# Apprentissage automatique de données massives bathymétriques pour l'optimisation de systèmes de levé hydrographique

Julian Le Deunf

DOCTORAT BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

# THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPÉRIEURE MINES-TÉLÉCOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Informatique*

Par

**Julian LE DEUNF**

## Apprentissage automatique de données massives bathymétriques pour l'optimisation des systèmes de levé hydrographique

**Thèse présentée et soutenue à Brest, le 12/12/2022**
**Unité de recherche : UMR CNRS 6285 Lab-STICC**
**Thèse N° : 2022IMTA0333**

### Rapporteurs avant soutenance :

| | |
|---|---|
| Nour-Eddin EL FAOUZI | Professeur, université Gustave Eiffel (Lyon) |
| Sylvie DANIEL | Professeure titulaire et directrice du baccalauréat en génie géomatique, université Laval (Québec) |

### Composition du Jury :

| | | |
|---|---|---|
| Président : | Nour-Eddin EL FAOUZI | Professeur, université Gustave Eiffel (Lyon) |
| Examinateurs : | Sylvie DANIEL | Professeure, université Laval (Québec) |
| | Jean-Guy NISTAD | Ingénieur hydrographe, Bundesamt für Seeschifffahrt und Hydrographie (Hamburg) |
| | Thierry SCHMITT | Docteur en géologie marine, Shom (Encadrant, Brest) |
| | Nathalie DEBÈSE | Enseignant-Chercheur, Lab-STICC UMR CNRS 6285, ENSTA Bretagne (Brest) |
| Dir. de thèse : | Romain BILLOT | Professeur, Lab-STICC UMR CNRS 6285, IMT Atlantique (Plouzané) |

### Invité(s) :

Steve OUDOT Maître de conférence Inria Saclay (Saclay)

# Remerciements

Je tiens à adresser mes plus sincères remerciements à toutes celles et tous ceux qui ont contribué à la réalisation de cette thèse.

Tout d'abord, je tiens à remercier très chaleureusement mon directeur de thèse, le professeur Romain Billot de l'IMT Atlantique, pour son soutien constant et pour m'avoir suivi dans cette aventure atypique, sans connaître ni le métier d'hydrographe ni la donnée étudiée. Je remercie également mes encadrants de thèse, le docteur Thierry Schmitt du Shom et la docteure Nathalie Debèse de l'ENSTA Bretagne pour leurs nombreux conseils et les échanges sur la thématique de recherche.

Je remercie également les membres du jury, la professeure Sylvie Daniel et le professeur Nour-Eddin El Faouzi, pour leur disponibilité et leurs commentaires constructifs lors de la soutenance de ma thèse mais également le temps pris pour la relecture précieuse de ce manuscrit. Merci aussi aux autres membres du jury, Jean-Guy Nistad et Steve Oudot.

Je suis également reconnaissant envers mes collègues du Shom (Morvan, Olivier, Christophe, Ronan, Yvan, Thibauld, Gaël) pour leur soutien et les discussions stimulantes autour d'un café, et tout particulièrement Yves-Marie et Yann pour avoir soutenu ce projet de thèse de la genèse à son aboutissement. Ils ont fait preuve d'une direction toujours bienveillante et ont su comprendre que l'intérêt du Shom passait par la recherche appliquée et les partenariats innovants.

Je remercie aussi l'équipe DECIDE de l'IMT Atlantique pour leur accueil et tout particulièrement les membres du défi qualité, John, Laurent, Arwa et Patrick. Ce défi aura grandement contribué à ce projet de recherche et à ce manuscrit.

Mes pensées vont aussi à mes amis les plus proches, Charly, Justine, Loïc, Milena, Vandier, Victor, Claire, Ewen, Justin, Antoine et Clément qui ont toujours su me motiver et m'encourager. Les moments partagés ont été une bouffée d'oxygène bienvenue depuis que je vous connais.

Je remercie également ma famille pour leur amour et leur soutien indéfectible pendant cette période intense et pour m'avoir constamment poussé à me dépasser

et donner le meilleur de moi-même. Tout particulièrement, je remercie ma sœur Morane dont les relectures nombreuses auront amélioré grandement ce manuscrit.

Enfin, je serai éternellement reconnaissant envers ma femme, Emily, pour sa présence constante, sa confiance en moi et son travail de l'ombre pour cette thèse. Elle a été une source inépuisable de motivation et de soutien et m'a donné le plus beau des cadeaux, notre fille Calypso. J'ai été chanceux de pouvoir compter sur elle tout au long de ces quatre années.

Ce travail n'aurait pas été possible sans l'aide et le soutien de toutes ces personnes. Je leur en suis infiniment reconnaissant.

*Trop de connaissance ne facilite pas les plus simples décisions.*
Franck Herbert - *Les enfants de Dune*

# Acronymes

**ACP** Analyse en Composantes Principales. 75

**AL** Altimétrie Littorale. 124, 146–148, 153, 156, 160, 222, 226, 233

**ASV** Autonomous Surface Vehicle. 163

**AUV** Autonomous Underwater Vehicle. 72, 163

**BDBS** Base de Données Bathymétriques du Shom. 27, 29, 35, 175, 216, 221

**BDD** Base De Données. 23, 29, 58, 164, 218, 233

**CHOF** Capacité Hydrographique et Océanographique Future. 29, 33, 111, 217

**CM** Carte Marine. 17, 19, 21, 26, 41, 110, 111, 149, 205, 216, 217, 219, 233

**CTD** Conductivity Temperature Depth. 56, 189

**CUBE** Combined Uncertainty and Bathymetry Estimator. 91, 99, 110, 214

**DBSCAN** Density-Based Spatial Clustering of Applications with Noise. 74, 100, 112, 113, 118, 157, 173

**DGA** Direction Générale de l'Armement. 109, 119, 160, 231

**DL** Deep Learning. 133, 144

**ENC** Electronic Navigational Chart. 219, 231

**ETL** Extract-Transform-Load. 180, 225

**GNSS** Global Navigation Satellite System. 53, 54, 60, 63, 85, 86, 187, 231

**IMU** Inertial Measurement Unit (station inertielle). 52, 60, 62, 85, 86, 181

**LiDAR** Light Detection And Ranging. 17–21, 27, 29, 46, 49–51, 63, 67, 68, 84, 86, 90, 93, 96, 97, 100, 105, 106, 110, 121, 123–126, 128, 131, 138–142, 147, 155–157, 159, 160, 167, 175, 181, 187, 189, 191, 204–206, 210, 214, 215, 226, 232, 233

**MAD** Median Absolute Deviation. 19, 112, 114, 115, 143, 151–153

**MAE** Mean Absolute Error. 135, 136

**MCDA** Multiple-Criteria Decision Analysis. 161, 181, 183–185, 191, 197, 206, 207, 211, 213, 215, 216, 221, 222, 225

**ML** Machine Learning. 24, 40, 67, 70, 77, 96, 111, 112, 116–120, 124, 128, 131, 153–156, 159, 222, 223, 231

**MLP** MultiLayer Perceptron. 20, 131, 133, 134, 136

**MNB** Modèle Numérique de Bathymétrie. 86, 99, 100, 102, 103, 128, 143

**MNT** Modèle Numérique de Terrain. 21, 23, 156, 157, 164, 182, 186, 187, 214, 216–219, 229

**NOAA** National Oceanic and Atmospheric Administration. 91, 99, 216

**OEM** Ondes ÉlectroMagnétiques. 39, 50, 51, 53, 57, 63, 66, 67, 81, 84

**OHI** Organisation Hydrographique Internationale. 22, 40, 43, 162, 163, 224, 231

**OMI** Organisation Maritime Internationale. 231

**PNH** Programme National d'Hydrographie. 212, 218, 220

**PTD** Projet de Technologies de Défense. 160, 231

**RETEX** RETour d'EXpérience. 155

**RF** Random Forest. 20, 24, 78, 116, 119, 131–133, 136, 139, 144, 226

**SBES** Single Beam Echo-Sounder (sondeur mono-faisceau). 46, 47, 49, 56, 60, 61, 64, 65, 100, 187, 189, 191, 193, 204, 206

**SDB** Satellite-Derived Bathymetry. 46, 51, 187, 189, 191, 231

**SH** Service Hydrographique. 26, 60, 88, 91, 99, 108, 110, 111, 121, 162, 211, 216, 221, 222, 224, 225, 231, 232

**Shom** Service Hydrographique et Océanographique de la Marine. 17, 18, 21–23, 25–33, 35, 41, 44, 45, 47, 49, 51, 56, 58, 60, 63, 68, 85, 91, 108–111, 123–125, 147, 149, 159, 160, 164, 166, 170, 187, 190, 192, 193, 202, 207, 210–212, 216–218, 221, 222, 224, 226, 231–234

**SI** Système d'Information. 148, 210

**SIG** Système(s) d'Informations Géographiques. 168, 219

**SME** Surface Minimale Englobante. 22, 45, 164–166, 168–170, 172, 173, 175, 176, 178–180, 209, 210, 218, 221

**SMF** Sondeur Multi-Faisceau. 17, 19, 27–30, 46–51, 56, 59–65, 67, 83–87, 90, 93, 96, 97, 100, 101, 106, 107, 109–112, 116, 119, 124, 157, 160, 181, 187, 189, 191–193, 205, 206, 215, 218, 226, 233

**SonaL** Sonar à balayage Latéral. 18, 59, 60, 191, 193, 204, 206

**SONAR** SOund Navigation And Ranging. 27

**ToMATo** Topological Mode Analysis Tool. 74, 157

**TPU** Total Propagated Uncertainty. 86, 99, 112–114, 207

**ZEE** Zone Économique Exclusive. 216

# GLOSSAIRE

**amer** En navigation maritime, un amer est un point de repère fixe et repérable sans ambiguïté utilisé pour la navigation maritime et clairement identifié sur une carte marine pour se positionner et réaliser des alignements. 41

**bathymétrie** La bathymétrie est la science de la mesure des profondeurs de l'océan, elle est l'équivalent de la topographie pour le milieu marin. 26, 32, 45, 193

**classification** Méthode d'analyse de données qui regroupe des algorithmes d'apprentissage supervisé adaptés aux données qualitatives. L'objectif est d'apprendre (autrement dit de trouver) la relation qui lie une variable d'intérêt, de type qualitative, aux autres variables observées, éventuellement dans un but de prédiction. On utilise la classification lorsque la variable d'intérêt est qualitative, c'est-à-dire qu'elle prend ses valeurs dans un espace qui ne possède pas de métrique naturelle. 77, 78

**clustering** Le *clustering*, classification non supervisée ou partitionnement de données en français, est une méthode en analyse des données qui vise à diviser un ensemble de données en différents groupes homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes. 73, 74, 113, 157

**descripteur** Un descripteur (*feature* en anglais) est une manière de décrire un jeu de données. L'ensemble des descripteurs est donc constitué des variables prédictives d'entrée du processus d'apprentissage. Dans la littérature, on trouve également le terme "caractéristique". 67, 72

**géodésie** La géodésie est la science et technique qui ont pour objet l'étude de la forme de la terre (géoïde et la détermination de ses dimensions, au moyen de mesures telles que la triangulation, la trilatération, le nivellement, les observations gravimétriques, les observations par satellites...). 53

**géophysique** La géophysique est l'ensemble des sciences qui étudient les caractéristiques et les propriétés physiques de la terre. 42

**hyperparamètre** Paramètre de l'algorithme dont la valeur est fixée avant le début du processus d'apprentissage. 20, 130, 135–137

**IA** L'intelligence artificielle est le champ interdisciplinaire théorique et pratique qui a pour objet la compréhension de mécanismes de la cognition et de la réflexion, et leur imitation par un dispositif matériel et logiciel, à des fins d'assistance ou de substitution à des activités humaines (définition du journal officiel). 69, 70, 133, 159, 160, 221–224, 232

**isobathe** Sur une carte marine, une isobathe, ou courbe de profondeur est une ligne joignant des points d'égale profondeur ; c'est donc une courbe de niveau, indiquant la profondeur d'une surface au-dessous du niveau de l'eau. 164

**minutes** Les minutes de bathymétrie sont des données bathymétriques valides qui permettent d'établir les cartes marines. Elles recensent uniquement les sondes, sous forme de radiales, effectuées par les bateaux chargés des levés. 220, 231

**non-supervisé** L'apprentissage non supervisé est un problème d'apprentissage automatique. Il s'agit, pour un logiciel, de trouver des structures sous-jacentes à partir de données non étiquetées. 70, 72, 73, 81, 97, 98, 118, 123, 124, 145, 146, 157

**océanographie** L'océanographie est l'ensemble des disciplines scientifiques ayant pour objet l'étude de l'océan et de ses limites, et qui comprend, notamment, la physique et la dynamique des masses d'eau, la chimie de l'eau de mer, la biologie marine et la géologie sous-marine. Au sens strict, le terme "océanographie" s'applique uniquement aux études relatives à la description du milieu marin, tandis que le terme océanologie s'applique à l'ensemble des études se rapportant plus ou moins directement à l'océan. 42

**pilonnement** Mouvement oscillatoire de montée et descente de l'ensemble d'un navire dû à l'action de la houle. 52

**réduction de dimension** La réduction de dimension consiste à prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension. Cette opération est cruciale pour l'apprentissage machine afin de lutter contre le fléau de la dimension, terme inventé par Richard Bellman [1] et qui revient souvent à tirer des inférences d'un nombre réduit d'expériences dans un espace de possibilités de dimension élevée. 73, 75

**régression** La régression est un ensemble de méthodes statistiques utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres. On utilise le terme régression lorsque la variable d'intérêt est quantitative. 77, 78

**sous-apprentissage** Le terme sous-apprentissage définit une situation dans laquelle un modèle n'a pas été capable d'apprendre assez à partir de l'ensemble d'apprentissage. Ce mauvais ajustement aux données entraîne de très mauvaises performances en généralisation. Le sous-apprentissage est souvent du à un manque de données, ou l'utilisation d'une règle de décision, modèle, mal adapté au problème ou trop simple. 19, 79–81, 116, 118

**supervisé** L'apprentissage supervisé (supervised learning) est une tâche d'apprentissage machine consistant à apprendre une fonction de prédiction à partir d'exemples annotés (étiquetés), au contraire de l'apprentissage non-supervisé. 70, 72, 75–77, 81, 83, 96, 97, 123, 146

**surapprentissage** Le terme surapprentissage désigne une situation dans laquelle un modèle présentera une très faible erreur d'apprentissage mais une erreur plus grande sur de nouvelles données. Ainsi, le modèle utilisé est trop précis, avec un biais très faible sur la base d'apprentissage, mais aura du mal à se généraliser (forte variance). 18, 75, 79, 80

**sédimentologie** La sédimentologie est la science qui traite de la description, de la classification, et de l'étude des sédiments et des roches sédimentaires. 58

**thermocline** Une thermocline est une couche d'eau de la mer dans laquelle la température décroît rapidement avec la profondeur. Elle sépare la couche de mélange superficielle des eaux sous-jacentes. Selon les zones, elle peut être permanente ou saisonnière. 57

**transducteur** Le transudcteur est un système de conversion d'un type d'énergie en une autre. Dans le cas de l'acoustique, le transducteur convertit l'énergie acoustique en énergie électrique ou inversement. 64

**voxel** Le voxel (mot-valise créé en contractant « volume » et « pixel ») stocke une information physique (couleur, densité, intensité, etc.) d'un point d'un volume sur un maillage régulier en 3D (il est l'équivalent 3D du pixel 2D). 106, 129, 130

# Sommaire

# TABLE DES FIGURES

# Liste des tableaux

# INTRODUCTION GÉNÉRALE

## L'origine du Shom et le traitement de la donnée bathymétrique

La plus ancienne carte marine connue est la *carte pisane*, datant du XIII$^{e}$ siècle, qui doit son nom à son lieu de découverte, Pise. Aujourd'hui conservée à la Bibliothèque Nationale de France, elle représente les côtes de la Méditerranée. Confectionnée à Gênes, elle n'a ni projection apparente ni coordonnées géographiques ni mesures bathymétriques. En France, les premiers documents nautiques remontent à la fin du XV$^{e}$ siècle. À cette époque, Dieppe en Normandie (voir [2], [3]) est le berceau d'une école d'hydrographie de grande renommée en Europe du Nord. Les pilotes dieppois sont à l'origine des premières cartes marines françaises. S'inspirant de l'école de Dieppe, Jean-Baptiste Colbert créa en 1661 des établissements analogues dans les principaux ports du royaume. Dirigés par les maîtres dieppois, les travaux réalisés furent publiés en 1693 sous la forme d'un atlas de cartes, le Neptune François (voir [2]-[4]), qui connut un grand succès international. Poursuivant l'œuvre de Colbert, un arrêt du conseil de la Marine du 19 novembre 1720 crée le Dépôt des cartes, plans et journaux de la Marine, ancêtre du Shom. La figure 1 est un exemple d'une carte marine des côtes françaises réalisée sous la direction de Charles-François Beautemps-Beaupré, ingénieur hydrographe, père de l'hydrographie moderne [3], [5], qui a donné son nom à un célèbre navire contemporain de la Marine nationale.

Un décret du 13 janvier 1886 transforme le Dépôt général de la Marine en Service Hydrographique de la Marine (SHM, voir [3]). Il est rattaché à l'état-major de la Marine et dirigé par un ingénieur hydrographe en chef. Un décret du 25 mai 1971 modifie l'appellation et l'organisation du Service qui devient le Shom [3], [4]. Le service est décentralisé et un établissement principal est construit à Brest. Le Shom, qui a ainsi fêté en 2020 ses 300 ans d'existence, est un Établissement Public Administratif (EPA - depuis 2007) sous tutelle du Ministère des Armées qui

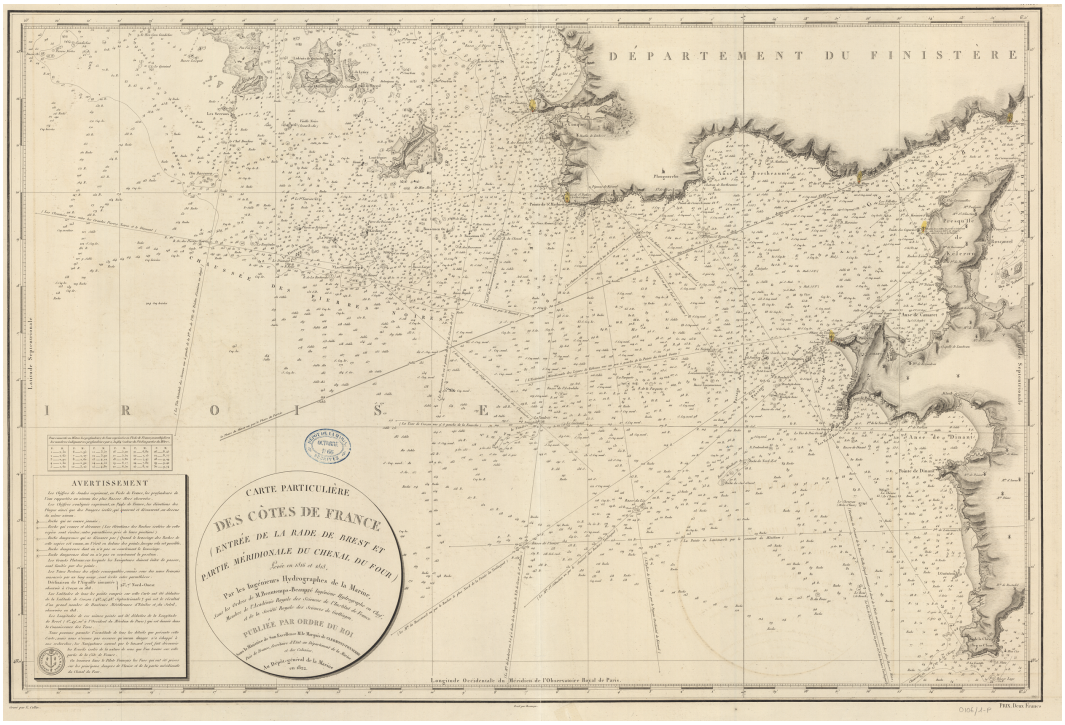


FIGURE 1 – CM particulière des côtes de France (entrée de la rade de Brest et partie méridionale du chenal du Four), publication de 1822 ©Shom.

emploie plus de 500 personnes dont 120 hydrographes. Il est l'opérateur public pour l'information géographique maritime et littorale de référence. Il a pour mission de connaître et de décrire l'environnement physique marin dans ses relations avec l'atmosphère, avec les fonds marins et les zones littorales, d'en prévoir l'évolution et d'assurer la diffusion des informations correspondantes. La figure 2 présente ainsi les trois activités principales du Shom pour mener à bien cette mission.

— **Hydrographie nationale** : sa première mission, en tant que Service Hydrographique (SH) national, est de recueillir, d'archiver et de diffuser les informations nécessaires à la navigation maritime de surface (voir convention SOLAS, *Safety Of Life At Sea*, de 1974 [6]), notamment la bathymétrie, et ce dans les eaux sous juridiction française et dans les zones placées sous la responsabilité cartographique de la France (via des conventions bilatérales) ;
— **Soutien à la défense** : caractérisé par l'expertise apportée par le Shom dans les domaines hydro-océanographiques à la Direction générale de l'Armement (DGA) et par ses capacités de soutien opérationnel des forces ;
— **Soutien aux politiques publiques de la mer et du littoral** : par lequel


FIGURE 2 – Les trois activités du Shom.

le Shom valorise ses données patrimoniales et son expertise en les mettant à la disposition des pouvoirs publics, et plus généralement de tous les acteurs de la mer et du littoral.

Le Shom assure ainsi des levés hydrographiques dans le cadre de ses opérations, grâce à divers moyens d'acquisition des données brutes, tels que le SMF [7], le LiDAR bathymétrique, le sondeur sédimentaire, le SOund Navigation And Ranging (SONAR) latéral [7], l'imagerie satellitaire, le courantomètre ou bien le marégraphe, pour ne citer que les capteurs les plus couramment utilisés aujourd'hui. Afin de traiter l'ensemble des informations récoltées par ces senseurs, différentes méthodes de traitement des données existent en fonction du type de capteur et du produit final attendu. La quasi-totalité des produits bathymétriques établis par le Shom garantissent la sécurité de la navigation maritime. Tous les levés bathymétriques du Shom et leurs métadonnées associées sont archivés, sans hiérarchie particulière (au sens pile de données non classée), dans la Base de Données Bathymétriques du Shom (BDBS), qui compte actuellement plus de 11 400 levés.

Les méthodes de traitement de toutes ces données ont ainsi évolué au cours du temps et ont su s'adapter aux volumes et contraintes liées à l'amélioration continue des capteurs. Nous sommes donc passés, pour la donnée bathymétrique, des sondeurs à plomb dont l'échantillonnage spatial est de très faible densité (moins de 10 sondes au km$^2$) à des SMF haute densité ou des LiDAR bathymétriques de

dernière génération permettant d'acquérir plusieurs milliers de sondes par seconde (quelques dizaines de sondes au $m^2$ pour des profondeurs d'eau inférieures à 20 mètres). La figure 3 montre ainsi la différence de densité entre le plomb de sonde et le SMF par 20 mètres de fond.

