



**ÉCOLE DES MINES
DE PARIS**



T H È S E

pour obtenir le grade de
Docteur de l'École des Mines de Paris
Spécialité «Géostatistique»

présentée et soutenue publiquement

par

Mireille BOULEAU

le 18 février 2005

**COMBINAISON GÉOSTATISTIQUE DE L'ACOUSTIQUE ET DES
CAPTURES DANS LES CAMPAGNES SCIENTIFIQUES DE PÊCHE
PAR CHALUTAGE**

Directeur de thèse : Nicolas Bez

Jury

Mme Liliane	BEL	Examineur
MM. Nicolas	BEZ	Examineur
Francis	LALOË	Examineur
Pascal	MONESTIEZ	Rapporteur
Pierre	PETITGAS	Rapporteur
Jacques	RIVOIRARD	Président

Remerciements

Je tiens à remercier Nicolas Bez pour avoir dirigé cette thèse. Sa disponibilité et son enthousiasme m'ont permis de réaliser ce travail dans des conditions tout à fait agréables. Il m'a conduit à tirer le meilleur parti de mes capacités par sa perspicacité et son pragmatisme.

Je remercie également tous les membres du centre de géostatistique pour leur accueil. Je remercie en particulier Jacques Rivoirard pour ses nombreuses suggestions et idées, Didier Renard, pour sa patience face à mes lacunes en programmation, ainsi que Chantal de Fouquet, Christian Lajaunie, Christian Lantuejoul, Hans Wackernagel, François Geoffroy et Jean-Paul Chilès pour leur disponibilité. Leurs remarques et conseils durant mes répétitions de thèse m'ont été d'une grande aide. Merci à Nathalie et Isabelle pour leur faculté incroyable à se souvenir de tout ce que j'ai tendance à oublier et pour leur gentillesse.

Un grand merci à Françoise pour son immense ouverture d'esprit, culturelle et humaine, et pour ses nombreux encouragements. Merci également à Sarah et Marta, pour m'avoir accueillie dans leur bureau et pour leur nombreux conseils.

Je remercie mes rapporteurs Pascal Monestiez et Pierre Petitgas pour l'attention et l'intérêt qu'ils ont portés à mon manuscrit et pour la pertinence de leurs remarques. La curiosité et le recul de Pierre m'ont plusieurs fois permis de considérer mes recherches sous un angle nouveau, ce qui était pour moi particulièrement stimulant. Je remercie également Francis Laloë et Liliane Bel pour le temps et l'attention qu'ils ont portés à ce travail.

Merci à tous les membres du projet CATEFA (Dave Reid, Doug Beare, Tony Greig, Jacques Massé, Yves Vérin, Suzanna Deville, Steve Mackinson, Vidar Hjellvik, Olav Rune Godø, ...) pour les enrichissantes discussions qui se sont déroulées lors des groupes de travail, pour leur écoute et leur enthousiasme.

Je remercie tout spécialement Jacques Massé et Yves Vérin sans qui les congrès ne se seraient probablement pas déroulés avec autant de bonne humeur. J'en profite également pour exprimer toute ma reconnaissance à Yves ainsi qu'à l'équipage de la Thalassa pour m'avoir accueillie à deux reprises à bord de leur yacht géant, pour une expérience pour le moins...houleuse, mais dans tous les cas fort sympathique. Merci enfin à Jean-Paul Delpech pour ses relectures, ses conseils et ses encouragements amicaux.

Sans énumérer les noms de toutes les personnes et amis qui m'ont d'une manière ou d'une autre aidée et encouragée durant ces 3 ans, je résumerai par : un grand merci à tous les thésards, stagiaires ou étudiants en général, qui m'ont permis par leurs propres expériences d'éviter bon nombre de déboires ! Cette thèse leur doit sans aucun doute beaucoup.

Je remercie Mathias, pour m'avoir si régulièrement informée de la météo scandinave, Sylvain pour son soutien gastronomique, Mathilde, Julie, Gab, David, Raffi, Simon et Anne, Adam, Marie et les autres, pour toutes les petites discussions de cafè, et dont la bonne humeur communicative doit certainement se retrouver dans ce mémoire.

Enfin, le plus grand des merci à Julien pour m'avoir aidée, conseillée, relue, corrigée, rassurée, soutenue, encouragée... et à qui je compte bien renvoyer la balle.

La meilleure façon de ne pas avancer,
c'est de suivre une idée fixe.
Jacques Prévert.

Table des matières

Glossaire	iii
Introduction	vi
1 Contexte et problématique	1
1.1 Présentation des données	2
1.1.1 Les données acoustiques	2
1.1.2 Les données de capture	5
1.1.3 Plans d'échantillonnage et caractéristiques de chaque chantier	7
1.1.4 Remarques sur les différences entre les deux types de mesures	10
1.1.5 Mise en place d'hypothèses de cohérence dans les données	17
1.2 Statistiques globales	19
1.2.1 Distributions asymétriques	19
1.2.2 Localisation des valeurs fortes	22
1.3 Conclusion et choix méthodologiques	22
2 Quelques modèles bivariés	27
2.1 Introduction	28
2.2 Analyse structurale	32
2.2.1 Sur le sous-ensemble de points en isotopie	33
2.2.2 Sur tous les points	37
2.3 Cokrigeage	39
2.4 Le Modèle à résidu	48
2.4.1 Correction de la variance	49
2.4.2 Vérification des hypothèses et estimation du paramètre	54
2.4.3 Cokrigeage collocalisé	59
2.4.4 Impact du choix du modèle	60
2.4.5 Vérification a posteriori de l'hypothèse d'autokrigeabilité ou comparaison avec une estimation classique de cokrigeage	62
2.5 Le modèle en corrélation intrinsèque	66
2.6 Le modèle à dérive externe	69
2.7 Estimation non-linéaire de l'acoustique	76
2.7.1 Krigeage disjonctif	77
2.7.2 Espérance conditionnelle	81
2.7.3 Comment combiner une estimation non-linéaire de l'acoustique avec la variable de pêche?	81

3 Étude de la verticale	87
3.1 Contexte	88
3.1.1 Liens entre layers	89
3.1.2 Modélisation 3D	89
3.2 Sommes de layers	93
3.2.1 Coefficient de corrélation de Pearson	94
3.2.2 Coefficient de corrélation de Spearman	94
3.3 Estimation d'un support vertical local	97
3.4 Régression de layers	101
3.5 Analyse canonique	104
3.6 Min/Max Autocorrelation Factors	107
3.6.1 Principe des MAF	107
3.6.2 Application	109
3.7 Approche par classification de profils verticaux	115
3.7.1 Principe	115
3.7.2 Modélisation	117
3.8 Synthèse	130
4 Analyse comparative des résultats	133
4.1 Analyse du gain apporté par l'acoustique	134
4.1.1 Cohérences nécessaires à la comparaison de modèles	134
4.1.2 Récapitulation des modèles utilisés	135
4.1.3 Résultats	137
4.1.4 Conclusion : l'acoustique améliore t-elle les estimations d'abondance ?	163
4.2 Effet de l'échantillonnage	165
4.2.1 L'engin de pêche	165
4.2.2 Densité d'échantillonnage	166
4.2.3 Comparaison des plans d'échantillonnage	168
4.2.4 Discussion	170
Conclusion	173
Bibliographie	176
ANNEXES	180
A Problèmes liés au seuillage	181
A.1 Contexte	181
A.2 Seuillage brut	182
A.3 Acoustique sans pélagique	183
A.4 Sélection sur la température	183
B Compléments sur les M.A.F	185

Glossaire

Ce glossaire vise à préciser le sens de certains termes spécifiques tels qu'ils sont employés dans le présent manuscrit. Certaines expressions peuvent avoir usuellement un sens différent ou plus général. Plusieurs acronymes et symboles utilisés dans le texte sont également explicités.

ABONDANCE :

Quantité de poissons présents dans la zone d'étude. Elle est estimée soit par capture, et alors généralement mesurée par ENASC (mais parfois également en kg ou en nombre d'individus), soit par sondage acoustique et alors mesurée en NASC ($m^2/m.n^2$).

ACFM :

Advisory Committee on Fishery Management

ACP :

Analyse en Composantes Principales

BENTHIQUE :

Adjectif qui qualifie les espèces inféodées au fond pour l'accomplissement des fonctions biologiques (nourriture, reproduction). Les espèces benthiques les plus courantes de cette étude sont les espèces à morphologie aplatie comme la plie, la sole et la limande.

BIOMASSE :

Masse totale de poissons présents dans un milieu donné, toutes espèces confondues, mesurée par unité de surface ou de volume.

CAPTURE :

Ensemble des poissons capturés lors d'un coup de chalut. Elle concerne généralement un groupe d'espèces (capture démersale). Elle peut être mesurée par ENASC ($m^2/m.n^2$), en poids (kg) ou en nombre.

CAPTURABLE :

Expression qualifiant la possibilité à un instant donné et pour une espèce spécifique d'être pêchée par le chalut de fond.

CATEFA :

Combining Acoustic and Trawl data for Estimating Fish Abundance (projet européen 2001-2004)

CEFAS :
Centre for the environment fisheries and aquaculture science

CIEM :
Conseil International pour l'Exploration de la Mer

DÉMERSAL :
Adjectif qui qualifie l'ensemble des espèces qui vivent sur le fond où à son voisinage, sans pour autant en être complètement dépendant dans l'accomplissement des fonctions biologiques. La morue, l'églefin et le merlan sont les espèces démersales les plus courantes de cette étude.

dB :
décibel. Les échos reçus par le sondeur acoustique sont seuillés à -70dB.

ENASC (ou NASC-équivalent) :
Equivalent Nautical Area Scattering Coefficient (représente la réponse acoustique fictive susceptible d'être renvoyée par une capture donnée)

ESDU :
Elementary Sampling Distance Unit. Représente la distance, d'intégration des données acoustiques le long de la route du navire.

FRS-MARLAB :
Fisheries Research Services, Marine Laboratory (Aberdeen, Écosse)

IBTS :
International Bottom Trawl Surveys (Campagnes scientifiques de chalutage de fond dont le protocole est coordonné par le CIEM).

ICES :
International Council for the Exploration of the Sea (= CIEM)

IFREMER :
Institut Français de Recherche pour l'Exploitation de la MER

MAF :
Min/Max Autocorrélation Factors

m.n :
mille nautique. Correspond à la distance entre deux points de même longitude, dont les latitudes diffèrent d'un angle d'une minute (= 1.852 km).

NASC = Nautical Area Scattering Coefficient (représente la grandeur physique mesurée par acoustique. Elle est mesurée en $m^2/m.n^2$)

PÉLAGIQUE :
Adjectif qui qualifie les espèces qui se trouvent en pleine eau, ne dépendant aucunement du fond

pour l'accomplissement de leurs fonctions biologiques. Le hareng et le sprat sont parmi les espèces pélagiques les plus courantes de cette étude.

RADIALES :

Désigne très précisément ici l'ensemble des localisations des enregistrements acoustiques effectués au cours d'une campagne, regroupe les stations et les transects.

STATIONS :

Ensemble des localisations des coups de chalut effectués au cours d'une campagne. En chaque station est disponible simultanément une information obtenue par capture ainsi qu'un enregistrement acoustique.

STOCK :

Utilisé ici dans le sens de biomasse

TRANSECTS :

Ensemble des localisations des sondages acoustiques enregistrés uniquement entre les stations, c'est à dire lorsque le chalut est hors de l'eau. Cette définition est tout à fait spécifique à cette étude.

VESSIE NATATOIRE :

Sac membraneux présent chez beaucoup de poissons osseux qui peut se remplir de gaz et sert à leur équilibre dans l'eau. L'écho renvoyé au sondeur acoustique par le poisson est en partie proportionnel au volume de cette poche.

Introduction

Dans la perspective d'un développement durable, la gestion des pêches vise à assurer la pérennité de l'exploitation des ressources en poissons. Cet objectif de gestion passe par la connaissance de l'état des stocks exploités ainsi que l'effort de pêche qu'ils supportent.

L'estimation des stocks est depuis longtemps réalisée à partir d'observations par prélèvements d'individus dans leur milieu naturel. On distingue classiquement les estimations dites directes basées sur les données scientifiques, des estimations indirectes fondées sur les captures commerciales. En particulier, l'évaluation du stock en espèces démersales se fait, en grande partie, par des campagnes de chalutage de fond. Ces campagnes scientifiques suivent un protocole pré-établi déterminant le plan d'échantillonnage suivi ainsi que la taille et la capacité des engins de pêche utilisés. Cette technique d'observation est souvent lourde en temps et en moyens.

Une alternative moins coûteuse consiste à procéder par sondage acoustique. Le sondeur acoustique est un émetteur-récepteur d'ondes sonores fixé sous la coque du navire, capable de repérer la présence de poissons de la plupart des espèces dans la colonne d'eau. Le gain en temps est considérable puisque les données peuvent être enregistrées en continu, même lorsque le navire est en grande vitesse. Étant non intrusif, ce nouvel outil permet de collecter des observations de la ressource sans la perturber (excepté par le bruit du bateau lui-même). Il permet également de positionner la hauteur des échos reçus dans la colonne d'eau.

Ces aspects sont des avantages majeurs par rapport aux observations obtenues par chalutage.

La morphologie des bancs, l'intensité de leur réponse acoustique ainsi que divers paramètres environnementaux ont inspiré plusieurs axes de recherche visant à permettre l'identification des espèces perçues par le sondeur. Cependant, au stade actuel du développement de cette technique, l'identification est impossible de façon systématique.

Les prélèvements par chalutage restent donc, à l'heure actuelle, le principal outil pour les évaluations de stocks des espèces démersales lors des campagnes scientifiques. En revanche, l'acoustique, couramment utilisée pour les espèces pélagiques, y est peu exploitée. En effet, l'acoustique pose des problèmes de confusion entre les échos renvoyés par les poissons proches du fond et par le fond lui-même.

Dans ce contexte, l'objet de cette thèse est d'évaluer dans quelle mesure les informations acoustiques, collectées de façon quasi-continue le long de la route du navire, peuvent améliorer les estimations d'abondance faites par chalutage de fond.

La variable cible est la variable de pêche, tandis que l'acoustique est utilisée comme variable auxiliaire. Lors des campagnes acoustiques, par exemple, les captures ne sont utilisées à l'inverse que pour identifier les espèces d'une abondance mesurée par le sondeur. Ce choix a donc un impact fort sur les résultats, l'abondance estimée est ici une abondance *capturable*.

Cette problématique s'insère au sein du projet européen CATEFA (Combining Acoustic and Trawl

data for Estimatif Fish Abundance) regroupant quatre principales approches méthodologiques : les régressions linéaires généralisées, les réseaux de neurones, la logique floue et la géostatistique. C'est cette dernière approche qui est utilisée dans ce travail de thèse.

La géostatistique est un ensemble de méthodes probabilistes spécifiquement développées pour l'analyse des données spatialement corrélées. Les prélèvements peuvent être corrélés lorsque l'abondance est spatialement structurée et lorsque le plan d'échantillonnage est non aléatoire. La statistique classique, en revanche, suppose souvent l'indépendance des observations. La géostatistique est donc particulièrement adaptée à l'étude des données acoustiques, localisées de façon systématique le long de la route du navire.

Le cadre méthodologique général de ce travail a consisté à dégager une relation statistique à partir des données. Il a été choisi, au commencement de ce projet, de ne s'appuyer sur aucune connaissance biologique ou physique connue qui ne serait pas portée par les présentes données.

L'objectif a été de développer un modèle susceptible de s'adapter à l'ensemble des configurations de données disponibles. Pour cela, l'approche a consisté à rechercher tout d'abord la relation la plus simple possible pour représenter le lien existant entre les deux variables dans toutes les zones et pour chacune des années.

Cette relation a ensuite été complexifiée en fonction des spécificités de chaque campagne, lorsque la relation initiale ne permettait pas une définition satisfaisante du problème.

Ce mémoire se divise en quatre parties.

La première partie présente de façon précise les deux types de données (l'acoustique et les captures par chalutage de fond) ainsi que les campagnes utilisées. Un certain nombre de difficultés intrinsèques aux données sont ainsi dégagées. Ce chapitre introduit également plusieurs hypothèses nécessaires à la modélisation.

Dans la deuxième partie est analysé le lien structural existant entre les deux variables. L'application de plusieurs modèles, dans une configuration où la variable acoustique est échantillonnée de façon beaucoup plus dense que la variable de pêche, est présentée. En particulier, le rôle joué par la variable auxiliaire, la variable acoustique, est analysé pour chacun des modèles. Certaines difficultés liées à la mise en pratique de ces modèles dans une configuration de données aux distributions très asymétriques sont également étudiées.

L'acoustique est en fait disponible par couche depuis le fond jusqu'à la surface. La troisième partie de ce travail traite de cette dimension verticale. Les observations obtenues dans les différentes couches sont très corrélées et redondantes. Ce chapitre vise à intégrer cette information de façon optimale pour se ramener à un contexte bivariable tel qu'il est présenté dans le second chapitre.

Enfin, la dernière partie compare les résultats des différentes approches étudiées précédemment à ceux de modèles simples n'utilisant pas l'information acoustique. L'amélioration apportée par l'utilisation conjointe du chalut de fond et du sondeur acoustique dans les estimations d'abondance est discutée pour chacune des zones d'étude : la mer du Nord, la mer d'Irlande et la mer de Barents, en fonction, en particulier, des différents plans d'échantillonnage.

Chapitre 1

Contexte et problématique

Des poissons d'une même marmite, on
ne peut tirer qu'un seul bouillon.
Proverbe malgache

Ce chapitre introductif vise à cadrer la problématique générale énoncée précédemment en dégagant un certain nombre de difficultés intrinsèques aux données.

Il commence par une description précise des techniques d'échantillonnage par chalutage et par acoustique.

Plusieurs vérifications nécessaires à la modélisation sont ensuite mises en place. Après régularisation, la cohérence des données acoustiques pendant et en dehors des chalutages est testée. L'impact des mouvements locaux de poissons est en particulier discuté.

Enfin, des statistiques élémentaires ainsi que l'étude des distributions des données présentent en deuxième partie le contexte très asymétrique qui s'ajoute à la problématique principale. Les localisations spatiales des valeurs fortes indiquent, par exemple, sur une simple observation des données, les campagnes pour lesquelles le lien acoustique-capture sera très difficile à dégager.

This first chapter aims to narrow the general problem presented earlier and identifies a number of difficulties intrinsic to the data.

Firstly, the two sampling techniques, i.e. acoustic and trawl surveys, are accurately described. Several assumptions required for the modelling are then checked. The consistency between the acoustic data collected away from the trawl stations with that collected during the trawling operations is tested, for equivalent supports. In particular, the fish reactions effects are discussed.

Then, in a second section, statistical data summaries show the very skewed distribution context of the problem. For example, by simply looking at the spatial location of the high values in some surveys, it is obvious that the link acoustic-trawl will be very difficult to establish.

1.1 Présentation des données

Les données utilisées proviennent de campagnes scientifiques réalisées dans trois mers différentes : la mer du Nord, la mer d'Irlande et la mer de Barents. Chacune des campagnes collecte à la fois des données de chalutage de fond et des données acoustiques.

1.1.1 Les données acoustiques

Utilisée depuis le début des années 60, la méthode de l'écho-intégration consiste à mesurer la quantité de poisson à partir d'ondes acoustiques. Elle permet d'échantillonner de façon dense des zones de grande taille à coût relativement faible. Comme elle ne limite pas la vitesse du bateau, l'effet des déplacements de poissons est très faible. Cependant cette technologie ne permet pas de distinguer les espèces, elle ne mesure qu'une abondance totale.

Certains poissons contiennent une vessie natatoire, sorte de poche d'air qui réfléchit les ondes sonores. Le sondeur est placé sur la coque immergée du navire. Il émet des ondes et perçoit les échos réfléchis par les poissons. Le volume ainsi échantillonné est un cône d'angle 7° , sa surface d'incidence sur le sous-sol est donc d'autant plus importante que la mer est profonde. Pour une profondeur de 40 m, par exemple, la base du cône est un cercle de près de 5 m de diamètre alors que pour une profondeur de 100 m, le diamètre est de 12 m.

Une description précise de la technique acoustique peut être trouvée dans MacLennan et Simmonds (1992, [39]) ainsi que dans Johansson et Mitson (1983, [35]). Pour une longueur d'onde donnée, la réponse d'un poisson dépend principalement de la taille de sa vessie natatoire et de sa position spatiale. En particulier, son inclinaison par rapport à l'horizontale ainsi que le fait d'être isolé ou en banc modifie son écho. La grandeur physique mesurée par acoustique est un indice d'abondance égal au rapport entre la surface réfléchissante des poissons et la surface échantillonnée. Elle est mesurée en $m^2/m.n^2$ et appelée NASC (Nautical Area Scattering Coefficient, MacLennan et al. 2002 [38]). La figure 1.1 montre un exemple d'échos acoustiques enregistrés par le sondeur.

Tout comme les poissons, le plancton, les algues et le fond marin envoient un écho acoustique qui va être perçu par le sondeur. Les échos en dessous d'une certaine intensité sont donc considérés comme du bruit et seuillés. Pour les données utilisées dans ce projet, le seuil a été fixé à -70 décibels, considérant que pour un seuil de -80 dB on enregistrerait trop de plancton et que pour un seuil de -60 dB on ne parvenait plus à percevoir les poissons de petite taille.

Les pertes d'intensité d'une onde dues directement à la longueur de son parcours sont corrigées en fonction de l'intervalle de temps qui sépare son émission de sa réception par le sondeur.

Les intégrations de fond

Les échos envoyés par le fond sous-marin posent un problème délicat. En effet celui-ci envoie généralement un écho très fort, qui permet ainsi de le détecter. Mais comment le distinguer de bancs de poissons d'intensité comparables, ou plus faibles, collés au fond de la mer ? C'est une question difficile d'autant plus importante que l'acoustique vise ici à être combinée à des données de chalutage de fond, c'est à dire des captures proches du fond de la mer.

L'acoustique est couramment utilisée dans les campagnes pélagiques, dans lesquelles on ne s'inté-

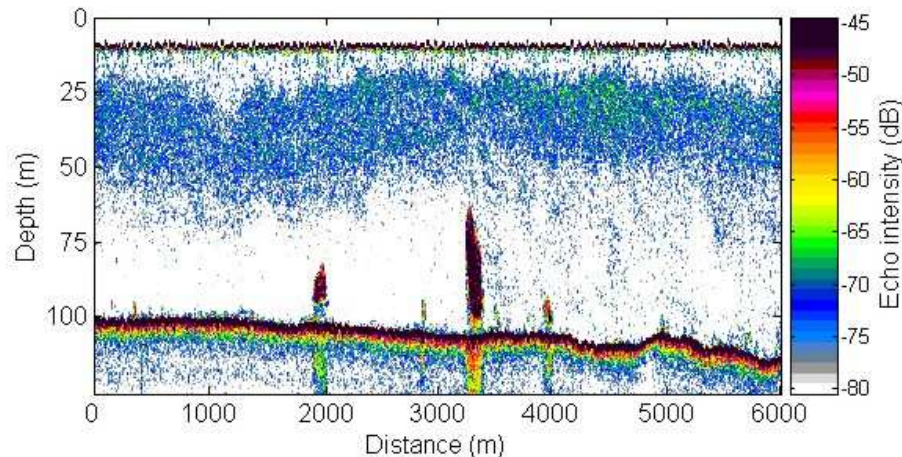


FIG. 1.1 – Exemple d'échos enregistrés par le sondeur acoustique, source : Marine Laboratory of Aberdeen

resse qu'aux couches d'eau intermédiaires. On se prémunit alors de toute erreur liée à l'intégration de détections du fond. Le cas étudié ici demande au contraire une précision importante des couches les plus profondes. A ce point de vue, ce projet a permis, pour les experts acousticiens, de développer des algorithmes de scrutation de ces données visant à améliorer les problèmes liés à la détection de fond (Greig et Reid 2004, [31], MacLennan et al. 2004, [37]).

Il y a deux erreurs possibles : soit considérer du fond comme étant du poisson, soit considérer du poisson comme étant du fond (cf. figure 1.2). La première erreur est très grave. Il suffit qu'un tout petit peu de fond soit considéré comme étant du poisson pour que des valeurs énormes soient générées qui vont influencer lourdement les statistiques. L'autre risque est d'oublier certains bancs de poissons. Aussi importante soit elle, cette erreur a beaucoup plus de chance de passer inaperçue. Ayant un impact particulièrement grave sur les résultats, chacune des valeurs fortes a été systématiquement vérifiée. C'est un travail manuel laborieux réalisé en collaboration avec les experts acousticiens au cas par cas. On s'est aperçu en effet qu'environ une valeur extrême sur deux provenait d'une erreur.

Les valeurs fortes restantes dans les bases de données sont donc considérées comme représentant effectivement du poisson et n'ont plus été remises en cause (cf. annexe A).

Échantillonnage

Les données acoustiques sont découpées par couches (ou layers) numérotées depuis le fond jusqu'à la surface. Il y a d'autant plus de couches que la mer est profonde. Les dix premières couches font 1 m de haut, les suivantes font 10 m de haut. Les couches utilisées ici sont référencées par rapport au fond de la mer (cf. figure 1.3).

En raison des problèmes d'intégration de fond, les données de la première couche sont fournies de deux façons : avec et sans "backstep". Le backstep est une couche de 10 à 40 cm, fixe ou variable

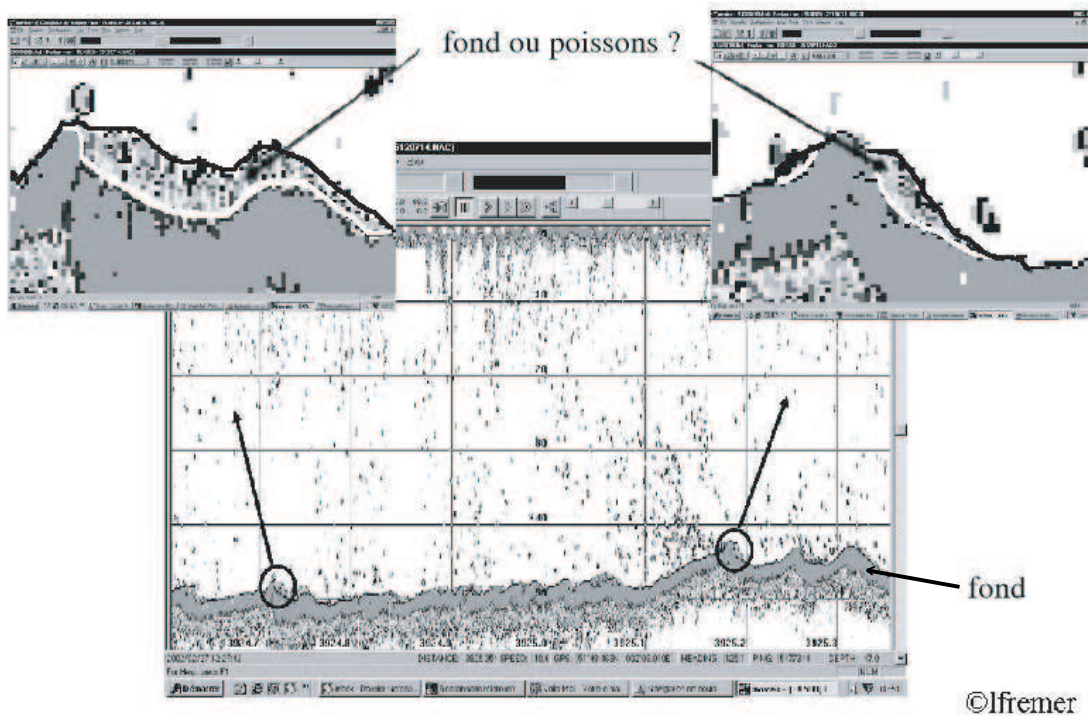


FIG. 1.2 – Exemple d’une détection de fond délicate, logiciel MOVIES, source IFREMER

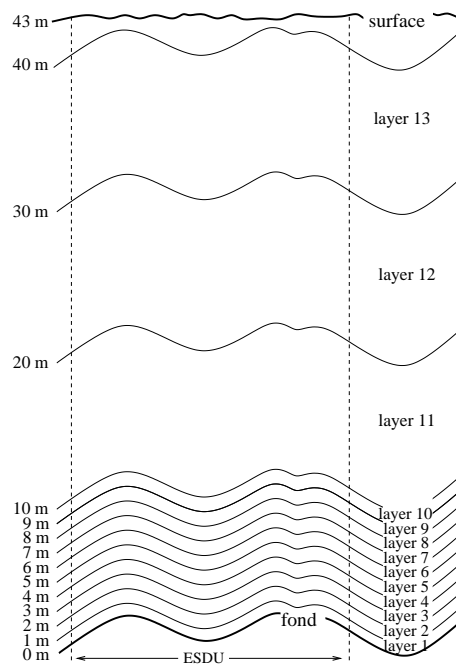


FIG. 1.3 – Découpage vertical de la colonne d’eau

selon les campagnes, dans laquelle les données sont considérées par les acousticiens comme incertaines. Cette zone a été mise de côté pour la plupart des explorations.

On appelle "station", la localisation du navire lors d'un coup de chalut. En stations le bateau va à 3 ou 4 noeuds et collecte à la fois des données de chalutage et des données acoustiques (cf. figure 1.11). Un coup de chalut mesure en moyenne 1.5 m.n (mille nautique) en Norvège, 1.8 m.n en France, 2 m.n en Écosse et en Angleterre et 3 m.n en Irlande. L'acoustique est intégrée sur toute la durée du coup de chalut. Elle est ensuite divisée par la longueur de la station pour obtenir une unité de densité de surface.

On appelle "transects" les enregistrements acoustiques situés entre les stations et "radiales" l'ensemble des localisations des enregistrements acoustiques (transects +stations). Le long des transects, le bateau va généralement à 8 ou 9 noeuds selon les conditions météorologiques. Les données sont intégrées par ESDU (elementary sampling distance unit) fixes pour une campagne donnée : 1 m.n en Norvège, 0.1 m.n pour la France et 0.5 m.n pour l'Écosse, l'Angleterre et l'Irlande.

1.1.2 Les données de capture

Le chalut est un filet lesté par deux panneaux en métal qui raclent le fond de la mer. Il a une ouverture de 3 à 6 mètres de haut sur près de 20 mètres de large. Un capteur permet de mesurer pour chaque coup de chalut la hauteur du filet et la distance entre les panneaux. Les fûnes (câbles qui tirent le chalut, cf. figure 1.4) sont filés sur des distances allant de 3 à 5 fois la profondeur de la mer. Donc plus la mer est profonde, plus le chalut est loin derrière le bateau. De même la distance entre les panneaux dépend de la profondeur et du type de chalut, elle peut varier de 30 à 90 m selon les campagnes.

La durée des coups de chalut est standardisée au sein de chaque campagne. Par convention, c'est seulement quand la position du filet ainsi que celle des panneaux sont stables que l'on commence à mesurer le temps de pêche.

Une fois capturés, les poissons sont triés par espèce, comptés, pesés et mesurés. Les variables directement disponibles sont pour chaque coup de chalut et pour chaque espèce :

- le nombre
- le poids (en kg)
- la racine carrée de la moyenne des carrés des longueurs (Root Mean Square)

Ces variables sont disponibles pour chacune des espèces capturées mais aussi par regroupement d'espèces. Selon leur type de comportement, on distingue trois groupes d'espèces, les pélagiques, les démersaux et les poissons benthiques. Les pélagiques sont des poissons vivant en pleines eaux et souvent en bancs assez denses. Dans les mers étudiées ici il s'agit principalement du hareng et du sprat. Les démersaux sont des poissons vivant plus proches du fond, principalement ici la morue, l'églefin, le merlan. Enfin les poissons benthiques regroupent plusieurs espèces vivant collés au fond comme la plie, la raie ou la limande. Les poissons benthiques à morphologie aplatie seront très peu utilisés dans cette étude en raison de la très faible capacité des sondeurs acoustiques à les enregistrer. En effet, le fait de vivre collé au fond les rend quasiment indétectables. Les autres espèces

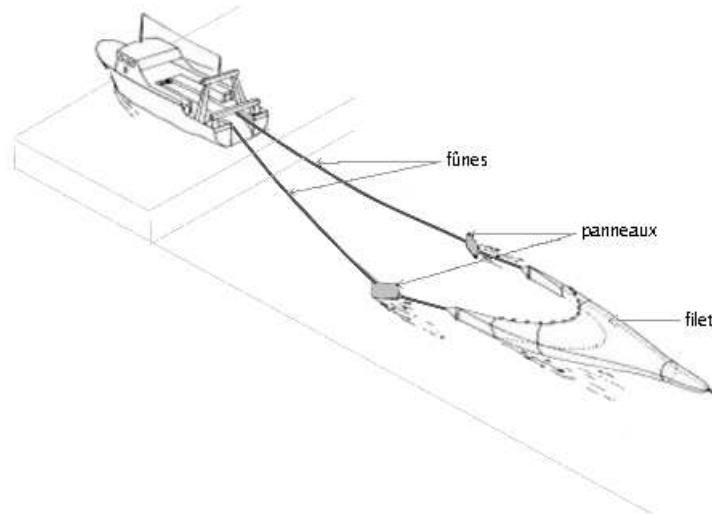


FIG. 1.4 – Description du chalut de fond.

benthiques (comme la petite roussette et le grondin) ont pour la plupart des campagnes été agrégées dans un groupe "divers", en raison de leur présence extrêmement rare.

Il a été vérifié dans les données que les groupes d'espèces correspondaient effectivement à des classes de poissons de comportements semblables, du point de vue des variables de captures. En effet, on voit par A.C.P que les espèces se regroupent nettement en trois ensembles qui correspondent aux trois catégories déterminées par les biologistes. Le premier axe principal distingue les poissons plats des autres espèces tandis que les deuxième et troisième permettent de séparer les démersaux des pélagiques.

Le souci de cette étude étant de trouver une relation simple et la plus robuste possible entre les données acoustiques et les données de captures, j'ai utilisé, pour la plupart des approches, les données agrégées par groupes d'espèces : les pélagiques et les démersaux. Des recherches de modélisation spécifiques par espèce ont été effectuées par l'IMR (Norvège) et le Marine Laboratory (Écosse) qui se sont intéressés aux deux ou trois espèces les plus courantes. Étant donné que ces régions ne comptent que trois ou quatre espèces principales, les résultats restent généralement similaires.

Le NASC-équivalent (ENASC)

Pour chacune des espèces capturées, une valeur acoustique équivalente est calculée, on l'appelle le NASC-équivalent et la note ENASC. Elle correspond à la réponse acoustique moyenne qui pourrait être enregistrée par le sondeur pour un groupe de poissons du type de celui pêché. Elle est calculée pour chaque espèce à partir de la distribution des longueurs moyennes des captures, par la relation suivante (MacLennan et al. 2002 [38]) :

$$ENASC(espece, x) = \frac{4\pi}{A_{m.n}(x)} \sum n_L(espece, x) L_{RMS}(espece, x) \cdot 10^{\frac{b(espece)}{10}} \quad (1.1)$$

où L_{RMS} est la racine carrée de la moyenne des carrés des longueurs (Root Mean Square), $b(espece)$ est une constante d'intégration (sans unité) spécifique à chacune des espèces, n_L est le nombre de poissons de la taille L et $A_{m.n}$ représente la surface chalutée en $m.n^2$. Les longueurs sont arrondies au demi-centimètre inférieur lors de la mensuration, la variable de longueur est donc discrète. La constante d'intégration de chaque espèce est obtenue expérimentalement à partir de sondages sur des poissons en captivité.

Le ENASC d'un groupe d'espèce est calculé en sommant les ENASC obtenus pour chacune des espèces du groupe pondéré par le nombre d'individus capturés de chaque espèce.

1.1.3 Plans d'échantillonnage et caractéristiques de chaque chantier

La mer du Nord, la mer d'Irlande et la mer de Barents sont trois régions ayant des caractéristiques bien distinctes. Les données de captures ne répondent pas aux mêmes plans d'échantillonnage; les espèces, la bathymétrie, la géographie, leur exploitation et l'évolution que suit la biomasse font de ces trois mers des configurations très différentes de la même problématique.

Aussi, bien que le souci de cette étude était de dégager un lien acoustique-capture le plus général possible, chacune des campagnes a été traitée séparément.

La table 1.1 (p. 11) synthétise les principales caractéristiques des campagnes disponibles.

Mer du Nord

La mer du Nord recouvre une zone assez étendue et très disparate. La profondeur va de 25 m au sud pour atteindre 150 m au large de l'Écosse. Les estimations montrent une réduction importante des stocks de morue et de merlan, qui sont actuellement en dessous du seuil de renouvellement (rapport ACFM, octobre 2003, [3]). Le haddock reste constant, tandis que le hareng, longtemps en dessous du seuil de renouvellement semble réaugmenter depuis peu.

La zone échantillonnée se situe entre Boulogne-sur-mer et les îles Shetlands en s'arrêtant à l'est à la fosse norvégienne (cf. figure 1.5). On trouve de gros bancs de harengs et de sprats au sud alors que le nord est essentiellement composé d'espèces démersales : autour des côtes écossaises, les espèces les plus capturées sont l'églefin, le merlan et le tacaud norvégien. La taille des poissons est également différente entre le Nord et le Sud. On trouve de grandes quantités de petits poissons au sud (zone de pont) et de plus gros poissons au Nord.

Les données utilisées proviennent des campagnes scientifiques IBTS (International Bottom Trawl Surveys) auxquelles ont été ajoutées les enregistrements acoustiques collectés entre et durant les chalutages. Dans les années 80, le Conseil International pour l'Exploration de la Mer (CIEM) a décidé de standardiser les programmes d'estimation d'abondance réalisés par chaque pays (cf. Vérin 1995 [69], Simmonds et al. 1999, [67]). Les engins de pêche (Chalut de fond à grande ouverture verticale), la vitesse des bateaux (4 noeuds lors des chalutages et entre 8 et 10 noeuds ailleurs) et les temps de chalutage (30 min.) sont donc identiques à toutes les campagnes. Les chalutages suivent un plan d'échantillonnage aléatoire stratifié. La mer du Nord a été divisée en rectangles de 30' de latitude sur 1° de longitude au sein desquels deux traits de chalut sont effectués au hasard. Chaque campagne compte entre 60 et 80 coups de chalut. L'acoustique est collectée entre les chaluts, sauf durant la nuit où les sondeurs sont arrêtés. En effet, une spécificité des campagnes IBTS est qu'aucune collecte (acoustique et chalutage) n'est réalisée de nuit, pour éviter les problèmes de

biais dû au comportement nyctéméral des espèces. La distance entre les panneaux va de 60 m en moyenne au sud à 90 m au Nord. L'ouverture verticale du chalut fait environ 5 mètres.

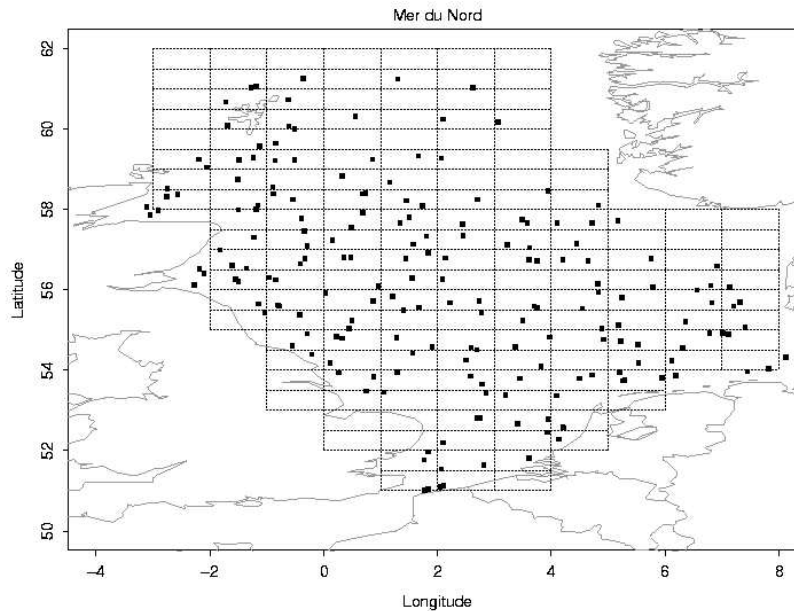


FIG. 1.5 – Plan d'échantillonnage aléatoire stratifié des stations de mer du Nord (campagne 2002). Les lignes pointillées délimitent les strates fixées par le CIEM.

Les campagnes sont réalisées par trois instituts différents (figure 1.6) :

- IFREMER : l'institut français de recherche pour l'exploitation de la mer effectue ses campagnes principalement dans le sud de la mer du Nord entre le 51° N et le 56° N. Les campagnes sont disponibles pour les années 2002 et 2003 (février).
- le CEFAS : Centre for the environment fisheries and aquaculture science (Angleterre). Ses campagnes se situent entre le 52° N et le 62° N. Les campagnes 2000, 2001 et 2002 sont disponibles (août-septembre).
- le Marine Laboratory of Aberdeen (FRS-MARLAB, Écosse) dont les campagnes sont regroupées autour de l'Écosse, entre le 57° N et le 62° N. Les campagnes disponibles sont 1999, 2000, 2002 et 2003 (janvier-février).

Malgré l'homogénéité des données fournies par chacun des trois pays, celles-ci ont généralement été traitées séparément en raison en particulier des différences d'espèces, de saisons et de profondeur entre le sud et le nord de la zone.

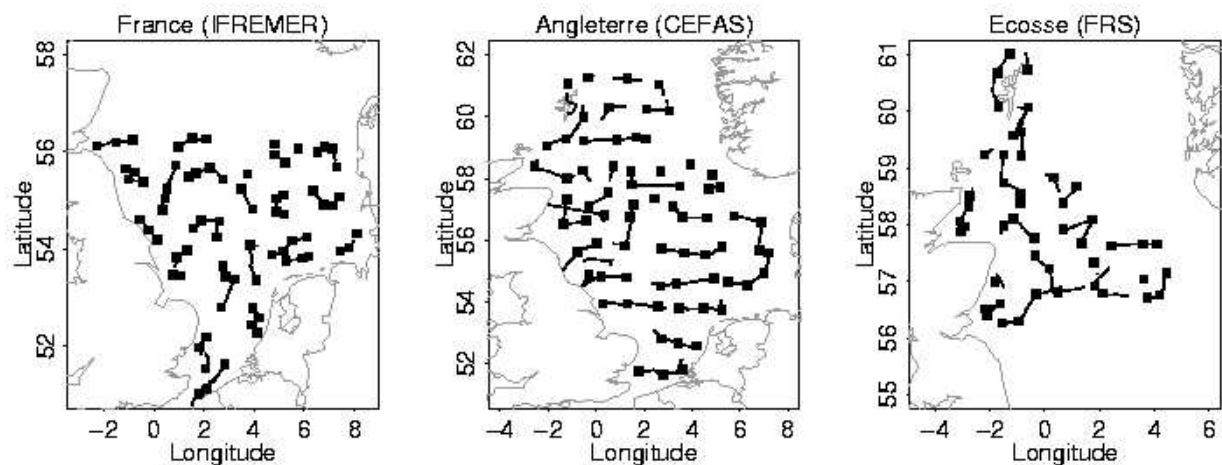


FIG. 1.6 – Stations et transects des campagnes de chacun des trois pays de mer du Nord (campagnes 2002). Les carrés représentent les stations et les traits les transects.

Mer d'Irlande

La mer d'Irlande est une petite mer dont les profondeurs varient entre 20 et 150 m. L'état des ressources indique une augmentation importante de l'églefin depuis les années 1990, tandis que les stocks de morue, de merlan et de plie sont estimés en dessous du seuil de renouvellement (ACFM, octobre 2003 [2]).

Les chalutages suivent un plan d'échantillonnage aléatoire stratifié par zone de profondeur et de substrats (sable/gravier) semblables. En tout, neuf strates sont ainsi délimitées dans chacune desquelles les chalutages sont répartis au hasard (cf. figure 1.7). Les campagnes ne comptent qu'entre 20 et 30 coups de chalut par an. Le chalut fait 3 mètres d'ouverture verticale, alors que la distance entre les panneaux varie entre 35 et 40 m. Les campagnes, réalisées par la Queen's University of Belfast (QUB), sont disponibles pour les années 2000 (hiver), 2001 (automne) et 2002 (hiver). Les données par espèce ne sont pas disponibles, seules les données agrégées pour les trois principaux groupes sont fournies. Les démersaux représentent environ 65% des captures en poids et 80% en ENASC. Les pélagiques représentent 20% en poids et 20% en ENASC. La plupart de l'abondance est concentrée sur le plateau continental.

Mer de Barents

La mer de Barents est une mer très vaste dont la profondeur varie de 30 à 800 m de fond et se situe généralement autour de 300 m. La zone échantillonnée est entre le 15° E et le 50° E de longitude et va de la côte norvégienne au sud jusqu'au 76° N (cf. figure 1.8). Le stock de morue semble augmenter en taille et en poids mais stagner en nombre depuis 2000, tandis que le capelan est estimé à un niveau satisfaisant (survey report octobre 2002 [1]).

Six campagnes, effectuées entre janvier et mars par l'Institute of Marine Research de Bergen (IMR),

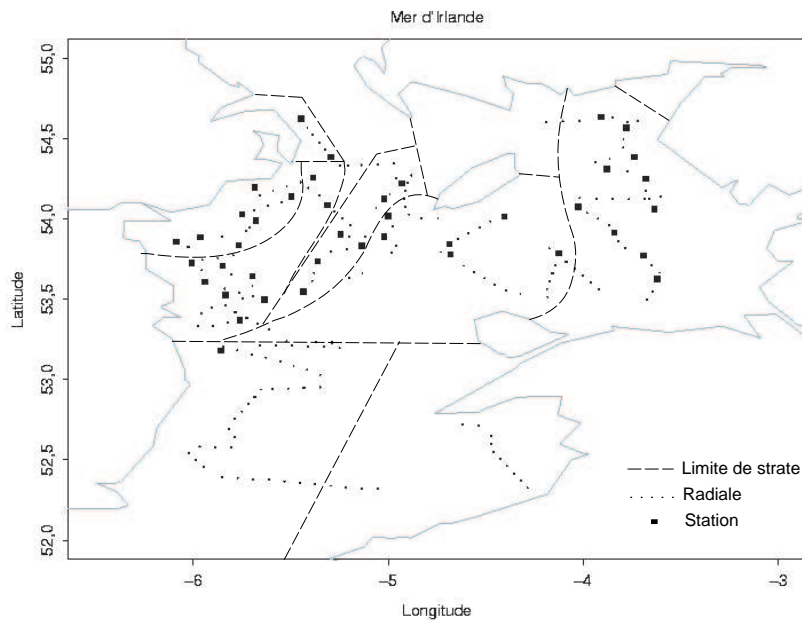


FIG. 1.7 – Plan d'échantillonnage aléatoire stratifié par la profondeur et le substrat sous-marin des stations en mer d'Irlande (campagne 2002). Les strates sont indiquées en pointillé.

sont disponibles : 1997, 1998, 1999, 2000, 2001 et 2002. Elles répondent à un plan d'échantillonnage quasi-régulier avec un trait de chalut environ tous les 15 à 20 miles nautiques. Chaque campagne réunit entre 200 et 300 stations. Les sondeurs acoustiques fonctionnent de jour et de nuit sans arrêt sur tout le trajet du bateau. Le chalut a en moyenne une ouverture de 4 m de haut pour une distance entre les panneaux de 40 à 50 m. Les espèces les plus couramment capturées (dans ces données) sont des démersaux : la morue, l'églefin, le capelan et le tacaud norvégien.

1.1.4 Remarques sur les différences entre les deux types de mesures

L'acoustique et le chalutage sont deux méthodes très différentes de mesurer l'abondance. Plusieurs phénomènes rendent leur relation plus bruitée.

Tout d'abord, le chalut de fond, par nature, capture des poissons vivant plutôt au fond de la mer, or le premier mètre au dessus du fond est justement difficile à percevoir par acoustique (cf. partie 1.1.1).

Ensuite, le chalut est placé loin derrière le bateau alors que le sondeur est directement fixé sous le navire. Ils mesurent donc des données légèrement décalées dans le temps (cf. figure 1.9). Comme le bateau se déplace à quatre noeuds, ce retard va d'environ 40 secondes quand la profondeur est de 30 mètres à plus de 7 min quand la profondeur dépasse 300 mètres. En mer de Barents, où les profondeurs sont les plus grandes, le décalage varie entre 3 et 15 minutes, c'est à dire entre un dixième et la moitié de la durée de chalutage.

Mer - Institut	Année	Mois	Nb de stations	Nb d'ESDU en radiales (après régularisation)	Distance chalutée moyenne (m.n)	Distance entre les panneaux(m)	Ouverture verticale (m)	Profondeur
Barents	1997	02-03	176	5209	1.5	53.6	4.7	143-699
Barents	1998	02	198	5135	1.53	49.6	4.4	63-720
Barents	1999	01-02	223	5567	1.49	47.3	4.8	104-480
Barents	2000	01-02	302	7680	1.42	48.2	4.1	58-550
Barents	2001	01-03	300	7666	1.49	48.9	4.3	55-487
Barents	2002	01-03	287	7383	1.44	50.1	4.0	63-542
Mer du Nord - FRS	1999	01-02	44	468	1.8	83.7	4.25	45-150
Mer du Nord - FRS	2000	01-02	46	351	2.01	87.1	4.38	48-144
Mer du Nord - FRS	2002	01-02	47	430	1.98	81.1	4.6	49-150
Mer du Nord - FRS	2003	01-02	46	303	1.9	80.7	4.3	50-138
Mer du Nord - CEFAS	2000	08-09	71	1038	1.98	78.8	5.6	24-178
Mer du Nord - CEFAS	2001	08-09	70	883	2.01	80.7	5.1	24-211
Mer du Nord - CEFAS	2002	02	70	1140	1.98	77.1	5.7	24-84
Mer du Nord - IFREMER	2002	02	77	440	1.83	65.8	5.0	9-88
Mer du Nord - IFREMER	2003	02	82	722	1.89	61.2	5.4	14-90
Irlande	2000	03	37	110	2.90	35.2	3	26-106
Irlande	2001	10	34	236	2.70	35.5	3	23-90
Irlande	2002	03	41	173	2.85	35.8	3	24-102

TAB. 1.1 – Caractéristiques des différentes campagnes disponibles

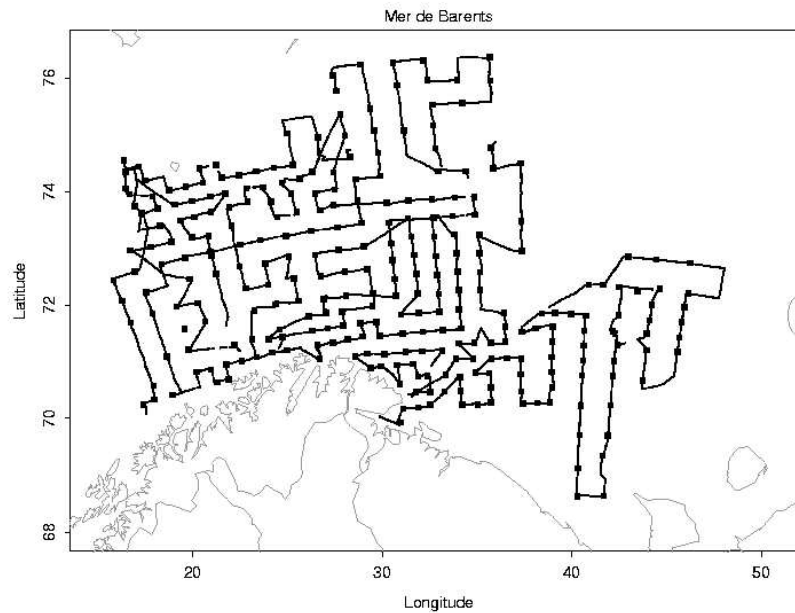


FIG. 1.8 – Plan d'échantillonnage quasi-régulier des stations en mer de Barents (campagne 2002). Les carrés représentent les emplacements des stations reliées par les transects.

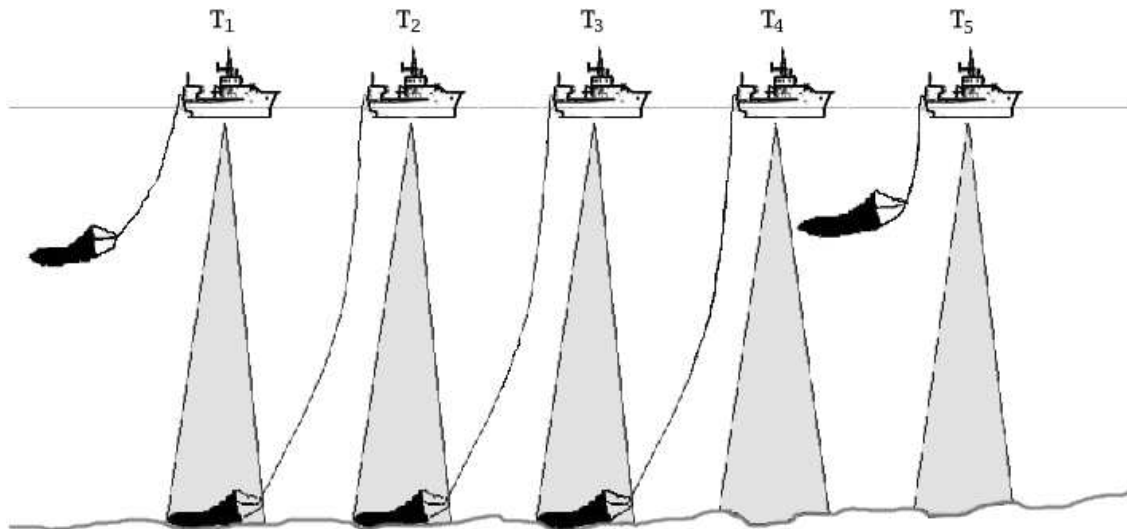


FIG. 1.9 – Décalage entre le chalut et le cône acoustique au sein d'une station. Les figures représentent de gauche à droite plusieurs instants du déroulement d'un chalutage. La position du cône acoustique à l'instant T correspond à celle du chalut à l'instant $T+1$.

Enfin lorsque les courants marins sont forts, le chalut est souvent entraîné et se retrouve alors décalé latéralement par rapport à la route du navire (cf. figure 1.10).

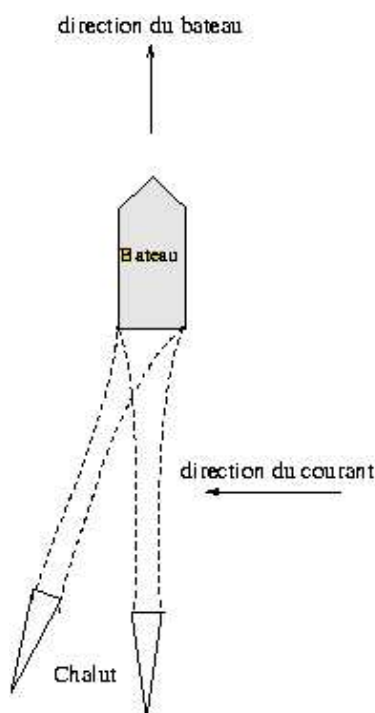


FIG. 1.10 – Selon les courants, le chalut peut se décaler par rapport à la route du navire.

Ainsi, à moins d'une grande homogénéité spatiale, ce qui est capturé ne correspond jamais exactement à ce qui est enregistré par le cône acoustique. Plus les poissons seront réactifs et auront des distributions irrégulières, plus la corrélation entre l'acoustique et les captures sera faible.

Les différentes situations susceptibles de se produire en stations peuvent se résumer dans le tableau récapitulatif suivant :

	Acoustique = 0	Acoustique > 0
Capture = 0	abondance nulle	erreur de type 1
Capture > 0	erreur de type 2	abondance positive

L'erreur de type 1 se produit lorsque le sondeur acoustique enregistre la présence de poissons mais que rien n'est collecté par le chalut. Elle est a priori due aux mouvements latéraux ou verticaux de poissons qui leur ont permis d'échapper au chalut. Il peut s'agir aussi bien d'une abondance capturable, que d'une abondance impossible à pêcher par ce type de chalut, ou dans ce type de conditions.

L'erreur de type 2 correspond au cas où rien n'est perçu par le cône acoustique alors qu'une abondance locale non négligeable est obtenue par pêche. Ce phénomène peut être expliqué par les mouvements latéraux de poissons, qui ont évité le cône acoustique, par la présence de poissons plats ou

de poissons très collés au fond et répondant mal à l'acoustique. Ce type d'erreur est favorisé lorsque l'ouverture horizontale du chalut est plus grande que le diamètre du cône acoustique. Par contre il s'agit évidemment d'une abondance capturable, qu'il est important de prendre en compte dans les estimations.

Réactivité des poissons et effet de support

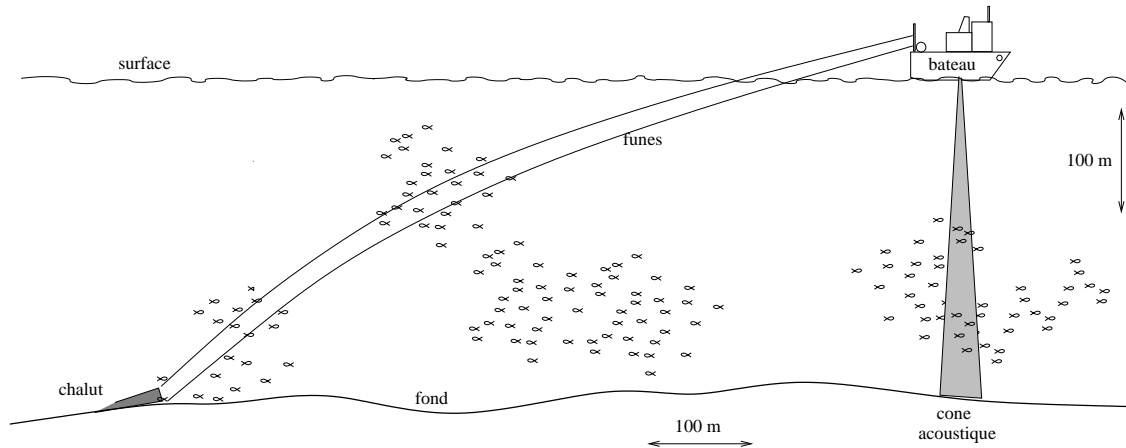


FIG. 1.11 – Représentation proportionnelle de la taille du cône acoustique et de la distance entre le bateau et le chalut en Norvège pour une profondeur de 300 mètres.

Les deux instruments de mesures ont un impact important sur les poissons. Les poissons réagissent d'abord à la présence du navire. Mitson et Knudsen (2003 [46]) montrent par exemple l'impact du bruit d'engins scientifiques diesel sur un banc de harengs. Les poissons réagissent aussi au bruit provoqué par le chalut. Cette réaction a été étudiée en détail (Michalsen et al. 1999 [45], Handegard et al. 2003 [33], Walsh 1996 [72], Godøet al. 1999 [24]), elle semble très variable, dépend beaucoup de l'espèce et des conditions extérieures. Pris de peur, les poissons ont souvent tendance à plonger vers le fond de la mer. Ce phénomène explique par exemple que l'on pêche avec un chalut de 4 à 5 mètres de haut des espèces vivant habituellement beaucoup plus haut dans la colonne d'eau. Les poissons peuvent aussi réagir en se déplaçant latéralement. Le support effectif des données de captures est donc a priori inconnu tout comme, mais dans une moindre mesure, celui de l'acoustique.

Si l'on néglige cet effet, le volume échantillonné par le chalut est égal au produit de son ouverture par la distance chalutée. La distance chalutée est connue précisément pour chaque chalutage. Elle est en moyenne égale à 2 m.n en mer du Nord, 1.5 m.n en mer de Barents et 3 m.n en mer d'Irlande. L'ouverture verticale du chalut est également enregistrée à chaque prélèvement. Elle varie autour de 3 m en Irlande, 4 en mer de Barents et 5 en mer du Nord (cf. partie 1.1.3). Les figures 1.11 et 1.12 sont des représentations à l'échelle du processus de collecte. Enfin, le support horizontal du chalut peut soit correspondre à la distance entre les panneaux soit à la largeur du filet lui-même. C'est généralement la distance entre les panneaux qui est retenue, considérant comme rare qu'un poisson placé entre les deux panneaux parvienne à s'échapper.

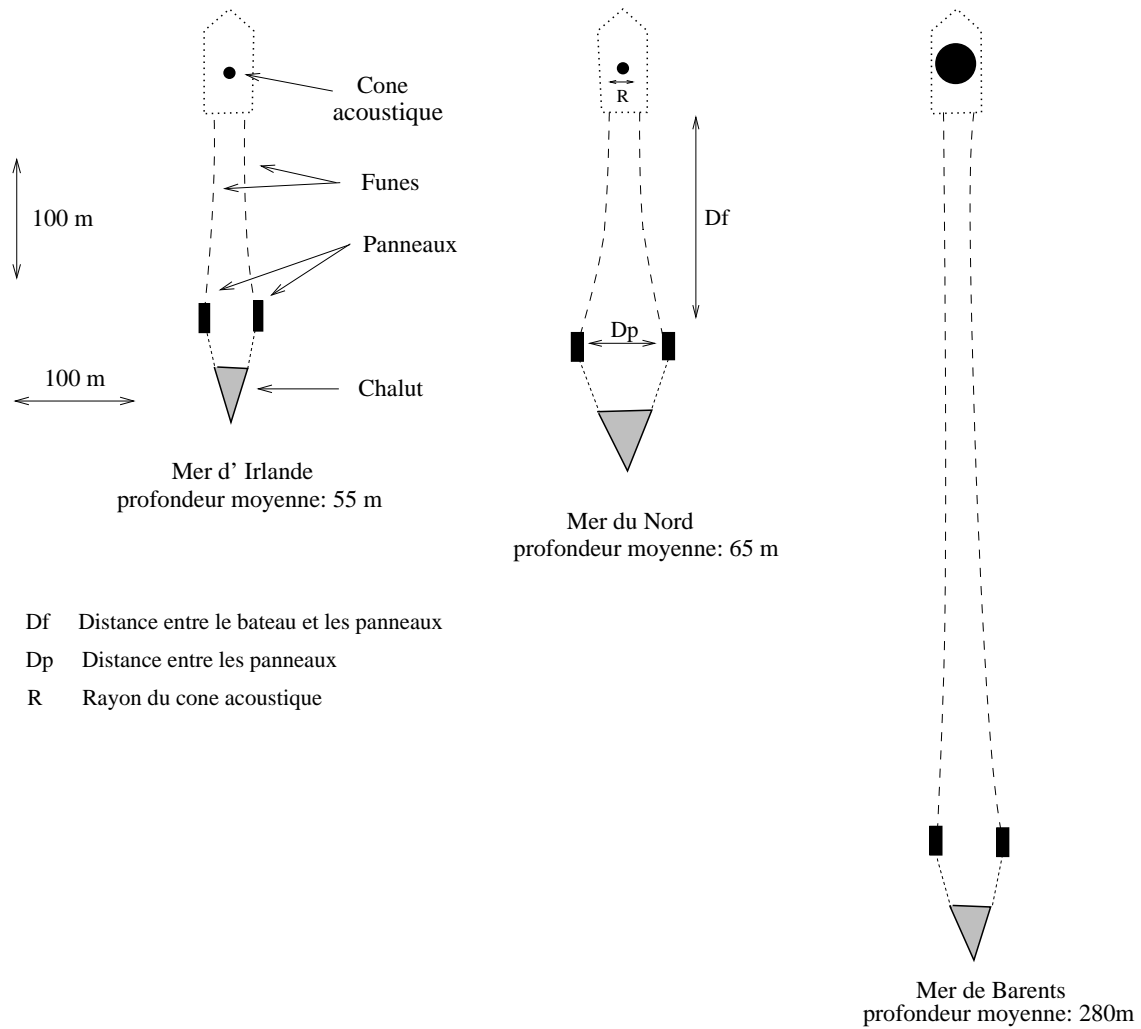


FIG. 1.12 – Représentation proportionnelle des différences de taille entre le cône acoustique et à distance entre les panneaux pour les profondeurs moyennes des trois mers. La taille du cercle noir représente la taille de la surface d'incidence du cône acoustique pour la profondeur correspondante.

On peut donc calculer un support moyen du chalut pour chaque prélèvement :

$$V_{chalut}(x) = h_{filet}(x) \times d_{panneaux}(x) \times d_{chalutee}(x)$$

où h_{filet} est la hauteur du filet, $d_{panneaux}$ est la distance entre les panneaux et $d_{chalutee}$ est la distance chalutée.

De la même façon, le volume échantillonné par l'acoustique, en négligeant le mouvement des poissons, est égal au volume du cône acoustique. Il peut donc pour un layer i d'une profondeur à sa base p_i être approché par une pyramide tronquée de base rectangulaire (en négligeant l'effet conique des bords) :

$$V_{acoustique}(x) = \frac{1}{3} \times d_{chalutee}(x) \times 2 \times \tan\left(\frac{7}{2}\right)[p_i^2 - p_{i-1}^2]$$

Si l'on néglige le mouvement des poissons, le rapport de support entre les deux types de données, en considérant une hauteur d'acoustique égale à la hauteur du filet, est proportionnel au rapport entre le diamètre du cône pour la profondeur sondée et la distance entre les panneaux. C'est en mer du Nord que la différence de support est la plus grande, l'ouverture du chalut est de 5 à 12 fois plus large que le diamètre du cône acoustique du nord au sud (cf. figure 1.13). En mer d'Irlande le support des captures est de 4 à 6 fois plus grand (figure 1.14), alors qu'en mer de Barents les supports sont de dimensions équivalentes, il arrive même que le cône acoustique soit plus large que la distance entre les panneaux (figure 1.15).

La question du support vertical effectif du chalut lorsque l'on prend en compte les éventuels déplacements de poissons est traitée dans le chapitre 3.2.

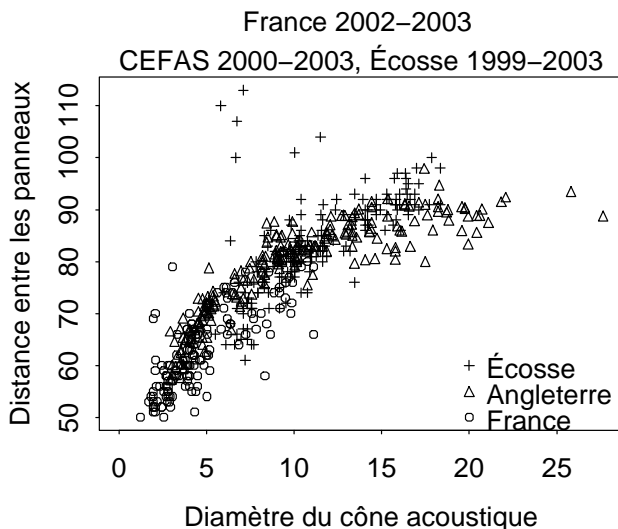


FIG. 1.13 – Nuage de corrélation entre le diamètre du cône acoustique et la distance entre les panneaux pour les stations de mer du Nord. Les unités de mesure des axes sont en mètres.

Ainsi, les plans d'échantillonnage, les supports des prélèvements, les différences d'espèces et d'environnement ont nécessité que les trois mers étudiées soient traitées séparément. Le lien entre l'acoustique et les captures sera, par exemple, beaucoup plus difficile à dégager en mer du Nord

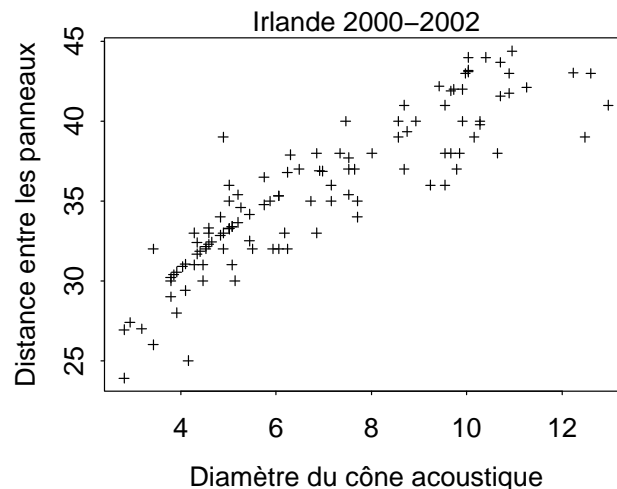


FIG. 1.14 – Nuage de corrélation entre le diamètre du cône acoustique et la distance entre les panneaux pour les stations de mer d’Irlande.

(cf. chapitre 4). L’impact des plans d’échantillonnage sur les résultats des modèles est discuté en dernière partie.

1.1.5 Mise en place d’hypothèses de cohérence dans les données

Cette partie vise à établir quelques vérifications sur les données et à s’assurer de plusieurs hypothèses qui seront par la suite supposées comme acquises tout au long de l’étude.

Régularisation des données acoustiques

Le but de cette thèse étant d’utiliser la corrélation acoustique-capture observée en stations pour l’extrapoler aux données acoustiques observées le long des transects, il est important de s’assurer que l’acoustique en stations et en transects sont sur des supports équivalents. En effet, si une correspondance entre l’acoustique et les captures est mise en place à une certaine échelle, celle de la station, cela ne nous informe en rien sur l’état de ce lien pour un support inférieur, sauf dans le cas particulier du modèle en corrélation intrinsèque, où la corrélation est indépendante du support. Bien que les données acoustiques le long des transects soient connues à un support souvent très fin, ce n’est qu’au support des stations qu’elles ont été utilisées : si le support des données en transects peut être agrégé, le support d’un coup de chalut peut difficilement être "découpé".

Pour les campagnes françaises, on passe par exemple d’ESDU de 0.1 m.n en transects à 1.8 m.n en stations. Même si les moyennes sont du même ordre de grandeur (puisque les données sont fournies par unité de surface), les données des transects décrivent un phénomène beaucoup plus détaillé mais aussi beaucoup plus variable que les données en stations. Les données acoustiques collectées le

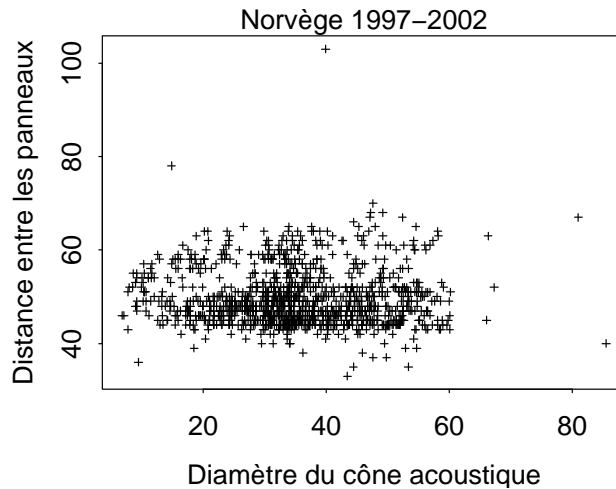


FIG. 1.15 – Nuage de corrélation entre le diamètre du cône acoustique et la distance entre les panneaux pour les stations norvégiennes de 1997 à 2002.

long des transects ont donc été régularisées à un support égal à la moyenne des distances chalutées au sein de la campagne. Ce support, comme on l'a vu, est de 2 m.n en mer du Nord et 3 m.n en mer d'Irlande. En mer de Barents, les stations mesurant en moyenne 1.5 m.n, et les données des transects étant disponibles tous les 1 m.n, aucune régularisation n'a été effectuée.

Ces régularisations laissent des chutes en bout de transect. En effet si un transect mesure 10 m.n et qu'il est régularisée par blocs de 3 m.n, une chute de 1 m.n restera en bout de transect. La règle opérée a été la suivante, si la chute était inférieure à la moitié du support de régularisation, la chute était intégrée au bloc précédent, si la chute était supérieure, elle constitue un bloc à elle seule. Les données acoustiques régularisées ne sont donc pas toutes exactement au même support.

Compatibilité de l'acoustique en stations et en transects

Comme on l'a vu, la présence du chalut en mer est un élément perturbateur pour les poissons. On peut donc se demander si l'acoustique observée en stations et celle observée le long des transects, une fois mise à des supports comparables peuvent être considérées comme une seule et même variable collectée en des points différents de l'espace. Une fois encore, cette hypothèse est indispensable si l'on veut par la suite utiliser les données des transects pour extrapoler le phénomène observé en stations. Une étude précise des moyennes, des variances et des structures locales en fonction de la profondeur a été effectuée et montre qu'aucune différence systématique ne peut être dégagé (Bez et al., 2004 [7]).

Les poissons semblent pourtant bien réagir à la présence du chalut en mer et se déplacer durant la station. Une autre approche montre en effet que le chalut collecte, en Norvège, des poissons très haut dans la colonne d'eau (cf. partie 3.2).

Ce phénomène s'explique si l'on distingue deux types d'effets : celui du bateau et celui du chalut.

Les résultats obtenus sont en effet cohérents si l'on suppose que le passage du bateau n'a qu'un effet négligeable sur le comportement des poissons alors que l'approche du chalut est très perturbatrice et tend à faire plonger les poissons vers le fond de la mer. Comme on l'a vu, le chalut est souvent loin derrière le bateau. L'effet retard (cf. figure 1.9) peut aller jusqu'à 15 minutes. Il est donc possible que les poissons ne réagissent à l'arrivée du chalut qu'après que le cône acoustique soit passé. Les données acoustiques n'enregistrent alors aucune perturbation et l'acoustique en stations est tout à fait cohérente avec l'acoustique enregistrée le long des transects.

1.2 Statistiques globales

Les statistiques élémentaires montrent des variations des moyennes inter-annuelles en Norvège assez faibles, surtout pour les pélagiques (cf. tableau 1.2). Les coefficients de variation (écart-type/moyenne) par contre sont très grands (entre 2 et 9). Les variances des variables sont particulièrement grandes en mer du Nord. On voit que les coefficients de variation dépassent souvent 3 voire plus, ce qui indique des distributions très dispersées autour de la moyenne. Enfin, on peut voir que de très fortes valeurs, d'un ordre de grandeur jusqu'à plus de 1000 fois la moyenne des données sont parfois rencontrées aussi bien par l'acoustique que par chalutage.

En mer d'Irlande, enfin, les coefficients de variations sont plus faibles dans l'ensemble, même s'ils restent supérieurs à 1.

1.2.1 Distributions asymétriques

Comme souvent pour ce qui est de données halieutiques, les distributions des variables sont très dissymétriques. En mer du Nord, pour les campagnes françaises, 3% des stations concentrent à la fois 65% de l'énergie totale collectée en stations et 26% des captures totales de pélagiques. En mer de Barents, sur 5 des 6 campagnes disponibles, ce sont les données de captures les plus asymétriques : 60% des captures totales se trouvent dans seulement 3% des stations. Il y a peu de valeurs nulles à proprement parler mais énormément de valeurs très proches de zéro.

Petitgas (1991, [50]), Guiblin (1997, [32]) et Bez (1997, [6]) remarquent des distributions similaires sur des données acoustiques de harengs et des captures par chalutage de fond de merlu. Le problème de l'asymétrie des distributions vient accentuer la problématique d'hétérotopie posée. Il existe en géostatistique de nombreuses méthodes pour traiter de telles distributions, consistant généralement à effectuer une transformation des variables initiales, par exemple lognormale, ou par un découpage en indicatrices. En effet tout modèle linéaire, géostatistique ou non, est très sensible aux valeurs fortes qui, bien que rares, ont des comportements mal connus, différents des autres valeurs, qui influent énormément toute statistique et cachent les comportements des valeurs basses. Cependant, ce sont elles qui font l'abondance. Il est donc délicat de les mettre de côté pour ne s'intéresser qu'aux valeurs faibles ou moyennes. Le choix de la valeur de seuillage est difficile (cf. annexe A). Décider qu'au delà d'une certaine limite, il ne s'agit plus du même phénomène est vite assez arbitraire pour un non spécialiste.

