LPUE data discretized on a  $0.05^\circ \times 0.05^\circ$  grid (i.e. 5.5 km x 3 km) on a monthly time step (see SM C.2). This combination requires:

1. to identify the fishing locations within the VMS data. This is realized based on a speed threshold similarly as in common data processing methods (Hintzen et al., 2012).
2. to reallocate the logbook declaration on the related VMS fishing locations. This reallocation is realized individually for each fishing vessel trajectory by uniformly reallocating the landings on all fishing locations. The link between both data sources is realized through the combination 'vessel identifier x statistical rectangle'

x fishing trip x day x gear'. LPUE are then computed by simply dividing the reallocated landings by the related fishing time.

## Scientific Data

We also integrated scientific data in the analysis. For whiting and squids, we used the survey data from the EVHOE survey. The Orhago survey was used for sole (ICES, 2020 - see SM C.3, Figure C.3). The data were extracted from the DATRAS database on the period 2010 - 2018. Only the mature fraction of the survey catches were kept in the analysis to make it comparable with commercial data.

Orhago is an annual beam trawl survey occurring in November and designed to assess sole stock status in the Bay of Biscay. Each year 50 stations are sampled within 4 strata all along the Bay of Biscay. Note that this survey is mainly coastal and does not sample offshore areas. EVHOE is an annual bottom trawl survey occurring in late October, November and early December with a stratified sampling plan. It is designed for demersal fishes in the Bay of Biscay and in the Celtic Sea. In the Bay of Biscay, 80 to 90 sampling hauls are recorded each year.

### 4.2.3 Spatio-temporal integrated model

Alglave et al. (2022) developed a hierarchical integrated statistical model to infer spatial distribution of fish density through scientific survey data and commercial data. It is structured in 4 layers:

- the latent field that represents biomass spatial distribution, and that is the main target of the inferences;
- the observations from scientific surveys and commercial declarations that are considered as direct zero-inflated observations of the latent field at the registered fishing locations;
- the fishing sampling intensity that relates fishing locations to the latent field and model explicitly the PS of commercial fleets towards areas of higher biomass;
- unknown parameters that control the shape of the biomass latent field and the sampling process.

This first model was purely spatial as no temporal dimension was included in the model. In this paper, we extend the framework by incorporating a temporal component to model the evolution of the latent field of biomass across the monthly time steps (Figure

4.2).

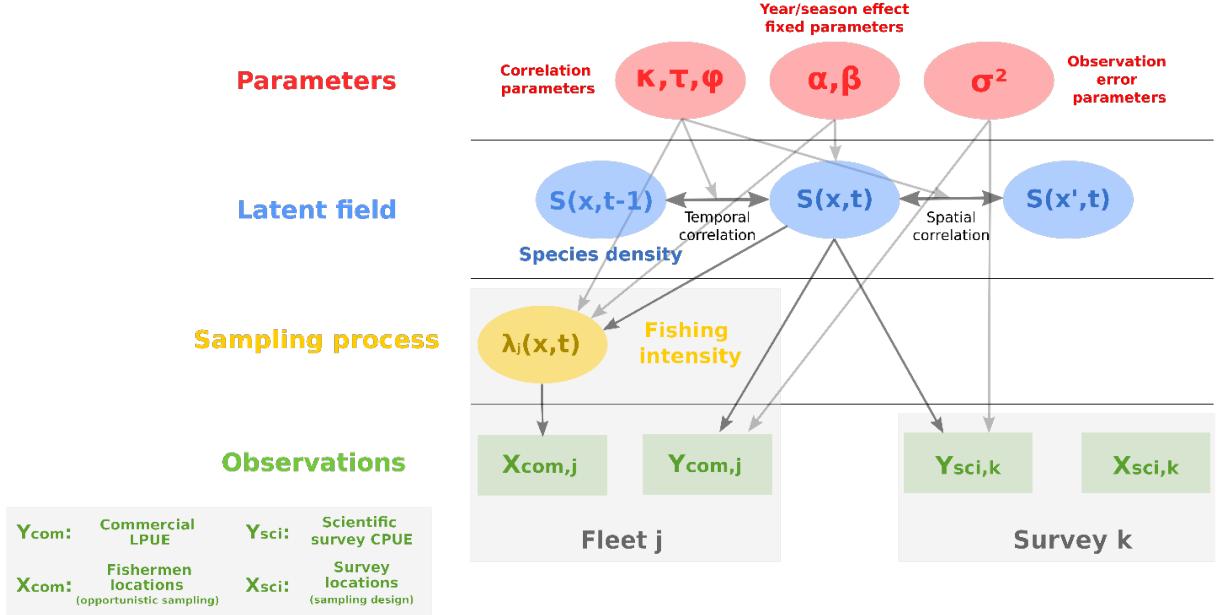


Figure 4.2 – Diagram of the integrated spatio-temporal model.

### Biomass field

As a notable extension of Alglave et al. (2022), the biomass field ( $S$ ) (eq. 4.1) is modeled as a spatio-temporal Gaussian Random Field (GRF) through a log link as:

$$\log(S(x, t)) = \alpha_S(t) + \delta(x, t) \quad (4.1)$$

where  $x \in \mathcal{D} \subset \mathbb{R}^2$  stands for the spatial locations and  $t \in [1, T]$  for the monthly time steps.  $S(x, t)$  is in the same unit as the data (here kg/hr fished as data are CPUE for survey trawls or LPUE for commercial landing declarations). The term  $\alpha_S(t)$  is a time varying intercept modeled as a fixed effect and  $(\delta)$  is a GRF spatio-temporal process which represents the spatio-temporal correlation structure of the biomass field. As commercial data may not always cover the full area, the temporal correlation component allows to interpolate between time-steps. Here, the spatio-temporal term has a classical stationary first-order autoregressive form (eq. 4.2) following (Cameletti et al., 2013):

$$\delta(x, t) = \varphi \cdot \delta(x, t - 1) + \omega(x, t) \quad \text{for } t = 2, \dots, T \quad (4.2)$$

The autocorrelation coefficient  $\varphi$  is a scalar with  $\varphi \in ] -1, 1[$ ,  $\omega(x, t)$  represents the spatial innovation and is modeled as a 0 mean GRF (with no temporal correlation). Spatial random effects are parameterized through a range parameter  $\rho$  that corresponds to the distance at which spatial autocorrelation falls below 0.1.

Note that no covariate is included in the latent field to keep the model as simple as possible. If any, the covariates effects are captured through the spatio-temporal term  $\delta(x, t)$ . Similarly, the intercept  $\alpha_S(t)$  was modeled through a simple fixed effect but more complex specifications including some seasonal, yearly and interaction effects could be adopted such as in Thorson et al. (2020).

### **Sampling process for the commercial fishing points**

As the scientific survey sampling plan is designed independently from the biomass field, scientific sampling locations do not need to be modeled explicitly (Diggle, Menezes, and Su, 2010). By contrasts, the dependence between the fishing locations and the biomass field has to be modeled to capture preferential sampling. We extended the model proposed by Alglave et al. (2022) to account for temporal variations in PS. Because the fishing behavior is potentially different among fishing fleets, PS is modeled specifically for each fleet  $j$ . Observed fishing locations are modeled as an inhomogeneous point process ( $X_{comj}$ ) (see Figure 4.2) whose intensity  $\lambda_j(x, t)$  (eq. 4.3) controls the expected number of fishing points within a given area:

$$\log(\lambda_j(x, t)) = \alpha_{Xj}(t) + b_j(t) \cdot \log(S(x, t)) + \eta_j(x, t) \quad (4.3)$$

where:

- the time varying intercept  $\alpha_{Xj}(t)$  quantifies the average fishing intensity on the whole area; as the biomass intercept  $\alpha_S(t)$  (eq. 4.1), it is modeled as a fixed effect;
- the time varying  $b_j(t)$  quantifies the strength of PS; it is modeled as a fixed effect too. If  $b_j(t) = 0$ , then PS is null. If  $b_j(t) > 0$ , then PS occurs and the greater, the stronger PS. Alternatively,  $b_j(t) < 0$  means that fishermen have a repulsive behavior towards the resource.
- the pure spatial GRF ( $\eta_j$ ) captures the remaining spatial variability in the fishing point pattern not captured by the PS term (for instance, dependence of the fishing locations towards management regulations, distribution of other targeted species, habits/tradition).

## Observation process

All observations ( $Y$ ) for both scientific and commercial data of any fleet  $j$  are assumed all mutually independent conditionally on the latent field of biomass and the sampling locations. As data (both scientific and commercial) eventually present a high proportion of zero values, we model the observations through a Poisson-link zero-inflated model introduced by Thorson (2018) and already used in Alglave et al. (2022). The observation model explicitly considers that each fleet can have its own catchability and its own zero inflation parameter.

The probability to obtain catch  $y_i$  conditionally on the location  $x_i$  (with  $i$  the observation index), the time-step  $t_i$ , the biomass field value  $S(x_i, t_i)$  and the fleet  $j$  is expressed as follow:

$$P(Y_i = y_i | x_i, t_i, S(x_i, t_i), j) = \begin{cases} p_i & \text{if } y_i = 0 \\ (1 - p_i) \cdot L(y_i, \frac{\mu_j(x_i, t_i)}{(1-p_i)}, \sigma_j^2) & \text{if } y_i > 0 \end{cases} \quad (4.4)$$

$$p_i = \exp(-e^{\xi_j} \cdot \mu_j(x_i, t_i)) \quad (4.5)$$

$\mu_j(x_i, t_i) = q_j \cdot S(x_i, t_i)$  is the expected catch of fleet  $j$  at location  $x_i$  and time step  $t_i$ . It is the product of the latent field value  $S(x_i, t_i)$  and of the relative catchability coefficient of fleet  $j$  denoted  $q_j$ .  $\xi_j$  is a zero-inflation parameter controlling the proportion of zero in the data,  $\sigma_j^2$  is the observation variance when the catch is positive.

Equation 4.5 shows the two components that compose the probability to observe a catch  $y_i$ :

- the probability to obtain a zero catch ( $y_i = 0$ ). It is modeled as a Bernoulli variable with probability  $p_i = \exp(-e^{\xi_j} \cdot \mu_j(x_i, t_i)) \cdot p_i$  is equivalent to the probability to obtain a 0 value with a Poisson distribution of intensity  $e^{\xi_j} \cdot \mu_j(x_i, t_i)$ . The value of  $\xi_j$  controls the intensity of the zero inflation, when it increases the amount of zero decreases). Then the probability to obtain a positive catch is given by  $1 - p_i$ .
- the value of the positive catch is modeled through a lognormal distribution  $L$  with expected value  $\mu_j(x_i, t_i)/(1 - p_i)$  and observation error  $\sigma_j^2$ . The standardization by  $(1 - p_i)$  allows to keep the expectancy of the observation model to  $\mu_j(x)$ .

The catchabilities  $q_j$  are not identifiable per se and some additional constraints need to be set to estimate the relative catchability of each fleet (Alglave et al., 2022). To ensure

identifiability, one fleet catchability is set as reference level (e.g.  $q_{ref} = 1$ , here OTB\_DEF was used as the reference fleet) and the other fleets' catchabilities are estimated relatively to the reference fleet through the equation:

$$\mu_j(x, t) = q_j \cdot S(x, t) \quad (4.6)$$

## Maximum likelihood estimation

The estimation of the spatio-temporal model is achieved through maximum likelihood estimation. We used the Stochastic Partial Differential Equation (SPDE) approach that allows to benefit from the nice computational properties of Gaussian Markov Random Fields while working on a continuous domain (Lindgren, Rue, and Lindström, 2011), as well as Template Model Builder (TMB - Kristensen et al. (2016)) which benefits from the Laplace approximation, automatic differentiation and sparse matrix computation technics for a fast estimation of the model through maximum likelihood estimation. Details on estimation are provided in SM C.4, C.5 and C.6.

### 4.2.4 Evaluating the interest of integrating multiple fleets

Integrating several fleets in inference allows to cover the whole area (Figure 4.1) and is expected to improve inferences. To illustrate the value of integrating the data from multiple fleets within a single integrated model, we compared the spatial predictions obtained by fitting the model to all available data with those obtained by integrating only one fleet. In addition, we investigated if integrating all the fleets in inference increased the correlation between scientific data and model predictions.

We also compared the coefficient of variation of the prediction between each model (for November 2018).

Note that scientific data was systematically integrated into inference (either in the integrated model or in the single-fleet models). However, due to the low sample size compared with intensive ‘VMS x logbook’ data (about 80 scientific samples each year in November compared with 17000 samples per month on average), they have very low contribution to inference. This was extensively discussed in Alglave et al. (2022). Here they mainly provide some standardized and reference data to assess the performance of the framework.

#### **4.2.5 Evaluating the value of modeling preferential sampling**

##### **Comparing the inferences with and without PS**

We first assessed the impact of PS on the distribution of biomass by comparing estimations obtained from integrated models (i.e. models fitted to all data sources) accounting for PS with those obtained when ignoring PS. We computed the log-likelihood related to each data source (commercial and scientific data) to assess if there is an improvement in model goodness-of-fit when accounting or not for PS. Note that fitting a model without PS is straightforward as it only requires to remove the sampling process component from the likelihood function.

##### **Interpreting the intensity of preferential sampling**

The estimates of PS parameters  $b_j(t)$  in (eq. 4.3) may bring valuable information on the dynamics of the fishery as they inform on the strength of the relationship between commercial sampling distribution and species distribution. We investigate the variability of the PS parameters among the three species and the different fleets. Then, focusing on the sole case study, we highlight the insights brought by the model on the temporal evolution of PS and its seasonal variations.

The estimates of PS parameters may bring valuable information on the dynamics of the fishery as they inform on the strength of the relationship between commercial sampling distribution and species distribution. We investigate the variability of the PS parameters ( $\mathbf{b}$ ) by representing the variability of the different  $\hat{\mathbf{b}}$  parameters for the three case studies and the different fleets. Then, focusing on the sole case study, we highlight the insights brought by the model on the temporal evolution of PS and its seasonal variations.

#### **4.2.6 Investigating spatio-temporal dynamics and identifying re-production grounds**

The spatio-temporal model provides some insight on the temporal dynamics of species distribution both at inter- and intra-annual levels. We applied a method to identify recurrent aggregation areas from the maps of abundance inferred at each time steps.

## Aggregation index

We used the Getis and Ord index ( $G_d$ ) (Getis and Ord, 1992; Ord and Getis, 1995) to determine persistent aggregation areas (see for instance Milisenda et al. (2021)). The generalized version of the Getis and Ord index is given in Bivand and Wong (2018) and Ord and Getis (1995). Basically, ( $G_d$ ) is a normalized version of the ratio between the sum of the log-biomass  $\log S(x, t)$  within a fixed neighborhood  $d$  and the sum of  $\log S(x, t)$  on the entire area (for a fixed time step) (Getis and Ord, 1992). We computed these indices on  $\log S(x, t)$  so that the variable used to compute ( $G_d$ ) are Gaussian, which makes ( $G_d$ ) Gaussian too. In the application, we used a neighborhood distance  $d = 7.5$  km which defines a small neighborhood of 8 cells (the direct neighbors of each cell grid) and allows to identify very localized aggregation areas. Positive values for the aggregation index  $G_d(x, t)$  indicates that location  $x$  falls within a local patch of high values while negative  $G_d(x, t)$  indicates that location  $x$  falls within a local patch of low values. Near 0 values  $G_d(x, t)$ , indicates that location  $x$  does not fall in some local aggregation patch. As ( $G_d$ ) follows a standardized Gaussian distribution, the comparison between the value of the index and the quantiles of a standard Gaussian distribution can be used to evaluate whether or not the latent field of biomass fall within a statistically significant high or low aggregation patch. We used the quantile 99% (2.58) as a threshold to ensure a high level of significance for patch detection (only local patch of positive values are considered) and applied the Bonferroni correction to account for the multiple statistical tests that are conducted.

Then, we define the persistence indices  $IP(x, m)$  as the proportion of times a point  $x$  falls significantly within an aggregation area for a specific month/season  $m$  (can be either a month or several months) among several years. Areas marked with high values of  $IP(x, m)$  are persistent aggregation areas throughout the time series.

## Confronting the results with the available literature

Persistent aggregation areas derived from our model during the reproduction period (as defined from the literature) were interpreted as potential recurrent reproduction grounds. We compare those inferences with the information of reproduction ground available from the literature (for sole and whiting).

Arbault, Camus, and Bec (1986) investigated the reproduction of sole along the Bay of Biscay based on several egg surveys occurring in 1982. Five surveys were conducted between January and May. Egg density was sampled in different locations from Hendaye

to Pointe du Raz ( $43^{\circ}30'N$ - $48^{\circ}N$ ) and allowed to map the distribution of egg production on the full study domain. The peak of reproduction occurred in February; thus we compare the maps obtained from the February survey with the persistence index obtained from our model in February.

For whiting, only two EVHOE trawl surveys occurred during spring (considered as the reproduction period of whiting) between 1987 and 1992 in the Bay of Biscay (Houise and Forest, 1993). For each haul, the individuals were counted and aged. Individual of two years and older were considered mature. We compare the distribution of mature individuals obtained with these surveys and the index of persistence from our model during spring (March to May).

No available information exists regarding the reproduction grounds of squids in this area, however the study from Moreno et al. (2002) investigated the reproduction period for *Loligo vulgaris* in the Eastern Atlantic and highlighted that their reproduction falls in winter and spring with a peak from January to April. We compute the persistence index for this period to identify the spatial aggregation patterns that emerge from the model outputs and that could be considered as spawning grounds.

To assess whether the aggregation patterns within the reproduction period are stable over the time period, we iteratively computed the persistence index over a 5-year mobile time-span while pushing forward one year each time.

## 4.3 Results

### 4.3.1 Assessing the contribution of each data sources to inference

Results highlight how combining several commercial fleets in the framework brings a better picture of the spatial distribution on the whole domain. For instance, when comparing model predictions with survey data (for the month of the survey), integrating several fleets into the analysis improves correlation with scientific data (Figure 4.3). It also reduces the standard deviation of the predictions on the full domain (SM C.7).

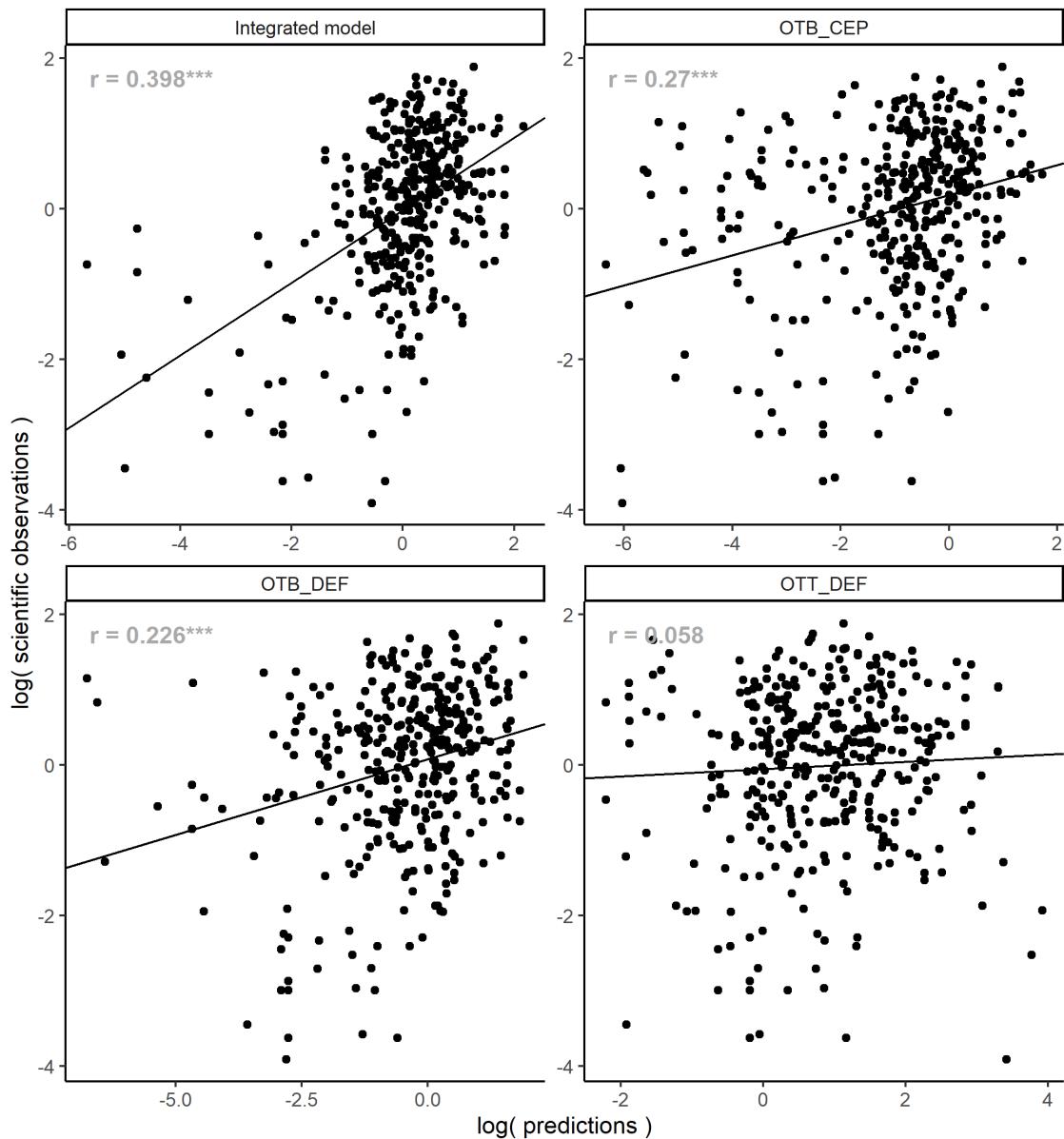


Figure 4.3 – Sole case study. Comparison between the observed scientific CPUE (y-axis) and the corresponding model predictions (x-axis) on the month of the survey, based on model integrating data from one commercial fleet only (either OTB\_CEP, OTB\_DEF, OTT\_DEF) or from all commercial fleets (Integrated model). x-axis: model predictions. y-axis: scientific data observations (CPUE in kg/hour). Black line: linear regression ' $\log(\text{scientific observations}) \sim \log(\text{model predictions})$ '. r: Spearman correlation coefficient. Scientific data are integrated to inference for all models. \*\*\* stands for the level of

significance.

When looking at the predictions within the spatial range of the fleets, single-fleet models logically provide similar spatial predictions compared with the integrated model (Figure 4.4; red dots). However, when using single fleet data, predictions realized outside the spatial range of the fleet largely depart from the ones realized through the integrated models (black dots, Figure 4.4), emphasizing the contribution of the other fleets to improve inferences on areas poorly covered by single fleet. This is particularly evidenced with the OTB\_CEP and OTT\_DEF fleets that partially cover the study area compared with OTB\_DEF that better cover the whole study area (Figure 4.1).

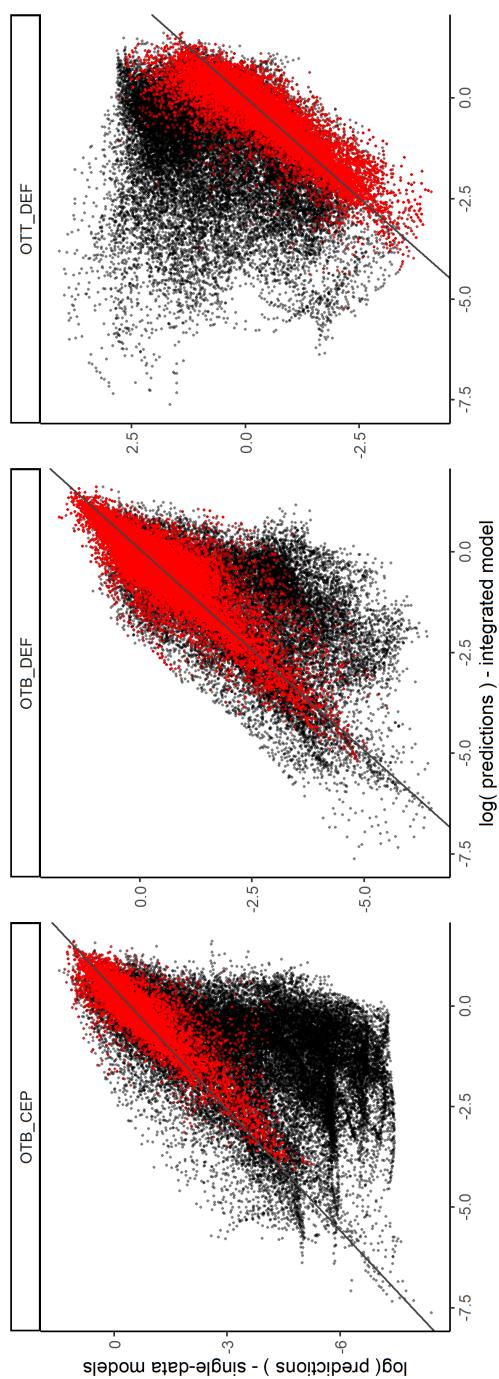


Figure 4.4 – Sole case study. Comparison between predictions from the integrated model (using all fishing fleets) and the model integrating only one commercial fleet for the 12 months of year 2018. Left: OTB\_CEP fleet, middle: OTB\_DEF fleet, right: OTT\_DEF fleet. x-axis: integrated model predictions. y-axis: single-fleet model predictions. The prediction values are log-scaled. Red points: predictions within the sampling area of the related fleets (i.e. the cells sampled by the fleet). Black points: predictions outside the sampling area of the related fleets. Black line:  $x = y$  axis. Note that the intercept of the x-y line has been scaled to account for differences in the intercept values between models. Scientific data are integrated to inference for all models.

### 4.3.2 Interpreting estimates of preferential sampling intensity

Estimates of the PS intensity ( $b_j(t)$  parameters in eq. 4.3) for the different species, the different fleets and the different time steps provide information on the targeting behavior that are consistent with expertise. Estimates of  $b_j(t)$  are positive for each species and each fleet (Figure 4.5, left column). For squids, PS is the strongest for OTB\_CEP followed by OTB\_DEF and OTT\_DEF. This is consistent with the expert knowledge of the targeting behavior of these fleets: OTB\_CEP target cephalopods and catch on average 15% of squids while OTB\_DEF and OTT\_DEF catch respectively 5% and 1% of squid). A similar pattern can be identified for whiting ( $b_{OTB\_CEP} > b_{OTB\_DEF} > b_{OTT\_DEF}$ ); this is consistent with species spatial distribution as whiting (like squids) are found in coastal areas (Figure 4.7) where the OTB\_CEP fleet is preferentially operating (Figure 4.1). For sole, the strength of PS is on average higher for OTB\_CEP and OTT\_DEF than for OTB\_DEF but with less contrast between the three commercial fleets which is also consistent with expertise as those three fleet target this high commercial value species.

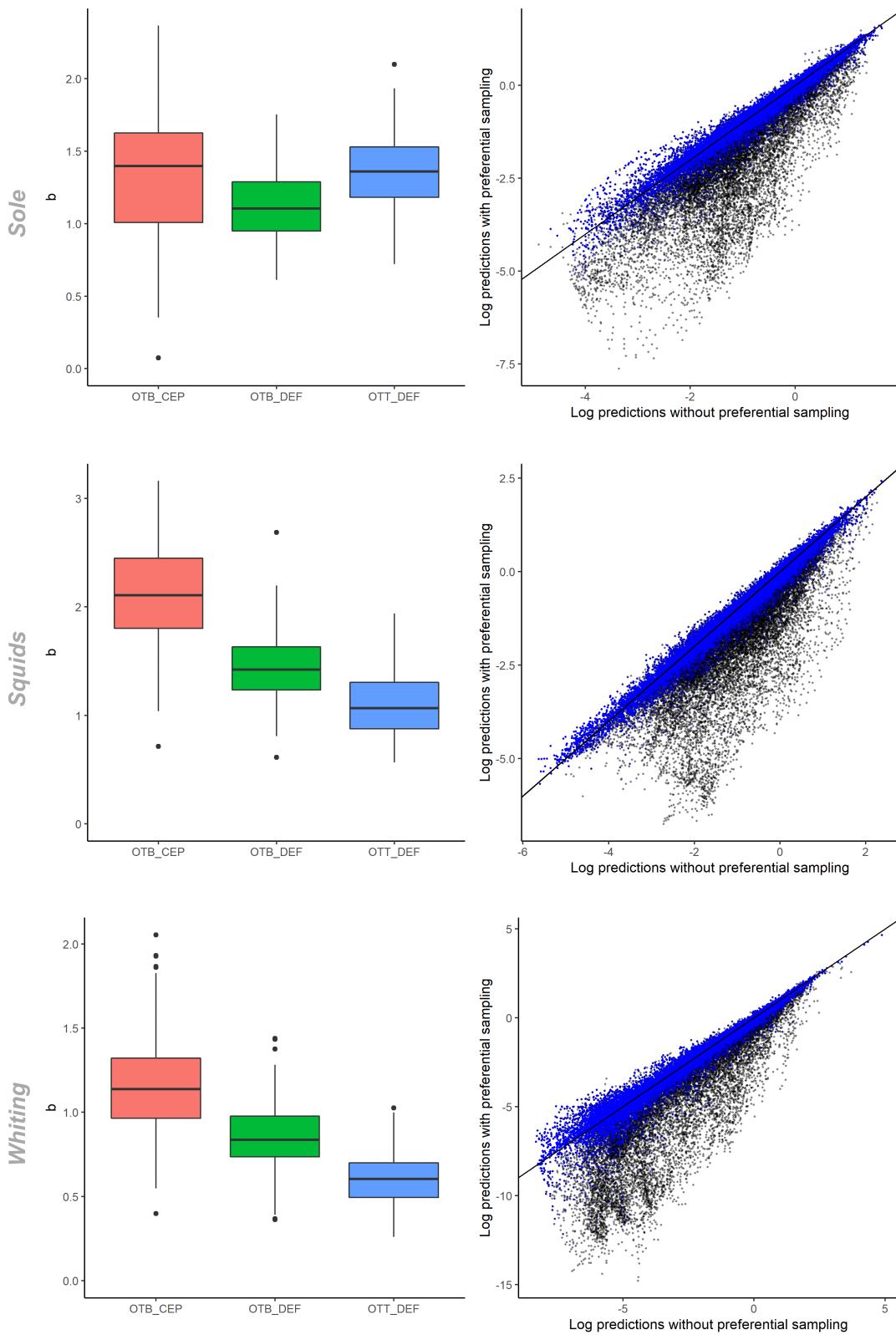


Figure 4.5 – Estimates of PS parameters for each commercial fleet (left) and effect of PS on model outputs (right). Left: boxplot represent the variability of maximum likelihood estimates of parameters  $b$  across the monthly time steps. Right: log-predictions of the integrated model accounting for PS (y-axis) versus log-predictions of the integrated model ignoring PS (x-axis) for the 12 months of year 2018. Blue points: predictions within the sampling area of the commercial fleets (i.e. the cells sampled by commercial fleets). Black point: predictions outside the sampling area of the commercial fleets. Black line:  $x = y$  axis.

Interestingly, some of the  $b$  parameters time series emphasize seasonal patterns (Figure 4.6, top). For instance in the sole case study for the OTB\_CEP fleet, the  $b$  parameters are higher in summer and autumn emphasizing relatively stronger PS, while being lower in winter and early spring (but see section 4.3.3 below for a more detailed interpretation of this seasonality pattern).

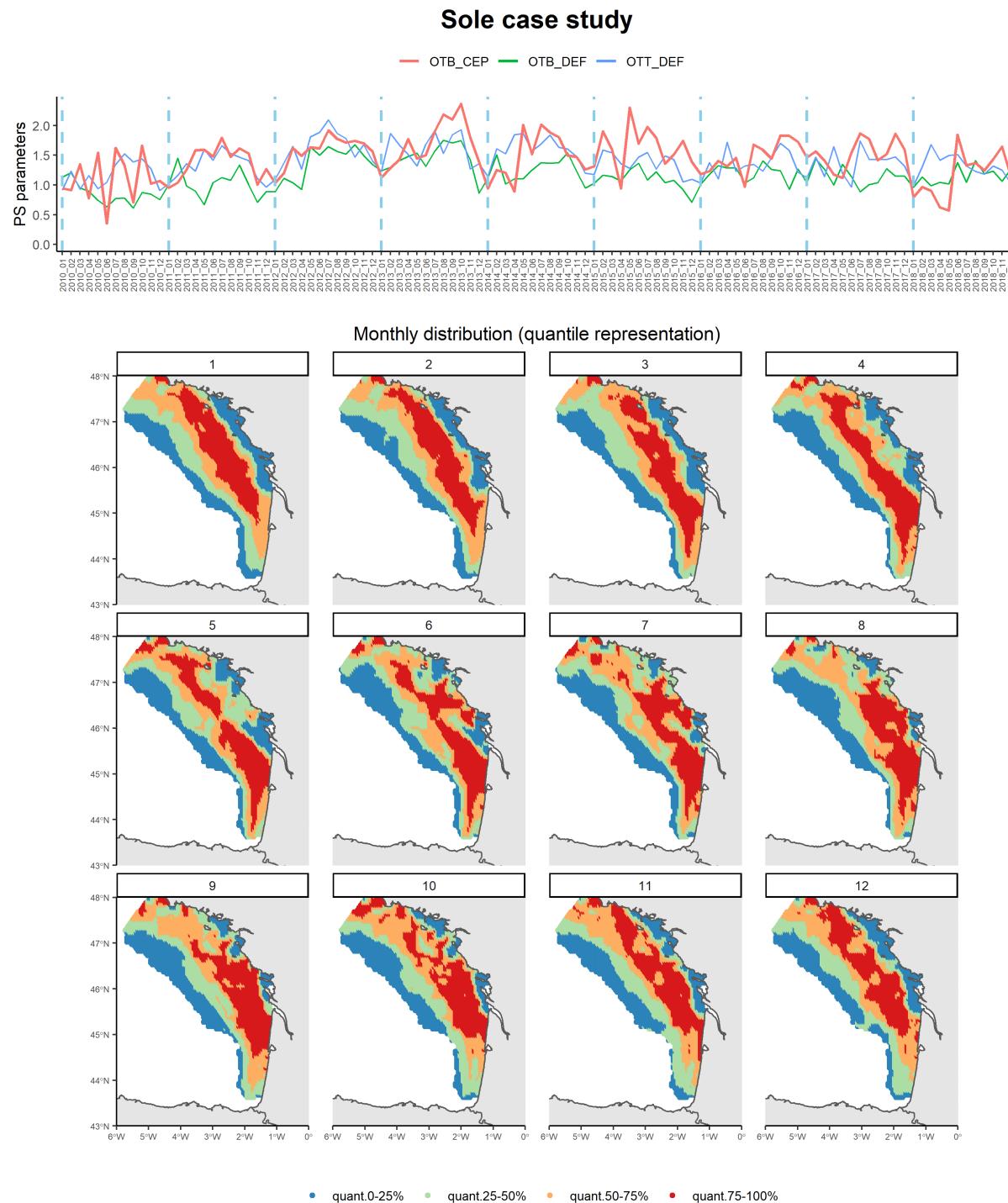


Figure 4.6 – Sole case study. (Top) Temporal evolution of the  $b$  parameters for the three commercial fleets fitted to the integrated model. Blue vertical lines: January. (Bottom)

tom) Monthly biomass distribution averaged on the full period. Only quantile values are represented. Model predictions come from the integrated model accounting for PS.

### 4.3.3 Evaluating the influence of preferential sampling on spatial distribution

Because estimates of  $b$  are positive, spatial density of fishing points is positively correlated with biomass density. Then logically, ignoring PS leads to a positive bias (i.e. overestimation) in biomass estimates in areas not sampled by the commercial fleets compared to estimates obtained while considering (Figure 4.5, right column, black points), but does not strongly affect predictions in locations within the range of the fleets (blue points). Considering PS only slightly improves the fit of the model to the data. For the Sole case study, some improvement of the likelihood occurred for both the likelihood associated with the commercial and the scientific data (Table 4.1). For whiting and squids, there are no strong modifications in both scientific and commercial likelihoods.

Table 4.1 – Ratio between the negative log-likelihood values (either commercial or scientific) from the integrated model accounting for preferential sampling and the integrated model ignoring preferential sampling.

Species	Negative log-likelihood ratio	
	Scientific data	Commercial data
Sole	0.97	0.92
Squids	1.00	1.01
Whiting	0.99	1.00

n.b. The ratio between negative log-likelihoods ( $-\log(lkl)$ ) is given as:  $r = \frac{-\log(lkl_{PS})}{-\log(lkl_{noPS})}$ . If  $r < 1$ , the model accounting for PS better fits the data than the model ignoring PS (noPS).

### 4.3.4 Investigating spatio-temporal dynamics of fish biomass

Results provide biomass density maps on a monthly time step that emphasize seasonal distribution patterns and from which aggregation index were calculated. The temporal correlation parameter ( $\varphi$ ) is estimated around 0.8 for all the species emphasizing strong

between months temporal correlations in the biomass field values. The range parameters are estimated to 55 km for sole and squids while being estimated to 67 km for whiting emphasizing wider spatial autocorrelation for this species.

Concerning the sole case study, model predictions highlight the relatively offshore distribution from November to April and a more coastal distribution from June to October suggesting some offshore-coastal migrations between these 2 periods (Figure 6, bottom). In particular, the migration in June/July is associated with a contraction of the sole distribution around the Vendée coast, the Gironde Estuary and the Landes coast (45.5°N-46°N) while the migration in November leads to an expansion of the species distribution towards the offshore areas all along the Bay of Biscay. Interestingly such seasonality coincides with the seasonality of PS intensity for the OTB\_CEP (Figure 6, Top). Higher PS parameter values are associated with a coastal distribution of sole while lower values corresponds to offshore distribution of sole.

Similar maps can be computed for the other species and are presented in SM C.9.

#### **4.3.5 Aggregation index and reproduction grounds**

For both sole and squids, areas of persistent aggregation areas during the spawning period exhibit strong patterns that match the available knowledge of reproduction grounds. For sole, the aggregation areas globally match with the observed area of maximum egg concentration (Figure 4.7), although the spawning grounds identified by egg maps are slightly further East of those identified by our method. This slight discrepancy could be interpreted as an effect of the larval drift as the maps provided by Arbault, Camus, and Bec (1986) are concentration of eggs and not reproduction grounds per se. This is consistent with the simulation analysis of Ramzi et al. (2001) showing that the eggs and larval drift in this area of the Bay of Biscay is oriented to the East. Overall, these aggregation areas are stable over time (Figure 4.8).

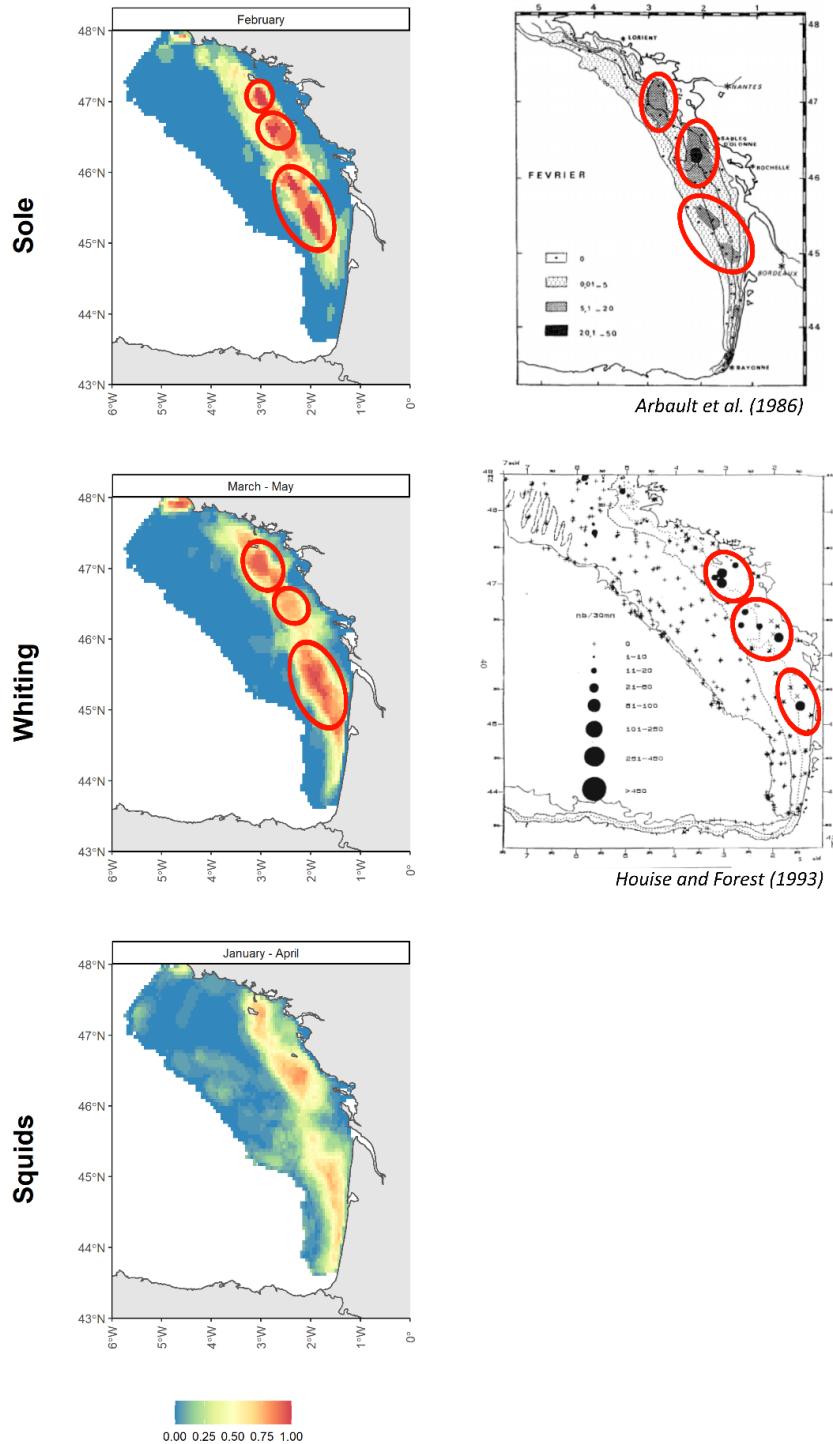


Figure 4.7 – Left: index of persistence during the reproduction period of sole (February), whiting (March-May) and squids (January-April). Reproduction period defined from ecological expertise. Index were computed from 2010 to 2018. Right: literature information on reproduction grounds when available. For sole, the map represents egg concentration from an egg and larvae survey conducted in 1982 (Arbault et al., 1986). For whiting, the map represents records of age-2+ whiting (i.e. mature individuals), from two spring trawl surveys that occurred between 1987 and 1992 (Houise and Forest, 1993). Model predictions come from the integrated model accounting for PS.

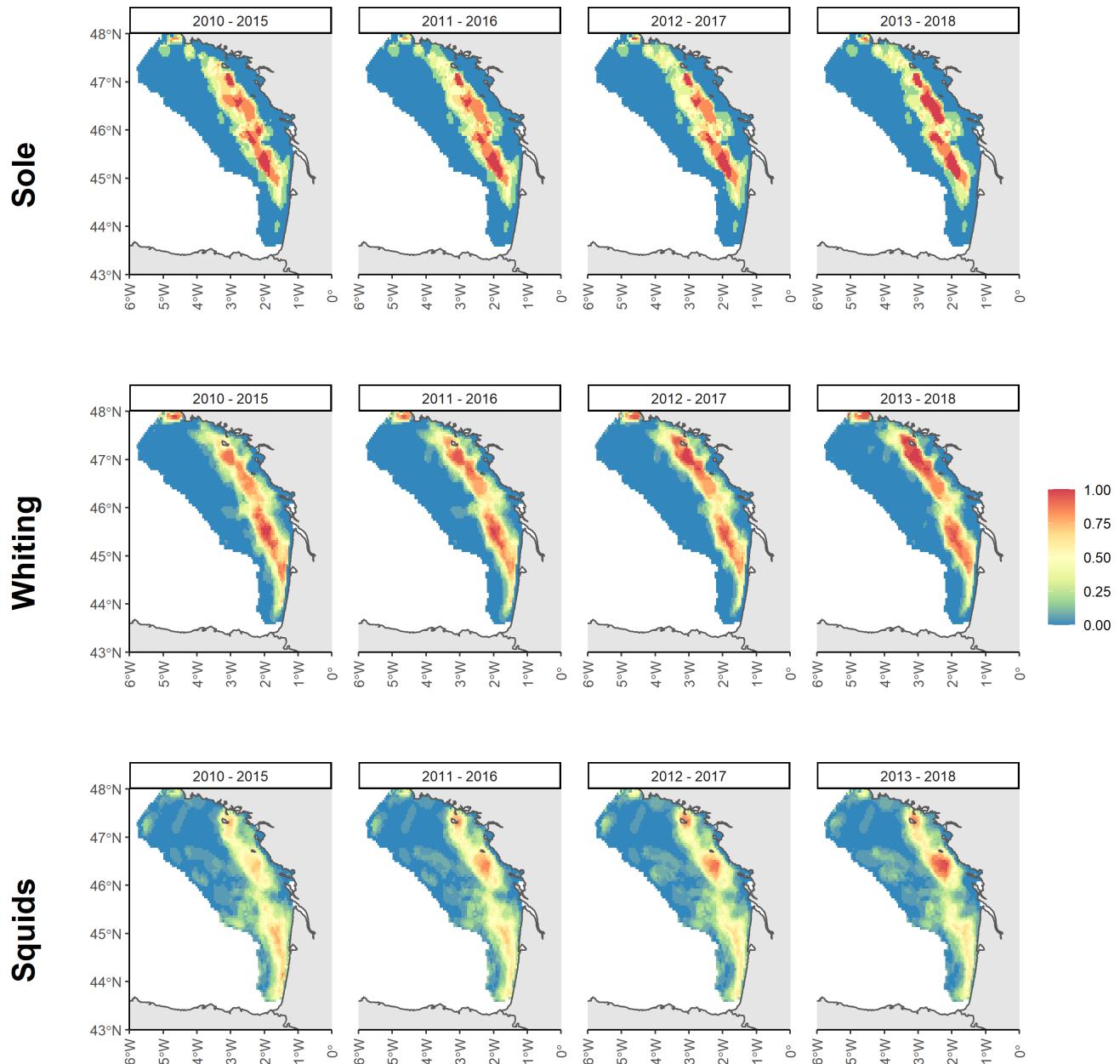


Figure 4.8 – Persistence indices within the reproduction period computed on a 5-year mobile time-span for each 3 species (5-year time span indicated on the top of each map). Model predictions come from the integrated model accounting for PS.

For whiting, similar patterns can be identified during the reproduction period (Figure 4.7); they match with previous studies investigating the spatial distribution of mature whittings (Houise and Forest, 1993). In particular, the Northern ( $3^{\circ}\text{W}$ - $47^{\circ}\text{N}$ ) and the Southern ( $2^{\circ}\text{W}$ - $45.5^{\circ}\text{N}$ ) aggregation patches are almost systematically significantly considered as aggregation areas (aggregation index equals 1) while the other middle one ( $2.5^{\circ}\text{W}$ - $46.5^{\circ}\text{N}$ ) is classified as an aggregation area that appears less frequently over the years. An additional persistent aggregation area can be identified in the North of the Bay of Biscay ( $4.5^{\circ}\text{W}$ - $48^{\circ}\text{N}$ ) suggesting that reproduction may also occur in this area which was not identified in the report of Houise and Forest (1993). Interestingly the Northern aggregation area ( $3^{\circ}\text{W}$ - $47^{\circ}\text{N}$ ) is more intense at the end of the period (Figure 4.8).

For squids, no information related to any reproduction ground exists in the literature, only the time period of the reproduction is known (the peak fall between January to April). On this time period, some persistent aggregation areas can be evidenced in coastal areas (Figure 4.7) along the Vendée coast ( $2.5^{\circ}\text{W}$ - $46.5^{\circ}\text{N}$ ), the Landes coasts ( $1.5^{\circ}\text{W}$ - $44^{\circ}\text{N}$  to  $45^{\circ}\text{N}$ ) and around Belle-Île-en-Mer ( $3^{\circ}\text{W}$ - $47.25^{\circ}\text{N}$ ). Interestingly the two Northern aggregation areas are more intense at the end of the time series compared to the beginning of the time series (Figure 4.8).

Maps of persistent aggregation areas are available for all month and evidence some other aggregation areas outside of the reproduction period (SM C.10). For instance, for sole a persistent patch can be identified offshore the Gironde Estuary ( $1.5^{\circ}\text{W}$  –  $45.5^{\circ}\text{N}$ ) from August to December.

## 4.4 Discussion

### 4.4.1 Main findings

In this paper, we develop a framework to infer fish spatio-temporal distribution on a monthly time step while combining scientific survey data and commercial catch declarations from several fleets. Commercial catch data constitute a valuable data source that complements scientific survey or onboard sampling programs by providing much higher spatio-temporal sampling density. Those complementary sources of data were integrated through a spatio-temporal hierarchical model taking into account spatio-temporal variation within the biomass field and PS on a monthly time step. We fitted the model to ‘VMS x logbook’ data filtered and processed over the period 2010-2018 for 3 demersal

species (sole, squids and whiting) in the Bay of Biscay.

We emphasize the benefit of integrating several spatially complementary fleets to infer fish distribution throughout the year. We demonstrate how the within year dynamics of the PS parameters can be interpreted with regards to the joint dynamics of species distribution and fishing distribution and to the overall targeting behavior of the fleets (e.g. OTB\_CEP for the squids case study). Even though PS parameters are not fishing intention per se (Bourdaud et al., 2019), these could advantageously complement information provided by landing profiles to estimate the targeting behavior of any group of vessels (either métier/fleet or any group that would seem appropriate).

Interestingly, although interpretation of the PS parameters provide insight into the spatio-temporal fleet dynamics, accounting for PS in the inferences does not significantly improve model fitting even when some fleets emphasize strong PS (e.g. squids, OTB\_CEP). These results contrasts with Alglave et al. (2022), and could result from the integration of several fleets in the analysis that allow a full coverage of the area. Indeed, in Alglave et al. (2022), the fleet emphasizing strong PS only covered a restricted (and coastal) part of the area. As introducing PS mainly affects inferences on poorly sampled areas, predictions in the offshore areas where mostly affected. Here, as the fleets are all estimated to have a positive PS and cover the whole area, PS only downscale the predictions in the few areas that remain unsampled.

Filtering the mature fraction of the population in both the scientific and the commercial data makes possible to infer the spatio-temporal distribution of the fraction of the biomass that could be potentially mature through the year on a monthly time step. We developed an index to infer aggregation areas of the potentially mature fraction of the biomass that are persistent across years. When calculated on a temporal window pre-defined following the available information on the reproduction period for each species, the aggregation index enables to identify the main recurrent spatial aggregation areas within the reproduction period. Results demonstrate that the recurrent aggregation areas identified from our method for Sole and Whiting were highly consistent with those already identified in the literature. Our results demonstrate how the aggregation index can provide new insights on the spawning grounds for species like squids for which no information on the spawning grounds is available on the literature. Areas of high aggregation persistent across years were identified during the expected period of reproduction and could be interpreted as spawning grounds. This opens perspectives for applying more systematically the approach for species where no information of reproduction grounds is

available to fill the gaps in our knowledge with minimum cost (Delage and Le Pape, 2016; Regimbart, Guitton, and Le Pape, 2018).

#### **4.4.2 Combining our results with other data sources to refine the identification of spawning grounds**

Persistent aggregation areas should be considered as potential spawning areas rather than actual spawning areas. Indeed, although the mature fraction of the biomass was filtered in the data, our maps do not directly inform whether individuals are actually reproducing or not. The outputs of the model provide maps of the mature fraction of the population (i.e. the individuals that can reproduce) and not the spawning fraction of the population (i.e. the individuals that are actually spawning). However, by focusing on temporal window identified as reproduction period in the literature, we limit the risk of misinterpreting the aggregation areas as reproduction areas.

Our results can also be used to help gathering additional data to identify reproduction grounds. Typically, our maps could be of great help to design surveys to record eggs, larvae and spawning individuals that would provide direct information of species reproduction (Fox et al., 2008). Because developing such additional surveys would be highly expensive, our maps could provide valuable a priori information to optimize the survey design and potentially find a compromise between the cost, the spatial extent, the temporal coverage of the survey and the accuracy of the expected estimates/predictions. Similar ideas were already applied to the sole case study to investigate more precisely the space-time variation of sole reproduction. Arbault, Camus, and Bec (1986) work provided a priori information of reproduction grounds that allowed to design more localized surveys to study inter- and intra-annual variability of one specific sole spawning area (Petitgas, 1997). Several statistical methods have been developed since and are suitable to optimize such adaptive sampling design; see for instance the recent work of Leach et al. (2021).

Our results could also be combined with fishermen expert knowledge (Yochum, Starr, and Wendt, 2011) to complement our knowledge of fish reproduction (Delage and Le Pape, 2016). For instance, Bezerra et al. (2021) and Silvano et al. (2006) proved the usefulness of fishers knowledge to determine the temporality of fish spawning and to identify some spawning grounds by crossing the information of aggregation areas provided by several fishermen. These were proved complementary with scientific data as they can be available at low cost and provide local knowledge of fish ecology.

#### **4.4.3 Limits and perspectives for the approach**

Our framework has several limitations that are all material for future research avenues.

First, the model remains relatively simple with regards to all the temporal processes that actually occur within a fishery. It is both a strength and a weakness: the model remains relatively generic, but one might want to extend it further to account for other temporal and spatio-temporal processes affecting fisheries dynamics. For instance, we opted for a non-seasonal representation of the model. One could make it seasonal by decomposing the intercepts  $\alpha_S(t)$  and  $\alpha_{X,j}(t)$  as well as the random effects  $\delta(x, t)$  and  $\eta(x, t)$  into yearly and seasonal terms in addition to some ‘season x year’ interaction terms as performed in Thorson et al. (2020). In their work, such specification mainly allowed to provide information over the time-steps where data was lacking. In the configuration of our case studies, data is available for all time steps and have a relatively good coverage of the study domain. Consequently, even though it provides a nice conceptual view of seasonality, complexifying our model in that direction should not deeply modify our inference of the biomass field. Alternatively, our framework could integrate orthogonal spatio-temporal terms in the latent field to capture the main mode of variability of the biomass field (Thorson, Ciannelli, and Litzow, 2020). Such orthogonal terms would allow to capture the main spatial patterns that structure the latent field as well as their variation in time. These could prove very useful to identify the structuring processes that affect species distribution and could give a valuable insight in the space-time dynamics of the species. Another exciting research avenue would consist in integrating population dynamics in the latent field of biomass (Cao et al., 2020). This would require to refine further the demographic resolution of the ‘VMS x logbook’ data (see for instance Azevedo and Silva (2020)), but once done, it would give access to huge data for inferring the space-time dynamics of fish populations. Finally, our model considers fishermen preferentially sample areas where the biomass is higher (preferential sampling), but does not consider any other drivers and specifically the temporal and spatio-temporal relations that can affect fishers behavior. These can be highly complex and may depend on the distribution of the resource, tradition/habits, management regulations (Abbott, Haynie, and Reimer, 2015; Girardin et al., 2017; Salas and Gaertner, 2004; Hintzen, 2021). These drivers are rarely studied in both space and time (although see Tidd et al. (2015)). Our framework could allow to jointly model the dynamics of the species, the distribution of the effort, the link that relates species distribution and effort in space/time and all the other spatial and/or temporal drivers that affect the distribution of fishing effort. For instance, we could relate

the fishing intensity to the biomass field from the previous time steps, or alternatively consider that the locational choice depends on the catches of the previous time steps. Adding such covariates and spatio-temporal dependencies in the sampling equation (eq. 4.3) will probably not modify the overall pattern of biomass distribution, but it would make possible to quantify the drivers of fishermen behavior and give valuable insight to the fishery dynamics.

Including discards would potentially improve our approach. Indeed, logbook data are landings declarations data which means they inform on the landings and not on the true catch. Thus, by assuming the landings per unit effort are proportional to the biomass, we make the hypothesis that the discard rate is constant in space and time and does not affect model predictions. This should not be a problem for sole and squids as the discards are low and TAC have not been really binding during the studied period. However, the issue might be more stringent for whiting and/or other species with a high and non-stationary level of discards. Integrating discards data in the analysis could help solving this issue. Stock et al. (2019) and Yan et al. (2022) used observer data to model bycatch in both space and time and Breivik, Storvik, and Nedreaas (2017) used bycatch data from onboard surveys to predict the temporal evolution of bycatch realized in the full commercial data. Similarly, we could integrate into the same analysis the logbook and the observer data by assuming that the catch of observer represents the sum of landings (which is also observed in the logbook data) and discards (which is unobserved in the logbook data). This way, the discards information available from observer data would be shared with the logbook data and would allow correcting for the missing portion of catch declarations data while possibly accounting for possible space or time variation in the discard rate.

Our analysis relies on the hypothesis that the spawning season is known a priori. Extending the approach to infer the spawning season based on the temporal dynamics of the aggregation patterns could improve our knowledge of species spatio-temporal distribution. In particular, identifying the main species phenological phases and their consistency (or shift) in time is crucial in the context of global change (Thorson et al., 2020). In our study, we computed the aggregation index on a predefined temporal window based on literature assumed to be the reproduction period (Arbault, Camus, and Bec, 1986; Houise and Forest, 1993; Moreno et al., 2002). Several methods exist and could be adapted to extract the spatial patterns that shape model outputs, their related temporal variation and identify the main phenological phases that characterize species distribution (e.g. reproduction, feeding - see for instance Empirical Orthogonal Functions or Prin-

cipal Oscillation Patterns - Cressie and Wikle (2015) and Wikle, Zammit-Mangion, and Cressie (2019)). While the approach we adopted in the manuscript requires to know the reproduction period of the species and would be inappropriate in a context of a changing reproduction time-span, those alternative methods would not require any a priori. Hence, these methods would be more appropriate to identify phenological modifications in species life cycle in response to climate change. Applying those kind of methods to the huge amount of data available from mandatory declarations, would make possible to track the effect of climate change on fish phenology at a monthly/seasonal scale, while it is generally only possible at a yearly time step through scientific survey data (Maureaud et al., 2020).

Last, confidentiality remains a major limitation to the massive use of VMS data (Hintzen, 2021). Indeed, there are strong confidentiality constraints on these data due to the huge information available on fishermen fishing grounds. Few countries are now giving free access to their data (e.g. Norway), but in most cases administrative procedures to get access to the data remain a burden and still constitute a limitation for the use of 'VMS x logbook' data for routine operational use.

#### **4.4.4 Future use for Marine Spatial Planning**

Our model has potential applications for Marine Spatial Planning (MSP). Janßen et al. (2018) highlighted that one of the main requirements for implementing MSP is the availability of fine-scale information on species distribution and of their essential habitats. Here we propose a method which can provide such information for the fraction of the population available through catch declarations (i.e. mainly the adult fraction and in some cases part of the juvenile fraction). This knowledge is required to design Marine Protected Areas (MPA – see for instance Lambert et al. (2017) or Loiselle et al. (2003)), Fishery Conservation Zones (Delage and Le Pape, 2016; Regimbert, Guitton, and Le Pape, 2018), or alternatively identify areas that should be kept for fishing in a context where many other human activities are competing in space and time with fishing (Campbell et al., 2014; Bastardie et al., 2015). This would require to integrate our results into bio-economic models in order to evaluate alternative management regulations and assess their tradeoffs in regards to all the sets of ecosystem services provided through activities such as fishing, aquaculture, energy, shipping, recreation and conservation (Nielsen et al., 2018).



# INFERRING FINE SCALE WILD SPECIES DISTRIBUTION FROM SPATIALLY AGGREGATED DATA

---

The two first chapters demonstrated how crossing catch declarations data with Vessel Monitoring System data provide an additional extensive data source to infer fish spatio-temporal distribution. However, the models were developed based on the very strong hypothesis of a proportional reallocation of the logbook data over VMS locations (fishing locations) along a fishing sequence. Such procedure is efficient and pragmatic. However, raw data are available at a rough spatial scale, and the uniform reallocation of catches over VMS locations may artificially introduce some spatial information that is far from the real (but unobserved) spatial repartition of catches over a fishing sequence. Even though it is a standard procedure, such a proportional reallocation could introduce deleterious bias in inference and could artificially reduce uncertainty estimation on parameters.

In the third chapter, we developed an alternative model integrating several data sources that do not have the same spatial resolution and enable to relax the hypothesis of uniform reallocation of catches over the VMS locations. Such issue is often referred as change of support (COS) or downscaling/upscaling (Cressie and Wikle, 2015). In this chapter, we introduce an approach that allows to tackle the COS issue for possibly complex data (i.e. highly zero-inflated positive continuous data) and to reconcile properly the spatial scales between catch declarations and scientific survey data. Even though the application is mainly oriented towards fisheries science, the overall ideas can find applications in any other domains where massive datasets have rough spatial resolution, but need to be used to infer fine-scale processes (e.g. health or climate science).

This chapter led to a submission in the Journal of the Royal Statistical Society: Series C.

## **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 handle properly 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 observations of ecological processes can be complex data (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 to infer fine scale species distribution in combination with scientific survey data that are exactly geolocalized. We compare (1) a rough but standard way to refine the resolution of declarations by proportionaly reallocating declarations on fishing locations and (2) a model that handle COS by matching the probability distribution of the aggregated observations with the unobserved geolocalized catch observations. 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 to 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. We argue that this approach is a valuable contribution for a wider use of spatially aggregated data in spatial ecology to make fine scale inferences of species distribution and to properly integrate datasets that do not have the same spatial resolution.

## 5.1 Introduction

### 5.1.1 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). A major objective of these studies aims at understanding the main underlying drivers of the ecological processes of interest. For instance, when investigating species distribution one will try to find the main covariates (e.g. substrate, temperature, depth/altitude) that shape species distribution (Guisan and Zimmermann, 2000; Planque et al., 2011).

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 et al., 2019; 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 the process under study. 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 prove highly valuable to map wildlife species distribution. A common example of such data sources are commercial catch declaration data in fisheries science. 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 et al., 2019). For instance, for fishermen declaration the administrative units are statistical rectangle with  $0.5^\circ \times 1^\circ$  resolution (Hintzen et al., 2012).

To refine the spatial resolution of these data, it is standard to apply rough hypothesis through geoprocessing techniques based on proportional allocation or centroid smoothing (Gotway and Young, 2007; Hintzen et al., 2012; Gotway and Young, 2007). Typically, to refine the spatial resolution of fishermen declarations, it is standard to proportionally reallocate the catch declaration on the related fishing geolocations available through Vessel Monitoring System (Hintzen et al., 2012; Gerritsen and Lordan, 2011). This enables to provide exactly located catch observations (called hereafter reallocated catches) and to combine these with high resolution data such as scientific survey catch (Alglave et al., 2022).

However, these methods do not explicitly represent the observation process. Overall, consequences of such procedures are hard to predict and can result in strong and deleterious biases (Gotway and Young, 2007; Pacifici et al., 2019). These biases do not only affect the mean and the variance, but also any statistics that would be derived from these estimates. Typically, in the example presented above, proportional reallocation is an ad-hoc procedure that could strongly homogenize the catch. This could lead to a loss of information in inference and to uncertainty under-estimation by assuming the reallocated catches (with exact geolocation) are the actual records while the actual ones are the aggregated declarations (coarse resolution).

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.

### **5.1.2 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 (Lajaunie and Wackernagel, 2000; Gotway and Young, 2002; Gelfand, 2010). It is often also referred as ‘downscaling/upscaling’ or Modifiable areal unit problem (MAUP) in the literature (Cressie and Wikle, 2015). 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 et al., 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.

### **5.1.3 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 et al., 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  $0.5^\circ \times 1^\circ$ ) 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., 2012). However, the consequence of this procedure on inference has never been explored. In particular, one can suspect that this could lead to strong homogenization of the catch, to deleterious bias on model parameters and consequently to a loss of information for inference. 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 5.2).

Then, we assess the method through 2 sets of simulations (Section 5.3):

- A first simulation set at the scale of a single statistical rectangle. Our aim is to explore the base properties of our framework: are the estimates unbiased? How imprecise are estimates when data is roughly reallocated? What is the gain of our alternative approach? For these simulations, the model describing fish distribution is simply considered to arise from a known covariate.
- A second simulation set parameterized so as to get closer to a realistic case study. 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) and illustrate the main results that are consistent with simulations (Section 5.4).

## 5.2 A spatialized catch model for aggregated data

In Alglave et al. (2022), the authors 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.

Let  $D \subset \mathbb{R}^2$  be a spatial domain and  $(S) = (S(x), x \in D)$  a spatial random process 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 5.1) where  $(\delta) = (\delta(x), x \in D)$  is a zero mean GRF whose variance-covariance is specified through a Matérn covariance function and  $(\Gamma) = (\Gamma(x), x \in D)$  a field of covariate. As the catch are zero-inflated positive continuous data, following Thorson (2018), the authors model a catch at a given site  $x_i$  with a mixture of a Dirac mass at 0 and a Log Normal distribution (Equation 5.1), the proportion of the mixture being defined conditionally on the random field  $(S)$ .

$$Y_i | S(x_i), x_i \stackrel{\text{ind}}{\sim} \mathcal{M}_Y(p_i, \mu_i, \sigma^2), \quad (5.1)$$

with  $p_i := \exp(-e^\xi S(x_i))$  the proportion of the mixture,  $e^\xi$  a parameter controlling zero-inflation,  $\mu_i := \frac{S(x_i)}{1-p_i}$  the expected catch when positive (on the natural-scale) and  $\sigma^2$  a transformation of its variance (on the log-scale). This parametrization of the mixture model corresponds to equations 5.2 (see the SM D for more details):

$$\begin{aligned}\mathbb{P}(Y_i = 0 | S(x_i), x_i) &= p_i = \exp(-e^\xi S(x_i)) \\ \mathbb{E}(Y_i | Y_i > 0, S(x_i), x_i) &= \mu_i = \frac{S(x_i)}{1 - p_i}, \\ \text{Var}(Y_i | Y_i > 0, S(x_i), x_i) &= \mu_i^2(e^{\sigma^2} - 1),\end{aligned}\quad (5.2)$$

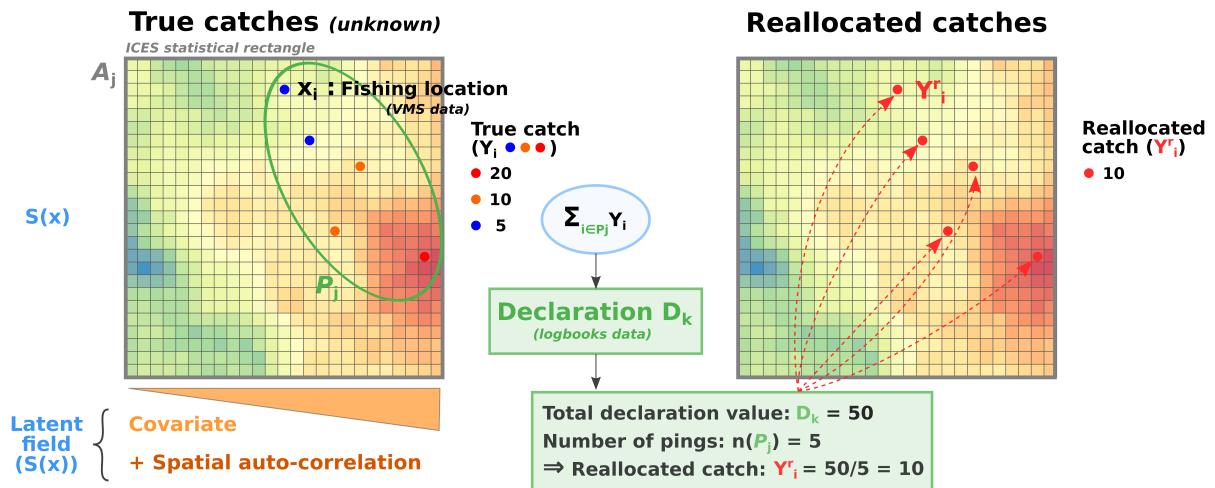


Figure 5.1 – Schematic representation of the reallocation process. The biomass field (the background field) depends on a covariate and a spatial random effect. The covariate is the  $x$  axis. It has a positive effect on biomass values (i.e. biomass is higher on the right of the grid than on the left). 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 catches 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  $Y_i^r$ .

In this approach all fishing locations  $x_i$  and the corresponding catch are supposed to be known. This is classically the case with scientific survey for which every catch is

recorded as well as its geolocation.

However, fishermen do not declare this precise information in logbook, 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 catches  $Y_i$  realized in the administrative unit  $\mathcal{A}_D$  associated with the catch declaration  $D$  (Equation 5.3), more formally:

$$D = \sum_{i|x_i \in \mathcal{A}_D} Y_i \quad (5.3)$$

Since the original model proposed in Alglave et al. (2022) is defined for exactly geolocated catches, those catches are derived from the logbook data combined with the VMS data. This is classically done in fisheries science by using a uniform reallocation process (Hintzen et al., 2012; 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 catch  $Y_i^r := D_k/m_k$ . As noted by Alglave et al. (2022), this process has several drawbacks. First, as a consequence of the reallocation process, the reconstructed individual catches tend to exhibit smoother patterns than the original catches. Second, the actual amount of data is the total number of catch declarations while the number of data 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 catch declarations  $D_k$  instead of the reconstructed individual catch  $Y_i^r$ . As we aim to propose a model compatible with the original one, we are led to specify the distribution of  $D_k$  which consists of a sum of  $m_k$  random variables following a mixture distribution. This distribution has no known analytical form. However, the catch declaration data also exhibit some zero-inflation (a null declaration means that for this declaration and this species the catch is null, but catch is positive for other species of the declaration) 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 catch declaration and its variance (Equation

5.4):

$$D_k | \mathbf{S}_{\mathcal{P}_k}, \mathcal{P}_k \sim \mathcal{M}_D(p_k^D, \mu_k^D, \sigma_k^{2,D}) \quad (5.4)$$

with  $\mathcal{P}_k = (1, \dots, i, \dots, m_k)$  the index of the individual catches belonging to the  $k^{th}$  declaration  $D_k = \sum_{i \in \mathcal{P}_k} Y_i$ ,  $(x_1, \dots, x_i, \dots, x_{m_k})$  is the list of all the fishing positions of the  $k^{th}$  declaration,  $\mathbf{S}_{\mathcal{P}_k}$  are the latent field values at fishing positions  $\mathcal{P}_k$ ,  $\mu_k^D$  the expected positive biomass,  $p_k^D$  the proportion of the mixture and  $\sigma_k^{2,D}$  the variance parameter.

In order to relate the individual observation level  $Y$  and the catch 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  $Y_1, \dots, Y_{m_k}$  are independent conditionally on  $\mathbf{S}_{\mathcal{P}_k}$ , the probability of a zero-declaration  $\mathbb{P}(D_k = 0)$  is obtained by simply multiplying the probability to obtain a zero-punctual observation  $\mathbb{P}(Y_i = 0)$  to all fishing points  $i \in \mathcal{P}_k$  (Equation 5.5).

$$\mathbb{P}(D_k = 0) = \prod_{i \in \mathcal{P}_k} \mathbb{P}(Y_i = 0) \quad (5.5)$$

$$= \exp \left\{ - \sum_{i \in \mathcal{P}_k} e^{\xi} \cdot S(x_i) \right\} = p_k^D$$

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 5.6 and SM D). It is straightforward to prove that

$$\begin{aligned} \mathbb{E}(D_k | D_k > 0) &= \frac{\sum_{i \in \mathcal{P}_k} \mathbb{E}(Y_i)}{1 - p_k^D} = \frac{\sum_{i \in \mathcal{P}_k} S(x_i)}{1 - p_k^D} \\ \mathbb{V}ar(D_k | D_k > 0) &= \frac{\sum_{i \in \mathcal{P}_k} \mathbb{V}ar(Y_i)}{1 - p_k^D} - \frac{\pi_k}{(1 - p_k^D)^2} \mathbb{E}(D_k)^2 \end{aligned} \quad (5.6)$$

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

$$\text{and } p_i = \mathbb{P}(Y_i = 0)$$

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  $\text{Var}(D_k|D_k > 0)$ . We propose to use the same family of distribution for  $D_k|D_k > 0$  as the one used for the individual catch  $Y_i|Y_i > 0$  i.e. a Lognormal distribution. This is an approximation that we discuss later.

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 catch data  $Y^r$ .

Finally, note that the scientific data is considered known at each fishing location and thus they can be simply modeled as standard individual data  $Y_i$ .

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).

We have finally three alternatives to estimate the spatial field of biomass from catch data:

- a baseline ideal configuration (or gold standard) when the locations of commercial catch (and eventually the scientific catch) are precisely geolocalized. This ideal situation, with no actual application, will be named **Spatial Model** and it is used as a reference for the comparison between the two alternatives.
- the original model fitted with commercial reallocated individual catch (and potentially few precisely geolocalized scientific data) as done in Alglave et al. (2022). This approach will be referred as **Reallocated 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. This approach is named **Declaration Model**.

### 5.3 Simulation studies

To assess the drawbacks and the advantages of the different approaches, we conduct two different simulation studies. First, we explore the base properties of the different models by a study at a single statistical rectangle, based on commercial data and with a very simple spatial latent field which only depends on one covariate (with no spatial random effect). These simulations will be referred as **single-square simulations**.

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

In these two sets of simulation studies, there is a unique covariate, modeled as a continuous GRF that we suppose known at each point of the grid (when present the spatial random component is also a GRF but its values are not considered to be known).

The covariate effect is fixed to  $\beta_S = 2$  and the intercept is also fixed to  $\mu = 2$ . Regarding commercial data, the number of fishing pings per declaration is fixed to 10 as it is the average number of fishing locations for a single declaration in real data. All parameterizations are detailed in the Table 5.1.

Table 5.1 – Parameter values for the simulations

Parameters	Single-square simulations	Multiple-square simulations
$\mu$	2	2
$\beta_S$	2	2
Range of $\delta$	–	0.6 ( $\approx 50$ km)
Marginal variance of $\delta$	–	1
Range of $\Gamma$	10 (cells)	1.5 ( $\approx 120$ km)
Marginal variance of $\Gamma$	0.5	0.5
$\xi_{com}$	-1	-1
$\sigma_{com}$	1	1
$k_{com}$	–	1
$\xi_{sci}$	–	0
$\sigma_{sci}$	–	0.8

N.b.  $k_{com}$  is the relative difference in catchability between scientific data and commercial data (i.e. it

is a scaling parameter). This parameter captures the difference in catch efficiency between scientific and commercial gears.

The locations of the individual commercial catch 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 of side 7 (approximately the distance of a trawl haul). At each fishing position, a catch 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 2 metrics:

- the MSPE (Equation 5.7) which quantifies the accuracy of the spatial predictions of the latent field over the spatial domain ( $n$  is the number of locations over the grid).

$$MSPE = \frac{\sum_i^n (S(x_i) - \hat{S}(x_i))^2}{n} \quad (5.7)$$

- the estimates of the parameter  $\beta_S$  which quantifies the species-habitat relationship.

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

### 5.3.1 Single-square simulations

Two important variables may affect the accuracy of model outputs: the amount of commercial data and the number of fishing zones explored and aggregated within a catch 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 catches

are spatially close. This situation corresponds to a declaration associated with only one fishing zone. 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 5.2).

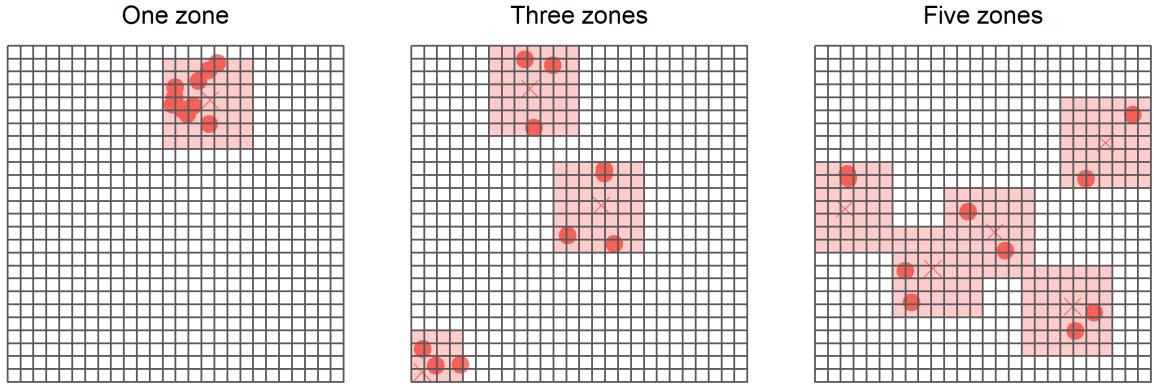


Figure 5.2 – Simulations of 10 fishing points within 1, 3 and 5 fishing zones. The full grid corresponds to a statistical rectangle. Cross are the centroid of the fishing zones. A declaration declared at the level of the statistical rectangle would be uniformly reallocated over these fishing points.

For the single square simulation, in addition to the metrics introduced before ( $MSPE$  and the species-habitat parameter  $\hat{\beta}_S$ ), we also assess 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  $\hat{\xi}$ .

The results are presented in Figures 5.3 and 5.4.

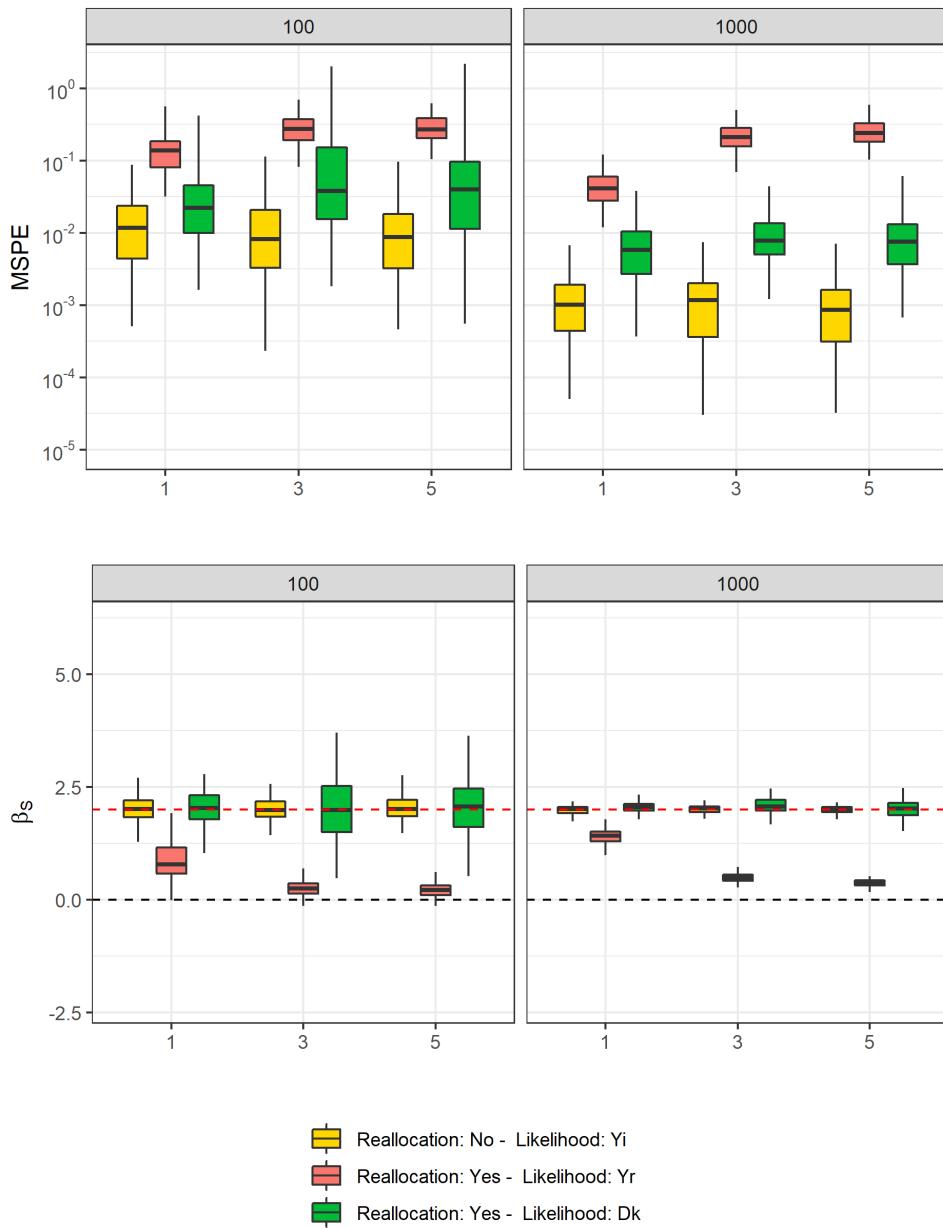


Figure 5.3 – Performance metrics for single-square simulations with a total of 100 or 1000 fishing positions in columns.  $MSPE = \frac{\sum_i^n (S(x_i) - \hat{S}(x_i))^2}{n}$  is the mean squared prediction error and  $\hat{\beta}_S$  is the species-habitat relationship parameter. The number of fishing 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 5.2.

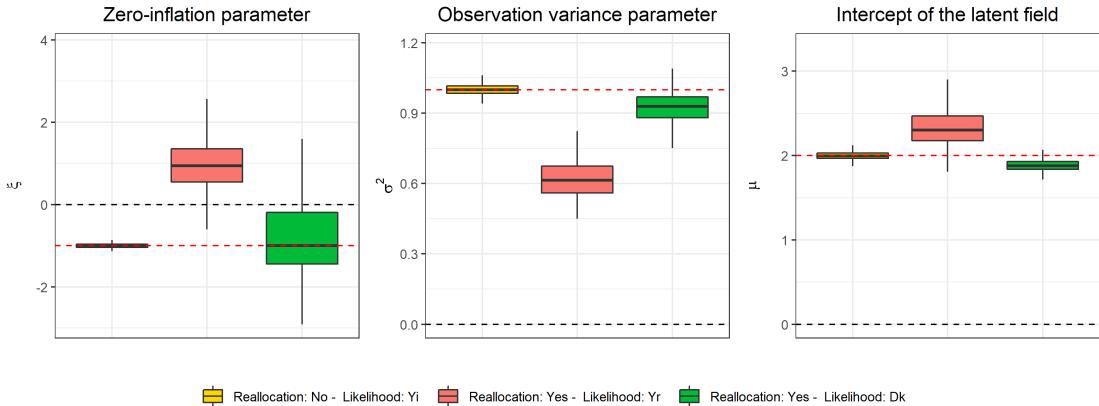


Figure 5.4 – Parameters relative bias for single-square simulations. **Reallocation:** data are or are not reallocated in simulations. **Likelihood:** the likelihood is computed on exact individual observations  $Y_i$ , on reallocated observations  $Y_i^r$  or on catch declarations  $D_k$ . Gold: gold standard. Red: uniform reallocation ( $Y_i^r$  model). Green: model-based reallocation ( $D_k$  model). Only the simulations with 1000 fishing positions are represented. Black line: zero value. Red line: parameter true value.

Figure 5.3 highlights that the reallocation process has a major effect on predictions and estimates accuracy. As expected, the reallocation process conducts to a 10 to 200 times decrease in accuracy for spatial predictions when fitting the Reallocated Model (MSPE gold compared to red boxplots). 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. Figure 5.4 shows the over-estimation of the zero-inflation parameter ( $\xi$ ) when using the Reallocated Model. When  $\xi$  increases, the amount of zero in the data decreases. Then, an overestimation of the  $\xi$  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 catches  $Y_i$  associated with the same declaration  $D_k$  is non-zero, uniform reallocation will lead to a positive catch for each reallocated individual catch  $Y^r$ . Consequently, this will tend to decrease the proportion of zero and will lead to the over-estimation of the  $\xi$  parameter.

The observation variance ( $\sigma$ ) is underestimated i.e. the data are estimated to be less noisy than they actually are. The intercept of the latent field ( $\mu$ ) is slightly over-estimated (Figure 5.4).

The Declaration Model allows to recover the species-habitat relationship and to improve the accuracy of the spatial predictions (Figure 5.3) 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 catch declarations. Other parameters (observation variance, intercept) are also better estimated than with the Reallocated Model even though they remain slightly biased (Figure 5.4). This alternative model has some convergence issues (Table 5.2) 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).

Table 5.2 – 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

### 5.3.2 Multiple-square simulations

As mentioned before, we propose another set of simulations designed to be closer from the case study. 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. The covariate is simulated with wider autocorrelation than the random effect (Table 5.1). 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 5.5A). To tailor the case study, we simulate 3000 of fishing positions grouped in 300 declarations (10 individual catches

per declaration). Commercial data may not cover the full area, and consequently we allow the commercial samples to cover only 2/3 of the area. 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 the 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 5.5A). Scientific observations are simulated following the observation equation of  $\mathcal{M}_Y$  (with specific parameters for scientific data - Table 5.1).

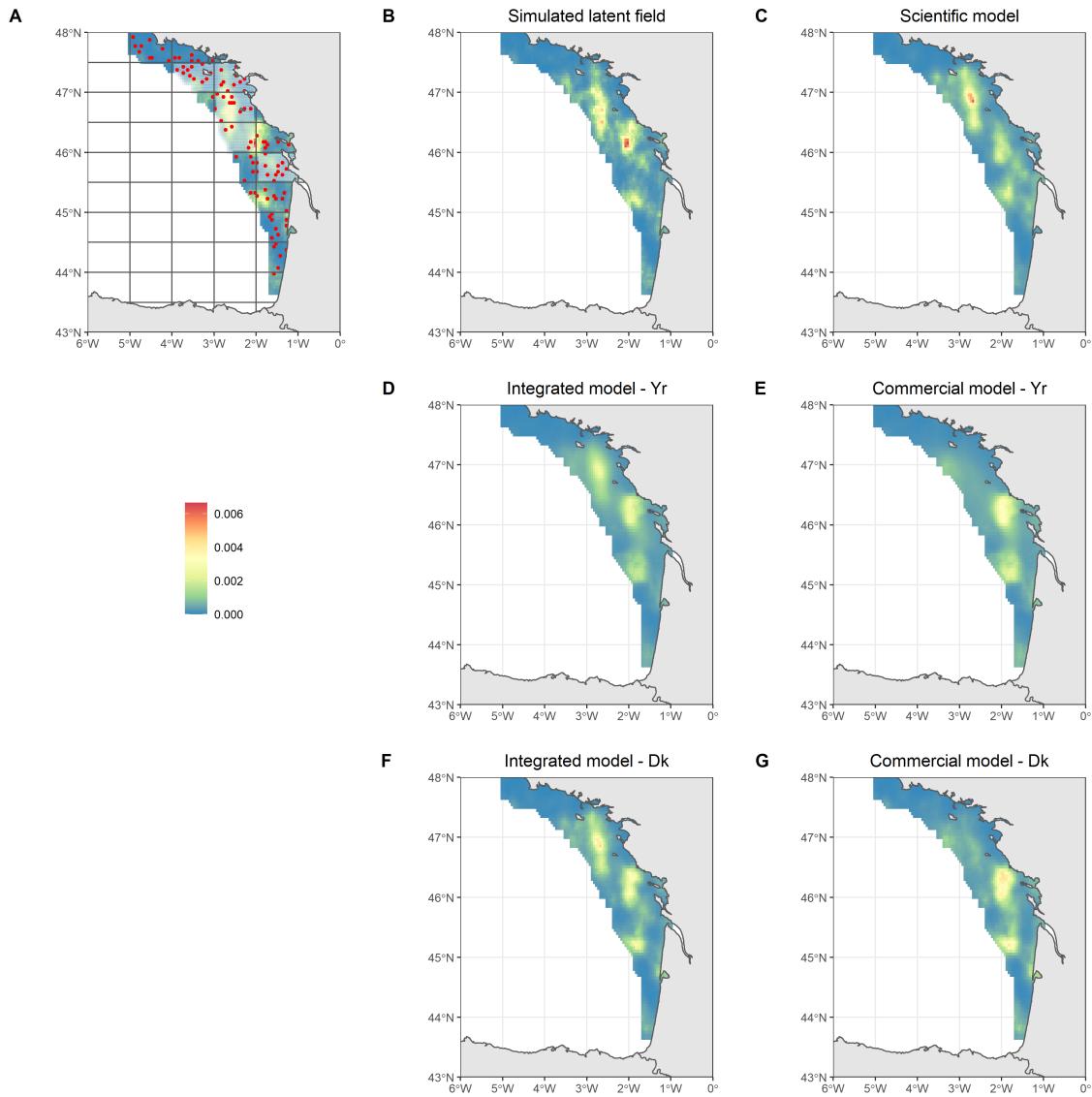


Figure 5.5 – Relative distribution of simulated/estimated biomass field. A: Simulated biomass field with scientific samples (red) and statistical rectangles. The rectangles that have not been sampled by commercial data are the transparent rectangles. They represent 1/3 of the full area. B: simulated biomass field. C: biomass field from the scientific-based model.  $Y_i^r$ : 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.

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).

In addition to the 2 metrics introduced at the beginning of the section (*MSPE* and species-habitat parameter  $\beta_S$ ), 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 catch level are always smaller than their single-data counterparts. This can be well illustrated from Figure 5.5. Integrating scientific and commercial data allows (1) to capture the hotspot missed by commercial data through scientific data and (2) to 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 5.6) compared to the Declaration Model.

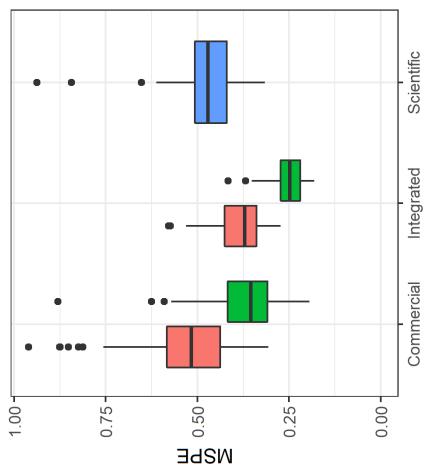
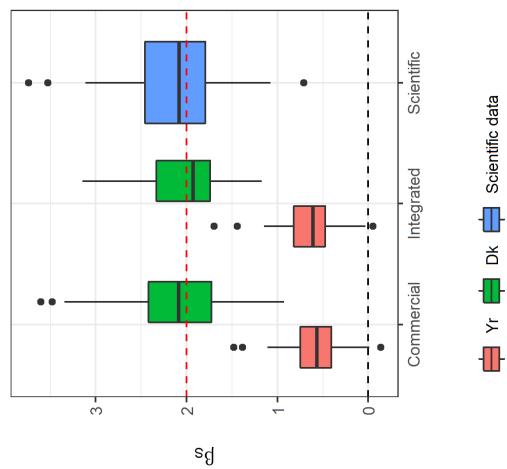
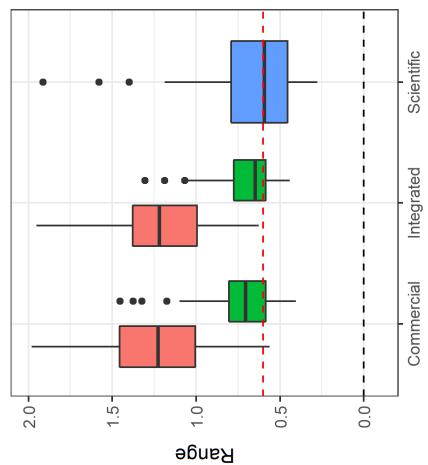


Figure 5.6 – Performance metric for the multiple-square simulations. Red line: true value for the range and the species-habitat parameter ( $\beta_S$ ). Red: uniform reallocation ( $Y_i^r$  model). Green: model-based reallocation ( $D_k$  model). Blue: scientific-based model.

Interestingly, in addition to the species-habitat relationship, uniform reallocation also affects the spatial autocorrelation terms such as 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. However, this goes with some difficulty in convergence as only 75% of the model built on catch declarations converge (Table 5.3).

Table 5.3 – 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

## 5.4 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 papers (Alglave et al., 2022) and is not developed further here. Scientific data were extracted from the DATRAS database for the Orhago beam trawl survey (Coupeau and Biais, 2019; ICES, 2018a). 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, (2) the integrated Reallocated model fitted to both

scientific data and reallocated individual catch data and (3) the integrated Declaration Model fitted to both scientific and declaration data.

The integrated Declaration Model faced convergence issues (some of the parameters were hardly estimated e.g. the range parameter). To ease convergence, we integrated in the analysis onboard observer data 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  $\mu$ , covariate effect  $\beta_S$ , range and marginal variance) and finally let them free in the following phases of estimation.

Consistently with simulations, the Declaration Model emphasize differences with the Reallocated Model in both parameters estimates and spatial pattern of the species distribution (Figures 5.7, 5.8). In particular, the substrate effect is recovered in the Declaration Model and fall in the same range as the scientific-based estimate (Figures 5.7). The zero-inflation parameter  $\xi$  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).