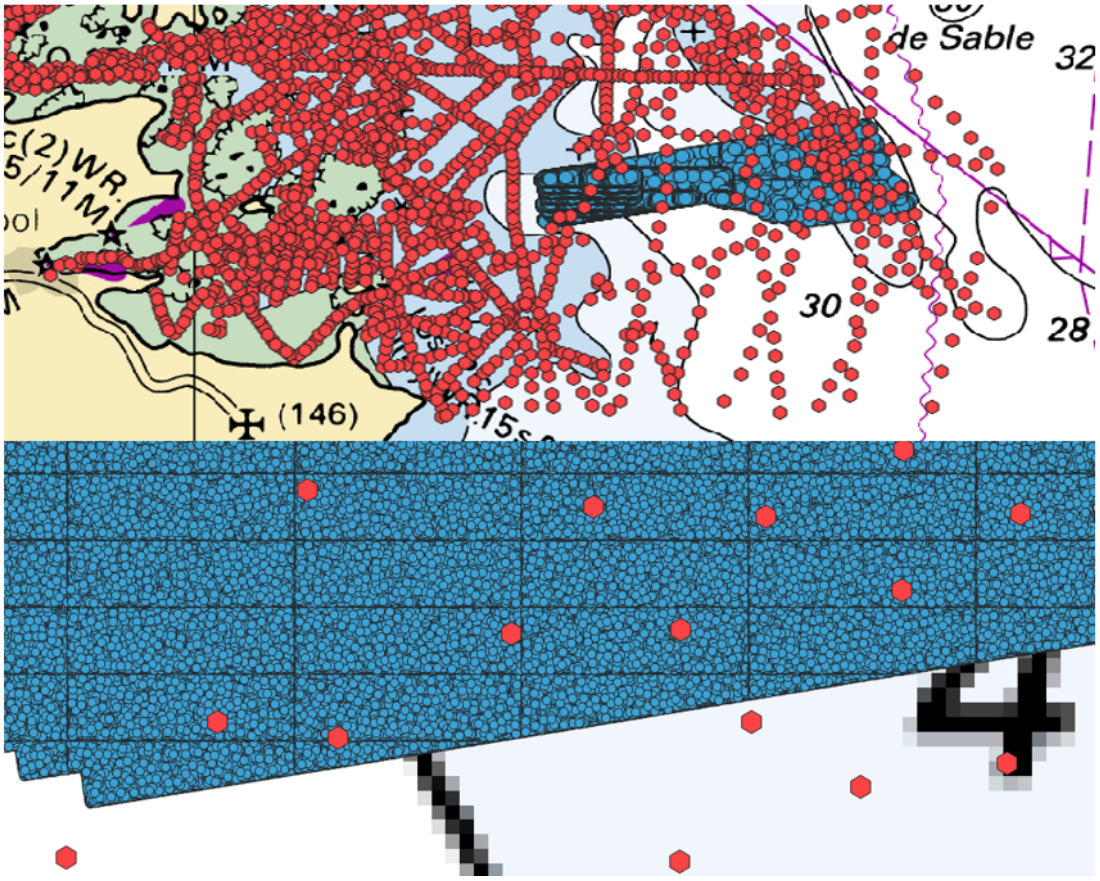


FIGURE 3 – Différence de densité entre le levé S183100100 au plomb de sonde en rouge et le levé SMF S202204500 en bleu. Vue globale sur la partie haute et zoom sur la partie basse de la figure.

Comme le montre la figure 4, le volume de données stockées dans la base de données bathymétriques du Shom (considérée comme représentative d'une grande organisation traitant des ensembles de données bathymétriques) a été multiplié par 10 sur une période de 6 ans, tendance également observée par d'autres organisations similaires comme le service hydrographique canadien ou anglais. Or actuellement, le ratio d'effort entre l'acquisition de la donnée bathymétrique et son post-traitement (manuel ou semi-automatique) est d'environ 4.5 jours par jour d'acquisition pour la donnée SMF et de 6 jours par jour d'acquisition pour la

donnée LiDAR (très variable en fonction du type de zones acquises). Il est donc important de s'intéresser à l'automatisation de la phase de traitement. Aujourd'hui, l'invalidation ou non des sondes ne repose que sur la décision d'un hydrographe qualifié (comme préconisé par [8]). Cette augmentation du volume de données bathymétriques à traiter est un phénomène mondial [9] qui s'accélère. Les décisions politiques internationales, comme l'initiative des Nations Unies pour la Décennie des Nations Unies des sciences océaniques au service du développement durable (2021-2030), sont un des moteurs de cette accélération. L'objectif de cette initiative est de mobiliser des acteurs du monde entier autour d'un cadre commun pour des objectifs de développement durable pour l'océan. Le projet Seabed 2030 voir [10], de la Nippon Foundation-GEBCO s'inscrit totalement dans cette démarche avec la quête d'une cartographie complète des océans du monde d'ici 2030. Elle permettrait ainsi de combler la disparité de connaissances qui subsiste actuellement entre les abysses de la terre et la surface de Mars. A l'échelle nationale, le ministère des Armées témoigne également d'une volonté forte dans ce domaine en élaborant une stratégie ministérielle de maîtrise des fonds marins [11] mais également via le programme d'armement Capacité Hydrographique et Océanographique Future (CHOF) qui sera structurant dans l'avenir du Shom.

La figure 4 met en avant une accélération importante du volume de données acquises à partir de 2010 et l'arrivée des SMF haute densité permettant d'acquérir beaucoup plus de sondes par ping (passant de 128 sondes par ping à 800). Le volume de l'année 2019 est partiellement représentatif car il reste des levés à valider et à intégrer dans la BDBS. Cette durée avant l'intégration d'un levé hydro-océanographique dans les BDD métiers du Shom varie entre quelques mois et plus de deux ans en fonction de la complexité des levés et du volume de données à traiter et contrôler. Ce volume est particulièrement important en 2016 de par l'intégration de levés LiDAR bathymétriques volumineux.

La nécessité de disposer de méthodes de détection des valeurs aberrantes pour traiter et contrôler semi-automatiquement les données bathymétriques redevient une priorité. Par ailleurs, dans le cadre de la sécurité de la navigation, il est fondamental de conserver un niveau important de compréhension des méthodes mises en place afin d'assurer l'adhésion des experts qui exploiteront ces nouvelles techniques tout en conservant la capacité d'expliquer les choix réalisés par les algorithmes (améliorant également la confiance dans les outils).

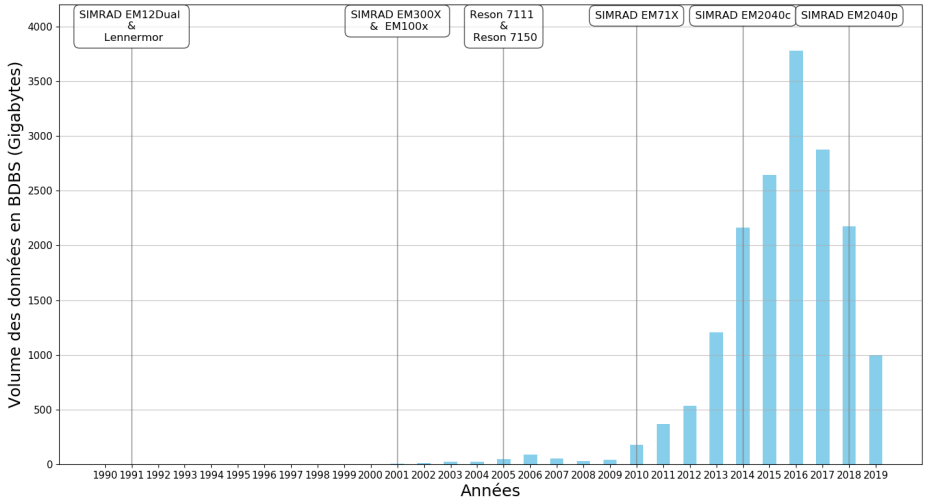


FIGURE 4 – Volumes des données bathymétriques SMF traitées par an et acquises par les moyens du Shom.

De plus, une fois la donnée traitée, une partie importante du travail de l'hydrographe consiste à rédiger un rapport synthétisant les résultats de ce traitement afin de fournir un livrable intégrable dans une base de donnée métier ou pour réaliser un produit bathymétrique. Ces tâches comprennent du contrôle, de la mise en forme de données et de la validation. Elles s'avèrent nécessaires pour assurer un haut niveau de qualité et permettre une intégration simple des données bathymétriques dans les autres chaînes de traitement et dans l'ensemble des processus de production de la connaissance du Shom. Il est donc critique d'améliorer cette extraction de la qualité dans les levés hydrographiques afin que les opérateurs puissent se concentrer sur l'interprétation des données et la création de produits innovants.

# Problématique et objectifs des travaux de recherche

L'explosion du volume de données bathymétriques, couplée à la démocratisation des algorithmes d'intelligence artificielle au cours des dernières années, a particulièrement motivé le lancement de nos travaux de recherche, qui visent à accompagner la transformation numérique inévitable du métier de l'hydrographe,

en proposant des solutions concrètes de traitement et gestion de la donnée bathymétrique. Plus précisément, notre travail de thèse vise à répondre à plusieurs questions de recherche :

— Comment faciliter le travail des opérateurs dans la détection des sondes aberrantes ?
— Quelles sont les méthodes d'intelligence artificielle que nous pouvons mettre en place pour cet objectif et qui seront acceptées par les hydrographes ?
— Comment détecter de façon fidèle la couverture géographique d'un levé hydro-océanographique en prenant en compte l'incertitude de la mesure associée et la densité du lot de données bathymétriques ?
— Comment intégrer des modèles de qualité de la donnée et de préférences pour évaluer la qualité d'un levé hydro-océanographique ?

Les axes majeurs de recherche de cette thèse ont été définis en réponse aux besoins et problématiques variés du Shom. Le premier axe est celui du traitement des données bathymétriques avec des techniques à la complexité variable (du non supervisé statistique au supervisé ainsi que les méthodes d'apprentissage machine ensemblistes). Le second axe concerne la détermination de la qualité d'une campagne hydrographique avec l'ensemble des éléments qui la compose (donnée bathymétrique, océanographique et sédimentologique par exemple). Dans cet axe, un travail sur la recherche des surfaces englobantes d'un nuage de points bathymétriques a été réalisé : le but était de calculer, pour un levé hydrographique, l'emprise géographique la plus restreinte possible car utilisée dans les processus de déconfliction des différents lots bathymétriques. Les processus de déconfliction permettent de sélectionner la donnée bathymétrique de meilleure qualité pour la réalisation d'une surface bathymétrique de référence. Une seconde activité sur l'évaluation de la qualité d'une campagne de levés par le biais d'un modèle d'aide à la décision multicritère a également été réalisée. Ce manuscrit est donc dual, intrinsèquement et de par les thèmes abordés. Il comprend une composante métier forte autour de l'hydrographie et de la donnée bathymétrique, mais surtout une composante informatique primordiale dans la transformation du métier d'hydrographe afin de l'aider, dans le futur, à gérer au mieux ces flux de données toujours plus importants.

## Contexte des travaux, enjeux académiques et industriels

Cette thèse de quatre ans a été réalisée en parallèle d'un poste d'ingénieur d'études et de développement dans l'amélioration des chaînes de traitement et d'exploitation en bathymétrie. Un écosystème de travail s'est donc mis en place au sein du département bathymétrie du Shom et j'ai été intégré dans l'équipe pluridisciplinaire DECIDE du Lab-STIC (Laboratoire des Sciences et Techniques de l'Information de la Communication et de la Connaissance), UMR 6285. La figure 5 montre le positionnement de l'équipe DECIDE sur trois axes de recherche : données, informations et décisions, permettant de relever les défis scientifiques rencontrés par l'équipe.

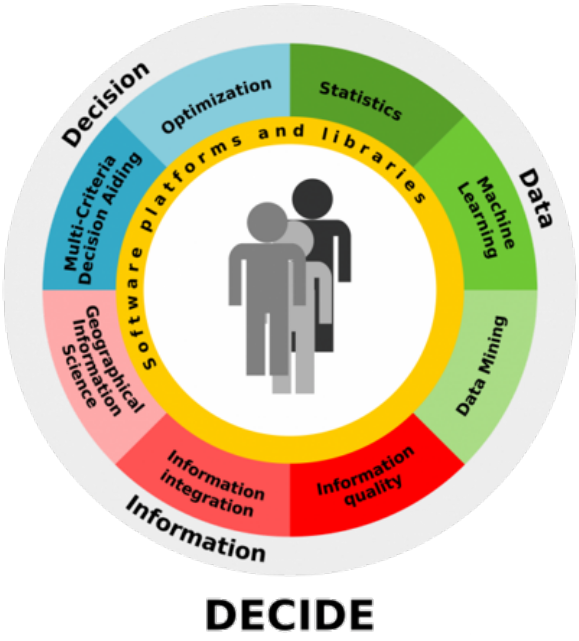


FIGURE 5 – Les trois axes de recherche de l'équipe DECIDE avec l'humain au centre du processus de décision. ©Équipe DECIDE Lab-STICC, UMR CNRS 6285.

Les enjeux et retombées de ces travaux de recherche sont multiples :

— **Sur le plan académique**, nous proposons des solutions informatiques et algorithmiques innovantes dans la gestion et la modélisation de la donnée bathymétrique au service de l'amélioration de la chaîne de valeur bathymétrique ;

— **Sur le plan industriel**, nous visons à renforcer le dispositif global du Shom face à l'augmentation du volume de données bathymétriques à traiter, dès à présent et à venir (renouvellement des capacités d'acquisition hydrographique et élargissement de son spectre par le programme CHOF), tout particulièrement dans les zones littorales qui sont les zones les plus complexes à appréhender et qui produisent le plus de données bathymétriques (dû à la densité importante des capteurs dans les faibles profondeurs).

## Organisation du manuscrit : une structuration multi-échelle de la donnée à la décision

La première partie du manuscrit décrit les concepts généraux autour des levés hydrographiques, des capteurs mis en place dans le cadre de ces levés (et leurs fondamentaux physiques) et des méthodes de traitement de donnée. L'objectif de cette partie est de permettre aux lecteurs non spécialistes du domaine de l'hydrographie-océanographie ou de la science informatique de détenir les éléments leur permettant d'appréhender ce manuscrit dans son entièreté.

Ensuite, de la donnée brute à l'intégration dans des bases de données métier en vue de l'élaboration d'un produit, un ensemble d'opérations de complexité et d'automatisation variables a été engagé pour extraire le maximum de valeurs de la donnée acquise sur le terrain. En effet, dans le cadre des opérations embarquées, les jours d'acquisition sont peu nombreux et les possibilités de revisite d'une zone géographique se comptent en dizaines d'années. Dans ce travail de recherche, nous avons étudié les capacités offertes par les sciences informatiques pour automatiser de nombreuses tâches réalisées manuellement par les hydrographes. Nous avons ainsi structuré ce document en fonction du recul pris par rapport à la donnée et le niveau de transformation et d'intelligence qui lui est apposé, adoptant une structuration multi-échelle du travail effectué comme présentée sur la figure 6.

Si nous prenons l'exemple du levé S201902600 de la figure 6, nous observons sur la partie de gauche une partie du levé bathymétrique ainsi que d'autres zones ponctuelles d'acquisition de données comme la position de carottages, ou de mouillages d'hydrophones. Ce levé a donc accumulé tout un ensemble de données brutes à traiter, contrôler et qualifier.

Si nous nous concentrons sur la donnée bathymétrique prise individuellement

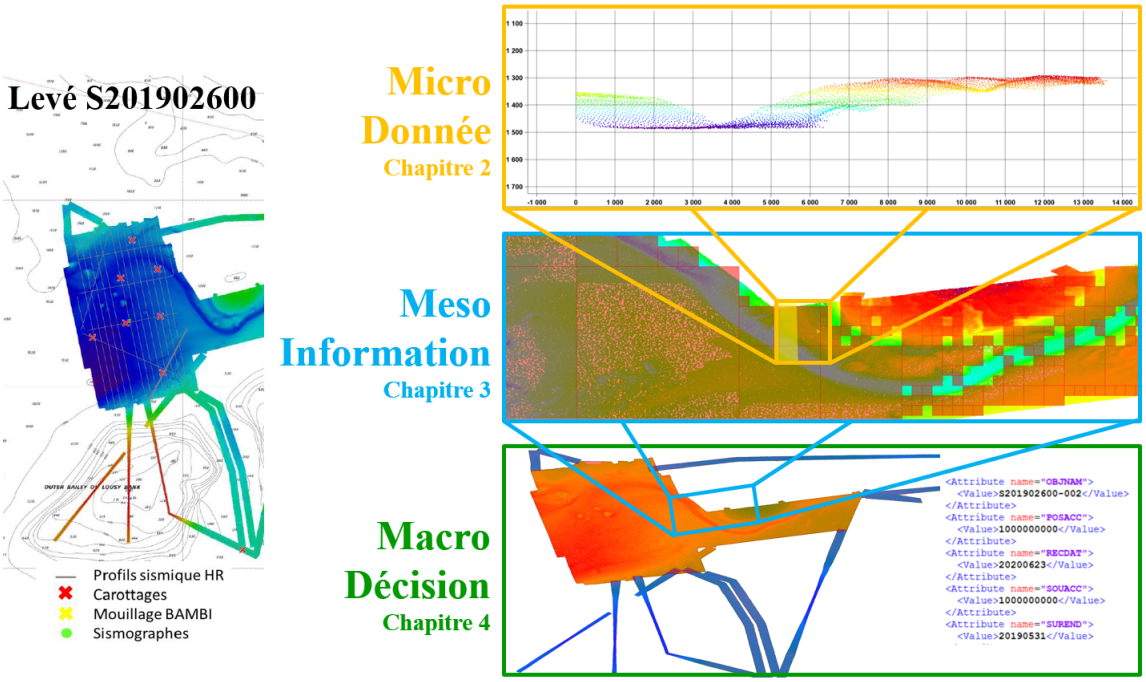


FIGURE 6 – La structuration multi-échelle : donnée, information, décision.

(au niveau de la sonde), nous nous plaçons au niveau microscopique traité dans le chapitre 2 (voir la partie supérieure droite de la figure 6). Une part importante de ce chapitre traitera de l'état de l'art autour des méthodes de traitement de la donnée bathymétrique ainsi que l'importance du choix/définition des descripteurs. Dans ce niveau est également étudiée la qualité associée à la donnée, permettant ainsi d'avoir pour chaque sonde collectée une mesure de fidélité et de justesse, en lien avec les capteurs utilisés et le protocole d'acquisition.

Une fois cette donnée bathymétrique étudiée, il est important d'extraire l'information de la donnée permettant de la contrôler et de la qualifier. Nous remontons à l'échelle mésoscopique définie comme l'augmentation de la donnée, sa qualité structurée et sa métadonnée, traité dans le chapitre 3 (voir la partie au centre à droite de la figure 6). Pour ce niveau méso, nous nous intéressons à l'extraction et la synthèse de l'information à partir des données (depuis le niveau micro, chapitre 2). Les données sont ainsi considérées comme un groupe et structurées de façon à créer un ensemble cohérent possédant des descripteurs/caractéristiques et des dimensions communes.

Dès que les données et les informations sont en notre possession, nous pouvons

les utiliser pour réaliser des prises de décisions associées à ces données hydro-océanographiques, comme construire l'emprise du levé présentée en bas à droite de la figure 6 avec les métadonnées du levé associées. Cela correspond au niveau macroscopique présenté dans le chapitre 4. Ce niveau s'intéresse à la décision issue de la fusion des données et des informations collectées. Il étudie également sa qualité au niveau du levé afin de pouvoir juger de la pertinence d'un ensemble de données bathymétriques, et améliorer la prise de décision associée à la donnée.

Enfin, nous terminerons ce manuscrit par un chapitre de discussion sur les sujets abordés dans les différentes parties. Nous présenterons ainsi le projet Téthys du Shom qui a pour objectif de produire une surface de référence bathymétrique fiabilisée et déconflictée à partir des sondes validées. Nous étudierons les interconnexions entre les trois niveaux d'échelle et les impacts métiers et méthodologiques de l'intelligence artificielle dans la chaîne de valeur bathymétrique.

**Synthèse du chapitre**

L'objectif de ce travail de recherche est d'accompagner le Shom dans sa transition numérique appliquée à la gestion des données bathymétriques du capteur au produit final et son intégration dans la BDBS. Durant ces quatre années de recherche ont été étudiées des solutions informatiques et algorithmiques de complexités variables en fonction du besoin mais également de l'interprétabilité requise pour la prise de décision.

# Publications

## Articles de revue

J. Le Deunf, N. Debèse, T. Schmitt et al., « A Review of Data Cleaning Approaches in a Hydrographic Framework with a Focus on Bathymetric Multibeam Echosounder Datasets », *Geosciences*, t. 10, 7, 2020, ISSN : 2076-3263. DOI : 10.3390/geosciences10070254. adresse : https://www.mdpi.com/2076-3263/10/7/254

M. Malik, A. C. G. Schimel, G. Masetti et al., « Results from the First Phase of the Seafloor Backscatter Processing Software Inter-Comparison Project », *Geosciences*, t. 9, *12*, 2019, issn : 2076-3263. doi : 10.3390/geosciences9120516. adresse : https://www.mdpi.com/2076-3263/9/12/516

## Conférences internationales à comité de lecture avec actes

J. Le Deunf, R. Mishra, Y. Pastol et al., « Seabed prediction from airborne topo-bathymetric lidar point cloud using machine learning approaches », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-9. doi : 10.23919/OCEANS44145.2021.9706113

J. Le Deunf, A. Khannoussi, L. Lecornu et al., « Automatic Data Quality Assessment of Hydrographic Surveys Taking Into Account Experts' Preferences », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-10. doi : 10.23919/OCEANS44145.2021.9705772

J. Le Deunf, R. Jarno, Y. Keramoal et al., « Téthys : automating a data workflow compiling over 300 years of bathymetric information », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-6. doi : 10.23919/OCEANS44145.2021.9705727

M. Michel, J. L. Deunf, N. Debese et al., « Multibeam outlier detection by clustering and topological persistence approach, ToMATo algorithm », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-8. doi : 10.23919/OCEANS44145.2021.9705930

J. Le Deunf, T. Schmitt et Y. Keramoal, *A pragmatical algorithm to compute the convex envelope of bathymetric surveys at variable resolutions*, International Cartographic Conference 2021 (ICC 2021), Florence, 2021

J. Le Deunf, N. Debèse, T. Schmitt et al., « Outlier detection for Multibeam echo sounder (MBES) data : from past to present », in *OCEANS 2019 - Marseille*, 2019, p. 1-10. doi : 10.1109/OCEANSE.2019.8867321

## Conférences internationales

J. Le Deunf, A. Ben-ammar, Y. Pastol et al., *Cleaning Outliers Bathymetric Lidar Data Using Machine Learning Techniques : Shom's Feedback*, Canadian Hydrographic Conference 2022 (CHC 2022), Ottawa, 2022

T. Schmitt, J. Le Deunf, M. Fally et al., *Automating the generation of the bathymetric surface reference through expert rules and ETL : the Tethys project*, Canadian Hydrographic Conference 2022 (CHC 2022), Ottawa, 2022

A. Khannoussi, J. Le Deunf, P. Meyer et al., *Integrating user preferences in the automatic quality assessment of hydrographic surveys*, 92nd Meeting of EURO Working Group on Multicriteria Decision Aiding "Multi-Criteria Decision Analysis as a transdisciplinary science", Cracow, sept. 2021

J. Le Deunf, *Traiter la bathymétrie avec l'aide de l'intelligence artificielle*, Seat Tech Week 2020, Virtuelle, oct. 2020

J. Le Deunf, J. Eddadsi, R. Billot et al., *Automated detection of bathymetric outliers*, Canadian Hydrographic Conference 2020 (CHC 2020), Québec, 2020

J. Le Deunf, *Prospects for innovation at Shom*, Vecteur 2019, Rimouski, 2019

## Conférences nationales

J. Le Deunf, *Traitement lidar bathymétrique utilisant des techniques d'apprentissage automatique pour une prédiction automatique du fond marin : application aux données de la Bretagne*, Journée de l'information scientifique et technique du Shom, Brest, juin 2022

T. Schmitt, J. Le Deunf, Y. Keramoal et al., *L'automatisation des processus de traitement et l'intelligence artificielle au service de la Bathymétrie*, Journées techniques de l'Afhy 2022, Lyon, 2022

Y. Pastol, J. Le Deunf, C. Salvaterra et al., *Levés par lidar bathymétrique aéroporté, dans le cadre du projet Litto3D®, et traitement des données par IA au Shom*, merIGeo 2020, Virtuelle, nov. 2020

J. Le Deunf, *Détection automatisée de sondes aberrantes, l'approche machine learning pour le traitement de la bathymétrie*, 29eme Journées de la Recherche de l'IGN, Virtuelle, oct. 2020

J. Le Deunf, N. Debèse, T. Schmitt et al., *Détection automatisée de sondes aberrantes*, Journée de l'information scientifique et technique du Shom, Brest, mars 2019

## Rapports d'études

J. Le Deunf, « Rapport du projet TraitLIA : Traitement semi-automatique de nuages de points issus de lidar bathymétrique aéroporté. », Shom, rapp. tech. n°18 DOPS / STM / BATHY / NP, avr. 2022

J. Le Deunf, « Utilisation de l'algorithme CUBE au Shom », Shom, Rapport technique n°27 DOPS / MIP / BATHY / NP, mai 2019

## Articles de magazine

J. Le Deunf, M. Legris et J. McManus, « Automatic Calibration for MBES Offsets », *Hydro International*, t. 25, *3*, p. 40, 42, oct. 2020

R. Legouge, G. André, J. Le Deunf et al., « Utilisation d'infrastructures géodésiques mondiales pour la réalisation nationale », *Revue XYZ*, t. 158, mars 2019, Publisher : Association Française de Topographie. adresse : `https://hal.archives-ouvertes.fr/hal-02317630`

Chapitre 1

# Généralités

**Synthèse du chapitre**

Ce chapitre présente les éléments nécessaires à la compréhension de la suite du manuscrit soit : les systèmes composant un système de levés hydro-océanographiques, les fondamentaux de la propagation acoustique et optique sous-marine, et les éléments de base de la science des données et de l'apprentissage machine.