Les distributions semblent souvent obéir à des lois lognormales (cf. figure 1.16). Cependant, notre

Campagne				Acoustique (somme des layers 1 à 10) en stations				Démersaux (ENASC)				Pélagiques (ENASC)			
Pays	Année	Nb de stations	Nb d'ESDU en radiales	Max	Moy	Var	CV	Max	Moy	Var	CV	Max	Moy	Var	CV
Barents	1997	176	5209	488	20.4	1860	2.10	1417.2	71.07	13472.5	1.63	36.07	2.42	39.58	2.59
Barents	1998	198	5135	4198	33.5	88380	8.9	468.5	41.72	2732	1.25	873.6	5.62	3844.9	11.02
Barents	1999	223	5567	1650	24.7	18199	5.4	1114.3	44.6	11774	2.42	4281	23.2	82138	12.3
Barents	2000	302	7680	1866	29.5	22363	5.0	380.3	38.5	2205	1.22	588	6.5	1680	6.2
Barents	2001	300	7666	508	24.8	2353.7	1.95	1318.3	56.3	9950	1.77	514	10.3	2385	4.7
Barents	2002	287	7383	1605	21.6	10154	4.65	239	47.8	1975	0.92	679	5.41	1688	7.6
Mer du N.- FRS	1999	44	468	918.3	31.6	18483	4.29	197.6	23.7	1526.9	1.64	28.32	1.37	21.51	3.38
Mer du N.- FRS	2000	46	351	9580.3	241	2157837	6.09	185.4	37.68	1035.3	0.85	51.12	1.91	63.14	4.14
Mer du N.- FRS	2002	47	430	865.4	34.59	15768.3	3.63	182.16	31.07	1265.6	1.14	47.3	2.38	63.3	3.34
Mer du N.- FRS	2003	46	303	1018.7	67.378	42675.6	3.06	220.2	28.28	1268	1.26	12.29	1.05	5.78	2.27
Mer du N.- CEFAS	2000	71	1038	4850.7	607.2	662911	1.34	40	3.99	31.38	1.40	10.9	0.81	3.1	2.18
Mer du N.- CEFAS	2001	70	883	1056,8	219.6	49923	1.01	19.2	3.51	12.9	1.02	25.3	0.91	10.56	3.56
Mer du N.- CEFAS	2002	70	1140	7344.2	848.8	2328812	1.79	4.59	0.87	0.86	1.06	31.4	0.8	14.53	4.33
Mer du N.- IFRE-MER	2002	77	440	724.9	58	11262	1.83	1030.8	57	16061.6	2.22	629.5	37.9	8588	2.44
Mer du N.- IFRE-MER	2003	82	722	7139.2	232.3	1046415	4,4	786.3	69.7	20223.5	2.04	345.3	28.9	3056.4	1.9
Irlande	2000	37	110	1910	205.3	146188	1.86	227	54.2	2566	0.9	426	24.4	4868	2.8
Irlande	2001	34	236	4794	825.6	1336403	1.4	930	133.9	30106	1.3	940	62.6	28515	2.7
Irlande	2002	41	173	553	55.9	11300	1.9	227	49.7	2611	1.0	419	20.9	4373	3.1

TAB. 1.2 – Statistiques élémentaires de chacune des campagnes.

problème est une approche multivariable et le lognormal à deux dimensions devient difficile à vérifier. La transformation lognormale utilisée dans la figure 1.16 est $f(z) = \log(1 + z)$, ce qui permet d'éviter le problème des valeurs nulles qui restent alors inchangées. Le lognormal translaté : $f(z) = \log(1 + \frac{z}{\text{moyenne}(z)})$ permet de relativiser l'impact des valeurs fortes tout en déformant moins les données qu'une transformation lognormale brute. Cependant, elle ne permet pas ici de réduire assez la dispersion des données (cf. figure 1.16). Cette méthode a donc été très peu utilisée mis à part dans les explorations de données pour mettre en évidence certains comportements spécifiques ou pour permettre de comparer les modèles proposés avec une approche différente.

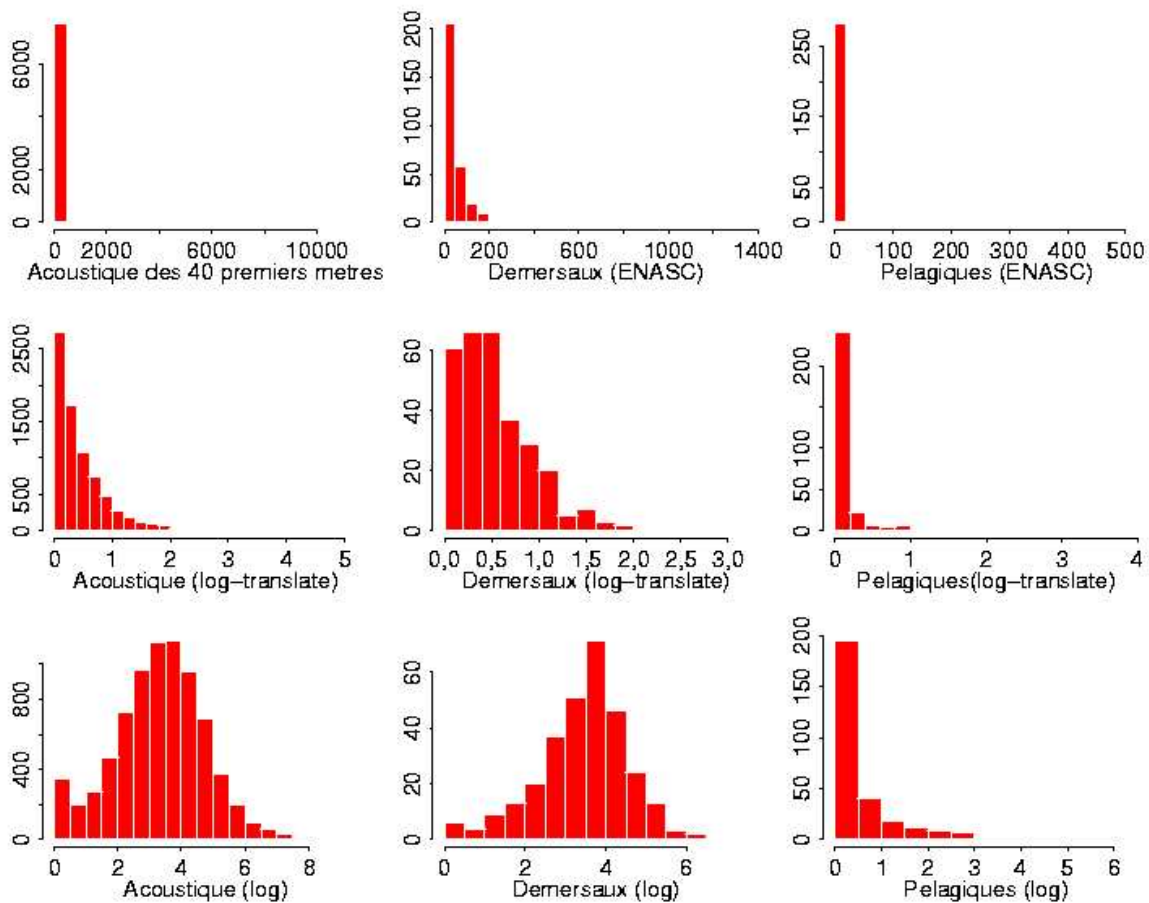


FIG. 1.16 – Histogrammes de l'acoustique en radiales, NASC-équivalent des démersaux et du NASC-équivalent des pélagiques en brut (première ligne), en log translaté : $\log(1+Z/m)$ (ligne centrale) et en log : $\log(1+Z)$ (ligne du bas) en Norvège 2001.

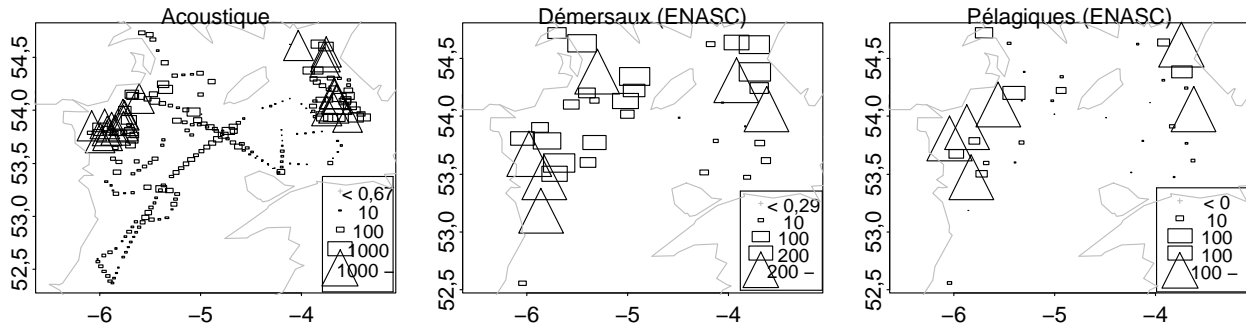


FIG. 1.17 – Représentation proportionnelle de l’acoustique en radiales (somme des 5 premiers layers) à gauche, des captures démersales (ENASC) au milieu et pélagiques (ENASC) à droite. Les rectangles sont d’autant plus grands que la valeur enregistrée est forte. Les valeurs les plus fortes sont indiquées par les triangles. Irlande 2001

1.2.2 Localisation des valeurs fortes

Une grande part de l’abondance étant contenue dans quelques points, il est important de voir si ces valeurs se situent dans les mêmes zones géographiques pour l’acoustique et le chalutage. Lorsque les valeurs extrêmes de chaque variable ne sont pas cooccurrentes ponctuellement, mais apparaissent dans les mêmes zones, on peut s’intéresser à une échelle plus grossière de régions riches ou pauvres. Si les deux types de mesure ne permettent pas de localiser le stock dans les mêmes régions, il sera en revanche difficile de les combiner.

En Irlande, où l’abondance se situe principalement sur le plateau continental, les données acoustiques et les données de captures sont assez cohérentes (cf. figure 1.17). En mer de Barents, les zones riches de l’acoustiques semblent également bien répondre aux zones riches observées par les captures démersales (cf. figure 1.18). En mer du Nord, par contre, les valeurs fortes observées par acoustique et par capture sont localisées en des zones très différentes de l’espace (cf. figure 1.19). L’inférence d’un modèle combinant les deux sources d’information sera donc beaucoup plus difficile à mettre en oeuvre en mer du Nord où les deux types de données correspondent mal.

1.3 Conclusion et choix méthodologiques

L’objectif de mon travail est d’améliorer les estimations d’abondances obtenues actuellement par chalutage de fond en utilisant l’information acoustique connue de façon beaucoup plus dense. Pour cela des données de mer d’Irlande, de mer du Nord et de mer de Barents sont disponibles. Les données collectées dans chacune de ces trois mers présentent des caractéristiques statistiques, écologiques et environnementales très différentes. En mer de Barents, la taille des campagnes permet de réunir un nombre très important de données. L’acoustique y semble relativement bien corrélée aux captures. En mer du Nord, par contre, les valeurs fortes de l’acoustique correspondent souvent assez peu spatialement aux régions indiquées comme étant riches par les données de capture. Les deux types de données fournissent alors des informations contradictoires difficiles à combiner. En

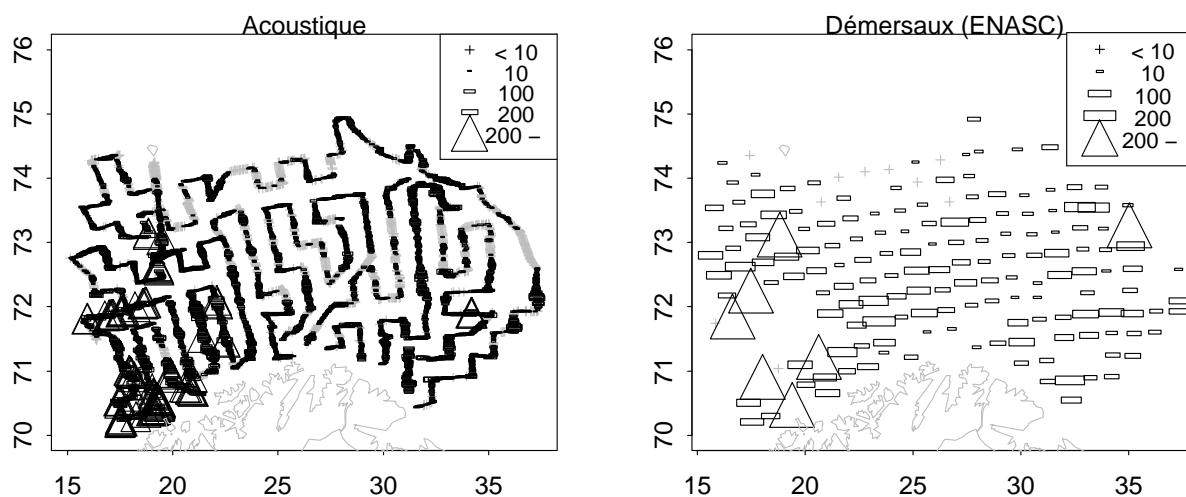


FIG. 1.18 – Représentation proportionnelle de l’acoustique en radiales (somme des 5 premiers layers) à gauche, des captures démersales (ENASC) à droite. Les rectangles sont d’autant plus grands que la valeur enregistrée est forte. Les valeurs les plus fortes sont indiquées par les triangles. Norvège 1997.

mer d’Irlande, enfin, les enregistrements acoustiques semblent assez bien corrélées aux captures pélagiques, mais le nombre de stations disponibles pour chaque campagne est souvent faible, rendant l’inférence d’un modèle statistique parfois instable. L’ensemble de ces campagnes présentent des données d’acoustique et de captures ayant des distributions statistiques fortement asymétriques. Matheron (1978, [42]) indique quatre critères que doit vérifier un modèle satisfaisant : celui-ci doit être opérationnel, spécifiable, compatible avec les données et efficace. Il insiste en particulier sur le *principe d’économie* indiquant qu’un modèle doit être suffisamment spécifié pour résoudre la problématique, mais il est inutile qu’il le soit trop.

Dans ce contexte, j’ai cherché à établir un modèle permettant de s’adapter à l’ensemble des campagnes. Le modèle doit être suffisamment simple et général pour couvrir l’ensemble des configurations présentes dans les données.

Malgré l’asymétrie des distributions statistiques, j’ai essayé d’établir un lien entre les variables brutes, sans les transformer (par passage au logarithme, par exemple), non seulement en raison des difficultés liées au retour à une estimation de la variable brute, mais aussi parce que la transformation des données n’améliore pas la corrélation entre les deux variables de façon notable. En mer du Nord, par exemple, l’analyse des tableaux de contingence croisée montre en particulier que les valeurs fortes de captures ne correspondent pas à des valeurs fortes d’acoustique. Il en est de même avec les valeurs moyennes et faibles. Dans cette zone, le lien croisé acoustique-capture ne pourra être amélioré par une transformation des données (par une fonction croissante).

Il est cependant important de garder à l’esprit cette asymétrie au long de l’étude. L’influence des valeurs fortes, représentant une grande part de l’abondance totale, explique certains choix méthodologiques, comme le découpage en indicatrices de la variable acoustique (cf. partie 2.7) et permet de mieux comprendre et interpréter les résultats.

La suite de la thèse se découpe en trois parties. Le chapitre suivant étudie le lien structural entre la variable acoustique et la variable de pêche ainsi que le comportement de différents modèles bivariés.

L'acoustique est en fait disponible par couches sur toute la hauteur d'eau. Le troisième chapitre étudie donc le lien entre acoustique et captures le long de la verticale. Ce chapitre vise à simplifier la modélisation pour se ramener à une configuration bivariable, telle qu'elle a été vue dans le deuxième chapitre. Les résultats présentés dans ce troisième chapitre ont donc été réalisés avant les développements du deuxième chapitre. J'ai cependant choisi de les présenter après en raison de la particularité de cette étude de la verticale aux jeux de données utilisés.

Un quatrième chapitre résume ensuite les différents résultats obtenus et compare les performances de différents modèles.

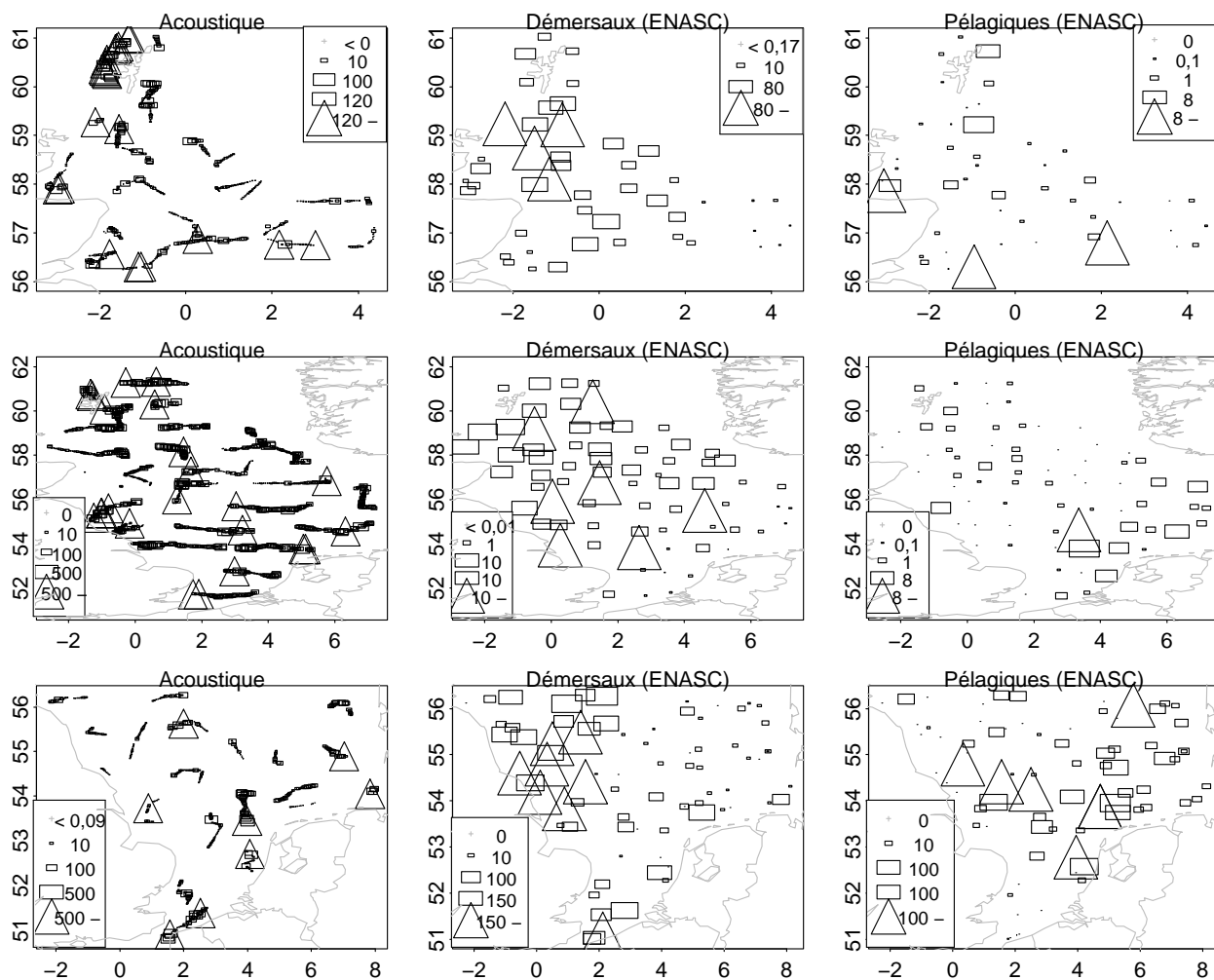


FIG. 1.19 – Représentation proportionnelle de l'acoustique en radiales (somme des 5 premiers layers) à gauche, des captures démersales (ENASC) au milieu et pélagiques (ENASC) à droite. Les rectangles sont d'autant plus grands que la valeur enregistrée est forte. Les valeurs les plus fortes sont indiquées par les triangles. Écosse 2002 (haut), CEFAS 2000 (milieu), IFREMER 2002 (bas).

Chapitre 2

Rôle de la variable acoustique dans différents modèles bivariables

“... it seems logical to advance that above combined arguments along with experimental results described in our work are likely to support the hypothesis of a semi-linear multi-stable multi-switching net back feed-work organization of the YR [Yelling Reponse] whose a tentative anatomical model can therefore be proposed.”
Georges Perec, *Experimental demonstration of the tomatotopic organization in the soprano, (Cantratrix sopranica L.)*

Ce chapitre étudie le comportement de différents modèles dans une configuration bivariable où la variable d'intérêt est la variable de pêche et la variable auxiliaire les enregistrements acoustiques.

Dans toute cette partie, la variable acoustique utilisée est la somme des 40 premiers mètres au-dessus du fond pour les campagnes norvégiennes, la somme des 4 premiers mètres pour les campagnes irlandaises et la somme de 5 premiers mètres pour les campagnes en mer du Nord. Ce choix est justifié dans le chapitre suivant par une étude détaillée du lien entre chaque couche acoustique et la variable de pêche.

La variable acoustique est échantillonnée de façon beaucoup plus dense que la variable de pêche. Après une analyse structurale de chacune des deux variables et de leur lien spatial pour ce type d'échantillonnage, la mise en pratique du cokrigeage est analysée pour différentes configurations de structures rencontrées dans les données : le cas où la structure croisée entre les variables est proportionnelle à la structure de la variable auxiliaire (modèle à résidu) et lorsque toutes les structures sont proportionnelles (corrélation intrinsèque). Plusieurs résultats présentés dans cette partie ont fait l'objet d'un article (cf. Bouleau et Bez 2004, [9]).

Les résultats sont également confrontés avec ceux du krigeage à dérive externe, assez proche analytiquement du modèle à résidu et très utilisé en pratique pour ce type d'échantillonnage.

Enfin, un modèle basé sur une estimation non-linéaire de la dérive est présenté pour palier la forte asymétrie des données, en particulier pour les campagnes en mer du Nord. Certaines observations réalisées lors de cette dernière partie ont également été présentées lors d'une conférence (cf. Bouleau et al. 2003, [10])

In this chapter, different bivariate models are studied. The target variable is the trawl catches when the acoustic recordings represent the auxiliary variable. Throughout this section, the acoustic variable used is the sum of the first 40 meters above the bottom for the Norwegian surveys, the sum of the first 4 meters above the bottom for the Irish surveys, and the sum of the first 5 meters above the bottom in the North Sea. This choice is justified in the next chapter by a detailed study of the relationship between the catch and the acoustic layers.

The acoustic variable is very densely sampled compared to the catch variable. Firstly, a structural analysis is carried out for each variable and for their spatial correlation, for this kind of sampling. Then, practical problems about the adjustment of cokriging are discussed for different spatial structures configurations observed in the data : the case where the spatial correlation between the two variables is proportional to the spatial structure of the acoustic (Markov-type model) and the case where all the spatial structures are proportional (model in intrinsic correlation).

Some of the results presented here have been summarized in a paper (cf. Bouleau and Bez 2004, [9]).

The results are also compared with those produced by the external drift model, which is very similar analytically to the Markov-type model and often used for this type of sampling.

Then, a non-linear method is used in order to reduce the outliers impact, in particular for the North Sea surveys. The problematics raised in this section were presented at a conference (cf. Bouleau and al. 2003, [10]).

Notations

$P(x)$ est la variable principale mesurée au point x (la variable de pêche), $x \in S$, l'ensemble des stations.

$A(x)$ est la variable auxiliaire, dans le cas d'étude la variable acoustique, $x \in R$, l'ensemble des radiales.

L'ensemble des stations est inclus dans l'ensemble des radiales : $S \subset R$. Le nombre de stations est noté n_S et le nombre de radiales n_R . Lorsqu'une distinction est nécessaire la variable acoustique restreinte aux points en stations est notée $A_S(x)$, $x \in S$ et la variable acoustique sur l'ensemble des points restants (les transects) $A_{\bar{S}}(x)$, $x \in \bar{S}$ avec $S \cup \bar{S} = R$ et $n_S + n_{\bar{S}} = n_R$. Ces notations pourront également concerner la variable principale pour distinguer les valeurs connues en stations P_S des valeurs estimées entre les stations, le long des transects, $P_{\bar{S}}$.

Attention, la notation \bar{Z} signifie la moyenne arithmétique de Z , lorsque Z est une variable aléatoire, mais \bar{S} , lorsque S est un ensemble (comme les stations) signifie l'ensemble complémentaire de S .

2.1 Introduction

Traditionnellement, trois types d'échantillonnages sont distingués en géostatistique (Wackernagel 1998 [70]) :

- le cas entièrement hétérotopique où les deux variables sont mesurées sur des ensembles disjoints
- le cas partiellement hétérotopique : il existe certains points où les deux variables sont disponibles
- le cas isotopique où les deux variables sont échantillonnées aux même points.

On se place ici dans un cas particulier de la situation partiellement hétérotopique pour laquelle l'ensemble de points où la variable d'intérêt est disponible, est inclus dans l'ensemble de points associé à la variable auxiliaire. Ce type d'échantillonnage est très courant, en particulier lorsqu'une variable auxiliaire corrélée à la variable d'intérêt peut être échantillonnée de façon très peu coûteuse. Cochran (1977 [16]) présente ce type de sondage sous le nom de "double sampling". L'objectif est d'améliorer l'estimation de la variable d'intérêt en utilisant la corrélation avec la variable auxiliaire mieux connue.

En géostatistique, en cas de sous-échantillonnage de la variable principale, le cokrigeage permet généralement d'améliorer l'estimation grâce aux informations supplémentaires apportées par la variable auxiliaire (Goovaerts 1998 [30]). Le krigeage à dérive externe et le cokrigeage collocalisé sont également des modèles largement utilisés lorsque la variable auxiliaire est connue de façon quasi-exhaustive. Ces modèles seront ici comparés pour différentes configurations de données.

La figure 2.1 montre l'évolution des enregistrements acoustiques le long d'un transect, c'est à dire d'un segment choisi sur le trajet du bateau, auxquelles sont superposées les données de captures démersales (en ENASC). Les localisations des coups de chalut sont indiquées par des lignes verticales. Quatre configurations différentes prélevées au sein de la même campagne (Norvège 1998) illustrent les différents comportements possibles de chacune des deux variables. Sur le transect 1, l'acoustique est très variable et les stations semblent passer au travers des variations locales, tout comme sur le transect 2 où une seule station (placée entre le 37° et 38°) tombent à la fois sur un maxima local de l'acoustique et des captures. Les transects 3 et 4 montrent des évolutions plus lentes de l'acoustique. Le transect 3 présente une dérive nord-sud et le transect 4 une zone riche d'extension géographique d'environ 30 m.n, comme le confirment les variogrammes expérimentaux correspondants (figure 2.2). La tendance indiquée par l'acoustique se retrouve globalement sur les captures pour le transect 3, alors que les reliefs du transect 4 se superposent mal, en raison d'une seule station (vers le 73°) largement sous-évaluée par rapport à l'acoustique. L'aspect très erratique des données acoustiques par rapport aux données de captures est lié à l'échantillonnage. Les deux variables ont été mises à des supports équivalents (cf. partie 1.1.5), il est donc naturel de penser que si la variable de pêche était plus échantillonnée, elle atteindrait une variabilité équivalente à celle de l'acoustique.

Ces quatre exemples montrent différentes configurations possibles des données. Ils seront repris en exemple tout au long du chapitre pour illustrer les comportements des différents modèles.

L'échantillonnage n'étant pas préférentiel, les stations peuvent passer au travers des zones d'abondance locale. L'acoustique apporte alors une information riche qu'il est important de prendre en compte. Il peut également arriver que les valeurs en stations de capture et d'acoustique ne soient pas localement cohérentes, en particulier à cause des déplacements de poissons (cf. partie 1.1.4). Dans ce cas, il convient de savoir si l'on préfère un modèle qui privilégiera l'information acoustique ou l'information de pêche. La variable cible étant la variable de pêche, la quantité que l'on cherche à estimer est une abondance *capturable*. D'un autre côté, les valeurs acoustiques ne semblent généralement pas provenir d'une erreur du sondeur, une structure réelle semble exister le long des transects. Les valeurs fortes, comme les faibles, sont fiables. Un bon modèle doit permettre de reproduire les zones d'abondances locales enregistrées par l'acoustique tout en utilisant l'information de capture pour recalibrer l'estimation pour obtenir une abondance capturable. Le choix du rôle donné à chacune des variables dans le modèle est donc une question primordiale qu'il convient de contrôler. Les quatre exemples mentionnés plus haut montrent la difficulté inhérente aux données de caler un modèle combinant les deux sources d'informations.

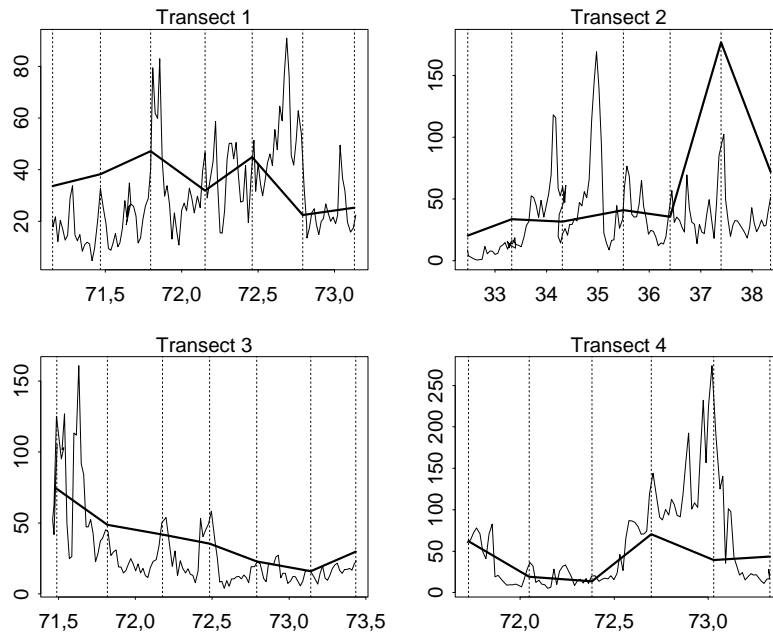


FIG. 2.1 – Représentation de l'acoustique en radiales et des captures de démersaux en ENASC (en gras) pour 4 sections de transect au sein de la même campagne. Norvège 1998

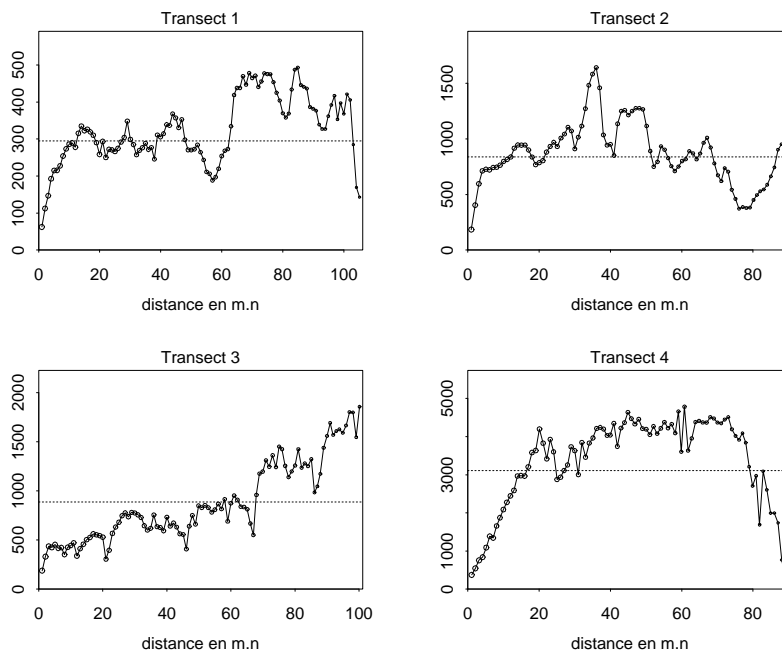


FIG. 2.2 – Variogrammes le long de la route du navire pour les 4 configurations précédentes. Norvège 1998

Dans ce chapitre, je commence par décrire une situation de régression classique dans laquelle toutes les observations sont supposées indépendantes. Le cas général du cokrigeage est présenté ensuite, permettant de prendre en compte des structures spatiales simples et croisées complexes. Deux cas de simplifications sont développés : le modèle à résidu, correspondant au cas où l'acoustique est la variable maîtresse, sa structure étant proportionnelle à la structure croisée avec les captures ; et le modèle en corrélation intrinsèque dans lequel toutes les structures sont proportionnelles les unes aux autres. Le rôle de la variable acoustique dans ce cas est alors tout à fait similaire au cas de régression introduit au début du chapitre. Le krigeage à dérive externe, très utilisé en pratique dans les cas où la variable auxiliaire est connue de façon très dense, est alors mis en parallèle avec le modèle à résidu, qui en est formellement proche. Dans une deuxième partie, une approche par krigeage à dérive externe dans lequel la dérive acoustique est estimée de façon non linéaire (par découpage en indicatrices) est présentée. Cette méthode permet d'obtenir une estimation plus robuste de l'acoustique et d'obtenir ainsi des résultats dans des zones géographiques où les données sont très dissymétriques, comme la mer du Nord.

La régression linéaire pour une variable cible sous-échantillonnée

Il est intéressant dans un premier temps de considérer une régression linéaire classique ne prenant pas en compte les corrélations spatiales. Ce cas équivaut à un modèle géostatistique dont les structures de chaque variable ainsi que leur croisée seraient pépitiées.

Si l'on considère le modèle de régression linéaire suivant :

$$P_i = \alpha A_{S,i} + \beta + \varepsilon_i$$

avec $\mathbb{E}(\varepsilon_i) = 0$; $\text{Var}(\varepsilon_i) = \sigma^2$; et $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ si $i \neq j$, $i, j \in S$.

Ce modèle extrêmement simple permet de prédire la quantité pêchée au point où seule l'acoustique est disponible.

Si l'on estime α et β par moindres carrés, on obtient :

$$\hat{\alpha} = \frac{\sum_i^{n_S} (A_{S,i} - \bar{A}_S)(P_i - \bar{P}_S)}{\sum_i^{n_S} (A_{S,i} - \bar{A}_S)^2}$$

et

$$\hat{\beta} = \bar{P}_S - \hat{\alpha} \bar{A}_S$$

avec $\bar{A}_S = \frac{1}{n_S} \sum_i^{n_S} A_{S,i}$ et $\bar{P}_S = \frac{1}{n_S} \sum_i^{n_S} P_i$ les moyennes arithmétiques de A et P sur l'ensemble S. Les coefficients sont estimés sur la sous-population pour laquelle les deux variables sont disponibles. On peut alors prédire P en un point où seule A est connue, c'est à dire le long d'un transect, par la relation :

$$\hat{P}_{\bar{S},i} = \hat{\alpha} A_{\bar{S},i} + \hat{\beta}$$

en remplaçant $\hat{\beta}$ par sa valeur on obtient :

$$\hat{P}_{\bar{S},i} = \bar{P}_S + \hat{\alpha}(A_{\bar{S},i} - \bar{A}_S)$$

Si l'on s'intéresse à l'estimation de la moyenne des captures sur l'ensemble des échantillons, on se sert de la régression établie précédemment. L'estimation globale de la moyenne s'écrit donc :

$$\begin{aligned}\hat{M} &= \frac{1}{n_R} \sum_{i \in R} \hat{P}_i \\ &= \frac{1}{n_R} \sum_{i \in R} (\bar{P}_S + \hat{\alpha}(A_{R,i} - \bar{A}_S)) \\ &= \bar{P}_S + \hat{\alpha}(\bar{A}_R - \bar{A}_S)\end{aligned}\tag{2.1}$$

Notons que cette estimation s'obtient en appliquant le modèle de régression sur tous les points : aussi bien ceux où les captures ne sont pas connues que pour les stations, où une valeur exacte de pêche est pourtant disponible. La régression n'est pas une interpolation exacte. Si l'on calculait un estimateur de la moyenne \hat{M} en utilisant les données de pêches là où elles sont disponibles et le modèle de régression ailleurs (le long des transects), la formule de l'estimateur serait alors différente. On aurait :

$$\hat{M} = \bar{P}_S + \hat{\alpha} \frac{n_{\bar{S}}}{n_R} (\bar{A}_{\bar{S}} - \bar{A}_S)$$

Dans les applications, c'est la formulation usuelle de Cochran 2.1 qui a été utilisée.

L'estimation de la moyenne de P sur l'échantillon total est égale dans ce cas de figure à sa moyenne sur le sous échantillon S plus un terme correctif estimé à partir de la variable auxiliaire A. La variance de cet estimateur peut-être approchée par (Cochran 1977 [16]) :

$$V(\hat{M}) = \frac{S_P^2}{n_S} (1 - \rho_S^2) + \frac{S_P^2}{n_R} \rho_S^2\tag{2.2}$$

avec $S_P^2 = \frac{1}{n_S - 1} \sum_i (P_i - \bar{P})^2$ la variance empirique de P et ρ_S le coefficient de corrélation empirique entre P et A, calculé sur S.

On voit dans cette formule que si le coefficient de corrélation entre les deux variables est grand (positif ou négatif), un point d'échantillonnage supplémentaire de l'acoustique apportera une baisse significative de la variance, alors que si ρ_S est faible, le nombre de points disponibles le long des transects importe peu.

Ainsi, pour une estimation par régression linéaire classique, en supposant les deux variables pépitiqes, l'acoustique joue un rôle correcteur permettant de réduire l'erreur faite lorsque l'on utilise les seules données en stations (cf. équation 2.1).

2.2 Analyse structurale

Cette partie étudie la relation spatiale existant entre la variable de pêche et la variable acoustique. Deux approches semblent possibles pour étudier ce lien : se restreindre à l'intersection des deux ensembles de points, c'est à dire analyser le lien existant sur la zone où les deux variables sont isotopiques ; ou utiliser des outils descriptifs hétérotopiques qui permettent d'analyser toute l'information disponible. Dans le premier cas, le lien établi en stations sera ensuite très extrapolé, ce qui induit une hypothèse forte d'homogénéité des variables sur l'ensemble du champ. Cette approche

devient délicate à mettre en oeuvre lorsque l'intersection des deux ensembles est petite. Cependant, pour plusieurs modèles, c'est l'adéquation entre les données aux points de stations qui va déterminer le poids donné à la variable auxiliaire.

2.2.1 Sur le sous-ensemble de points en isotopie

De nombreux outils permettent de comprendre les liens entre les deux variables étudiées lorsque celles-ci sont échantillonnées aux mêmes points.

Les nuages de corrélations montrent des corrélations assez fortes en Norvège 1997, 1998 et 2001, en raison de valeurs fortes alignées au reste du nuage (figure 2.3). Sur les nuages de corrélation des variables log-transformées, on voit un lien linéaire bruité, mais net, entre les deux variables pour les données de la mer de Barents (cf. figure 2.4). En mer d'Irlande, l'acoustique est mieux corrélée avec les pélagiques, qui pourtant ne représentent pas la part la plus importante des captures. En mer du Nord, les nuages de corrélation ne montrent aucun lien que ce soit sur les variables brutes ou les variables log-transformées.

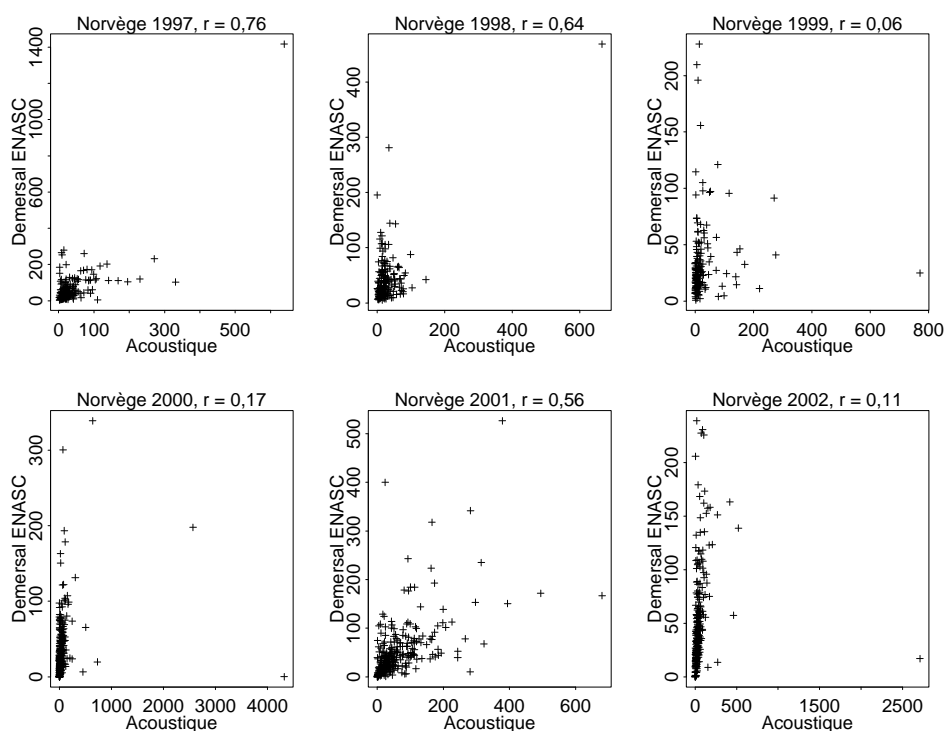


FIG. 2.3 – Nuages de corrélations de l'acoustique avec les démersaux (en ENASC), Norvège.

Le variogramme croisé ne peut comparer les variables que si elles sont échantillonnées aux mêmes points. Pour certains modèles, comme le modèle en corrélation intrinsèque, il est important de comparer les structures spatiales simples et croisées, pour savoir s'il est raisonnable ou non de les supposer proportionnelles. Il peut être utile alors de comparer les structures sur l'ensemble de points en isotopie, en plus d'une analyse du lien sur la totalité des points à cause de la difficulté

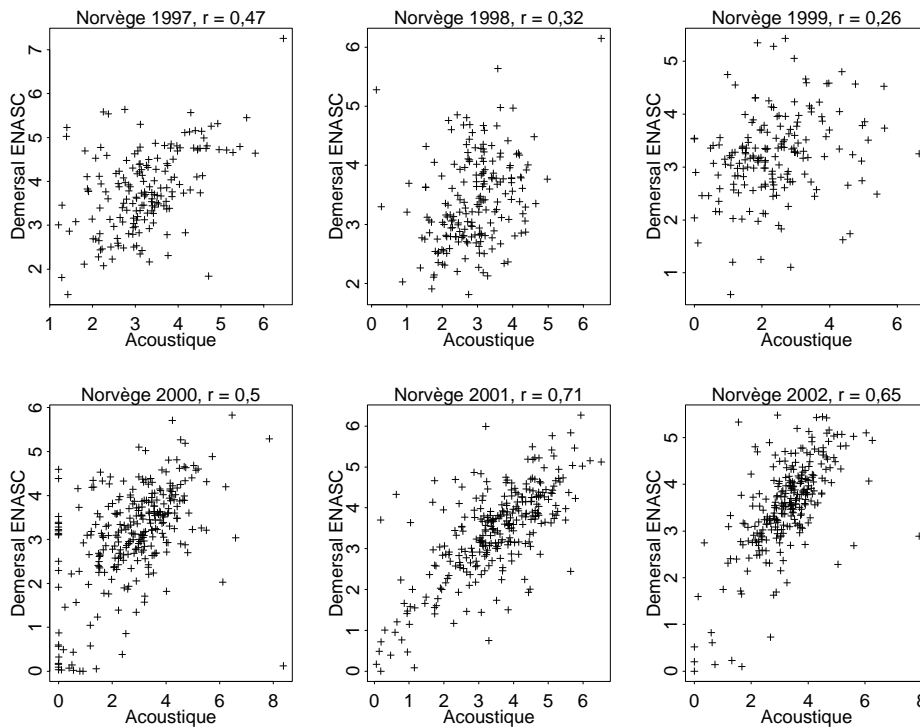


FIG. 2.4 – Nuages de corrélations de l’acoustique avec les démersaux (en ENASC), Norvège. Les variables sont transformées par passage au LOG (équivalent à la figure 2.3 après passage au LOG)

à comparer des variogrammes effectués à des échelles différentes (pas de 1 m.n pour les radiales et 20 m.n pour les stations). Savoir que les deux variables se comportent de façon très similaire sur le sous-ensemble des stations, peut-être une aide utile à la décision.

En raison des valeurs fortes, les variogrammes simples et croisés sont assez bruités (figure 2.5). Le variogramme faisant intervenir les carrés des écarts de valeurs entre les points d’une certaine distance, il est très sensible aux valeurs fortes.

Ici, les variogrammes tracés dans plusieurs directions ne présentant pas de différence nette, les modèles ont été développés de façon omnidirectionnelle.

Les variogrammes de l’acoustique montrent souvent une structure très petite portée. En Irlande en particulier des portées de l’ordre de 3 à 4 m.n se dégagent nettement. Cependant, pour un besoin de cohérence avec le support irréductible des captures, les données acoustiques ont été régularisées à un pas rendant ce type de structure invisible (3 m.n). La variable cible étant la capture, une estimation sur des blocs inférieurs à son support n’aurait pas réellement de sens (cf. partie 1.1.5). De façon générale sur l’ensemble des campagnes, l’acoustique présente une structure petite portée entre 10 et 20 m.n, parfois accompagnée d’un comportement linéaire, en particulier en Norvège, en raison de zones riches en bordure du champ. Ces valeurs sont liées à des phénomènes d’agrégation de bancs dans les zones où la température avoisine zéro (cf. annexe A). Toutefois pour plusieurs campagnes en mer du Nord, en particulier dans la zone sud, correspondant aux campagnes françaises, aucune portée même courte ne parvient à se dégager. Même après une transformation des variables en log, ou log-translaté, les variogrammes sont complètement pépitiqes. Certaines valeurs

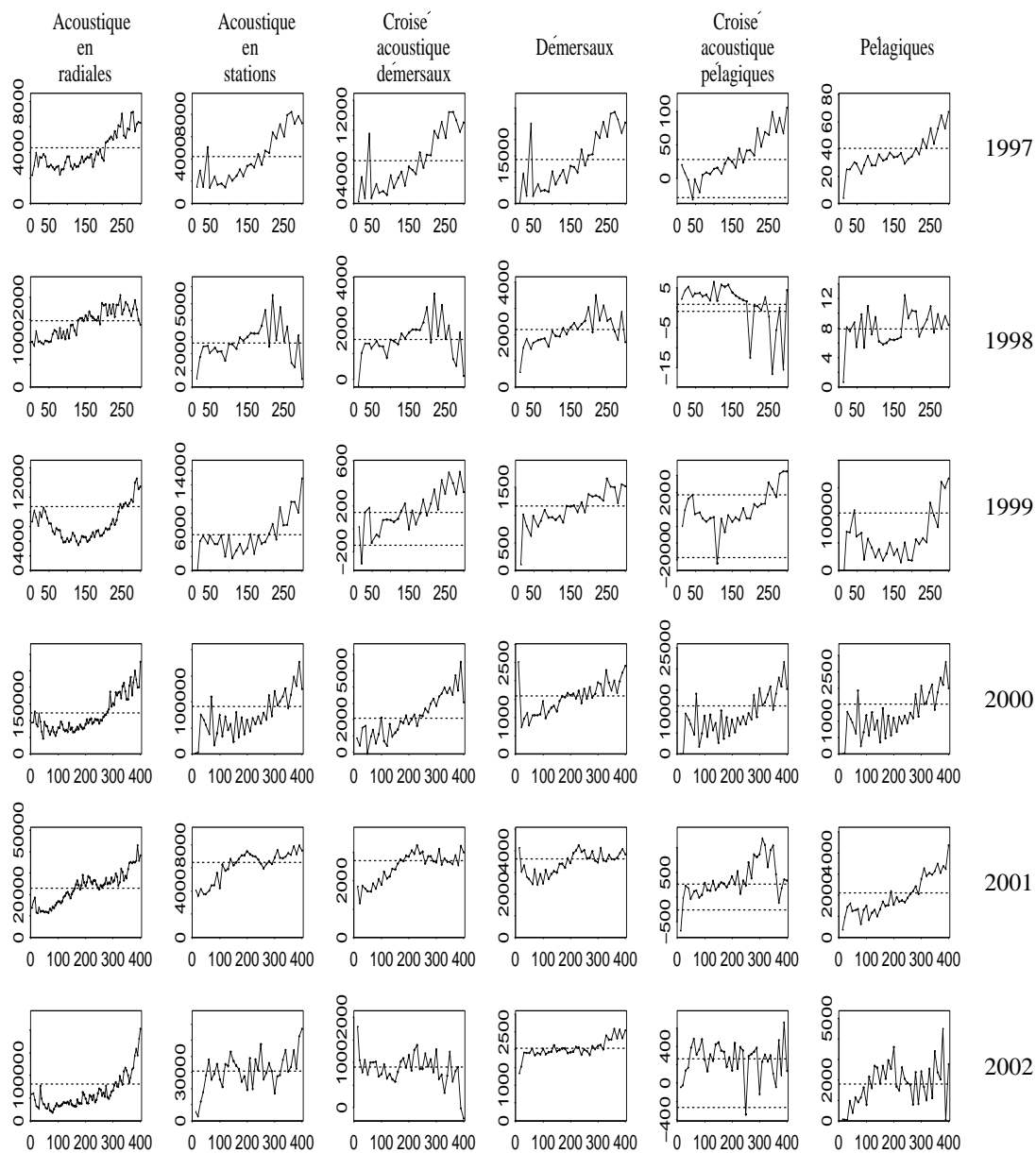


FIG. 2.5 – Variogrammes de l'acoustique (colonne 1 et 2) et variogrammes croisés avec les démersaux en ENASC (colonne 3) et avec les pélagiques en ENASC (colonne 5), pour chacune des six campagnes norvégiennes 1997-2002. Les distances sont en milles nautiques

extrêmement fortes continuent à masquer la structure des valeurs plus faibles. Il faut seuiller extrêmement bas pour s'apercevoir qu'il existe une classe de valeurs faibles structurées spatialement. Malheureusement cette classe pèse peu dans l'abondance totale.

Il n'y a qu'en mer de Barents que les variogrammes croisés présentent une structure nette entre l'acoustique et les démersaux, mais aussi avec les captures pélagiques pour les années 1997, 2000 et 2001. La structure est généralement une structure longue portée de l'ordre de 250 à 300 m.n.

Lorsque les variables sont très asymétriques, une étude de la relation entre les indicatrices de ces mêmes variables pour certains niveaux de seuillage renseigne sur les liens non-linéaires.

Le tableau de contingence des données en stations entre l'acoustique et les démersaux permet ainsi de voir qu'en Norvège, la plupart des stations riches à l'acoustique sont aussi des stations riches en capture. Les stations *pauvres* et *moyennes* correspondent également entre elles. Les cases de la diagonale du tableau sont affectées de poids importants, excepté en 1998 où le lien est moins net (tableaux non fournis). De même en Irlande, le lien entre l'acoustique et les pélagiques est fort. Par contre, en mer du Nord, les tableaux de contingence ne permettent pas de dégager de lien clair que ce soit avec les pélagiques ou les démersaux, ni même en descendant au niveau de l'espèce, aussi bien dans la zone Nord échantillonnée par l'Écosse que la zone sud. Ainsi, alors qu'un lien existe en Norvège et en Irlande, malgré le bruit induit par l'asymétrie des données, les stations riches et pauvres de la mer du Nord ne semblent pas les mêmes pour chacune des deux variables.

Cependant, les tableaux de contingence ne tiennent compte que du lien entre la capture et l'acoustique enregistrée exactement au même point. L'absence de corrélation en mer du Nord pourrait être due à un effet de décalage local entre l'acoustique et les captures. Pour cela les mêmes tableaux de contingence ont été calculés en moyennant l'acoustique sur une fenêtre entourant la station, mais les résultats obtenus restent très similaires. Ces observations viennent compléter les comparaisons des localisations des valeurs fortes faites dans la partie précédente (cf. partie 1.2.2). Non seulement les valeurs fortes de l'acoustique et les valeurs fortes des captures ne sont pas vraiment situées dans les mêmes zones de la mer du Nord, mais les valeurs moyennes et faibles non plus. Alors que la mauvaise correspondance des valeurs fortes pourrait s'expliquer par un phénomène de non capturabilité des gros bancs, le fait que les deux variables soient très peu corrélées quelque soit le niveau des valeurs rend inévitablement toute modélisation délicate.

Saporta (1990 [61]) utilise pour décrire le lien entre une variable quantitative (comme les captures) et une variable catégorielle (comme l'acoustique mise sous forme d'indicatrices avec k différents niveaux de seuillage) la relation suivante :

$$\rho_{A,P}^2 = \frac{\text{variance empirique les moyennes } \bar{P}_k}{\text{variance totale de P}}$$

où les moyennes \bar{P}_k sont les moyennes de P calculées sur les sous-ensembles déterminés par les indicatrices acoustiques. Si il n'y a aucune corrélation entre les variables, les différentes moyennes \bar{P}_k sont quasiment égales et le coefficient est faible. Réciproquement, quand la dépendance est totale, les deux variances sont presque égales et le coefficient est proche de 1. De façon générale ce coefficient a tendance à être assez faible pour l'ensemble des campagnes, excepté en Norvège 2000-2002 avec les démersaux ainsi qu'en Irlande avec les pélagiques.

2.2.2 Sur tous les points

Pour une analyse plus complète il convient de connaître les liens spatiaux simples et croisés sur l'intégralité des données disponibles.

Les nuages de points décalés d'une certaine distance h fixée entre les captures (en stations) et l'acoustique en radiales, permettent de voir comment évolue le lien entre les variables dans l'espace. En Norvège, le lien linéaire a tendance à diminuer très rapidement lorsque la distance augmente (figure 2.6). Le lag de distance est de 1 m.n. La corrélation entre les deux variables est donc locale, c'est à dire largement inférieure à la distance entre les stations. De même, en mer d'Irlande, le nuage se déstructure dès le deuxième pas de distance, la structure est de très faible portée, inférieure à la distance entre les stations.

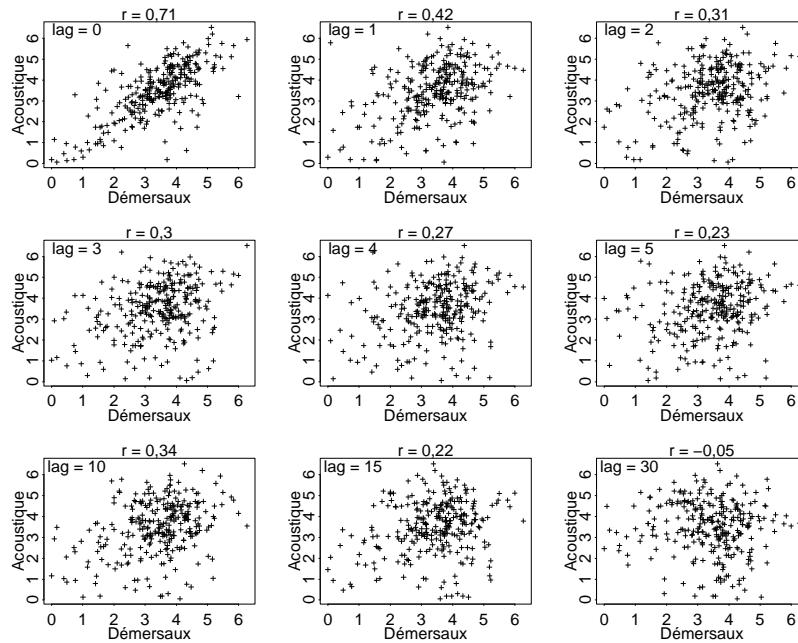


FIG. 2.6 – Nuages de corrélation entre les captures démersales en stations et l'acoustique enregistrée autour de la station, le long de la route du navire. Les deux variables sont en LOG. Un pas de distance = 1 m.n (le pas = 0 correspond à la corrélation en stations), Norvège 2001.

La covariance croisée entre deux variables s'écrit : $Cov[P(x), A(y)]$

Dans le cas où les variables sont stationnaires au second ordre, la fonction ne dépend plus que de la distance entre les points : $C(h) = Cov[P(x)A(x+h)]$. La covariance non-centrée est généralement considérée comme plus robuste parce qu'elle ne nécessite pas l'estimation des moyennes. La covariance croisée n'est pas nécessairement positive et maximale en zéro. Le corrélogramme croisé permet de réduire l'impact des différences de variances entre les deux variables (Goovaerts 1997 [29]) :

$$\rho_{P,A}(h) = corr[P(x), A(x+h)]$$

La corrélation croisée nécessite également l'estimation des moyennes et des variances. On voit sur

la figure 2.7 que les corrélogrammes sont plus réguliers que la covariance. Ils permettent un gain en stabilité. Le corrélogramme, contrairement à la covariance croisée est nécessairement maximal en zéro ce qui facilite l'étude de la structure.

Ici les corrélogrammes ont été tracés le long de la route du bateau, ce qui a tendance à augmenter la structure spatiale des données (Guiblin 1997 [32]). Une structure très courte portée de l'acoustique entre 5 et 15 m.n se dégage de nouveau pour la plupart des campagnes. Cette structure est également parfois visible, en particulier en mer d'Irlande et en Norvège, sur les corrélogrammes croisés avec les captures. Les nuages de corrélations translats confirment que le lien linéaire entre l'acoustique et les captures démersales se déstructure entre 10 et 15 m.n en Norvège. La taille de cette petite structure est de l'ordre de grandeur de celle des agrégations de bancs (Reid 2000, [55], Petitgas 2003 [52]). Malgré l'échelle de définition grossière variant entre 1 et 3 m.n, l'acoustique permet tout de même de percevoir les agrégats de taille moyenne, invisibles à partir des seules données en stations ; alors que le chalut ne permet que de délimiter les grandes zones riches des zones plus faibles. L'acoustique apporte une information supplémentaire riche, puisque une structure plus petite que la distance entre stations existe.

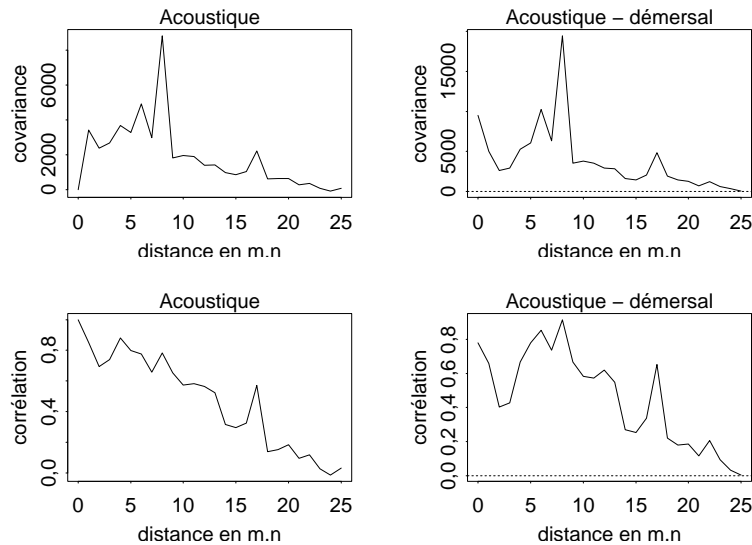


FIG. 2.7 – Covariances (haut) et corrélogrammes (bas) simples (colonne de gauche) et croisés (colonne de droite) de l'acoustique et des démersaux en ENASC, l'axe des abscisses représente la distance en m.n le long de la route du navire. Norvège 1997

2.3 Cokrigeage

Le cokrigeage est une estimation par combinaison linéaire des données disponibles de chacune des deux variables :

$$P(x_0) = \sum_{i \in S} \lambda_i^P P(x_i) + \sum_{j \in R} \lambda_j^A A(x_j) + \lambda_0$$

où les poids λ_i^P et λ_j^A sont estimés par moindres carrés, en imposant des contraintes de non biais. Plus de détails peuvent être trouvés dans Matheron (1970 [41]) ou Wackernagel (1998 [70]). Les hypothèses dépendent du modèle choisi : s'il s'agit d'un variogramme, les incréments de chaque variable doivent être stationnaires d'ordre 2 ; s'il s'agit d'une covariance, les variables elles mêmes doivent être stationnaires. Dans le premier cas, la variance de cokrigeage s'estime par :

$$\sigma_{CoK}^2 = \sum_{i \in S} \hat{\lambda}_i^P \gamma_P(x_i - x_0) + \sum_{j \in R} \hat{\lambda}_j^A \gamma_{A,P}(x_j - x_0) + \mu_{CoK} \quad (2.3)$$

où μ_{CoK} est le paramètre de Lagrange associé à la contrainte de non-biais sur P. La variance de cokrigeage est nécessairement inférieure ou égale à la variance de krigeage. Les poids du cokrigeage s'obtenant en minimisant la variance de l'erreur, si la variable auxiliaire ne permet pas de d'améliorer l'estimation, elle est affectée de poids nuls.

Le cokrigeage est un interpolateur exact qui prend en compte toutes les données disponibles (éventuellement au sein d'un voisinage glissant défini de façon ad hoc) pour l'estimation du point cible. Bien qu'optimal au sens des moindres carrés, le cokrigeage est souvent difficile à interpréter. La compréhension *physique* de l'estimation passe par l'interprétation de chaque structure du modèle, en fonction du type d'échantillonnage. Les poids de la combinaison linéaire changeant pour chacun des points cibles, une approche par cokrigeage ne permet pas d'écrire une relation formelle liant les deux variables. C'est la principale limite de ce modèle (ou sa force selon le point de vue adopté, le cokrigeage peut en effet sembler plus *objectif* puisqu'il ne se limite pas aux modèles interprétables). En effet, la modélisation recherchée généralement par les halieutes vise non seulement à fournir une estimation précise en chaque point, mais également à établir une relation permettant de mieux comprendre le phénomène naturel étudié ainsi que le rôle joué par chacune des deux variables dans l'estimation.

Le cokrigeage dépend de trois structures : la structure propre à chaque variable ainsi que leur structure croisée. Si les deux variables sont orthogonales, la structure croisée est pépitique et le cokrigeage équivaut au krigeage ordinaire de la variable cible. Lorsque la structure croisée est proportionnelle à la structure de l'acoustique, le cokrigeage se factorise (partie 2.4) : la variable acoustique prédomine dans le modèle et un résidu permet de corriger l'estimation. Lorsque les trois structures sont proportionnelles, on se retrouve en situation de corrélation intrinsèque (partie 2.5) : l'estimation revient alors au krigeage ordinaire des seules données de captures, auquel s'ajoute un facteur de correction locale apportée par l'acoustique.

Exemple 1

L'exemple suivant permet de comprendre sur un cas très simple le comportement des poids de cokrigeage en fonction des structures de chacune des variables. Un exemple similaire est présenté dans Goovaerts (1997 [29]).

On suppose que l'on dispose d'une seule station (S) pour laquelle est disponible une donnée de capture et une donnée acoustique ainsi que de deux points situés le long des transects (R_1, R_2) pour lesquels seule l'acoustique est disponible. On cherche à estimer successivement trois points $x_0 = (x_{0,1}, x_{0,2}, x_{0,3})$ par cokrigage simple, c'est à dire en supposant que la moyenne de chaque variable est connue. Aucune contrainte n'est alors imposée aux poids.

Les points sont ordonnés selon la géométrie suivante :

Les trois points cibles sont $x_{0,1}, x_{0,2}$ et $x_{0,3}$.

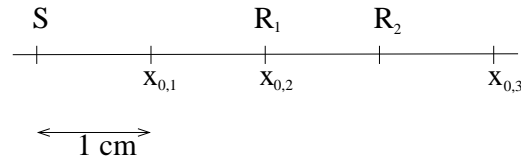


FIG. 2.8 – Représentation de la localisation des points de l'exemple 1, à une seule station S et deux points de transect R_1 et R_2

On considère également trois configurations différentes de structures :

- configuration 1, la proportion d'effet de pépité de P varie :

$$\begin{aligned} C_p(h) &= \phi \text{ pep}(h) + (1 - \phi) \text{ sph}(h/5) \\ C_{A,P}(h) &= 0.4 \text{ sph}(h/5) \\ C_A(h) &= 0.1 \text{ pep}(h) + 0.9 \text{ sph}(h/5) \\ &\text{avec } \phi \in [0; 0.82] \end{aligned}$$

- configuration 2, le coefficient de corrélation entre les deux variables varie :

$$\begin{aligned} C_p(h) &= 0.3 \text{ pep}(h) + 0.7 \text{ sph}(h/5) \\ C_{A,P}(h) &= \rho_{AP} \text{ sph}(h/5) \\ C_A(h) &= 0.1 \text{ pep}(h) + 0.9 \text{ sph}(h/5) \\ &\text{avec } \rho_{AP} \in [0; 0.79] \end{aligned}$$

- configuration 3, la proportion d'effet de pépité de A varie :

$$\begin{aligned} C_p(h) &= 0.3 \text{ pep}(h) + 0.7 \text{ sph}(h/5) \\ C_{A,P}(h) &= 0.4 \text{ sph}(h/5) \\ C_A(h) &= \psi \text{ pep}(h) + (1 - \psi) \text{ sph}(h/5) \\ &\text{avec } \psi \in [0; 0.77] \end{aligned}$$

Le coefficient ϕ mesure la proportion d'effet de pépité de la variable principale, et le paramètre ψ celui de la variable auxiliaire. Le paramètre ρ_{AP} représente la corrélation entre les deux variables puisque les variances de celles-ci sont égales à 1. Les intervalles imposés aux paramètres des modèles visent à assurer que chaque matrice de corégionalisation soit définie positive, une condition nécessaire d'admissibilité du modèle (Wackernagel 1998 [70]).

La figure 2.9 montre l'évolution des poids de krigeage de chacun des points dans chacune des configurations de modèle présentées ci-dessus ; lorsque l'on fait varier les paramètres ϕ , ψ , et ρ_{AP} dans l'intervalle qui leur est associé. Les deux dernières colonnes montrent l'évolution de la variance de cokrigeage ainsi que le ratio égal à la somme des valeurs absolues des poids accordés à l'acoustique sur celle de la donnée de capture :

$$\text{ratio} = \frac{\sum_{j \in R} |\lambda_j^A|}{\sum_{i \in S} |\lambda_i^P|} \quad (2.4)$$

Il permet de mesurer l'influence relative de la variable auxiliaire lorsque les deux variables ont la même variance (Goovaerts 1997 [29]).

Au fur et à mesure que la variable de capture se déstructure, c'est à dire quand la proportion de l'effet de pépité augmente, le poids accordé à l'information de pêche en stations diminue linéairement, alors que le poids de l'acoustique au même point S augmente tout en restant quasi-constant en R_1 et R_2 . Par contre, lorsque l'on augmente le coefficient de corrélation, le poids de la variable de capture diminue beaucoup plus lentement et de façon non-linéaire, alors que l'acoustique en R_1 et R_2 augmente linéairement. Lorsque l'acoustique se déstructure, le poids de la capture n'augmente pas, il a même tendance à diminuer légèrement, alors que la répartition des poids de l'acoustique tend à s'uniformiser entre les points S, R_1 et R_2 . Le ratio d'influence de l'acoustique est nul ou négligeable, sauf lorsque l'on estime le point le plus éloigné de la station. Il est intéressant de voir que le ratio n'évolue pas de manière linéaire. Lorsque la variable de capture devient très déstructurée, ou que le coefficient de corrélation entre les deux variables est grand, l'estimation des données éloignées des stations néglige rapidement l'information de capture. Enfin, l'effet de pépité de l'acoustique ne semble pas avoir d'influence sur la variance de cokrigeage, alors que celui de la capture modifie la variance de l'estimation des deux premiers points $x_{0,1}$ et $x_{0,2}$. Pour un niveau de corrélation assez bas, fixé à 0.4, le point $x_{0,2}$ placé en R_1 n'est mieux estimé que $x_{0,1}$ que si l'effet de pépité dépasse 50%. Le coefficient de corrélation permet quant à lui de faire diminuer la variance de cokrigeage de tous les points. C'est la variance d'estimation du point $x_{0,2}$ qui diminue le plus vite. Lorsque le coefficient de corrélation est supérieur à 0.5, il devient le point le mieux estimé.

Cet exemple permet également de visualiser l'effet écran (Wackernagel 1998 [70]) : le poids de R_1 pour l'estimation de $x_{0,3}$ est quasi nul en raison de la présence de R_2 . Le poids de l'acoustique en S quant à lui est souvent négatif.

Ainsi, quelque soit la configuration, l'acoustique a un poids important dès qu'il s'agit d'estimer un point loin d'une station. L'effet de pépité de P a peu d'impact sur les poids accordés à l'acoustique hors stations. Le coefficient de corrélation, par contre, les fait varier linéairement.

Exemple 2

L'exemple suivant vise à mesurer l'impact des modèles structuraux précédents (configurations 1, 2 et 3 de l'exemple précédent) sur les quatre transects présentés en introduction du chapitre (cf. figure 2.1). La portée de la composante sphérique a cependant été changée. Elle est fixée à 60 m.n, ce qui correspond approximativement à la distance entre trois stations. La figure 2.10 montre l'évolution de l'estimation de chacun des quatre transects pour différentes proportions d'effet de pépité pour le modèle de la variable cible, la variable de pêche. Les estimations sont assez semblables entre elles lorsque les données acoustiques en stations sont compatibles, comme pour les transects 1 et

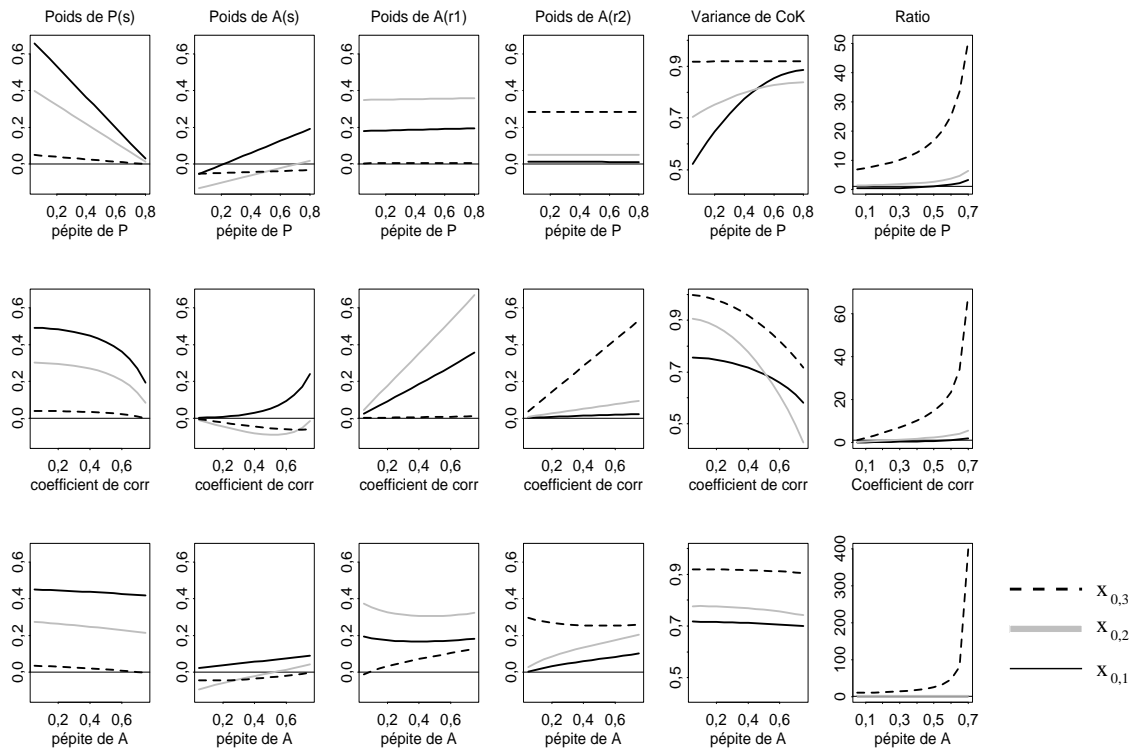


FIG. 2.9 – Évolution des poids de cokrigage de chaque point (les 4 premières colonnes), de la variance de cokrigage (5e colonne) ainsi que du ratio des poids accordés à l'acoustique sur le poids accordés à la donnée de capture (6e colonne, cf. équation 2.4) pour chacune des trois configurations de structure : augmentation de la proportion d'effet de pépite de la variable de capture (en haut), augmentation du coefficient de corrélation entre les deux variables (milieu) et augmentation de la proportion d'effet de pépite de la variable auxiliaire (bas). La ligne pointillée correspond à l'estimation du point $x_{0,3}$, la ligne grise au point cible $x_{0,2}$ et la ligne noire au point $x_{0,1}$. Une ligne horizontale correspondant aux poids nuls est également tracée sur les graphiques des 4 premières colonnes.

3. Par contre le modèle est assez sensible à l'effet de pépité de la variable de captures lorsque les valeurs en stations de celle-ci diffèrent des valeurs acoustiques, comme pour le transect 2 et, dans une moindre mesure, pour le transect 4. Plus la variable cible est déstructurée, plus l'estimation se rapproche des valeurs acoustiques.

Lorsque l'on augmente la proportion d'effet de pépité de la variable auxiliaire (figure 2.11) l'estimation se lisse localement pour ne suivre que les grandes tendances indiquées par les valeurs acoustiques. Une fois encore ce sont pour les zones d'incompatibilités entre les deux variables que les différences sont les plus importantes. Par contre il est intéressant de voir que lorsque le modèle choisi pour la variable acoustique est très pépitique, le relief général de l'estimation est plus accentué, bien que les variations locales de l'acoustique soient effacées. L'impact des grandes zones riches de l'acoustique sur l'estimation est plus fort, alors que celui des valeurs fortes isolées est moindre. Sur le transect 4, la zone forte observée par l'information acoustique est localement contredite par la valeur de pêche (avant dernière station). Cette zone riche est mieux représentée lorsque l'acoustique est très pépitique, par contre l'estimation devient négative sur le reste du transect.

Lorsque c'est le coefficient de corrélation qui varie (cf. figure 2.12), l'estimation tend à mieux suivre les détails apportés par l'acoustique tout en respectant les valeurs de captures. Les valeurs de l'estimation en stations ne changent quasiment pas. C'est principalement l'estimation entre les stations qui tend à mieux se coller aux données acoustiques. L'estimation est donc plus fluctuante dans l'espace.

Ainsi, lorsque les données de captures suivent les données acoustiques (comme sur le transect 3) qui sont régulières dans l'espace, les différentes configurations de modèles ont relativement peu d'impact sur l'estimation par cokrigeage le long des transects.

Lorsqu'il y a contradiction en stations entre la donnée de capture et la donnée acoustique, l'estimation est alors très sensible à l'effet de pépité du modèle associé à la variable de capture. Si la proportion d'effet de pépité des captures est importante, l'estimation aura tendance à suivre l'information acoustique en cas d'erreur de type 1 aussi bien qu'en cas d'erreur de type 2 (cf. partie 1.1.4).

Application

Un cokrigeage ordinaire a été effectué pour deux campagnes irlandaises (2000 et 2002). Le modèle choisi est un modèle de corégionalisation à deux structures comprenant une pépité et un sphérique. En 2000, l'effet de pépité du modèle choisi représente 26% de la variable de capture et 4% pour l'acoustique, la portée du sphérique étant de 22 m.n pour les deux variables et le coefficient de corrélation entre les deux variables est de 0.7. En 2002, la pépité représente 53% du modèle de capture et seulement 3% de l'acoustique. La portée du sphérique est de 35 m.n et le coefficient de corrélation entre les variables de 0.5. L'estimation permet un gain en variance, inhérente au cokrigeage par rapport au krigeage et également l'apport de certains détails et de contrastes sur les cartes d'interpolation (figure 2.13). Le modèle de cokrigeage permet de mieux délimiter, grâce à l'information disponible le long des transects, les zones riches et pauvres. En 2000, par exemple, l'influence d'une station à l'est de la zone d'étude est fortement réduite par le cokrigeage.

Cependant, les cartes restent très semblables. Une étude comparative détaillée des résultats des modèles visant à évaluer le gain en précision et en robustesse apporté par l'information acoustique sera présentée dans le dernier chapitre.