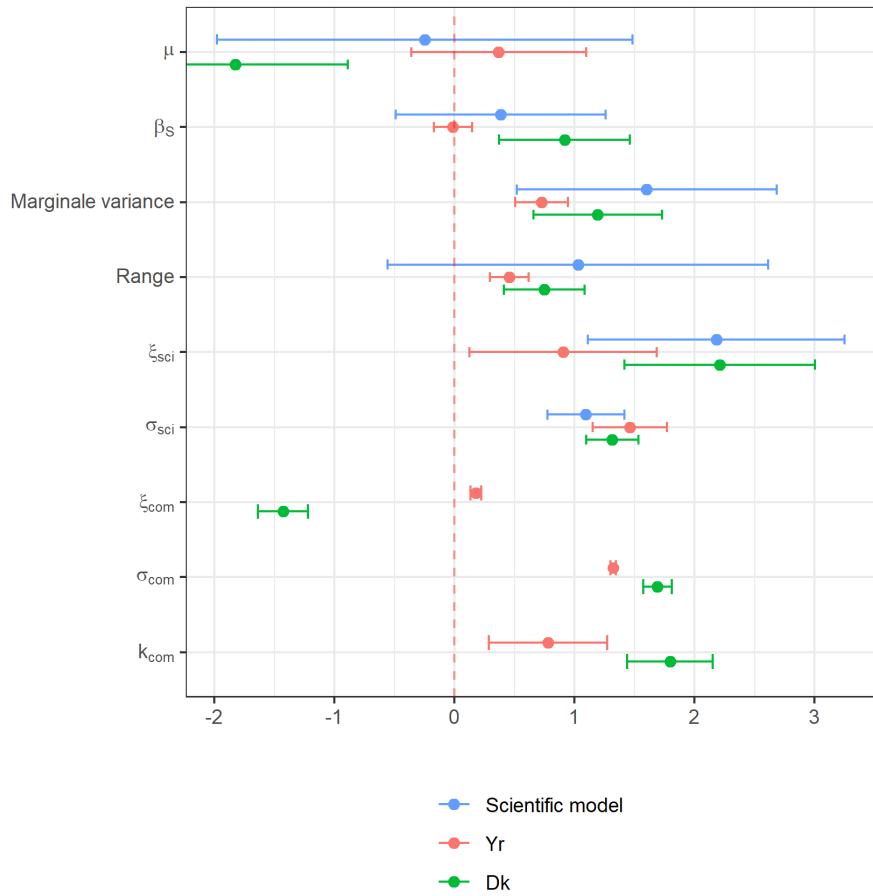


Figure 5.7 – Parameters obtained with the scientific-based model, the integrated model fitted on reallocated catch  $Y_i^r$  and the integrated model fitted on catch declarations  $D_k$ .  $k_{com}$  is the relative difference in catchability between scientific data and commercial data (i.e. it is a scaling parameter). This parameter captures the difference in catch efficiency between scientific and commercial gears.

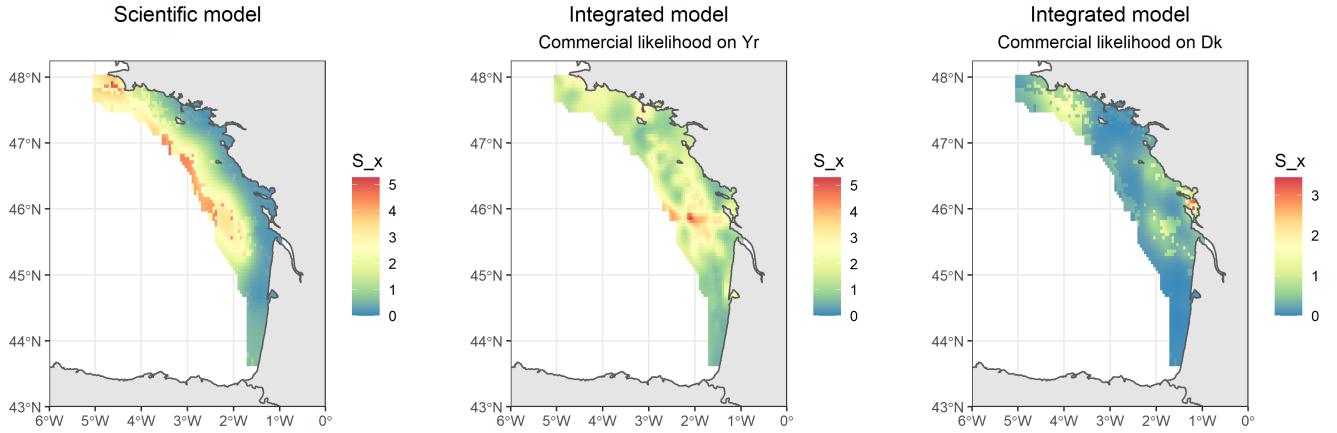


Figure 5.8 – Maps obtained from the scientific-based model (left), the integrated model fitted on reallocated catch  $Y_i^r$  (center), the integrated model fitted on catch declarations  $D_k$  (right).

In addition, uncertainty is also revised when fitting the model at the declaration level. For instance, when comparing the Reallocated Model to the Declaration Model, the confidence intervals of  $\beta_S$ , the marginal variance, the range,  $\xi_{com}$ ,  $\sigma_{com}$  are much wider. This emphasizes that uncertainty is probably underestimated in the Reallocated Model compared with the Declaration Model. When comparing the scientific-based model and the integrated Declaration Model, some parameters are more precisely estimated than when only scientific data feed the model. For instance, while in the scientific-based model the substrate effect was not significant, in the integrated model built on the declarations, substrate is significant and the confidence interval is smaller.

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  $\mu$  is revised upwards when building the likelihood on the individual precisely geolocalized catch and revised downwards when estimated with the Reallocated Model. This is consistent with the simulations results, see Figure 5.4.

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.

## 5.5 Discussion

### 5.5.1 The benefit of a statistical approach for change of support

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 (Cressie and Wikle, 2015; Wakefield and Lyons, 2010). Such methods are key to integrate data that have differing spatial resolution and to make fine inference on spatial processes and specifically on species distribution (Pacifici et al., 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 catches on fishing locations biases the parameter estimation and tends 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 nicely 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 allow to capture the main patterns of species distribution through the spatial random effect, reallocation leads to the loss of the species-habitat relationship (parameters estimates are close to 0). Furthermore, uncertainty intervals are very narrow emphasizing 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, a method that accounts for COS is key to estimate properly the species-habitat relationship from catch 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.

### **5.5.2 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 catch declarations data) arise from a shared latent process (species distribution) and that, while scientific data are recorded at their exact locations, commercial declarations are recorded at a rough scale and are a convolution of exact location catches. 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 case of 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 catch level. Considering that a convolution of zero-inflated lognormal distribution follows a zero-inflated lognormal is an approximation that can be questioned. We showed that this approximation is reasonably good in our context therefore we kept this approach as we built up on previous work (Alglave et al., 2022). However, exploring observation models that verify additive property as the Gamma distribution would be an interesting perspective for the future.

Finally, another approach that is common in the COS literature is ‘Block kriging’ (Gelfand, Zhu, and Carlin, 2001; Gelfand, 2010; Pacifici 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 catch declarations. It

takes over 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 occurring 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 kriging would imply to lose sparsity by integrating over block areas  $B$ .

### **5.5.3 Future perspectives for the framework**

More and more declarative data are now becoming available in the field of ecology, epidemiology and environmental 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 to 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 et al., 2021) as well as species misspecification (Botella et al., 2018) can arise. 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 developed to correct for the several potential deleterious bias that can arise in non-standardized data (Dobson et al., 2020).



# DISCUSSION

---

The reference data to map fish distribution and identify essential fish habitats are mainly standardized **scientific survey data and onboard observer data** (Pennino et al., 2016; Rufener et al., 2021). Both provide direct observation of the catch with the exact location of the records. However, scientific surveys only occur once or twice a year and they can mismatch key species life cycle stages (Hilborn and Walters, 1992). Additionally, the sampling intensity of both scientific surveys and onboard observer data is in general relatively low due to technical constraints and consequently they may provide inaccurate estimates of species distribution (Alglave et al., 2022; ICES, 2005). Furthermore, onboard observer data are preferentially sampled in areas of higher biomass which requires to account for PS in modeling (Pennino et al., 2019).

By contrast, **commercial declarations** are intensive datasets regrouping the daily catch declarations from the fishermen at the scale of statistical rectangles (Hintzen, 2021). When combined with fishing locations available through **VMS data**, it is possible to refine the spatial resolution of the catches and produce fine scale maps of catch distribution (Hintzen et al., 2012).

During my PhD, I developed spatial and spatio-temporal statistical models that integrate commercial catch declarations data with scientific survey data to produce species distribution maps at fine spatio-temporal resolution. This raises major methodological challenges:

- The integration of several data sources: several data sources need to be combined within a hierarchical model to infer a shared latent process (here fish distribution – chapter 3 – 4).
- Accounting for preferential sampling: fishermen behavior is generally directed towards areas of higher biomass and such spatial targeting behavior needs to be handled in the model (chapter 3 – 4).
- The problem of change of support: the different data sources do not have the same spatial resolution (or support) and this needs to be accounted for in the model to

---

reconcile the different spatial scale of the data (chapter 5).

The predictions produced by the model were used to investigate fish spatio-temporal distribution at a monthly resolution over several years for several species of the Bay of Biscay (Chapter 4). Our approach provides maps of the mature fraction of the population (i.e. the individuals that can potentially reproduce); by focusing on the temporal window identified as reproduction period in the literature, the aggregation areas that are identified from our model can be interpreted as reproduction areas. These aggregation areas were proved to be consistent with the knowledge on spawning areas available from the literature which demonstrates the potential of our approach to identify (as a first step) fish spawning grounds. By contrast, scientific survey data alone would only have allowed to compute one map per year, and would not have been informative about reproduction areas as the related surveys fall outside the reproduction period.

Note that a big challenge of this PhD lies in the capacity to combine two field of expertise:

1. the expertise required in fisheries science to properly handle the data, filter the fleets of interest and select the species for which the approach is suitable
2. the strong statistical and technical skills that are required to understand, build and fit the model to the data. The following points keep illustrating this challenge.

In the next paragraph, I highlight some general considerations about the actual framework and the methodological issues that are tackled in this manuscript. I also develop future possible use and developments for the framework.

## 6.1 Challenge of data integration

Chapter 3 and 4 of the PhD illustrate the interest of statistical integrated modeling to integrate multiple data sources in a single inferential framework. However, it also exemplifies the challenges and the difficulties of such data integration especially when data are highly unbalanced or when they do not have the same spatial resolution.

### 6.1.1 Integrating highly unbalanced datasets

The integration of **highly unbalanced data** is a major challenge in the integrated species distribution model literature (Fletcher et al., 2019). From a theoretical point of view, if the model is correctly specified then the balance of the influence of the different

---

data sources in inference depends upon the sample size and upon the variance of the observation processes associated with each data source. The observation variance parameters controls the level of information that is conveyed from the data to the latent field and then balance the information brought by each data source.

In the explored case study (demersal case studies) of chapters 3 and 4, due to the massive amount of commercial data as well as the preprocessing methods used to refine the resolution of logbooks, **commercial observations are much more voluminous than survey data** and consequently they dominate inference. If we consider scientific and commercial data arise from the same observation model (i.e. commercial data are modeled similarly as scientific data and one observation of commercial data is considered to provide the same quantity of information as scientific data), scientific data will always have low contribution in the case of demersal species. Indeed, commercial ‘VMS x logbook’ sample size may easily be tenfold bottom trawl surveys sample size. In practice, bottom trawl surveys usually provide a hundred observations, while ‘VMS x logbook’ data easily provide thousands of observations (Cf. Figure 1.2 in introduction and SM B.3.2).

A first option (suggested by the review of chapter 3) is to give more weight to scientific data by adopting a **data-weighting approach**. Data-weighting methods are very common in stock assessment methods – also much debated (Wang and Maunder, 2017) – as they intend to modify the relative influence of the data sources by assigning or estimating a weight for each data source (Francis, 2017; Punt, 2017; Punt, 2019). Weighting can be achieved outside the model (1) by fixing the weights before fitting the model or (2) by estimating them within the modeling framework as any other parameters (Francis, 2014). Another option consists in (1) fitting the model iteratively, (2) modifying the data weights at each iteration following the results of the previous run and (3) stopping the process after a fixed number of iteration or until the data weights do not change significantly (see Fletcher et al. (2019)). This option requires to fit the model several times (which is not always realistic for operational purposes) and does not allow to propagate uncertainty related to data weighting into estimates. Overall, these are ad-hoc approaches and questions always remain regarding the meaning and the interpretation of the weights used to balance the data.

Another possibility to increase the weight of scientific data (or more appropriately to down weight commercial data) is to **revise the observation process of commercial data**. By doing so, one could expect to correct for model misspecification (Wang and Maunder, 2017). This is a more satisfying solution as most overweighting issues arise in

---

reality from model misspecification which can lead to data inconsistencies. However, it can be particularly challenging to handle model misspecification as there is usually a lack of understanding and strong uncertainty on the ecological processes that actually occur in the latent field and the observation process. In the third chapter, we pointed out that revisiting the observation process of the commercial data and handling COS within the integrated framework allows to decrease the weight of commercial data and to increase the contribution of scientific data in inference while recovering the estimates of the species-habitat relationship. Then, this looks as a better research avenue to investigate how to properly balance the different datasets in inference.

Note also that **other configurations of data exist**. For instance, pelagic acoustic surveys provide much larger sample size as they are based on continuous acoustic observations where the order of magnitude is about 2000 samples per survey (Doray et al., 2018). During a Master thesis that I co-supervised with Etienne Rivot and Marie-Pierre Etienne, Quemper (2021) proposed a similar integrated approach as Alglave et al. (2022), but applied the framework to a pelagic fish (sardine in the Bay of Biscay – Figure 6.1, SM E.2). He outlined that in this case the contribution of the different data sources in the inference was reversed: scientific data mainly contribute to inference during the time steps of the survey and commercial fleets bring very little information to inference as their spatial range is quite narrow and restricted to the coastal area.

Such concentration of the fishing activity in coastal areas between May and October is a consequence of market demands. Indeed, sardines must be fished, sold and transformed the same day and fishermen are then constrained to fish in coastal areas near their harbor. Furthermore, canneries require that sardines have fatty content between 8 and 10%. Consequently, the fishing period begin in May (after reproduction) when fatty content is high enough to commercialize them and end at the beginning of winter in October/November (Bandarra et al., 1997).

This sector effect implies that fishermen only sample data during specific seasons (end spring to autumn), in very constrained areas (coastal areas) and then provide limited information to map sardine distribution. This case study contrasts the case study that was used in the first chapter and clearly illustrates that commercial data may not always provide an informative source of information to map species distribution in particular when the sampling effort is very constrained in space in relation to very specific fisheries context.

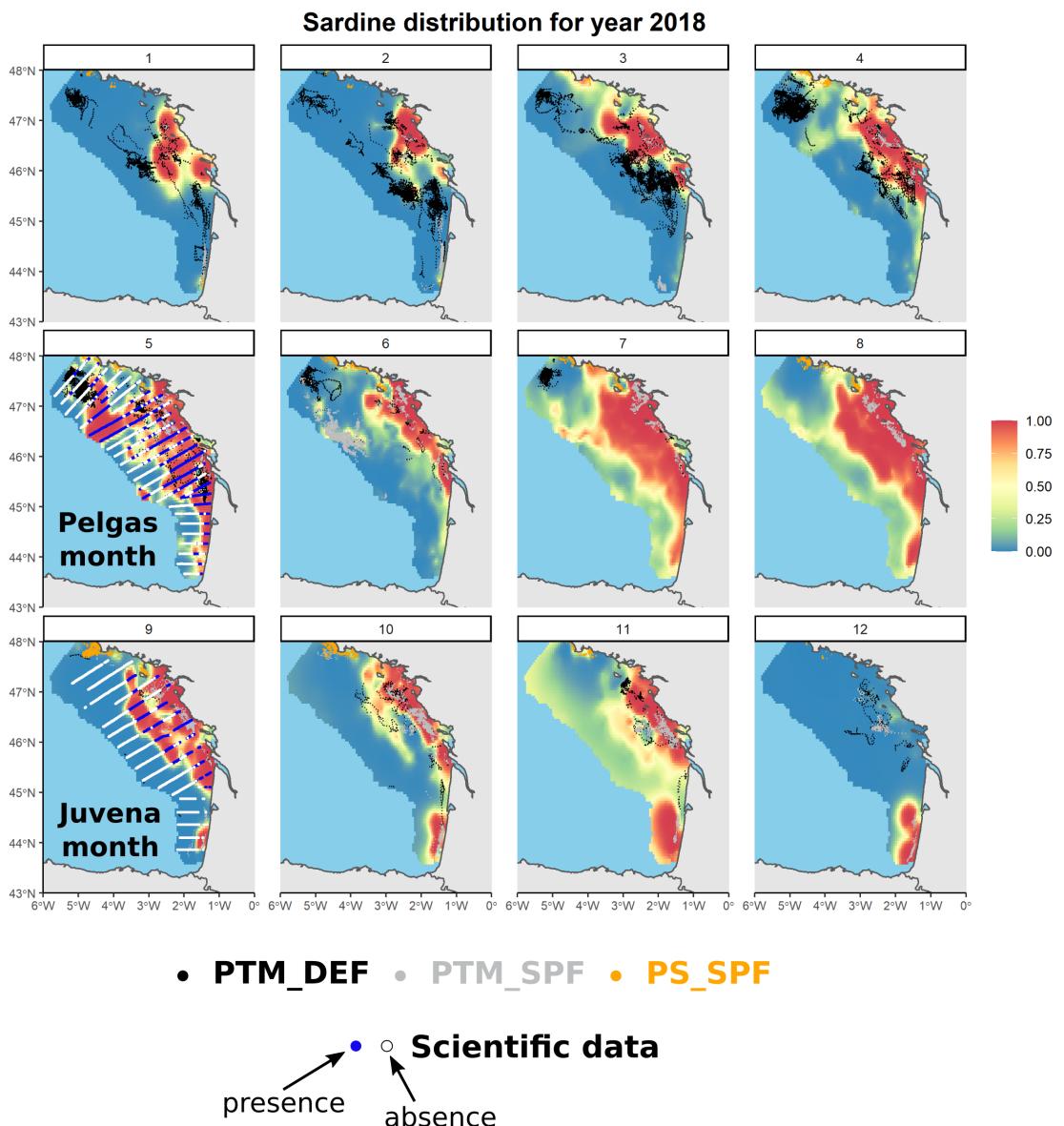


Figure 6.1 – Sardine spatio-temporal monthly distribution for year 2018. Species distribution maps are presence-absence fields. Two acoustic surveys are integrated into inference: the PELGAS survey that occurs in May and the Juvena survey that occurs in September. Three commercial fleets are integrated into inference: pelagic trawls targeting demersal fish (PTM\_DEF), pelagic trawls targeting pelagic fish (PTM\_SPF) and purse seine targeting pelagic fish (PS\_SPF).

---

### 6.1.2 Challenge of change of support

Finally, perhaps the biggest challenge when integrating declarations data with scientific records is to **reconcile the spatial scales of the different data sets**. We demonstrated that a rough reallocation of catch on VMS data allows to represent species distribution relatively reliably through the spatial random effect. However, as a direct consequence of the uniform reallocation of catches on fishing points, inferences on the spatial field of relative density can be smoothed, which may seriously blur the species-habitat relationship. This is a strong limitation as a main focus of species distribution modeling is to define the habitat preferences of species (Guisan and Zimmermann, 2000). We developed a model that reconciles different spatial scales and that allows to recover the species-habitat preferences. By handling properly COS in statistical modeling, we were able to produce more accurate inference of spatial fields.

In **fisheries science**, a huge amount of aggregated data are becoming available through catch declarations data and handling COS for these data will be crucial to make fine scale inference based on these data. COS approaches usually require additional punctual observations to improve accuracy of the model outputs and ease convergence. It is expected that new available data such as haul-level logbook data (Plet-Hansen, Bastardie, and Ulrich, 2020) will provide additional punctual data to feed the model and then will possibly ease the operational use of the framework.

The problematic of COS is a problematic that goes far beyond the specific applications addressed in this thesis. For example, in **ecological applications**, atlas data are often aggregated over large areas and one has to refine resolution of these data to combine them with other fine-scale data sources. See for instance a floristic case study in Trivedi et al. (2008) where rough scale data are compared to fine scale data in their ability to produce similar species-habitat relationship. They outlined a strong bias in species-habitat relationship from the atlas aggregated data that induces a strong over-estimation of climate change effect by providing misleading species-habitat relationship estimates. Typically, the COS issue may explain partly these discrepancies and accounting for COS could correct this bias.

Furthermore, as mentioned in the introduction, **in many other fields of research** declarative datasets are aggregated over large scales while they could be used for fine scale analysis and where properly handling the difference of spatial resolution is a key challenge. Typically, administrative healthcare data (Young and Gotway, 2007), environmental data or air pollution data (Berrocal, Gelfand, and Holland, 2010a; Berrocal, Gelfand, and

---

Holland, 2010b) are data sets that can face COS issues. However, the use of COS is still mostly constrained to the statistical literature and operational applications of COS remains relatively uncommon. There is consequently a great need to develop operational tools to handle COS for spatial analysis. Such tools would possibly greatly help for a wider use of declarative data in research and expertise.

## 6.2 Enhancing integrated ecosystem assessment

Fine scale species distribution maps provided by the model have sufficiently fine spatio-temporal resolution to identify essential habitats, and could therefore be used to inform Marine Spatial Planning for a more integrated ecosystem management (Janßen et al., 2018).

### 6.2.1 Identifying essential fish habitats

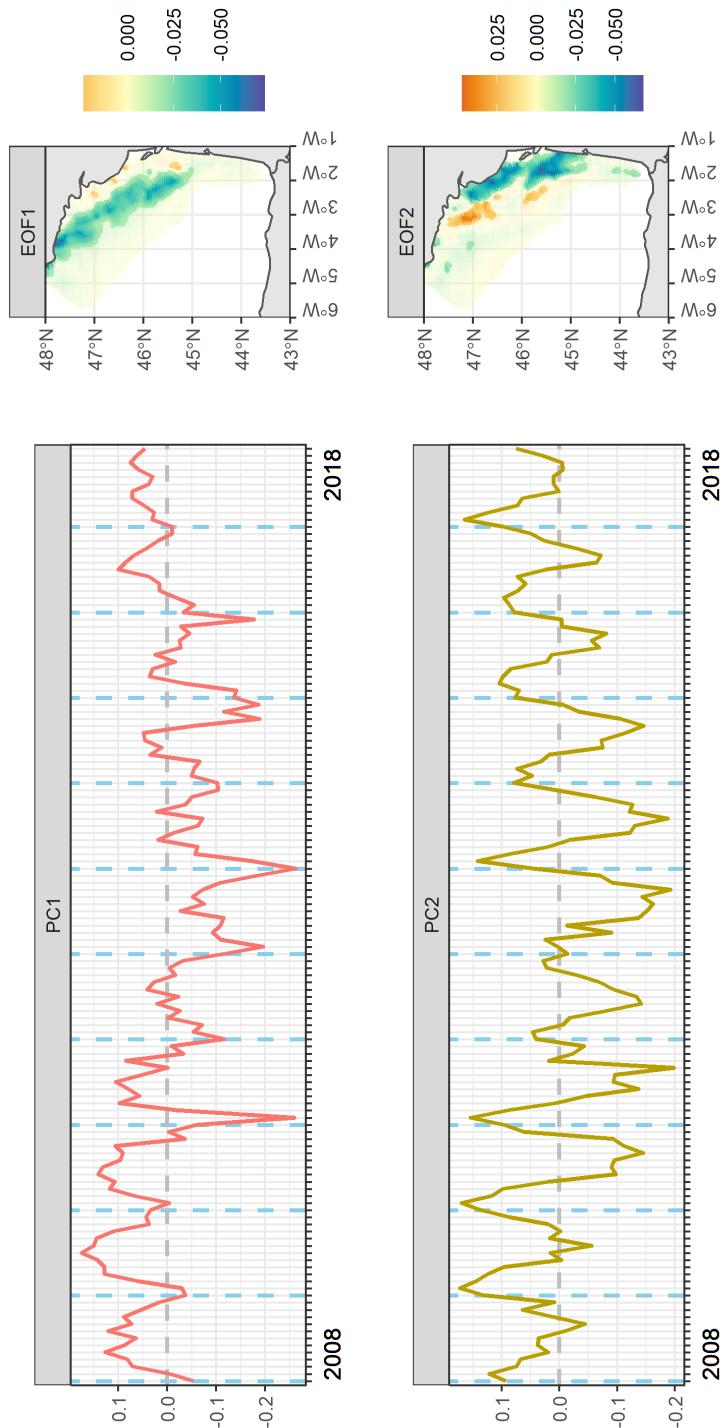
Huge information on fish essential habitats can be extracted from the model outputs. We demonstrated that the model allows to identify **aggregation areas during the spawning season** and that they match with the information on spawning grounds available from literature. Extensive discussion was provided in chapter 4 to emphasize that these outputs should be interpreted with care and that other complementary studies should be conducted to support the interpretation of the aggregation areas as spawning areas e.g. qualitative surveys among fishermen (Silvano et al., 2006; Bezerra et al., 2021) or direct observation of spawning areas through egg and larvae surveys (Fox et al., 2008).

Additional analysis realized a posteriori on model outputs using spatio-temporal data analysis methods (Figures 6.2; SM E.1) illustrate how model outputs can be used to visualize the main **spatio-temporal patterns** that structure the spatial predictions and their temporal variability for several key species of the Bay of Biscay and the Celtic Sea (hake, sole and sea bass). Strong seasonal patterns can be identified and can be related to reproduction migrations between the feeding areas (summer period) and the reproduction areas (winter period).

Exciting perspectives of this work would consist in applying a similar method to several species to identify essential habitats at the scale of fish community. Furthermore, integrating hydrological and planktonic variables in the analysis would enable to investigate spatio-temporal variability of other compartments of the ecosystem. Similarly as

---

is performed by Petitgas et al. (2018) at a yearly time step, such analysis performed at a monthly time step would allow to identify ecosystem components with their related intra-annual and inter-annual variability. This is typically the type of products that are required for the integrated assessment of ecosystems (Woillez et al., 2010; Levin et al., 2014; Möllmann et al., 2014).



---

Figure 6.2 – Two first EOF maps and principal component time series (PC) for sole of the Bay of Biscay. Blue dashed line: January. For EOF maps, only the locations that have a significant contribution to the dimension are plotted. Interpretation of the plot is as follow: when the principal component (PC) is positive (resp. negative), then sole is mostly distributed in red areas (resp. blue) on the related EOF map. One can see on PC1 and PC2 strong seasonal signals which can be interpreted as migration patterns between offshore reproduction areas in winter (blue areas on EOF1 and red areas on EOF2) and coastal feeding areas in summer (blue areas on EOF2).

### 6.2.2 Towards seasonal spatio-temporal population dynamics

Fisheries science is primarily (but not only) interested in qualifying the effect of exploitation on population dynamics (Hilborn and Walters, 1992; Gascuel, 2015). Typically, **stock assessment models** seek to capture fish population dynamics by modeling recruitment, growth and exploitation processes. These models form the basis of the present management of harvested marine resources in Europe. The overall advice procedure, the models and the data that feed the full management process are designed on a yearly time step at the scale of rough ICES areas without accounting for spatial heterogeneity of fish distribution and fishing effort.

Progressively, models are becoming spatially explicit to **overcome the ‘well-mixed single-stock’ paradigm** of most common stock assessment models (Ulrich et al., 2013; Punt et al., 2020; Archambault et al., 2016; Archambault et al., 2018). Although these are generally not yet operational, these kinds of models are considered as basis for a new generation of stock assessment frameworks (Cao et al., 2020). For now, these are fitted on data with yearly time-step (Olmos et al., n.d.), but our results exemplify that the huge amount of data available through ‘VMS x logbook’ data offers possibility to refine the temporal accuracy of these models and to describe fish spatio-temporal dynamics at an infra-annual time step (e.g. month or quarter). Such model would typically allow to investigate which segment of the population is over/under-exploited, when (which season/month?) and how does exploitation interact with the key species life cycle stages and the related essential fish habitats. This would require to introduce population dynamics in the latent field of the actual framework. Typically, the equations could be based on similar equations as in **the model proposed by Cao et al. (2020)** which reproduces:

1. The population dynamics. Abundance at size  $\mathbf{n}_{x,t}$  for location  $x$  and time step  $t$  is

---

modeled as a function of:

- the abundance at size of the previous time step  $\mathbf{n}_{x,t-1}$
- a natural mortality at size vector  $\mathbf{m}_{x,t-1}$
- a fishing mortality at size vector  $\mathbf{f}_{x,t-1}$  and selectivity at size  $\mathbf{v}$
- a growth transition matrix  $\mathbf{G}$  modeling the transition from smaller size classes to bigger size classes
- a recruitment at size term  $\mathbf{r}_{x,t}$  modeling the new individuals in a size class

$$g(\mathbf{n}_{x,t}) = \mathbf{G}(\mathbf{n}_{x,t-1} \cdot \exp(-\mathbf{m}_{x,t-1} - \mathbf{v}\mathbf{f}_{x,t-1})) + \mathbf{r}_{x,t}$$

2. The fishing pressure. Similarly as in the Baranov equation, catches at size vector  $\mathbf{c}_{x,t}$  for location  $x$  and year  $t$  depend on fishing mortality at size  $\mathbf{f}_{x,t}$ , natural mortality at size  $\mathbf{m}_{x,t}$  and the abundance at size  $\mathbf{n}_{x,t}$ .

$$\mathbf{c}_{x,t} = (\mathbf{v}\mathbf{f}_{x,t}) / (\mathbf{v}\mathbf{f}_{x,t} + \mathbf{m}_{x,t}) \cdot (1 - \exp(-\mathbf{m}_{x,t} - \mathbf{v}\mathbf{f}_{x,t})) \cdot \mathbf{n}_{x,t}$$

Recent extensions of this type of framework also explicitly include movement based on combined survey data, commercial data or mark-recapture data (Hulson et al., 2013; Thorson et al., 2021b).

Overall, moving to an infra-annual time-step for these models would provide great tools to investigate fish spatio-temporal dynamics at fine spatio-temporal resolution.

### 6.2.3 Perspectives for an implementation in Marine Spatial Planning

Our results can help designing **Marine Protected Areas (MPA)** or spatio-seasonal closures specifically with the perspective of **protecting spawning grounds** of species that aggregates during the reproduction season (Grüss et al., 2019; Biggs et al., 2021). Protecting the spawning grounds to preserve population renewal is an old idea in fisheries science:

1. because individuals can be vulnerable during this period due to their aggregating behavior leading to population collapse if exploitation is not regulated properly on this time span (Sadovy and Domeier, 2005).
2. because these are crucial areas for species life cycle and degradation of these habitats could decrease the population reproductive capacity (Seitz et al., 2014).

---

Protecting spawning areas is typically the approach that is invoked by the STECF (Scientific, Technical and Economic Committee for Fisheries) working group focusing on spatio-seasonal closures for the demersal species of the Western Mediterranean Sea. In that context, the model has been applied on French, Spanish and Italian ‘VMS x logbook’ data for hake (SSB and juveniles - *Merluccius merluccius*), red mullet (*Mullus barbatus*) and shrimps (*Aristaeomorpha foliacea* and *Aristeus antennatus* – Figure 6.3 for some species distribution maps produced in this working group). The outputs of the model aim at designing fishing closure scenarios that will be assessed through simulation modeling tools for potential implementation as regulation measures.

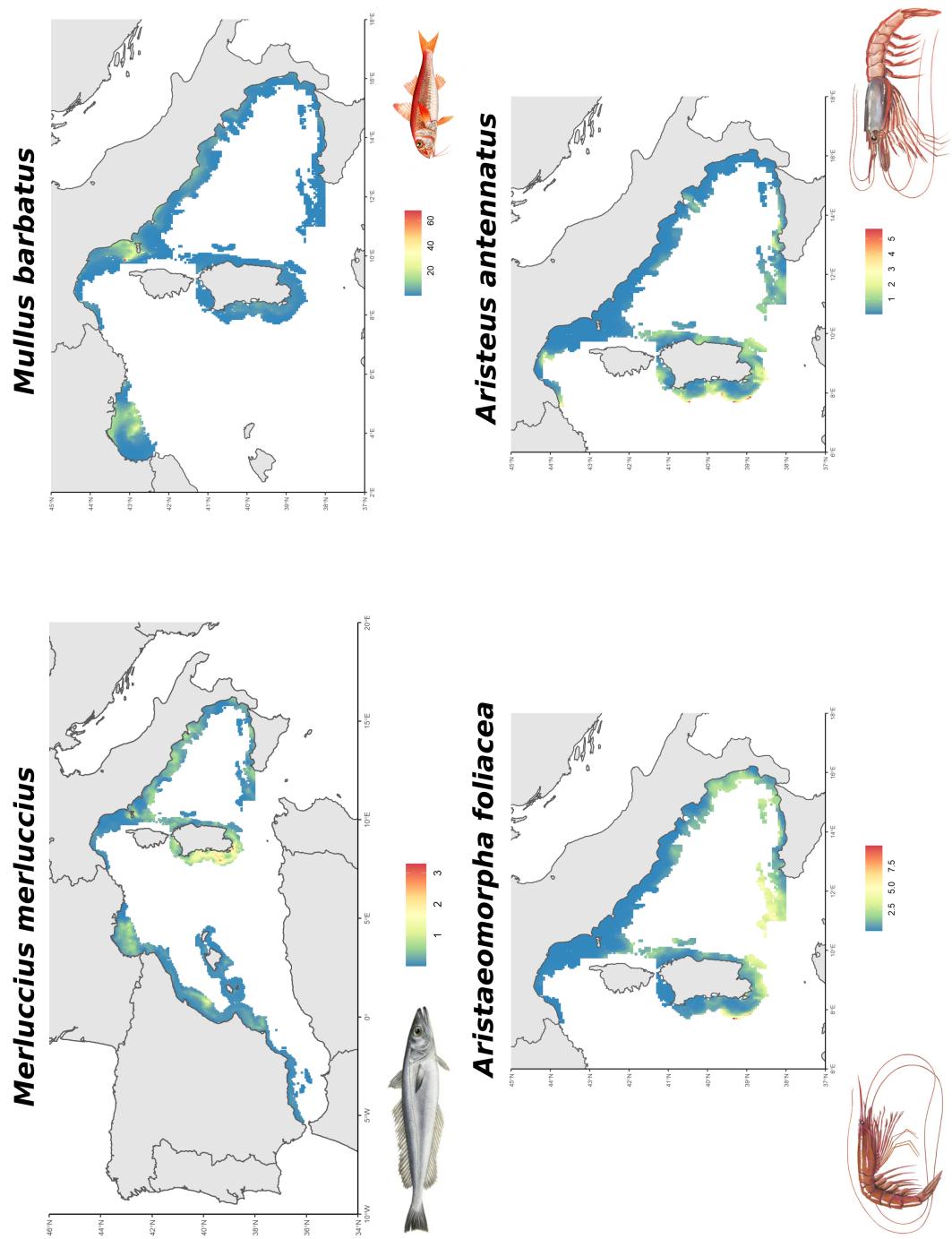


Figure 6.3 – Average spatial distribution for the demersal species of the Western Mediterranean Sea. Unit: for hake, densities in kg/km<sup>2</sup>; for other species, CPUE in kg/hour. See SM E.3 for the full report.

---

Although modeling works emphasize the benefit of MPAs designed to protect fish spawning aggregations on fish abundance and recovery, **strong uncertainties remain regarding concrete and empirical benefits of these MPA**. Indeed, these may have poor effect on fish productivity or recovery due to lack of enforcement, inadequate design and insufficient time since MPA creation (Grüss et al., 2014). Furthermore, it is recognized that implementing MPA on reproduction grounds can strongly affect fishermen income by preventing them to access important fishing areas. This is particularly true for temperate waters as fish spawning areas are generally much more extensive than in tropical waters. Then, great care should be taken to ensure the economic viability of such measures as closing these areas could be very costly to fishermen on the short run while gains on the long run would be very hypothetical (Smith et al., 2010).

Furthermore, **real efficiency of MPA may also be diminished by modification in fish distribution**. For instance, the Plaice Box experience in the North Sea illustrates how an unexpected movement of plaice juveniles outside the MPA made it ineffective while there was a common consensus among fishermen, stakeholders and researchers for its implementation (Beare et al., 2013). Even though the Plaice Box was designed for nursery grounds, similar processes may affect other stages of life cycle (e.g. reproduction season and location). This is especially true in the context of climate change as species phenology (and specifically reproduction timing and location) may be affected by environmental modifications (Pendleton et al., 2022). In such context, only a regular assessment of the closure areas and adaptive management measures may be efficient to face these challenges.

The outputs provided by the model (information on essential habitats and species distribution) could also be combined with spatial management assessment tools such as Displace (Bastardie, Nielsen, and Miethe, 2014) or ISIS-Fish (Mahévas and Pelletier, 2004) to fairly **assess the trade-off between multiple scenarios of the marine space use**. This is particularly true in a context of high competition for the marine space. Indeed, fishermen target certain hotspots in space and they can hardly move their activity without reducing the viability of their business (Hintzen, 2021). Marine spatial planning needs to identify and protect the most crucial fishing locations for the economical viability of the fisheries and evaluate the best way to exploit these areas in terms of economic, social, ecological effects to limit as much as possible the non-desired effect of conflicting cross-sectors (Fock, 2008; Jennings and Lee, 2012; Campbell et al., 2014). For instance, fishing activities and other activities exploiting the marine ecosystem can be mutually exclusive (e.g. marine renewable energy platforms, shipping routes) or

---

in contradiction with management objectives of fisheries (e.g. marine conservation areas) requiring then a fair trade-off between the different activities (Bastardie et al., 2015). The behavioral shifts that could arise from area closures or from competition for space could have strong non-expected deleterious effect on non-targeted species and other life-stages while affecting the adaptive capacity of fishermen by constraining their activity to very limited areas, forcing them to leave the sector in case of crisis (Suuronen, Jounela, and Tschernij, 2010).

## 6.3 Improving the realism of fishermen targeting behavior

### 6.3.1 Modeling targeting behavior as multifactorial process

Another major challenge when using commercial data is related to **the preferential distribution of fishing locations** in areas of higher biomass. We demonstrated that such behavior leads to bias in spatial predictions if not accounted for in inference. The so-built framework provides a method to account and to quantify for PS in the context of fishery data. In this case, targeting behavior arises from PS towards the species of interest, but also arises from many other factors such as tradition, physical constraints (e.g. distance to the coast), management regulations (e.g. MPA).

Quantifying the relation between fishermen behavior and resource distribution is a key issue in the fisheries literature as understanding the relationship between fishermen and resource strongly determines the way management regulations can be implemented (Tidd et al., 2015).

The framework offers a starting point to model fishermen preferences towards areas of higher fish biomass. From a modeling perspective, we adopted a simple **point process modeling** approach and assumed the fishing locations  $\mathbf{X}$  follow a standard inhomogeneous spatial point process with intensity  $\lambda(x)$  where intensity depends on a combination of species distribution and on additional processes modeled through a spatial random term.

$$\mathbf{X} \sim \mathcal{IPP}(\lambda(x)) \quad \text{and} \quad \lambda(x) = \alpha_X + b \cdot \log(S(x)) + \eta(x)$$

where  $\lambda(x)$  is the spatial smoothed representation of fishermen preferences (fishing

---

intensity) towards location  $x$ ,  $\alpha_X$  is the intercept of the sampling intensity,  $b$  is the preferential sampling parameter,  $S(x)$  is fish biomass,  $\eta(x)$  is the spatial random effect capturing remaining variability of fishing intensity. As stated in chapter 3 and 4, such model strongly simplifies the relationship between fishing locations and fish biomass. Indeed, the functional relationship may be much more complex and could vary in space (Conn, Thorson, and Johnson, 2017) or may be related to multiple interacting factors that affect fishermen spatial preferences (e.g. distance to the coast, management regulation, economic factors - Girardin et al. (2017)). Existing research perspectives would consist in incorporating new explicative factors or by modeling some temporal lags in the sampling process to consider that information from other time steps affect the actual sampling positions of fishermen (see for instance the work of Vermard et al. (2008)).

### 6.3.2 Modeling fishing locations as a non-random trajectory

In addition, considering that fishing locations arise from a **simple point process is another limitation** of the framework that could be relaxed in future development. In particular, such process assumes that conditionally on  $\lambda(x)$  all fishing positions are sampled independently. Fishing locations correspond more realistically to realizations of a trajectory where locations are conditionally dependent on previous locations and whose direction is conditional on fishermen preferences towards some habitats.

Several methods are suited to study the relation between the movement of a forager – here fishermen – and the distribution of the resource it uses – here fish (Gloaguen, 2015). An appealing option to define spatial use of a forager is to define the fishing position  $X_t$  at time  $t$  through a **stochastic differential equation (SDE)**. For instance, Gloaguen, Etienne, and Le Corff (2018) proposed a SDE framework where fishing positions  $X_t$  are modeled as:

$$dX_t = b_\eta(X_t) dt + \gamma dW_t \quad \text{and} \quad X_0 = x_0$$

where  $(W_t)_{0 \leq t \leq T}$  is a standard Brownian motion on  $\mathbb{R}^2$ ,  $\gamma \in \mathbb{R}_+^*$  is a diffusion parameter and  $b_\eta(X_t)$  is a drift term.  $b_\eta(X_t)$  is a potential map controlling the foragers preference towards some specific area. It is defined as:

---


$$\begin{aligned}
b_\eta(x) &:= \nabla P_\eta(x), \\
P_\eta(x) &= \sum_{i=1}^K \pi_k \varphi_k^\eta(x) \\
\varphi_k^\eta(x) &:= \exp \left\{ -\frac{1}{2} (x - \mu_k)^T \mathbf{C}_k (x - \mu_k) \right\}
\end{aligned}$$

The potential map  $P_\eta(x)$  is modeled as a mixture of Gaussian distribution with  $K$  components. For each  $k^{th}$  component,  $\pi_k \in \mathbb{R}^+$  is the relative weight of the component ( $\sum_{k=1}^K \pi_k = 1$ ),  $\mu_k \in \mathbb{R}^2$  is its centre,  $\mathbf{C}_k \in \mathcal{S}_2^+$  is its information matrix which is defined on a  $2 \times 2$  symmetric positive definite matrices. Michelot et al. (2019) proposed an extension of the framework to include environmental covariates.

In such framework, movement is supposed to arise from the attractiveness of some areas, which is modeled as a potential function  $P_\eta(x)$ .  $P_\eta(x)$  is a multimodal surface with  $K$  components.  $K$  is fixed while  $\pi_k$ ,  $\mu_k$  and  $\mathbf{C}_k$  are estimated.

Even though **this framework** would allow to model fishermen trajectory towards their preferential areas, it **would not distinguish between the preference towards the resource and other factors influencing fishermen targeting**. To overcome this limitation, one pragmatic approach would be to consider that species distribution is known and use this as a covariate in the trajectory framework (see for instance Michelot et al. (2019)). This would require to proceed in a two-step estimation (first estimate species distribution and second estimate fishermen preferences towards biomass while taking species distribution as covariate). However, in such approach, fishermen preferences and species distribution would not be estimated simultaneously. Alternatively, a more challenging approach would be to infer the potential function representing fishermen preferences through two surface layers: one related to fish distribution representing the preference of fishermen towards fish biomass and another representing the remaining fishermen habitat preferences that are not related to fish distribution. This may be challenging as the potential surface would not be only a simple mixture of Gaussian distributions, but a mixture of 2 continuous layers representing 2 different types of habitat preferences (one for resource distribution, one for other additional factors). Instead of modeling fishermen preferences through a Point process, this would give an alternative way to look at preferential sampling based on trajectory modeling. Furthermore, extending the SDE approach to model that fishermen movement arise from a set of variables – i.e. some being related to fish distribution and others being related to additional factors

---

independent from the resource – looks as a nice extension of the existing framework.

### 6.3.3 Adapting the framework to mixed fisheries

Most fisheries are per se plurispecific as fishermen simultaneously catch a set of species. These types of fisheries are commonly referred to as mixed fisheries (Ulrich, 2021). These are characterized by technical interactions i.e. the fact that fishers catch a set of distinct species with distinct size, some being targeted, others being by-caught and some being unwanted or not allowed to be sold. These lasts are most often discarded at sea, even though regulations are being implemented to forbid such behavior (i.e. the landing obligation – Borges et al. (2018)).

#### Moving to multispecies modeling

##### Modeling species joint-distribution

Moving to a multi-specific framework would constitute a valuable extension of the model. **Accounting for species correlation in space and time** is now becoming standard through joint multispecies distribution models (Clark et al., 2014; Pollock et al., 2014; Thorson et al., 2015b; Thorson et al., 2016b). Such approaches allow to model spatial cross-correlations among species and reproduce the fact that some species are positively correlated because they share the same habitats (co-occurrence) or may be negatively correlated due to competition relationship. This typically allows to separate species that are related together and to constitute communities based on their spatial co-occurrence. For instance, Dolder, Thorson, and Minto (2018) identified assemblages of species being caught together within a mixed fishery to provide the knowledge for understanding which species should be managed together in such fisheries.

##### Modeling multi-species targeting

Another aspect of the multi-specific aspect of fisheries is the **multi-specific dimension of fishermen targeting behavior**. Understanding how is structured fishermen spatial targeting relatively to the targeted species distribution is key in properly managing mixed fisheries (Bourdaud et al., 2019). Implementing a multispecies component in the sampling process would typically allow to reproduce the multispecies dimension of targeting by quantifying the spatial preference of fishermen for a set of species instead of a single species. This way, one could quantify the species that are preferentially targeted

---

(PS parameter  $b > 0$ ) and those that are avoided (PS parameter  $b < 0$ ). Such knowledge (by-caught and discarded species distribution as well as fishermen targeting) is typically required to manage mixed fisheries in order to identify which area should be avoided at which period because unwanted catch may be present (Bellido et al., 2019).

## Integrating discards into inference

Last, a blind spot of catch declarations are discards as they are usually not recorded in logbook data. **Discards constitute a major issue in mixed fisheries** and understanding what control discarding is crucial in mitigating the deleterious impact it has on fish populations and on ecosystems. This is particularly challenging because contrarily to targeted species where some spatial and temporal patterns can be identified in relation to clear targeting behavior or seasonal activity of fishing, discards dynamics are much more erratic and depend on many factors such as (Rochet and Trenkel, 2005):

1. environmental factors (space, time and environmental variables)
2. fishing effort (amount and spatiotemporal distribution and gear characteristics)
3. value of the caught species
4. hold capacity, trip duration
5. actual management regulations

**Discards are often mapped based on onboard observer data** (Yan et al., 2022) and following the species under study these can represent an important part of the catches and they can face strong spatial variations. Consequently, a non-negligible part of the catches may be missing from logbook data which may bias the maps of species distribution produced by the model.

Then, there would be a need **to integrate observer data in the framework to inform the discarded portion of the catch**. Considering observer data inform the full catch while logbook data only inform landings would allow to transfer the information on discards from the onboard observer data to the logbook data (see similar ideas in Breivik, Storvik, and Nedreaas (2017) and Stock et al. (2019)).

For instance, we could specify the model so that:

— the catch  $Y_{catch}$  is a sum of landings  $Y_{landings}$  and discards  $Y_{discards}$ :

$$Y_{catch} = Y_{discards} + Y_{landings}$$

- 
- that the onboard data allow to provide observations on both  $Y_{catch}$ ,  $Y_{discards}$  and  $Y_{landings}$  while logbook data only provide observations on  $Y_{landings}$
  - that  $Y_{catch}$  is informative of species distribution through the equation:

$$Y_{catch}|S(x), x \sim M_Y(p, S(x), \sigma^2)$$

with  $M_Y$  the probability distribution,  $p$  the probability to obtain a null observation,  $S(x)$  the latent field value and  $\sigma^2$  the variance parameter. By parameterizing the relationship between  $Y_{catch}$ ,  $Y_{discards}$  and  $Y_{landings}$  and integrating both observer data and logbook data in inference, one should be able to infer simultaneously  $S(x)$  and the  $Y_{discards}$  that are not declared in logbook data. For highly discarded species (especially when the discards have a contrasted spatial distribution), such extension could strongly modify the maps obtained from the spatio-temporal model.

## 6.4 Conclusion

The spatio-temporal model developed in this thesis provide a strong basis for identifying essential habitats and investigating the spatio-temporal patterns that structure fish distribution. These results could find application in Marine Spatial Planning to identify areas that should be protected for ensuring both species renewal (in particular spawning areas) and fishermen income. Inclusion of population dynamics in the framework as well as a multispecies dimension to reproduce mixed fisheries processes looks as a natural step forward and these could find huge audience and application in fisheries science.

Overall, an important conclusion of these methodological considerations is the care that should be taken when using fishermen declarations data to infer fish spatial distribution. There is a great need to specify spatial models that respect the observation scale of the data and its sampling characteristics in order to make accurate inference on species distribution at a fine scale.

Finally, the main contribution of this work goes beyond the domain of fisheries science. The methodological challenges that we tackled may find application in other fields of spatial statistics when intending to use data that face PS and that are aggregated over rough scales. Development of operational tools to give access to these methods is a need for a wider use of these data in research and expertise.

# BIBLIOGRAPHY

---

- Abbott, Joshua, A. Haynie, and Matthew Reimer (2015), « Hidden Flexibility: Institutions, Incentives, and the Margins of Selectivity in Fishing », *in: Land Economics* 91, pp. 169–195.
- Aeberhard, William H, Joanna Mills Flemming, and Anders Nielsen (2018), « Review of state-space models for fisheries science », *in: Annual Review of Statistics and Its Application* 5, pp. 215–235.
- Alglave, Baptiste et al. (2022), « Combining scientific survey and commercial catch data to map fish distribution », *in: ICES Journal of Marine Science*, fsac032, ISSN: 1054-3139.
- Arbault, Par Suzanne, P. Camus, and C. le Bec (1986), « Estimation du stock de sole (*Solea vulgaris*, Quensel 1806) dans le Golfe de Gascogne à partir de la production d'œufs », *in: Journal of Applied Ichthyology* 2.4, pp. 145–156, ISSN: 1439-0426.
- Archambault, Benoit et al. (2016), « Adult-mediated connectivity affects inferences on population dynamics and stock assessment of nursery-dependent fish populations », *in: Fisheries Research* 181, pp. 198–213.
- Archambault, Benoit et al. (2018), « Using a spatially structured life cycle model to assess the influence of multiple stressors on an exploited coastal-nursery-dependent population », *in: Estuarine, Coastal and Shelf Science* 201, Publisher: Elsevier, pp. 95–104.
- Auger-Méthé, Marie et al. (2021), « A guide to state-space modeling of ecological time series », *in: Ecological Monographs*.
- Azevedo, Manuela and Cristina Silva (2020), « 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, pp. 2933–2944, ISSN: 1054-3139.
- Baddeley, Adrian, Ege Rubak, and Rolf Turner (2015), *Spatial point patterns: methodology and applications with R*, CRC press.

- 
- Bakka, Haakon (2018), « How to solve the stochastic partial differential equation that gives a Matern random field using the finite element method », *in: arXiv preprint arXiv:1803.03765*.
- Bakka, Haakon et al. (2018), « Spatial modeling with R-INLA: A review », *in: Wiley Interdisciplinary Reviews: Computational Statistics* 10.6, e1443.
- Bandarra, NM et al. (1997), « Seasonal changes in lipid composition of sardine (*Sardina pilchardus*) », *in: Journal of food science* 62.1, pp. 40–42.
- Banerjee, Sudipto, Bradley P. Carlin, and Alan E. Gelfand (2014), *Hierarchical modeling and analysis for spatial data*, CRC press.
- Bastardie, Francois, J. Rasmus Nielsen, and Tanja Miethe (2014), « DISPLACE: a dynamic, individual-based model for spatial fishing planning and effort displacement—integrating underlying fish population models », *in: Canadian Journal of Fisheries and Aquatic Sciences* 71.3, pp. 366–386.
- Bastardie, Francois et al. (2015), « Competition for marine space: modelling the Baltic Sea fisheries and effort displacement under spatial restrictions », *in: ICES Journal of Marine Science* 72.3, pp. 824–840, ISSN: 1054-3139.
- Bastardie et al. (2010), « Detailed mapping of fishing effort and landings by coupling fishing logbooks with satellite-recorded vessel geo-location », *in: Fisheries Research* 106.1, pp. 41–53.
- Bauder, Javan M. et al. (2021), « Mismatched spatial scales can limit the utility of citizen science data for estimating wildlife-habitat relationships », *in: Ecological Research* 36.1, pp. 87–96.
- Beare, Doug et al. (2013), « Evaluating the effect of fishery closures: lessons learnt from the Plaice Box », *in: Journal of Sea Research* 84, Publisher: Elsevier, pp. 49–60.
- Bellido, José M. et al. (2019), « A marine spatial planning approach to minimize discards: Challenges and opportunities of the Landing Obligation in European waters », *in: The European Landing Obligation*, Publisher: Springer Cham, p. 239.
- Berg, Astrid (1999), *Implementing and enforcing European fisheries law: the implementation and the enforcement of the Common Fisheries Policy in the Netherlands and in the United Kingdom*, Martinus Nijhoff Publishers.
- Berrocal, Veronica J., Alan E. Gelfand, and David M. Holland (2010a), « A bivariate space-time downscaler under space and time misalignment », *in: The annals of applied statistics* 4.4, p. 1942.

- 
- (2010b), « A spatio-temporal downscaler for output from numerical models », *in: Journal of agricultural, biological, and environmental statistics* 15.2, pp. 176–197.
- Bertrand, Sophie et al. (2005), « Lévy trajectories of Peruvian purse-seiners as an indicator of the spatial distribution of anchovy (*Engraulis ringens*) », *in: ICES Journal of Marine Science* 62.3, pp. 477–482.
- Bez, Nicolas, Didier Renard, and Dedah Ahmed-Babou (2022), « Empirical Orthogonal Maps (EOM) and Principal Spatial Patterns: Illustration for Octopus Distribution Off Mauritania Over the Period 1987–2017 », *in: Mathematical Geosciences*, pp. 1–16.
- Bezerra, Inajara Marques et al. (2021), « Spatial and temporal patterns of spawning aggregations of fish from the Epinephelidae and Lutjanidae families: An analysis by the local ecological knowledge of fishermen in the Tropical Southwestern Atlantic », *in: Fisheries Research* 239, p. 105937.
- Biais, Gérard (2003), *ORHAGO*, Publisher: Sismar.
- Biggs, Christopher R. et al. (2021), « The importance of spawning behavior in understanding the vulnerability of exploited marine fishes in the US Gulf of Mexico », *in: PeerJ* 9, e11814.
- Billet, Norbert et al. (2021), *Evaluation des fermetures spatio-temporelles mises en oeuvre à partir du 1er janvier 2020 pour la pêche au chalut en mer Méditerranée*, Ifremer.
- Bivand, Roger S. and David W. S. Wong (2018), « Comparing implementations of global and local indicators of spatial association », *in: TEST* 27.3, pp. 716–748, ISSN: 1863-8260.
- Board, Ocean Studies and National Research Council (2000), *Improving the collection, management, and use of marine fisheries data*, National Academies Press.
- Borges, Lisa et al. (2018), *Conflicts and trade-offs in implementing the Common Fisheries Policy (CFP) discard policy*, DiscardLess Deliverable Report 7.3, Publisher: DiscardLess.
- Botella, Christophe et al. (2018), « Species distribution modeling based on the automated identification of citizen observations », *in: Applications in Plant Sciences* 6.2, e1029.
- Botella, Christophe et al. (2021), « Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data », *in: Methods in Ecology and Evolution* 12.5, pp. 933–945.
- Bourdaud, Pierre et al. (2017), « Inferring the annual, seasonal, and spatial distributions of marine species from complementary research and commercial vessels' catch rates », *in: ICES Journal of Marine Science* 74.9, pp. 2415–2426, ISSN: 1054-3139.

- 
- Bourdaud, Pierre et al. (2019), « Improving the interpretation of fishing effort and pressures in mixed fisheries using spatial overlap metrics », *in: Canadian Journal of Fisheries and Aquatic Sciences* 76.4, pp. 586–596, ISSN: 0706-652X, 1205-7533.
- Breivik, Olav Nikolai, Geir Storvik, and Kjell Nedreaas (2017), « Latent Gaussian models to predict historical bycatch in commercial fishery », *in: Fisheries Research* 185, pp. 62–72.
- Brown, Elliot J. et al. (2018), « Conflicts in the coastal zone: human impacts on commercially important fish species utilizing coastal habitat », *in: ICES Journal of Marine Science* 75.4, Publisher: Oxford University Press, pp. 1203–1213.
- Cameletti, Michela, Rosaria Ignaccolo, and Stefano Bande (2011), « Comparing spatio-temporal models for particulate matter in Piemonte », *in: Environmetrics* 22.8, pp. 985–996.
- Cameletti, Michela et al. (2013), « Spatio-temporal modeling of particulate matter concentration through the SPDE approach », *in: AStA Advances in Statistical Analysis* 97.2, pp. 109–131, ISSN: 1863-818X.
- Campbell, Maria S. et al. (2014), « Mapping fisheries for marine spatial planning: Gear-specific vessel monitoring system (VMS), marine conservation and offshore renewable energy », *in: Marine Policy* 45, pp. 293–300, ISSN: 0308-597X.
- Cao, Jie et al. (2020), « A novel spatiotemporal stock assessment framework to better address fine-scale species distributions: development and simulation testing », *in: Fish and Fisheries* 21.2, pp. 350–367.
- Cariou, Thibault et al. (2021), « Comparison of the spatiotemporal distribution of three flatfish species in the Seine estuary nursery grounds », *in: Estuarine, Coastal and Shelf Science* 259, p. 107471.
- Cheung, William WL et al. (2009), « Projecting global marine biodiversity impacts under climate change scenarios », *in: Fish and fisheries* 10.3, Publisher: Wiley Online Library, pp. 235–251.
- Clark, James S. et al. (2014), « More than the sum of the parts: forest climate response from joint species distribution models », *in: Ecological Applications* 24.5, Publisher: Wiley Online Library, pp. 990–999.
- Cochran, William G. (1977), *Sampling techniques*, 3rd edn., John Wiley and Sons.
- Coggins Jr, Lewis G, Nathan M Bacheler, and Daniel C Gwinn (2014), « Occupancy models for monitoring marine fish: a Bayesian hierarchical approach to model imperfect detection with a novel gear combination », *in: PLoS One* 9.9, e108302.

- 
- Conn, Paul B. (2010), « Hierarchical analysis of multiple noisy abundance indices », *in: Canadian Journal of Fisheries and Aquatic Sciences* 67.1, pp. 108–120.
- Conn, Thorson, and Johnson (2017), « Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage », *in: Methods in Ecology and Evolution* 8.11, pp. 1535–1546.
- Conrad, Cathy C and Krista G Hilchey (2011), « A review of citizen science and community-based environmental monitoring: issues and opportunities », *in: Environmental monitoring and assessment* 176.1, pp. 273–291.
- Constantin, Alexandre, Mathieu Fauvel, and Stéphane Girard (2021), « Joint supervised classification and reconstruction of irregularly sampled satellite image times series », *in: IEEE Transactions on Geoscience and Remote Sensing* 60, pp. 1–13.
- Cornou, Anne-Sophie et al. (2021), *Captures et rejets des métiers de pêche français - Résultats des observations à bord des navires de pêche professionnelle en 2019*, Ifremer.
- Coupeau, Yann and Gérard Biais (2019), *ORHAGO 19*, Publisher: Sismer.
- Cressie, Noel and Christopher K. Wikle (2015), *Statistics for spatio-temporal data*, John Wiley and Sons.
- Cressie, Noel AC (1993), « Statistics for spatial data. John Willy and Sons », *in: Inc., New York*.
- Dambrine, Chloé et al. (2021), « Characterising Essential Fish Habitat using spatio-temporal analysis of fishery data: A case study of the European seabass spawning areas », *in: Fisheries oceanography* 30.4, pp. 413–428.
- Delage, Nicolas and Olivier Le Pape (2016), *Inventaire des zones fonctionnelles pour les ressources halieutiques dans les eaux sous souveraineté française. Première partie: définitions, critères d'importance et méthode pour déterminer des zones d'importance à protéger en priorité*, Rapport de recherche, Rennes: Pôle halieutique AGROCAMPUZ OUEST, p. 36.
- Demanèche, Sébastien et al. (2013), *Projet SACROIS*, STH/LBH/SACROIS, Ifremer / DPMA.
- Deporte, Nicolas et al. (2012), « Regional métier definition: a comparative investigation of statistical methods using a workflow applied to international otter trawl fisheries in the North Sea », *in: ICES Journal of Marine Science* 69.2, Publisher: Oxford University Press, pp. 331–342.

- 
- Di Stefano, Marine et al. (2022), « Insights into the spatio-temporal variability of spawning in a territorial coastal fish by combining observations, modelling and literature review », *in: Fisheries Oceanography*.
- Diggle, Menezes, and Su (2010), « Geostatistical inference under preferential sampling », *in: Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59.2, pp. 191–232.
- Diggle, Peter J. (2013), *Statistical analysis of spatial and spatio-temporal point patterns*, CRC press.
- Dobson, Andrew DM et al. (2020), « Making messy data work for conservation », *in: One Earth* 2.5, pp. 455–465.
- Dolder, Paul J., James T. Thorson, and Cólín Minto (2018), « Spatial separation of catches in highly mixed fisheries », *in: Scientific reports* 8.1, Publisher: Nature Publishing Group, pp. 1–11.
- Doray, Mathieu et al. (2018), « The PELGAS survey: ship-based integrated monitoring of the Bay of Biscay pelagic ecosystem », *in: Progress in Oceanography* 166, pp. 15–29.
- Ducharme-Barth, Nicholas D. et al. (2022), « Impacts of fisheries-dependent spatial sampling patterns on catch-per-unit-effort standardization: A simulation study and fishery application », *in: Fisheries Research* 246, p. 106169, ISSN: 0165-7836.
- Eales, James and James E. Wilen (1986), « An examination of fishing location choice in the pink shrimp fishery », *in: Marine Resource Economics* 2.4, Publisher: Crane Russak and Company, Inc., pp. 331–351.
- EC (July 23, 1987), *Council Regulation (EEC) No 2241/87 of 23 July 1987 establishing certain control measures for fishing activities*.
- (1993), *Council Regulation (EEC) No 2847/93 of 12 October 1993 establishing a control system applicable to the common fisheries policy*.
- Erisman, Brad E. et al. (2020), « Balancing conservation and utilization in spawning aggregation fisheries: a trade-off analysis of an overexploited marine fish », *in: ICES Journal of Marine Science* 77.1, Publisher: Oxford University Press, pp. 148–161.
- Etienne, Marie-Pierre and Pierre Gloaguen (2022), « Trajectory Reconstruction and Behavior Identification Using Geolocation Data », *in: Statistical Approaches for Hidden Variables in Ecology*, John Wiley and Sons, Ltd, chap. 1, pp. 1–25, ISBN: 9781119902799.
- Farley, Scott S. et al. (2018), « Situating ecology as a big-data science: Current advances, challenges, and solutions », *in: BioScience* 68.8, pp. 563–576.

- 
- Feldman, Mariano J et al. (2021), « Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review », *in: PLoS One* 16.3, e0234587.
- Ferraris, Jocelyne (2002), *Fishing fleet profiling methodology*, 423, Food and Agriculture Org.
- Finley, Andrew O., Sudipto Banerjee, and Bruce D. Cook (2014), « Bayesian hierarchical models for spatially misaligned data in R », *in: Methods in Ecology and Evolution* 5.6, pp. 514–523.
- Fithian, William et al. (2015), « Bias correction in species distribution models: pooling survey and collection data for multiple species », *in: Methods in Ecology and Evolution* 6.4, Publisher: Wiley Online Library, pp. 424–438.
- Fletcher, Robert J. et al. (2019), « A practical guide for combining data to model species distributions », *in: Ecology* 100.6, e02710, ISSN: 1939-9170.
- Fock, Heino O. (2008), « Fisheries in the context of marine spatial planning: defining principal areas for fisheries in the German EEZ », *in: Marine Policy* 32.4, pp. 728–739.
- Fournier, David A et al. (2012), « AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models », *in: Optimization Methods and Software* 27.2, pp. 233–249.
- Fox, Clive J. et al. (2008), « Mapping the spawning grounds of North Sea cod (*Gadus morhua*) by direct and indirect means », *in: Proceedings of the Royal Society B: Biological Sciences* 275.1642, pp. 1543–1548.
- Francis, RIC Chris (2014), « Replacing the multinomial in stock assessment models: A first step », *in: Fisheries Research* 151, pp. 70–84.
- (2017), « Revisiting data weighting in fisheries stock assessment models », *in: Fisheries Research* 192, pp. 5–15.
- Garrigues, Sébastien, Denis Allard, and Frédéric Baret (2008), « Modeling temporal changes in surface spatial heterogeneity over an agricultural site », *in: Remote Sensing of Environment* 112.2, pp. 588–602.
- Gascuel, D. (2015), *Dynamique des populations et gestion des stocks halieutiques*, Rennes: Agrocampus Ouest / Campus numérique ENVAM éd., p. 116.
- Gelfand, Alan E. (2010), « Misaligned Spatial Data; The Change of Support Problem », *in: Handbook of spatial statistics* 29, pp. 495–515.

- 
- Gelfand, Alan E. and Shinichiro Shirota (2019), « Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data », *in: Ecological Monographs* 89.3, e01372.
- Gelfand, Alan E. et al. (2010), *Handbook of spatial statistics*, CRC press.
- Gelfand, Sahu, and Holland (2012), « On the effect of preferential sampling in spatial prediction », *in: Environmetrics* 23.7, pp. 565–578.
- Gelfand, Zhu, and Carlin (2001), « On the change of support problem for spatio-temporal data », *in: Biostatistics* 2.1, pp. 31–45.
- Gerritsen, Hans and Colm Lordan (2011), « 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, pp. 245–252.
- Getis, A. and JK Ord (1992), « The analysis of spatial association by use of distance statistics », *in: Geographical Analysis*.
- Gilbert, Neil A. et al. (2021), « Integrating harvest and camera trap data in species distribution models », *in: Biological Conservation* 258, p. 109147.
- Gimenez, Olivier et al. (2014), « Statistical ecology comes of age », *in: Biology Letters* 10.12, Publisher: Royal Society, p. 20140698.
- Gimenez, Olivier et al. (2022), « Studying Species Demography and Distribution in Natural Conditions: Hidden Markov Models », *in: Statistical Approaches for Hidden Variables in Ecology*, John Wiley and Sons, Ltd, chap. 3, pp. 47–67, ISBN: 9781119902799.
- Girardin, Raphaël et al. (2017), « Thirty years of fleet dynamics modelling using discrete-choice models: What have we learned? », *in: Fish and Fisheries* 18.4, pp. 638–655, ISSN: 1467-2979.
- Gloaguen, Pierre (2015), « Modélisation mécaniste et stochastique des trajectoires pour l’halieutique », PhD thesis, Nantes, 202 pp.
- Gloaguen, Pierre, Marie-Pierre Etienne, and Sylvain Le Corff (2018), « Stochastic differential equation based on a multimodal potential to model movement data in ecology », *in: Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.3, pp. 599–619.
- Gotway, Carol A. and Linda J. Young (2002), « Combining incompatible spatial data », *in: Journal of the American Statistical Association* 97.458, pp. 632–648.
- (2007), « A geostatistical approach to linking geographically aggregated data from different sources », *in: Journal of Computational and Graphical Statistics* 16.1, pp. 115–135.

- 
- Grémillet, David, Damien Chevallier, and Christophe Guinet (2022), « Big data approaches to the spatial ecology and conservation of marine megafauna », *in: ICES Journal of Marine Science*.
- Grüss, Arnaud et al. (2014), « Conservation and fisheries effects of spawning aggregation marine protected areas: what we know, where we should go, and what we need to get there », *in: ICES Journal of Marine Science* 71.7, Publisher: Oxford University Press, pp. 1515–1534.
- Grüss, Arnaud et al. (2019), « Protecting juveniles, spawners or both: A practical statistical modelling approach for the design of marine protected areas », *in: Journal of Applied Ecology* 56.10, pp. 2328–2339, ISSN: 1365-2664.
- Grüss, Arnaud et al. (2020), « Spatio-temporal analyses of marine predator diets from data-rich and data-limited systems », *in: Fish and Fisheries* 21.4, pp. 718–739, ISSN: 1467-2979.
- Guisan, Antoine and Niklaus E. Zimmermann (2000), « Predictive habitat distribution models in ecology », *in: Ecological modelling* 135.2, Publisher: Elsevier, pp. 147–186.
- Hampton, Stephanie E. et al. (2013), « Big data and the future of ecology », *in: Frontiers in Ecology and the Environment* 11.3, pp. 156–162.
- Haynie, Alan C., Robert L. Hicks, and Kurt E. Schnier (2009), « Common property, information, and cooperation: Commercial fishing in the Bering Sea », *in: Ecological Economics*, Special Section: Analyzing the global human appropriation of net primary production - processes, trajectories, implications 69.2, pp. 406–413, ISSN: 0921-8009.
- Hefley, Trevor J, Brian M Brost, and Mevin B Hooten (2017), « Bias correction of bounded location errors in presence-only data », *in: Methods in Ecology and Evolution* 8.11, pp. 1566–1573.
- Hiddink, Jan G. et al. (2006), « Cumulative impacts of seabed trawl disturbance on benthic biomass, production, and species richness in different habitats », *in: Canadian journal of fisheries and aquatic sciences* 63.4, pp. 721–736.
- Hilborn, R. and C. J. Walters, eds. (1992), *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*, Springer US, ISBN: 978-0-412-02271-5.
- Hintzen, Niels T. (2021), « Zooming into small-scale fishing patterns: The use of vessel monitoring by satellite in fisheries science », PhD thesis, Wageningen University.
- Hintzen, Niels T. et al. (2021), « Quantifying habitat preference of bottom trawling gear », *in: ICES Journal of Marine Science* 78.1, pp. 172–184.

- 
- Hintzen et al. (2012), « VMStools: Open-source software for the processing, analysis and visualisation of fisheries logbook and VMS data », in: *Fisheries Research* 115, Publisher: Elsevier, pp. 31–43.
- Houise, C. and A. Forest (1993), *Etude de la population du Merlan (*Merlangius merlangius L.*) du Golfe de Gascogne*, Ifremer.
- Hovgård, Holger and Hans Lassen (2008), *Manual on estimation of selectivity for gillnet and longline gears in abundance surveys*, vol. 397, Food and Agriculture Org.
- Hulson, Peter-John F. et al. (2013), « Spatial modeling of Bering Sea walleye pollock with integrated age-structured assessment models in a changing environment », in: *Canadian Journal of Fisheries and Aquatic Sciences* 70.9, Publisher: NRC Research Press, pp. 1402–1416.
- Hussey, Nigel E et al. (2015), « Aquatic animal telemetry: a panoramic window into the underwater world », in: *Science* 348.6240, p. 1255642.
- ICES (2005), *Report of the Workshop on Survey Design and Data Analysis (WKSAD)*, Sète, France.
- (2013), *Report of the Working Group on the Assessment of Southern Shelf Stocks of Hake, Monk and Megrime (WGHMM)*, Copenhagen, Denmark, p. 727.
- (2015), *Manual for the International Bottom Trawl Surveys*. Series of ICES Survey Protocols.
- (2017), *Report of the Working Group on Commercial Catches (WGCATCH)*, Oostende, Belgium, p. 141.
- (2018a), *Report of the Working Group on Beam Trawl Surveys (WGBEAM)*, Galway, Ireland, p. 121.
- (2018b), *Sole (*Solea solea*) in divisions 8.a–b (northern and central Bay of Biscay)*, Advice, p. 8.
- (2019a), *Sole (*Solea solea*) in divisions 8.a–b (northern and central Bay of Biscay)*, Advice, p. 8.
- (2019b), *Whiting (*Merlangius merlangus*) in Subarea 8 and Division 9.a (Bay of Biscay and Atlantic Iberian waters)*.
- (2020a), *International Bottom Trawl Survey Working Group (IBTSWG)*, ICES Scientific Reports, Publisher: ICES, p. 197.
- (2020b), *Working Group for the Bay of Biscay and the Iberian Waters Ecoregion (WGBIE)*, ICES Scientific Reports, Publisher: ICES, p. 845.

- 
- (2020c), *Working Group on Cephalopod Fisheries and Life History (WGCEPH)*, Publisher: ICES.
- Isaac, Nick JB et al. (2020), « Data integration for large-scale models of species distributions », *in: Trends in ecology and evolution* 35.1, pp. 56–67.
- Janßen, Holger et al. (2018), « Integration of fisheries into marine spatial planning: Quo vadis? », *in: Estuarine, Coastal and Shelf Science* 201, pp. 105–113.
- Jennings, Simon and Janette Lee (2012), « Defining fishing grounds with vessel monitoring system data », *in: ICES Journal of Marine Science* 69.1, pp. 51–63, ISSN: 1054-3139.
- Jullum, Martin (2020), « Investigating mesh-based approximation methods for the normalization constant in the log Gaussian Cox process likelihood », *in: Stat* 9.1, e285.
- Kai, Mikihiko et al. (2017), « Spatiotemporal variation in size-structured populations using fishery data: an application to shortfin mako (*Isurus oxyrinchus*) in the Pacific Ocean », *in: Canadian Journal of Fisheries and Aquatic Sciences* 74.11, pp. 1765–1780.
- Karcher et al. (2016), « Quantifying and mitigating the effect of preferential sampling on phylodynamic inference », *in: PLoS computational biology* 12.3, e1004789.
- Kays, Roland et al. (2015), « Terrestrial animal tracking as an eye on life and planet », *in: Science* 348.6240, aaa2478.
- Kim, Yongku and L. Mark Berliner (2016), « Change of spatiotemporal scale in dynamic models », *in: Computational Statistics and Data Analysis* 101, pp. 80–92.
- Kraainski, Elias et al. (2018), *Advanced spatial modeling with stochastic partial differential equations using R and INLA*, Chapman and Hall/CRC.
- Kristensen, Kasper et al. (2014), « Estimating spatio-temporal dynamics of size-structured populations », *in: Canadian Journal of Fisheries and Aquatic Sciences* 71.2, ed. by Josef Michael Jech, pp. 326–336, ISSN: 0706-652X, 1205-7533.
- Kristensen, Kasper et al. (2016), « TMB: Automatic Differentiation and Laplace Approximation », *in: Journal of Statistical Software* 70.1, pp. 1–21, ISSN: 1548-7660.
- Lajaunie, C and H Wackernagel (2000), *Geostatistical approaches to change of support problems-theoretical framework*.
- Lambert, C. et al. (2017), « Habitat modelling predictions highlight seasonal relevance of Marine Protected Areas for marine megafauna », *in: Deep Sea Research Part II: Topical Studies in Oceanography* 141, pp. 262–274.
- Laptikhovsky, Vladimir et al. (2022), « Spatial and temporal variability of spawning and nursery grounds of *Loligo forbesii* and *Loligo vulgaris* squids in ecoregions of Celtic

- 
- Seas and Greater North Sea », *in: ICES Journal of Marine Science* 79.6, pp. 1918–1930.
- Lauret, Valentin et al. (2021), « Using single visits into integrated occupancy models to make the most of existing monitoring programs », *in: Ecology* 102.12, e03535.
- Le Pape, Olivier et al. (2003), « Quantitative description of habitat suitability for the juvenile common sole (*Solea solea*, L.) in the Bay of Biscay (France) and the contribution of different habitats to the adult population », *in: Journal of Sea Research*, Proceedings of the Fifth International Symposium on Flatfish Ecology, Part I 50.2, pp. 139–149, ISSN: 1385-1101.
- Leach, Clinton B. et al. (2021), « Recursive Bayesian computation facilitates adaptive optimal design in ecological studies », *in: Ecology*, e03573.
- Lecomte, Jean-Baptiste et al. (2013), « Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume », *in: Methods in Ecology and Evolution* 4.12, pp. 1159–1166.
- Lehuta, S and Y Verma (2022), « Contrasting impacts of the landing obligation at fleet scale: impact assessment of mitigation scenarios in the Eastern English Channel », *in: ICES Journal of Marine Science*.
- Levin, Phillip S et al. (2014), « Guidance for implementation of integrated ecosystem assessments: a US perspective », *in: ICES Journal of Marine Science* 71.5, pp. 1198–1204.
- Lindgren, Finn, Haavard Rue, and Johan Lindström (2011), « 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, Publisher: Wiley Online Library, pp. 423–498.
- Lindgren, Finn and Håvard Rue (2015), « Bayesian Spatial Modelling with R-INLA », *in: Journal of Statistical Software* 63.1, pp. 1–25, ISSN: 1548-7660.
- Loiselle, Bette A. et al. (2003), « Avoiding pitfalls of using species distribution models in conservation planning », *in: Conservation biology* 17.6, pp. 1591–1600.
- Lokers, Rob et al. (2016), « Analysis of Big Data technologies for use in agro-environmental science », *in: Environmental Modelling and Software* 84, pp. 494–504.
- Lorenz, Edward N (1956), *Empirical orthogonal functions and statistical weather prediction*, vol. 1, Massachusetts Institute of Technology, Department of Meteorology Cambridge.

- 
- MacPhail, Victoria J and Sheila R Colla (2020), « Power of the people: A review of citizen science programs for conservation », *in: Biological Conservation* 249, p. 108739.
- Mahévas, Stéphanie and Dominique Pelletier (2004), « ISIS-Fish, a generic and spatially explicit simulation tool for evaluating the impact of management measures on fisheries dynamics », *in: Ecological Modelling* 171.1, pp. 65–84, ISSN: 0304-3800.
- Martínez-Minaya, Joaquín et al. (2018), « Species distribution modeling: a statistical review with focus in spatio-temporal issues », *in: Stochastic environmental research and risk assessment* 32.11, Publisher: Springer, pp. 3227–3244.
- Matheron, G (1985), « Change of support for diffusion-type random functions », *in: Journal of the International Association for Mathematical Geology* 17.2, pp. 137–165.
- Maureaud, Aurore et al. (2020), « Are we ready to track climate-driven shifts in marine species across international boundaries? - A global survey of scientific bottom trawl data », *in: Global Change Biology*, gcb.15404, ISSN: 1354-1013, 1365-2486.
- Mérillet, Laurène et al. (2022), « Effects of life-history traits and network topological characteristics on the robustness of marine food webs », *in: Global Ecology and Conservation* 34, e02048.
- Michelot, Théo et al. (2019), « The Langevin diffusion as a continuous-time model of animal movement and habitat selection », *in: Methods in ecology and evolution* 10.11, Publisher: Wiley Online Library, pp. 1894–1907.
- Milisenda, Giacomo et al. (2021), « Identifying Persistent Hot Spot Areas of Undersized Fish and Crustaceans in Southern European Waters: Implication for Fishery Management Under the Discard Ban Regulation », *in: Frontiers in Marine Science* 8, p. 60, ISSN: 2296-7745.
- Miller, David AW et al. (2019), « The recent past and promising future for data integration methods to estimate species' distributions », *in: Methods in Ecology and Evolution* 10.1, pp. 22–37.
- Möllmann, Christian et al. (2014), « Implementing ecosystem-based fisheries management: from single-species to integrated ecosystem assessment and advice for Baltic Sea fish stocks », *in: ICES Journal of Marine Science* 71.5, pp. 1187–1197.
- Morel, Maryan et al. (2020), « ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection », *in: Biostatistics* 21.4, pp. 758–774.
- Moreno, A. et al. (2002), « Biological variation of *Loligo vulgaris* (Cephalopoda: Loliginidae) in the eastern Atlantic and Mediterranean », *in: Bulletin of Marine Science* 71.1, pp. 515–534.

- 
- Moriarty, Meadhbh et al. (2020), « Combining fisheries surveys to inform marine species distribution modelling », *in: ICES Journal of Marine Science* 77.2, Publisher: Oxford University Press, pp. 539–552.
- Mugglin, Andrew S., Bradley P. Carlin, and Alan E. Gelfand (2000), « Fully Model-Based Approaches for Spatially Misaligned Data », *in: Journal of the American Statistical Association* 95.451, pp. 877–887, ISSN: 0162-1459.
- Murawski, Steven A. et al. (2005), « Effort distribution and catch patterns adjacent to temperate MPAs », *in: ICES Journal of Marine Science* 62.6, pp. 1150–1167, ISSN: 1095-9289, 1054-3139.
- Murray, Lee G. et al. (2013), « 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, pp. 1330–1340.
- Nathan, Ran et al. (2022), « Big-data approaches lead to an increased understanding of the ecology of animal movement », *in: Science* 375.6582, eabg1780.
- Nielsen (2015), *Methods for integrated use of fisheries research survey information in understanding marine fish population ecology and better management advice: improving methods for evaluation of research survey information under consideration of survey fish detection and catch efficiency*, Wageningen University.
- Nielsen, J. Rasmus et al. (2018), « Integrated ecological–economic fisheries models—Evaluation, review and challenges for implementation », *in: Fish and Fisheries* 19.1, pp. 1–29.
- Okamura, Hiroshi et al. (2018), « Target-based catch-per-unit-effort standardization in multispecies fisheries », *in: Canadian Journal of Fisheries and Aquatic Sciences* 75.3, pp. 452–463, ISSN: 0706-652X, 1205-7533.
- Olmos, Maxime et al. (n.d.), « Resolving population processes in spatiotemporal population dynamics models: eastern Bering Sea snow crab as a case study », *in: ()*.
- Ord, J. K. and Arthur Getis (1995), « Local Spatial Autocorrelation Statistics: Distributional Issues and an Application », *in: Geographical Analysis* 27.4, pp. 286–306, ISSN: 1538-4632.
- Paci, Lucia et al. (2020), « Spatial hedonic modelling adjusted for preferential sampling », *in: Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183.1, pp. 169–192.
- Pacifi, Krishna et al. (2017), « Integrating multiple data sources in species distribution modeling: a framework for data fusion », *in: Ecology* 98.3, pp. 840–850.

- 
- Pacifci, Krishna et al. (2019), « Resolving misaligned spatial data with integrated species distribution models », *in: Ecology* 100.6, e02709.
- Papaïx, Julien et al. (2022), « Inferring Mechanistic Models in Spatial Ecology Using a Mechanistic-Statistical Approach », *in: Statistical Approaches for Hidden Variables in Ecology*, John Wiley and Sons, Ltd, chap. 4, pp. 69–95, ISBN: 9781119902799.
- Paradinas, I et al. (2015), « Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas », *in: Marine Ecology Progress Series* 528, pp. 245–255, ISSN: 0171-8630, 1616-1599.
- Parent, Eric and Etienne Rivot (2012), *Introduction to hierarchical Bayesian modeling for ecological data*, CRC Press.
- Parker, Ryan J., Brian J. Reich, and Stephan R. Sain (2015), « A multiresolution approach to estimating the value added by regional climate models », *in: Journal of Climate* 28.22, pp. 8873–8887.
- Pati, Reich, and Dunson (2011), « Bayesian geostatistical modelling with informative sampling locations », *in: Biometrika* 98.1, pp. 35–48.
- Pecquerie, Laure, Pierre Petitgas, and Sebastiaan ALM Kooijman (2009), « Modeling fish growth and reproduction in the context of the Dynamic Energy Budget theory to predict environmental impact on anchovy spawning duration », *in: Journal of Sea Research* 62.2-3, pp. 93–105.
- Pedersen, Søren Anker, Heino O. Fock, and Anne F. Sell (July 1, 2009), « Mapping fisheries in the German exclusive economic zone with special reference to offshore Natura 2000 sites », *in: Marine Policy* 33.4, pp. 571–590, ISSN: 0308-597X.
- Pelletier, Dominique and Jocelyne Ferraris (2000), « A multivariate approach for defining fishing tactics from commercial catch and effort data », *in: Canadian Journal of Fisheries and Aquatic Sciences* 57.1, Publisher: NRC Research Press, pp. 51–65.
- Pendleton, Daniel E. et al. (2022), « Decadal-scale phenology and seasonal climate drivers of migratory baleen whales in a rapidly warming marine ecosystem », *in: Global Change Biology*, Publisher: Wiley Online Library.
- Pennino, M. Grazia et al. (Oct. 1, 2013), « Modeling sensitive elasmobranch habitats », *in: Journal of Sea Research*, Main results from the XVII Iberian Symposium of Marine Biology Studies 83, pp. 209–218, ISSN: 1385-1101.
- Pennino, Maria Grazia et al. (2016), « Fishery-dependent and-independent data lead to consistent estimations of essential habitats », *in: ICES Journal of Marine Science* 73.9, Publisher: Oxford University Press, pp. 2302–2310.

- 
- Pennino et al. (2019), « Accounting for preferential sampling in species distribution models », *in: Ecology and evolution* 9.1, pp. 653–663.
- Peterson, Cassidy D et al. (2021), « Reconciling conflicting survey indices of abundance prior to stock assessment », *in: ICES Journal of Marine Science* 78.9, pp. 3101–3120, ISSN: 1054-3139.
- Petitgas, Pierre (1997), « Sole egg distributions in space and time characterised by a geostatistical model and its estimation variance », *in: ICES Journal of Marine Science* 54.2, pp. 213–225.
- Petitgas, Pierre et al. (2018), « Indicator-based geostatistical models for mapping fish survey data », *in: Mathematical Geosciences* 50.2, pp. 187–208.
- Petitgas, Pierre et al. (2020), « Analysing Temporal Variability in Spatial Distributions Using Min–Max Autocorrelation Factors: Sardine Eggs in the Bay of Biscay », *in: Mathematical Geosciences* 52.3, pp. 337–354.
- Peyrard, Nathalie, Stéphane Robin, and Olivier Gimenez (2022), « Front matter », *in: Statistical Approaches for Hidden Variables in Ecology*, John Wiley and Sons, Ltd, chap. Introduction, pp. i–xvii, ISBN: 9781119902799.
- Pinto, Cecilia et al. (2019), « Combining multiple data sets to unravel the spatiotemporal dynamics of a data-limited fish stock », *in: Canadian Journal of Fisheries and Aquatic Sciences* 76.8, Publisher: NRC Research Press, pp. 1338–1349.
- Planque, Benjamin et al. (2011), « Understanding what controls the spatial distribution of fish populations using a multi-model approach », *in: Fisheries Oceanography* 20.1, pp. 1–17, ISSN: 1365-2419.
- Plet-Hansen, Kristian S., François Bastardie, and Clara Ulrich (2020), « The value of commercial fish size distribution recorded at haul by haul compared to trip by trip », *in: ICES Journal of Marine Science* 77.7, pp. 2729–2740.
- Poggi, Sylvain et al. (Jan. 1, 2021), « Chapter Seven - How can models foster the transition towards future agricultural landscapes? », *in: Advances in Ecological Research*, ed. by David A. Bohan and Adam J. Vanbergen, vol. 64, The Future of Agricultural Landscapes, Part II, Academic Press, pp. 305–368.
- Pollock, Laura J. et al. (2014), « Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM) », *in: Methods in Ecology and Evolution* 5.5, pp. 397–406, ISSN: 2041-210X.
- Poos, Jan-Jaap and Adriaan D Rijnsdorp (2007), « An "experiment" on effort allocation of fishing vessels: the role of interference competition and area specialization », *in:*

- 
- Canadian Journal of Fisheries and Aquatic Sciences* 64.2, pp. 304–313, ISSN: 0706-652X, 1205-7533.
- Poulard, Jean-Charles (2001), « Distribution of hake (*Merluccius merluccius*, Linnaeus, 1758) in the Bay of Biscay and the Celtic sea from the analysis of French commercial data », *in: Fisheries Research* 50.1-2, pp. 173–187.
- Punt, André E. (2017), « Some insights into data weighting in integrated stock assessments », *in: Fisheries Research*, Data conflict and weighting, likelihood functions, and process error 192, pp. 52–65, ISSN: 0165-7836.
- (2019), « Spatial stock assessment methods: A viewpoint on current issues and assumptions », *in: Fisheries Research* 213, pp. 132–143, ISSN: 0165-7836.
- Punt, André E. et al. (2020), « Essential features of the next-generation integrated fisheries stock assessment package: A perspective », *in: Fisheries Research* 229, p. 105617, ISSN: 0165-7836.
- Qiu, Wanfei and Peter JS Jones (2013), « The emerging policy landscape for marine spatial planning in Europe », *in: Marine Policy* 39, pp. 182–190.
- Quemper, Florian (2021), *Modélisation de la distribution spatiale de la sardine du Golfe de Gascogne (*Sardina pilchardus*) par intégration de données commerciales et scientifiques: enjeux et limites*. Rennes: Institut Agro.
- Ramzi, Azeddine et al. (2001), « Modelling and numerical simulations of larval migration of the sole (*Solea solea* (L.)) of the Bay of Biscay. Part 2: numerical simulations », *in: Oceanologica acta* 24.2, pp. 113–124.
- Regimbart, Amélie, Jérôme Guittot, and Olivier Le Pape (2018), *Zones fonctionnelles pour les ressources halieutiques dans les eaux sous souveraineté française. Deuxième partie : Inventaire. Rapport d'étude. Les publications du Pôle halieutique A.* 46, Rennes: Pôle halieutique AGROCAMPUZ OUEST.
- Reich, Brian J., Howard H. Chang, and Kristen M. Foley (2014), « A spectral method for spatial downscaling », *in: Biometrics* 70.4, pp. 932–942.
- Renner, Ian W., Julie Louvrier, and Olivier Gimenez (2019), « 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, pp. 2118–2128.
- Renner, Ian W. et al. (2015), « Point process models for presence-only analysis », *in: Methods in Ecology and Evolution* 6.4, pp. 366–379, ISSN: 2041-210X.

- 
- Rivoirard, Jacques et al. (2008), *Geostatistics for estimating fish abundance*, John Wiley and Sons.
- Rivot, Etienne et al. (2004), « A Bayesian state-space modelling framework for fitting a salmon stage-structured population dynamic model to multiple time series of field data », in: *Ecological Modelling* 179.4, pp. 463–485.
- Rochet, Marie-Joëlle and Verena M Trenkel (2005), « Factors for the variability of discards: assumptions and field evidence », in: *Canadian Journal of Fisheries and Aquatic Sciences* 62.1, pp. 224–235.
- Rochette, S. et al. (2010), « Effect of nursery habitat degradation on flatfish population: Application to Solea solea in the Eastern Channel (Western Europe) », in: *Journal of sea Research* 64.1, Publisher: Elsevier, pp. 34–44.
- Rue, Havard and Leonhard Held (2005), *Gaussian Markov random fields: theory and applications*, CRC press.
- Rue, Håvard, Sara Martino, and Nicolas Chopin (2009), « Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations », in: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2, pp. 319–392.
- Rufener, Marie-Christine (2020), « Integrating commercial fisheries and scientific survey data: Advances, new tools and applications to model the fish and fishery dynamics », PhD thesis, Denmark, DTU Aqua: National Institute of Aquatic Resources, 209 pp.
- Rufener et al. (2021), « Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species », in: *Ecological Applications*, e02453.
- Sadovy, Yvonne and Michael Domeier (2005), « Are aggregation-fisheries sustainable? Reef fish fisheries as a case study », in: *Coral reefs* 24.2, pp. 254–262.
- Salas, Silvia and Daniel Gaertner (2004), « The behavioural dynamics of fishers: management implications », in: *Fish and Fisheries* 5.2, pp. 153–167, ISSN: 1467-2979.
- Saunders, Sarah P. et al. (2019), « Disentangling data discrepancies with integrated population models », in: *Ecology* 100.6, e02714, ISSN: 1939-9170.
- Schaub, Michael and Fitsum Abadi (2011), « Integrated population models: a novel analysis framework for deeper insights into population dynamics », in: *Journal of Ornithology* 152.1, Publisher: Springer, pp. 227–237.