En 2022, nous connaissons bien mieux les reliefs de la Lune (à 384 400 km de la Terre) ou de Mars (à 78 millions km de la Terre) que ceux de notre propre planète. Cet écart considérable dans la connaissance cartographique de ces corps célestes et le manque d'observations directes des fonds océaniques s'expliquent principalement par des limites technologiques. En effet, pour mesurer la surface de la Terre, de la Lune ou même de Mars, des moyens passifs (sans émission de signaux) ou actifs (avec émission de signaux), via des Ondes ÉlectroMagnétiques (OEM) émises et/ou captées par des capteurs installés sur des satellites, permettent de mesurer les échos générés sur ces surfaces. Les OEM parcourent sans difficulté de longues distances dans l'air ou l'espace et sont peu absorbées dans ces milieux. De plus, les satellites peuvent se déplacer très rapidement et couvrir de très grandes distances, offrant des capacités de cartographie importantes. En revanche, ces OEM se propagent beaucoup plus difficilement dans l'eau et sont très rapidement absorbées, voir la section 1.2.2. C'est pourquoi les capteurs acoustiques actifs sont

le seul moyen, à grande échelle, d'atteindre les plus grandes profondeurs marines afin de cartographier les fonds marins. Ainsi, le point le plus haut de la Terre, le Mont Everest, a été mesuré et validé par une équipe sino-népalaise en 2020 à 8848.86m [35] (incertitude centimètrique par rapport à un géoïde déterminé par l'international *height reference system*). À l'inverse, le point le plus profond de la terre est bien moins connu, la méta-analyse [36] estimant une profondeur de 10 984m $\pm$ 25m (incertitude à $2\sigma$) avec une incertitude sur la position horizontale de la mesure bathymétrique comprise entre 20 et 25m (à $2\sigma$ également).

Le présent chapitre décrira les systèmes employés dans le cadre des levés hydro-océanographiques, les principales particularités liées à ce type d'acquisition et les méthodologies mises en place pour traiter et qualifier la donnée issue de ces capteurs. Dans un second temps, nous présenterons les fondamentaux physiques associés aux capteurs de la chaîne d'acquisition bathymétrique. Nous ferons ainsi un point spécifique sur la physique associée à l'acoustique et l'optique sous-marine. Enfin, nous décrirons les méthodes d'apprentissage machine, ou *Machine Learning* (ML) en anglais, et passerons en revue tous les fondamentaux techniques afin de définir les termes clés et s'entendre sur une base commune pour la suite du manuscrit.

## 1.1 Systèmes de levé hydro-océanographique

Les systèmes de levés hydro-océanographiques sont l'ensemble des moyens technologiques et des méthodologies mis en place pour la réalisation de levés hydro-océanographiques respectant les critères de qualité attendus par le prescripteur de levé, synthétisés notamment par les normes hydrographiques établies par l'OHI et tout particulièrement la norme S-44 [8] pour les levés hydrographiques.

Les systèmes à mettre en œuvre au cours d'un levé hydro-océanographique, et le choix du paramétrage qui leur est associé, sont déterminés par les besoins finaux qui spécifient ce levé et par les résultats attendus en termes de capacité de détection, d'incertitude, de couverture, de densité et de qualité des données fournies. Ce sont les besoins et le contexte associé au levé (reconnaissance d'une zone, voie d'accès ou chenal, profil de plage, géomorphologie, exploration du sous-sol...) qui détermineront les efforts, en terme d'acquisitions et traitements, à mettre en place au cours d'un levé.

Au cours du temps, la représentation des fonds marins a grandement évolué en lien direct avec l'amélioration des moyens d'acquisition, de traitement et de visualisation des données hydro-océanographiques. Les premiers outils mis en place pour mesurer les fonds marins ont été les sondeurs mécaniques : des perches pour les très petits fonds, ou un plomb attaché à une corde graduée qui mollit lorsqu'elle rencontre le fond. Ces méthodes étaient utilisées dans le cadre de l'élaboration du Neptune François [2]. Aujourd'hui encore, de nombreuses sondes issues de ce type de techniques sont encore présentes sur les cartes marines du Shom, comme le montre le diagramme des sources sur la figure 1.1 de la CM 7066 de l'île vierge à la pointe de Penmarc'h. La vidéo [37] est une reconstitution historique d'un levé hydrographique du début des années 1800 présentant ainsi le plomb de sonde suiffé (permettant d'avoir une information sur le sédiment de surface) et la technique de positionnement par triangulation d'amer à terre.

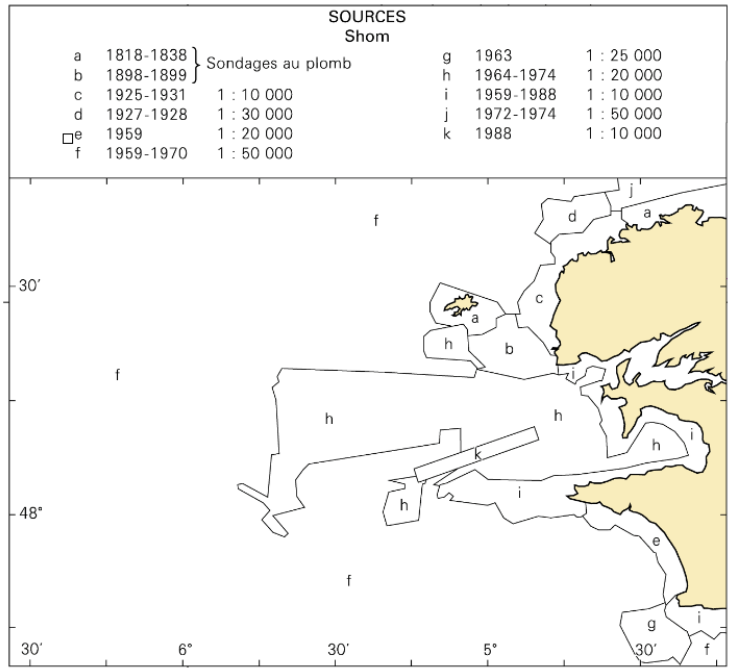


FIGURE 1.1 – Diagramme des sources de la CM 7066 de l'île vierge à la pointe de Penmarc'h.

### 1.1.1 Le levé hydro-océanographique

De manière générale, nous appelons levé toute campagne d'acquisition de données à la mer [38], [39], peu importe la finalité de cette campagne. Les levés mettent ainsi en place un panel très large de capteurs embarqués ou largués afin de mesurer au mieux l'environnement marin, que ce soit dans la colonne d'eau ou les fonds marins. Ainsi, un levé hydro-océanographique capture les données du fond marin, de la colonne d'eau et des aires terrestres adjacentes aux mers, océans, lacs, rivières, bassins portuaires, et autres étendues d'eau existantes sur la terre. Au sens strict, un levé hydro-océanographique est défini simplement comme une description physique des zones marines. Cependant, il peut inclure une large variété d'autres objectifs, comme la mesure des marées, des courants, de la gravité, du magnétisme terrestre et la détermination des propriétés physiques et chimiques de l'eau.

Classiquement, nous distinguons trois types de levés : les campagnes d'hydrographie, les campagnes d'océanographie et les levés géophysiques. Ces levés mettent en place des capteurs spécifiques qui sont fonction des objectifs finaux de la campagne d'acquisition.

1. Un levé hydrographique est axé sur la détermination de la profondeur, la détection des hauts-fonds dangereux pour la navigation, et la mise à jour de l'information nautique environnante pour l'entretien et/ou la création des documents nautiques (cartes, instructions nautiques, livre des feux) traitant de la zone levée. L'objectif principal de la plupart des levés hydrographiques est d'obtenir les données essentielles à la compilation des cartes marines en mettant l'accent sur les détails qui peuvent affecter la sécurité de la navigation. Les autres objectifs incluent l'acquisition des informations nécessaires à l'établissement des produits pour la navigation (hors cartes marines comme les instructions nautiques), à la gestion des zones côtières, à l'ingénierie et aux sciences marines. Ces levés sont ainsi essentiels pour la sécurisation du transport maritime d'autant plus qu'aujourd'hui, plus de 87% du commerce international dans le monde transite par la mer. Le commerce maritime constitue un élément central pour l'économie d'une nation. Beaucoup de régions et de ports dans le monde ne possèdent pas de couverture en cartes marines précises et adéquates. Or les cartes marines modernes sont nécessaires pour la sécurité de la navigation dans les eaux

d'un pays, le long de ses côtes et les entrées de ses ports. Ainsi, le manque de cartes marines appropriées entrave le développement du commerce maritime dans les eaux et les ports des nations concernées. Les cartes modernes, suivant la norme S-57 [40] de l'OHI, fournissent aussi l'information nécessaire à la création de systèmes de routages tels que ceux établis par les conventions internationales et qui répondent aux intérêts économiques des États côtiers. En 2020, l'OHI a publié la 6$^{\text{ème}}$ édition de la norme S-44 [8], l'objectif de cette nouvelle version de la publication spéciale est de fournir un ensemble de spécifications destinées à l'exécution des levés hydrographiques afin de recueillir des données qui seront essentiellement utilisées dans la compilation de cartes de navigation employées pour garantir la sécurité de la navigation et la protection de l'environnement marin. Ainsi, le but de cette norme est de fournir des exigences minimales à atteindre en termes de qualité et d'incertitude concernant les levés hydrographiques. Cette norme donne également des indications sur le contrôle qualité et le traitement associé aux données acquises qui aboutiront dans le produit final principal de ces données : les cartes marines.

2. Les campagnes d'océanographie sont dédiées à l'étude des phénomènes physiques (température, salinité, courants et clarté de l'eau essentiellement), chimiques (oxygène dissout, sels nutritifs...) ou biologiques (teneur en planctons, décompte d'espèces végétales ou animales) dans la colonne d'eau. Les mesures sont réalisées le long de radiales, destinées à obtenir une coupe de l'océan pour le phénomène étudié ou bien par stations, de sorte à observer la variation des conditions hydrologiques en un point donné au cours d'une période suffisamment longue. Ces campagnes intègrent en général également la mise à l'eau de flotteurs, de mouillages ou d'appareils de mesure autonomes restituant leurs données par satellites ou lors de leur relevage plusieurs mois après leurs installations. La diversité des capteurs mis en oeuvre au cours des campagnes d'océanographie est très importante. Elle comprend notamment des marégraphes, des courantomètres, des bathysondes...

3. Les levés géophysiques ont pour objectif d'étudier les phénomènes physiques terrestres qui caractérisent ou découlent de la nature et la morphologie du fond mais également l'étude de la croûte terrestre et son évolution au tra-

vers des millénaires. Ils mettent en œuvre des techniques similaires à celles des levés hydrographiques, complétés par des capteurs spécifiques notamment par la mesure de champs de pesanteur avec des gravimètres marins, comme le gravimètre absolu GIRAFE [41] développé en collaboration avec l'ONERA (Office National d'Études et de Recherches Aérospatiales) et le Shom, ou les mesures du champ magnétique terrestre par magnétomètre remorqué, comme utilisé dans cette étude de la baie de Saint-Malo [42].

La bathymétrie est souvent acquise dans le cadre de l'un de ces trois types de levé. Elle constitue en effet une donnée primordiale pour la compréhension de tous les phénomènes physiques de l'environnement marin. Dans ce manuscrit, nous désignerons sous le terme hydro-océanographique ces trois types de levés.

La réalisation d'un levé hydro-océanographique se décompose en plusieurs étapes :

1. Rédaction du cahier des charges (dénommé instruction technique au sein du Shom) : cette étape consiste à définir les besoins et le type de données à acquérir (incluant incertitude attendue, couverture...) ayant des impacts sur une emprise spatiale et temporelle.
2. Définition des travaux à réaliser (dénommé instruction particulière au sein du Shom) : le but de cette étape est de traduire en pratique la première étape, spécifier le mode de levé, les instruments à utiliser, la stratégie de levé et le paramétrage à mettre en place pour atteindre les spécifications.
3. Préparation de la campagne : l'objectif est de s'assurer que tous les éléments (techniques, logistiques et administratifs) soient prêts pour réaliser la campagne (autorisations de levés, logistique du matériel, capteurs ajustés).
4. Acquisition à la mer : une fois l'ensemble de la mission préparée, la campagne en mer peut être conduite. Les travaux en mer sont souvent complétés par des travaux à terre (contrôle de marégraphe, mesure d'amer, nivellement topographique).
5. Traitement des mesures : cette étape consiste à traiter l'ensemble des données acquises lors de la campagne hydro-océanographique et notamment valider les données bathymétriques comme présenté dans les chapitres 2 et 3.
6. Exploitation rapide des dangers nouveaux : après un contrôle rapide des données en vue d'y isoler les dangers nouveaux, ceux-ci sont envoyés vers

la cellule du Shom qui gère la diffusion des dangers urgents pour la navigation afin d'assurer une transmission vers les usagers de la mer (avis aux navigateurs).

7. Contrôle de l'ensemble de la campagne : un contrôle systématique est réalisé pour assurer la qualité des données acquises mais également vérifier la concordance entre les données acquises et ce que le prescripteur a spécifié dans le cahier des charges (étape 1). C'est une étape de qualification du levé.
8. Rédaction de la campagne : la rédaction d'une campagne consiste en une production des métadonnées homogènes et comparables intégrables dans une base de données et la rédaction du rapport du levé qui va archiver tous les éléments constituant la campagne hydro-océanographique, comme la détermination de la SME présentée dans la section 4.1.

Pour réaliser ces levés hydro-océanographiques, il est important de mettre en place des capteurs performants et adaptés aux attentes des besoins finaux de la campagne de mesure, tout en s'assurant de répondre au juste besoin et ne pas tomber dans le piège de la surqualification pouvant être très coûteuse. Dans la sous-section suivante, nous présenterons les différents capteurs et les données collectées lors de levés hydro-océanographiques.

### 1.1.2 Capteurs employés lors d'un levé hydro-océanographique

Les capteurs utilisés lors des levés hydro-océanographiques ont beaucoup évolué à travers le temps, comme le montrent les films [37] et [43] produits par le Shom. Dans le cadre de ce manuscrit, nous nous concentrerons sur les moyens actuellement utilisés au cours des campagnes du Shom. Ces capteurs sont à l'origine des paramètres de qualité qui seront présentés dans la section 4.2.2.

#### Mesurer la profondeur

Un des objectifs quasiment systématique des campagnes hydro-océanographiques est d'obtenir la bathymétrie des fonds marins. En fonction des attentes de la campagne d'acquisition, l'ingénieur en charge du levé détermine le besoin ou non d'explorer complètement le fond, les capacités de détection nécessaires, la résolution, la précision ou la densité de sondes requises, et par conséquent le type de capteurs

adapté au positionnement, à la bathymétrie et à l'imagerie. Pour la bathymétrie, différents capteurs (acoustiques/optiques et actifs/passifs) peuvent être mis en oeuvre dont les cinq principaux sont [38], [39] :

— le Single Beam Echo-Sounder (sondeur mono-faisceau) (SBES) ;
— le SMF ;
— le sonar interférométrique ;
— le LiDAR bathymétrique ;
— le satellite via le Satellite-Derived Bathymetry (SDB).

Dans le cas des mesures bathymétriques, essentiellement les exigences suivantes ont un impact direct sur les moyens matériels à mettre en oeuvre en cours de levé [8] :

— l'incertitude horizontale du positionnement des données de bathymétrie ;
— l'incertitude verticale de ces données ;
— la couverture totale ou partielle du fond ainsi que la densité de mesures ;
— la capacité de détection dans le cas d'une couverture totale.

Le SBES mesure la hauteur d'eau sous le capteur (perpendiculairement à la base de l'antenne). Le principe général est le suivant : l'antenne du sondeur émet une onde acoustique qui se réfléchit sur le fond marin. Le signal retour ainsi que le temps de trajet de l'onde sont enregistrés par le capteur, nous parlons alors d'empreinte insonifiée au sol. Ce signal est alors analysé et le premier écho (amplitude du signal retour lorsqu'il dépasse un seuil) correspond à la réflexion du signal sur le premier objet rencontré (dans la colonne d'eau ou bien le fond). Le temps de trajet aller-retour ($\Delta t$) ainsi mesuré est transformé en hauteur d'eau ($H$) en utilisant la célérité ($c$) du son dans la colonne d'eau, et donc en profondeur d'eau sous le capteur. Cette mesure est transformée en sonde en connaissant la position du capteur par rapport à la surface de l'eau via l'ensemble de la chaîne d'acquisition hydrographique. La hauteur mesurée est ainsi donnée par l'équation 1.1.

$$H = \frac{c\Delta t}{2} \tag{1.1}$$

La résolution (capacité à séparer deux objets) horizontale du sondeur correspond à la largeur de l'empreinte insonifiée au sol et elle est fonction de l'ouverture angulaire $\theta$ (spécificité technique du SBES) et la vitesse du porteur, voir la figure 1.2. Sur des fonds chahutés, voir la figure 1.2, le premier retour ne correspond pas

nécessairement au retour à la verticale du sondeur. Cette limitation va cependant dans le sens de la sécurité de la navigation car elle générera une sonde plus courte.

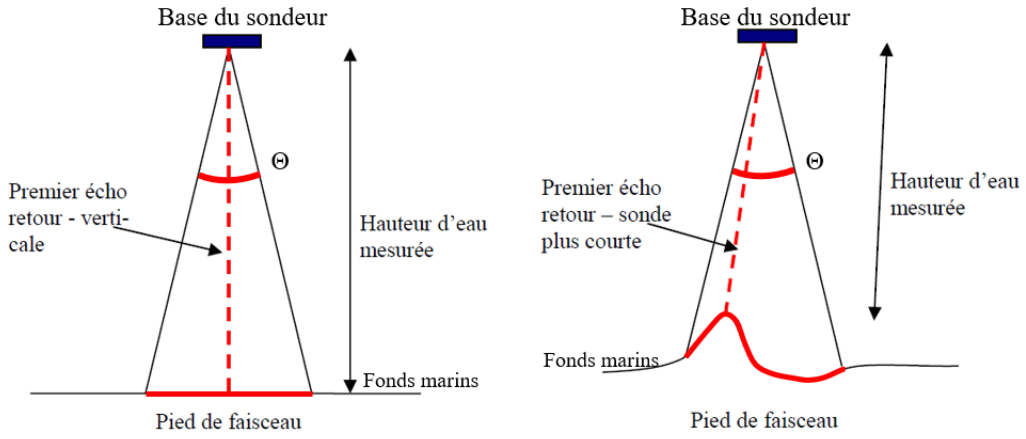


FIGURE 1.2 – Principe de fonctionnement du SBES sur fond plat à gauche et sur fond chahuté à droite.

Les deux principaux défauts de ce type de sondeur sont la limite d'échantillonnage latéral qui entraîne un manque de couverture et l'acquisition ponctuelle du sondeur. En effet, une seule sonde par empreinte au sol est retenue et l'ouverture angulaire $\theta$ est souvent élevée (sur les premiers modèles). Il en résulte une incertitude sur le positionnement exact de l'écho et une résolution horizontale faible. De plus, l'acquisition étant très ponctuelle à la verticale du sondeur, il est possible de ne pas échantillonner des reliefs plus hauts entre deux passages parallèles. La figure 1.3 montre ainsi la différence de couverture entre un levé au SBES et un levé au SMF.

Les SMF, voir [7], [39], ont permis de pallier ces deux limitations en augmentant le nombre de faisceaux latéraux afin de couvrir une plus grande surface. Cette surface couverte (*swath* en anglais) par un ensemble de faisceaux émis (*beam* en anglais) au même horodatage (*timestamp* en anglais) est appelé *ping*. Ce type de sondeur est arrivé au début des années 1990 au Shom avec le sondeur Thomson Lennermor et le Simrad EM12Dual. L'augmentation de cette zone insonifiée est clairement visible sur la figure 1.3. De plus, l'ouverture angulaire de chaque faisceau est beaucoup plus étroite que pour un SBES, la résolution de chaque empreinte au sol est donc significativement plus fine. Ainsi, la description du relief de la figure

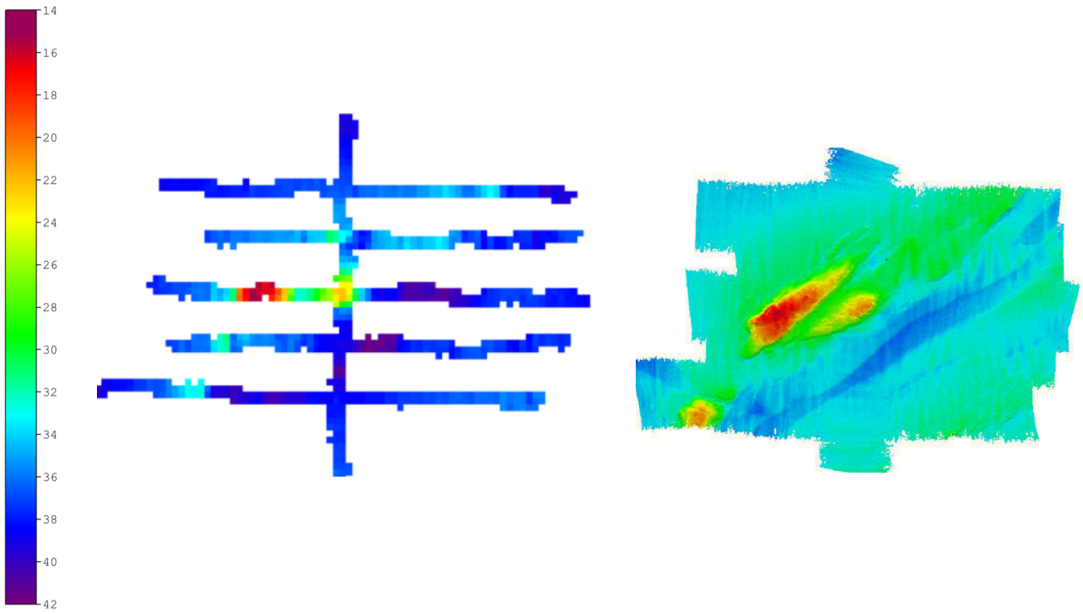


FIGURE 1.3 – Comparaison d'une même zone acquise au SBES à gauche et avec un SMF à droite.

1.3 est nettement supérieure, tant en termes de couverture que de résolution. La remontée de fond est décrite complètement. Enfin, ce type de capteur permet également de mesurer la réflectivité du fond (intensité du signal réverbéré), permettant d'obtenir une information surfacique sur la nature des fonds. La plupart d'entre eux possèdent des antennes en configuration "croix de Mills" pour la formation des faisceaux [7]. L'antenne d'émission du SMF forme un faisceau large latéralement (perpendiculairement à la route du navire) et étroit longitudinalement (dans le sens de l'avancement du navire), représenté en orange sur la figure 1.4. Les valeurs caractéristiques d'ouverture latérale de ce faisceau sont comprises en général entre 120° et 160°. Les ouvertures longitudinales sont en général comprises entre 0,5° et 4°.