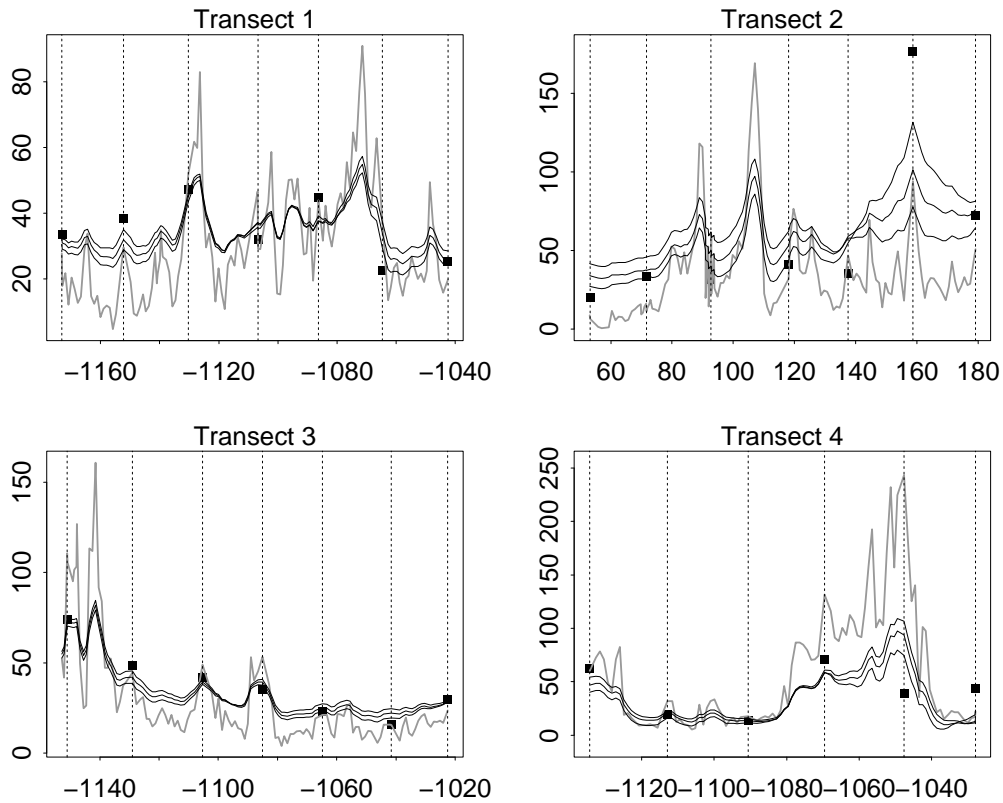


FIG. 2.10 – Impact de l'augmentation de l'effet de pépité de la variable cible sur le cokrigage pour quatre transects (proportion de pépité = 25%, 50% et 70%). L'effet de pépité de la variable auxiliaire est fixé à 10%, et le coefficient de corrélation entre les deux variables est de 0.5. Les lignes verticales représentent les localisations des stations. Les valeurs des captures, variable cible, sont indiquées par des carrés.

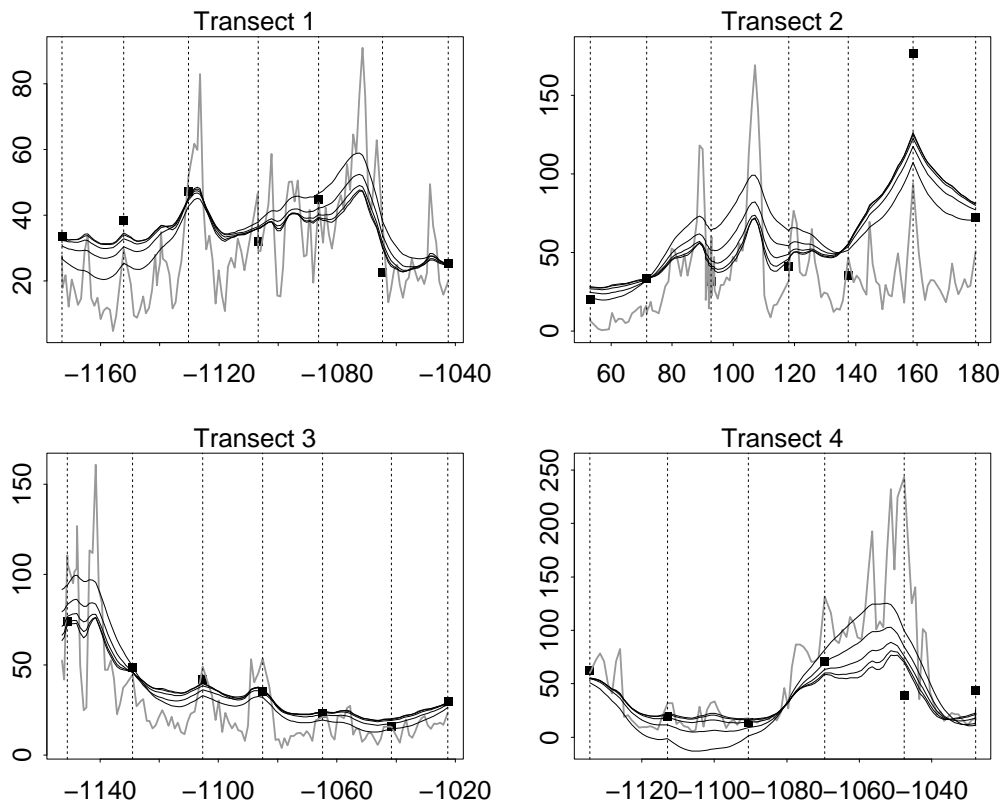


FIG. 2.11 – Impact de l'augmentation de l'effet de pépite de la variable auxiliaire sur le cokrigage pour quatre transects (pépite = 25%, 35%, 50%, 65% et 75%). L'effet de pépite de la variable cible est fixée à 30%, et le coefficient de corrélation entre les deux variables est de 0.5. Les données acoustiques sont tracées en gris. Les lignes verticales représentent les localisations des stations. Les valeurs des captures, variable cible, sont indiquées par des carrés.

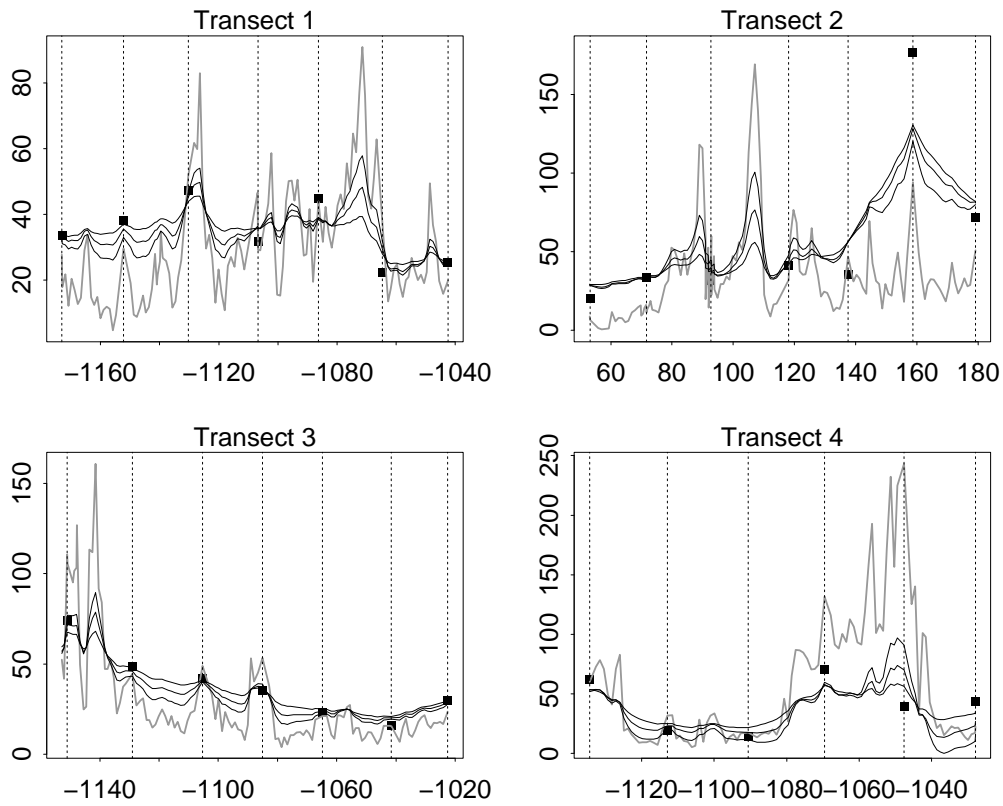


FIG. 2.12 – Impact de l’augmentation du coefficient de corrélation entre les deux variables sur le cokrigage des quatre transects ($\rho = 20\%$, 40% et 60%). La proportion d’effet de pépite de la variable cible est fixée à 30% alors que celui de la variable auxiliaire est de 10% .

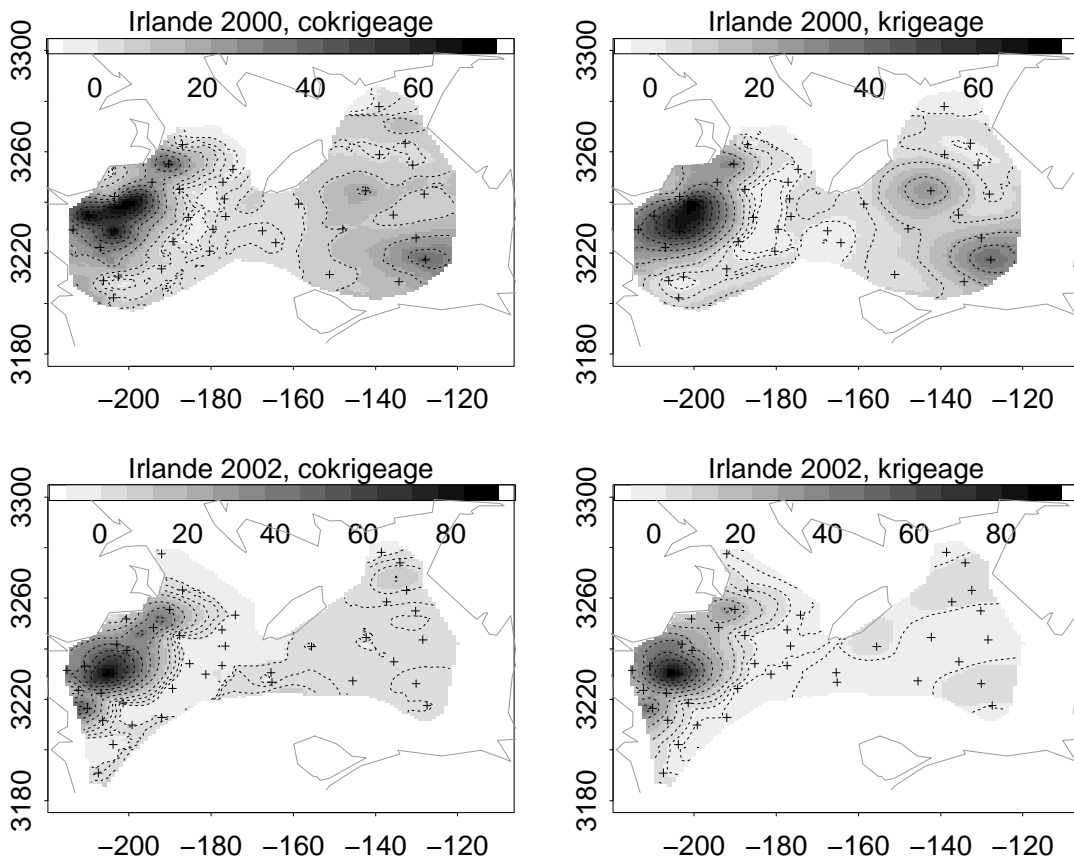


FIG. 2.13 – Cartes de cokrigage (gauche) des captures pélagiques (en ENASC) avec l’acoustique et cartes de krigeage ordinaire (droite) des captures pélagiques seules. Irlande 2000 (haut) et 2002 (bas). Les coordonnées de longitudes et de latitudes ont été converties en milles nautiques.

La mise en pratique du cokrigage nécessite qu’un modèle cohérent puisse être ajusté sur chacun des trois variogrammes expérimentaux. Lorsque les variables sont fortement asymétriques, les valeurs fortes brulent fortement les estimations linéaires, comme l’inférence du variogramme et du cokrigage. Une approche non-linéaire, consistant à découper la variable auxiliaire en indicatrice est discutée en dernière partie du chapitre. Une transformation permettant de normaliser les variables serait une autre approche possible, en effectuant par exemple un cokrigage lognormal. Cependant si une transformation de ce type est délicate à mettre en oeuvre sur une seule variable, elle demande des hypothèses sévères pour une estimation bivariable. Une comparaison des estimations obtenues sur l’acoustique par transformation en indicatrice et lognormale est présentée en dernière partie de ce chapitre.

Les deux parties suivantes traitent de simplifications liées aux structures spatiales croisées.

2.4 Cas où la variable auxiliaire est maîtresse : le modèle à résidu

La variable acoustique étant connue de façon très dense, un modèle simple consiste à supposer que l'estimation des captures est portée par la variable acoustique.

Lorsque la structure croisée entre les deux variables est proportionnelle à la structure de la variable auxiliaire :

$$C_{A,P}(h) = \alpha C_A(h) \quad (2.5)$$

le cokrigeage se simplifie. Le modèle est présenté ici pour des variables stationnaires qui admettent une covariance, mais il est directement transposable dans le cas intrinsèque pur (Rivoirard 2001 [57]). La variable auxiliaire étant autokrigeable, le modèle revient alors à considérer que la variable cible se décompose en une combinaison linéaire de la variable auxiliaire et d'un résidu (Rivoirard 2001 [57]) :

$$P(x) = \alpha A(x) + \beta + \varepsilon(x) \quad (2.6)$$

D'après la relation 2.5, le résidu est alors spatialement orthogonal à la variable auxiliaire :

$$\begin{aligned} Cov[A(x), \varepsilon(x+h)] &= Cov[A(x), P(x+h) - \alpha A(x+h)] \\ &= C_{A,P}(h) - \alpha C_A(h) \\ &= 0 \end{aligned}$$

Dans ce modèle, la variable auxiliaire joue un rôle prédominant dans l'estimation. C'est la variable maîtresse.

Dans le cas gaussien, $A(x+h)$ et $T(x)$ sont donc indépendants quand $A(x)$ est connue. C'est la raison pour laquelle, le modèle est également appelé modèle de "type Markov" (Chilès et Delfiner 1999, [15]). Quelque soit la loi de distribution des variables, le poids donné à $A(x+h)$ dans l'estimation de $T(x)$ est nul lorsque $A(x)$ est connu par effet écran, ce qui rappelle la propriété des processus de Markov.

Dans la littérature, le modèle de type Markov est une appellation générale désignant à la fois le modèle utilisé ici, pour lequel la structure croisée est proportionnelle à la structure de la variable auxiliaire, que le modèle pour lequel elle est proportionnelle à la structure de la variable cible (Shmaryan et Journel 1999 [63]). Ce dernier a des propriétés très différentes, en particulier en configuration de sous-échantillonnage de la variable d'intérêt. Un modèle dans lequel la variable cible serait maîtresse aurait ici moins d'intérêt, puisque c'est la variable auxiliaire qui est la mieux échantillonnée (il n'est d'ailleurs pas suggéré par les données).

On a donc la relation en stations entre les covariances :

$$C_P(h) = \alpha^2 C_{A_S}(h) + C_\varepsilon(h) \quad (2.7)$$

La notation $C_{A_S}(h)$ représente le modèle de covariance de l'acoustique en stations.

La variable auxiliaire étant spatialement orthogonale au résidu, le cokrigeage se décompose, pour le type d'échantillonnage étudié, en somme de deux krigeages :

$$P^{CoK} = \alpha A^K + \varepsilon^K$$

La variance de cokrigeage devient la somme des deux variances de krigeages :

$$\sigma_{CoK}^2(P) = \alpha^2 \sigma_K^2(A) + \sigma_K^2(\varepsilon) \quad (2.8)$$

Cette configuration particulière de structures spatiales permet de présenter le cokrigage sous une forme simple, facilement interprétable. La capture est linéairement liée à l'acoustique. Le coefficient β représente la différence moyenne entre ce qui est enregistré par le sondeur et ce qui est collecté dans le chalut. Le coefficient α représente le rapport moyen entre l'acoustique et la capture, une fois la constante β soustraite. Une valeur plus petite que 1 pour α signifie qu'une proportion de ce qui est enregistré par l'acoustique échappe au chalut, en raison de la différence de hauteur verticale par exemple. Mais on peut également imaginer une valeur de α plus grande que 1, en particulier en mer du Nord. Le chalut étant très grand par rapport au cône acoustique, les valeurs de pêches pourrait être en moyenne α fois plus grandes que les valeurs observées à l'acoustique.

Ce modèle bien que très simple semble donc tout à fait réaliste pour décrire la relation liant les deux variables entre elles.

2.4.1 Correction de la variance

La forte asymétrie des distributions étudiées ici s'accompagne d'une différence expérimentale entre la variance des données acoustiques observées aux seuls points de stations et la variance de l'intégralité des données acoustiques.

On considère le plan d'échantillonnage des campagnes norvégiennes en mer de Barents et l'on se place dans le domaine R à une dimension, formé par la route du navire. Les données en radiales découpent uniformément ce domaine par petits blocs unités de 1 m.n, représentés par leur centre. On peut considérer que les stations suivent un échantillonnage régulier au sein de ce domaine espacées chacune de 20 m.n avec une origine aléatoire. Le support d'une station est également le bloc unité de 1 m.n. La variance des données acoustiques en radiales est en moyenne égale à la variance de dispersion d'une station dans le domaine des radiales $D^2(. | R)$. La variance des données en stations est quant à elle égale en moyenne à la variance de dispersion d'un point dans l'ensemble formé par les stations $D^2(. | S)$. Or ces deux variances sont reliées par la relation de Krige (Matheron 1970 [41]) :

$$D^2(. | R) = D^2(. | S) + D^2(S | R) \quad (2.9)$$

Ces variances de dispersion s'entendent en moyenne et supposent que l'on dispose d'un grand nombre de réalisations de la variable acoustique suivant chacune le même plan d'échantillonnage.

Le troisième terme de la relation $D^2(S | R)$ représente la variance de dispersion des stations dans les radiales. Ce terme implique que la seule variance des données en stations $D^2(. | S)$ est un estimateur biaisé de la variance des données en radiales $D^2(. | R)$. Ce terme est d'autant plus important que la structure petite portée (inférieure à 20 m.n) des données est forte. Si les données sont pépitiées, ce terme dépend uniquement de la différence de taille entre l'ensemble des données de radiales et l'ensemble des données en stations. Il est égal, dans ce cas, à $\frac{1}{n_R} - \frac{1}{n_S}$ avec, rappelons-le, n_R le nombre de points en radiales et n_S le nombre de stations.

Les corrections de variance présentées dans cette partie supposent ce troisième terme $D^2(S | R)$ négligeable par rapport aux deux autres. La variance des données en stations est donc supposée être un estimateur sans biais de la variance des données en radiales. La différence expérimentale observée entre les deux variances est supposée être uniquement due aux fluctuations statistiques liées aux tailles de chaque échantillon et à la distribution des données.

De plus, les plans d'échantillonnage des campagnes IBTS rendent une telle relation délicate à établir. Les données n'étant pas enregistrées de nuit, la disposition des stations dans l'ensemble des radiales est particulière.

En situation de sous-échantillonnage de la variable d'intérêt, le but du modèle de type Markov présenté précédemment est d'utiliser la variable acoustique échantillonnée de façon dense, corrigée par une estimation du résidu orthogonal connue en stations. On s'aperçoit alors d'une hypothèse sous-jacente au modèle qui est que la variance de l'acoustique doit être la même en stations et sur l'ensemble des points de données. En effet la relation 2.7 relie la covariance de l'acoustique *en stations* à la covariance de la capture et du résidu. L'équation de la variance de cokrigeage 2.8 qui s'en déduit doit être corrigée pour utiliser, de façon cohérente, l'intégralité des données acoustiques.

Supposons que les données soient telles que la covariance expérimentale de l'acoustique en radiales est égale à la covariance expérimentale de l'acoustique en stations à un facteur multiplicatif près :

$$C_{A_S}^*(h) = k^2 C_{A_R}^*(h) \quad (2.10)$$

avec

$$k^2 = \frac{\text{Var}^*[A(x), x \in \{\text{Radiales}\}]}{\text{Var}^*[A(x), x \in \{\text{Stations}\}]}$$

L'hypothèse de proportionnalité des structures du modèle (équation 2.5) s'écrit :

$$C_{A,P}^*(h) = \alpha C_{A_S}^*(h) = \frac{\alpha}{k^2} C_{A_R}^*(h) \quad (2.11)$$

Il convient alors de décider si le palier du modèle de covariance de l'acoustique doit être ajusté sur la variance en radiales ou sur la variance en stations. Le palier du modèle ne modifie pas les poids de krigeage. Ceux-ci sont invariants à toute multiplication du modèle par une constante. Les estimations de krigeages calculées à partir de la covariance en stations $C_{A_S}(h)$ ou de la covariance en radiales $C_{A_R}(h)$ seront donc strictement égales. Par contre la variance de krigeage est directement proportionnelle au palier, celle établie à partir de la covariance en radiales sera donc k^2 fois supérieure à celle obtenue par la première.

Si le modèle utilisé pour le krigeage de l'acoustique est ajusté sur les données en stations, cette correction est inutile, le modèle est déjà cohérent. Toutefois, le plus naturel semble d'ajuster le modèle de covariance sur les données en radiales, puisque l'acoustique en radiales est connue de façon très dense, le modèle sera mieux ajusté, en particulier pour les petites distances. La relation 2.8 doit alors être corrigée :

$$\sigma_{C_{oK}}^2(P) = \sigma_K^2(\varepsilon) + \frac{\alpha^2}{k^2} \sigma_K^2(A_R) \quad (2.12)$$

On retrouve cette relation si l'on revient à la définition de la variance de cokrigeage 2.3, celle-ci s'écrit :

$$\begin{aligned}
\sigma_{CoK}^2 &= \sum_{i \in S} \hat{\lambda}_{CoK}^P(i) C_P(x_i - x_0) + \sum_{j \in R} \hat{\lambda}_{CoK}^A(j) C_{A,P}(x_j - x_0) + \mu_{CoK} \\
&= \sum_{i \in S} \hat{\lambda}_{CoK}^P(i) \left[\frac{\alpha^2}{k^2} C_{A_R}(x_i - x_0) + C_\varepsilon(x_i - x_0) \right] \\
&\quad + \sum_{j \in R} \hat{\lambda}_{CoK}^A(j) \left(\frac{\alpha}{k^2} C_{A_R}(x_j - x_0) \right) + \mu_{CoK} \text{ d'après 2.7 et 2.11} \\
&= \sum_{i \in S} \hat{\lambda}_{CoK}^P(i) C_\varepsilon(x_i - x_0) \\
&\quad + \frac{\alpha^2}{k^2} \left[\sum_{i \in S} \hat{\lambda}_{CoK}^P(i) C_{A_R}(x_i - x_0) + \sum_{j \in R} \frac{\hat{\lambda}_{CoK}^A(j)}{\alpha} C_{A_R}(x_j - x_0) \right] + \mu_{CoK} \\
&= \sigma_K^2(\varepsilon) + \frac{\alpha^2}{k^2} \sigma_K^2(A_R)
\end{aligned}$$

avec la correspondance des poids de krigeage du résidu et de l'acoustique et les poids de cokrigeage de l'acoustique avec la variable de pêche suivante :

$$\left\{ \begin{array}{l} \hat{\lambda}_K^\varepsilon(i) = \hat{\lambda}_{CoK}^P(i) \\ \hat{\lambda}_K^A(j) = \hat{\lambda}_{CoK}^P(j) + \frac{\hat{\lambda}_{CoK}^A(j)}{\alpha}, \text{ si } j \in S \\ \quad = \frac{\hat{\lambda}_{CoK}^A(j)}{\alpha} \text{ sinon} \\ \mu_{CoK} = \mu_K^\varepsilon + \frac{\mu_K^A}{\alpha^2} \end{array} \right.$$

Il est intéressant de constater que l'expression de la variance de cokrigeage telle qu'elle est exprimée dans l'équation 2.3 ne nécessite pas de renormalisation. En effet, la variance de cokrigeage s'exprime en fonction de la structure de la variable principale, la variable de pêche, ainsi que de la structure croisée. Or si le cokrigeage hétérotopique est formulé en covariance, le palier de la structure croisée est égal à la covariance croisée *en zéro*, c'est à dire en stations. Si il est formulé en variogramme, la question ne se pose pas, puisque celui-ci n'est défini que sur les points en isotopie, c'est à dire les stations. La variance de cokrigeage utilise donc bien un palier de modèle de capture ainsi qu'un palier de modèle croisé cohérents. La variance obtenue n'a pas besoin d'être renormée. La nécessité de corriger la variance de la variable acoustique dans le modèle à résidu vient de la décomposition du cokrigeage en somme de deux krigeages indépendants. La structure croisée ne joue pas directement de rôle dans le calcul lorsque la variable acoustique est auto-krigeable, mais permet d'assurer la cohérence des paliers des modèles.

La différence de variance entre les données en radiales et en stations est en pratique très fréquente. Les valeurs empiriques du coefficient k^2 sont indiquées dans la table 2.1.

Il est intéressant d'étudier le comportement du rapport de variances k^2 . Sa valeur dépend de la taille du sous-échantillon comparée à celle de l'échantillon principal, ainsi que de l'asymétrie de la distribution de la variable acoustique.

Institut	Année	Nb de données en stations	Nb de données en radiales	ratio : Nb en radiales/Nb en stations	k^2
IMR	1997	176	5209	30	1.33
IMR	1998	198	5135	25	1.83
IMR	1999	223	5567	24	2.23
IMR	2000	302	7680	25	1.35
IMR	2001	300	7666	25	3.55
IMR	2002	287	7383	25	2.65
FRS	1999	44	468	10	4.6
FRS	2000	46	351	7	3.2
FRS	2002	47	430	9	1.51
FRS	2003	46	303	6	0.53
CEFAS	2000	71	1038	14	0.08
CEFAS	2001	70	883	12	0.5
CEFAS	2002	70	1140	16	0.06
IFREMER	2002	77	440	5	1.91
IFREMER	2003	82	722	9	1.73
QUB	2000	37	110	3	0.3
QUB	2001	34	236	7	0.5
QUB	2002	41	173	4	0.23

TAB. 2.1 – Rapport entre la variance empirique de l'acoustique observée en radiales et en stations (coefficient k^2) pour chacune des campagnes

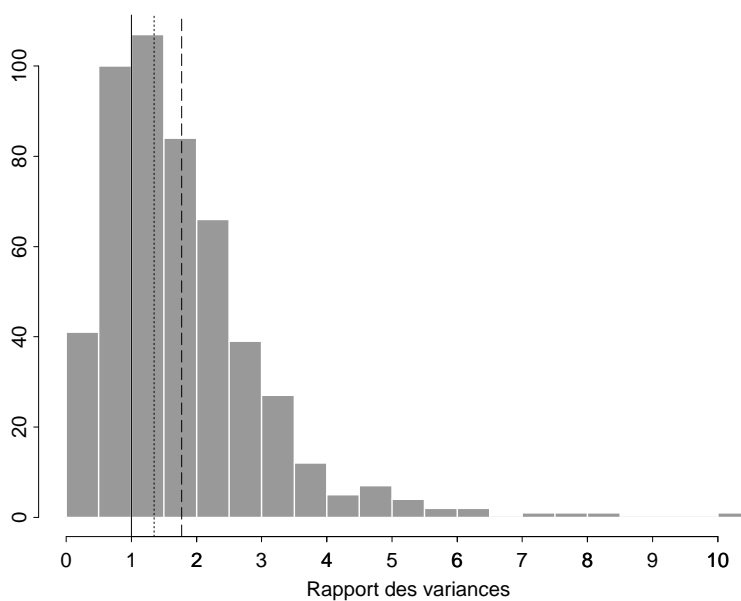


FIG. 2.14 – Histogramme du rapport entre la variance empirique obtenue sur l'échantillon principal (7000 points) et sur le sous-échantillon (300 points) lors de 500 simulations indépendantes d'une loi lognormale. La valeur empirique observée dans les données ($=1.35$) ainsi que la moyenne de l'histogramme ($=1.8$) sont indiquées par des lignes pointillées.

Pour illustrer ce phénomène, 500 échantillons de 7000 points ont été simulés suivant une loi lognormale telle que la variance et la moyenne correspondent aux valeurs empiriques observées pour une campagne choisie : Norvège 2000 ($\sigma^2=122\ 826$ et $m=54.05$). Parmi les 7000 valeurs obtenues, 300 ont été tirées aux hasard, pour représenter les stations au sein des radiales. Cette simulation suppose donc que les structures de chacune des variables sont pépitiqes. Dans ce cas, le terme de droite de la relation de Krige (équation 2.9) est négligeable. Il est ici de l'ordre de 10^{-3} en raison des tailles respectives des échantillons.

Dans plus de 80% des cas, le rapport k^2 entre les variances empiriques est supérieur à 1 (figure 2.14). En effet, le nombre de données en radiales étant 25 fois plus grand que le nombre de données en stations, les rares valeurs fortes simulées ont tendance à ne pas être tirées dans le sous ensemble des stations. Par contre, si certaines valeurs extrêmes s'y trouvent, la variance en stations devient très grande, amplifiée par le faible nombre de points et dépasse alors la variance en radiales.

Lorsque le ratio entre le nombre de données en radiales et le nombre de stations est petit, le phénomène a tendance à s'inverser, il y a alors quasiment autant de chance que les valeurs fortes se trouvent dans le sous-ensemble des stations. La variance des stations devient alors souvent supérieure à celle des radiales.

Par ailleurs, pour s'assurer que ce phénomène provient de la différence de taille respective des échantillons, la taille du sous-échantillon a été modifiée valant successivement 7000, 5000, 2000, 1000, 500, 300 et 100. Cette opération a été répétée environ 500 fois pour étudier de façon empirique la différence de variance entre le sous-échantillon et l'échantillon total de 7000 points. Les résultats sont montrés dans le graphique 2.15. Le rapport entre la variance de l'échantillon total et du sous-échantillon montre que les valeurs de k^2 observées sur les données réelles ne sont pas exceptionnelles. Le rapport a tendance à augmenter lorsque la taille du sous-échantillon diminue, mais son écart-type augmente également. Lorsque le sous-échantillon est très petit, sa variance a tendance à être assez instable. Elle est généralement faible, mais aussi plus sensible à la présence d'une valeur forte.

Ainsi, la différence de taille entre l'échantillon principal et le sous échantillon entraîne dans les cas de distributions asymétriques une différence de variance. Cette différence n'est donc pas particulière aux jeux de données étudiés ici.

2.4.2 Vérification des hypothèses et estimation du paramètre

Une variable est dite autokrigeable si le rapport du variogramme croisé avec l'autre variable sur son variogramme simple (Wackernagel 1998 [70]) :

$$\alpha(h) = \frac{C_{A,P}^*(h)}{C_A^*(h)}$$

ne dépend pas de h .

Cette hypothèse peut être testée à différentes échelles : soit sur l'ensemble des points disponibles, soit sur le sous ensemble de points en isotopie. Comme on l'a vu dans la partie précédente, il est possible que l'hypothèse soit vérifiée pour chacun des ensembles mais que les valeurs du rapport ne soit pas égales dans les deux cas. Avec les notations précédentes, si l'on utilise uniquement les données en stations, c'est le paramètre α qui est estimé alors que si c'est toute l'acoustique en radiales qui intervient dans les covariances simples et croisées, c'est directement le rapport $\frac{\alpha}{k}$ qui

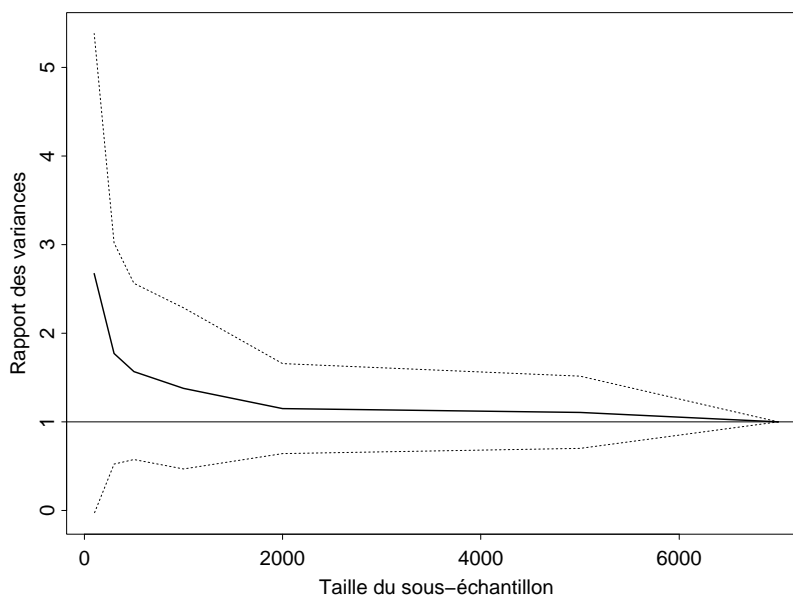


FIG. 2.15 – Évolution du rapport k^2 simulé pour une loi lognormale, lorsque la taille du sous-échantillon varie de 5000 à 100 points. La courbe noire représente la moyenne du coefficient k^2 et les lignes pointillées la moyenne plus ou moins un écart-type.

est évalué. Dans les deux cas, l'estimation de k^2 est nécessaire, soit pour corriger la variance de l'acoustique en radiales, soit pour estimer la variance du résidu en stations.

Le paramètre d'autokrigeabilité α permet d'obtenir le résidu en stations :

$$\varepsilon(x) = P(x) - \hat{\alpha}A_S(x) \text{ pour } x \in S$$

où $\hat{\alpha}$ est l'estimation de α . Cette relation permet au travers du résidu de passer d'une valeur acoustique à la valeur de pêche mesurée au même point, c'est à dire en stations. Il est donc naturel de la calculer sur ce même sous-ensemble de points. Cependant, utiliser les covariances entre les captures et l'acoustique en radiales permet d'utiliser plus de points et donc peut-être d'obtenir une estimation plus fiable de α , mais peut-être aussi d'introduire des variabilités locales. Carr (2003 [13]) montre de façon empirique pour un cas réel que l'hétérotopie induit une plus grande variabilité autour des paramètres d'autokrigeabilité.

Ici les deux approches ont été menées simultanément. La proportionnalité entre la structure croisée et la structure simple de l'acoustique a été vérifiée à la fois sur le sous ensemble de points en isotopie et sur l'ensemble du nuage de points. Sur 18 campagnes disponibles, 7 ont été considérées comme répondant à l'hypothèse du modèle à résidu, sur comparaison simultanée des variogrammes simples et croisés en stations et des corrélogrammes calculés sur tous les points le long de la route du navire. La figure 2.5 (p.35) montre les variogrammes simples de la variable acoustique et des croisés acoustiques captures pour les six campagnes norvégiennes. Pour les années 1997, 1998, 2000 et 2001 les structures simples et croisées semblent proportionnelles, comme le confirme les rapports de variogrammes relativement constants pour les différentes distances (figure 2.17). Les quatre cam-

pagnes sélectionnées sont celles indiquées en gras : 1997, 1998, 2000 et 2001. On peut voir sur la figure 2.16 que les corrélogrammes de ces mêmes années sont bien de formes similaires.

Quatre campagnes réalisées en mer de Barents et trois réalisées en mer du Nord ont ainsi été pré-sélectionnées sur comparaison *visuelle* des structures simples et croisées.

De façon plus quantitative, l'hypothèse d'autokrigeabilité est vérifiée si le rapport du variogramme croisé sur le variogramme simple est constant pour tous les pas de distances. Ce rapport fournit alors une estimation du paramètre α :

$$\hat{\alpha} = \frac{1}{N_h} \sum_h \frac{\gamma_{P,A}^*(h)}{\gamma_A^*(h)} \text{ pour } h \in \text{distances inter-station} \quad (2.13)$$

Les valeurs obtenues sont indiquées sur la figure 2.17 par les lignes horizontales. En raison des variabilités locales inhérentes aux données expérimentales, il est cependant difficile de juger si le rapport entre les variogrammes peut effectivement être considéré comme constant pour tous les pas de distance ou non. L'échelle de l'axe des ordonnées du graphe 2.17 doit être considérée avec attention. Il est délicat de s'imposer un critère, comme la somme des écarts à la moyenne, à partir duquel on rejette l'hypothèse. Le test a ici été effectué *à l'oeil* à partir de la ressemblance entre les structures simples et croisées de l'acoustique.

Deux questions sont extrêmement liées l'une à l'autre : la vérification de l'hypothèse d'autokrigeabilité de l'acoustique, qui consiste à vérifier que la structure croisée est bien proportionnelle à la structure simple de l'acoustique ; ainsi que la question de l'estimation du paramètre α , et de la qualité de cette estimation. En effet, l'estimation du paramètre α peut être mauvaise aussi bien parce que l'hypothèse de départ n'est pas satisfaite que pour des raisons de variabilité statistique.

Le coefficient α d'autokrigeabilité correspond également à la pente de la régression linéaire de $A(x)$ sur $P(x)$.

$$\hat{\alpha} = \frac{Cov^*(A(x), P(x))}{Var^*(A(x))}$$

A première vue, cette formulation est identique à la précédente 2.13, ne faisant intervenir que la covariance en zéro. L'avantage de la régression est de permettre l'estimation d'un indicateur de qualité de l'estimation de α . On s'aperçoit ainsi que l'estimation du coefficient a tendance à être tirée par une ou deux valeurs extrêmes (figure 2.18), à cause de la sensibilité très forte de la régression linéaire aux distributions asymétriques. Les valeurs obtenues pour le paramètre sont très similaires aux valeurs obtenues précédemment. Notons qu'un estimateur de qualité du même type que celui de la régression peut également être obtenu à partir des nuages variographiques présents autour des rapports de la figure 2.17.

Les valeurs prises par le paramètre α sont souvent plus petites que un et le paramètre bêta est quasi nul, indiquant qu'en moyenne le chalut ne collecte qu'une partie de l'abondance mesurée par l'acoustique (sur les 40 premiers mètres à partir du fond). Cependant, les différences d'années en années et l'influence de quelques valeurs fortes sur son estimation rend difficile une interprétation de sa valeur en tant que telle. Obtenir une estimation de α similaire pour l'ensemble des campagnes d'une même zone serait en effet particulièrement intéressant. Au delà des fluctuations statistiques, le paramètre α prendrait alors un sens *physique*. Si les assemblages halieutiques sont également stables d'années en années, le coefficient α peut représenter le rapport entre l'efficacité du chalut

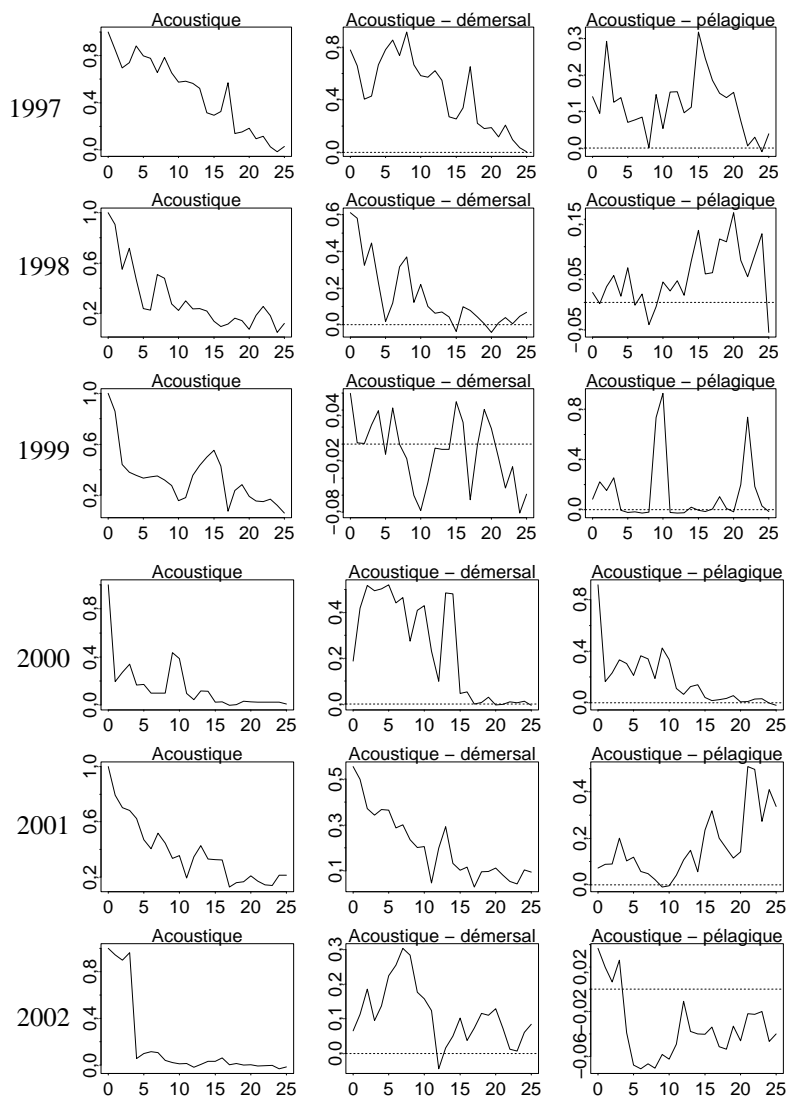


FIG. 2.16 – Corrélogrammes de l’acoustique (colonne de gauche) et corrélogrammes croisés avec les démersaux en ENASC (colonne du milieu) et avec les pélagiques en ENASC (colonne de droite), pour chacune des six campagnes norvégiennes. Les corrélogrammes ont été calculés le long de la route du bateau, de façon omnidirectionnelle. L’axe des abscisses représente l’éloignement de la station en m.n.

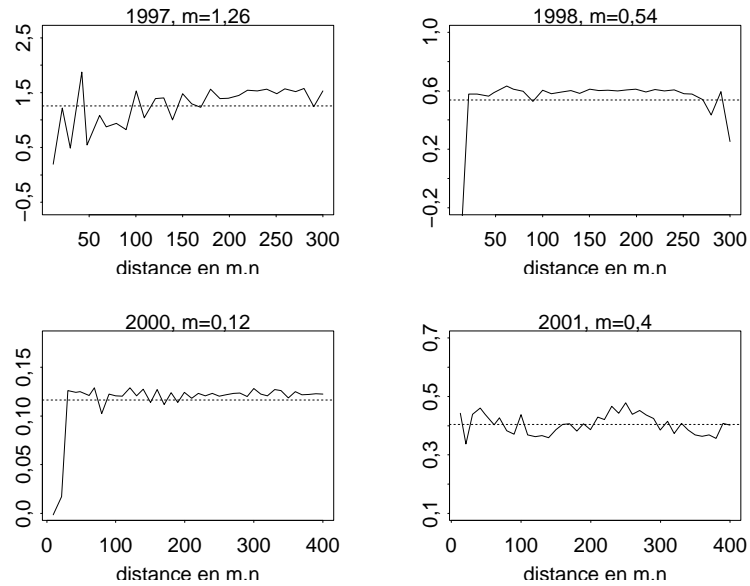


FIG. 2.17 – Rapports du variogramme croisé acoustique-capture sur le variogramme simple de l'acoustique, estimés en stations pour les 4 années sélectionnées en Norvège 1997, 1998, 2000 et 2001.

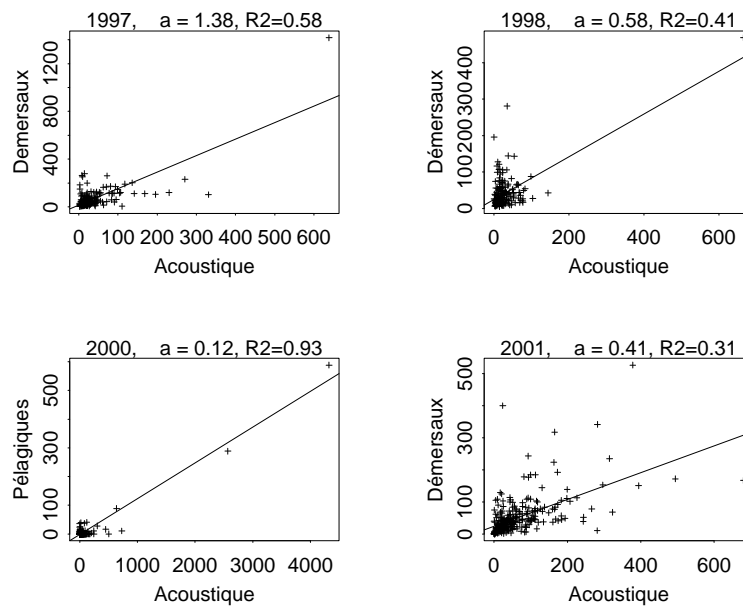


FIG. 2.18 – Nuages de points acoustique-capture en stations, auxquels sont superposés les droites de régression obtenues pour les 4 années sélectionnées en Norvège 1997, 1998, 2000 et 2001. Les valeurs de α correspondantes et les valeurs du R^2 de la régression sont indiquées sur les graphes.

de fond par rapport aux enregistrements acoustiques.

D'autre part, l'ordonnée à l'origine, représentant l'estimation du paramètre β est souvent quasiment nulle. Si l'on fait l'hypothèse supplémentaire que le coefficient β est nul, le chalut est alors strictement proportionnel à l'acoustique. La proportionnalité entre les deux variables entraîne que le paramètre α peut être estimé sur un sous-ensemble de points, en limitant ainsi l'influence des valeurs extrêmes. L'estimation de α peut être par exemple une moyenne d'estimation calculées sur différentes classes de valeurs. Toutefois, les valeurs obtenues par cette approche changent beaucoup d'une classe à l'autre et restent encore assez fluctuantes d'années en années.

La qualité de l'estimation du paramètre est très importante dans le modèle. Si le coefficient est surestimé, l'acoustique jouera un rôle trop prépondérant et la variance d'estimation risque d'être largement surestimée (la variance fait intervenir α au carré). Réciproquement si α est sous-estimé, le rôle de l'acoustique sera minimisé à tort par rapport au rôle du résidu et le modèle sera équivalent au krigeage ordinaire des données de captures.

Si l'hypothèse d'autokrigeabilité de l'acoustique est abusive, le modèle à résidu ne sera qu'une approximation du cokrigeage (approximation d'autant plus grossière que les hypothèses ne seront pas vérifiées). La variance d'estimation de α n'intervient pas dans l'estimation finale, il est tout de même important que la qualité de la régression soit prise en compte lors de l'interprétation des résultats.

2.4.3 Cokrigeage collocalisé

Lorsque la variable d'intérêt est largement sous échantillonnée par rapport à la variable auxiliaire, il est courant d'utiliser une approximation du cokrigeage simple, appelée le cokrigeage collocalisé, consistant à se fixer un voisinage restreint au seul point cible pour la variable auxiliaire (Xu et al., 1992 [73]). Goovaerts (1997 [29]) souligne de plus que le membre de gauche du système de cokrigeage peut devenir numériquement instable lorsque les corrélations à petites distances de la variable auxiliaire sont très fortes. Le krigeage collocalisé permet alors un gain en fiabilité important. En cokrigeage ordinaire, la somme des poids de la variable auxiliaire devant être nul, on parle rapidement de cokrigeage collocalisé multiple (Chilès et Delfiner 1999, [15]) dans lequel le voisinage est réduit au point cible ainsi qu'à l'ensemble des points en isotopie (les points pour lesquels la variable cible est connue).

Rivoirard (2001 [57]) montre que le modèle à résidu est le seul modèle pour lequel le cokrigeage est exactement égal au cokrigeage collocalisé (multiple), sans approximation.

En effet, si l'on se place en un point pour lequel l'acoustique est connue (hors stations) :

$$\begin{aligned} P^{cok}(x_0) &= \alpha A_{\bar{S}}^K(x_0) + \varepsilon^K(x_0) \\ &= \alpha A_{\bar{S}}(x_0) + \varepsilon^K(x_0) \\ &= \alpha A_{\bar{S}}(x_0) + \sum_{i \in S} \lambda_{\varepsilon}^K(x_i) \varepsilon(x_i) \end{aligned}$$

comme le krigeage est un interpolateur exact, l'estimation ne dépend que des données en stations (au travers du résidu) et de l'acoustique au point cible. Les autres données acoustiques en radiales sont affectées de poids nuls.

Si l'hypothèse d'autokrigeabilité de l'acoustique n'est pas pleinement vérifiée, le modèle à résidu devient une approximation du cokrigeage permettant de ne pas modéliser de structure croisée.

Cependant il convient de noter que le cokrigage collocalisé n'est utilisable que lorsque la variable auxiliaire est connue de façon exhaustive, alors que le modèle à résidu permet une estimation en tout point du champ.

2.4.4 Impact du choix du modèle

Dans le modèle à résidu présenté ci-dessus, l'estimation des captures est portée par l'acoustique et corrigée par un résidu permettant de réajuster les valeurs en stations pour obtenir un interpolateur exact de la variable de capture. Plus l'acoustique sera pépitique plus l'estimation sera proche d'un krigeage ordinaire des données de captures. Chilès et Delfiner (1999 [15]) soulignent en effet que $A(x)$ doit être préférablement *plus continue* que $P(x)$. Cette comparaison est délicate puisque les deux variables ne sont pas échantillonnées avec la même densité. La variabilité des captures en dessous de 15 m.n est inconnue. A l'échelle des stations, les variogrammes indiquent une continuité spatiale comparable pour les deux variables. Dans les cas présentés ici, le résidu présente fréquemment une structure longue portée, en particulier en Norvège, où ils semblent être dus aux valeurs fortes observées sur la bordure de la surface échantillonnée. Le résidu joue alors un rôle de recalage général de l'acoustique sur les tendances grandes distances des captures. Un résidu avec une structure très petite portée et une variable auxiliaire dotée d'un modèle de variogramme longue portée fournirait l'effet inverse : alors que la forme générale de l'estimation serait portée par l'acoustique, le résidu apporterait des ajustements locaux correspondant aux erreurs de mesures ou aux décalages ponctuels entre les deux variables en stations. Cependant, une fois de plus, une structure petite portée est délicate à dégager alors que les stations sont éloignées de plus de 15 m.n.

Étudions à nouveau l'exemple ne comprenant qu'une seule station et deux points situés le long des transects présenté précédemment (cf. figure 2.3) et les deux configurations de structures suivantes :

– configuration 1 :

$$\begin{aligned} C_p(h) &= \phi \text{ pep}(h) + (1 - \phi) \text{ sph}(h/5) \\ C_{A,P}(h) &= 0.4 \text{ sph}(h/5) \\ C_A(h) &= \text{ sph}(h/5) \\ &\text{avec } \phi \in [0; 0.82] \end{aligned}$$

– configuration 2 :

$$\begin{aligned} C_p(h) &= 0.3 \text{ pep}(h) + 0.7 \text{ sph}(h/5) \\ C_{A,P}(h) &= \rho_{AP} \text{ sph}(h/5) \\ C_A(h) &= \text{ sph}(h/5) \\ &\text{avec } \rho_{AP} \in [0; 0.79] \end{aligned}$$

Par rapport au cokrigage (figure 2.9) les effets d'écran sont légèrement plus marqués (cf figure 2.19). Quelque soit la proportion d'effet de pépite de P et la valeur du coefficient de corrélation entre les deux variables, le poids de $A(R_1)$ dans l'estimation de $x_{0,3}$ est négatif et le poids de $A(R_2)$ dans l'estimation de $x_{0,2}$ est nul (cokrigage collocalisé multiple). Le poids de $A(S)$ pour l'estimation de $x_{0,2}$ devient fortement négatif lorsque P est fortement structurée. Il atteint un minimum pour une corrélation de 0.6. Le poids total de l'acoustique devient très fort lorsque P est très pépitique mais passe vite en dessous de 1 pour l'estimation de $x_{0,1}$ et $x_{0,2}$ quand le coefficient de corrélation est inférieur à 0.4.

Chilès et Delfiner (1999 [15], p307) montrent, pour un exemple de modèle identique à la première

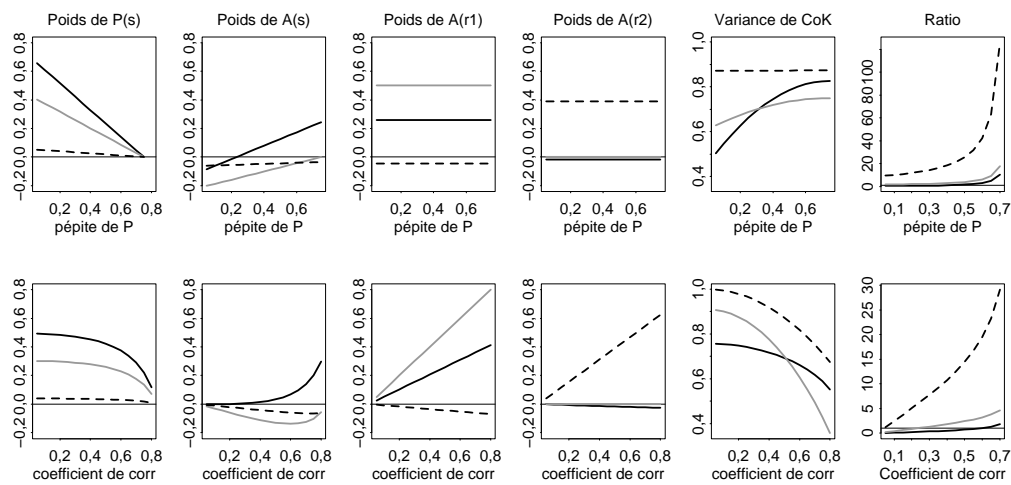


FIG. 2.19 – Évolution des poids de cokrigage (modèle de type Markov) de chaque point, de la variance de cokrigage ainsi que du ratio des poids accordés à l'acoustique sur le poids accordé à la donnée de capture (cf. équation 2.4) pour chacune des deux configurations de structure : augmentation de la proportion d'effet de pépite de la variable de capture (en haut) et augmentation du coefficient de corrélation entre les deux variables (en bas). La ligne pointillée correspond à l'estimation du point $x_{0,3}$, la ligne grise au point cible $x_{0,2}$ et la ligne noire au point $x_{0,1}$.

configuration mais pour un voisinage de points en isotopie, que le rapport entre la variance de cokrigeage et la variance d'un krigeage ordinaire de la variable cible est une fonction de l'effet de pépite qui décroît puis réaugmente, atteignant un minimum pour une proportion d'effet de pépite de 0.56%.

2.4.5 Vérification a posteriori de l'hypothèse d'autokrigeabilité ou comparaison avec une estimation classique de cokrigeage

Les résultats obtenus par le modèle à résidu ont été comparés à ceux obtenus par cokrigeage en utilisant les mêmes modèles. La seule différence réside dans l'inférence du variogramme croisé. Les modèles de variogrammes croisés utilisés en norvège sont indiqués dans la figure 2.20. Les modèles utilisés pour l'acoustique et les captures, sont eux, les mêmes que pour le modèle à résidu (ces modèles sont présentés pour la Norvège dans le chapitre 4, page 136, figure 4.1). Comme on l'a vu, les deux modèles sont équivalents dès que l'acoustique est autokrigeable. Mais l'hypothèse d'autokrigeabilité est parfois difficile à tester et le paramètre α peut-être assez mal estimé. Le modèle de type Markov est alors une approximation du cokrigeage qui permet de ne pas modéliser la structure croisée entre les variables. La comparaison avec un "véritable" cokrigeage est un moyen de tester l'hypothèse d'autokrigeabilité de façon a posteriori (cette démarche est ici pédagogique puisque l'intérêt d'un modèle à résidu est en partie le fait qu'il évite d'avoir à mettre en oeuvre un cokrigeage "complet").

On peut voir que les estimations obtenues par cokrigeage et par le modèle à résidu sont relativement semblables (figures 2.21 et 2.22). Une fois de plus, la seule différence entre les deux modèles réside dans la modélisation du variogramme croisé pour quatre campagnes pour lesquelles l'acoustique a été considérée comme étant autokrigeable. Les deux années pour lesquelles les différences sont les plus importantes semblent être 1998 et 2001. Les corrélogrammes simples et croisés (figure 2.16) semblaient cependant tout à fait similaires, tout comme les variogrammes (figure 2.5, p.35). La régression linéaire montrait que l'estimation était tirée par une valeur forte (figure 2.18), le R^2 n'était que de 0.32, mais les estimations des autres campagnes ne semblaient pas meilleures.

Les différences principales entre le cokrigeage et le modèle à résidu en 1998 et 2001 viennent de points de radiales pour lesquels le modèle à résidu fournit une estimation plus forte que le cokrigeage. Ces points correspondent à de fortes valeurs acoustiques enregistrées souvent en bordure du champ et ayant échappé à la sélection sur la température de fond (cf. annexe A). Le modèle à résidu a alors tendance à donner un poids supérieur à la variable acoustique qu'un cokrigeage ordinaire. Les estimations en ses points sont alors beaucoup plus élevées que celles obtenues par cokrigeage, en particulier en 2001.

L'estimation des covariances croisées est souvent incertaine (Matheron 1979 [43]). Les variables étant très asymétriques, les modèles sont, pour les deux approches, très sensibles aux valeurs fortes. Il est possible que les différences entre les résultats des deux modèles soient dues à des problèmes d'instabilité numérique, l'hypothèse d'autokrigeabilité semblant acceptable pour l'ensemble des quatre années.

Les trois tables 2.2, 2.3 et 2.4, comparent les résultats obtenus par validation croisée pour chacune des deux approches. Dans les deux cas, ce sont les deux valeurs acoustiques et de capture qui ont été supprimées pour chaque station, en vue de ré-estimer la valeur de capture et de la comparer avec la

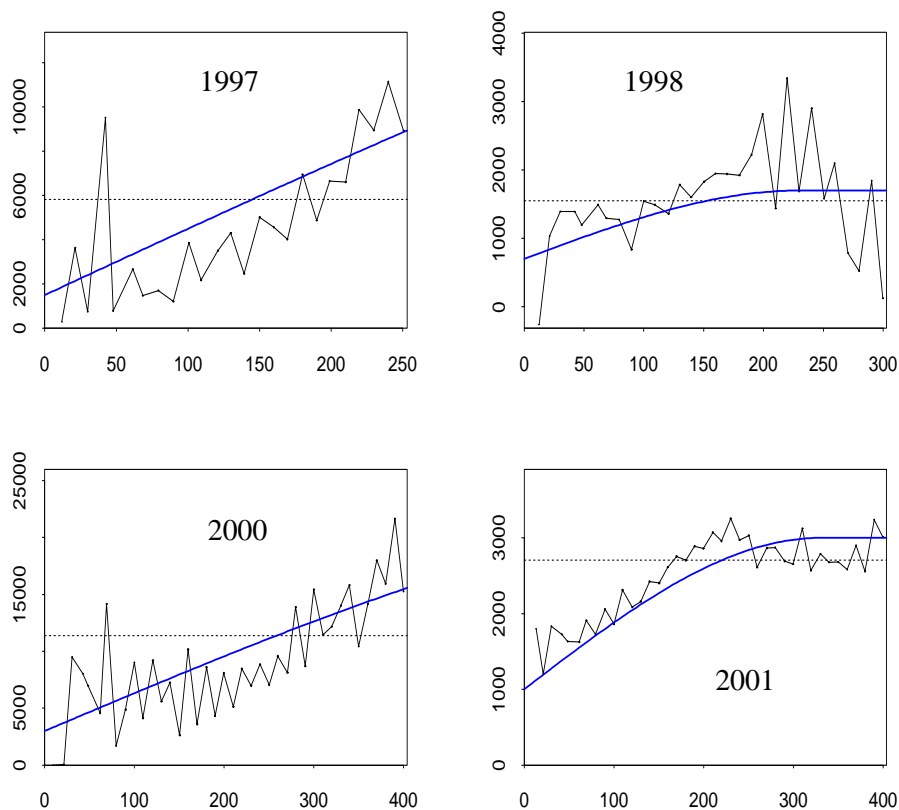


FIG. 2.20 – Variogrammes croisés acoustique-capture utilisés pour le cokrigage, Norvège 1997, 1998, 2000 et 2001.

Institut	Année	Modèle à résidu	Cokrigage
IMR	1997	10346.3	12731
IMR	1998	2002.1	2169
IMR	2000	1320.1	1466
IMR	2001	3020.3	3104

TAB. 2.2 – Moyenne des différences au carré entre les valeurs estimées en stations par validation croisée et les valeurs vraies pour chacun des modèles : $\frac{1}{N} \sum_i (P^*(x_i) - P(x_i))^2$

Institut	Année	Modèle à résidu	Cokrigage
IMR	1997	-0.0016	0.0017
IMR	1998	-0.00003	-0.079
IMR	2000	-0.0018	-0.046
IMR	2001	-0.0006	-0.015

TAB. 2.3 – Moyenne des erreurs de validation croisée standardisées : $\frac{1}{N} \sum_i \frac{P^*(x_i) - P(x_i)}{\sigma(x_i)}$

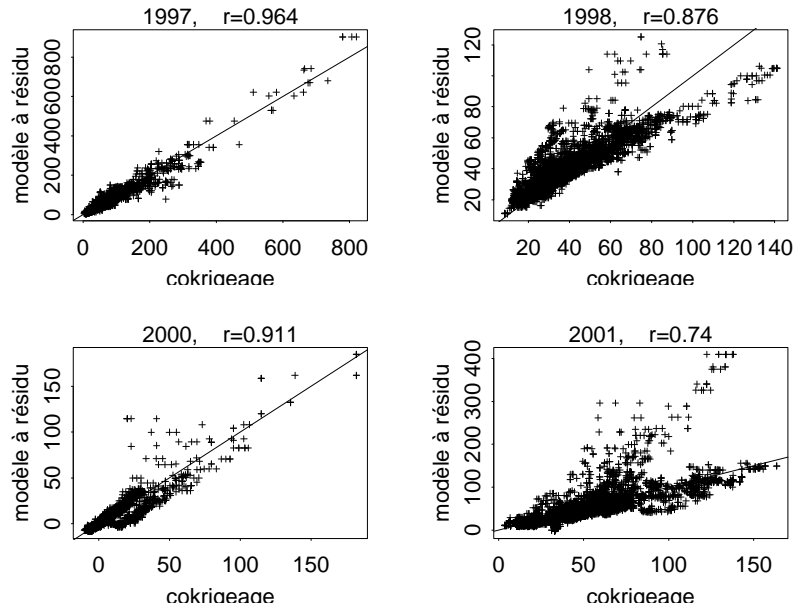


FIG. 2.21 – Nuage de corrélation entre les estimations de captures obtenues le long des transects par le modèle à résidu et par cokrigage, Norvège 1997, 1998, 2000 et 2001. Les droites tracées représentent la première bissectrice.

Institut	Année	Modèle à résidu	Cokrigage
IMR	1997	0.51	0.35
IMR	1998	0.2	0.22
IMR	2000	0.38	0.202
IMR	2001	0.41	0.37

TAB. 2.4 – Coefficient de corrélation entre les valeurs estimées en stations par validation croisée et les valeurs vraies pour chacun des modèles : $\rho(P^*(x_i), P(x_i))$

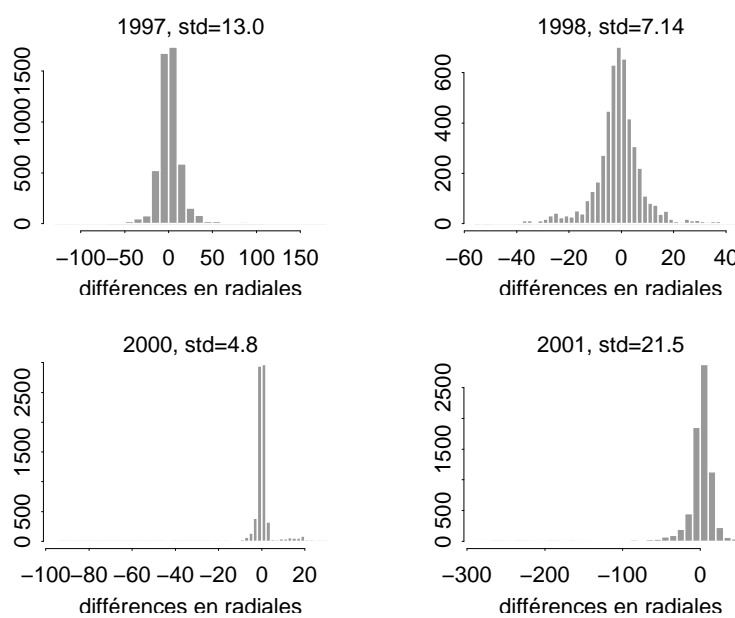


FIG. 2.22 – Histogramme des différences entre les estimations de captures obtenues le long des transects par le modèle à résidu et par cokrigage, Norvège 1997, 1998, 2000 et 2001.

valeur vraie. Il est intéressant de constater que le modèle à résidu permet d'obtenir une meilleure ré-estimation des points. Il semble que le fait de ne pas modéliser la structure croisée permet de gagner en robustesse, alors que les estimations restent de façon générale, très semblables aux valeurs obtenues par une estimation de cokrigeage classique.

2.5 Cas où les deux variables ont le même rôle : le modèle en corrélation intrinsèque

Le modèle en corrélation intrinsèque est un cas particulier du modèle à résidu, et donc également un cas particulier du cokrigeage, pour lequel la variable cible et la variable auxiliaire ont même structure (à un facteur multiplicatif près). On a donc :

$$\frac{C_P(h)}{\sigma_T^2} = \frac{C_A(h)}{\sigma_A^2} = \frac{C_\varepsilon(h)}{\sigma_\varepsilon^2} \quad (2.14)$$

Toutes les structures simples et croisées sont proportionnelles les unes aux autres. D'après la relation $C_P(h) = \alpha^2 C_A(h) + C_\varepsilon(h)$ on voit immédiatement que si la structure de la variable cible est proportionnelle à la structure de l'acoustique, elles sont nécessairement proportionnelles à la structure du résidu. Réciproquement, si toutes les structures sont proportionnelles, on se trouve dans le cas d'un modèle à résidu et cette même relation entraîne la proportionnalité de la structure du résidu avec celle des autre structures. Le cas isotopique est simple : le cokrigeage de toute variable est égal à son krigeage propre. Dans le cas hétérotopique, si l'on suppose que l'on a un modèle cohérent entre l'acoustique en radiales et l'acoustique en stations (par exemple que le modèle inféré dans le krigeage de l'acoustique en radiales est calé sur le palier de l'acoustique calculé en stations, voir partie 2.4.1) le cokrigeage s'écrit :

$$\begin{aligned} P^{CoK} &= \alpha A_R^{CoK} + \beta + \varepsilon^{CoK} \\ &= \alpha A_R^K + \beta + (P - \alpha A_S - \beta)^{CoK} \end{aligned}$$

Comme les captures et l'acoustique en stations sont en isotopie, toutes les variables sont autokrigeables :

$$(P - \alpha A_S - \beta)^{CoK} = P^K - \alpha A_S^K - \beta$$

On a donc la relation :

$$P^{CoK} = P^K + \alpha(A_R^K - A_S^K) \quad (2.15)$$

Le cokrigeage de la variable de capture est alors égal à son krigeage propre plus un terme correcteur proportionnel à la différence entre le krigeage de l'acoustique en radiales et le krigeage de l'acoustique en stations. Le seul gain pouvant être apporté par l'acoustique réside dans les détails présents dans l'interpolation réalisée avec les données de radiales mais inaccessibles pour les interpolations basées sur les seules données de stations. Il est intéressant de voir que dans la formulation du modèle à résidu (équation 2.6), l'estimation semble portée par l'acoustique et corrigée par le résidu. Lorsque les deux variables ont même structure, le modèle peut s'écrire différemment insistant alors au contraire sur le rôle correcteur de l'acoustique.

La similitude entre cette relation et celle présentée en introduction du chapitre sur l'estimation de la moyenne par régression linéaire, en situation de sous-échantillonnage de la variable cible, est intéressante. Mis à part que l'estimation de la moyenne tient ici compte de l'autocorrélation des

variables, les deux approches sont tout à fait équivalentes. Le rôle donné à chaque variable par le cokrigeage est le même que pour une régression linéaire (cf. p. 31).

La variance de cokrigeage s'écrit :

$$\begin{aligned}\sigma_{C_oK}^2(P) &= \alpha^2 \sigma_K^2(A_R) + \sigma_K^2(\varepsilon) \\ &= \alpha^2 \sigma_K^2(A_R) + (1 - \alpha^2 \frac{\sigma_{A_S}^2}{\sigma_P^2}) \sigma_K^2(P)\end{aligned}$$

Or le paramètre α peut s'écrire en fonction du coefficient de corrélation entre les deux variables ρ :

$$\alpha^2 = \frac{C_{A,P}^2(0)}{(\sigma_{A_S}^2)^2} = \rho^2 \frac{\sigma_P^2}{\sigma_{A_S}^2}$$

On a donc la relation entre la variance de cokrigeage et les variances de krigeage propre à chaque variable :

$$\sigma_{C_oK}^2(P) = \rho^2 \frac{\sigma_P^2}{\sigma_{A_S}^2} \sigma_K^2(A_R) + (1 - \rho^2) \sigma_K^2(P) \quad (2.16)$$

Si le palier du modèle utilisé pour krigeage de l'acoustique en radiales n'a pas été renormé au niveau du palier observé sur les stations, cette formule doit être corrigée par le coefficient k^2 (cf. partie 2.4.1) :

$$\sigma_{C_oK}^2(P) = \frac{\rho^2}{k^2} \frac{\sigma_P^2}{\sigma_{A_S}^2} \sigma_K^2(A_R) + (1 - \rho^2) \sigma_K^2(P)$$

On voit par cette relation que la variance de cokrigeage réalise un compromis entre la variance de krigeage de la capture et la variance de krigeage de l'acoustique au travers du coefficient de corrélation. Plus celui-ci est faible, plus le cokrigeage et sa variance seront proches du krigeage ordinaire des captures. Une fois encore, la variance de l'estimateur de cokrigeage est similaire à l'expression de la variance de l'estimateur de la moyenne lors de la régression linéaire (équation 2.2, p. 32).

Étudions à nouveau l'exemple ne comprenant qu'une seule station et deux points hors stations présenté précédemment (cf. figure 2.19) et un modèle de corrélation intrinsèque dans lequel varie le coefficient de corrélation entre les deux variables :

$$\begin{aligned}C_p(h) &= \text{sph}(h/5) \\ C_{A,P}(h) &= \rho_{AP} \text{sph}(h/5) \\ C_A(h) &= \text{sph}(h/5) \\ &\text{avec } \rho_{AP} \in [0; 0.9]\end{aligned}$$

On voit sur la figure 2.23 que le poids de la donnée de capture est constant et positif, quelque soit le coefficient de corrélation et le point cible, alors que pour un cokrigeage standard il tendait à diminuer pour les valeurs fortes du coefficient de corrélation. L'effet écran est très fort : le poids de R_2 est nul pour l'estimation de $x_{0,2}$ et $x_{0,1}$ et réciproquement le poids de R_1 est négatif pour l'estimation de $x_{0,3}$. Le ratio augmente linéairement pour chacun des points estimés, alors que la croissance s'accélérait pour les configurations de structure précédentes. Le poids de l'acoustique en stations est lui toujours négatif. Il est intéressant de voir que pour l'estimation de $x_{0,1}$ le ratio est toujours inférieur à 1 quelque soit le coefficient de corrélation. Le point $x_{0,1}$ est équidistant de la station et de R_1 , mais le modèle fait pourtant jouer un rôle privilégié à la donnée de capture.

Le modèle fait jouer un rôle plus important à la variable cible, la variable de capture, que le modèle

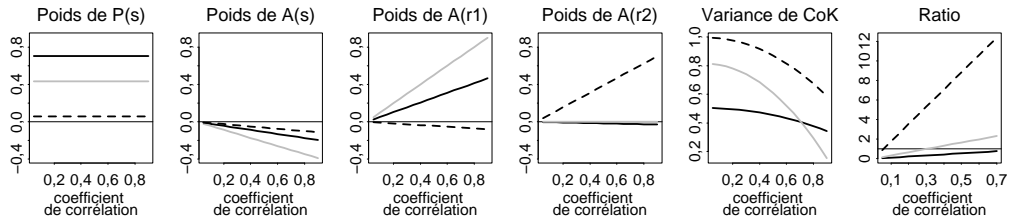


FIG. 2.23 – Évolution des poids de cokrigage de chaque point, de la variance de cokrigage ainsi que du ratio des poids accordés à l’acoustique sur le poids accordé à la donnée de capture (cf. équation 2.4) pour une configuration de structure où les deux variables sont en corrélation intrinsèque et où le coefficient de corrélation entre les deux variables varie. La ligne pointillée correspond à l’estimation du point $x_{0,3}$, la ligne grise au point cible $x_{0,2}$ et la ligne noire au point $x_{0,1}$.

à résidu. Il semble que l’acoustique ne joue vraiment un rôle important dans l’estimation que pour les points très éloignés des stations.

Tester l’hypothèse de proportionnalité entre les captures et l’acoustique est délicat puisque, une fois de plus, les deux variables ne sont pas disponibles aux mêmes résolutions. Pour 3 des 4 campagnes jugées correspondre à l’hypothèse d’autokrigeabilité de l’acoustique en Norvège, l’hypothèse de corrélation intrinsèque a également été considérée comme acceptable. Naturellement ici encore, si l’hypothèse est abusive, l’estimation obtenue ne sera qu’une approximation du cokrigage, permettant de négliger la modélisation des structures croisées. On voit sur les variogrammes simples et croisés (cf. figure 2.5, p. 35) de l’acoustique et des captures que les structures empiriques sont très proches. La figure 2.24 compare pour les quatre campagnes pour lesquelles l’acoustique a été considérée comme autokrigeable, le modèle choisi pour le résidu et le modèle imposé pour celui-ci si l’on suppose de plus que les variables sont en corrélation intrinsèque. En 2001, l’hypothèse de corrélation intrinsèque semble trop forte, le nouveau modèle s’ajuste mal au variogramme empirique. Par contre, pour les trois autres campagnes, le modèle de variogramme choisi pour le modèle à résidu est extrêmement proche du modèle en corrélation intrinsèque.

Il est alors intéressant de visualiser pour les trois campagnes où l’hypothèse de corrélation intrinsèque est acceptable, la carte des corrections apportées par l’acoustique, qui correspond à la différence entre l’estimation par krigeage ordinaire en utilisant toutes les données acoustiques et celle n’utilisant que l’acoustique en stations (figure 2.25). On voit que les corrections sont assez locales. L’acoustique permet également de réduire l’influence de quelques valeurs fortes qui ont tendance à créer des zones fortes assez étalées dans le modèle monovarié, comme dans la zone sud-est pour la campagne 2000. Leur impact est minimisé par l’information additionnelle des données en radiales.

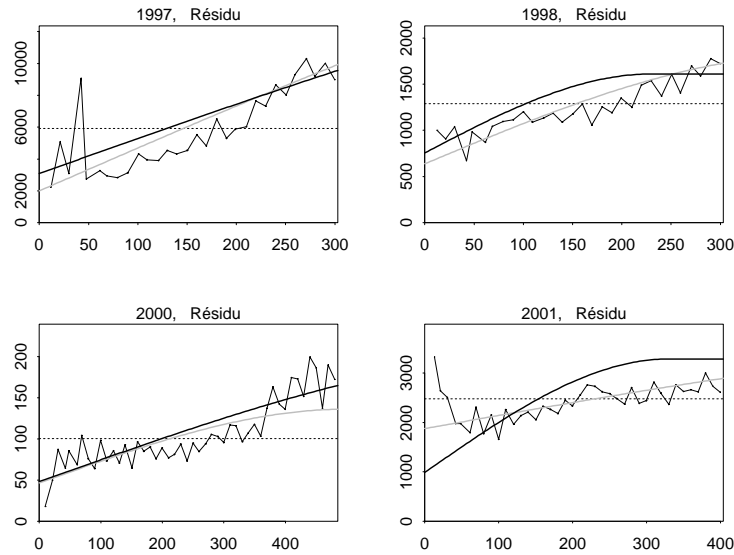


FIG. 2.24 – Superposition aux variogrammes empiriques des résidus des modèles ajustés lors de l’approche de type Markov (courbe grise), et lorsque l’on impose de plus une hypothèse de corrélation intrinsèque entre les variables (courbe noire), Norvège 1997, 1998, 2000 et 2001.