- 
- Schmittner, Rolland A. (1999), « Essential fish habitat: opportunities and challenges for the next millennium », *in: American Fisheries Society Symposium*, vol. 22, Issue: 3, p. 10.
- Seitz, Rochelle D et al. (2014), « Ecological value of coastal habitats for commercially and ecologically important species », *in: ICES Journal of Marine Science* 71.3, pp. 648–665.
- Shaddick, Gavin and James V. Zidek (2014), « A case study in preferential sampling: Long term monitoring of air pollution in the UK », *in: Spatial statistics* 9, pp. 51–65.
- Shaddick, Gavin, James V. Zidek, and Yi Liu (2016), « Mitigating the effects of preferentially selected monitoring sites for environmental policy and health risk analysis », *in: Spatial and spatio-temporal epidemiology* 18, pp. 44–52.
- Silvano, Renato AM et al. (2006), « When does this fish spawn? Fishermen's local knowledge of migration and reproduction of Brazilian coastal fishes », *in: Environmental Biology of fishes* 76.2, pp. 371–386.
- Simpson, Daniel et al. (2016), « Going off grid: Computationally efficient inference for log-Gaussian Cox processes », *in: Biometrika* 103.1, pp. 49–70.
- Skaug, Hans J. and David A. Fournier (2006), « Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models », *in: Computational Statistics and Data Analysis* 51.2, pp. 699–709.
- Smith, Martin D. et al. (2010), « Political economy of marine reserves: Understanding the role of opportunity costs », *in: Proceedings of the National Academy of Sciences* 107.43, Publisher: National Acad Sciences, pp. 18300–18305.
- Stelzenmüller, Vanessa, Stuart I. Rogers, and Craig M. Mills (2008), « Spatio-temporal patterns of fishing pressure on UK marine landscapes, and their implications for spatial planning and management », *in: ICES Journal of Marine Science* 65.6, pp. 1081–1091, ISSN: 1095-9289, 1054-3139.
- Stephens, Andi and Alec MacCall (2004), « A multispecies approach to subsetting logbook data for purposes of estimating CPUE », *in: Fisheries Research* 70.2, pp. 299–310, ISSN: 01657836.
- Stock, Brian C et al. (2019), « The utility of spatial model-based estimators of unobserved bycatch », *in: ICES Journal of Marine Science* 76.1, pp. 255–267, ISSN: 1054-3139.
- Sullivan, Brian L et al. (2014), « The eBird enterprise: An integrated approach to development and application of citizen science », *in: Biological conservation* 169, pp. 31–40.

- 
- Suuronen, Petri, Pekka Jounela, and Vesa Tschernij (2010), « Fishermen responses on marine protected areas in the Baltic cod fishery », *in: Marine Policy* 34.2, pp. 237–243, ISSN: 0308-597X.
- Thorson, James T. (2015), « Spatio-temporal variation in fish condition is not consistently explained by density, temperature, or season for California Current groundfishes », *in: Marine Ecology Progress Series* 526, pp. 101–112.
- (2018), « 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, Publisher: NRC Research Press, pp. 1369–1382.
- (2019), « Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments », *in: Fisheries Research* 210, pp. 143–161, ISSN: 0165-7836.
- Thorson, James T., Grant Adams, and Kirstin Holsman (2019), « Spatio-temporal models of intermediate complexity for ecosystem assessments: A new tool for spatial fisheries management », *in: Fish and Fisheries* 20.6, pp. 1083–1099, ISSN: 1467-2979.
- Thorson, James T., Lorenzo Ciannelli, and Michael A. Litzow (2020), « Defining indices of ecosystem variability using biological samples of fish communities: A generalization of empirical orthogonal functions », *in: Progress in Oceanography* 181, p. 102244, ISSN: 0079-6611.
- Thorson, James T., Jason Jannot, and Kayleigh Somers (2017), « Using spatio-temporal models of population growth and movement to monitor overlap between human impacts and fish populations », *in: Journal of Applied Ecology* 54.2, pp. 577–587, ISSN: 1365-2664.
- Thorson, James T. and Kasper Kristensen (2016), « Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples », *in: Fisheries Research* 175, Publisher: Elsevier, pp. 66–74.
- Thorson, James T. and Eric J. Ward (2014), « Accounting for vessel effects when standardizing catch rates from cooperative surveys », *in: Fisheries Research* 155, pp. 168–176, ISSN: 0165-7836.
- Thorson, James T. et al. (2015a), « Spatial delay-difference models for estimating spatiotemporal variation in juvenile production and population abundance », *in: Canadian journal of fisheries and aquatic sciences* 72.12, pp. 1897–1915.

- 
- Thorson, James T. et al. (2015b), « Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range », *in: Methods in Ecology and Evolution* 6.6, Publisher: Wiley Online Library, pp. 627–637.
- Thorson, James T. et al. (2016a), « Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data », *in: Canadian Journal of Fisheries and Aquatic Sciences* 74.11, pp. 1794–1807.
- Thorson, James T. et al. (2016b), « Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring », *in: Global Ecology and Biogeography* 25.9, pp. 1144–1158, ISSN: 1466-8238.
- Thorson, James T. et al. (2017), « Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution », *in: Fisheries Research* 192, pp. 84–93.
- Thorson, James T. et al. (2020), « Seasonal and interannual variation in spatio-temporal models for index standardization and phenology studies », *in: ICES Journal of Marine Science* 77.5, pp. 1879–1892.
- Thorson, James T. et al. (2021a), « The surprising sensitivity of index scale to delta-model assumptions: Recommendations for model-based index standardization », *in: Fisheries Research* 233, p. 105745, ISSN: 0165-7836.
- Thorson et al. (2021b), « Estimating fine-scale movement rates and habitat preferences using multiple data sources », *in: Fish and Fisheries* 22.6, pp. 1359–1376.
- Tidd, A and Steve Warnes (2006), « Species distributions from English Celtic Sea ground-fish surveys, 1992–2003 », *in:*
- Tidd, Alex N. et al. (2015), « Fishing for Space: Fine-Scale Multi-Sector Maritime Activities Influence Fisher Location Choice », *in: PLOS ONE* 10.1, e0116335, ISSN: 1932-6203.
- Trenkel, Verena M. et al. (2013), « Testing CPUE-derived spatial occupancy as an indicator for stock abundance: application to deep-sea stocks », *in: Aquatic living resources* 26.4, Publisher: EDP Sciences, pp. 319–332.
- Trivedi, Mandar R et al. (2008), « Spatial scale affects bioclimate model projections of climate change impacts on mountain plants », *in: Global change biology* 14.5, pp. 1089–1103.
- Ulrich, Clara (2021), « Mixed fisheries, bycatch and discards in the Common Fishery Policy. Examples from the demersal fisheries in the North Sea », HDR, Université de Nantes, 119 pp.

- 
- Ulrich, Clara et al. (2013), « Variability and connectivity of plaice populations from the Eastern North Sea to the Western Baltic Sea, and implications for assessment and management », *in: Journal of Sea Research* 84, pp. 40–48.
- Vermard, Youen et al. (2008), « A dynamic model of the Bay of Biscay pelagic fleet simulating fishing trip choice: the response to the closure of the European anchovy (*Engraulis encrasicolus*) fishery in 2005 », *in: Canadian Journal of Fisheries and Aquatic Sciences* 65.11, pp. 2444–2453, ISSN: 0706-652X.
- Vermard, Youen et al. (2010), « Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models », *in: Ecological Modelling* 221.15, pp. 1757–1769.
- Vigneau, J (2009), *Common tool for raising and estimating properties of statistical estimates derived from the Data Collection Regulation (COST) - Final report*.
- Wakefield and Lyons (2010), « Spatial aggregation and the ecological fallacy », *in: Handbook of spatial statistics* 541, p. 558.
- Walker, Emily and Nicolas Bez (2010), « A pioneer validation of a state-space model of vessel trajectories (VMS) with observers' data », *in: Ecological Modelling* 221.17, pp. 2008–2017.
- Wang, Sheng-Ping and Mark N. Maunder (2017), « Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? », *in: Fisheries Research* 192, pp. 41–51.
- Warton, Renner, and Ramp (2013), « Model-based control of observer bias for the analysis of presence-only data in ecology », *in: PloS one* 8.11, e79168.
- Watson, Joe, James V. Zidek, and Gavin Shaddick (2019), « A general theory for preferential sampling in environmental networks », *in: The Annals of Applied Statistics* 13.4, pp. 2662–2700.
- Wikle, Christopher K. and L. Mark Berliner (2005), « Combining information across spatial scales », *in: Technometrics* 47.1, pp. 80–91.
- Wikle, Christopher K., Zammit-Mangion, and Cressie (2019), *Spatio-temporal Statistics with R*, CRC Press.
- Winker, Henning, Sven E. Kerwath, and Colin G. Attwood (2013), « Comparison of two approaches to standardize catch-per-unit-effort for targeting behaviour in a multispecies hand-line fishery », *in: Fisheries Research* 139, pp. 118–131, ISSN: 01657836.
- Witman, Jon D. and Kaustuv Roy (2009), *Marine Macroecology*, University of Chicago Press, 442 pp., ISBN: 978-0-226-90414-6.

- 
- Woillez, Mathieu et al. (2010), « Statistical monitoring of spatial patterns of environmental indices for integrated ecosystem assessment: application to the Bay of Biscay pelagic zone », *in: Progress in Oceanography* 87.1-4, pp. 83–93.
- Yan, Yuan et al. (2022), « Spatiotemporal modeling of bycatch data: methods and a practical guide through a case study in a Canadian Arctic fishery », *in: Canadian Journal of Fisheries and Aquatic Sciences* 79.1, pp. 148–158.
- Yochum, Noëlle, Richard M. Starr, and Dean E. Wendt (2011), « Utilizing fishermen knowledge and expertise: keys to success for collaborative fisheries research », *in: Fisheries* 36.12, pp. 593–605.
- Young, Linda J. and Carol A. Gotway (2007), « Linking spatial data from different sources: the effects of change of support », *in: Stochastic Environmental Research and Risk Assessment* 21.5, pp. 589–600.
- Zidek, James V., Gavin Shaddick, and Carolyn G. Taylor (2014), « Reducing estimation bias in adaptively changing monitoring networks with preferential site selection », *in: The Annals of Applied Statistics* 8.3, pp. 1640–1670.
- Zipkin, Elise F., Brian D. Inouye, and Steven R. Beissinger (2019), « Innovations in data integration for modeling populations », *in: Ecology* 100.6, e02713, ISSN: 1939-9170.
- Zipkin, Elise F. and Sarah P. Saunders (2018), « Synthesizing multiple data types for biological conservation using integrated population models », *in: Biological Conservation* 217, pp. 240–250, ISSN: 0006-3207.

---

## **Supplementary material**

# STATISTICAL TOOLS FOR SPATIAL AND SPATIO-TEMPORAL MODELING

---

## A.1 Properties of spatio-temporal covariance functions

### Non-negative and positive-definiteness

A function  $\{\mathcal{C}(u, v) : u, v \in D\}$  defined on  $D \times D$  is said to be **non-negative-definite**, if for any complex numbers  $\{a_i : i = 1, \dots, m\}$ , any  $\{u_i : i = 1, \dots, m\}$  in  $D$ , and any integer  $m$ , we have

$$\sum_{i=1}^m \sum_{j=1}^m a_i \bar{a}_j \mathcal{C}(u_i, u_j) \geq 0$$

where  $\bar{a}$  denotes the complex conjugate of  $a$ .

To be *valid*, a covariance function must be non-negative definite.

A function is **positive-definite** when the inequality below is strictly positive whenever  $(a_1, \dots, a_m)'$  is a nonzero vector.

For the following properties, we rewrite  $u$  as  $u = (s; t)$  and  $\mathcal{C}(u_i, u_j) \equiv \mathcal{C}((x_i, t_i), (x_j, t_j))$ .

In our case as the time is assumed discrete, the domain  $D$  is a subset of  $\mathbb{R}^2 \times [1, T]$  with  $T$  the number of time steps.

### Stationarity

There are 2 main kinds of stationarity: strong and second-order (or weak) stationarity.

-  $\delta(x, t)$  is strongly stationary when the two probability measures defining  $\delta(x; t)$  and  $\delta(x + h; t + \tau)$  are equivalent for all  $h \in \mathbb{R}^d$  and  $t \in \mathbb{R}$ .

-  $\delta(x, t)$  is second-order (or weakly) stationary when it has a constant expectation and a stationary covariance function.

---

### Isotropy

Spatial isotropy corresponds to  $\text{Cov}(\delta(x; t), \delta(s; r)) \equiv \mathcal{C}(\|x - s\|; t, r)$

### Separability

$\delta(x, t)$  has a separable spatio-temporal covariance function if:

$$\mathcal{C}((x, t), (s, r)) = \mathcal{C}^{(x)}(x, s) \cdot \mathcal{C}^{(t)}(t, r), s, x \in \mathbb{R}^d, t, r \in \mathbb{R}$$

where  $\mathcal{C}^{(x)}$  and  $\mathcal{C}^{(t)}$  are respectively spatial and temporal covariance functions.

When assuming spatial and temporal stationarity, the expression can be rewritten as  $\mathcal{C}(h; \tau) = \mathcal{C}^{(x)}(h) \cdot \mathcal{C}^{(t)}(\tau), h \in \mathbb{R}^d, \tau \in \mathbb{R}$

### Full symmetry

$\delta(x, t)$  is fully symmetric if for all  $(s, x) \in \mathbb{R}^d$  and  $(t, r) \in \mathbb{R}$ :

$$\text{Cov}(\delta(x, t), \delta(s, r)) = \text{Cov}(\delta(x, r), \delta(s, t))$$

In general, when a process is separate then it is fully symmetric.

### Intrinsic

$\delta(x, t)$  is intrinsic if for all  $(s, x) \in \mathbb{R}^d$ ,  $h \in \mathbb{R}$  and  $(t, r) \in \mathbb{R}$ ,  $\delta(x + h, t + \tau) + \delta(x, t)$  is second-ordered stationary (constant expectation and stationary covariance).

Equivalently, by introducing the spatio-temporal covariogram  $\text{Var}(Y(s; t) - Y(x; r)) \equiv 2\gamma(s, x; t, r)$  (with  $\gamma()$  the semivariogram) and its stationary version  $2\gamma(h, \tau); h \in \mathbb{R}$ ,  $\delta(x, t)$  is intrinsically stationary if its expectation is constant and its variogram is stationary.

The class of intrinsically stationary process contains the class of second-order stationary processes.

## A.2 Automatic differentiation

In general, there is no closed form for the marginal likelihood of most spatio-temporal models; this requires to maximize the likelihood numerically. One standard method to derive numerically complex functions is Automatic Differentiation (AD). AD consists in (1) decomposing a complex function into a set of elementary operations and (2) applying the chain rule on these elementary operations to compute the derivative of the function.

---

Higher order derivatives can be obtained by applying iteratively first order AD.

$$\text{Chain rule: } \frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}$$

For instance, as Fournier et al. (2012) let's introduce the simple linear regression model  $y_i = a + b \cdot x_i + \epsilon_i$  with  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  and  $i \in \llbracket 1, n \rrbracket$ . Estimating  $a$  and  $b$  through the method of least squares consists generally in minimizing the following formula:

$$RSS = \sum_{i=1}^n RSS_i = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2$$

Deriving the first order derivative of such equation relatively to  $a$  and  $b$  leads to:

$$\frac{\partial RSS}{\partial a} = \sum_{i=1}^n 2(y_i - (a + b \cdot x_i)) = 0$$

$$\frac{\partial RSS}{\partial b} = \sum_{i=1}^n 2 \cdot x_i (y_i - (a + b \cdot x_i)) = 0$$

Alternatively, AD consists in decomposing RSS into a set of unary operation so that:

$$u_1 = b \cdot x_i$$

$$u_2 = a + u_1$$

$$u_3 = y_i - u_2$$

$$u_4 = u_3^2$$

$$RSS_i = u_4$$

These equations can be used to reformulate the partial derivative of  $RSS$ :

$$\frac{\partial RSS}{\partial a} = \sum_{i=1}^n \frac{\partial RSS_i}{\partial a} = \sum_{i=1}^n \frac{\partial RSS_i}{\partial u_4} \frac{\partial u_4}{\partial u_3} \frac{\partial u_3}{\partial u_2} \frac{\partial u_2}{\partial a}$$

$$\frac{\partial RSS}{\partial b} = \sum_{i=1}^n \frac{\partial RSS_i}{\partial b} = \sum_{i=1}^n \frac{\partial RSS_i}{\partial u_4} \frac{\partial u_4}{\partial u_3} \frac{\partial u_3}{\partial u_2} \frac{\partial u_2}{\partial u_1} \frac{\partial u_1}{\partial b}$$

These are equivalent to the formulas obtained through the least square methods.

Here the model equation is relatively simple (it is a standard linear regression) and AD will not bring much in deriving the derivative of the function, but when the equations are more complex and depend on many parameters, AD can be of great interest.

---

There are two modes of AD: forward and reverse. The forward mode is the most intuitive way to perform the calculation. It consists in derivating the function from the independent variables ( $a$  and  $b$ ) to the output variable by simply applying the chain rule as in the previous example.

The reverse mode consists in (1) evaluating the function and storing all the intermediate values ('forward sweep') and (2) finding the derivatives of the output variable with respect to all intermediate variables through the chain rule ( $\frac{\partial \text{RSS}}{\partial u_1}$ ,  $\frac{\partial \text{RSS}}{\partial u_2}$ ,  $\frac{\partial \text{RSS}}{\partial u_3}$ ,  $\frac{\partial \text{RSS}}{\partial u_4}$  - 'reverse sweep'). Once done, the gradients are easily computed

$$\frac{\partial \text{RSS}}{\partial a} = \frac{\partial \text{RSS}}{\partial u_2} \frac{\partial u_2}{\partial a}$$

$$\frac{\partial \text{RSS}}{\partial b} = \frac{\partial \text{RSS}}{\partial u_1} \frac{\partial u_1}{\partial b}$$

. The reverse mode can be 4 times faster than the forward mode and most AD algorithms are based on this mode (this is what is commonly referred as the 'cheap gradient principle').

Note that AD is also commonly used in machine learning community but is often referred as 'back propagation'.

## A.3 Laplace approximation

Let's introduce the multivariate Gaussian distribution of a variable:

$$f_{MG}(\boldsymbol{\delta}) = \frac{1}{(2\pi)^{q/2}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{\delta}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}-\boldsymbol{\mu})}$$

with  $q$  the number of random variables in  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_q)$ ,  $\boldsymbol{\mu}$  the average of  $\boldsymbol{\delta}$ ,  $\boldsymbol{\Sigma}$  the variance-covariance matrix (positive-definite) and  $|.|$  its determinant.

By approximating  $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$  through Taylor series around the maximum  $\hat{\boldsymbol{\theta}}_\theta = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}(\ell(\boldsymbol{\theta}, \boldsymbol{\delta}))$ , we obtain:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) \approx \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) - \frac{1}{2} (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)^T (-\ell''_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\theta}, \boldsymbol{\delta})|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_\theta}) (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)$$

We rewrite the marginal likelihood of  $\boldsymbol{\theta}$ :

---


$$\begin{aligned}
L_M(\boldsymbol{\theta}) &= \int_{\mathbb{R}^q} L(\boldsymbol{\theta}, \boldsymbol{\delta}) d\boldsymbol{\delta} \\
&= \int_{\mathbb{R}^q} e^{\ell(\boldsymbol{\theta}, \boldsymbol{\delta})} d\boldsymbol{\delta}
\end{aligned}$$

Approximation through Taylor expansion at the mode  $\hat{\boldsymbol{\delta}}_\theta$

$$\approx \int_{\mathbb{R}^q} e^{\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) - \frac{1}{2}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)^t (-\ell''_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\theta}, \boldsymbol{\delta})|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_\theta})(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)} d\boldsymbol{\delta} = L_M^*(\boldsymbol{\theta})$$

$$\begin{aligned}
L_M^*(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) \int_{\mathbb{R}^q} e^{-\frac{1}{2}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)^t (-\ell''_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\theta}, \boldsymbol{\delta})|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_\theta})(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)} d\boldsymbol{\delta} \\
&\text{As } -\ell''_{\boldsymbol{\delta}\boldsymbol{\delta}}(\boldsymbol{\theta}, \boldsymbol{\delta})|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_\theta} = H(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1} \text{ for} \\
&\text{a multivariate Gaussian distribution} \\
&= L(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) \int_{\mathbb{R}^q} e^{-\frac{1}{2}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_\theta)} d\boldsymbol{\delta} \\
&= L(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) \int_{\mathbb{R}^q} f_{MG}(\boldsymbol{\delta}) \sqrt{|\boldsymbol{\Sigma}|} (2\pi)^{q/2} d\boldsymbol{\delta} \quad (\text{Cf. formula of } f_{MG}(\boldsymbol{\delta})) \\
&= L(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) \sqrt{|\boldsymbol{\Sigma}|} (2\pi)^{q/2} \int_{\mathbb{R}^q} f_{MG}(\boldsymbol{\delta}) d\boldsymbol{\delta} \\
&\text{and finally as } \int_{\mathbb{R}^q} f_{MG}(\boldsymbol{\delta}) d\boldsymbol{\delta} = 1 \\
&= L(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) |\boldsymbol{\Sigma}|^{1/2} (2\pi)^{q/2} = L(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}_\theta) |\boldsymbol{H}(\boldsymbol{\theta})|^{-1/2} (2\pi)^{q/2}
\end{aligned}$$

## A.4 Some remarks on the matrix of the SPDE approach

The Gaussian weights  $w_k$  defined on the mesh nodes are parameterized as a GMRF  $\boldsymbol{\delta}^*$  through a precision matrix  $\mathbf{Q}$ . Assuming the mesh geometry, the precision matrices  $\mathbf{Q}$  defined in Figure 2.2 can be re-expressed through  $3 n \times n$  – matrices  $\mathbf{C}, \mathbf{G}, \mathbf{K}$ :

$$\begin{aligned}
C_{ij} &= \langle \psi_i \cdot \psi_j \rangle \\
G_{ij} &= \langle \nabla \psi_i \cdot \nabla \psi_j \rangle \\
(K_\kappa)_{ij} &= \kappa^2 \cdot C_{ij} + G_{ij}
\end{aligned}$$

where  $\psi_i$  and  $\psi_j$  are the basis functions values at mesh nodes  $i$  and  $j$ .  $\nabla$  is the gradient and  $\langle . \rangle$  is the inner product.

The precision matrix  $\mathbf{Q}$  depends on  $\alpha = \nu + d/2$  (with  $d = 2$  in two dimensions) and on  $\kappa$  so that:

---


$$\begin{aligned}\mathbf{Q}_{1,\kappa} &= \mathbf{K}_\kappa = \kappa^2 \mathbf{C} + \mathbf{G} \quad \text{for } \alpha = 1 \\ \mathbf{Q}_{2,\kappa} &= \mathbf{K}_\kappa \mathbf{C}^{-1} \mathbf{K}_\kappa = \kappa^4 \mathbf{C} + 2\kappa^2 \mathbf{G} + \mathbf{G} \mathbf{C}^{-1} \mathbf{G} \quad \text{for } \alpha = 2 \\ \mathbf{Q}_{\alpha,\kappa} &= \mathbf{K}_\kappa \mathbf{C}^{-1} \mathbf{Q}_{\alpha-2,\kappa} \mathbf{C}^{-1} \mathbf{K}_\kappa \quad \text{for } \alpha = 3, 4, \dots\end{aligned}$$

Several remarks can be made on  $\mathbf{Q}$  and the projector matrix  $\mathbf{A}$ :

- The matrix that is used to relate the Gaussian weights (i.e. the random field values at the mesh nodes) to the values of the random field within the triangles is usually referred as the projector matrix  $\mathbf{A}$ . It is a  $m \times n$  matrix (with  $m$  the number of data points and  $n$  the number of mesh nodes) filled with the values of the basis functions at each position of the  $m$  data points (see Krainski et al. (2018) for illustration).
- $\mathbf{C}$  and  $\mathbf{G}$  are easy to compute as they only depend on the basis functions (i.e. the mesh geometry). They only need to be computed when building the mesh and they will remain fixed during the fitting procedure. Then, the expression of  $\mathbf{Q}$  depends on only one parameter  $\kappa$ .
- $\mathbf{C}$  is dense and can be reformulated as the diagonal matrix  $\tilde{\mathbf{C}}$  with  $\tilde{C}_{i,i} = \langle \psi_i, 1 \rangle$  so that resulting matrix  $\mathbf{Q}$  remains sparse.
- Both  $\mathbf{G}$  and  $\tilde{\mathbf{C}}$  matrices reflect some characteristics of the mesh nodes ( $\mathbf{G}$  represents the connectivity of the mesh node while  $\tilde{\mathbf{C}}$  represents the area of the dual polygon associated with the node). By contrast, the projector matrix  $\mathbf{A}$  reflects the relative position of the data points from the mesh nodes.
- Lindgren, Rue, and Lindström (2011) generalized the expression of  $\mathbf{Q}$  so that  $\alpha$  (or  $\nu$ ) can take fractional values. This allows to match some standard covariance functions such as the exponential ones (for  $\alpha = 3/2$  or  $\nu = 0.5$ ).

# COMBINING SCIENTIFIC SURVEY AND COMMERCIAL CATCH DATA TO MAP FISH DISTRIBUTION

---

## B.1 Modeling framework

This supplementary material provides all the notations used in the 1st chapter (B.1.1), how the sampling process density function is discretized (1.2), the formula of a dimensionless spatial metric we developed to quantify the strength of preferential sampling (1.3), some comments about the method of inference (maximum likelihood estimation through TMB) (1.4) and the theory that underlies the consistency check (1.5).

### B.1.1 Notations

Table B.1: List of model variables, parameters, indices and metrics.

Symbol	Name
<b>Indices</b>	
$x$	Grid cell index
$n$	Number of grid cells
$com$	Commercial
$sci$	Scientific
$i$	Observation index
$j$	Index related to the different data sources or fleets in the observation and the sampling equations
$m$	Size of the training sample

---

$m'$

Size of the validation sample

**Latent field**

$S(x)$

Latent field of relative biomass

$\alpha_S$

Intercept of the process equation

$\Gamma_S(x)^T$

Covariates of the process equation at point  $x$

$\beta_S$

Fixed parameters of the biomass field equation (species-habitat relationship estimates)

$\delta(x)$

Spatial random effect of the biomass field equation

$M(x, x'; \kappa, \phi)$

Matérn correlation function with shape  $\kappa$  and scale  $\phi$

$\rho$

Range

**Sampling process for data source  $j$**

**subject to preferential sampling**

$X_{comj}$

Inhomogeneous Poisson point process representing VMS points identified as fishing  
Intensity of  $X_{comj}$

$\lambda_j(x)$

Intercept of the commercial sampling equation

$\alpha_{Xj}$

Spatial random effect of the commercial sampling equation

$\eta_j(x)$

**Observations**

$Y_i$  (or  $y_i$  when  $Y_i$  is realized)

Observations

$\xi_j$

Expected rate of decrease of null catch with increasing latent field values

$\mu_j(x)$

Catch expectancy at point  $x$

$\sigma_j^2$

Observation error - variance of the observation process

---

$q_j$	Relative catchability
$k_j$	Scaling factor
$n_{scientific}$	Number of scientific data
$n_{commercial}$	Number of commercial data
<b>Metrics</b>	
$MSPE$	Mean square prediction error between simulated latent field $S(x)$ and estimated latent field $\hat{S}(x)$ (only in simulations)
$MSPE_{fit}$	Goodness-of-fit metric - Mean square prediction error between fitted observations and predicted observations (only in case studies)
$PCV$	Predictive capacity metric – Predictive cross validation – Mean square error between validation observations and predicted observations
$T_j(x)$	Spatial targeting metric (dimensionless)

---

### B.1.2 Simplification of the density function of an inhomogeneous Poisson point process on a discrete domain

Following Diggle (2013) and assuming that the density of the inhomogeneous point process  $\lambda(x)$  is piecewise constant in each cell grid, the density function with respect to the Poisson measure is given by:

$$\begin{aligned}\log(f(x)) &= \int_D (1 - \lambda(u)) du + \sum_i \log(\lambda(u_i)) \quad (\text{base expression of the log-likelihood}) \\ &= \sum_{x \in D} (1 - \lambda(x)) \cdot \zeta_x + \sum_i \log(\lambda(x_i)) \quad (\text{piecewise constant approximation}) \\ &= \sum_{x \in D} (1 - \lambda(x)) \cdot \zeta_x + \sum_{x \in D} c_x \cdot \log(\lambda(x)) \quad (\text{change of variable of the second term: pass from the observation level } i \text{ to the cell level } x)\end{aligned}$$

where  $f(x)$  is the density function of the inhomogeneous Poisson point process,  $D$  the domain,  $u$  the continuous space locations,  $x$  the discrete space locations (i.e. grid cells),  $i$  the index for commercial samples,  $u_i$  or  $x_i$  the locations (resp. continuous or discrete) of the  $i^{th}$  commercial sample,  $\lambda(u)$  and  $\lambda(x)$  the value of  $\lambda$  at point  $u$  or at grid cell  $x$ ,  $\zeta_x$  the surface of a grid cell (here 1),  $c_x$  the number of fishing points in a cell  $x$ .

---

### B.1.3 Targeting metric

The parameter  $b_j$  is not dimensionless as its value depends on the scale of the latent field. Thus, its value may not necessarily be representative of the strength of PS, particularly when comparing case studies where latent field values do not fall in the same range of values. To quantify the degree of PS in a cell independently of the latent field dimension, we build a dimensionless metric  $T_j(x)$  that quantifies how much a cell is over or under-sampled compared with the case where sampling would have been non-preferential.  $T_j(x)$  is calculated as the ratio between the expected proportion of fishing points in the cell  $x$  with and without PS (i.e.  $b_j$  is either fixed to its maximum likelihood estimates ( $\hat{b}_j$ ) or fixed to 0). Note that  $\lambda_j^*(x)$  is dimensionless for any choice of  $b_j$ , hence  $T_j(x)$  is dimensionless as well.

$$\begin{aligned}\lambda_j^*(x) &= \lambda_j(x) / \sum_x \lambda_j(x) \\ T_j(x) &= \frac{\lambda_j^*(x)}{\lambda_{j,b_j=0}^*(x)}\end{aligned}$$

---

### B.1.4 Using TMB for maximum likelihood estimation

Parameters and spatial random effects are estimated through maximum likelihood methods with the package TMB that is highly efficient to maximize the likelihood of hierarchical random effect models (Kristensen et al., 2016).

Following the Stochastic Partial Differential Equations (SPDE) approach (Lindgren, Rue, and Lindström, 2011), the computational complexity is reduced by approximating GRF with Gaussian Markov Random Fields (GMRF). The resulting sparse precision matrices are quickly calculated in TMB - see details of the methods in Kristensen et al. (2016). Because estimates of the biomass latent field are primarily in the log scale, we use epsilon bias-correction to avoid retransformation bias when calculating the latent field values and the estimates of total biomass (Thorson and Kristensen, 2016).

---

### B.1.5 Consistency check

This validation step is proposed as a way to check whether or not the commercial data used in the IM lead to inconsistent estimates with the scientific data (considered as a reference dataset). It was developed in Rufener (2020) and Rufener et al. (2021) for models combining scientific survey data and commercial data (in this case, data collected by onboard observers). The validation is built on a likelihood ratio test where the null hypothesis states that parameters estimated with the scientific data alone are equal to the parameters estimated with both data sources. There is a significant inconsistency between the two models when the null hypothesis is rejected.

Following the notation in Rufener (2020), let's denote  $\boldsymbol{\theta}_{latent}$  the vector of the shared fixed parameters from the latent process,  $\boldsymbol{\theta}_{sci}$  the vector of the fixed parameters related to the observation process of the scientific data,  $\boldsymbol{\theta}_{com}$  the vector of the fixed parameters related to the observation process of the commercial data,  $\mathbb{P}_\theta(\boldsymbol{\lambda})$  the probability density of the shared random effect  $\boldsymbol{\lambda}$ . When fitting scientific data only, the likelihood of the model can be formulated as  $L_{sci}(\boldsymbol{\theta}_{latent}, \boldsymbol{\theta}_{sci}) = \int L_{sci}(\boldsymbol{\theta}_{sci}|\boldsymbol{\lambda}) \cdot \mathbb{P}_\theta(\boldsymbol{\lambda}) \cdot d\lambda$ . Related estimates will be noted with the subscript 1. When fitting both data sources, it can be written as  $L_{both}(\boldsymbol{\theta}_{latent}, \boldsymbol{\theta}_{sci}, \boldsymbol{\theta}_{com}) = \int L_{sci}(\boldsymbol{\theta}_{sci}|\boldsymbol{\lambda}) \cdot L_{com}(\boldsymbol{\theta}_{com}|\boldsymbol{\lambda}) \cdot \mathbb{P}_\theta(\boldsymbol{\lambda}) \cdot d\lambda$ . Related estimates will be noted with the subscript 2.

The validation check for fixed parameters consists in:

- 1a. fitting the scientific data only and estimating  $\hat{\boldsymbol{\theta}}_{latent}^{(1)}, \hat{\boldsymbol{\theta}}_{sci}^{(1)}$ .
- 2a. fitting the IM and estimating the parameters  $\hat{\boldsymbol{\theta}}_{latent}^{(2)}, \hat{\boldsymbol{\theta}}_{sci}^{(2)}$  and  $\hat{\boldsymbol{\theta}}_{com}^{(2)}$ .
- 3a. compute the quantity  $t = 2 \cdot (\log L_{sci}(\hat{\boldsymbol{\theta}}_{latent}^{(1)}, \hat{\boldsymbol{\theta}}_{sci}^{(1)}) - \log L_{sci}(\hat{\boldsymbol{\theta}}_{latent}^{(2)}, \hat{\boldsymbol{\theta}}_{sci}^{(2)}))$  where  $\log L_{sci}(\hat{\boldsymbol{\theta}}_{latent}^{(2)}, \hat{\boldsymbol{\theta}}_{sci}^{(2)})$  is the likelihood value of the scientific-based model computed with the estimated values of the IM.  $t \sim \chi^2(df)$  and  $df = \dim(\boldsymbol{\theta}_{latent}) + \dim(\boldsymbol{\theta}_{sci})$ .

Since the null hypothesis corresponds to the case where both model parameters are equal, if the p-value is greater than 0.05 then the estimates of the IM are considered to fall within the confidence region of the estimates of the scientific model. High p-values support consistency between the IM outputs and scientific data while small p-values indicate lack of consistency.

A similar test can be applied with random effects:

- 1b. with the scientific-based model, obtain the most probable random effects  $\hat{\boldsymbol{\lambda}}^{(1)}$
- 2b. with the IM, obtain the most probable random effect  $\hat{\boldsymbol{\lambda}}^{(2)}$

- 
- 3b. compute the same statistical test as 3a with the estimated parameters and random effects

Note that when comparing the IM and the scientific-based model, the catchability coefficient of scientific data  $q_{sci}$  should have the same value as in the IM (eq. 4.6 - 3.6). Then, when fitting scientific data only, the catchability coefficient of scientific data is scaled with the scaling factor  $k_f$  estimated within the IM.

---

## B.2 Simulations material and methods

This supplementary material provides all the additional material and methods for the simulation-estimation experiments. Equations numbers refer to equations in the main text. All simulation-estimation codes are on gitlab and will be given access on request at the address: `baptiste.alglave@agrocampus-ouest.fr`.

### B.2.1 Description of simulations

Simulations are conducted using base function of the R environment for the latent field and catch observations. The point process was simulated using the package spatstat (Baddeley, Rubak, and Turner, 2015). The simulation codes were based on those provided by Conn, Thorson, and Johnson (2017). The spatial domain is a 25 x 25 cells' grid separated in 4 strata (Figure B.1).

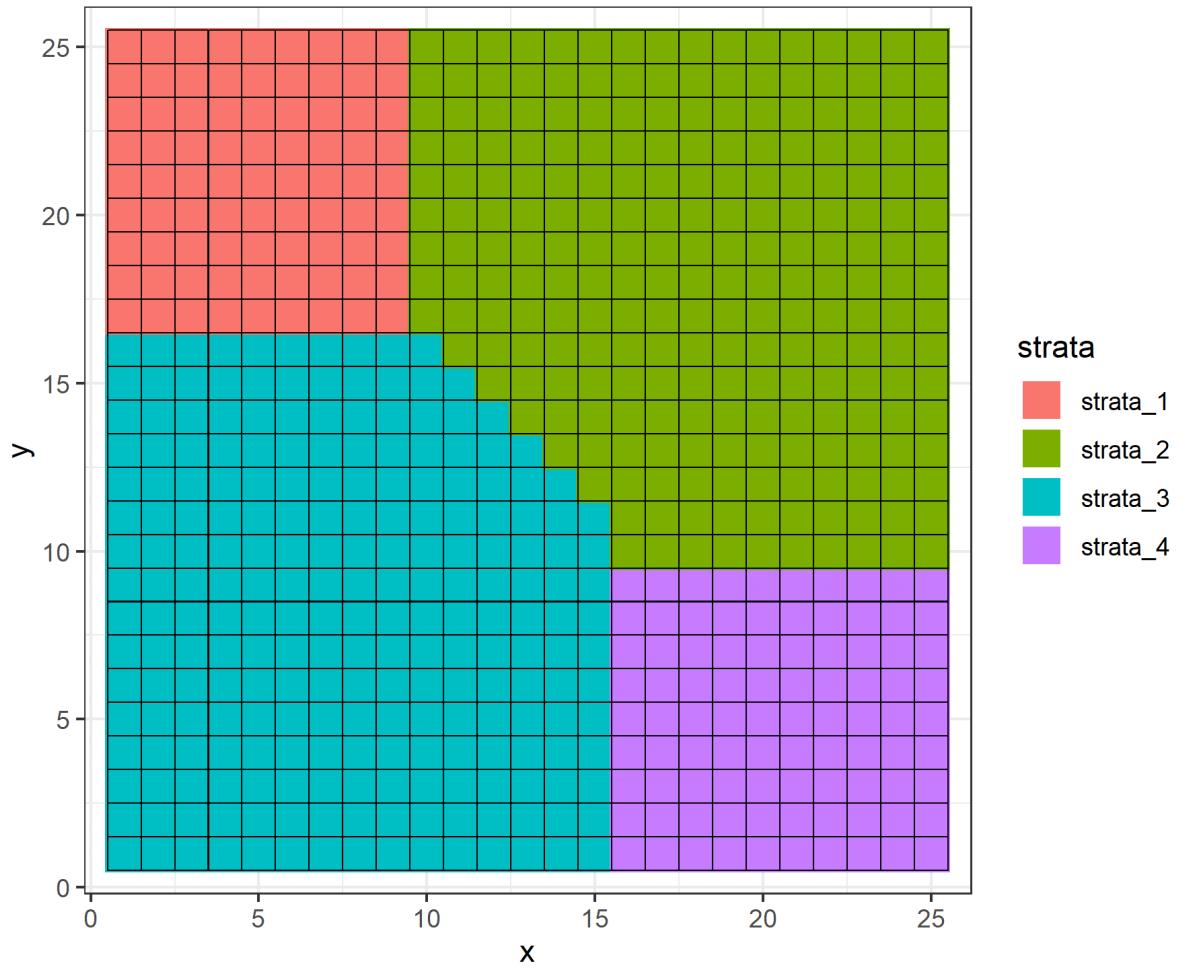


Figure B.1 – Spatial domain of the simulations. Colors: strata.

The latent field is composed of (eq. 3.1 - 3.2 in the main text):

- an intercept  $\alpha_S$  (arbitrarily fixed)
- a continuous covariate  $\Gamma_S(x)$ : the covariate is simulated as a GRF with range = 10 and marginal variance = 1. The effect of the covariate  $\beta_S$  is randomly sampled in a uniform distribution defined on  $[-0.5; 0.5]$ .
- a random effect  $\delta(x)$  parameterized as a GRF with range = 10 and marginal

---

variance = 1.

The GRF covariance is a Matérn function. GRFs are simulated with the codes provided by Krainski et al. (2018) in chapter 4, the paragraph concerning the Matérn covariance.

Scientific data points are randomly sampled within each strata following a stratified sampling plan. The sample size in each strata is proportional to the strata area, and data points are randomly sampled within the corresponding strata. One cell is never sampled twice.

Once the positions are pulled, scientific catches are simulated conditionally on the latent field values through the zero-inflated distribution described in (eq. 3.3 -3.4). The zero-inflation parameter  $\xi_j$  is set to 0 so that 10% of the scientific observations are zero values. The catchability of scientific data  $q_{sci}$  and the observation variance  $\sigma_{sci}^2$  are both set to 1.

Commercial data points are simulated with the ‘rpoint’ function (spatstat package). Sampling intensity  $\lambda_j(x)$  is specified through (eq. 3.4). It is composed of an intercept  $\alpha_{Xj}$  (fixed), the PS component  $b \cdot \log(S(x))$  ( $b$  parameter takes the value 0, 1, 3) and a spatial random effect  $\eta_j(x)$ .  $\eta_j(x)$  is set to 0 for Q1, Q2, Q3. For Q4,  $\eta_j(x)$  is set to tailor the sole case study. The range of  $\eta_j(x)$  is set to 40 (4 times the range of  $\delta$ ), the marginal variance is set to 5 (5 times the marginal variance of  $\delta$ ). Note that ‘rpoint’ simulates a set of points defined on a continuous domain. As our framework is defined on a discrete domain, the data points are discretized over the grid.

Similarly to scientific data, for each data point, catch data are simulated conditionally on the latent field value following the zero-inflated distribution described in (eq. 3.5-3.6). The zero-inflation parameter is set to -1 so that 30% of commercial observations are zero values. The catchability of commercial data  $q_{com}$  and the observation variance  $\sigma_{com}^2$  are set to 1.

---

## B.3 Case studies material and methods

This supplementary material provides all the additional material and methods that concern the case studies.

### B.3.1 Spatial grids

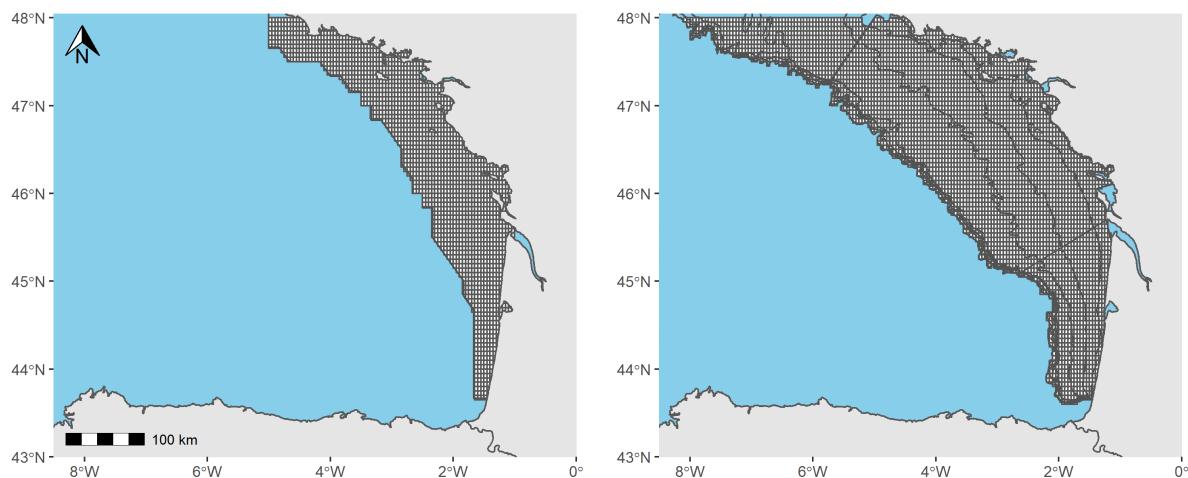


Figure B.2 – Spatial grids for the case studies. Cells dimension:  $0.05^\circ \times 0.05^\circ$ . Left: grid based on Orhago survey domain. Right: grid based on EVHOE survey domain.

The spatial domains were obtained from the shapefiles of the Orhago and the EVHOE surveys. Both domains were discretized in  $0.05^\circ \times 0.05^\circ$  grids and then were used to discretize the data to associate a spatial location to a grid cell. The domain discretization and the association between the spatial locations and the grid cells are achieved by using the GIS functionalities of R available in the package raster (<https://cran.r-project.org/web/packages/raster/index.html>) and sp (<https://cran.r-project.org/web/packages/sp/index.html>).

---

### B.3.2 Sampling intensity of European bottom trawl surveys

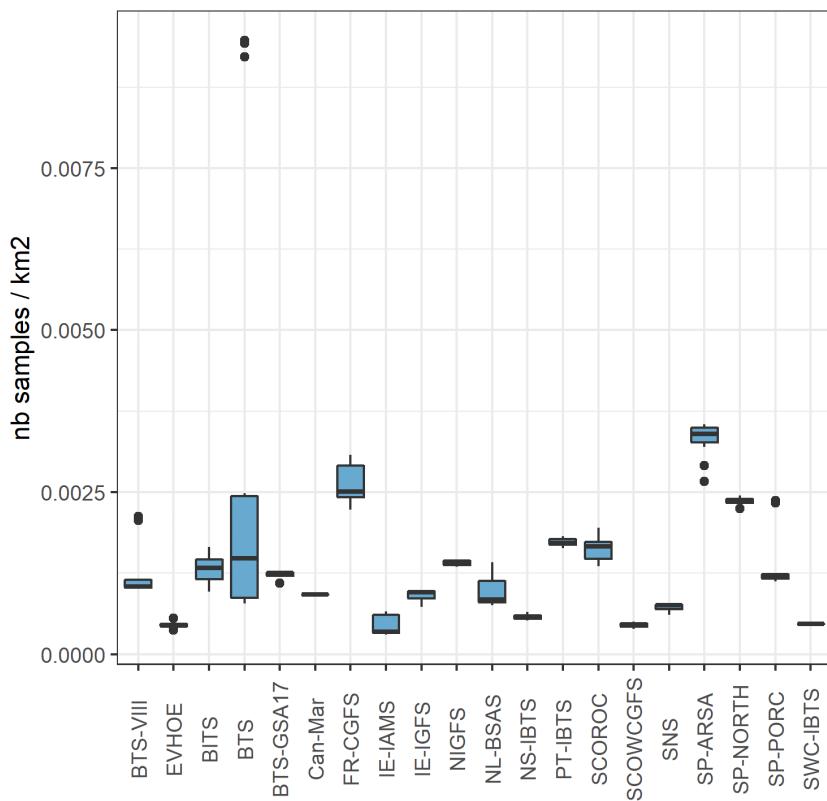


Figure B.3 – Sampling density of the main European trawl surveys. Source: DATRAS. Unit: number of fishing points per km<sup>2</sup>. Orhago survey is BTS-VIII. In comparison, sampling intensity of ‘VMS x logbooks’ data in the case studies ranges from 0.15 to 0.25 samples/km<sup>2</sup>.

---

### B.3.3 Species and stock description

The common sole is a benthic nursery-dependent flatfish widely distributed over the eastern Atlantic Coast from Senegal to southern Norway and along the Mediterranean rim. The Bay of Biscay stock (Divisions VIIa,b) is mainly exploited by French vessels which realize 90% of total catch for this stock (ICES, 2019a). As discards are assumed to be at very low levels for that stock (1.7% discard rate), landings are considered representative of catches (ICES, 2018b). Hake is a benthic fish distributed over the Northeast Atlantic shelf, from Norway to Mauritania (ICES, 2013; ICES, 2020b). In the Bay of Biscay (Divisions VIIa,b,d) and the Celtic Sea (Division IIIa, Subareas IV, VI and VII), Spain accounts for 60% of landings followed by France which accounts for 25%. Discards are estimated at a relatively low level (10 %). Squids concern a mixture of three species that are merged in landings data and in stock assessment: *Loligo vulgaris* (Lamarck, 1798) and *Loligo forbesii* (Steenstrup, 1856) and *Alloteuthis subulata* (Lamarck, 1798). They all have a benthic existence as adults. In the Bay of Biscay (Divisions VIIa,b), French vessels account for 95% of landings (ICES, 2020c). It is a data-poor stock, which may potentially benefit from an integrated approach. Discard rates are not well documented but appear to be relatively small from 2016 to 2018 (less than 10%).

---

### B.3.4 Scientific data

All scientific data from the Orhago and EVHOE survey were extracted from the DATRAS database (<https://www.ices.dk/data/data-portals/Pages/DATRAS.aspx>) with the package ‘icesDatras’ (<https://cran.r-project.org/web/packages/icesDatras/index.html>).

#### **Orhago**

Orhago is an annual winter beam trawl survey occurring in November designed to assess sole stock status in the Bay of Biscay (Figure 3.3 in the main text). Mesh size of the beam trawl equals 50 mm. 49 hauls are distributed within four strata all along the Bay of Biscay following a random stratified sampling plan for 2017/2018. Catch weights are available for each sampled species.

#### **EVHOE**

EVHOE is an annual bottom trawl survey occurring in late October and November. It is designed for demersal fishes in the Bay of Biscay and in the Celtic Sea (only the data in the Bay of Biscay is considered here). The gear used is a GOV 36/47 with mesh size equals to 20 mm (ICES, 2015). In the Bay of Biscay, 86 hauls are distributed within 14 strata following a stratified sampling plan for 2014 and 2015 (Figure 3.3 in the main text).

Note that in both surveys, the standardized fishing time for each haul is 30 minutes.

---

### B.3.5 Commercial data

Fishing vessels have been gradually equipped with VMS since 2005 and today all boats above 12 m provide geolocalised data with a one-hour time step. A procedure combining VMS and catch declaration data were used to allocate catches to VMS positions (with the same methodology as Gerritsen and Lordan (2011), Hintzen et al. (2012), and Murray et al. (2013)), and discretize data on the grid (see the next section SM B.3.6). As for survey data, the catches are standardized by fishing effort to compute CPUE (in kg/hour). All parameters are supposed to be homogeneous within a fishing fleet  $j$  (either scientific or commercial). Thus, pre-processing of the data must be carried out a priori (e.g. before using the data in the IM) to maximize homogeneity within each fleet. For each species, we filtered commercial data for ‘bottom trawlers’ as they cover a wide part of the study area (Figure 3.3 in the main text) and provide easy to compute and reliable CPUE. Indeed, trawlers are characterized by a more linear relationship between catch and fishing time making fishing time a good proxy of effort to compute CPUE representative of biomass (Hovgêrd and Lassen, 2008). For hake and sole, we filtered the métier targeting demersal fish (called OTB\_DEF) and for squids, the métier targeting cephalopods (called OTB\_CEP). For both métier, the mesh size is greater than 70 mm. The common sole case study was used to illustrate the capacity of the model to integrate multiple commercial fishing fleets. A simple clustering was carried out to partition OTB\_DEF vessels in smaller and more homogeneous units exhibiting a certain homogeneity regarding targeting behavior (towards sole only) and technical characteristics. Following methods proposed by several authors (Pelletier and Ferraris, 2000; Ferraris, 2002; Stephens and MacCall, 2004; Deporte et al., 2012; Winker, Kerwath, and Attwood, 2013; Okamura et al., 2018), a PCA coupled with a hierarchical clustering (HCPC) were conducted on commercial data (see SM B.3.9 below for more details) and showed that the OTB\_DEF métier could be partitioned in two different fleets:

- a first fleet composed of small vessels for which sole represents a large proportion of catch (8.3% in average). These vessels fish between latitude 44°N to 46°N (from Oléron to the south of the Bay of Biscay).
- a second fleet composed of larger vessels for which sole represents a small proportion of catch (2.0% in average). These vessels fish all along the Bay of Biscay. We then treat these two fleets as separate fisheries for common sole.

---

### B.3.6 Building commercial data

Raw data are directly issued from landing declaration (logbook data) and VMS data (GPS position and speed data of fishing vessels) made available by the French administration and the SIH (Système d'Information Halieutique - <https://sih.ifremer.fr/>). Logbook data consist in catches declarations by the fishers for each fishing trip, day, gear, fishing zone (ICES statistical rectangles). Logbook data are controlled and consolidated through the SACROIS algorithm (Demanèche et al., 2013) by comparing declaration data to other data flows (landing declarations and sales data are used to control declared quantity and declared species names, VMS data is used to control catch spatial distribution). To get a high resolution map of catch distribution, SACROIS data (catch per ICES rectangle, fishing trip, day and gear) are then distributed on a finer spatial scale based on VMS data following a methodology similar to the one developed by Hintzen et al. (2012). This spatial allocation is based on two standard hypotheses (Gerritsen and Lordan, 2011; Murray et al., 2013):

#### **Identifying fishing sequences from VMS positions**

To identify fishing sequences from VMS data, Ifremer uses the AlgoPesca Software dedicated to the treatment of geo-localized data (Demanèche et al., 2013). This algorithm identifies fishing activities based on an average speed threshold of 4.5 knots and removes “false positive” fishing position associated for example to low speed while approaching harbor.

#### **Distribution of catches on fishing sequences**

To distribute catch on fishing sequence identified in VMS positions of a given fish trip, daily catches are uniformly distributed on all positions previously identified as fishing. Catch at a point  $x$  are then standardized by the related fishing effort (here the fishing time) so that data is expressed in CPUE to be representative of biomass. Given the space discretization of the grid, some commercial catch of the same fishing operation can fall in the same defined cell. In this case, catch and effort are summed at the cell level and CPUE are computed with resulting catch and effort.

---

### B.3.7 Habitat covariates

#### Depth

Depth was extracted from the GEBCO website (<https://download.gebco.net/> - resolution:  $0.004^\circ$ ) and sediment from the EMODNET platform (<https://www.emodnet-seabedhabitats.eu/access-data/download-data/> - resolution:  $0.05^\circ$ ). Depth does not have a linear effect on scientific observations (see SM B.3.8) and was then considered as a category variable: we divided depth into 2 classes in each case study ( $]-\infty, -50m]$ :  $]-50m, -0m]$  for sole and  $]-\infty, -100m]$ :  $]-100m, -0m]$  for hake and squids) that include almost the same number of scientific observations and that have a relatively homogeneous effect on scientific observations.

#### Sediment

Sediment types were aggregated into three classes: sand/coarse substrate, mud/fine sediment and rock. There is no "rock" level for scientific data as the scientific vessels do not sample on rock substrate.

#### Identifiability constraint

The identifiability constraint for the latent field equation is set with the sand/coarse substrate level and the depth class  $]-\infty, -10m]$  for hake,  $]-\infty, -50m]$  for sole and  $]-100m, 0]$  for squids. This combination of factor levels is referred as 'baseline' in the following.

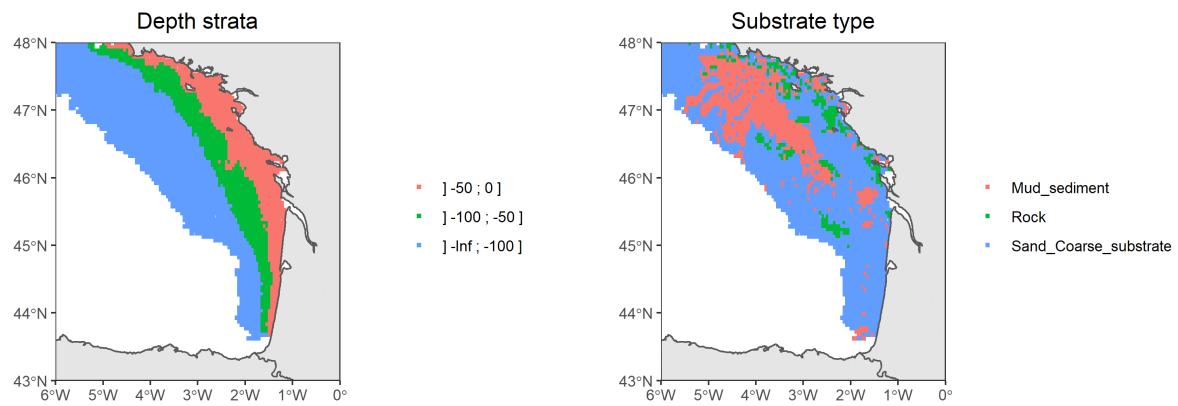
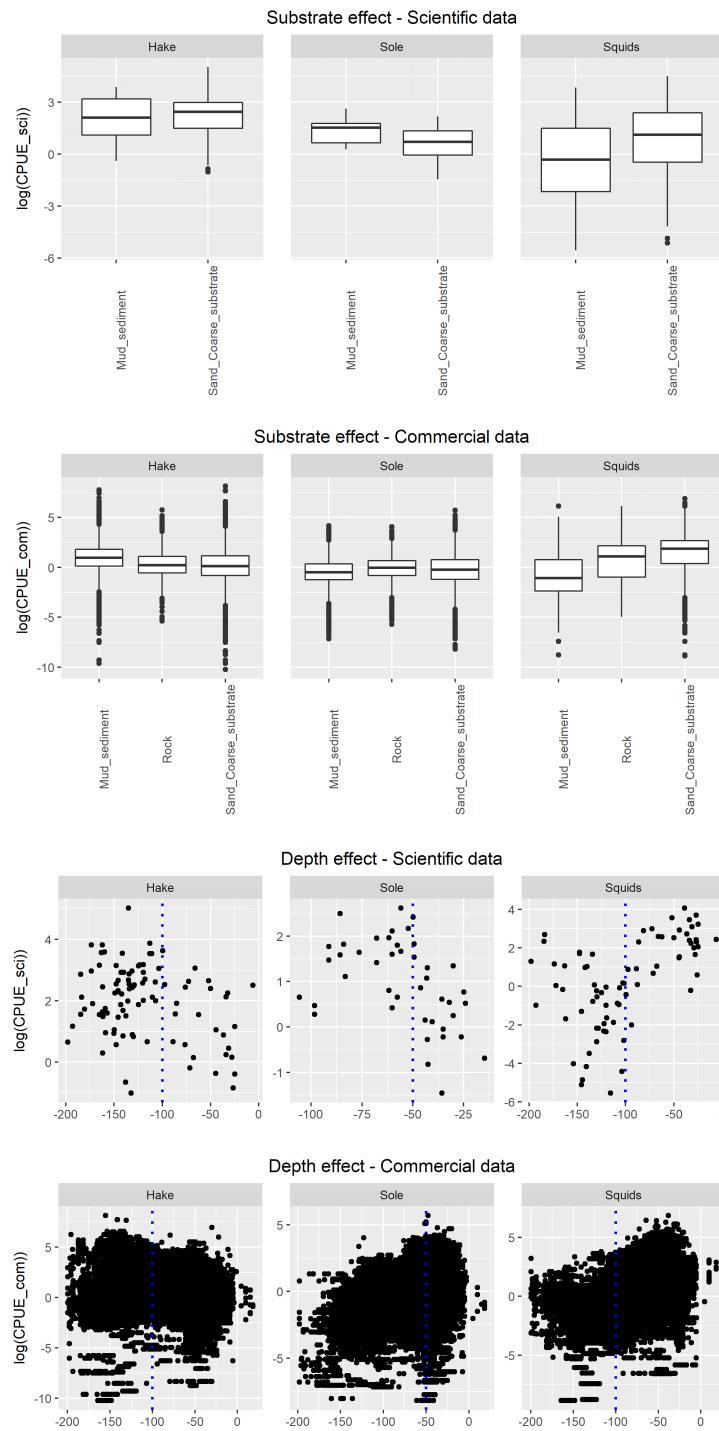


Figure B.4 – Maps of covariates. Left: depth strata limits in meters. Right: substrate type

### B.3.8 Effect of substrate and depth on scientific and commercial observations



---

Figure B.5 – Effect of depth (meter) and substrate type on log-scaled scientific and commercial observations for each case study. Dashed blue line: limit of the depth factor levels. Data were aggregated over years to produce the plots; Hake and squid plots were obtained on 2014 and 2015 data; Sole plots were obtained on 2017 and 2018.

---

### B.3.9 Fleet structure analysis: results of the PCA and the HCPC conducted on commercial data

The PCA and the HCPC were performed with the package FactoMineR (<https://cran.r-project.org/web/packages/FactoMineR/index.html>). They are conducted on a dataset containing 57 individuals (the vessels constituting the fleet) and 4 quantitative variables: proportion of sole in catch (prop\_spp), vessel length (VE\_LEN), mean latitude of the points identified as fishing for each vessel, mean distance to the coast of the points identified as fishing for each vessel. Proportion of sole in landings aims to quantify fishers targeting behavior toward sole. We used the Ward method with Euclidian metric to perform the HCPC.

The graph of variables is presented in Figure B.6, the clustering tree is presented in Figure B.7 and the graph of individuals with corresponding clusters is presented in Figure B.8.

From the inertia gain graph (loss of inertia inter class - Figure B.7), we retained a partition in 2 clusters. This partition separates small vessels characterized by relatively high proportion of sole and large vessels characterized by small proportion of sole. The first cluster is called "fleet\_0", vessels mainly fish in the south of the Bay of Biscay. The second cluster is called "fleet\_1", vessels fish all along the Bay of Biscay.

Other variables could have been integrated in the analysis (landing harbor, departure harbor, vessel power, etc.), in this case the multivariate analysis would have been an MFA (Multiple Factor Analysis). However, integrating those variables doesn't strongly modify the actual partitioning.

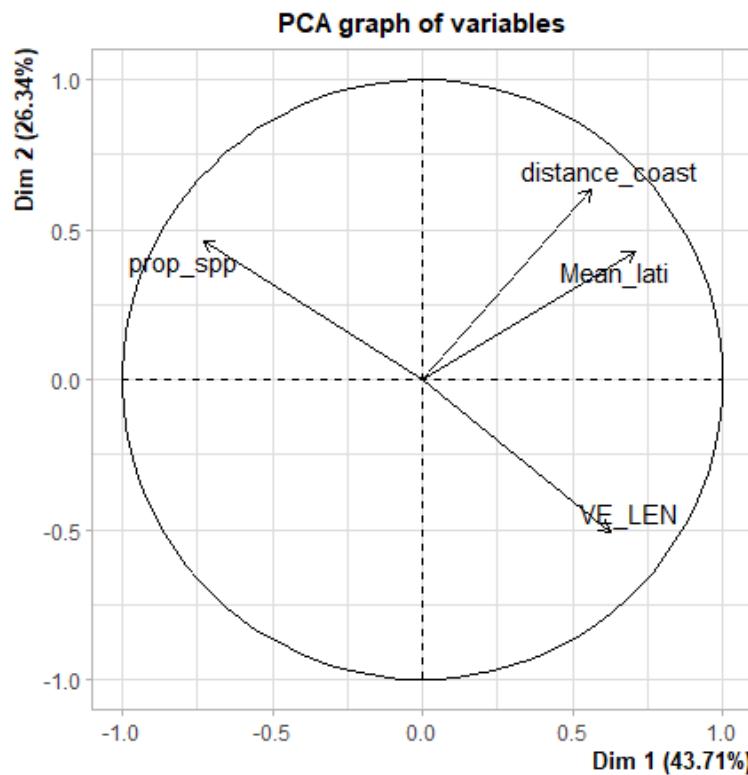


Figure B.6 – PCA - Graph of variables. prop\_spp: proportion of sole, VE\_LEN: vessel length, Mean\_lati: mean latitude of the points identified as fishing for each vessel, distance\_coast: mean distance to the coast of the points identified as fishing for each vessel.

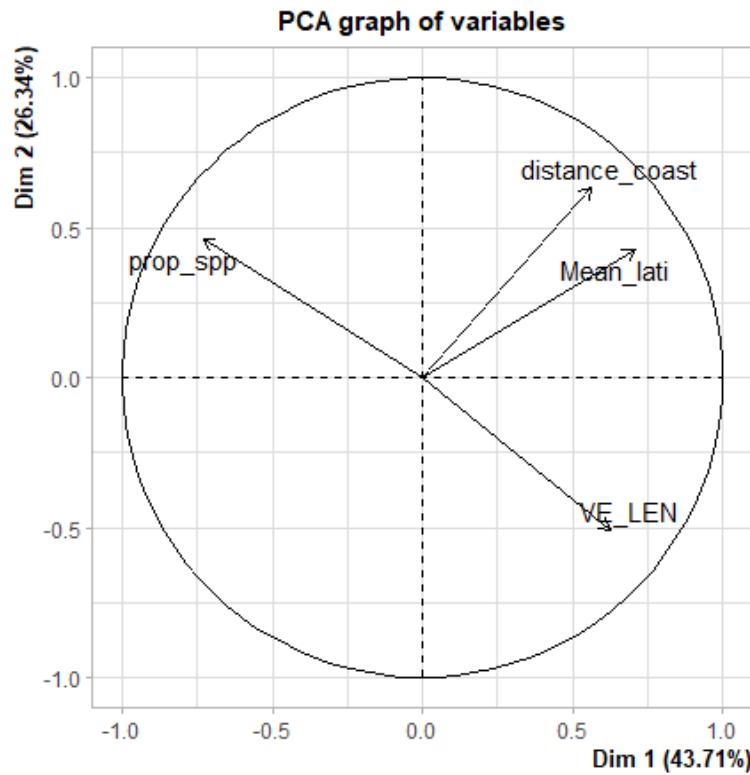


Figure B.7 – Clustering tree.

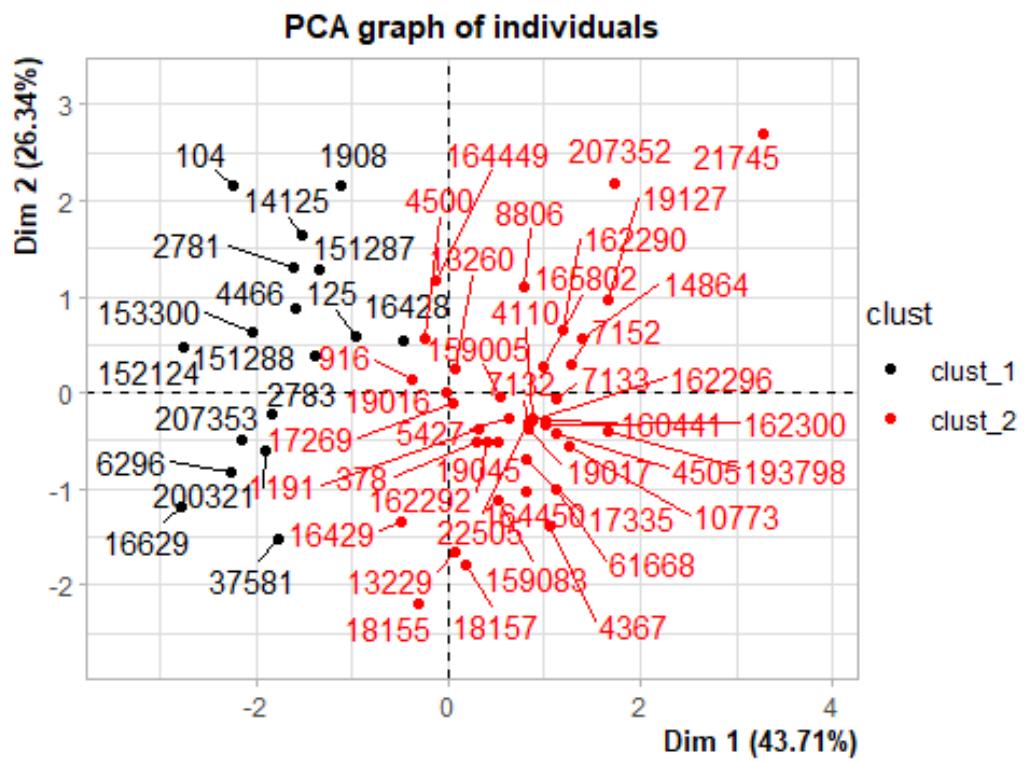


Figure B.8 – Graph of individuals (individual = fleets).

---

### B.3.10 Goodness-of-fit and predictive capacity metrics

To compare model performances, we assess model goodness-of-fit and predictive capacity through a 5-fold cross validation. We evaluate models based on 2 criteria: the mean square prediction error for goodness-of-fit ( $MSPE_{fit}$ ) and the predictive k-fold cross-validation for predictive capacity ( $PCV$ ):

$$MSPE_{fit} = \frac{1}{m} \sum_{j=1}^m \{y_j - E(Y(x_j))\}^2$$

$$PCV = \frac{1}{m'} \sum_{j'=1}^{m'} \{y_{j'} - E(Y(x_{j'}) | y_{-m'})\}^2 \quad (\text{B.1})$$

where,  $Y$  stands for the observations as random variable, while  $y$  designs their realization.  $y_{-m'}$  denotes the training sample (i.e the whole dataset with all validation observations  $y_{j'}$  removed),  $m$  is the training sample size and  $m'$  the validation sample size.  $x$  are the locations of the fitted or the validation observations.

For both metrics, the lower the values, the better the model fits/predicts the data. Note that in the case study the  $MSPE_{fit}$  is computed by confronting real observations to predicted observations while in simulations/estimations, the  $MSPE$  is computed by confronting simulated latent field values to estimated latent field values.

These metrics compare model predictions to observations (either fitted observations with  $MSPE_{fit}$  or validation observations with  $PCV$ ). In the case of IM, those metrics can be computed separately for each source of information (e.g., scientific data and commercial data). Thus for each model, we can compute 4 metrics :

- $MSPE_{survey}$ : the goodness-of-fit metric of the IM towards scientific data.
- $MSPE_{vms}$ : the goodness-of-fit metric of the IM towards commercial data.
- $PCV_{survey}$ : the predictive capacity metric of the IM towards scientific data.
- $PCV_{vms}$ : the predictive capacity metric of the IM towards commercial data.

Furthermore, all metrics can be computed for each set of k-fold validation (here, 5 sets of validation). All of these five values are plotted when analyzing model goodness-of-fit and predictive capacity (e.g. in SM B.5.10).

## B.4 Simulation results

This supplementary material provides all the additional results that concern the simulations.

### B.4.1 Relative bias of covariates effect estimates

Estimates are unbiased for all models and data configurations. That commercial data cover or not the full area does not affect the species-habitat relationship estimates. Accounting for PS or not does not affect species-habitat relationship estimates.

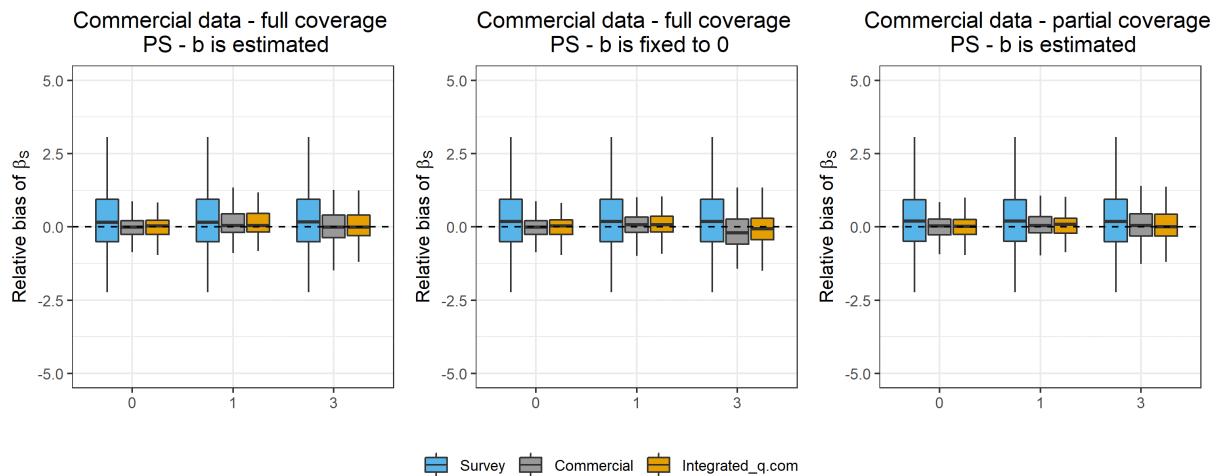


Figure B.9 – Relative bias of the covariates effect estimate  $\hat{\beta}_S$ . In all configurations, simulations are conducted for 3 levels of preferential sampling (x-axis:  $b = 0, b = 1, b = 3$ ). Colors: data sources used in the integrated model for inferences. Number of scientific samples: 50. Number of commercial samples: 3000. Boxplots represent the variability among the 100 replicates.

## B.4.2 Consistency check for simulations-estimations

### Effect of sample size on consistency tests (Q1)

Regarding the fixed effect test, almost all simulations successfully passed the consistency check (p-values fall above the 0.05 threshold). The more the datasets are unbalanced, the more the p-values decrease; but even when both data are very unbalanced (low scientific sample and high commercial sample - com.L\_sci.S) only 3% of the p-values fall below the 0.05 threshold for the fixed effect test (the test wrongly rejects consistency).

For the random effect test, the results are a bit more balanced as 10% of the p-values fall below the 0.05 threshold when data size are very unbalanced (com.L\_sci.S). The test is an asymptotic test (true for a high number of observations) and then the small scientific sample size (50) might explain that the test performs worse in this case.

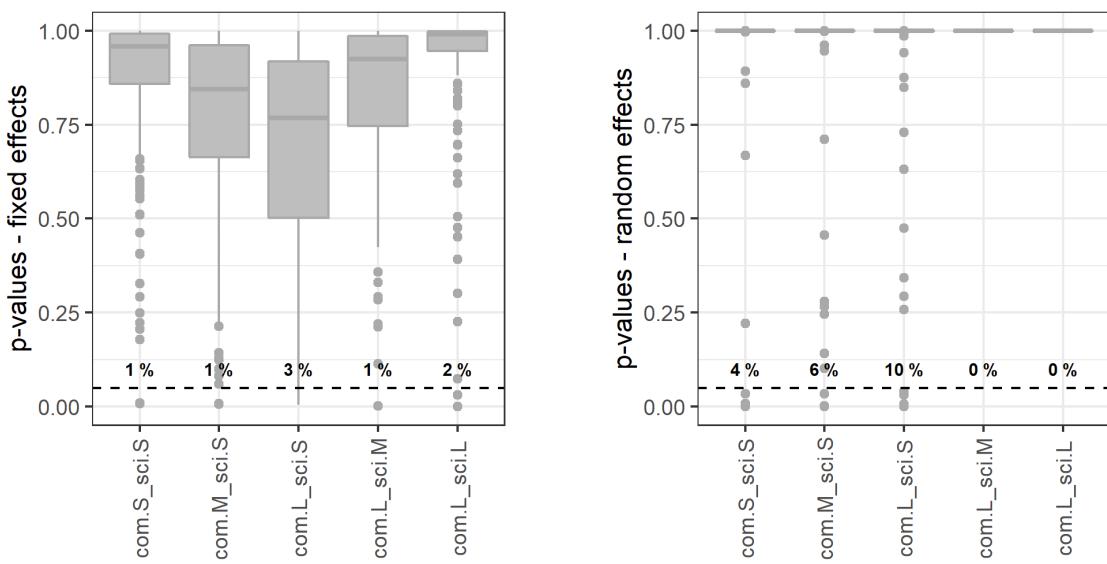


Figure B.10 – Results of the consistency check on fixed (left) and random effects (right) for several commercial and scientific sample sizes. Black lines: 0.05 threshold value. Percentage values: percentage of tests falling below the 0.05 threshold. x-axis: 5 combinations of commercial and scientific sample size. ‘com’ stands for commercial, ‘sci’ stands for scientific, S stands for small sample size (50), M stands for middle sample size (400), L stands for large sample size (3000).

---

### **Consistency tests for alternative simulation scenarios and model configurations (Q2.3.4)**

As previously, for fixed effect, almost all p-values fall above the 0.05 threshold for the fixed effect test. When regarding the random effect, the results are a bit worse as 10% of p-values fall below the 0.05 threshold for the base case, Q2 (partial coverage of commercial data) and Q3 (estimation does not account for PS).

Interestingly, for Q4 (other processes than PS affect sampling locations), the number of p-values falling below the 0.05 threshold is smaller for the random effect test. This can be related to the higher contribution of scientific data in inference leading to increased consistency with the scientific-based estimates.

Finally, the model misspecifications we assessed in simulations (Q2 and Q3) do not strongly affect the consistency check outcomes. The percentage of p-values falling below the 0.05 threshold for the fixed and random effect tests are similar in the base case and in Q2 or Q3.

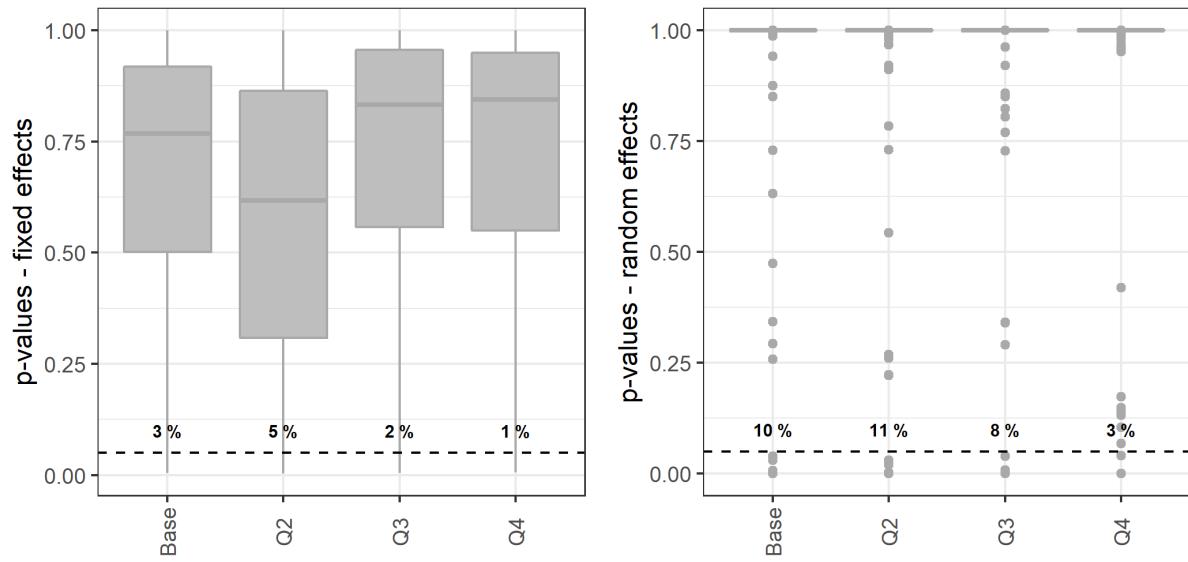


Figure B.11 – Results of the consistency check on fixed (left) and random effects (right) for different simulation scenario and model configuration. Black lines: 0.05 threshold value. Percentage values: percentage of consistency checks falling below the 0.05 threshold. Base: no discrepancy between simulation and estimation. Q2: available commercial data do not cover a  $9 \times 9$  zone of the grid. Q3:  $b$  is arbitrarily fixed to 0 in the estimation models. Q4: data is simulated with a random effect  $\eta$  in the sampling intensity process.

---

### B.4.3 Relative bias of range estimates when increasing sample size

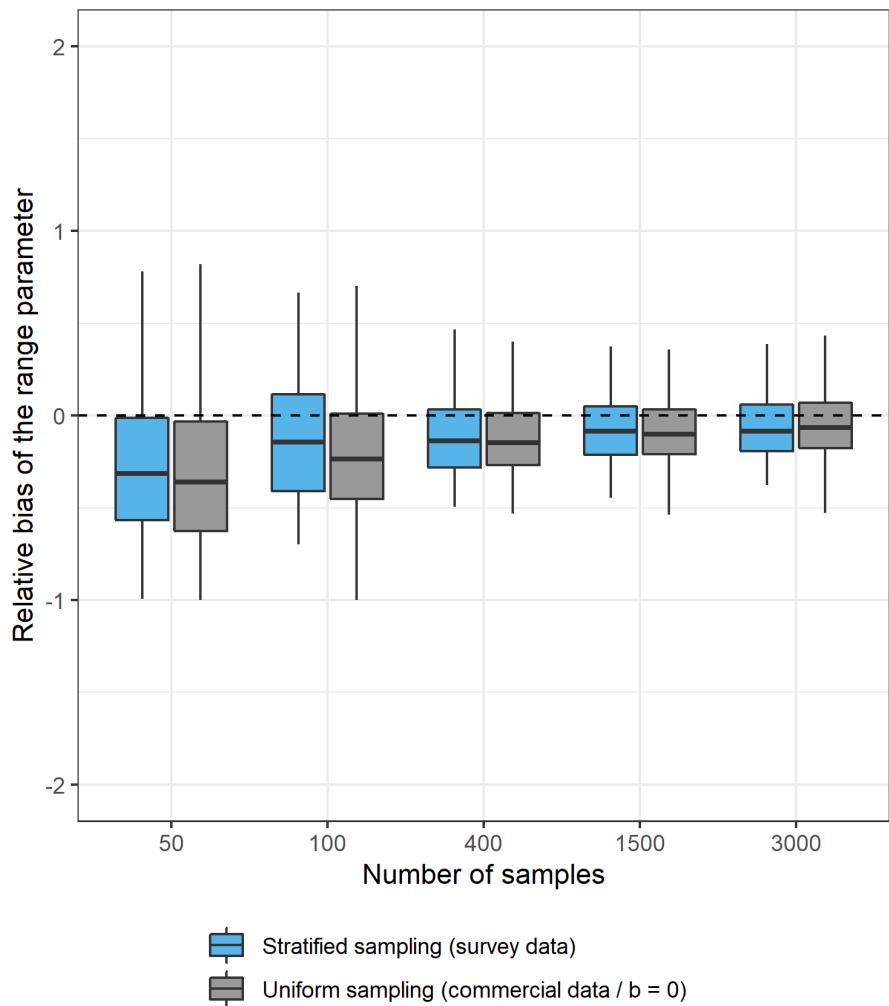


Figure B.12 – Relative bias of the range  $\hat{\rho}$  of the latent field random effect  $\delta(x)$  for increasing number of observations and two alternative sampling designs. Stratified sampling is simulated with scientific data (random stratified plan described in the manuscript). Uniform sampling is simulated with commercial data assuming  $b = 0$  and  $\eta(x) = 0$  for all  $x$ . Commercial data cover the full spatial domain.

---

#### B.4.4 Comparison between simulations and predictions for strong preferential sampling ( $b = 3$ )

The scientific-based model provides a smooth picture of species biomass distribution as there are only few scientific samples (50) and scientific samples mismatch some localized high-density areas (Figure B.13). Integrating the commercial data in inference allows to better capture the fine scale patterns and the localized hotspot.

The areas where sampling is high are well predicted by both the integrated and the commercial-based model and the overall spatial patterns are the same when accounting or not for PS (Figure B.13). When ignoring PS in estimation, the main differences between simulated and estimated values are localized in poorly sampled areas. Fitting commercial data alone overestimates the latent field values in those areas (Figure B.14, bottom left – this is the difference between the bottom middle and the upper middle panel of Figure B.13). Integrating scientific data in the analysis corrects in part this bias and improve predictions accuracy (Figure B.14, bottom right and see related MSPE in Figure 5, 3<sup>rd</sup> row, 3<sup>rd</sup> column).

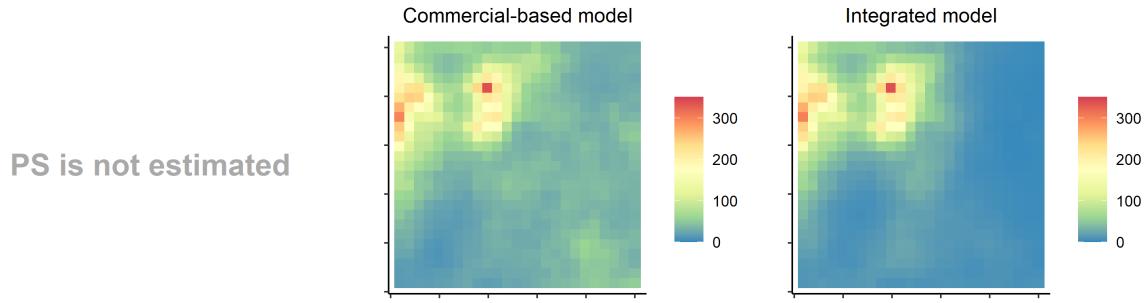
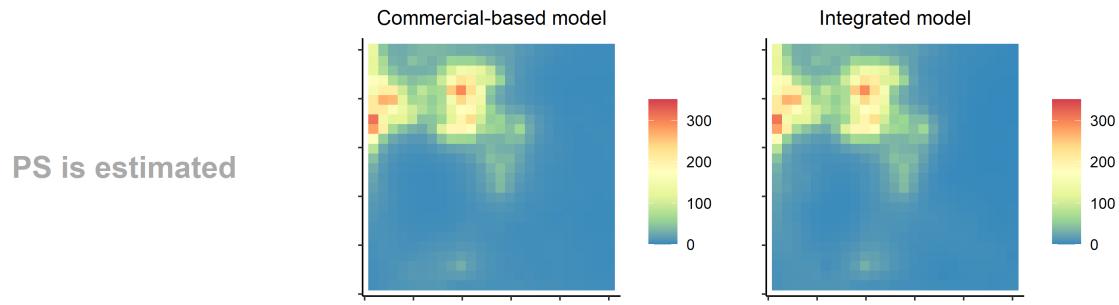
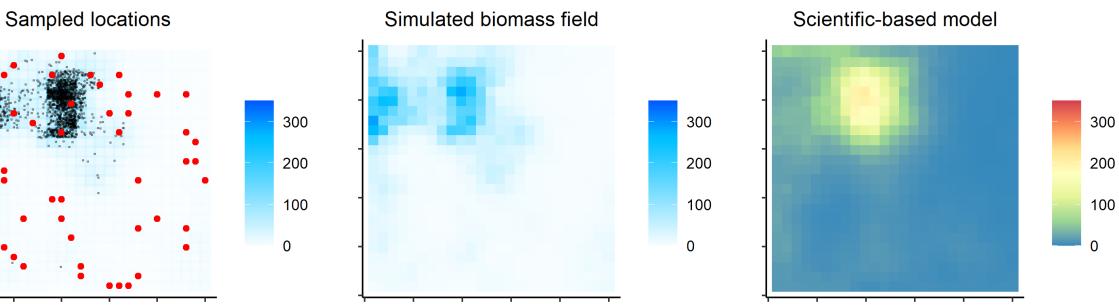


Figure B.13 – Simulated biomass field, sampled locations and model predictions of the biomass field. The number of scientific samples is fixed to 50. Red dots: scientific data points. Black dots: commercial data points. The number of commercial samples is fixed to 3000, preferential sampling is strong ( $b = 3$ ).

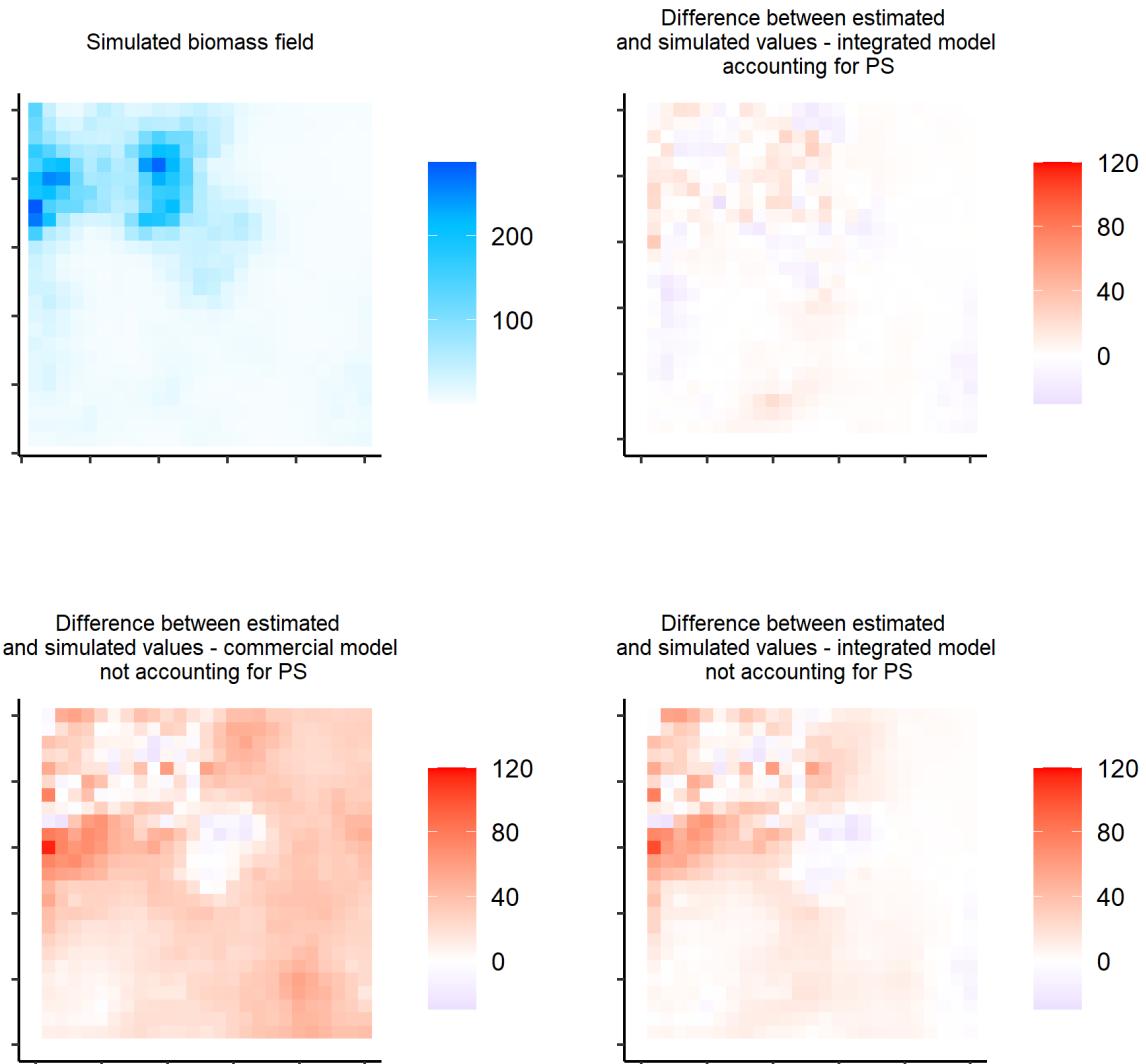


Figure B.14 – Comparison of simulated and estimated biomass field. Data is simulated with strong PS ( $b = 3$ ). Difference is given in the form (estimated - simulated).

---

### B.4.5 Fitting time for simulation-estimation

In Figure B.15, we compare the fitting time of 3 alternative model configurations regarding the sampling process:

- ‘est\_b’: the full model. The sampling process contributes to the likelihood.  $b$ ,  $\eta(x)$  and the intercept of the point process  $\alpha_{X_j}$  are estimated.
- ‘no\_samp\_process’: the sampling process does not contribute to the likelihood.
- ‘fix\_b’: an intermediate case where the sampling process contributes to the likelihood but  $b$  is fixed to 0 (only  $\eta(x)$  and  $\alpha_{X_j}$  are estimated). By comparing its computation time to ‘est\_b’ (the full model), we assess what is the computational cost of estimating the relationship between the sampling process and the biomass process.

Accounting for preferential sampling in addition to other additional processes (‘est\_b’ vs. ‘no\_samp\_process’) in inference multiplies by 4 the fitting time compared with a standard model which does not account for the sampling process. Furthermore, the estimation of the parameter  $b$  is quite expensive as the fitting time between ‘fix\_b’ and ‘est\_b’ is multiplied by 1.5 on average.

Finally, the main difference in computation time does not seem to come from the amount of data that is fitted but from the complexity of the model which is fitted. No strong differences in fitting time can be evidenced when comparing the survey-based model (50 scientific samples) with the commercial-based model (3000 commercial observations) or the IM (3000 commercial + 50 scientific samples) not accounting for PS (‘no\_samp\_process’).

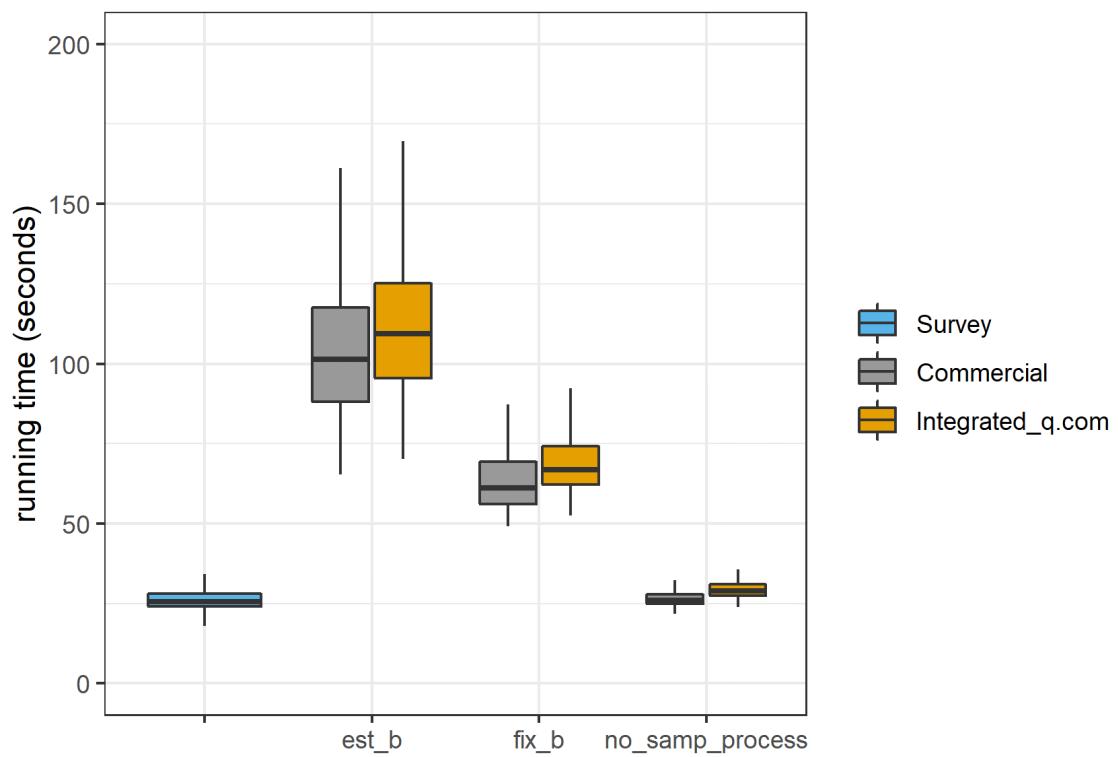


Figure B.15 – Simulation-estimation. Fitting time in seconds for different model configurations. The number of scientific samples is fixed to 50. The number of commercial samples is fixed to 3000.