Après émission, le signal se propage dans la colonne d'eau et se réfléchit (à la verticale) ou est réverbéré (en incidence oblique) par le fond. Le signal ainsi renvoyé est reçu par l'antenne de réception du sondeur. Cette antenne de réception traite le signal reçu et forme les faisceaux, représentés en bleu sur la figure 1.4. Pour chaque faisceau formé, le signal est analysé par un algorithme de détection du fond calculant le temps aller-retour de l'onde sonore entre le sondeur et le fond

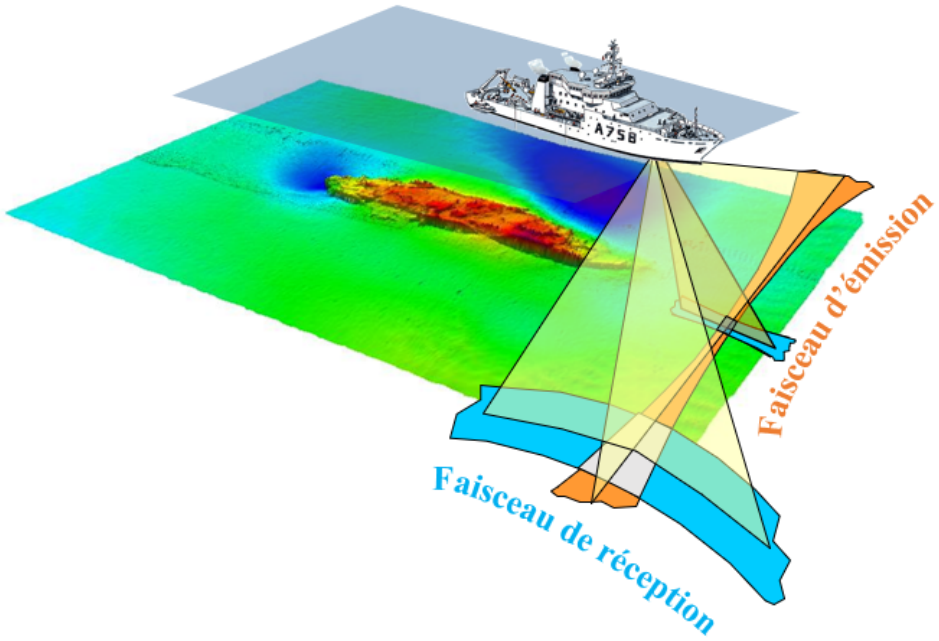


FIGURE 1.4 – Principe de fonctionnement du SMF - les mesures bathymétriques sont les intersections entre les faisceaux d'émissions (orange) et les faisceaux de réceptions (bleu), ©Shom.

(son écho). Comme pour le SBES, une fois que nous avons connaissance du temps d'aller-retour de l'onde acoustique, nous pouvons facilement en déduire la hauteur d'eau à partir du profil de célérité du son dans l'eau le long de la colonne d'eau. Il est important de mesurer cette célérité sur toute la colonne d'eau, car lorsque la célérité change (changement dû principalement à des variations de salinité ou de température), l'onde acoustique est soumise à des phénomènes de réfraction selon la loi géométrique de Snell-Descartes, voir [7]. La propagation de l'onde acoustique dans le milieu marin sera développée dans la section 1.2.1. Ainsi, nous pouvons, pour chaque faisceau et via l'équation 1.1 (en prenant en compte l'angle d'incidence pour les faisceaux hors nadir), en déduire sa hauteur d'eau associée. La dernière étape consiste à définir le positionnement absolu des sondes. Pour cela, il faut connaître la position du capteur et son orientation, c'est-à-dire en pratique : la position du navire, la position des antennes (d'émission et réception) par rapport au point de référence du navire et l'attitude du navire afin de positionner correctement le sondeur dans l'espace.

Le LiDAR, voir [44], est un système optique actif, son fonctionnement pour

obtenir de la donnée bathymétrique est très proche du SMF si ce n'est qu'il utilise une OEM plutôt qu'une onde acoustique. Ainsi, plutôt que la célérité du son dans l'eau pour le capteur LiDAR, nous utilisons la vitesse de la lumière (approximativement 0.3m par nano-seconde dans l'air). Les premières utilisations du LiDAR remontent aux années 60 pour la détection des sous-marins, puis, avec l'amélioration des techniques de positionnement à la fin des années 90, la technologie a été utilisée pour réaliser des campagnes hydrographiques de l'interface terre-mer, couplant ainsi bathymétrie avec topographie quel que soit le niveau de marée. Ce système émet des impulsions laser à deux longueurs d'ondes (vert et proche infrarouge (PIR)) par balayage d'une fauchée perpendiculaire à l'axe de vol. Il capte le retour des deux impulsions lasers, la première réfléchie par la surface de l'eau et la seconde par le fond marin. La différence de temps mesurée entre les deux retours est convertie en distance (de la même façon qu'avec l'équation 1.1). On peut calculer ainsi de la hauteur de la colonne d'eau avec cette différence et avec le signal vert on dispose de l'élévation du fond. Ce type de capteur bathymétrique émet environ 35 000 impulsions par seconde pour le laser vert et 500 000 impulsions par seconde pour le laser IR (pour les systèmes récents). Il permet d'effectuer des levés bathymétriques en zones côtières dangereuses pour la navigation dans un temps très court et pour des coûts d'acquisition plus faibles qu'un déploiement naval [39]. Malgré ces nombreux avantages, le LiDAR bathymétrique présente néanmoins un inconvénient majeur : il ne peut être utilisé sur tous les océans du globe quelle que soit sa gamme de profondeur. En effet, la turbidité de l'eau doit être faible pour s'assurer que l'onde optique ne soit pas réfléchie sur des particules en suspensions cachant le fond. De plus, l'absorption rapide de l'OEM dans la colonne d'eau ne permet pas d'atteindre des grandes profondeurs (45/50m maximum dans des eaux très claires type lagon australien [44]).

La figure 1.5 compare ainsi les zones de levé où les capteurs LiDAR topographiques, LiDAR bathymétriques et SMF peuvent être utilisés, ainsi que la vitesse d'acquisition $V$, leurs différentes altitudes $A$ par rapport à la surface d'eau et la largeur de fauchée $\theta$ couverte pour chaque technique. Nous pouvons ainsi constater la complémentarité de chaque technique pour acquérir de la donnée sur les zones immergées, d'estran et émergées. Il existe également une complémentarité en terme de capacité à atteindre une exigence sur les incertitudes verticale et horizontale de la sonde, et sur la densité de sonde sur une même zone de levé.

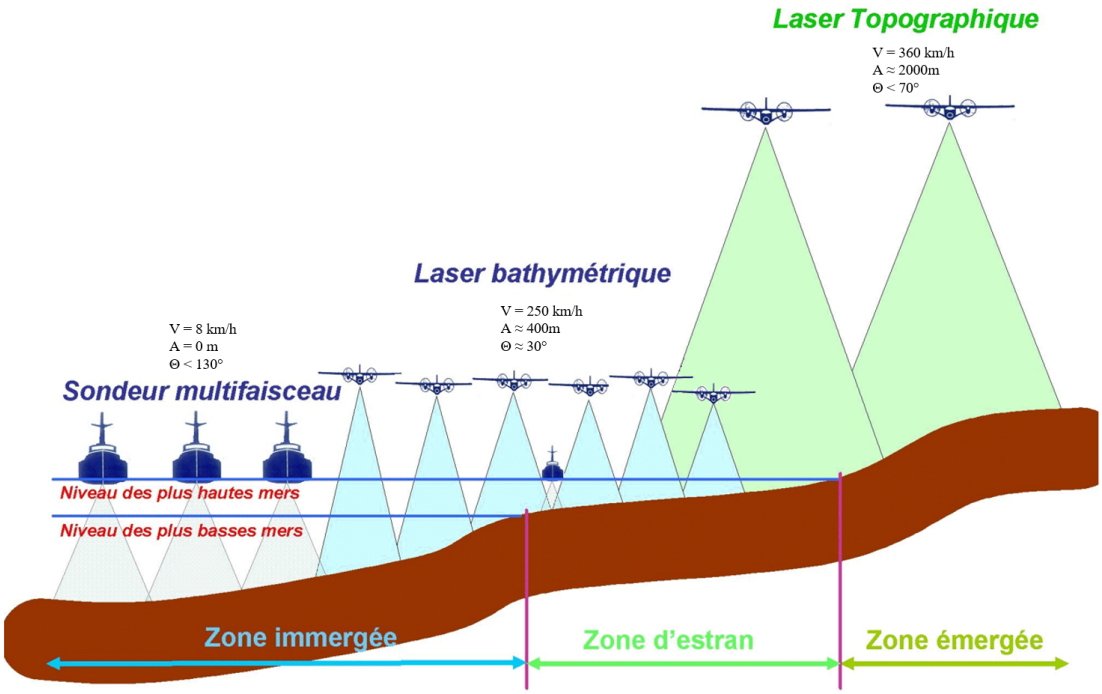


FIGURE 1.5 – Complémentarité des techniques LiDAR et SMF pour l'acquisition de données topo-bathymétriques, ©Shom.

La SDB, voir [45], est un système optique passif qui repose également sur les OEM émises par le soleil et captées par un satellite après une réflexion sur le fond marin. L'énergie lumineuse reçue par le capteur satellitaire est inversement proportionnelle à la profondeur après application des corrections atmosphériques et des corrections liées à la colonne d'eau comme présenté sur la figure 1.6. Ces capteurs fonctionnent sur des spectres panchromatiques (dans le spectre du visible et information sur une large bande spectrale), multispectraux (dans le spectre du visible et l'infrarouge sur des bandes spectrales plus étroites) et hyperspectraux (dans le spectre du visible et l'infrarouge sur des bandes spectrales beaucoup plus étroites). La SDB est la méthode la plus récente de sondage des eaux peu profondes. Elle permet un accès rapide aux données bathymétriques et permet de réaliser des économies importantes car elle nécessite très peu de mobilisation, voire aucune pour certaines méthodes se basant uniquement sur des modèles d'inversion et non des modèles empiriques. Néanmoins, l'incertitude sur la mesure bathymétrique est aujourd'hui encore importante. Par moins de 10m de fond, cette incertitude est métrique pour la SDB, alors qu'elle est décimétrique pour les méthodes acous-

tiques.

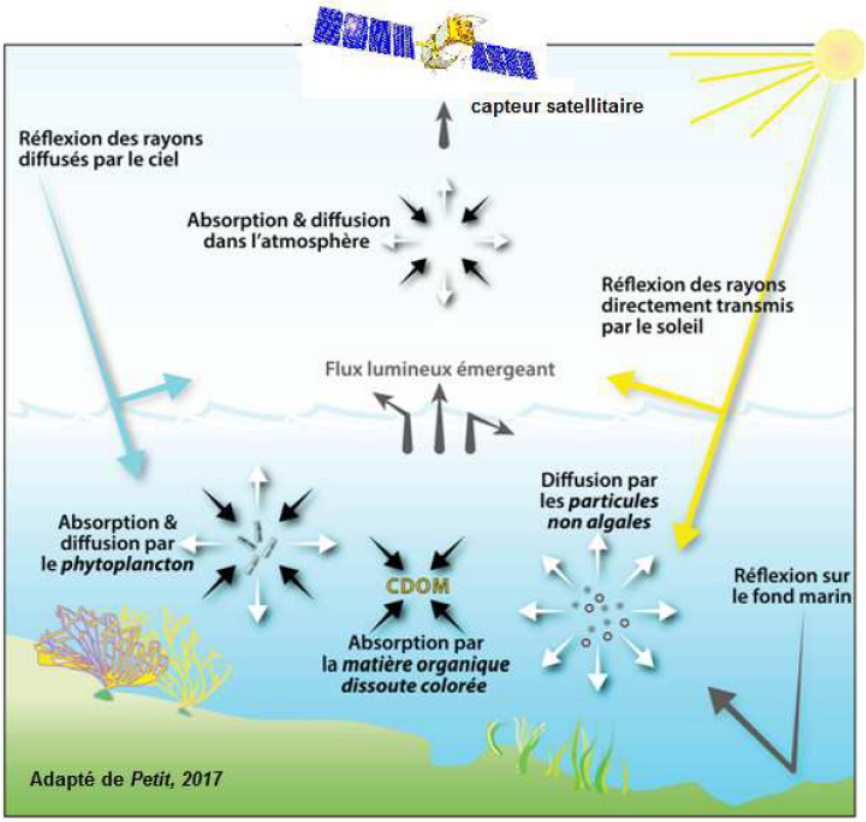


FIGURE 1.6 – Principe de fonctionnement de la SDB et les différents phénomènes d'absorption et de réflexion dans la colonne d'eau et l'atmosphère.

### Mesurer l'attitude du porteur

Mesurer l'attitude du porteur est essentiel pour compenser ses mouvements et les dépointages (angles formés entre les faisceaux et le repère du porteur) des faisceaux lumineux ou acoustiques associés. Cette information est ainsi cruciale dans l'équation de géoréférencement des sondes bathymétriques présentée à la section 1.1.3. L'attitude et/ou l'orientation sont calculées à l'aide d'une centrale d'attitude ou une Inertial Measurement Unit (station inertielle) (IMU) qui mesure des vitesses (linéaires ou angulaires) et des accélérations. Par intégration de ces mesures, nous obtenons les données de cap $\psi$, de roulis $\phi$, de tangage $\theta$ et de pilonnement en fonction du temps. La centrale d'attitude permet ainsi de mesurer 4 des 6 degrés de liberté du porteur : ses 3 rotations et son mouvement vertical.

Les données des accéléromètres sont traitées par un filtre passe-bas afin de supprimer les variations à haute fréquence mesurées sur la verticale apparente et causées par la houle, les girations ou les variations soudaines de vitesse. De même, les données des capteurs de vitesse angulaire sont traitées par un filtre passe-haut afin de supprimer les mouvements à basse fréquence. Au final, les données de sortie des filtres correspondent à l'attitude du bâtiment pour les fréquences voisines du seuil de coupure choisi (des fréquences de coupure de 5 à 20 secondes sont considérées comme normales). Au cours du temps, surviennent une déviation apparente de la verticale et des erreurs sur les mesures angulaires de roulis, de tangage et de lacets. Les capteurs inertiels sont ainsi considérés comme des capteurs biaisés dans le temps. C'est pourquoi il est nécessaire de les coupler à des capteurs Global Navigation Satellite System (GNSS) pour améliorer les capacités des capteurs d'attitude.

### Mesurer la position

La mesure de la position avec une précision répétable constitue le problème central du référencement géographique des données terrestres et la principale fonction de la géodésie, voir [38]. Pour réussir à positionner les mesures acquises durant un levé hydro-océanographique, les opérateurs utilisent le système de positionnement GNSS. Le système de positionnement GNSS est basé sur la réception d'OEM envoyées par une constellation de satellites artificiels orbitant autour de la Terre. Les constellations classiquement utilisées aujourd'hui sont le système GPS (*Global Positionning System*) américain, le système GLONASS russe et le système Galileo européen. Chaque système comporte trois composantes fonctionnelles principales :

— le segment spatial : ce segment est constitué d'une constellation d'au moins 24 satellites. Ce nombre peut être temporairement supérieur, de nouveaux satellites étant régulièrement lancés pour remplacer les plus anciens, dont la durée de vie est de quelques années. Ils sont placés dans 6 plans orbitaux inclinés de 55 ° sur l'équateur. Sur chaque plan évoluent donc 4 satellites, sur des orbites quasi-circulaires à une altitude d'environ 20 000 km, ce qui permet la réception, en tout lieu et à tout moment, d'un minimum de 4 satellites (minimum nécessaire pour une mesure fidèle de la position). Le temps de parcours d'une orbite est de 11 h 58 min, une constellation peut donc être observée à l'identique deux fois par jour dans un même lieu, et

chaque jour 4 minutes plus tôt que le jour précédent ;

— le segment commande et contrôle : ce segment est composé de stations au sol qui enregistrent les signaux GNSS reçus, effectuent des mesures météorologiques, puis transmettent toutes les données acquises vers une station principale. Celle-ci analyse les données, prépare les manœuvres des satellites pour les aligner sur leur trajectoire, procède à l'ajustement des horloges et élabore les messages de navigation. Ces messages sont ensuite envoyés vers chaque satellite. Le segment commande et contrôle calcule ainsi de façon très précise la position des satellites sur leur orbite et diffusent ces informations, appelées éphémérides, vers le segment utilisateur afin de réaliser des positionnements centimétriques ;

— le segment utilisateur : ce segment regroupe tous les utilisateurs des services GNSS, qu'ils soient civils ou militaires. Il ne subsiste actuellement plus de différence entre la qualité du service proposé au grand public et celle du service proposé aux militaires. Jusqu'en 2000, le signal GPS était volontairement brouillé par la défense américaine afin de détériorer la précision du positionnement pour les utilisateurs non-américains et/ou civils.

Le positionnement par GNSS est assuré par intersection de distances satellite-récepteur GNSS dans la référence d'un système géodésique, qui est défini dans un repère cartésien ayant pour origine le centre de la Terre et auquel est associé une ellipsoïde (représentation mathématique de la géométrie de la Terre) de référence. Si les positions des satellites dans ce système de référence sont connues, les coordonnées d'un point inconnu peuvent être reliées à celles des satellites par la mesure d'un nombre suffisant de distances entre ces derniers et le centre de phase de l'antenne d'un récepteur placé sur le point recherché. Cette méthode repose ainsi sur la trilatération et nécessite 4 satellites minimum, 3 pour la mesure de la position et 1 pour la synchronisation des différentes horloges : celle du récepteur et des satellites captés par le récepteur.

### Mesurer la marée

Les variations du niveau de la mer par rapport au niveau moyen sont dues principalement aux effets météorologiques et aux phénomènes de marée [46]. Les deux caractéristiques fondamentales de la marée sont le marnage (ou amplitude de la marée, la différence de hauteur d'eau entre la basse mer et la pleine mer) et la

période (le temps entre la basse (ou haute) mer et la haute (ou basse) mer suivante). Cette période définit le régime de marée qui varie d'une zone à l'autre, comme le montre la figure 1.7. Il peut ainsi être diurne (une marée par jour), semi-diurne comme le long des côtes en France métropolitaine (2 cycles de marée par jour), semi-diurne à égalité diurne (2 cycles de marée par jour mais avec une différence importante de marnage dans une même journée) ou de type mixte (tantôt une marée diurne et tantôt une marée semi-diurne à inégalité diurne).

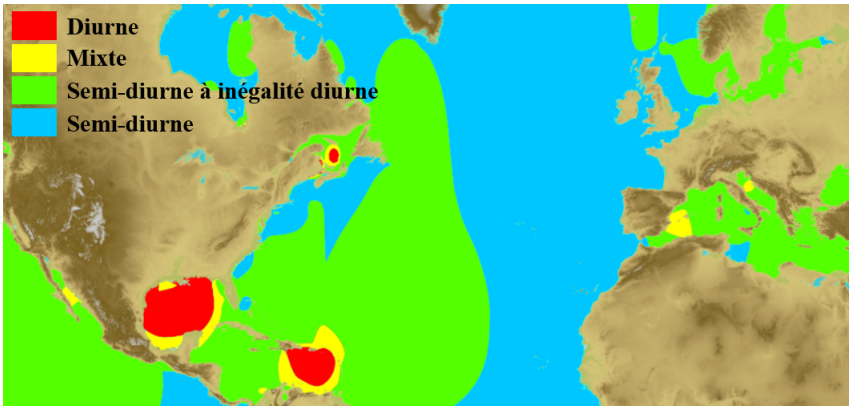


FIGURE 1.7 – Répartition des quatre types de marée en Atlantique nord.

Au cours d'une campagne hydro-océanographique, il est essentiel de mesurer la marée afin de réduire la profondeur des sondes mesurées par les sondeurs bathymétriques. L'objectif de la réduction est de rapporter toutes les mesures à un niveau de référence et aussi tenir compte de la hauteur d'eau provoquée par la marée à l'instant du sondage. En effet, les cartes marines représentent des sondes réduites au niveau de la plus basse des basses-mers astronomiques (plus basse-mer de coefficient 120) appelé zéro hydrographique ou zéro des cartes. Ces observations sont effectuées par l'intermédiaire de marégraphes sur la zone du levé. Deux types de marégraphe sont utilisé : les marégraphes immergés et les Marégraphes Côtiers Numériques (MCN). Le marégraphe immergé mesure la pression au-dessus de son capteur et en déduit la hauteur d'eau correspondante. Les données mesurées par ce type de marégraphe correspondent à des pressions. Nous devons donc observer simultanément l'évolution de la pression atmosphérique pour s'affranchir de ces variations. De plus, le marégraphe mesurant le poids de l'eau, nous devons prendre en compte la densité de l'eau de mer et son évolution dans le temps pour le calcul de la hauteur d'eau finale. Le marégraphe immergé doit ensuite être calé

afin de ramener la mesure de la hauteur d'eau au zéro hydrographique. Pour le caler, nous utilisons une mesure de marée à terre (via une échelle de marée ou de tirant d'air par exemple) ou la mesure d'un MCN installé sur les côtes françaises. Le MCN mesure des tirants d'air par ultrasons dans un puits adapté. Les données sont transmises numériquement au Shom par l'appareil et sont disponibles en temps peu différé. La section 4.2.2 présente une mesure de marée au cours d'un phénomène de tsunami exceptionnel et les paramètres de qualités que nous pouvons extraire de ce type d'acquisition.

### Mesurer la célérité du son dans l'eau

Comme présenté dans [7], [39] et par l'équation 1.1, la célérité est une mesure essentielle pour transformer le traitement de l'onde acoustique du SMF (tout particulièrement dans le phénomène de réfraction) ou SBES en une mesure de profondeur des fonds marins. Nous pouvons calculer la célérité du son dans l'eau de manière empirique, via l'équation 1.2 proposée dans [47] :

$$\begin{aligned} C(Z,T,S) = {} & 1449.05 + T[4.57 - T(0.0521 - 0.00023T)] \\ & + [1.333 - T(0.0126 - 0.00009T)](S-35) \\ & + 16.3Z[1 - 0.0026\cos 2\phi] + 0.18[Z(1 - 0.0026\cos 2\phi)]^2 \end{aligned} \tag{1.2}$$

Où $T$ est la température en °C, $S$ la salinité en parties pour mille (‰), $Z$ la profondeur en kilomètres (qui représente le phénomène physique de pression de la colonne d'eau) et $\phi$ la latitude en degré. Afin de mesurer ces éléments, une sonde Conductivity Temperature Depth (CTD), des thermistances ou un profileur de célérité (qui mesure directement la célérité dans la colonne d'eau) peuvent être utilisés. La célérité du son dans l'eau varie avec l'élasticité et la densité du milieu qui dépendent de :

— la salinité ($S$) : elle est la mesure de la quantité de sel et autres minéraux dissous dans l'eau de mer. Elle est usuellement exprimée en parties pour mille (‰). La salinité moyenne de l'eau de mer se situe aux environs de 35‰ et une variation de 1‰ se traduit par une variation approximative de la célérité de 1,3 m/s ;

— la pression ($P$) : elle est fonction de la profondeur, et le taux de variation de la célérité est d'environ 1,6 m/s pour 10 atmosphères, soit approximativement 100 mètres d'eau. Dans les grandes profondeurs, la pression joue

un rôle prépondérant sur le changement de la célérité ;

— la température ($T$) : celle de surface varie en fonction de la position géographique, des saisons et de l'heure de la journée [48]. Quant à la distribution du champ de température, elle est complexe et ne peut être prédite avec une précision suffisante pour les levés hydrographiques. Les variations de température au sein de la colonne d'eau sont également très complexes. La mesure des profondeurs est très sensible aux variations de célérité. Une variation de température d'un degré Celsius se traduit approximativement par une variation de 4,5 m/s de la célérité. La variation de température est le facteur dominant pour la variation de célérité dans la couche de mélange, couche à la surface des océans qui est chaude et homogène en température. Au-dessous de la thermocline, c'est la pression qui exerce l'influence principale.

**Mesurer le courant**

Lors des levés hydro-océanographiques, il est important de mouiller et de relever des courantomètres car leurs mesures permettent d'alimenter les instructions nautiques, les annuaires et les atlas de courant. Ces documents sont essentiels à la navigation, notamment dans les zones de forts courants et pour les navires de faible vitesse (afin d'éviter d'être déporté vers des zones dangereuses).