2.6 Cas où la variable auxiliaire est utilisée en contrainte de forme

Très largement utilisé, le modèle à dérive externe n’est pas un cas particulier du cokrigeage. Il est cependant très proche du modèle à résidu. Il est basé sur un lien linéaire entre la variable auxiliaire et la moyenne locale de la variable cible. L’hypothèse forte de ce modèle est de supposer la variable auxiliaire *déterministe*. Celle-ci doit également être disponible de façon exhaustive sur l’ensemble du champ étudié, ou plus précisément, sur l’ensemble constitué des points à estimer et des points où la variable cible est connue. Le procédé est le même qu’un krigeage ordinaire de la variable principale mis à part que des contraintes sont ajoutées aux poids de façon à imposer la relation (Wackernagel 1998 [70], Goovaerts 1997 [29]) :

$$\mathbb{E}(P(x)) = \alpha A(x) + \beta$$

La notation de l’acoustique $A(x)$ par une majuscule est ici abusive puisque il ne s’agit pas d’une fonction aléatoire. Cependant, elle est conservée pour plus de cohérence avec les chapitres précédents. La variable de pêche s’écrit alors comme la somme d’une dérive à paramètres inconnus et d’un résidu :

$$P(x) = \alpha A(x) + \beta + \varepsilon(x)$$

$A(x)$ étant déterministe on a :

$$P^{KeD}(x) = \alpha A(x) + \beta + \varepsilon^K(x)$$

L’estimation est une combinaison linéaire des seules données de captures, elle ne dépend que de la structure du résidu qui est supposée être la même que la structure de la variable cible. L’acoustique

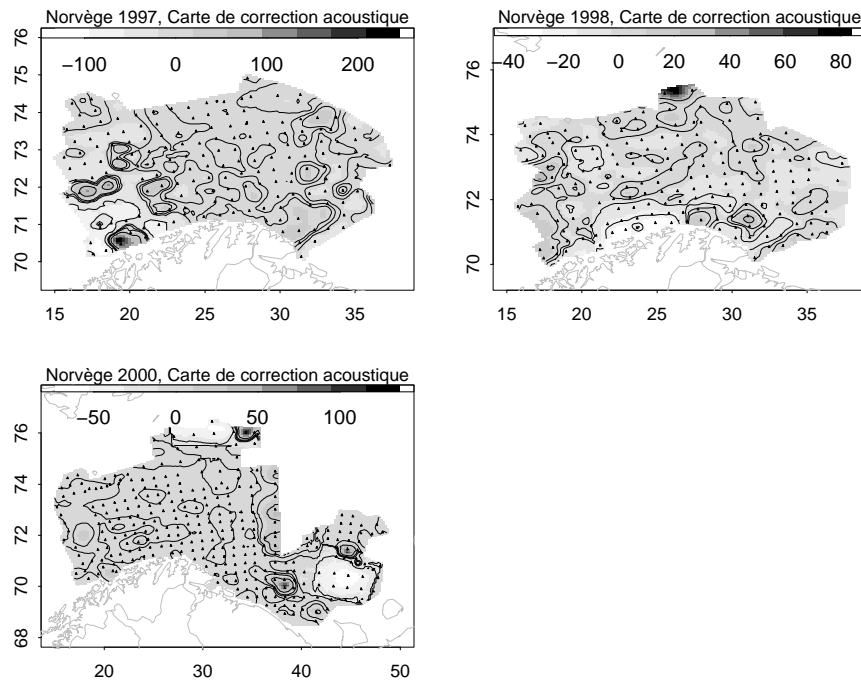


FIG. 2.25 – Correction apportée par l'acoustique dans le modèle en corrélation intrinsèque : $\alpha(A_R^K - A_S^K)$, Norvège 1997, 1998, 2000.

impose la forme que l'estimation doit suivre. Les deux paramètres α et β ne sont pas estimés directement mais filtrés par le système. Celui-ci assure juste que les poids optimaux vérifient la contrainte. La variance d'estimation est donc nécessairement plus grande qu'une variance de krigeage, puisque le système est le même avec des contraintes plus fortes. Au mieux la variance du krigeage à dérive externe sera égale à la variance d'un krigeage ordinaire.

En pratique, l'acoustique n'est pas disponible sur l'intégralité du champ étudié. Le modèle utilisé est donc différent du modèle théorique puisque l'acoustique y est remplacée par son krigeage :

$$P^{KeD}(x) = \alpha A^K(x) + \beta + \varepsilon^K(x)$$

Supposer la variable auxiliaire déterministe alors qu'il s'agit d'une interpolation est assez abusif. La variance de cette estimation n'est en outre pas prise en compte dans l'estimation de $P(x)$.

Lorsque le résidu est orthogonal à l'acoustique (cette fois-ci considérée donc comme variable aléatoire) et que les estimations sont effectuées à voisinage unique, le krigeage à dérive externe est équivalent au modèle de type Markov où la variable acoustique est maîtresse (Rivoirard 2002 [58]). Cependant, le fait de supposer l'acoustique comme étant déterministe permet d'éviter artificiellement toute hypothèse de lien structural entre la variable auxiliaire et la variable cible.

Lorsque le krigeage à dérive externe est effectué en voisinage glissant, les deux paramètres α et β peuvent changer avec le voisinage. Leurs valeurs, filtrées par le système, sont inconnues, tout comme celles du résidu. Une hypothèse d'orthogonalité est donc en pratique difficile à vérifier.

Par contre, les paramètres pouvant changer avec le voisinage, le modèle à dérive externe offre une plus grande souplesse que le modèle à résidu.

L'interprétation du modèle passe donc par l'estimation des paramètres, en particulier du paramètre multiplicatif de la dérive α . Plus α est grand localement plus le poids donné à l'acoustique est important. L'estimation de α s'obtient simplement en remplaçant le membre de droite du système de krigeage par un vecteur nul, sauf sur sa dernière composante, égale à 1 (Matheron 1970, [41], Castelier 1993, [14]). L'estimateur du paramètre α est une combinaison linéaire des données de captures :

$$\hat{\alpha} = \sum_{i \in S} \lambda_i P(i)$$

De la même façon, le résidu peut être estimé pour tester postérieurement les hypothèses. L'estimateur du résidu s'obtient en annulant les deux derniers termes du membre de droite du système de krigeage. Mais on n'obtient ainsi qu'une estimation du résidu ou des paramètres avec une certaine précision ; leurs vraies valeurs restent inconnues.

De plus, aucune contrainte n'est imposée sur les paramètres, ceux-ci peuvent donc devenir localement négatifs. Lorsque la variable cible suit localement une tendance contraire à l'acoustique, le krigeage à dérive externe aura tendance à mettre un poids fortement négatif à la dérive. Dans la plupart des applications, il semble absurde que le paramètre de la dérive puisse localement changer de signe. Le poids doit pouvoir devenir localement nul, si les deux variables fournissent des informations trop contradictoires. Si le poids est négatif, la variable de pêche ne suit non pas la forme indiquée par les enregistrements acoustique, mais son contraire, ce qui n'a aucun sens.

Une possibilité pour contraindre la positivité du paramètre α serait l'approche suivante :

1. évaluer l'estimation α^* du paramètre α
2. seuiller l'estimation à zéro : $\alpha^{**} = \alpha^* \mathbb{1}_{\alpha^* \geq 0}$
3. faire le krigeage ordinaire du résidu $P(x) - \alpha^{**} A(x)$ (pour éviter l'estimation de β)

4. en déduire l'estimation $P^*(x)$

Le risque est de multiplier les approximations en remplaçant les valeurs par leurs estimations lorsque celles-ci sont peu précises. En pratique l'estimation obtenue est assez différente des résultats du modèle de krigeage à dérive externe.

En mer de Barents, pour les quatre années pour lesquelles le modèle à résidu était vérifié, c'est à dire où la variable acoustique était autokrigeable, un krigeage à dérive externe a également été effectué dans le but de comparer les résultats obtenus ainsi que les valeurs du coefficient multiplicatif. Le krigeage de l'acoustique ainsi que le krigeage du résidu, avaient été effectués en voisinage glissant dans le modèle à résidu (cf. chapitre précédent). Le krigeage à dérive externe a également été effectué en utilisant le même krigeage pour l'acoustique. Par contre l'estimation du résidu est différente, en particulier parce que les valeurs des paramètres changent avec le voisinage. Il est intéressant de localiser les zones où l'acoustique a le plus d'influence et de confronter les résultats du krigeage à dérive externe avec ceux du modèle à résidu.

Les variances d'estimation ne sont pas vraiment comparables telles quelles, puisque par construction, le krigeage à dérive externe a une variance plus grande. Les outils de validation croisée ont donc été utilisés comme critères de choix.

Il est intéressant de voir que le paramètre multiplicatif, ou plutôt son estimation, est très variable spatialement pour la plupart des campagnes. Il est également souvent négatif.

Pour les années 1997, 1998 et 2001, une étude plus précise de l'impact du choix du voisinage dans l'estimation a été effectuée afin d'analyser comment converge le paramètre multiplicatif au fur et à mesure que l'on élargit le voisinage. Pour cela le krigeage à dérive externe a été effectué avec un voisinage carré variant de 100 m.n de côté jusqu'à devenir unique. Selon les campagnes, le coefficient multiplicatif de la dérive converge différemment. En 1997, lorsque le voisinage est petit, le coefficient varie beaucoup sur l'ensemble de la zone d'étude (figure 2.26). Au fur et à mesure que le voisinage augmente, le coefficient semble se stabiliser autour de 0.5 sur la majeure partie de la zone, excepté au sud-ouest où sa valeur reste élevée. En voisinage unique, le coefficient atteint 2.16. Alors que sa valeur semblait stabilisée sur la plus grande partie de la mer pour les voisinages moyens, elle réaugmente assez brusquement en raison de l'influence de la zone sud-ouest. Vraisemblablement, cette zone est marquée par une bonne adéquation entre les valeurs de l'acoustique et les valeurs de capture et c'est de plus la zone la plus riche de la campagne. L'influence de ces quelques stations est très importante dans l'estimation du paramètre.

Pour les autres campagnes, la convergence est plus symétrique, le coefficient se stabilise autour de la valeur la plus fréquente de l'ensemble du champ (figure 2.27).

Il est intéressant de voir que la valeur la plus fréquente spatialement pour un voisinage de taille moyenne est relativement stable d'années en années et semble se situer entre 0.3 et 0.5. Cette valeur peut-être considérée comme une estimation plus robuste du paramètre d'autokrigeabilité de l'acoustique.

Il est également intéressant d'étudier le voisinage permettant d'avoir le modèle le plus précis, d'après un critère de validation croisée. La figure 2.28 montre que pour trois critères différents de comparaison des valeurs estimées par validation croisée avec les valeurs vraies observées en stations, les résultats sont meilleurs à voisinage unique. Galli et Meunier (1987 [22]) observent sur une étude similaire de krigeage de données de puits d'un réservoir de gaz avec une interpolation de données sismiques en dérive externe, que les meilleurs résultats par validation croisée sont obtenus à voisinage

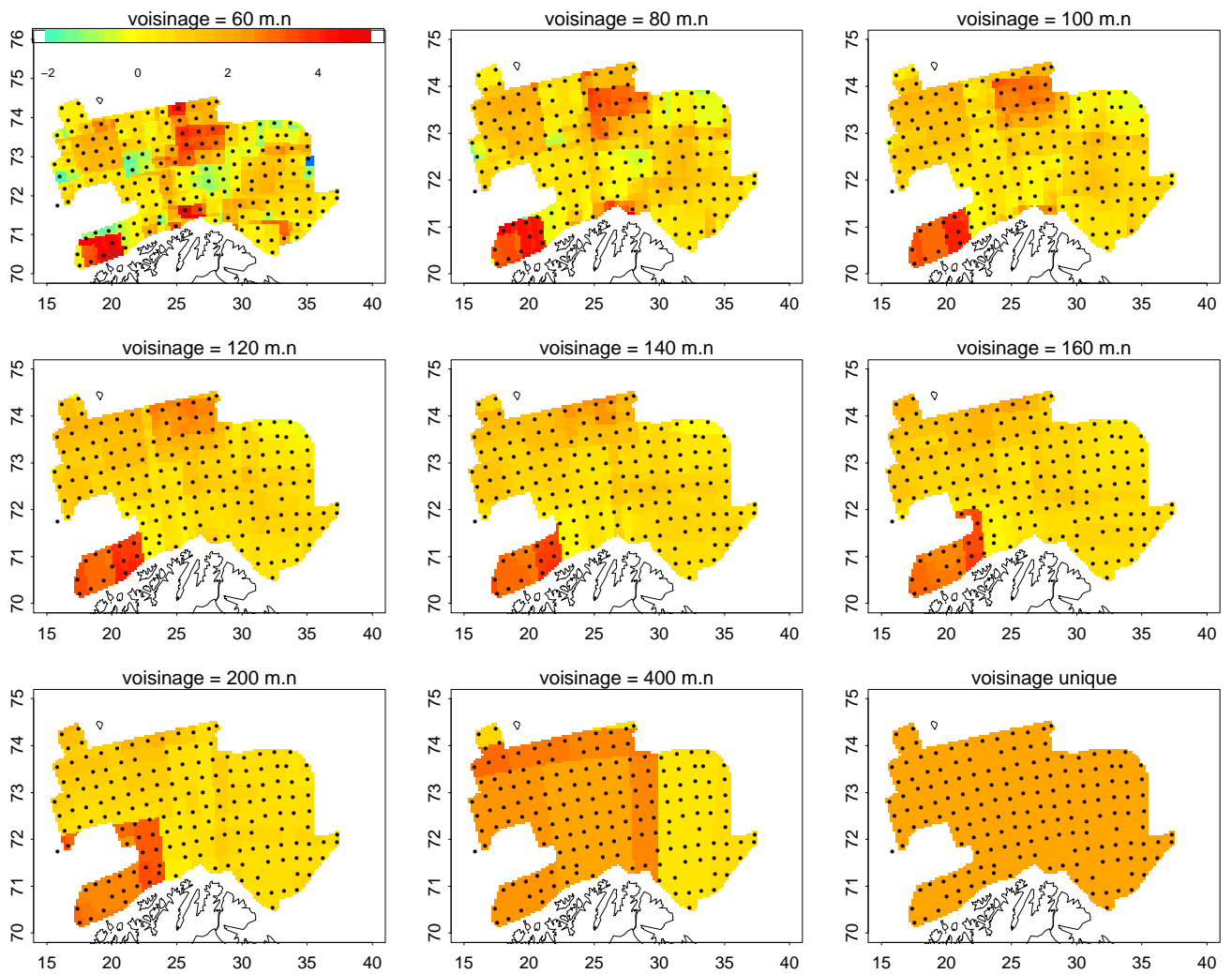


FIG. 2.26 – Estimation du coefficient multiplicatif de la dérive en fonction de la taille du voisinage, dans le krigeage à dérive externe, Norvège 1997.

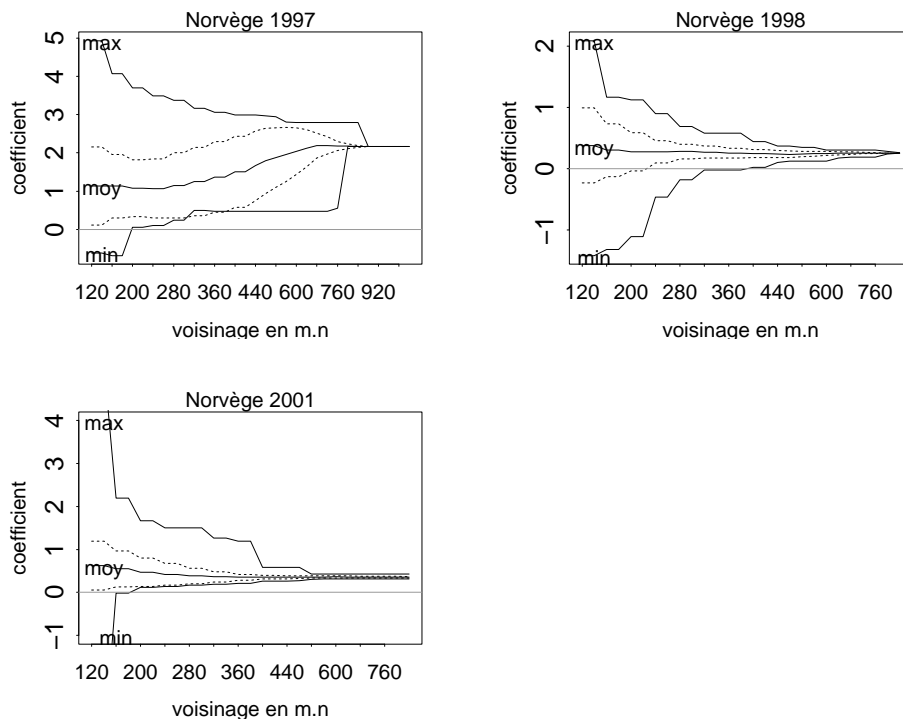


FIG. 2.27 – Convergence du coefficient multiplicatif de la dérive en fonction de la taille du voisinage, dans le krigeage à dérive externe, Norvège 1997, 1998, 2001. Les lignes pointillées représentent la moyenne du coefficient plus ou moins un écart-type.

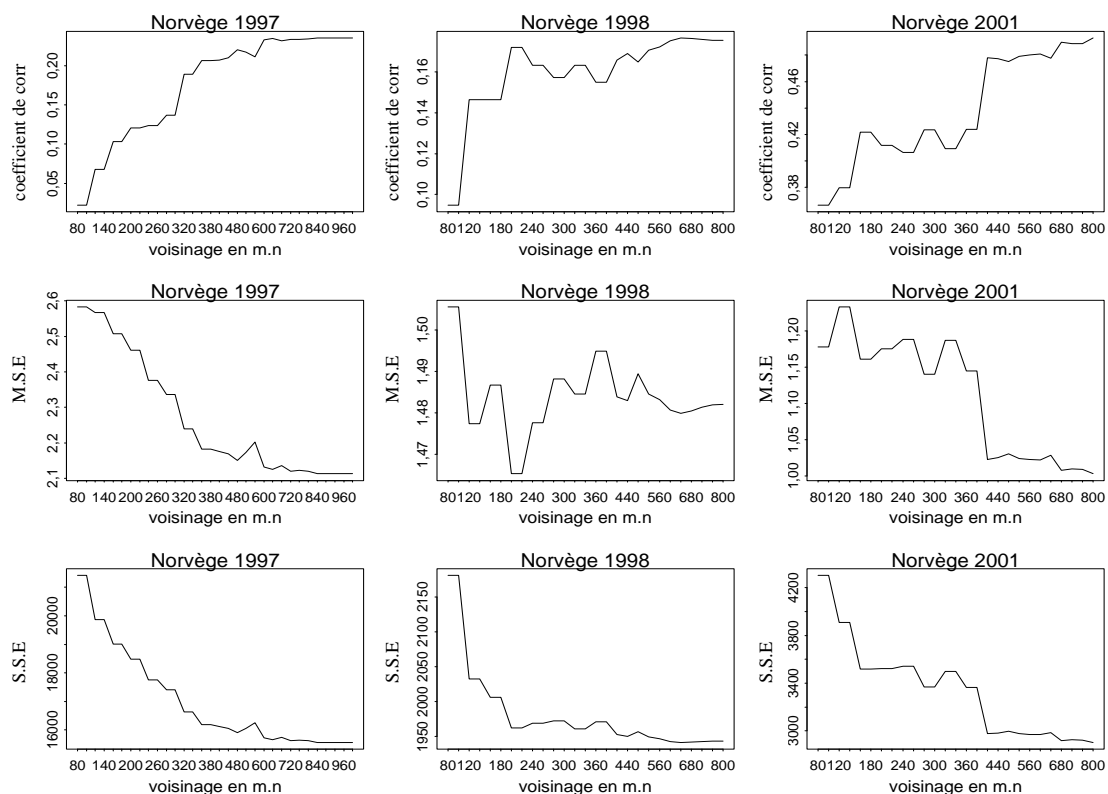


FIG. 2.28 – Évolution de trois critères de validation croisée en fonction de la taille du voisinage dans le krigeage à dérive externe : le coefficient de corrélation entre les valeurs estimées par validation croisée et les valeurs vraies (1ère ligne), la moyenne des erreurs quadratiques standardisées (2e ligne) et la somme des erreurs quadratiques (3e ligne), Norvège 1997, 1998, 2001.

unique.

Il est possible de montrer que le krigeage à dérive externe est équivalent à une régression linéaire où le produit scalaire est celui induit par l'inverse de la matrice de covariance de la variable cible (Castelier 1993, [14]). Tout comme la régression linéaire, le krigeage à dérive externe est donc très sensible au nombre de points sur lesquels est faite l'estimation. Plus le voisinage est grand, plus celle-ci est susceptible d'être robuste.

2.7 Estimation non-linéaire de l'acoustique

La souplesse du krigeage à dérive externe a certains avantages. Le lien fonctionnel existant entre la variable cible et la variable auxiliaire étant ignoré, la dérive peut être issue de différents processus d'estimation. Évidemment, la rigueur mathématique de l'approche devient alors discutable. Mais l'estimation peut cependant gagner en robustesse tout en restant simple.

L'acoustique est connue de façon très dense sur l'ensemble du champ. Cependant, comme cela a été vu à plusieurs reprises, ses valeurs suivent une distribution très asymétrique. Quelques rares valeurs fortes cachent le comportement des valeurs plus faibles et modifient considérablement les estimations. Les valeurs fortes sont très peu nombreuses mais représentent à elles seules une part importante de l'abondance totale. En mer du Nord, par exemple, plus de 65% de l'énergie totale a été collectée dans seulement 3% des stations. Le modèle d'estimation est fortement dépendant de la probabilité de rencontrer ou non de telles valeurs au cours de la campagne. Il est donc important d'étudier l'agencement dans l'espace de ces valeurs par rapport aux valeurs plus faibles, dans le but d'obtenir une estimation plus robuste.

Le codage en indicatrice consiste à découper la distribution statistique de la variable en différents intervalles. Le krigeage disjonctif repose sur une estimation spatiale cohérente de ces indicatrices. De plus, comme on l'a vu dans la partie traitant de l'analyse structurale, l'acoustique découpée en classes est assez bien corrélée aux captures, en particulier en mer de Barents et en mer d'Irlande. Les variogrammes d'indicatrices, en mer du Nord et en mer d'Irlande, montrent que les classes pauvres ont des structures longues portées et moins d'effet de pépite que les classes riches, qui sont très déstructurées.

Dans la pratique, pour une compréhension plus facile des structures d'indicatrices, celles-ci ne correspondent pas à des intervalles mais à des classes cumulées. On étudie donc les indicatrices : $\mathbb{1}_{A(x) \geq c_1}$, $\mathbb{1}_{A(x) \geq c_2}$, $\mathbb{1}_{A(x) \geq c_3}$, etc... où $c_1 \leq c_2 \leq c_3$ sont les coupures choisies.

De plus, il est intéressant de constater que le découpage en indicatrice permet un gain important en robustesse. Les structures sont très stables d'années en années indiquant que la géométrie spatiale de chacune des classes est assez constante au fil des ans (figure 2.29). Ce phénomène s'observe même pour des zones où les structures brutes sont très difficiles à observer, comme la mer du Nord. Enfin, lorsque l'on modifie légèrement la coupure, on s'aperçoit que la structure est assez robuste. Le palier a tendance à changer, mais la portée est la même pour une coupure variant de plus ou moins 10%, montrant que l'approche est assez fiable et n'est pas trop dépendante du choix arbitraire des coupures (Bouleau 2003, [8]).

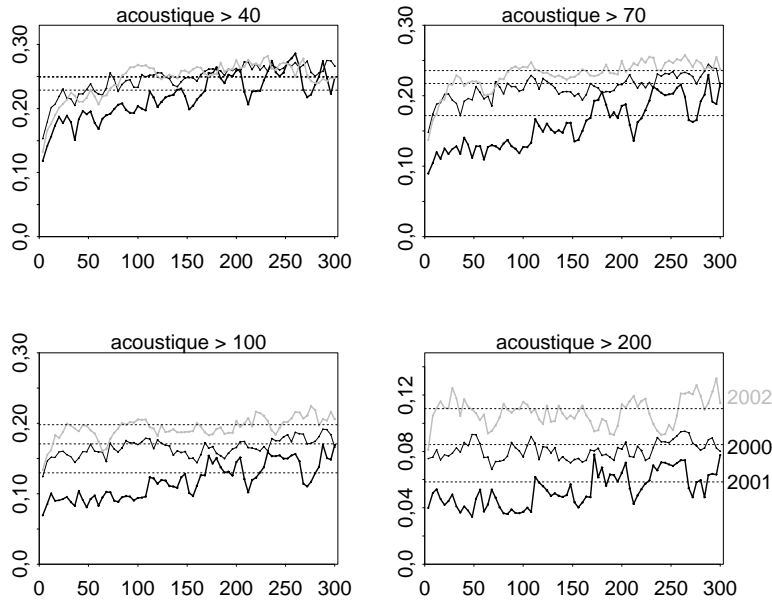


FIG. 2.29 – Variogrammes empiriques des indicatrices de classes en fonction des années, données CEFAS 2000, 2001, 2002

2.7.1 Krigeage disjonctif

Il y a un effet de bord (définition, cf. Rivoirard 1994, [56]) entre deux indicatrices lorsque la probabilité d'appartenir à une classe sachant que l'on est dans une autre augmente avec la distance. Il y a un effet de bord dans l'intervalle I si les bordures de I sont en moyenne plus pauvres que le centre de I (Rivoirard 1994, [56]), c'est à dire s'il faut, pour aller d'une classe pauvre à une classe riche, traverser une classe moyenne. L'étude de ces effets de bord permet donc de mieux comprendre la distribution spatiale des données et la relation entre les différentes classes.

La probabilité que la variable d'intérêt dépasse une coupure donnée sachant que l'on entre dans la classe correspondant à la coupure inférieure est égale au rapport entre le variogramme croisé et le variogramme simple de l'indicatrice :

$$\mathbb{P}[A(x) \geq c_2 | A(x) \geq c_1, A(x+h) < c_1] = \frac{\gamma_{12}(h)}{\gamma_1(h)}$$

où $\gamma_{12}(h)$ est le variogramme croisé des deux indicatrices $\mathbb{1}_{A(x) \geq c_1}$ et $\mathbb{1}_{A(x) \geq c_2}$ et $\gamma_1(h)$ est le variogramme simple de $\mathbb{1}_{A(x) \geq c_1}$.

En l'absence d'effet de bord, un modèle simple consiste à effectuer un changement de variables permettant d'obtenir des facteurs orthogonaux : le modèle à résidus orthogonaux d'indicatrices. Les

résidus sont calculés de la façon suivante :

$$\begin{aligned} H_0(A(x)) &= 1 \\ H_1(A(x)) &= \frac{\mathbb{1}_{A(x) \geq c_1}}{\mathbb{E}[\mathbb{1}_{A(x) \geq c_1}]} - 1 \\ H_2(A(x)) &= \frac{\mathbb{1}_{A(x) \geq c_2}}{\mathbb{E}[\mathbb{1}_{A(x) \geq c_2}]} - \frac{\mathbb{1}_{A(x) \geq c_1}}{\mathbb{E}[\mathbb{1}_{A(x) \geq c_1}]} \\ H_3(A(x)) &= \frac{\mathbb{1}_{A(x) \geq c_3}}{\mathbb{E}[\mathbb{1}_{A(x) \geq c_3}]} - \frac{\mathbb{1}_{A(x) \geq c_2}}{\mathbb{E}[\mathbb{1}_{A(x) \geq c_2}]} \\ &\text{etc...} \end{aligned}$$

Ils sont, par construction, orthogonaux. Leur cokrigeage est donc égal à leur krigeage, ce qui permet de simplifier les calculs. Petitgas (1993, [51]) utilise ce type de modèle pour l'étude d'une prospection acoustique du hareng sur la côte norvégienne. Dans ce modèle, en effet, la portée des variogrammes diminue pour les hautes coupures, ce qui semble parfaitement représenter la distribution des données acoustiques. Une caractéristique de ces données est le fait que les zones de valeurs fortes sont souvent assez peu dispersées spatialement et peuvent apparaître au milieu d'une zone pauvre, sans transition par une classe moyenne, ce qui correspond précisément à ce modèle.

En mer du Nord, des effets de bords sont observés sur les classes pauvres, mais pas entre les classes moyennes et riches. Négliger la propriété diffusive des classes pauvres en utilisant qu'une seule indicatrice pour ces différentes classes structurées de valeurs basses revient à remplacer les valeurs faibles par leur moyenne. Cependant, cette approximation n'aura qu'une influence faible dans l'abondance totale. Le plus important pour l'estimation du stock est de modéliser les zones riches qui, elles, sont sans effet de bord. Quelque soit la campagne étudiée, il est possible de déterminer deux à trois coupures fournissant des classes d'effectifs comparables (autour de 15 à 20%) sans effet de bord. Seule l'année 2003 (pour les campagnes françaises et écossaises) présente des effets de bord jusqu'à des coupures assez hautes (figure 2.30). De 1999 à 2002, les campagnes françaises, anglaises et écossaises présentent chacune cet effet de déstructuration des hautes teneurs et l'absence d'effet de bord des classes moyennes et riches, tout comme les trois campagnes disponibles en mer d'Irlande (2000, 2001 et 2002). Par contre les campagnes en mer de Barents sont beaucoup plus structurées et montrent des effets de bords entre toutes les classes pauvres, moyennes et riches, indiquant une distribution spatiale des données beaucoup plus diffuse. Le modèle à résidu d'indicateurs est donc mal adapté à décrire les données norvégiennes. Cependant celles-ci sont suffisamment structurées pour ne pas nécessiter de découpage en indicatrices (des modèles satisfaisants ont été obtenus sur les variables brutes dans les parties précédentes).

Généralement, pour les campagnes en mer du Nord et en mer d'Irlande, le dernier résidu d'indicateur est pépétique (figure 2.31). Cela signifie qu'à partir d'un certain niveau élevé de valeurs, seul le contour des zones est structuré, mais les valeurs elles mêmes au sein de la zone sont distribuées de façon pépétique.

Cette approche fournit des cartes de probabilité d'occurrence de valeurs d'une classe donnée. Elle ne permet pas d'obtenir directement une carte d'estimation de l'abondance acoustique. Une estimation d'abondance moyenne a donc été calculée en partant de la relation de décomposition par indicatrice d'une variable aléatoire discrète :

$$\mathbb{E}[Z(x)] = \sum_{j=1}^J c_j \mathbb{1}_{Z(x)=c_j}$$

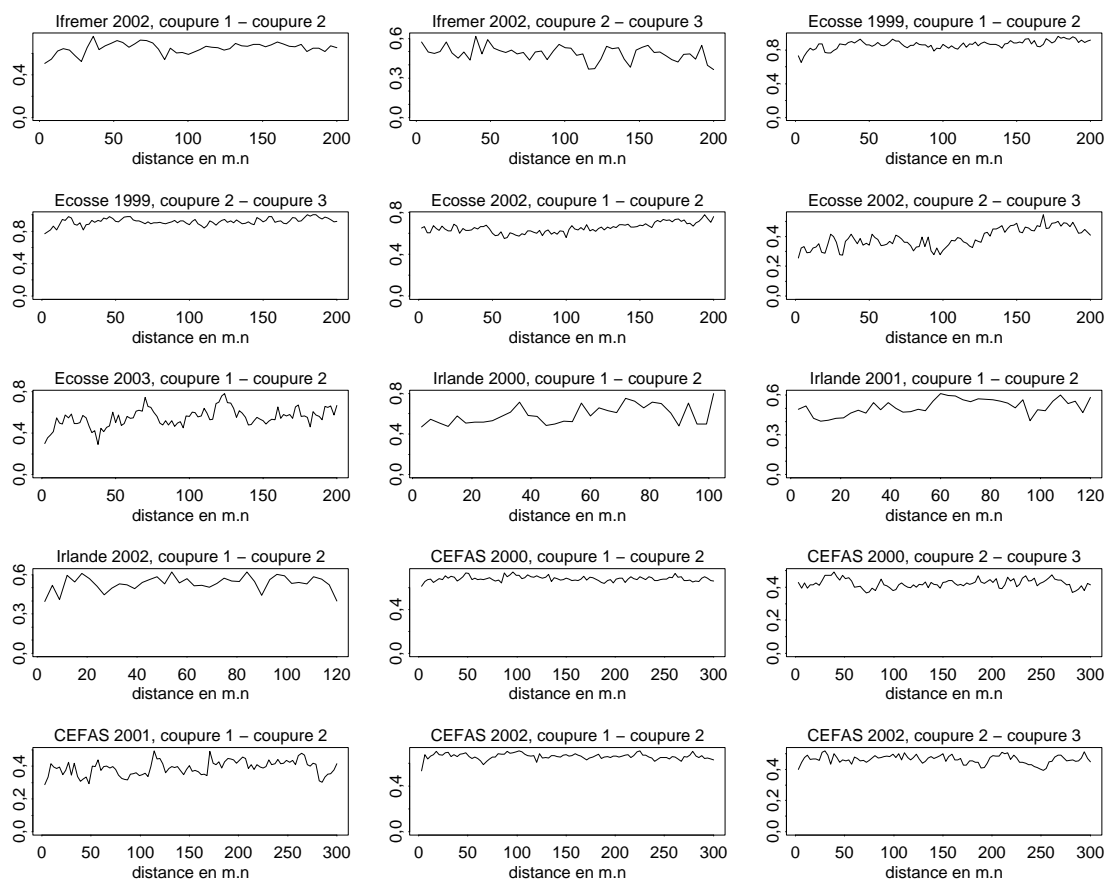


FIG. 2.30 – Effets de bord entre les classes (rapport du variogramme croisé d'indicatrices sur le variogramme simple) pour chacune des campagnes. De haut à gauche à en bas à droite : IFREMER 2002 (3 coupures), Écosse 1999 (3 coupures), 2002 (3 coupures), 2003 (2 coupures), Irlande 2000, 2001 et 2002 (2 coupures) et CEFAS 2000 (3 coupures), 2001 (2 coupures) et 2002 (3 coupures).

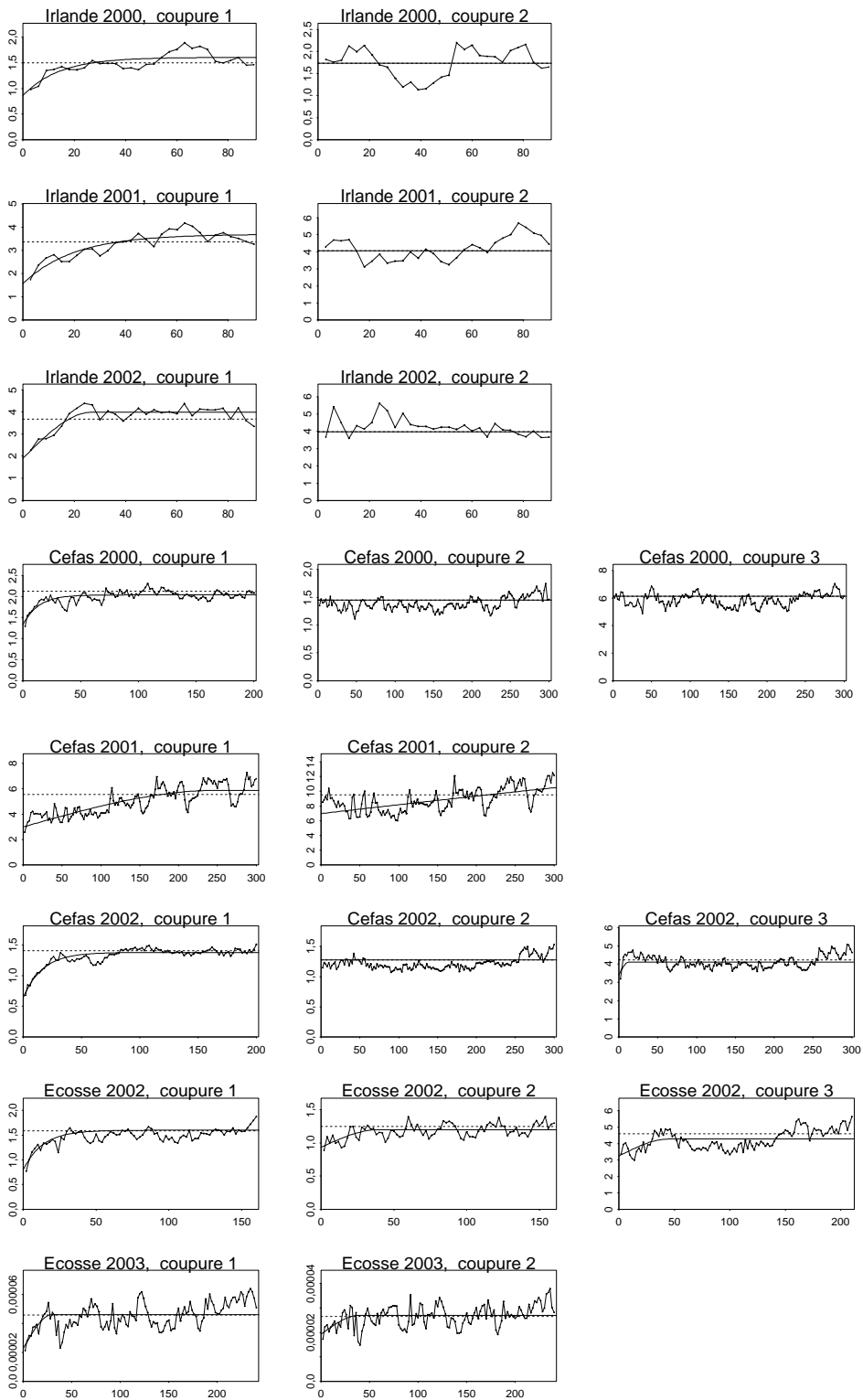


FIG. 2.31 – Variogrammes empiriques et modélisés des résidus d'indicatrices. Irlande 2000, 2001 et 2002, CEFAS 2000, 2001 et 2002 et Écosse 2002 et 2003

La variable acoustique ne prend pas ses valeurs dans \mathbb{N} . Cependant, on peut l'approcher par l'estimateur :

$$\widehat{A}(x) = \sum_{j=1}^{J-1} m_j (\mathbb{1}_{c_j \leq A(x) < c_{j+1}})^{KD}$$

avec m_j la moyenne de $A(x)$ sur l'intervalle $[c_j, c_{j+1}[$.

L'estimation sera d'autant plus précise que le nombre de classes J sera élevé. Mais ici seulement entre trois et quatre classes, selon les campagnes, ont été utilisées. Cependant, malgré la grossièreté de l'approximation, cette approche permet l'estimation d'une abondance acoustique pour les campagnes où la déstructuration des valeurs fortes était trop importante pour permettre l'inférence d'un variogramme et donc un krigeage des données brutes.

2.7.2 Espérance conditionnelle

Une autre approche permettant l'estimation d'une indicatrice de $A(x)$ en cohérence avec l'ensemble de la distribution de A est d'utiliser la loi statistique de cette variable. L'histogramme de l'acoustique log-transformée semble suivre une loi normale de façon assez précise, pour les campagnes anglaises. Le test de la droite d'Henry, ainsi qu'un test de Kolmogorov le confirment. Il est par contre assez difficile de tester la lognormalité spatiale de l'ensemble de la distribution de A , bien que les nuages de corrélation log-transformés décalés semblent assez elliptiques.

Si l'on suppose donc que l'acoustique est une variable lognormale, il est possible d'estimer une indicatrice par espérance conditionnelle.

On pose pour $A(x) > 0$ (les données acoustiques ne sont en pratique jamais strictement nulles) :

$$Z(x) = \log(A(x))$$

On a alors :

$$A(x) \geq c \Leftrightarrow Z(x) \geq \log(c)$$

par croissance monotone de la fonction logarithme. On peut alors estimer (Chilès et Delfiner 1999, [15]) :

$$\mathbb{P}(A(x) \geq c | Z(x_1) = z_1, Z(x_2) = z_2, Z(x_3) = z_3, \dots) = 1 - G \left[\frac{\log(c) - Z^{KS}(x)}{\sigma_{KS}(x)} \right]$$

où G est la fonction de répartition de la loi normale et Z^{KS} l'estimation par krigeage simple de Z et σ_{KS} son écart-type.

Comme on peut le voir sur la figure 2.32 les résultats obtenus par espérance conditionnelle et par modèle à résidus orthogonaux d'indicatrices sont semblables, malgré des différences locales pouvant aller de 0 à 40%.

2.7.3 Comment combiner une estimation non-linéaire de l'acoustique avec la variable de pêche ?

Les approches non-linéaires présentées dans les deux paragraphes précédents permettent d'obtenir une estimation plus robuste de l'acoustique, en particulier dans les régions, comme la mer du

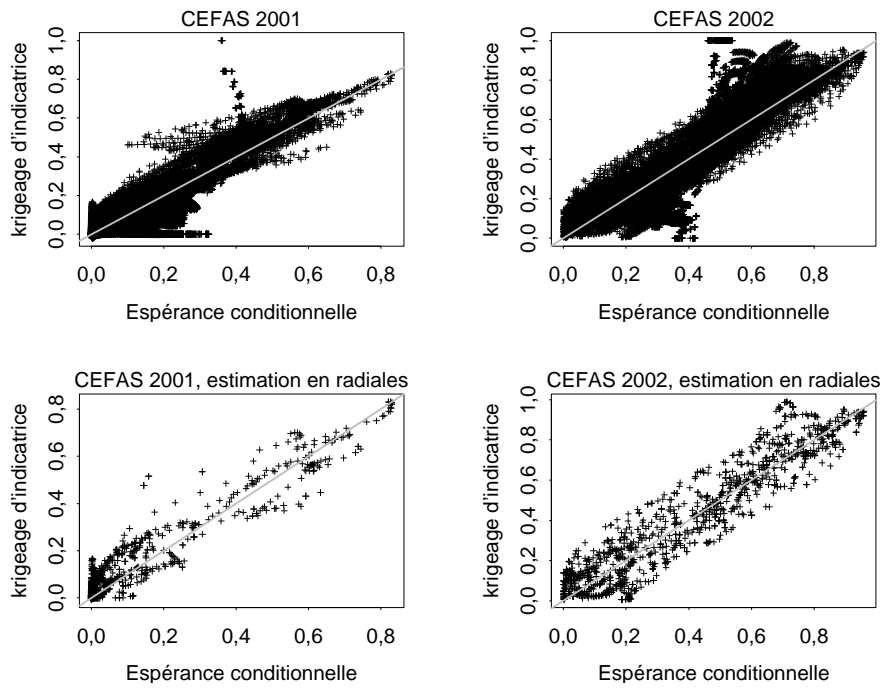


FIG. 2.32 – Différences entre l'estimation par modèle à résidu d'indicatrices et par espérance conditionnelle de l'acoustique log-transformée, CEFAS 2001 et CEFAS 2002, pour l'ensemble de la grille d'estimation (en haut) et uniquement pour les points en radiales (bas).

Nord, où la déstructuration des fortes valeurs est trop importante pour permettre d'inférer un modèle sur les données brutes. Cependant, l'objectif de cette étude est de combiner les données acoustiques avec les données de chalutage et non d'établir un modèle à partir des seules données acoustiques. Se pose alors la question : comment utiliser l'estimation non-linéaire des données acoustiques obtenue précédemment pour améliorer l'estimation des données de pêche ?

Souplesse du modèle à dérive externe

De ce point de vue, le modèle à dérive externe offre une certaine souplesse. Le rôle de la variable auxiliaire est d'imposer une contrainte de forme à l'estimation des captures. Ses valeurs peuvent donc provenir d'origines diverses. A l'extrême, le statisticien peut choisir d'estimer la variable cible avec *n'importe quelles* valeurs en dérive externe, puisque les structures croisées entre la variable cible et la variable auxiliaire sont ignorées. Il est alors, en quelque sorte, le seul arbitre de la légitimité de son choix.

L'idée est donc de supposer que les captures suivent en moyenne la forme provenant de l'interpolation des données acoustiques.

En pratique les résultats sont assez peu satisfaisants. Bien que la carte d'estimation utilisant l'acoustique en dérive externe soit moins lisse (figure 2.33), le coefficient multiplicatif de la dérive externe (voir partie 2.6) est souvent négatif, indiquant que les données de pêche contredisent au contraire souvent la carte acoustique, en particulier en Écosse et pour certaines campagnes anglaises. En voisinage unique, cependant, le coefficient devient généralement positif. Les incohérences locales entre les deux sources d'information tendent à s'effacer.

Autres approches possibles et discussion

Goovaerts (1997 [29]) propose plusieurs approches pour combiner une variable continue avec une variable découpée en indicatrices, ou plus généralement une variable catégorielle. Le krigeage intra-classe, par exemple, consisterait à utiliser les classes acoustiques pour délimiter des régions de l'espace et d'inférer un variogramme des captures au sein de chaque zone, en effectuant des krigeages simples des données de capture au sein de chaque zone, avec comme moyenne, la moyenne de la zone. Ces approches, mises en place pour des données géologiques, sont délicates à appliquer ici, en raison de la géométrie spatiale très discontinue des classes acoustiques. De plus, la variable auxiliaire ne sert dans ce modèle qu'à tracer des contours de zones, ses valeurs ne sont pas prises en compte.

Le variogramme croisé entre la variable de capture et une indicatrice de l'acoustique, divisé par le variogramme simple de cette même indicatrice permet de mesurer l'écart moyen entre deux valeurs de captures distantes de h de part et d'autre de la frontière définie par l'indicatrice. En Norvège, ces rapports de variogrammes sont uniquement structurés pour les classes moyennes. Pour les campagnes anglaises, un effet de bord est visible en 2000 (cf. figure 2.34). Pour les démersaux, l'écart moyen diminue avec la distance pour chacune des coupures alors que pour les pélagiques il semble au contraire augmenter. Mais en 2001 et 2002, les tendances sont moins nettes.

D'un autre côté, la variable de capture, elle aussi, suit une distribution très asymétrique souvent

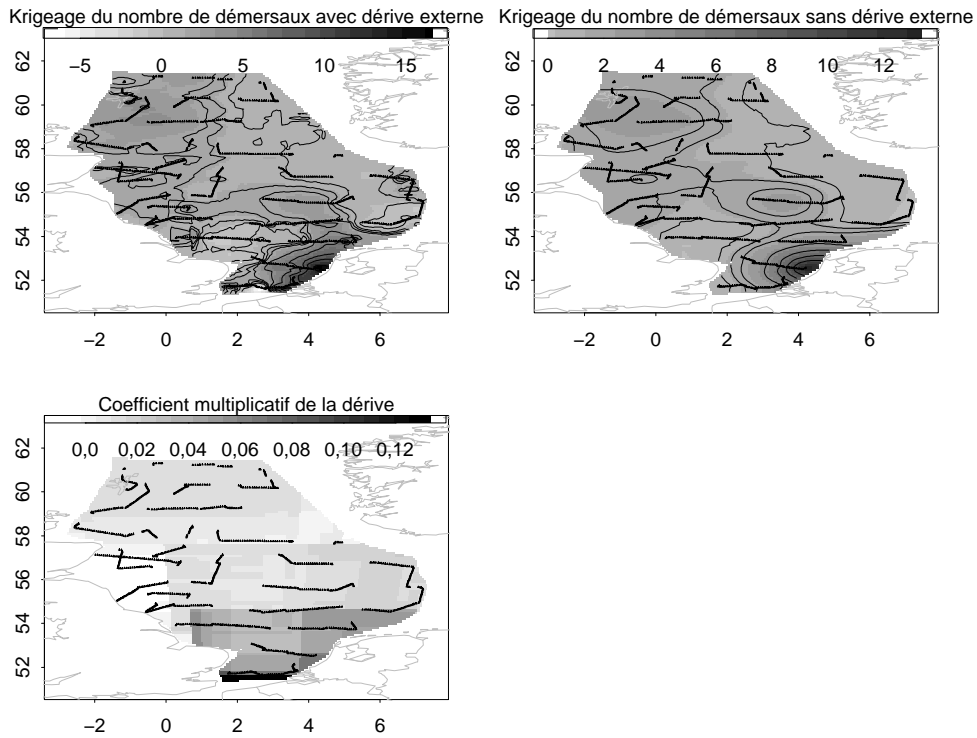


FIG. 2.33 – Comparaison du krigeage des données de captures (démersaux+pélagiques) avec l'estimation non-linéaire de l'acoustique en dérive externe (en haut à gauche) et sans dérive externe; par krigeage ordinaire (en haut à droite) et coefficient multiplicatif de la dérive, CEFAS 2002

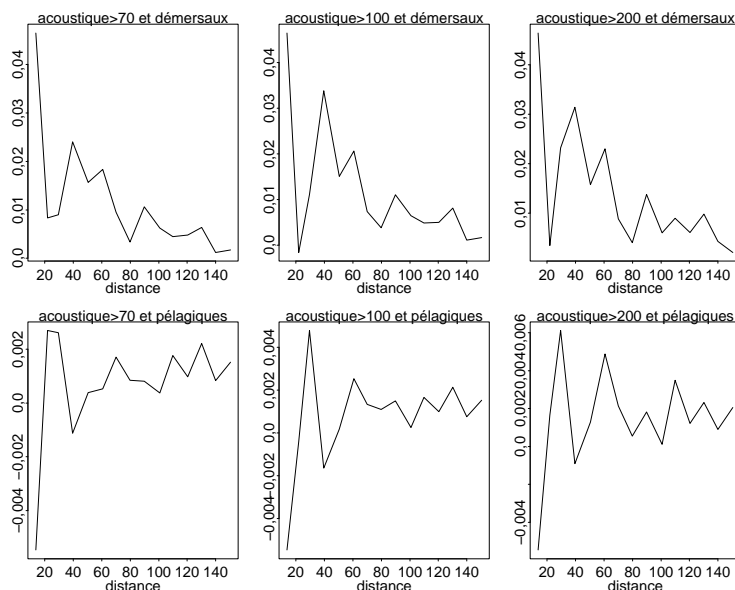


FIG. 2.34 – Rapport entre variogrammes croisés capture-indicatrices acoustiques sur le variogramme simple de l'indicatrice. La première ligne correspond aux démersaux et la deuxième aux pélagiques, CEFAS 2000

déstructurée par les valeurs fortes. La méthode de dérive externe, par exemple nécessite l'inférence d'un modèle sur le variogramme empirique des captures ou du résidu. Celui-ci a souvent une grande composante pépitique. Un modèle permettant une estimation non-linéaire cohérente à partir du découpage en classes de l'acoustique ainsi que des classes de captures serait sûrement la possibilité d'obtenir une estimation plus robuste encore et applicable sur l'intégralité des campagnes.

L'étude des tableaux de contingence croisées (voir partie d'analyse structurale 2.2) montre qu'une telle approche serait particulièrement prometteuse en Irlande, où la corrélation entre les classes de chacune des deux variables était bonne avec les pélagiques. En mer du Nord, par contre, la correspondance entre les zones riches d'après l'information acoustique et les zones riches d'après l'information de capture correspondent assez mal, indiquant qu'une méthode plus avancée dans le découpage en indicatrices est inutile, quelque soit le groupe d'espèces.

De plus, le nombre de stations est souvent assez faible, ce qui rend le découpage en classe assez délicat sur cette variable. Chaque classe ne comprendra que très peu de valeurs, en particulier en mer d'Irlande où une campagne ne contient qu'une trentaine de stations.

Chapitre 3

Étude de la verticale : Recherche d'une méthode d'intégration des données acoustiques

Transcrivez un modèle, on vous traite de plagiaire, mais si vous en copiez cent, vous voilà bientôt docteur.

Michel Serres, *Le Tiers-Instruit*

L'information acoustique est disponible par couche sur toute la hauteur d'eau (cf. figure 1.3). Or les modèles acoustique-capture vus dans le chapitre précédent deviennent vite complexes lorsque l'on multiplie le nombre de variables. Le but de cette partie est donc de simplifier le système multivariable pour que la modélisation puisse être abordée le mieux possible.

La première partie de ce chapitre place l'analyse dans son contexte. Les layers, d'autant plus nombreux que la mer est profonde, sont corrélés linéairement entre eux et ont des distributions très asymétriques. De plus le niveau de finesse d'estimation recherché rend une approche 3D laborieuse.

Ensuite deux types d'approches sont étudiées :

- soit réduire le nombre de variables en condensant toute l'information en une combinaison de layers (somme ou régression de layers)*
- soit orthogonaliser les variables acoustiques, le système se simplifiant alors par la disparition de liens croisés (MAF).*

Enfin, une approche plus explicative et plus proche des données visant à ne conserver du profil acoustique que deux variables : son intensité et sa forme, est présentée.

Une synthèse comparant les différentes méthodes entre elles est proposée en conclusion du chapitre.

The acoustic information available for the whole water column is divided into layers (cf. schema 1.3). Now, the combined acoustic and trawl models developed in the last chapter become complex when the variables are numerous. The aim of this chapter is then to simplify the multivariate system to make the modelling as easy as possible. The first part of the chapter introduces the problem : the layers, which are more numerous as the sea is deeper, are well correlated linearly, and the statistical distributions of the variables are highly skewed. Then, two kinds of methods are considered :

- *reducing the number of variables by summarizing the information into a combination of layers (sum or regression approach)*
- *building new factors, combination of layers (spatially) uncorrelated. The problem is the simplified because the cross structure between the variables disappears (Min/Max autocorrelations factors).*

Then, a more illustrative approach, closer to the data, is discussed. The idea is to summarize the acoustic profile (the acoustic distribution along the water column) into 2 variables only : the shape and the total intensity of the profile. A synthesis, comparing all the methods, is proposed as a conclusion of the chapter.

Notations

On note toujours la variable de pêche $P(x)$ et $A(x) = (A_1(x), \dots, A_k(x))$ l'ensemble des variables acoustiques (les variables auxiliaires) si l'on considère l'ensemble des couches disponibles. Ces différentes variables sont modélisées par des fonctions aléatoires indexées par l'espace. L'indice i est donc un indice de couches, la première étant la plus proche du fond et la dernière la plus proche de la surface (cf. figure 1.3). Le nombre de couches k n'est pas fixe, il dépend de la profondeur. Pour des notations plus rigoureuses, il conviendrait donc de le noter k_x . Cependant, le chalut raclant le fond de la mer, les couches acoustiques les plus proches de la surface ne seront en pratique jamais utilisées. On a donc considéré, pour plus de simplicité dans les notations, k fixe.

3.1 Contexte

Le modèle de cokrigage tel qu'il a été présenté dans le chapitre précédent dans le cas bivariable se complexifie beaucoup lorsque l'on multiplie les variables auxiliaires.

L'estimation de la variable principale au point x_0 s'écrit :

$$P(x_0) = \sum_{x \in R} \sum_{i \in \text{layers}} \lambda_{ix}^A A_i(x) + \sum_{x \in S} \lambda_x^P P(x)$$

Le système de cokrigage s'écrit donc ($C(h)$ désigne la fonction de covariance) :

$$\begin{bmatrix} C_A & C_{AP} & 1 & 0 \\ C_{PA} & C_P & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \lambda^A \\ \lambda^P \\ \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} C_{AP}(x_0) \\ C_P(x_0) \\ 1 \\ 0 \end{bmatrix} \quad (3.1)$$

avec la sous matrice :

$$\begin{bmatrix} C_A & C_{AP} \\ C_{PA} & C_P \end{bmatrix} = \begin{bmatrix} C_{A_1} & C_{A_1 A_2} & C_{A_1 A_3} & \cdots & C_{A_1 A_k} & C_{A_1 P} \\ C_{A_2 A_1} & C_{A_2} & C_{A_2 A_3} & \cdots & C_{A_2 A_k} & C_{A_2 P} \\ C_{A_3 A_1} & C_{A_3 A_2} & C_{A_3} & \cdots & C_{A_3 A_k} & C_{A_3 P} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{A_k A_1} & C_{A_k A_2} & C_{A_k A_3} & \cdots & C_{A_k} & C_{A_k P} \\ C_{PA_1} & C_{PA_2} & C_{PA_3} & \cdots & C_{PA_k} & C_P \end{bmatrix}$$

Le système devient vite laborieux lorsque la mer est profonde et que le nombre de couches acoustiques augmente. Il est alors utile, voire nécessaire, de simplifier le système en diminuant le nombre de variables ou en les orthogonalisant pour supprimer les structures croisées. De plus, les incertitudes qui entourent l'estimation des covariances croisées rend le cokrigage d'un grand nombre de variables assez incertain en pratique (Matheron, 1979 [43]).

3.1.1 Liens entre layers

Les layers acoustiques sont très corrélés linéairement entre eux, quelque soit la campagne étudiée (cf. figure 3.1) pour les données en radiales après régularisation. De façon générale les 10 premiers layers de 1 m de haut ont des coefficients de corrélations supérieurs à 0.9 et les nuages de corrélations confirment un lien linéaire très fort et très proche de la première bissectrice. Par contre, même s'il subsiste, ce lien est nettement moins marqué pour layers suivants de 10 m de haut.

La réduction du système est donc aussi justifiée par la nécessité de limiter la redondance d'information d'un layer à l'autre. La colinéarité peut, par ailleurs, rendre instable le calcul numérique de certaines méthodes linéaires comme la régression (cf. partie 3.4).

Les variogrammes croisés montrent que les différentes couches sont également corrélées spatialement (fig. 3.3). La présence de corrélation spatiale entre les layers est d'une certaine manière assez rassurante. Elle montre que l'acoustique a une régularité spatiale importante, même si celle-ci est de courte portée. Enfin, on remarque que les 10 premiers layers (de 1 mètre de haut) se rapprochent d'un modèle linéaire de corégionalisation (figures 3.2 et 3.3).

Le modèle linéaire de corégionalisation à deux composantes est un modèle géostatistique simple dans lequel chaque variable se décompose en une combinaison linéaire de deux composantes spatialement non corrélées l'une à l'autre. Les composantes utilisées sont normées et centrées.

$$A_i(x) = b_{1,i}Y_1(x) + b_{2,i}Y_2(x) + Cte \text{ avec } cov(Y_1(x), Y_2(x+h)) = 0 \forall x \forall h$$

où chacune des composantes représente une structure. Ici par exemple on choisit Y_1 pépitique, $\gamma_{Y_1} = \text{Pépité}(h)$ et Y_2 représentant la structure moyenne portée (environ 200 m.n), avec un variogramme par exemple sphérique : $\gamma_{Y_2} = \text{Sphérique}(h/200)$. Toutes les structures simples et croisées des variables initiales sont alors des combinaisons linéaires des deux structures de base : γ_{Y_1} et γ_{Y_2} . Le modèle de cokrigage des variables se simplifie alors nettement. Les deux composantes Y_1 et Y_2 étant non corrélées, elles sont autokrigeables.

Cependant, seuls les 10 premiers layers correspondent à un modèle linéaire de corégionalisation ; les couches supérieures sont beaucoup plus fluctuantes. Cette approche n'a donc pas été poursuivie.

3.1.2 Modélisation 3D

Passer d'un modèle multivariable à un modèle 3D en ajoutant une coordonnée de profondeur peut être une solution efficace pour conserver toute l'information disponible en prenant compte de la corrélation entre les points. Cependant la modélisation 3D est extrêmement lourde pour une taille de champ comme celle des zones étudiées ici ramenée à l'échelle d'estimation voulue (le mile nautique en longitude et latitude et le mètre en profondeur). Les campagnes en mer de Barents couvrent des zones supérieures à 520×320 m.n sur 300 m de profondeur, ce qui fait près de 50

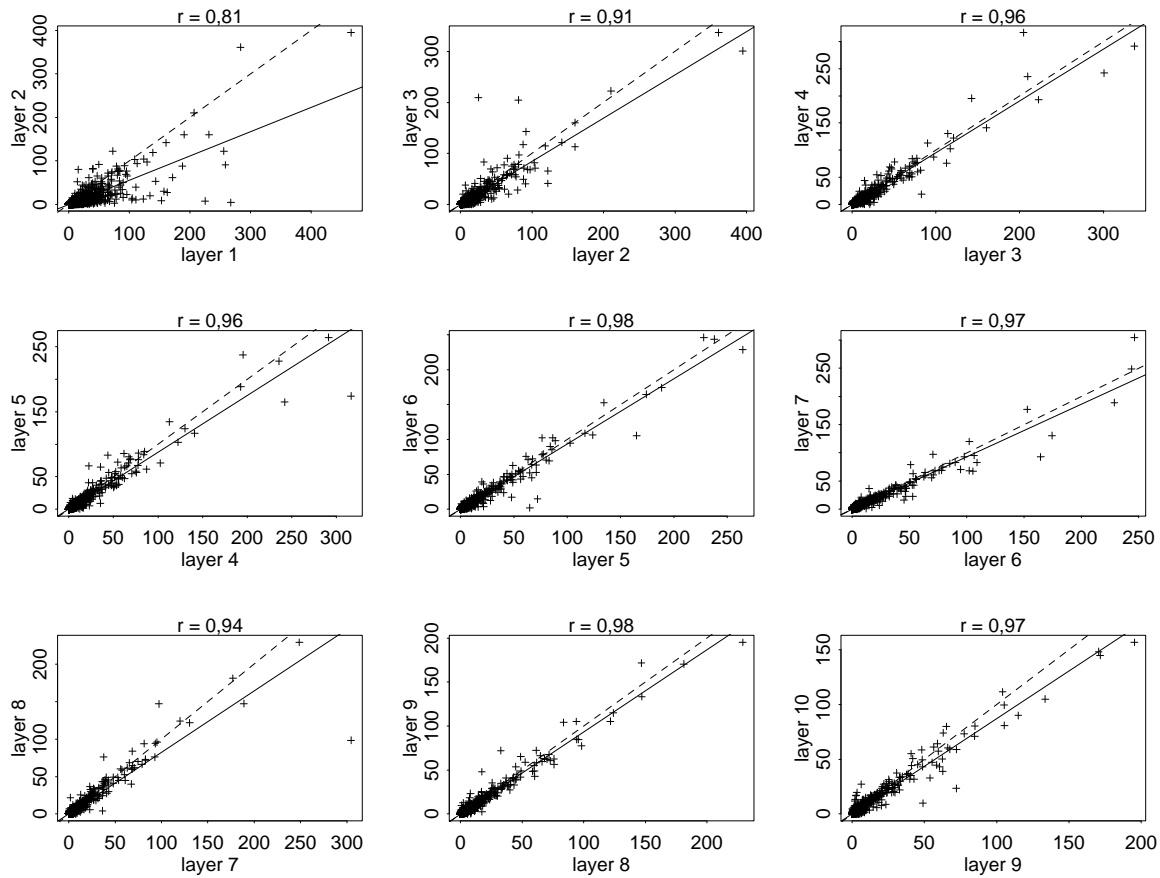


FIG. 3.1 – Nuages de corrélation d'un layer au suivant, pour les 10 premiers mètres au dessus du fond, Norvège 2001. La ligne pointillée représente la première bissectrice et la ligne pleine la droite de régression du nuage de point.

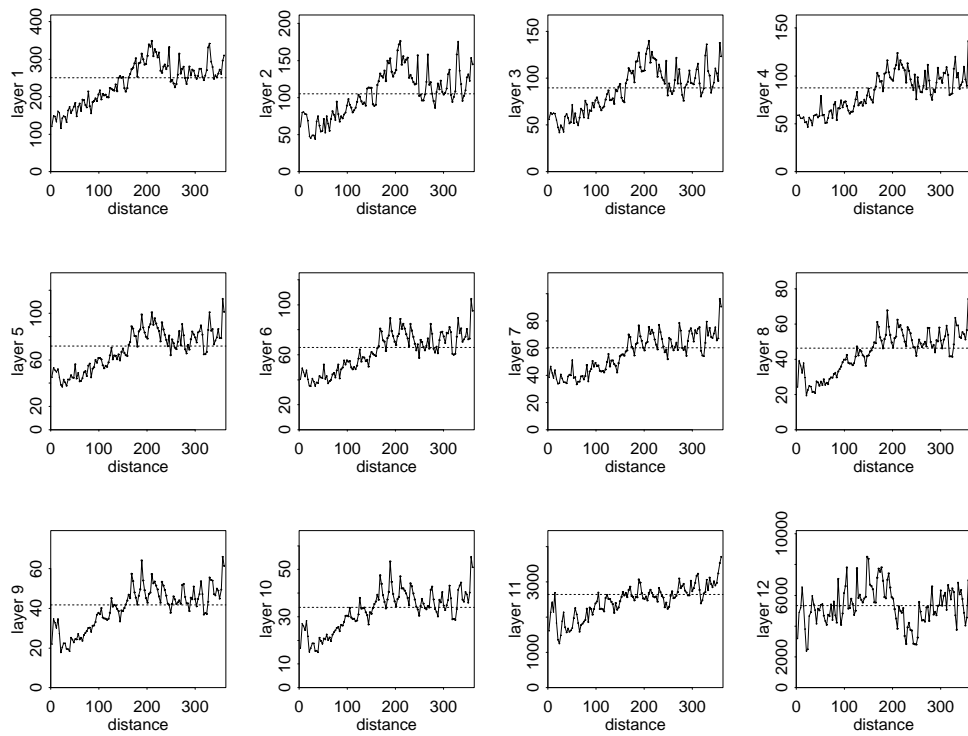


FIG. 3.2 – Variogrammes des layers, Norvège 2001.

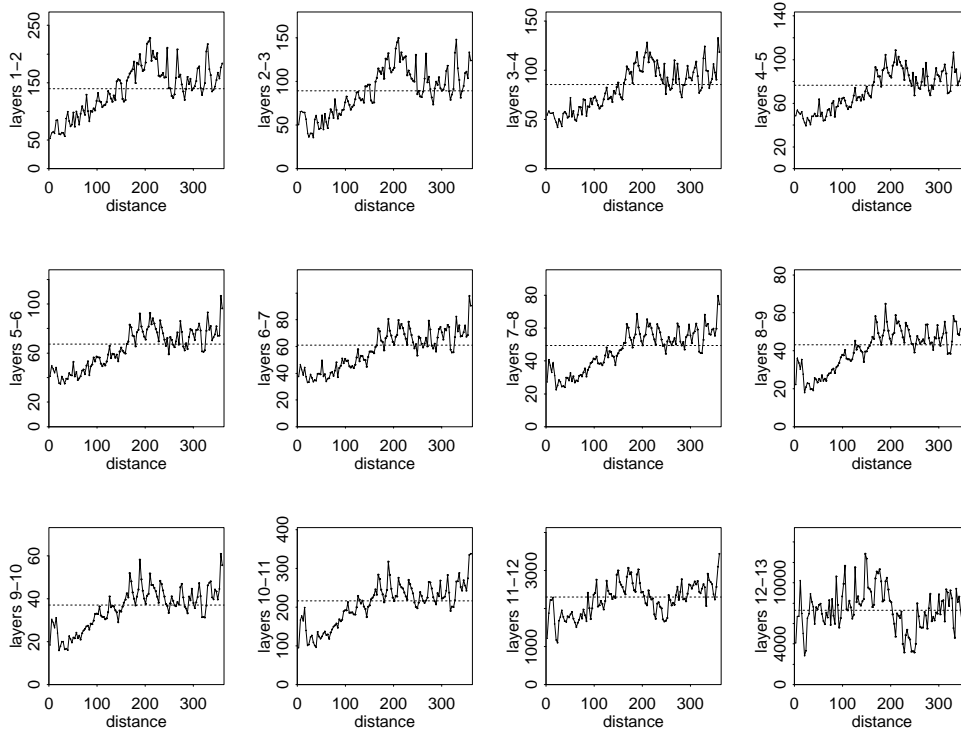


FIG. 3.3 – Variogrammes croisés d'un layer au suivant, Norvège 2001.

millions de noeuds de grille.

De plus, le variogramme vertical des données acoustiques montre une assez forte dérive (fig. 3.4 pour la Norvège 1997). Le fait de ne pouvoir supposer la stationnarité verticale complexifie le modèle.

Le peu d'effet de pépité des variogrammes verticaux montrent une fois de plus la corrélation importante entre les layers, en particulier pour les layers de 1 m dont le variogramme montre une faible pente à l'origine, ce qui indique une grande continuité spatiale des réalisations de la variable acoustique (à l'échelle du mètre).

Enfin, l'objectif étant de croiser l'acoustique avec des variables de captures sans véritable dimension verticale, la modélisation 3D semble trop lourde en hypothèses pour être ici véritablement pertinente. Elle pourrait par contre être intéressante pour l'élaboration d'un modèle purement acoustique. Gerlotto et al. (1999, [23]) et Simard et al. (2003 [65]) montrent des exemples de modèles géostatistiques 3D pour des estimations de champs acoustiques.

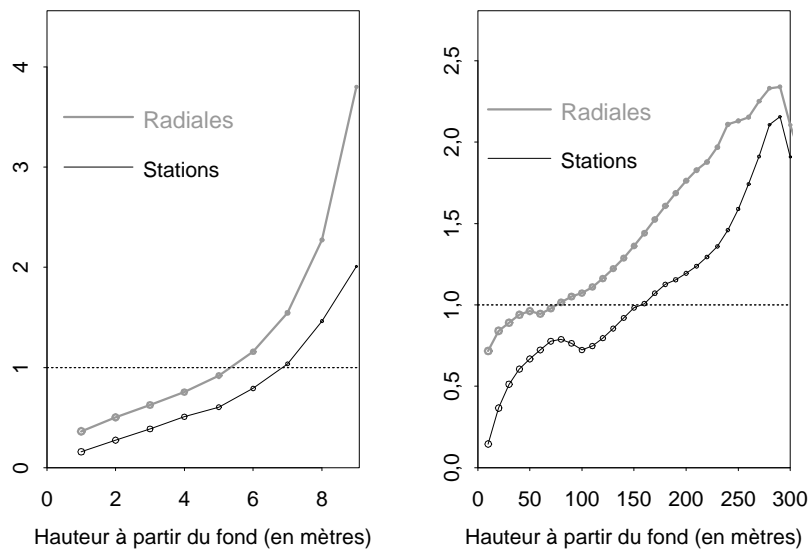


FIG. 3.4 – Variogrammes verticaux normés des données acoustiques en Norvège 1997 pour les 10 premières couches de 1 m (gauche) et pour les couches suivantes de 10 m de haut (droite).

3.2 Sommes de layers

Le chalut ne fait que 3 ou 4 mètres de haut (cf. partie 1.1.3). Cependant, il est connu des pêcheurs et des biologistes que de nombreuses espèces plongent vers le fond de la mer à l'approche du bateau ou de l'engin de pêche, ou des deux (Handegard et al. 2003 [33], Michalsen et al. 1999 [45], Godøet al. 1999 [24]). Le support vertical réel du coup de chalut n'est donc pas connu.

A priori différente pour chaque espèce, cette grandeur est intéressante pour la connaissance du phénomène de pêche. L'acoustique, justement disponible sur toute la hauteur d'eau permet de la quantifier en étudiant la hauteur d'eau permettant une corrélation acoustique-capture maximale.

Les estimations d'abondance obtenues par chalutage, peuvent ainsi être associées non seulement à une surface équivalente (la surface chalutée) mais aussi à un volume, bien que le support horizontal soit en fait incertain, bruité par les éventuels déplacements latéraux de poissons.

3.2.1 Coefficient de corrélation de Pearson

Le coefficient de corrélation (dit de Pearson) mesure le lien linéaire entre deux variables (cf. Saporta 1990 [61]).

$$\rho(Z_1, Z_2) = \frac{\text{cov}(Z_1, Z_2)}{\sigma_{Z_1} \sigma_{Z_2}}$$

où σ_{Z_1} et σ_{Z_2} sont les écarts-types de Z_1 et Z_2 .

La hauteur d'eau optimale recherchée doit alors vérifier :

$$\max_n \rho(P(x), \sum_{i=1}^n A_i(x))$$

La figure 3.5 montre l'évolution du coefficient de corrélation entre le nombre de démersaux pêchés et l'acoustique au fur et à mesure que l'on somme les couches à partir du fond, en mer de Barents entre 1997 et 2002. La hauteur optimale varie selon les années de 1 m à 200 m au dessus du fond, en raison de la grande sensibilité du coefficient de corrélation aux valeurs extrêmes, les données étant très asymétriques.

3.2.2 Coefficient de corrélation de Spearman

Pour diminuer l'impact des valeurs fortes, la même relation a été établie avec le coefficient de corrélation de rang. Initialement introduits pour des variables non numériques dont on ne dispose que de l'ordre, les coefficients de corrélation de rang ont l'avantage de ne pas dépendre de la distribution des données, ils mesurent les concordances entre les classements des variables. Ici, on associe à chaque station un rang dans la distribution des données acoustiques ainsi qu'un rang dans la distribution des données de captures. Les deux suites formées des rangs de chacune des stations sont ensuite comparées par la relation suivante :

$$\rho_{rang}(Z_1, Z_2) = \frac{\text{cov}(r_1, r_2)}{\sigma_{r_1} \sigma_{r_2}}$$

où r_1 et r_2 sont les rangs de chacun des objet selon la variable Z_1 ou la variable Z_2 , σ_{r_1} et σ_{r_2} sont leur écarts-types.

Si $\rho_{rang}(x, y) = 1$ les deux classements sont identiques, s'il est égal à -1, ils sont inversés et s'il est égal à 0, ils sont orthogonaux. Pour les autres valeurs, un test de concordance selon la taille de l'échantillon peut être calculé à partir de la table de Spearman (Saporta 1990 [61]). Ce coefficient a l'avantage d'être invariant pour toute transformation monotone croissante des variables et est

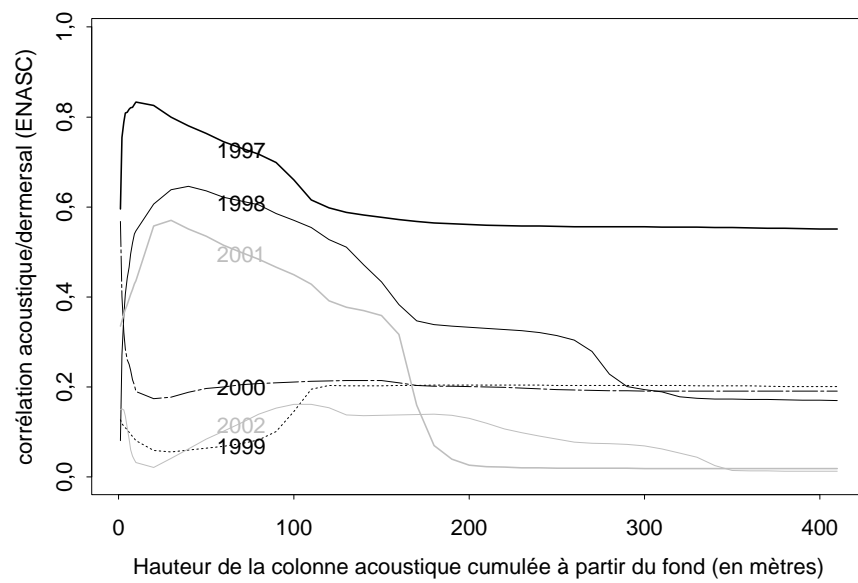


FIG. 3.5 – Évolution de la corrélation entre les démersaux (en ENASC) et la somme cumulée de layers acoustiques pour les stations norvégiennes entre 1997 et 2002.

insensible aux valeurs extrêmes.

Comme précédemment, le problème s'écrit donc :

$$\max_n \rho_{rang}(P(x), \sum_{i=1}^n A_i(x))$$

Comme on peut le voir sur la figure 3.6, la hauteur d'eau optimale obtenue est alors très stable pour toutes les années et se situe autour de 40 m au-dessus du fond pour les démersaux en Norvège. Hjellvik (2003 [34]) observe des hauteurs comparables pour la morue en travaillant sur les variables log-transformées et en distinguant plusieurs heures du jour. Le coefficient de Spearman permet ici un gain important en robustesse.

En mer d'Irlande, ce sont avec les espèces pélagiques que l'acoustique est la mieux corrélée. La hauteur d'eau optimale se situe autour de 4 ou 5 m au dessus du fond. En mer du Nord, les corrélations sont mauvaises et les courbes d'évolution sont très plates, de sorte qu'une hauteur optimale est difficile à déterminer. J'ai alors choisi de sommer les 5 premières couches, correspondant à la hauteur d'ouverture du filet.

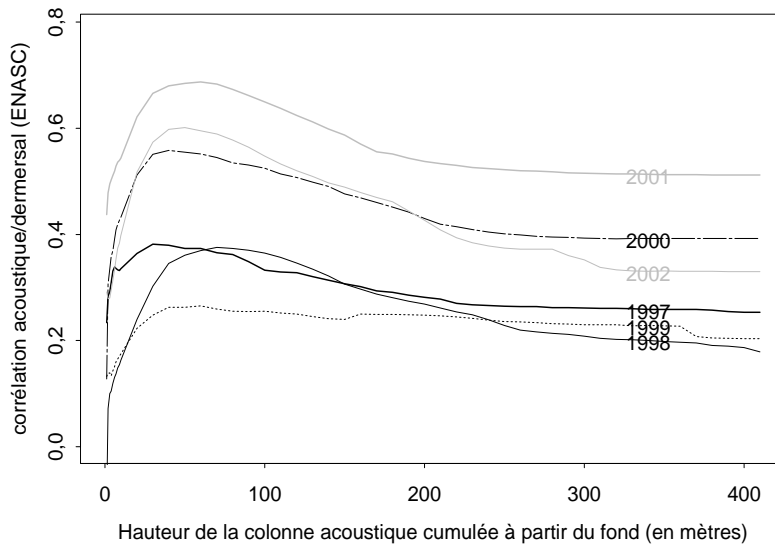


FIG. 3.6 – Évolution de la corrélation de rang (Spearman) entre les démersaux (en ENASC) et la somme cumulée de layers acoustiques pour les stations norvégiennes entre 1997 et 2002.