---

## B.5 Case studies results

This supplementary material provides all detailed results that concern the case studies.

### B.5.1 Consistency check for each case study

Table B.2: p-values obtained from the consistency check on fixed parameters (consistency\_fixed) and random effects (consistency\_random).

Species	Year	Fleet	PS	Consistency_fixed	Consistency_random
Squids	2015	1.fleet	b_est	0.0675	1.0000
Squids	2014	1.fleet	b_est	<b>0.0001</b>	1.0000
Squids	2015	1.fleet	b_fix	<b>0.0483</b>	1.0000
Squids	2014	1.fleet	b_fix	<b>0.0000</b>	1.0000
Hake	2015	1.fleet	b_est	0.2704	1.0000
Hake	2014	1.fleet	b_est	0.5410	1.0000
Hake	2015	1.fleet	b_fix	0.2704	1.0000
Hake	2014	1.fleet	b_fix	0.8435	1.0000
Sole	2017	1.fleet	b_fix	0.6988	1.0000
Sole	2018	1.fleet	b_fix	0.0777	<b>0.0000</b>
Sole	2017	2.fleets	b_est	0.6218	1.0000
Sole	2018	2.fleets	b_est	0.1977	<b>0.0000</b>
Sole	2017	1.fleet	b_est	0.5001	1.0000
Sole	2018	1.fleet	b_est	0.2047	<b>0.0000</b>

fleet: number of fleets considered in the model;

PS: estimation of the preferential sampling parameter,

b\_est – preferential sampling is accounted for and b is estimated,

b\_fix – preferential sampling is ignored and b is fixed to 0.

Some case studies have a p-value below 0.05 for one of the two statistical test. The case studies with a very low p-value for one of the two tests are not analyzed in this work (sole in 2018, Loliginidae in 2014). In 2015, the squids case study has relatively low p-values ( $\sim 0.05$ ) for the consistency check on fixed parameters (when ignoring PS it is inferior to 0.05 and when accounting for PS it is slightly above 0.05).

### B.5.2 Species-habitat relationship

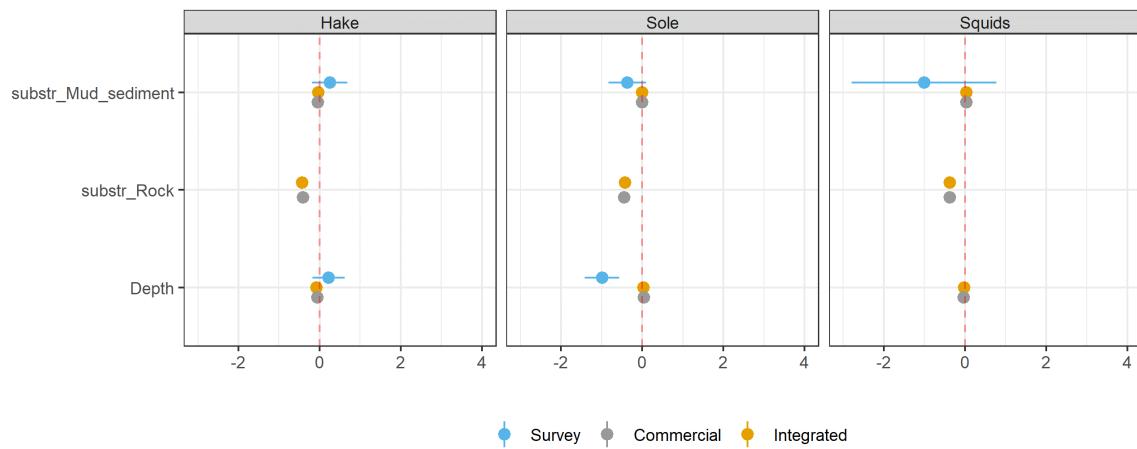


Figure B.16 – Estimates of the species-habitat relationship estimates for each case study. The bars represent 95% confidence intervals. Both commercial-based and integrated models account for PS. For both models, some standard deviation estimates are very low and thus confidence intervals are very tiny and do not clearly appear graphically. Baseline: sole – sand and  $] -\infty, -50]$  m, Hake – sand and  $] -\infty, -100]$  m, squids – sand and  $] -100, 0]$  m substrate and the deeper depth class. substr: substrat type. Depth: depth effect, it corresponds to the modality  $] -50, 0]$  m for sole,  $] -100, 0]$  m for hake and  $] -\infty, -100]$  m for squids. There is no rock substrate for scientific models as the two surveys do not sample on rocky habitats. For squids, the estimates of the depth factor level falls outside the plot and is hardly estimated with scientific data only (the standard deviation is very high) thus it was not presented on the plot.

For sole and squids, the scientific-based predictions are shaped by the covariates effect (see also SM B.5.3 for the relative contribution of the covariates and random effect in biomass distribution) even though estimates of the species-habitat relationship are quite

---

uncertain. The depth effect in particular shapes the squid and the sole spatial distribution (see Figure 3.6 in the main text for more evidence).

Conversely, for the integrated models estimates the species-habitat relationship are very close to 0 and the variability of biomass distribution is mainly captured by the random effect (SM B.5.3).

### B.5.3 Proportion of variance of the latent field ( $\log(S(x))$ ) explained by the random effect $\delta(x)$ and the covariates ( $\Gamma_S(x)^T \cdot \beta_S$ )

Table B.3: Proportion of variance explained in  $\log(S(x))$  by the random effect  $\delta(x)$  and the covariates for each model and each case study (Cf. eq. 3.1 in the main manuscript)

Species	Year	Model	$\delta(x)$	Covariates
Hake	2014	Commercial	0.98	0.02
Hake	2014	Integrated	0.96	0.04
Hake	2014	Survey	0.88	0.12
Sole	2017	Commercial	0.87	0.13
Sole	2017	Integrated	0.88	0.12
Sole	2017	Survey	0.00	1.00
Squids	2015	Commercial	0.99	0.01
Squids	2015	Integrated	1.00	0.00
Squids	2015	Survey	0.10	0.90
n.b. Both commercial-based and integrated models account for PS.				

### B.5.4 Spatial correlation parameter

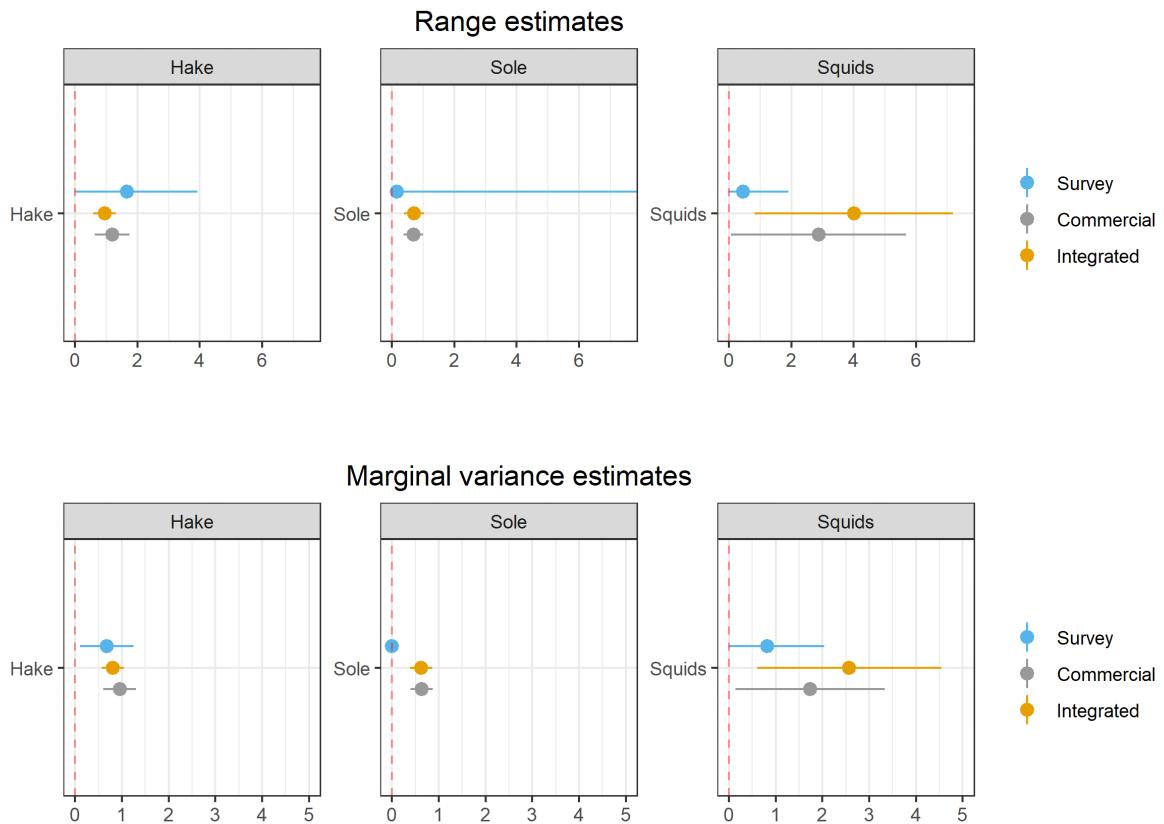


Figure B.17 – Range and marginal variance estimates of the latent field random effect  $\delta(x)$  for each case study. The commercial-based and the integrated model account for PS. Uncertainty is given in 95% confidence intervals (for the sole case study, the standard errors were poorly estimated - either a NA or very large value - and are not plotted on the figure).

For hake, the range and the marginal variance of the integrated model falls within the confidence interval of the range estimated with scientific data alone. For squids, the range of the integrated model falls outside the confidence interval of the survey-based model. This can be related to the relatively low p-value of the consistency test for fixed effects (SM B.5.1).

For sole and squids, both parameters estimates are very low (i.e. very flat and un-

---

structured spatial random effect) and for sole especially they are both poorly estimated (standards deviations have either very large or NA values). In this cases, the survey data allow to capture the covariate effect but not the local spatial correlation structures. All the variability of the predictions will be driven by the covariates effects (see Figure 3.6 in the main manuscript, SM B.5.3).

---

### B.5.5 Information brought by commercial data

The following plots present log-predictions of biomass and parameters estimates of the integrated models versus the ones of the models fitted to scientific data only (Figure B.18 - Figure B.19). They inform what brings the commercial data to inference in comparison with scientific data and they should be considered separately from the consistency check presented above.

If all the points of the following plot follow the  $x = y$  axis, then commercial data do not bring any information to inference and thus is useless.

On the contrary, some deviation from the  $x = y$  axis emphasizes that commercial data bring new information to inference. Whether this additional information is misleading is informed by the p-values of the consistency checks (SM B.2). If the p-values of the tests are above 0.05, then the information brought by commercial data should not be rejected as the fixed/random parameters of the integrated model fall within the confidence interval of the scientific-based model.

As mentioned in the ‘results’ section, there are strong differences between predictions from the integrated models and the scientific-based models emphasizing commercial data bring lot of information to inference.

Regarding estimates, the parameters related to the shape of the random effect  $\delta(x)$  (logtau\_lf and logkappa\_lf) are different in the integrated and the scientific-based model as commercial data strongly revise the overall shape of the random effect.

This is particularly true for the sole case study; predictions based on scientific data alone mainly match covariates levels (see Figure 6 in the main text to get a better idea of sole spatial distribution). By contrast, commercial data better capture spatial correlation structures due to a much higher sampling effort and to a tighter sampling (see Figure 3.3 in the main text).

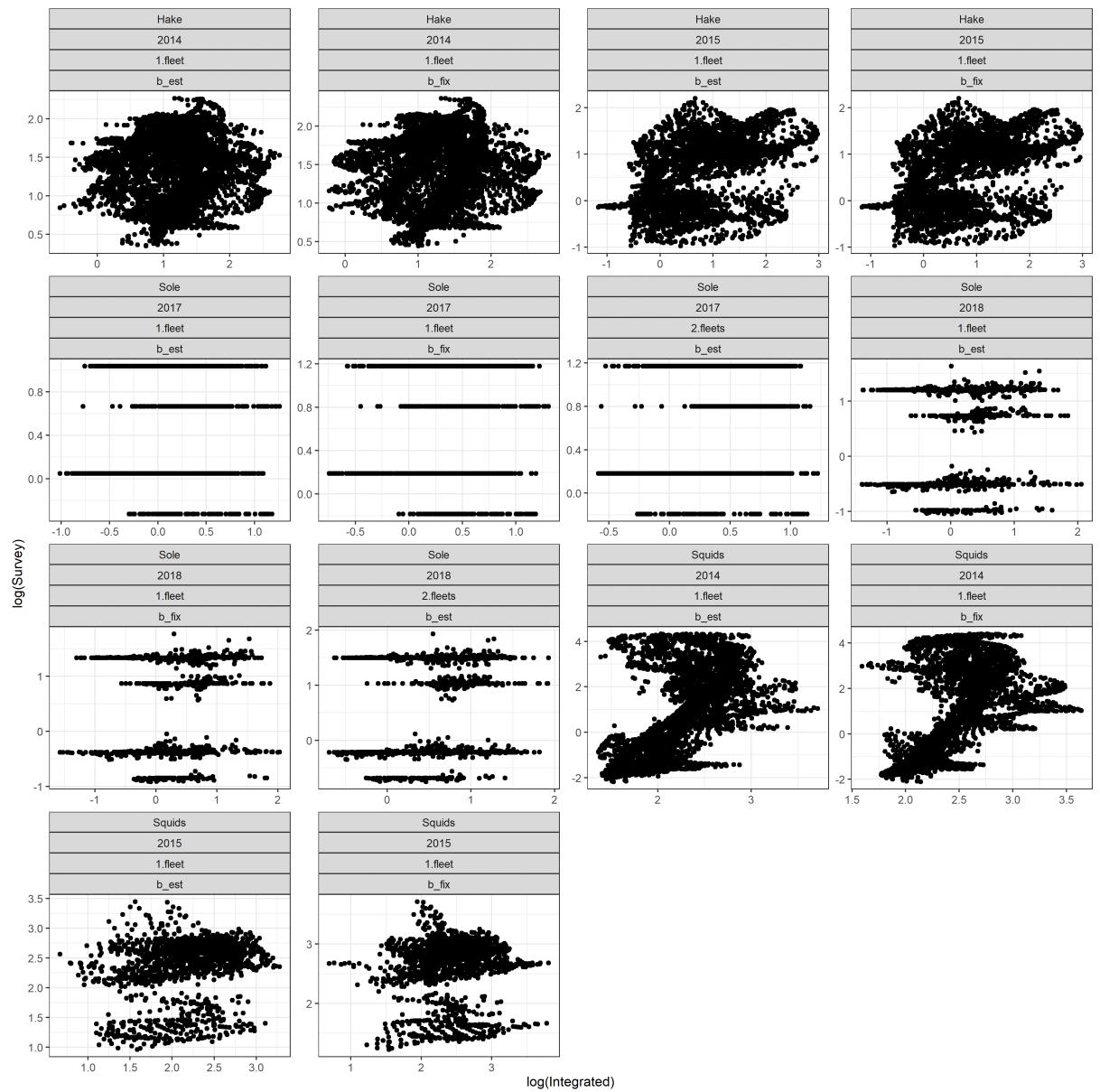


Figure B.18 – Log biomass predictions of the integrated model against log biomass predictions of the survey-based model. Title 1<sup>st</sup> line: species name. Title 2<sup>nd</sup> line: year. Title 3<sup>rd</sup> line: number of fleets considered in the model; Title 4<sup>th</sup> line: estimation of the preferential sampling parameter, b\_est – preferential sampling is accounted for and b is estimated, b\_fix – preferential sampling is ignored and b is fixed to 0;

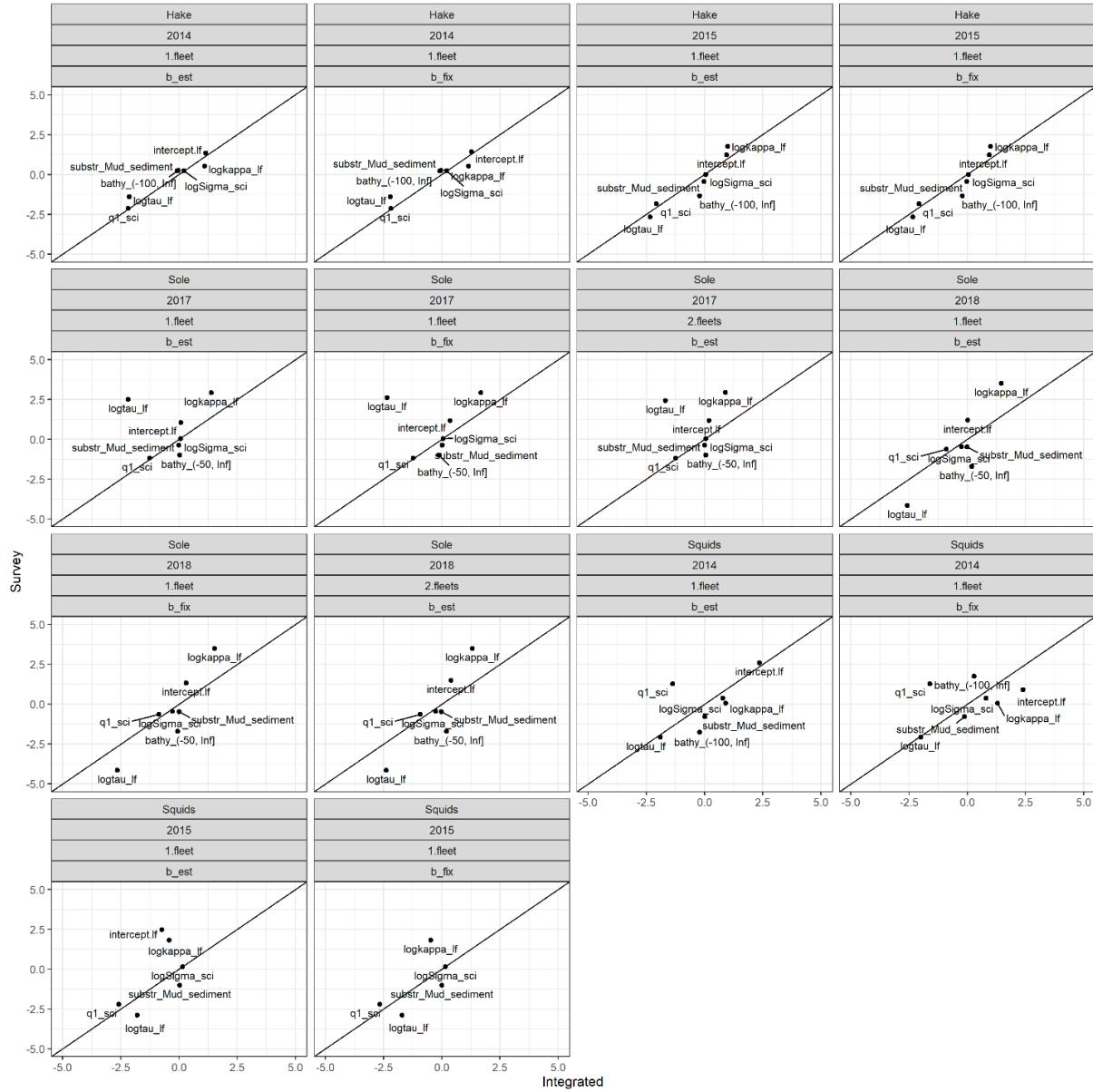


Figure B.19 – Estimates of the integrated model against estimates of the survey-based model. Black line:  $x = y$ . Title 1<sup>st</sup> line: species name. Title 2<sup>nd</sup> line: year. Title 3<sup>rd</sup> line: number of fleets considered in the model; Title 4<sup>th</sup> line: estimation of the preferential sampling parameter, b\_est – preferential sampling is accounted for and  $b$  is estimated, b\_fix – preferential sampling is ignored and  $b$  is fixed to 0;

---

### B.5.6 Contribution of scientific data to the integrated model spatial predictions

This figure should be compared with Figure 3.3 in the main text which provides the spatial distribution of sampling for both commercial and scientific data.

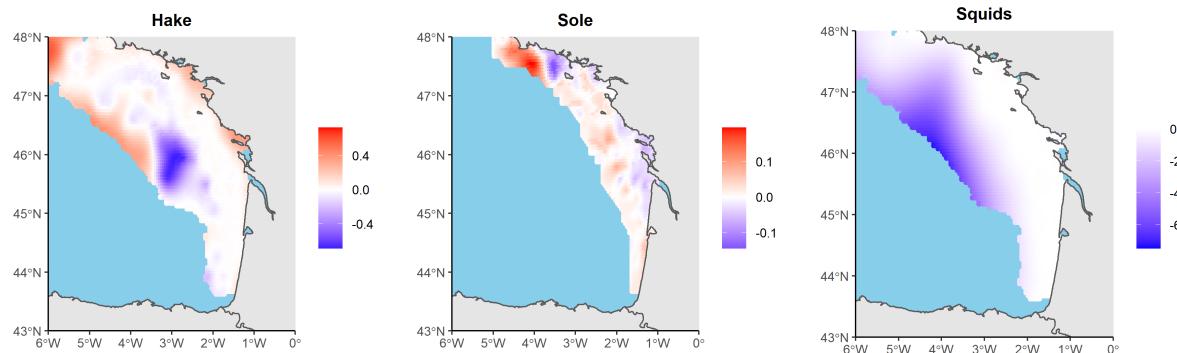


Figure B.20 – Contribution of scientific data to spatial predictions for each case study. The plot represents the relative difference between the integrated model and the commercial-based model accounting for preferential sampling. The relative difference is under the form:  $(\text{integrated} - \text{commercial}) / \text{integrated}$ . Blue scale: areas downscaled by scientific data, red scale: areas where biomass estimates increases when integrating scientific data.

### B.5.7 Targeting metric $T_j(x)$ maps

Following plots illustrate the targeting metric for each case study (described in SM B.1.3). It provides a dimensionless metric to quantify how much the cells are over/under sampled compared with a situation where PS is null.

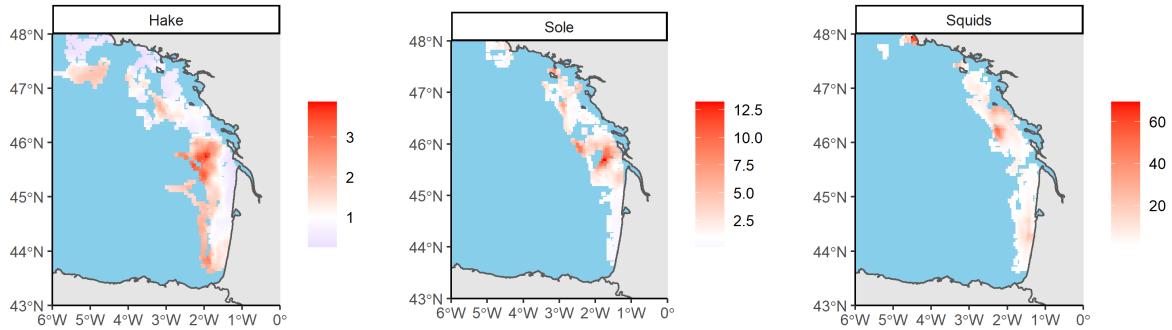


Figure B.21 – Targeting metric  $T_j(x)$  from the integrated model for each case study. Blue:  $T_j(x) < 1$ . Red:  $T_j(x) > 1$ . Only the areas where the commercial fleets sampled at least once are used to plot the targeting metric as the targeting metric is only meaningful in these areas.

Here consistently with  $b$  estimates, the squids case study emphasizes strong PS ( $T_j(x)$  goes up to 100 in some areas) followed by sole ( $T_j(x)$  goes up to 12.5) and hake ( $T_j(x)$  goes up to 4, the values are relatively small compared to the other case studies which highlights a low PS behavior).

---

### B.5.8 Random effect $\eta_j(x)$ of the sampling process

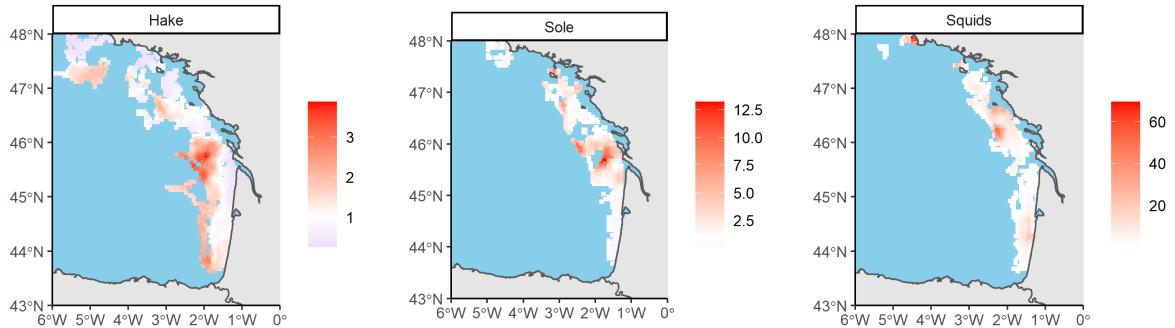


Figure B.22 – Sampling process random effect  $\eta$  for each case study. Model: integrated model accounting for PS and assuming 1 fleet.

---

### B.5.9 Proportion of variance of fishing intensity ( $\log(\lambda_j(x))$ ) explained by the random effect $\eta_j(x)$ and the preferential sampling term $b_j \cdot \log(S(x))$ .

Table B.4: Proportion of variance explained in  $\log(\lambda_j(x))$  by the random effect  $\eta_j(x)$  and the preferential sampling term  $b_j \cdot \log(S(x))$  for each model and each case study (Cf. eq. 3.4 in the main text).

Species	Year	Model	$\eta_j(x)$	$b_j \cdot \log(S(x))$
Hake	2014	Commercial	0.94	0.06
Hake	2014	Integrated	0.97	0.03
Sole	2017	Commercial	0.84	0.16
Sole	2017	Integrated	0.84	0.16
Squids	2015	Commercial	0.41	0.59
Squids	2015	Integrated	0.25	0.75

n.b. Only the results obtained with the 1-fleet models accounting for PS are presented

---

### B.5.10 Comparison of goodness-of-fit and predictive capacity metrics for models accounting or not for preferential sampling

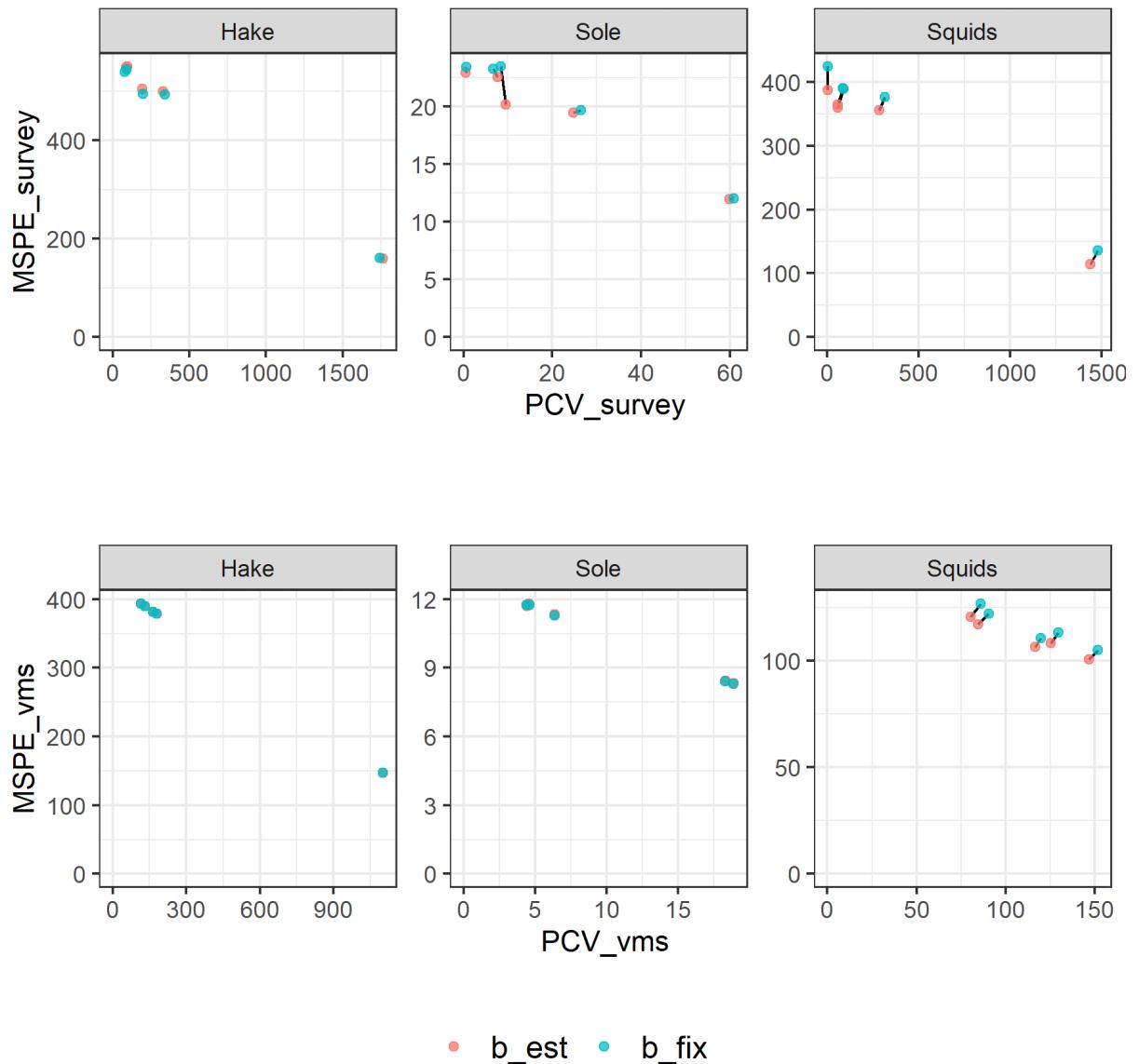


Figure B.23 – Case studies - Comparison of the validation metrics when accounting for PS ( $b$  is estimated - red) and when ignoring PS ( $b$  is fixed at 0 - blue). Only the inte-

---

grated models were used to build these plots. MSPE\_fit: mean squared prediction error, goodness-of-fit metric. PCV: predictive cross validation, predictive quality metric. The lower the values, the better the model fits/predicts the data. See SM B.3.10 for more intuition on the metrics.

---

### B.5.11 Comparison of goodness-of-fit and predictive capacity metrics for models considering 1 fleet or 2 distinct fleets

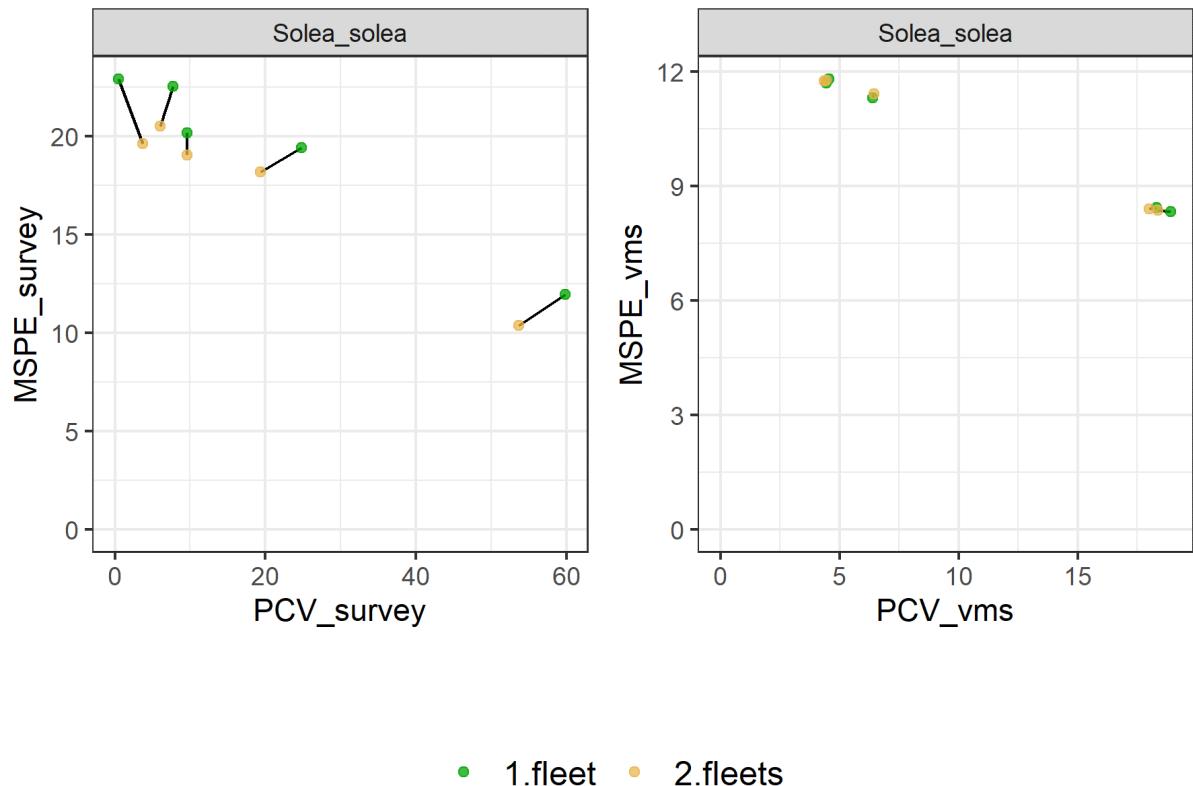


Figure B.24 – Sole case study (2017) - Comparison of the validation metrics obtained with the 1-fleet model (green) and the 2-fleets model (brown). Only the integrated models accounting for PS were used to build these plots. MSPE\_fit: mean squared prediction error, goodness-of-fit metric. PCV: predictive cross validation, predictive quality metric. The lower the values, the better the model fits/predicts the data. See SM B.3.10 for more intuition on the metrics.

---

### B.5.12 Comparison of spatial prediction for models considering 1 fleet or 2 distinct fleets

In the 1-fleet model, biomass aggregates mainly in 2 points at locations 46°N/2°W and 46°N/2.5°W. In the 2-fleet models, biomass is more evenly distributed on the study area.

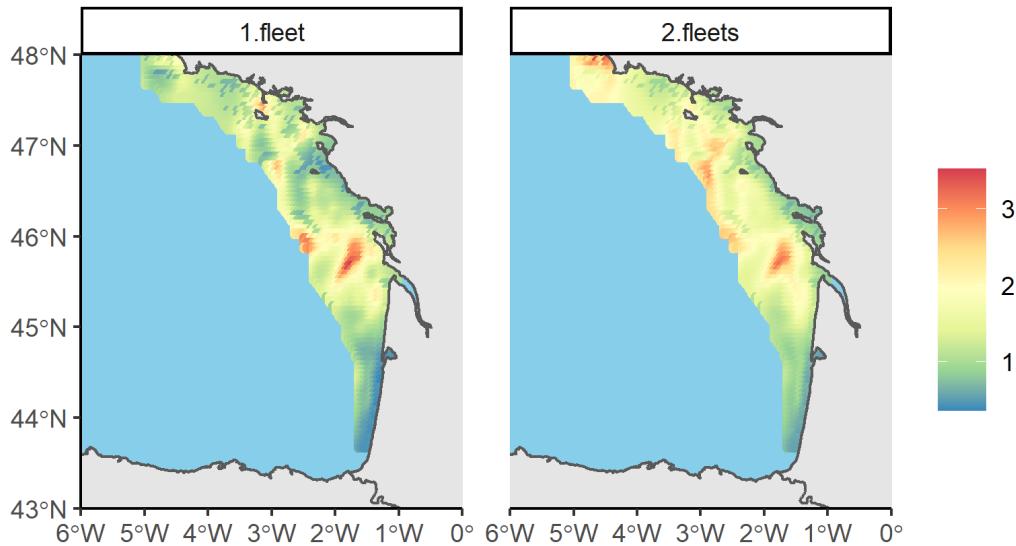


Figure B.25 – Sole case study (2017) - Estimation of the relative biomass for the 1-fleet model (left) and the 2-fleets model (right). Estimates are obtained from the integrated model accounting for PS.

### B.5.13 Maps and related uncertainty

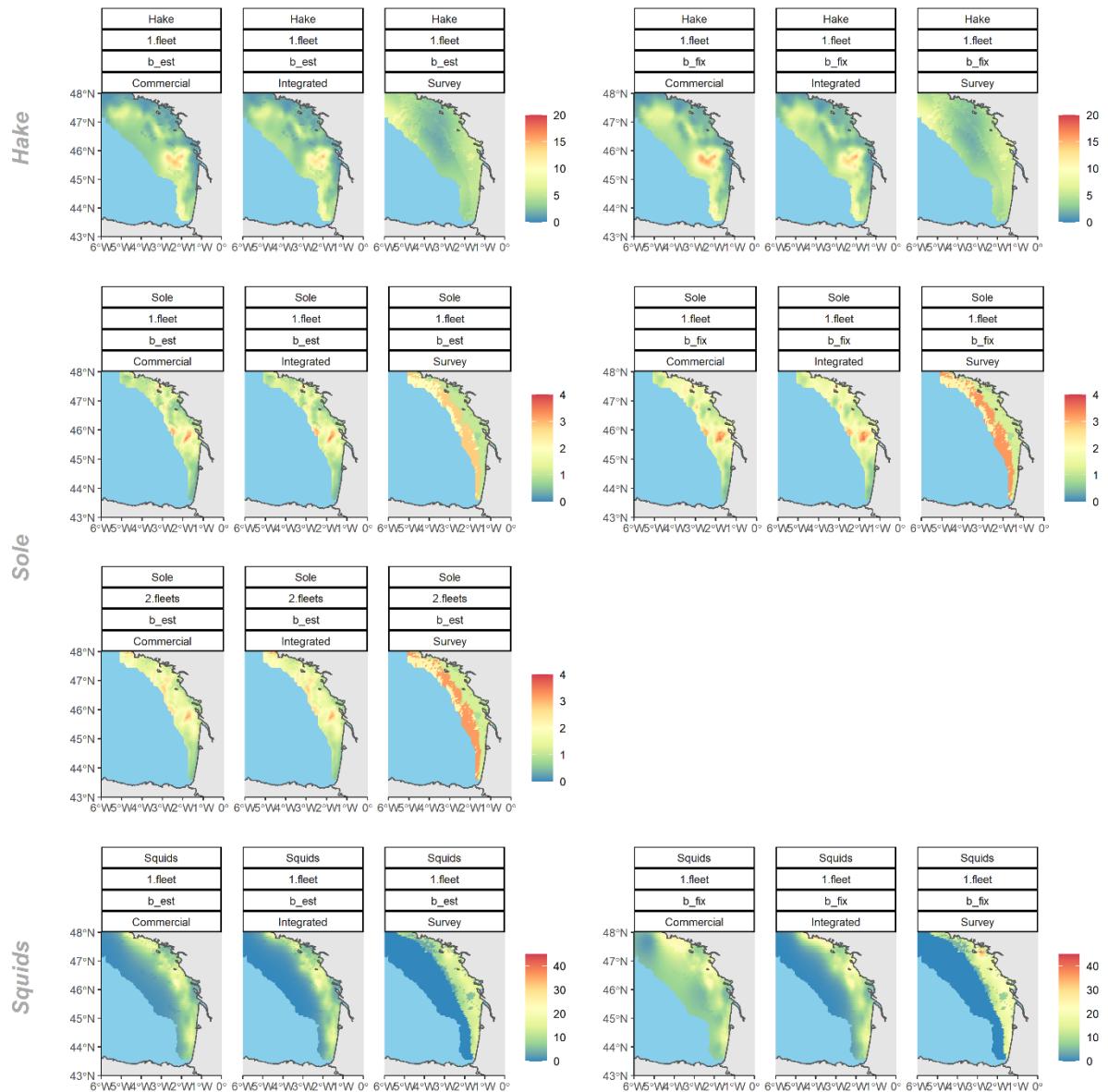


Figure B.26 – Spatial predictions obtained for all case studies and all model configurations. Title 1st line: species name; Title 2nd line: number of fleets considered in the model; Title 3rd line: estimation of the preferential sampling parameter,  $b_{\text{est}}$  – preferential sampling is accounted for and  $b$  is estimated,  $b_{\text{fix}}$  – preferential sampling is ignored and  $b$  is fixed to 0; Title 4th line: fitted data source.

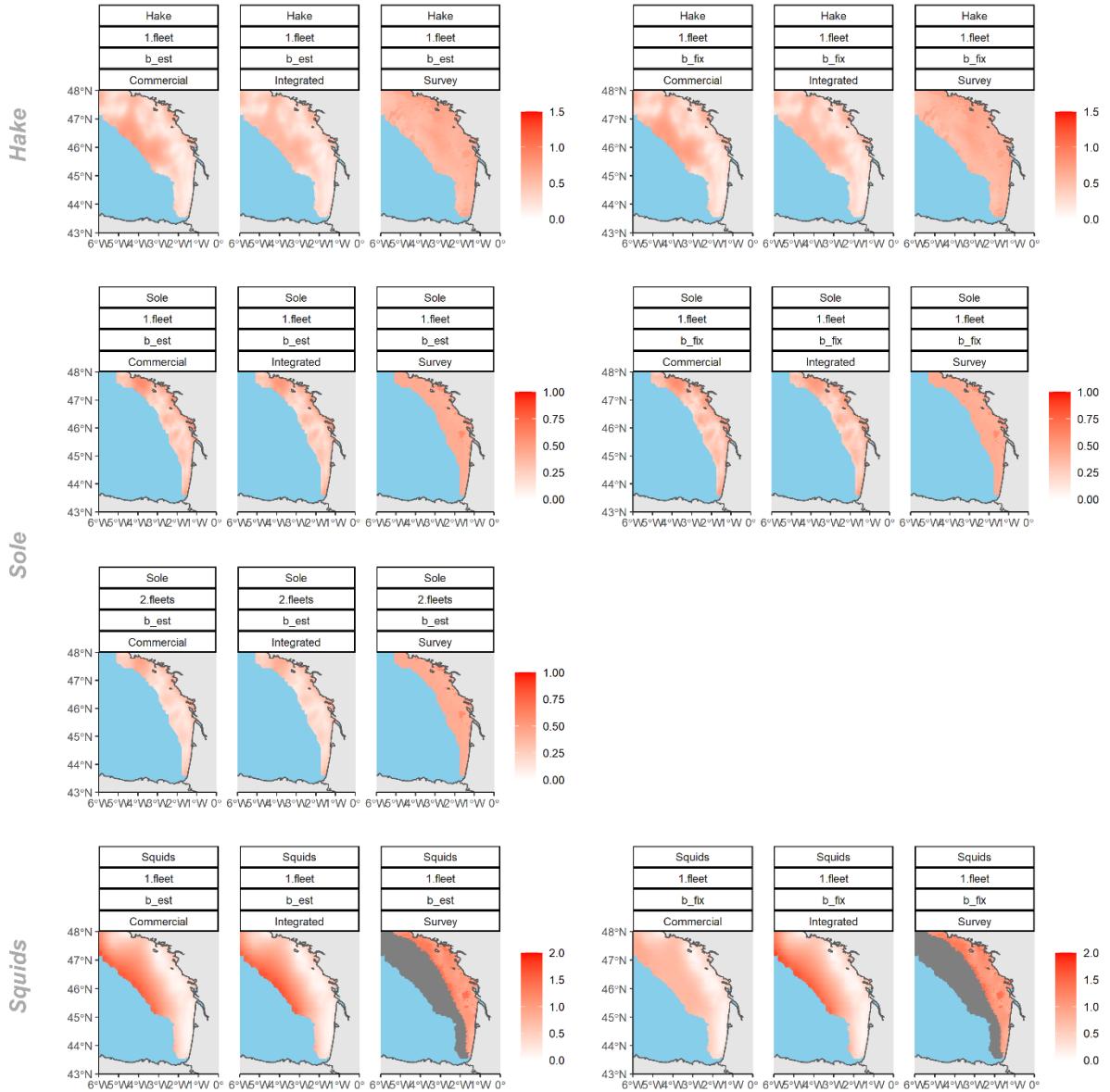


Figure B.27 – Uncertainty of the spatial predictions obtained for all case studies and all model configurations. Uncertainty is expressed in coefficient of variation. Title 1st line: species name; Title 2nd line: number of fleets considered in the model; Title 3rd line: estimation of the preferential sampling parameter, *b*\_est – preferential sampling is accounted for and *b* is estimated, *b*\_fix – preferential sampling is ignored and *b* is fixed to 0; Title 4th line: fitted data source; Grey values are none-estimated uncertainty values. Specifically for squids, all scientific observations of the deepest strata are 0 values. As

---

0 values are poorly informative of species distribution in this model, this factor level is poorly estimated and the related standard deviation are very high.

# IDENTIFYING MATURE FISH AGGREGATION AREAS DURING SPAWNING SEASON BY COMBINING CATCH DECLARATIONS AND SCIENTIFIC SURVEY DATA

---

## C.1 Filtering the mature fraction from landings

The filtering of the mature fraction of the landings is described through the formula:

$$Land_m(CC) = Land(CC) \cdot p_m(CC)$$

$$p_m(CC) = \sum_l p_l(CC, l) \cdot MO(l)$$

where  $Land$  are the landings and  $Land_m$  the mature fraction of the landings.

By merging sales notes and logbooks, a comprehensive part of the landings can be expressed by commercial size categories  $CC$  so as landings can be written as  $Land(CC)$  (SACROIS - Demanèche et al. (2013)). These commercial categories are regularly sampled to derive length structure of each commercial category. These data are often used in stock assessment routines to obtain catch-at-length data (ICES, 2017). The proportion of length class  $l$  within commercial category  $CC$  is denoted  $p_l(CC, l)$ . To compute the mature proportion of the corresponding commercial category  $p_m(CC)$ ,  $p_l(CC, l)$  is combined with the proportion of mature individuals  $MO(l)$  for a specific length class  $l$  available through maturity ogive. Once the mature proportions per commercial category are available, they can be crossed with landings  $Land(CC)$  to obtain the mature fraction of the landings

---

$Land_m(CC)$ .

In the case studies, auction data are taken from ObsVentes data (Vigneau, 2009). The sampling of these data is designed on a quarterly basis. The maturity ogive is assumed constant over the full period and the size distribution within each catch category is assumed to vary on a quarterly time step (sampling of demographic data is designed by quarters). Overall, the proportions of mature individuals within each commercial category fall between 75% to 100% for both sole and whiting (see Figure C.1). When data are lacking for a specific quarter (because sampling is missing for certain quarters and commercial category), the missing data is replaced by the average of the mature proportion for the related catch category over the period.

For sole, 742 individuals were used to compute the maturity ogive, 2357 individuals were used for whiting. Maturity stage are assessed by visual assessment (macroscopic assessment). The commercial classes are considered relatively stable over time (Figure C.1 – there is no strong variation of size between each category).

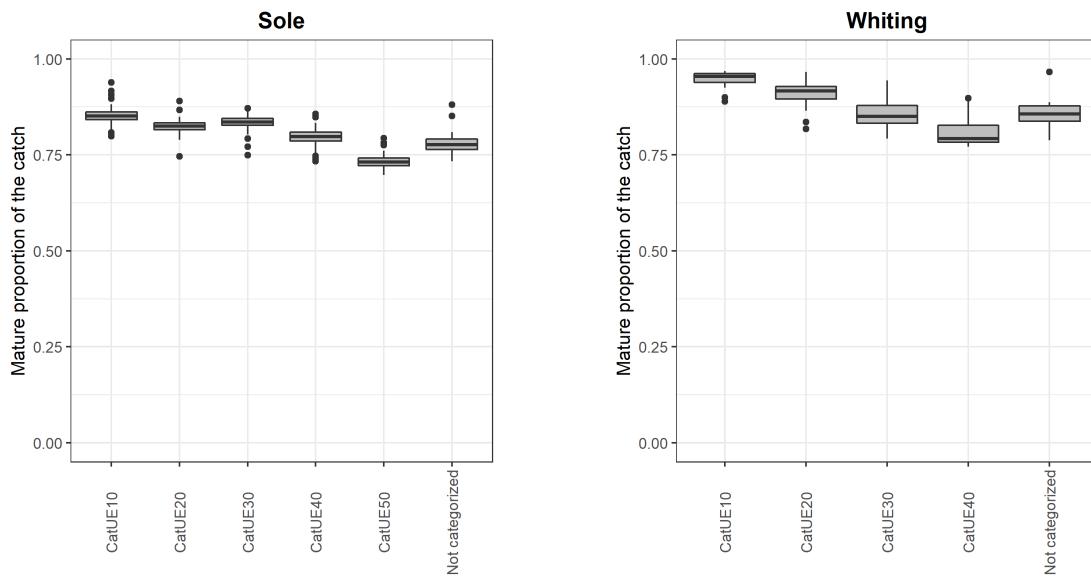


Figure C.1 – Mature proportion of the catch per commercial size category on the whole time series (each point is the record of a quarter). Only the commercial categories where more than 10 individuals have been observed are plotted.

---

## C.2 Discretization grid

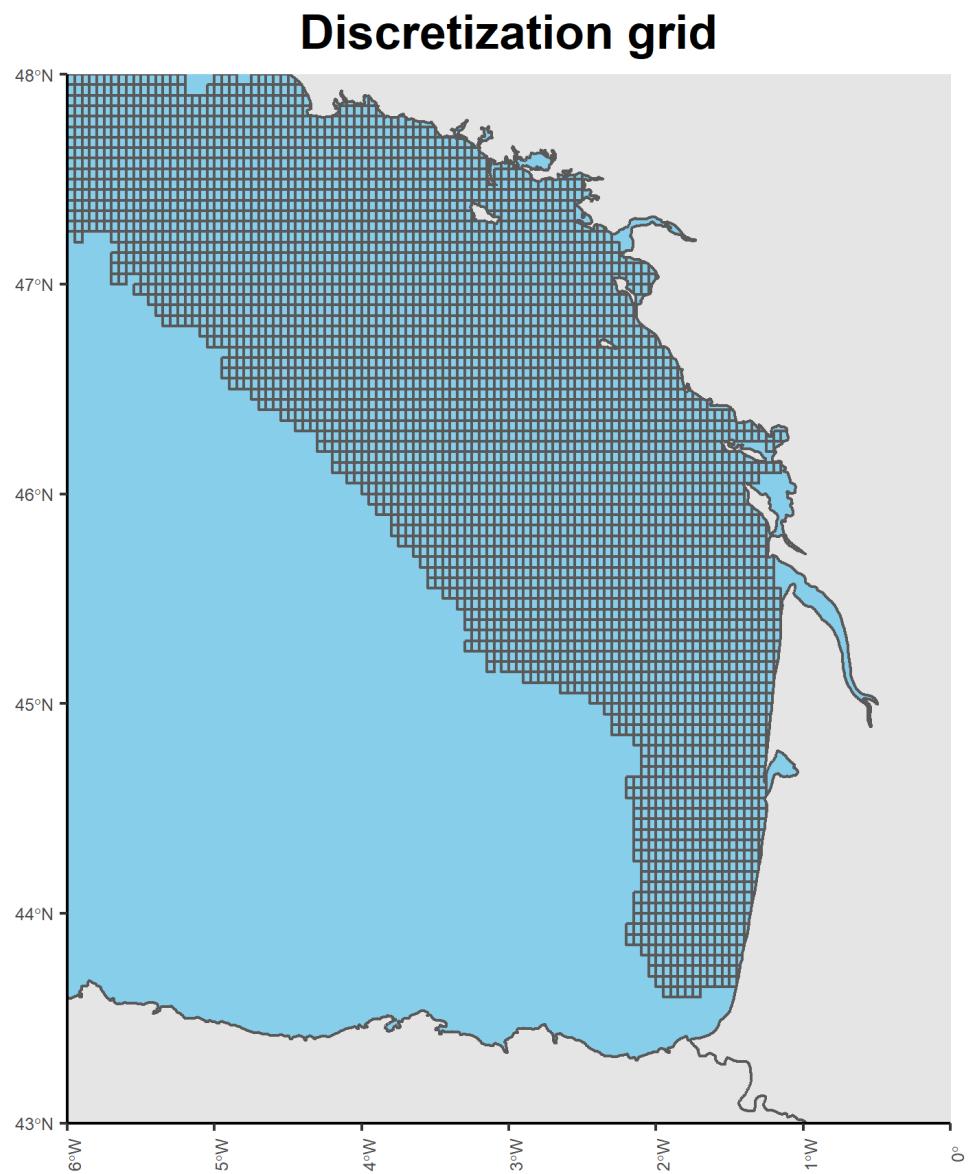


Figure C.2 – Grid used to discretize commercial data, integrate the point process constant and compute the biomass predictions. Resolution:  $0.05^\circ \times 0.05^\circ$ .

---

### C.3 Survey sampling locations

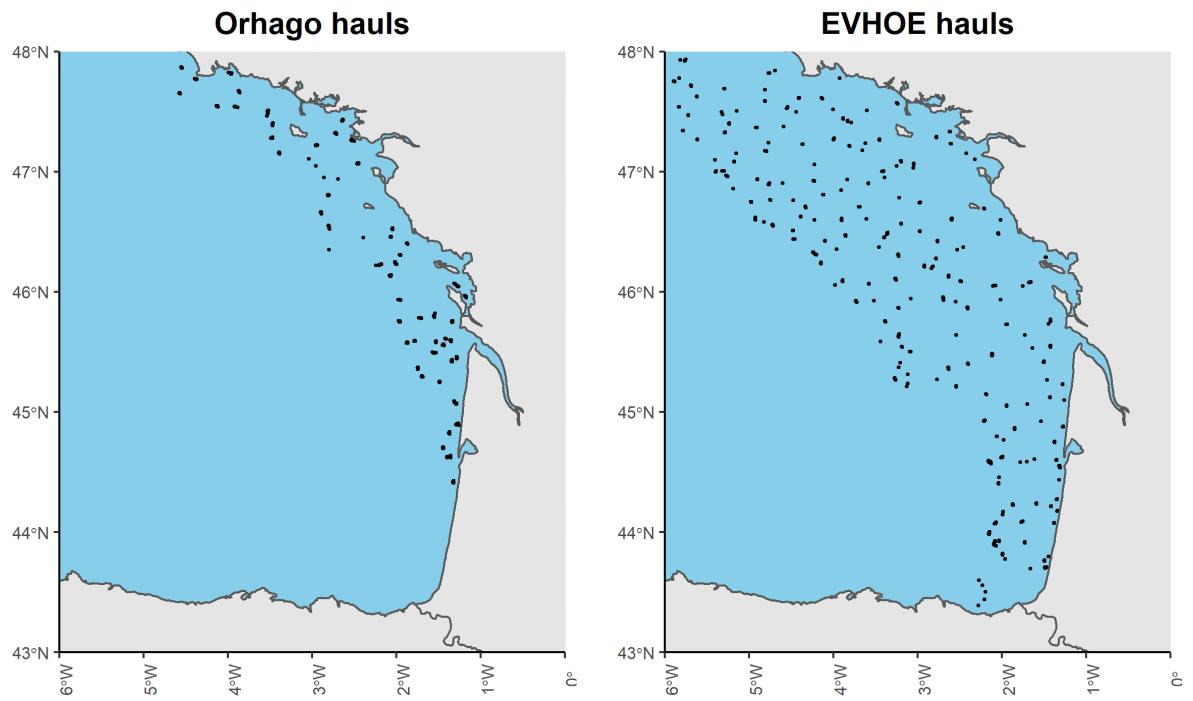


Figure C.3 – Spatial distribution of the fishing hauls of the Orhago and the EVHOE surveys from 2010 to 2018.

---

## C.4 The SPDE approach

Estimating  $\Sigma$  is performed through the SPDE approach, which proves efficient to estimate correlation among points when the size of the covariance matrix becomes large.

Estimating  $\Sigma$  at every spatial location can be computationally challenging when the dimension of  $\Sigma$  increases. A solution to overcome this computational burden was provided by Lindgren, Rue, and Lindström (2011) through the SPDE approach. Instead of modeling the random effect as a GRF, the random effect  $\omega$  is represented as a Markovian representation of the GRF (GMRF) on the nodes of a triangulated mesh (e.g. see Figure C.4).  $\omega \sim GF(0, \mathbf{Q}^{-1})$ , with  $\mathbf{Q}$  the precision matrix which benefits from the sparse property of GMRF. The link between the random effect values estimated at each mesh node and the observations on the continuous space (defined on  $\mathcal{D} \subset \mathbb{R}^2$ ) is realized through linear interpolation. For extended details on the SPDE approach, GRF and the Matérn function, refer to Lindgren, Rue, and Lindström (2011) and Cameletti et al. (2013).

To compute fine-scale spatial predictions of the biomass field, the latent field values obtained at the mesh nodes are interpolated on a discrete grid with much finer resolution (see SM C.2, Figure C.2) as done in other packages such as VAST (Thorson, 2019).

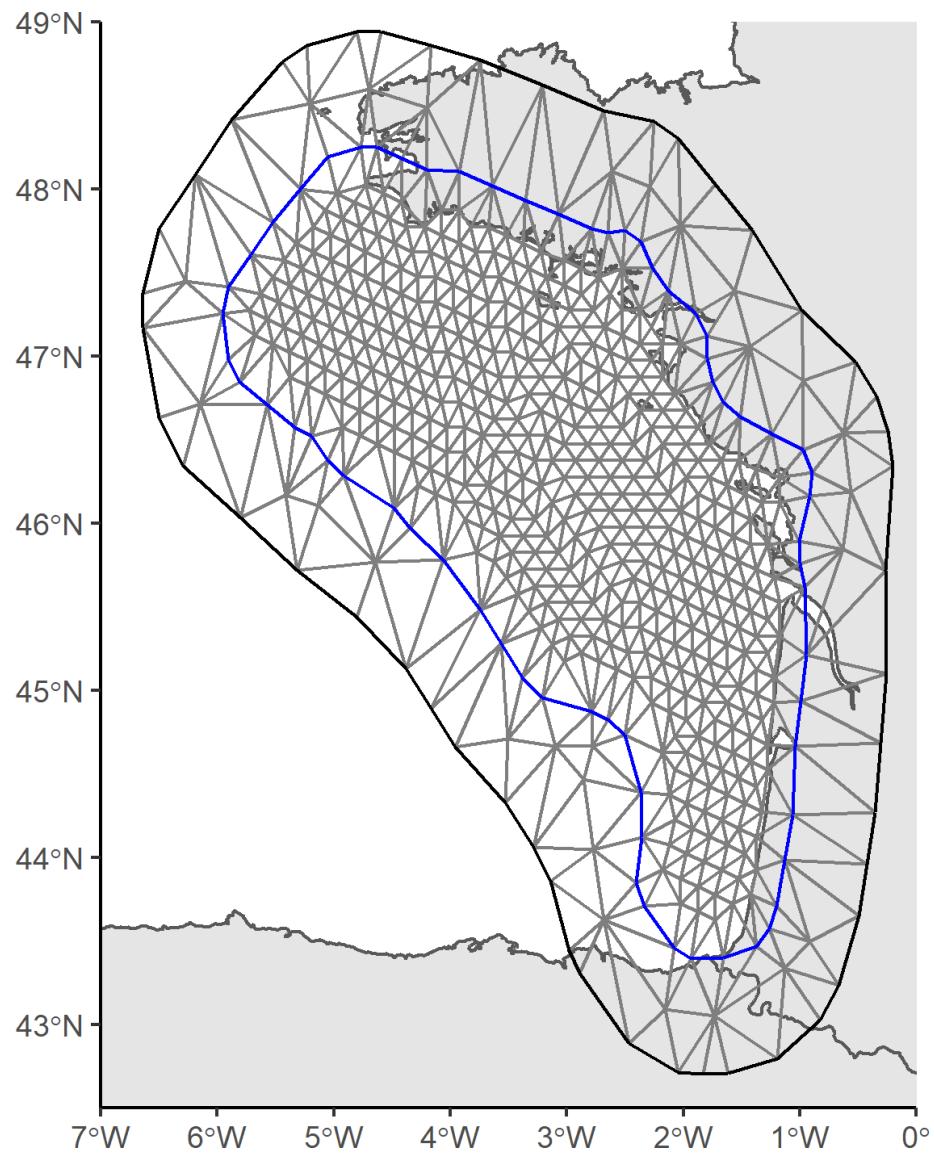


Figure C.4 – Mesh used to estimate the spatio-temporal random effects  $\epsilon(x, t)$  and  $\eta(x, t)$ .  
The mesh was computed with the R-INLA package.

---

## C.5 Estimating the point process

The log-likelihood of the point process is expressed as  $\log(\pi[\mathbf{X}_{com,j}]) = \sum_{i=1}^m \log(\lambda_j(x_i) - \int_{\mathcal{D}} \lambda_j(x) dx)$  with  $m$  the number of observations. The term  $\int_{\mathcal{D}} \lambda_j(x) dx$  (also called the ‘normalization constant’) cannot be calculated explicitly and must be computed numerically (Renner et al., 2015). A common procedure consists in integrating  $\lambda_j(x, t)$  over the study domain through a method referred as ‘quadrature’; a set of ‘quadrature points’ are selected on the area and are used to approximate the log-likelihood through a weighted sum of the intensity  $\lambda_j$  at each quadrature points. Concretely, the log-likelihood is re-expressed as  $\log(\pi[\mathbf{X}_{com,j}]) \approx \sum_{i=1}^m \log \lambda(x_i) - \sum_{k=1}^n w_k \cdot \lambda(x_k)$  with  $\mathbf{w} = w_1, \dots, w_n$  the quadrature weights and  $x_k, k \in [1, n]$  the related quadrature points. For simplicity, we opted for a fine regular quadrature which makes all  $w_k$  equal (the quadrature points are the centroid of the cell grid of Figure C.2). Note that more time-efficient methods exist (Jullum, 2020; Simpson et al., 2016), but will not be explored here as the one proposed by Renner et al. (2015) is a simple, stable and standard method for estimating the normalization constant.

## C.6 Maximum likelihood estimation

As the random effects are in the logscale, the estimates of the biomass field and sampling intensity may be biased. We used the epsilon bias-correction to mitigate this bias (Thorson and Kristensen, 2016). Standard deviations were computed through the  $\delta$ -method. However, both methods imply an increasing computation time as number of time steps and the number/size of random effects increases. For this reason, bias correction and  $\delta$ -method were performed only for the biomass field values we explicitly map in the results.

Still, fitting the model on the full time series can be computationally intensive. This is particularly true when the number of time steps and the number of fleets increases. To overcome this issue, we considered each year as a block and fitted each year (12 time steps) separately before merging the outputs of each block to reconstruct the full time series. We used the Datarmor supercomputer (Ifremer, 115 Gb available for each node, sequential fitting) to fit each block.

## C.7 Biomass predictions and related coefficient of variation for November 2018

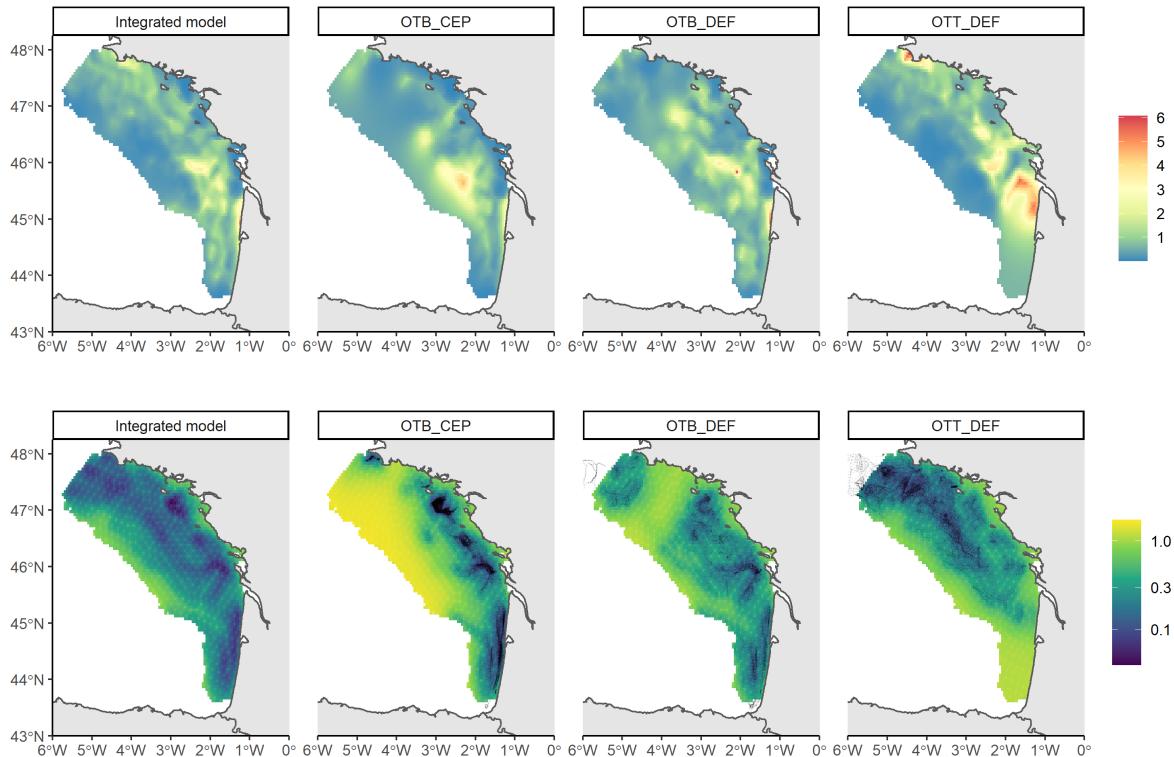


Figure C.5 – Sole case study. Biomass field spatial predictions (top) and related coefficient of variation (bottom) obtained for November 2018. Integrated model: model combining all data sources. OTB\_CEP: model fitted to OTB\_CEP fleet. OTB\_DEF: model fitted to OTB\_DEF fleet. OTT\_DEF: model fitted to OTT\_DEF fleet. Black dots: fishing pings of each fleet.

Integrating the data from all the fleets allows a better coverage of the whole area and provide more accurate predictions on the full study domain. Predictions based on single-fleet models have high standard errors outside the fleet's sampling area while, when all fleets are integrated, standard deviation is drastically reduced in these areas (Figure C.5).

---

## C.8 Spatial predictions with and without PS (November 2018)

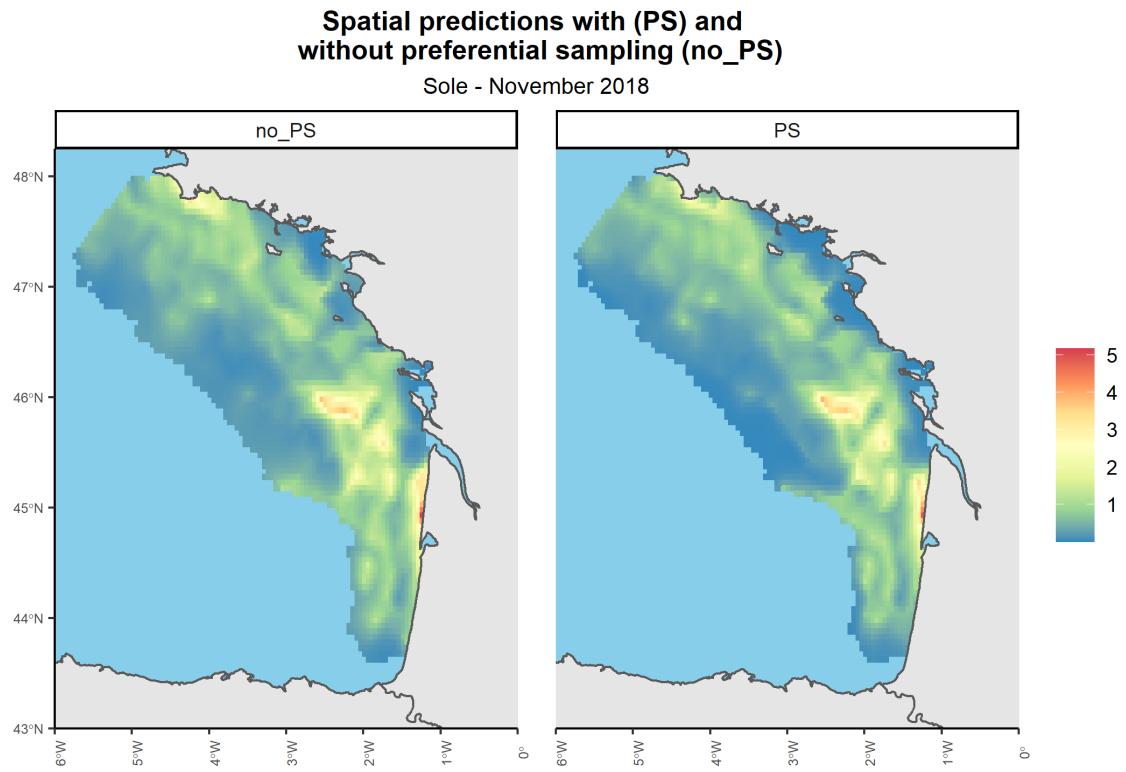


Figure C.6 – Sole. Biomass predictions in November 2018 for an integrated model accounting for PS or not (no\_PS). Unit: CPUE in kg/hour.

---

**Spatial predictions with (PS) and  
without preferential sampling (no\_PS)**  
Whiting - November 2018

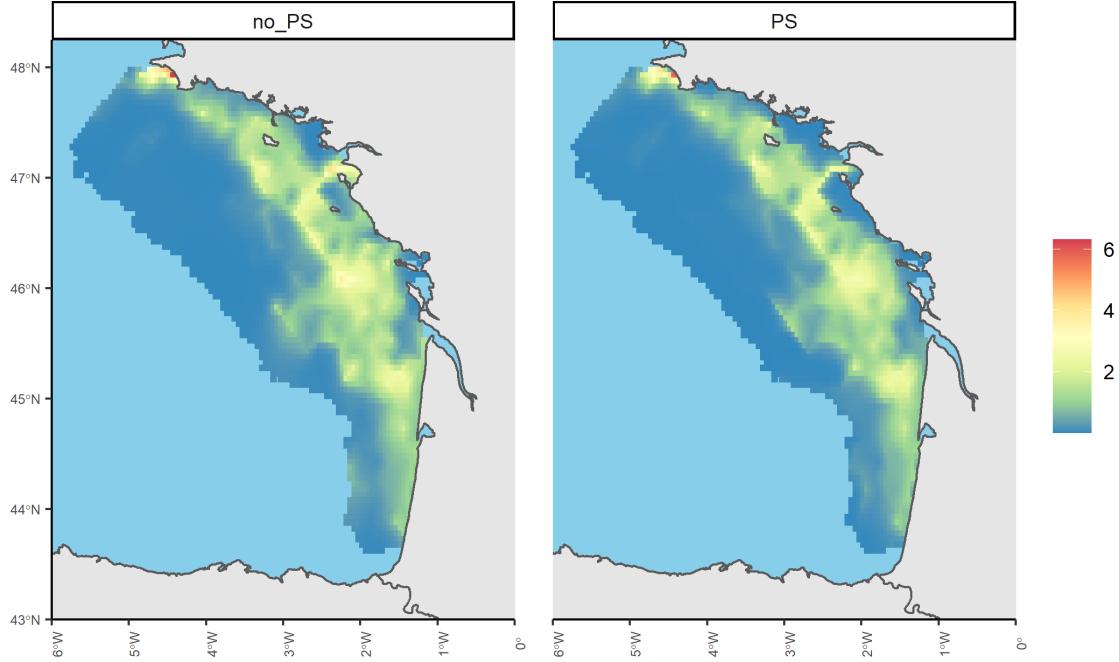


Figure C.7 – Whiting. Biomass predictions in November 2018 for an integrated model accounting for PS or not (no\_PS). Unit: CPUE in kg/hour.

---

**Spatial predictions with (PS) and  
without preferential sampling (no\_PS)**  
Squids - November 2018

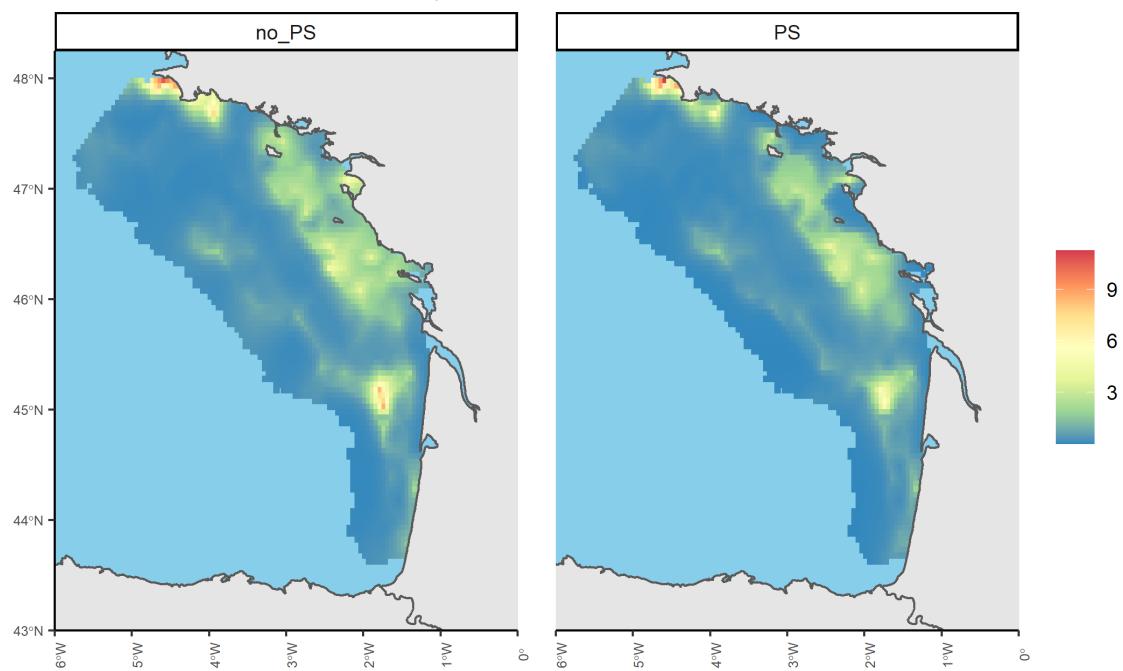


Figure C.8 – Squids. Biomass predictions in November 2018 for an integrated model accounting for PS or not (no\_PS). Unit: CPUE in kg/hour.

---

## C.9 Monthly average biomass predictions

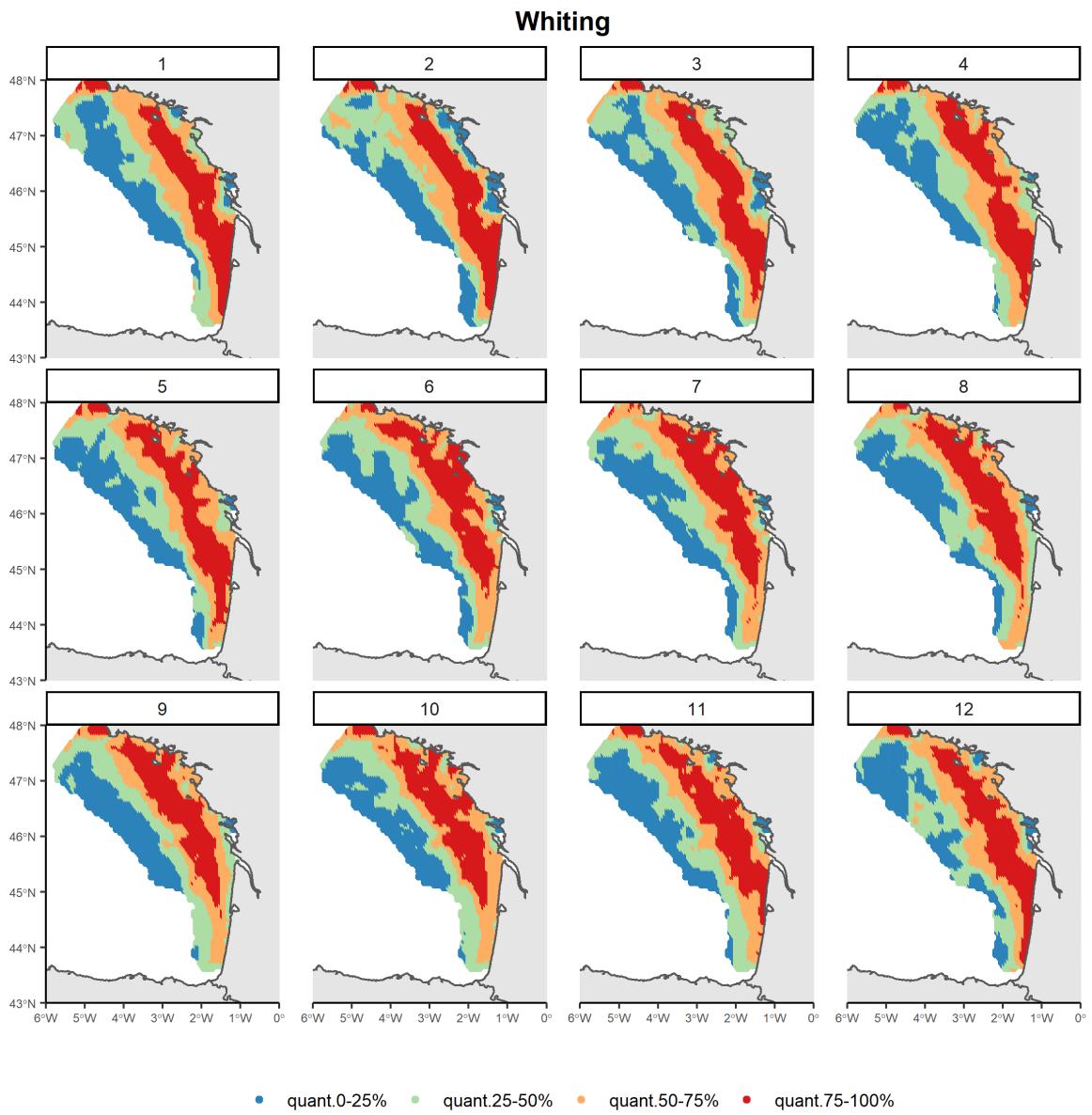


Figure C.9 – Whiting. Monthly biomass distribution averaged on the full period. Only quantile values are represented.

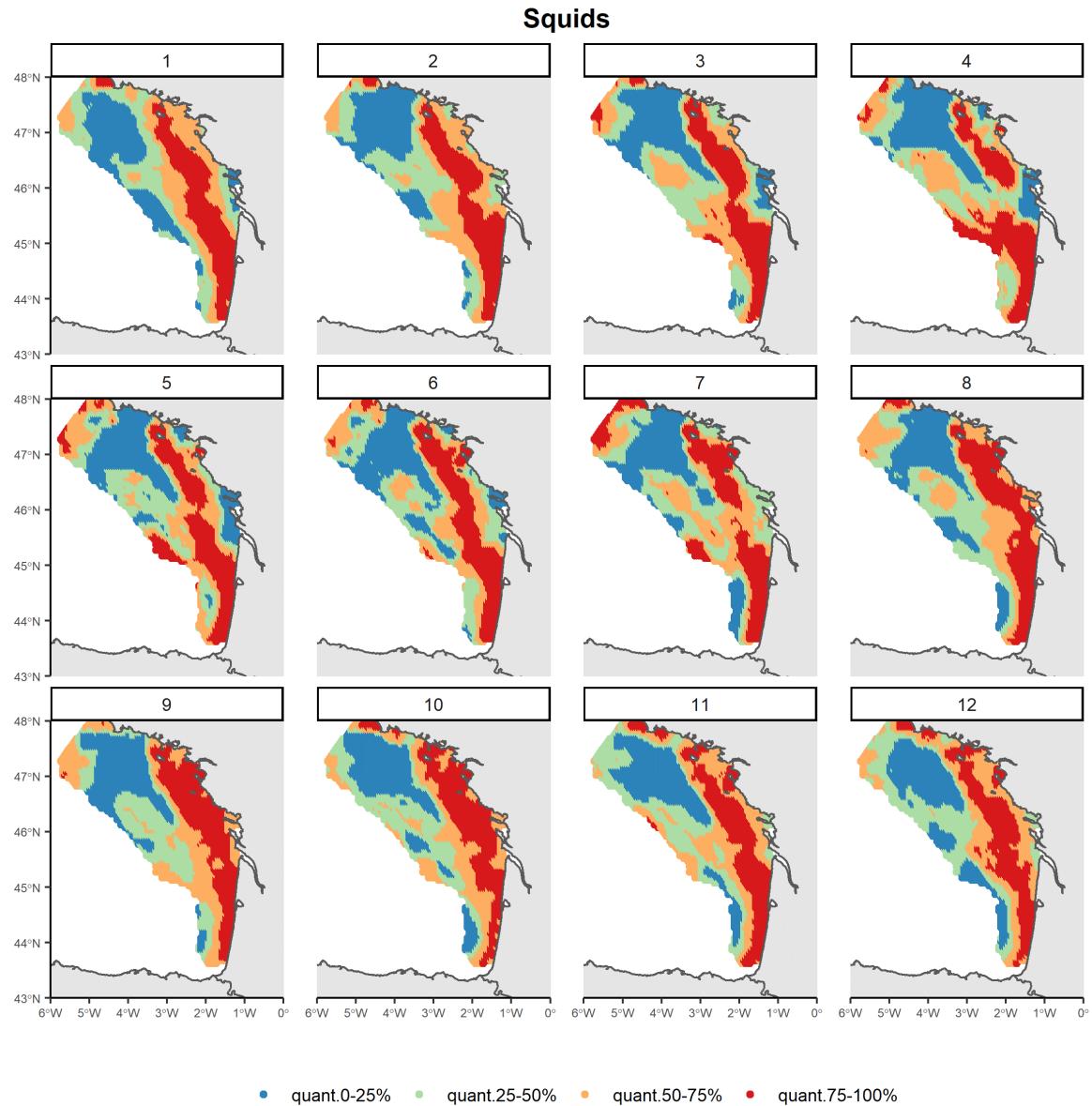


Figure C.10 – Squids. Monthly biomass distribution averaged on the full period. Only quantile values are represented.

---

## C.10 Persistence index maps

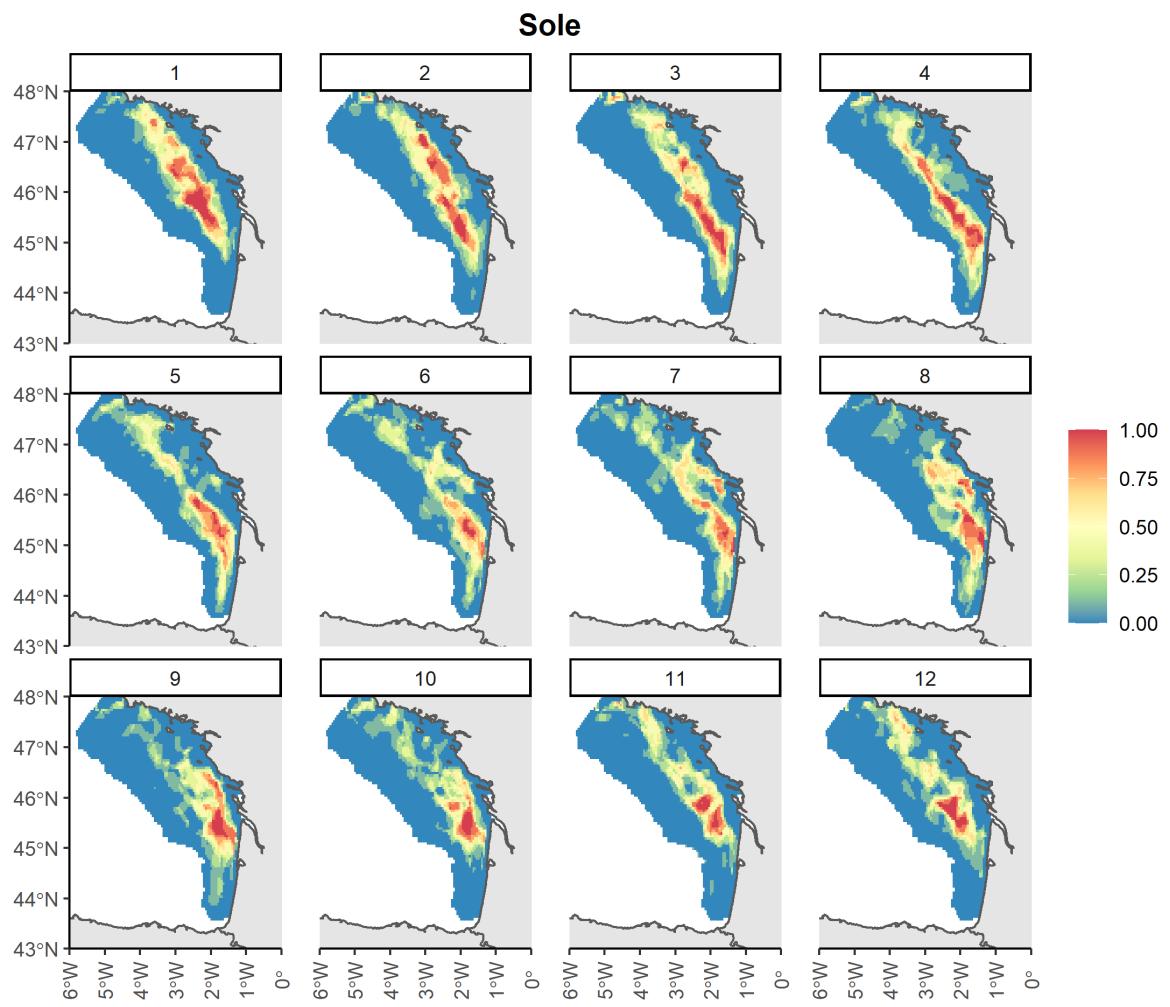


Figure C.11 – Sole. Monthly persistence indices. Aggregation over 2010-2018.

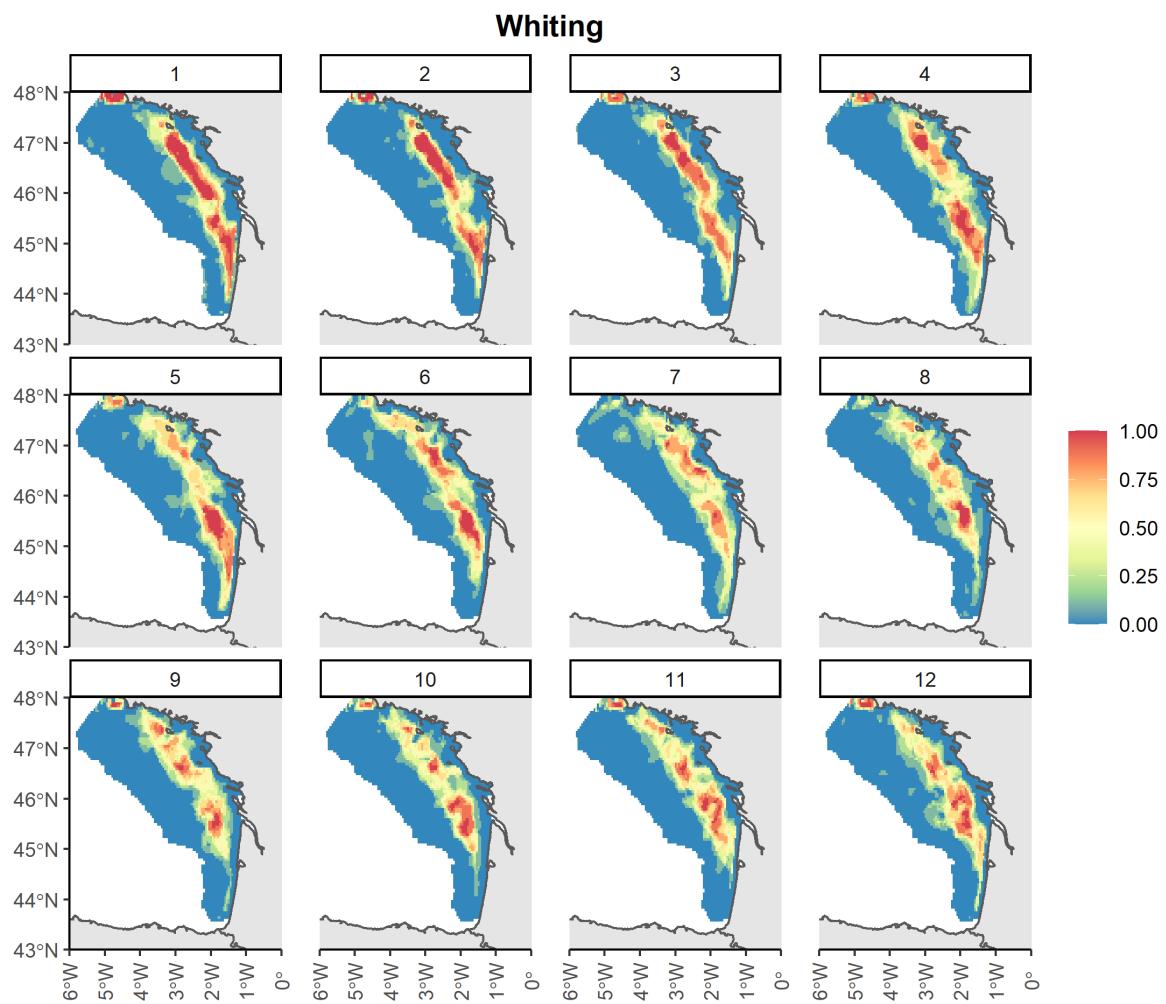


Figure C.12 – Whiting. Monthly persistence indices. Aggregation over 2010-2018.

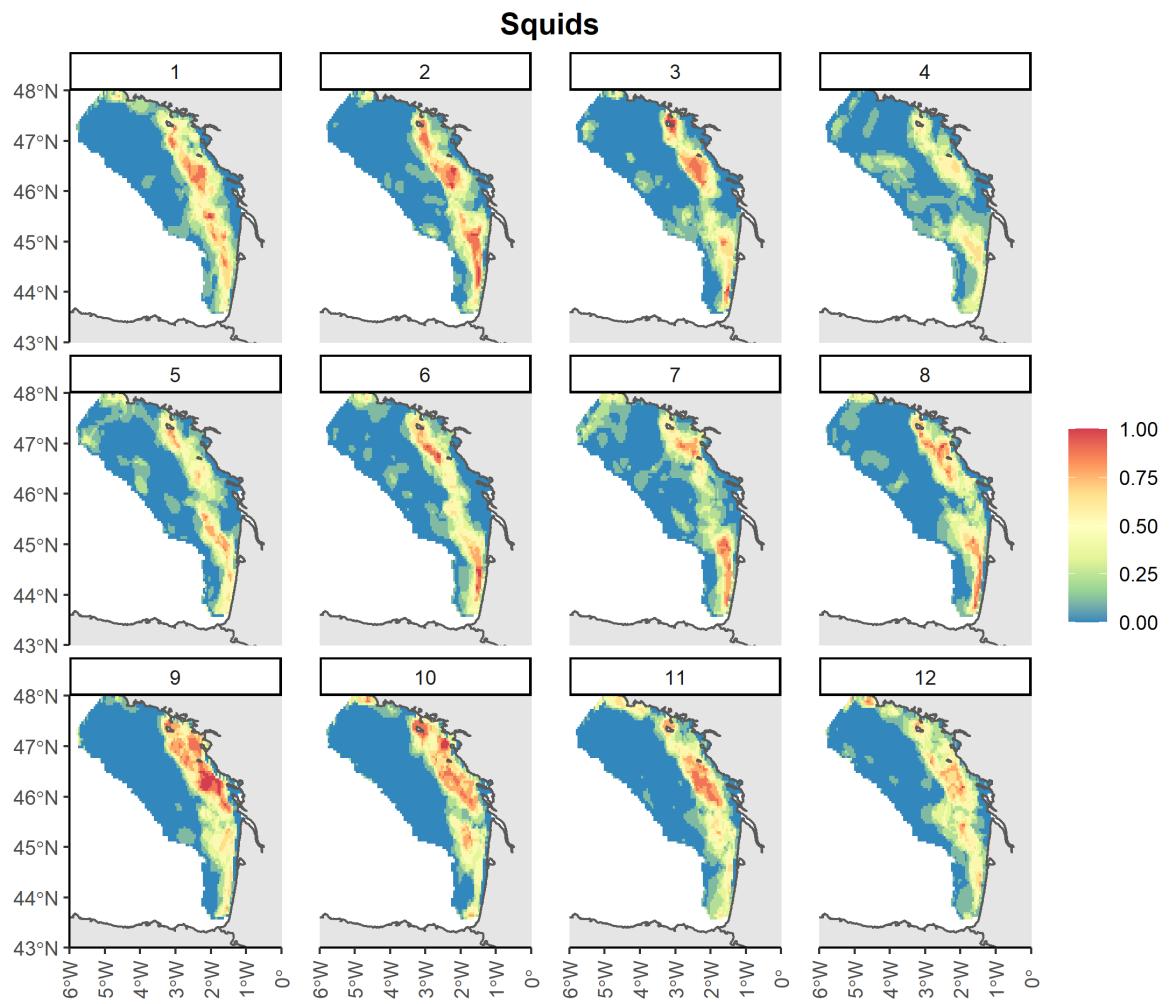


Figure C.13 – Squids. Monthly persistence indices. Aggregation over 2010-2018.