Les courants sont des mouvements, principalement horizontaux, des masses d'eau qui peuvent être induits par la marée ou par d'autres phénomènes. Les courants de marée sont causés par l'interaction gravitationnelle entre le Soleil, la Lune et la Terre et font partie du même mouvement général des océans qui se manifeste par les marées. Les autres types de courant, non directement liés à la marée, incluent les courants permanents du système de circulation océanique ainsi que les courants temporaires causés par les variations météorologiques. En océanographie physique, les mesures de courant peuvent être réalisées au point fixe, en déplacement vertical ou en déplacement horizontal. Actuellement, les courantomètres modernes, ou ADCP (*Acoustic Doppler Current Profilers*), utilisent l'effet doppler pour mesurer les profils de courant dans la colonne d'eau à partir d'un appareil mouillé au fond ou installé sur la coque d'un navire. Des instruments basés sur des OEM permettent également via des radars hautes-fréquences d'obtenir le courant de surface à partir de capteurs à terre. Ces mesures ponctuelles permettent

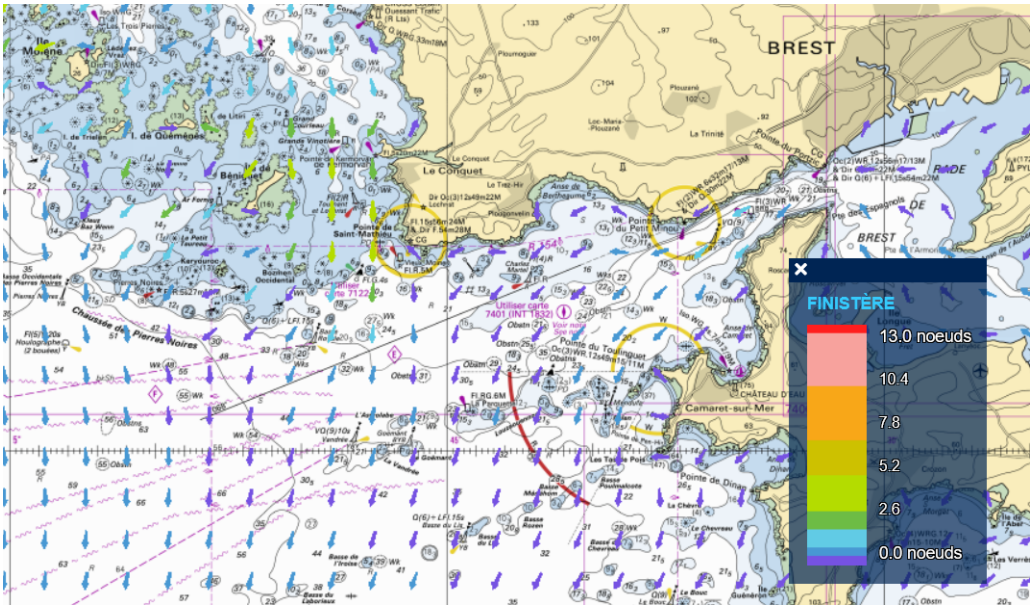


FIGURE 1.8 – Extrait du portail data.shom.fr présentant les courants de surface dans le goulet de Brest du 18/08/2022 à 13h15, ©Shom.

ensuite de calculer des prédictions sur un plus grand domaine (spatial et/ou temporel) via le même type d'étude harmonique que pour le phénomène de marée. La figure 1.8 montre ainsi la prédiction pour les courants de surface dans le goulet de Brest le 18/08/2022 à 13h15. Les courants ont un effet important sur le transport sédimentaire.

### Caractériser les sédiments

La sédimentologie, voir [49], est l'étude de la nature des fonds et des déplacements sédimentaires, déjà effective dès le XIX$^{\text{ème}}$ siècle. La carte de la figure 1 présente ainsi des annotations concernant le type de fond comme *Roche* ou *Sable*, de même la vidéo [37] montre un moyen de déterminer le type de sédiment via un plomb de sonde suiffé. Les données utilisées en sédimentologie proviennent de systèmes d'imagerie des fonds, de prélèvements et de classification. Au Shom, elles sont intégrées dans une BDD et servent à la réalisation de produits comme les cartes G (cartes décrivant les caractéristiques sédimentaires du fond marin).

Les observations et les prélèvements renseignent sur la nature des fonds, autant d'un point de vue qualitatif (texture du sédiment) que quantitatif (granulométrie). La figure 1.9 montre ainsi un prélèvement sédimentaire par benne, une étude gra-

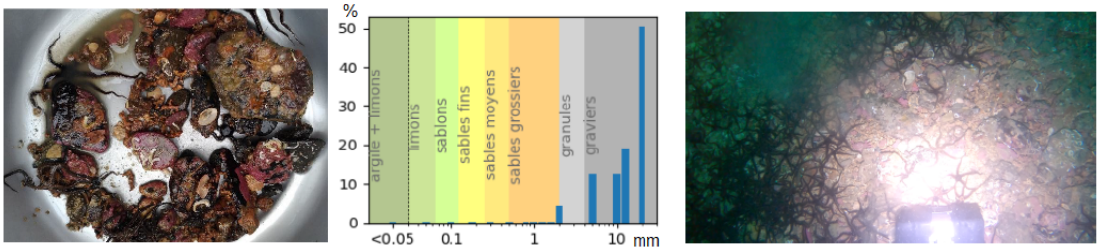


FIGURE 1.9 – Prélèvements sédimentaires, étude granulométrique (pourcentage du type de sédiment en fonction de la taille du sédiment) et image du fond sur une zone de prélèvement en rade de Brest.

nulométrique de cet échantillon et l'image du fond dans lequel l'échantillon a été prélevé.

Le SonaL est un émetteur-récepteur d'ondes acoustiques qui présente l'avantage d'insonifier les fonds sur une large fauchée. Le sonar latéral possède deux transducteurs latéraux, voir 1.2.1, qui émettent chacun une onde dans un faisceau de largeur 50°, incliné de 10° ou 20° par rapport à l'horizontale. Les faisceaux sont très fins dans la direction longitudinale. L'onde se réfléchit sur le fond et une partie de l'énergie revient vers les antennes du sonar. Le SonaL enregistre l'intensité du signal retour en fonction du temps. Selon la nature du fond (rugosité, morphologie...) rencontré par l'onde, l'intensité du signal (visualisée sous forme de nuances de gris) varie : en général un substrat dur sera foncé (ex : roche) et un sédiment meuble sera plus clair (ex : vase), voir la figure 1.10. Il en résulte une imagerie représentant la morphologie du fond. Le SonaL permet une bonne reconnaissance des structures rocheuses et sédimentaires (champ de mégarides, dunes de sable, rubans sableux etc.) mais ne permet pas une identification précise de la nature des sédiments. Le sonar latéral, en plus de produire une imagerie du fond, est utilisé pour ses capacités de détection. Étant remorqué près du fond, les faisceaux émis sont plus rasants que ceux émis par un SMF et vont donc se réfléchir plus clairement sur les obstructions, même petites. Ainsi, une obstruction est facilement identifiable sur l'imagerie SonaL grâce à l'ombre créée par l'objet. En revanche, les données recueillies par le SonaL ne sont pas positionnées précisément (notamment car les engins sont très souvent tractés) et ne permettent pas la cotation et le positionnement fidèle des obstructions repérées.

FIGURE 1.10 – Photo d'un SonaL tracté du Shom à gauche et imagerie à droite issue du même SonaL.

### 1.1.3 Le système de levé bathymétrique

Actuellement, dans les SH, une très grande partie du temps des opérateurs est consacrée à l'analyse, au traitement au le contrôle des données bathymétriques, comme présenté dans l'introduction. Il est donc important de comprendre comment sont générées ces données bathymétriques à partir des capteurs présentés dans la section précédente. Ainsi, la figure 1.11 présente l'ensemble de la chaîne de valeur des données bathymétriques permettant de transformer une mesure de profondeur au niveau du capteur SMF en une sonde géoréférencée au niveau du zéro hydrographique (voir [39]). Ces sondes géoréférencées une fois acquises doivent être traitées : ce traitement fera l'objet des chapitres 2 et 3.

Pour calculer une sonde géoréférencée au zéro hydrographique, les SMF et les SBES sont généralement associés à des capteurs complémentaires pour l'acquisition de données, avec pour objectif final de déterminer une estimation de la profondeur. Ces capteurs sont utilisés pour connaître la position et l'attitude du vecteur de mesure à tout moment de l'acquisition. L'ensemble de ces éléments est appelé système de levés bathymétriques SMF (en soit un système de systèmes). Il est généralement composé de trois systèmes :

— le capteur de position GNSS, qui mesure la position du centre de phase du récepteur GNSS pour le porteur dans le repère de navigation $(n)$ ;
— le capteur d'attitude IMU, qui donne l'orientation de l'IMU au centre du capteur. Ce capteur est également le point de référence du système de levés bathymétriques (souvent positionné au point tranquille du navire soit le

métacentre de carène correspondant au centre de gravité à la première mise à l'eau du navire) par rapport au repère navire $(bI)$ ;

— la mesure de la profondeur au niveau du centre acoustique du SMF ou du SBES dans le repère sondeur $(bS)$.

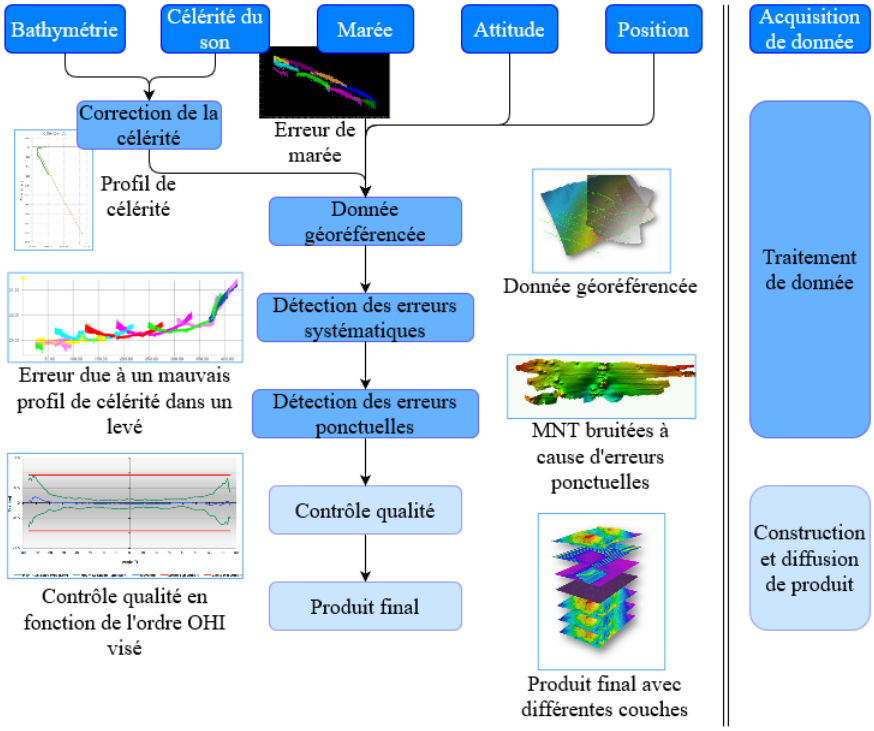


FIGURE 1.11 – Chaîne de valeur des données bathymétriques.

Le repère de navigation $(n)$ est défini par rapport à un ellipsoïde, l'axe $X$ est orienté vers le nord géodésique, l'axe $Y$ vers l'est et l'axe $Z$ suit la normale à l'ellipsoïde (repère direct).

Le repère navire $(bI)$ est fixé au corps de la plate-forme utilisée pour l'acquisition des données. L'axe $u$ indiquera la direction vers l'avant, l'axe $v$ la direction vers la droite et l'axe $w$ est orienté vers le bas (perpendiculaire au plan $(u, v)$. Désignons par $\phi$, $\theta$, $\psi$, les angles d'Euler avec les conventions de signe suivantes que nous appliquerons (voir la figure 1.12 qui représente le repère navire) :

— $\phi \geq 0$ : roulis, côté tribord vers le bas

— $\theta \geq 0$ : tangage, proue vers le haut

— $\psi \geq 0$ : cap, tourne dans le sens horaire

Le repère sondeur $(bS)$ est attaché au capteur avec l'axe $x$ qui indiquera la

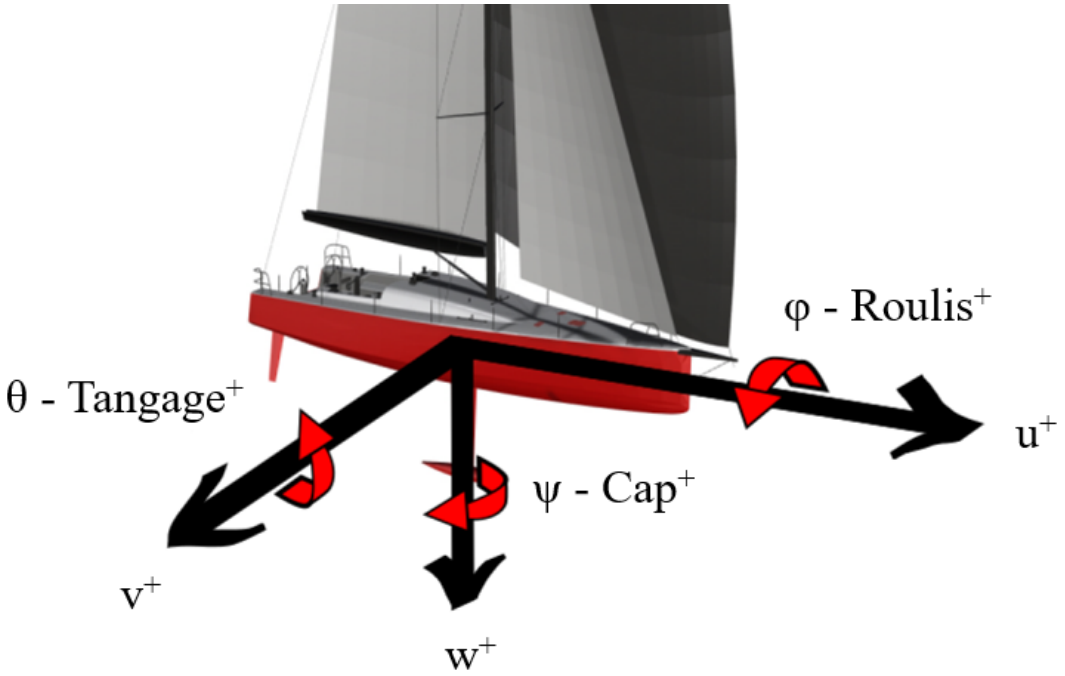


FIGURE 1.12 – Représentation du repère navire $bI$.

direction vers l'avant, l'axe $y$ la direction vers la droite et l'axe $z$ orienté vers le bas. Ce repère est très proche du repère navire mais peut être désaligné (limite de l'intégration) et comme indiqué plus haut, les centres des repères ne sont pas identiques. Dans ce repère, les coordonnées d'une sonde $r_{bS}$ s'expriment par 1.3 :

$$r_{bS} = \begin{pmatrix} 0 \\ -r_i \sin \alpha_i \\ r_i \cos \alpha_i \end{pmatrix} \tag{1.3}$$

Ce vecteur $r_{bS}$ est calculé à partir de l'angle de dépointage du faisceau (l'angle formé entre l'axe $z$ et le faisceau acoustique aussi appelé angle d'incidence) $\alpha_i$ et $r_i$ la distance oblique obtenue à partir du temps de propagation de l'onde acoustique et du profil de célérité adéquat.

Ainsi, désignons maintenant par $C_{bS}^{bI}$ la matrice de passage (matrice composée de trois rotations) entre le repère $(bS)$ et le repère $(bI)$, voir [50] pour son expression mathématique. Cette transformation représente le désalignement entre le SMF et l'IMU. Par conséquent, le vecteur $C_{bS}^{bI} r_{bS}$ est la coordonnée d'une sonde SMF dans le repère $(bI)$ de l'IMU. Considérons maintenant l'emplacement du centre de phase

du récepteur GNSS, ses coordonnées dans le référentiel de navigation sont désignées par $P_n$. De la même façon, $C_{bI}^{n}$ désigne la matrice de passage entre le repère $(bI)$ et le repère $(n)$. Compte tenu du fait que le centre acoustique du SMF ne coïncide pas avec le point de référence du système de levés, nous désignerons par $a_{bI}$ le bras de levier entre ces deux points. Ses coordonnées se situent dans le repère navire $(bI)$. Enfin, un terme correctif $Z_corr$ sur la verticale doit être ajouté à toutes les sondes pour les amener à une même référence verticale (prise en compte du phénomène de marée ou rattachement au zéro hydrographique pour les données à l'ellipsoïde), voir 1.4 :

$$Z_{corr} = \begin{pmatrix} 0 \\ 0 \\ z_{ZH} \end{pmatrix} \tag{1.4}$$

Avec l'ensemble de ces éléments, une sonde peut être ramenée dans le référentiel $(n)$, et les retours réels géoréférencés $X_n$ du SMF sont intégrés dans un référentiel géodésique absolu grâce à l'équation de géoréférencement 1.5, voir [51].

$$X_n = P_n + C_{bI}^{n}(C_{bS}^{bI} r_{bS} + a_{bI}) + Z_{corr} \tag{1.5}$$

## 1.2 Les fondamentaux de la propagation acoustique et optique sous-marine

Il est important de comprendre la physique de la propagation sous-marine des ondes acoustiques et OEM afin de mieux identifier et comprendre les sources potentielles d'erreurs associées, qui seront présentées dans le chapitre suivant. Les phénomènes physiques permettant l'acquisition de données SMF et LiDAR bathymétriques sont de natures différentes (comme présenté dans la section 1.1.2 et dans les sections suivantes). Ne seront présentés dans cette section que les éléments concernant le SMF et LiDAR actuellement en production au Shom. Néanmoins, les géométries produisant les nuages de points sont similaires sur certains aspects (comme nous le verrons dans la section 2.1 et 3.1).

### 1.2.1 L'acoustique

Les systèmes SMF et SBES fonctionnent à partir d'un transducteur piézo-électrique. Traditionnellement, un système qui convertit l'énergie électrique en énergie acoustique est appelé émetteur ou source acoustique alors que le contraire (acoustique vers électrique) est appelé hydrophone. Ainsi, en appliquant une tension aux bornes du matériau (dans le cas des sondeurs, il s'agit de céramique synthétique), celui-ci se déforme permettant de créer une onde acoustique. À l'inverse, si nous appliquons une contrainte physique sur le matériau, alors une charge électrique apparaîtra [52]. Les performances de ces transducteurs en émission sont caractérisées par :

— le rendement : soit la transmission de la puissance de l'onde émise sans perte, associée au gain d'envoi (puissance) de l'onde ;
— la directivité : soit la capacité à pointer l'énergie acoustique dans une direction spécifique en atténuant les autres directions non privilégiées (lobes secondaires), associée à la géométrie du transducteur d'émission (l'antenne d'émission) ;
— la largeur de bande : soit la capacité à émettre sur une large bande de fréquence sans transformer l'onde acoustique (en phase et en amplitude), associée également à la géométrie du transducteur d'émission.

Et en réception :

— la sensibilité : seuil de détection des ondes acoustiques (pour détecter des signaux très faibles) ;
— l'incertitude : fidélité de la mesure, pour s'assurer d'avoir peu de distorsion ;
— la directivité : soit la capacité à sélectionner les ondes acoustiques dans une direction spécifique, associée à la géométrie du transducteur de réception (antenne de réception) ;
— la largeur de bande : soit la capacité à écouter sur une large bande de fréquence, associée également à la géométrie du transducteur de réception.

L'équation du sonar permet de comparer le signal émis et la sensibilité du capteur de réception afin de détecter une cible. Nous réalisons ainsi un bilan d'énergie (NRJ) pour estimer si une cible est détectable. De manière simplifiée, nous pouvons décrire l'équation du sonar comme $NRJ_{signal}/NRJ_{bruit} > Seuil$, soit exprimé en décibels : $NRJ_{signal} - NRJ_{bruit} > Seuil_{log}$ [53]. Les phénomènes qui vont avoir une influence sur les performances sont donc les suivants [7], [52] :

— le système sonar ;
— le milieu de propagation ;
— la cible à détecter.

L'équation 1.6 du sonar actif (qui représente le cas du SMF et du SBES) décrite précédemment devient alors [52] :

$$SL - 2TL + TS - NL + DI + PG > DT \tag{1.6}$$

Les termes de l'équation du sonar associés au système sonar sont le *Source Level (SL)* soit la puissance (ou énergie) émise par le sondeur, le *Directivity Index (DI)* soit le gain d'antenne associée à sa direction privilégiée, le *Processing Gain (PG)* soit le gain de traitement du signal, et le *Detection Threshold (DT)* soit le seuil de détection nécessaire à l'antenne de réception.

Le milieu de propagation va lui aussi avoir un impact sur les performances de détection de la cible. Il va ainsi provoquer des pertes liées à l'absorption dans l'eau et le sédiment des ondes acoustiques ainsi que des dispersions géométriques qu'on notera *Transmission Loss (TL)* soit les pertes en transmission. Les réverbérations qui seraient dues aux trajets multiples sont corrigées par le *DI*. De plus, des signaux parasites viennent perturber la détection et sont notés *Noise Level (NL)*, correspondant au bruit du fond de la mer provenant de sources abiotiques, anthropiques ou bien biologiques [54].

Enfin, le terme associé à la cible est le *Target Strength (TS)* qui correspond à la puissance de l'écho de la cible. Le *TS* dépend de la nature de la cible (la force du retour de l'onde acoustique sera différente si elle touche une roche ou du sable), de sa dimension et du type d'onde émise (angle d'incidence et fréquence utilisée). Le *TS* caractérise le pouvoir réfléchissant lorsque l'onde acoustique va entrer en contact avec la cible. À partir de cette grandeur, nous pouvons retrouver la force du signal rétrodiffusé aussi appelé *backscatter* en anglais.

Les pertes de propagations *TL* sont présentes à l'aller et au retour de l'onde acoustique dans la colonne d'eau et sont donc comptées deux fois.

Lorsque le retour d'énergie atteint un pic au niveau de l'antenne de réception du sondeur, la détection du fond est réalisée. Comme précisé plus haut, nous pouvons récupérer le *backscatter* permettant de caractériser le type de sédiment sur lequel l'onde acoustique s'est réfléchie. La figure 1.13 permet ainsi de montrer différents niveaux de *backscatter* (entre -15db et -30db) et d'observer différents

faciès géoacoustiques ainsi que la présence de roches (les points les plus foncés), [8], [38].

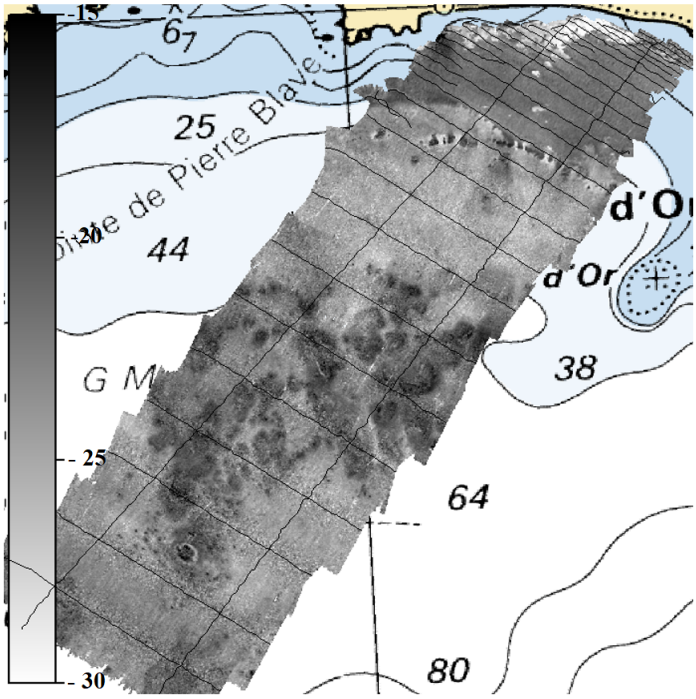


FIGURE 1.13 – Carte du *backscatter* (signal rétrodiffusé) au large de l'île d'Or, où nous pouvons observer des roches à un niveau de -15db.