Il serait possible que les 40 mètres qu'il faille intégrer permettent en fait de compenser la petite largeur du cône acoustique par rapport à la largeur du chalut. Cependant, en Norvège, pour une profondeur moyenne de 300 m, le cône acoustique fait environ 36 m de diamètre à sa base tandis que les panneaux du chalut sont écartés de 40 à 50 mètres environ. Les supports sont donc de tailles comparables et c'est ici vraiment le mouvement vertical des poissons qui semble être quantifié.

Par contre, en mer du Nord, l'écart entre les supports horizontaux peut expliquer qu'aucune corrélation systématique n'ait pu être dégagée (pour une profondeur d'environ 70 m, le diamètre de base du cône acoustique est de 8.5 m alors qu'en mer du Nord, la distance entre les panneaux dépasse

généralement les 60 m, cf. partie 1.1.3).

En plus d'être (par construction) corrélée aux captures, la somme de layers ainsi obtenue a l'avantage d'être structurée dans l'espace (cf. figure 3.7). La somme est une fonction assez régularisante, souvent même mieux structurée que les variables initiales.

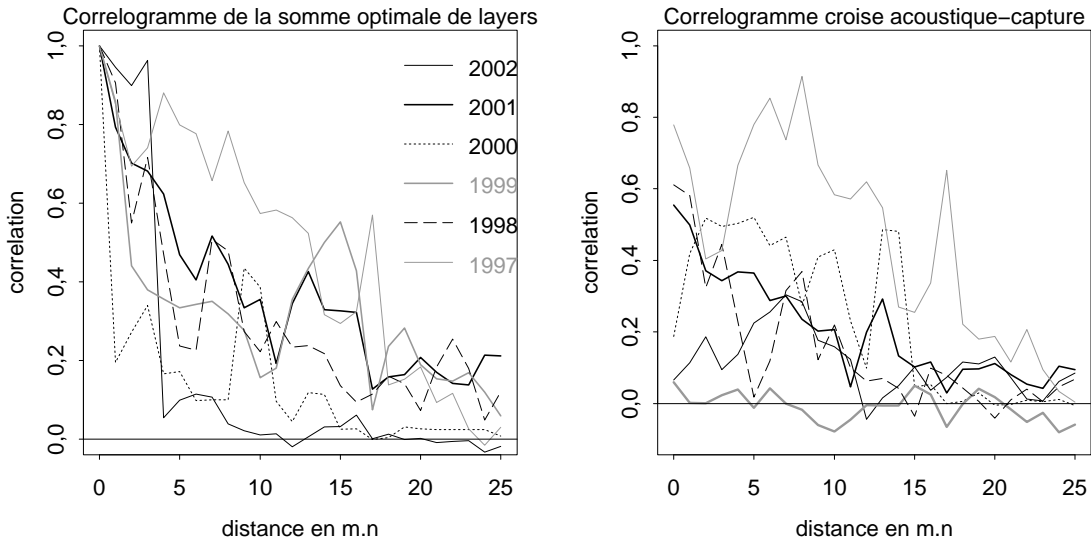


FIG. 3.7 – Correlogramme simple de la somme des 20 premiers layers acoustiques (gauche) et correlogramme croisé avec les démersaux en ENASC (droite), en Norvège 1997 à 2002. Le correlogramme représente l'évolution du coefficient de corrélation entre deux variables éloignées de la distance h . Ici ils ont été calculés en 1D, c'est à dire le long de la route du navire.

3.3 Estimation d'un support vertical local

Il peut également être intéressant d'étudier en chaque station le nombre de layers qu'il faut sommer pour retrouver une grandeur comparable à celle observée dans le chalut au même point. Cette hauteur n'a de sens que si la capture est exprimée en ENASC, c'est à dire dans la même unité que l'acoustique. De plus elle suppose que les poissons pêchés sont proches du fond, ou plus précisément que l'abondance des couches acoustiques est de moins en moins susceptible d'être capturée au fur et à mesure que l'on monte dans la colonne d'eau. Cette hypothèse est fautive pour les espèces pélagiques dont l'abondance est généralement décollée du fond mais elle semble réaliste pour les espèces démersales qui sont de loin les espèces les plus présentes dans le chalut de fond. On la suppose vraie également pour l'ensemble de l'abondance capturable (pélagiques + démersaux + poissons plats). En mer du Nord, où la largeur du cône acoustique est extrêmement petite par rapport à celle du chalut, la signification d'une telle hauteur est discutable. Par contre, en mer de Barents où les largeurs sont comparables, cette hauteur représente la hauteur d'efficacité locale du chalut, lorsque l'on néglige les déplacements latéraux de poissons pouvant intervenir entre le passage

du sondeur acoustique et celui du chalut (cf. partie 1.1.3).

Cette hauteur s'obtient par : $H(x) \in \{0, \dots, k - 1\}$ telle que

$$\sum_{i=1}^{H(x)} A_i(x) \leq P(x) < \sum_{i=1}^{H(x)+1} A_i(x)$$

Puisque l'acoustique enregistre indistinctement toutes les espèces, $P(x)$ représente ici l'ensemble des captures en ENASC, toutes espèces confondues. Mais une fois de plus, l'abondance pêchée par le chalut de fond en mer de Barents est à plus de 97% démersale.

Les stations pour lesquelles la valeur de capture reste supérieure à la somme de l'intégralité de la hauteur d'eau ont été mises de côté. Elles correspondent aux cas où la capture en un certain point est supérieure aux enregistrements du sondeur (erreur de type 2, cf. partie 1.1.4). Il peut s'agir de stations pour lesquelles un déplacement latéral important de poissons a eu lieu, ou de stations riches en poissons plats, collés au fond et quasiment invisible à l'acoustique. Il est difficile de dégager des facteurs explicatifs de l'évitement du cône acoustique par les espèces capturées. Les stations pour lesquelles $P(x)$ est supérieure à l'acoustique intégrée sur la totalité de la colonne d'eau ne semblent pas spécifiques à un groupe d'espèces particulier. Leur apparition ne semble pas non plus être liée à la température, la salinité ou l'heure du jour ; ni favorisée par la profondeur, de grandes distances entre les portes du chalut, la hauteur du filet ou la longueur des câbles.

Ces stations ne sont heureusement pas les plus fréquentes mais représentent tout de même une part non négligeable des données (entre 10% et 30% des stations selon les années).

Pour les autres stations, il est donc possible de trouver un nombre de couches telles que l'acoustique cumulée sur toute cette hauteur soit très proche de la valeur de capture observée dans le chalut. Les valeurs les plus fréquentes se situent autour de 13 et 14, ce qui correspond à 40 ou 50 m au dessus du fond, c'est à dire la hauteur d'eau optimale obtenue dans la partie précédente (cf. partie 3.2). Les espèces démersales étant très largement dominantes (en ENASC) dans les captures, il est normal que l'on retrouve en moyenne la somme de layers qui maximise la corrélation avec les captures démersales.

Les histogrammes de cette hauteur équivalente montrent que la plupart des valeurs se situent entre couches 11 à 18 (20 et 90 mètres au dessus du fond).

De plus, cette hauteur ne semble pas corrélée avec l'intensité des captures. Cette observation est rassurante. Si les très grosses captures correspondaient systématiquement à des hauteurs équivalentes d'acoustique très grandes, l'approche n'aurait pas de sens. Les grosses valeurs de pêche proviennent plus probablement d'un banc concentré de poissons qui se fait capturer que d'une hauteur d'efficacité du chalut supérieure à l'ordinaire. Une telle corrélation refléterait donc des déplacements latéraux de poissons trop importants pour être négligés. Les captures les plus fortes correspondent au contraire à des hauteurs équivalentes d'acoustique autour de 10 mètres en moyenne.

En revanche, cette hauteur locale d'efficacité du chalut est corrélée avec la profondeur. Les coefficients de corrélation varient entre 0.24 et 0.53 selon les campagnes. Le lien est assez bruité mais la tendance est nette, surtout en 1997, 1999, 2000 et 2002 (cf. figure 3.8). Plus la mer est profonde plus la hauteur d'efficacité correspondante a tendance à être grande. A priori lorsque la mer est profonde, le bruit du navire perturbe moins les poissons, mais par contre l'effet du chalut est le même. Comme le chalut est plus loin derrière le bateau, il est possible que les poissons aient plus de temps pour réagir (jusqu'à 15 minutes) et que des poissons placés plus haut dans la colonne

d'eau se retrouvent ainsi capturés. L'abondance est peut-être aussi plus décollée du fond lorsque la mer est profonde. Mais le chalut étant plus loin, il est également possible que les mouvements latéraux soient de plus en plus importants et que l'effet observé ne soit qu'un effet compensatoire. On somme plus de couches pour obtenir une abondance qui viendrait d'ailleurs. Enfin, les espèces et leurs regroupements ne sont souvent pas les mêmes pour les petites et grandes profondeurs. Les espèces pourraient donc être plus réactives par grand fond.

Enfin, excepté en 1997, cette hauteur équivalente est structurée dans l'espace, ce qui permet de la cartographier. L'effet de pépite est important, il varie de 30 à 50%, suivi d'une structure d'une portée de 100 m.n en 1998 et 1999 à 200 m.n en 2001 et 2002 (cf. figure 3.9). La structure est linéaire en 2000. Ces longues portées montrent que des géométries assez grandes sont présentes (cf. figure 3.10). Bien que liées à la profondeur, ces cartes ont des motifs assez différents de la carte de la bathymétrie. Les reliefs sont assez réguliers, en particulier en 1998 ou 1999 où des zones de hautes efficacités se dégagent nettement. Il serait intéressant d'approfondir cette étude pour comprendre les raisons de la présence de telles zones. Les variables environnementales et les variables de captures disponibles dans les bases de données n'ont pas permis de dégager de facteurs explicatifs clairs.

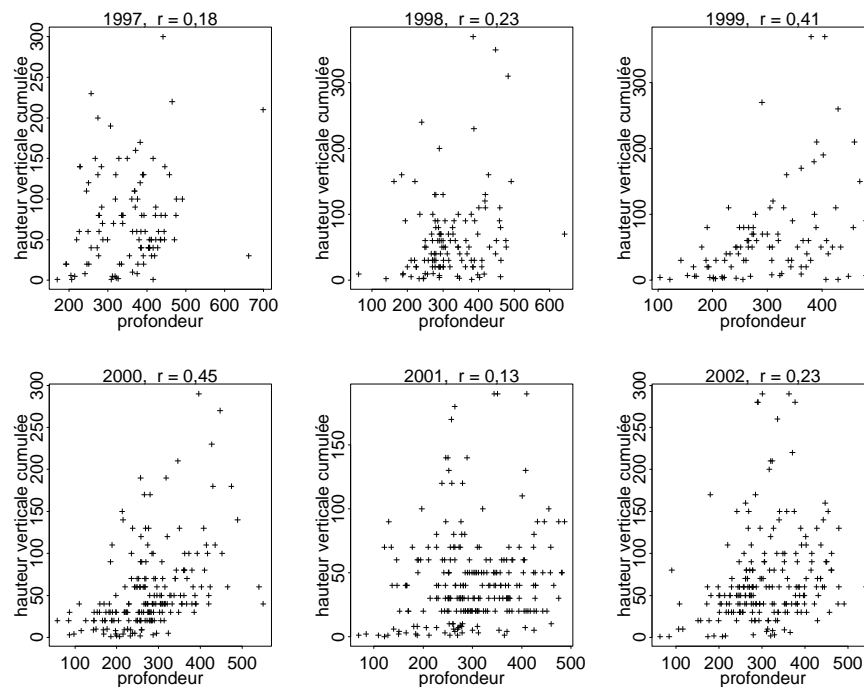


FIG. 3.8 – Hauteur locale de l'acoustique permettant d'obtenir des valeurs équivalentes aux captures en fonction de la profondeur, en mer de Barents.

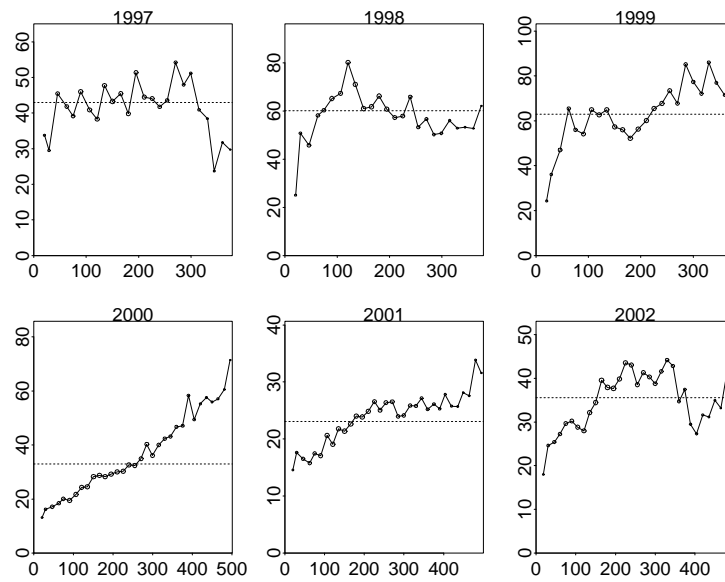


FIG. 3.9 – Variogrammes expérimentaux du support vertical local.

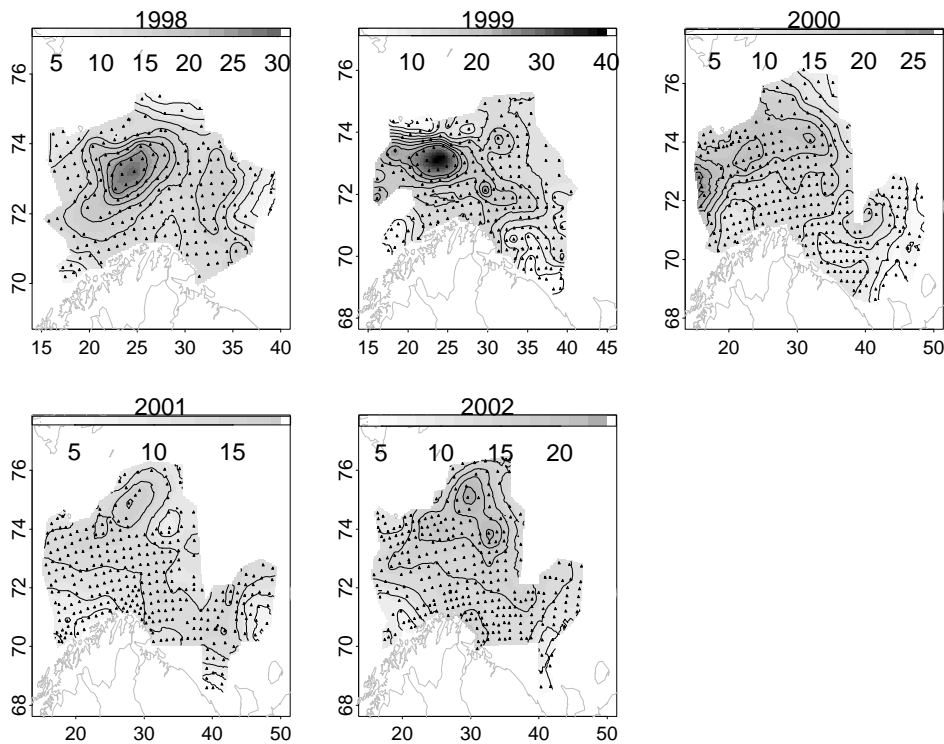


FIG. 3.10 – Cartes de krigeage des hauteurs équivalentes de l'acoustique entre 1998 et 2002 (en 1997, le variogramme a été considéré comme pépitique). L'échelle de gris est la même pour toutes les cartes.

3.4 Régression de layers

Plutôt que de s'imposer une somme de layers, une approche plus générale consiste à chercher une combinaison linéaire de couches acoustiques, en optimisant toujours la corrélation avec les captures par un critère de moindres carrés.

Le problème s'écrit alors : trouver $\alpha \in \mathfrak{R}^k$, tel que $\|P(x) - \beta - {}^t\alpha A(x)\|^2$ soit minimale.

Pour cela plusieurs régressions linéaires ont été effectuées. La principale difficulté vient de la très forte colinéarité entre les 10 premières couches (fig. 3.1). Les coefficients de la régression linéaire classiques sont alors mal identifiables parce que la matrice des variables explicatives n'est pas de rang maximal, plusieurs de ses valeurs propres sont quasi-nulles et si l'inverse de la matrice tAA existe, son calcul est très instable et la variance des paramètres très grande. Une possibilité est alors d'effectuer un changement de variables. En mer de Barents, la profondeur étant importante, la solution la plus simple consiste à sommer les 10 premières couches et ainsi de n'effectuer la régression que sur des couches de 10 m de haut, beaucoup moins corrélées entre elles. Cette approche est délicate en mer du Nord et en mer d'Irlande, parce que les profondeurs sont plus faibles et que la plus grande part de l'abondance se situe dans les tous premiers mètres. Pour cela, des régressions sur des accroissements de layers ont également été testées, ceux-ci semblant linéairement orthogonaux à la vue des nuages de corrélation.

Le problème s'écrit alors :

Trouver les $\tilde{\alpha}_i$ tels que $\|P(x) - \tilde{\beta} - \sum_{i=1}^{k-1} \tilde{\alpha}_i (A_{i+1}(x) - A_i(x))\|^2$ soit minimale.

Il est alors possible de revenir aux variables initiales en effectuant le changement de variables dans l'autre sens. La correspondance entre les coefficients obtenus sur les accroissements et ceux sur les variables brutes est simple :

$$\begin{cases} \beta = \tilde{\beta} \\ \alpha_1 = -\tilde{\alpha}_1 \\ \dots \\ \alpha_i = \tilde{\alpha}_{i-1} - \tilde{\alpha}_i \\ \dots \end{cases}$$

avec les α_i les coefficients des layers et les $\tilde{\alpha}_i$ les coefficients sur les accroissements associés.

Ces régressions, bien qu'assurant des coefficients de corrélation acoustique-capture élevés, sont peu robustes car elles sont très sensibles aux valeurs fortes. Les poids obtenus varient beaucoup d'une couche à la suivante et d'une année à l'autre. Les résultats sont donc difficilement interprétables (cf. fig. 3.11 et fig. 3.12).

On voit en effet que les coefficients changent quasiment systématiquement de signe d'un layer au suivant, peut-être par un effet de compensation à cause de la corrélation entre les couches (pourtant plus faibles sur les layers de 10 m de haut). On peut voir également en Norvège que les régressions sur les layers de 10 m de haut fournissent des coefficients semblables à ceux obtenus par les régressions sur des accroissements de layers, mis à part pour les 10 premiers mètres (fig. 3.12). Le poids de la première couche allant de 1 à 10 m au dessus du fond est souvent estimé très proche de zéro par la régression sur couches de 10 m ; alors que la méthode sur accroissements de layers fournit des coefficients très grands et très variables. En mer du Nord, les coefficients des 10 premiers mètres sont également très instables tandis que les couches supérieures ont des poids quasi nuls

(cf. fig. 3.11). Les régressions effectuées sur les pélagiques, espèces qui génèrent souvent quelques rares bancs extrêmement denses, semblent être encore plus instables que celles obtenues pour les démersaux.

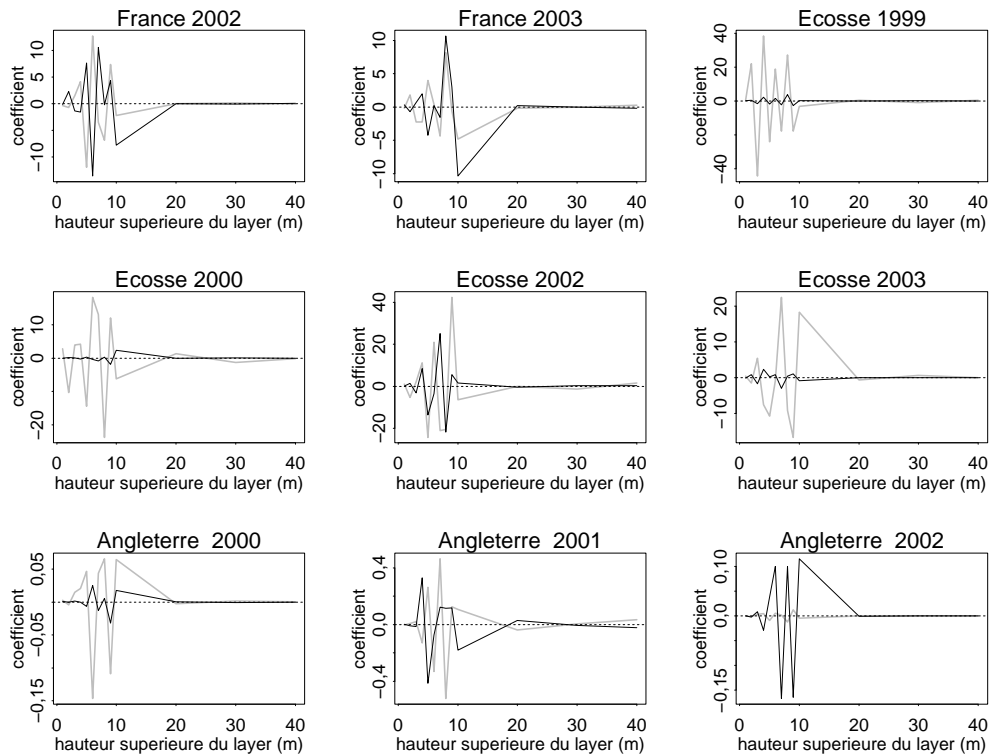


FIG. 3.11 – Composition des combinaisons linéaires de layers obtenues par régression sur les démersaux en ENASC (en noir) et les pélagiques en ENASC (en gris), pour les campagnes en mer du Nord.

Aussi bien en mer de Barents qu'en mer du Nord, les combinaisons linéaires obtenues par régression sont très peu structurées, les corrélogrammes des combinaisons linéaires obtenues sont quasi-pépitiques, excepté certaines années comme en 1997 en Norvège. De même les structures croisées acoustique-capture semblent moins bien structurées que les structures croisées avec la somme de layers obtenue précédemment. Il semble que la régression linéaire, en optimisant la corrélation acoustique-capture en stations est un effet destructurant du point de vue spatial (cf. figure 3.13).

Il peut en effet arriver qu'une valeur forte enregistrée à l'acoustique à un layer particulier corresponde justement à une capture importante. Le poids de ce layer va être alors prépondérant dans la régression. Réciproquement un autre layer peut se retrouver doté d'un coefficient fortement négatif en raison de valeurs fortes décalées par rapport aux captures. La combinaison linéaire obtenue sera alors totalement dénuée de sens.

De plus, même en dehors de toute configuration asymétrique, la régression, maximisant un critère

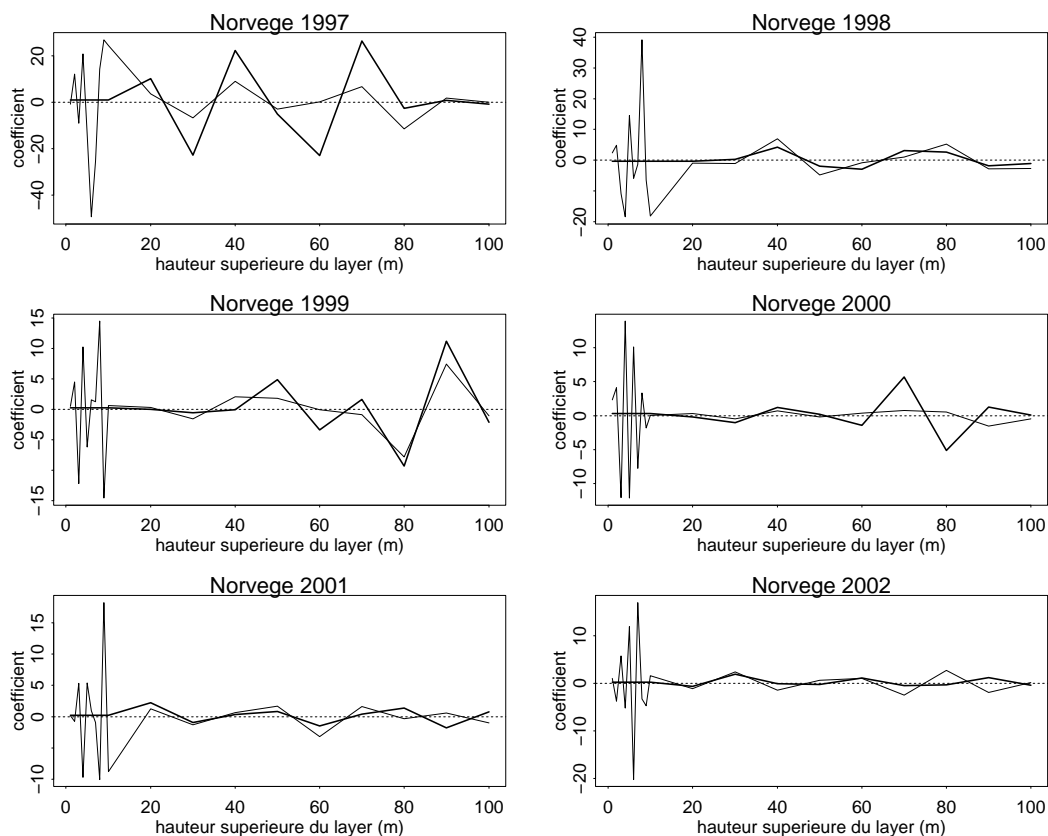


FIG. 3.12 – Composition des combinaisons linéaires de layers obtenues par régressions sur les démersaux en ENASC à partir d'accroissements de layers (courbe fine) ou de layers de 10m de haut (en gras), Norvège 1997 à 2002.

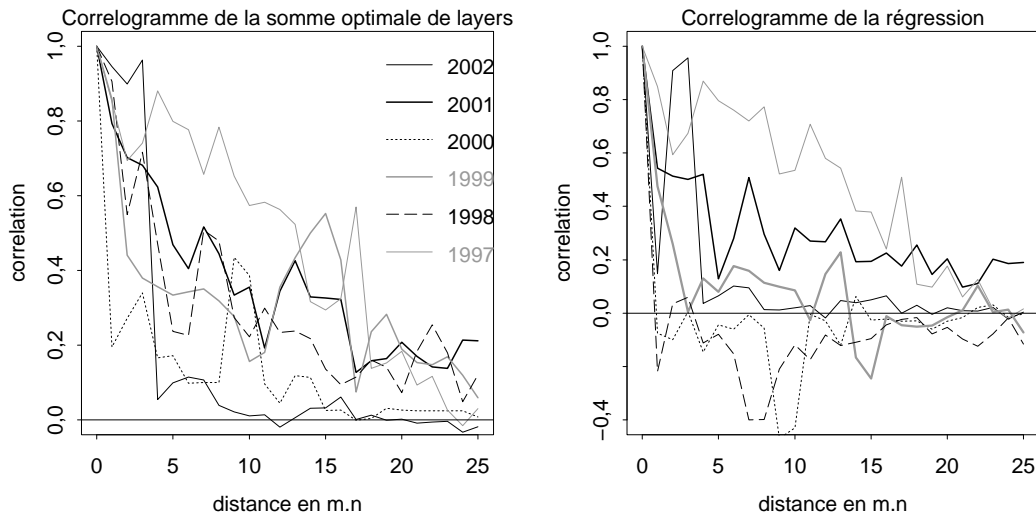


FIG. 3.13 – Corrélogramme de la somme optimale de layer acoustiques (gauche) et de la régression des layers (droite).

en stations, n'a a priori aucune raison d'être spatialement structurée, du moins plus que les variables initiales.

Ainsi l'approche par somme de layers, par rapport à la régression, a l'avantage d'être plus facilement interprétable. Elle s'avère aussi plus robuste entre les années et mieux structurée spatialement.

3.5 Analyse canonique

Une approche encore plus générale consiste à ne plus rechercher une combinaison de couches acoustiques qui maximise la corrélation avec une seule variable de capture, mais de s'ouvrir à toutes les variables de captures disponibles (cf. partie 1.1.2) pour étudier les combinaisons linéaires optimales de chaque groupe.

Introduite dans les années 30, l'analyse canonique permet d'étudier les relations linéaires entre deux groupes de variables observés sur le même échantillon. Plus précisément, on cherche une combinaison linéaire des variables du premier groupe et une combinaison linéaire des variables du deuxième groupe qui soient les plus corrélées possible. Ici, par exemple on s'intéresse à la combinaison linéaire de variables acoustiques et la combinaison linéaire de variables de captures de corrélation maximale. C'est une problématique très générale qui regroupe non-seulement les méthodes de régressions multiples (quand un des deux groupes est réduit à une seule variable) mais aussi d'analyse des correspondances (quand les variables d'un groupe sont des indicatrices d'une même variable).

Si l'on considère toutes les variables de captures disponibles notées : $P = (P_1, P_2, \dots, P_p)$.

Il peut s'agir par exemple de toutes les espèces pêchées en ENASC, pour voir quels layers sont liés à quelles espèces, ou au contraire des différentes variables d'une même espèce comme le poids, le nombre et la longueur pour comparer la combinaison linéaire obtenue à la formule du NASC-équivalent.

Les couches acoustiques sont notées comme précédemment $A = (A_1, \dots, A_k)$. Le principe est donc ici de trouver deux vecteurs α^1 et β^1 tels que : $\rho(\alpha^1 A(x), \beta^1 P(x))$ soit maximale.

Lorsque ce premier couple (α^1, β^1) est trouvé, on cherche alors un deuxième couple (α^2, β^2) tel que non seulement $\rho(\alpha^2 A(x), \beta^2 P(x))$ soit maximale mais aussi tel que : $\rho(\alpha^1 A(x), \alpha^2 A(x)) = 0$ et $\rho(\beta^1 P(x), \beta^2 P(x)) = 0$.

On peut montrer (Diday et al. 1982, [21]) que les facteurs $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ et $\beta = (\beta_1, \beta_2, \dots, \beta_k)$ sont réciproquement les vecteurs propres des matrices $C_A^{-1} C_{AP} C_P^{-1} C_{PA}$ et $C_P^{-1} C_{PA} C_A^{-1} C_{AP}$, avec C_A la matrice $\in \mathbb{R}^k \times \mathbb{R}^k$ de variance-covariance des variables acoustiques, avec C_P la matrice $\in \mathbb{R}^p \times \mathbb{R}^p$ de variance-covariance des variables de captures et C_{AP} et C_{PA} les matrices $\in \mathbb{R}^k \times \mathbb{R}^p$ et $\in \mathbb{R}^p \times \mathbb{R}^k$ de variance-covariance croisées.

Les facteurs obtenus sont orthogonaux en zéro, mais pas spatialement. L'idée pour simplifier le modèle multivariable serait donc de ne conserver que les premiers facteurs, les plus significatifs qui sont orthogonaux (pas spatialement mais aux mêmes points).

Les résultats obtenus en Norvège sont indiqués sur la figure 3.14. Le premier facteur des captures laisse souvent se dégager une seule espèce : le lieu en 1997 et 1998, le capelan en 1999, le hareng en 2000, le merlan en 2001 ; alors que le premier facteur de l'acoustique est généralement un groupe de layers aux pondérations assez régulières. De façon générale ce sont les premières couches qui ont les coefficients les plus forts.

L'analyse canonique appliquée sur les différentes variables de mesure (nombre, poids, taille, NASC-équivalent) ont tendance à dégager le nombre de pélagiques, face à des coefficients négatifs pour les layers. Le nombre est la variable de mesure la plus sensible aux outliers, surtout pour les pélagiques. Je pense donc que cette dernière observation est liée une fois de plus à la sensibilité de l'approche aux valeurs fortes qui influent énormément les corrélations linéaires.

Sur les données françaises, l'analyse canonique sur les espèces indique une forte corrélation entre les layers 11 et 12 (20 à 30 mètres au dessus du fond) et le sprat associé négativement au merlan. Le sud de la mer du Nord comprend en effet des gros bancs de pélagiques qui parviennent parfois à se faire capturer par les chaluts de fond. En Écosse, le même balancement demersal/pélagique est visible entre les deux groupes merlan-tacaud et hareng-sprat.

De façon générale, les résultats varient beaucoup d'une année à l'autre et sont assez instables.

L'analyse canonique est peu utilisée en analyse des données en raison de la difficulté à interpréter et utiliser les résultats obtenus. Bouroche et Saporta (1980 [12]) montrent un exemple où les liens qui se dégagent nettement entre les variables par l'analyse canonique pouvaient finalement être directement vus sur les données brutes. De plus, comme pour la régression linéaire qui en est un cas particulier, l'analyse canonique appliquée en stations n'a aucune raison de présenter une quelconque structure spatiale sur les données situées le long des transects.

Dans le cadre d'un modèle linéaire de corégionalisation, l'analyse canonique des matrices de chaque échelle fournit des résultats permettant de décrire les liens spatiaux entre deux groupes de variables (Wackernagel et al. 1989 [71]). Goovaerts (1994, [28]) utilise cette technique pour montrer dans

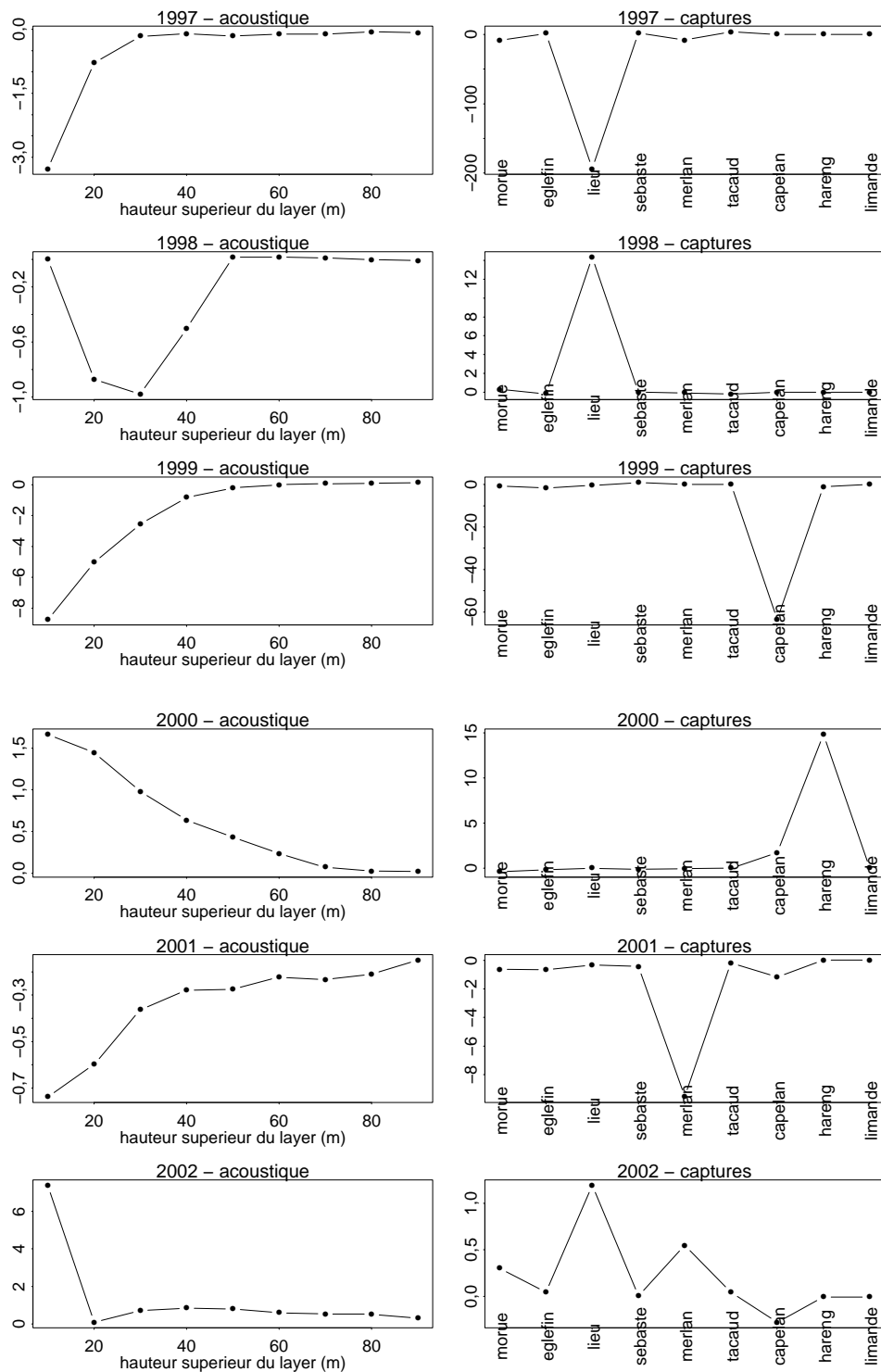


FIG. 3.14 – Coefficients du premier facteur de l'analyse canonique entre les layers de 10 m (les 10 premiers layers sont sommés) et les principales espèces (en ENASC), en Norvège 1997-2002. L'axe des ordonnées est parfois négatif, ce qui peut modifier l'interprétation des résultats.

un cas précis comment les liens entre les deux groupes de variables, géologiques et végétales, évoluent selon l'échelle de structure. La méthode se révèle particulièrement intéressante pour décrire et comprendre le phénomène étudié, même si les facteurs obtenus sont difficilement utilisables pour la modélisation.

3.6 Min/Max Autocorrelation Factors

Les facteurs obtenus par l'analyse canonique, comme toutes les méthodes d'analyse de données classiques, sont orthogonaux pour un point fixé mais pas spatialement. L'approche décrite maintenant consiste à rechercher des transformations linéaires des variables acoustiques qui soient spatialement orthogonales. Le nombre de variables reste alors le même, mais le système de cokrigage se simplifie parce que les structures croisées entre les nouvelles variables disparaissent et le cokrigage équivaut alors au krigeage de chacun des facteurs.

Rappel sur l'ACP normée

Le principe de l'analyse en composantes principales est de trouver les combinaisons linéaires orthonormées des variables qui maximisent l'inertie du nuage de points des données acoustiques dans l'espace formé par les k layers. L'avantage est de permettre de projeter un nuage de points de dimension k sur un espace de dimension 2 ou 3 (généralement on ne retient que 2 ou 3 facteurs) tel que la projection du nuage de point initial soit la plus étalée possible.

On montre que ces facteurs sont les vecteurs propres de la matrice de variance multipliés par la métrique choisie, celle-ci définie la distance que l'on a choisie pour déterminer quelles variables sont proches ou éloignées. Soit α un facteur, on a :

$$C(0)M\alpha = \lambda\alpha \quad (3.2)$$

avec $C(0) = {}^tADA$ où $C(0)$ est la matrice de variance des données, D est la matrice diagonale des poids que l'on veut mettre sur chaque point, M la métrique choisie, c'est une matrice définie-positive. Généralement, on choisit pour métrique une matrice diagonale ce qui revient à pondérer les caractères. La plus fréquemment utilisée est la métrique $M = \text{diag}(\frac{1}{\sigma_i^2})$ ce qui revient à diviser chacune des variables par son écart type. L'avantage est que les variables sont alors sans dimension et les résultats de l'ACP ne dépendent plus des variances de chacune d'elles. On fait alors une ACP normée, c'est à dire que l'on maximise la corrélation entre les données et la composante principale.

3.6.1 Principe des MAF

La méthode des Min/Max autocorrelation factors (MAF) est une méthode visant à orthogonaliser spatialement un ensemble de variables aléatoires autocorrélées. Elle a été introduite par Switzer et Green (1984 [68]) dans le but d'extraire le bruit en imagerie multivariable en éliminant les facteurs pépitiques. La méthode classique de décomposition en composantes principales fournit des facteurs orthogonaux points à points, mais ne tient pas compte de la corrélation spatiale entre les facteurs. L'idée ici est de trouver des combinaisons linéaires des variables initiales maximisant la corrélation

entre les facteurs à un pas de distance Δ fixé qui, de plus, obéissent à la contrainte d'être orthogonales en zéro et en Δ . C'est cette propriété de fournir des facteurs spatialement orthogonaux qui est la plus intéressante ici. Cependant, extraire les facteurs pépitiqes et les facteurs les plus structurés permet de simplifier encore le système, puisque non seulement les structures croisées n'auront plus à être modélisées, mais les facteurs pépitiqes non plus.

Problème

Soit $A = (A_1, \dots, A_k)$ l'ensemble des variables initiales (par exemple les variables acoustiques). On se place dans le cas stationnaire.

Alors les Y_i tels que $Y_i = {}^t\alpha^i A(x)$ sont les Min/Max autocorrelation factors avec α^i vérifiant :

$$\begin{aligned} \rho[Y_1(x), Y_1(x + \Delta)] &= \rho[{}^t\alpha^1 A(x), {}^t\alpha^1 A(x + \Delta)] = \max_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x + \Delta)] \\ \rho[Y_2(x), Y_2(x + \Delta)] &= \max_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x + \Delta)] \\ \text{avec } \rho[{}^t\alpha_i A(x), {}^t\alpha_j A(x + \Delta)] &= 0 \text{ pour } j < i \\ \dots \\ \rho[Y_k(x), Y_k(x + \Delta)] &= \min_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x + \Delta)] \end{aligned}$$

Le problème peut également s'écrire matriciellement :

Trouver α_i solution de :

$$\max_{\alpha} \frac{{}^t\alpha C(\Delta)\alpha}{{}^t\alpha C(0)\alpha} \quad (3.3)$$

$$\text{avec } {}^t\alpha_i C(\Delta)\alpha_j = 0 \text{ pour } i \neq j$$

Remarque

Notons que la matrice de corrélation peut s'écrire :

$$\begin{aligned} \rho[Y_i(x), Y_i(x + \Delta)] &= \frac{Cov[Y_i(x), Y_i(x + \Delta)]}{Cov[Y_i(x), Y_i(x)]} \\ &= 1 - \frac{Cov[Y_i(x), Y_i(x)] - Cov[Y_i(x), Y_i(x + \Delta)]}{Cov[Y_i(x), Y_i(x)]} \\ &= 1 - \frac{1}{2} \frac{Cov[Y_i(x) - Y_i(x + \Delta), Y_i(x) - Y_i(x + \Delta)]}{Cov[Y_i(x), Y_i(x)]} \\ &= 1 - \frac{1}{2} \frac{{}^t\alpha_i Cov[A(x) - A(x + \Delta), A(x) - A(x + \Delta)]\alpha_i}{{}^t\alpha_i Cov[A(x), A(x)]\alpha_i} \end{aligned}$$

Il est donc équivalent de maximiser $\frac{{}^t\alpha C(\Delta)\alpha}{{}^t\alpha C(0)\alpha}$ ou de minimiser $\frac{{}^t\alpha \Gamma(\Delta)\alpha}{{}^t\alpha C(0)\alpha}$ où $\Gamma(\Delta)$ est égal à deux fois la matrice variogramme de A au pas Δ . Dans les articles cités, on trouvera la plupart du temps des notations faisant intervenir la matrice Γ au lieu de la matrice de covariance.

Solution

Les α_i solutions des équations 3.3 sont les vecteurs propres du produit de la matrice de covariance au pas Δ par l'inverse de la matrice de variance (preuve en annexe B).

$${}^t\alpha_i C(\Delta)C(0)^{-1} = \Lambda^t\alpha_i \quad (3.4)$$

avec $C(\Delta) = Cov[A(x), A(x + \Delta)]$ et $C(0) = Cov[A(x), A(x)]$, la matrice de variance de A.

Une démonstration de ce résultat a été développée et est proposée en annexe B. Une autre preuve peut être trouvée dans Berman 1985 [5]. D'autres développements, comme l'application des MAF pour deux pas de distances Δ , peuvent, par exemple, être trouvés dans la récente thèse de E. Sicard (2004, [64]).

On voit par cette notation que les MAF ont une formulation très proche d'une ACP classique. Ils proviennent, non pas des vecteurs propres de la matrice de covariance au pas zéro, mais de celle au pas Δ renormée par celle au pas zéro. Les deux approches sont donc très voisines.

3.6.2 Application

La méthode des Min/Max autocorrelation factors a été appliquée à différents types de données géologiques (Desbarats et Dimitrakopoulos 2000 [20], Desbarats 2001 [19]). Elle a été utilisée ici sur les données acoustiques de façon à obtenir des combinaisons linéaires de layers sans corrélation spatiale. Les structures croisées entre les variables acoustiques étant alors minimisées, le système multivariable se simplifie largement. Le pas Δ choisi est le pas minimum entre deux données (cf. annexe B), soit 1 m.n en mer de Barents, 2 m.n en mer du Nord et 3 m.n en mer d'Irlande.

En Norvège, où la mer est profonde, les MAF ont été appliqués sur les layers de 10 m de haut, en sommant les 10 premiers layers de 1 m. Dans les autres mers, moins profondes, les 10 premiers mètres étant assez proches d'un modèle linéaire de corégionalisation à deux structures (une pépite + une structure moyenne portée), on se retrouve dans la configuration décrite en annexe B.

Les coefficients des facteurs les plus structurés et les plus bruités changent d'années en années. Une fois de plus cette approche est linéaire et donc sensible aux valeurs fortes, ce qui explique peut-être les problèmes de stabilité rencontrés lors de l'application aux données. On voit en effet que les compositions des MAF changent de signe de façon quasi systématique entre les layers (cf. figure 3.15). Aucune structure logique semble se dégager des compositions qui sont difficilement interprétables. Il est intéressant de voir que les facteurs de l'ACP sont au contraire des combinaisons assez régulières de couches, en particulier les premiers facteurs (i.e. ceux associés à une partie significative de l'inertie). On voit par exemple (figure 3.18) en Norvège 1997, que le premier facteur de l'ACP donne un poids d'autant plus important à la couche que celle-ci est élevée dans la colonne d'eau. Le facteur suivant oppose les couches élevées (au dessus de 80 m) aux couches plus basses. Le troisième facteur donne au contraire un poids plus important aux layers intermédiaires, autour de 60 m au dessus du fond. Les facteurs suivants, associés à une faible part de l'inertie, ont des coefficients beaucoup plus désordonnés entre les couches. On observe des résultats assez similaires pour les autres années.

On voit sur les figures 3.16 et 3.17 que les facteurs des MAF sont bien rangés du plus structuré au

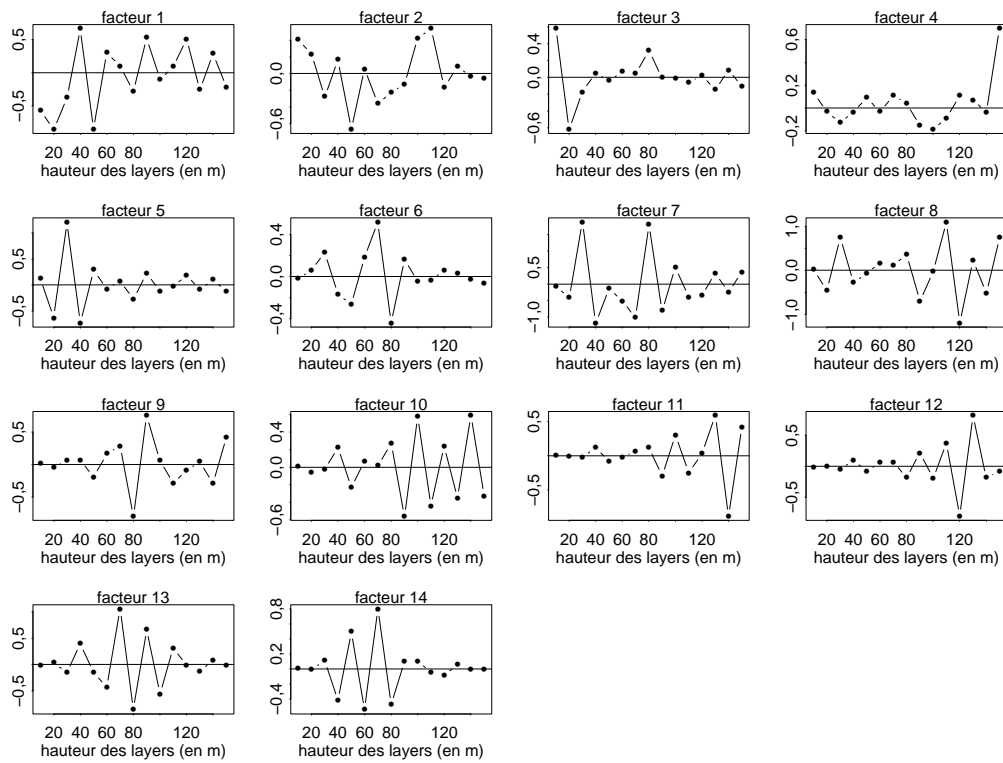


FIG. 3.15 – Composition des MAF, Norvège 1997.

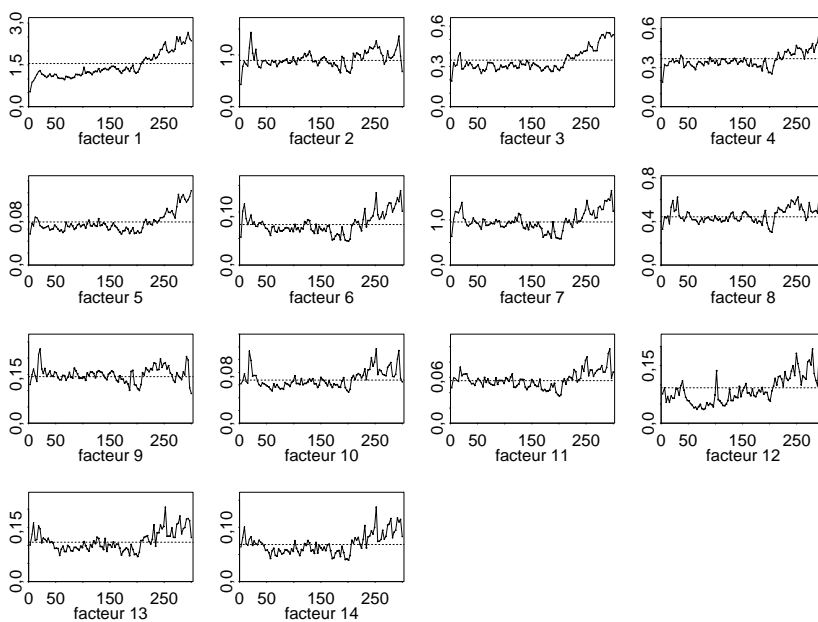


FIG. 3.16 – Variogrammes des MAF, Norvège 1997.

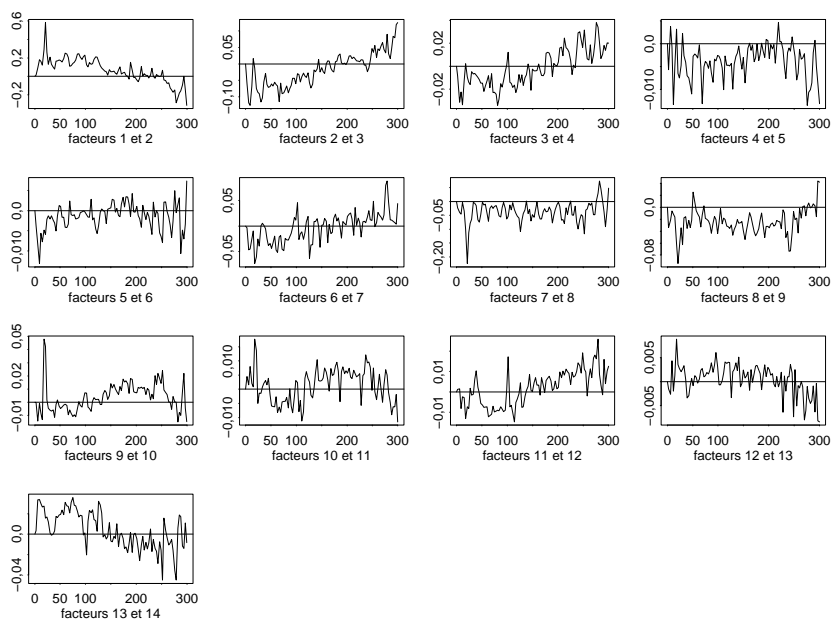


FIG. 3.17 – Variogrammes croisés entre MAF successifs, Norvège 1997.

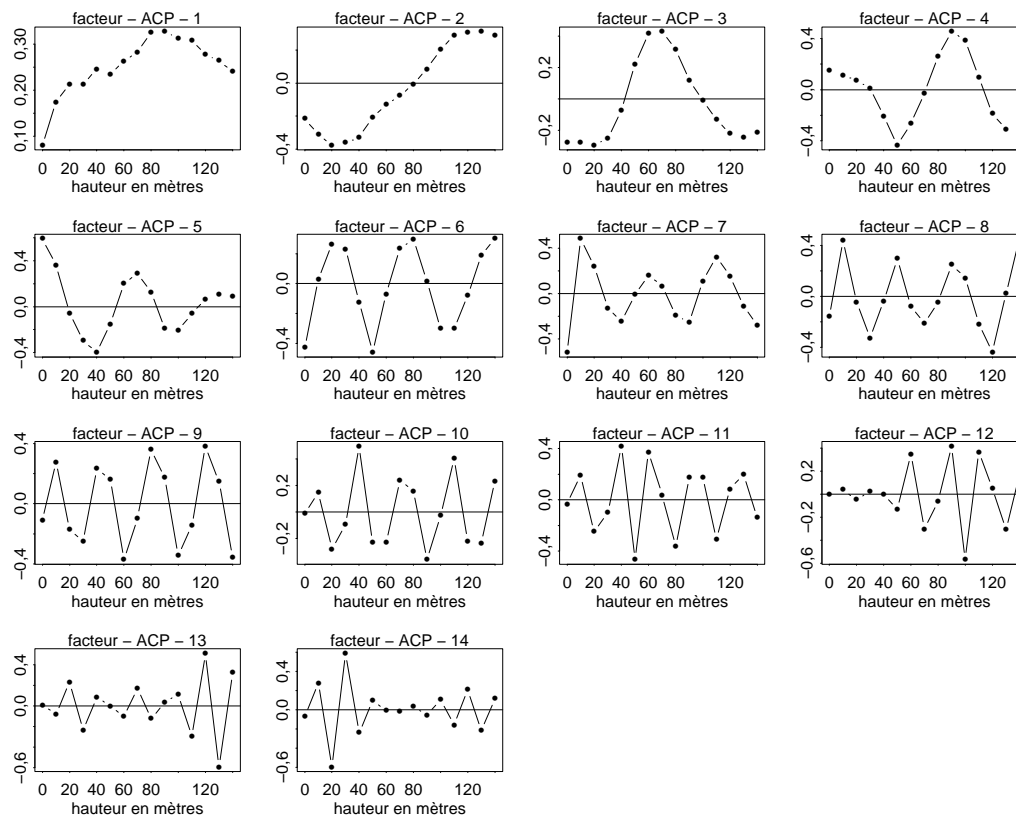


FIG. 3.18 – Composition des facteurs de l'ACP, Norvège 1997.

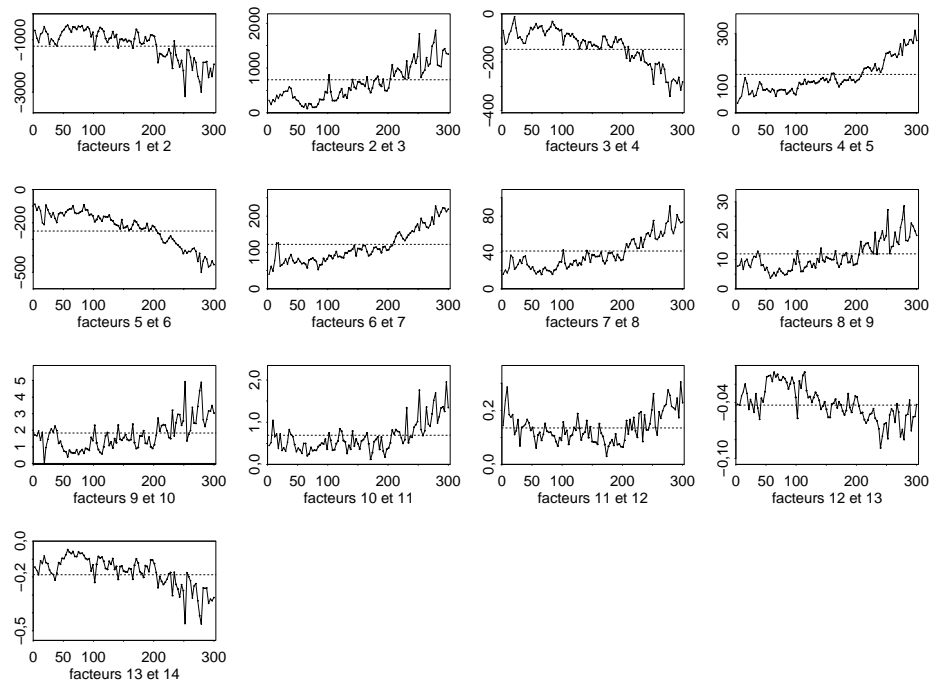


FIG. 3.19 – Variogrammes croisés des facteurs de l'ACP, Norvège 1997.

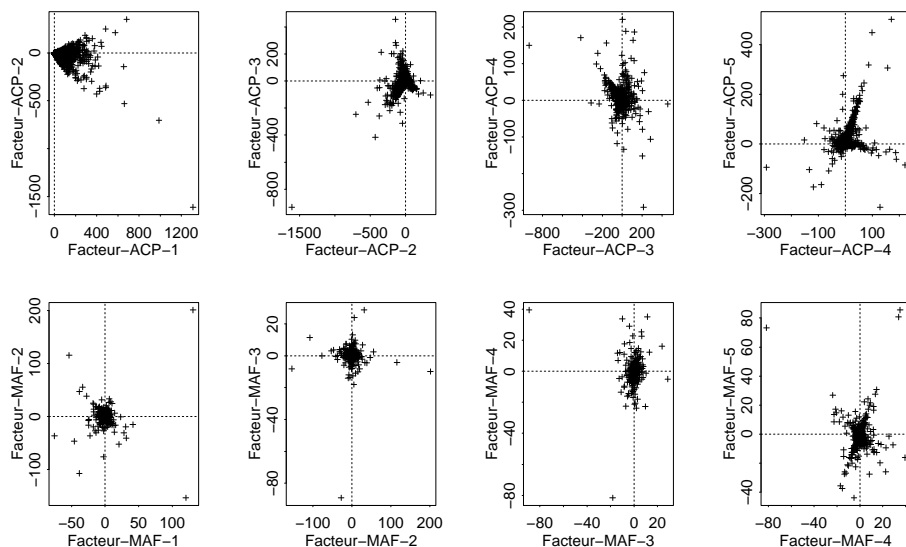


FIG. 3.20 – Comparaison des nuages de corrélations en zéro des 4 premiers facteurs de l'ACP (ligne du haut) et des 4 premiers facteurs des MAF (ligne du bas), Norvège 1997.

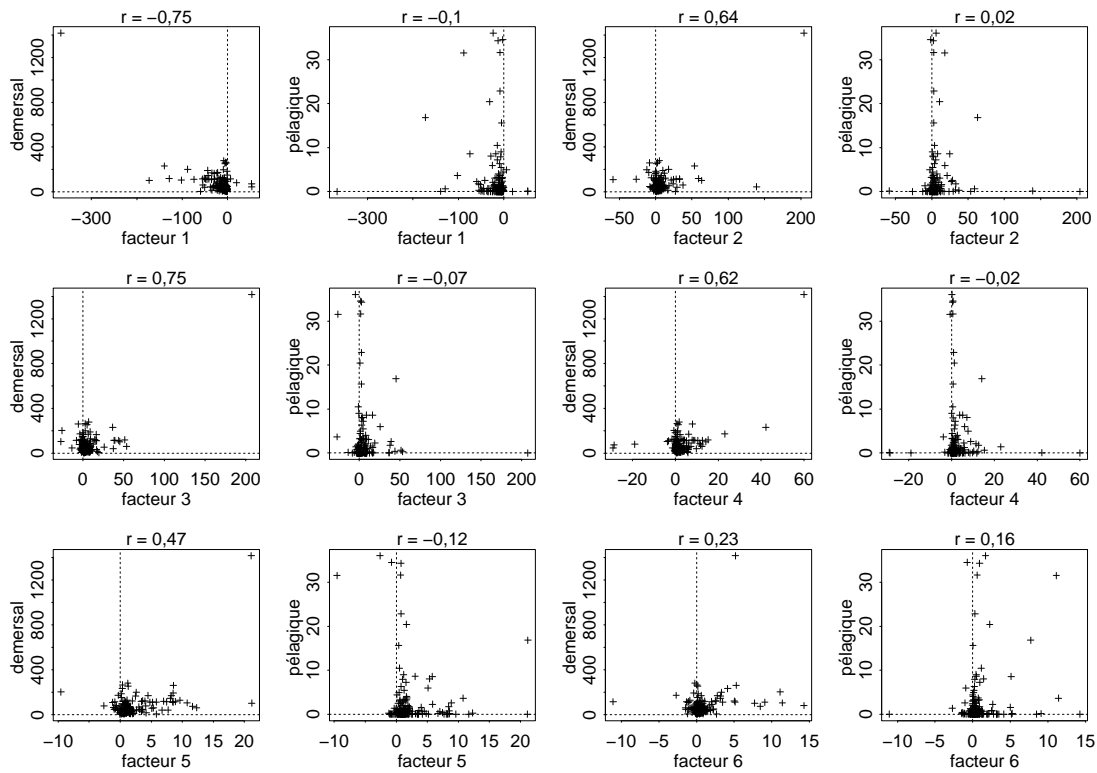


FIG. 3.21 – Nuages de corrélation entre les facteurs des MAF et les démersaux (première colonne à gauche et troisième colonne) ou des pélagiques (deuxième et quatrième colonnes) en ENASC, stations Norvège 1997.

plus bruité (l'effet de pépite est de plus en plus grand) et que leur structure croisée est négligeable. Les facteurs de l'ACP ont des variogrammes croisés plus structurés (figure 3.19). Bien que les nuages de corrélation confirme que les facteurs sont orthogonaux en zéro aussi bien pour l'ACP que pour les MAF (figure 3.20), les facteurs de l'ACP ont plus de corrélation spatiale que les facteurs des MAF.

Ainsi les facteurs des MAF ont l'avantage de fournir de nouvelles variables acoustiques à la fois bien structurées (les facteurs de bruits peuvent être mis de côté, Switzer et Green 1984 [68]) et quasiment orthogonaux spatialement. Ils sont cependant difficilement interprétables. Par ailleurs, dans l'optique visée ici qui est de construire un modèle acoustique-capture, les MAF ont le désavantage majeur, tout comme les facteurs de l'ACP, de ne pas être corrélés aux captures aussi bien en zéro (cf. les nuages de corrélations, figure 3.21) que spatialement.

Une telle approche pourrait permettre de construire un modèle acoustique pour lui-même de façon simple. Une fois les variables orthogonalisées, le système de cokrigeage revient au krigage de chacun des facteurs. Il suffit alors de revenir aux variables pour avoir un modèle complet de l'acoustique.

Par contre, quand on croise l'acoustique aux captures, le cokrigeage multivariable n'est que partiellement simplifié puisque seule la sous matrice C_A de l'équation 3.1 devient diagonale.

3.7 Approche par classification de profils verticaux

3.7.1 Principe

L'approche par classification de layers diffère totalement des méthodes présentées précédemment. Son but est toujours de simplifier le système multivariable en cherchant une variable acoustique synthétique qui permette d'interpoler de façon optimale les captures. Mais le principe est d'origine plus "halieutique" que les méthodes précédentes, classiques en analyse des données.

L'idée est de chercher des caractéristiques permettant de distinguer les profils acoustiques caractéristiques de certaines espèces ou groupes d'espèces.

Cette méthode se rapproche de la "scrutation" des données effectuées par les acousticiens. Les scientifiques experts des campagnes acoustiques ont l'habitude d'analyser attentivement toute l'information acoustique, de façon extrêmement fine. Plusieurs organismes (comme IFREMER avec MOVIES+) ont développé des outils permettant de classer automatiquement chaque banc de poissons en fonction d'un certain nombre de paramètres. Reid et al. (2000, [55]) préconisent trois types de paramètres : les paramètres morphologiques comme la longueur du banc, sa hauteur, sa surface, son périmètre ou la régularité de son contour ; des critères énergétiques comme l'intensité acoustique du banc, sa densité, sa variance interne ; ainsi que des facteurs de positions : les coordonnées géographiques du banc (latitude, longitude, profondeur), l'heure, la date, la hauteur au dessus du fond, etc... Tous ces paramètres semblent influencer fortement la caractérisation des bancs (Reid et al. 2000 [55], Petitgas et al. 2001 [54]).

L'identification d'un banc, muni de tous ces paramètres, à une espèce ou un groupe d'espèces est cependant délicate. Simmonds et al. (1996 [66]) montrent des résultats encourageants pour des espèces en captivité. Cependant, même si l'allure des bancs détectés par l'acoustique semble assez stable d'années en années et selon les zones géographiques (Muiño et al. 2003 [48]) et ne semble pas dépendre de l'état des stocks (Petitgas et al. 2001 [54]), leur attribution à une espèce semble

par contre assez sensible (Reid et al. 2000 [55]). Massé et al. (1996 [40]) soulignent par exemple les différences de comportements et les modifications de formes selon les heures du jour, en particulier pour les bancs mixtes (sprats / anchois).

Traditionnellement, l'identification du banc se fait donc au cas par cas par les experts acousticiens en fonction de captures ponctuelles auxquelles sont attribuées des zones d'influences.

Les données disponibles ici ne permettent pas une analyse aussi fine. L'information acoustique n'est pas disponible par banc mais par ESDU (voir définition dans la partie 1.1.1) et par couches verticales. Le rapport ICES (Reid and al. 2000 [55]) estime grâce à une analyse petite échelle des portées des variogrammes, que le diamètre d'un banc se situe autour de 18 m, et une agrégation de banc autour de 15 km (9,3 m.n). Petitgas (2003, [52]) observe des distances comparables de l'ordre de 10 km. On se rapproche donc dans notre cas d'une étude à l'échelle de l'agrégation de bancs. Petitgas (1996, [53]) étudie une classification à une échelle comparable en définissant un certain nombre d'images *types* au niveau d'un ESDU (ESDU formé de petits bancs, de bancs pelucheux, dispersés, condensés, etc), mais cette typologie nécessite tout de même une connaissance sur la distribution à l'intérieur de l'ESDU, non disponible dans les données utilisées dans ce travail.

L'approche présentée ici vise donc à se rendre compte à quel point ce type de classification de formes acoustiques peut être simplifiée pour être applicable de façon complètement automatique et très générale.

Plusieurs pistes ont été tentées comme l'étude de la hauteur du centre de gravité de la colonne acoustique pour les stations riches en démersaux. Une ACP a été effectuée pour dégager les variables les plus pertinentes parmi la position du centre de gravité de la colonne, sa variance, la proportion des premiers layers dans l'énergie totale de la colonne d'eau, l'intensité absolue de la colonne d'eau, et ces mêmes variables mesurées avant, pendant et après le coups de chalut pour tenir compte également de l'effet des déplacements de poissons. Malheureusement si certaines tendances ont pu être dégagées, les variables choisies ne semblaient pas très discriminantes.

La méthode développée part finalement de l'observation suivante : si l'on s'intéresse à l'espèce ou le groupe d'espèce le plus abondant de la zone d'étude, le profil acoustique moyen des stations qui en sont très riches est très différent du profil vertical moyen des stations qui en sont très pauvres (cf. figures 3.22, 3.23, 3.24). En partie, la force de cette propriété provient de ce qu'elle est vérifiée quelque soit la campagne étudiée.

Deux phénomènes sont observés :

- l'évolution de la forme avec la densité de l'espèce principale est progressive, elle semble s'accroître jusqu'à une forme maximale à partir de laquelle elle converge ou revient vers la moyenne selon les cas ;
- la forme générale correspond dans l'ensemble à la théorie biologique selon laquelle les démersaux vivent plus au fond que les pélagiques.

Les deux profils extrêmes servent de "profils-type" pour la classification. Chaque profil est affecté à une classe, soit "station de type pélagique" soit "station de type non-pélagique" (si le groupe d'espèce d'intérêt est les pélagiques), en le comparant à chacun des deux profils types précédemment établis, par un critère de minimisation des écarts quadratiques. Pour cela tous les profils acoustiques sont normalisés, c'est à dire divisés par leur intensité totale, pour travailler uniquement sur leur forme. On associe également à chaque profil une classe d'intensité selon que l'acoustique totale de la colonne d'eau est supérieure ou inférieure à la moyenne de la campagne.

On ne dispose alors plus que de 2 variables pour décrire un profil acoustique :

- son intensité, $I(x) = i_1$ ou i_2 selon que le profil est de somme totale supérieure ou non à la moyenne de la campagne
- sa forme, $F(x) = f_1$ ou f_2 selon que la forme du profil une fois normalisé est plus proche du profil de type de l'espèce considérée ou du groupe d'espèces complémentaire (démersal/non démersal, pélagique/non pélagique ou morue/autre, etc), au sens des moindres carrés.

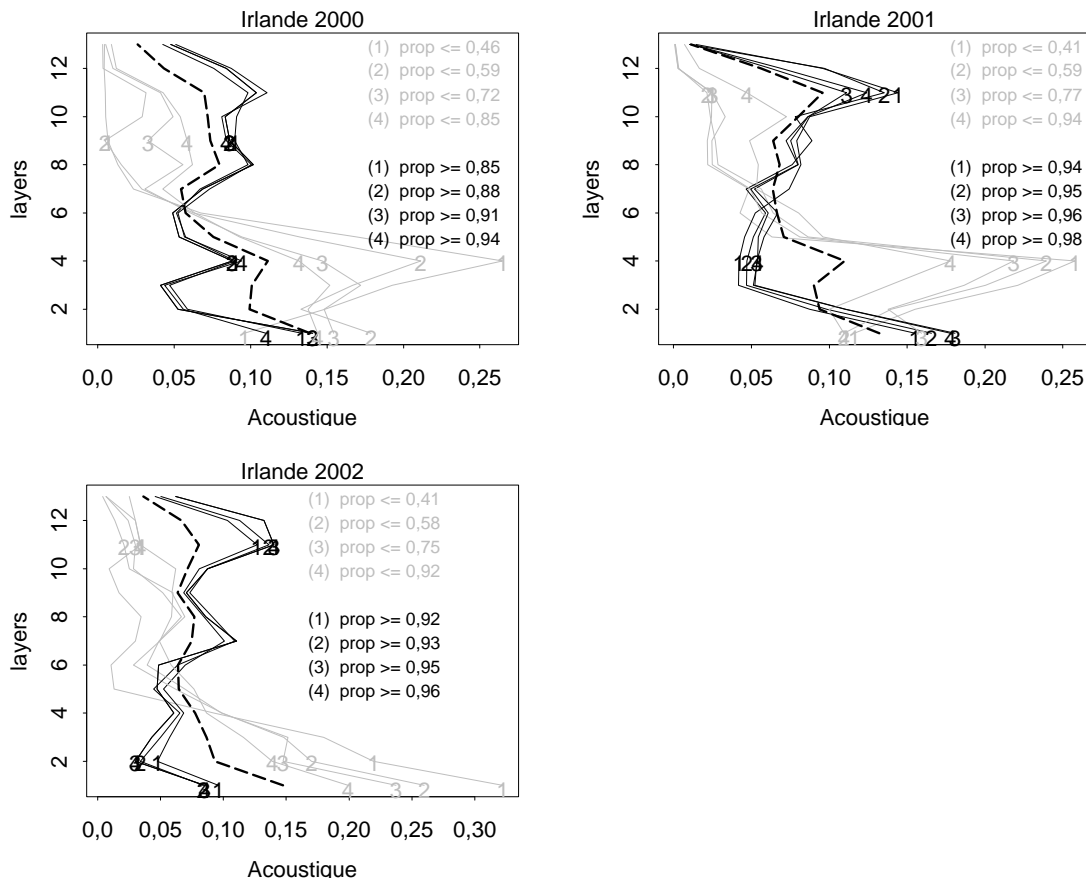


FIG. 3.22 – Évolution du profil acoustique normé en fonction de la proportion de démersaux pêchés, Irlande 2000, 2001 et 2002. La courbe pointillée représente le profil moyen sur toutes les stations. Les profils gris, les stations contenant une faible proportion de démersaux et les profils noirs, les stations contenant une forte proportion de démersaux.

L'ambition de cette méthode est de modéliser le lien acoustique-capture uniquement à partir de ces deux indicatrices et d'évaluer ses performances et ses limites.

3.7.2 Modélisation

La modélisation peut se résumer en quelques étapes :

1. Définition des deux formes "types" de profils, comme par exemple : pélagique/non pélagique,

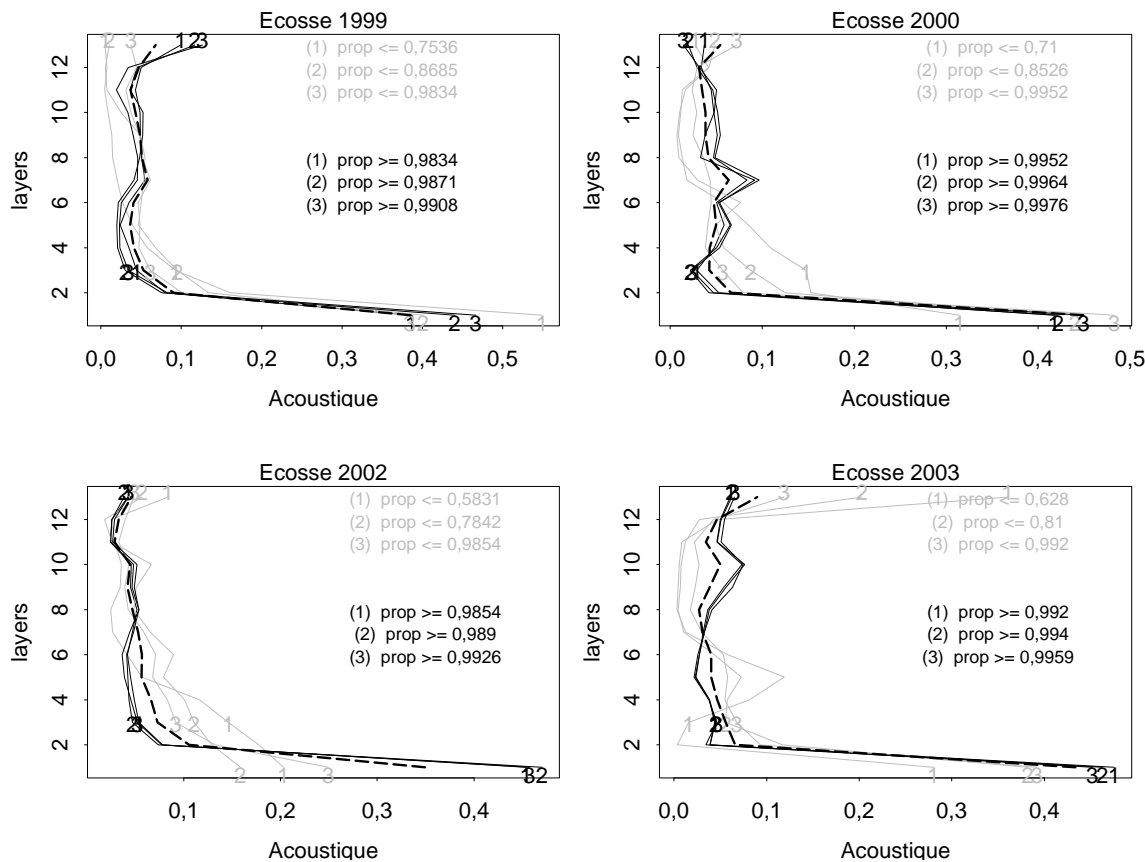


FIG. 3.23 – Évolution du profil acoustique normé en fonction de la proportion de démersaux pêchés, Écosse 1999, 2000, 2002 et 2003. La courbe pointillée représente le profil moyen sur toutes les stations. Les profils gris, les stations contenant une faible proportion de démersaux et les profils noirs, les stations contenant une forte proportion de démersaux.