# INFERRING FINE SCALE WILD SPECIES DISTRIBUTION FROM SPATIALLY AGGREGATED DATA

---

## D.1 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 realized at one fishing position  $x_i$  (known through VMS data).

We note:

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

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

## D.2 Reparameterization of the Lognormal distribution

The Lognormal distribution is usually 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:

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

### D.3 $D_k$ 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.

$$Z_i|S(x_i), x_i \sim \mathcal{L}\left(\frac{S(x_i)}{1 - p_i}, \sigma^2\right)$$

Here,  $Y_i$ ,  $C_i$  and  $Z_i$  are observations of a latent field ( $S$ ) at a sampled point  $x_i$ . In the following, both  $Y_i$  and  $D_k$  are supposed to be conditional on the latent field  $\mathbf{S}$  and on the related fishing positions (either  $x_i$  or  $\mathcal{P}_k$ ).

### D.4 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:

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



---

## D.5 Expectation of a positive declaration

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

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

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

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

## D.6 Variance of a positive declaration

The variance then can be expressed as:

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

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

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

---

## D.7 Sum up of the main formulas

The main formulas of the model can be summarised as follows:

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

- The probability to obtain a zero declaration

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

- 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:

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

# DISCUSSION

---

## E.1 The value of ‘VMS x logbook’ data to explore fish spatio-seasonal patterns and species phenology at fine spatio-temporal scale

### E.1.1 Material and methods

#### Case studies

We selected as case studies.:

- sole in the Bay of Biscay
- sea bass in the Bay of Biscay
- hake in both the Bay of Biscay and the Celtic Sea

These three case studies illustrate the progressive development of the use of catch declarations data to describe fish phenology and identify fish essential habitats.

First, sole represents a case where poor information are available to map fish distribution throughout the year. Information on spawning areas are available from an old egg and larvae survey (Arbault, Camus, and Bec, 1986) and a yearly beam trawl survey (Orhago) occurs each November and allows to sample few observations on the population (about 50 per year) to compute abundance index for the stock assessment (Biais, 2003). However, no additional data is available to map intra-annual species distribution and the knowledge of species distribution throughout the year remains very limited.

Second, hake illustrates the case where the analysis of logbooks has enabled to investigate the spatio-temporal distribution of mature individuals throughout the year at the scale of ICES rectangles. These were interpreted in terms of fish phenology and reproduction migrations (Poulard, 2001). Furthermore, studies investigating the spatial distribution of mature individuals during reproduction are available to identify potential reproduction areas (Tidd and Warnes, 2006).

---

More recently, Dambrine et al. (2021) characterized spawning areas of seabass at fine spatial scale based on logbooks data combined with VMS data. They identified aggregation areas from January to April based on occurrence data and interpreted these as reproduction areas. However, they only focused their analysis on the reproduction period and they did not investigate the whole year spatio-temporal variability.

### Aim of the analysis

To go one step further, we propose to investigate the spatio-temporal variability of these species throughout the year based on monthly species distribution maps built on ‘VMS x logbooks’ data. Our goal is to provide a generic and synthetic approach to analyze the key spatio-temporal patterns that structure species distribution on the full year. We aim at identifying potential long-term trends in species distribution, seasonal phases that structure species distribution and eventually the punctual events that affected species distribution in the previous years. Our final goal is to interpret these signals in terms of fish phenology and then to identify some key functional zones of the species (i.e. fish essential habitats).

### Data and model

The spatio-temporal models were fitted between 2008 and 2018. We filter the mature fraction of the landings to map the potential breeders of the population and interpret the model outputs in terms of reproduction phenology.

The models integrate several trawl fleets that ensure good coverage of the area. For hake we selected 3 bottom trawl fleets targeting demersal fish (OTB\_DEF\_>=70\_0, OTB\_DEF\_100\_119\_0, OTB\_DEF\_70\_99\_0) and one otter trawl fleet targeting demersal fish (OTT\_DEF\_>=70\_0); for sole we selected a bottom trawl targeting demersal fish (OTB\_DEF\_>=70\_0), a bottom trawl targeting cephalopods (OTB\_CEP\_>=70\_0), and an otter trawl targeting demersal fish (OTT\_DEF\_>=70\_0); for seabass, one bottom trawl targeting cephalopods (OTB\_CEP\_>=70\_0) and demersal fish (OTB\_DEF\_>=70\_0) and one pelagic trawl fleet targeting demersal fish (PTM\_DEF\_>=70\_0).

Preferential sampling was not accounted for in inference as this does not modify the overall patterns of species distribution and it would lead to increased computation time for little gain regarding spatial predictions (Cf. chapter 4).

---

## Decomposing model outputs: Empirical Orthogonal Factors (EOF)

### Base formulation of EOF

Empirical Orthogonal Functions is a method that has been developed by Lorenz (1956) for weather forecasting applications. The original aim of the technic was to deduce from a set of spatio-temporal maps a smaller set of maps that best describe and summarize the spatio-temporal process of interest. The main idea is to define the spatio-temporal process  $S(x, t)$  as a linear combination of spatial patterns  $\mathbf{p}_m$  (named EOF) each related to temporal indices  $\alpha_m(t)$  (Equation E.1). These temporal indices indicate when the spatio-temporal process is distributed following their related spatial pattern.

$$S(x, t) = \sum_{m=1}^M \alpha_m(t) \cdot p_m(x) + r^M(x, t) \quad (\text{E.1})$$

where  $r^M(x, t)$  is the residual variation not captured by the  $M$  modes of variation  $\alpha_m(t) \cdot p(m, x)$ .  $x \in \llbracket 1, n \rrbracket, t \in \llbracket 1, T \rrbracket$ .

To estimate the temporal components  $\alpha_m(t)$  and the spatial patterns  $p(m, x)$ , some criteria need to be set. A natural choice is to minimize the residual variation to best capture the variability of  $S(x, t)$  (i.e.  $R^M = \sum_{m=1}^M (r_m^M)^2$  is minimized) and set orthogonal constraints between the modes of variability so that each mode is the 'best' representation of variability from a statistical point of view (Equations E.2, E.3). This last criteria is further discussed in the discussion.

$$\sum_{x=1}^N p_m(x) \cdot p_j(x) = \delta_{mj} \equiv \begin{cases} 1 & \text{if } m = j \\ 0 & \text{if } m \neq j \end{cases} \quad (\text{E.2})$$

$$T \cdot \overline{\alpha_m^* \alpha_j^*} = a_m \delta_{mj} \quad (\text{E.3})$$

with  $a_m \geq a_{m+1} \geq 0$ ,  $\overline{(\ )}$  denoting the time average and  $( )^*$  a departure from the time average.

### Matrix formulation

Let's write these equations in matrix terms by introducing the  $T \times N$  matrix  $\mathbf{S}, \mathbf{S}^*$ ,

---

$\mathbf{Q}$ ,  $\mathbf{Q}^*$ . They respectively refer to  $S(x, t)$ ,  $S^*(x, t)$ ,  $\alpha_m(t)$  and  $\alpha_m^*(t)$ . We denote  $\mathbf{Y}$  as a square matrix of order  $N$  with elements  $p_m(x)$ . Then, the problem can be reformulated as:

$$\mathbf{S} = \mathbf{Q}\mathbf{Y} \quad (\text{E.4})$$

$$\mathbf{Y}\mathbf{Y}^T = \mathbf{I} \quad (\text{E.5})$$

$$\mathbf{Q}^{*T}\mathbf{Q} = \mathbf{D} \quad (\text{E.6})$$

with  $()^T$  the transpose,  $\mathbf{I}$  the identity,  $\mathbf{D}$  a diagonal matrix with diagonal being equal to  $a_m/T$ .

Finally, by introducing  $\mathbf{A} \equiv \mathbf{Q}^{*T}\mathbf{Q}^*$  (whose elements are proportional to the covariance of  $\alpha_m(t)$ ), we can rewrite these equations as:

$$\mathbf{Y}\mathbf{A}\mathbf{Y}^T = \mathbf{D} \quad (\text{E.7})$$

This is a standard "eigenvalue-eigenvector" problem. As a consequence, a link can be done with standard multivariate analysis such as PCA.

Thus, in addition to the spatial patterns (and the related temporal components) that appear in the EOF formulations (Equation E.1) and that can be obtained by diagonalizing the problem in Equation E.7, additional analysis can be performed to obtain similar visualization as in PCA (e.g. plot of individuals and variables, contribution of locations and time steps to the several dimensions) possibly coupled with standard clustering analysis.

## Analysis

We filter the number of dimensions for each species based on the graph of the variance captured by each dimension (Figure E.1). As is commonly done in PCA analysis, we cut the graph at the dimension where there is a drop in the variance explained. For sole, we filtered the six first dimensions. For seabass, we select the first dimension only. For hake in the Bay of Biscay as well as for hake in the Celtic sea, we filtered the two first dimensions.

Then, we analyze the spatio-temporal model outputs by presenting the spatial patterns  $p_m(x)$  (which are the eigen-vectors in the diagonalization). These can be seen either as

---

maps that structure  $\mathbf{S}$  (e.g. Figure E.2) or either as standard variable plot as in classical PCA analysis (Cf. Figure E.6, center figure).

The temporal index  $\alpha_m(t)$  can be either seen as time series (e.g. Figure E.2) or as a classical plot of individuals on the first components of the PCA (Cf. Figure E.6, left figure).

When confronted together, the spatial representations and the related temporal indices inform how the spatial patterns that structure species distribution vary in time. When the loading factor  $\alpha_m(t)$  are positive (resp. negative), then species distribution  $\mathbf{S}$  at time step  $t$  is distributed following the EOF map  $p_m(x)$  (resp.  $-p_m(x)$ ). Then the evolution of  $\alpha_m(t)$  tells how  $\mathbf{S}$  evolve in time. Such patterns can be interpreted in regards to the spatial ecology of the species.

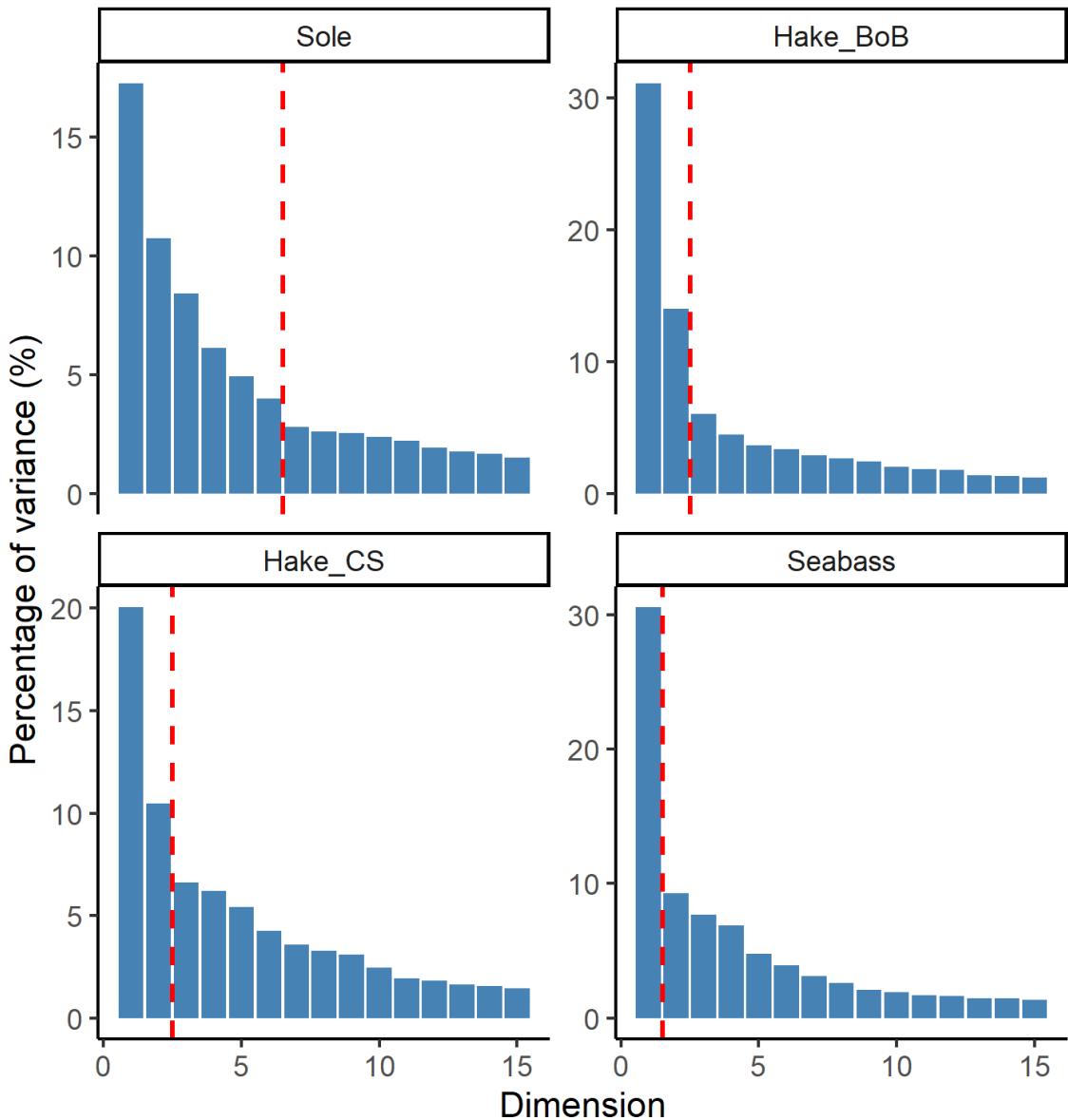


Figure E.1 – Graph of variance explained by each dimension for each species. Dashed line: cutting level of the EOF.

### E.1.2 Results

#### Spatio-temporal representation

All species evidence clear seasonal patterns (Figure E.2, E.3, E.4, E.5); In all cases, the 2 first EOF dimensions capture a seasonal signal.

---

For sole, some periodic signal is evidenced highlighting high biomass in offshore areas in winter (December to April) and relatively coastal distribution in summer. This is consistent with the results highlighted in chapter 4 (see section D) and in Arbault, Camus, and Bec (1986) where sole is evidenced to reproduce in relatively offshore areas.

For seabass, a strong positive anomaly occurs seasonally each January/February. High biomass zones are mainly localized along and off shore the Vendée coast and up to the plateau of rochebonne (i.e. all the locations that are not blurred in Figure E.3). This is consistent with the paper from Dambrine et al. (2021) that emphasizes that both areas and months that we emphasize corresponds to the reproduction period and area of seabass. In addition to their work, we provide information on biomass densities (not only on occurrence density) which can help to order the aggregation areas of fish and to identify where most fish aggregate.

For hake, similar (but weaker) patterns can be evidenced on the first two EOF (Figure E.4). In the Bay of Biscay, hake has a more coastal distribution in winter (January/February/March) compared with summer. For hake in the Celtic Sea, there is a strong signal on the first EOF (Figure E.5) in the center of the Celtic Sea occurring in summer, a period where no information is available from the literature to inform which process could occur here. When looking at the second EOF some high biomass patterns come back seasonally on the Eastern of the Celtic Sea and match with the distribution of mature individuals from old surveys focusing on reproduction (Tidd and Warnes, 2006).

Astonishingly, the high biomass patterns on the Eastern of the Celtic Sea occur slightly later than in the Eastern of the Bay of Biscay. This could be related to a later reproduction period in the Northward areas due to temperature differences between these two areas.

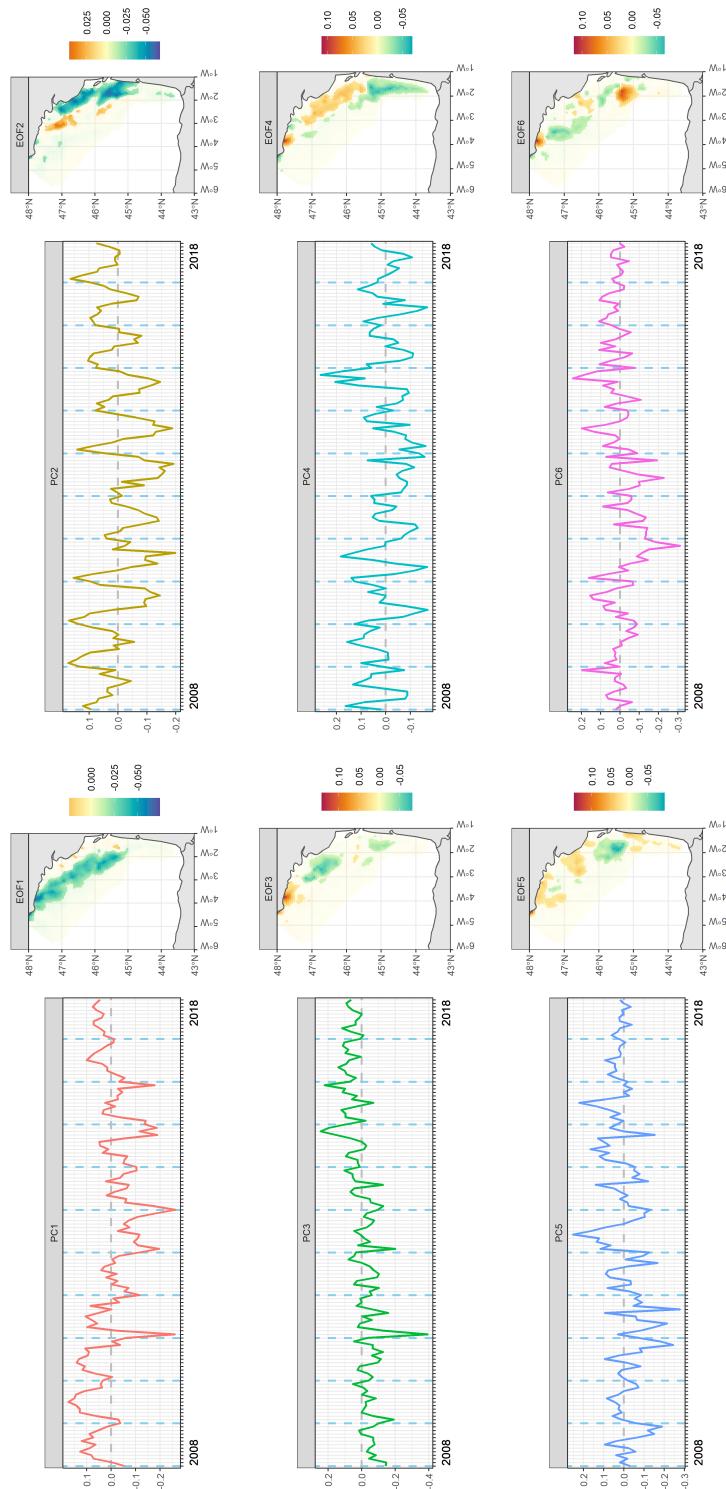


Figure E.2 – Sole. Six first EOF maps and time-series for sole. Blue dashed line: January. For EOF maps, the locations that do not have a significant contribution to the dimension are blurred.

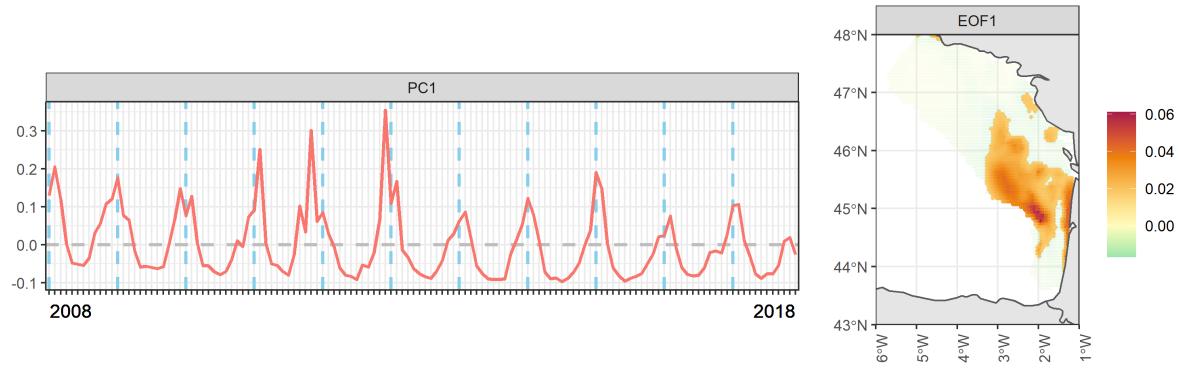


Figure E.3 – Sea bass in the Bay of Biscay. First EOF map and time-series. Blue dashed line: January. For EOF maps, the locations that do not have a significant contribution to the dimension are blurred.

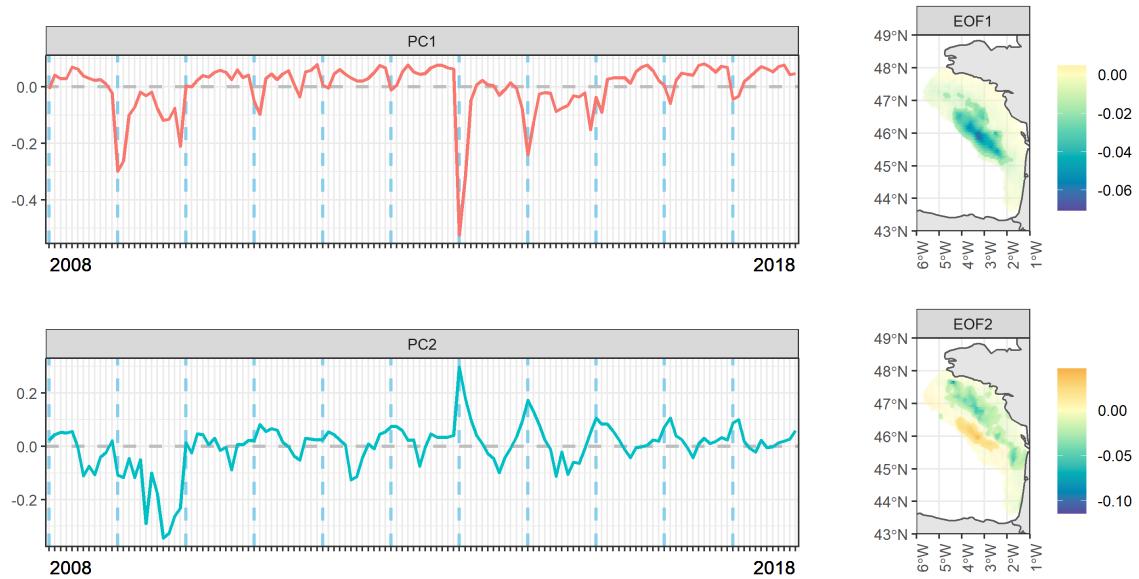


Figure E.4 – Hake in the Bay of Biscay. Two first EOF maps and time-series. Blue dashed line: January. For EOF maps, the locations that do not have a significant contribution to the dimension are blurred.

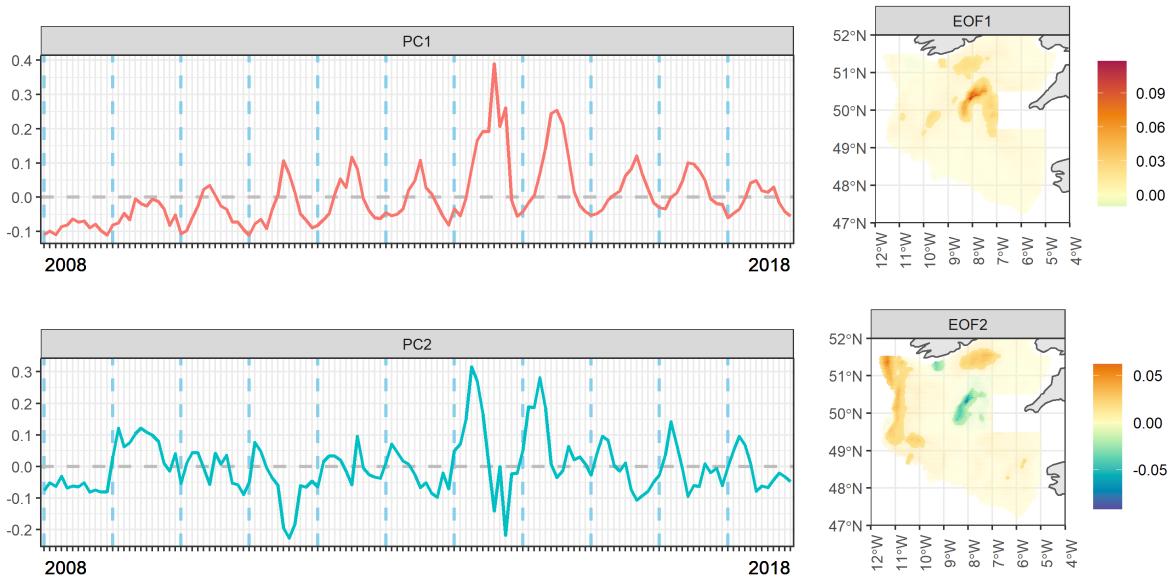


Figure E.5 – Hake in the Celtic Sea. Two first EOF maps and time-series. Blue dashed line: January. For EOF maps, the locations that do not have a significant contribution to the dimension are blurred.

### PCA representation

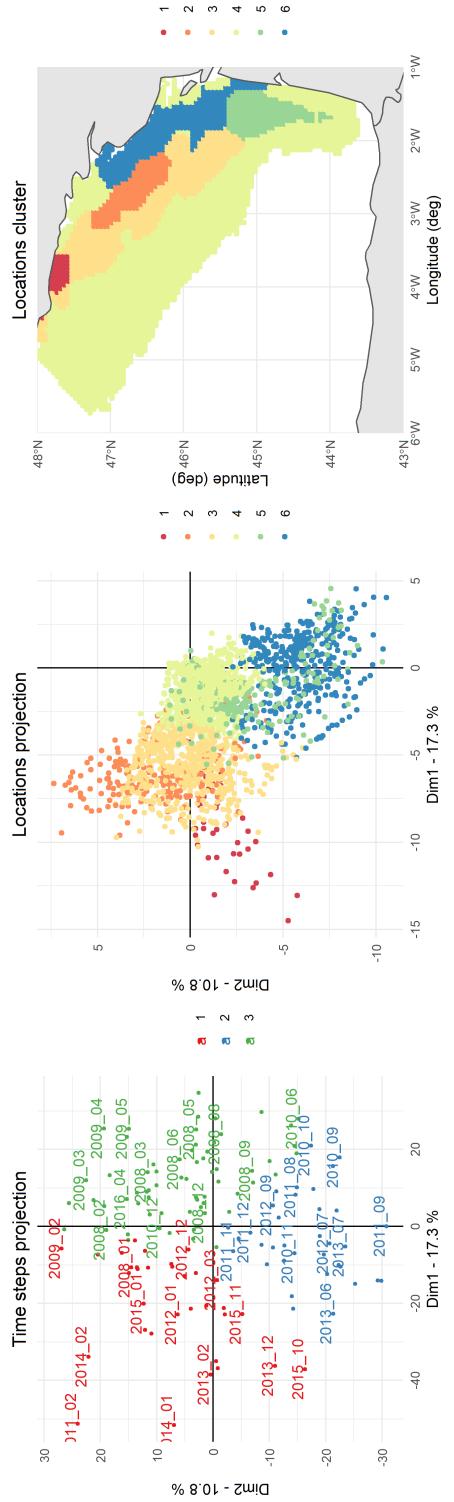
The PCA representation of the spatial and temporal patterns coupled with hierarchical clustering allows to identify sets of points and time steps that can be interpreted as combination of seasons and functional zones. This is illustrated with the sole case study where Figure E.6 and E.8 highlights that 3 clusters of time steps can be grouped into 3 seasons:

1. the red cluster (Figure E.6, left) that regroups winter months (December to February - Figure E.7). This can be related to the cluster 1, 2 and 3 on the graph of locations (center figure) and maps of clusters (right figure) which corresponds to areas where sole reproduce (see chapter 4 and Arbault, Camus, and Bec (1986)).
2. the green cluster (Figure E.6, left) mainly corresponds to spring and early summer months (March to June - Figure E.7). Sole progressively go back to coastal areas.
3. the blue cluster (Figure E.6, left) mainly corresponds to autumn and spring distribution (July to November - Figure E.7). Sole has a more coastal distribution (biomass of sole are high in the blue area cluster on the right plot of Figure E.6)

---

and biomass also aggregates in the hotspots of the South of the Bay of Biscay (green cluster in the right plot of Figure E.6).

Then related 3 areas could be interpreted as functional zones (right plot in Figure E.6): (1) a reproduction area (the orange and yellow clusters) where biomass aggregates mainly from December to January, (2) a coastal area where biomass is high in autumn and winter (the blue cluster) and (3) an area where biomass is always high but specifically in autumn and winter (green cluster). These two last areas could be interpreted as feeding areas: sole feed in these areas during autumn and early winter before going back to reproduction grounds each year from December to February. The green cluster locations could also be interpreted as the arrival of juveniles in the mature population.



---

Figure E.6 – (Left) Projection of the time steps indices on the 2 first dimensions of the EOF. Color: cluster identified through HAC analysis. All the points are related to their time step. (Center) Projection of the locations on the 2 first dimensions. Color: cluster identified through HAC analysis. (Right) Spatial representation of the clusters.

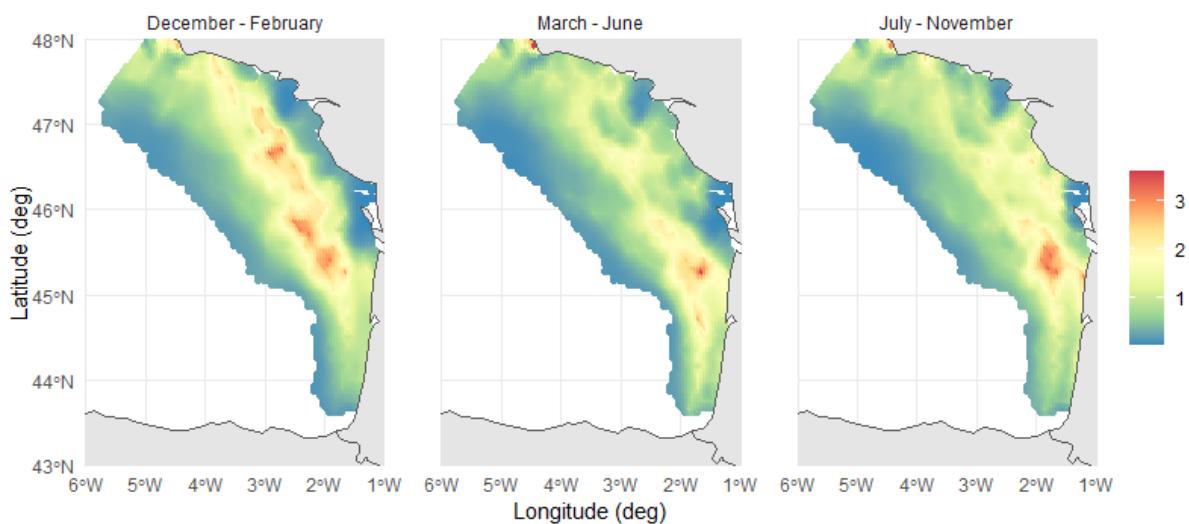


Figure E.7 – Mean pattern for each seasonal cluster identified in Figure E.6. December to February corresponds to the red cluster; March to June corresponds to the green cluster; July to November corresponds to the blue cluster.

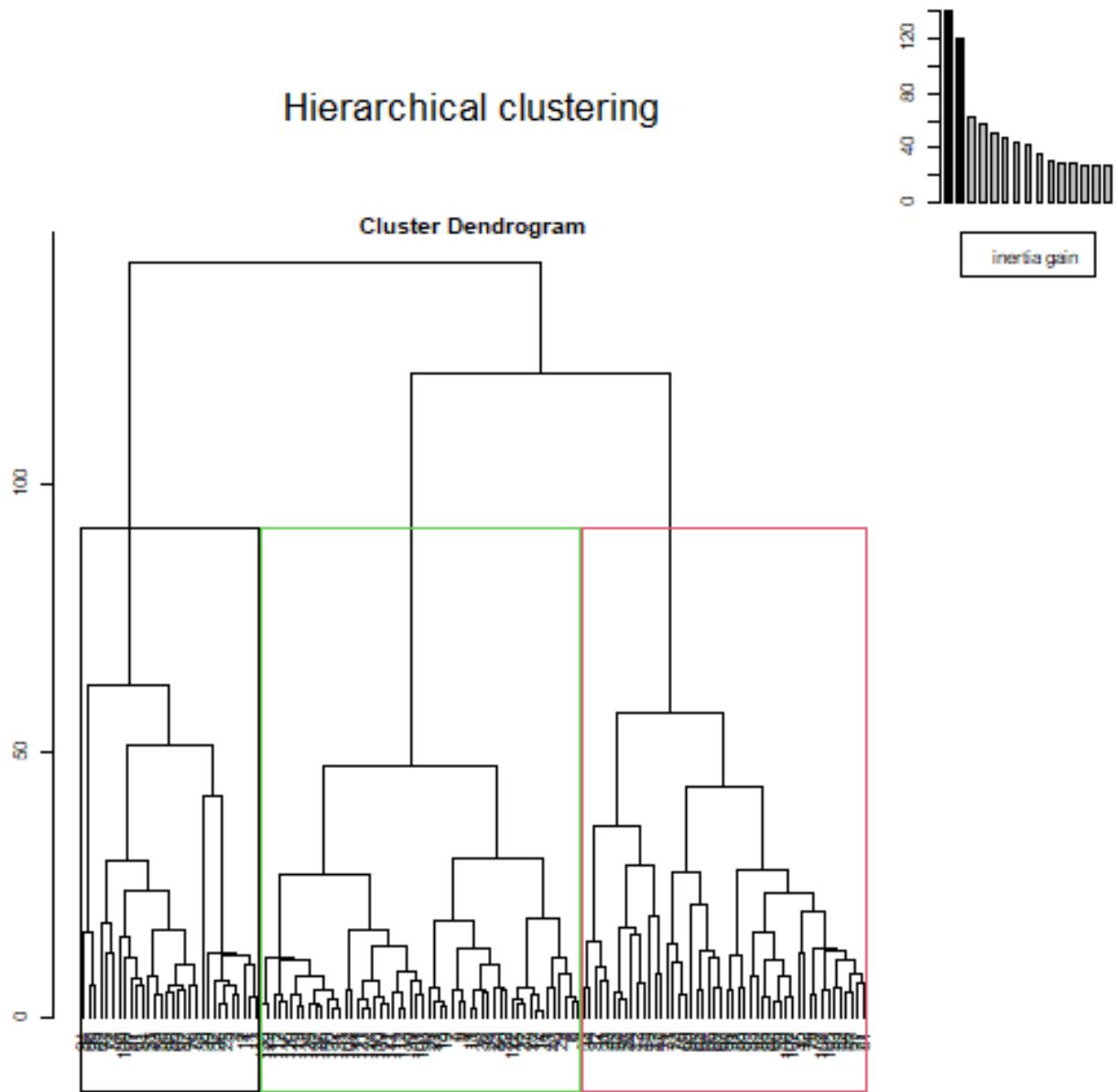


Figure E.8 – Sole. Clustering tree for the time steps.

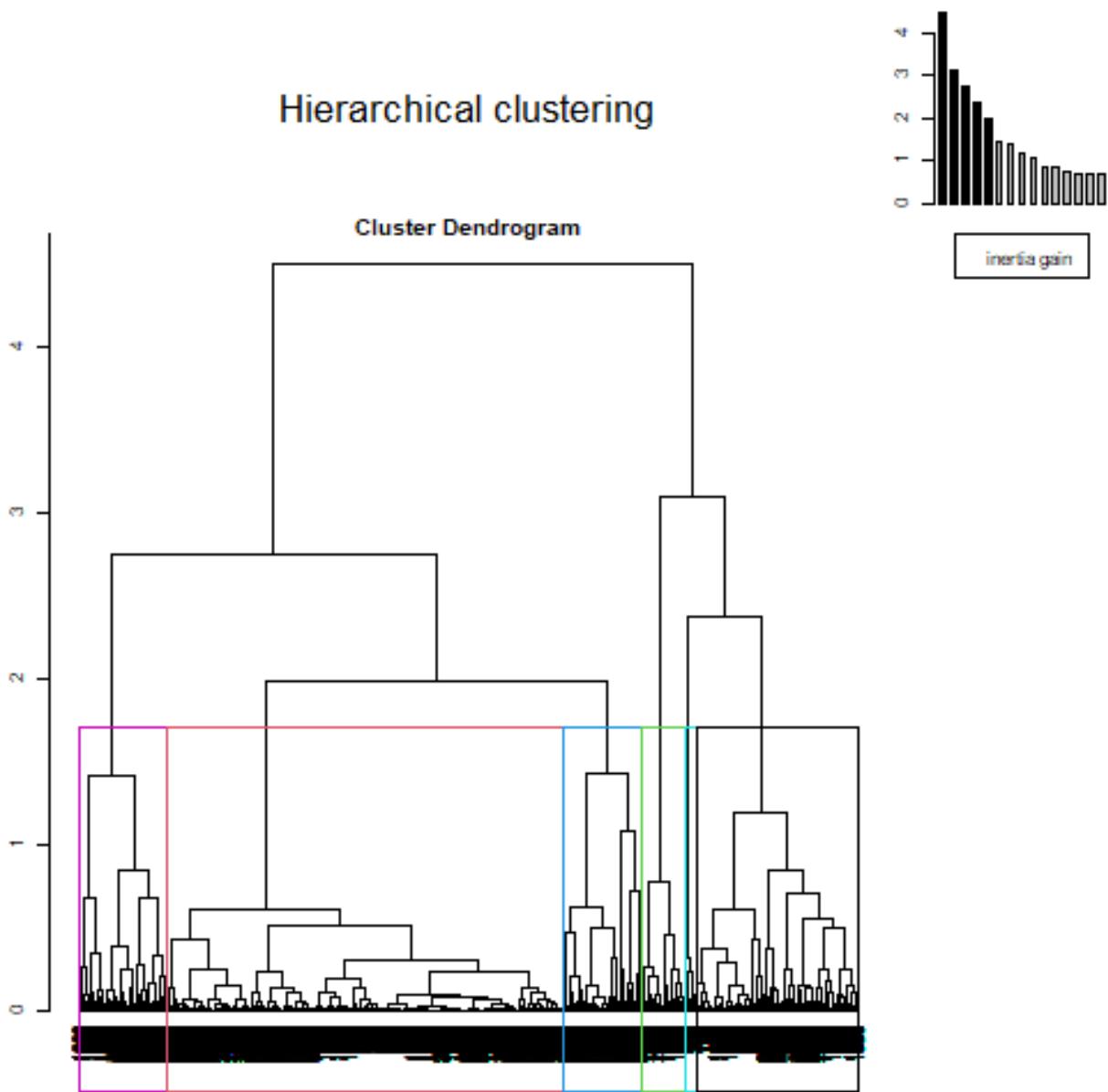


Figure E.9 – Sole. Clustering tree for the locations.

### E.1.3 Discussion

- ‘VMS x logbooks’ provide information to infer species distribution at a monthly time step. This makes possible to identify spatio-seasonal patterns and to interpret these in terms of fish functional zones for sole. Similar results are available for the other species.

- 
- Applying a similar approach for several species would enable to identify spatio-seasonal patterns that are common to these species and possibly identify essential habitats for a set of species.
  - Our approach raises methodological challenges as the signals of the different maps and time series are sometimes redundant (i.e. several time series emphasize the same seasonal pattern), but one would like to better synthesize the information of this signal in a single time-series and a single map. Methods exist to better disentangle the spatial signal from a set of maps. These are based on geostatistical methods e.g. MAF and EOM (Bez, Renard, and Ahmed-Babou, 2022). However, these are strongly sensitive to the choice of the neighborhood and they do not account for temporal correlation. Developing methods that better disentangle the signal without the limitations of MAF would be a valuable contribution to spatial statistics. This could be possible by modifying the constraints of EOF by specifying orthogonality constraints that accounts for spatio-temporal correlations.



---

## E.2 Modeling the spatial distribution of the sardine (*Sardina pilchardus*) in the Bay of Biscay by integrating commercial and scientific data: challenges and limits

**Année universitaire :** 2020 - 2021

**Diplôme :** Ingénieur agronome

**Spécialisation :**

« Sciences Halieutiques et Aquacoles »  
préparé à Agrocampus Ouest Rennes

**Option :**

« Ressources et Ecosystèmes Halieutiques »

### Mémoire de fin d'études

- d'ingénieur d'AGROCAMPUS OUEST (École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage), école interne de L'institut Agro (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement)
- de master d'AGROCAMPUS OUEST (École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage), école interne de L'institut Agro (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement)
- de Montpellier SupAgro (étudiant arrivé en M2)
- d'ingénieur d'AgroSup Dijon (Institut national supérieur des sciences agronomiques, de l'alimentation et de l'environnement)

### Modélisation de la distribution spatiale de la sardine du Golfe de Gascogne (*Sardina pilchardus*) par intégration de données commerciales et scientifiques : enjeux et limites.

Par : Florian QUEMPER



**Soutenu à Rennes le 14/09/2021**

**Devant le jury composé de :**

Président : Olivier Le Pape

Maître de stage : Baptiste Alglave, Etienne Rivot,  
Marie-Pierre Etienne

Enseignant référent : Olivier Le Pape

Autres membres du jury (Nom, Qualité) :

Hélène Peltier – Ingénierie de recherche à Pelagis  
Nicolas Bez – Directeur de Recherche, IRD

*Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST*

Ce document est soumis aux conditions d'utilisation «Paternité-Pas d'Utilisation Commerciale-Pas de Modification 4.0 France» disponible en ligne <http://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr>



## Confidentialité

Non  Oui      si oui :  1 an  5 ans  10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible<sup>(1)</sup>.

Date et signature du maître de stage<sup>(2)</sup> : 28/09/2021  
(ou de l'étudiant-entrepreneur) 

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

## Droits d'auteur

L'auteur<sup>(3)</sup> Nom Prénom   
autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

Oui  Non

Si oui, il autorise

- la diffusion papier du mémoire uniquement(4)
- la diffusion papier du mémoire et la diffusion électronique du résumé
- la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif)  accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur :  28/09/2021

## Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

Oui  Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

- la diffusion papier du mémoire uniquement(4)
- la diffusion papier du mémoire et la diffusion électronique du résumé
- la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :  28 Sept 2021

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.

(2) Signature et cachet de l'organisme

(3) Auteur = étudiant qui réalise son mémoire de fin d'études

(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option) sera signalée dans les bases de données documentaires sans le résumé

---

*« Si les mouettes suivent le chalutier, c'est parce qu'elles pensent que des sardines seront jetées à la mer »*

E. Cantona – 1995

---

## Remerciements

Je tiens dans un premier temps à remercier sincèrement Baptiste Alglave pour l'aide précieuse qu'il m'a apporté tout au long de ce stage. Toujours présent pour répondre à mes questions (même les plus idiotes, surtout les plus idiotes), ça aurait sans aucun doute été d'une bien plus grande difficulté sans toi. Encore merci Baptiste. Franchement j'insiste sans plaisanter, c'est vraiment super agréable de travailler avec toi, t'es patient, humble et quand même vachement doué, même au rugby tu donnes de bons conseils !

Je souhaite ensuite remercier Etienne Rivot et Marie Pierre Etienne, également pour leur suivi, leurs propositions et conseils avisés qui nous ont permis à moi et Baptiste d'avancer dans le stage le plus sereinement possible. De même, je remercie les différentes personnes qui ont pu nous apporter leur aide tout au long de ce stage ainsi que ceux nous ayant fourni les différents jeux de données : Mathieu Doray, Thomas Opitz, Youenn Vermand, Mathieu Woillez et Guillermo Boyra.

Vient ensuite l'ensemble des stagiaires, on ne se sera pas beaucoup vu les 3 premiers mois grâce à ce bienheureux confinement. Malheureusement nous avons ensuite été forcés de nous côtoyer, de sortir et d'aller au bar. On aura même un peu dansé.

Pas toujours facile de vous sortir de votre timidité, en particulier toi Gasparccio, franchement t'abuses il faut tout le temps te pousser. En plus depuis que je t'ai montré « Me llaman calle » tu t'arrêtes plus de la chanter. Je reparle de toi un peu plus loin quand je cause de la Cocoflo'c et de La Rochelle.

Un p'ti coucou à Lulu d'ses morts, j'ai hâte de recevoir un autre de tes messages le 14 juillet 2022 à midi, jour béni de la naissance de ton petit frère, Paul, et nous expliquant ô combien tu es farcie comme une tomate après avoir galement festoyer en admirant un feu d'artifice, à tel point qu'il t'est désormais impossible d'avaler le moindre aliment au restaurant. Dommage.

La bizette à Daphneux, à qui je décerne le titre de « révélation alcoolisée de ces 6 mois de stage » tellement ta furtive mais ô combien remarquable apparition aura été marquante dans la courte période de ce stage. Franchement c'est avec plaisir qu'on remet le couvert, d'ailleurs je vais de ce pas t'envoyer un message pour te motiver à venir claquer une canouche chez les amis de Lulu demain soir. Allez Zag comme dirais l'autre.

Plus brièvement, je fais un petit salut depuis cette page aux autres stagiaires, Florian, Jérémy et Sully (je sens que je vais bientôt plus avoir de place, en fait il faudrait que je fasse un rapport de remerciements)

Ensuite je serre chaleureusement la main à mes 2 collocs de la « Versailles sans bouchon des chiottes loc' », dont l'un a été mon collègue aux archives du labo, le bien sus-nommé Thomas. En vrai c'était sympa d'être avec toi dans le cagibi du bat' Halieut, on aura quand même pas mal discuté et tu m'auras aussi filé quelques conseils et idées, c'était chouette ! Dommage que tu ais trouvé l'amour durant ce stage, on se sera moins cotoyé que prévu. M'enfin je rentrais tous les week-ends aussi donc ça n'a pas aidé. Bref ! Une grosse bisse au second colloc', l'heureux Quentining, je ne dirais rien de plus que : « La Carambole ! ». Si je vais dire un peu plus en fait, c'était là encore un plaisir de passer du temps à discuter avec toi (je me souviens notamment de cette soirée où je suis rentré à 1h du mat de l'agro et tu es sorti de ta chambre pour discuter, sacré toi tiens). Je pense qu'il m'arrivera de revenir toquer à ta porte en disant « Oh Quenting... ? ».

Je garde un dernier petit paragraphe, parce que j'ai été trop – Bordel Mikaëla j'allais t'oublier ! Enfin bon de toute façon tu restes dans le couloir – je disais donc trop exhaustif peut-être, pour les copains de l'agro Rennes et Dijon - en particulier mon chto Rem's et mon Coco qui comme moi ont quitté les terres Dijonnaises pour rejoindre ACO, avec qui on aura fait les loubards et qui ont été des véritables soutiens sans s'en rendre compte, j'ai plus de place mais vous méritez une page chacun pour toutes nos aventures - de prépa, de PeB, et ma famille évidemment, et en particulier mon père, qui m'a épaulé tout au long de ce stage, ma mère, mes 3 frères et mes grands-parents (mention gâteaux pour Mamignonne !). Je ne veux ni exagérer et faire plus d'une page de remerciements ni trop détailler mes sentiments pour ma famille et mes amis proches ici, mais sincèrement, merci. Je vous porte dans mon cœur.

Alllez La Rochellllee !

Une pensée pour Jérôme Quinquis.

---

## Table des matières

1.	Introduction .....	1
1.1.	Généralités : distribution d'espèce et SDM.....	1
1.2.	Données scientifiques et données commerciales.....	2
1.3.	Des modèles hiérarchiques pour intégrer différentes sources de données .....	3
2.	Matériel et méthode.....	5
2.1.	Description du cas d'étude : la sardine du Golfe de Gascogne. ....	5
2.2.	Données .....	6
2.2.1.	Données scientifiques issues de campagnes acoustiques .....	6
2.2.2.	Données commerciales (logbooks & VMS).....	7
2.3.	Spécificités du cas d'étude et conséquence pour la modélisation .....	9
2.4.	Traitement des données .....	10
2.5.	Modèle intégré hiérarchique.....	10
2.5.1.	Champ latent de présence-absence.....	11
2.5.2.	Modèle d'observation de la présence/absence.....	12
2.5.3.	Modéliser l'échantillonnage préférentiel.....	12
2.5.4.	Simplification induite par les contraintes d'INLA .....	13
2.5.5.	Outil d'inférence : R-INLA .....	13
2.5.6.	Métriques de validation .....	15
2.5.7.	Démarche de valorisation du modèle .....	15
3.	Résultats .....	16
3.1.	Analyse à l'échelle d'un mois : mai 2018 .....	16
3.1.1.	Apport des différentes sources de données à l'inférence.....	16
3.1.2.	Intensité de l'échantillonnage préférentiel .....	19
3.1.3.	Ajout des co-variables environnementales.....	20
3.1.4.	Comparaison des capacités prédictives selon les différents modèles.....	22
3.2.	Analyse spatio-temporelle .....	23
4.	Discussion .....	28
4.1.	Une inférence à l'échelle du mois dirigée par la donnée scientifique .....	28
4.2.	Capacité prédictive du modèle .....	29
4.3.	Mise en perspective des cartes de distribution avec les connaissances déjà disponibles .....	29
4.4.	Limites de l'étude .....	30
4.4.1.	Spécificités de l'activité de pêche à la sardine dans le Golfe de Gascogne.....	30
4.4.2.	Limites de la modélisation de l'échantillonnage préférentiel.....	31
4.4.3.	Estimation de l'effet des covariables environnementales.....	31
4.4.4.	Biais liés aux données. ....	32
5.	Conclusion.....	33
	Références .....	35
	Annexes.....	43

---

## Liste des Figures

**Figure 1 :** Carte du Golfe de Gascogne et délimitation du domaine étudié.

**Figure 2 :** Observations (présence/absence) de *S. pilchardus* par la campagne scientifique PELGAS au cours du mois de Mai 2018 et par la campagne scientifique JUVENA au cours du mois de Septembre 2018.

**Figure 3 :** Observations (présence/absence) de *S. pilchardus* par les 3 flottilles commerciales entre janvier et décembre 2018

**Figure 4 :** Diagramme du modèle spatio-temporel hiérarchique intégré (d'après Alglave et al ; under review)

**Figure 5 :** Maille d'interpolation triangulaire et points de données dans le GdG (Mai 2018).

**Figure 6 :** Cartographies des probabilités de présence et des observations par les flottilles commerciales et la campagne scientifique PELGAS en Mai 2018.

**Figure 7 :** Cartographies des probabilités de présence (colonne de gauche) et écarts-types associés (colonne de droite) au mois de Mai 2018 obtenues à partir des données scientifiques, des données commerciales et dans un modèle intégré les combinant.

**Figure 8 :** Cartes d'inférence du modèle avec prise en compte d'un échantillonnage préférentiel dans le fonctionnement des flottilles commerciales, les écarts types associés et de la différence avec le modèle intégré ne prenant pas en compte d'E.P. (E.P – Intégré)

**Figure 9 :** Effet des co-variables environnementales (Bathymétrie, SST et Chlorophylle A) dans un modèle prenant en compte des données satellitaires ou issues des sorties du modèle physico-biogéochimique POLCOM-ERSEM.

**Figure 9 :** Cartes d'inférences en prenant en compte les co-variables environnementales (Bathymétrie, SST et Chlorophylle A issue des données satellitaires), les écarts types associés et de la différence avec le modèle intégré ‘simple’. (Covariables – Intégré).

**Figure 11 :** Capacités prédictives des différents modèles par source de données, AUC à gauche et CPO à droite.

**Figure 12 :** Cartographie mensuelle de la probabilité de présence de la sardine dans le GdG en 2018

**Figure 13 :** Cartographie mensuelle de l'écart-type de la probabilité de présence de la sardine dans le GdG en 2018.

**Figure 14 :** Score AUC du modèle spatio-temporel par mois et source de donnée

**Figure 15 :** AUC moyen par mois et source de donnée (2009-2018) et écarts types. Le trait rouge correspond à une AUC de 0.70.

**Figure 16 :** Cartographies de la probabilité de présence de la sardine et d'écarts-types associés dans le GdG entre mai et octobre 2016

---

## Liste des Tableaux

**Tableau 1 :** Intérêts et limites de l'approche intégrée

**Tableau 2 :** Différences entre cas d'étude pélagique et benthico-démersaux

## Liste des Annexes

**Annexe 1 :** Analyse exploratoire des données commerciales : Spatialisation des CPUE en Mai 2018 de la flottille PTM\_SPF.

**Annexe 2 :** Analyse exploratoire des données commerciales : Spatialisation des CPUE en Mai 2018 de la flottille PTM\_DEF.

**Annexe 3 :** Analyse exploratoire des données commerciales : Spatialisation des CPUE en Mai 2018 de la flottille PS\_SPF.

**Annexe 4 :** Séries temporelles des profils de débarquements entre 2008-2018 à un pas de temps mensuel pour les 3 flottilles commerciales.

**Annexe 5 :** Cartographie de la SST en Mai 2018 (données issues du modèle biogéochimique POLCOM-ERSEM à gauche et satellitaires à droite).

**Annexe 6 :** Cartographie de la chlorophylle A en Mai 2018 (données issues du modèle biogéochimique POLCOM-ERSEM à gauche et satellitaires à droite).

**Annexe 7 :** Cartographie de la bathymétrie en Mai 2018.

**Annexe 8 :** Fonctionnement des PC priors et récapitulatif des paramétrisations testées

**Annexe 8 :** Tableau récapitulatif des différentes modélisations testées (toutes les combinaisons n'ont pas été testées).

**Annexe 9 :** Maillages utilisées

**Annexe 10 :** Inférences obtenues pour chaque variables environnementale (SST, chlorophylle A, bathymétrie).

**Annexe 11 :** Comparaison effet des conditions environnementaux avec ou non une prise en compte de l'effet aléatoire (données satellitaires).

**Annexe 14 :** Courbes ROC et AUC pour chaque source de données et modèles.

**Annexe 15 :** Cartographies des CPO pour chaque source de donnée issue du modèle « donnée seule ». Les CPO sont représentées sur les cartes d'inférences respectives.

**Annexe 16 :** Cartographies des CPO pour chaque source de donnée et issue du modèle « donnée seule », et différence avec la CPO de ce modèle.

**Annexe 18 :** Scores AUC des modèles spatio-temporels par mois et source de donnée entre 2009 et 2018. Le trait rouge correspond à une AUC de 0.70

**Annexe 19 :** Cartographies de probabilité de présence de la sardine dans le GdG en 2009.

**Annexe 20 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2009.

---

**Annexe 21** : Cartographies de probabilité de présence de la sardine dans le GdG en 2010

**Annexe 22** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2010.

**Annexe 23** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2011.

**Annexe 24** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2011.

**Annexe 25** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2012.

**Annexe 26** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2012.

**Annexe 27** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2013

**Annexe 28** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2013.

**Annexe 29** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2014

**Annexe 30** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2014.

**Annexe 31** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2015.

**Annexe 32** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2015.

**Annexe 33** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2016

**Annexe 34** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2016.

**Annexe 35** : Cartographies de la probabilité de présence de la sardine dans le GdG en 2017.

**Annexe 36** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2017.

**Annexe 37** : Cartographies moyennes de la probabilité de présence de la sardine dans le GdG entre 2009-2018

**Annexe 38** : Cartographies moyennes des écarts types de la probabilité de présence de la sardine dans le GdG entre 2009-2018.

**Annexe 39** : Cartographies de probabilités de présence d'œufs de sardines entre 2000 et 2004 (Bellier et al, 2007), à gauche cartes de présence moyenne et coefficients de variations à droite.

**Annexe 40** : Cartographies d'abondance d'œufs de sardines (2000-2016) (Huret et al, 2018), à gauche cartes de d'abondances moyennes et coefficients de variations à droite.

**Annexe 41** : Cartographies de biomasse de sardines (2000-2015) (Doray et al, 2017a), à gauche cartes de présence moyenne et coefficients de variations à droite.

---

**Annexe 42 :** Distribution moyenne des abondances d'œufs de sardines réparties en 2 groupes à partir d'un cluster sur 13 MAFS et exemples de 4 cas représentatifs des 2 groupes (G1 au-dessus : 2000 et 2008 ; G2 en-dessous : 2010 et 2017) après krigeage à partir du modèle MAF (Données PELGAS 2000- 2017) (Petitgas *et al.*, 2020).

---

## Liste des abréviations

- ADNe : Acide Désoxyribo-Nucléique environnemental
- AUC : Area Under the Curve
- AMP : Aires Marines Protégées
- CPO : Conditional Predictive Ordinate
- CPUE : Captures Par Unité d'Effort
- EMR : Energies Marines Renouvelables
- ESDU : Unité horizontale d'échantillonnage élémentaire, "Elementary Sampling Distance Unit", un ESDU correspond à un mile nautique d'échantillonnage acoustique
- GdG : Golfe de Gascogne
- GPS : Système de Positionnement Global
- IFREMER : Institut Français de Recherche pour l'Exploitation de la Mer
- INLA : Integrated Nested Laplace Approximation
- MCMC : Markov Chain Monte Carlo
- PELGAS : PELagique GAScogne
- PTM\_DEF : Pelagic Pair Trawl Demersal pelagic Fish
- PTM\_SPF : Pelagic Pair Trawl Small Pelagic Fish
- PS\_SPF : Purse Seine Small Pelagic Fish
- ROC : Receiver operating characteristic
- SDM : Species Distribution Model
- SST : Sea Surface Temperature
- VMS : Vessel Monitoring Systems

---

## 1. Introduction

### 1.1. Généralités : distribution d'espèce et SDM

La bonne compréhension de la distribution des espèces marines est fondamentale pour la compréhension du fonctionnement des populations et des écosystèmes. Elle présente des enjeux majeurs pour la gestion et la conservation des espèces marines (Gaston, 2000; Lamoreux *et al.*, 2006). L'établissement de plans de gestion spatialisés (Wright and Kyhn, 2015), notamment les AMP, les parcs éoliens en mer (EMR) ou encore la délimitation de zones de pêches (Hunter *et al.*, 2006; Meyer *et al.*, 2007) nécessitent une connaissance fine de la distribution spatiale et saisonnière des organismes marins de manière générale et des stocks exploités en particulier (Booth, 2000).

Dans les écosystèmes marins, de nombreux facteurs affectent la répartition des poissons dans le temps et dans l'espace (Whittaker *et al.*, 1973) et rendent la détermination des aires de distribution complexe. En effet, les besoins d'une espèce varient selon un ensemble de conditions biotiques et abiotiques (habitat physique, relations trophiques, conditions physico-chimiques) desquelles vont dépendre leur distribution spatiale. Par ailleurs, leur cycle de vie se divise en différentes phases, correspondant chacune à des besoins précis et peut s'accompagner de changements d'habitats au cours de l'ontogénie (Delage et Le Pape, 2016). Enfin, l'homme est l'un des principaux prédateurs des espèces marines et les impacts liés aux activités anthropiques peuvent considérablement modifier la structure des communautés de poissons et leur répartition (Worm *et al.*, 2009). Notamment, l'exploitation par la pêche n'a pas lieu au hasard dans l'espace et le temps. Les pêcheurs sélectionnent le lieu et le moment de la pêche en fonction de leurs connaissances de l'écologie des espèces et des règles qui régulent l'activité de pêche (Branch *et al.*, 2006; Vermaud *et al.*, 2008; van Putten *et al.*, 2012). En conséquence, la distribution des espèces marines peut potentiellement changer sur des périodes relativement courtes, le changement de distribution reflétant l'effet conjoint de l'activité de pêche, de l'environnement et du cycle biologique des espèces.

Les modèles de distribution d'espèces (SDM) sont des outils essentiels en écologie marine pour améliorer notre connaissance de la distribution des espèces et de la relation entre espèces et habitats. Ils permettent notamment de comprendre l'influence de l'environnement et des pressions anthropiques sur la distribution des espèces (Marshall *et al.*, 2014). Aujourd'hui, le développement des nouvelles technologies et techniques d'informations comme les ADNe (Bálint *et al.*, 2018), les suivis par acoustique passive (Gibb *et al.*, 2019) ou encore la multiplication des campagnes de science participative (August *et al.*, 2015), permets d'accéder à une grande diversité de sources de données, et présente de nouvelles opportunités pour cartographier les espèces. Cependant, la diversité des sources de données peut rendre leur intégration dans un seul et même modèle complexe (i.e. les données ne sont pas nécessairement de même nature – e.g. données de comptage, de présence-absence, de présence seulement). Un challenge important consiste à développer de nouvelles méthodologies permettant d'intégrer ces différentes sources de données afin d'en tirer une information synthétique sur la distribution des espèces (Isaac *et al.*, 2020).

---

## 1.2. Données scientifiques et données commerciales

En halieutique, les données de référence pour cartographier les espèces marines sont généralement issues de campagnes océanographiques. Ces campagnes scientifiques ont pour but de récolter des données représentatives des populations étudiées via un plan d'échantillonnage et une capturabilité standardisée. Cependant, ces campagnes ont un coût élevé et par conséquent, ont généralement une couverture temporelle restreinte (en général, elles ont lieu une à deux fois par an toujours à la même période) (Hilborn and Walters, 1992; ICES, 2005; Nielsen, 2015). Elles ont donc une utilité limitée pour étudier la dynamique intra-annuelle des espèces étudiées.

De nouvelles sources de données sont aujourd’hui rendues disponibles par le biais des dispositifs de surveillance des navires de pêche (Vessel Monitoring Systems – VMS). D’abord mis en place au Portugal en 1988 afin de suivre l’activité des pêcheurs suite à la dégradation des stocks (Navigs, 2005), le système VMS s’est rapidement généralisé depuis le début du 21<sup>ème</sup> siècle (Marzuki, 2017). Ces dispositifs, aujourd’hui installés à bord de tous les navires sous pavillon européen de plus de 12 m, transmettent des données (position GPS, cap, vitesse, identification du navire...) par communication satellitaire toutes les heures aux centres de surveillance des pêches. Les points VMS (ou pings) peuvent ensuite être utilisés pour calculer un effort de pêche en identifiant les points VMS émis lors de l’activité de pêche sur la base de différents critères ou méthodes (Vermaud *et al.*, 2008; Bez *et al.*, 2011; Gerritsen and Lordan, 2011; Hintzen *et al.*, 2012; Murray *et al.*, 2013).

En parallèle à l’acquisition de ces données VMS, les données de déclaration de captures sont déclarées dans les livres de bord des navires de pêche (données « logbooks »), dans lesquels les pêcheurs indiquent les quantités capturées par espèce à l’échelle d’un carré statistique (de 1° longitude par 0.5° de latitude) pour chaque journée de pêche, chaque engin utilisé et chaque marée. Cependant, la résolution spatiale de ces données (le carré statistique) demeure assez grossière ce qui limite leur utilisation pour des analyses à fine échelle spatiale (Marrs *et al.*, 2002; Lee *et al.*, 2010).

Chaque donnée de déclaration de pêche (quantités capturées par espèce) peut être associée à un ensemble de pings VMS identifiés en pêche (position de pêche des navires de pêche). Ainsi, en ventilant les données logbooks sur les pings VMS correspondant, il est possible de passer d’une image de la distribution des captures et des CPUE grossière, définie sur un domaine discret (i.e. les carrés statistiques) à une résolution spatio-temporelle fine et continue dans l'espace (Gerritsen and Lordan, 2011 ; Murray *et al.*, 2013). Ainsi, le croisement des données VMS et des données logbooks constituent une source de données additionnelle pour décrire la distribution des ressources exploitées. Elle est notamment susceptible de fournir de l’information en dehors des périodes couvertes par la donnée scientifique.

Les données commerciales présentent cependant un inconvénient majeur : les pêcheurs vont avoir tendance à cibler des zones poissonneuses où les abondances sont les plus élevées (Conn *et al.*, 2017; Pennino *et al.*, 2019; Rufener, 2020) afin de maximiser leurs captures et leurs revenus (Vermaud *et al.*, 2008). Ce phénomène, appelé “échantillonnage préférentiel”, introduit un biais systématique dans les estimations et les prédictions de modèles qui peuvent entraîner une surestimation de l’abondance dans les zones où l’échantillonnage est faible (Conn *et al.*, 2017; Pennino *et al.*, 2019; Alglave *et al.*, under review).

---

Par ailleurs, d'autres composantes peuvent intervenir dans les prises de décisions des pêcheurs et la distribution de l'effort de pêche (Salas and Gaertner, 2004; Abbott *et al.*, 2015; Girardin *et al.*, 2017; Stephenson *et al.*, 2018). En particulier, leur activité de pêche peut être fortement structurée par les exigences du marché et des effets filières limitant la dispersion de l'activité de pêche à des saisons et/ou des zones plus ou moins restreintes.

### 1.3. Des modèles hiérarchiques pour intégrer différentes sources de données

L'intérêt de développer des modèles hiérarchiques intégrés pour combiner différentes sources de données est maintenant bien établi en écologie (Schaub *et al.*, 2007; Zipkin *et al.*, 2019) et en sciences halieutiques en particulier (Maunder and Punt, 2013; Rochette *et al.*, 2013; Archambault *et al.*, 2018; Punt *et al.*, 2020; Rufener, 2020). En effet, en considérant que différentes sources de données sont des réalisations d'un même processus sous-jacent modélisé sur la base de variables latentes non directement observées, il est possible de combiner différentes sources de données dans un même modèle et ainsi d'améliorer l'estimation des paramètres du modèle, d'étendre des séries temporelles, d'élargir l'aire d'étude par rapport à une analyse séparée de différentes sources de données (Moriarty *et al.*, 2020).

Plusieurs auteurs ont déjà développé des modèles combinant données commerciales et données scientifiques pour tirer profit des deux sources de données - et ainsi prédire la distribution des espèces d'intérêt halieutique à une résolution spatio-temporelle plus fine (Alglave *et al.*, under review; Rufener, 2020; Gonzalez *et al.*, 2021). En particulier, le modèle développé par Alglave *et al.* (under review) permet de combiner les données scientifiques issues de campagnes de chalutage et les données 'logbooks x VMS', tout en prenant en compte le comportement de ciblage des pêcheurs. Cette approche a été appliquée à plusieurs espèces démersales du Golfe de Gascogne (en particulier la sole ainsi que d'autres espèces telles que la baudroie, le merlu, l'encrenet). Il en ressort que les inférences dépendent principalement des données commerciales, en comparaison à la donnée scientifique qui n'a qu'une faible influence du fait du nombre restreint d'échantillons (en ordre de grandeur, 3000 échantillons commerciaux contre 50 échantillons scientifiques). De plus, il apparaît que les engins filtrés pour décrire la distribution des espèces d'intérêt (ici les chalutiers démersaux) ont un comportement opportuniste et n'ont pas un comportement de ciblage élevé pour une espèce en particulier (la proportion des espèces majoritaires dans les débarquements n'excède pas 20 – 25 % en général). Ainsi, le comportement de ciblage des pêcheurs, et en particulier l'effet de l'échantillonnage préférentiel vis-à-vis de la ressource, n'a que peu d'impact sur les prédictions spatiales des modèles.

Dans ce travail, nous proposons d'explorer l'intérêt du développement de modèles de distribution d'espèces intégrés combinant données scientifiques et donnée 'VMS x logbooks' pour des espèces de petits pélagiques comme l'anchois ou la sardine. Les petits poissons pélagiques (e.g. sardine, anchois, sprat) jouent un rôle écologique clé dans les écosystèmes côtiers, en transférant l'énergie du plancton aux niveaux trophiques supérieurs (Cury *et al.*, 2000; Spitz *et al.*, 2018). Par ailleurs, leur position trophique relativement basse dans le réseau alimentaire marin, ainsi que leur courte durée de vie et leur stratégie de reproduction consistant à produire de grandes quantités d'œufs pélagiques au cours d'une saison de frai prolongée, les rendent très dépendants de l'environnement (Hunter, and Alheit, 1995; Bakun, 1996).

Par ailleurs, la caractérisation de leur distribution soulève des enjeux méthodologiques importants. Les petits poissons pélagiques sont très mobiles par rapport aux espèces démersales et benthiques et effectuent des migrations plus fréquentes et plus importantes, à des échelles

---

allant de quelques jours à quelques semaines et de quelques dizaines de mètres à quelques dizaines de kilomètres, afin de faire face aux fluctuations plus importantes des conditions hydrobiologiques de l'écosystème pélagique (Mackinson, 1999; Robinson, 2004; Barange *et al.*, 2005). Enfin, ils font l'objet d'un ciblage fort de la part des flottilles qui les exploitent, ce qui renforce l'intérêt de développer des méthodes permettant de prendre en compte explicitement l'échantillonnage préférentiel pour l'intégration des données commerciales.

Le cas d'étude abordé dans ce travail est la sardine européenne (*Sardina pilchardus*) dans le Golfe de Gascogne. La sardine est un clupéidé sténotherme au comportement gréginaire qui habite généralement des eaux dont la température varie de 8°C à 24°C et la salinité de 30 à 38 psu (Coombs *et al.*, 2006; Petitgas *et al.*, 2006; Stratoudakis *et al.*, 2007). Suivant son âge et son état physiologique elle recherche des conditions environnementales différentes, rendant sa répartition dans le GdG discontinue et variable.

L'étude des populations de sardine en GdG a commencé au début du siècle précédent, mais n'a été l'objet d'un échantillonnage systématique le long du plateau continental qu'à partir des années 1960 (cartographies de larves et des œufs de sardine - L'Herrou, 1967; Arbault and Lacroix, 1970). Ces premières campagnes ont d'abord été utilisées par Arbault et Lacroix (1970, 1977) pour réaliser les premières cartographies des zones de frayères de sardine en apportant des indications sur les secteurs de ponte, la saison de frai ainsi que les conditions favorables du milieu. Ces premières études présentent des résultats sur l'ensemble de l'année (une carte par saison) et sont les seules disponibles à ce jour. Par la suite les cartographies ont été réalisées à partir des données issues de la campagne PELGAS qui a lieu au printemps de chaque année (la campagne est centrée sur le mois de mai). La donnée scientifique pour décrire la répartition de la sardine est donc essentiellement limitée au mois de mai et, à notre connaissance, aucun travail n'a cherché à inférer la distribution de la sardine en dehors de cette période depuis les années 1970.

Ainsi ce travail a pour objectif de construire et d'ajuster un modèle combinant données scientifiques (PELGAS au printemps et JUVENA en automne) et données 'VMS x logbooks' pour inférer la distribution de la sardine du GdG sur l'ensemble de l'année. Ce travail doit nous permettre :

- D'appliquer un modèle intégré à des cas de d'étude différents des espèces étudiées jusqu'à maintenant (i.e. des espèces benthico-démersales) et de voir dans quelle mesure les conclusions sont modifiées pour le cas d'espèces pélagiques.
- D'explorer l'apport des données commerciales pour inférer la distribution de la sardine en dehors de la période de campagne.

Le modèle est développé sur la base du cadre de modélisation intégré décrit par Alglave *et al.* (under review), et adapté pour son application à un cas d'étude de petit pélagique.

Dans un premier temps, ce modèle est développé pour inférer la distribution sur un seul pas de temps (pas de temps mensuel) afin d'en évaluer plusieurs caractéristiques :

- (i) la contribution des différentes sources de données dans l'inférence ;
- (ii) l'effet de la prise en compte de l'échantillonnage préférentiel des flottilles commerciales ;
- (iii) l'apport de covariables environnementales sur les prédictions du modèle.

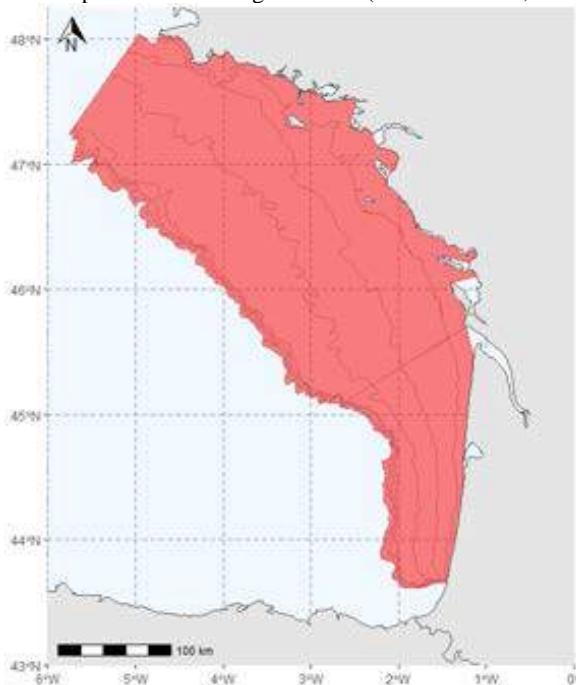
Dans un second temps, nous avons étendu les analyses à la période 2009-2018 (toujours sur un pas de temps mensuel) pour explorer la variabilité temporelle de la distribution spatiale de la sardine.

---

## 2. Matériel et méthode.

### 2.1. Description du cas d'étude : la sardine du Golfe de Gascogne.

La sardine européenne, *Sardina pilchardus* (Walbaum, 1792) est une espèce pélagique appartenant à l'ordre des clupéiformes, la famille des *Clupeidae*, et la sous-famille des *Alosinae*. L'aire de répartition de la sardine s'étend des côtes mauritaniennes (Furnestin and Furnestin, 1959; Ettahiri *et al.*, 2003) jusqu'aux eaux de la Norvège en incluant la mer Méditerranée et la Mer Noire (Parrish *et al.*, 1989). Les sardines vivent en bancs très denses. Elles sont proches du fond le jour et elles remontent en surface en se dispersant la nuit (Whitehead, 1985). On la retrouve au-dessus du plateau continental jusqu'à des profondeurs de 120 m, mais elle abonde surtout le jour entre 30 et 55 m et la nuit entre 15 et 40 m de la surface (Quero *et al.*, 1989). La sardine effectue des déplacements saisonniers de faible amplitude régies par des besoins nutritionnels, reproducteurs ou en lien avec les conditions thermiques. Elle migre en automne vers le large et se rapproche des côtes au printemps. Selon la saison, l'âge et l'état sexuel elle réalise également des déplacements le long des côtes (Tosello-Bancal, 1994).



**Figure 1 :** Carte du Golfe de Gascogne et délimitation du domaine étudié.

A ce jour, les études suggèrent des populations auto-suffisantes (uniques et sans échanges avec les autres) distinctes en Manche/Mer Celtique et dans le Golfe de Gascogne (Gatti *et al.*, 2018; Lavialle *et al.*, 2019). En revanche la comparaison des tailles et structures en âge semblent indiquer un lien entre les cohortes observées en Mer Cantabrienne et les recrutements dans le GdG, suggérant ainsi l'existence d'une connectivité entre ces zones (Silva *et al.*, 2009; Silva *et al.*, 2019). Néanmoins, en l'absence de certitudes sur l'intensité des flux entre ces zones, le

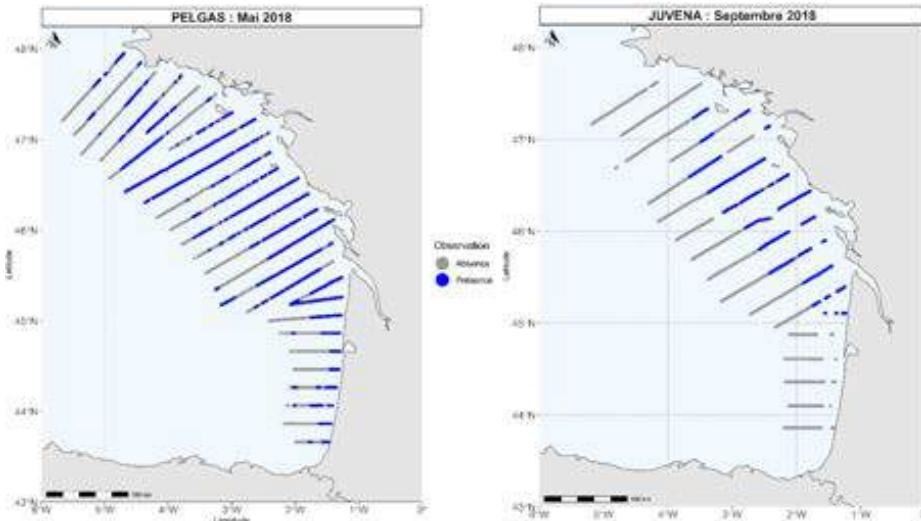
CIEM réalise une évaluation en considérant un stock de sardine indépendant pour le GdG (ICES, 2016). Elle n'est actuellement soumise à aucune réglementation commerciale, en dehors d'une taille minimale de capture fixée à 11 cm. Surpêchée depuis 2014, sa dernière évaluation indique que ce stock a enfin atteint le bon état écologique en 2019, avec 24 400 T capturées. (ICES, 2020)

Cette étude se limite à la zone dite française du GdG (Figure 1). Cette zone correspond au domaine échantillonné par la campagne scientifique PELGAS. Elle comprend le plateau continental situé entre la pointe du Raz et la fosse du Cap Breton.

## 2.2. Données

### 2.2.1. Données scientifiques issues de campagnes acoustiques

Deux campagnes scientifiques ont lieu chaque année dans le GdG (Doray *et al.*, 2018b; Boyra *et al.*, 2020). Leur objectif principal est d'estimer la biomasse et de prédire le recrutement des petits pélagiques. PELGAS a lieu au printemps depuis 2000. Elle est centrée sur le mois de mai et est réalisée par l'IFREMER à bord du RV "Thalassa". La campagne JUVENA a lieu depuis 2003 et est centrée sur le mois de septembre (Figure 2). Elle a un focus sur les populations de juvéniles d'anchois et couvre les côtes espagnole et française. Seuls les échantillons situés sur le domaine d'étude ont été utilisés dans l'analyse (i.e. les échantillons situés dans le GdG). Son plan d'échantillonnage est adaptatif et est donc susceptible de varier entre les années : les limites de la zone échantillonnée sont redéfinies chaque année afin de couvrir l'ensemble de l'aire de distribution de l'anchois, connue via les données commerciales et les précédentes campagnes scientifiques.



**Figure 2 :** Observations (présence/absence) de *S. pilchardus* par la campagne scientifique PELGAS au cours du mois de Mai 2018 et par la campagne scientifique JUVENA au cours du mois de Septembre 2018.

PELGAS et JUVENA sont des campagnes acoustiques qui utilisent un écho-sondeur pour émettre de courtes impulsions électriques, transmises sous forme d'impulsions ultrasonores, vers les fonds marins. En présence de poissons, une partie de l'énergie est rétrodiffusée vers l'émetteur. L'écho-intégration permet d'évaluer la biomasse présente dans une zone à partir du

---

cumul de l'énergie acoustique rétrodiffusée par l'ensemble des cibles présentes (MacLennan *et al.*, 2002). Ces écho-intégrations sont réalisées à l'échelle d'une ESDU (Unité horizontale d'échantillonnage élémentaire, "Elementary Sampling Distance Unit") qui correspond à un mile nautique d'échantillonnage acoustique. Les densités acoustiques, issues de l'écho-intégration, correspondent à un assemblage d'espèces.

Pour déterminer la composition en espèce et en taille des ESDU, des chalutages pélagiques sont effectués régulièrement (2 à 3 chalutages par jour) et la composition en espèce et en taille des coups de chaluts sont réalloués aux écho-intégrales par ESDU soit : (i) par utilisation d'un trait de chalut de référence, (ii) en définissant des régions où les assemblages espèces/tailles sont homogènes, (iii) ou à l'aide d'un modèle géostatistique, une estimation est obtenue aux noeuds d'une grille et est suivie d'une interpolation (Doray *et al.*, 2010).

La majorité des poissons pélagiques étant souvent hors de portée du sondeur (dans une couche d'eau entre la surface et 10 mètres d'immersion) pendant les périodes de nuit, la prospection acoustique se réalise essentiellement de jour (Doray *et al.*, 2018b; Boyra *et al.*, 2020).

Pour les données PELGAS, on obtient ainsi des biomasses et des abondances par espèce, classe de taille, et ESDU. Pour JUVENA seule une biomasse par espèce et ESDU est indiquée.

La taille réglementaire de capture de la sardine étant fixée à 11 cm, seules les observations scientifiques supérieures à cette taille sont conservées pour la suite des analyses si le jeu de données le permet.

#### 2.2.2. Données commerciales (logbooks & VMS)

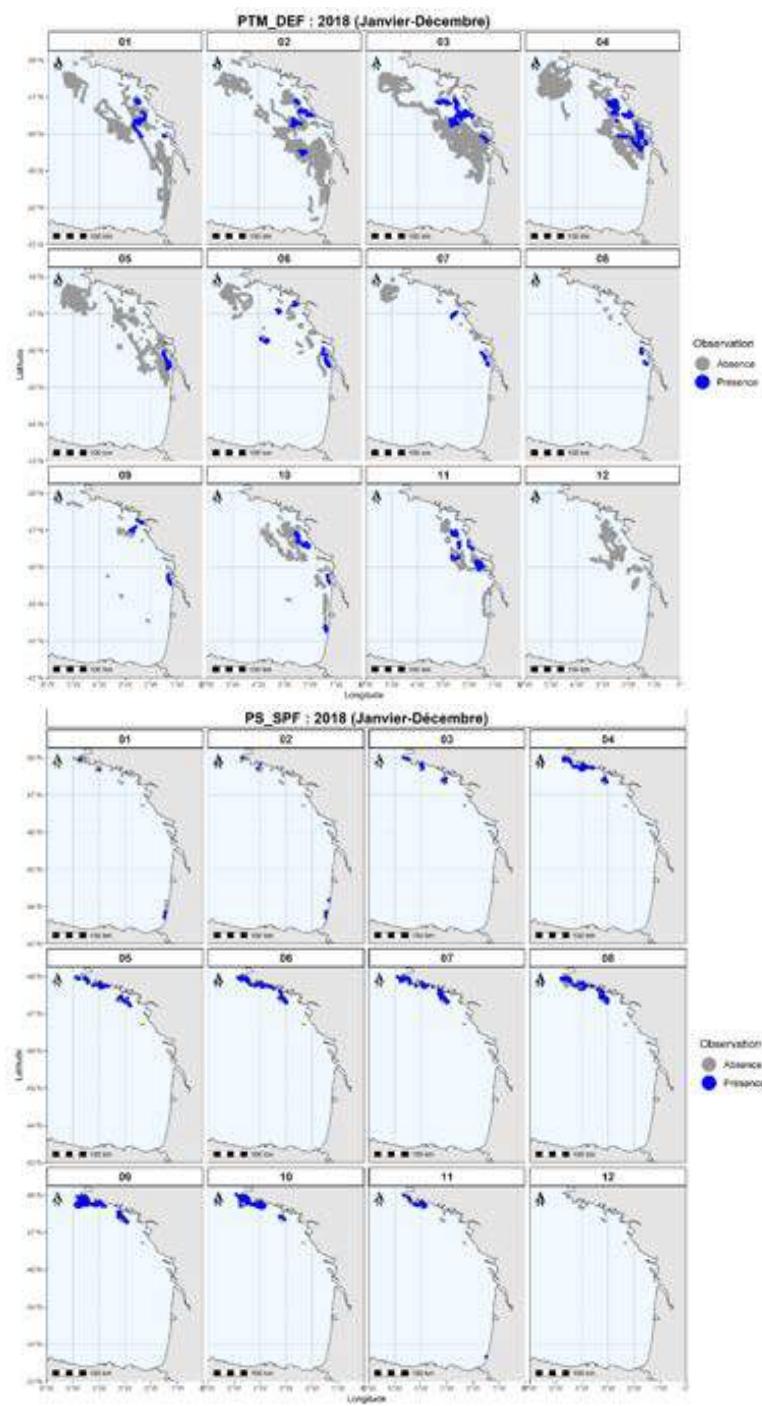
Le système de surveillance des navires par satellite (VMS) fournit à intervalles réguliers (toutes les heures) des données sur la position, la direction et la vitesse des navires aux autorités de pêche. Il est obligatoire pour les navires de pêche professionnelle de plus de 12 mètres, sous pavillon de l'Union européenne, depuis le 1er janvier 2012, et permet un suivi de l'activité de pêche à une résolution spatio-temporelle très fine.

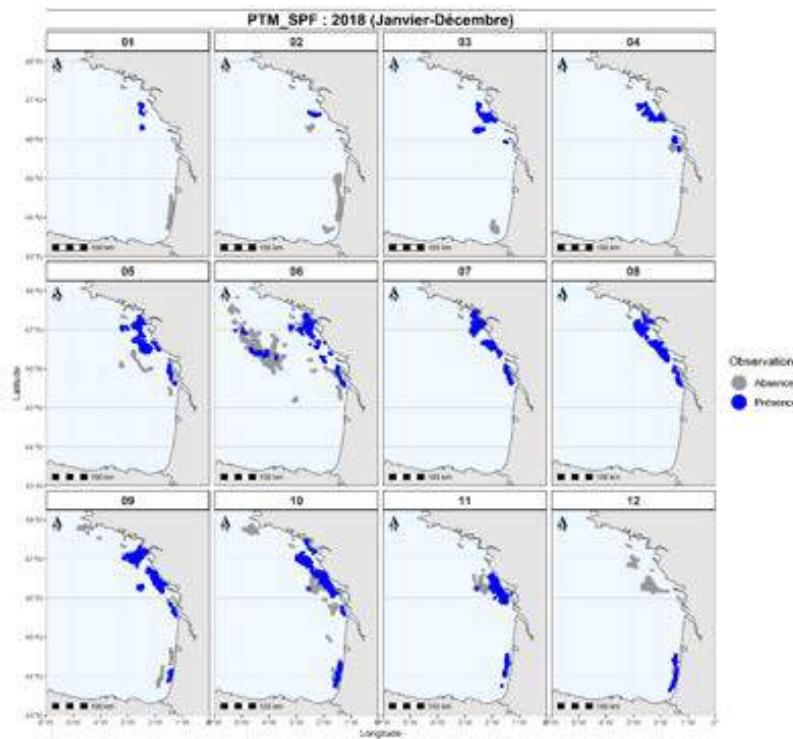
Ces informations spatialisées sont couplées aux données de débarquements (logbooks) suivant la méthodologie décrite par Hintzen *et al.* (2012). Afin de classifier les navires en flottilles homogènes, une analyse exploratoire des captures, des efforts et des CPUE a été réalisée sur la période 2009 – 2018 à un pas de temps mensuel pour tous les métiers ayant pêché au moins une fois la sardine (Annexes 1 à 4).

A la suite de cette analyse, les données commerciales sont ainsi réunies en groupes de navires utilisant les mêmes engins et ayant un comportement de ciblage relativement homogène (i.e. groupe d'espèce pêchée similaire). Ces groupes de navires sont appelés flottilles dans la suite du rapport. Seules les flottilles pêchant sur le domaine d'étude et présentant une proportion de présence de sardine non négligeable dans les captures (> 1% de présence) sont conservées pour l'analyse (Annexe 4). Finalement, les données commerciales ont été regroupées en 3 flottilles :

- les chaluts à bœufs pélagiques ciblant les petits pélagiques (PTM\_SPF) ;
- les chaluts à bœuf pélagiques ciblant les espèces démersales (PTM\_DEF) ;
- les senneurs ciblant les petits pélagiques (PS\_SPF).

Ces 3 flottilles sont marquées par une forte saisonnalité et présentent une activité essentiellement côtière (Figure 3). La conséquence du regroupement de PTM\_SPF et PTM\_DEF en une seule flottille a été étudiée au cours du stage.





**Figure 2 :** Observations (présence/absence) de *S. pilchardus* par les 3 flottilles commerciales entre janvier et décembre 2018

### 2.3. Spécificités du cas d'étude et conséquence pour la modélisation

Dans le cas des espèces pélagiques, les CPUE issues des flottilles commerciales ne peuvent pas être considérées comme des indices d'abondance proportionnels à la biomasse. En effet, les poissons ont des comportements plus ou moins grégaires, durant les phases de fraie ou d'alimentation par exemple, et certaines espèces se déplacent en bancs de tailles importantes et très denses. Ces comportements d'agrégation déterminent en partie l'espace occupé par les individus d'un stock (Pitcher, 1980). L'organisation en bancs est commune chez les poissons, Burgess et Shaw (1979) estiment que 80% des poissons ont un tel comportement au cours de leur cycle de vie, en particulier au stade juvénile. Ce comportement est prévalant chez les espèces pélagiques, comme les scombridés, les carangidés ou encore les clupéidés, et cela à des conséquences majeures sur le fonctionnement des pêches associées à leur exploitation. En effet, dans ce cas, la relation entre biomasse et captures-par-unité d'effort n'est plus linéaire. Du fait du comportement d'agrégation de la ressource, les CPUE restent relativement stables quel que soit le niveau de biomasse sous-jacent (ce phénomène est aussi connu sous le nom d'hyperstabilité des CPUE - Pitcher, 1995; Fréon et Misund, 1999).

Dans ce cas, les données de CPUE commerciales ne peuvent pas être utilisées directement pour modéliser la distribution de la biomasse de sardine. Par conséquent, les données commerciales et les données scientifiques sont dégradées en données de présence-absence et le cadre de modélisation développé pour intégrer ces données doit être adapté en conséquence.

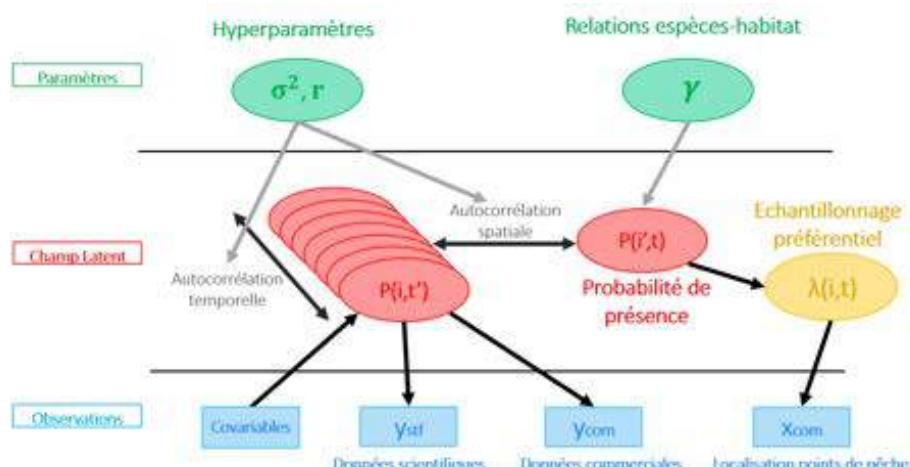
## 2.4. Traitement des données

Les données scientifiques sont filtrées pour conserver uniquement les sardines d'une taille supérieure à 11 cm (possible uniquement avec les données PELGAS). Elles sont ensuite groupées par 'Espèce x ESDU x Date' (à chaque ESDU correspond une unique position GPS). Si une biomasse non nulle est mesurée, on indique une présence (pres = 1), dans le cas contraire on indique une absence (pres = 0). Quand aucune information n'est spécifiée pour la sardine, mais que des biomasses sont mesurées pour d'autres espèces, on ajoute une absence pour cette ESDU. Enfin, toutes les observations sont considérées comme étant réalisées sur le mois de calibration de la campagne (i.e. mai pour PELGAS et septembre pour JUVENA)

Les données commerciales de captures sont traitées en suivant la méthode développée par Hintzen et al (2012). D'une part, on filtre les ping VMS pour conserver uniquement pings émis lors de l'activité de pêche (vitesse moyenne supérieur à 4.5 noeuds – algorithme AlgoPesca développé par l'IFREMER). Les déclarations de pêche (logbooks) sont déclarées à l'échelle d'une séquence de pêche (i.e. la combinaison 'carré statistique x engin utilisé x marée x jour'). Les quantités déclarées sont réallouées de façon uniforme sur les pings VMS de la séquence de pêche correspondante. Les présence/absence sont calculées de la même manière que précédemment (biomasse non nulle/nulle).

## 2.5. Modèle intégré hiérarchique

Le modèle hiérarchique est structuré en 3 couches de variables reliées entre elles par des relations de dépendances probabilistes : (1) la couche des observations (scientifiques et commerciales) ; les observations sont conditionnelles au (2) champ latent (non observé) ; du fait de la nature des observations (présence-absence), le champ latent est un champ de présence-absence ; (3) les hyperparamètres contrôlent la structure du champ latent.