L'intensité de l'onde acoustique associée à une détection de fond s'avère donc une mesure très pertinente pour la suppression des données aberrantes car, assez naturellement, les sondes les plus aberrantes présentent souvent une intensité plus faible (et sont majoritairement filtrées par le sondeur directement).

### 1.2.2 L'optique

Comme les ondes acoustiques, la propagation des OEM dans l'eau de mer est fonction de la température, de la pression, de la salinité et des particules en suspension dans la colonne d'eau. En effet, l'eau de mer est jusqu'à un certain point, transparente à la lumière. Par conditions idéales et sans matière en suspension, l'atténuation est fonction des phénomènes d'absorption et de dispersion des OEM. Pour les fenêtres infrarouges et visibles du spectre électromagnétique, la transparence de l'eau de mer dépend de la quantité de matières en suspension. C'est elle

qui fixe la limite d'utilisation du LiDAR lors de levés bathymétriques. La profondeur maximale d'utilisation du LiDAR (pour le laser vert) est approximativement 2 à 3 fois la profondeur observée avec un disque de Secchi [38]. Le disque de Secchi, voir figure 1.14, est un dispositif permettant de caractériser expérimentalement la transparence/turbidité de la colonne d'eau. C'est pour cette raison que les levés LiDAR bathymétriques sont peu adaptés aux estuaires ayant une dynamique sédimentaire importante comme l'estuaire de la Gironde.

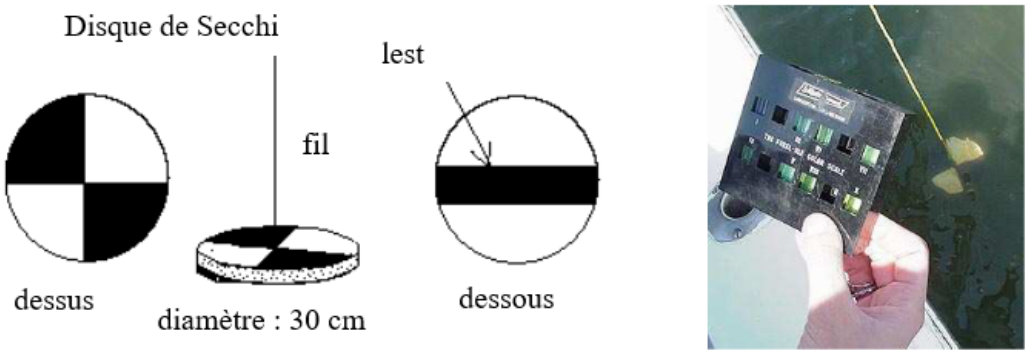


FIGURE 1.14 – Marques de peinture sur le disque de Secchi à gauche et utilisation sur le terrain à droite.

Comme montré sur la figure 1.15, le LiDAR exploite un signal IR, qui se réfléchit sur la surface de l'eau permettant ainsi de déterminer la distance avion-surface d'eau, et un signal vert qui pénètre davantage dans la colonne d'eau et qui permet de déduire la distance $H$ entre la surface et le fond marin grâce à une mesure fidèle de la différence de temps entre les deux retours [44]. La surface et le fond sont détectés grâce à un pic d'intensité au niveau du capteur LiDAR. Comme pour le SMF, avec l'intensité acoustique (en prenant bien en compte l'angle d'incidence de l'onde), l'intensité lumineuse associée à chaque sonde est un bon indicateur du type de sonde étudié (aberrante, surface d'eau, fond). Il s'agit donc d'un descripteur potentiel pour notre approche ML future.

L'OEM laser verte change de milieu et passe de l'atmosphère au milieu marin : l'onde est donc réfractée au changement de milieu et suit la loi de Snell-Descartes suivante : $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$, avec $n_1$ (respectivement $n_2$) l'indice de réfraction de l'air (respectivement de l'eau), $\theta_1$ l'angle d'incidence pris entre la normale au point d'incidence et le rayon incident, et $\theta_2$ l'angle de réfraction pris entre la normale au point d'incidence et le rayon réfracté.

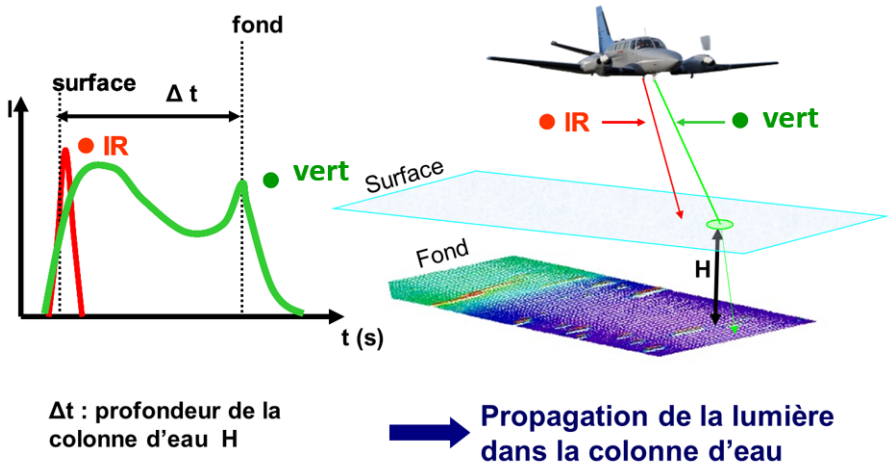


FIGURE 1.15 – Principe de fonctionnement du LiDAR - le retour de l'intensité lumineuse en fonction du temps à gauche et une représentation des lasers vert et PIR émis depuis l'avion, ©Shom.

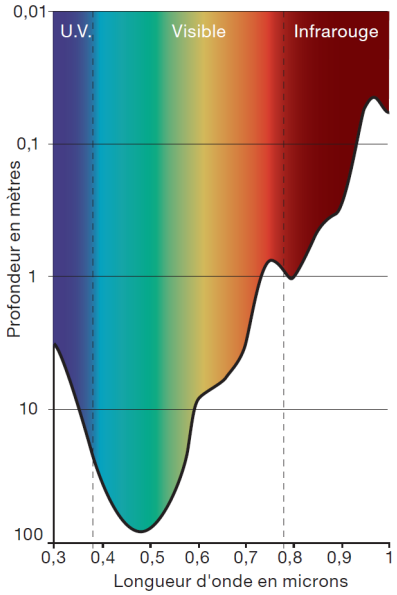


FIGURE 1.16 – Profondeur de pénétration du rayonnement solaire dans l'eau pure en fonction de sa longueur d'onde, issue de [55].

La figure 1.16 représente la profondeur de pénétration du rayonnement solaire dans de l'eau pure, définie comme la profondeur à laquelle il ne reste plus que 1% de l'énergie incidente [55]. Ainsi, ce graphe explique l'intérêt du laser vert dans le cadre de la mesure bathymétrique. En effet, le spectre de la lumière allant du jaune

(longueur d'onde 0.6 microns) à l'IR (au dessus de 0.8 microns), pénètre très peu dans la colonne d'eau (plus de rayonnement au-dessus d'un mètre de profondeur, voir 1.16). À l'inverse, les longueurs d'onde autour du vert (aux environs de 0.5 microns) pénètrent beaucoup mieux dans la colonne d'eau, jusqu'à presque 100 mètres mais dans une eau sans particule en suspension [55].

## 1.3 La science des données et l'apprentissage machine

*I do not fear computers. I fear the lack of them.*

Isaac Asimov au travers de cette citation nous offre sa vision de l'avenir et de l'impact des ordinateurs dans le futur. Or, nous ne pouvons que constater aujourd'hui la véracité de cette vision au vu de l'omniprésence des ordinateurs dans notre quotidien que ce soit pour notre travail, nos loisirs ou dans nos activités journalières. L'arrivée de la Covid-19 n'a fait qu'accélerer cette utilisation en favorisant le travail à distance et dématérialisé.

Dans le contexte des sciences informatiques, les ordinateurs ont toujours détenu et détiendrons toujours une place centrale. Nous pouvons également constater ces dernières années que la puissance de calcul et le volume de données toujours plus important ont provoqué un changement de dogme dans la manière de traiter les données et de les transformer en information. Ainsi, l'apprentissage automatique a opéré un changement complet de paradigme, voir la figure 1.17, par rapport à la précédente génération d'Intelligence Artificielle (IA) : pour les systèmes experts, l'approche se veut désormais inductive : il ne s'agira plus pour un informaticien de coder les règles à la main mais de laisser les ordinateurs les découvrir par corrélation et classification, sur la base d'une quantité massive de données. Autrement dit, l'objectif de l'apprentissage automatique n'est pas réellement d'acquérir des connaissances déjà formalisées mais de comprendre la structure des données et de l'intégrer dans des modèles, dans le but notamment d'automatiser des tâches.

L'IA a connu un nouvel essor depuis 2010, notamment grâce aux algorithmes dits d'apprentissage automatique et tout particulièrement l'apprentissage profond. Deux facteurs sont à l'origine de ce nouvel engouement des chercheurs et des industries informatiques : l'accès à des volumes massifs de données et l'amélioration des capacités de calcul des processeurs des cartes graphiques informatiques pour

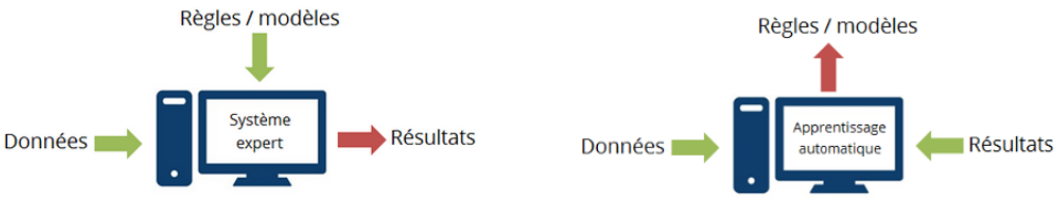


FIGURE 1.17 – Présentation du système expert à gauche et de l'apprentissage automatique ou machine à droite, crédit Conseil de l'Europe.

accélérer le calcul des algorithmes d'apprentissage. De fait, l'actuelle « révolution » de l'IA ne provient donc pas d'une découverte de la recherche fondamentale mais de la possibilité d'exploiter avec efficacité des fondements relativement anciens, tels que l'inférence bayésienne (XVIII$^{\text{ème}}$ siècle) ou les neurones formels (par Warren McCulloch et Walter Pitts dès 1943) par l'une des sous-classes de l'apprentissage automatique, l'apprentissage profond (ou *deep learning*).

Plus spécifiquement, le ML (apprentissage machine ou apprentissage automatique en français) est un champ d'étude de l'IA. Il regroupe un ensemble de techniques permettant l'extraction de connaissances sous la forme de modèles directement à partir des données. Les champs d'application de ces méthodes sont très vastes et comprennent notamment la perception de l'environnement (vision, reconnaissance d'objets, reconnaissance de caractères), les moteurs de recherche ou processus de recommandation, la locomotion de robots, l'aide au diagnostic (médical, bio-informatique), la détection de fraudes, l'analyse financières, etc. [56], [57].

Classiquement, le ML est divisé en trois branches : l'apprentissage non-supervisé (*unsupervised learning* en anglais), l'apprentissage supervisé (*supervised learning* en anglais) et l'apprentissage par renforcement (*reinforcement learning* en anglais) [57], comme présenté sur la figure 1.18. On parle également d'apprentissage semi-supervisé qui se situe ainsi entre l'apprentissage supervisé et l'apprentissage non-supervisé. L'apprentissage non-supervisé et l'apprentissage supervisé seront détaillés dans les deux prochaines sous-sections.

Contrairement à l'apprentissage supervisé et non-supervisé, l'apprentissage par renforcement ne repose pas sur un jeu de données statiques, mais sur une succession d'expériences dans un environnement dynamique. Les points de données, ou

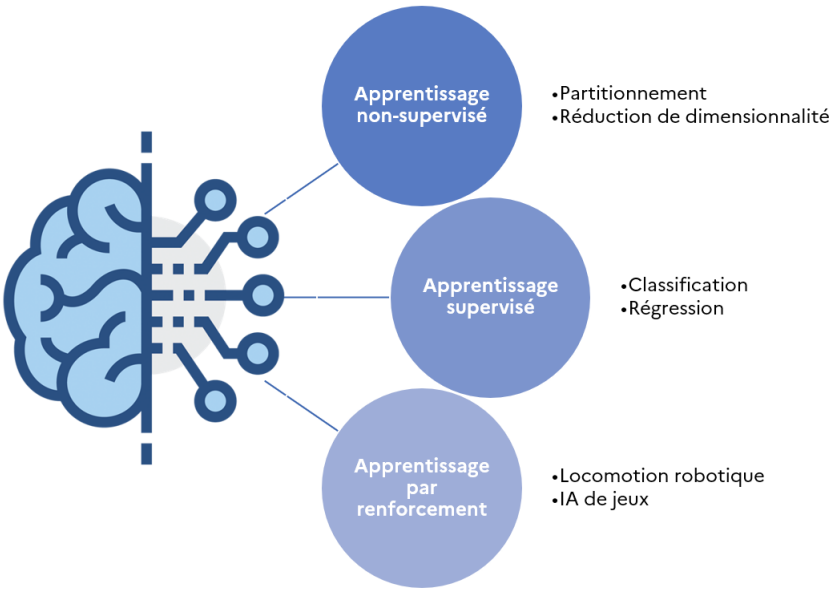


FIGURE 1.18 – Les trois branches de l'apprentissage machine.

expériences, sont recueillis lors des interactions entre un agent et son environnement, sur la base d'un apprentissage par essai-erreur. Le fondement du mécanisme d'apprentissage reflète de nombreux scénarios du monde réel. Appliquons la terminologie "apprentissage par renforcement" au dressage d'un chien 1.19. Le but de l'apprentissage dans ce cas est d'entraîner le chien (l'agent) à accomplir une tâche dans un environnement qui englobe à la fois l'animal et le dresseur, voir [56].

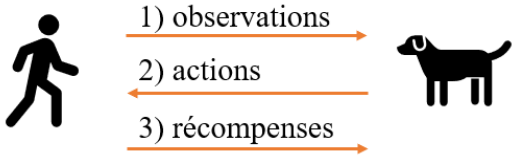


FIGURE 1.19 – Les trois étapes de l'apprentissage par renforcement.

Les trois étapes de l'apprentissage par renforcement sont :

— étape 1, pour commencer, le dresseur donne un ordre ou un signal que le chien observe (observation) ;
— étape 2, le chien réagit à cette observation, c'est l'action ;
— étape 3, le dresseur récompense, *reward* dans la terminologie anglaise (positivement si c'est l'action attendue, négativement sinon).

Il est probable qu'en début d'apprentissage le chien réagisse de manière plus ou moins aléatoire, comme rouler sur le dos lorsque nous lui disons "assis", parce qu'il essaie d'associer des observations spécifiques à des actions et à des récompenses. Cette mise en correspondance, ou *mapping*, entre observations et actions, est appelée « politique ». Du point de vue du chien, l'idéal serait de réagir correctement à chaque demande, afin de recevoir un maximum de friandises. Le but de l'apprentissage par renforcement est donc de « paramétrer » la politique du chien afin qu'il apprenne des comportements qui maximiseront la récompense. À la fin de l'apprentissage, le chien doit être capable d'observer son maître et de faire ce qui est demandé, par exemple s'asseoir lorsqu'on lui dit "assis", en s'appuyant sur la politique interne qu'il a développé. À ce stade, les friandises sont les bienvenues mais devraient théoriquement ne plus être nécessaires. L'apprentissage est terminé.

L'apprentissage par renforcement n'a pas encore été exploité dans le traitement des levés hydro-océanographiques mais il fait l'objet de publications dans le cadre de l'automatisation de l'acquisition des données comme pour la navigation autonome des Autonomous Underwater Vehicle (AUV) [58]. Néanmoins, son utilisation fait partie des perspectives d'études associées au traitement des données bathymétriques comme présenté dans la section 5.

Les méthodes d'apprentissage supervisées et non-supervisées sont fondées sur le même mode de fonctionnement inductif, de l'expérimentation à l'acquisition de connaissances. C'est pourquoi ces méthodes requièrent un volume important de données afin d'être confrontées plusieurs fois au même cas d'apprentissage et de capitaliser sur cette redondance d'informations. Nous retrouvons trois éléments clés dans ce contexte :

— $X$ : l'ensemble des observations, échantillons, exemples ;
— $y$ : l'ensemble des étiquettes, labels, groupes ;
— $\mathcal{F}$ : la fonction à apprendre tel que $\mathcal{F}(X) = y$

De cette façon, tout processus d'apprentissage passe par ces cinq étapes [56], [57] :

1. Récupération d'échantillons bruts (*dataset* en anglais) ;
2. Sélection de descripteurs / caractéristiques (*features* en anglais) ;
3. Choix du modèle / méthode d'apprentissage ;

4. Entraînement / apprentissage (détermination des paramètres de $\mathcal{F}$, *training* en anglais) ;
5. Évaluation de l'entraînement, qualification du modèle (procédure de *testing* et *validation* en anglais).

Ce processus est souvent itératif car à l'issue de l'évaluation (étape 5), nous pouvons modifier les descripteurs pour essayer d'améliorer le résultat général, ou nous pouvons décider de modifier la méthode d'apprentissage ou contraindre différement les paramètres pour comparer l'évaluation finale sur les échantillons bruts de départ. Nous pouvons également changer le *dataset* d'entrée pour évaluer la généralisation du modèle.

### 1.3.1 Les méthodes non-supervisées

Les méthodes non-supervisées sont utilisées lorsque l'étiquette sur les observations $X$ n'est pas connue. L'apprentissage se fait donc sans "professeur". Dans certains cas, il peut être complexe d'obtenir des étiquettes sur des jeux de données importants car cette action est coûteuse en temps. La constitution d'une base d'apprentissage dans laquelle est associée à chaque observation une valeur à prédire constitue de ce fait un enjeu important de l'apprentissage machine. Ce type de méthode a été mise en place dans la section 2.4 et la section 3.2.3.

Conscient de l'importance de la base d'apprentissage, Google a, par exemple, acheté *Recaptacha* en 2009 : un système de détection automatisé qui permet de constituer des bases de données d'images étiquetées via un test de Turing inversé pour s'assurer qu'un site web ne se trouve pas en présence d'un robot informatique (*bot* en anglais). Grâce à ce type de technologie, des millions d'utilisateurs participent à la génération d'une base de connaissance (reconnaissance d'objet dans une image), là où une personne n'aurait pu aboutir à un tel résultat.

L'apprentissage non-supervisé s'attaque à deux problématiques importantes de la science de donnée : le partitionnement de données ou *clustering* en anglais et la réduction de dimension.

#### Clustering

Le clustering est une méthode utilisée en analyse des données qui vise à diviser un ensemble de données en différents groupes (ou *cluster* en anglais) homogènes.

En effet, les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (degré de similarité en mathématiques basé sur une différence de distances) que nous définissons en introduisant des mesures et classes de distance entre objets.

La figure 1.20 représente schématiquement ce qui est réalisé lors d'un partitionnement de donnée. Nous passons ainsi d'une donnée brute, sans étiquette, à une donnée intégrée dans un sous-ensemble avec, pour étiquette, le groupe auquel elle appartient. Le paramètre souvent essentiel de ce type de méthode est le nombre de groupes $K$ qui devra être formé par le partitionnement de donnée.

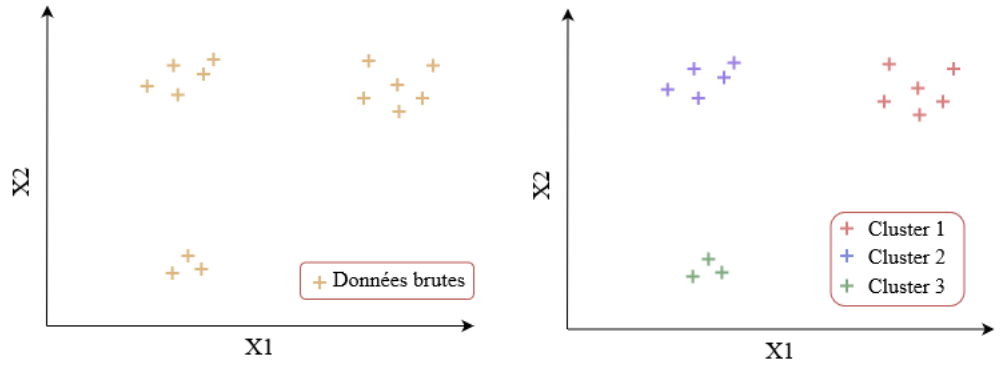


FIGURE 1.20 – Partitionnement de donnée pour $K = 3$ groupes, à gauche la donnée brute et à droite la donnée partitionnée.

Il existe de nombreuses méthodes pour réaliser un clustering. Nous pouvons citer par exemple l'algorithme K-means [59] qui sépare les échantillons en $K$ groupes de variance égale, en minimisant un critère appelé l'inertie, qui mesure la cohérence interne des clusters. Dans le chapitre 2, les algorithmes de clustering Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [60] et Topological Mode Analysis Tool (ToMATo) [61] seront présentés en détail dans le cadre du traitement de données bathymétriques.

Cette famille d'algorithmes présente plusieurs intérêts comme :

— améliorer la compréhension d'un jeu de données (analyse exploratoire) afin d'identifier des utilisateurs aux profils similaires ;

— faciliter la représentation et la visualisation des données en limitant par exemple la visualisation à un représentant (barycentre par exemple) du cluster ;

— inférer des connaissances, comme l'identification des images représentant le même objet (sans pourtant pouvoir étiqueter directement le type d'objet).

### Réduction de la dimensionnalité

La réduction de dimension consiste à transformer la donnée d'un espace de grandes dimensions en un espace de faibles dimensions de sorte que la nouvelle représentation conserve les propriétés les plus significatives des données originales et idéalement proches de leur dimension intrinsèque. Travailler dans des espaces de grandes dimensions peut être complexe pour de nombreuses raisons comme le coût de calcul, la visualisation difficile et la complexification de l'apprentissage. Ainsi, les données brutes sont souvent clairsemées en raison du "fléau de la dimension", terme inventé par Richard Bellman en 1961 [1] pour désigner l'occurence de divers phénomènes lorsque nous cherchons à analyser ou organiser des données dans des espaces de grande dimension alors qu'ils n'ont pas lieu dans des espaces de dimension moindre. En effet, en grande dimension, nous avons besoin de beaucoup plus de points pour couvrir tout l'espace et les observations sont loin les unes des autres. Il est donc difficile de trouver ce qu'elles peuvent avoir de commun ou de différent.

La réduction de la dimensionnalité est donc une méthode très importante pour l'apprentissage supervisé qui permet notamment d'améliorer la qualité de l'apprentissage. En utilisant moins de variables, nous pouvons construire des modèles plus simples et ainsi éviter le piège du surapprentissage.

Classiquement, la réduction de la dimension repose sur deux techniques : la sélection de descripteurs pertinents (afin de ne conserver que les descripteurs ayant le plus d'impact pour un apprentissage) et la projection des caractéristiques. Une des méthodes classiques pour la projection des caractéristiques est l'Analyse en Composantes Principales (ACP) [62]. Cette méthode consiste à transformer des variables corrélées entre elles en nouvelles variables le plus décorrélées possibles les unes des autres. Ces nouvelles variables sont nommées composantes principales ou axes principaux. Le but de l'ACP est de trouver une nouvelle base orthonormée dans laquelle représenter les données, telle que la variance des données selon ces nouveaux axes soit maximisée. Les composantes principales sont les vecteurs propres de la matrice de covariance des données, classés par ordre décroissant de valeur propre correspondante.

La figure 1.21 montre le passage d'un ensemble de points de la dimension 2 à la dimension 1. Nous voyons sur la partie gauche de la figure, les données dans l'ensemble des descripteurs $(X_1, X_2)$ et la représentation des composantes principales associées aux données $(X_1^{'}, X_2^{'})$. Nous pouvons ensuite projeter les données sur la première dimension qui est beaucoup plus représentative que la seconde. Cela permet ensuite de ne conserver qu'un seul descripteur mais qui porte une information pertinente, $X_1^{'}$ pour l'analyse de donnée comme présenté sur la partie droite de la figure 1.21.