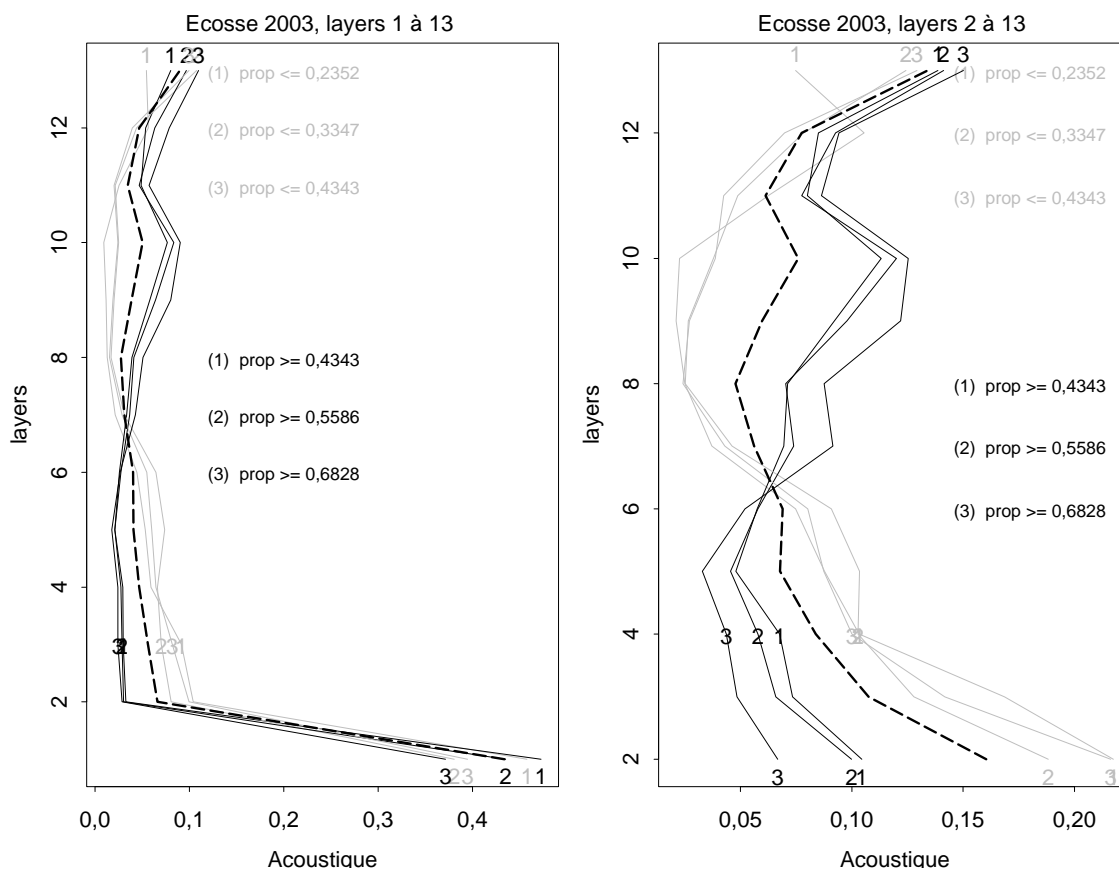


FIG. 3.24 – Évolution du profil acoustique normé en fonction de la proportion d'églefins pêchés, Écosse 2003.

démersal/non démersal, morue/autre, etc....

2. Classification des profils normalisés et attribution à chaque colonne acoustique d'une classe de forme et d'une classe d'intensité.
3. Estimation des proportions moyennes de démersaux (ou de l'espèce particulière considérée) pour chacune des 4 classes :

$$p_{\text{espece}}(i, f) = \frac{1}{N_{if}} \sum_{x \in \text{stations}} \frac{P_{\text{espece}}(x)}{P_{\text{total}}(x)} \mathbb{1}_{I(x)=i} \mathbb{1}_{F(x)=f}$$

où $P_{\text{espece}}(x)$ représente les captures d'une espèce donnée en x et $P_{\text{total}}(x)$ l'intégralité de la pêche pour cette même station.

4. Krigeage de chacune des indicatrices $\mathbb{1}_{I(x)=i}$ et $\mathbb{1}_{F(x)=f}$. On obtient alors quatre cartes d'estimation d'une probabilité d'appartenance à chacune des classes. S'il y avait plus que deux classes d'intensité, il conviendrait alors d'effectuer un cokrigeage d'indicatrices, ou d'utiliser un modèle assurant une cohérence entre les classes.
5. Estimation de la proportion de l'espèce étudiée par pondération de chacun des krigeages en fonction des proportions de l'espèce dans chacune des quatre classes. La proportion s'estime donc par :

$$p_{\text{espece}}^*(x) = \sum_{i=1}^2 \sum_{f=1}^2 p_{\text{espece}}(i, f) (\mathbb{1}_{I(x)=i})^K (\mathbb{1}_{F(x)=f})^K \quad (3.5)$$

Vérification des hypothèses

Une fois la classification faite sur les données en stations, il est important de vérifier que l'on retrouve bien en moyenne les proportions voulues : que le groupe correspondant à la forme démersal a bien en moyenne une proportion plus forte en démersal que l'autre groupe. Pour cela, des boxplot, des tests de Student, et de Wilcoxon (plus robuste à la non-normalité) ont été effectués sur chacune des populations (fig. 3.25 et tables 3.1 et 3.2).

Le test de Student teste l'égalité des moyennes de deux échantillons de loi normale. Sur le même principe, le test de Wilcoxon se base sur la somme pondérée des rangs des différences entre les deux échantillons pour n'importe quel type de lois.

On peut constater que de manière générale la proportion en démersal des stations de la première forme est jugée significativement supérieure à la proportion en démersal des stations de forme 2 (tests à niveaux de significativité de 5%, cf. tableaux 3.1 et 3.2). Les profils démersaux sont également d'intensité souvent plus faible que les profils acoustiques.

Cependant il convient de vérifier qu'il ne s'agisse pas d'une tautologie. Étant donné que l'on classe les profils selon que leur forme se rapproche ou non de celle observée dans les stations très riches ou très pauvres en démersaux, il est normal que le groupe du premier type soit plus riche en démersaux puisque par construction, les stations les plus riches y sont. Il est important de vérifier qu'il existe une différence significative de proportion parmi le sous-groupe de stations qui n'ont pas servi à déterminer les profils types. Il s'agit donc des stations de concentration plus ou moins moyennes. Or l'analyse des proportions montre qu'en particulier en Norvège (table 3.1) et en Irlande (table 3.2), les tests sont toujours vérifiés, même si les p-values sont moins discriminantes.

Une hypothèse sous-jacente à la modélisation est que les deux indicatrices d'intensité et de forme sont indépendantes. Cette hypothèse signifie qu'il ne faut pas, par exemple, que la quantité de poissons présents influe sur leur position verticale. Cette hypothèse assez forte a été vérifiée par un test du χ^2 qui permet de tester l'indépendance de deux caractères (Dacunha-Castelle et Duflo 1982 [17]). Les résultats ont montré que pour toutes les campagnes, l'hypothèse d'indépendance peut être acceptée, sauf en Irlande où le peu de données en stations rend le test délicat. Il convient également de tester spatialement cette hypothèse en traçant le variogramme croisé des deux indicatrices (figure 3.26).

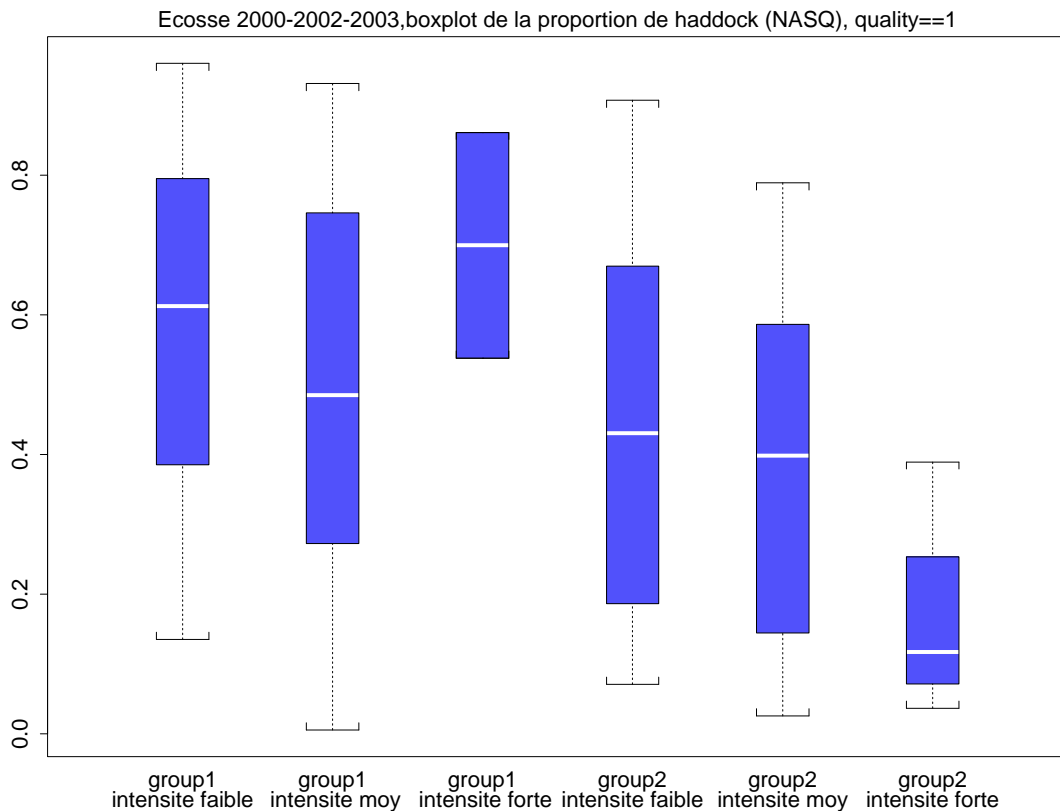


FIG. 3.25 – Boxplot de la proportion d'églefins pour 2 sortes de formes de profils et 3 niveaux d'intensité, Écosse 2003.

Résultats

Étude des profils verticaux

Il est assez général de voir que les stations riches en démersaux ont une intensité en moyenne plus faible que les stations riches en pélagiques.

Année	Variable	Forme=1, Intensité faible	Forme=0, Intensité faible	Forme=1, Intensité forte	Forme=0, Intensité forte
1997	Nb de points	38	50	50	38
1997	Proportion moy. de morue	0.443	0.380	0.508	0.292
1997	Écart-type de la prop.	0.223	0.289	0.305	0.238
1997	P-value Student	0.128		0.000	
1997	P-value Wilcoxon	0.048		0.000	
1998	Nb de points	32	67	39	60
1998	Proportion moy.de morue	0.730	0.508	0.536	0.431
1998	Écart-type de la prop.	0.238	0.241	0.237	0.268
1998	P-value Student	0		0.022	
1998	P-value Wilcoxon	0		0.020	
2000	Nb de points	53	75	54	97
2000	Proportion moy. de morue	0.492	0.392	0.419	0.348
2000	Écart-type de la prop.	0.279	0.270	0.295	0.267
2000	P-value Student	0.023		0.073	
2000	P-value Wilcoxon	0.035		0.076	
2001	Nb de points	63	86	46	104
2001	Proportion moy. de morue	0.496	0.394	0.373	0.247
2001	Écart-type de la prop.	0.279	0.252	0.275	0.206
2001	P-value Student	0.011		0.003	
2001	P-value Wilcoxon	0.011		0.005	
2002	Nb de points	49	86	51	92
2002	Proportion moy. de morue	0.480	0.321	0.471	0.308
2002	Écart-type de la proportion de morue	0.241	0.206	0.254	0.197
2002	P-value Student	0		0	
2002	P-value Wilcoxon	0		0	

TAB. 3.1 – Statistiques et tests sur chacune des 4 classes après classification des profils acoustiques en stations, Norvège. La forme 1 correspond aux profils acoustiques de type morue, la forme 0 aux autres. Les tests testent l'hypothèse H_0 : "proportion 1 = proportion 2" contre H_1 : "proportion 1 > proportion 2".

Variable	Forme=1, Intensité faible	Forme=0, Intensité faible	Forme=1, Intensité forte	Forme=0, In- tensité forte
Nb de points	43	19	28	33
Proportion de démersaux	0.919	0.791	0.900	0.614
Écart-type de la prop.	0.099	0.253	0.135	0.279
P-value Student	0.023		0.000	
P-value Wilcoxon	0.029		0.000	

TAB. 3.2 – Statistiques et tests sur chacune des 4 classes après classification des profils acoustiques en stations, Irlande. En raison du peu de points, les stations de toutes les années sont confondues 1997 et 2000-2002. La forme 1 correspond aux profils acoustiques de type démersal, la forme 0 aux autres (plutôt de type pélagiques). Les tests testent l'hypothèse H_0 : "proportion 1 = proportion 2" contre H_1 : "proportion 1 > proportion 2".

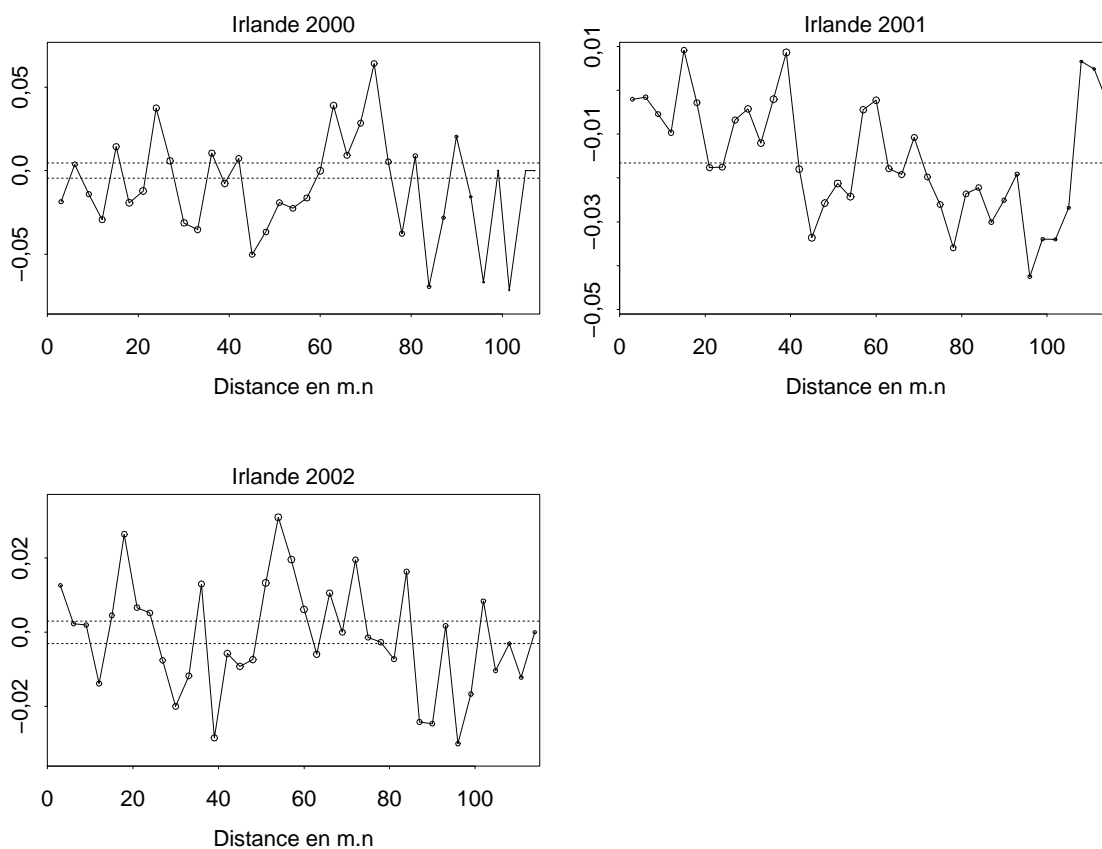


FIG. 3.26 – Variogrammes croisés des deux indicatrices de formes et d'intensité en Irlande 2000, 2001 et 2002.

En Norvège, on s'aperçoit que le groupe démersal contient des espèces ayant des comportements très différents les uns des autres. Les stations riches en merlan correspondent souvent à des profils acoustiques plus riches sur les couches de 10 à 100 mètres que de 0 à 10 mètres. Le tacaud semble souvent assez décollé du fond et assez haut dans la colonne d'eau. La morue, par contre, correspond à des stations riches dans les tous premiers mètres. A cause de cette discordance, les stations riches et pauvres en démersaux n'ont pas des profils très discriminants. En Norvège, la méthode a donc été appliquée à la morue qui est l'espèce la plus répandue de mer de Barents.

En mer du Nord et en mer d'Irlande, les profils démersaux/pélagiques sont souvent nettement différents.

On observe que les profils de type démersaux sont souvent très riches dans les deux premiers mètres. En Irlande (figure 3.22) les pélagiques se situent dans les tous premiers mètres tandis que les démersaux correspondent à des profils acoustiques homogènes sur toute la hauteur (après normalisation). Lorsque l'on trace les écart-types autour des profils types on voit cependant que les zones de valeurs se superposent, parfois dans une assez large mesure, en particulier pour les campagnes anglaises. La typologie est alors bruitée et l'erreur d'affectation en sera d'autant plus grande.

Dans les données françaises, les espèces les plus capturées sont le merlan suivi du hareng et du sprat. Il est intéressant de voir que le merlan se situe plus bas dans la colonne d'eau pour les intensités faibles que pour les intensités fortes (layers 4-5 au lieu de 1-2). Le hareng et le sprat déterminent une forme assez discriminante très riche dans les layers 9 et 10.

Dans données écossaises, les démersaux prennent une part très importante de chaque pêche. Les coups de chalut parmi les plus faiblement chargés en démersaux de la campagne en comptent tout de même encore plus de 50%. On peut par contre obtenir des profils quasiment purs à 100% en démersaux. On voit que les différences entre les profils acoustiques des stations riches et "pauvres" en démersaux changent d'années en années et se situent dans les 3 premiers mètres au dessus du fond (figure 3.23). Le profil démersal pur semble (excepté en 1999 où les différences ne sont pas nettes) très riche dans le premier mètre puis assez homogène sur toute la hauteur. Tandis que les profils des stations moins riches en démersaux sont moins riches dans le premier mètre, mais plus dans les layers 2 à 4.

Si l'on se concentre sur l'églefin, on s'aperçoit que les profils des deux classes se distinguent très nettement, surtout si l'on supprime le premier layer (figure 3.24). L'églefin ne semble pas se situer dans les premiers mètres au dessus du fond, mais légèrement plus haut dans la colonne d'eau.

Krigeage des indicatrices de forme et d'intensité

Les deux indicatrices de forme et d'intensité sont bien structurées spatialement (cf. figure 3.27). En mer de Barents, on peut voir que le rang et la portée des variogrammes sont très stables dans le temps. Les deux indicatrices étant de plus orthogonales, leur cokrigeage est équivalent à leur krigeage. La nette structure et le nombre de données disponibles (toutes les données en radiales) permettent d'obtenir une estimation par krigeage avec des variances relativement faibles.

La figure 3.28 montre les deux cartes de krigeage obtenues pour chaque indicatrice pour la campagne 2002. La probabilité d'occurrence d'un profil d'intensité supérieur à la moyenne est très concentrée dans le nord et le sud-est de la zone. La forme de profil associé à la morue à par contre une probabilité

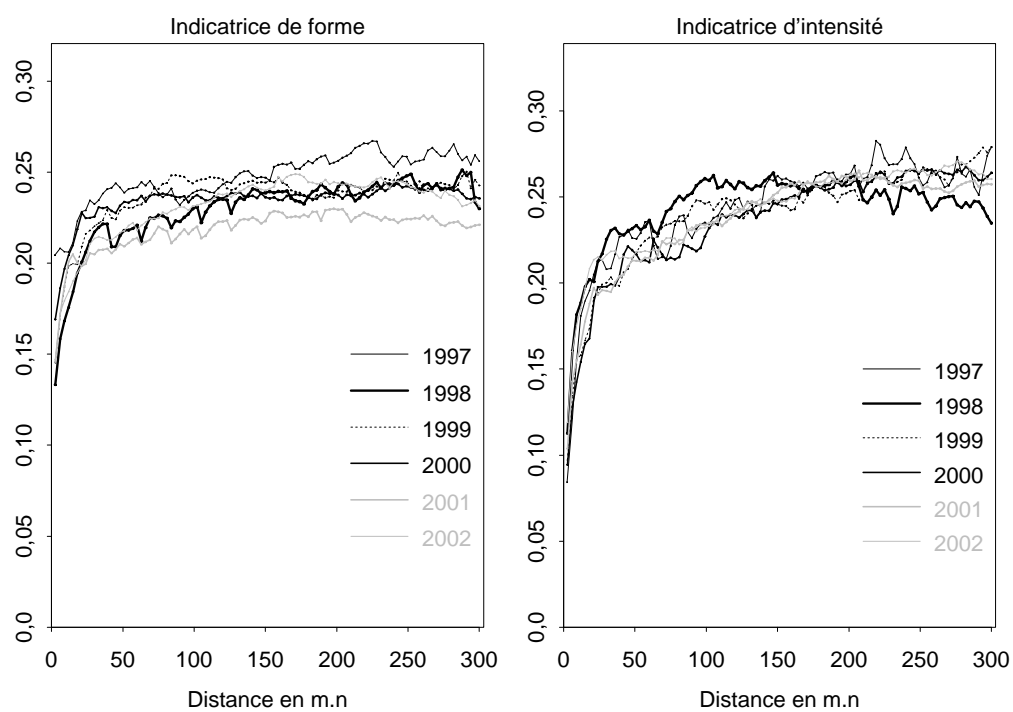


FIG. 3.27 – Variogrammes des deux indicatrices de formes et d'intensité, pour la morue, Norvège 1997 à 2002.

plus dispersée spatialement.

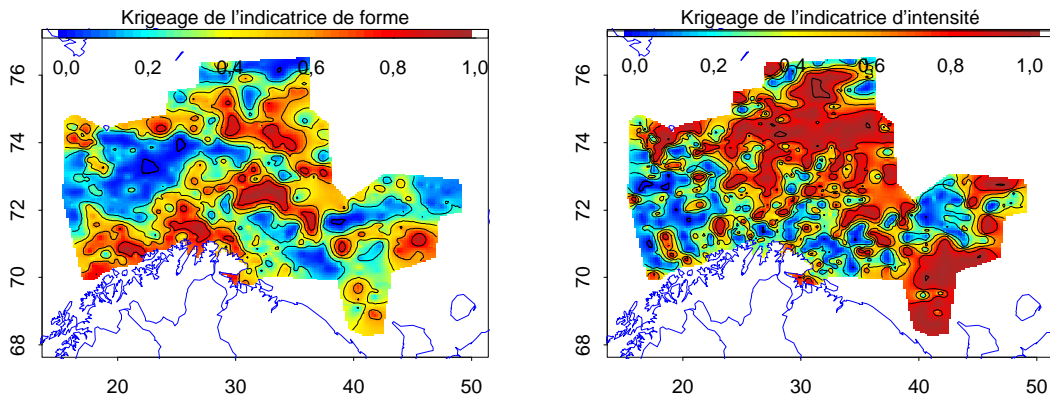


FIG. 3.28 – Krigages des deux indicatrices de forme (à gauche) et d'intensité (à droite) dans la méthode de classification appliquée à la proportion de morue, Norvège 2002.

La carte d'estimation obtenue par pondération des krigages d'indicatrices (cf. équation 3.5) permet de retrouver le relief des données, mais de façon moins contrastée. On voit par exemple qu'en Irlande les données se superposent assez bien à la carte d'estimation obtenue (figure 3.29). En Norvège, l'allure générale des zones riches et pauvres se retrouve assez bien par la méthode de classification, excepté certaines zones, comme le sud-ouest qui est sur-estimé par rapport au krigage des véritables proportions de morues observées (figure 3.30).

La carte d'estimation obtenue par classification est très proche de la carte de krigage de l'indicatrice de forme (figure 3.28). Il est intéressant de voir que le sud-ouest et l'extrême est de la carte obtenue par classification (zones réciproquement sur et sous évaluées) auraient pu être plus proches des vraies valeurs de proportions si plus de poids avait été accordé à l'indicatrice d'intensité (qui est justement plus proche du krigage de proportions dans ces zones).

La méthode par classification des profils acoustiques n'a été obtenue qu'à partir des données acoustiques. Les données de pêche ne sont utilisées qu'indirectement pour définir les critères de la classification. Il est donc assez remarquable que plusieurs reliefs soient retrouvés. Les cartes permettent d'obtenir beaucoup de détails dans les estimations en particulier en Norvège (figure 3.30, ligne du haut). Le krigage des proportions de morue en stations est très lisse (l'effet de pépité du modèle de variogramme est d'environ 60%). La méthode fournit une proportion d'abondance également dans des zones où aucun chalutage n'a été effectué, comme dans le sud de la mer d'Irlande.

Cependant l'estimation est par construction extrêmement lissante. C'est la principale limite de cette approche. En effet, si p_{11} , p_{12} , p_{21} et p_{22} sont les quatre proportions de chacune des quatre classes, les valeurs de proportions estimées ne pourront pas être inférieures à la plus petite des quatre, ni supérieures à la plus grande (cf. équation 3.5). Lorsque la carte d'estimation obtenue par classification est mise à la même échelle de couleur que la carte de krigage des proportions (figure 3.30) ligne du bas), elle s'avère en effet beaucoup plus plate, bien que les lignes de niveaux soient plus sinueuses.

La formule 3.5 permet d'estimer la proportion d'abondance à partir des proportions moyennes

de chaque classe. Mais c'est en fait toute une distribution statistique empirique qui est disponible, pas uniquement la moyenne. Une perspective intéressante serait alors de simuler différentes valeurs possibles de proportions. Cette méthode permettrait d'utiliser toute la variabilité des proportions de chaque classe et d'être ainsi beaucoup moins lissante. Au lieu de se fixer quatre constantes, on s'intéresserait alors à quatre lois, ou plus raisonnablement une seule loi avec différents paramètres. Lorsqu'un profil est attribué à une des quatre classes par la classification, une proportion pourrait alors être tirée suivant la loi correspondante.

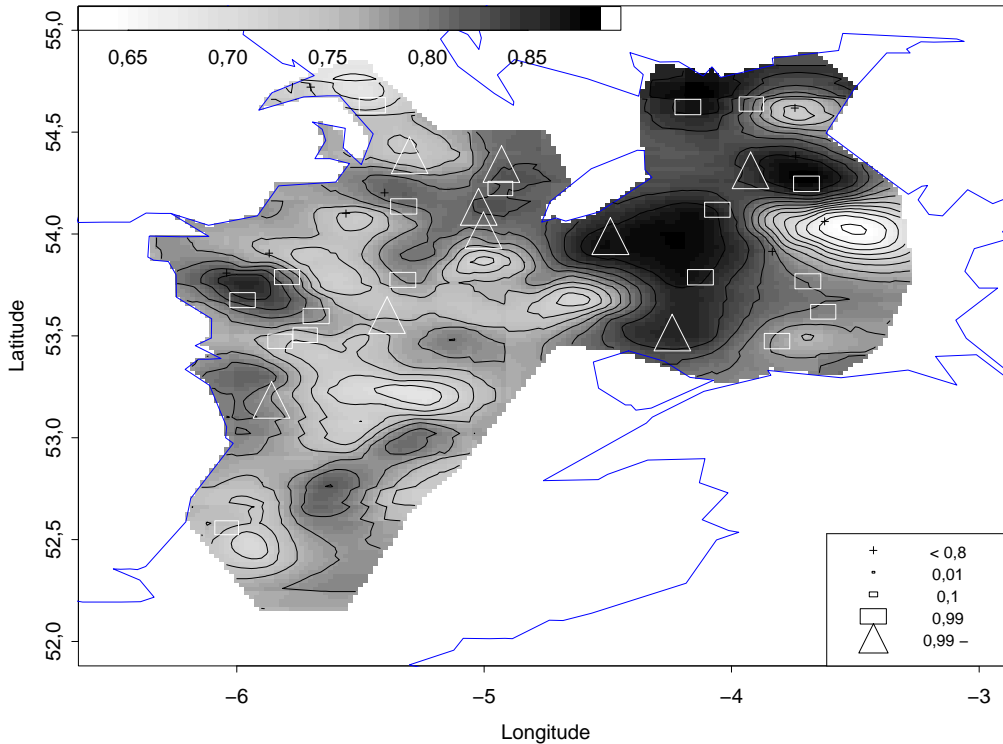


FIG. 3.29 – Carte de proportion de démersaux obtenue par la méthode de classification de profils acoustiques, Irlande 2001.

Discussion

Autres approches envisagées

D'autres approches similaires visant à améliorer la finesse du modèle ont été testées sans obtenir de résultats satisfaisants. Plutôt que de s'imposer une variable de forme de profil égale à 0 ou 1 selon que le profil étudié se rapproche d'un profil de type démersal ou d'un profil de type pélagique (ou d'une espèce particulière), on peut estimer deux paramètres α et β tels que la forme du profil soit une combinaison linéaire des deux profils types.

$$F(x) = \alpha \cdot f_1 + \beta \cdot f_2$$

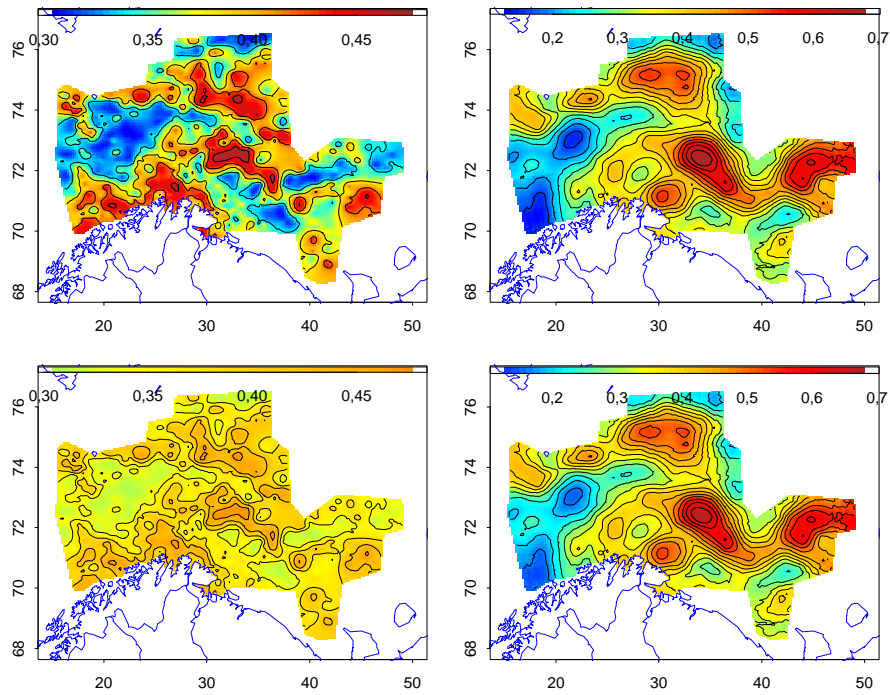


FIG. 3.30 – Carte de proportion de morue obtenue par la méthode de classification de profils acoustiques (colonne de gauche) et krigage des proportions de morue observées en stations (colonne de droite). Les deux cartes sont à des échelles de couleur différentes sur la première ligne et aux mêmes échelles sur la ligne du bas. Norvège 2002.

où f_1 et f_2 sont les deux formes types choisies. Ce qui donne en imposant une somme des coefficients égale à 1 :

$$F(x) = \alpha \cdot f_1 + (1 - \alpha) \cdot f_2$$

Le coefficient peut être estimé en minimisant la somme des écarts quadratiques.

On espère alors que le coefficient α soit d'autant plus fort quand la proportion de captures démersales de la station étudiée est grande. Cependant la corrélation entre ce coefficient et la proportion de l'espèce s'avère la plupart du temps mauvaise. Bien que les profils proches de la forme type démersal soient en moyenne plus riches en démersal et réciproquement les stations pauvres en démersal soient plus proches d'un profil de type pélagique, un profil "entre les deux" ne correspond pas nécessairement à une station où les proportions des deux espèces sont égales.

J'ai également essayé de prendre en compte l'erreur de classification, c'est à dire la somme des erreurs quadratiques, en tant qu'erreur de mesure dans le krigeage par exemple, ou en ajoutant une classe "indéterminée". Mais les estimations ne sont pas meilleures. L'ajout d'une erreur de mesure rend la carte de krigeage encore plus lissée (cf. krigeage avec erreur de mesure, Chilès et Delfiner 1999, [15]).

Les profils types sont définis en effectuant des moyennes sur les stations ayant une concentration forte en une espèce donnée. Mais ces profils ne sont pas 100% mono-espèce. La classification sur les profils légèrement extrapolés, c'est à dire accentués dans leur tendance, a donc été testée. Mais l'attribution par moindres carrés d'une forme à un des deux groupes est en fait peu sensible à de faibles variations de chacune des formes types, pourvu qu'elles gardent les mêmes allures générales. Ce ne sont que les indéterminations sur quelques profils très proches de la moyenne qui peuvent être levées.

Estimation d'une proportion

L'approche fournit une carte de proportion d'abondance et non une carte d'abondance absolue. C'est à mon sens la principale limite de la méthode.

Obtenir à partir de l'estimation des proportions une carte d'abondance n'est pas évident. Il serait possible par exemple de multiplier la carte obtenue par une estimation de l'acoustique cumulée sur toute la hauteur d'eau. Bien qu'il s'agisse d'une grandeur en ENASC, c'est à dire en unité acoustique, la proportion estimée correspond à l'interpolation d'une proportion observée *dans le chalut*, c'est à dire capturable. La multiplier par l'acoustique cumulée sur toute la hauteur d'eau n'a pas forcément de sens. Une autre solution consiste à interpoler le total des captures en ENASC pour la multiplier à la carte de proportion. On obtient alors une abondance capturable.

Cependant le produit de cartes d'interpolation a tendance naturellement à multiplier les erreurs et l'abondance totale pêchée n'étant disponible qu'en stations, la carte finale d'abondance risque de perdre en précision.

En pratique, les cartes de proportions ou d'abondances obtenues par cette méthode sont satisfaisantes, en particulier en Irlande, bien que les cartes ont moins de détails que les cartes de proportions.

Conclusion

Le rapport ICES (Reid et al. 2000 [55]) indique en perspective d'une approche par réseaux de neurones que la phase d'apprentissage ne devrait peut-être pas être faite uniquement sur des captures mono-espèces, mais également sur les stations mixtes. Il est ici aussi possible qu'augmenter le nombre de modalités de l'indicatrice de forme pour permettre une meilleure identification des

bancs. Il faudrait alors établir un modèle disjonctif prenant en compte les effets de bords entre les modalités.

Cependant, le but était ici de se rendre compte si une modélisation du type de celle des classifications de bancs faites par les acousticiens pouvait fournir des résultats intéressants à une échelle plus grossière. Quitte à affiner le modèle, il faudrait alors peut-être revenir à l'échelle du banc pour étudier par exemple la mise en pratique d'une méthode d'identification telle que celle décrite par Simmonds et al. (1996 [66]) en milieu naturel.

Je pense cependant que les résultats obtenus ici sont encourageants et montrent qu'il existe bien une identification des formes acoustiques possible à l'échelle étudiée. Les estimations obtenues, bien que par construction trop lissantes, permettent de retrouver les reliefs généraux observés sur les captures. Des simulations prenant en compte toute la distribution des proportions au sein de chaque classe permettraient, je pense, de fournir des cartes beaucoup plus réalistes rendant compte de toute la variabilité et de fournir des cartes de variances empiriques de l'estimation pour quantifier la précision effective du modèle.

3.8 Synthèse

Un certain nombre de critères ont permis de classer les différentes analyses précédemment décrites et de choisir celle semblant la mieux appropriée aux besoins de l'étude :

- La corrélation entre la ou les variables acoustiques obtenues et les captures
- Le niveau de simplification du modèle en vue du cokrigeage, c'est à dire le nombre de variables et leur orthogonalisation
- La structure spatiale (l'auto-corrélation) des variables
- L'interprétabilité des résultats : il m'a semblé important que les nouvelles variables acoustiques obtenues soient des combinaisons de layers ayant un sens physique satisfaisant
- La sensibilité de la transformation aux valeurs fortes : il est important étant donné l'asymétrie des distributions que les variables transformées soient robustes.

Les données sont telles que la corrélation acoustique-capture est de façon générale assez faible et que les variables sont assez peu structurées (voir chapitre 1 partie 1.2). L'idée directrice de cette étude verticale des couches acoustiques visait à tirer le meilleur parti possible de cette configuration pour pouvoir établir un modèle combiné acoustique-capture.

Le schéma 3.31 classe toutes les approches décrites selon trois critères.

Comme on l'a vu, la méthode par régression maximise la corrélation avec la variable de capture (coefficients de corrélation autour de 0.7) mais fournit une combinaison de layers spatialement peu structurée dans l'espace par rapport à une simple somme de couches à partir du fond. Celle-ci par contre, en plus d'être structurée spatialement et corrélée aux captures, est robuste. En Norvège elle se stabilise autour des 40 mètres au dessus du fond pour les six années étudiées.

L'analyse canonique montre qu'en général une seule espèce se dégage des facteurs. Il semble donc suffisant de travailler par espèce ou groupes d'espèces plutôt que de conserver toutes les variables de pêche. Par ailleurs l'analyse canonique s'est avérée, comme toutes les méthodes linéaires, très sensible aux valeurs fortes. Il aurait été possible pour limiter cet impact et de permettre de dégager

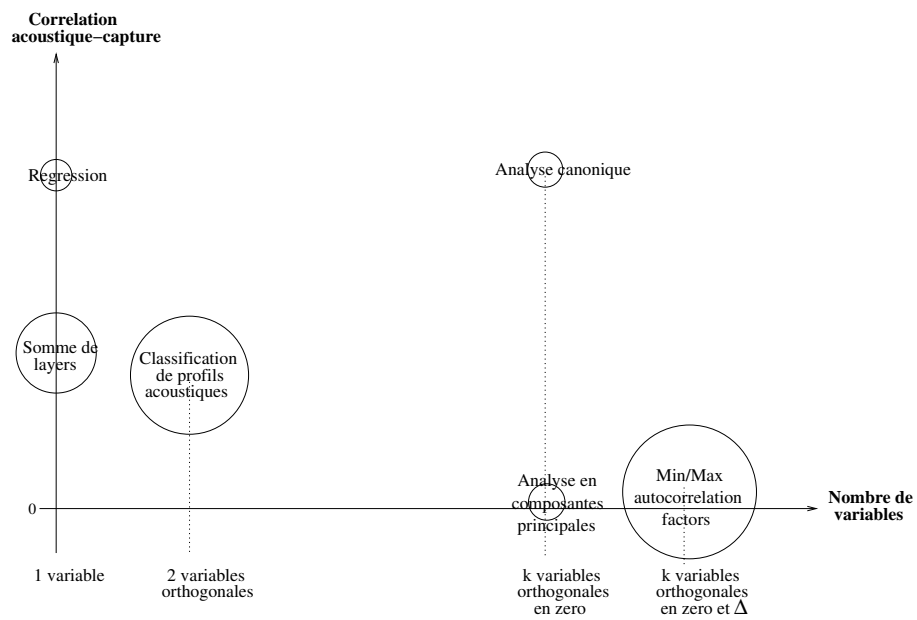


FIG. 3.31 – Schéma classant les différentes approches selon trois critères : la corrélation acoustique-capture obtenue et le nombre de variables après transformation. Les cercles représentent le niveau de structure spatiale (un moins la proportion d'effet de pépité) des nouvelles variables.

des liens non-linéaires d'effectuer une analyse factorielle multiple. Mais les facteurs obtenus étant des combinaisons linéaires de modalités d'indicatrices sur des variables différentes, il s'agit d'une méthode plus descriptive qu'opérationnelle pour la modélisation.

Les méthodes d'orthogonalisation comme l'ACP ou les MAF se sont avérées assez prometteuses pour l'élaboration d'un modèle acoustique mais finalement ici peu corrélés aux captures.

Une méthode du même type que l'analyse canonique, permettant de maximiser la corrélation avec des captures, dont les facteurs seraient spatialement orthogonaux, comme les MAF, permettrait une modélisation intéressante.

Enfin, la méthode par classification de profils verticaux permettant de ne conserver que deux variables indépendantes : la forme et l'intensité, semble prometteuse bien que limitée par l'échelle à laquelle sont disponibles les données de cette étude.

Finalement, la somme des 40 premiers mètres au-dessus du fond en mer de Barents, la somme des 4 premiers mètres en mer d'Irlande ont été les variables choisies pour établir le modèle combiné acoustique-capture présenté dans le chapitre précédent. Aucune hauteur stable n'ayant pu être dégagée en mer du Nord, c'est la hauteur d'ouverture du filet qui a été choisie, soit 5 mètres.

On se retrouve alors dans une problématique bivariante hétérotopique, avec des variables très asymétriques, telle qu'elle a été présentée en chapitre 2.

Chapitre 4

Analyse comparative des résultats obtenus par différents modèles dans chacune des trois mers

Il n'y a pas de problèmes résolus, il n'y a que des problèmes plus ou moins résolus
Henri Poincaré

Ce dernier chapitre vise à discuter l'amélioration apportée par l'information acoustique dans les différents modèles bivariés étudiés précédemment. Pour cela, les estimations obtenues par des modèles utilisant l'information acoustique sont comparées à celles de modèles plus simples utilisant uniquement les données de chalutages. Après avoir présenté certaines cohérences nécessaires à la comparaison, la précision de chaque estimation est estimée selon différents critères de variance d'estimation et de validation croisée.

Les résultats obtenus sont très différents selon les zones d'étude. En mer du Nord, l'apport de l'acoustique dans les estimations d'abondance est souvent quasi-nulle, alors que le gain est important en mer de Barents. Dans la deuxième partie du chapitre, l'impact des deux plans d'échantillonnage (aléatoire stratifié en mer du Nord et régulier en mer de Barents) est mis en parallèle avec les résultats observés. Plusieurs résultats présentés dans cette partie ont fait l'objet d'une communication à un congrès (cf. Bouleau et al. 2004, [11]).

The aim of the last chapter is to discuss the improvement brought by the acoustic information in the different bivariate models presented earlier.

For that, the outputs of each model using both acoustic and catch, are compared to the outputs of models using only the catch information. After an introduction about some assumptions needed to compare the models, the accuracy of each model is assessed using different criteria of estimation variance and cross validation.

The results obtained vary a lot from one sampled area to another. In the North Sea, the additional acoustic information does not improve the estimations, while the enhancement is very important in the Barents Sea. In the second part of the chapter, the impact of the sampling design (random stratified in the North Sea and regular in the Barents Sea) and the results observed are compared. Most of the results have been presented at a conference (cf Bouleau and al. 2004, [11]).

4.1 Analyse du gain apporté par l'acoustique

Ce dernier chapitre vise à quantifier l'amélioration apportée par l'acoustique aux estimations d'abondance, en particulier en terme de réduction de la variance de l'estimation. Pour cela les modèles bivariés combinant acoustique et chalutage ont été comparés à des modèles n'utilisant que l'information de capture.

Il est assez délicat de comparer deux modèles différents. Une multitude de paramètres peuvent rendre un modèle plus attractif qu'un autre, plus opératoire, plus performant. Un modèle peut être satisfaisant parce qu'il fournit des estimations précises, parce qu'il met en relation différentes variables et permet ainsi de comprendre un phénomène physique ou biologique, parce qu'il est robuste d'années en années et s'adapte à différentes configurations de données, etc. Il y a beaucoup de critères possibles pour son évaluation aussi bien statistiques que qualitatifs, voire subjectifs.

Cette partie se contente d'évaluer l'apport de l'acoustique sur des critères statistiques classiques : comparaison des cartes d'estimation, étude du gain en variance et de la capacité du modèle à réestimer une valeur connue (validation croisée).

4.1.1 Cohérences nécessaires à la comparaison de modèles

Les résultats de deux modèles différents sont intrinsèquement liés aux hypothèses choisies. Le cokrigeage, par exemple, fournit une variance d'estimation nécessairement inférieure à celle d'un krigeage ordinaire. Par construction, si la variable auxiliaire ne permet aucune réduction de variance, celle-ci se trouve affectée d'un poids nul.

Pour mesurer le gain en variance apporté par la variable auxiliaire, il convient de comparer les résultats du cokrigeage et du krigeage des captures obtenues avec le même modèle de variogramme pour les captures, ou la comparaison n'aurait pas de sens. Si les modèles de variogrammes (ou de covariances) ne sont pas cohérents, les variances d'estimation ne sont plus comparables.

Dans la plupart des cas naturellement, le modèle de variogramme utilisé pour les captures dans le modèle bivariable est optimal. C'est exactement le modèle qui aurait été choisi si aucune connaissance supplémentaire n'avait été disponible. Ce n'est cependant pas toujours le cas. Certaines hypothèses de simplification de modèles multivariés, comme le choix d'un modèle linéaire de corégionalisation, peut mener à choisir un modèle de variogramme légèrement différent pour les captures. Dans le modèle à résidu (voir partie 2.4), la contrainte de cohérence entre les variogrammes entraîne que le modèle choisi pour les captures n'est pas parfaitement ajusté au variogramme empirique de ces captures. En effet, dans ce modèle, l'estimation du cokrigeage ne nécessite pas directement l'inférence du variogramme des captures, mais uniquement celle de la variable acoustique et d'un résidu. Le modèle des captures s'obtient par la relation suivante :

$$\gamma_T(h) = \alpha^2 \cdot \gamma_A(h) + \gamma_\epsilon(h)$$

Il est alors possible que pour des raisons de variabilité statistique (lors de l'inférence de chacun des deux variogrammes et lors de l'estimation du paramètre multiplicatif, voir partie 2.4.2), le modèle de capture sous-jacent ne s'ajuste pas aussi bien au variogramme empirique des données de captures

que le serait un modèle calé indépendamment de toute contrainte de cohérence. Une fois de plus, cela signifie que l'autokrigeabilité de l'acoustique n'est pas pleinement vérifiée. Le gain en variance dû à l'acoustique est alors abusif, puisque si seules les données de captures étaient disponibles ce serait probablement un autre modèle que aurait été choisi. Une grande attention a été portée dans le second chapitre sur le contrôle des hypothèses du modèle à résidu et la qualité de l'estimation du paramètre multiplicatif. Dans la pratique, le modèle de variogramme sous-jacent obtenu s'ajuste heureusement bien au variogramme empirique. Il est important de garder à l'esprit qu'analyser l'amélioration apportée par l'information acoustique n'a de sens que si les hypothèses du modèle sont vraies et que le modèle choisi est porté de façon robuste par les données. Naturellement, plus le modèle choisi est simple, plus il est facile de contrôler l'ensemble de ces facteurs. Malgré tout, comme cela a été vu dans la partie 2.4.5 du chapitre 2, les résultats obtenus par modèle à résidu sont très semblables à ceux obtenus par un cokrigeage classique, utilisant la structure croisée acoustique-capture, mais les tests de fiabilité obtenus par validation croisée sont meilleurs pour le modèle simplifié.

On considère donc dans les résultats présentés ici que le modèle des captures est optimal, que le même modèle aurait été choisi si aucune information acoustique n'avait été disponible.

Le recalage de la variance de l'acoustique en radiales au niveau de celle des stations, vue dans le cadre du modèle à résidu (partie 2.4.1), est également une cohérence qui n'est nécessaire que pour la comparaison de deux modèles. Comme cela a été discuté, la variance de l'acoustique observée sur l'ensemble des observations en radiales est plus réaliste, plus fiable que la variance observée sur les seules données en stations. Mais celle-ci n'est pas comparable avec celle d'un modèle utilisant 40 fois moins de données.

Les résultats obtenus par des modèles bivariés ont ainsi été comparés avec ceux de modèles monovariés cohérents du point de vue des variogrammes et des variances.

4.1.2 Récapitulation des modèles utilisés

Deux types de modèles combinant acoustique et captures ont été utilisés : le cokrigeage (noté de façon abrégée : CoK) et le krigeage des captures avec une interpolation de l'acoustique en dérive externe, celle-ci estimée par krigeage ordinaire (le modèle est alors noté K.D.E) ou de façon non linéaire (noté K.D.E non lin.). L'estimation non-linéaire de l'acoustique n'a été effectuée que lorsque que la déstructuration des valeurs fortes était trop importante pour permettre une modélisation directe. Pour chaque campagne, un krigeage ordinaire des captures (K.O) a également été effectué pour être comparé avec les modèles bivariés. Lorsque les données de captures étaient trop bruitées pour pouvoir inférer un modèle, n'étant alors généralement pas assez nombreuses pour permettre d'établir un modèle disjonctif robuste, une simple estimation par régression linéaire en situation de sous-échantillonnage (cf. partie 2.1) a été réalisée et comparée avec la moyenne des données de captures.

En mer de Barents, comme cela a été vu en partie 2.4, les données correspondaient à un modèle de cokrigeage dans lequel l'acoustique est autokrigeable (le modèle à résidu, noté M.R) pour quatre des six années disponibles. Les modèles de variogrammes utilisés pour ces campagnes sont indiqués dans le graphique 4.1. Pour les deux autres années (1999 et 2002), la structure croisée

entre les deux variables était trop bruitée pour être modélisée, mais un krigeage avec la variable acoustique en dérive externe a tout de même été appliqué (cf. tableau 4.1).

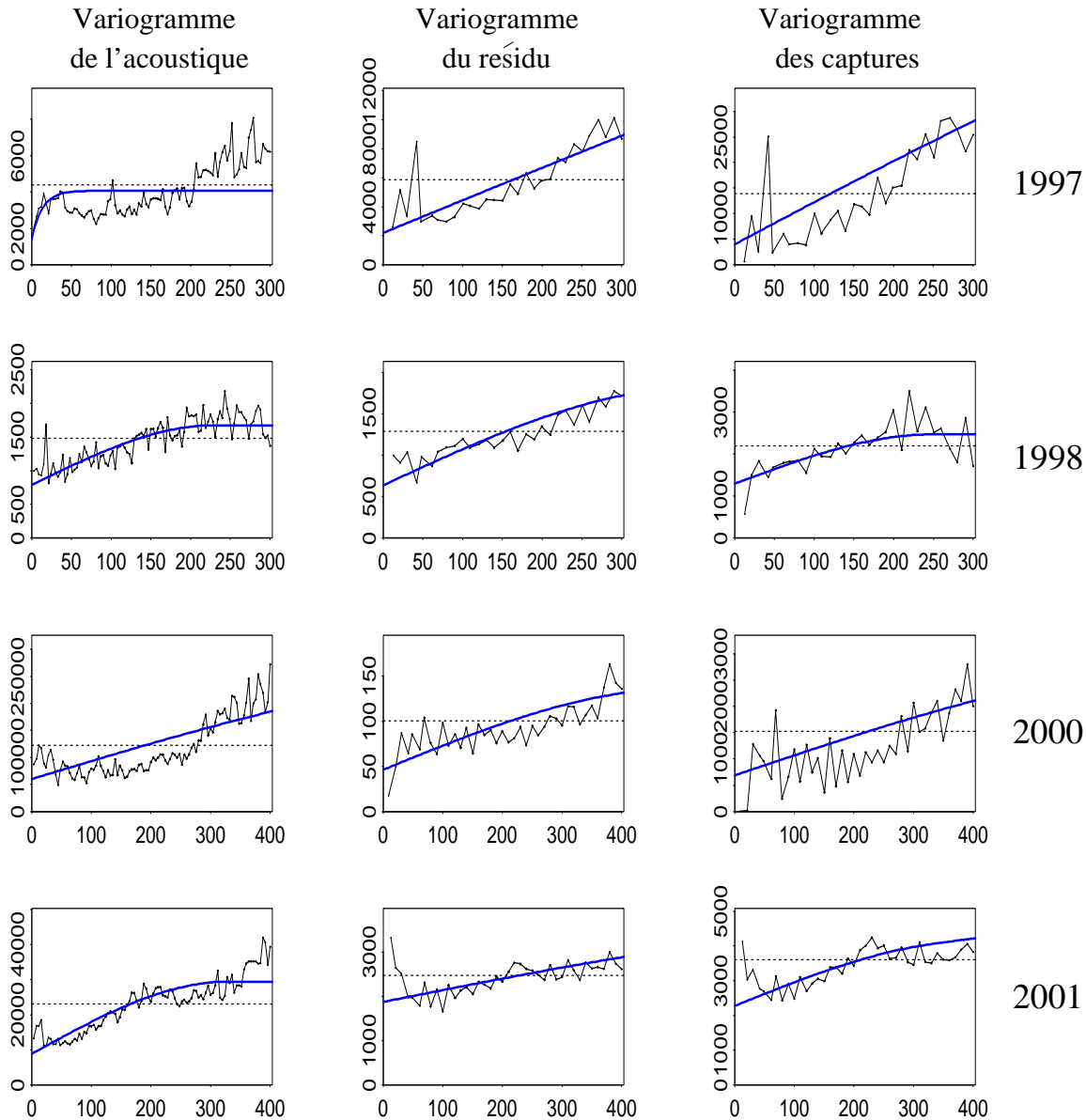


FIG. 4.1 – Modèles de variogrammes utilisés dans le modèles à résidu pour l'acoustique (à gauche), le résidu(au centre) et le modèle utilisé pour le krigeage ordinaire des captures correspondant (à droite), en Norvège, 1997, 1998, 2000 et 2001.

En mer du Nord, les campagnes écossaises 1999, 2002 et 2003 répondaient à des modèles de co-krigeage (modèle à résidu) si quelques valeurs fortes étaient mises de côté (2 données en radiales et une station pour la campagne 1999, 4 données en radiales et 2 stations en 2002, 2 données en radiales

et 2 stations en 2003). Par contre l'ensemble des données a été utilisé pour une estimation avec un krigeage disjonctif de l'acoustique en dérive externe, mais sur les variables de captures mesurées en nombre et non en ENASC (la déstructuration par les valeurs fortes étant moins importante).

Les enregistrements acoustiques des campagnes anglaises étaient de structure trop pépitiqes pour être utilisés tels quels. L'acoustique a été découpée en indicatrices et estimée par krigeage disjonctif (cf. partie 2.7). Cette estimation a ensuite été utilisée en dérive externe pour le krigeage des captures.

En Irlande, enfin, les années 2000 et 2002 répondaient à un cokrigeage (modèle linéaire de corégionalisation à deux structures, cf. partie 2.3). Mais comme en Écosse, l'émergence de structure dans les données n'a été possible qu'en supprimant entre une et quatre valeurs fortes en stations.

Pour l'ensemble des campagnes, une simple régression point à point extrapolée aux radiales a été utilisée pour fournir un indice combinant acoustique et chalutage. Celle-ci ne nécessitant pas la modélisation d'une structure spatiale puisqu'elle suppose les points indépendants, elle a permis de fournir une estimation comparable pour l'intégralité des campagnes disponibles.

Lorsqu'une sélection a été effectuée, certaines données ont été complètement supprimées de la modélisation. En Norvège, comme dans tout ce mémoire, les données où la température était inférieure à zéro ont ainsi été mises de côté (cf. A). En mer du Nord et en mer d'Irlande, si certaines valeurs ont été supprimées pour mettre en place un modèle de cokrigeage, toutes les données ont été prises en compte dans les krigeages avec estimation non-linéaire de la dérive.

L'ensemble des modèles utilisés pour chacune des campagnes est synthétisé dans le tableau 4.1, page 138.

4.1.3 Résultats

Les cartes d'estimation et de variance

De façon générale, les cartes d'abondance des modèles combinés sont moins lisses que les cartes d'estimation des modèles n'utilisant que les données de captures. Bien que les "pattern" régionaux soient les mêmes, l'information disponible en radiales permet de prendre en compte des zones de richesse locale, invisibles à partir des seules données en stations.

En mer de Barents, les cartes d'estimation obtenues par modèle à résidu présentent en effet nettement plus de détails que les cartes obtenues par le modèle de capture. Les reliefs sont du même ordre de grandeur et globalement similaires, mais de petites zones d'abondances locales émergent entre les stations grâce au modèle combiné. En 1997, par exemple (cf. figure 4.2) pour chaque estimation la région sud-ouest est la plus riche, mais les lignes de niveau du modèle n'utilisant que les données de captures sont extrêmement lisses et les reliefs sont moins contrastés que le modèle de cokrigeage. En 1998 les deux approches fournissent des cartes assez semblables. En 2000, l'acoustique permet de réduire l'influence d'une valeur forte isolée dans le sud-est en délimitant de façon plus précise la zone riche. En 2001, l'information acoustique permet de dégager toute une région riche à l'est où les valeurs fortes se situent entre les stations, alors que la carte obtenue à partir des seules données

Institut	Année	Variable cible	Sélection	Modèle mono- variable de capture	Modèles bivariables acoustique-captures		
				Krigeage ordinaire	Cokrigeage modèle à résidu	Krigeage à dérive externe	Krigeage à dérive externe (non-linéaire)
IMR	1997	Dem.(enasc)	temp.> 0	K.O.	M.R.	K.D.E.	-
IMR	1998	Dem.(enasc)	temp.> 0	K.O.	M.R.	K.D.E.	-
IMR	1999	Dem.(enasc)	temp.> 0	K.O.	-	K.D.E.	-
IMR	2000	Pel.(enasc)	temp.> 0	K.O.	M.R.	K.D.E.	-
IMR	2001	Dem.(enasc)	temp.> 0	K.O.	M.R.	K.D.E.	-
IMR	2002	Dem.(enasc)	temp.> 0	K.O.	-	K.D.E.	-
FRS	1999	Pel.(enasc)	2 Rad. +1 St.	K.O.	M.R.	K.D.E.	-
FRS	2000	-	-	-	-	-	-
FRS	2002	Dem.(enasc)	4 Rad. + 1 St.	K.O.	M.R.	K.D.E.	-
		Dem.(num)	-	K.O.	-	-	K.D.E non lin.
FRS	2003	Dem.(enasc)	2 Rad. + 2 St.	K.O.	M.R.	K.D.E.	-
		Dem.(num)	-	K.O.	-	-	K.D.E non lin.
CEFAS	2000	-	-	-	-	-	-
CEFAS	2001	Dem.(enasc)	-	K.O.	-	-	K.D.E non lin.
CEFAS	2002	Dem.(enasc)	-	K.O.	-	-	K.D.E non lin.
IFREMER	2002	-	-	-	-	-	-
IFREMER	2003	-	-	-	-	-	-
QUB	2000	Pel.(enasc)	1 St.	K.O.	CoK	K.D.E.	-
		Pel.(enasc)	-	K.O.	-	-	K.D.E non lin.
QUB	2001	-	-	-	-	-	-
QUB	2002	Pel.(enasc)	1 St.	K.O.	CoK	K.D.E.	-
		Pel.(enasc)	-	K.O.	-	-	K.D.E non lin.

TAB. 4.1 – Tableaux récapitulatif des modèles utilisés pour chacune des campagnes disponibles

de captures est quasiment plate.

Le modèle à dérive externe (K.D.E), assez semblable en théorie au modèle à résidu mis à part le fait que les coefficients α et β peuvent changer avec le voisinage (voir chapitre 2.6) fournit des estimations qui semblent assez instables. Les estimations passent en effet de valeurs très fortes à des valeurs largement négatives en peu de distance, comme c'est le cas par exemple en 1997 (figure 4.2). En 1998 les estimations du modèle à dérive externe descendent même ponctuellement en dessous de -100, alors que sur le reste du champ les valeurs sont très proches de celles fournies par les deux autres modèles. De façon générale, les estimations obtenues dans les zones pauvres sont rapidement négatives alors que celles obtenues dans les zones riches sont souvent largement supérieures à celles fournies par les deux autres modèles. Pour la campagne réalisée en 2000 (figure 4.3), par exemple, les estimations montent jusqu'à 1500 alors que la valeur la plus forte par le modèle à résidu est de 200 et de 120 par krigeage ordinaire des captures. Ce contraste très fort peut paraître plus réaliste puisque il semble moins lisser les données que le krigeage et le cokrigeage, mais cela peut également être dû à une grande instabilité du modèle qui est très sensible aux valeurs extrêmes et peut fournir des estimations aberrantes.

Par construction, le modèle à dérive externe fournit une variance d'estimation plus grande qu'un krigeage ordinaire, en raison de la contrainte supplémentaire imposées aux poids de krigeage. De la même façon, la variance de cokrigeage est nécessairement inférieure à la variance de krigeage. La comparaison des variances d'estimation obtenues par chacun des trois modèles est de ce point de vue sans surprise. Mais il est tout de même intéressant de connaître l'ordre de grandeur de l'écart des variances obtenu par les différentes approches. Les résultats sont montrés dans la figure 4.4. La variance de cokrigeage fait apparaître les transects qui sont de variance inférieure. En effet, la variance de krigeage de l'acoustique y étant nulle, la variance de cokrigeage est égale à la variance du résidu. Le cokrigeage permet en moyenne une baisse de variance de 25% en 1997, de 30% en 1998, 17% en 2000 et de 6% en 2001 par rapport à la variance moyenne de krigeage.

En Écosse, les données sont moins bien structurées qu'en mer de Barents et un seuillage de quelques valeurs a été nécessaire pour faire émerger une structure dans les données acoustiques brutes. Les quelques valeurs les plus fortes ne sont donc pas présentes dans l'estimation. En 1999 (cf. figure 4.5) les captures étant extrêmement bruitées, seul le cokrigeage combinant acoustique et capture montre du relief, le krigeage à dérive externe et le krigeage des captures sont quasiment constants sur l'ensemble du champ. Seule l'acoustique plus structurée en raison de sa plus haute définition permet d'obtenir un contour des zones riches. En 2002, Le modèle à résidu localise une zone riche ($\sim 250 m^2/m.n^2$) près des îles Shetland alors que par le modèle à dérive externe et le krigeage ordinaire des captures, la zone la plus riche se situe plus au sud et vaut autour de $120 m^2/m.n^2$. En 2003 le modèle à résidu fournit une carte d'estimation très semblable à celle obtenue à partir des captures. Le gain en variance obtenu par cokrigeage dans le modèle combiné est minime (cf. figure 4.6). La variance d'estimation du krigeage à dérive externe est également très proche de celles des deux autres modèles.

Pour les campagnes 2002 et 2003, une estimation par krigeage disjonctif de l'acoustique a été réalisée à partir de l'ensemble des données disponibles (sans sélection) et utilisée en dérive externe d'un krigeage des captures démersales mesurées en nombre (modèle K.D.E non lin.). Celles-ci semblaient en effet mieux structurées que les mesures converties en ENASC et permettaient ainsi d'éviter le seuillage. En 2002, les résultats obtenus sont les mêmes que ceux d'un simple krigeage des captures, tandis que des reliefs diffèrent en 2003, en particulier dans la partie nord-ouest de la campagne (figure 4.7). La perte de variance due à la dérive externe est quasiment nulle dans chacune des campagnes.

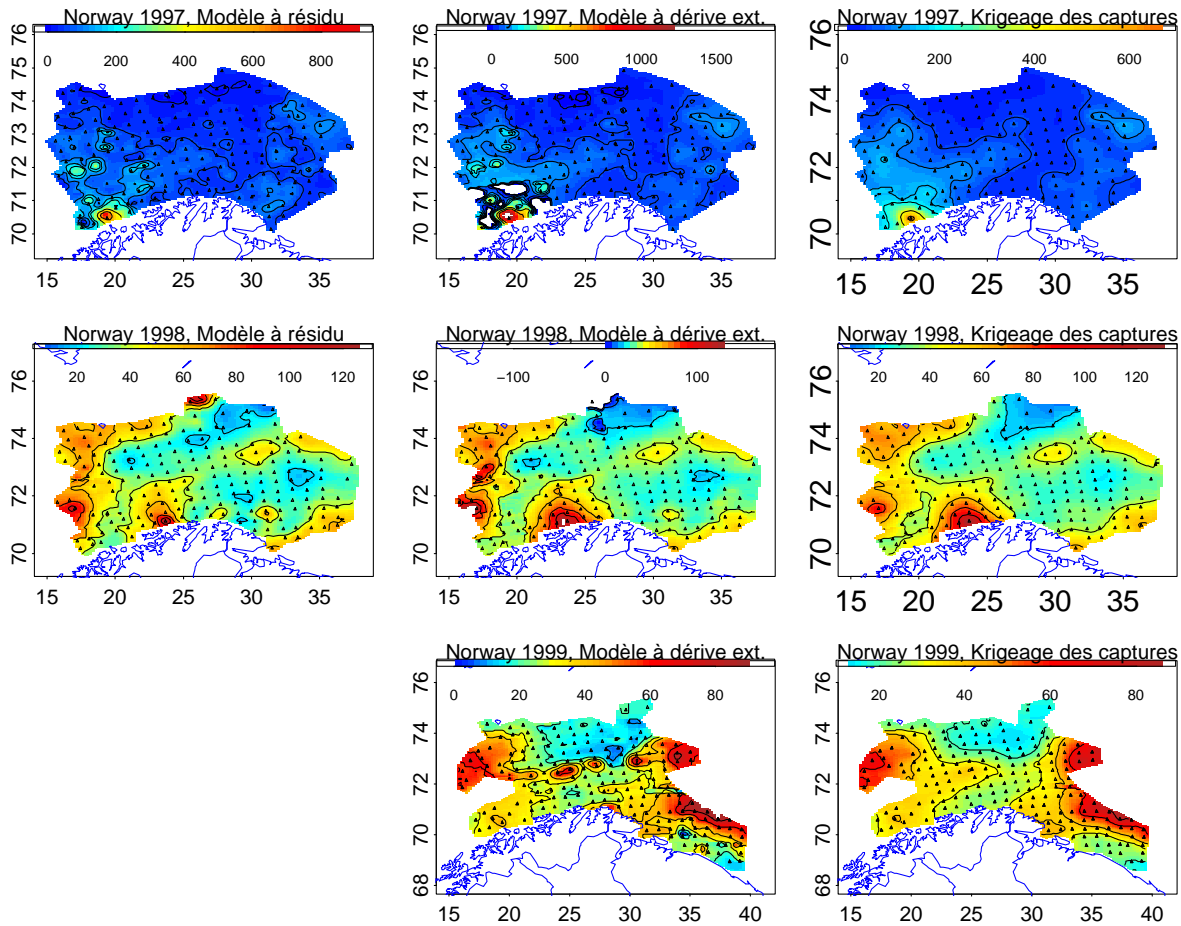


FIG. 4.2 – Cartes d'estimation obtenues par cokrigage (modèle à résidu) sur la colonne de gauche, par krigage à dérive externe (colonne du milieu) et par un krigage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour chaque campagne de mer de Barents de 1997 (ligne du haut) à 1999 (ligne du bas). Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 3 modèles.

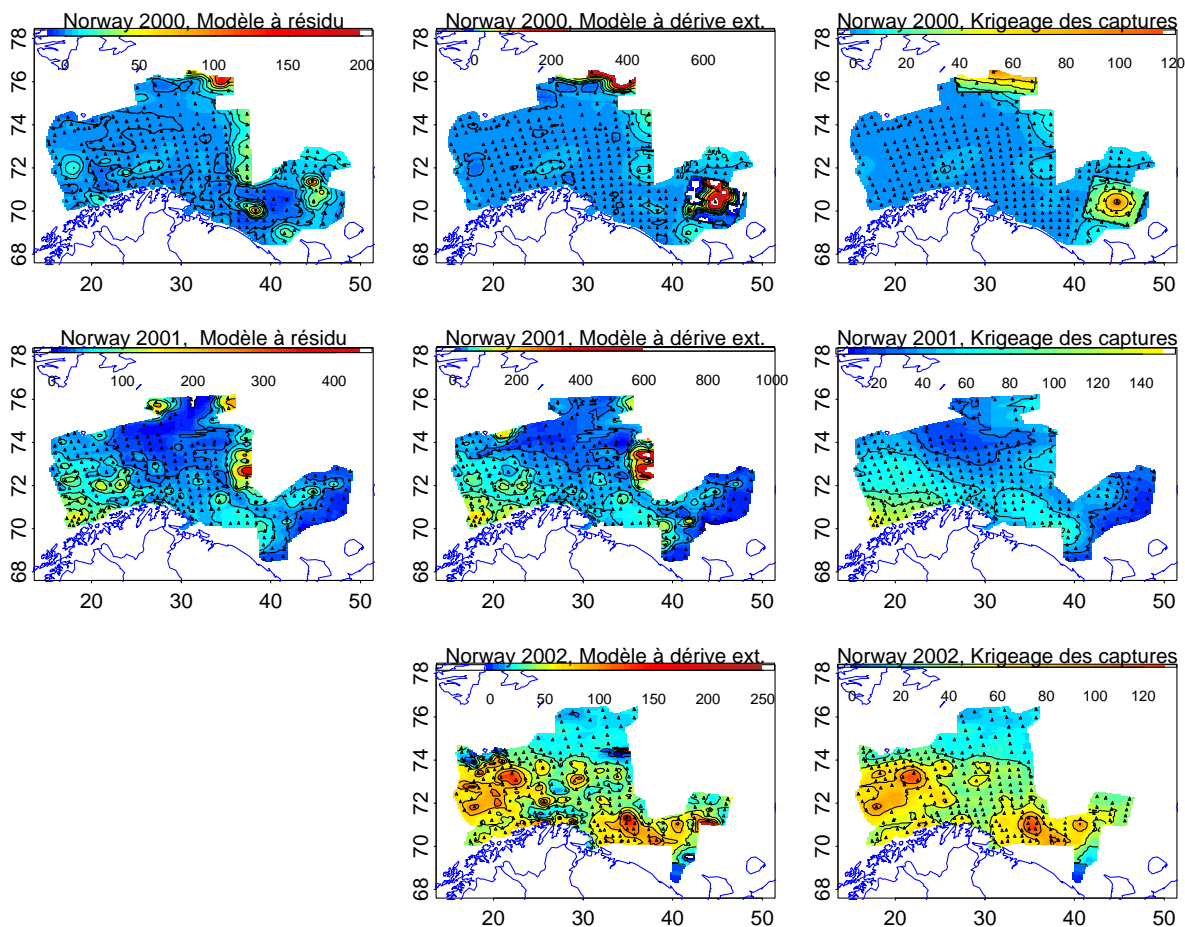


FIG. 4.3 – Cartes d'estimation obtenues par cokrigage (modèle à résidu) sur la colonne de gauche, par krigage à dérive externe (colonne du milieu) et par un krigage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour chaque campagne de mer de Barents de 2000 (ligne du haut) à 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 3 modèles.

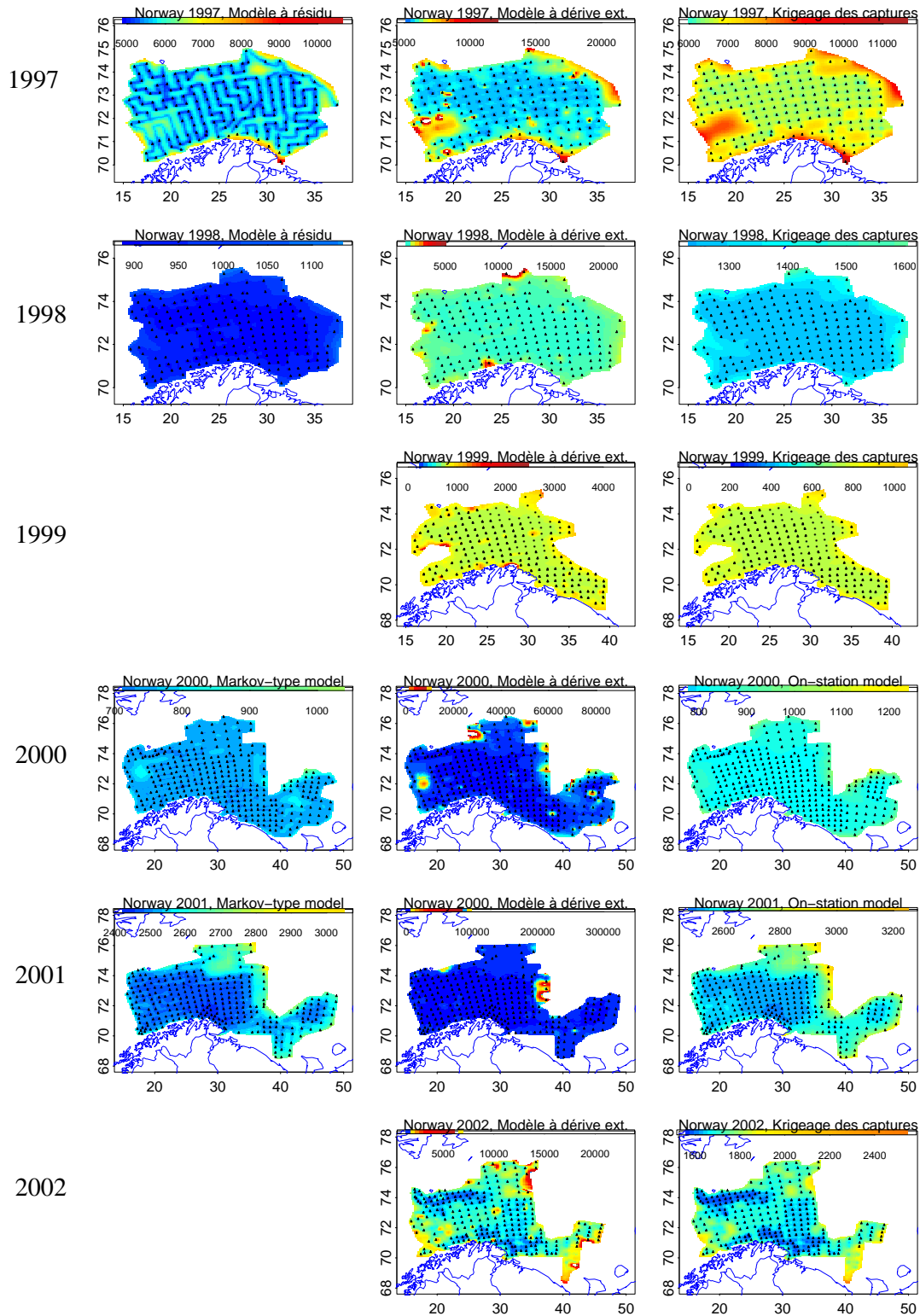


FIG. 4.4 – Cartes de variance d'estimation obtenues par cokrigeage (modèle à résidu) sur la colonne de gauche, par krigeage à dérive externe (colonne du milieu) et par un krigeage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour chaque campagne de mer de Barents de 1997 (ligne du haut) à 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur est la même pour les 3 modèles.

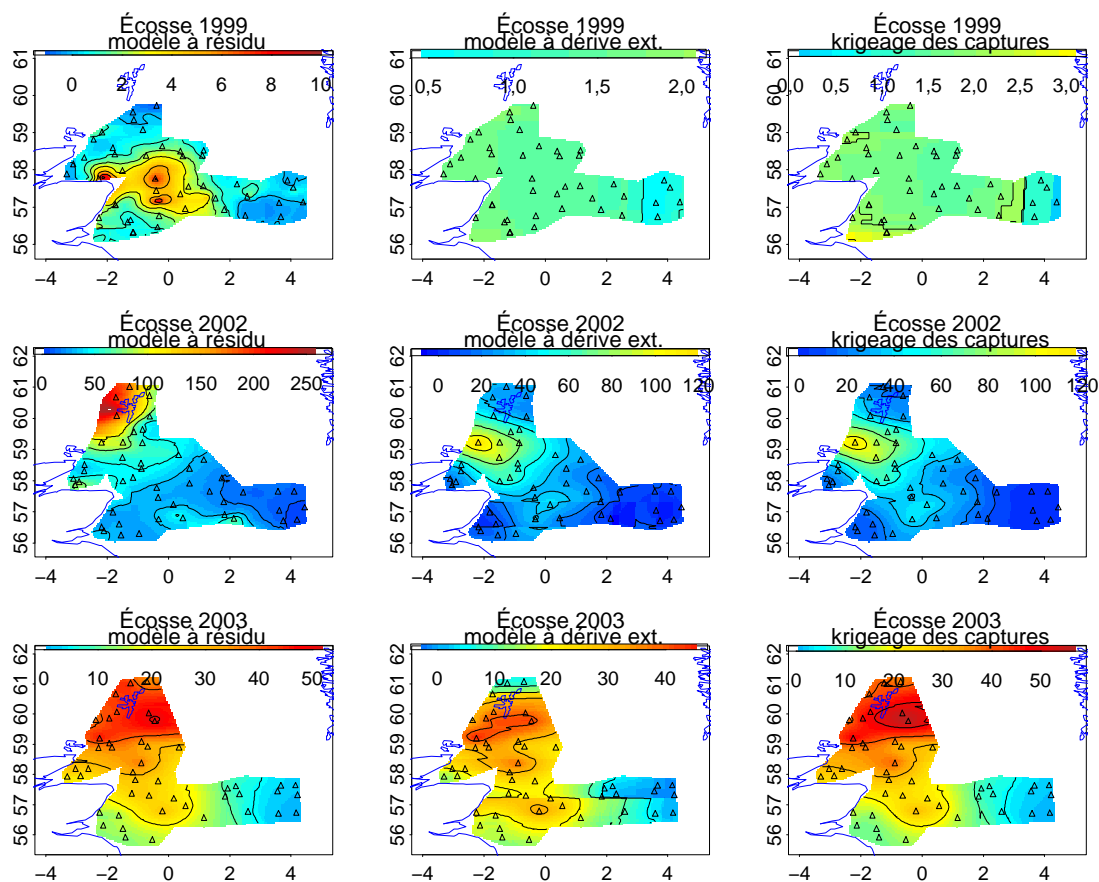


FIG. 4.5 – Cartes d'estimation obtenues par cokrigage (modèle à résidu) sur la colonne de gauche, par krigeage à dérive externe (colonne du milieu) et par un krigeage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour les campagnes écossaises 1999 (ligne du haut), 2002 (milieu) et 2003 (ligne du bas). Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 3 modèles.