**Figure 3 :** Diagramme du modèle spatio-temporel hiérarchique intégré (d'après Alglave et al ; under review)

Dans ce modèle, les équations d'observation permettent de relier les différences sources de données (données de présence absence commerciales et scientifique) au champ latent de

---

présence absence. Le modèle permet de représenter explicitement le fait que la position spatiale des points d'échantillonnage commerciaux dépend du champ latent de présence/absence (Figure 4). En effet, les pêcheurs tendent à cibler des zones où la ressource est présente (i.e. l'échantillonnage est dit préférentiel – Diggle *et al.*, 2010) ce qui peut introduire un biais dans les prédictions si l'échantillonnage de la donnée commerciale et le lien avec la variable à modéliser n'est pas pris en compte dans la méthode d'inférence.

### 2.5.1. Champ latent de présence-absence

Soit une variable aléatoire  $Z(i,t)$  décrivant la présence de sardine au site  $i$  au pas de temps  $t$ , elle prend la valeur 1 ou 0 selon qu'une sardine est présente ( $Z(i,t) = 1$ ) ou pas ( $Z(i,t) = 0$ ). La variable aléatoire  $W_{i,t}$  suit donc une loi de Bernoulli de paramètre  $P_{i,t}$  (Eq. 1).

$$Z(i,t) \sim \text{Bernoulli}(P(i,t)) \quad (\text{Eq. 1})$$

$P(i,t)$  représente la probabilité de présence de sardines au site  $i$  au temps  $t$ .

Cette probabilité de présence est modélisée par un champ aléatoire gaussien (dans l'échelle du logit) sur un domaine discret. La valeur du champ latent au point  $i$  est exprimée comme une combinaison linéaire qui dépend de :

- Un intercept  $\mu$ , représentant la valeur moyenne du champ de présence/absence de sardine.
- Des co-variables environnementales  $C$  et un paramètre  $\gamma$  qui capture la relation espèce-habitat (effets fixes) ; Les  $k$  covariables  $C_k(i,t)$  avec la fonction  $f_k$  modélisant la relation espèce-habitat (possiblement non-linéaire). Dans le cas présent, la bathymétrie, la chlorophylle a et la température de surface (SST) ont été inclus dans le modèle (Zwolinski *et al.*, 2010 ; Doray, *com. pers.*). Les modèles ont été ajustés avec des données satellitaires de Copernic ainsi qu'avec les sorties du modèle physico-biogéochimiques POLCOM-ERSEM (SST et chlorophylle a) dont la résolution était plus fine. Les cartes de ces covariables au mois de mai 2018 sont disponibles en Annexes 5 à 7. Les covariables ont été normalisées pour l'ajustement afin de faciliter la convergence du modèle.
- Un effet fixe Mois  $V(t)$
- Un effet aléatoire spatio-temporel  $U(i,t)$  qui capture les autocorrélations spatiale et temporelle (Eq. 2). L'effet spatio-temporel est un champ aléatoire gaussien de moyenne zéro et dont la matrice de covariance suit une fonction de corrélation de Matérn caractérisée par un paramètre de portée  $r$  – la distance à laquelle la corrélation entre deux points est égal à 0.1 – et un paramètre de variance  $\sigma^2$  (Eq. 3). La corrélation temporelle est modélisée à l'aide d'un processus autorégressif d'ordre 1 (Blangiardo *et al.*, 2013, Krainski *et al.*, 2019).

$$\text{Logit}(P(i,t)) = \mu + V(t) + \sum_k f_k(C_k(i,t)) + U(i,t) \quad (\text{Eq. 2})$$

$$U(i,t) \sim \text{GMRF}(0, \Sigma(\sigma^2, r)) \quad (\text{Eq. 3})$$

---

Afin d'assurer l'identifiabilité du modèle, les paramètres associés au mois de mai ( $V(t_{mai})$  - mois de PELGAS) et à la campagne PELGAS  $\alpha(j_{PELGAS})$  sont fixés à 0 et leur effet est capturé par l'intercept  $\mu$ .

### 2.5.2. Modèle d'observation de la présence/absence

Soit la variable aléatoire  $Y(i, j, t)$  représentant l'observation d'un individu au site  $i$  au temps  $t$  par la flottille (commerciale ou scientifique)  $j$ . Le processus d'observation  $Y(i, j, t)$  est ainsi conditionnel au processus écologique  $Z(i, t)$ . Une sardine ne peut être observée au site  $i$  et au temps  $t$  que si elle est présente ( $Z(i, t) = 1$ ), mais la probabilité qu'elle soit détectée quand elle est présente n'est pas forcément 1 mais dépend de la détectabilité de la flottille  $j$  qu'on modélise à l'aide d'une probabilité de détection noté  $\delta(i, j, t)$  et dépendant éventuellement de covariables  $C'$  (Eq. 4).

$$P(Y(i, j, t) = y_{i,j,t}) = \begin{cases} P'(i, j, t) & : \text{Si } y_{i,j,t} = 1 \text{ (présence)} \\ 1 - P'(i, j, t) & : \text{Si } y_{i,j,t} = 0 \text{ (absence)} \end{cases} \quad (\text{Eq. 4})$$

$$P'(i, j, t) = Z(i, t) * \delta(i, j, t) \quad (\text{Eq. 5})$$

$$\text{Logit}(\delta(i, j, t)) = C_k'(i, j) * \rho \quad (\text{Eq. 6})$$

### 2.5.3. Modéliser l'échantillonnage préférentiel

L'échantillonnage des données commerciales ne suit pas de protocole standardisé et, en particulier, les points de pêches peuvent être préférentiellement répartis dans les zones où la probabilité de présence des espèces cibles est plus forte. Ce processus est appelé 'échantillonnage préférentiel' (Diggle *et al.*, 2010), il peut biaiser les sorties des méthodes d'inférence et conduire à la surestimation des prédictions dans les zones sous-échantillonées. Pour intégrer l'échantillonnage préférentiel à l'inférence, nous modélisons de façon explicite la distribution des points de pêche dans le modèle au travers d'un processus ponctuel poissonnien. Cette modélisation fait intervenir un second terme dans la vraisemblance (en plus du terme de vraisemblance associé à l'observation de la présence/absence) qui intègre la position GPS des points de pêche au travers de ce processus ponctuel poissonnien.

Notons  $x_{com,j}$  les points où les bateaux de la flottille  $j$  sont identifiés en pêche (Eq. 9).  $x_{com,j}$  suit un processus ponctuel non homogène de Poisson décrit par une intensité  $\lambda(i)$  variant dans l'espace.  $\lambda(i)$  correspond à l'intensité de pêche au point  $i$ . Cette intensité est modélisée comme une combinaison log-linéaire de l'effet aléatoire  $U$ , multiplié par un paramètre  $b_j$ , et un effet spatial résiduel  $W$  de même paramétrisation que  $U$  capturant l'effet de covariables non pris en compte par le terme associé à l'échantillonnage préférentiel (Eq. 10). Le paramètre  $b_j$  quantifie la force de l'échantillonnage préférentiel. Ainsi, une valeur de  $b_j = 0$  indique que l'échantillonnage de la donnée commerciale est indépendant du champ latent et  $b_j > 0$  indique que l'échantillonnage est préférentiel.

$$x_{com,j} \sim IPP(\lambda(i)) \quad (\text{Eq. 9})$$

$$\text{Log}(\lambda(i)) = \mu + b_j * U(i, j) + W(i, j) \quad (\text{Eq. 10})$$

---

L'échantillonnage préférentiel est analysé sur un seul pas de temps (en mai) afin d'identifier son effet sur les prédictions du modèle.

#### 2.5.4. Simplification induite par les contraintes d'INLA

L'inférence a été réalisée grâce au logiciel INLA (voir ci-dessous). Les contraintes de codage liées à INLA rendent impossible la séparation explicite des équations de processus (Eq. 1-3) et des équations d'observation (Eq. 4-6). Ainsi, la variable latente intermédiaire de présence absence  $Z$  ne peut pas être modélisée explicitement. Le modèle final intègre les facteurs écologiques responsables de la variabilité du champ latent de présence/absence et les facteurs de la variation de la détectabilité en une seule équation :

$$P(Y(i,j,t) = y_{i,j,t}) = \begin{cases} P(i,j,t) & : \text{Si } y_{i,j,t} = 1 \text{ (présence)} \\ 1 - P(i,j,t) & : \text{Si } y_{i,j,t} = 0 \text{ (absence)} \end{cases} \quad (\text{Eq. 7})$$

$$\text{logit}(P(i,j,t)) = \mu + \alpha(j) + V(t) + \sum_k f_k(C_k(i,t)) + U(i,t) \quad (\text{Eq. 8})$$

où  $\alpha(j)$  représente directement l'effet source de données capturant les différences de détectabilité entre les différentes sources de données ou les différentes flottilles

#### 2.5.5. Outil d'inférence : R-INLA

L'inférence a été réalisé dans un cadre bayésien grâce à l'outil R-INLA et au package Inlabru. R-INLA (Integrated Nested Laplace Approximation) est un outil d'inférence bayésien introduit par Rue *et al.* (2009). Il offre une alternative aux méthodes d'estimation des distributions a posteriori par échantillonnage de type MCMC pour l'estimation de champs gaussiens avec pour principal avantage d'être notablement plus rapide.

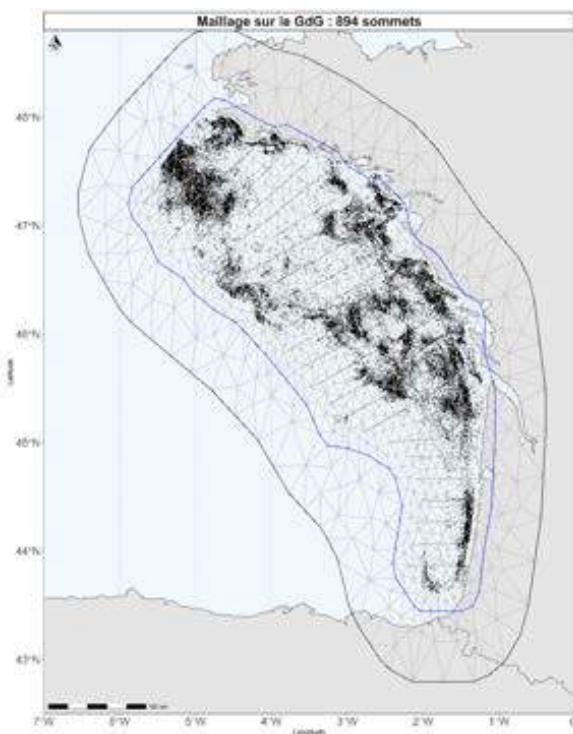
R-INLA combine 3 approximations distinctes faisant appel à l'approche SPDE (Stochastic Partial Differential Equations) (Approximation 1 et 2 - Lindgren *et al.*, 2011) et à l'approximation de type INLA (Approximation 3 - Rue *et al.*, 2009) :

- (1) Il approxime un champ gaussien par un champ Gauss-Markov (lorsque le champ gaussien admet comme fonction de corrélation une fonction de Matérn). Cette approximation permet d'obtenir une représentation creuse de la matrice de précision  $Q$  (propriétés markoviennes) et de simplifier les calculs matriciels liés à l'estimation de l'effet aléatoire spatial (Lindgren *et al.*, 2011).
- (2) Il diminue la résolution spatiale à laquelle est estimée la structure spatiale du champ latent en le modélisant à l'échelle d'une maille triangulaire creuse (Figure 5). Le lien entre les points de données (définis sur un domaine continu) et l'effet aléatoire (définie aux sommets de la maille) est réalisé via une interpolation linéaire des valeurs de l'effet aléatoire sur les points de données (Krainski *et al.*, 2019). Cette approximation permet d'estimer l'effet aléatoire sur un nombre de points plus faible (les nœuds de la maille) que si l'effet aléatoire avait dû être estimé en chaque point de données (en particulier lorsque les points de données sont nombreux). Différentes mailles ont été testées afin d'évaluer la sensibilité des sorties du modèle à la résolution de la maille (Annexe 8).

- (3) Il simplifie l'estimation des distributions a posteriori conjointe du champ latent  $U$  et des hyperparamètres  $\theta$  notée  $\pi(U, \theta|y)$  en mobilisant des méthodes d'approximations de Laplace. Au lieu d'utiliser des méthodes d'échantillonnage de type MCMC souvent très longues notamment quand le nombre de dimensions augmente, R-INLA approxime les distributions marginales à posteriori du champ latent  $\pi(U_i|y) = \int \pi(U_i|\theta, y)\pi(\theta|y)d\theta$  et des hyperparamètres  $\pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j}$  en simplifiant  $\pi(\theta|y)$  et  $\pi(U_i|\theta, y)$  via des approximations de Laplace emboîtées.

Dans le modèle spatio-temporel, la domaine spatial est défini par une SPDE et un modèle autorégressif d'ordre 1 (i.e. AR(1)) pour modéliser l'auto-corrélation temporelle (Blangiardo *et al.*, 2013, Krainski *et al.*, 2019). La relation espèce habitat, possiblement non-linéaire, est modélisée par une SPDE à une dimension.

Les sorties n'étant pas sensible aux choix des priors, ils sont fixés par défaut pour la suite de l'analyse (Annexe 8). Par ailleurs, différentes mailles ont été testées afin de sélectionner celles assurant le meilleur compromis qualité d'inférence/temps de calcul (Annexe 9).



**Figure 5 :** Maille d'interpolation triangulaire et points de données dans le GdG (mai 2018).

---

### 2.5.6. Métriques de validation

La performance prédictive des modèles est évaluée pour chaque source de donnée  $j$  par analyse de la courbe ROC, l'AUC associée et la CPO.

Une courbe ROC (receiver operating curve characteristic) est un graphique représentant les performances d'un modèle de classification. Cette courbe trace le taux de vrais positifs (les présences observées prédites correctement) en fonction du taux de faux positifs (les absences observées faussement prédites par des présences). A partir de cette courbe on peut calculer l'AUC (« Area Under Curve » ou "aire sous la courbe ROC") qui correspond à l'intégrale sous la courbe ROC. L'AUC prend des valeurs entre 0 et 1. Un modèle qui prédit systématiquement l'inverse de ce qui est observé à une AUC de 0. Un modèle dont toutes les prédictions sont justes à une AUC de 1. Un modèle qui classifie les observations au hasard aura une AUC de 0.5.

La CPO (Conditional predictive ordinate) correspond à un test de validation croisée basé sur la densité *a posteriori* de la donnée  $y_i$  quand le modèle est ajusté sur toute les données à l'exception de  $y_i$  (Gómez-Rubio, 2020).

$$CPO_i = \pi(y_i | y_{-i}) \quad (\text{Eq. 10})$$

Cette densité prend une valeur comprise entre 0 et 1 selon qu'elle est mal ( $CPO_i = 0$ ) ou bien prédite ( $CPO_i = 1$ ).

Nous résumons cette métrique en calculant la Log-CPO pour chaque source de donnée (i.e. en ne considérant que les  $n_j$  observations issues d'une des sources de données). Plus la valeur de la LCPO est faible, plus la capacité prédictive du modèle est élevée pour cette source de donnée.

$$LCPO_j = - \sum_{i=1}^{n_j} \log(CPO_i) \quad (\text{Eq. 11})$$

### 2.5.7. Démarche de valorisation du modèle

Dans un premier temps, l'approche de modélisation intégrée a été appliquée à un seul pas de temps de la campagne PELGAS (mai 2018) afin de comparer différentes configurations de modèles et d'évaluer (1) l'apport des différentes sources de données dans l'inférence, (2) l'impact de l'échantillonnage préférentiel sur l'inférence et (3) l'apport des co-variables environnementales dans le modèle.

- (1) La contribution des différentes sources de données à l'inférence est étudiée en comparant les sorties du modèle intégrant toutes les sources de données aux modèles ajustés soit à la donnée scientifique seule, soit à la donnée commerciale seule. Pour cette partie, les trois flottilles commerciales sont intégrées dans un modèle « commercial ».
- (2) Pour évaluer l'effet de l'échantillonnage préférentiel sur l'inférence, nous comparons un modèle prenant en compte l'échantillonnage préférentiel avec un modèle ne prenant pas en compte l'échantillonnage préférentiel. Dans ces configurations, toutes les sources de données sont intégrées.

---

(3) L'ajout de covariables au modèle est réalisé en deux étapes. Dans un premier temps, nous ajustons le modèle avec les covariables sans prendre en compte l'effet aléatoire spatial. Dans un second temps, nous ajustons un modèle prenant en compte les covariables et l'effet aléatoire spatial. Cette approche en 2 étapes permet d'appréhender les problèmes d'identifiabilité de la relation espèce-habitat lorsque le modèle est potentiellement sur-paramétré.

La comparaison des différentes configurations de modèle sur un seul pas de temps a permis de choisir la configuration qui réalise le meilleur compromis entre qualité de l'inférence, temps de calcul et complexité du modèle pour l'ajuster sur l'ensemble des pas de temps des années 2009 à 2018.

### 3. Résultats

#### 3.1. Analyse à l'échelle d'un mois : mai 2018

Pour chaque sous partie nous présentons deux types de cartes - la probabilité de présence prédictive par le modèle et les écarts-types associés – ainsi que les performances prédictives de chaque modèle. Une comparaison de ces performances est réalisée dans une dernière partie. Un troisième type de carte, étudiant les différences absolues des probabilités de présence inférées, compare les configurations testées (ajout de l'échantillonnage préférentiel et des co-variables) avec le modèle intégré de référence. En annexe figurent : les courbes ROC, les cartes de capacités prédictives des CPO, les cartes d'inférences de 2009 à 2018 et les scores AUC associés et moyens sur la période (Annexes 14 à 36).

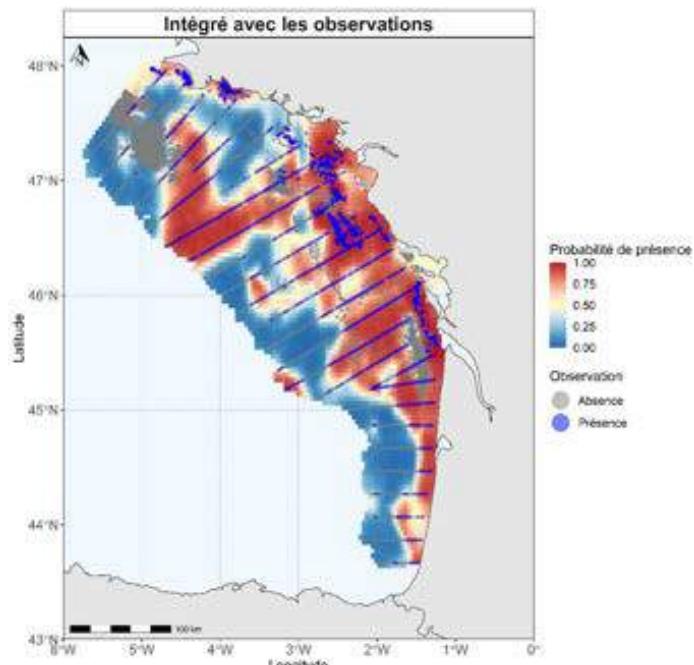
##### 3.1.1. Apport des différentes sources de données à l'inférence

Le modèle intégré produit des cartes de présence-absence en cohérence avec les différentes sources de données (Figure 6). Dans les zones où les observations sont positives, les probabilités de présence sont élevées. Dans les zones où les observations sont nulles, les probabilités de présence sont faibles. La bonne concordance entre les observations de présence-absence et les prédictions du modèle apporte une première validation qualitative de la performance du modèle.

Les cartes de prédictions ajustées aux différentes sources de données montrent que les données scientifiques apportent le plus d'information à l'inférence (Figure 7). En effet, la carte de distribution obtenue à partir du modèle intégré est presque superposable à celle obtenue à partir des données scientifiques seules. Comme les données scientifiques couvrent l'ensemble du domaine d'étude et bénéficient d'un échantillonnage dense, elles sont suffisantes pour décrire les patrons de distribution de la sardine sur l'ensemble du GdG. Les écarts-types sont globalement faibles sur l'ensemble de l'aire d'étude avec des valeurs légèrement plus élevées entre les transects du plan d'échantillonnage ainsi que dans les zones de transitions de présence/absence.

Les données commerciales sont beaucoup plus concentrées dans l'espace et essentiellement restreintes à une bande très côtière. Lorsqu'elles sont la seule source de donnée utilisée, elles ne permettent pas de capturer les bancs de sardine au large et les prédictions en dehors du rayon d'action des flottilles commerciales ne sont pas fiables. Les écarts types du modèle ajusté à la donnée commerciale seule sont très élevés au large et faibles dans le rayon d'action des flottilles.

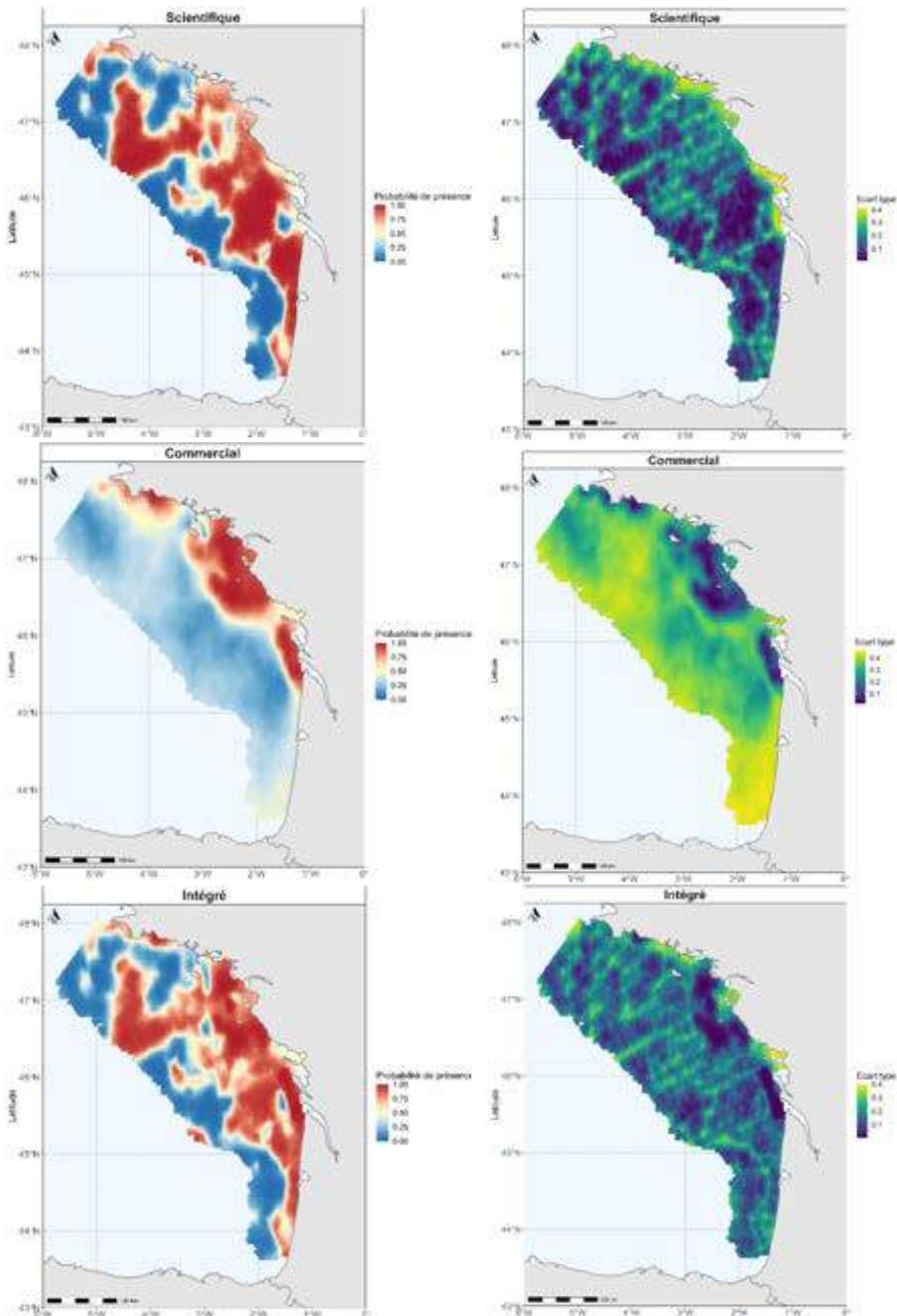
Dans le modèle intégré, les données commerciales apportent de l'information en complément des données scientifiques. Elles apportent de l'information sur les zones côtières, mais ne modifient que légèrement les patrons de distributions obtenus avec la donnée scientifique seule, en précisant certaines tâches de présences côtières mal décelées par les données scientifiques (e.g. au Nord de la carte).



**Figure 6 :** Cartographies des probabilités de présence et des observations par les flottilles commerciales et la campagne scientifique PELGAS en Mai 2018.

Le modèle ajusté sur une seule source de donnée a systématiquement d'excellentes capacités prédictives avec des AUC toujours supérieures à 0.90 (Figure 11). De même, les modèles intégrés ont aussi d'excellentes performances prédictives, à l'exception des prédictions des données PTM\_SPF pour lesquelles les scores sont autour de 0.8, ce qui demeure de très bons scores.

Concernant les LCPO, il apparaît ici que l'intégration des données commerciales dans le modèle en plus des données scientifiques améliore les prédictions de données commerciales au détriment des observations scientifiques (plus la LCPO est faible, meilleur est la capacité prédictive du modèle) (Figure 11).

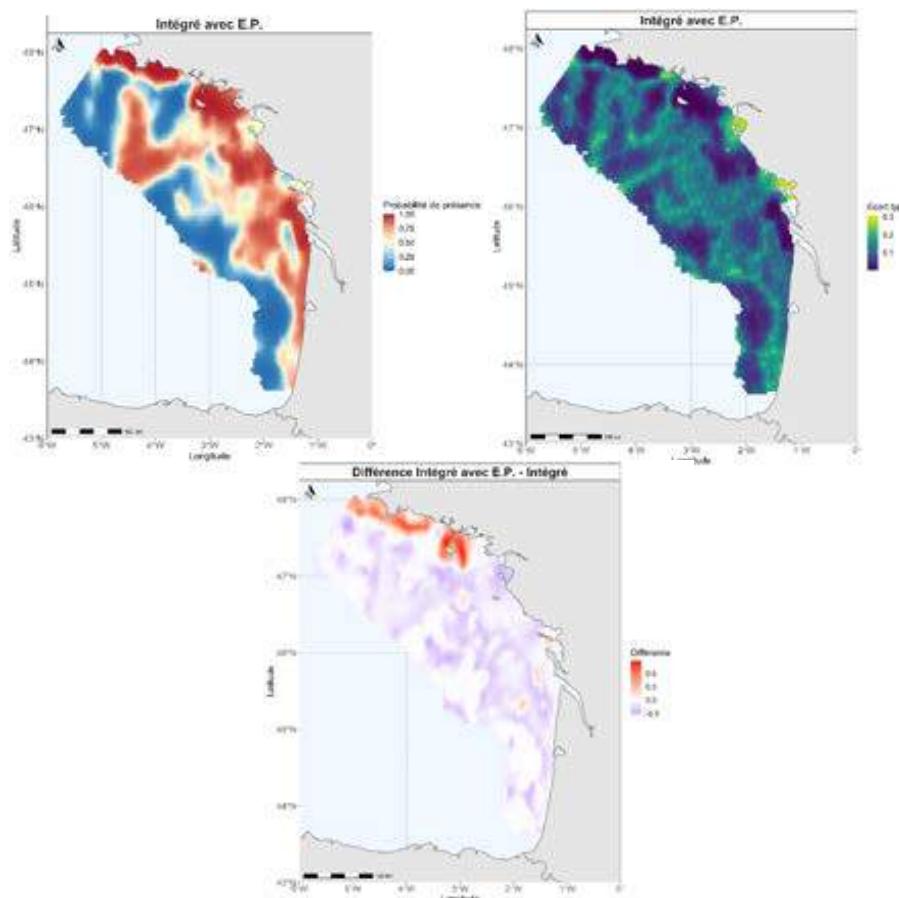


**Figure 7 :** Cartographies des probabilités de présence (colonne de gauche) et écarts-types associés (colonne de droite) au mois de Mai 2018 obtenues à partir des données scientifiques, des données commerciales et dans un modèle intégré les combinant.

### 3.1.2. Intensité de l'échantillonnage préférentiel

L'échantillonnage préférentiel semble avoir peu d'influence sur les inférences car les cartes d'inférences obtenues avec ou sans E.P. sont similaires. Cependant, le modèle avec E.P. rend mieux compte du fonctionnement des pêcheries.

Les résultats suggèrent des intensités de ciblage différentes en fonction des trois flottilles commerciales, avec un ciblage plus marqué pour les senneurs. Ainsi logiquement, la prise en compte d'un échantillonnage préférentiel modifie légèrement les inférences dans les zones côtières où sont présents les senneurs (au Sud de la Bretagne, de la Pointe du Raz à Belle-Île-en-Mer) (Figure 8). En effet, le paramètre  $b_j$  estimé (qui contrôle l'intensité de l'échantillonnage préférentiel) prend une valeur de 1.48 pour les senneurs (PS\_SPF), tandis qu'il est estimé à 0.37 et -0.48 respectivement pour les flottilles PTM\_SPF et PTM\_DEF. Cela suggère l'existence d'un échantillonnage préférentiel uniquement pour les flottilles



**Figure 8 :** Cartes d'inférence du modèle avec prise en compte d'un échantillonnage préférentiel dans le fonctionnement des flottilles commerciales, les écarts types associés et de la différence avec le modèle intégré ne prenant pas en compte d'E.P. (E.P – Intégré)

---

commerciales ciblant explicitement les petits pélagiques (senneurs et PTM\_SPF), et un échantillonnage plus fort pour les senneurs en comparaison aux chaluts en bœufs pélagiques. Ces résultats sont consistants avec le fonctionnement des pêcheries étudiées.

La prise en compte de l'E.P. permet de réduire les écarts-types dans le rayon d'action des flottilles. En effet, dans le cas où l'échantillonnage préférentiel est pris en compte dans l'inférence, la distribution des points de pêche contribue à l'inférence et ils apportent de l'information dans les zones échantillonnées, ce qui diminue l'écarts-types dans ces zones. De plus, la prise en compte de l'E.P. réduit les probabilités de présence sur l'ensemble du GdG, à l'exception des zones d'activités des senneurs, notamment dans les zones de transition entre les fortes et les faibles probabilités de présence (Fig. 8, carte du bas). Cette diminution est due à l'absence de points de pêche des bolincheurs qui concentrent leur activité en Bretagne Sud. Le modèle considère ainsi que les probabilités de présence sont plus faibles sur les autres secteurs.

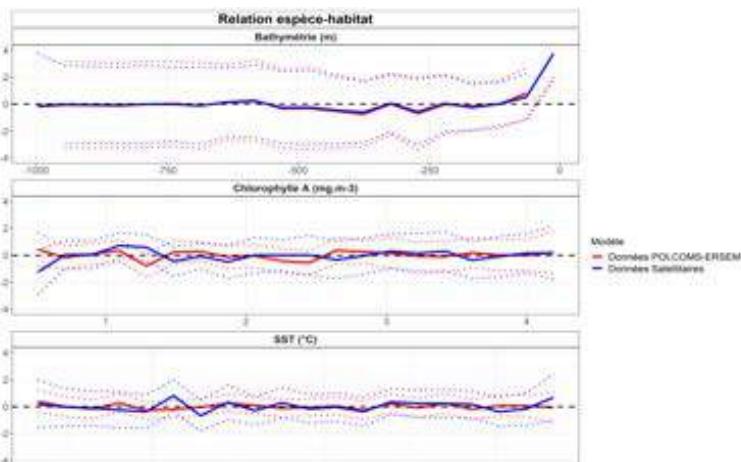
Notons que l'analyse avec seulement deux flottilles (les flottilles PTM\_DEF et PTM\_SPF étaient regroupés en une seule flottille PTM) ne laissait apercevoir aucunes modifications dans les inférences (les résultats ne figurent pas dans ce manuscrit). Les  $b_j$  estimés prenaient des valeurs proches de zéro et la force de l'échantillonnage préférentiel estimé était relativement faible. La distinction de PTM en 2 segments plus homogènes permet de mieux estimer l'intensité du ciblage. Pour autant les patrons de distributions de la sardine sont globalement similaires en fusionnant ou non les flottilles PTM\_DEF et PTM\_SPF.

En réduisant le domaine à la frange côtière (50 km de distance du littoral) les estimations des paramètres de ciblage sont modifiées ( $b_{PS\_SPF} = 1.05$  ;  $b_{PTM\_SPF} = 0.49$  ;  $b_{PTM\_DEF} = -0.17$ ). Cela suggère que d'autres facteurs, notamment la distance à la côte, entrent en compte dans la répartition des points de pêches  $x_{com}$ . En particulier on observe que  $b_{PS\_SPF}$  diminue tandis que les autres valeurs augmentent : tous les points de pêche des senneurs étant côtiers, la réduction de la zone diminue l'influence de cette répartition sur la force du ciblage. Au contraire, les autres flottilles ont des activités qui s'éloignent plus vers le large et pour lesquelles les observations indiquent des absences plus nombreuses. En restreignant la zone d'étude on augmente l'importance de l'échantillonnage préférentiel.

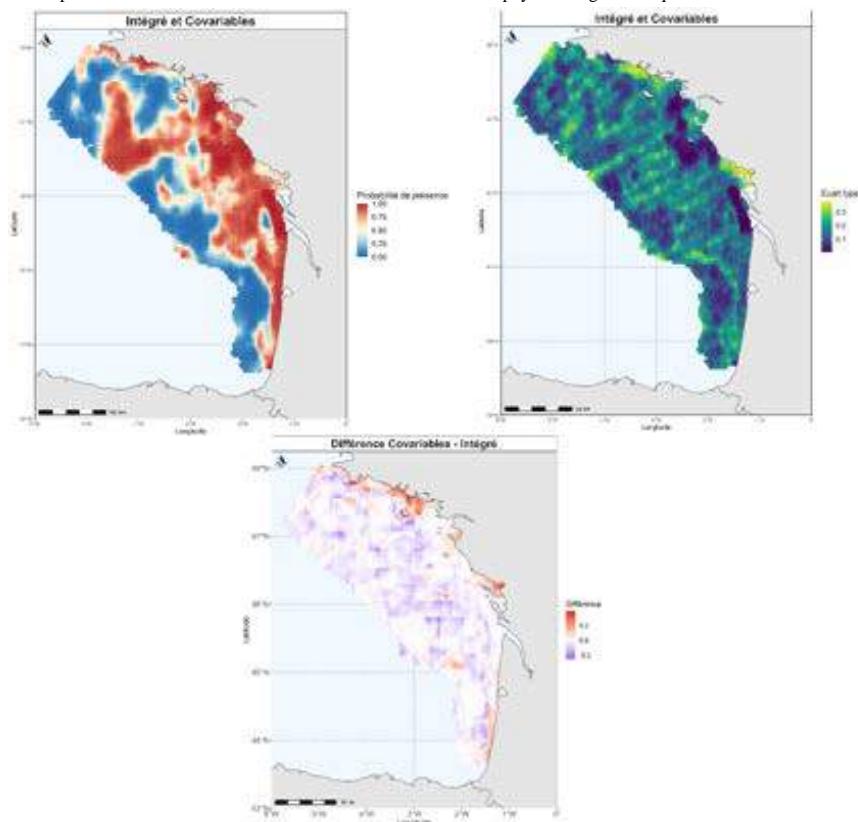
Cependant, malgré l'intérêt de la prise en compte de l'E.P. pour une meilleure représentation de la distribution spatiale des captures commerciales, cette dernière entraîne une diminution de l'AUC et une augmentation des LCPO en comparaison au modèle intégré et donc à de moins bonnes capacités prédictives des observations de présence/absence (Figure 11). Ainsi, en considérant ces résultats et les temps de calculs nécessaires à l'ajustement des modèles prenant en compte l'échantillonnage préférentiel, nous choisissons de ne pas le mettre en place dans le modèle spatio-temporel.

### 3.1.3. Ajout des co-variables environnementales.

L'ajout de covariables environnementales dans le modèle a peu d'effet sur les prédictions du modèle et les résultats obtenus sont peu sensibles à la source des données considérée, qu'elles soient satellitaires ou issues du modèle physico-biogéochimique (Figure 9). Les relations espèces-habitat permettent de mettre en évidence un effet bathymétrie pour les faibles profondeurs et des effets SST et chlorophylle A relativement faibles. Les cartes d'inférences sont présentées en Annexe 11. Les relations espèces-habitat sont similaires pour toutes les covariables même en retirant l'effet aléatoire (Annexe 12). Les cartographies obtenues par le modèle intégré avec ou sans effet des covariables sont similaires (Figure 10).



**Figure 9 :** Effet des co-variables environnementales (Bathymétrie, SST et Chlorophylle A) dans un modèle prenant en compte des données satellitaires ou issues des sorties du modèle physico-biogéochimique POLCOM-ERSEM.



**Figure 10 :** Cartes d'inférences en prenant en compte les co-variables environnementales (Bathymétrie, SST et Chlorophylle A issue des données satellitaires), les écarts types associés et de la différence avec le modèle intégré "simple". (Covariables – Intégré).

La principale différence provient d'une frange littorale de forte probabilité, issue de la prise en compte de la bathymétrie. Finalement, l'ajout de covariables environnementales dans le modèle n'améliore pas les capacités prédictives du modèle (Figure 11).

Les difficultés d'identifiabilité du modèle, le faible apport des covariables dans l'inférence et le risque d'une confusion d'effet avec les facteurs affectant la distribution des points pêche nous ont conduit à laisser de côté les covariables lors de l'ajustement du modèle spatio-temporel. L'effet des covariables sur les prédictions du modèle spatio-temporel sont étudiées à posteriori de l'ajustement du modèle spatio-temporel. Les résultats sont présentés en annexe (13).

#### 3.1.4. Comparaison des capacités prédictives selon les différents modèles.

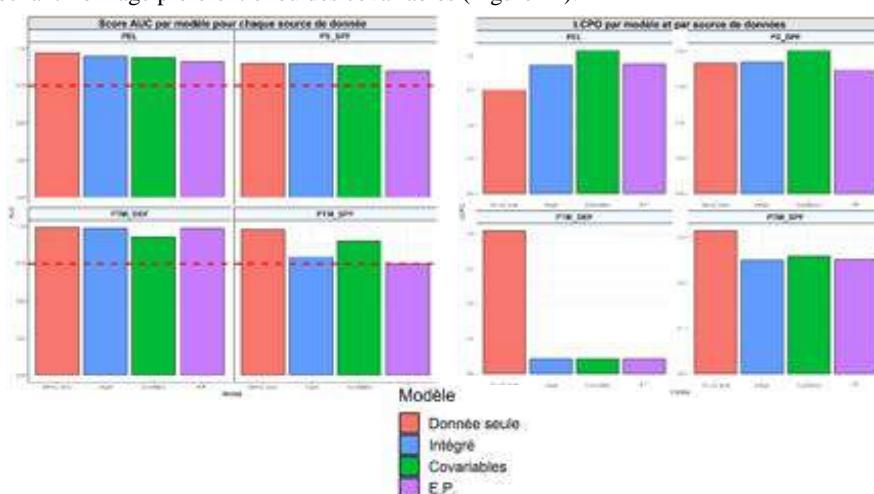
Les modèles prenant en compte un seul jeu de données présentent le plus souvent les meilleures capacités prédictives (Figure 11). Ce résultat est attendu. En effet, le modèle intégré fait une synthèse de l'ensemble des données, ce qui explique le fait qu'il soit moins bien ajusté à chaque jeu de données pris en compte séparément.

La comparaison des AUC entre les différents modèles laisse apparaître qu'elles ont tendance à diminuer avec une augmentation de la complexité du modèle selon l'ordre suivant :

$$AUC(\text{Donnée seule}) > AUC(\text{Intégré}) > AUC(\text{Covariables}) > AUC(\text{E.P.})$$

Néanmoins, dans tous les cas l'AUC reste supérieure à 0.75, ce qui correspond à de bonnes capacités prédictives. Notons un changement dans cet ordre pour la prédiction des données issues des chaluts en bœufs pélagiques (PTM\_SPF et PTM\_DEF) lorsque les covariables environnementales sont prises en compte.

L'analyse des LCPO est plus nuancée : pour les données scientifiques et PS\_SPF, la qualité prédictive diminue avec la complexité du modèle, de la même manière que pour les AUC. En revanche pour les flottilles PTM\_SPF et PTM\_DEF, la performance prédictive est meilleure avec un modèle intégré et les différences sont très faibles en prenant en compte un échantillonnage préférentiel ou des covariables (Figure 11).



**Figure 11 :** Capacités prédictives des différents modèles par source de données, AUC à gauche et CPO à droite.

---

Au global, la prise en compte de plusieurs sources de données dans un modèle ne modifie que légèrement sa capacité prédictive, qui demeure néanmoins bonne. Dans le détail, cette réduction n'est pas systématique et varie selon la source de données (e.g. LCPO de PELGAS). La complexification du modèle via la mise en place de l'échantillonnage préférentiel ou de covariables accentue le plus souvent ces différences tout en réduisant globalement les performances prédictives du modèle.

### 3.2. Analyse spatio-temporelle

Le modèle spatio-temporel (Eq 7 et 8) est ajusté sur un modèle intégré « simple » combinant différentes sources de données sur une année à un pas de temps mensuel (campagnes PELGAS en mai, JUVENA en septembre, et trois flottilles commerciales disposant de données chaque mois), mais sans processus d'échantillonnage préférentiel et sans covariables.

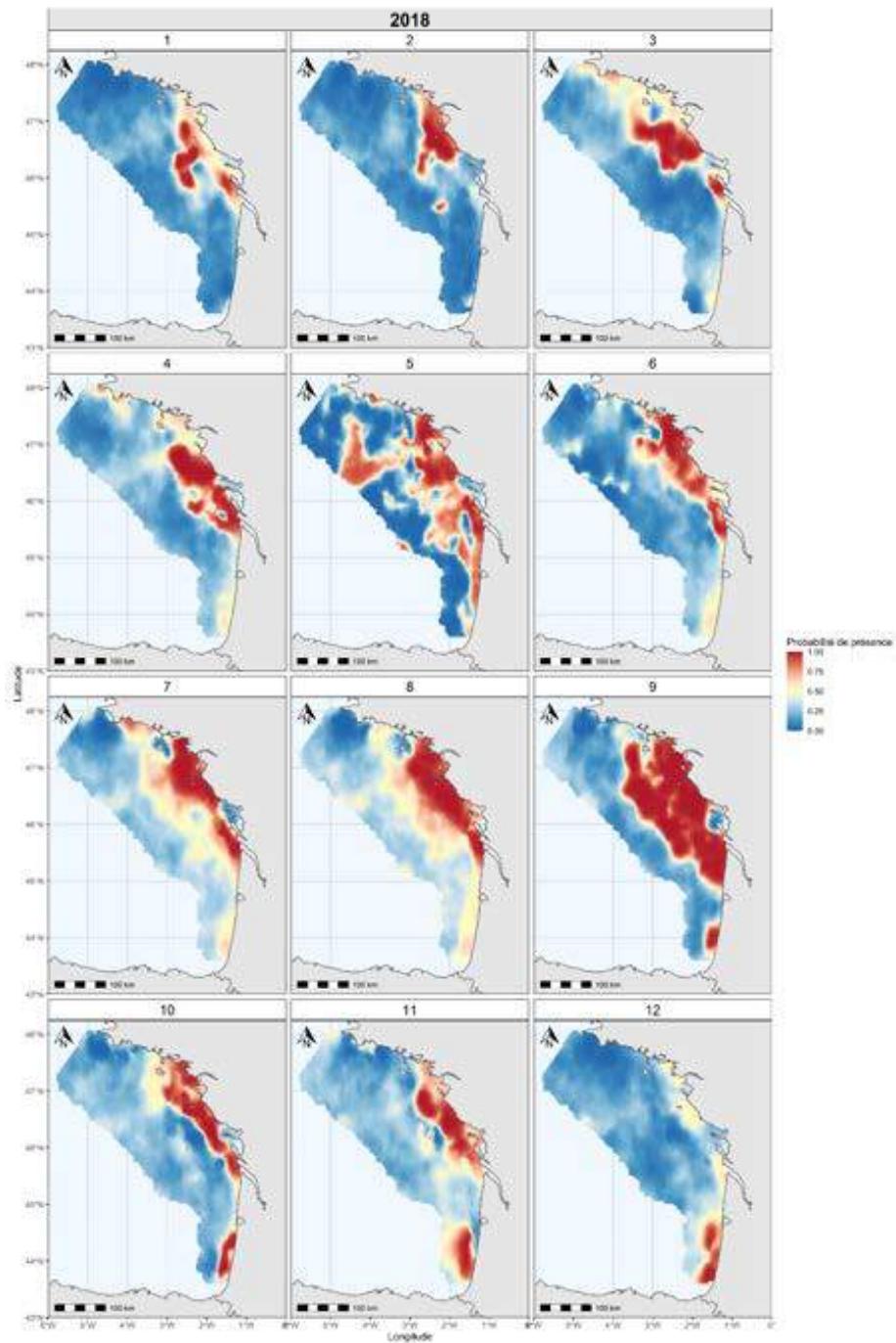
Les cartographies inférées indiquent des présences de sardine principalement dans les zones côtières (Figure 12, pour l'année 2018 uniquement). Les zones de présences élevées sont situées entre la Baie de Vilaine et l'Estuaire de la Gironde ainsi qu'au niveau du Bassin de Parentis au Sud. Ces patrons de présences s'étendent sur le plateau continental en Mai et Septembre, qui correspondent aux périodes durant lesquelles sont réalisées les échantillonnages par les campagnes scientifiques. En hiver, les zones de présences de sardines sont très réduites dans l'espace du fait de la faible quantité de données commerciales disponibles sur cette période.

Pour les mois où seules les données commerciales sont disponibles (tous les mois sauf mai et septembre), les zones échantillonnes par les flottilles commerciales sont restreintes à la bande côtière et ne permettent pas d'obtenir des inférences fiables sur l'ensemble du GdG. La prédiction de la présence/absence au-delà de la frange côtière fréquentée par les pêcheurs ciblant ces espèces n'est pas interprétable. L'analyse des cartographies des écarts types associés révèlent des zones très vastes où les écarts-types atteignant des valeurs élevées (autour de 0.40 - Figure 13).

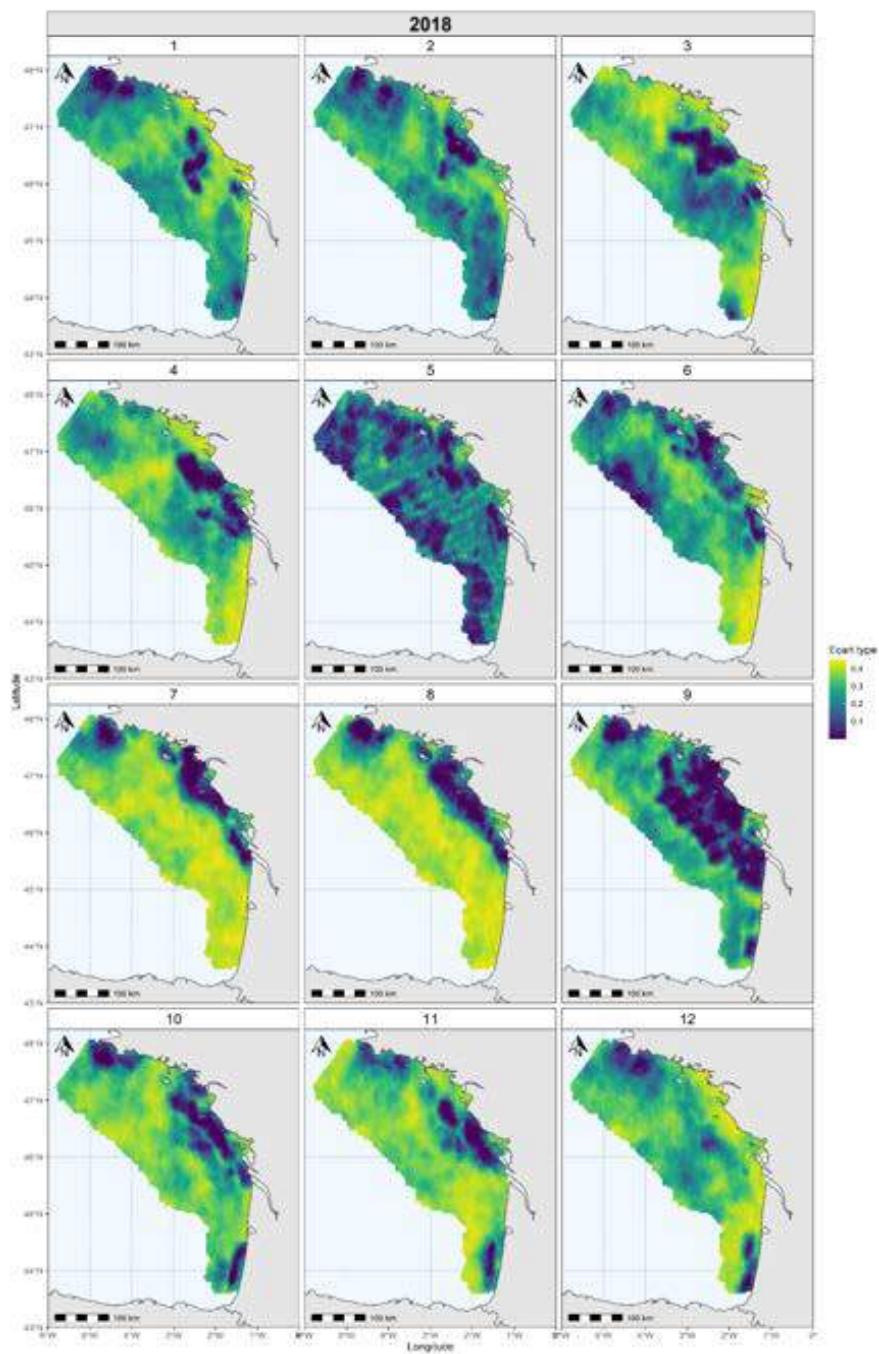
Concernant la capacité prédictive du modèle spatio-temporel, les AUC sont très élevées (presque systématiquement supérieures à 0.90, des valeurs minimales à 0.75) (Figure 14). Notons que pour certains mois aucune AUC n'a pu être calculée, les observations étant uniquement constituées soit de présences soit d'absences.

Cette approche est étendue à la période 2009-2018 (chaque année est traitée indépendamment des autres ; cartes en Annexes 19 à 36). Les analyses sont semblables et les observations commerciales sont localisées, la saisonnalité de l'activité de pêche est plus marquée sur le début de la série chronologique. Des scores AUC élevés sont également obtenus pour chaque année et source de données avec, dans 95% des cas, des valeurs d'AUC supérieures à 0.70. (Figure 15 et Annexes 17 et 18)

Cependant, il est intéressant de noter que la moitié des années étudiées (2010, 2012, 2013, 2014, 2016) n'identifient aucun patron de distribution de sardine au large malgré un échantillonnage dans cette zone (Figure 16 pour exemple avec l'année 2016). De plus, pour ces années, les patrons sont similaires à ceux décrits par la donnée commerciale seule sur la période printemps-automne (Figure 16). Ainsi, dans certains cas, la donnée commerciale semble permettre de décrire une partie de la distribution de la sardine malgré sa restriction à des zones côtières et identifier une stabilité des patrons de présence.



**Figure 12 :** Cartographie mensuelle de la probabilité de présence de la sardine dans le GdG en 2018



**Figure 13 :** Cartographie mensuelle de l'écart-type de la probabilité de présence de la sardine dans le GdG en 2018.

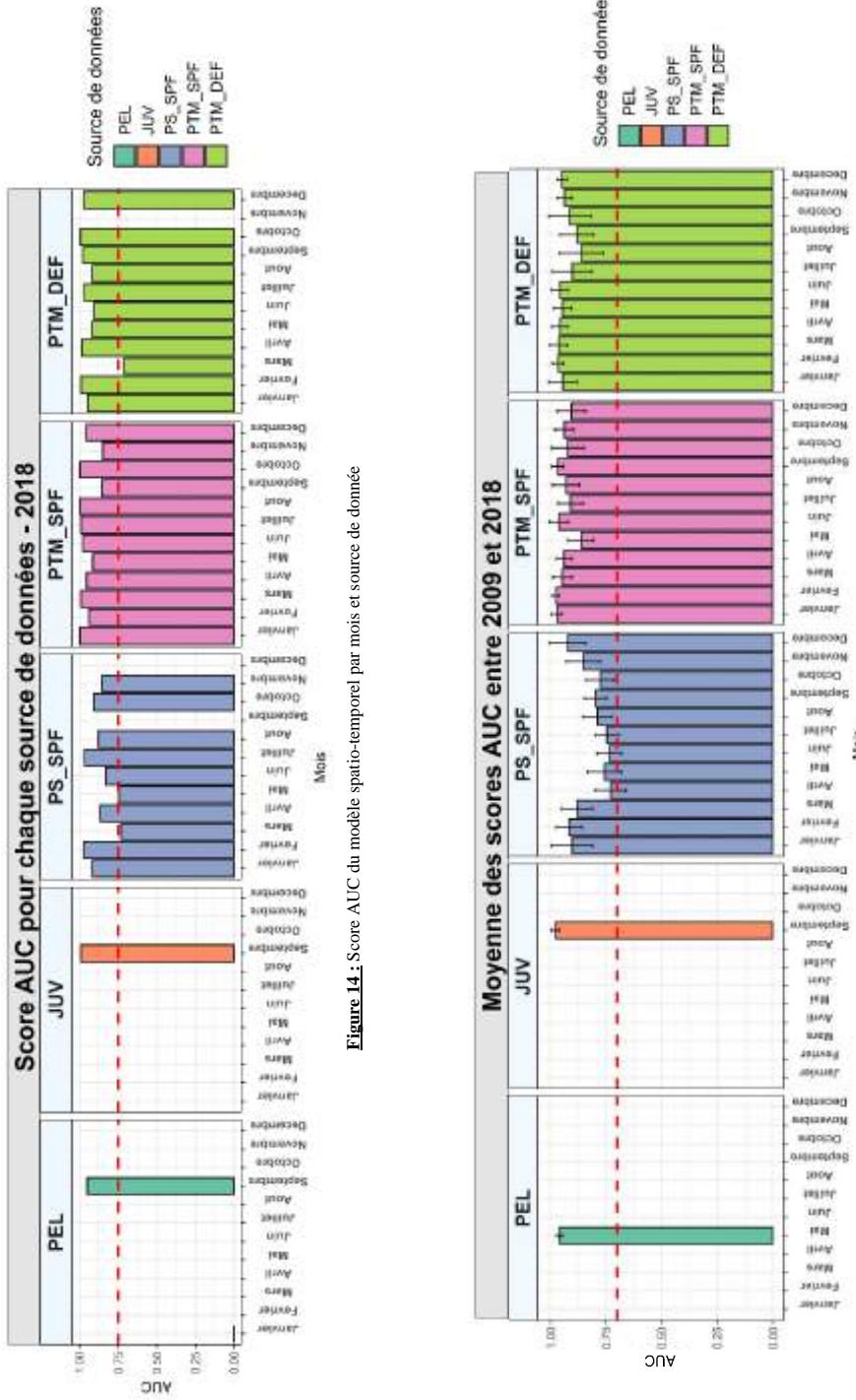
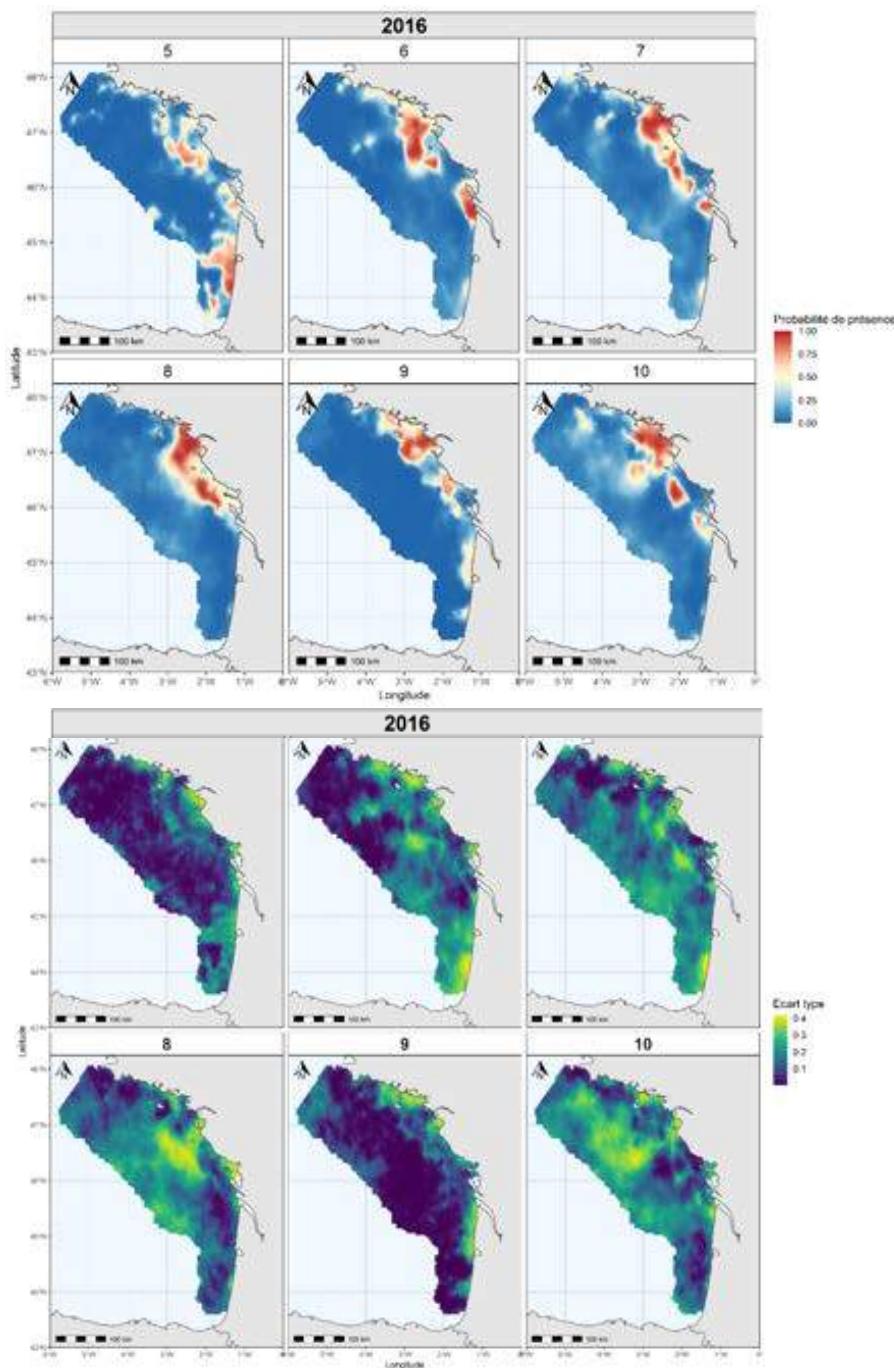


Figure 14 : Score AUC du modèle spatio-temporel par mois et source de donnée

Figure 15 : AUC moyené par mois et source de donnée (2009-2018) et écarts types. Le trait rouge correspond à une AUC de 0.70.



**Figure 16 :** Cartographies de la probabilité de présence de la sardine et d'écart-types associés dans le GdG entre mai et octobre 2016

---

## 4. Discussion

Dans ce travail, nous avons développé un cadre de modélisation permettant de combiner des données issues de campagnes scientifiques et commerciales pour inférer la distribution de la sardine (carte de probabilité de présence) dans le Golfe de Gascogne. Le modèle a été construit en adaptant une approche de modélisation initialement développée pour des espèces démersales (Alglave *et al.*, under review). L'analyse a été conduite d'abord sur un seul pas de temps, puis à l'échelle d'une année entière dans un second temps en ajoutant une composante temporelle au modèle. Les résultats permettent de mettre en évidence des différences notables avec les cas d'étude benthico-démersaux. Ils révèlent l'intérêt de combiner différentes sources de données pour inférer la distribution de la sardine sur l'ensemble de l'année, mais aussi les limites de l'approche liées en partie au fonctionnement des pêcheries de petits pélagiques.

### 4.1. Une inférence à l'échelle du mois dirigée par la donnée scientifique

Les analyses menées sur les mois où la donnée scientifique est disponible (i.e. les mois de mai et septembre) montrent que pour ces deux mois, les inférences sur la distribution de la sardine sont dominées par les données de campagnes scientifiques. L'intégration des données commerciales permet de compléter l'information fournie par les données scientifiques dans des zones côtières très localisées.

Ces résultats sont différents de ceux obtenus sur des espèces démersales par Alglave *et al* (under review) qui indiquent au contraire que les données commerciales dominent les inférences. Ces différences sont liées à la taille relative des deux sources de données ainsi qu'à la couverture spatiale des données commerciales.

La première différence essentielle concerne la taille des jeux de données scientifique et commercial. Dans le cas de la sardine, les données scientifiques sont issues d'un échantillonnage acoustique (PELGAS et JUVENA), alors que les données scientifiques utilisées par Alglave *et al* (under review) sont issues de données des campagnes EVHOE obtenues par traits de chaluts. Ainsi le nombre d'observations est très différent : les campagnes acoustiques fournissent plus de 1500 observations (1800 en 2018 pour PELGAS et 1600 pour JUVENA) pour un mois, contre 150 observations pour les campagnes EVHOE. En effet, la majorité des plans d'échantillonnages des campagnes scientifiques sont mis en œuvre pour fournir des indices d'abondance sans biais ou des données biologiques pour l'évaluation des stocks, plutôt que pour décrire précisément la distribution des espèces d'intérêt halieutique. Dans le cas des poissons pélagiques comme la sardine, les campagnes acoustiques permettent de couvrir de larges aires d'études avec un échantillonnage dense sur une courte période, fournissant des estimations de biomasses tout en renseignant précisément sur la répartition de la spatiale (Georgakarakos *et al.*, 2011).

Une seconde différence est liée à la répartition spatiale des données commerciales. Dans le cas des pêcheries démersales traitées dans Alglave *et al.*, les pêcheries commerciales couvrent un large domaine spatial comparable à la couverture spatiale des données de campagne. Dans le cas de la sardine, les coups de pêche se concentrent exclusivement dans la zone côtière (et sur des zones déjà bien couvertes par les données scientifiques). Ainsi, l'essentiel de la zone de distribution de la sardine n'est pas couvert par la pêcherie commerciale, alors qu'elle est bien couverte par les données scientifiques.

---

#### 4.2. Capacité prédictive du modèle

Les cartographies de la répartition spatiale de la sardine semblent précises dans les zones où la donnée est disponible. Les écarts-types associés aux probabilités de présences sont relativement faibles sur les secteurs où la donnée est disponible. De plus, les scores AUC sont élevés (AUC > 0.70) pour toutes les sources de ce qui laisse penser que le modèle produit des prédictions fiables (Hosmer *et al.*, 2013).

Cependant l'utilisation de l'AUC pour justifier de la qualité des sorties du modèle est discutable puisque l'AUC est calculé à partir de la donnée ajustée au modèle. Par conséquent, l'AUC peut conduire à surestimer les performances du modèle (Jiménez-Valverde *et al.*, 2008; Lobo *et al.*, 2008). En particulier, Lobo *et al.* (2008) ont remis en question la fiabilité de l'AUC comme mesure unique de la performance du modèle.

#### 4.3. Mise en perspective des cartes de distribution avec les connaissances déjà disponibles

La comparaison des patrons de distribution avec ceux obtenus par les différentes études révèlent une certaine cohérence (Annexe 37 à 40). En étudiant la distribution spatiale des zones de fraies de la sardine dans le GdG, Bellier *et al.*, (2007) identifient 3 sites récurrents : (1) du sud-ouest de la Bretagne, au nord de l'estuaire de la Gironde (2) au sud de l'estuaire de la Gironde dans le Bassin de Parentis et (3) sur le rebord du plateau continental entre 46°N - 47°30'N et -6°W - 4°50'W. Ces sites correspondent à des zones de plus fortes probabilités de présences d'œufs de sardines et de faibles variances. Les 2 premiers sites, les plus côtiers, sont également observés et décrits par Doray *et al.* (2018a) et Huret *et al.* (2018) en termes de biomasses de sardines et d'abondances d'œufs. De même, Petitgas *et al.* (2020) complètent et confirment ces analyses, d'une part en identifiant les mêmes zones de fortes abondances d'œufs et en suggérant un changement dans leur distribution, se rapprochant des zones côtières à partir de 2009. Au cours de cette étude nous avons pu identifier des zones semblables de fortes probabilités de présences : Pointe du Raz au Plateau de Rochebonne (du Sud-Ouest de la Bretagne au Nord de l'estuaire de Gironde (1)) et au niveau du Bassin de Parentis (Au nord de l'estuaire de la Gironde (2)) au mois de mai, mais également dans une certaine mesure durant la période printemps-automne pour chaque année étudiée.

Le troisième site identifié par Bellier *et al.* (2007) correspond également à une zone de fortes probabilités de présences, cette dernière n'est cependant observable qu'à partir des campagnes scientifiques, et uniquement certaines années (2009, 2015 et 2018). Ces résultats sont donc aussi en cohérence avec les conclusions de Petitgas *et al.* (2020) qui rapporte que les distributions de la sardine se rapprochent de ces zones côtières.

Pour autant, il est important de rappeler que ces différentes études portent sur les œufs de sardines (en abondance ou présence) ou sur la biomasse de sardines et pas sur la probabilité de présence.

---

#### 4.4. Limites de l'étude

##### 4.4.1. Spécificités de l'activité de pêche à la sardine dans le Golfe de Gascogne

L'exploitation de la sardine par les flottilles commerciales se caractérise par un lien très fort entre l'activité de pêche et les demandes du marché, avec pour conséquence une forte saisonnalité ainsi qu'une activité de pêche très localisée dans l'espace et côtière. A cela s'ajoute l'hyperstabilité des CPUE qui rend impossible le calcul d'indices d'abondances ou de biomasses à partir des captures commerciales. Ainsi, les captures de sardines par les professionnels de la pêche renseignent sur la **présence** du poisson **commercialisable** et sur des **zones restreintes géographiquement et temporellement**. Elles ne donnent aucune indications sur le reste de la population en dehors de leur période de pêche. Ces limites de la donnée commerciale ont des implications fortes sur les prédictions du modèle intégré et sur les interprétations qu'on peut en tirer.

La concentration de l'activité de pêche dans des zones côtières est une conséquence des demandes du marché. En effet, les sardines doivent être vendues fraîche le jour même pour être transformées dans la journée, ainsi les conserveries fonctionnent souvent par contrats ou de gré à gré, et commandent un tonnage précis dans un délai court. Les pêcheurs sont donc incités à ne capturer que la commande passée par les conserveries dans un délai très court garantissant une réponse rapide à la demande en maximisant la fraîcheur (et la qualité) du poisson. Aussi, quand ils ciblent la sardine, ils ont intérêts à privilégier les zones poissonneuses les moins éloignées, même si les abondances sont élevées au large. Pour cette raison, les données d'observations de sardines, en particulier, celles des senneurs (PS\_SPF), se concentrent dans des zones très précises peu éloignées des zones de débarquements et de traitements dans les conserveries :

- Bolincheurs : Nord de la Bretagne et Bassin de Parentis – Douarnenez ; Lorient ; Quiberon ; Saint-Jean-de-Luz –
- Chaluts en bœufs pélagiques : Baie de Villaine jusqu'au Plateau de Rochebonne et Bassin de Parentis – La Turballe ; Saint-Gilles-Croix-de-Vie ; La Cotinière ; Saint-Jean-de-Luz –

Par ailleurs, les inférences sont également restreintes par la saisonnalité de l'activité de pêche, marquée par une faible activité commerciale pendant les mois d'hiver. Cela est à mettre en lien avec les exigences des conserveries sur la condition de la sardine. Afin de garantir les qualités organoleptiques du produit transformé, les conserveries exigent chez la sardine un taux de matière grasse supérieur à 8-10%. Or ce taux varie fortement au cours de l'année, avec des minimums observés en mars-avril, qui coïncident avec la période de fraie et à partir de laquelle ce taux augmente jusqu'à atteindre des valeurs maximales en Septembre-Octobre (Bandarra *et al.*, 1997). En exigeant un taux de matière grasse élevé, les conserveries entretiennent un effet filière sur l'exploitation de la sardine qui va directement influer sur la période d'exploitation de la sardine. La saison d'activité commence ainsi vers mai, lorsque le taux de graisse de la sardine atteint un seuil suffisamment élevé pour les conservereurs, et s'achève au début de l'hiver (Doray, *com. pers.*). C'est cet effet filière qui explique la faible quantité de données commerciales en hiver.

Enfin, il est intéressant de relever que les observations issues des données commerciales renseignent sur le poisson commercialisable. Au-delà de la restriction de la réglementation sur

---

la taille minimale, les exigences du marché portent également sur la taille de la sardine et sa bonne condition. En conséquence, certains navires préfèrent parfois ne garder que les individus les plus gros et rejettent ainsi ceux de petites tailles (Cornou *et al.*, 2021). En effet, on observe des rejets importants dans l'activité des chaluts pélagiques : la sardine y représente 80% des rejets (soit 20% des captures totales) avec près de 50% d'individus rejetés alors qu'ils ont une taille au-delà de la limite réglementaire de 11 cm. Ainsi, chez les chaluts pélagiques, 90 % des sardines d'une taille inférieure à 13.5 cm sont rejetées.

#### 4.4.2. Limites de la modélisation de l'échantillonnage préférentiel

Le modèle détecte un échantillonnage préférentiel pour 2 des 3 flottilles, ce qui est cohérent avec le fonctionnement des pêcheries d'espèces pélagiques qui les ciblent particulièrement (Vermard *et al.*, 2008). En effet, les sardines du Golfe de Gascogne représentent respectivement 90% et 35% des captures et des débarquements des chaluts pélagiques et des senneurs en 2019, avec des probabilités d'occurrence de 90% et 50% (Guerineau *et al.*, 2010; Cornou *et al.*, 2021).

Toutefois, le fait de le prendre en compte dans l'inférence n'a que peu d'influence sur les patrons de distributions de présence de la sardine, ce qui est consistant avec les résultats d'Alglave et al (under review). La prise en compte de l'échantillonnage préférentiel influence légèrement les inférences sur la distribution de la sardine en révélant des zones de forte probabilité de présence dans les aires où l'activité de pêche est élevée.

Cependant, l'intégration de l'échantillonnage préférentiel tel que modélisé dans ce travail reste très discutable et n'a finalement que peu d'intérêt dans l'objectif de valoriser la donnée commerciale pour inférer la distribution de la sardine pendant les mois où aucune donnée scientifique n'est disponible. En effet, la paramétrisation de l'EP proposé dans ce travail suppose directement que la fréquentation spatiale d'une zone par les pêcheurs dépend de la probabilité de présence de la sardine dans cette zone. Ainsi, l'EP interprète une zone non fréquentée comme une zone de faible probabilité de présence. Ce résultat apparaît dans les cartographies issues de la modélisation avec échantillonnage préférentiel pendant le mois de mai, la probabilité de présence est légèrement réduite dans les zones avec une probabilité de présence moyenne (aux alentours de 0.5) sur l'ensemble du Golfe de Gascogne. En effet, les aires où aucun pêcheur ne se rend, vont être considérée comme moins poissonneuses, ce qui affecte les inférences. Or, dans le cas de cette pêcherie, les contraintes de fonctionnement de la filière précédemment citées sont telles que les zones éloignées de principaux points de débarquement ne sont pas fréquentées non pas en raison de l'absence de sardine, mais pour d'autres raisons liées à la nécessité de fournir un produit frais dans des délais brefs. Ainsi, pour les mois sans données scientifiques, la prise en compte de l'EP tel que paramétré dans ce travail conduirait à la conclusion (fausse) que les sardines sont uniquement cantonnées aux zones de pêche.

Une perspective claire pour ce travail serait de développer la modélisation de l'échantillonnage préférentiel de façon à prendre en compte l'ensemble des facteurs qui expliquent la répartition des coups de pêche, comme la distance à la côte et/ou aux criées et conserveries par exemple.

#### 4.4.3. Estimation de l'effet des covariables environnementales

L'ajout des covariables dans le modèle n'a pas permis de mettre en évidence d'effets importants des covariables sur les probabilités de présences de sardine dans le GdG. Ces résultats semblent contradictoires avec les connaissances disponibles sur les petits pélagiques. En effet, leur

---

distribution spatiale est supposée être très dépendante des conditions environnementales (Schickele *et al.*, 2020). Cependant, Planque *et al.* (2007) observent que l'influence des facteurs hydrographiques sur la distribution de la sardine en mai semble moins prononcée que pour d'autres espèces pélagiques, notamment l'anchois. Ils remarquent ainsi que tous les facteurs hydrographiques semblent avoir un degré d'influence similaire sur la distribution de la sardine, ce qui suggère une plus grande tolérance aux conditions environnementales. Ce résultat est cohérent avec le comportement de la sardine qui est connue pour nager sur de grandes distances (Parrish *et al.*, 1981; Doston and Griffith, 1996), avoir une distribution spatiale plus fragmentée (Barange and Hampton, 1997) et être généralement plus flexible sur le plan environnemental que d'autres petits pélagiques comme l'anchois (Bakun and Broad, 2003). Ces caractéristiques peuvent ainsi mener à de grandes variations dans sa distribution spatiale d'une année à l'autre, comme observées par Doray *et al.* (2018a) qui rapportent que des sardines ont été trouvées dans la plupart des zones du Golfe de Gascogne sans que sa distribution ne soit restreinte à un habitat spécifique. De même l'ICES (2010) décrit que *S. pilchardus* présente une grande extension spatiale dans les eaux de l'Atlantique Nord-Est, où les adultes effectuent des migrations à grande échelle englobant une large gamme de conditions hydrologiques. Ces résultats suggèrent que les distributions de *S. pilchardus* pourraient être moins influencées par les conditions hydrologiques locales du GdG, et plus probablement déterminées par des processus à plus grande échelle (Stratoudakis *et al.*, 2004). Aussi, il est probable que les conditions environnementales ne soient pas assez contrastées sur le domaine d'étude pour exercer un contrôle clair et détectable sur la distribution de la sardine.

De même, les gammes de températures de surface décrites comme optimales par d'Arbault et Lacroix (1977) et Planque et al (2007) correspondent aux valeurs de SST testées dans notre cas d'étude (12-15°C), ce qui peut expliquer les faibles effets de la SST, la gamme de températures testées étant trop restreintes.

Enfin il est important de rappeler que le modèle se base sur des données de présence-absence (des '0' et des '1') qui sont moins informatives que des données d'abondances ou de biomasses, ce qui pourrait également expliquer le faible effet des covariables.

#### 4.4.4. Biais liés aux données.

Une première limite concerne les données commerciales et la réallocation des déclarations de pêche aux pings VMS. Pour obtenir la donnée de déclaration de pêche à la résolution des pings VMS, les données de pêche sont réallouées de façon uniformes sur les pings VMS correspondants. Dans le cas d'un modèle de présence/absence, dès qu'une déclaration de pêche est positive, l'ensemble des pings associés seront considérés comme des présences ce qui pourrait conduire à la surestimation de la probabilité de présence.

Une seconde limite concerne la campagne d'échantillonnage JUVENA. D'une part, le jeu de données obtenu ne permettait pas de distinguer les classes de tailles et donc de filtrer les individus d'une taille inférieure à la réglementation en vigueur (11 cm). D'autres part, son échantillonnage adaptif n'est pas identique dans le temps et ne couvre pas exactement l'aire d'étude ici analysée (correspondant à l'aire d'étude de PELGAS, qui est identique chaque année).

---

## 5. Conclusion

L'approche développée a permis d'étudier l'application d'un modèle intégrée combinant données de campagnes acoustiques et données 'VMS x logbooks' pour inférer des cartes de probabilités de présence/absence de la sardine du Golfe de Gascogne.

Pour les mois de mai (PELGAS) et septembre (JUVENA) pour lesquels des campagnes scientifiques acoustiques sont disponibles, les cartes reflètent essentiellement l'information portée par les données scientifiques. Elles permettent d'identifier des patrons de distribution cohérents avec la littérature. Sur les pas de temps non couverts par les données scientifiques, les données commerciales apportent de l'information dans le rayon d'action des flottilles restreint à une bande côtière. Mais une grande partie de la zone d'étude n'est pas recouverte par les flottilles commerciales et les cartes obtenues fournissent une image partielle de la distribution de la sardine.

Pour autant, l'approche a permis de produire des cartographies sur un pas de temps mensuel entre 2009 et 2018. Cela représente un apport significatif par rapport aux dernières cartes qui recouvrent l'ensemble de l'année et qui datent des années 70. De plus, les patrons de distributions côtiers semblent présenter des similitudes entre les mois de campagnes (mai et septembre) et sur la période printemps-automne ce qui laisse penser que les zones de fortes probabilités de présences de la sardine identifiées par les études précédentes (uniquement sur le mois de mai) perdurent sur les mois d'été.

L'approche de modélisation intégrée entreprise dans ce travail présente un certain nombre d'intérêt, mais est aussi marquée par de nombreuses limites qui sont résumées dans le tableau 1.

**Tableau 1 : Intérêts et limites de l'approche intégrée**

	<b>Intérêts</b>	<b>Limites</b>
<b>Données commerciales 'VMS x logbooks'</b>	Données disponibles en dehors des périodes de campagne	Activités commerciales restreinte dans l'espace (concentrée dans des zones côtières) et dans le temps (peu d'activité en hiver)
		Captures règlementées et structurée par la demande
		Réallocations des débarquements aux données VMS par carré statistique
		Rejets des petits individus
<b>Données scientifiques</b>		Hyperstabilité des captures
	Observations scientifiques nombreuses, échantillonnage dense et recouvrant l'ensemble du domaine d'étude	Uniquement pendant les mois de mai (PELGAS) et septembre (JUVENA)
<b>Qualité d'ajustement du modèle</b>	AUC élevée	Modèle de présence-absence

	Bon comportement en validation croisée	Pas de possibilité de dériver des indices d'abondance continus (pas de CPUE)
<b>Estimations</b>	Comportement des flottilles consistants avec la connaissance disponible pour les seneurs et les chaluts ciblant ( $b>0$ – ciblage élevé)	Faible effet des covariables environnementales sur les inférences, pas de relation espèce-habitat
	Distributions inférées cohérentes avec la littérature (en Mai).	Cartes de distribution biaisées en dehors des périodes de campagne.

Il est alors intéressant de comparer l'approche méthodologique mise en œuvre pour cartographier l'abondance d'espèces benthico-démersale dans Alglave *et al.* (under review). Dans cette approche sur la sardine, les caractéristiques de la pêcherie font que les données commerciales sont finalement peu informatives de la distribution de la sardine. Les données scientifiques restent donc la source d'information à privilégier.

A l'inverse, dans le cas des espèces benthico-démersales traitées dans Alglave *et al.* (under review), les données commerciales sélectionnées pour cartographier la distribution sont plus informatives : seuls les chalutiers de fonds sont utilisés – OTB, OTT et ceux-ci ont une couverture spatiale plus large et un comportement de ciblage opportuniste. A l'inverse du cas de la sardine, l'information portée par les données commerciales domine celle portée par les campagnes scientifiques. En effet, la taille des échantillons issus des campagnes de chalutage scientifique est beaucoup plus réduite. Le tableau 2 résume les différences entre les 2 types de cas d'étude.

**Tableau 2 :** Différences entre cas d'étude pélagique et benthico-démersaux

Caractéristique	Espèces benthico-démersales (Sole, Merlu et calamars)	Espèces pélagiques
<b>Flottilles commerciales</b>	Ciblage faible sur un nombre élevé d'espèces (pêcheries plurispecifiques)	Ciblage fort sur quelques espèces (pêcherie monospécifique pour les chaluts pélagiques)
	Plusieurs centaines de navires	Quelques dizaines de navires
	Rayon d'action large	Rayon d'action restreint dans l'espace
	Comportement opportuniste	Saisonnalité très marquée
<b>Flottilles scientifiques</b>	Faible taille d'échantillon (~150 traits de chalut au maximum)	Forte densité d'échantillonnage (~1500 observations acoustiques associées à des traits de chaluts)

---

## Références

- Abbott JK, Haynie AC, Reimer MN** (2015) Hidden Flexibility: Institutions, Incentives, and the Margins of Selectivity in Fishing. *Land Econ* **91**: 169–195
- Alglave B, Rivot E, Etienne MP, Woillez M, Thor J, Vermaud Y** (in press) Integrated framework accounting for preferential sampling to infer fish spatial distribution. *ICES J. Mar. Sci.*
- Arbault S, Lacroix N** (1970) Quatre ans de mesures volumétriques de plancton total dans le Golfe de Gascogne (1964-1967). *Rev Trav L’Institut Pêch Marit* **34**: 59–68
- Arbault S, Lacroix N** (1977) Œufs et larves de Clupéidés et Engraulidés dans le golfe de Gascogne (1969 - 1973). Distribution des frayères. Relations entre les facteurs du milieu et de la reproduction. *Rev Trav L’Institut Pêch Marit* **41**: 227–254
- Archambault B, Rivot E, Savina M, Le Pape O** (2018) Using a spatially structured life cycle model to assess the influence of multiple stressors on an exploited coastal-nursery-dependent population. *Estuar Coast Shelf Sci* **201**: 95–104
- August T, Harvey M, Lightfoot P, Kilbey D, Papadopoulos T, Jepson P** (2015) Emerging technologies for biological recording. *Biol J Linn Soc* **115**: 731–749
- Bakun A** (1996) Patterns in the Ocean: Ocean Processes and Marine Population Dynamics. San Diego: University of California Sea Grant , USA, in co-operation with Centro de Investigaciones Biologicas de Noroeste. La Paz, Baja California Sur, Mexico. 323 pp.
- Bakun A, Broad K** (2003) Environmental ‘loopholes’ and fish population dynamics: comparative pattern recognition with focus on El Niño effects in the Pacific: *Environmental loopholes and fish population dynamics*. *Fish Oceanogr* **12**: 458–473
- Bálint M, Pfenniger M, Grossart H-P, Taberlet P, Vellend M, Leibold MA, Englund G, Bowler D** (2018) Environmental DNA Time Series in Ecology. *Trends Ecol Evol* **33**: 945–957
- Bandarra NM, Batista I, Nunes ML, Empis JM, Christie WW** (1997) Seasonal Changes in Lipid Composition of Sardine (*Sardina pilchardus*). *J Food Sci* **62**: 40–42
- Barange M, Coetzee JC, Twatwa NM** (2005) Strategies of space occupation by anchovy and sardine in the southern Benguela: the role of stock size and intra-species competition. *ICES J Mar Sci* **62**: 645–654
- Barange M, Hampton I** (1997) Spatial structure of co-occurring anchovy and sardine populations from acoustic data: implications for survey design. *Fish Oceanogr* **6**: 94–108
- Bellier E, Planque B, Petitgas P** (2007) Historical fluctuations in spawning location of anchovy (*Engraulis encrasicolus*) and sardine (*Sardina pilchardus*) in the Bay of Biscay during 1967?73 and 2000?2004. *Fish Oceanogr* **16**: 1–15

- 
- Bez N, Walker E, Gaertner D, Rivoirard J, Gaspar P** (2011) Fishing activity of tuna purse seiners estimated from vessel monitoring system (VMS) data. *Can J Fish Aquat Sci* **68**: 1998–2010
- Blangiardo M, Cameletti M, Baio G, Rue H** (2013) Spatial and spatio-temporal models with R-INLA. *Spat Spatiotemporal Epidemiol* **4** : 33 - 49
- Booth A** (2000) Incorporating the spatial component of fisheries data into stock assessment models. *ICES J Mar Sci* **57**: 858–865
- Boyra G, Rico I, Udane Martínez** (2020) Acoustic surveying of anchovy Juveniles in the Bay of Biscay: JUVENA 2020 Survey Report. doi: 10.13140/RG.2.2.36115.09768
- Branch TA, Hilborn R, Haynie AC, Fay G, Flynn L, Griffiths J, Marshall KN, Randall JK, Scheuerell JM, Ward EJ, et al** (2006) Fleet dynamics and fishermen behavior: lessons for fisheries managers. *Can J Fish Aquat Sci* **63**: 1647–1668
- Burgess JW, Shaw E** (1979) Development and ecology of fish schooling. *Oceanus* **27**: 11–17
- Conn PB, Thorson JT, Johnson DS** (2017) Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods Ecol Evol* **8**: 1535–1546
- Coombs SH, Smyth TJ, Conway DVP, Halliday NC, Bernal M, Stratoudakis Y, Alvarez P** (2006) Spawning season and temperature relationships for sardine (*Sardina pilchardus*) in the eastern North Atlantic. *J Mar Biol Assoc U K* **86**: 1245–1252
- Cornou A-S, Quinio-Scavinner M, Sagan J, Cloatre T, Dubroca L, Billet N, Roy EL, Chassanite A, Boiron-Leroy A, Martin-Baillet V** (2021) Captures et rejets des métiers de pêche français. Résultats des observations à bord des navires de pêche professionnelle en 2019. Obsmer. 544
- Cury P, Bakun A, Crawford RJM, Jarre A, Quinones RA, Shannon LJ, Verheye HM** (2000) Small pelagics in upwelling systems: patterns of interactionand structural changes in ““wasp-waist”” ecosystems. *ICES J Mar Sci* **57**: 603–618
- Delage N, Le Pape O** (2016) Inventaire des zones fonctionnelles pour les ressources halieutiques dans les eaux sous souveraineté française. Première partie: définitions, critères d’importance et méthode pour déterminer des zones d’importance à protéger en priorité. Pôle halieutique AGROCAMPUS OUEST, 2016. 36
- Diggle PJ, Menezes R, Su T** (2010) Geostatistical inference under preferential sampling. *J R Stat Soc Ser C Appl Stat* **59**: 191–232
- Doray M, Hervy C, Huret M, Petitgas P** (2018a) Spring habitats of small pelagic fish communities in the Bay of Biscay. *Prog Oceanogr* **166**: 88–108
- Doray M, Masse J, Petitgas P** (2010) Pelagic fish stock assessment by acoustic methods at Ifremer. R.INT. DOP/DCN/EMH 10- 02. 18
- Doray M, Petitgas P, Romagnan JB, Huret M, Duhamel E, Dupuy C, Spitz J, Authier M, Sanchez F, Berger L, et al** (2018b) The PELGAS survey: Ship-based integrated

- 
- monitoring of the Bay of Biscay pelagic ecosystem. *Prog Oceanogr* **166**: 15–29
- Doston RC, Griffith DA** (1996) A high-speed midwater trawl for collecting coastal pelagic fishes. *CalCOFI Rep* **37**: 134–139
- Ettahiri O, Berraho A, Vidy G, Ramdani M, Do chi T** (2003) Observation on the spawning of Sardina and Sardinella off the south Moroccan Atlantic coast (21–26°N). *Fish Res* **60**: 207–222
- Fréon P, Misund OA** (1999) Dynamics of pelagic fish distribution and behaviour: effects on fisheries and stock assessment. *Fishing News Books*, Oxford
- Furnestin J, Furnestin M-L** (1959) La reproduction de la sardine et de l'anchovy des côtes atlantiques du Maroc (saisons et aires de ponte). *Rev Trav Inst Pêch Marit* **23**: 79–104
- Gaston KJ** (2000) Global patterns in biodiversity. *Nature* **405**: 220–227
- Gatti P, Cominassi L, Duhamel E, Grellier P, Le Delliou H, Le Mestre S, Petitgas P, Rabiller M, Spitz J, Huret M** (2018) Bioenergetic condition of anchovy and sardine in the Bay of Biscay and English Channel. *Prog Oceanogr* **166**: 129–138
- Georgakarakos S, Trygonis V, Haralabous J** (2011) Accuracy of Acoustic Methods in Fish Stock Assessment Surveys. *Sonar Syst*. doi: 10.5772/18631
- Gerritsen H, Lordan C** (2011) Integrating vessel monitoring systems (VMS) data with daily catch data from logbooks to explore the spatial distribution of catch and effort at high resolution. *ICES J Mar Sci* **68**: 245–252
- Gibb R, Browning E, Glover-Kapfer P, Jones KE** (2019) Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods Ecol Evol* **10**: 169–185
- Girardin R, Hamon KG, Pinnegar J, Poos JJ, Thébaud O, Tidd A, Vermaire Y, Marchal P** (2017) Thirty years of fleet dynamics modelling using discrete-choice models: What have we learned? *Fish Fish* **18**: 638–655
- Gómez-Rubio, V** (2020) Bayesian Inference with INLA. Chapman & Hall/CRC Press. Boca Raton, FL.
- Gonzalez GM, Wiff R, Marshal CT, Cornullier T** (2021) Estimating spatio-temporal distribution of fish and gear selectivity functions from pooled scientific survey and commercial fishing data. *Fish Res* **243**: 11
- Guerineau L, Rochet M-J, Peronnet I** (2010) Panorama des rejets dans les pêches françaises. 49
- Hilborn R, Walters CJ** (1992) Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty., Springer US.
- Hintzen NT, Bastardie F, Beare D, Piet GJ, Ulrich C, Deporte N, Egekvist J, Degel H** (2012) VMStools: Open-source software for the processing, analysis and visualisation of fisheries logbook and VMS data. *Fish Res* **115–116**: 31–43

- 
- Hosmer DW, Lemeshow S, Sturdivant RX** (2013) Applied logistic regression, Third edition. Wiley, Hoboken, New Jersey
- Hunter E, Berry F, Buckley AA, Stewart C, Metcalfe JD** (2006) Seasonal migration of thornback rays and implications for closure management: Ray migration and closure management. *J Appl Ecol* **43**: 710–720
- Hunter, JR, Alheit J** (1995) International GLOBEC Small Pelagic Fishes and Climate Change program. Report of the First Planning Meeting, La Paz, Mexico, June 20-24, 1994.. GLOBEC Rep. 8, 72 pp.
- Huret M, Bourriau P, Doray M, Gohin F, Petitgas P** (2018) Survey timing vs. ecosystem scheduling: Degree-days to underpin observed interannual variability in marine ecosystems. *Prog Oceanogr* **166**: 30–40
- ICES** (2005) Report of the Workshop on Survey Design and Data Analysis (WKSAD), Sète, France. ICES CM 2005/B:07. 170
- ICES** (2016) Report of the Working Group on Southern Horse Mackerel, Anchovy and Sardine (WGHANSA), 24–29 June 2016, Lorient, France. ICES CM 2016/ACOM:17. 588 pp.
- ICES** (2020) Sardine (*Sardina pilchardus*) in divisions 8.a-b and 8.d (Bay of Biscay). Rep ICES Advis Comm 2020 ICES Advice 2020 Pil278abd. doi: 10.17895/ICES.ADVICE.5906
- Isaac NJB, Jarzyna MA, Keil P, Dambly LI, Boersch-Supan PH, Browning E, Freeman SN, Golding N, Guillera-Arroita G, Henrys PA, et al** (2020) Data Integration for Large-Scale Models of Species Distributions. *Trends Ecol Evol* **35**: 56–67
- Jiménez-Valverde A, Lobo JM, Hortal J** (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Divers Distrib* **14**: 885–890
- Krainski ET, Gomez-Rubio V, Bakka H, Lenzi A, Castro-Camilo D, Simpson D, Lindgren F, Rue H** (2019) Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. Chapman & Hall/CRC Press, Boca Raton, FL.
- Lamoreux JF, Morrison JC, Ricketts TH, Olson DM, Dinerstein E, McKnight MW, Shugart HH** (2006) Global tests of biodiversity concordance and the importance of endemism. *Nature* **440**: 212–214
- Lavialle G, Duhamel E, Véron M** (2019) Preliminary results on the comparison of the sardine growth between the channel part (>48°N; Douarnenez Bay) and the south part of the bay of Biscay Sardine.
- Lee J, South AB, Jennings S** (2010) Developing reliable, repeatable, and accessible methods to provide high-resolution estimates of fishing-effort distributions from vessel monitoring system (VMS) data. *ICES J Mar Sci* **67**: 1260–1271
- L'Herrou R** (1967) Répartition des oeufs et larves de sardine dans le Golfe de Gascogne et sur le plateau celtique (mai 1966 ; février et mai 1967). *CM 1967 J 14*
- Lindgren F, Rue H, Lindström J** (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach:

- 
- Link between Gaussian Fields and Gaussian Markov Random Fields. *J R Stat Soc Ser B Stat Methodol* **73**: 423–498
- Lobo JM, Jiménez-Valverde A, Real R** (2008) AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* **17**: 145–151
- Mackinson S** (1999) Cross-scale observations on distribution and behavioural dynamics of ocean feeding Norwegian spring-spawning herring (*Clupea harengus* L.). *ICES J Mar Sci* **56**: 613–626
- Maclellan DN, Fernandes PG, Dalen J** (2002) A consistent approach to definitions and symbols in fisheries acoustics. *ICES J Mar Sci* **59**: 365–369
- Marrs SJ, Tuck ID, Atkinson RJA, Stevenson TDI, Hall C** (2002) Position data loggers and logbooks as tools in fisheries research: results of a pilot study and some recommendations. *Fish Res* **58**: 109–117
- Marshall CE, Glegg GA, Howell KL** (2014) Species distribution modelling to support marine conservation planning: The next steps. *Mar Policy* **45**: 330–332
- Marzuki MI** (2017) VMS data analyses and modeling for the monitoring and surveillance of Indonesian fisheries. *Computer Vision and Pattern Recognition [cs.CV]*. Ecole nationale supérieure Mines-Télécom Atlantique
- Maunder MN, Punt AE** (2013) A review of integrated analysis in fisheries stock assessment. *Fish Res* **142**: 61–74
- Meyer C, Holland K, Papastamatiou Y** (2007) Seasonal and diel movements of giant trevally *Caranx ignobilis* at remote Hawaiian atolls: implications for the design of Marine Protected Areas. *Mar Ecol Prog Ser* **333**: 13–25
- Moriarty M, Sethi SA, Pedreschi D, Smeltz TS, McGonigle C, Harris BP, Wolf N, Greenstreet SPR** (2020) Combining fisheries surveys to inform marine species distribution modelling. *ICES J Mar Sci* **77**: 539–552
- Murray LG, Hinz H, Hold N, Kaiser MJ** (2013) The effectiveness of using CPUE data derived from Vessel Monitoring Systems and fisheries logbooks to estimate scallop biomass. *ICES J Mar Sci* **70**: 1330–1340
- Navigs robert** (2005) Fishing Vessel Monitoring Systems: Past, Present and Future. The High Seas Task Force OECD. Paris.
- Nielsen JR** (2015) Methods for integrated use of fisheries research survey information in understanding marine fish population ecology and better management advice: improving methods for evaluation of research survey information under consideration of survey fish detection and catch efficiency.
- Parrish RH, Nelson CS, Bakun A** (1981) Transport Mechanisms and Reproductive Success of Fishes in the California Current. *Biol Oceanogr* **1**: 175–203
- Parrish RH, Serra R, Grant WS** (1989) The Monotypic Sardines, *Sardina* and *Sardinops*: Their Taxonomy, Distribution, Stock Structure, and Zoogeography. *Can J Fish Aquat*

- Pennino MG, Paradinas I, Illian JB, Muñoz F, Bellido JM, López-Quílez A, Conesa D** (2019) Accounting for preferential sampling in species distribution models. *Ecol Evol* **9**: 653–663
- Petitgas P, Masse J, Bourriaud Paul, Beillois P, Delmas D, Herblard A, Koueta N, Froidefond JM, Santos M** (2006) Hydro-plankton characteristics and their relationship with sardine and anchovy distributions on the French shelf of the Bay of Biscay. *Sci Mar* **70**: 161–172
- Petitgas P, Renard D, Desassis N, Huret M, Romagnan J-B, Doray M, Woillez M, Rivoirard J** (2020) Analysing Temporal Variability in Spatial Distributions Using Min–Max Autocorrelation Factors: Sardine Eggs in the Bay of Biscay. *Math Geosci* **52**: 337–354
- Pitcher TJ** (1980) Some ecological consequences of fish school volumes. *Freshw Biol* **10**: 539–544
- Pitcher TJ** (1995) The impact of pelagic fish behavior on fisheries. *Sci Mar* **59**: 295–306
- Planque B, Bellier E, Lazure P** (2007) Modelling potential spawning habitat of sardine (*Sardina pilchardus*) and anchovy (*Engraulis encrasicolus*) in the Bay of Biscay. *Fish Oceanogr* **16**: 16–30
- Punt AE, Dunn A, Elvarsson BP, Hampton J, Hoyle SD, Maunder MN, Methot RD, Nielsen A** (2020) Essential features of the next-generation integrated fisheries stock assessment package: A perspective. *Fish Res* **229**: 105617
- van Putten IE, Kulmala S, Thébaud O, Dowling N, Hamon KG, Hutton T, Pascoe S** (2012) Theories and behavioural drivers underlying fleet dynamics models: Theories and behavioural drivers. *Fish Fish* **13**: 216–235
- Quero J-C, Dardignac J, Vayne J-J** (1989) Les poissons du golfe de Gascogne.
- Robinson CJ** (2004) Responses of the northern anchovy to the dynamics of the pelagic environment: identification of fish behaviours that may leave the population under risk of overexploitation. *J Fish Biol* **64**: 1072–1087
- Rochette S, Le Pape O, Vigneau J, Rivot E** (2013) A hierarchical Bayesian model for embedding larval drift and habitat models in integrated life cycles for exploited fish. *Ecol Appl* **23**: 1659–1676
- Rue H, Martino S, Chopin N** (2009) Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *J R Stat Soc Ser B* **71**: 319–392
- Rufener M-C** (2020) Rufener, M.-C. 2020. Integrating commercial fisheries and scientific survey data: Advances, new tools and applications to model the fish and fishery dynamics. National Institute of Aquatic Resources, Denmark, DTU Aqua. 209 pp.
- Salas S, Gaertner D** (2004) The behavioural dynamics of fishers: management implications.