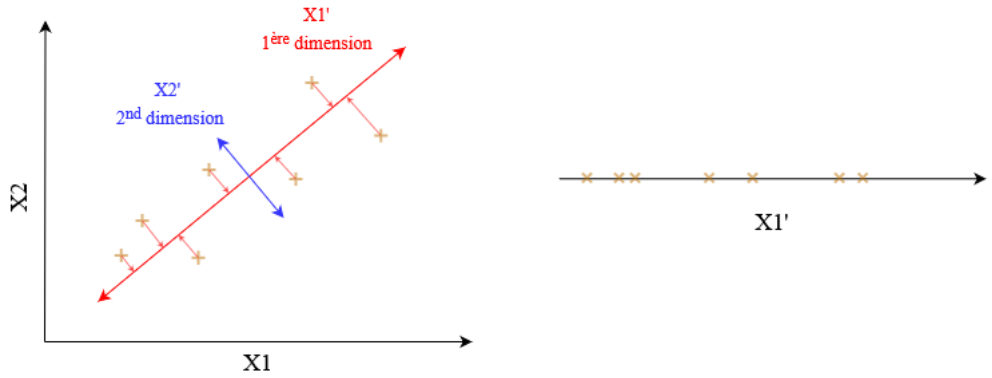


FIGURE 1.21 – Réduction de la dimension 2 à la dimension 1 pour un ensemble de points.

Les application des méthodes non-supervisées sont nombreuses. Nous retrouvons entre autres la fouille de données (*Data Mining* en anglais), l'analyse de documents et de profils de clients dans le cadre de l'informatique décisionnelle (*business intelligence* en anglais) ou même la création de descripteurs pour des méthodes supervisées [56], [57].

### 1.3.2 Les méthodes supervisées

Cette approche consiste à apprendre une fonction de prédiction à partir d'exemples annotés. Ainsi, la machine apprend à réaliser une tâche spécifique en étudiant des exemples de cette tâche. Par exemple, elle peut apprendre à classer des photos de chiens et de chats après que nous lui ayons montré des millions de photos de chiens et chats. Ou encore, elle peut apprendre à traduire le français en chinois après avoir vu des millions d'exemples de traduction français-chinois. D'une manière générale, la machine peut apprendre une relation en ayant analysé des millions d'exemples

d'associations [57]. Ce type de méthode a été mis en place dans la section 2.4 et le chapitre 3.

La figure 1.22 décrit une version très simpliste des méthodes supervisées. Toutefois, cette vision présente un message clé : le modèle prédictif doit visualiser un ensemble de données assez représentatives pour que le modèle généré puisse se généraliser. L'auteur fait probablement également référence au concept *garbage in - garbage out* selon lequel des données d'entrée défectueuses ou mauvaises produisent des sorties de mauvaise qualité qui ne répondent pas aux attentes de l'utilisateur.

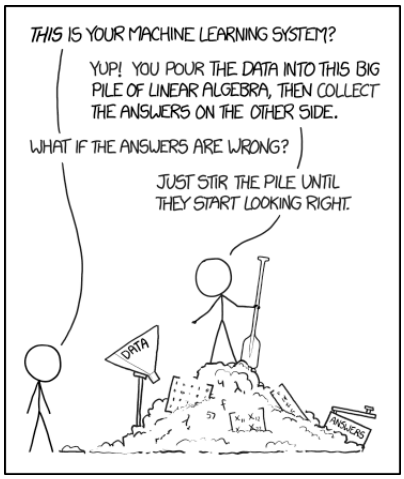


FIGURE 1.22 – Le machine learning vu par l'artiste Randall Munroe, auteur d'xkcd. Sous licence CC BY-NC 2.5.

L'apprentissage supervisé utilise une fonction de coût afin de mesurer la performance du modèle et donc l'erreur entre la prédiction et son observation réelle. En ML, il existe beaucoup de métriques pour mesurer ces erreurs [56], [57]. L'objectif de l'apprentissage est donc de minimiser globalement cette erreur sur l'ensemble du jeu de données utilisé pour l'apprentissage. L'apprentissage supervisé s'attaque à deux problématiques importantes de la science de donnée : les problèmes de régression et les problèmes de classification.

### La régression

Dans l'apprentissage supervisé, le but de la régression est d'estimer une valeur (quantitative) de sortie à partir des valeurs d'un ensemble de caractéristiques en entrée, par exemple déterminer le prix d'une maison en se fondant sur sa surface, le nombre d'étages, son emplacement, etc. Ainsi, l'ensemble des solutions $y$ est infini.

Parmi les algorithmes de régression, nous retrouvons entre autres la régression linéaire [57] ou les arbres de décisions [63]. La figure 1.23 représente une régression pour un descripteur $X$ et une observation $y$, les croix oranges étant la donnée d'apprentissage et la courbe rouge le modèle pour le descripteur $X$. L'étoile rouge représente la prédiction pour une nouvelle donnée dont la valeur est représentée par la croix rouge pour le descripteur $X$.

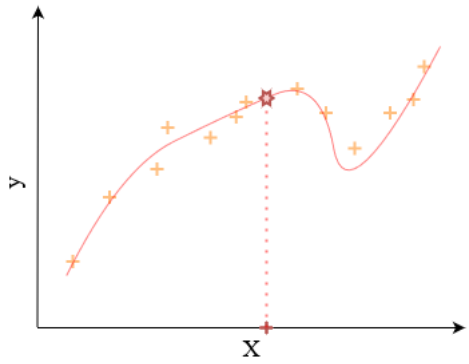


FIGURE 1.23 – Régression sur un jeu de données et prédiction d'une nouvelle valeur.

### La classification

Les méthodes de classification séparent un ensemble d'échantillons d'entrée en différentes classes (qualitative) de sortie, par exemple l'attribution d'un courrier électronique donné à la classe "spam" ou "non spam". Ainsi, l'ensemble des solutions $y$ est fini. Parmi les algorithmes de classification, nous retrouvons notamment les RF [64] ou les machines à vecteurs de support (*support-vector machines* en anglais) [65]. La figure 1.24 représente une classification pour les descripteurs $X_1$ et $X_2$ : les étoiles vertes et les cercles bleus symbolisent les données d'apprentissages, et la courbe orange la frontière de séparation entre les deux classes (résultat de la modélisation). Toute nouvelle donnée qui tomberait dans l'espace vert ou bleu serait directement classifiée en fonction de son espace d'appartenance. Ainsi, l'étoile rouge qui représente une nouvelle donnée pour les descripteurs $X_1$ et $X_2$ sera étiquetée dans la classe bleue.

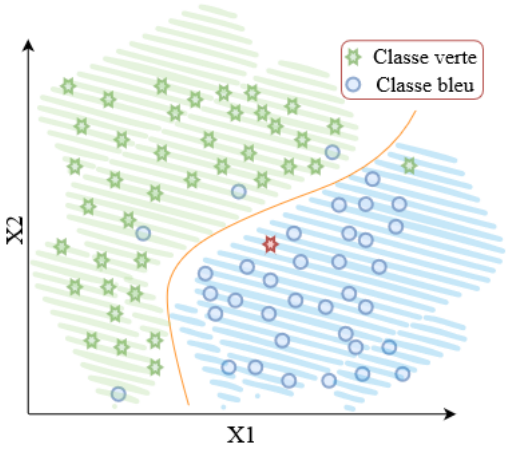


FIGURE 1.24 – Classification binaire sur un jeu de données comportant deux descripteurs $X_1$ et $X_2$.

### 1.3.3 Compromis biais-variance

Pour les méthodes supervisées, il est important d'optimiser les paramètres utilisés dans le modèle de prédiction pour éviter les problèmes de surapprentissage et de sous-apprentissage [66] et d'adapter le compromis biais-variance.

Le surapprentissage (*overfitting* en anglais) et le sous-apprentissage sont des concepts utilisés dans l'apprentissage supervisé pour décrire comment l'algorithme d'apprentissage automatique s'adapte aux données d'apprentissage.

En statistique, le surapprentissage signifie qu'une analyse statistique ou un modèle est trop proche de ses données d'apprentissage, par conséquent, il est possible que le modèle prédise les nouvelles données avec une erreur importante. Le surapprentissage empêche le modèle de prédire correctement les données provenant de l'ensemble de développement et de l'ensemble de test. Le modèle est trop proche des données de l'ensemble d'apprentissage, même si des valeurs aberrantes peuvent s'y trouver. La figure 1.25 (à gauche) représente ainsi le phénomène de surapprentissage dans un problème de classification. Nous voyons que le modèle essaye de séparer au maximum les deux classes mais se complexifie par la même occasion. De même, dans le cas d'une régression (voir 1.25 à droite), le modèle essaye de passer par tous les points de données afin d'avoir la justesse (au sens statistique) la plus importante.

Des termes statistiques spécifiques peuvent être employés pour analyser plus

en détail le comportement de l'algorithme : si le modèle est en sous-apprentissage, on dit que l'algorithme présente un biais élevé et une faible variance ; s'il est en surapprentissage, il présente une variance élevée et un biais faible. Au cours de chaque tâche d'apprentissage automatique, un compromis doit être trouvé entre biais et variance. Dans l'apprentissage automatique traditionnel, le modèle final ne peut pas présenter simultanément un très faible biais et une très faible variance. L'objectif est de trouver un équilibre optimal du biais et de la variance n'entraînant pas de sous ou surapprentissage.

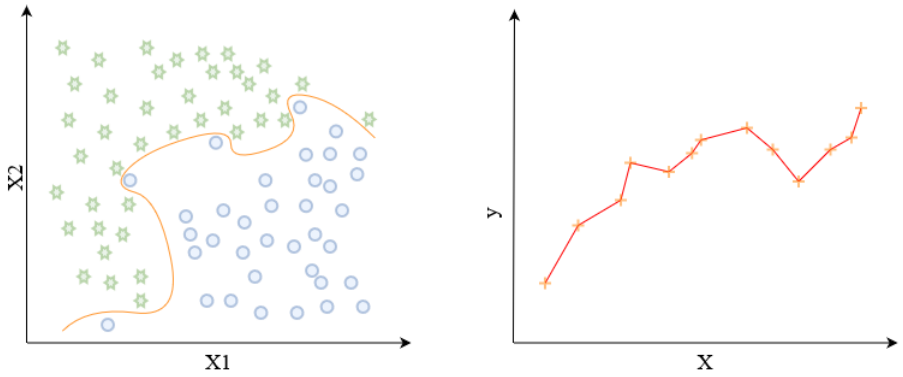


FIGURE 1.25 – Schéma représentant le phénomène de surapprentissage, à gauche dans le cas d'une classification et à droite dans le cas d'une régression.

Un modèle peut également être en sous-apprentissage s'il est trop simple pour la tâche d'apprentissage. Le recours à un algorithme avec une modélisation plus complexe (avec plus de descripteurs par exemple) est alors nécessaire, incluant plus de non-linéarités. Le sous-apprentissage signifie que le modèle ne s'adapte pas suffisamment aux données d'apprentissage et n'apprend pas correctement la tâche. Il n'est capable de récupérer ni étiquettes correctes, ni données d'apprentissage, ni données de validation/test. La figure 1.26 (à gauche) représente ainsi le phénomène de sous-apprentissage dans un problème de classification. Nous constatons que le modèle est très simple et rencontre donc beaucoup de difficultés à réaliser la tâche de classification. Pour le cas de la régression (à droite sur la figure 1.26), nous remarquons que la courbe de prédiction n'arrive pas à modéliser parfaitement les données d'entrées et que, pour certains points, l'erreur de prédiction peut être importante.

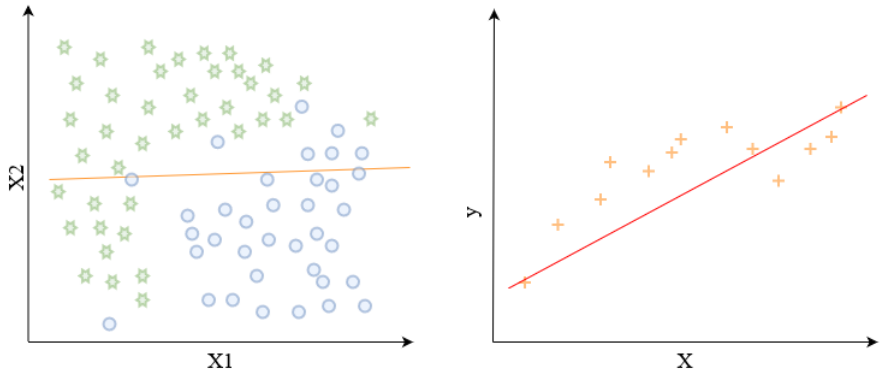


FIGURE 1.26 – Schéma représentant le phénomène de sous-apprentissage, à gauche dans le cas d'une classification et à droite dans le cas d'une régression.

## 1.4 Conclusion

Afin d'assurer une bonne compréhension dans la suite du manuscrit des termes et des méthodes employées, ce chapitre a présenté les éléments clés suivants :

- le système d'acquisition de levés hydro-océanographiques (en tant que système de systèmes) et les capteurs utilisés dans le cadre de ces levés et tout particulièrement les capteurs bathymétriques qui acquièrent des sondes que nous allons traiter dans le chapitre 2 et 3 ;
- la manière dont se propagent les ondes acoustiques et OEM dans l'environnement sous-marin ;
- la science des données avec les méthodes non-supervisées et supervisées.

Chapitre 2

# L'échelle micro : la sonde comme donnée

**Synthèse du chapitre**

Ce chapitre s'intéresse de plus près à la donnée et plus spécifiquement la donnée bathymétrique. Nous nous plaçons donc au niveau d'échelle le plus bas, voir la figure 6 de l'introduction, afin de comprendre quelles sont les caractéristiques de cette donnée et quelles techniques employer pour la traiter et la manipuler efficacement. Ce chapitre présente deux contributions principales :

- Un travail de revue de littérature exhaustif dans le domaine du traitement de données bathymétriques, qui présente une taxonomie originale des méthodes de détection de sondes aberrantes pour les données SMF et LIDAR.
- Une méthode de détection de données aberrantes dans un cadre supervisé, qui repose sur une combinaison d'un ensemble de descripteurs identifiés dans l'état de l'art. Des résultats sur un jeu de données réel SMF sont présentés à la fin de ce chapitre.

## 2.1 La sonde bathymétrique

Comme présenté dans la section 1.1.2, une mesure bathymétrique peut être issue de différents capteurs. Dans le cadre de ce manuscrit, nous allons nous intéresser tout particulièrement aux capteurs SMF et LiDAR qui sont aujourd'hui les deux capteurs les plus utilisés dans la communauté des levés hydrographiques pour cartographier les fonds marins et qui sont particulièrement complémentaires comme présentés sur la figure 1.5. Un levé bathymétrique (que ce soit SMF ou LiDAR) est constitué d'une collection de points qui sont les sondes (abus de langage accepté pour le LiDAR). Chaque sonde est définie par un triplet (x, y et z) où x et y représentent les coordonnées géographiques ou projetées sur un plan horizontal, et z la profondeur mesurée. Ce triplet est donc la position de la sonde dans un système de coordonnées géospatiales.

Dans le cas du SMF, comme présenté dans la section 1.1.2, les échos du fond sont détectés à l'intersection des faisceaux de transmission et de réception avec le fond. Dans un souci de simplification des termes, l'intersection des faisceaux d'émission et de réception est désignée sous le terme de "faisceau" (*beam* en anglais). Au cours d'une même impulsion acoustique, plusieurs faisceaux sont formés afin de constituer un *ping* (la zone orange perpendiculaire à l'avancée du bateau dans la figure 1.4). Cette émission synchronisée crée une proximité temporelle entre les faisceaux réceptionnés au même *timestamp*. Dans le cas du LiDAR, et grâce à la beaucoup plus grande célérité de l'OEM, les faisceaux lasers sont envoyés les uns à la suite des autres de façon décorrélée. Néanmoins, plusieurs retours dans la courbe d'intensité sont conservés (données exportées en retours multiples) pour le LiDAR, si le retour le plus important était le seul conservé (comme ce qui est réalisé pour le SMF), alors nous risquerions de ne garder que les retours associés à la surface de l'eau, comme présenté dans la figure 1.15. La liste des attributs (position) de chaque sonde peut être complétée par l'intégration d'informations caractérisant la mesure acoustique, par exemple, le facteur de qualité de la donnée SMF (*QF - Quality Factor*), défini dans [67]. De plus, pendant que le système acquiert les mesures de profondeur, il fournit également une mesure de la puissance du signal retour (*backscatter* en anglais) pour le SMF et intensité (*intensity* en anglais) pour le LiDAR. La texture (ou rugosité du fond marin) et la composition géologique sont deux caractéristiques qui influencent fortement la puissance du signal retour (voir

la section 1.1.3). Dans le cas du SMF, il est également possible de pousser plus loin la description du contexte environnemental de la mesure en exploitant le stockage des données dans la colonne d'eau [54]. Toutefois, de par leur volume important et la réduction de la couverture (60° maximum), les données de la colonne d'eau ne sont pas systématiquement enregistrées.

De plus, comme indiqué dans la section 1.1, les systèmes d'acquisition bathymétrique ne sont qu'une composante de la combinaison des capteurs d'un système de levé hydrographique. Ainsi, à chaque valeur de sonde (position et profondeur) sont associées toutes les données annexes issues du signal de chaque capteur comme le roulis, le cap et le tangage provenant de l'IMU. Concernant ces données complémentaires, et durant la phase de post-traitement, chacune des séries temporelles par capteur est contrôlée et traitée individuellement si nécessaire (perte longue du signal GNSS entraînant une dégradation de la position) [8].

Enfin, durant l'acquisition en mer, les données bathymétriques sont acquises en suivant certaines trajectoires planifiées en fonction des caractéristiques du capteur, de la morphologie du fond marin (la profondeur locale mais également de toute la zone à lever) et des spécifications du projet [8] (l'instruction particulière pour le Shom définie dans la section 1.1.1). En effet, l'espacement des lignes est ajusté de façon à garantir la couverture de données bathymétriques (paramètre de qualité important d'un levé bathymétrique qui sera développé dans la section 4.2.2). À chaque sonde sont également associées des métadonnées provenant de la ligne acquise (nom de la ligne, *timestamp* du début de la ligne).

De par la complexité du processus associé à l'acquisition de la donnée bathymétrique, le format des fichiers des capteurs bathymétriques stocke toutes les mesures nécessaires au calcul de la position de chaque sonde depuis la donnée brute (durée du parcours aller/retour et angle de pointage du faisceau, position, attitude du porteur, intensité...). Grâce à cette information, il est possible :

1. d'intégrer dans le post-traitement n'importe quel type de mesure indisponible au moment du levé, exemple : marée observée rattachée au zéro hydrographique ;
2. de recalculer la position d'une sonde dans le cas d'un dysfonctionnement capteur non détecté durant l'acquisition ou la disponibilité de mesures auxiliaires de meilleures qualités, exemple : amélioration de la mesure de positionnement via post-traitement GNSS avec les éphémérides finales des

satellites GNSS ;

3. d'assister un hydrographe dans la prise de décision durant le processus de traitement des sondes mais également pendant la qualification du levé.

Finalement, chaque sonde bathymétrique est enrichie d'un ensemble de données supplémentaires permettant de mieux décrire celle-ci (en plus de sa position dans l'espace) et détaillant le processus d'acquisition ou la donnée des capteurs auxiliaires. La plupart des algorithmes de traitement de données bathymétriques exploitent seulement la cohérence spatiale et/ou temporelle de la donnée bathymétrique. Cependant certains exploitent ces attributs additionnels. Pour le SMF par exemple, Kammerer *et al.* [68] suggèrent d'utiliser l'information *backscatter* pour le traitement des sondes bathymétriques issues du SMF. Calder et Mayer, dans l'article [69], combinent la profondeur avec les incertitudes horizontales (récupérées de la Total Propagated Uncertainty (TPU) [70]) dans leur estimateur statistique bathymétrique. Ces mêmes valeurs d'incertitudes sont présentées dans la section 4.2.2. Dans un contexte tout autre, Ladner *et al.* [71] utilisent des informations *a priori* issues d'un Modèle Numérique de Bathymétrie (MNB) considéré comme la référence, aussi appelé vérité terrain (*ground truth* en anglais), pour nettoyer des jeux de données bathymétriques anciens. Dans le cas du LiDAR, Lowel et Calder [72] exploitent des descripteurs présents dans les données brutes (tel que le nombre et le numéro de retour du signal, l'incertitude de mesure de position ou d'attitude, l'angle d'incidence, etc.) afin de réaliser leur modèle d'apprentissage supervisé et ce afin de classer les sondes comme données bathymétriques ou non.

## 2.2 Les erreurs dans la donnée bathymétrique

Les données bathymétriques s'accompagnent d'erreurs aléatoires, systématiques et ponctuelles [39]. La figure 2.1 présente ces trois types d'erreur pour une coupe bathymétrique issue d'un SMF avec notamment la présence d'un biais pour l'angle de montage en roulis entre l'IMU et le SMF (erreur systématique).

Les erreurs aléatoires sont intrinsèques au processus de mesure. Les limites acceptables d'incertitude de la mesure sont estimées et généralement évaluées selon la publication spéciale n°44 de l'IHO correspondant à la norme sur les levés hydrographiques [8]. Pour de nombreux capteurs, nous parlons de sensibilité de la mesure [73] et nous assimilons ces erreurs au bruit de mesure, soit l'ensemble

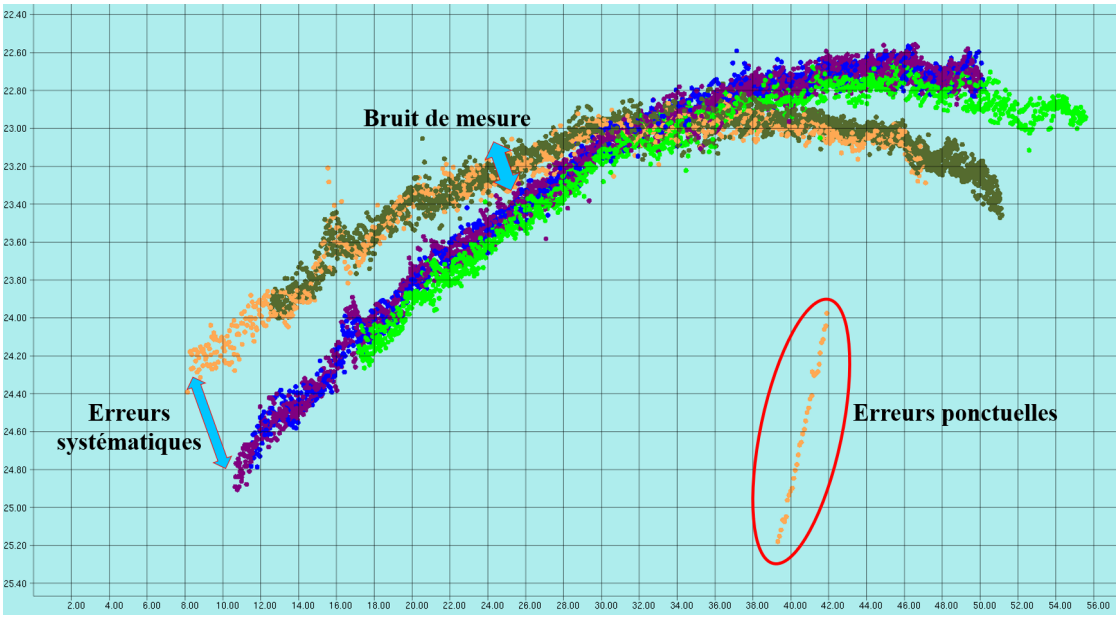


FIGURE 2.1 – Les trois types d'erreurs représentées sur une coupe bathymétrique. Les sondes sont colorées en fonction des lignes de levés. Les données bathymétriques ont été acquises lors d'une campagne d'ajustage en rade de Brest à l'aide d'un SMF Kongsberg/SIMRAD EM2040C.

des signaux parasites qui se superposent au signal que nous cherchons à obtenir au moyen d'une mesure d'un phénomène physique. Dans le cas des mesures bathymétriques, nous employons le terme de "tapis de sondes" pour nommer cette erreur qui est caractérisée par son épaisseur (une valeur en centimètre correspond à l'écart-type à $2\sigma$).