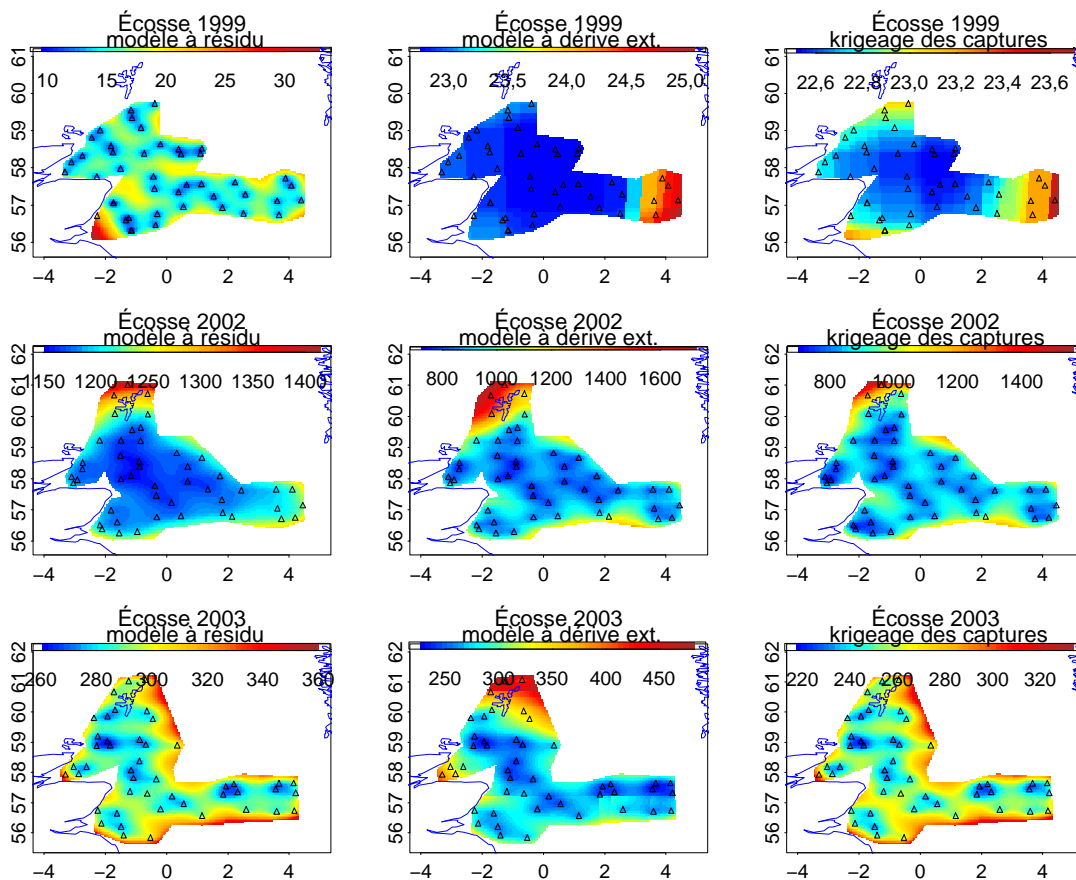


FIG. 4.6 – Cartes de variance d'estimation obtenues par cokrigage (modèle à résidu) sur la colonne de gauche, par krigeage à dérive externe (colonne du milieu) et par un krigeage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour les campagnes écossaises 1999 (ligne du haut), 2002 (milieu) et 2003 (ligne du bas). Pour chaque campagne l'échelle de couleur est la même pour les 3 modèles.

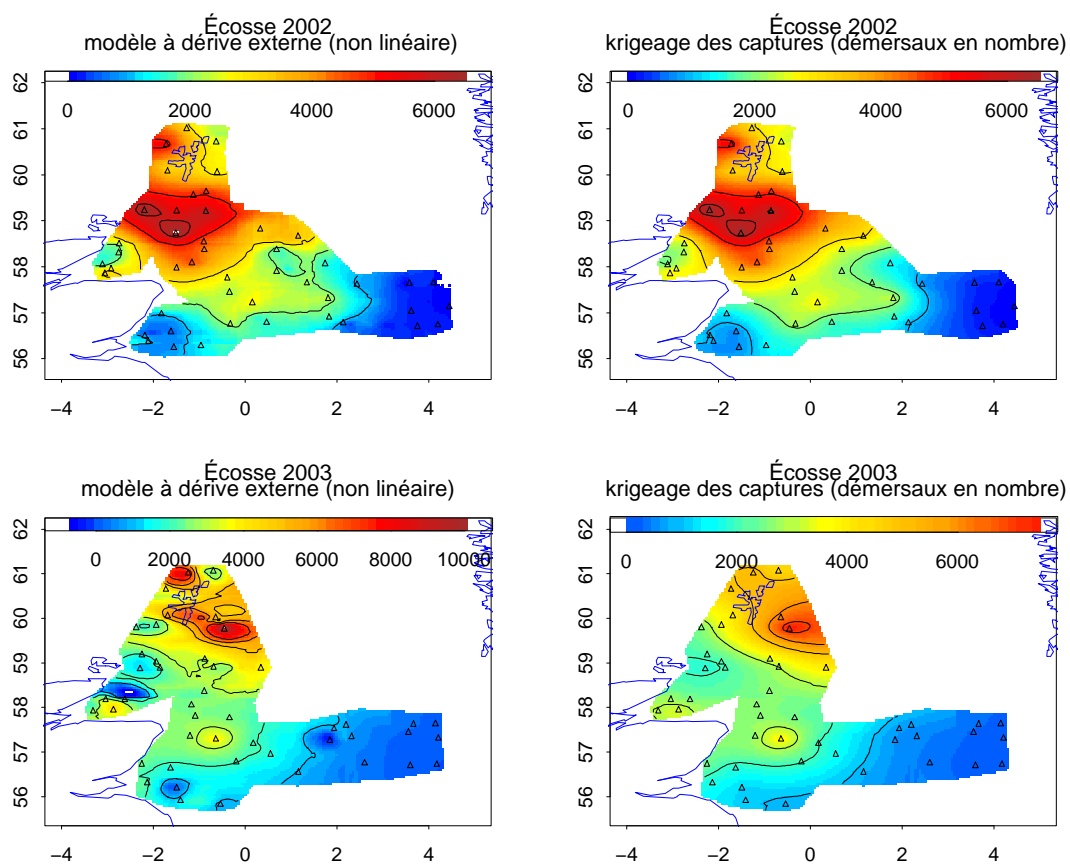


FIG. 4.7 – Cartes d'estimation obtenues par krigeage des captures utilisant une estimation **non linéaire** (krigeage disjonctif) de l'acoustique en dérive externe (colonne de gauche) et sans dérive externe (colonne de droite), pour les campagnes réalisées en Écosse en 2002 et 2003 (ligne du bas). La variable cible est le **nombre** de démersaux capturés, **sans sélection** sur les données. Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 2 modèles.

Pour les campagnes réalisées par le CEFAS, le modèle combiné (K.D.E non lin.) fournit en 2001 des résultats quasiment identiques au modèle mono-variable n'utilisant que les captures. Le poids donné à l'acoustique dans le krigeage à dérive externe est très proche de zéro. En 2002, le modèle combiné permet de délimiter une zone riche au nord-ouest (cf. figure 4.8). Mais si l'on estime le coefficient multiplicatif de la dérive acoustique dans le krigeage à dérive externe, celui-ci est souvent négatif, ce qui n'a pas de sens (cf. partie 2.6). En 2001, la perte de variance intrinsèque au krigeage à dérive externe est assez réduite. Sur la majeure partie de la zone étudiée les variances d'estimation sont assez semblables. En 2002, par contre la variance du krigeage des captures est environ $2/3$ de celle obtenue avec la dérive externe (cf. figure 4.9).

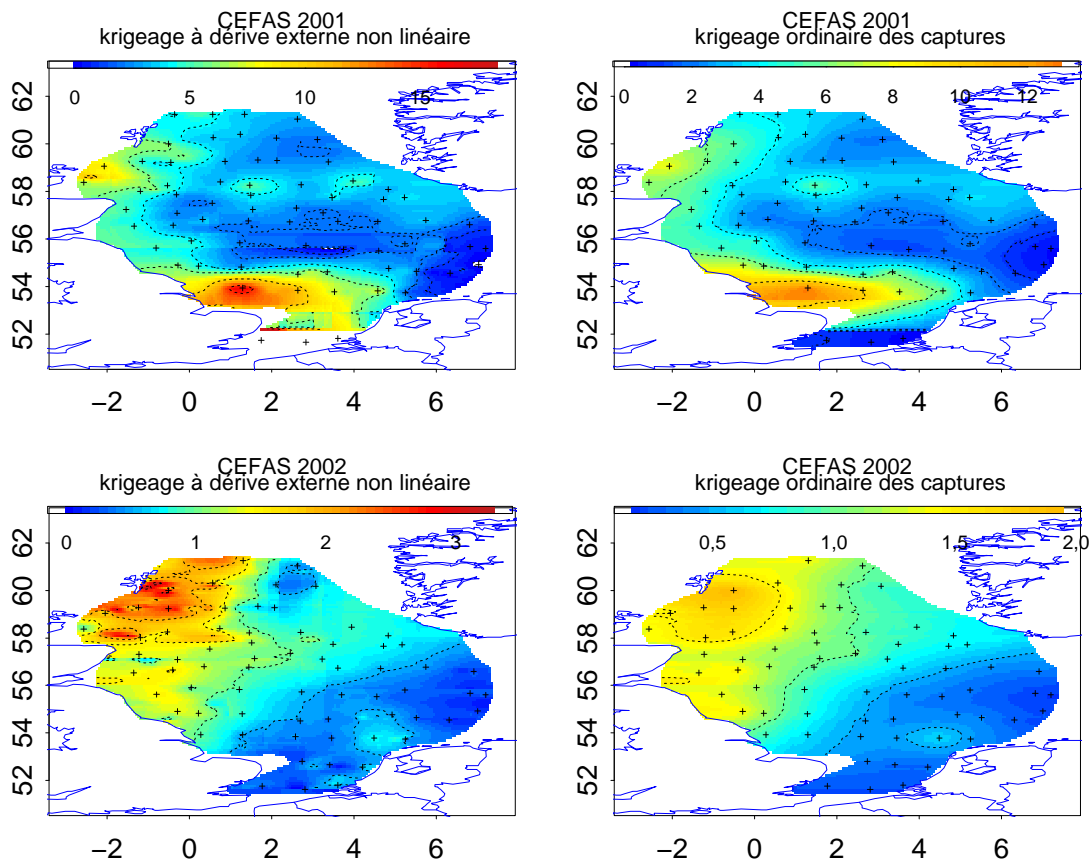


FIG. 4.8 – Cartes d'estimation obtenues par krigeage des captures utilisant une estimation non linéaire (krigeage disjonctif) de l'acoustique en dérive externe (colonne de gauche) et sans dérive externe (colonne de droite), pour les campagnes réalisées par le CEFAS en 2001 et 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 2 modèles.

En Irlande, enfin, les estimations obtenues par cokrigeage (après seuillage d'une seule donnée en station) sont extrêmement proches de celles obtenues à partir des seules données de captures. La zone d'influence de certaines valeurs fortes est parfois réduite, comme au nord-est en 2000 (cf. figure 4.10), grâce aux enregistrements effectués entre les stations. Les zones riches semblent donc

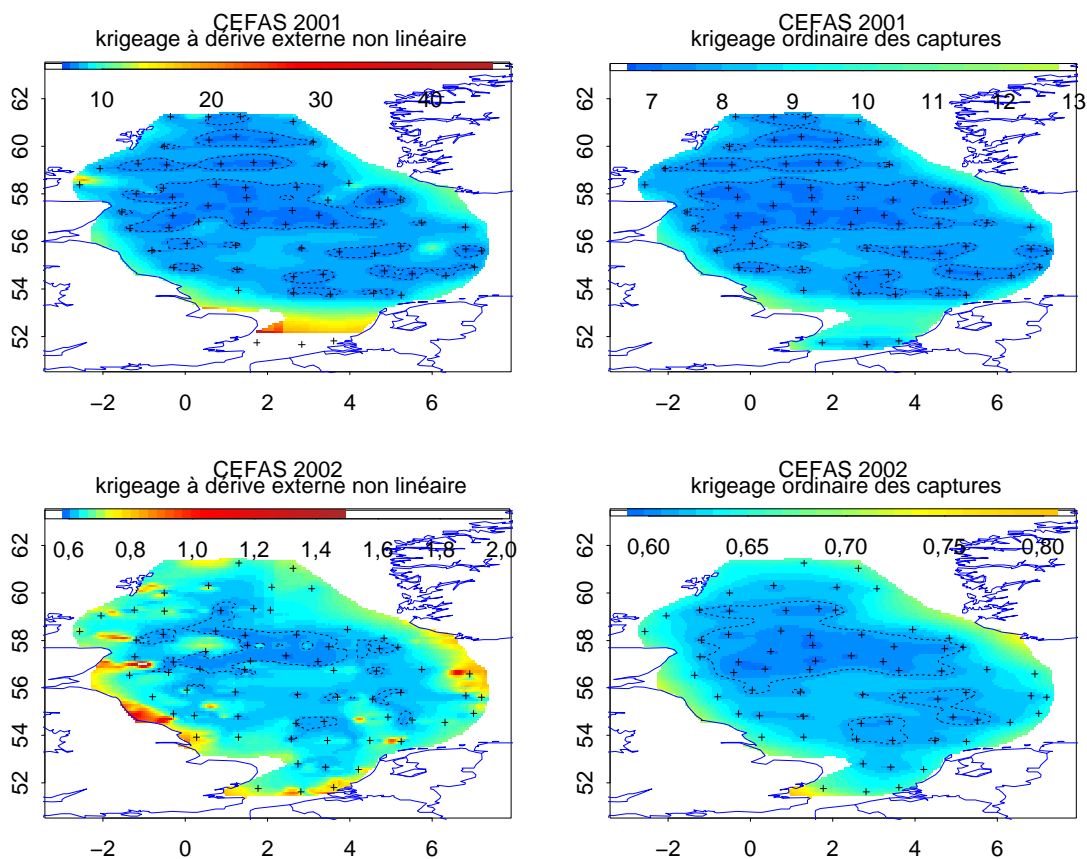


FIG. 4.9 – Cartes de variances d'estimation obtenues par krigeage des captures utilisant une estimation non linéaire (krigeage disjonctif) de l'acoustique en dérive externe (colonne de gauche) et sans dérive externe (colonne de droite), pour les campagnes réalisées par le CEFAS en 2001 et 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur est la même pour les 2 modèles.

mieux délimitées.

Le modèle combiné effectué sur l'ensemble des points (sans sélection) par krigeage disjonctif de l'acoustique offre par contre des cartes beaucoup plus contrastées qu'un modèle monovarié, en particulier en 2002 (cf. figure 4.12). Il est intéressant de voir qu'en 2002, le krigeage ordinaire des données de captures sans sélection fournit une carte d'estimation très lisse. La valeur forte, présente dans la partie nord-ouest tire l'estimation et cache tout autre relief présent dans les données. Le modèle à dérive externe (K.D.E non lin.) permet de réduire considérablement l'influence de cette valeur et d'obtenir un relief très semblable à celui obtenu par cokrigeage sur les données seuillées de cette valeur forte. Par contre les cartes de variance obtenues sur l'ensemble des données sont beaucoup plus fortes que celles obtenues par les modèles sur données seuillées (cf. figures 4.11 et 4.13). Elles varient autour de 4000 avant seuillage et autour de 10 après. Une fois de plus les transects se dessinent assez nettement sur les cartes de variance obtenues par cokrigeage, en particulier en 2000.

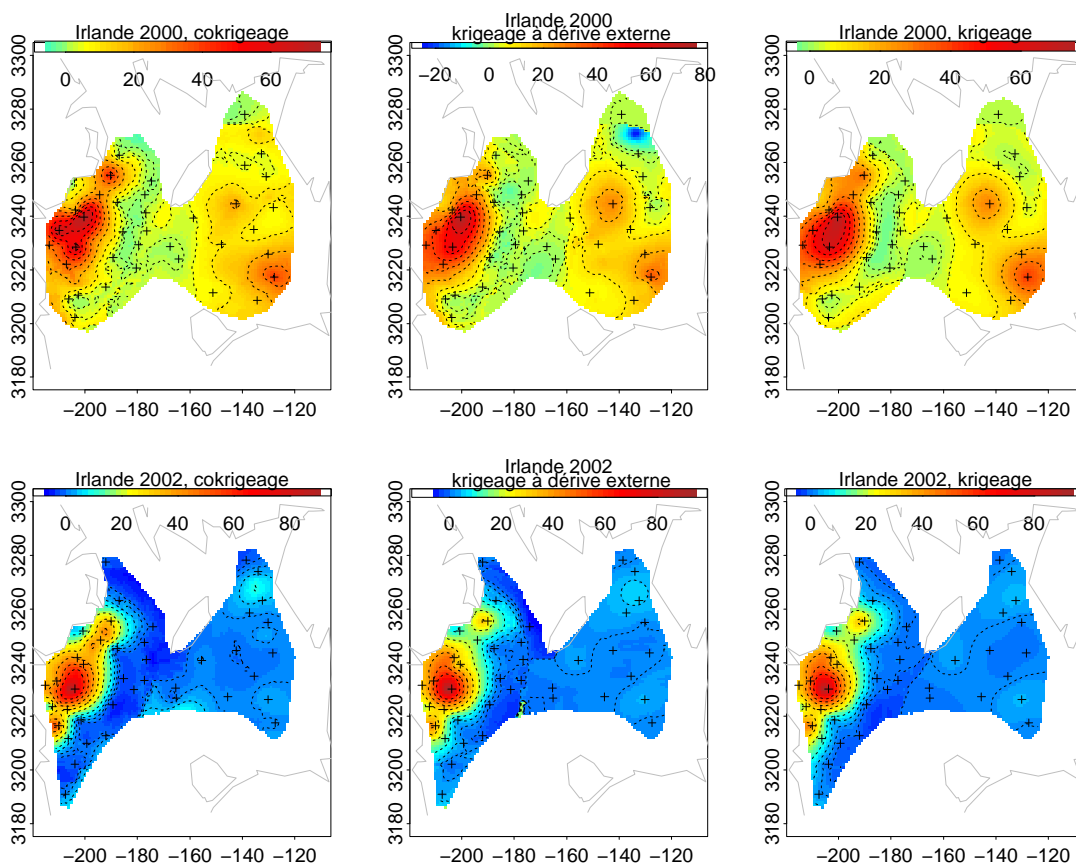


FIG. 4.10 – Cartes d'estimation obtenues par cokrigeage (modèle à résidu) sur la colonne de gauche, par krigeage à dérive externe (colonne du milieu) et par un krigeage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour les campagnes irlandaises 2000 (ligne du haut) et 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 3 modèles.

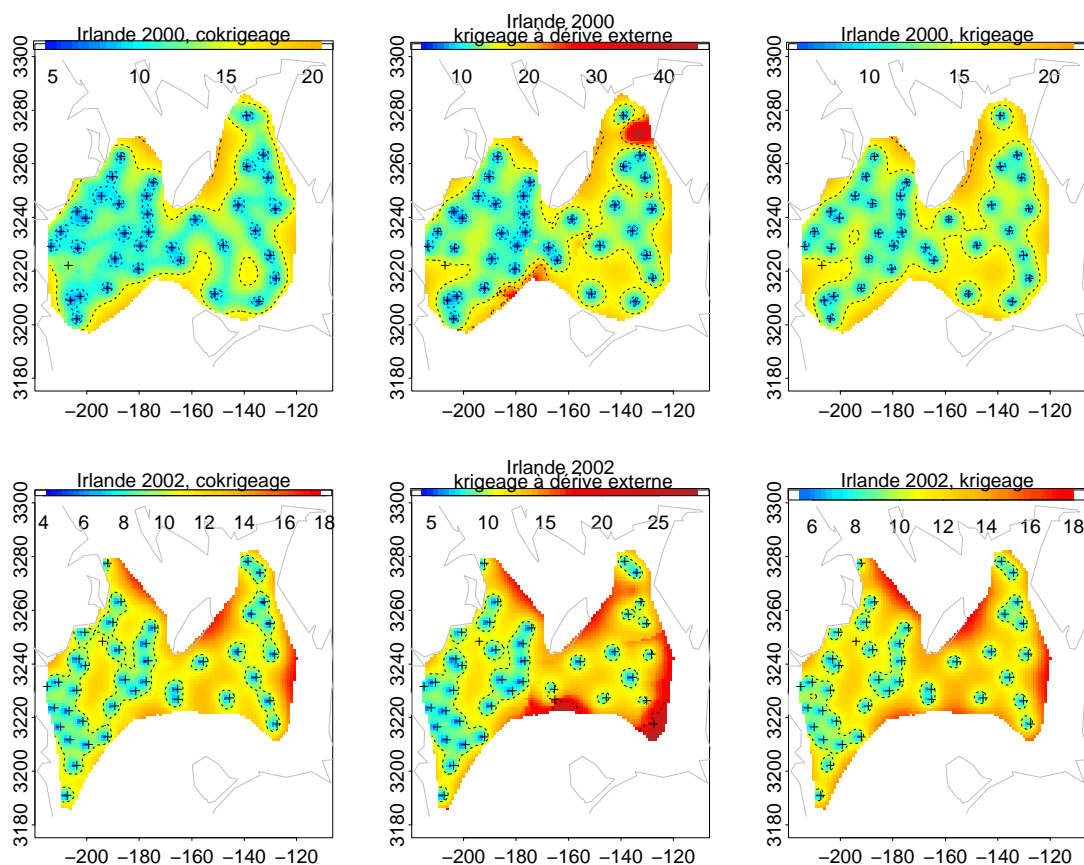


FIG. 4.11 – Cartes de variances d'estimation obtenues par cokrigage (modèle à résidu) sur la colonne de gauche, par krigeage à dérive externe (colonne du milieu) et par un krigeage ordinaire des données de captures n'utilisant pas l'acoustique (colonne de droite), pour les campagnes irlandaises 2000 (ligne du haut) et 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur est la même pour les 3 modèles.

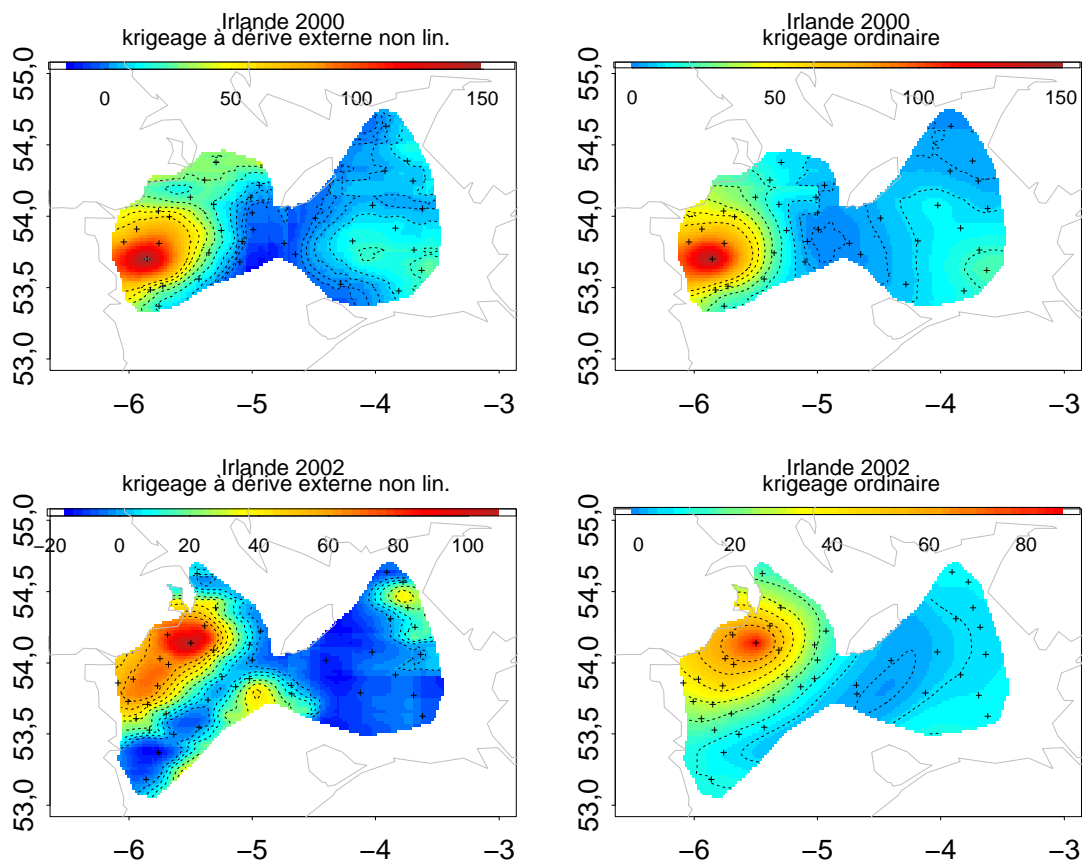


FIG. 4.12 – Cartes d'estimation obtenues par krigeage des captures utilisant une estimation non linéaire (krigeage disjonctif) de l'acoustique en dérive externe (colonne de gauche) et sans dérive externe (colonne de droite), pour les campagnes réalisées par l'Irlande en 2000 et 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur et les lignes de niveaux sont les mêmes pour les 2 modèles.

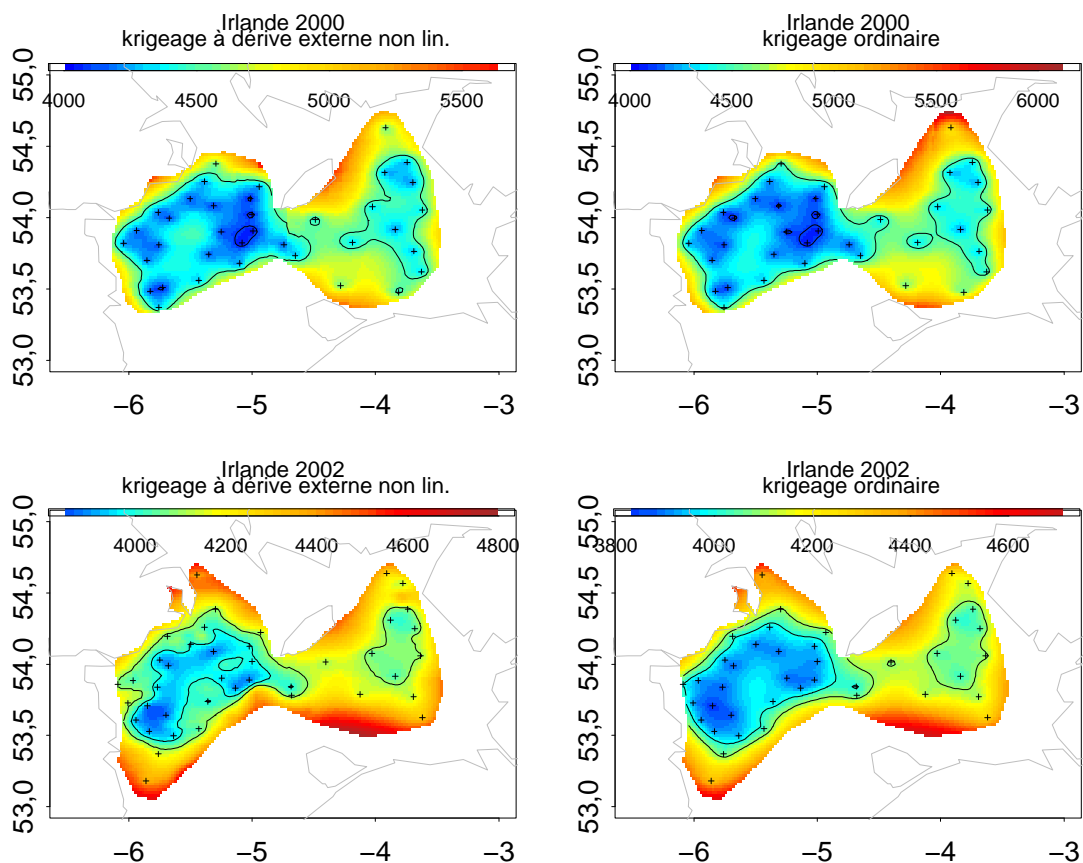


FIG. 4.13 – Cartes de variances d'estimation obtenues par krigeage des captures utilisant une estimation non linéaire (krigeage disjonctif) de l'acoustique en dérive externe (colonne de gauche) et sans dérive externe (colonne de droite), pour les campagnes réalisées par l'Irlande en 2000 et 2002 (ligne du bas). Pour chaque campagne l'échelle de couleur est la même pour les 2 modèles.

Les indices d'abondance et les variances d'estimation globales

Traditionnellement, l'estimation des stocks de poissons à partir des campagnes scientifiques de chalutage de fond vise à produire un indice d'abondance, représentant l'état du stock sur l'ensemble de la zone étudiée. Cet indice peut par exemple être représenté par la moyenne des données, ou par l'estimation à partir d'un modèle choisi de la valeur moyenne sur un polygone délimitant la zone d'étude.

L'enregistrement des données acoustiques peut être globalement considéré comme bénéfique pour les estimations globales pour deux raisons principales :

- soit l'indice obtenu est différent de celui obtenu par les seules données de pêche et est ainsi mieux représentatif de l'état véritable du stock
- soit l'indice obtenu est identique au précédent, mais sa variance est nettement inférieure. Cela signifie que l'estimation globale obtenue à partir des seules données en stations peut être assez bien représentative de l'état réel du stock, mais l'information acoustique permet un gain important en précision. L'indice est alors connu de façon (beaucoup) plus sûre.

Les modèles utilisés dans cette thèse rendent le premier cas impossible. En effet, que l'on n'utilise que les captures, ou au contraire, un modèle combinant acoustique et chalutage, l'estimation de la moyenne globale sera la même, calée sur la moyenne empirique des données de chalut.

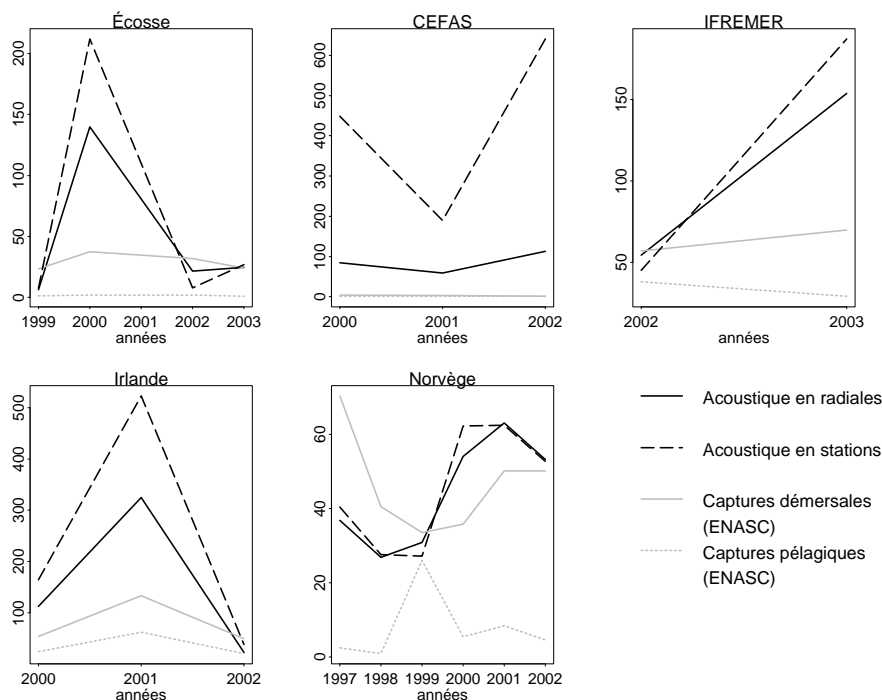


FIG. 4.14 – Evolution inter-annuelle des moyennes de l'acoustique collectée en radiales et en stations, ainsi que des captures démersales et pélagiques pour chacune des campagnes

Pays	Année	Moyenne		Variance		CV en %	
		Cokrigeage	Krigeage	Cokrigeage	Krigeage	Cokrigeage	Krigeage
Norvège	1997	65.5	70.3	15.4	27.1	6	7
Norvège	1998	39.1	40.5	3.4	5.6	5	6
Norvège	2000	6.7	5.4	0.5	2.6	10	30
Norvège	2001	50.2	50.2	7.0	8.8	5	6
Écosse	1999	1.2	1.4	0.1	0.5	30	52
Écosse	2002	50.7	31.8	22.1	12.5	9	11
Écosse	2003	23.9	24.2	4.3	3.9	9	8
Irlande	2000	11.7*	12.5	2.0*	2.2	12*	12
Irlande	2002	4.9*	6.0	2.3*	3.9	31*	33

TAB. 4.2 – Tableau récapitulatif des indices obtenus par cokrigeage des captures et de l'acoustique et krigeage seul des captures sur un polygone (moyenne, variance globale et coefficient de variation $= \frac{\sigma}{m}$). *En Irlande, le cokrigeage est effectué avec modélisation de la structure croisée (cokrigeage usuel) alors qu'en Écosse et Norvège, la simplification du modèle à résidu est utilisée.

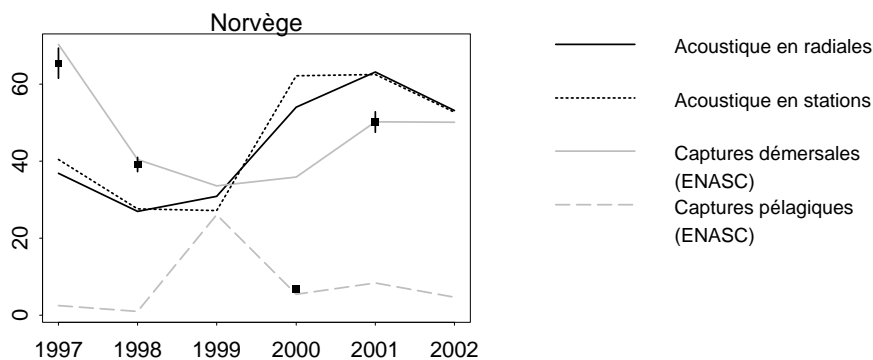


FIG. 4.15 – Evolution inter-annuelle des moyennes de l'acoustique et des captures démersales et pélagiques en Norvège, auxquelles sont superposées les estimations moyennes globales obtenues par cokrigeage (représentées par des carrés). La variable cible dans le cokrigeage est les captures démersales en 1997, 1998 et 2001 et pélagiques et 2000. Les segments autour des moyennes estimées par cokrigeage représentent l'écart-type de l'estimation.

Les deux types de données ont tout d'abord des moyennes qui ne varient pas de la même façon selon les années (cf. figure 4.14). Excepté pour quelques campagnes, l'acoustique a une moyenne souvent plus élevée que les captures, même lorsque l'on somme les pêches démersales et pélagiques, en particulier l'acoustique en stations. Cette différence est surprenante en mer du Nord où le chalut est plus large que le cône acoustique, mais s'explique par certaines valeurs extrêmes parfois enregistrées par le sondeur qui gonflent la moyenne acoustique. Il est possible que certains bancs de poisson de taille importante ne puissent pas être capturés. Une autre explication serait la sous-évaluation des captures de gros bancs lors de la transformation en valeur acoustique équivalente (ENASC). Les captures démersales et pélagiques des campagnes anglaises sont quasiment nulles comparées aux enregistrements acoustiques. En Irlande et en France pour les démersaux, l'évolution des moyennes est la même pour l'acoustique et les pêches, bien que l'acoustique marque les tendances inter-annuelles de façon plus accentuée. En Écosse, l'acoustique fournit une moyenne très proche de celle des captures démersales, excepté pour l'année 2000, pour laquelle une valeur de l'ordre de 10 000 fois la moyenne des données a été enregistrée par les sondeurs, alors que seule une faible proportion en a été capturée. En Norvège, enfin, l'évolution de l'acoustique en stations et en radiales sont très proches. Il est intéressant de voir que les captures démersales sont en moyenne supérieures à l'acoustique jusqu'en 1999 où la tendance s'inverse. Cette année correspond à un pic dans les captures pélagiques qui sont multipliées par cinq par rapport aux autres années. C'est aussi l'année des plus faibles captures démersales. A partir de l'année 1999, l'échantillonnage des campagnes en mer de Barents s'est élargi. La surface couverte par les données a augmenté de près d'un tiers. Il est possible que les nouvelles régions échantillonnées, en particulier à l'est reconnues comme étant plus froides et plus riches en poissons, expliquent en partie l'augmentation de la moyenne acoustique. Mais il est étonnant que cette augmentation ne soit pas suivie par les données de pêche, en particulier en 2000. Depuis que des campagnes collectant à la fois des données de captures et de chalutage sont effectuées (1981 en mer de Barents), deux indices sont généralement estimés séparément : un indice acoustique estimé à partir des seuls enregistrements des sondeurs et un indice de capture, calculé à partir des données de pêche. Ces deux indices sont considérés comme des estimations indépendantes d'une même population. Ces deux sources d'information sont utilisées pour se confronter l'une à l'autre lors de l'étude des variations inter-annuelles du stock (Godøet Wespestad, 1993 [25]). La différence existante au sein des données entre la moyenne des captures et celle de l'acoustique (figure 4.14) montre la difficulté d'une telle approche et l'importance de produire un indice combiné à partir d'un modèle basé sur les corrélations locales.

Ces observations pouvaient être faites dès la description des données lors du premier chapitre, puisqu'elles ne font intervenir aucune modélisation et ne sont que des statistiques de bases sur les différentes campagnes. J'ai cependant choisi de les présenter dans ce dernier chapitre pour pouvoir mieux interpréter les valeurs obtenues par différents modèles.

Pour évaluer et comparer les abondances estimées par différents modèles, ceux utilisant l'information acoustique et ceux n'utilisant que l'information de capture, il est pratique d'estimer la moyenne sur un polygone ainsi que la variance de cette moyenne, appelée variance globale. Un polygone a donc été tracé autour de chacune des campagnes de façon à s'ajuster au mieux à la zone échantillonnée pour laisser le moins de surface possible à l'extrapolation. Le polygone n'est donc pas exactement le même d'une année à l'autre, mais par contre pour une année donnée, le même polygone a servi à l'estimation de chaque modèle.

Que ce soit pour le cokrigeage ou pour le krigeage ordinaire, le krigeage de la moyenne s'obtient en remplaçant le membre de droite du système de cokrigeage (ou de krigeage) par le vecteur nul, mis à part les termes correspondant aux contraintes de non biais qui restent les mêmes : la somme des

poids de la variable cible est égal à 1 tandis que les poids de la variable auxiliaire pour le cokrigeage doivent être de somme nulle.

L'estimation de la moyenne ne permet pas de prendre en compte une dérive externe. Seules les approches par cokrigeage sont donc évaluées et comparées aux résultats obtenus par krigeages monovariés.

Comme on peut le voir dans le tableau 4.2, les moyennes obtenues par le modèle combiné et le modèle monovarié sont approximativement les mêmes. La variance globale obtenue est par contre systématiquement inférieure par cokrigeage, excepté en Écosse, où l'instabilité due à l'influence de valeurs extrêmes est très forte. En effet, en mer de Barents et en Écosse, le modèle à résidu étant utilisé en simplification du cokrigeage, il peut donc arriver que pour des raisons d'instabilité statistique, la variance d'estimation soit supérieure par cokrigeage que par krigeage mono-variable. La baisse de variance varie entre 20 et 97% en Norvège, avec une moyenne autour de 50%, et entre 10 et 40% en Irlande. L'acoustique apporte donc un gain en précision non négligeable du point de vue des estimations globales, puisqu'il permet de réduire en moyenne la variance de moitié en Norvège et d'un quart en Irlande.

Enfin, la méthode de régression linéaire pour les cas de sous-échantillonnage de la variable cible (voir partie 2.1) a été utilisée pour estimer des moyennes combinant acoustique et chalutage sur l'ensemble du champ, pour chacune des campagnes. Cette approche ne fournit pas de carte d'estimation, mais une moyenne globale qui peut être comparée à la moyenne des données de captures. Les résultats sont indiqués dans le tableau 4.3. Les moyennes obtenues sont très proches de la moyenne des données de pêche. Comme cela a été vu, le gain en variance est directement lié à la valeur du coefficient de corrélation entre les deux variables. Ce coefficient est souvent assez faible, excepté pour les démersaux en Norvège ainsi que certaines campagnes pour lesquelles le coefficient est fortement influencé par une valeur forte. Les variances d'estimation obtenues sont toujours supérieures aux variances obtenues par krigeage et cokrigeage. Cette approche équivaut à un modèle géostatistique où les deux variables sont pépitiées, ainsi que leur covariance croisée. L'utilité est ici de fournir un indice moyen et une variance d'estimation, pour permettre d'évaluer malgré l'absence de structure spatiale, le gain apporté par l'acoustique en terme de variance d'estimation. La variance de l'estimateur de la moyenne, calculé par régression (équation 2.2, page 32), ne peut être qu'inférieure à celle de l'estimateur de moyenne ordinaire. Celle-ci s'écrit :

$$Var(\bar{P}) = Var \left[\frac{1}{N} \sum_{i=1}^N P_i \right] = \frac{Var(p_i)}{N} \quad (4.1)$$

La diminution de variance obtenue par régression, entre le modèle combiné et le modèle monovarié, est directement liée au coefficient de corrélation entre l'acoustique et les captures. En Norvège, elle varie de 50 à 10% pour les démersaux et de 5 à 15% pour les pélagiques, excepté en 2000 où elle n'est que de 7% pour les démersaux, mais atteint 90% pour les pélagiques. En Irlande la baisse de variance atteint 60% en 2002 avec les pélagiques, mais ne dépasse pas 10% les autres années. En mer du Nord, les gains les plus importants sont en Écosse où entre 25 et 75% de baisse est obtenue entre 2000 et 2003. Il n'y a quasiment aucun gain en variance, par contre, pour les campagnes réalisées par le CEFAS et IFREMER. Évidemment, tout comme le coefficient de corrélation duquel elles se déduisent, les variances obtenues sont très sensibles aux valeurs fortes.

Pays	Année	Démersaux				Pélagiques			
		ρ	Moy	Var	$var(\hat{P}_{dem})$	ρ	Moy	Var	$var(\hat{P}_{pel})$
Norvège	1997	0.7	65.4	41.8	82.2	0.1	2.4	0.2	0.2
Norvège	1998	0.6	40.1	6.7	11.6	0	0.9	0	0
Norvège	1999	0.1	33.6	5.2	6.7	0.1	34.9	466.8	598.5
Norvège	2000	0.2	35.7	5.1	5.5	0.9	4.4	0.5	5.3
Norvège	2001	0.6	50.5	8.4	12.6	0.1	8.4	7	7.4
Norvège	2002	0.1	50.1	7	7.8	0	4.6	6.2	6.8
Irlande	2000	0	53.9	71.2	71.3	0.3	21.7	128.2	135.2
Irlande	2001	0.3	120.9	791.2	860.2	0.3	48.5	732.7	814.7
Irlande	2002	0.1	49.0	65.0	65.3	0.9	10.6	41.3	109.3
Écosse	1999	0.1	24.7	35.2	35.5	0	1.4	0.5	0.5
Écosse	2000	0	37.7	23.5	23.5	0.9	1.5	0.3	1.4
Écosse	2002	0	31	27.4	27.5	0.5	2.6	1.1	1.4
Écosse	2003	0	28.2	28.2	28.2	0.5	0.9	0.1	0.1
CEFAS	2000	0.2	3.2	0.4	0.4	0.2	0.6	0	0
CEFAS	2001	-0.1	3.8	0.2	0.2	0.1	0.6	0.1	0.1
CEFAS	2002	0	0.9	0	0	0.2	0.6	0.2	0.2
IFREMER	2002	0	56.6	208.4	208.6	0	37.5	11.4	11.5
IFREMER	2003	0	70	249.1	249.6	0	29	37.7	37.7

TAB. 4.3 – Tableau récapitulatif des résultats obtenus par régression linéaire en situation de sous-échantillonnage de la variable d'intérêt. Pour les démersaux et les pélagiques sont indiqués : le coefficient de corrélation entre les captures et l'acoustique (ρ), la moyenne obtenue sur l'ensemble du champ par la méthode de régression (Moy), la variance de l'estimation (Var) ainsi que la variance de l'estimateur de la moyenne pour les captures (= variance des données divisée par le nombre de stations)

Validation croisée

La validation croisée est un outil classique pour évaluer la capacité d'un modèle à retrouver des valeurs connues et pour comparer deux modèles différents. Le principe consiste à supprimer une à une chacune des données disponibles, de réestimer chaque point supprimé à partir des données restantes puis de comparer l'estimation obtenue avec la vraie valeur mise de côté. La pertinence de la validation croisée s'appuie sur trois propriétés de l'estimateur de krigeage : celui-ci est sans biais, de variance minimale et surtout il est un interpolateur exact. Ce dernier point justifie le besoin de supprimer une donnée et de la re-estimer pour évaluer le modèle. La qualité d'un modèle de type régression, par exemple, réside souvent en l'étude directe des résidus, inexistant par krigeage. Naturellement, les modèles ayant les meilleurs résultats par validation croisée ne sont pas nécessairement ceux fournissant les meilleures estimations aux points non échantillonnés (Goovaerts 1997 [29]).

Pour les modèles de cokrigeage, les deux valeurs de l'acoustique et de capture disponibles en une station ont été supprimées pour réestimer la donnée de capture. Pour les krigeages à dérive externe, l'estimation dépend des données de captures ainsi que des valeurs de la dérive aux points de stations et au point cible. On ne peut pas estimer la valeur en un point si la valeur de la dérive y est inconnue. Pour ces modèles, seule la donnée de captures a donc été supprimée en stations, la valeur de la dérive a été conservée.

Plusieurs critères ou indices peuvent ensuite être utilisés pour comparer l'ensemble des valeurs vraies (les données) observées en stations avec l'ensemble des estimations obtenues en ces mêmes points par validation croisée.

Plusieurs indices sont présents dans la littérature (voir par exemple Saito et Goovaerts 2000 [60], Chilès et Delfiner 1999 [15], Moyeed et Papritz 2000 [47]). Parmi eux les sept indices suivant ont été sélectionnés : (\bar{Z} est la moyenne empirique des données, $Z^*(x_i)$ l'estimation au point x_i par validation croisée et $\sigma^*(x_i)$ l'écart-type d'estimation du modèle au point x_i) :

1. le biais empirique, ou l'erreur moyenne :

$$M.E = \frac{1}{N} \sum_{i=1}^N [Z^*(x_i) - Z(x_i)]$$

2. la moyenne des erreurs absolues :

$$M.A.E = \frac{1}{N} \sum_{i=1}^N |Z^*(x_i) - Z(x_i)|$$

3. la moyenne des erreurs relatives :

$$M.R.E = \frac{1}{N} \sum_{i=1}^N \frac{|Z^*(x_i) - Z(x_i)|}{Z(x_i)}$$

4. la moyenne des erreurs au carré (mean square error) :

$$M.S.E = \frac{1}{N} \sum_{i=1}^N (Z^*(x_i) - Z(x_i))^2$$

5. la moyenne des carrés des erreurs standardisées (mean standardized square error) :

$$M.S.S.E = \frac{1}{N} \sum_{i=1}^N \frac{(Z^*(x_i) - Z(x_i))^2}{\sigma^{*2}(x_i)}$$

6. le biais relatif :

$$rME = \frac{M.E}{\bar{Z}} = \frac{\frac{1}{N} \sum_{i=1}^N (Z^*(x_i) - Z(x_i))}{\frac{1}{N} \sum_{i=1}^N Z(x_i)}$$

7. la moyenne relative des carrés des erreurs (relative mean square error) :

$$rMSE = \frac{M.S.E}{\frac{1}{N} \sum_{i=1}^N (\bar{Z} - Z(x_i))^2} = \frac{\sum_{i=1}^N (Z^*(x_i) - Z(x_i))^2}{\sum_{i=1}^N (\bar{Z} - Z(x_i))^2}$$

La moyenne des erreurs absolues (M.A.E) a l'avantage d'être moins sensible aux valeurs fortes que la moyenne des carrés des erreurs (M.S.E). La moyenne des carrés des erreurs standardisées par la variance de krigeage, ou de cokrigeage (M.S.S.E) doit être proche de 1. Une tolérance de plus ou moins $3\sqrt{\frac{2}{N}}$ basée sur la valeur du paramètre dans le cas gaussien est communément utilisée (Chilès et Delfiner 1999, [15]). Cette tolérance avantage donc le modèle basé sur les captures qui n'utilise que peu de données : elle varie entre 0.3 (pour la Norvège) à 0.7 (Irlande) pour les données en stations et de 0.06 (Norvège) à 0.3 (Irlande) pour les données en radiales.

La moyenne des erreurs relatives (M.R.E) est robuste aux valeurs extrêmes, puisque l'erreur est divisée par la valeur à estimer, mais elle interdit les données nulles. Étant donné la distribution statistique des données étudiées, certains de ces indices, comme le M.E, le M.A.E ou le M.S.E sont en effet très dépendants de la façon dont le modèle estime les quelques valeurs extrêmes du champ. Il est en effet important que l'estimation des valeurs les plus fortes soit la plus précise possible, puisque celles-ci représentent à elles seules une partie importante de l'abondance totale. Les deux derniers indices (rME et rMSE) mesurent le biais (M.E) et la moyenne des erreurs au carré (M.S.E) réciproquement rapportés à la moyenne et la variance empirique des données. Cela permet en particulier de comparer des valeurs obtenues pour des campagnes différentes ou pour des classes de valeurs différentes au sein d'une même campagne. Il peut être intéressant, par exemple d'étudier la moyenne des erreurs de validation croisée pour trois classes de valeurs, les valeurs faibles, les valeurs moyennes et les valeurs fortes. Il est possible que certains modèles, par exemple, soient mieux adaptés à réestimer les valeurs fortes que les valeurs faibles ou inversement (Saito et Goovaerts 2000 [60]).

Les résultats obtenus pour l'ensemble des campagnes (tableau 4.4) souvent assez proches pour les modèles combinant acoustique et chalutage et les modèles monovariabiles correspondants.

En Écosse, les résultats ne sont pas très satisfaisants. Le krigeage des captures, excepté pour la campagne 2003, semble obtenir de meilleures estimations que le cokrigeage et le krigeage à dérive externe. Le modèle de krigeage avec une estimation non linéaire de l'acoustique en dérive externe semble fournir des résultats légèrement meilleurs que le modèle mono-variable pour les campagnes du CEFAS, mais strictement égaux pour les campagnes en mer d'Irlande. Le cokrigeage permet par contre de gagner en précision pour les données irlandaises de 2000 et 2002.

En mer de Barents, les résultats de la validation croisée sont à nouveau assez semblables entre les modèles. Le cokrigeage fournit des estimations systématiquement meilleures du point de vue de la moyenne des erreurs absolues (MAE), de la moyenne des carrés des erreurs (MSE), la moyenne des carrés des erreurs standardisées (MSSE) ainsi que la moyenne relative des carrés des erreurs

Pays	Année	Variable cible	Modèle	ME	MAE	MRE	MSE	MSSE	rME	rMSE
Écosse	1999	pèl.(enasc)	cok	-0.2	2.5	65.9	30.9	1.4	-0.2	1.4
Écosse	1999	pèl.(enasc)	ked	0.2	2.2	52.8	23.6	1.0	0.1	1.1
Écosse	1999	pèl.(enasc)	ko	0.2	2.2	52.9	23.4	1.0	0.2	1.1
Écosse	2002	dem.(enasc)	cok	13.4	31.2	6.5	2076	0.4	0.4	1.5
Écosse	2002	dem.(enasc)	ked	1.1	21.5	3.9	1014	0.9	0	0.8
Écosse	2002	dem.(enasc)	ko	0.5	19.9	2.1	900.6	0.9	0	0.7
Écosse	2003	dem.(enasc)	cok	-0.3	11.8	1.0	292	1.0	0	0.6
Écosse	2003	dem.(enasc)	ked	0.3	11.1	0.9	288	1.0	0	0.6
Écosse	2003	dem.(enasc)	ko	0.2	11.7	1.0	314	1.2	0	0.7
Écosse	2002	dem.(num)	ked nonlin	-12.5	1736	2.6	6.10 ⁶	2.2	0	0.9
Écosse	2002	dem.(num)	ko	17.4	1658	2.6	5.10 ⁶	2.05	0	0.8
Écosse	2003	dem.(num)	ked nonlin	-2.5	1464	1.5	5.10 ⁶	1.1	-0.1	0.8
Écosse	2003	dem.(num)	ko	0.2	1299.3	1.1	5.10 ⁶	1.4	0	0.8
CEFAS	2001	dem.(enasc)	ked nonlin.	-0.1	2.4	21.5	10.7	1.2	0	0.8
CEFAS	2001	dem.(enasc)	ko	-0.1	2.4	16.9	11.2	1.3	0	0.9
CEFAS	2002	dem.(enasc)	ked nonlin.	0	0.6	0.9	0.7	1.1	0	0.8
CEFAS	2002	dem.(enasc)	ko	0	0.6	1	0.8	1.2	0	0.9
Irlande	2000	pèl.(enasc)	cok	-0.4	10.7	8.4	238	1.5	0	0.6
Irlande	2000	pèl.(enasc)	ked	5.3	14.5	4.1	1122	1.4	0.4	2.9
Irlande	2000	pèl.(enasc)	ko	0.5	11.1	9.8	362	1.4	0	0.9
Irlande	2002	pèl.(enasc)	cok	0.6	10.4	6.9	315	1.95	0	0.7
Irlande	2002	pèl.(enasc)	ked	0.7	10.6	6.8	330	1.91	0	0.8
Irlande	2002	pèl.(enasc)	ko	0.6	10.3	6.9	323	1.80	0	0.8
Irlande	2000	pèl.(NASC)	ked nonlin.	1	28.8	16.3	4819	1	0	0.9
Irlande	2000	pèl.(NASC)	ko	0.7	29.2	14.9	4996	1	0	1
Irlande	2002	pèl.(NASC)	ked nonlin.	-0.3	28.5	18.4	4449	1	0	1
Irlande	2002	pèl.(NASC)	ko	0.4	25.4	20.1	4481	1	0	1

TAB. 4.4 – Tableau récapitulatif des indices obtenus par validation croisée par les différents modèles utilisés pour chacune des campagnes disponibles en mer du Nord et mer d'Irlande

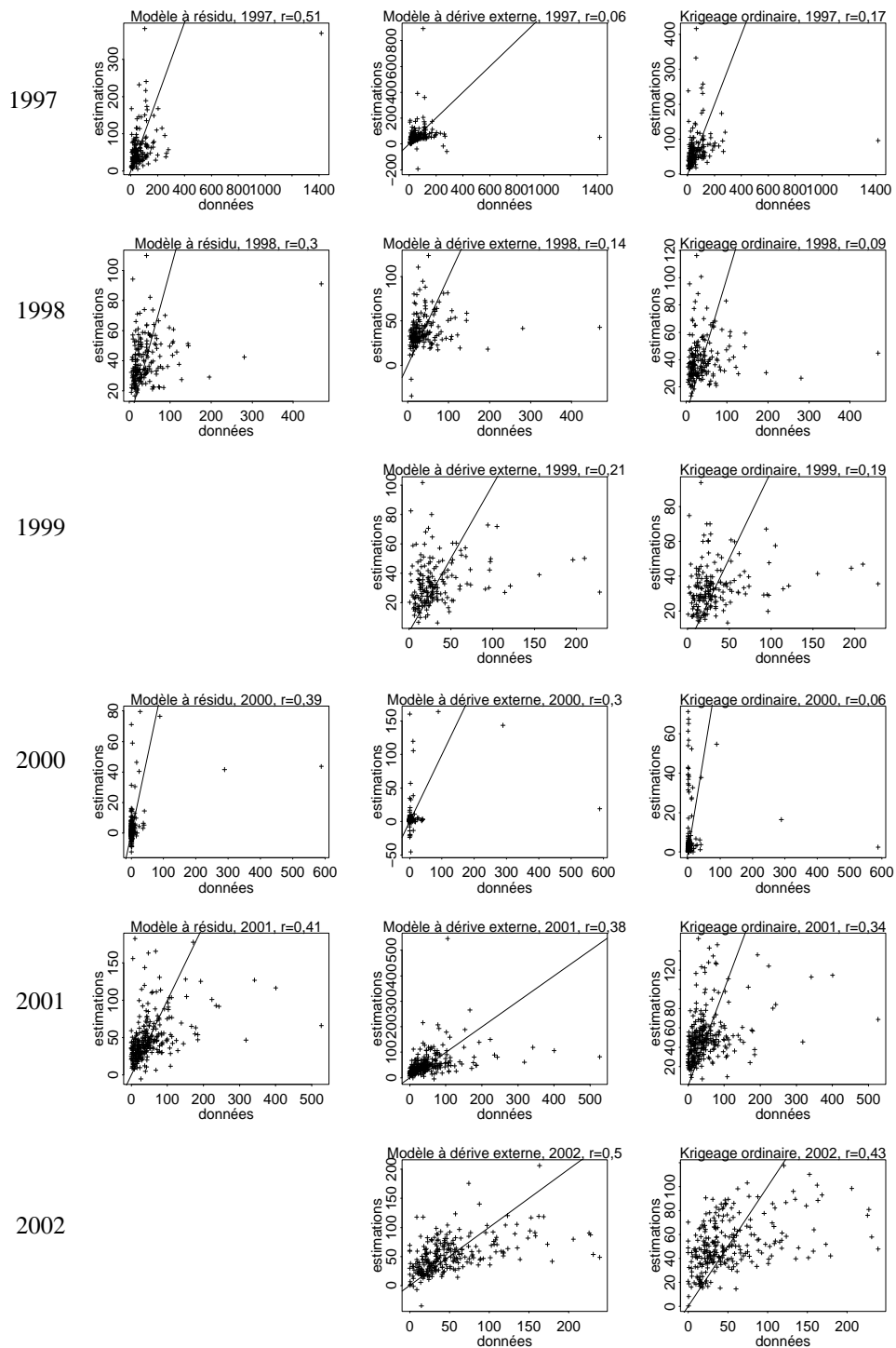


FIG. 4.16 – Nuages de corrélations entre les valeurs observées (les données) et les valeurs estimées par validation croisée par les trois modèles : cokrigeage, krigeage à dérive externe et krigeage ordinaire des captures pour chacune des campagnes réalisées par la Norvège. La première bissectrice et le coefficient de corrélation sont indiqués sur les figures.

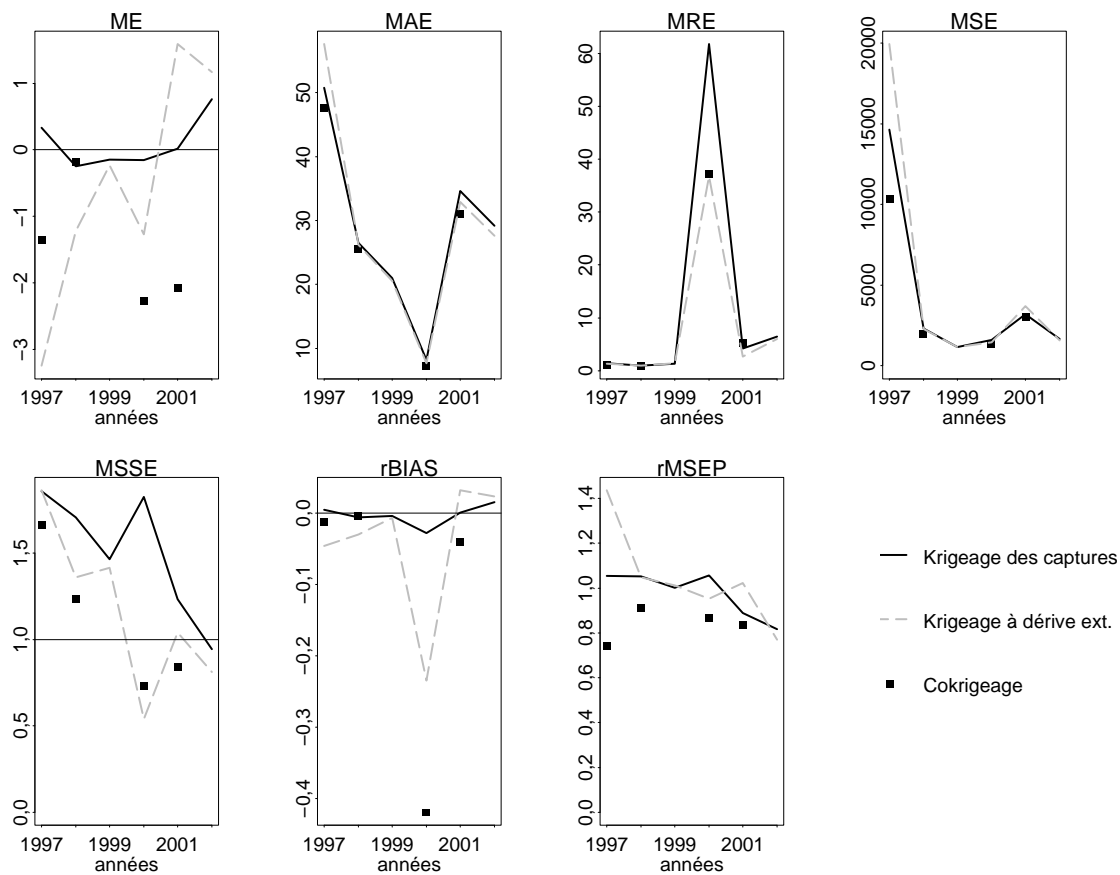


FIG. 4.17 – Évolution des indicateurs de précision évalués par validation croisée pour chaque campagne en mer de Barents (ME, MAE, MRE, MSE, MSSE, rME et rMSE).

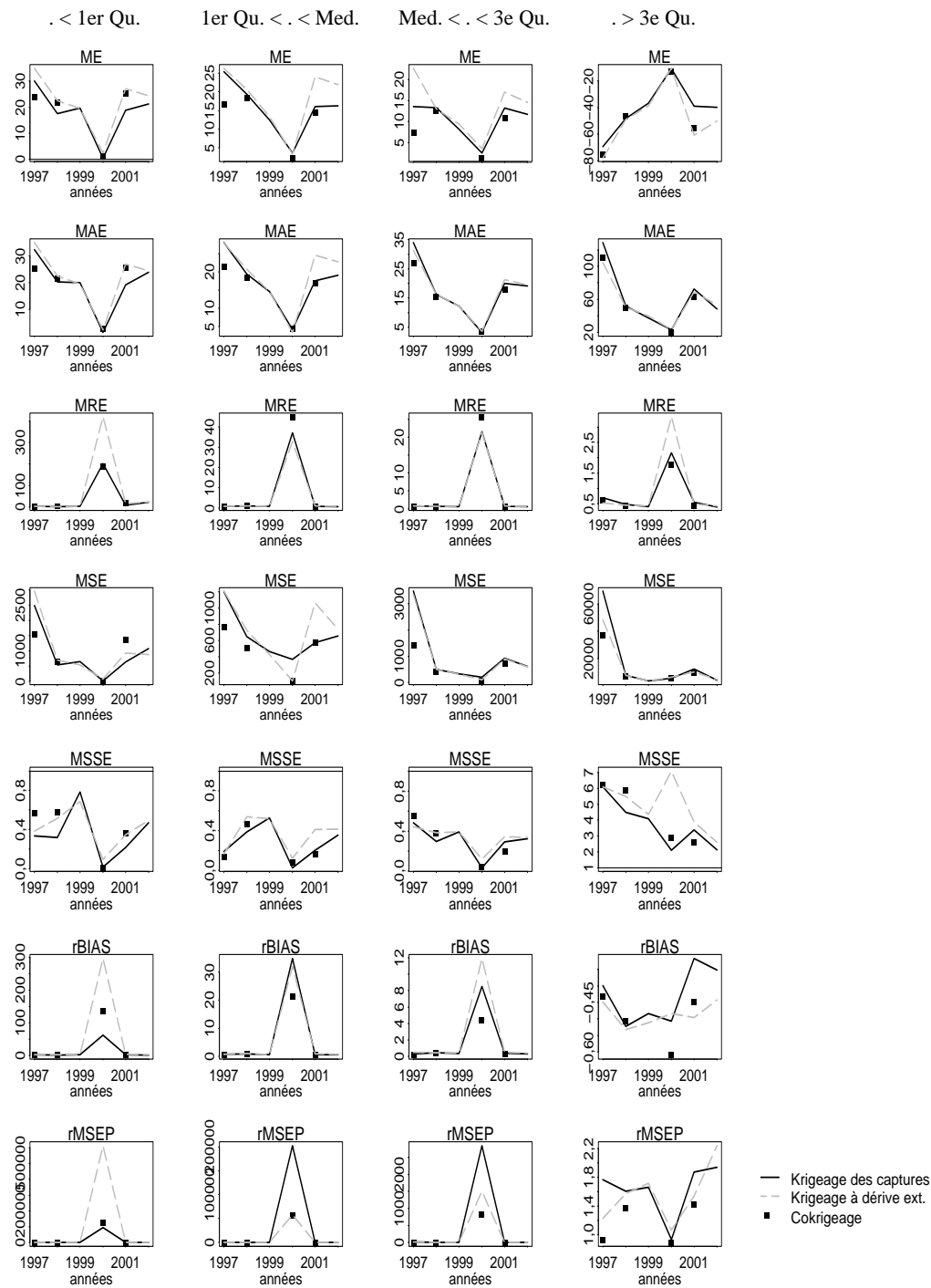


FIG. 4.18 – Évolution des indicateurs de précision évalués par validation croisée pour chaque campagne en mer de Barents (ME, MAE, MRE, MSE, MSSE, rME et rMSE), selon 4 classes de valeurs : la première colonne correspond aux données inférieures au 1er quartile, la deuxième celles entre le 1er quartile et la médiane, la troisième celles entre la médiane et le 3e quartile, puis celles supérieures au 3e quartile dans la colonne de droite.

(rMSE) ; bien que le gain apporté par le cokrigeage semble petit comparé aux différences d'ordre de grandeur inter-annuel (figure 4.17). La figure 4.18 montre l'évolution de ces mêmes indicateurs en fonction du niveau des données ré-estimées dans leur répartition statistique. Il est intéressant de voir si le modèle combiné et le modèle mono-variable ont la même précision quand il s'agit d'estimer des valeurs faibles ou fortes. On voit ainsi que pour plusieurs indicateurs la tendance s'inverse entre les différentes classes. Pour les indicateurs de biais (ME et rME) le cokrigeage semble meilleur pour les classes moyennes. Pour les indicateurs faisant intervenir la valeur absolue des erreurs ou le carré (MAE, MSE, rMSE) les estimations de cokrigeage semblent plus précises pour les valeurs moyennes et fortes.

4.1.4 Conclusion : l'acoustique améliore t-elle les estimations d'abondance ?

Les enregistrements acoustiques collectés sur tout le trajet du navire apportent incontestablement des informations permettant de délimiter plus précisément les zones riches et pauvres et de tenir compte de zones locales d'abondance ou de pauvreté situées entre les points de chalutage et donc invisibles à partir des seules données de captures. Cet effet est compensé par l'estimation de zones d'abondance le long des transects, si bien que les indices d'abondance globale obtenus sont similaires qu'ils soient calculés avec ou sans acoustique (cf. tableau 4.2).

Le cokrigeage permet un gain en variance inhérent à ce modèle puisque la variable auxiliaire est affectée d'un poids nul si elle ne permet pas une diminution de variance. Ce gain s'avère particulièrement important en Norvège en raison de la prise en compte de 30 fois plus de données. De plus, lors de l'estimation globale sur un polygone regroupant l'ensemble de la zone d'étude, la diminution de variance est de l'ordre de 50%. Les études de validation croisée montrent enfin une diminution de l'erreur de re-estimation grâce au cokrigeage, en particulier lorsque l'on s'intéresse à la moyenne des erreurs quadratiques, standardisées ou non.

En Irlande, les résultats obtenus sont assez semblables. L'acoustique permet d'apporter des détails dans les cartes d'estimation et de réduire la variance. Celle-ci baisse de 10% en 2000 et de 40% en 2002 lors de l'estimation globale de la zone d'étude. Par contre, les erreurs de validation croisée sont assez similaires pour les modèles combinés et les modèles mono-variables.

En mer du Nord, enfin, les résultats sont assez décevants. Le krigeage ordinaire fournit à plusieurs points de vue des résultats meilleurs que les modèles combinés avec l'acoustique. Bien que les cartes d'estimation soient notablement moins lisses et plus contrastées, les résultats de la validation croisée favorisent le choix d'un modèle n'utilisant que l'information de pêche.

Les variances obtenues par régressions linéaires ou grâce aux modèles combinés montrent que l'acoustique permet une diminution de variance des estimations, inhérente à l'ajout d'une variable quelle qu'elle soit dans un modèle, pourvu que celle-ci soit corrélée à la variable cible. Il est important ensuite de savoir si ce gain en variance est suffisamment important pour justifier les "dépenses" engendrées par la collecte de l'acoustique au cours des campagnes de chalutage de fond et celle engendrées par leur analyse statistique.

Allocation optimale

Cochran (1977, [16]) propose une technique simple pour estimer si la baisse de variance est rentable ou non en situation de double échantillonnage, avec des données indépendantes.

La méthode consiste à supposer que l'institut effectuant les campagnes dispose d'une somme fixée \mathbb{C} . La collecte de données de la variable auxiliaire est alors dite rentable, si la variance de l'estimateur de moyenne obtenue par régression (équation 2.2, page 32) sous contrainte de respecter le coût maximal \mathbb{C} est inférieure à la variance de l'estimateur de moyenne classique (équation 4.1). L'évaluation se déroule donc de la sorte :

1. On suppose que le coût de collecte d'une donnée de capture est c_P et que le coût de collecte d'une donnée acoustique est c_A . L'acoustique est naturellement de coût inférieur aux captures $c_A < c_P$. Le coût total d'une campagne s'écrit alors : $n_S c_P + n_R c_R = \mathbb{C}$, où n_S est le nombre de stations et n_R le nombre de radiales (l'acoustique étant également collecté en stations, l'intégralité des points est pris en compte).
2. On recherche alors la variance minimale que peut prendre l'estimateur obtenu par régression sous la contrainte de coût. On résout donc le problème, trouver n_S et n_R tels que :

$$\begin{aligned} \text{Min } V_r(\hat{M}) &= \frac{S_P^2}{n_S}(1 - \rho^2) + \frac{S_P^2}{n_R}\rho^2 \\ \text{sc. } N_S c_P + N_R c_R &= \mathbb{C} \end{aligned}$$

On obtient alors à l'optimum la relation suivante :

$$\frac{n_S}{n_R} = \sqrt{\frac{c_R(1 - \rho^2)}{c_P \rho^2}} \quad (4.2)$$

Ce qui donne en remplaçant dans $V_r(\hat{M})$:

$$V_{rmin} = \frac{S_P \left(\sqrt{c(1 - \rho^2)} + \sqrt{c_R \rho^2} \right)^2}{\mathbb{C}}$$

3. Si l'on n'avait pas collecté l'acoustique mais uniquement les données en stations, on aurait

$$n_S c_P = \mathbb{C}$$

et

$$\text{Var}(\bar{P}) = \frac{S_P}{n_S} = \frac{c_P S_P}{\mathbb{C}}$$

4. La collecte de l'acoustique est alors rentable si :

$$V_{rmin} \leq \text{Var}(\bar{P}) \Leftrightarrow \rho^2 \geq \frac{4(c_P/c_R)}{(1 + c_P/c_R)^2}$$

Cette relation permet donc de définir pour une corrélation donnée entre les variables de capture et d'acoustique, le ratio maximal des coûts pour que l'opération de sondage acoustique soit rentable. Il est ensuite possible d'en déduire un ratio n_R/n_P optimal à partir de l'équation 4.2.

Ainsi, s'il existe une corrélation ρ de 0.5 entre l'acoustique et les captures, le sondage acoustique n'est *rentable* que si l'enregistrement d'une donnée acoustique coûte dix fois moins qu'un coup de chalut et qu'au moins le double de données sont collectées. Si le coefficient de corrélation est de 0.7, le prix d'une donnée acoustique doit être 6 moins que celui d'un chalutage. Si l'on considère que les

valeurs de coefficient de corrélation observées dans les données sont représentatives de la corrélation réelle existante entre les deux variables, l'enregistrement d'une donnée acoustique devrait être entre 10 et 80 fois moins chère (selon les années) qu'un coup de chalut en Norvège lors de l'estimation des démersaux et 800 fois moins chère pour l'estimation des pélagiques, pour être rentable. En mer du Nord, pour l'estimation des pélagiques, le coût d'une donnée acoustique devrait être environ 10 fois moins chère qu'un coup de chalut pour les campagnes écossaises, 80 fois moins chère pour les campagnes anglaises et infiniment moins chère pour les campagnes françaises !

Cependant, le coefficient de corrélation étant très sensible aux quelques rares valeurs fortes, il est difficile de pouvoir réellement le considérer comme représentatif. Néanmoins, cette approche, bien que simpliste parce qu'elle ne tient pas compte du plan d'échantillonnage et de la corrélation entre les points de données, peut être un outil d'aide à la décision pour savoir s'il est utile ou non de continuer à collecter les données acoustique dans le but d'améliorer les estimations d'abondance par chalutage de fond. Il faudrait alors sûrement utiliser une estimation de la corrélation plus robuste, comme par exemple une moyenne inter-annuelle du coefficient de corrélation.

Le coût de l'enregistrement acoustique est souvent considéré comme quasiment nul pour un navire disposant déjà de sondeurs, bien que le coût du traitement des données soit également à prendre en compte. Naturellement, étant donné qu'il s'agit de campagnes scientifiques, de multiples intérêts autres que la baisse de variance de l'estimateur de moyenne peuvent justifier la collecte de l'acoustique lors des campagnes. De plus le coût financier n'est pas nécessairement le critère le plus pertinent pour évaluer le gain apporté par l'acoustique. Enfin, la variance de l'estimateur utilisée pour ce critère est largement supérieure à celles obtenues par des approches géostatistiques beaucoup plus performantes et justifiant pleinement l'utilisation de sondeurs.

4.2 Effet de l'échantillonnage

Comme cela a été décrit dans le premier chapitre, les campagnes réalisées dans chacune des trois mers suivent des plans d'échantillonnage différents. Les campagnes en mer de Barents suivent un plan d'échantillonnage régulier avec une station tous les 20 m.n tandis que les campagnes en mer du Nord suivent un plan d'échantillonnage aléatoire stratifié (cf. partie 1.1.3 page 7 et figures 1.5, 1.7 et 1.8). Le but de cette partie est de se rendre compte si la différence entre les modes d'échantillonnage peut expliquer, dans une certaine mesure, les différences entre les résultats obtenus.

4.2.1 L'engin de pêche

La première différence réside dans l'utilisation d'un chalut différent. L'ouverture entre les portes du chalut est, en mer du Nord, jusqu'à 12 fois plus large que le diamètre du cône acoustique, alors que les deux dimensions sont équivalentes en mer de Barents. Cette différence est considérable. Il est très probable qu'un nombre important de poissons ne passent pas dans le cône acoustique et soient tout de même capturés. Ceci est non seulement dû à la différence de profondeur entre les deux mers, mais aussi à l'utilisation pour les campagnes IBTS d'un chalut à grande ouverture. Elle pourrait cependant disparaître avec l'apparition récente des sondeurs multi-faisceaux qui permettent d'enregistrer les données acoustiques sur un cône extrêmement plus large. Les sondages acoustiques réalisés en mer du Nord pourraient ainsi enregistrer la totalité de ce qui entre dans le chalut.

Cette première différence entre les campagnes peut être une explication de la difficulté à obtenir

une bonne corrélation entre le chalut et l'acoustique en mer du Nord.