- Schaub M, Gimenez O, Sierro A, Arlettaz R** (2007) Use of Integrated Modeling to Enhance Estimates of Population Dynamics Obtained from Limited Data. *Conserv Biol* **21**: 945–955
- Schickele A, Leroy B, Beaugrand G, Goberville E, Hattab T, Francour P, Raybaud V** (2020) Modelling European small pelagic fish distribution: Methodological insights. *Ecol Model* **416**: 108902
- Silva A, Garrido S, Ibaibarriaga L, Pawlowski L, Riveiro I, Marques V, Ramos F, Duhamel E, Iglesias M, Bryère P, et al** (2019) Adult-mediated connectivity and spatial population structure of sardine in the Bay of Biscay and Iberian coast. *Deep Sea Res Part II Top Stud Oceanogr* **159**: 62–74
- Silva A, Skagen DW, Uriarte A, Masse J, Santos MB, Marques V, Carrera P, Beillois P, Pestana G, Porteiro C, et al** (2009) Geographic variability of sardine dynamics in the Iberian Biscay region. *ICES J Mar Sci J Cons* **66**: 495–508
- Spitz J, Ridoux V, Trites AW, Laran S, Authier M** (2018) Prey consumption by cetaceans reveals the importance of energy-rich food webs in the Bay of Biscay. *Prog Oceanogr* **166**: 148–158
- Stephenson F, Mill AC, Scott CL, Stewart GB, Grainger MJ, Polunin NVC, Fitzsimmons C** (2018) Socio-economic, technological and environmental drivers of spatio-temporal changes in fishing pressure. *Mar Policy* **88**: 189–203
- Stratoudakis Y, Coombs S, de Lanzós AL, Halliday N, Costas G, Caneco B, Franco C, Conway D, Santos MB, Silva A, et al** (2007) Sardine (*Sardina pilchardus*) spawning seasonality in European waters of the northeast Atlantic. *Mar Biol* **151**
- Tosello-Bancal F** (1994) L' évolution de la pêche de la sardine sur le littoral français. Thèse de doctorat : Géographie. Paris 4, Paris
- Vermard Y, Marchal P, Mahévas S, Thébaud O** (2008) A dynamic model of the Bay of Biscay pelagic fleet simulating fishing trip choice: the response to the closure of the European anchovy (*Engraulis encrasicolus*) fishery in 2005. *Can J Fish Aquat Sci* **65**: 2444–2453
- Whitehead P** (1985) Clupeoid fishes of the world. An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, anchovies and wolf herrings. Part 1 - Chirocentridae, Clupeidae and Pristigasteridae., Rome. Rome
- Whittaker RH, Levin SA, Root RB** (1973) Niche, Habitat, and Ecotope. *Am Nat* **107**: 321–338
- Worm B, Hilborn R, Baum JK, Branch TA, Collie JS, Costello C, Fogarty MJ, Fulton EA, Hutchings JA, Jennings S, et al** (2009) Rebuilding Global Fisheries. *Science* **325**: 578–585
- Wright AJ, Kyhn LA** (2015) Practical management of cumulative anthropogenic impacts with working marine examples: Practical Cumulative Impact Management. *Conserv Biol* **29**:

---

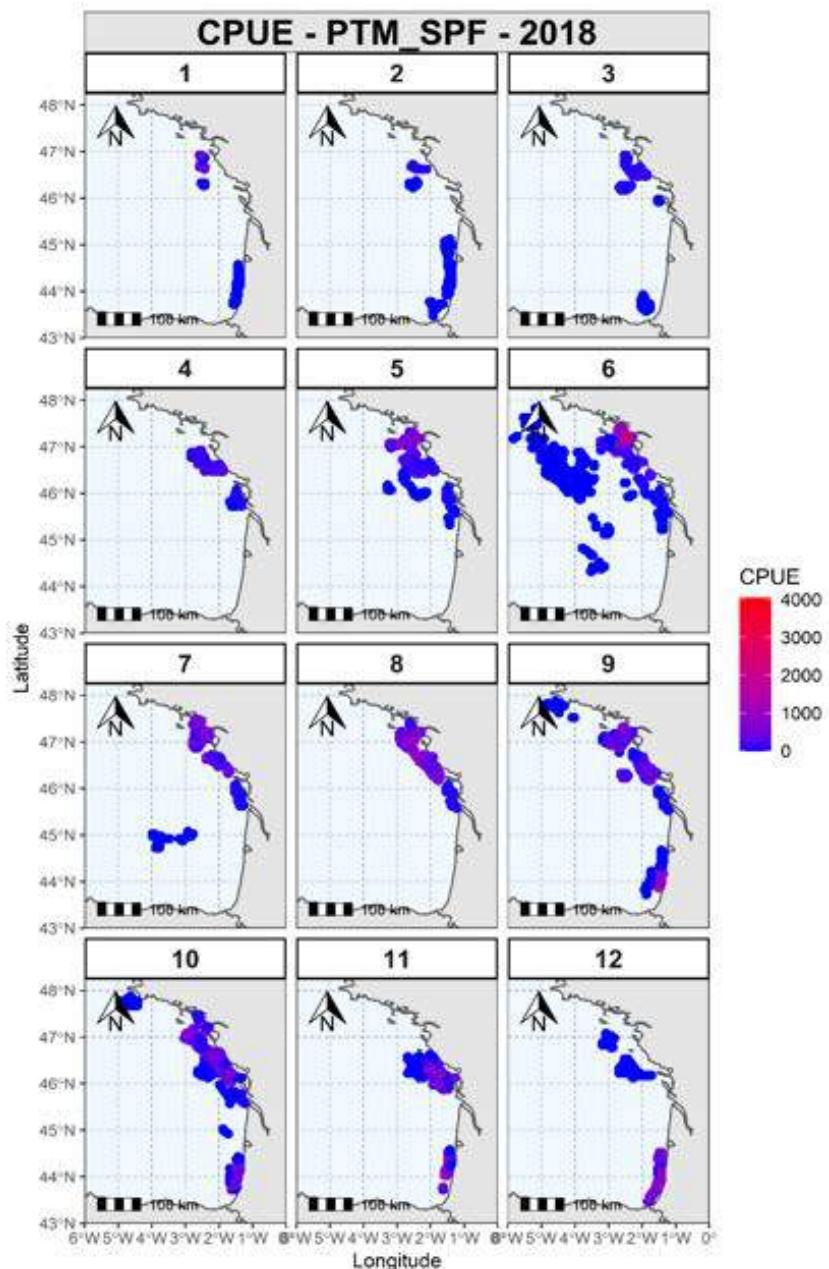
333–340

**Zipkin EF, Inouye BD, Beissinger SR** (2019) Innovations in data integration for modeling populations. *Ecology* e02713

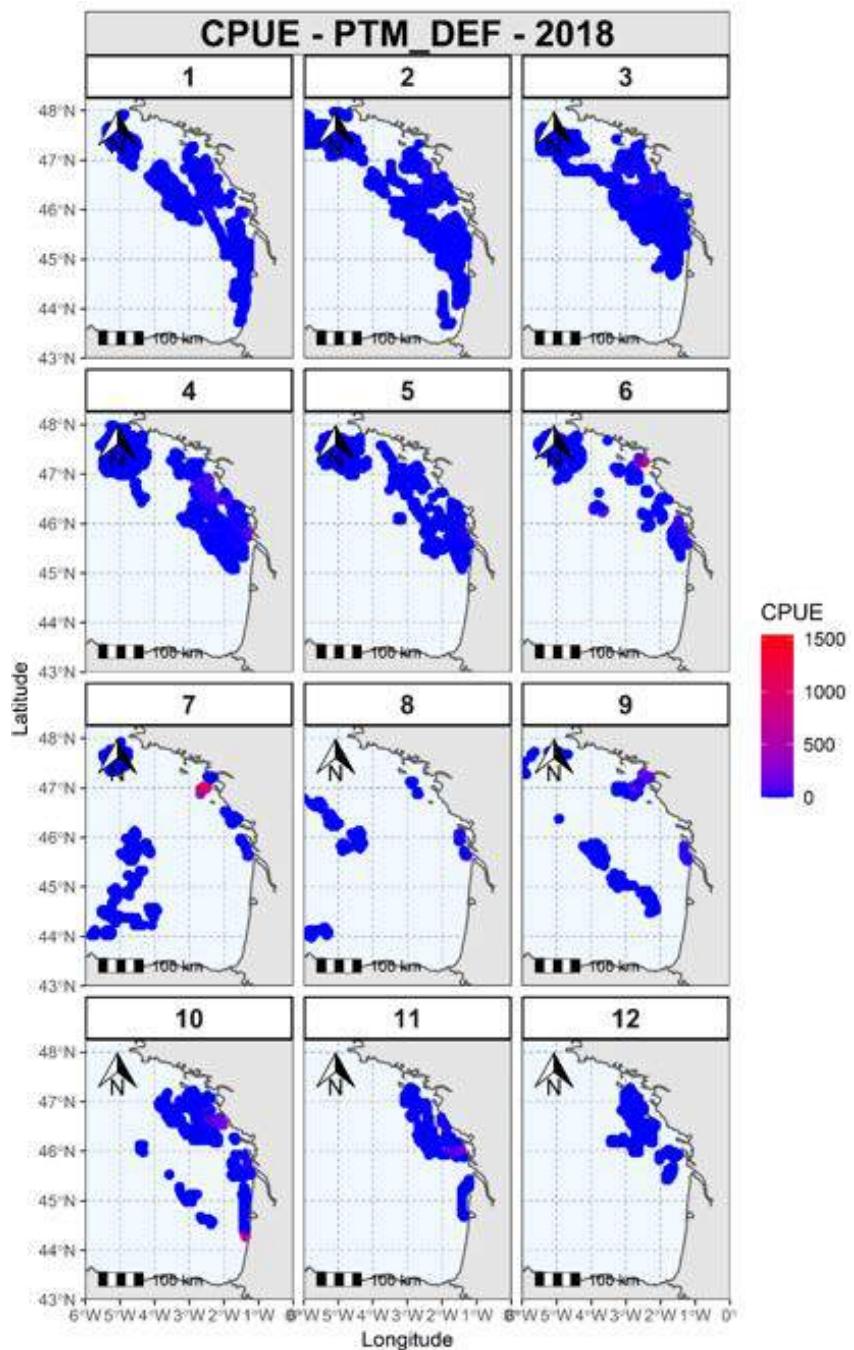
**Zwolinski JP, Oliveira PB, Quintino V, Stratoudakis Y** (2010) Sardine potential habitat and environmental forcing off western Portugal. *ICES J Mar Sci* **67**: 1553–1564

## Annexes

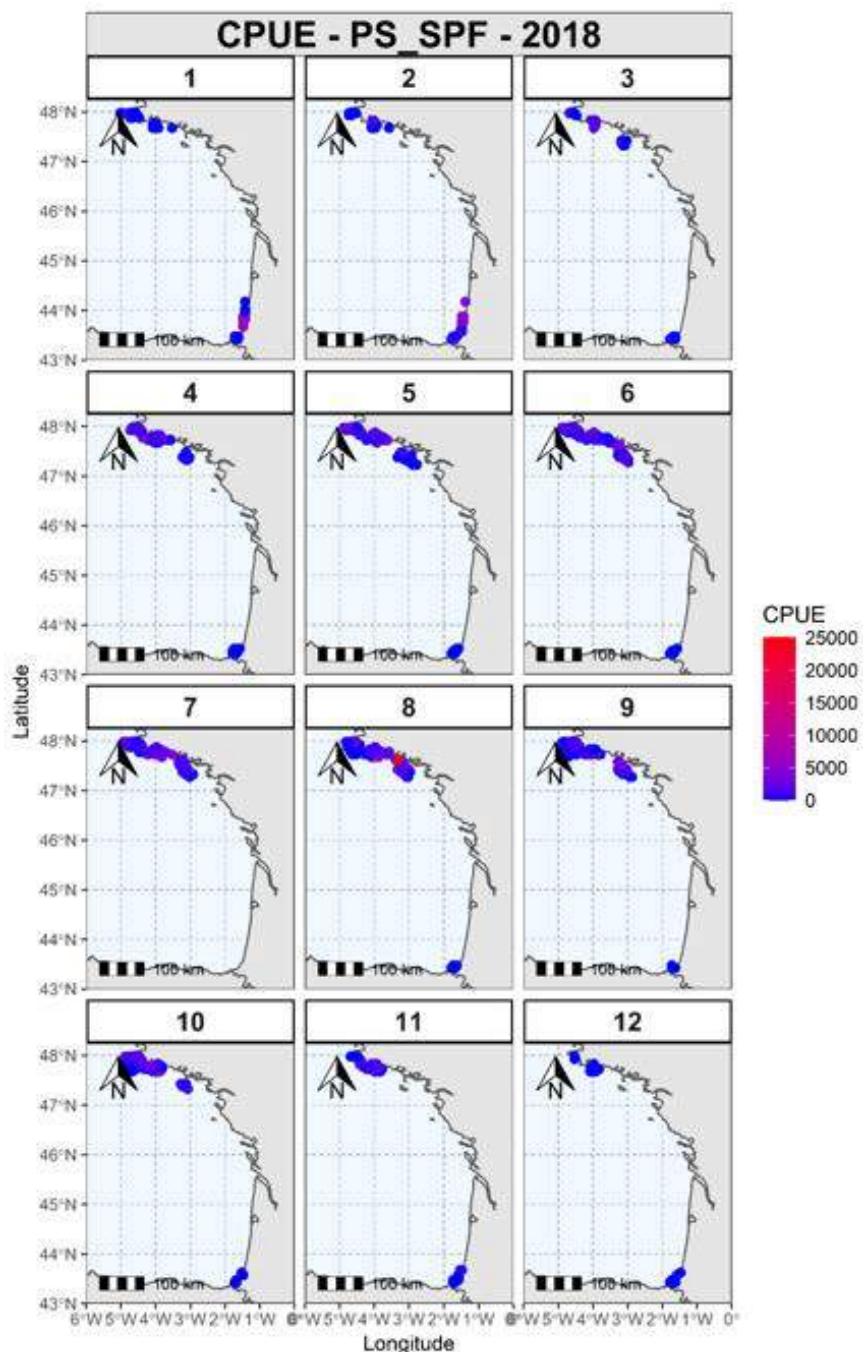
**Annexe 1 :** Analyse exploratoire des données commerciales : Spatialisation des CPUE en Mai 2018 de la flottille PTM\_SPF.



**Annexe 2 : Analyse exploratoire des données commerciales : Spatialisation des CPUE en Mai 2018 de la flottille PTM\_DEF.**

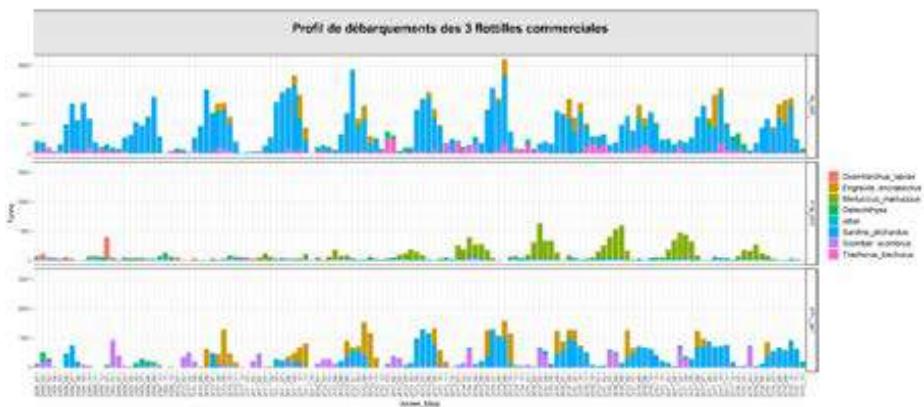


**Annexe 3 : Analyse exploratoire des données commerciales : Spatialisation des CPUE en Mai 2018 de la flottille PS\_SPF.**

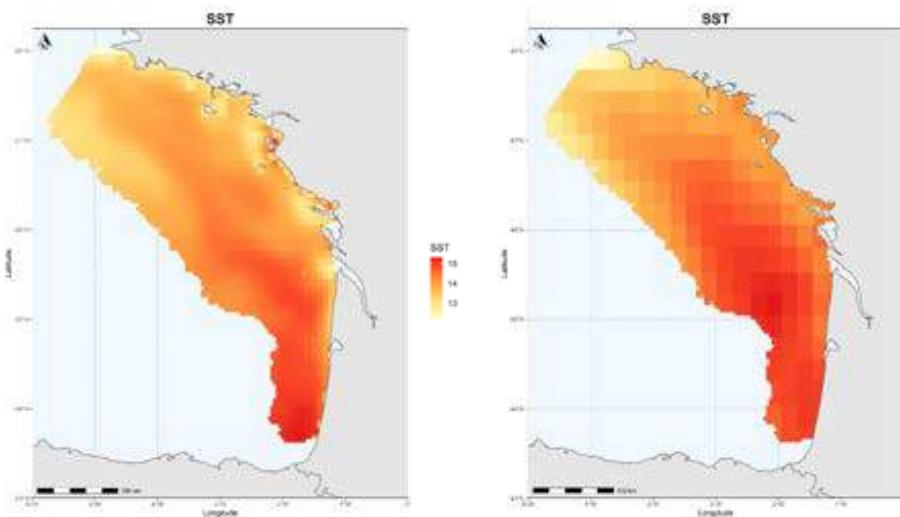


---

**Annexe 4 :** Séries temporelles des profils de débarquements entre 2008-2018 à un pas de temps mensuel pour les 3 flottilles commerciales.

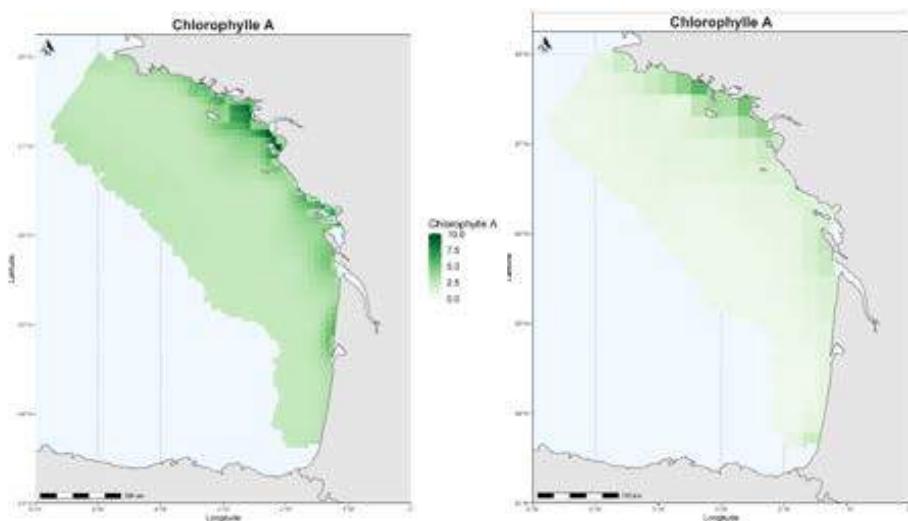


**Annexe 5 :** Cartographie de la SST en Mai 2018 (données issues du modèle biogéochimique POLCOM-ERSEM à gauche et satellitaires à droite).

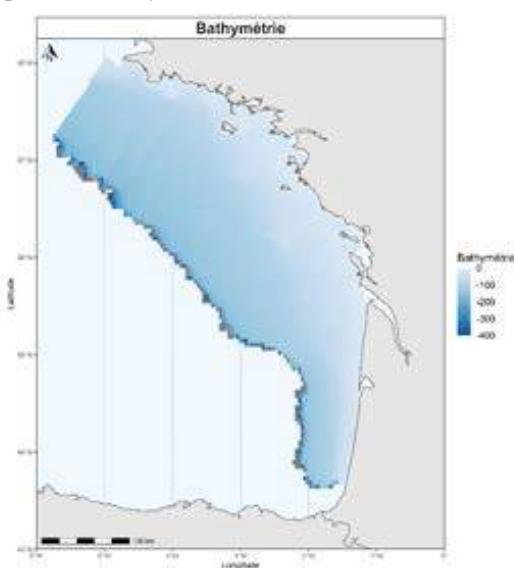


---

**Annexe 6 :** Cartographie de la chlorophylle A en Mai 2018 (données issues du modèle biogéochimique POLCOM-ERSEM à gauche et satellitaires à droite).

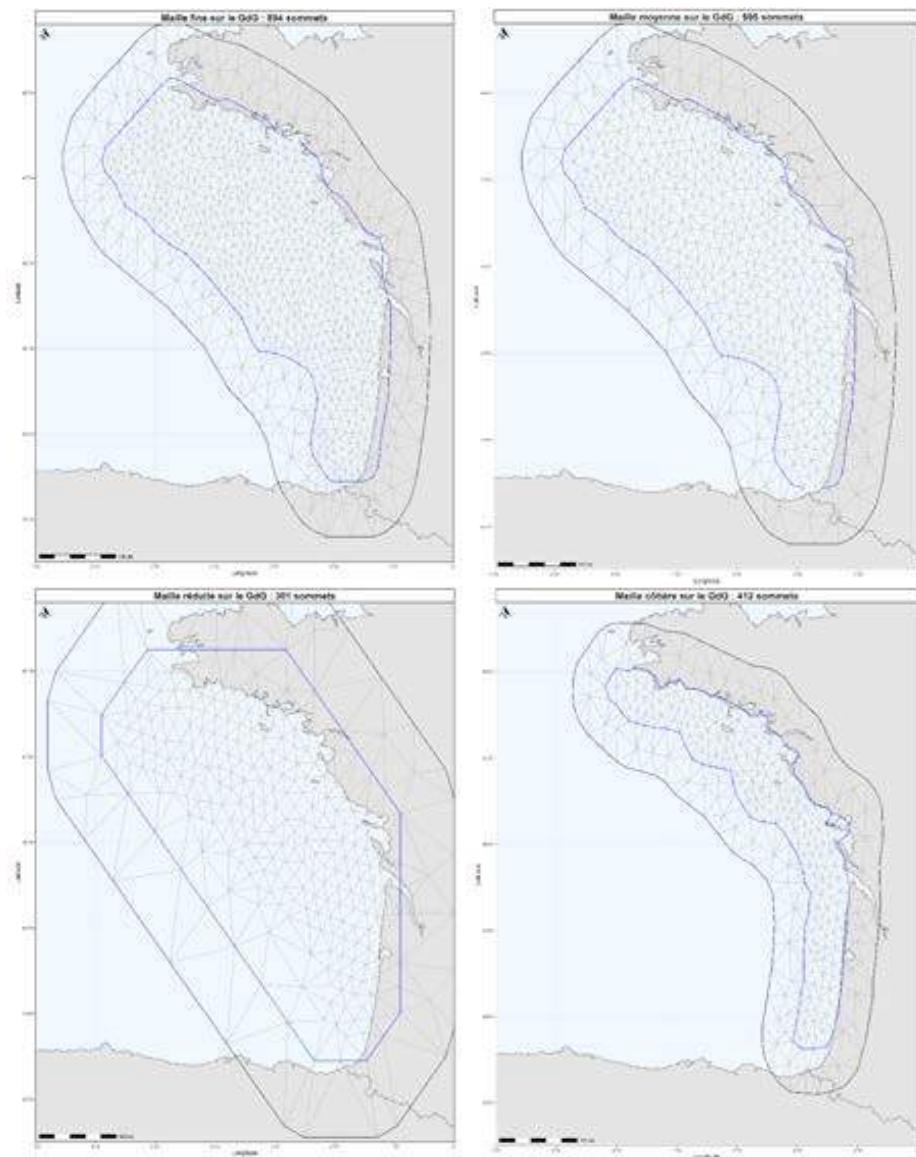


**Annexe 7 :** Cartographie de la bathymétrie en Mai 2018.



---

**Annexe 8 : Maillages utilisées**



---

**Annexe 9 :** Fonctionnement des PC priors et récapitulatif des paramétrisations testées

En suivant l'approche de Simpson *et al.* (2017), nous définissons les SPDE avec des PC priors (penalized complexity priors). La distribution des PC priors est unique pour chaque paramètre. Ils correspondent à la probabilité de s'écarte d'un modèle moins complexe. Pour le paramètre de la variance, le modèle moins complexe correspond à un  $\sigma^2 = 0$ , et pour le paramètre de distance celui-ci a une valeur infinie (soit dépendance parfaite i.e. moins de variabilité). La paramétrisation suit l'idée suivante : Soit  $\tau$  un paramètre, nous définissons  $(u, \alpha)$  tel que :

$$Prob(\tau > u) = \alpha, \quad u > 0, \quad 0 < \alpha < 1$$

La première représente la valeur du paramètre, la seconde est le quantile à partir duquel on dépasse la valeur choisie. Par exemple, nous posons pour la distance de corrélation qu'elle a 10% de chance d'être supérieure à 0.01°. Nous avons testé différentes paramétrisations afin d'étudier leur influence sur l'inférence du champ latent et la relation espèce habitat :

**Tableau 1 :** Paramétrisations testées pour l'effet aléatoire.

Hyperparamètre	Valeurs	Quantile
Distance de corrélation ( $r$ )	0.01	0.1
	2	0.1
Ecart-type ( $\sigma^2$ )	100	0.1
	1	0.1

**Tableau 2 :** Paramétrisations testées pour la relation espèce habitat.

Hyperparamètre	Valeurs	Quantile
Distance de corrélation ( $r$ )	0.01	0.1
	2	0.1
Ecart-type ( $\sigma^2$ )	100	0.1
	2	0.1

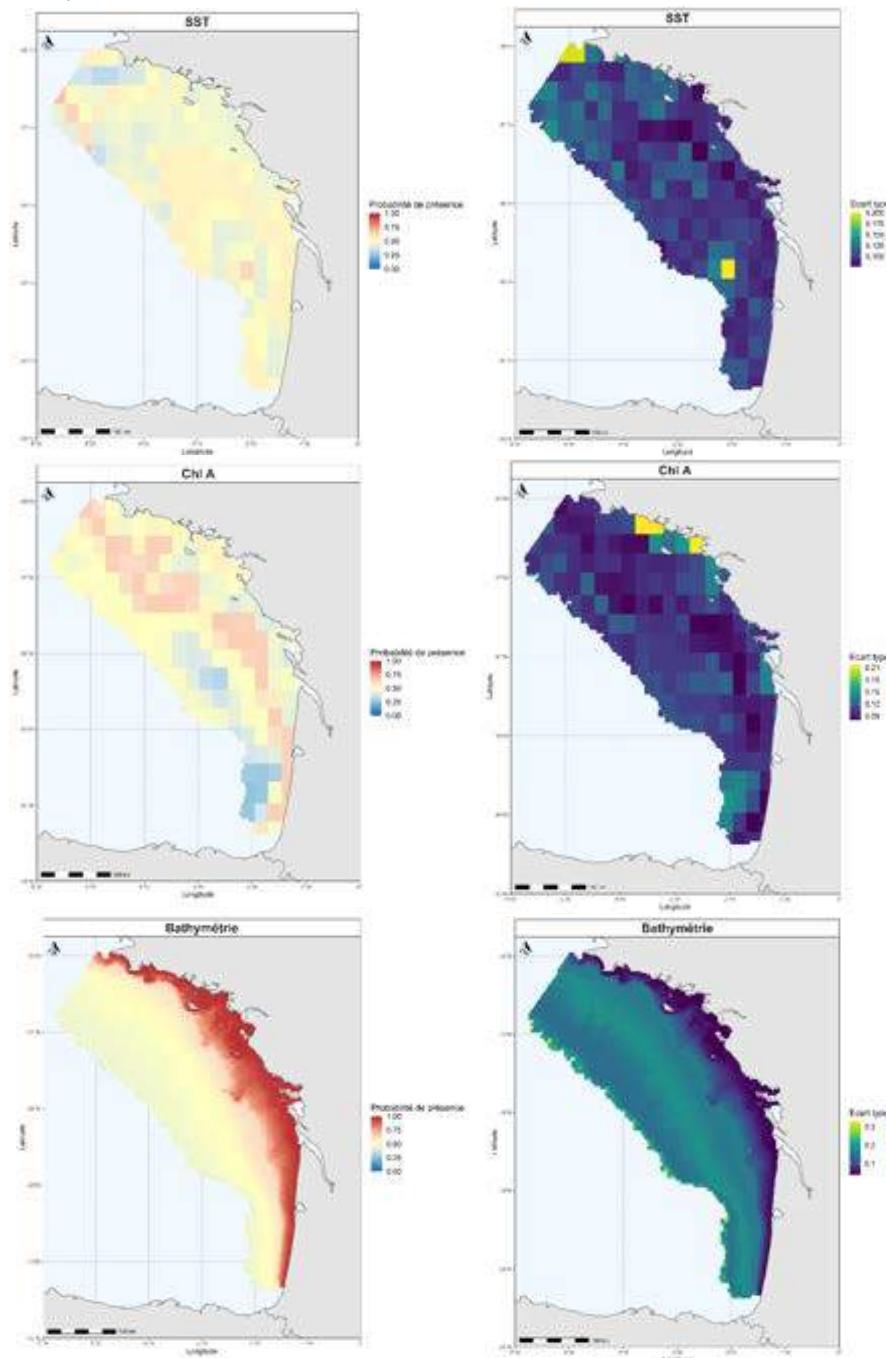
En parallèle, les priors des effets fixes sont par défaut des logGamma de paramétrisations (1, 5<sup>e-5</sup>).

---

**Annexe 10 :** Tableau récapitulatif des différentes modélisations testées (toutes les combinaisons n'ont pas été testées).

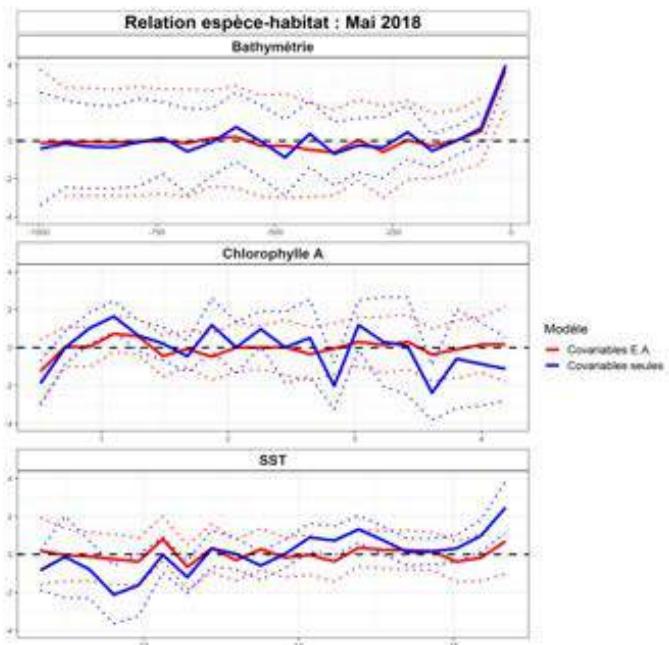
Modèle	Zone d'étude	Maille	Sources de données
Donnée seule	GdG	Réduite	2 survey (PELGAS + JUVENA) + 2 flottilles commerciales (PS + PTM)
Intégré	Frangé côtière (50km)	Moyenne	2 survey + 3 flottilles commerciales (PS_SPF + PTM_SPF + PTM_DEF)
Ech. Préférentiel		Fine	
Co-variables		Côtière	
Intégré annuel			

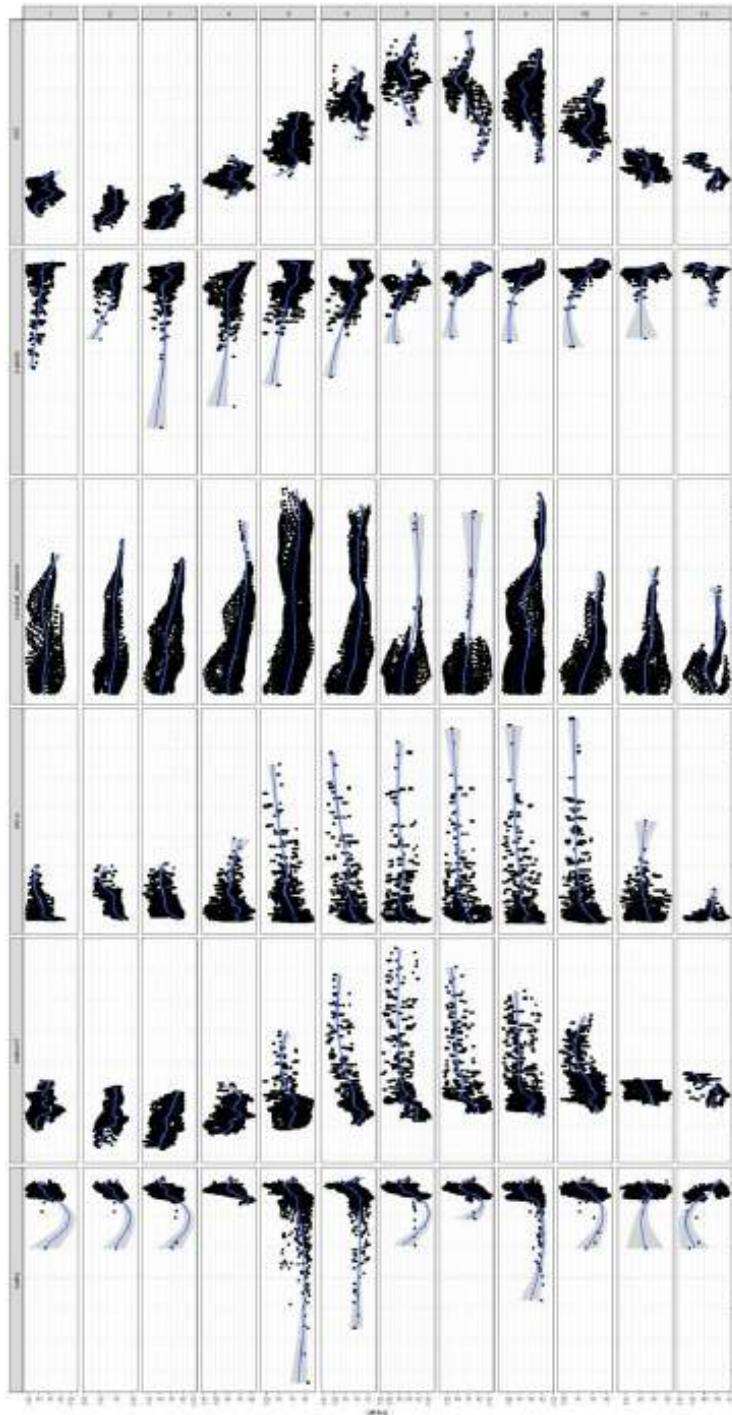
**Annexe 11 :** Inférences obtenues pour chaque variables environnementale (SST, chlorophylle A, bathymétrie).



---

**Annexe 12 :** Comparaison effet des conditions environnementaux avec ou non une prise en compte de l'effet aléatoire (données satellitaires).

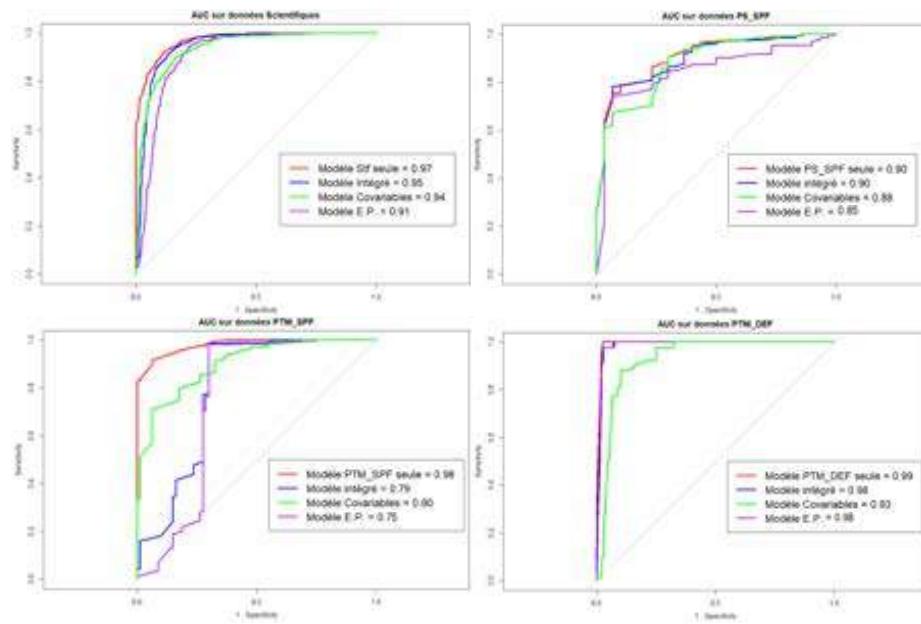




**Annexe 13 :** Estimation du champ aléatoire de présence-absence en fonction des co-variables environnementales par mois en 2018  
(bathymétrie, température de fond, distance à la côte, chlorophylle a, température de surface)

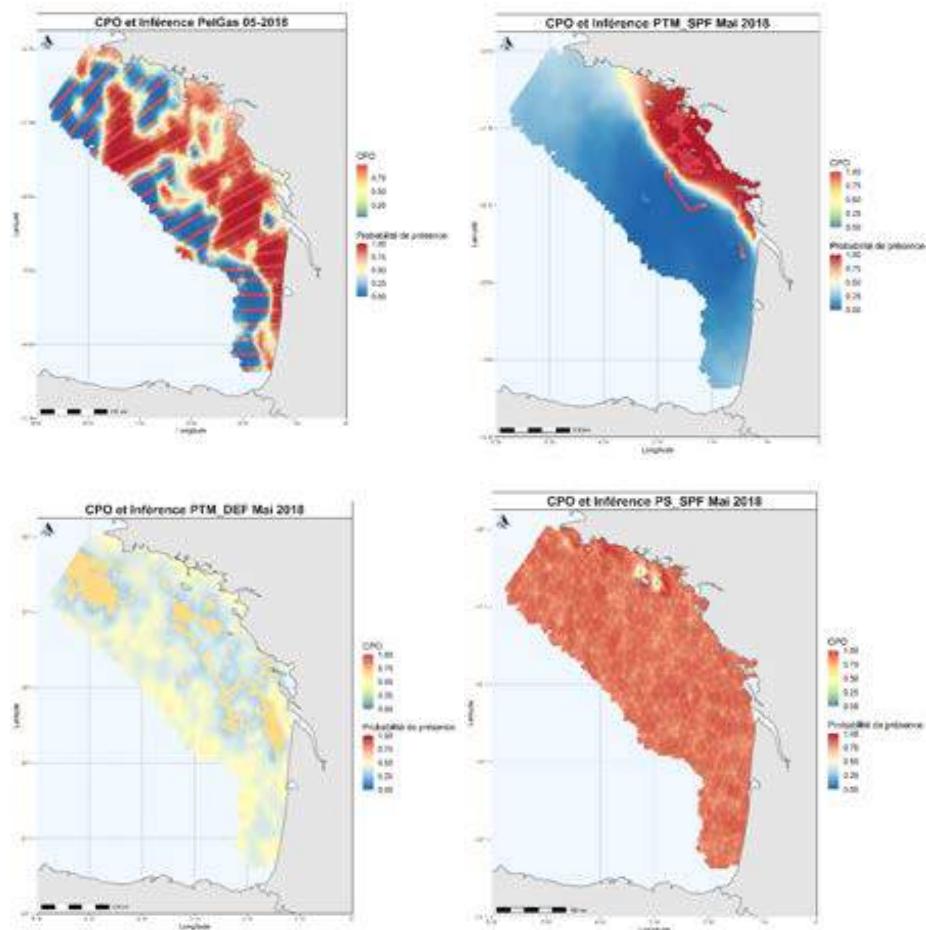
---

**Annexe 14 :** Courbes ROC et AUC pour chaque source de données et modèles.



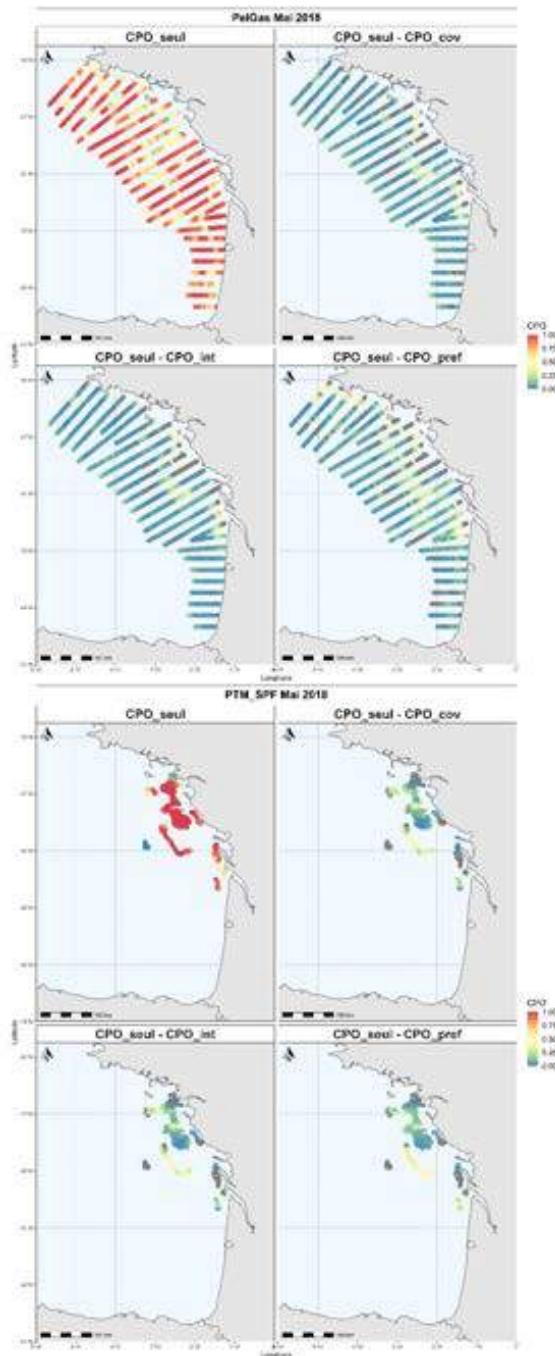
---

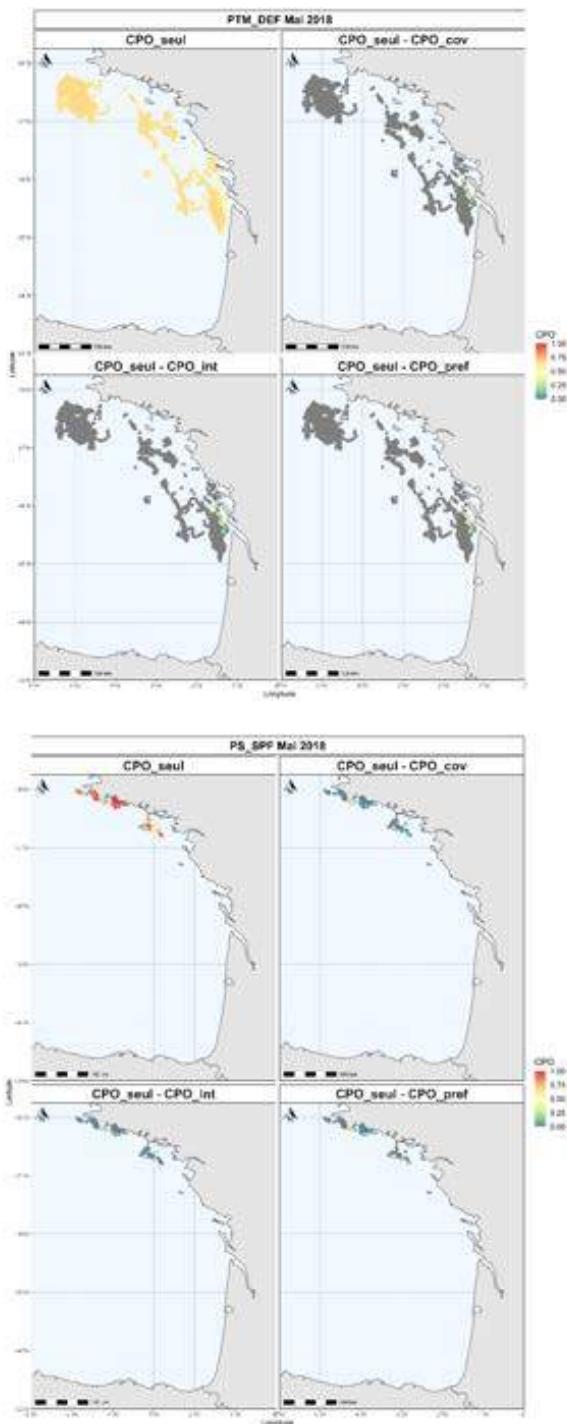
**Annexe 15 :** Cartographies des CPO pour chaque source de donnée issue du modèle « donnée seule ». Les CPO sont représentées sur les cartes d’inférences respectives.



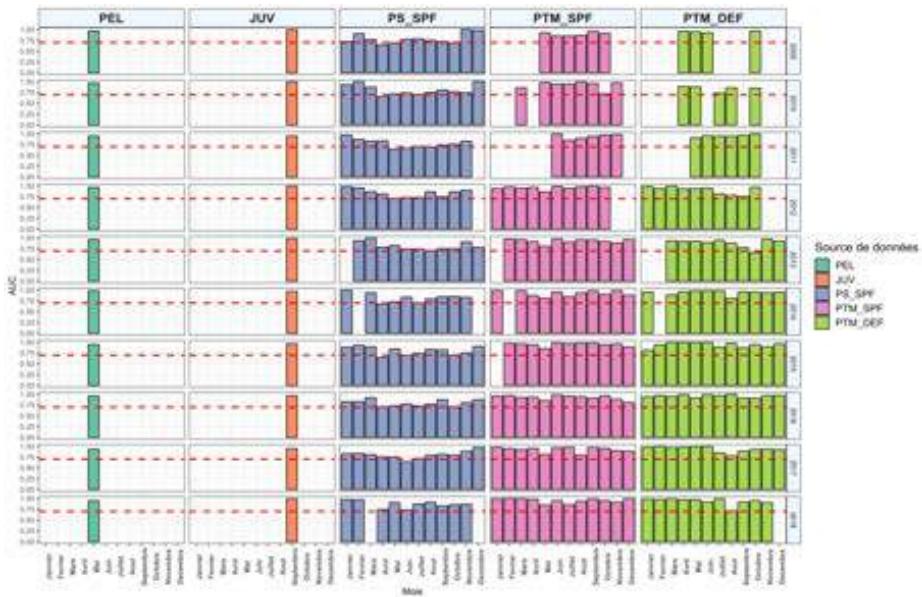
---

**Annexe 16 :** Cartographies des CPO pour chaque source de donnée et issue du modèle « donnée seule », et différence avec la CPO de ce modèle.

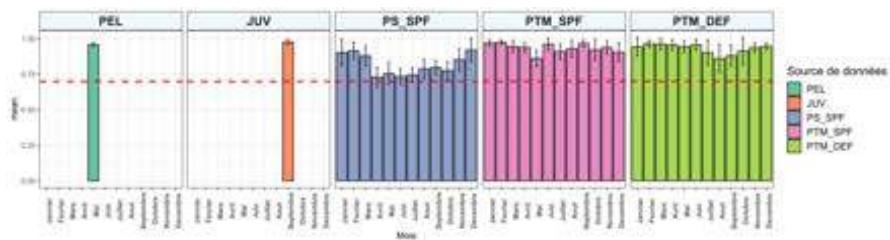




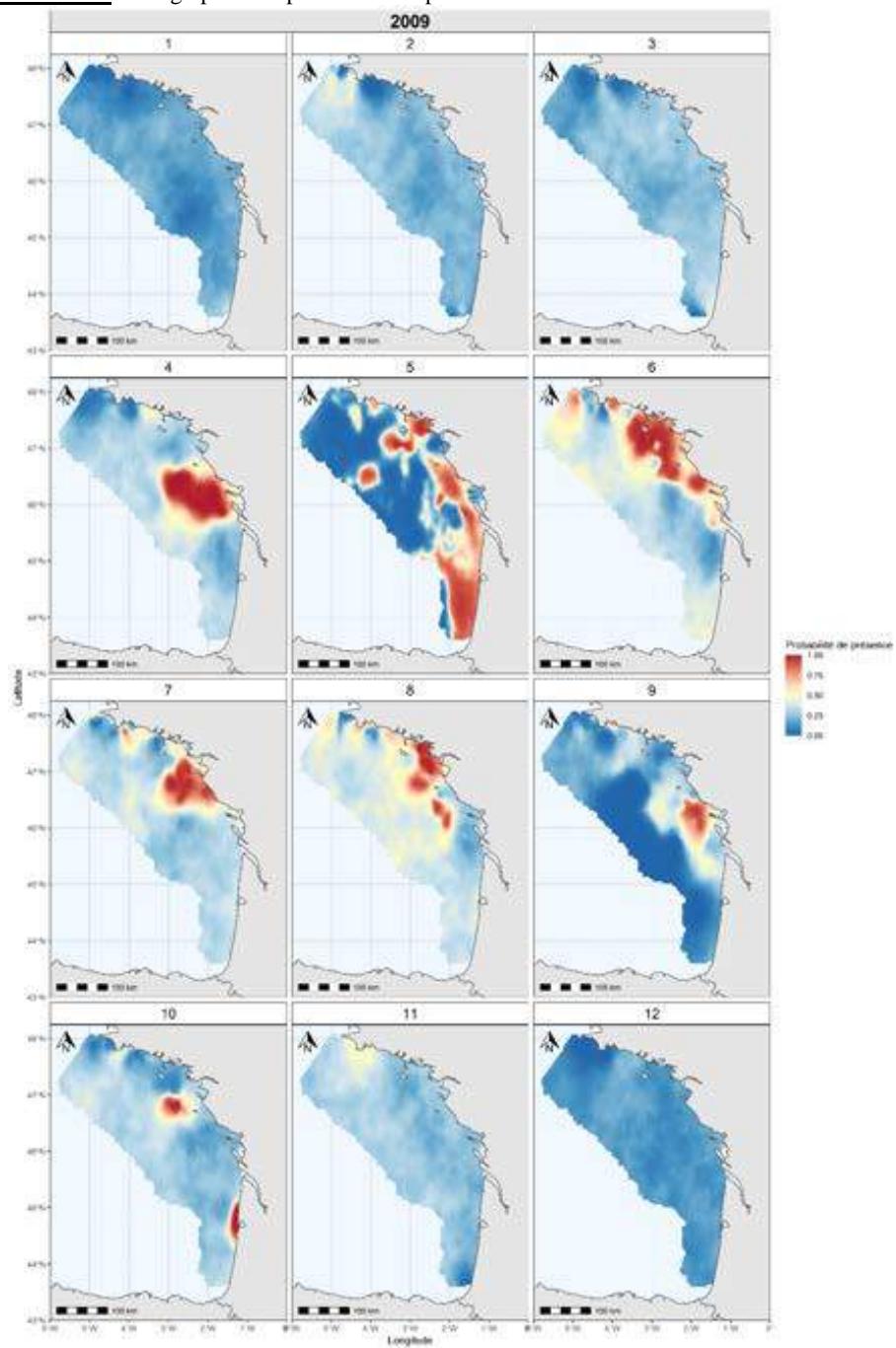
**Annexe 17 :** Scores AUC des modèles spatio-temporels par mois et source de donnée entre 2009 et 2018. Le trait rouge correspond à une AUC de 0.70



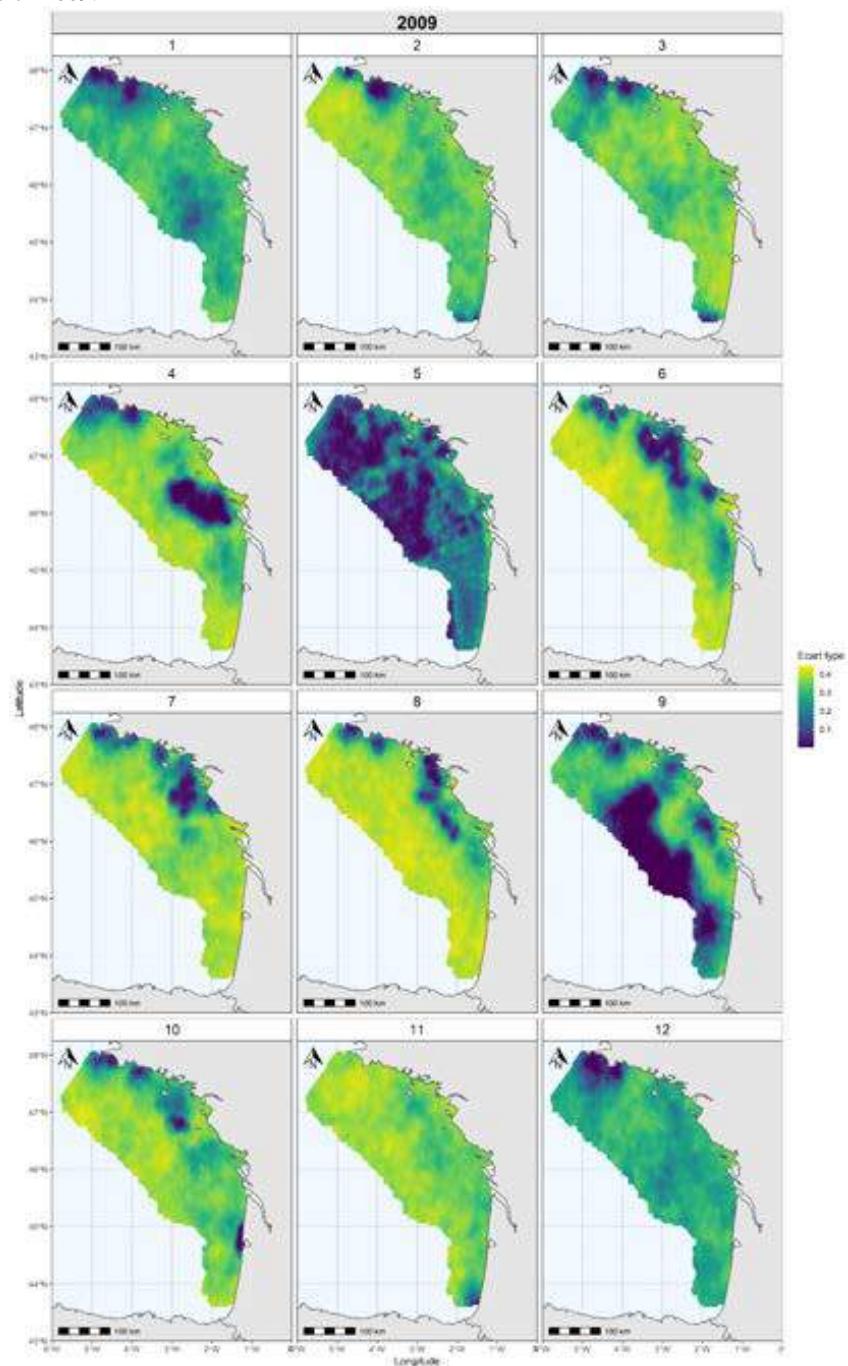
**Annexe 18 :** AUC moyen par mois et source de donnée (2009-2018) et écarts types. Le trait rouge correspond à une AUC de 0.70.



**Annexe 19 :** Cartographies de probabilité de présence de la sardine dans le GdG en 2009.

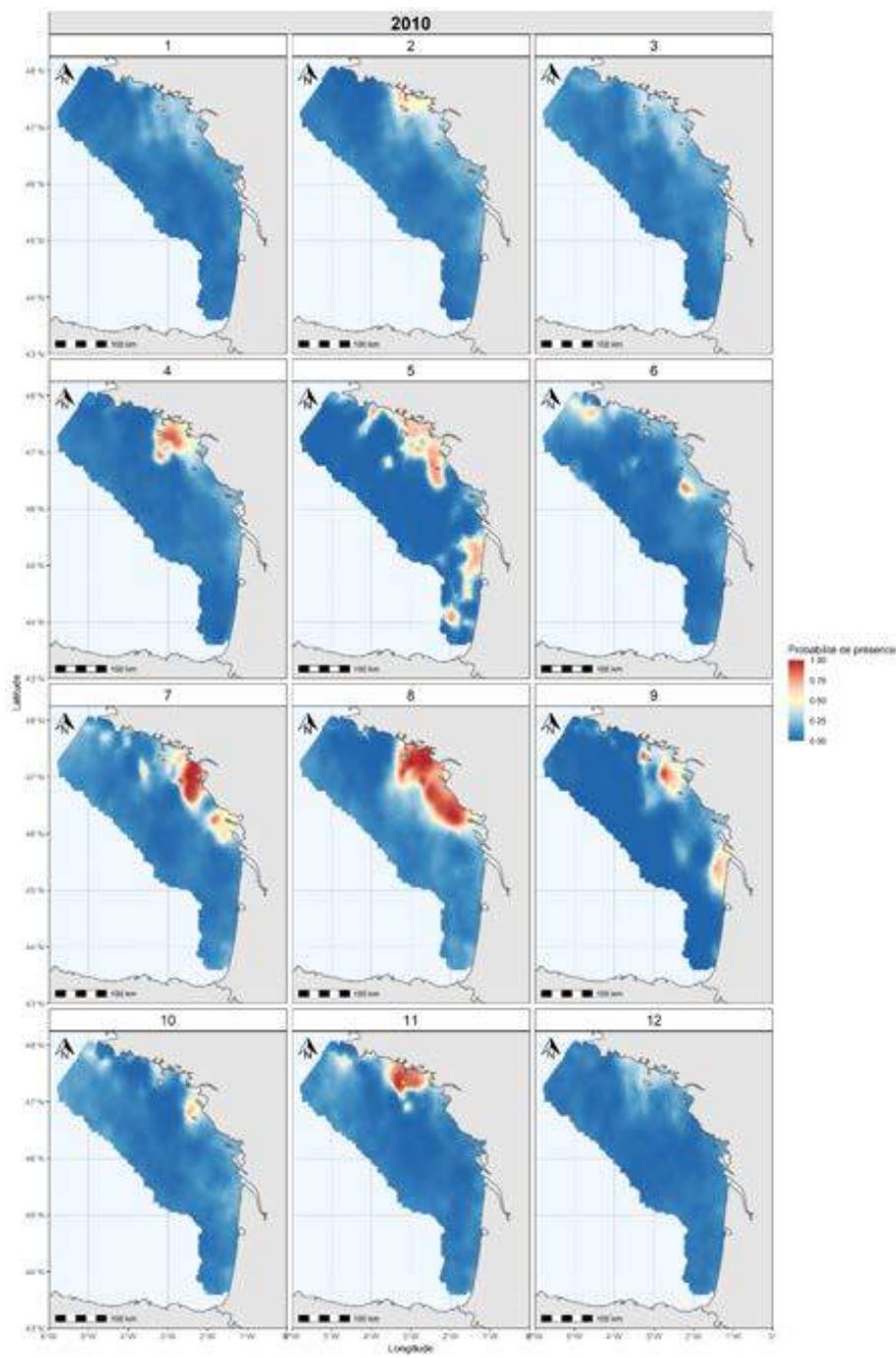


**Annexe 20 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2009.

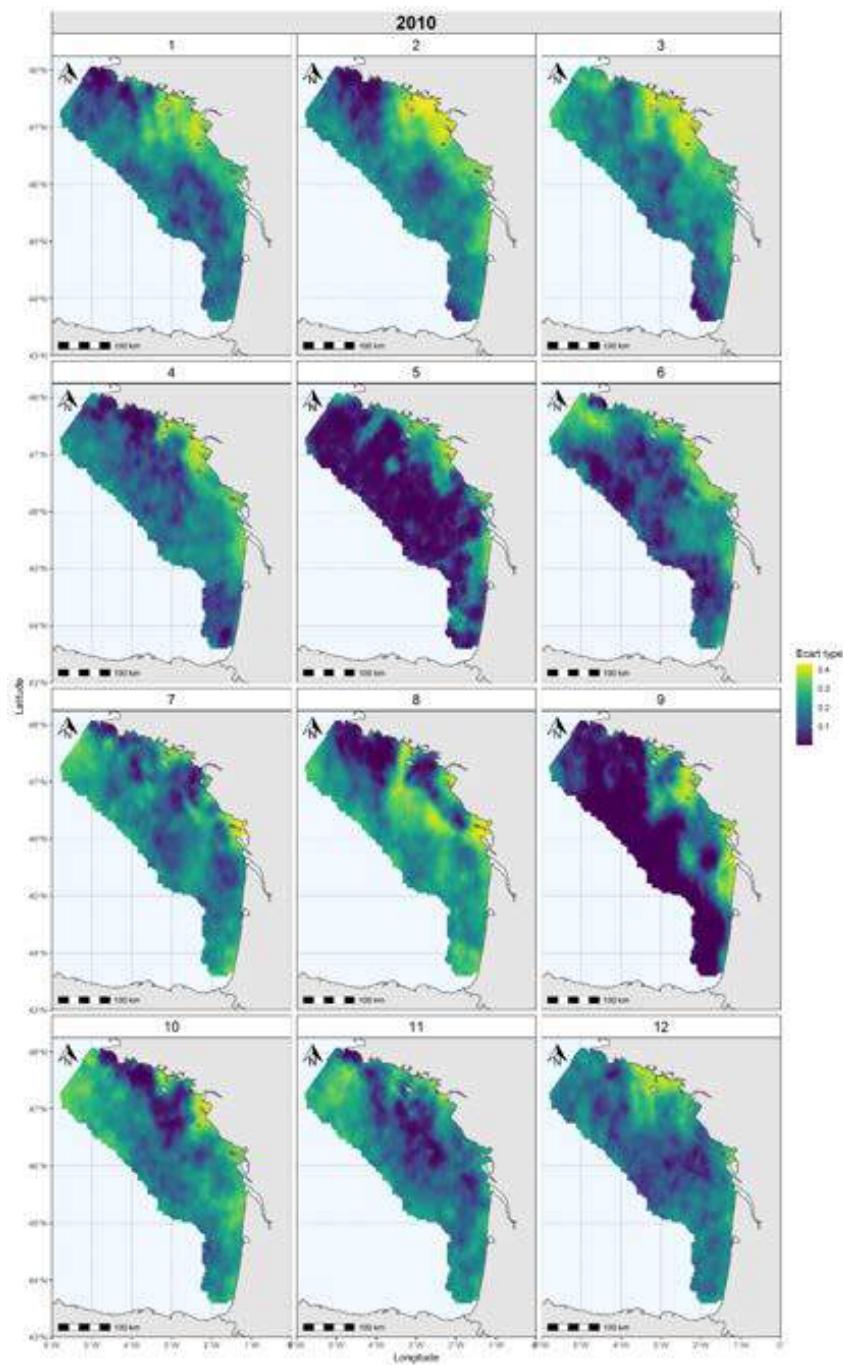


---

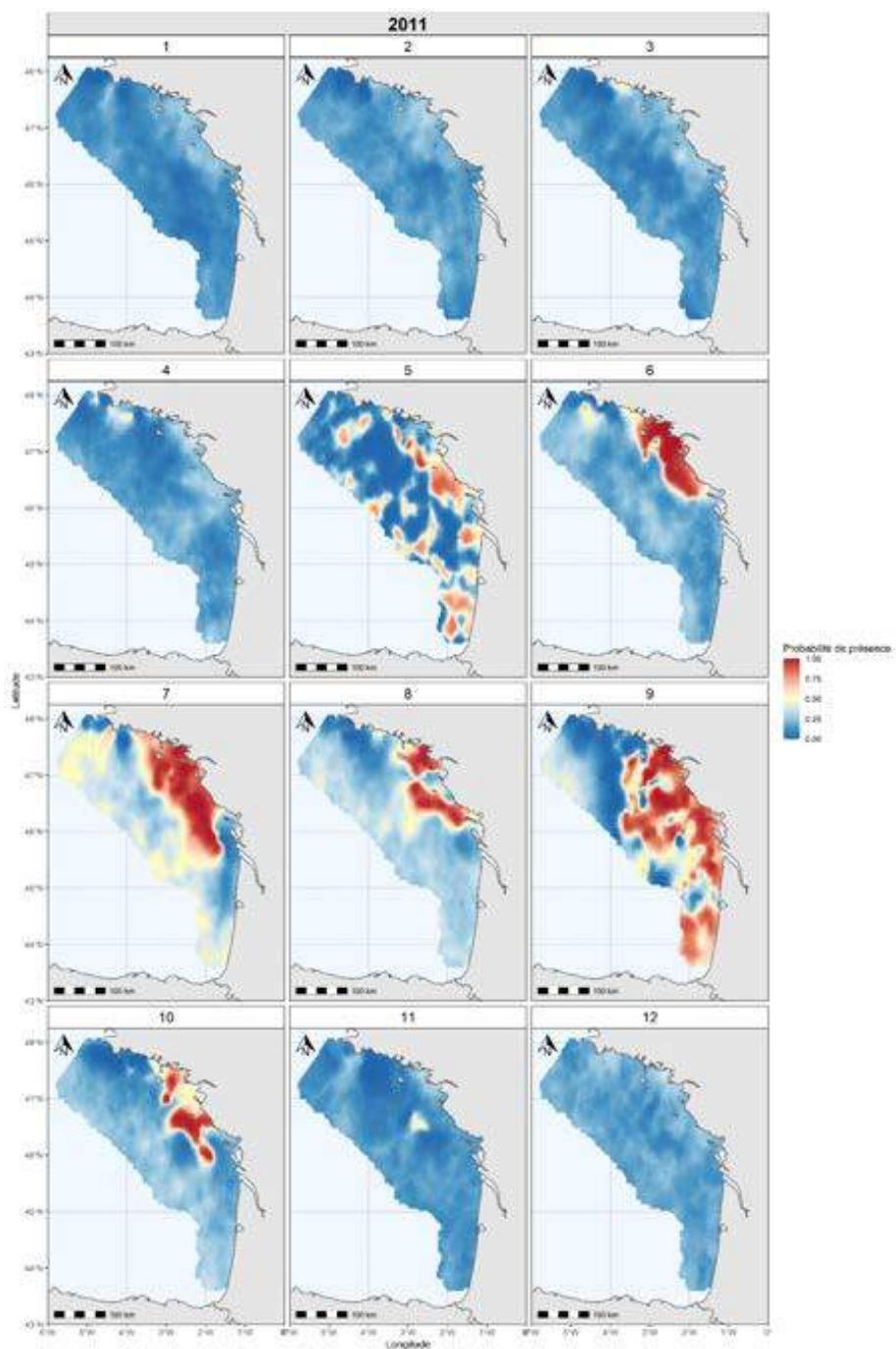
**Annexe 21 :** Cartographies de probabilité de présence de la sardine dans le GdG en 2010.



**Annexe 22 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2010.

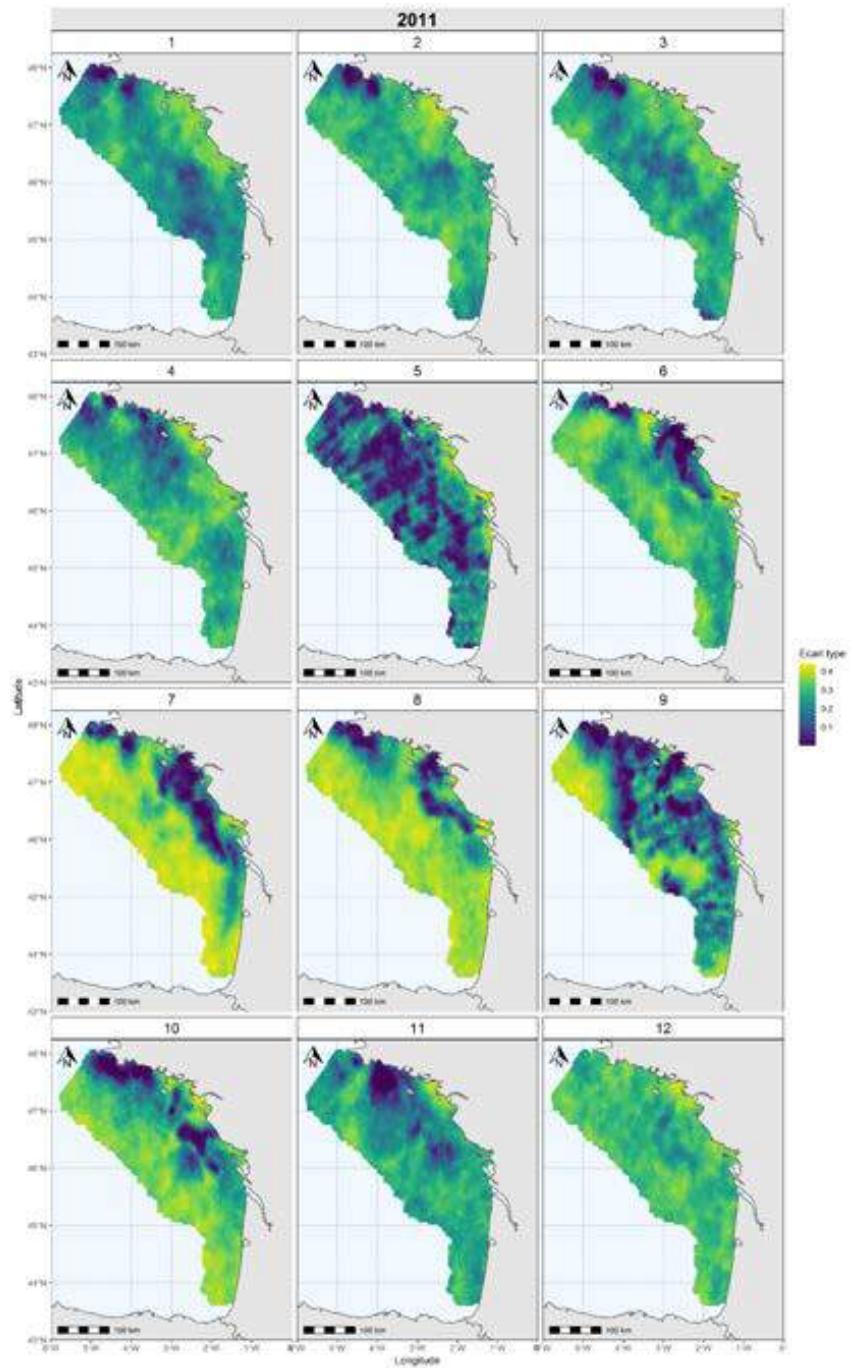


**Annexe 23 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2011.

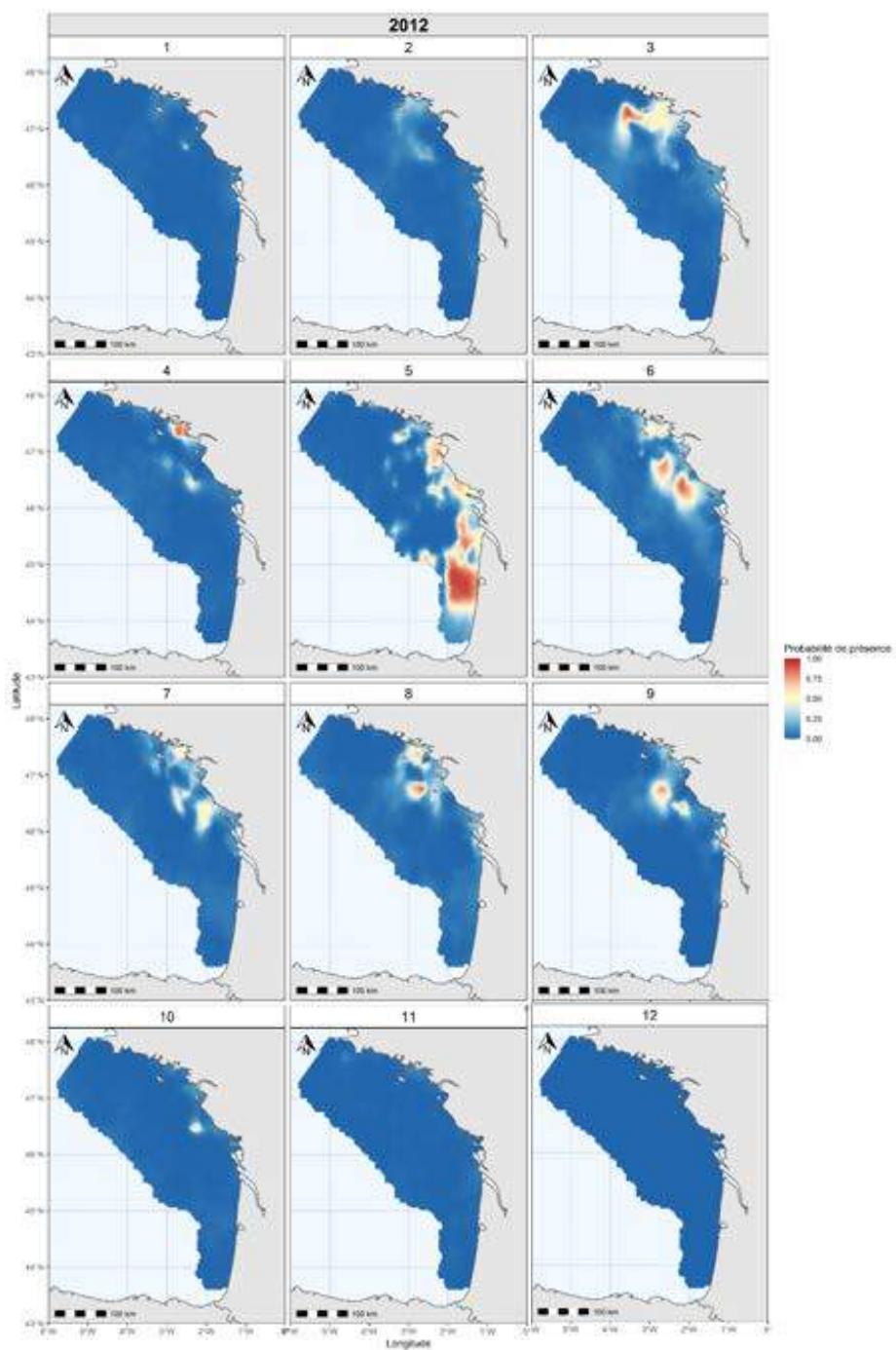


---

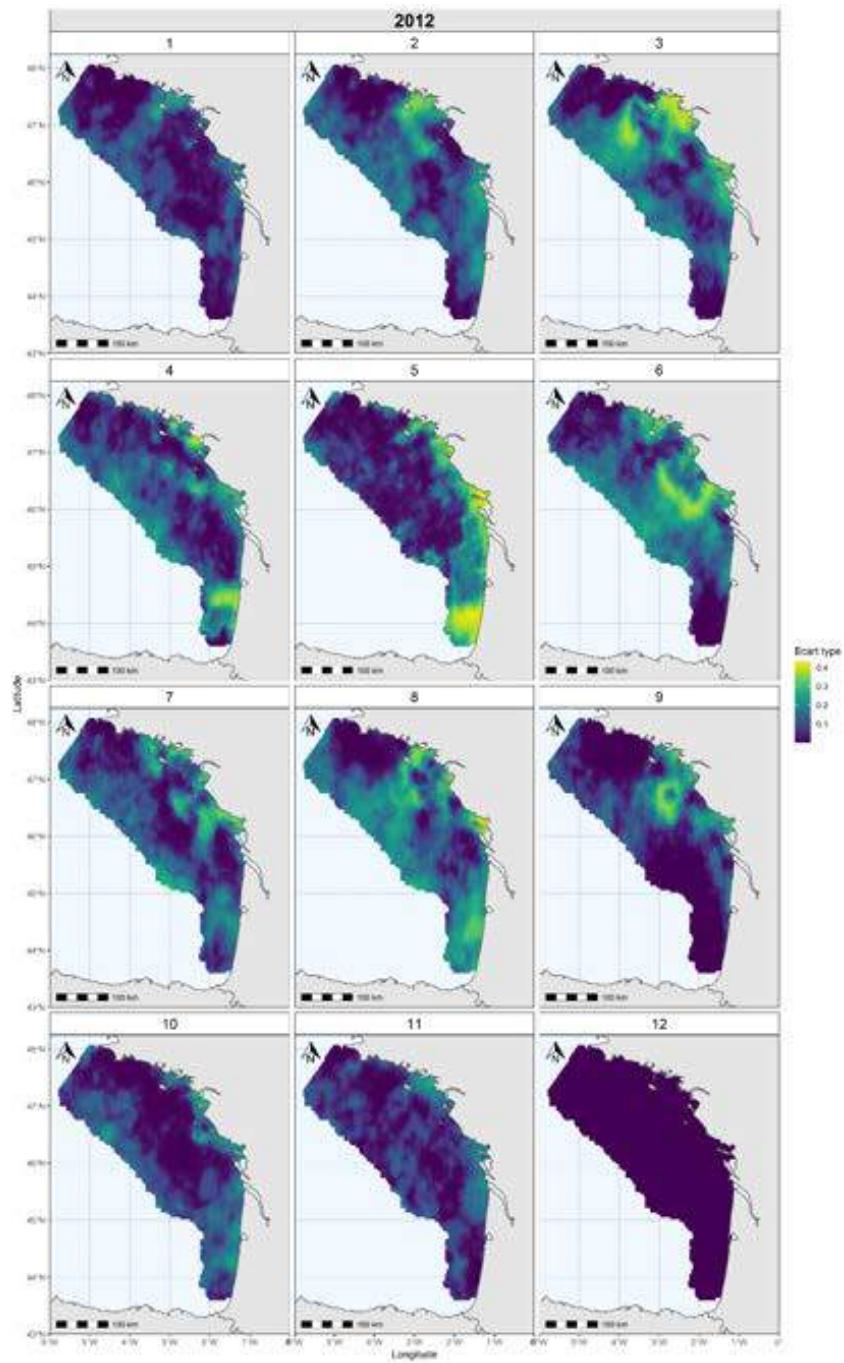
**Annexe 24 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2011.



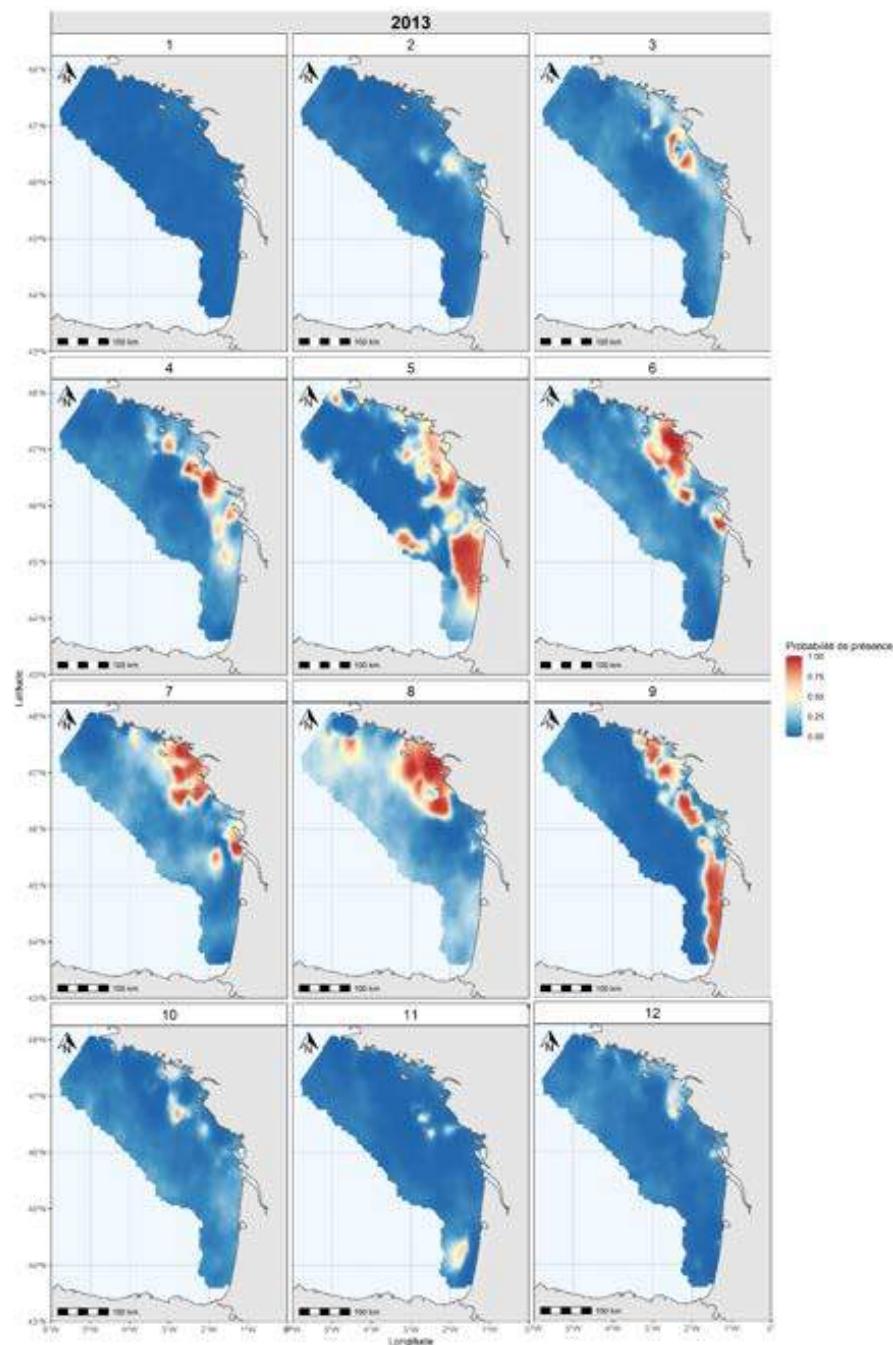
**Annexe 25 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2012.



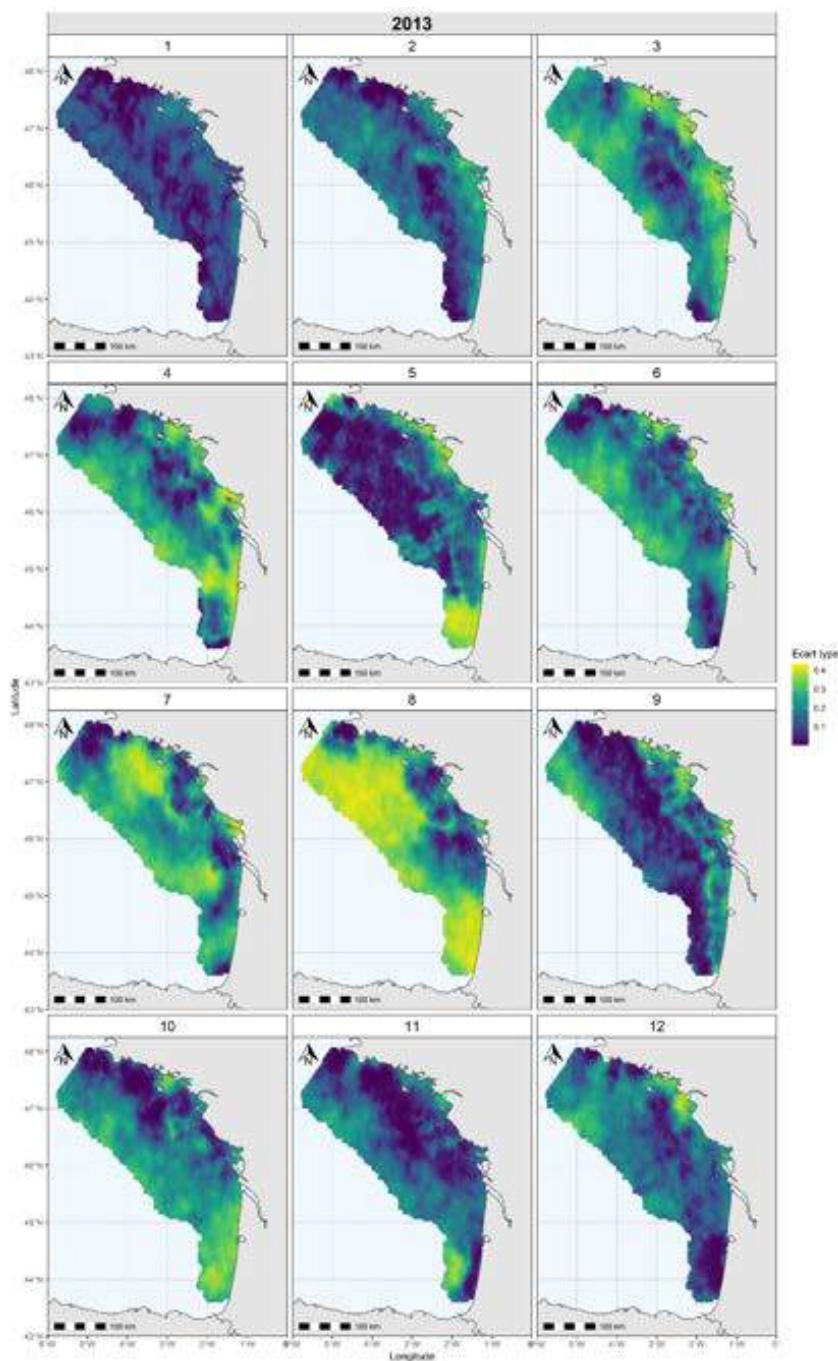
**Annexe 26 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2012.



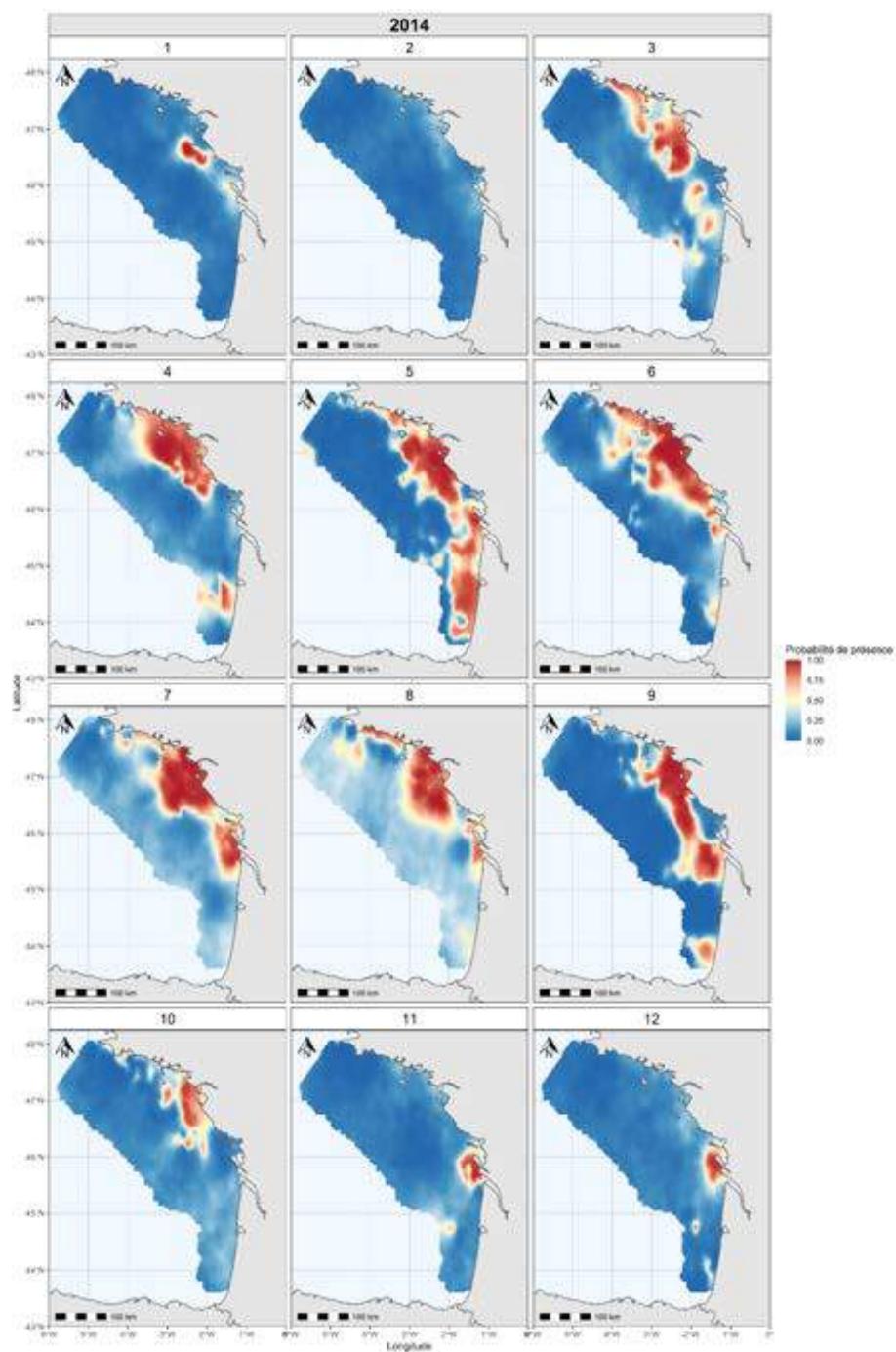
**Annexe 27 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2013.