4.2.2 Densité d'échantillonnage

Pour toute approche statistique, la qualité de l'estimation est liée au nombre de données disponibles. De ce point de vue, la différence entre les trois zones d'étude est frappante, puisqu'il y a environ 6000 données disponibles en radiales et 250 en stations en mer de Barents, 200 radiales et 35 stations en moyenne en mer d'Irlande et 630 radiales et 60 stations en mer du Nord. Le nombre de données n'est cependant représentatif que proportionnellement à la surface étudiée. Un polygone a donc été tracé pour chacune des trois zones de façon à s'ajuster au mieux à la zone échantillonnée. Les figures 4.19 et 4.20 représentent le ratio entre la surface du polygone en $m.n^2$ et le nombre de données respectivement en stations et en radiales. Cette grandeur représente la surface d'influence moyenne d'un échantillon ou encore, la surface équivalente ($S_{eq.}$) à chaque donnée si la campagne suivait un échantillonnage régulier.

La surface associée à une station est entre 400 et 600 $m.n^2$ en Norvège puisque les stations sont régulièrement espacées de 20 m.n. En raison de sa petite superficie, la densité de stations en mer d'Irlande est très forte et la $S_{eq.}$ se situe autour de 150 $m.n^2$ /station. La densité de stations pour les campagnes réalisées par l'Écosse et la France est exactement la même que celle de la mer de Barents ($S_{eq.} \approx 500m.n^2$ /station). Elle est par contre beaucoup plus faible pour les campagnes réalisées par le CEFAS où la $S_{eq.}$ se situe qu'autour de 1300 $m.n^2$ /station, soit plus du double.

A densités de chalutage et structures spatiales égales, la précision des modèles n'utilisant que l'information de capture devrait donc être similaires entre la Norvège et la mer du Nord (du moins pour les campagnes écossaises et françaises). En particulier, les indices d'abondance calculés actuellement par chalutage de fond sont estimés à partir de la même densité de données (excepté pour les campagnes du CEFAS).

Par contre, il est intéressant de constater que la densité des données de radiales n'est pas la même (figure 4.20). La $S_{eq.}$ en radiales est de 20 $m.n^2$ /radiale en mer de Barents, 30 $m.n^2$ en mer d'Irlande, 70 $m.n^2$ pour les campagnes écossaises et environ 90 $m.n^2$ pour les campagnes françaises et anglaises. Ceci vient principalement du fait que les données ne sont pas enregistrées de nuit pour les campagnes en mer du Nord (protocole IBTS), alors qu'elles le sont en mer d'Irlande et mer de Barents. Le comportement des poissons est en effet très différent entre le jour et la nuit et l'utilisation sans distinction de toutes les données peut fausser les estimations. La perte en densité d'échantillonnage est cependant plus importante que du simple au double. Le nombre de données en radiales devrait être multiplié par 4 pour atteindre la densité des campagnes norvégiennes, alors que la densité des stations est la même.

Le nombre de stations doit être important pour permettre d'établir une relation entre les captures et les enregistrements acoustiques. De ce point de vue, les trois zones ont des densités équivalentes. Mais c'est évidemment entre les stations que l'acoustique présente le plus d'intérêt : plus ces données seront nombreuses, plus l'acoustique sera susceptible d'améliorer les estimations d'abondance. Enfin, l'effort d'échantillonnage doit être d'autant plus important que les distributions statistiques des deux variables sont très asymétriques. Il est nécessaire de disposer d'un grand nombre de données pour avoir une bonne connaissance de la queue de l'histogramme qui représente la plus grande partie de l'abondance. En outre plus le nombre de données collectées est grand, plus l'influence des valeurs fortes sur les estimateurs linéaires sera réduite.

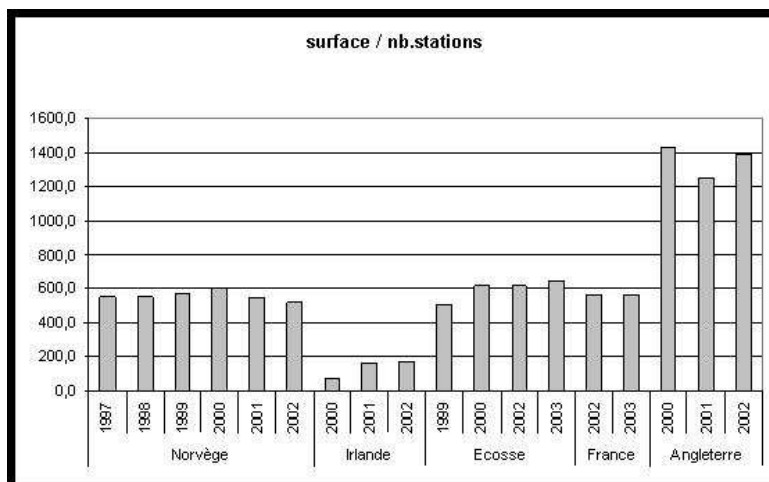


FIG. 4.19 – Densité moyenne de stations dans chaque zone d'étude

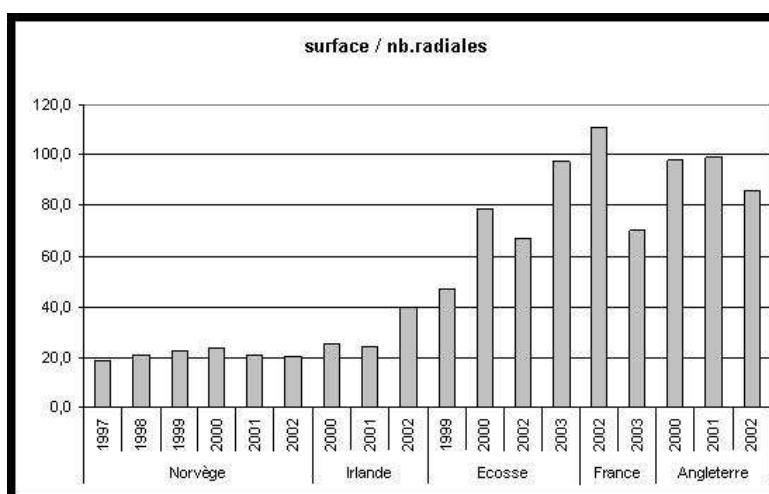


FIG. 4.20 – Densité moyenne de radiales dans chaque zone d'étude

Enfin, le nombre d'années de campagnes disponibles au sein d'une région est également un facteur déterminant pour ce type de données. Le lien acoustique-capture étant ténu et bruité, disposer d'un grand nombre d'observations permet d'extraire un signal du bruit ambiant. Dans le cadre de ce travail, six campagnes étaient disponibles en mer de Barents, dix en mer du Nord (quatre campagnes écossaises, trois anglaises, et deux campagnes françaises) et trois campagnes en mer d'Irlande. Les campagnes de mer du Nord réalisées par des instituts différents ont dû être traitées séparément, bien que répondant toutes au protocole IBTS en raison de la différence des espèces, de l'environnement et des écosystèmes. Il est important de disposer d'une série temporelle importante en raison des fluctuations inhérentes à ce type de données. En effet, sur les six campagnes mises à disposition par la Norvège, seulement quatre avaient des structures marquées permettant une estimation par modèle à résidu. Les campagnes 1999 et 2002 se sont avérées extrêmement bruitées. Il est possible que les quelques années de campagnes étudiées en mer du Nord (principalement 2000,

2002 et 2003) ne soient pas représentatives. Disposer d'un nombre suffisant de campagnes permet de mettre de côté les années exceptionnelles et de faire émerger au sein des autres des tendances significatives.

4.2.3 Comparaison des plans d'échantillonnage

Plusieurs auteurs ont comparé les deux plans de sondage les plus classiques en statistique spatiale :

- régulier : les échantillons sont répartis uniformément sur la zone d'étude
- aléatoire stratifié : la localisation de chaque échantillon est tirée au hasard au sein de chaque strate, toutes de même superficie, avec le même nombre de points dans chaque strate.

Par sa nature, le premier crée des données corrélées entre elles par leur localisation. Dans une certaine mesure, l'échantillonnage aléatoire stratifié aussi, puisque deux points ne peuvent pas appartenir au même carré, mais généralement de façon beaucoup moins marquée. La géostatistique est particulièrement apte à traiter des phénomènes de ce type. Naturellement les données issues d'un échantillonnage aléatoire peuvent aussi être corrélées, non pas par leur localisations qui sont indépendantes, mais parce qu'elles sont des prélèvements d'une biomasse spatialement structurée. L'étude des variances de dispersion (Delfiner 1979 [18]) montre que l'échantillonnage régulier fournit une variance d'estimation globale inférieure à celle d'un plan aléatoire stratifié. Si les données sont au centre de carrés, le plan d'échantillonnage régulier est deux fois plus précis qu'un plan aléatoire stratifié pour un variogramme linéaire. Pour un variogramme sphérique, le gain est d'autant plus important que la portée du variogramme est grande par rapport à la longueur du côté du carré. En particulier, pour une distance inter-station de 20 m.n, il faut que la structure des données ait une portée de plus de 2.5 m.n pour que l'échantillonnage régulier soit plus performant.

Si la disposition des stations suit un plan d'échantillonnage établi, celle des transects est plus délicate à comparer. Rivoirard et al (2000, [59]) montrent par des simulations que lorsque la distance entre les transects est supérieure à leur longueur, il est plus intéressant d'effectuer les transects en zigzag que parallèles. Mais reliant les stations entre elles, les transects suivent ici des échantillonnages trop hétéroclites pour répondre à ces caractéristiques.

Pour comparer de façon très simple les différents plans d'échantillonnage, j'ai effectué trois simulations par bandes tournantes d'une fonction aléatoire lognormale de paramètres équivalents à ceux observés en Norvège et en mer du Nord pour les captures démersales (en ENASC) à savoir une moyenne de 30 et une variance de 1500. La portée du modèle sphérique choisi a été fixée à 200 m.n. Chacune de ces trois simulations varie par sa proportion d'effet de pépité. La première a une structure comprenant 25% d'effet de pépité, la seconde 40% et la troisième 50% (figure 4.21). Ces simulations permettent d'obtenir une abondance fictive sur un champ de 700 m.n sur 500 m.n.

Sur ce champ, ont été reportées les coordonnées spatiales des radiales de chacune des campagnes, par projection (figure 4.22). Les localisations sont celles de la campagne 1997 pour la mer de Barents, 2002 en mer d'Irlande, 2003 en mer du Nord pour la zone écossaise, 2002 pour la zone anglaise et 2003 pour la zone française. On voit de nouveau sur cette figure la différence entre la densité des données des cinq zones d'étude. Les radiales norvégiennes forment un réseau très dense par rapport aux radiales des campagnes IBTS.

Des variogrammes expérimentaux ont alors été calculés à partir des valeurs des trois simulations

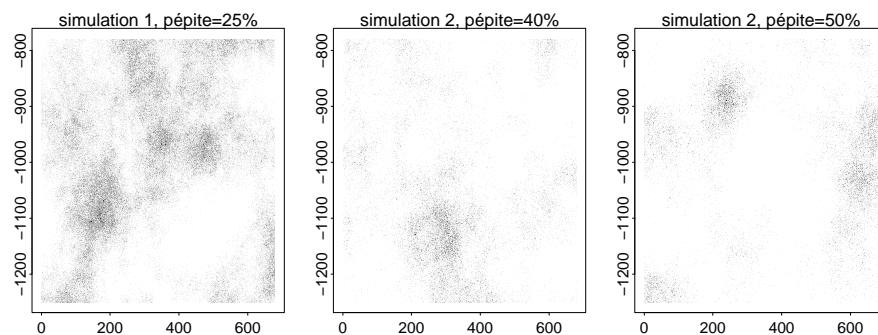


FIG. 4.21 – Trois simulations d’une variable lognormale suivant un modèle sphérique de portée 200 m.n avec une proportion d’effet de pépite (de gauche à droite) de 25%, 40% et 50%, la moyenne et la variance restant les mêmes.

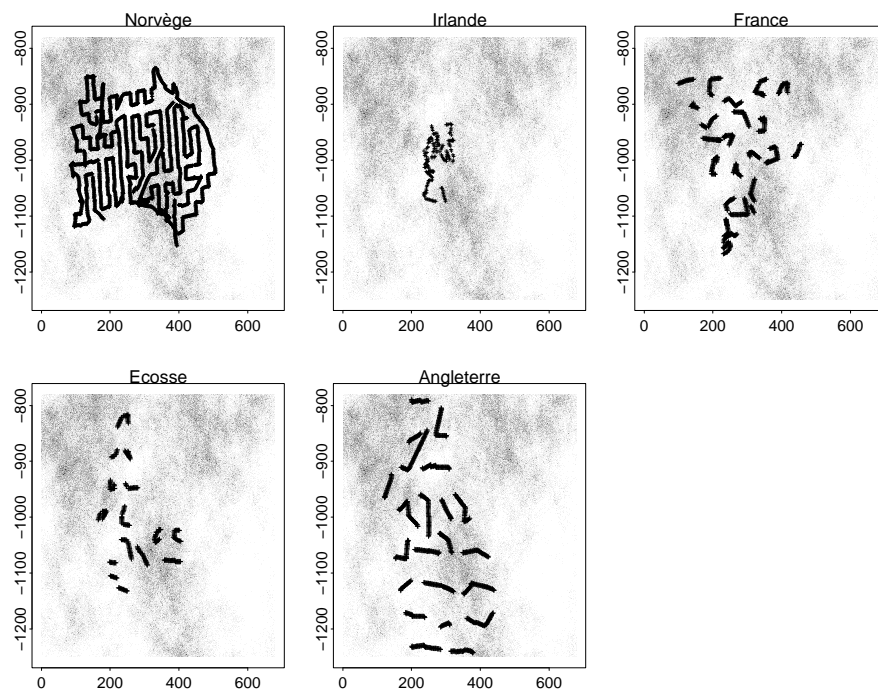


FIG. 4.22 – Localisations des radiales de chaque campagne dans le champ simulé après projection.

aux coordonnées obtenues (figure 4.23). On peut voir que les radiales de la campagne en mer de Barents permettent de dégager la structure pour chacune des trois simulations, malgré un bruit important pour les très petites distances pour la simulation la plus bruitée (50% d'effet de pépité). Les transects de la campagne en mer d'Irlande fournissent une structure bruitée mais assez stable pour les trois simulations, malgré la petite taille de la surface échantillonnée. Les transects des campagnes anglaises et écossaises en mer du Nord génèrent par contre des structures assez bruitées, surtout lorsque l'effet de pépité devient important.

Il serait intéressant d'effectuer de la même façon un plus grand nombre de simulations pour s'assurer de la généralité de ses observations.

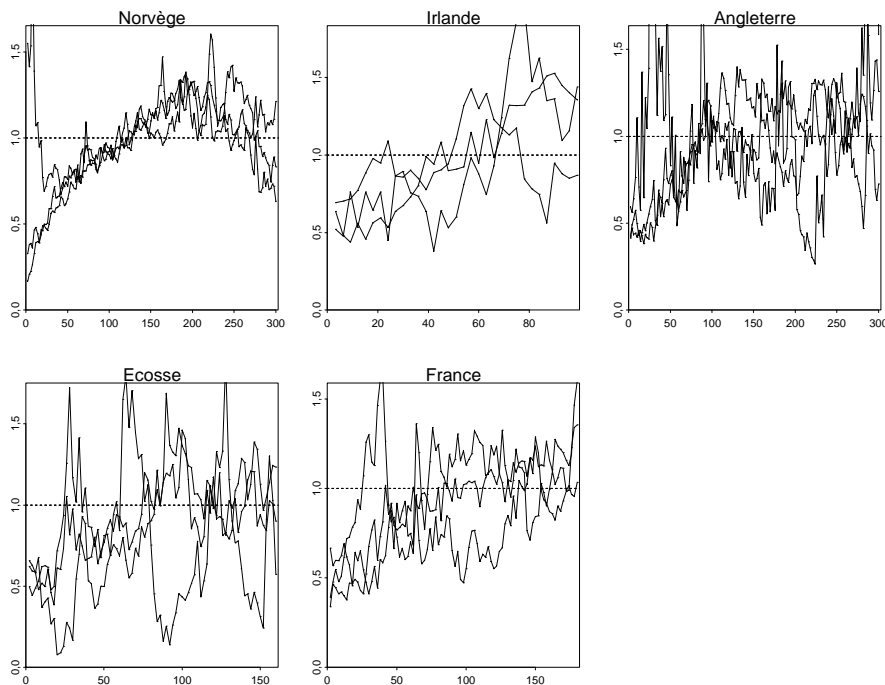


FIG. 4.23 – Pour chaque zone étudiée : variogrammes expérimentaux obtenus à partir des trois simulations présentées précédemment (les 3 variogrammes sont superposés) en prenant les coordonnées projetées des transects des campagnes.

4.2.4 Discussion

Disposer de données issues de trois mers différentes permet de se rendre compte de l'influence relative de facteurs environnementaux et biologiques sur la relation acoustique-chalet et de la complexité de la problématique. Quelque soit la méthodologie employée : par géostatistique comme je l'ai présenté, par réseaux de neurones, par logique floue, réalisés par le CEFAS, ou par modèles linéaires généralisés mis en oeuvre par le Marine Laboratory of Aberdeen, les conclusions convergent sur une meilleure corrélation acoustique-chalet pour les campagnes réalisées en mer de Barents que

pour celles réalisées en mer du Nord.

Cette différence s'explique probablement par l'addition de plusieurs phénomènes conjoints.

Tout d'abord, la densité de données doit être très importante pour obtenir une estimation satisfaisante. La répartition spatiale des poissons est par nature très fluctuante, il est donc nécessaire de collecter les données à pas fin. La présence de bancs très concentrés mais également très rares, générant des valeurs extrêmement fortes avec une très faible probabilité, rend très incertaines les estimations obtenues à partir de peu de données. De plus, les outils de mesure ne sont pas des représentations parfaites de la réalité. Le chalut, tout comme l'acoustique laisse échapper un grand nombre de poissons. Il est donc important de disposer de beaucoup de points de mesure, et en particulier de beaucoup de stations, pour pouvoir établir une relation fiable entre les deux variables, celle-ci étant nécessairement bruitée.

La différence de profondeur entre les deux mers (environ 300 m en mer de Barents et 65 m en mer du Nord) peut avoir plusieurs implications. La longueur des câbles qui tirent le chalut étant proportionnelle à la profondeur, le chalut est plus loin derrière le bateau, et donc derrière le cône acoustique, pour les mers profondes.

D'un autre côté, lorsque la mer est peu profonde, le bruit du bateau a tendance à plus perturber les poissons (Mitson et Knudsen, 2003 [46]), ce qui peut également bruite la relation acoustique-chalut.

La taille de l'ouverture horizontale du chalut comparée au diamètre du cône acoustique est aussi un facteur limitant la corrélation entre les deux variables en mer du Nord. Comme cela a été vu dans le premier chapitre, le cône acoustique est défini partout par un angle constant de 7° quelque soit la zone étudiée. Le diamètre du cercle de base est donc directement proportionnel à la profondeur. Par ailleurs, les chaluts de fond utilisés dans les campagnes IBTS sont de grande ouverture. La distance entre les panneaux varie également en fonction de la profondeur mais se situe autour de 80 m en mer du Nord et de environ 25 m en mer de Barents. C'est souvent la distance entre les panneaux qui est considérée comme la largeur d'efficacité du chalut (et non la largeur du filet lui-même) en raison de l'effet rassembleur de ces derniers. La largeur du support des captures est donc environ 10 fois plus large que celle de l'acoustique en mer du Nord alors qu'elle est de dimension comparable en mer de Barents.

Les espèces démersales vivent en Norvège plus décollées du fond qu'en mer du Nord où elles ont souvent des comportements similaires aux espèces benthiques. Le comportement semi-pélagiques des démersaux de la mer de Barents permet de mieux les percevoir à l'acoustique. Lorsque les poissons sont collés au fond, ils se trouvent dans la "dead zone" acoustique, il est impossible de les enregistrer.

Les écosystèmes peuvent également être plus complexes en mer du Nord qu'en mer de Barents. Certaines espèces ayant des répartitions spatiales et des comportements très différents se mélangent et forment une abondance assez peu structurée.

Il est délicat de comprendre pourquoi la relation acoustique-chalut est plus satisfaisante en mer de Barents qu'en mer du Nord. Il s'agit probablement de la combinaison de multiples raisons aussi bien statistiques que biologiques et environnementales. Les différences en terme d'échantillonnage, en particulier le nombre de stations et la collecte de l'acoustique durant la nuit pour disposer d'un jeu de données plus important, me semble tout de même le point le plus facile à mettre en oeuvre

actuellement pour étudier plus en détail ce phénomène. S'affranchir, dans une certaine mesure, de l'instabilité statistique permettra de mesurer l'impact des différences en terme de profondeur, de l'état des stocks ou de la complexité des écosystèmes.

Conclusion

La démarche méthodologique employée dans cette thèse a été de dégager une relation acoustique-capture à partir des données, tout en ne prenant en compte qu'un minimum d'hypothèses basées sur les connaissances biologiques ou physiques. L'objectif recherché a été d'améliorer les estimations de la variable de capture sous-échantillonnée en utilisant au mieux les enregistrements acoustiques, échantillonnés de façon plus dense, comme variable auxiliaire.

Pour cela, le rôle de chacune des deux variables a été étudié dans différents modèles géostatistiques, dans le but d'obtenir une relation suffisamment simple et robuste pour être portée par l'ensemble des données de chaque campagne.

Pour se ramener à cette configuration bivariable simple, une étude de l'évolution du lien acoustique-capture en fonction de la verticale a été menée parallèlement. Les différentes couches acoustiques ont ainsi été combinées en une seule variable synthétique et opératoire.

Enfin, les résultats des modèles obtenus ont été confrontés à ceux de modèles n'utilisant pas l'information acoustique, pour répondre concrètement à la question de l'amélioration apportée par le sondeur dans les campagnes de chalutage de fond.

Analyse verticale

L'analyse des différentes couches acoustiques a permis de déterminer l'efficacité verticale effective du chalut de fond en mer de Barents. La somme des 40 mètres au dessus du fond est la hauteur la mieux corrélée avec les captures démersales. Les espèces classiquement considérées comme démersales, peuvent être, en effet, plus ou moins décollées du fond. Cette hauteur, en outre, est très stable d'années en années.

Les espèces démersales capturées par le chalut de fond plongent donc en moyenne de 40 mètres entre le passage du cône acoustique et celui du chalut (entre 3 et 15 minutes).

Il est également possible, en mer de Barents, de mesurer une efficacité verticale locale du chalut de fond, si l'on suppose les déplacements latéraux de poissons négligeables. Il est intéressant de constater que cette hauteur locale est corrélée à la profondeur. Le chalut semble capturer les poissons placés en moyenne moins haut dans la colonne d'eau, lorsque la profondeur est faible.

En mer d'Irlande, où la profondeur moyenne n'est que de 60 m, l'efficacité verticale du chalut se situe justement dans les tous premiers mètres au dessus du fond. La corrélation entre les captures et l'acoustique diminue rapidement pour les couches supérieures.

En mer du Nord, par contre, les corrélations sont mauvaises et la hauteur moyenne d'efficacité du chalut de fond est difficile à établir.

La somme de couches acoustiques ainsi obtenue s'est avérée mieux structurée spatialement que les combinaisons de couches obtenues par d'autres approches, comme les régressions linéaires, trop sen-

sibles aux valeurs fortes. Il serait intéressant de comparer les résultats avec ceux de régressions plus robustes, utilisant également des corrélations non-linéaires de type Spearman.

Une approche alternative a également montré qu'il était possible de retrouver le relief des cartes de proportions d'espèces capturées par le chalut, simplement à partir de la forme et de l'intensité des profils acoustiques. Par construction, la méthode fournit des estimations lissées. Pour des développements futurs, toutefois, il semble possible d'utiliser l'ensemble de la distribution statistique disponible des proportions de captures, plutôt qu'uniquement la moyenne, lors de la pondération des cartes de krigeage d'indicatrices (cf. partie 3.7.2). Ceci permettrait de prendre en compte l'ensemble de la variabilité dans l'estimation et d'obtenir de façon empirique une carte de la variance d'estimation des proportions. Il pourrait également être intéressant de poursuivre cette approche en considérant plusieurs classes de formes et d'intensités. Il faudrait alors utiliser un modèle disjonctif pour s'assurer de la cohérence entre classes. Multiplier le nombre de classes serait un autre moyen de réduire le lissage inhérent à la méthode.

Modèles bivariés acoustique-capture

Le modèle à résidu est une simplification du cokrigeage dans le cas où l'estimation de la variable cible est portée par la variable auxiliaire et corrigée par un résidu spatialement indépendant :

$$P(x) = \alpha A(x) + \beta + \varepsilon(x)$$

Ce modèle est porté par les données pour 9 des 18 campagnes disponibles. Le coefficient multiplicatif α de la variable acoustique est souvent inférieur à 1 en mer de Barents, ce qui indique que les enregistrements acoustiques sont en moyenne supérieurs aux captures. Ce coefficient n'est malheureusement pas assez stable d'années en années pour proposer une interprétation physique à cette différence. Le paramètre constant β est souvent quasi-nul en mer de Barents, ce qui montre qu'il n'y a pas de biais systématique entre l'acoustique et les captures. Ceci montre, qu'en mer de Barents, il n'y a pas une partie de l'abondance qui échappe au chalut ou à l'acoustique de façon systématique.

Le modèle à dérive externe est assez proche du modèle à résidu. La principale différence est que les paramètres α et β peuvent changer avec le voisinage, lorsque l'estimation est réalisée en voisinage glissant. Il s'est avéré alors que le paramètre multiplicatif de la dérive acoustique peut devenir localement négatif, ce qui n'a pas de sens. Les meilleurs résultats ont donc généralement été obtenus à voisinage unique.

Une perspective intéressante serait de construire un modèle bivariable utilisant chacune des deux variables mise sous forme disjonctive. Les tableaux de contingence croisés montrent en effet une bonne corrélation entre les classes pauvres, moyennes et riches de l'acoustique et des captures en mer d'Irlande et de Barents. Goovaerts (1997, [29]) propose dans une configuration de ce type d'effectuer un cokrigeage complet des toutes les indicatrices, par exemple par un modèle de corégionalisation. Le nombre de variables peut dans ce cas devenir assez important et le nombre de structures croisées rendre le modèle assez laborieux à mettre en pratique (pour quatre classes de captures et d'acoustique, il y aurait 16 structures croisées). Plusieurs simplifications, comme le choix d'un voisinage collocalisé, sont alors possibles. On peut rechercher aussi des facteurs du modèle bivariable spatialement orthogonaux pour alléger le système. Cependant, la mise sous forme d'indicatrices de la variable de capture nécessite un nombre de stations assez grand et ne sera réellement applicable qu'en mer de Barents.

Améliorations apportées par l'acoustique

L'utilisation de l'acoustique dans un modèle de cokrigeage a permis d'obtenir des cartes beaucoup plus détaillées qu'un modèle n'utilisant que les informations de capture. L'information disponible le long des transects permet de mieux délimiter les zones riches et pauvres, en évitant par exemple qu'une valeur forte ait une zone d'influence spatialement surestimée.

Du point de vue de l'estimation globale, l'estimation d'abondance sur l'ensemble du champ d'étude, l'acoustique apporte principalement un gain en variance. Les moyennes globales ne sont pas sensiblement modifiées. La baisse de la variance d'estimation globale de 50% en moyenne en mer de Barents et entre 10 et 40% en mer d'Irlande. Le gain est par contre inégal et souvent assez décevant en mer du Nord.

Les estimations combinant acoustique et chalutage sont également plus précises que celles obtenues par un modèle mono-variable du point de vue de la moyenne des carrés des erreurs de validation croisée, standardisées ou non, en mer de Barents. En mer du Nord et en mer d'Irlande par contre, les résultats obtenus par validation croisée sont moins satisfaisants et incitent même parfois à choisir le modèle n'utilisant que les données de pêche. On ne peut, en effet, passer outre les incohérences qui se sont dégagées dans ces zones entre acoustique et chalutage. Les modèles développés, parce que portés par les données, assurent la prévalance de ces dernières sur les modèles.

Différences entre les trois zones d'études : la mer du Nord, la mer de Barents et la mer d'Irlande

Disposer de données issues de trois mers a permis de comparer les résultats; des perspectives d'étude visant à expliquer les différences observées peuvent alors être proposées.

L'acoustique est mieux corrélée aux captures démersales en mer de Barents qu'en mer du Nord. Quelque soit la méthodologie mise en oeuvre, un modèle combinant acoustique et chalutage est plus performant en mer de Barents. Les approches des autres centres de recherche participants au projet mènent toutes à cette même conclusion. Les méthodes par régressions généralisées, réseaux de neurones et logique floue montrent par exemple qu'en mer du Nord la localisation des points et l'heure du jour sont des variables plus explicatives de l'abondance que l'acoustique enregistrée en stations (Beare et al. 2004, [4], Neville et al. 2004, [49], Mackinson et al. 2004, [36]).

La différence de résultats obtenus entre les trois zones peut-être due à la différence d'échantillonnage. Les campagnes en mer de Barents ont une densité d'enregistrements acoustiques quatre fois supérieures aux campagnes réalisées en mer du Nord, en raison de la non-collecte des données de nuit. Étant donnée la variabilité du phénomène étudié, le nombre de données disponibles est très important pour permettre de distinguer le lien structural existant entre les deux variables du bruit et des erreurs de mesure intrinsèques aux deux sources d'observations.

Le diamètre de la surface d'incidence du cône acoustique est largement plus petit que l'ouverture horizontale du chalut en mer du Nord, alors que les deux supports sont de taille comparable en mer de Barents. La taille du cône est uniquement proportionnelle à la profondeur, puisque dans toutes les campagnes, l'angle du cône acoustique fait 7° . Les chaluts utilisés, en revanche, ne sont pas les mêmes entre les deux mers. En particulier dans le sud de la mer du Nord, la zone la moins profonde, échantillonnée par IFREMER, l'ouverture du chalut est douze fois plus grande que le diamètre du cône acoustique. La probabilité que l'abondance capturée échappe au sondeur est alors

grande. Le sud de la mer du Nord est aussi la zone où la corrélation entre les deux variables est la plus faible. Ce type de problème d'échantillonnage peut disparaître avec l'apparition récente des sondeurs multi-faisceaux. Ces sondeurs envoient simultanément plusieurs faisceaux acoustiques, ce qui permet d'agrandir considérablement la surface d'incidence et le support des enregistrements. Il sera donc beaucoup plus rare que les poissons capturés aient échappé au cône acoustique.

Le nombre de campagnes disponibles dans chaque zone est enfin, un autre point important. Il existe des années atypiques où les estimations sont moins précises parce que les données sont très bruitées, en raison des conditions météorologiques, par exemple. Il est alors utile de disposer d'une longue série temporelle de campagnes. Seule la mer de Barents réunissait dans ce projet six années de données disponibles.

La diversité des espèces et le niveau général du stock sont également des facteurs importants. Les captures réalisées en mer de Barents sont essentiellement démersales et particulièrement riches en morues. La mer du Nord comprend des bancs très denses de pélagiques, envoyant des échos acoustiques extrêmement forts qui ont tendance à influencer fortement les estimations. Ces bancs se rencontrent souvent hors stations et il est difficile de juger de leur capturabilité. Le sud de la mer du Nord, une fois de plus, est particulièrement riche en bancs de harengs et de sprats.

Les démersaux de la mer de Barents vivent souvent assez décollés du fond et sont ainsi souvent qualifiés de *semi-pélagiques*. Cette propriété permet de limiter les confusions d'échos entre les poissons et le fond. De plus, les poissons vivant plus haut dans la colonne d'eau, il est possible que le chalut de fond ne permette pas à lui seul une bonne estimation du stock en espèces démersales. L'acoustique est dans ce cas particulièrement utile pour compenser les éventuelles pertes verticales qui sous-estimeraient l'état du stock. En mer du Nord, par contre, les espèces démersales, sûrement en raison de la faible profondeur, vivent très proches du fond et sont ainsi plus difficilement détectables par acoustique. Mais le chalut permet peut-être alors de suffisamment bien estimer le stock, sans qu'il soit nécessaire de recourir à l'acoustique.

Il est probable que le support de l'ESDU auquel étaient disponibles les données de cette étude soit trop large pour permettre une bonne corrélation acoustique-chalut. Il a été impossible de descendre à l'échelle du banc. Certains bancs ou ensemble de bancs sont capturables et d'autre non. Une perspective prometteuse serait donc de ne plus intégrer les échos acoustique par ESDU mais d'analyser les échos à l'échelle du banc. Godøet al. ont ainsi montré que l'identification *subjective* des bancs par un expert acousticien permettait d'augmenter sensiblement les corrélations avec les captures (Godøet al. 2004, [26]). Il est possible d'identifier quelques espèces selon la morphologie des bancs, de façon plus ou moins systématique. Des recherches ont récemment montré, par exemple, que l'utilisation de plusieurs fréquences aidaient à identifier les espèces perçues par le sondeur.

Cette thèse a permis de montrer que la construction d'un indice d'abondance combinant acoustique et capture par chalutage de fond était non-seulement possible mais permettait, dans un certain nombre de cas, un gain très satisfaisant en précision et en robustesse. Il est probable que les progrès effectués dans la technologie acoustique, comme l'apparition des sondeurs multifaisceaux, l'utilisation de plusieurs bandes de fréquences, ainsi que les algorithmes plus performants de détection du fond, permettront d'obtenir des corrélations beaucoup plus fortes entre le chalut et l'acoustique et d'améliorer encore ces estimations.

Références

- [1] ANONYME : Report of the ices advisory committee on fishery management - stock in the northeast arctic (subareas i and ii). Rapport technique, ICES Cooperative Research Report, 2002.
- [2] ANONYME : Stock in the irish sea. Rapport technique, ACFM Report, Octobre 2003. ICES, p27-36.
- [3] ANONYME : Stock in the north sea - overview. Rapport technique, ACFM Report, Octobre 2003. ICES, p173-174.
- [4] D. J. BEARE, D. G. REID, A. GREIG, N. BEZ, V. HJELLVIK, O.R GODØ, M. BOULEAU, J. van der KOOIJ, S. NEVILLE et S. MACKINSON : Positive relationships between bottom trawl and acoustic data. *ICES C.M.2004/R :24*, 2004.
- [5] M. BERMAN : The statistical properties of three noise removal procedures for multichannel remutely sensed data. Consulting report nsw/85/31/mb9, CSIRO Division of Mathematics and Statistics, 1985.
- [6] N. BEZ : *Statistiques individuelles et géostatistique transitive en écologie halieutique*. Thèse de doctorat, Ecole des Mines de Paris, 1997.
- [7] N. BEZ, D. G. REID, M. BOULEAU, D. J. BEARE, S. NEVILLE, Y. VÉRIN, O.R GODØ et M. GERRITSEN : Insight on fish reaction to the presence of trawl from the comparison of acoustic data recorded during and between bottom trawl stations. *ICES C.M.2004/R :14*, 2004.
- [8] M. BOULEAU : Rapport d'avancement de deuxième année de thèse. Rapport technique, École des Mines de Paris, Centre de géostatistique, 2003.
- [9] M. BOULEAU et N. BEZ : Simplification of cokriging by a markov-type model : application to combine acoustic and trawl data in fish surveys. *GeoENV 2004*, 2004.
- [10] M. BOULEAU, N. BEZ, V. HJELLVIK, O. R. GODØ, D. G. REID, D. J. BEARE, T. GEIG, M. AMSTRONG, H. GERRITSEN, S. MACKINSON, J. Von der KOOIJ, Y. VÉRIN et J. MASSÉ : Differences between near bottom biomass spatial structures observed in the ireland sea, the north sea and the berents sea in the recent years. *ICES C.M. 2003/ Q. 17*, 2003.
- [11] M. BOULEAU, N. BEZ, D. G. REID, O. R. GODØ et H. GERRITSEN : Testing various geostatistical models to combine bottom trawl stations and acoustic data. *ICES C.M. 2004/ R : 28*, 2004.
- [12] J-M. BUROCHE et G. SAPORTA : *L'analyse des données*. Que Sais-Je? P.U.F, 1980. 127p.
- [13] J. R. CARR : Comments about autokrigeability of cokriging based on an empirical study. *Math. Géol.*, 35(2), Feb. 2003.
- [14] E. CASTELIER : Dérive externe et régression linéaire. Rapport technique, Centre de géostatistique, 1993.

- [15] J-P CHILÈS et P. DELFINER : *Geostatistics, Modelling spatial uncertainty*. Wiley, New York, 1999. 695p.
- [16] W. G. COCHRAN : *Sampling techniques*. Wiley, New York, 1977. 428p.
- [17] D. DACUNHA-CASTELLE et M. DUFLO : *Probabilités et statistiques - Problèmes à temps fixe*. Masson, Paris, 1982. 214p.
- [18] P. DELFINER : Basic introduction to geostatistics. Cours c-78, Centre de géostatistique, 1979. 179 p.
- [19] A. J. DESBARATS : Geostatistical modeling of regionalized grain-size distributions using min/max autocorrelation factors. In *Geoenv III*, P. Monestier et Al., pages 441–452. Kluwer A. P., 2001.
- [20] A. J. DESBARATS et R. DIMITRAKOPOULOS : Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. *Math. Geol.*, 32(8), 2000.
- [21] E. DIDAY, J. LEMAIRE, J. POUGET et F. TESTU : *Éléments d'analyse des données*. Dunod Informatique, 1982.
- [22] A. GALLI et G. MEUNIER : *Study of a gaz reservoir using external drift method*, chapitre Geostatistical case studies, pages 105–119. D. Reidel Pub. Comp., 1987.
- [23] F GERLOTTO, M. SORIA et P. FRÉON : From two dimensions to three : the use of multibeam sonar for a new approach in fisheries acoustics. *Can. J. Fish. Aquat.*, 1999.
- [24] O. R. GODØ, D. SOMERTON et A. TOTLAND : Fish behaviour during sampling as observed from free floating buoys-application for bottom trawl survey assessment. *ICES C.M. 1999/J :10*, 1999.
- [25] O. R. GODØ et V. G. WESPESTAD : Monitoring changes in abundance of gadoids with varying availability to trawl and acoustic surveys. *ICES Journal of Marine Science*, 50:39–51, 1993.
- [26] O.R. GODØ, V. HJELLVIK, A. GREIG, D. J. BEARE et J. van der KOOIJ : Can subjective evaluation of echograms improve correlation between bottom trawl and acoustic densities? *ICES C.M.2004/R :23*, 2004.
- [27] P. GOOVAERTS : Spatial orthogonality of the principal components computed from regionalised variables. *Math. Geol.*, 25(3), 1993.
- [28] P. GOOVAERTS : Study of spatial relationships between two sets of variables using multivariate geostatistics. *Geoderma*, 62:93–107, 1994.
- [29] P. GOOVAERTS : *Geostatistics for Natural Resources Evaluation*. Applied Geostatistics. Oxford, 1997. 483p.
- [30] P. GOOVAERTS : Ordinary cokriging revisited. *Math. Geol.*, 30(1), 1998.
- [31] A. GREIG et D. G. REID : Fish in the back step : can we be missing a signal from within this layer? *ICES C.M.2004/R :17*, 2004.
- [32] P. GUIBLIN : *Analyse géostatistique de campagnes (acoustique et chalutage) sur le hareng écossais*. Thèse de doctorat, Ecole des Mines de Paris, 1997.
- [33] N.O. HANDEGARD, K. MICHALSEN et D. TJØSTHEIM : Avoidance behaviour in cod (gadus morhua) to a bottom-trawling vessel. *Aquatic Living Ressources*, 2003.
- [34] V. HJELLVIK, K. MICHALSEN, A. AGLÉN et O. NAKKEN : An attempt at estimating the effective fishing height of the bottom trawl using acoustic survey recordings. *ICES Journal of Marine Science*, 60:967–979, 2003.
- [35] K.A. JOHANESSON et R.B. MITSON : Fisheries acoustics : a practical manual for aquatic biomass estimation. Rapport technique, FAO, 1983.

- [36] J. MACKINSON, S. van der Kooij et S. NEVILLE : The fuzzy relationship between acoustic and trawl surveys in the north sea. *ICES C.M.2004/R :04*, 2004.
- [37] D. N. MACLENNAN, P.J. COPLAN, E. ARMSTRONG et E.J. SIMMONDS : Experiments on the discrimination of fish and seabed echoes. *ICES Journal of Marine Science*, 2004.
- [38] D. N. MACLENNAN, P. G. FERNANDES et J. DALEN : A consistent approach to definitions and symbols in fisheries acoustic. *ICES Journal of Marine Science*, 59:365–369, 2002.
- [39] D. N. MACLENNAN et E.J. SIMMONDS : *Fisheries Acoustics*. Chapman & Hall, 1992.
- [40] J. MASSÉ, C. KOUTSIKOPOULOS et W. PATTY : The structure and spatial distribution of pelagic fish schools in multispecies clusters : an acoustic study. *ICES Journal of Marine Science*, 53:155–160, 1996.
- [41] G. MATHERON : La théorie des variables régionalisées et ses applications. Cahiers du CMM Fascicule 5, Ecole des Mines de Paris, 1970.
- [42] G. MATHERON : Estimer et choisir. Cahiers du CMM Fascicule 7, Ecole des Mines de Paris, 1978.
- [43] G. MATHERON : Recherche de simplification d'un problème de cokrigage. Rapport N-628, Centre de Géostatistique, 1979.
- [44] G. MATHERON : Pour une analyse krigeante des données régionalisées. Rapport N-732, Centre de Géostatistique, 1982.
- [45] K. MICHALSEN, A. AGLÉN, D. SOMMERTON, I. SVELLINEN et J.T. ØVREDAL : Quantifying the amount of fish unavailable to a bottom trawl by use of an upward looking transducer. *ICES C.M. 1999/J :08*, 1999.
- [46] R.B. MITSON et H.P. KNUDSEN : Causes and effects of underwater noise on fish abundance estimation. *Aquatic Living Resources*, 2003.
- [47] R.A. MOYEED et A. PAPRITZ : An empirical comparison of kriging methods for non-linear spatial point prediction. Rapport technique, Institute of Terrestrial Ecology soil Physics, ETH, Zürich, 2000.
- [48] R. MUIÑO, P. CARRERA, P. PETITGAS, D. J. BEARE, S. GEORGAKARAKOS, J. HARALAMBOUS, M. IGLESIAIS, B. LIORZOU, J. MASSÉ et D.G. REID : One the relation between schools, cluster of schools, and abundance in pelagic fish stocks. *ICES Journal of Marine Science*, 60:164–175, 2003.
- [49] S. NEVILLE, V HJELLVIK, S. MACKINSON et J. van der KOOIJ : Using neural network to combine acoustic and trawl data in the barents and the north seas. *ICES C.M.2004/R :05*, 2004.
- [50] P. PETITGAS : *Contributions géostatistiques à la biologie des pêches maritimes*. Thèse de doctorat, Ecole des Mines de Paris, 1991.
- [51] P. PETITGAS : Use of a disjunctive kriging to model areas of high pelagic fish density in acoustic fisheries surveys. *Aquat. Living Resour.*, 6:201–209, 1993.
- [52] P. PETITGAS : A method for the identification and characterization of clusters of schools along the transect lines of fisheries-acoustic surveys. *ICES Journal of Marine Science*, 60:872–884, 2003.
- [53] P. PETITGAS et J. LEVENEZ : Spatial organization of pelagic fish : echogram structure, spatio-temporal condition, and biomass in senegalese waters. *ICES Journal of Marine Science*, 53:147–153, 1996.
- [54] P. PETITGAS, D. REID, P. CARRERA, M. IGLESIAIS, B. GEORGAKARAKOS, B. LIORZOU et J. MASSÉ : One the relation between schools, cluster of schools, and abundance in pelagic

- fish stocks. *ICES Journal of Marine Science*, 58:1150–1160, 2001.
- [55] D.G. REID : Report on echo trace classification. Rapport des recherches collectives 238, ICES Coop. Res. Report, 2000.
- [56] J. RIVOIRARD : *Introduction to disjunctive kriging and nonlinear geostatistics*. Spatial Information Systems. Oxford : Clarendon Press., 1994. 181p.
- [57] J. RIVOIRARD : Which models for collocated cokriging? *Math. Geol.*, 33(2), 2001.
- [58] J. RIVOIRARD : On the structural link between variables in kriging with external drift. *Math. Geol.*, 34(7), 2002.
- [59] J. RIVOIRARD, J. SIMMONDS, K.G. FOOTE, P. FERNANDEZ et N. BEZ : *Geostatistics for estimating fish abundance*. Blackwell Science, 2000. 206 p.
- [60] H. SAITO et P. GOOVAERTS : Geostatistical interpolation of censored data in dioxin-contaminated site. *Environ. Sci. Technol.*, 34:4228–4235, 2000.
- [61] G. SAPORTA : *Probabilités, Analyse des données et Statistique*. Editions Technip, Paris, 1990. 493p.
- [62] D. E. SHAPIRO et P. SWITZER : Extracting time trend from multiple monitoring sites. Rapport technique, Department of statistics, Stanford Univerity, 1989.
- [63] L. E. SHMARYAN et A. G. JOURNAL : Two markov-type models and their application. *Math. Geol.*, 31(8), 1999.
- [64] E. SICARD : *Choix de composantes optimales pour l'analyse spatiale et la modélisation : application aux pluies mensuelles du Nordeste brésilien*. Thèse de doctorat, Université de Montpellier I, 2004.
- [65] Y. SIMARD, D. MARCOTTE et K. NARAGHI : Three-dimensional acoustic mapping and simulation of krill distribution in the saguenay - st. lawrence marine park whale feeling ground. *Aquatic Living Resources*, 2003.
- [66] E. J. SIMMONDS, F. ARMSTRONG et P. J. COPLAND : Species identification using wideband backscatter with neural network and discriminant analysis. *ICES Journal of Marine Science*, 53:189–195, 1996.
- [67] E. J. SIMMONDS, M. BAILAY, R. TORESEN, E. TORSTENSEN, E. PEDERSEN, E. GÖTZE, P. FERNANDES et A.S. COUPERUS : 1998 ices coordinated acoustic survey of ices divisions 3a, 6a, 4b and 6a(north). *ICES CM 1999/J :16*, 1999.
- [68] P. SWITZER et A. A. GREEN : Min/max autocorrelation factors for multivariate spatial imagery. Rapport technique, Department of Statistics, Stanford Univerity, 1984.
- [69] Y. VÉRIN : Analysis of the ibts trawl behaviours : possible enhanced standardization of trawling conditions. *ICES C.M. 1995/B.10*, 1995.
- [70] H. WACKERNAGEL : *Multivariate Geostatistics*. Springer, 2nd ed. édition, 1998. 291p.
- [71] H. WACKERNAGEL, P. PETITGAS et Y. TOUFFAIT : Overview of methods for coregionalization analysis. In *Geostatistics, Vol I*, Armstrong M., pages 409–420. Kluwer Academic Publishers, 1989.
- [72] S. WALSH : Efficiency of bottom sampling trawls in deriving survey abundance indices. *NAFO Sci. Coun. Studies*, 1996.
- [73] W. XU, T. TRAN, R. M. SRIVASTAVA et A. G. JOURNAL : Integrating seismic data in reservoir modeling : the collocated cokriging alternative. SPE paper 24742, 1992.

Annexe A

Étude de la pertinence des outliers et question du seuillage

Dans le cadre du projet CATEFA, plusieurs sortes de seuillages des données ont été envisagées et cette annexe vise à les présenter brièvement.

Le terme de seuillage sous-entend, dans cette étude, que les données supérieures à un seuil fixé n'ont pas été prises en compte dans la modélisation. Il existe des formes de seuillages où les données trop fortes sont ramenées au niveau du seuil choisi, mais cette approche n'a pas été considérée, ici.

A.1 Contexte

La nécessité du seuillage vient de la forte asymétrie des distributions statistiques de chacune des deux variables : l'acoustique et le chalutage. En mer du Nord, par exemple, plus de 65% de l'énergie totale et 26% du nombre total de pélagiques ont été collectés dans 3% des stations. En mer de Barents, sur 5 des 6 campagnes disponibles, 80% des pélagiques ont été pêchés dans seulement 3% des stations.

En ce qui concerne l'acoustique, on peut se demander si les valeurs fortes représentent des bancs de poissons ou des valeurs erronées. Il arrive en effet couramment que le sondeur interprète comme du poisson le fond de la mer et engendre ainsi des valeurs extrêmement élevées. Un long travail de vérification avec les laboratoires responsables de la collecte des données acoustiques a donc été effectué. Chaque valeur *anormalement* forte a été vérifiée en regardant son échogramme. Pour certaines campagnes, près d'une donnée sur trois vérifiées s'est avérée être erronée. Après ce contrôle systématique des valeurs les plus suspectes, les valeurs fortes restantes ont donc été définitivement considérées comme représentant du poisson. En discutant avec des acousticiens, il est paraît-il courant de voir près de la moitié de la biomasse totale répartie en quelques points de l'espace.

La figure A.1 représente, par exemple, l'échogramme d'un banc enregistré lors d'une campagne écosaise d'une valeur d'environ $140\,000\text{ m}^2/\text{m.n}^2$, alors que la moyenne des autres données enregistrées est de seulement $7\text{ m}^2/\text{m.n}^2$. Ce banc mesure environ 25 m de haut sur 40 m de long et fournit une réponse 20 000 fois supérieure aux autres.

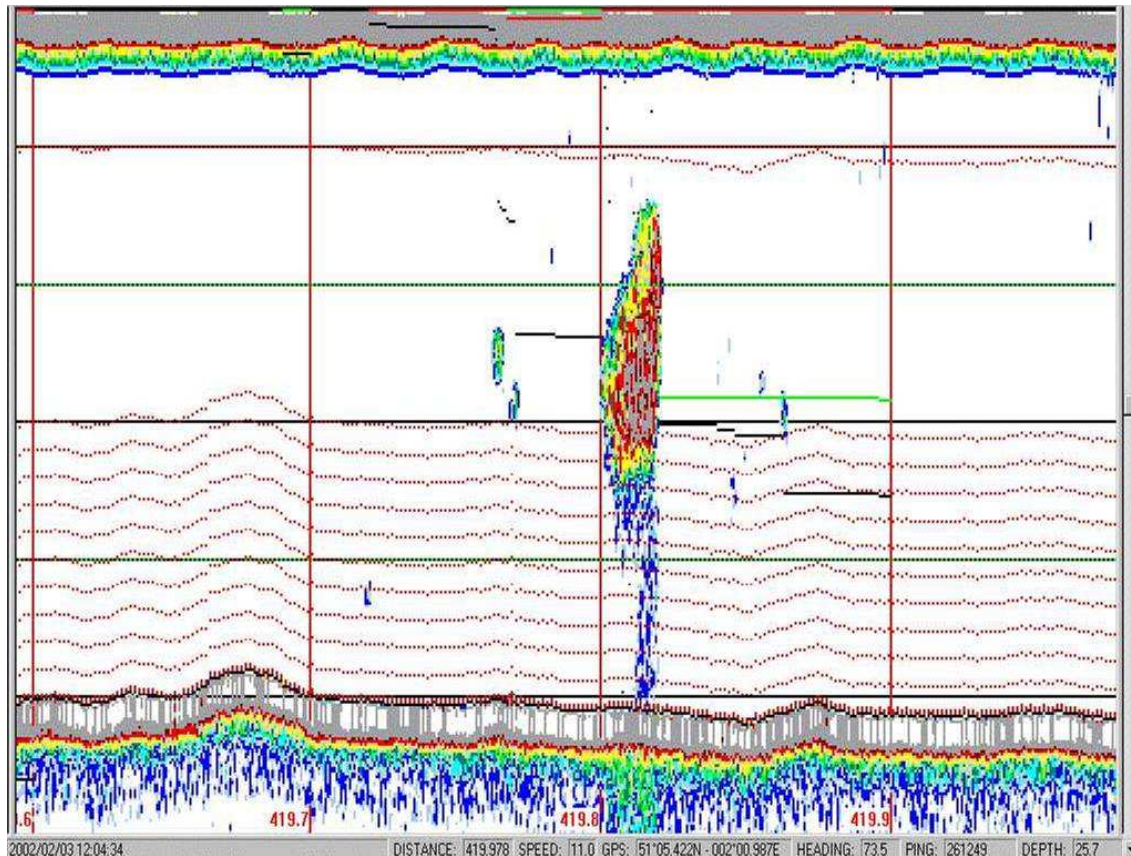


FIG. A.1 – Exemple de banc de poisson enregistré par acoustique fournissant une réponse 20 000 fois supérieure à la moyenne des données.

Le problème réside dans le fait que ces valeurs fortes semblent pouvoir apparaître dans n'importe quelle zone de la mer, aussi bien en zone riche que pauvre. Elles ont donc un effet destructurant très fort sur les variogrammes. Le comportement de ces bancs (a priori pélagiques) est assez mal connu et ne semblent pas lié de façon directe à l'état du stock (Petitgas et al. 2001 [54]). Il est, par exemple, difficile de savoir si ces bancs très concentrés de poissons sont capturables ou non. Des valeurs fortes apparaissent également dans les données de captures, mais restent d'un ordre de grandeur inférieur. Il est également délicat de les supprimer puisqu'il s'agit d'une part importante de la biomasse totale.

Enfin, même une fois mis à part les quelques bancs dont les valeurs sont très élevées, les distributions statistiques restent très asymétriques.

A.2 Seuillage brut

Plus on seuille à un niveau bas les données, que ce soit en mer de Barents ou en mer du Nord, plus l'acoustique devient structurée spatialement. Mais est-il intéressant d'estimer la répartition

spatiale des valeurs faibles, si une grande partie de l'abondance dépend de celle des valeurs les plus fortes ? La question du seuillage brut des données est toujours une question délicate en statistique, non seulement dans l'estimation du biais engendré mais aussi dans le choix souvent arbitraire du niveau de seuillage.

Cette approche a donc vite été abandonnée.

A.3 Acoustique sans pélagique

Une sélection moins arbitraire, permettant de conserver la partie de l'acoustique la plus susceptible d'être corrélée aux captures a alors été mise en place.

Elle consiste à analyser les échogrammes de façon systématique pour supprimer des bases de données les échos ayant des formes typiques de bancs pélagiques. Ces procédures ont été réalisées, non pas à la main, mais de façon systématique par des algorithmes de reconnaissance de forme. L'idée est alors que l'acoustique restante est de type démersal et donc sûrement mieux corrélée aux captures du chalut de fond. De plus, les valeurs extrêmes provenant souvent de bancs pélagiques sont ainsi mises de côté. Il y a cependant plusieurs inconvénients à cette approche. Un certain nombre de pélagiques sont capturés par le chalut de fond. Il est difficile de savoir d'après la forme du banc si celui-ci est capturable ou non. Les données seuillées de cette manière ne se sont le plus souvent pas avérées mieux structurées spatialement ou mieux corrélées aux captures. L'exception est l'Irlande où l'acoustique sans pélagique est très bien structurée spatialement contrairement aux données brutes. Cependant, la procédure a supprimé dans cette zone près de 90% de l'acoustique, alors que les espèces démersales sont importantes dans la zone. Cette approche n'a donc pas non plus été poursuivie.

A.4 Sélection sur la température

En mer de Barents, des valeurs très fortes étaient systématiquement observées en bordure du champ, en particulier à l'est. Les spécialistes norvégiens ont alors expliqué que cette spécificité provenait de la température de l'eau qui engendrait la création d'un écosystème particulier, particulièrement propice à certaines espèces. Cette zone étant considérée pour les experts halieutiques comme très spécifique, a été jugée comme non représentative dans le cadre de l'étude de la corrélation acoustique-chalut.

Les campagnes sont effectuées en mer de Barents entre janvier et mars, selon les années. Durant cette saison, la température de fond tombe dans cette partie en dessous de zéro. Il a donc été décidé que la zone où la température est négative serait mise de côté dans les bases de données.

Les données de température n'étant disponibles qu'en stations, un krigeage ordinaire a été utilisé pour délimiter cette zone de façon précise. Les variogrammes des mesures de température étaient assez réguliers d'années en années, généralement sans effet de pépité et avec une dérive linéaire. Les zones délimitées par le krigeage ont ainsi permis de supprimer la plus grande partie des valeurs fortes. Il s'est cependant avéré que certaines valeurs fortes se situaient *en bordure* de cette zone. Les contours obtenus sont indiqués dans la figure A.2 sur lesquels est superposée une représentation proportionnelle des données acoustiques. On voit que la plupart des valeurs fortes sont dans la surface délimitée, mais pas toutes. Les valeurs fortes restantes sont donc, une fois la sélection effectuée,

assez isolées. Ceci a tendance à provoquer une déstructuration encore plus forte. Il est cependant difficile de décider d'élargir arbitrairement la zone, en supprimant par exemple les points dès que la température est en dessous de 1°C . La sélection reviendrait alors à un seuillage brut des valeurs gênantes.

La sélection a donc été utilisée telle qu'elle. Elle présente en effet l'avantage de supprimer de façon non arbitraire une grande partie des valeurs fortes, très déstructurantes.

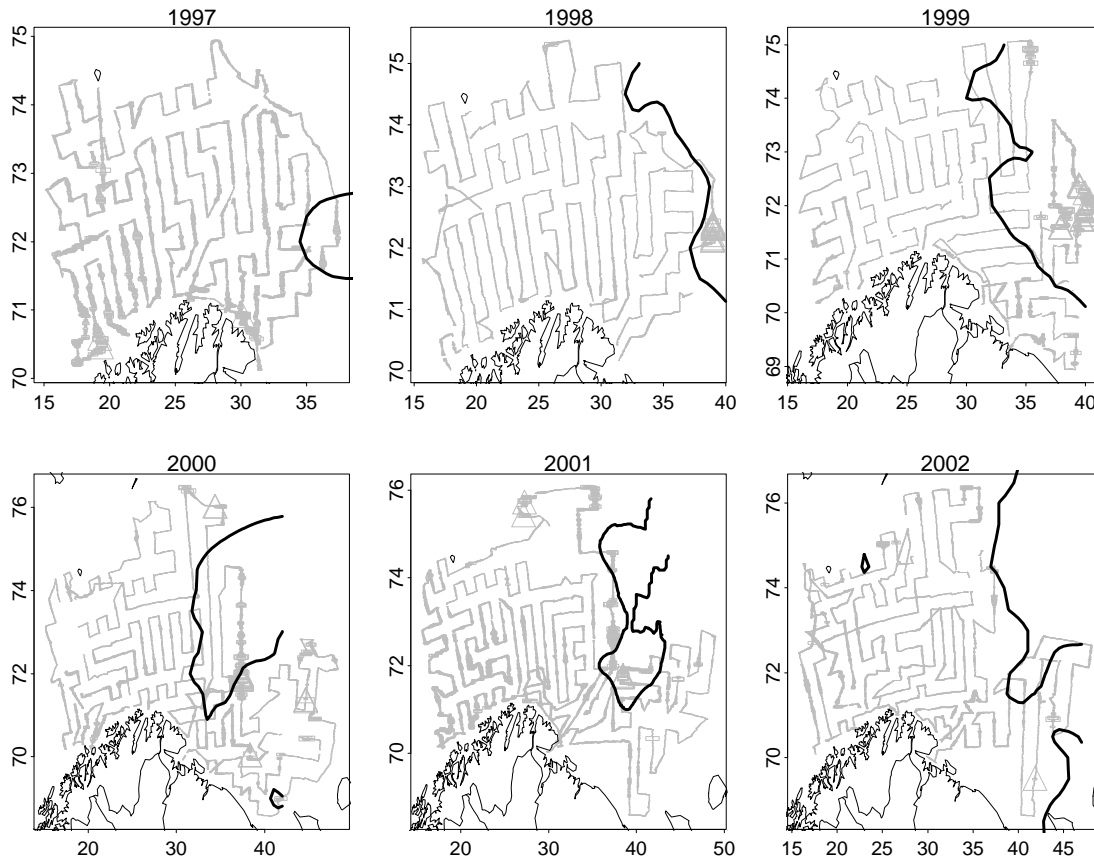


FIG. A.2 – Représentations proportionnelle des données acoustiques pour chacune des 6 campagnes disponibles en Norvège, avec la zone de températures négatives délimitée par la courbe grise.

Annexe B

Compléments sur la méthode des Min/Max autocorrelation factors

Soit $A = (A_1, \dots, A_k)$ l'ensemble des variables initiales (par exemple les variables acoustiques). On se place dans le cas stationnaire. On note la fonction de covariance : $C(h) = Cov[A(x), A(x+h)] \in \mathfrak{R}^p \times \mathfrak{R}^p$
Alors les Y_i tels que $Y_i = {}^t\alpha_i Z A(x)$ sont les Min/Max autocorrelation vectors avec α_i tel que :

$$\begin{aligned}\rho[Y_1(x), Y_1(x+\Delta)] &= \max_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x+\Delta)] \\ \rho[Y_2(x), Y_2(x+\Delta)] &= \max_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x+\Delta)] \\ \text{avec } \rho[{}^t\alpha_i A(x), {}^t\alpha_j A(x+\Delta)] &= 0 \text{ pour } j < i \\ \dots \\ \rho[Y_k(x), Y_k(x+\Delta)] &= \min_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x+\Delta)]\end{aligned}$$

Le problème peut également s'écrire matriciellement :
Trouver α_i tel que :

$$\max_{\alpha} \frac{{}^t\alpha C(\Delta)\alpha}{{}^t\alpha C(0)\alpha}$$

$$\text{avec } {}^t\alpha_i C(\Delta)\alpha_j = 0 \text{ pour } i \neq j$$

Les α_i sont alors les vecteurs propres du produit de la matrice de covariance au pas Δ par l'inverse de la matrice de variance, ils vérifient :

$${}^t\alpha_i C(\Delta)C(0)^{-1} = \Lambda^t\alpha_i \tag{B.1}$$

avec $C(\Delta) = Cov[A(x), A(x+\Delta)]$ et $C(0) = Cov[A(x), A(x)]$, la matrice de variance de A.

Preuve

Recherche du premier facteur

Le problème s'écrit donc :

$$\max_{\alpha} \rho[{}^t\alpha A(x), {}^t\alpha A(x + \Delta)] = \frac{{}^t\alpha C(\Delta)\alpha}{{}^t\alpha C(0)\alpha}$$

Si l'on se fixe en plus la contrainte de normalisation : ${}^t\alpha C(0)\alpha = 1$ la fonction de Lagrange se simplifie :

$$L(\alpha, \lambda) = {}^t\alpha C(\Delta)\alpha - \lambda({}^t\alpha C(0)\alpha - 1)$$

A l'optimum, on a :

$$\begin{aligned} \frac{\partial L(\alpha, \lambda)}{\partial \alpha} &= 0 \\ \Leftrightarrow \begin{cases} 2{}^t\alpha C(\Delta) = 2\lambda {}^t\alpha C(0) \\ {}^t\alpha C(0)\alpha = 1 \end{cases} \\ \Leftrightarrow \begin{cases} {}^t\alpha C(\Delta)C(0)^{-1} = \lambda {}^t\alpha \\ {}^t\alpha C(0)\alpha = 1 \end{cases} \end{aligned}$$

On retrouve bien l'équation (B.1).

α est un vecteur propre à gauche de la matrice $C(\Delta)C(0)^{-1}$. Cette matrice n'est pas nécessairement symétrique (puisque'il s'agit d'un produit de matrices), prendre ses vecteurs propres à gauche et à droite n'est donc pas toujours équivalent. De plus, comme on cherche à maximiser ${}^t\alpha C(\Delta)\alpha$, on prendra le vecteur propre associé à la plus grande valeur propre λ .

En effet, on a :

$$\begin{aligned} {}^t\alpha C(\Delta)C(0)^{-1} &= \lambda {}^t\alpha \\ \Leftrightarrow {}^t\alpha C(\Delta)C(0)^{-1}C(0)\alpha &= \lambda {}^t\alpha C(0)\alpha \text{ en multipliant à droite par } C(0)\alpha \\ \Leftrightarrow {}^t\alpha C(\Delta)\alpha &= \lambda, \text{ car } {}^t\alpha C(0)\alpha = 1 \end{aligned}$$

La covariance entre les facteurs au pas Δ est exactement égale à la valeur propre λ .

Recherche des autres facteurs

Par rapport au premier facteur, les suivants doivent en plus obéir à une deuxième contrainte, celle d'être sans corrélation (en 0 et au pas Δ) avec les facteurs précédents. Les matrices $C(\Delta)$ et $C(0)$ étant symétriques, on observe que si x_1 et x_2 sont deux vecteurs propres de $C(\Delta)C(0)^{-1}$ réciproquement associés aux valeurs propres λ_1 et λ_2 , alors nécessairement ${}^t x_1 C(0)x_2 = 0$ et ${}^t x_1 C(\Delta)x_2 = 0$. Le fait d'être un vecteur propre de $C(\Delta)C(0)^{-1}$ est donc non seulement nécessaire mais suffisant.

En effet :

$$\begin{aligned}
& \begin{cases} {}^t x_1 C(\Delta) C(0)^{-1} = \lambda_1 {}^t x_1 \\ {}^t x_2 C(\Delta) C(0)^{-1} = \lambda_2 {}^t x_2 \end{cases} \\
& \Leftrightarrow \begin{cases} {}^t x_1 C(\Delta) = \lambda_1 {}^t x_1 C(0) \\ C(\Delta) x_2 = \lambda_2 C(0) x_2 \end{cases} \\
& \Rightarrow \begin{cases} {}^t x_1 C(\Delta) x_2 = \lambda_1 {}^t x_1 C(0) x_2, \text{ en multipliant à droite par } x_2 \\ {}^t x_1 C(\Delta) x_2 = \lambda_2 {}^t x_1 C(0) x_2, \text{ en multipliant à gauche par } {}^t x_1 \end{cases} \\
& \Rightarrow \begin{cases} \text{soit } \lambda_1 = \lambda_2 \\ \text{soit } {}^t x_1 C(0) x_2 = 0 \end{cases}
\end{aligned}$$

Donc si $\lambda_1 \neq \lambda_2$ alors la covariance entre les facteurs en zéro $Cov[{}^t x_1 A(x), {}^t x_2 A(x)]$ est nulle et donc leur corrélation. On obtient de la même façon que si $\lambda_1 \neq \lambda_2$ alors ${}^t x_1 C(\Delta) x_2 = 0$.

Ainsi les vecteurs propres à gauche de la matrice non-symétrique $C(\Delta)C(0)^{-1}$ fournissent bien les coefficients recherchés des MAF.

Une démonstration similaire peut être trouvée dans Berman (1985, [5]).

Choix du pas de distance Δ

L'idée des MAF est de constituer non seulement un ensemble de facteurs les moins corrélés spatialement possible mais aussi de décomposer le signal initial en des combinaisons linéaires structurées et des combinaisons linéaires pépétiques. Il faut donc choisir un pas Δ tel qu'il permette le meilleur contraste possible entre elles. En général, Δ est donc choisi très petit, voire le plus petit possible. Le pas choisi dans l'application est celui de l'ESDU. Les premiers MAF vont ainsi maximiser la corrélation pour les toutes petites distances alors que les derniers MAF seront les combinaisons linéaires les moins corrélées pour cette même petite distance. Ceux-ci auront donc toutes les chances de correspondre au bruit contenu dans les données. Choisir un Δ plus grand que la distance minimale entre les données présente le risque que les facteurs obtenus n'aient pas les mêmes propriétés de corrélation pour les distances inférieures à Δ . Dans la pratique, il peut être intéressant de se fixer une tolérance autour de Δ , en particulier si les données ne suivent pas un échantillonnage régulier. On cherche alors à maximiser une corrélation moyenne sur un intervalle donné.

Application dans le cas de modèles multivariables simples et comparaison avec une ACP classique

Le modèle en corrélation intrinsèque

Si l'on se place dans la configuration très particulière d'un modèle multivariable en corrélation intrinsèque, toutes les variables ont la même structure à un coefficient multiplicatif près, ainsi que toute combinaison linéaire des variables. La matrice $C(\Delta)C(0)^{-1}$ est donc une constante fois la

matrice identité. On est dans la configuration où toutes les valeurs propres sont égales, les MAF sont donc égaux aux variables initiales, aucune transformation n'est effectuée.

Par contre une ACP classique, dans laquelle les facteurs sont les vecteurs propres de la matrice de variance-covariance $C(0)$ (multipliée par la matrice diagonale de la métrique), permet ici non seulement d'orthogonaliser les variables en zéro mais aussi à tout pas de distance (Goovaerts 1993 [27]).

Le modèle linéaire de corégionalisation

On se place dans le cas simple de deux composantes :

$$A = B_1 + B_2, \text{ non corrélées, on a donc } V = V_1 + V_2 \text{ et } \gamma = \gamma_1 + \gamma_2$$

où V , V_1 et V_2 sont les matrices de variance-covariance de A , B_1 et B_2 .

De plus, on note : $Cov[B_1(x), B_1(x+h)] = V_1 c_1(h)$ et $Cov[B_2(x), B_2(x+h)] = V_2 c_2(h)$ avec $c_1(h)$ et $c_2(h)$ les fonctions de covariance (scalaires) de B_1 et B_2 . On a alors :

$$Cov[A(x), A(x+h)] = V_1 c_1(h) + V_2 c_2(h) \quad (\text{B.2})$$

Ces notations établies, en posant $V_2 = V - V_1$ on obtient :

$$C(\Delta)C(0)^{-1} = c_2(\Delta)I + [c_1(\Delta) - c_2(\Delta)]V_1V^{-1} \quad (\text{B.3})$$

Les MAF sont donc les vecteurs propres de V_1V^{-1} associés aux valeurs propres $\Lambda = [c_1(\Delta) - c_2(\Delta)]\Lambda_1 + c_2(\Delta)$, où Λ_1 est la matrice diagonale des valeurs propres de V_1V^{-1} .

Notons que par symétrie on aurait pu poser $V_1 = V - V_2$, les MAF sont donc également les vecteurs propres de V_2V^{-1} associés aux valeurs propres $[c_1(\Delta) - c_2(\Delta)]\Lambda_2 + c_1(\Delta)$.

A partir de la relation B.2, si l'on considère un autre pas de distance Δ' on déduit que :

$$\begin{aligned} Cov[Y_i(x), Y_j(x + \Delta')] &= {}^t\alpha_i C(A(x), A(x + \Delta'))\alpha_j \\ &= {}^t\alpha_i (V_1 c_1(\Delta') + V_2 c_2(\Delta'))\alpha_j \\ &= {}^t\alpha_i (V_1 (c_1(\Delta') - c_2(\Delta')) + V c_2(\Delta'))\alpha_j \text{ avec } V_2 = V - V_1 \\ &= {}^t\alpha_i V_1 \alpha_j (c_1(\Delta') - c_2(\Delta')) + {}^t\alpha_i V \alpha_j c_2(\Delta') \\ &\text{avec } {}^t\alpha_i V \alpha_j = 0 \text{ si } i \neq j \text{ (} = 1 \text{ sinon)} \\ &\text{et } {}^t\alpha_i V_1 \alpha_j = 0 \text{ si } i \neq j \text{ (} = \lambda_i \text{ si } i=j) \end{aligned}$$

C'est une matrice diagonale. Donc $\forall \Delta'$, $\rho[Y_i(x), Y_j(x + \Delta)] = 0$, si $i \neq j$, les facteurs sont sans corrélation spatiale.

Si l'on considère un facteur particulier $Y_i(x) = {}^t\alpha_i A(x)$ et son variogramme

$$\begin{aligned} c_{Y_i}(h) &= c_2(h) + {}^t\alpha_i V_1 \alpha_i (c_1(h) - c_2(h)) \\ &= c_2(h) + \lambda_i (c_1(h) - c_2(h)) \\ &= (1 - \lambda_i) c_2(h) + \lambda_i c_1(h) \end{aligned} \quad (\text{B.4})$$

On voit bien dans cette dernière formulation que les MAF décomposent la structure de A en pondérant chacune des deux composantes par les valeurs propres associées. Si par exemple B_1 correspond à une structure courte portée (ou une pépite) et B_2 une composante grande portée, on voit que le premier MAF associé à la plus grande valeur propre aura quasiment la structure c_1 tandis que le dernier facteur aura une structure beaucoup plus proche de c_2 .

Pour l'ACP, les coefficients des facteurs seront les vecteurs propres de la matrice $(V_1 + V_2)$. Ils seront uniquement orthogonaux pour la distance $\Delta = 0$ mais pas spatialement. Goovaerts (1993 [27]) montre que dans plusieurs configurations les facteurs de l'ACP sont spatialement orthogonaux, comme lorsque la première structure est une pépite avec une matrice de corégionalisation diagonale, c'est à dire lorsque les bruits des variables sont indépendants, comme une erreur de mesure par exemple. Enfin, les facteurs de l'ACP ne sont pas ordonnés par structure mais en fonction de leur variance totale.

Introduite par Matheron sous le nom d'analyse krigeante (1982 [44]), l'analyse en composantes principales des matrices de corégionalisation de chaque échelle permet une description spatiale des relations entre les variables pour chaque "bande de fréquence", une illustration détaillée d'un cas réel est donnée par Goovaerts (1993 [27]) qui montre l'intérêt descriptif d'une telle approche permettant d'étudier l'évolution des corrélations entre les variables pour différentes échelles spatiales, par rapport aux méthodes classiques d'analyse des données.

D'autres propriétés des MAF (lissage, invariance,...) ont été étudiées par Shapiro et Switzer (1989 [62]). Desbarat (2001 [19]) étudie les différences de structures simples et croisées entre les facteurs de l'ACP et les MAF d'un point de vue empirique sur une application géologique.

Résumé

Actuellement l'état des stocks de poissons des mers européennes est en grande partie estimé à partir des campagnes scientifiques de chalutage de fond. Durant ces campagnes sont également collectés des enregistrements acoustiques reflétant la densité de poissons présente à chaque instant sous le navire.

Comment tirer partie de cette source d'information beaucoup plus dense pour améliorer les estimations d'abondance?

Prenant en compte les corrélations spatiales entre les points de mesure, la géostatistique s'est avérée un outil efficace pour évaluer les ressources halieutiques.

Dans une première partie, le comportement du cokrigeage bivariable pour une situation de sous-échantillonnage de la variable cible, est étudié pour différentes configurations de structures spatiales, en commentant le rôle joué dans le modèle par chacune des deux variables : la mesure acoustique et la donnée de pêche. L'application à des données, dont les fortes variabilités rendent souvent délicate l'inférence d'un variogramme, soulève plusieurs questions concrètes de mise en pratique.

L'acoustique fournit une information discrétisée sur toute la hauteur d'eau, alors que le chalut n'a pas de support vertical véritablement connu en raison des déplacements de poissons. Une étude détaillée de la dimension verticale fournie par les sondeurs permet de décrire et comprendre les relations entre les deux types de mesures grâce à plusieurs outils d'analyse de données. En mer de Barents, l'intégration de l'acoustique sur les 40 premiers mètres au-dessus du fond se révèle une variable aux propriétés satisfaisantes de robustesse et d'autocorrélation spatiale pour la modélisation.

Enfin, les performances des modèles sont comparées pour des données de trois mers différentes : la mer du Nord, la mer d'Irlande et la mer de Barents. Le gain en précision apportée par la variable auxiliaire, la variable acoustique, est analysé pour chaque plan d'échantillonnage. Alors que l'information acoustique permet une baisse de la variance d'estimation de 20 à 40 % en Norvège et en Irlande, les améliorations en mer du Nord restent minimales.

Summary

Demersal fish populations are regularly monitored with scientific bottom trawl surveys. It is now routine practice to have acoustic devices running during trawl surveys. How to take into account this additional and densely sampled information to improve abundance estimations?

Using the spatial correlation between the data locations, Geostatistics is an effective tool to assess fish biomass.

In the first part of the thesis, the different bivariate models are studied. A structural analysis is carried out for each variable and for their spatial correlation, for this kind of sampling. Then, practical problems about the fitting of cokriging are discussed for different spatial layouts observed in the data.

The acoustic information is available for the whole water column, while the effective vertical support of the catch variable is unknown, because of fish displacement. Using several data analysis tools, a detailed study of the vertical dimension provided by the sonar makes it possible to understand the relationship between the two variables. In the Barents Sea, the sum of the first 40 meters above the bottom turned into a robust variable with a high spatial autocorrelation.

Then, the outputs of each model are compared for data from three different areas : the North Sea, the Irish Sea and the Barents Sea. The accuracy gained by the auxiliary variable, the acoustics recordings, is analysed for each sampling design. The results obtained vary a lot from one sampled area to another. In the North Sea, the additional acoustic information does not improve the estimations, while the enhancement is very important in the Barents Sea.