**Annexe 28** : Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2013.

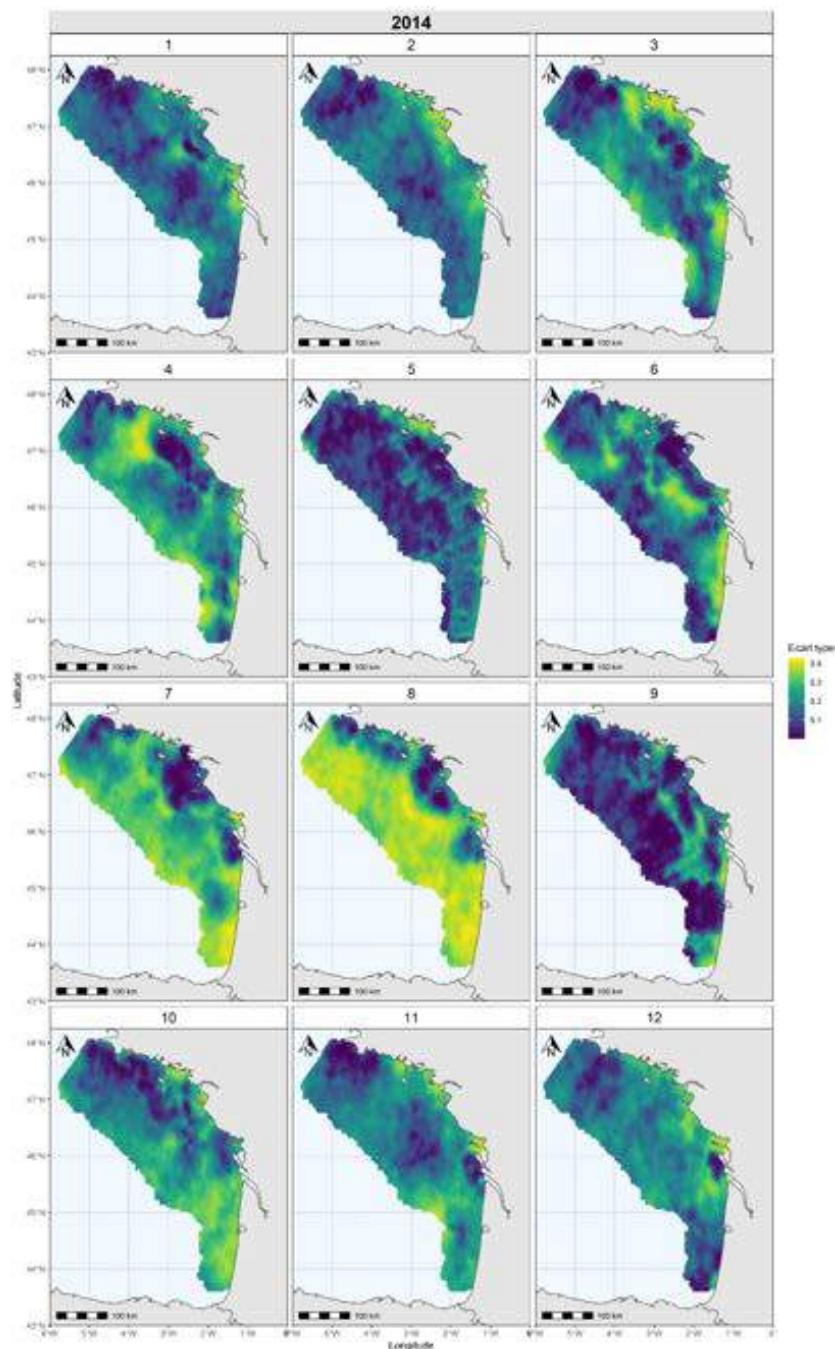


**Annexe 29 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2014.

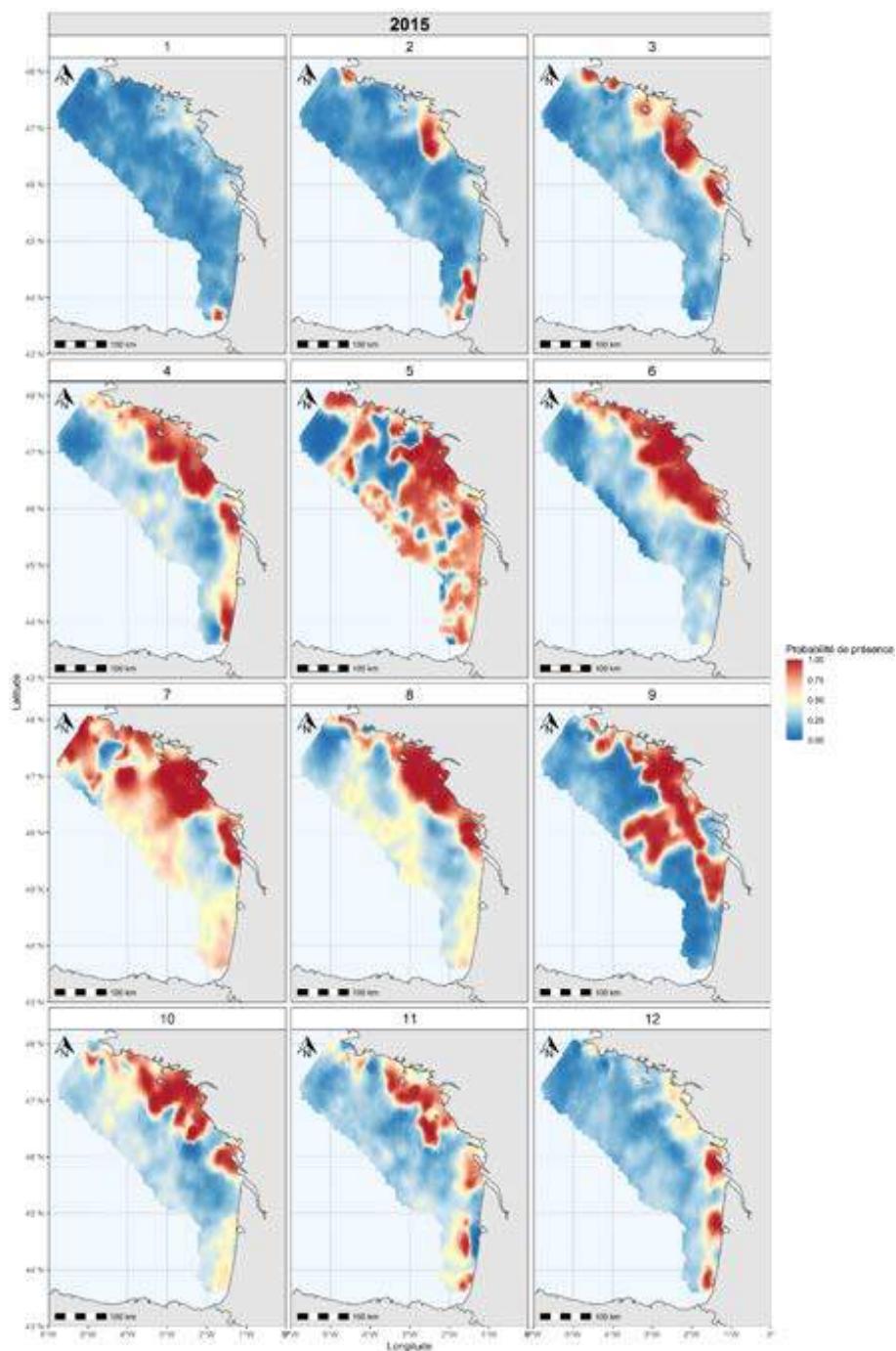


---

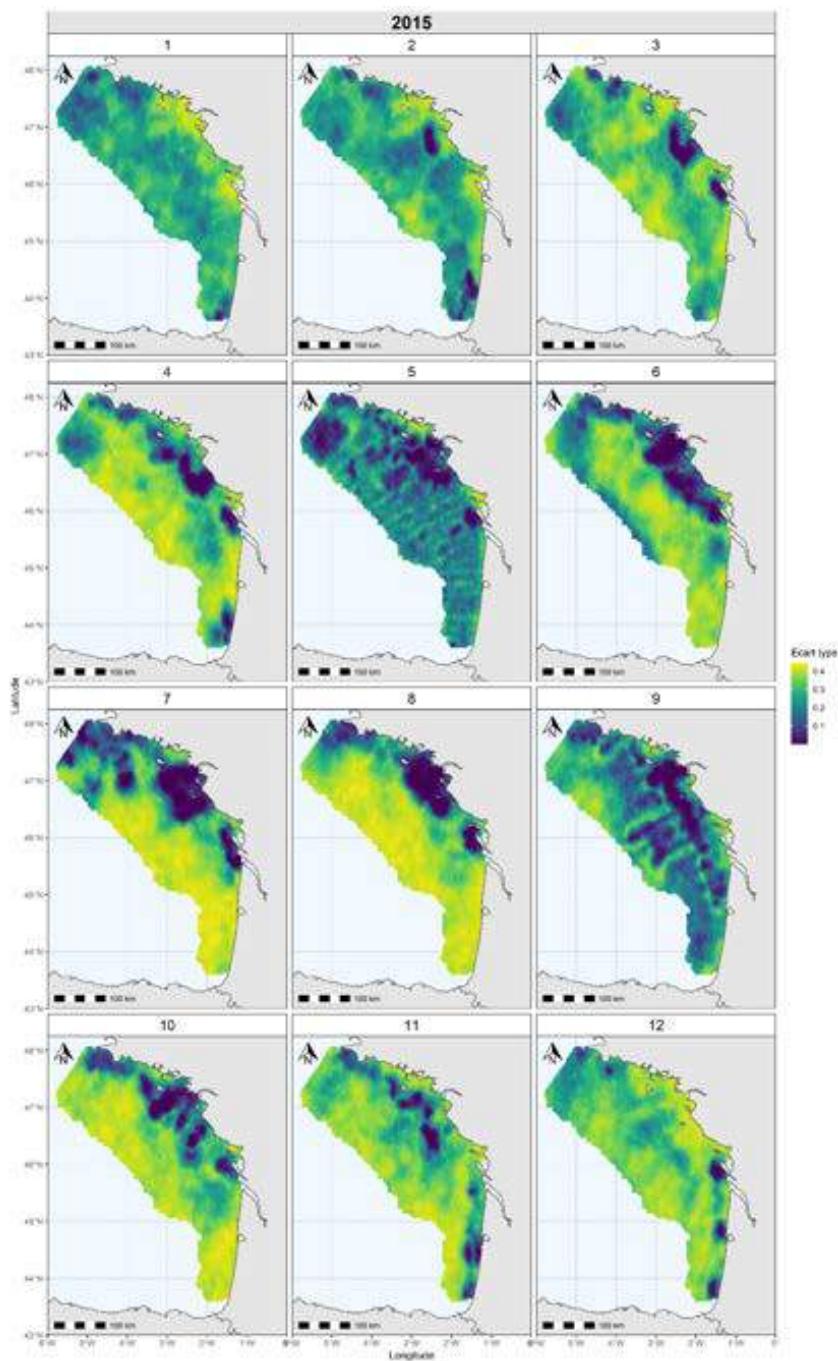
**Annexe 30 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2014.



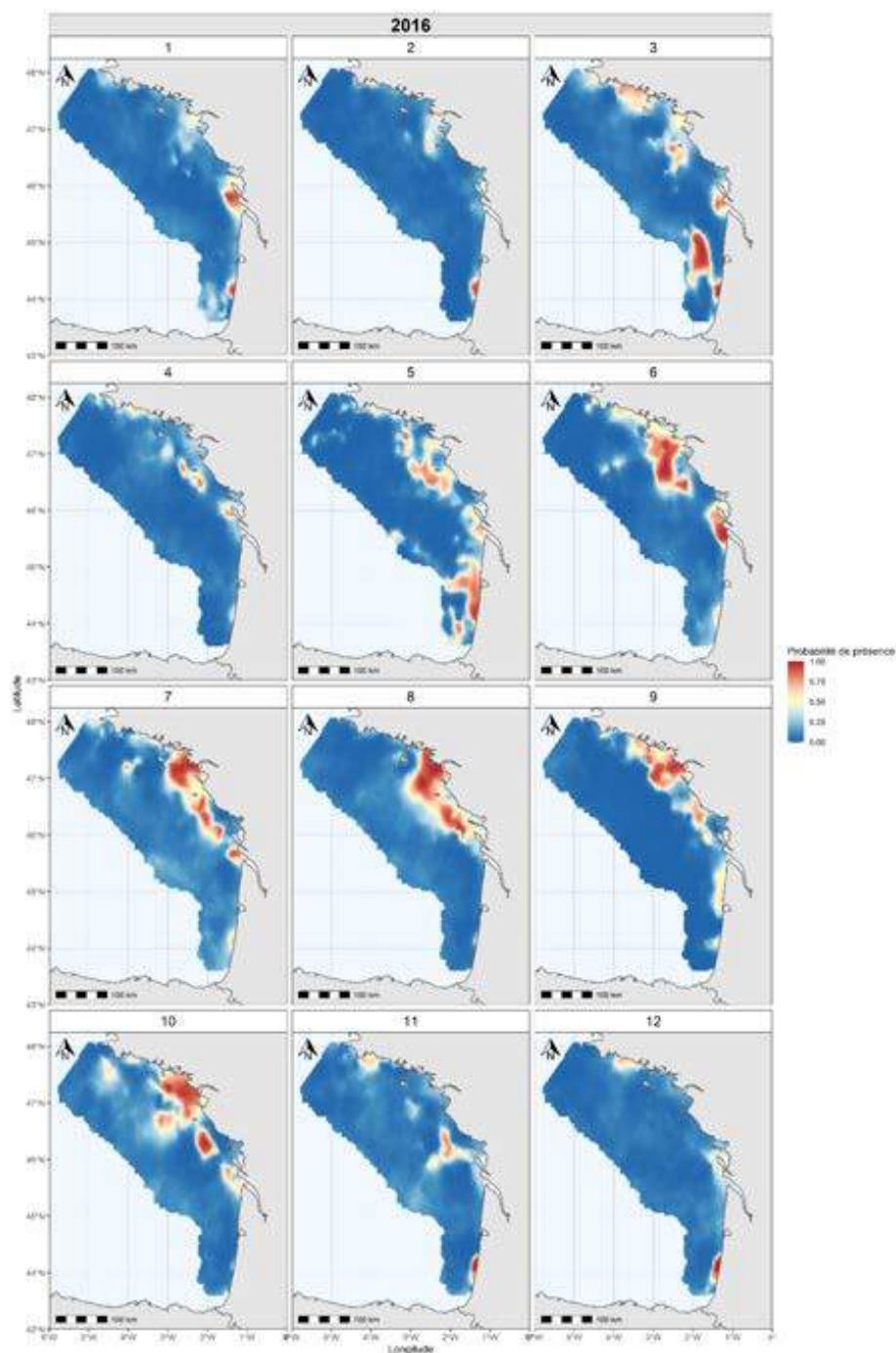
**Annexe 31 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2015.



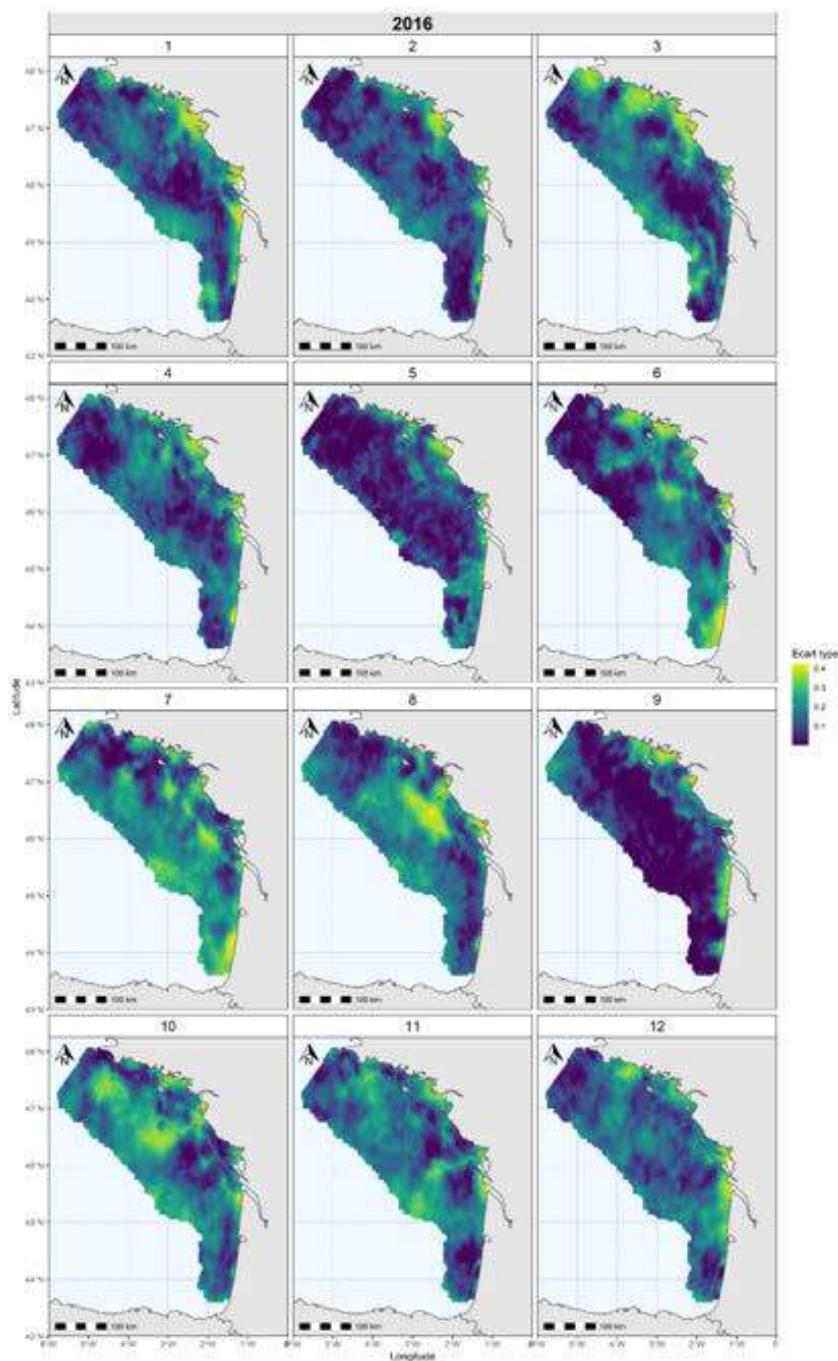
**Annexe 32 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2015.



**Annexe 33 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2016.

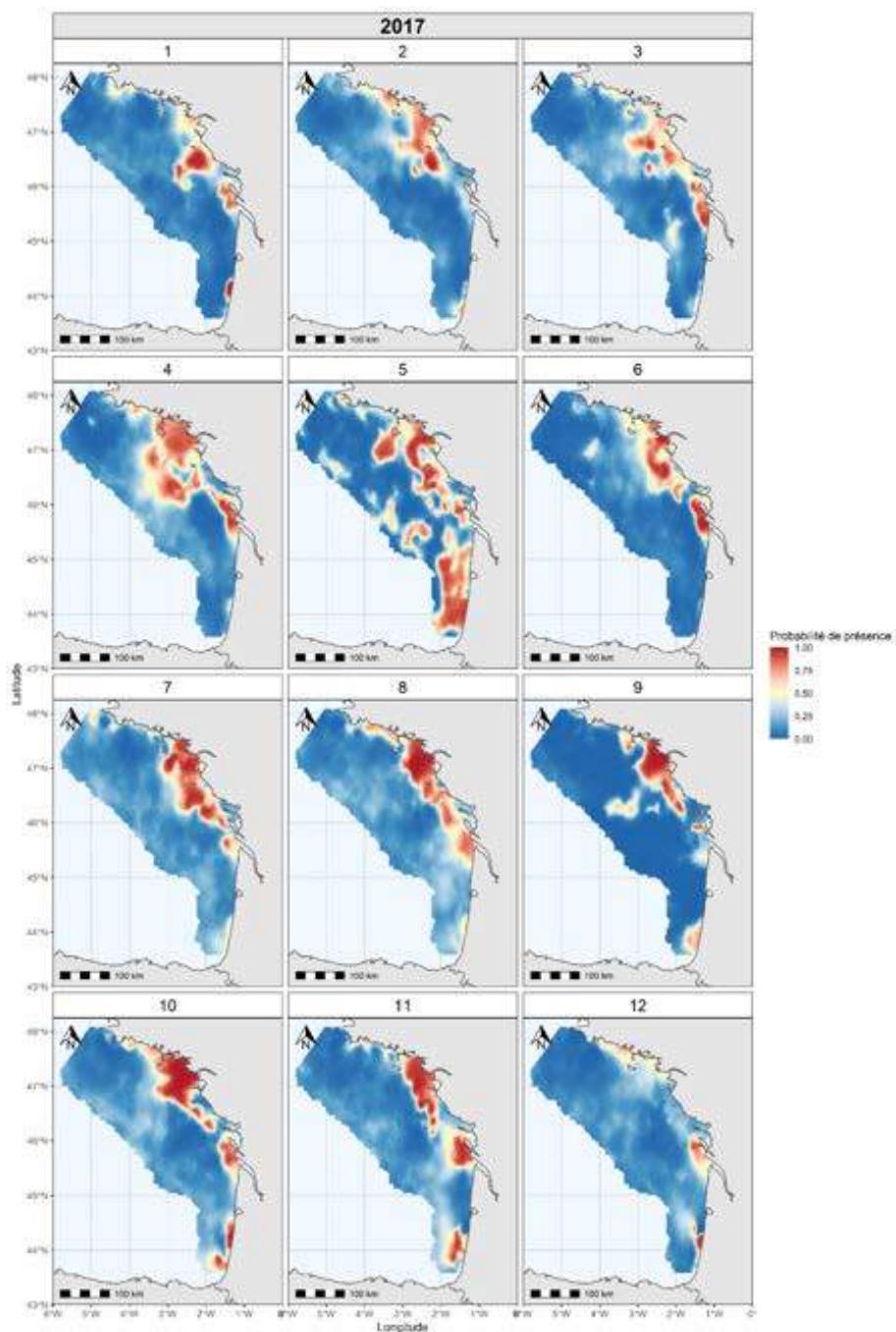


**Annexe 34 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2016.



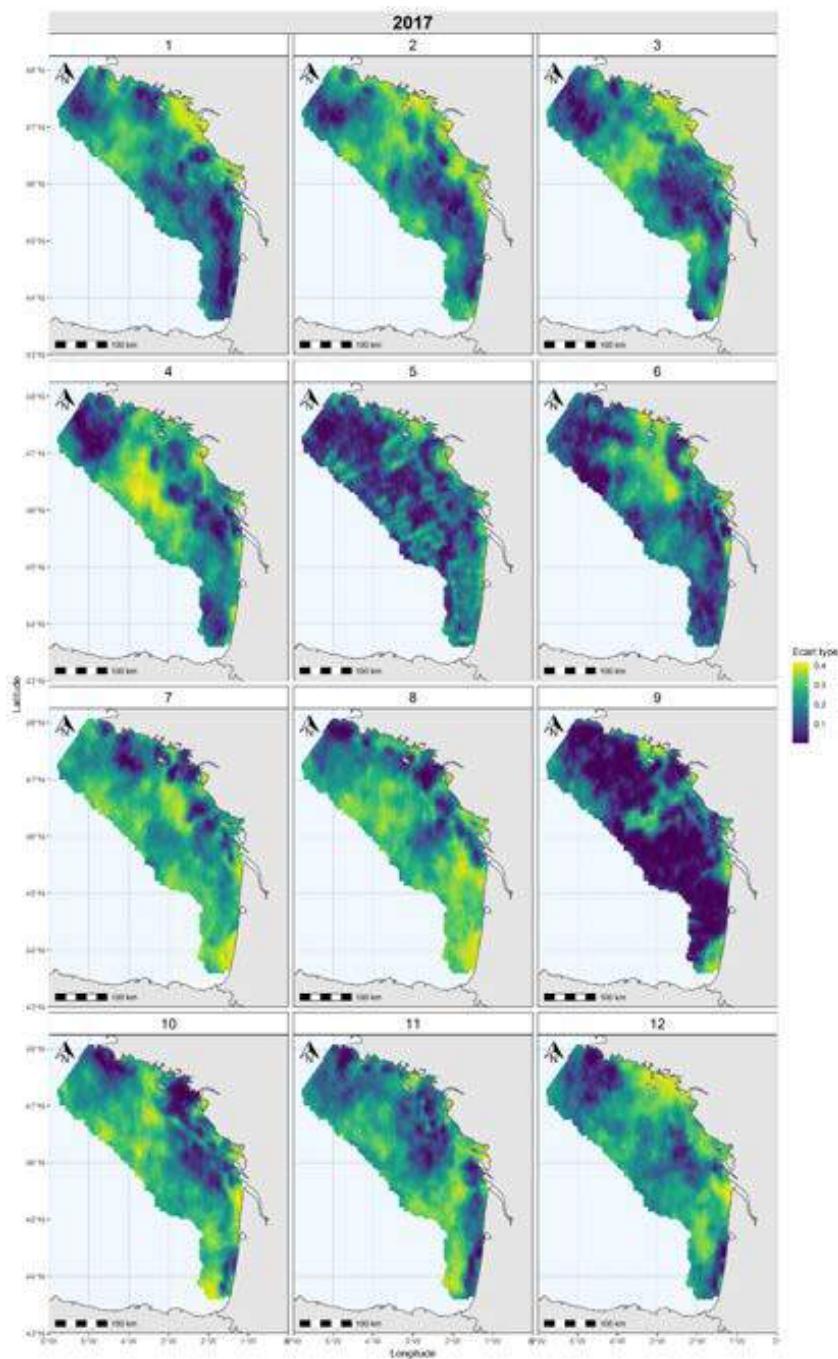
---

**Annexe 35 :** Cartographies de la probabilité de présence de la sardine dans le GdG en 2017.

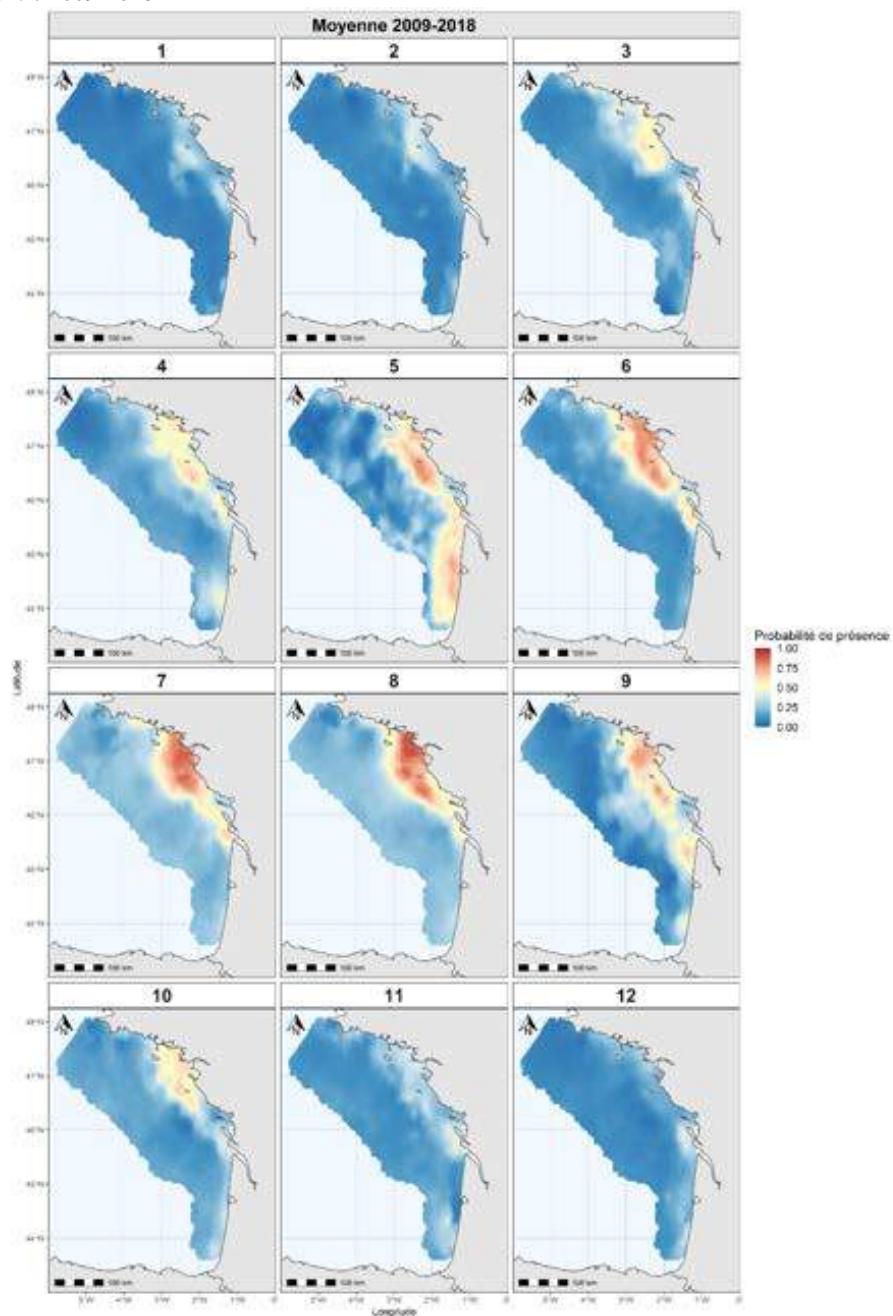


---

**Annexe 36 :** Cartographies des écarts types de la probabilité de présence de la sardine dans le GdG en 2017.

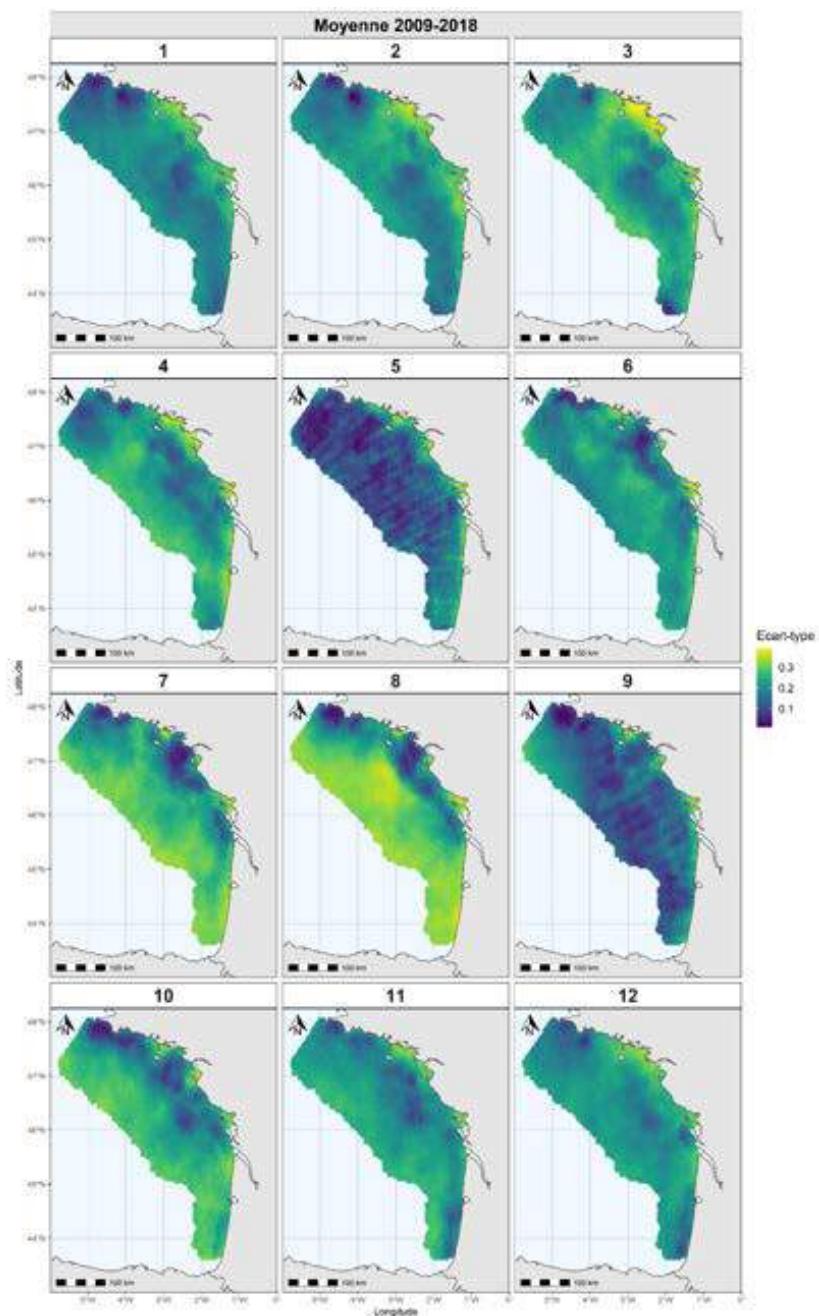


**Annexe 37 :** Cartographies moyennes de la probabilité de présence de la sardine dans le GdG entre 2009-2018

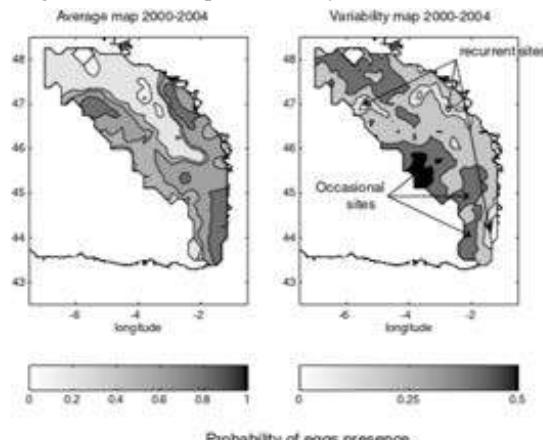


---

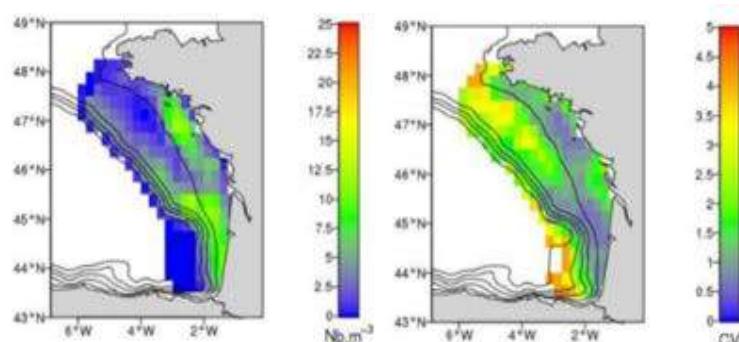
**Annexe 38 :** Cartographies moyennes des écarts types de la probabilité de présence de la sardine dans le GdG entre 2009-2018.



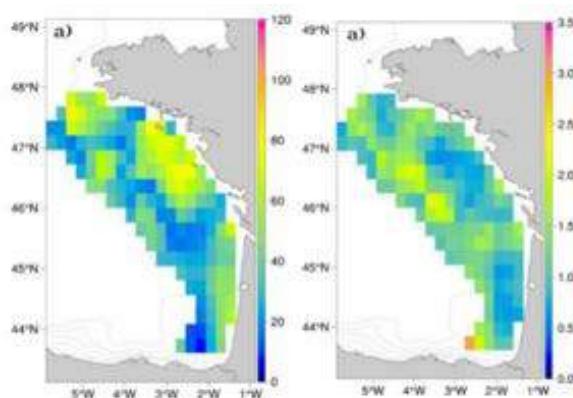
**Annexe 39 :** Cartographies de probabilités de présence d'œufs de sardines entre 2000 et 2004 (Bellier et al, 2007), à gauche cartes de présence moyenne et coefficients de variations à droite.



**Annexe 40 :** Cartographies d'abondance d'œufs de sardines (2000-2016) (Huret et al, 2018), à gauche cartes de d'abondances moyennes et coefficients de variations à droite.

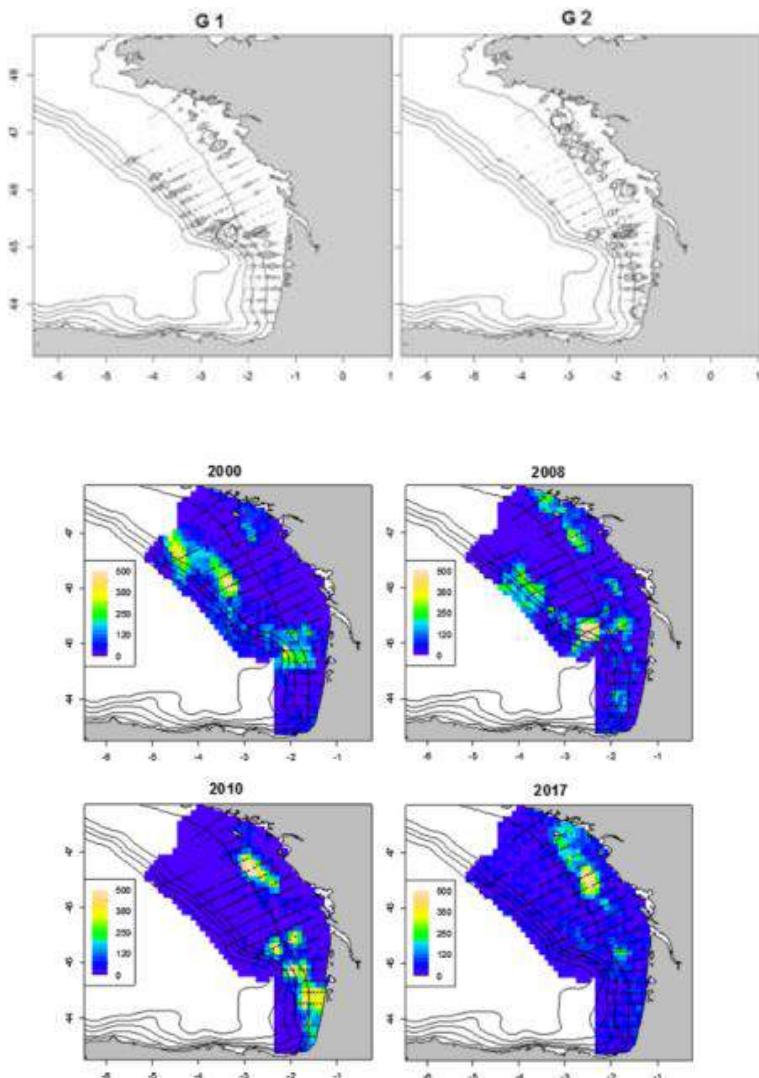


**Annexe 41 :** Cartographies de biomasse de sardines (2000-2015) (Doray et al, 2017a), à gauche cartes de présence moyenne et coefficients de variations à droite.



---

**Annexe 42 :** Distribution moyenne des abondances d'oeufs de sardines réparties en 2 groupes à partir d'un cluster sur 13 MAFS et exemples de 4 cas représentatifs des 2 groupes (G1 au-dessus : 2000 et 2008 ; G2 en-dessous : 2010 et 2017) après krigeage à partir du modèle MAF (Données PELGAS 2000- 2017) (Petitgas *et al.*, 2020).



 Diplôme : Ingénieur Agronome Spécialité : Sciences Halieutiques et Aquacoles (SHA) Spécialisation / option : Ressources et Ecosystèmes Aquatiques (REA) Enseignant référent : Olivier Le Pape	
Auteur(s) : Florian Quemper Date de naissance* : 11/10/1996 Nb pages : 34 Annexe(s) : 40 Année de soutenance : 2021	Organisme d'accueil : Institut Agro – Agrocampus Ouest Rennes Adresse : 65 rue de st Brieuc 35042 Rennes Maître de stage : Baptiste Alglave
<b>Titre français :</b> Modélisation de la distribution spatiale de la sardine du Golfe de Gascogne ( <i>Sardina pilchardus</i> ) par intégration de données commerciales et scientifiques : enjeux et limites.	
<b>Titre anglais :</b> Modelling the spatial distribution of sardine in the Bay of Biscay ( <i>Sardina pilchardus</i> ) by combining commercial and scientific data: issues and limits.	
<b>Résumé :</b> En vue d'assurer une gestion durable des ressources marines, une connaissance fine de la dynamique spatio-temporelle des espèces exploitées est nécessaire. Les campagnes scientifiques fournissent des données standardisées afin de nourrir les modèles de distribution d'espèce. Néanmoins, l'information apportée par ces données reste limitée puisque l'échantillonnage est restreint dans le temps et dans l'espace. Les données de débarquements (logbooks) couplées aux données de position des navires de pêche (VMS) constituent une source d'information additionnelle qui peut être mobilisée pour cartographier la distribution des espèces à une résolution spatio-temporelle fine. Le développement d'un modèle spatio-temporel combinant les deux sources de données et prenant en compte le comportement de ciblage des pêcheurs a déjà permis d'inférer la distribution de plusieurs espèces benthico-démersales du golfe de Gascogne (GdG - e.g. la sole, la baudroie, le merlan). Cette étude vise à évaluer l'intérêt de l'approche intégrée dans le cas d'une espèce pélagique (la sardine du GdG), dont l'écologie, la dynamique de flottille et la nature des données de campagne diffèrent des cas étudiés jusqu'à maintenant. Nous présentons les défis méthodologiques propres à l'application de l'approche intégrée au cas de la sardine du GdG - flottilles côtières, comportement de ciblage important, saisonnalité des pêches -, et le modèle utilisé pour combiner les différentes sources de données et cartographier la sardine. Sur les mois de campagne, les cartes sont essentiellement drivées par les données scientifiques. En dehors de ces mois, les données commerciales apportent de l'information dans le rayon d'action des flottilles. Toutefois, une grande partie de la zone d'étude n'est pas recouverte par les flottilles commerciales et les cartes obtenues fournissent une image partielle de la distribution de la sardine.	
<b>Abstract :</b> In order to ensure sustainable management of marine resources, detailed knowledge of the spatio-temporal dynamics of exploited species is necessary. Scientific surveys provide standardized data to feed species distribution models. However, the information provided by these data remains limited since sampling is restricted in time and space. Landings data (logbooks) coupled with vessel monitoring position data (VMS) provide an additional source of information that can be mobilized to map species distribution at a fine spatio-temporal resolution. The development of a spatio-temporal model combining the two data sources and taking into account the targeting behaviour of fishermen has already made it possible to infer the distribution of several benthic-demersal species in the Bay of Biscay (BoB - e.g. sole, anglerfish, whiting). This study aims to evaluate the interest of the integrated approach in the case of a pelagic species (the sardines in the BoB), whose ecology, fleet dynamics and the nature of the survey data are very different from the cases studied so far. We present the methodological challenges of applying the integrated approach to the sardines in the BoB - coastal fleets, strong targeting behaviour, seasonality of fisheries - and the model used to combine the different data sources and map the sardine. When the surveys occurs (May and September), the maps are essentially driven by scientific data and allow the identification of distribution patterns consistent with the literature. Outside these months, commercial data provide information within the range of the fleets. However, the commercial fleets do not cover a large part of the study area and the maps obtained provide a partial representation of sardines distribution.	
<b>Mots-clés :</b> <i>Sardina pilchardus</i> , Golfe de Gascogne, données VMS, données logbooks, PELGAS, JUVENA, modélisation spatiale et spatio-temporelle, INLA.	
<b>Key Words :</b> <i>Sardina pilchardus</i> , Bay of Biscay, VMS, logbook, PELGAS, JUVENA, spatial and temporal modelling, INLA.	

\* Élément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires

---

**E.3 Report of the ad-hoc contract for the preparation of STECF EWG 22-01 concerning closure areas to protect juveniles and spawners of all demersal stocks in western Mediterranean Sea**

---

**ad-hoc contract for the preparation of STECF  
EWG 22-01 concerning closure areas to protect  
juveniles and spawners of all demersal stocks  
in western Mediterranean Sea**

Baptiste Alglave, Youen Vermand  
22/02/2022

---

## **1 1 ToR**

1. Collate and analyse fisheries dependent data to identify the spatio-temporal distribution of juveniles and spawners of all demersal stocks (to the extent possible following the order presented in the Background), in EU waters of the Western Mediterranean Sea (GSAs 1-2-5-6-7-8-9-10-11).
2. After the above collating and identifying of hake spawning aggregations, the contractor shall draft a short report (20-page max) including detailed maps and GIS layers 2 to feed into the work of EWG 22-01 to develop an advice on the efficiency of the existing closure areas and the development of additional closure areas to protect juveniles and spawners of demersal stocks in the region.

---

## 2 2 data call

raw VMS data (reference years: Complete available time series up to 2020, including vessel id and date of the VMS emission), VMS data will be linked to logbook data. It will allow computing CPUE (kg of landings per fishing time) at fine spatial scale. Fishing operation should be identified from VMS data by member states in order to take into account national specificities in fishing patterns. If fishing operation are not available, fishing operation will be identified based on common speed threshold.

logbook data (reference years: The same time series requested for VMS data up to 2020): logbook data should contain identifiers (vessel number and date) to be able to link logbook and VMS data to compute CPUE. Métier (levels 4, 5 and 6) should be specified to identify the different fleets and fishing behaviours within the data. When available, landings should be reported separating mature/juvenile fish. If not available landings should be reported by commercial category and an estimate of the fraction of juvenile/mature fish by commercial category provided independently. If the proportion of juvenile/mature fish by commercial category is not available an average proportion will be applied by commercial category based on available data. Fine scale georeference CPUE including information of mature/juvenile fish will be used to infer the latent biomass of respectively juvenile/mature fish.

---

### 3 3 Context

#### 3.1 3.1 Background provided by the Commission

In adopting the western Mediterranean multi-annual management plan (West Med MAP), Member States agreed to implement several management measures, such as fishing effort reduction, closure areas and maximum catch limits, to secure the achievement of MSY by 1 January 2025 for all demersal stocks in the western Mediterranean. The work of the STECF expert working group will continue building on the previous evaluations by STECF expert working groups to look into (i) the implementation of maximum catch limits for deep-water shrimps and hake as well as (ii) the delineation of additional closure areas. Regarding closure areas, Article 11.1, alternatively Article 11.2, aims at protecting juveniles of European hake. All three concerned Member States adopted Article 11.3 and agreed to establish additional closure areas by 17 July 2021 and on the basis of best available scientific advice, where there is evidence of a high concentration of juvenile fish, below the minimum conservation reference size, and of spawning grounds of demersal stocks, in particular for the target stocks of the West Med MAP. In addition, France and Spain adopted in December 2020 targets of capture reductions of demersal stocks and committed to reduce between 15% and 25% the capture of juveniles and spawners in each GSA.

In order to implement closure areas to protect juveniles and spawners of these demersal stocks, knowledge on their spatial distribution is needed. Information on their spatial distribution is usually available through MEDITS surveys but is then limited in time. Spatial monitoring of commercial data (based on logbooks crossed with Vessel Monitoring Systems) can provide an additional extensive data source to inform fish spatial distribution. These data have however hardly been used because considered biased by fishermen behaviour.

(Alglave et al. 2022) developed a spatial hierarchical framework integrating both scientific and commercial data sources while accounting for preferential sampling (PS) of commercial data in order to map fish spatio-temporal distribution. The model was adapted to inform on spatio-temporal distribution of either: - Hake Juveniles and adults - Hake biomass - Red mullet biomass - blue and red shrimps based on data availability.

---

## 4 4 Data

### 4.1 4.1 Logbook data

Raw logbook data were provided anonymized.

#### 4.1.1 4.1.1 Merluccius merluccius (HKE)

Information on commercial landings were provided for 2015-2020 for France, 2017-2020 for Italy and 2018-2020 for Spain. Only France and Spain provided information by Commercial Categories. All landings could be allocated to different commercial categories for Spanish landings but a small fraction of the French landings were not allocated to commercial category.

Fig.4.1 and Fig.4.2 present the total French and Spanish landings and their share between commercial categories. Fig.4.3 presents total Italian landings.

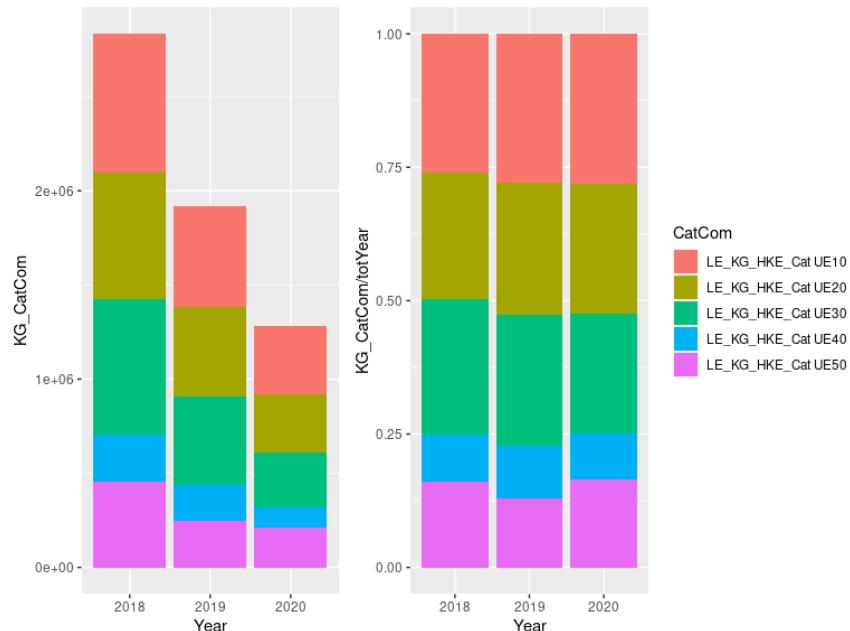


Figure 4.1: Spanish landings by commercial categories (left pannel: in kg, right pannel: in proportion of the total landings)

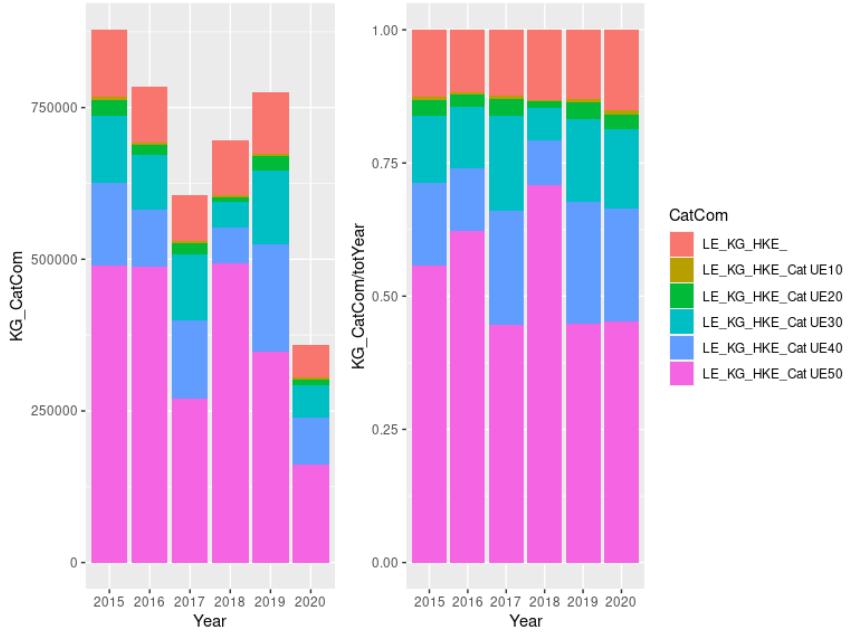


Figure 4.2: French landings by commercial categories (left pannel: in kg, right pannel: in proportion of the total landings)

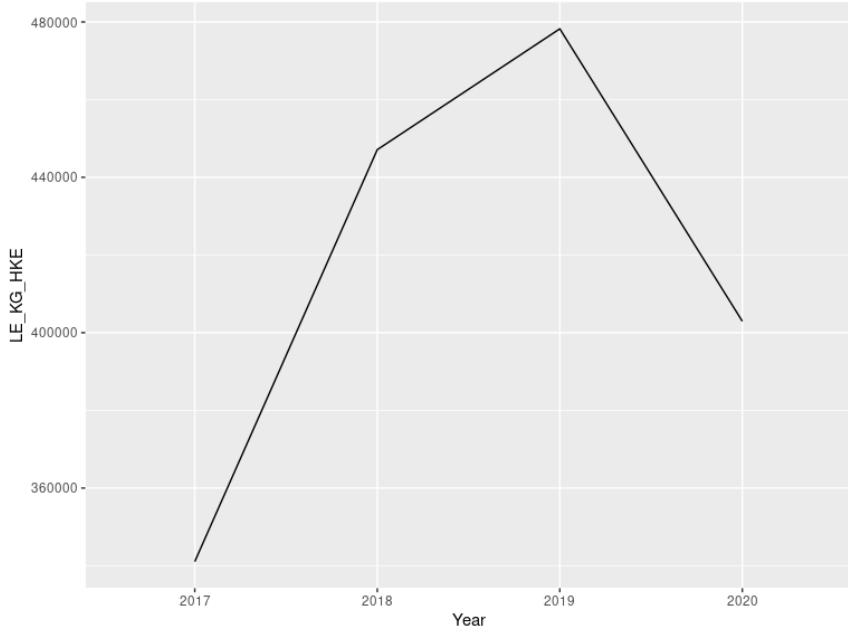


Figure 4.3: Italian landings in kg

Previous work (Billet et al. 2021), derived proportion of juvenile and adult by commercial category based on commercial sampling and a size of 29 cm to differentiate juveniles and adults (<http://www.fao.org/gfcm/data/safs>). (Billet et al. 2021) then computed the proportion of juveniles and adults in weight per commercial category for the demersal trawl fleet operating in GSA 7. These proportions were the only available data to compute juvenile/adult fraction for

landings data. They were used to divide landings in juveniles and adults of French and Spanish landings. Tab.4.1 presents the proportion of juveniles per commercial categories used to derive juveniles and adults landings from total French and Spanish landings per commercial categories.

Table 4.1: Observed juveniles percentages (in weight) per commercial category and year for demersal trawlers in GSA 7

<b>Species</b>	<b>Com- mer- cial.C- at- egory</b>	<b>Y2015</b>	<b>Y2016</b>	<b>Y2017</b>	<b>Y2018</b>	<b>Y2019</b>	<b>Y2020</b>
Merlu-	UE10	0.0	0.0	0.0	0.0	0.0	0.0
cius merluc-	cius						
Merlu-	UE20	5.2	0.0	0.0	1.1	0.0	1.5
cius merluc-	cius						
Merlu-	UE30	13.0	13.8	7.6	15.8	5.1	4.7
cius merluc-	cius						
Merlu-	UE40	17.6	20.0	16.7	29.6	7.7	26.2
cius merluc-	cius						
Merlu-	UE50	89.9	89.6	92.6	88.5	77.7	90.6
cius merluc-	cius						

#### **4.1.2 Aristeus antennatus (ARA) and Aristaeomorpha foliacea (ARS)**

For these two shrimps, no landing by commercial was available. The total landings were used as a proxy of the total shrimp biomass. Data from Italy were used.

Fig.4.4 presents the total Italian landings for these two different shrimps.

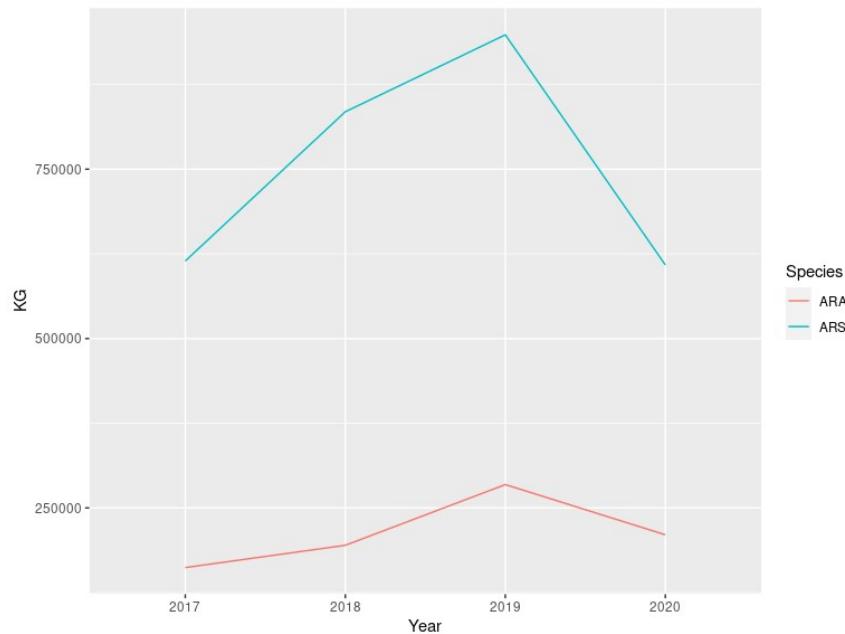


Figure 4.4: Italian landings in kg

#### 4.1.3 Mullus barbatus (MUT)

(Billet et al. 2021) shown that no juvenile was caught by demersal trawlers. Even if landings were provided by commercial categories for France, total landings were used as a proxy of the total biomass.

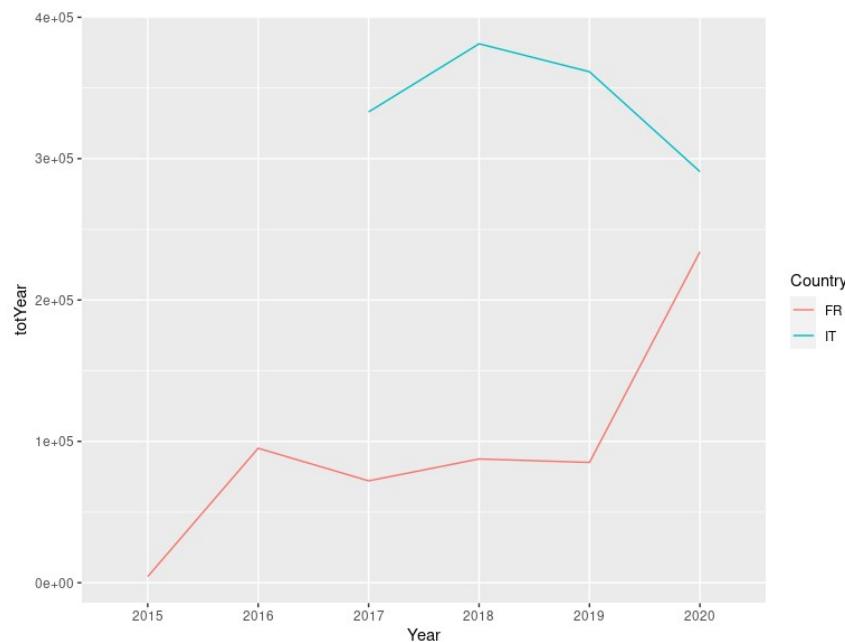


Figure 4.5: red mullet landings in kg

---

Fig.4.5 presents the total Italian and French red mullet landings. It should be noted that some mismatch in the French landings were observed with lower landings in the data used than in the data reported in other reports and used for assessment. However, as in the final model, only LPUE are used and not total landings, the impact might be mitigated. Some improvement in the input data might be necessary in the future.

## 4.2 VMS data

VMS data were provided with the same anonymization code and on the same historical series for the 3 countries. Each country defined the pings identified as “fishing” according to their own algorithm.

Using VMSTools package (Hintzen et al. 2012), logbooks and VMS data were merged to get a better spatialisation of effort and landing data. Landings and effort were used to compute Landings Per Unit of Effort (LPUE).

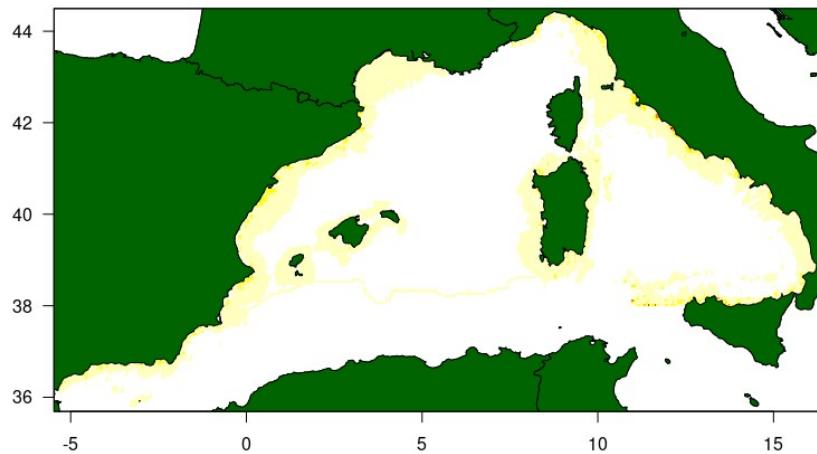


Figure 4.6: Sum of effort for France, Italy and Spain over the years 2019 and 2020

---

## 5 5 Method

Based on the methodology described in (Alglave et al. 2022), Species Distribution Models (SDM) were developed for *Merluccius merluccius*, *Aristeus antennatus* and *Aristaeomorpha foliacea* and *Mullus barbatus*. LPUEs for the different species are computed based on “spatialised” landings and effort data (commercial data and survey data). Both scientific and commercial observations (LPUE (in weights / unit effort)) are considered as proportional to the underlying biomass through a zero-inflated observation process. The model allows to interpolate biomass predictions on the entire area of study based on punctual observations of biomass. No covariates were included in the model as VMS-logbook data do not provide reliable estimates for the species-habitat relationship and only allow to identify spatial and spatio-temporal correlation structures (i.e. areas with relatively higher biomass - Alglave et al., in prep). Here we make the strong hypothesis that discarding behavior does not change the biomass perception due to spatial specific discarding events. As described in (Alglave et al. 2022), when commercial data far exceed scientific data, the latter bring little information to spatial predictions in the areas sampled by commercial data and integrating scientific data in inference will not modify the maps.

Then, when building the SDM the MEDITS data were not used to feed the model except in the case where both French and Italian data had to be combined (map of total biomass for *Mullus barbatus* and *Merluccius merluccius*). In these cases, there is no overlap between the French and the Italian fleets, and then MEDITS data allow to standardize country specific catchabilities by assuming a constant catchability for the MEDITS survey.

The model provides an estimate of the underlying biomass at each cell of the discretization grid at a monthly time step. Biomass are expressed in units of LPUE (kg/hr) except for the model where MEDITS data were used to intercalibrate (biomass of *Mullus barbatus* and *Merluccius merluccius*) where biomass are in the unit of the survey (kg/km<sup>2</sup>).

### 5.1 5.1 Maps for Juvenile and Adult fraction

For *Merluccius merluccius* and for France and Spain, it was possible to divide landings in Juvenile and Adult fraction using the commercial categories and commercial sampling. Models were then built based on Juveniles and Adults fraction in the areas covered by French and Spanish fleets.

### 5.2 5.2 Maps of total biomass

When it was not possible to share the landings on juvenile and adult fraction, landings were considered representative of the total biomass.

### 5.3 5.3 Plotting the maps

To plot the overall pattern distribution of the species, we averaged biomass distribution over the whole period (average pattern) and by quarter (quarterly patterns). To extract the areas with higher biomass, we plotted the 90% quantile as a threshold for hotspot identification (red points in the following figures). The model theoretically allows to estimate biomass on the whole domain even if no observation were made. However, outside the range of the fleets predictions were not considered to be reliable, only the predictions inside the spatial extent of the fleet are plotted. And maps were produced reducing geographical contours to the areas where effort was observed at least once on the full time series.

---

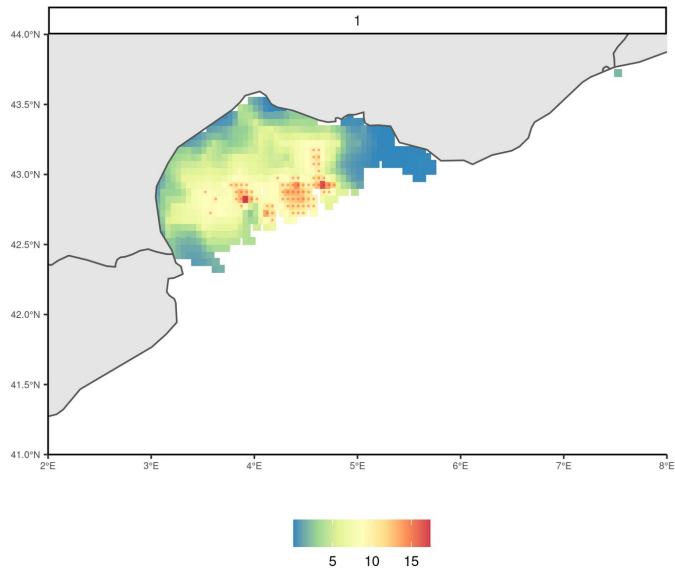
## 6 6 Results

### 6.1 6.1 **Merluccius merluccius**

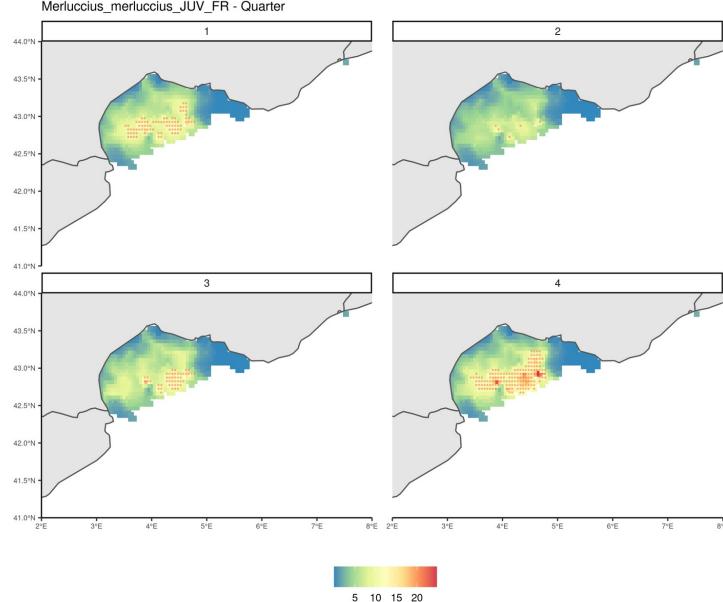
#### 6.1.1 6.1.1 **Juvenile/Adult distribution based on French landings**

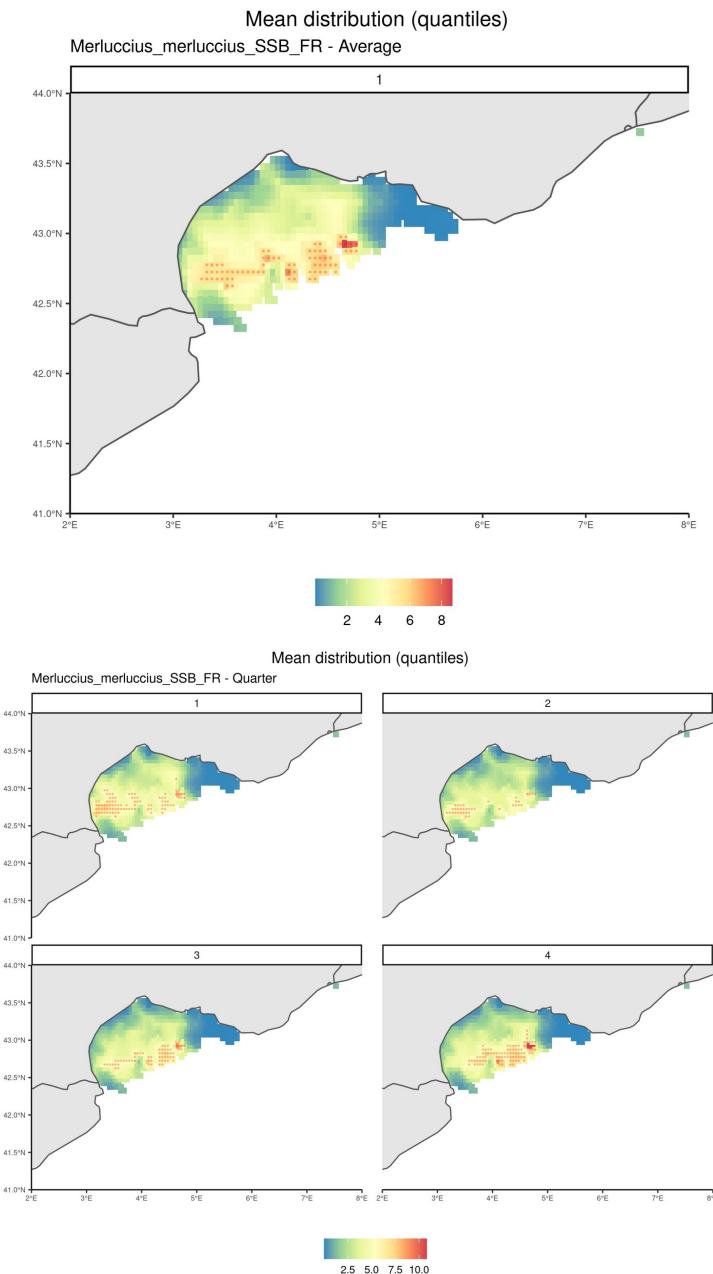
French landings from demersal trawls were divided into juveniles (JUV in subtitle) and adult (SSB in subtitle) landings based on the method described above. Estimated biomass distribution is presented in the following figures. Each time the yearly pattern and quarterly patterns are presented. Squares that are in the quantile 90% are overlaid with red points.

Mean distribution (quantiles)  
Merluccius\_merluccius\_JUV\_FR - Average



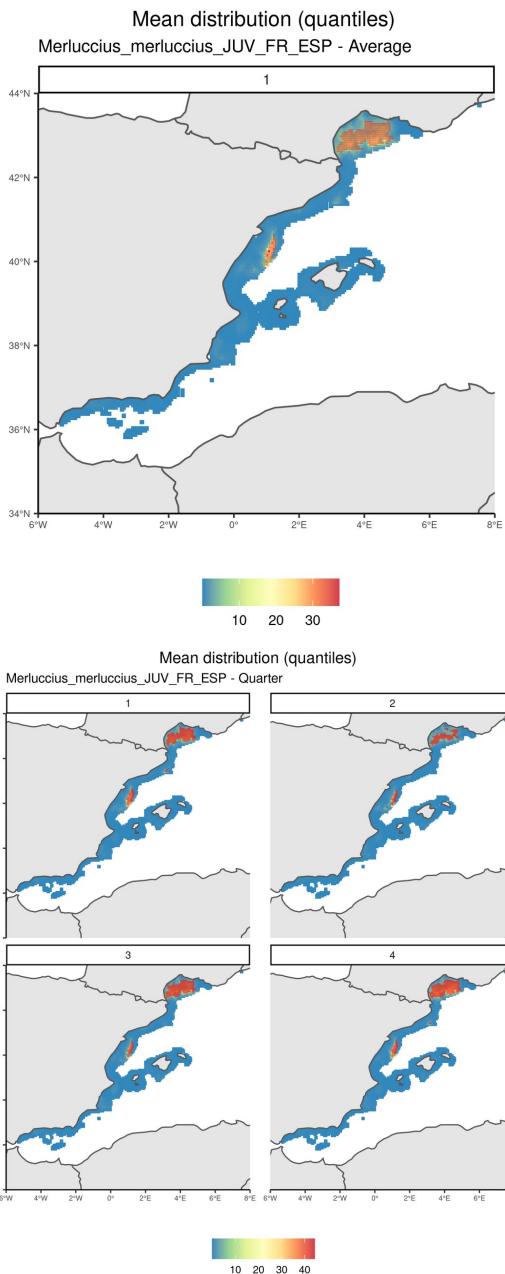
Mean distribution (quantiles)  
Merluccius\_merluccius\_JUV\_FR - Quarter

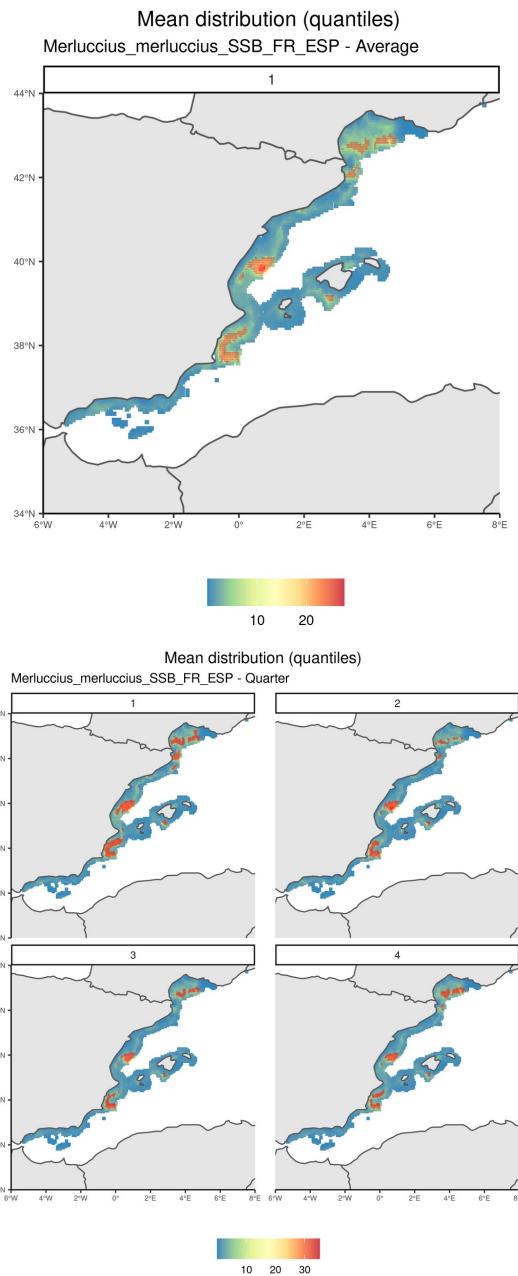




### 6.1.2 Juvenile/Adult distribution based on French and Spanish landings

French and Spanish landings from demersal trawls were divided into juveniles (JUV in subtitle) and adult (SSB in subtitle) landings based on the method described above. Estimated biomass distribution is presented in the following figures. Each time the yearly pattern and quarterly patterns are presented. Squares that are in the quantile 90% are overlaid with red points.





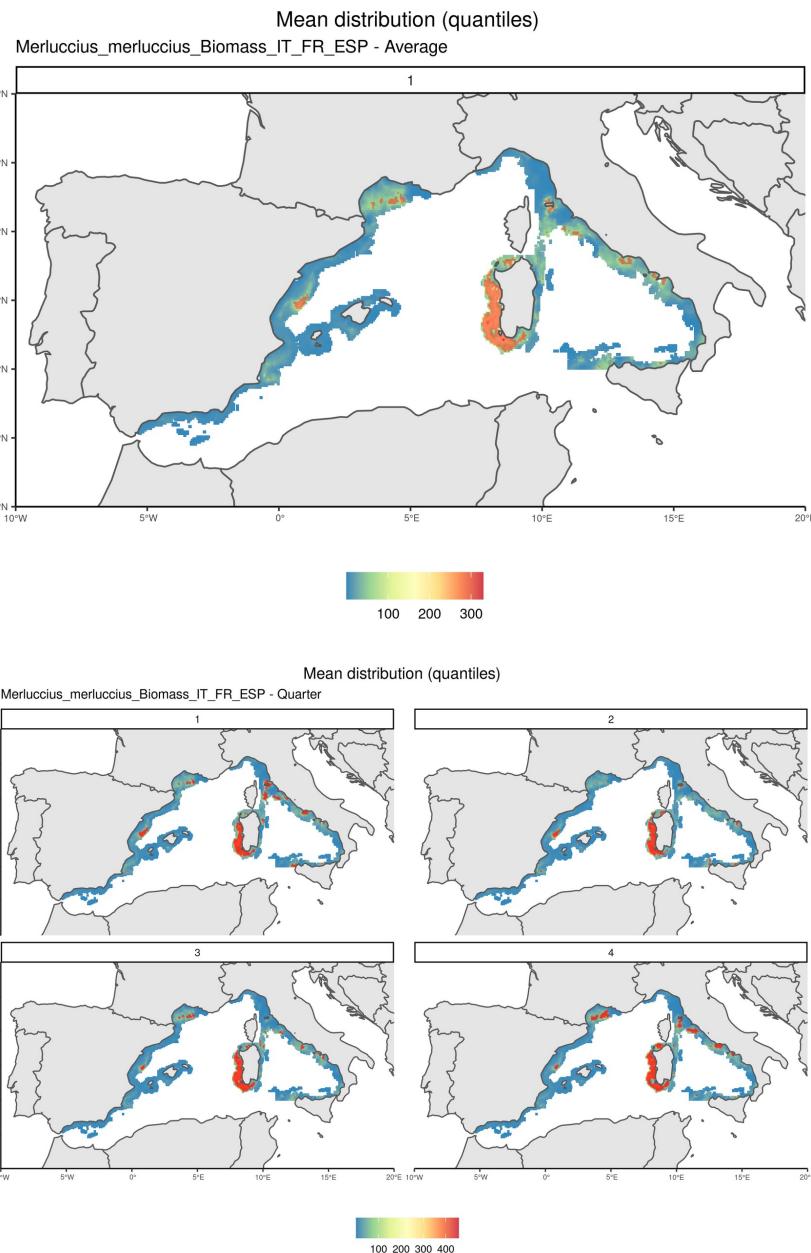
### 6.1.3 Biomass distribution based on French Italian and Spanish landings

Italian, French and Spanish landings from demersal trawls were merged to estimate 'total biomass.' MEDITS data were used to standardize country/gear catchability.

Estimated biomass distribution is presented in the following figures. Each time the yearly pattern and quarterly patterns are presented.

---

Squares that are in the quantile 90% are overlaid with red points.



## 6.2 6.2 Aristeus antennatus

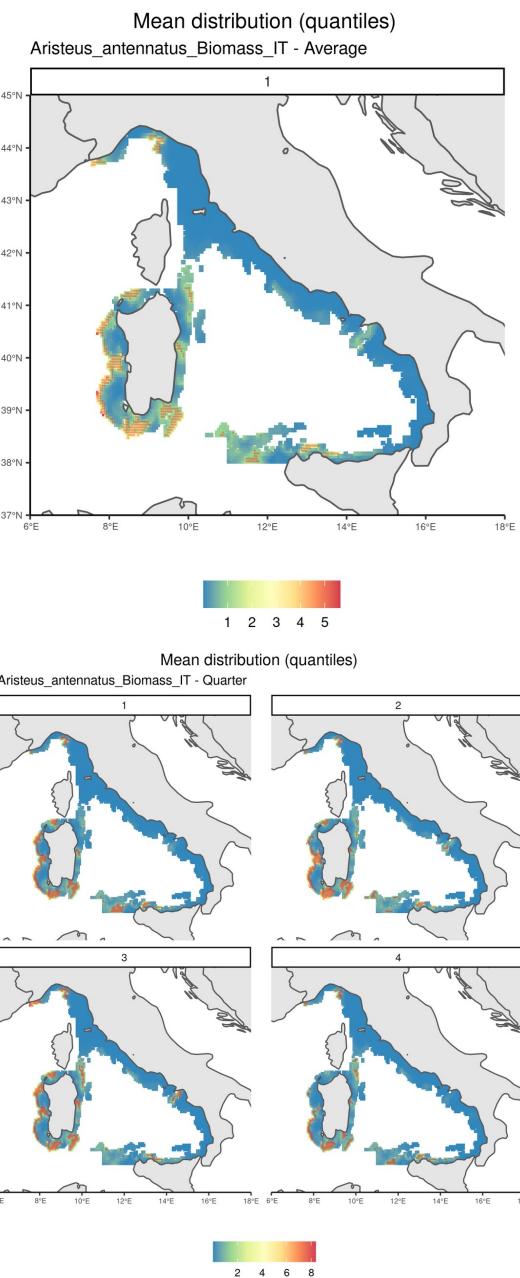
### 6.2.1 6.2.1 Biomass distribution based on Italian landings

Italian landings from demersal trawls were merged to estimate 'total biomass.'

---

Estimated biomass distribution is presented in the following figures. Each time the yearly pattern and quarterly patterns are presented.

Squares that are in the quantile 90% are overlaid with red points.



---

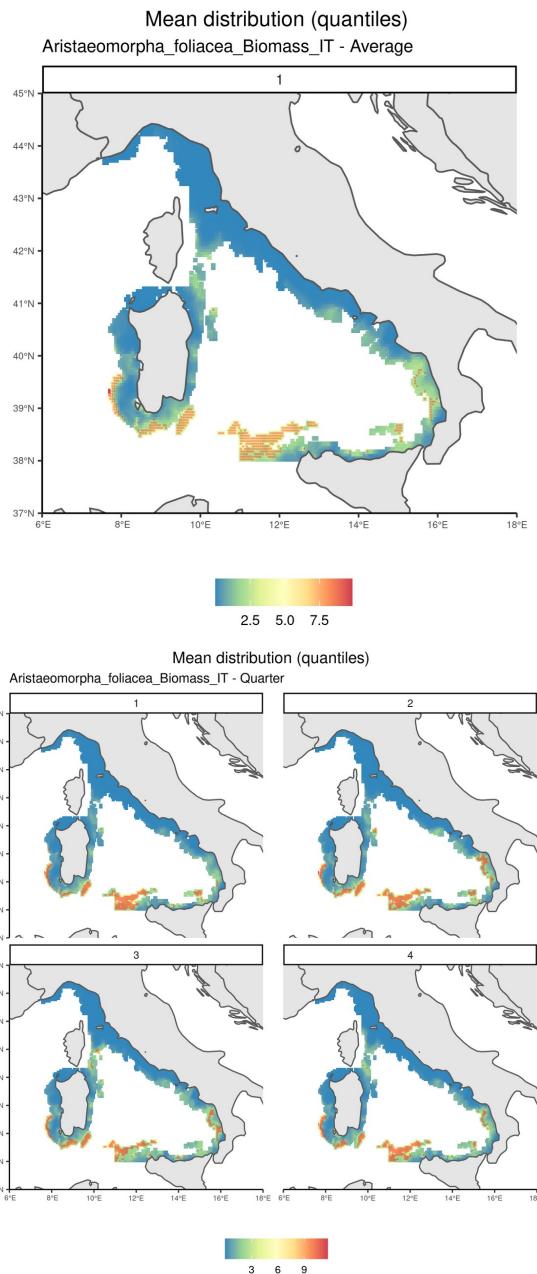
## **6.3      6.3      *Aristaeomorpha foliacea***

### **6.3.1    6.3.1    Biomass distribution based on Italian landings**

Italian landings from demersal trawls were merged to estimate ‘total biomass.’

Estimated biomass distribution is presented in the following figures. Each time the yearly pattern and quarterly patterns are presented.

Squares that are in the quantile 90% are overlaid with red points.



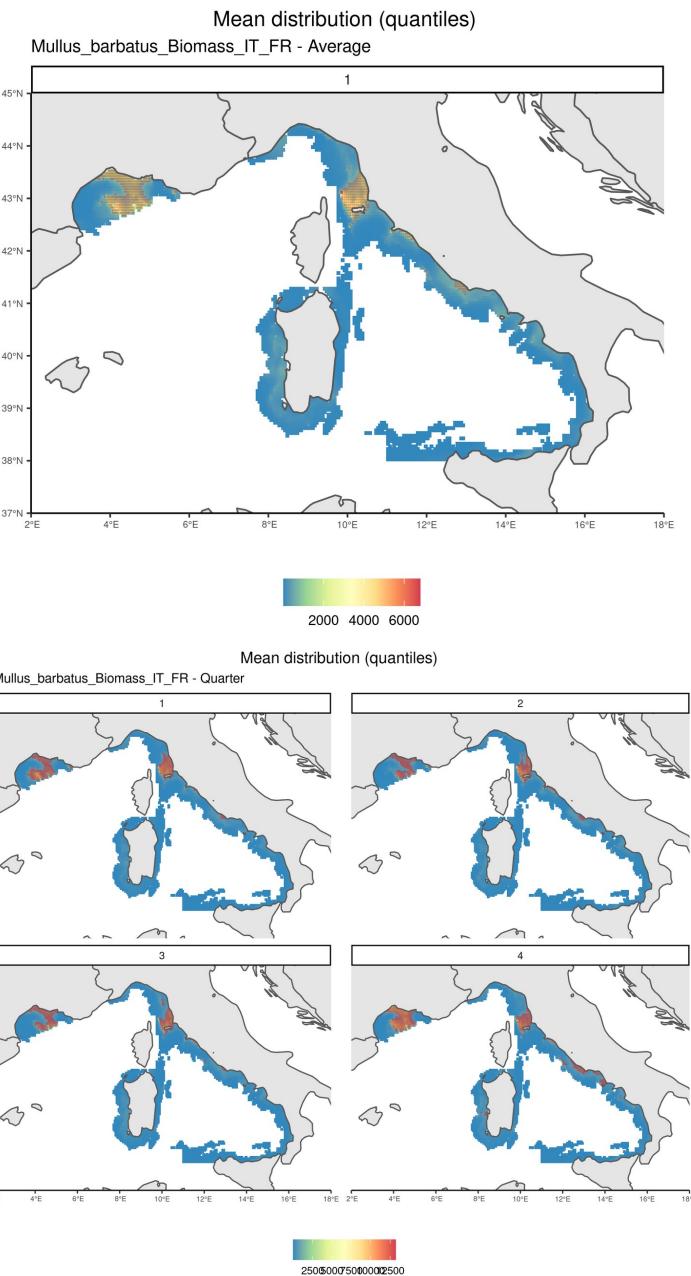
## 6.4 6.4 *Mullus barbatus*

### 6.4.1 6.4.1 Biomass distribution based on French and Italian landings

Italian, French and Spanish landings from demersal trawls were merged to estimate 'total biomass.' MEDITS data were used to standardize country/gear catchability.

Estimated biomass distribution is presented in the following figures. Each time the yearly pattern and quarterly patterns are presented.

Squares that are in the quantile 90% are overlaid with red points.



---

## 7.7 Limitations

- The study is based primarily on the reallocation of landings to GPS positions estimated to be fishing. It is known that fishermen can, within a day, explore a large number of areas resulting in very different catch profiles. These aggregations of different profiles into a single declaration can bias models of species spatial distribution. The maps produced must therefore be critically reviewed by experts familiar with the distribution of these species in the area.
- The study was conducted using bottom trawlers. Bottom trawlers were used because the calculation of LPUEs from VMS/logbook processing is routinely done on these gears, which may be more problematic on passive gears such as nets. However, by filtering on a gear category, part of the space may not have been represented and the distribution of species may not be fully mapped.
- For the distribution between juvenile and adult hake, only the French commercial category structures were available and were applied to the Spanish landings. Data on the Spanish and Italian category structures by commercial category would undoubtedly improve the distribution by stage.

---

## 8 8 Outputs to the group

All maps presented in this report were provided to the group in .Rdata format with a document explaining the different objects and attached to the report.

---

## 9 References

- Alglave, Baptiste, Etienne Rivot, Marie-Pierre Etienne, Mathieu Woillez, James T Thorson, and Youen Vermaud. 2022. "Combining Scientific Survey and Commercial Catch Data to Map Fish Distribution." *Ices Journal of Marine Science*. <https://doi.org/10.1093/icesjms/fsac032>.
- Billet, Norbert, Gregoire Certain, Jerome Bourjea, and Sandrine Vaz. 2021. "Evaluation Des Fermetures Spatio-Temporelles Mises En Oeuvre à Partir Du 1er Janvier 2020 Pour La pêche Au Chalut En Mer méditerranée." Expertises (Expertise).
- Hintzen, Niels T., Francois Bastardie, Doug Beare, Gerjan J. Piet, Clara Ulrich, Nicolas Deporte, Josefina Egekvist, and Henrik Degel. 2012. "VMStools: Open-Source Software for the Processing, Analysis and Visualisation of Fisheries Logbook and VMS Data." *Fisheries Research* 115-116: 31–43. <https://doi.org/10.1016/j.fishres.2011.11.007>.



**Titre :** Inférer la distribution spatio-temporelle des espèces d'intérêt halieutique et identifier leurs habitats essentiels: modéliser l'échantillonnage préférentiel et le changement de support pour intégrer des sources de données hétérogènes.

**Mots clés :** modélisation spatiale et spatio-temporelle, modèle hiérarchique, intégration de données, échantillonnage préférentiel, changement de support, zone fonctionnelle halieutique

**Résumé :** La cartographie de la répartition des espèces En les combinant aux données de géolocalisation des d'intérêts halieutiques et l'identification de leurs zones navires disponibles par le système de surveillance des fonctionnelles est cruciale pour assurer le renouvellement navires de pêche (VMS), les données de déclarations des espèces et pour l'aménagement de l'espace marin. peuvent permettre de compléter l'information apportée par Pour autant, la localisation des habitats essentiels des poissons, et plus particulièrement des frayères, reste incertaine pour de nombreuses espèces exploitées.

Les données de référence pour cartographier la distribution des espèces exploitées et identifier leurs frayères sont issues de campagnes scientifiques qui bénéficient d'un protocole d'échantillonnage standardisé. Ces campagnes ont généralement lieu une ou deux fois par an, elles prélèvent un nombre limité d'échantillons et elles peuvent ne pas correspondre à la période de reproduction des espèces étudiées. Elles sont donc limitées pour identifier les frayères des espèces d'intérêt halieutique.

Par ailleurs, les déclarations de capture des pêcheurs (logbook) fournissent des informations sur l'ensemble de l'année avec une densité d'échantillonnage supérieure à celle des données scientifiques.

Dans cette thèse, nous avons développé un modèle statistique qui permet de combiner les données commerciales et scientifiques pour inférer la distribution des espèces d'intérêt halieutique à une résolution spatio-temporelle fine. Le modèle permet de prendre en compte le comportement de ciblage des pêcheurs (échantillonnage préférentiel) et d'intégrer les données de déclarations qui sont définies à une résolution spatiale grossière pour inférer la distribution des espèces à une résolution fine (changement de support).

Les cartes de la distribution des espèces permettent d'identifier les zones d'agrégation pendant la saison de reproduction. Nous décrivons également les applications potentielles du cadre de modélisation pour l'aménagement de l'espace marin et les extensions qui pourraient être ajoutées à la version actuelle du modèle.

**Title :** Inferring fish spatio-temporal distribution and identifying essential habitats: tackling the challenge of preferential sampling and change of support to integrate heterogeneous data sources

**Keywords :** spatial and spatio-temporal modeling, hierarchical model, data integration, preferential sampling, change of support, fisheries functional zones

**Abstract :** Mapping fish distribution and identifying fish survey data as fishermen landings provide information on essential habitats grounds is key to ensure species renewal and manage the marine space. Information on the location of fish essential habitats and specifically of fish spawning grounds is still lacking for many harvested species.

The reference data to map fish distribution and identify spawning grounds are scientific survey data. These data benefit from a standardized sampling protocol. However, due to their costs, they also generally suffer from a low sampling density in space and time. In particular, they generally occur once or twice a year and they may mismatch fish reproduction.

Commercial declarations combined with Vessel Monitoring System data could prove highly valuable to complement the information brought by scientific

the full year with a much denser sampling density.

In this PhD, we developed an integrated statistical framework that allows to combine commercial and scientific data sources to infer fish distribution in space and time. Our approach accounts for fishermen targeting behavior towards areas of higher biomass (preferential sampling) and allows to infer fine scale species distribution based on spatially aggregated declarations data (change of support). We demonstrate the ability of the framework to produce monthly maps of fish distribution and to identify aggregation areas during reproduction season. We also outline the potential applications of the framework for Marine Spatial Planning and discuss several extensions that could be added to the actual model.