Les erreurs systématiques peuvent provenir de :

— l'oubli ou le mauvais réglage d'un paramètre dans le système de levé bathymétrique ;

— le mauvais étalonnage d'un capteur entraînant un biais systématique sur la donnée acquise ;

— l'application d'une procédure d'acquisition erronée.

La procédure d'ajustage, connue sous le nom de *patch test* [50], [70], qui précède chaque levé, ainsi que les bonnes pratiques de levés permettent de minimiser les effets des erreurs systématiques [8]. Le *patch test* classique découple l'étude des trois biais angulaires (*boresight angle* en anglais) : il commence par l'angle de roulis, suivi de l'angle de tangage, puis de l'angle de cap. Pour l'angle de roulis, un fond plat (permettant de minimiser l'effet d'une erreur sur les angles de tangage

et de cap) est sondé dans des directions opposées, ce qui permet de caractériser facilement l'effet de roulis comme montré sur la figure 2.2.

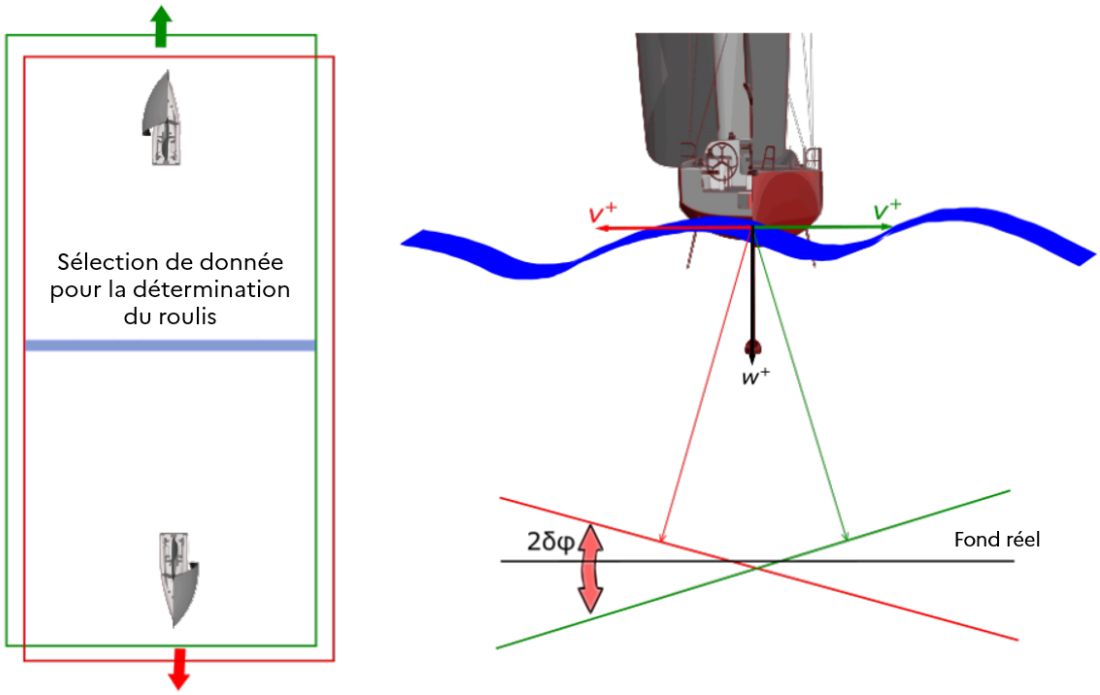


FIGURE 2.2 – Exemple d'acquisition du *patch test*. Gauche : vue de dessus du motif de lignes acquises pour mesurer le biais en roulis. Droite : effet du roulis sur deux lignes opposées sur un fond marin plat.

D'autres mesures métrologiques, comme la mesure des bras de leviers (définis dans la section 1.1.3), permettent également de réduire la présence d'erreurs systématiques qui peuvent être parfois très complexes à supprimer en post-traitement à cause de leur couplage [74].

Dans le cadre de ce manuscrit, nous nous sommes intéressés uniquement à la détection des erreurs ponctuelles afin de simplifier le travail des opérateurs. Bien que nous n'avons pas négligé l'effet des erreurs systématiques et du bruit aléatoire, nous avons émis également les deux hypothèses suivantes (qui se vérifient dans la très grande majorité des levés acquis par les SH) : les données ne comportent pas d'erreur systématique et le bruit de mesure est faible au regard des données aberrantes à détecter. Les procédures annuelles d'ajustage permettent aux SH de conforter ces deux hypothèses.

Que signifie le terme aberrant ? Intuitivement, nous pourrions qualifier une valeur aberrante de valeur non ordinaire, de mesure inhabituelle, d'une exception ou

comme le définit Hawkins : "une observation qui dévie tant d'une autre observation que l'on soupçonne qu'elle ait été générée par un mécanisme différent" [75]. Ainsi, la détection d'une donnée aberrante consiste à repérer dans un jeu de données les points dont les caractéristiques diffèrent de celles attendues.

La vaste littérature dédiée aux techniques de détection de données aberrantes offre un large panel de définitions pour le terme "donnée aberrante". La plupart de ces définitions sont néanmoins peu précises et ne fournissent qu'une liste des types de données aberrantes. Cependant, deux définitions formelles émergent. De l'avis de Davies et Gather [76], chaque point d'un jeu de données provient d'une unique distribution statistique. Les points qui appartiennent donc à la queue de la distribution sont considérés comme des données aberrantes. Une approche plus conventionnelle consiste à émettre le postulat que deux lois de distribution cohabitent : les bons échantillons, issus d'une des distributions, sont contaminés par les données aberrantes provenant de l'autre distribution. Cette introduction aux définitions statistiques met en avant une difficulté à définir ce terme dans un contexte générique.

En effet, la plupart du temps, cette définition est adaptée à la méthode utilisée pour la détection de donnée aberrante. Il existe donc autant de définitions que de façons d'aborder ce problème. Toutefois, de nombreuses publications montrent également que certains types de donnée aberrante sont plus communs que d'autres. Il est clair que les performances des algorithmes de détection dépendent du type de donnée aberrante à traiter. Certaines stratégies sont donc plus adéquates que d'autres dans la détection d'un type particulier de données aberrantes mais échouent à en découvrir d'autres types. De ce fait, la conception d'un algorithme de détection implique de connaître *a priori* les principaux types de donnée aberrante à rechercher.

En général, les données aberrantes sont divisées en deux groupes : celui des erreurs grossières ou données aberrantes lointaines (*global outlier* en anglais) et celui des erreurs du modèle ou erreurs locales (*contextual* ou *local outliers* en anglais). Cette distinction peut également être appliquée dans le contexte des données bathymétriques. D'un côté, même si les données aberrantes lointaines sont présentes dans les jeux de données bathymétriques, elles sont le plus souvent filtrées durant le processus d'acquisition (directement filtrées par le capteur). Mais leur présence, même en nombre limité, peut mettre en difficulté certaines des techniques de détec-

tion de donnée aberrante notamment lorsqu'elles sont combinées avec des erreurs locales. Néanmoins, lorsque les erreurs grossières sont détectées, elles peuvent être supprimées sans risque car elles ne sont pas cohérentes avec le fond marin et ne nécessitent donc pas une modélisation complexe pour les détecter. Elles sont principalement définies en utilisant des connaissances *a priori* exprimées sous forme de frontières maximales de l'espace acquis (comme la profondeur *a priori* ou la restriction d'extension géographique). D'un autre côté, les données aberrantes locales nécessitent une description beaucoup plus précise du fond marin pour s'assurer de leur détection. Ainsi, ce type de données aberrantes peut être intégré avec les *model failure*, mentionnés par Hampel *et al.* [77], ou avec les données aberrantes spatiales ou de relation comme mentionné par Planchon [78].

La distinction faite entre ces deux types d'erreurs ponctuelles dissocie artificiellement les données mauvaises dans un jeu de données. En réalité, des interactions existent entre ces deux types et conduisent à des effets indésirables comme le masquage ou des effets de contamination (*swamping effects* en anglais) lorsque des données aberrantes sont détectées. De façon intuitive, les effets de masquage apparaissent lorsqu'une donnée aberrante n'est pas détectée du fait de la présence d'une autre. Les effets d'interactions apparaissent quand une sonde valide est étiquetée comme une donnée aberrante de par la présence voisine de donnée aberrante. La détection des erreurs aberrantes dans les données bathymétriques est surtout concernée par la détection des erreurs locales.

## 2.3 L'état de l'art dans le traitement de la donnée bathymétrique

### 2.3.1 Les données aberrantes dans un contexte hydrographique

L'objectif de cette section est de proposer une vision d'ensemble structurée et claire des techniques de détection automatique (détection et suppression par l'algorithme) ou semi-automatique (détection par l'algorithme et suppression/validation par l'opérateur) des données aberrantes qui ont été développées et appliquées dans le cadre du traitement de la donnée bathymétrique SMF ou LiDAR. Cet exercice a fait l'objet d'un article de revue [12] déjà cité à plusieurs reprises et nous n'avons

pas trouvé dans la littérature d'autres articles réalisant un tel exercice de revue concernant le traitement des données bathymétriques. Il est donc repris et détaillé ici.

La détection des données aberrantes n'est pas un besoin nouveau. En effet, les statisticiens se sont penchés sur la question dès la fin du 19$^{e}$ siècle. Cet intérêt s'est ensuite développé au fil du temps, donnant naissance à une branche spécifique de la statistique. Aujourd'hui, la détection de valeurs aberrantes est un problème de recherche clairement identifié qui possède des implications dans une variété d'applications (par exemple la détection de fraude, le diagnostic médical ...). La profusion de manuscrits scientifiques consacrés à la détection de valeurs aberrantes, comme le livre d'Aggarwal [79] ou le chapitre du livre de Kotu [80], témoigne clairement de l'importance mais également de la complexité de cette tâche.

La détection de données aberrantes dans les jeux de données bathymétriques est une tâche cruciale. En effet, l'objectif final du traitement est de représenter l'information bathymétrique sur une carte nautique possédant un statut légal en cas de fortune de mer. Traiter la donnée bathymétrique dans ce contexte de sécurité de la navigation a conduit les SH à mettre en place des procédures manuelles de traitement et de contrôle de la donnée bathymétrique. Ce travail nécessite l'emploi de logiciels de visualisation dédiés à l'inspection détaillée de toutes les sondes. Cette tache est fastidieuse, chronophage et ne garantit pas que les obstructions seront toutes traitées correctement (subjectivité potentielle de ce type de traitement surtout après une journée complète de traitement).

Les procédures de nettoyage des données aberrantes des données bathymétriques ont été progressivement automatisées, depuis les premiers filtres implémentés par la National Oceanic and Atmospheric Administration (NOAA) ou le Shom [81], [82] jusqu'à l'algorithme Combined Uncertainty and Bathymetry Estimator (CUBE) [83] exploité encore aujourd'hui par plusieurs services hydrographiques dont le Shom. Toutefois, la principale difficulté réside dans le fait qu'il existe plusieurs explications physiques à l'apparition de ces données aberrantes, donc autant de réponses analytiques associées différentes, possiblement à modéliser.

Des fonds marins pouvant servir de vérité terrain (*ground truth* en anglais) ou bien de l'information bathymétrique indépendante et redondante sur une même zone, sont rarement voire jamais disponibles. Dans le peu de zone où la donnée bathymétrique est disponible, le processus de détection des données aberrantes

pourrait se fonder sur de précédents levés, à condition de prendre en compte les potentielles évolutions du fond marin (cas des matériaux mous tels que la migration de dunes, ou obstacles artificiels tel que les épaves) et l'évolution des caractéristiques des sondeurs utilisés. Aujourd'hui encore, 80% de la surface des océans n'a pas été cartographiée avec du matériel moderne et de haute résolution [10], [84]. Il est donc souvent impossible d'utiliser des techniques de détection de données aberrantes basées sur une comparaison avec la vérité terrain car notre connaissance des fonds marins demeure non seulement limitée mais également hétérogène.

Dans des environnements aussi complexes et critiques en termes de sécurité, définir une limite entre l'occurrence la plus probable du fond marin et les valeurs aberrantes devient particulièrement difficile. La principale difficulté est de distinguer les différents types de valeurs aberrantes dans des ensembles de données bruitées et donc de savoir précisément comment le processus de mesure a été effectué (en tenant compte des métadonnées et du rapport de levé). Cela n'est pas toujours possible, notamment lorsque nous considérons des données bathymétriques provenant de sources multiples (comme le *crowdsource bathymetry*) et/ou des jeux de données peu ou mal documentés. Les techniques de détection de la donnée aberrante sont nécessaires pour traiter une donnée bathymétrique de qualité variable, la plupart du temps, sans métadonnée associée.

Pour répondre à l'objectif de produire un état de l'art portant sur le traitement de donnée bathymétrique, nous avons adopté une approche taxonomique visant à catégoriser les techniques de détection des valeurs aberrantes, ce qui est courant dans les études de détection des valeurs aberrantes. Dans son approche générique, Chandola [85] identifie quatre caractéristiques majeures des techniques de détection de valeurs aberrantes permettant de réaliser une telle classification, qui sont établis sur :

— la donnée d'entrée ;
— le type de donnée aberrante ;
— le type de supervision ;
— la donnée de sortie pour la détection.

Même avec les récents progrès technologiques des capteurs bathymétriques, la gestion des erreurs dans un environnement aussi complexe reste un défi. Plusieurs facteurs contribuent à détériorer les mesures acoustiques ou même à générer des valeurs aberrantes. Il est courant de regrouper ces sources d'aberrations potentielles

en deux classes : 1) celles spécifiques à la plateforme et à son mouvement (bruit propre) ; 2) celles induites par l'environnement (bruit ambiant acoustique). Voici une liste non exhaustive de ces facteurs :

- le dysfonctionnement des capteurs ;
- la présence de bulles à la tête des transducteurs ;
- des multiples chemins de réflexion acoustique ;
- des interfaces acoustiques fortes dans la colonne d'eau ;
- des effets de lobes secondaires (liés à la géométrie de l'antenne d'émission) ;
- des mauvaises conditions météorologiques (faible rapport signal/bruit) ;
- des objets dans la colonne d'eau (comme des poissons, algues, panache hydrothermal) ;
- d'autres équipements créant des interférences...

Ainsi, comparé à un levé terrestre (technologie LiDAR avec laser PIR uniquement), le taux de donnée aberrante est plus élevé dans un levé bathymétrique en raison de la complexité du processus d'acquisition des données. Dans le contexte bathymétrique et pour la donnée issue de SMF, le pourcentage de données aberrantes est considéré comme inférieur à 1% par certains auteurs [86], [87]. Pour d'autres, il est inférieur à 10% [88] mais peut atteindre jusqu'à 25% pour certains tests [89]. Pour le capteur LiDAR bathymétrique, le pourcentage de données ne correspondant pas au fond est extrêmement variable et peut-être compris entre 60% et 95% (très dépendant du type de capteurs utilisé). Dans le cas de ce capteur, il y a un déséquilibre souvent fort entre la donnée de fond et les autres signaux (comme la surface d'eau). La grande variabilité du pourcentage de résultats aberrants observée dans la littérature peut s'expliquer par le large panel de capteurs utilisés ainsi que par la grande variabilité des conditions environnementales. Elle peut également s'expliquer par le contexte de traitement qui peut favoriser la conservation des sondes erronées. Dans le contexte de la cartographie nautique par exemple, en cas de sondes s'écartant de la représentation probable du fond marin, la sécurité de la navigation privilégiera la conservation des sondes, même anormales, les moins profondes dans le jeu de données. Enfin, si pour un jeu de données donné, le pourcentage de données aberrantes reste inconnu faute de vérité terrain, il est considéré par de nombreux auteurs comme faible [87], [90], [91]. Il est encore plus faible aujourd'hui avec les performances des nouveaux capteurs bathymétriques et plus particulièrement avec les améliorations liées aux algorithmes de détection de

fond, qui invalident des sondes en temps réel (mais qui ne repose que sur l'étude du signal brut).

Les valeurs aberrantes sont ainsi identifiées comme étant distinctes du fond marin. L'hypothèse de travail la plus récurrente dans la littérature concerne la topographie du fond marin, qui est censée varier de façon faible et continue par rapport aux résolutions horizontale et verticale des sondeurs multifaisceaux [87], [92]-[94]. Les sondes qui s'écartent trop de cette hypothèse sont alors identifiées comme aberrantes. Cependant, dans le domaine de l'hydrographie, une telle définition peut être ambiguë notamment dans le cadre de la caractérisation d'obstructions ou de haut-fonds qui, contrairement aux données aberrantes, devront être conservées dans le jeu de données afin d'être identifiées comme des risques locaux pour la sécurité de la navigation. Comme l'indique Hou [93], ces obstructions diffèrent des valeurs aberrantes lorsque des sondes voisines provenant de différents *pings* touchent la cible plusieurs fois. Contrairement aux objets structurés tels que les pipelines reposant sur un fond marin plat [94], il n'est pas si évident de distinguer une caractéristique réelle d'un groupe de valeurs aberrantes. La figure 2.3 présente quatre profils bathymétriques classés par complexité croissante de post-traitement dans le contexte de la sécurité de la navigation, et trois classes de valeurs aberrantes les plus fréquemment identifiées :

— les valeurs aberrantes isolées (figure 2.3b) ;
— les groupes structurés de valeurs aberrantes (Figures 2.3a, 2.3c et 2.3d) ;
— les groupes non structurés de valeurs aberrantes (figure 2.3d) ;

Certains auteurs mentionnent et proposent de ne détecter qu'un seul type d'erreur, à savoir les valeurs aberrantes isolées pour Arnold [95], les pics pour Eeg [96] Ferreira [97] et Bjorke [98] et le bruit impulsif pour Mann [99]. Calder, Du, Li, Lu et Motao [69], [89], [100]-[102] opposent à ce premier type de données aberrantes un second type constitué de groupes ou de grappes de données aberrantes, soulignant la difficulté de les traiter. Les *pings* isolés ou les faisceaux défectueux entrent dans cette catégorie. Hou [93] invalide les sondes appartenant à un *ping* erroné en utilisant un filtre spécifique, dénommé *bag ping* dans la section 2.4. Bottelier [103] et Arge [104] pointent un autre type de bruit structurel qui apparaît le long des trajectoires de levés. Ces rubans de points n'ont qu'une épaisseur de sonde dans le cas le plus simple mais peuvent être plus complexes à traiter lorsqu'ils sont denses, car ils deviennent localement indiscernables des pipelines [104]. Des valeurs

aberrantes consécutives peuvent produire des géométries facilement identifiables, en particulier pour un expert qui est capable de choisir la représentation la plus appropriée des données (c'est-à-dire le mode de visualisation par fauchée (dans le repère ping/beam) ou nuage de point géographique) pour les reconnaître (figure 2.3 c). D'une manière plus générale, les groupes non structurés de valeurs aberrantes peuvent nécessiter des investigations manuelles supplémentaires, principalement établis sur l'analyse des données de la colonne d'eau. Le problème de la rupture du gradient morphologique décrit par Calder [105] relève de ce type de valeurs aberrantes. Des pentes significatives apparaissent à travers les fauchées de sonar génèrant des faisceaux inconsistants qui sont cohérents spatialement et plus courts que le fond marin. De telles configurations de valeurs aberrantes sont particulièrement problématiques pour les experts et pour les procédures automatiques qui conduisent à un traitement asymétrique en fonction du contexte hydrographique, en traitant différemment les sondes détectées comme erronées selon qu'elles soient situées au-dessus ou en-dessous du fond marin environnant [87], [100], [102].

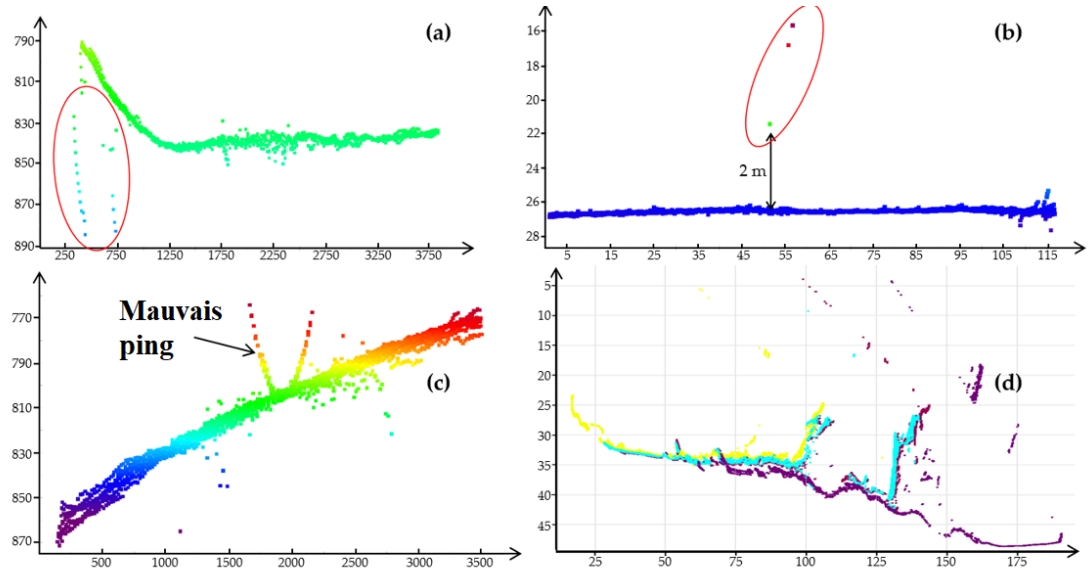


FIGURE 2.3 – Quatre exemples de données aberrantes présentées du cas le plus simple au cas le plus complexe dans le contexte de sécurité de la navigation. Les sondes sont colorés en fonction de leur profondeur pour les cas a) à c) et par ligne de levé pour le cas d).

### 2.3.2 Taxonomie des techniques de détection de données aberrantes

Le premier niveau de classification que nous retrouvons habituellement dans les méthodes de détection des valeurs aberrantes, comme l'article de Chandola [11], est le type de supervision. La littérature scientifique les regroupe dans les trois familles suivantes, suivant la typologie classique des algorithmes d'apprentissage automatique, voir la section 1.3 :

— supervisés : algorithmes générant une fonction prédictive pour un ensemble de données à partir de données préalablement étiquetées (en relation avec le problème à résoudre) ;
— non-supervisés : algorithmes traitant des données non étiquetées, en d'autres termes sans aucune connaissance au préalable de la donnée ;
— semi-supervisés : algorithmes fusionnant ces deux approches en déterminant une fonction de prédiction pour l'apprentissage avec une petite quantité de données étiquetées et une grande quantité de données non étiquetées.

Dans le domaine hydrographique, la majorité des algorithmes de détection des aberrations sont établis sur des méthodes non supervisées. Aucune méthode semi-supervisée n'a été identifiée et seules quatre méthodes supervisés fondées sur des méthodes ML de complexités variables existent pour l'instant. Nous remarquons également que les méthodes les plus récentes rentrent dans cette dernière catégorie [54], [72], [106], [107].

La figure 2.4 présente notre taxonomie des méthodes de détections des sondes aberrantes dans le contexte hydrographique. Nous pouvons constater très rapidement la prépondérance des algorithmes de traitement de données bathymétriques SMF (en bleu sur la figure) par rapport aux méthodes de traitement des données LiDAR (en vert sur la figure). En effet, les capteurs acoustiques ont été utilisés beaucoup plus tôt que les capteurs pour l'acquisition de donnée bathymétrique LiDAR et, de plus, la densité des sondes sur les capteurs LiDAR et le niveau de bruit important complexifient grandement l'utilisation de méthode automatique pour la détection des données aberrantes. Avec l'amélioration de la sensibilité des capteurs et l'augmentation du nombre de sondes acquises par les capteurs LiDAR bathymétrique ces dernières années, nous constatons l'arrivée d'algorithmes pour le traitement de données bathymétriques LiDAR que ce soit avec des techniques

supervisées [72], [107] ou non-supervisées [35]. Enfin, des méthodes automatiques pour le traitement du LiDAR terrestre sont beaucoup plus nombreuses, comme le montre cet article [108], mais inapplicables pour le LiDAR bathymétrique, les données aberrantes étant complètement différentes et le déséquilibre entre les données valides et aberrantes inversé, voir détails dans la section 3.3.1.

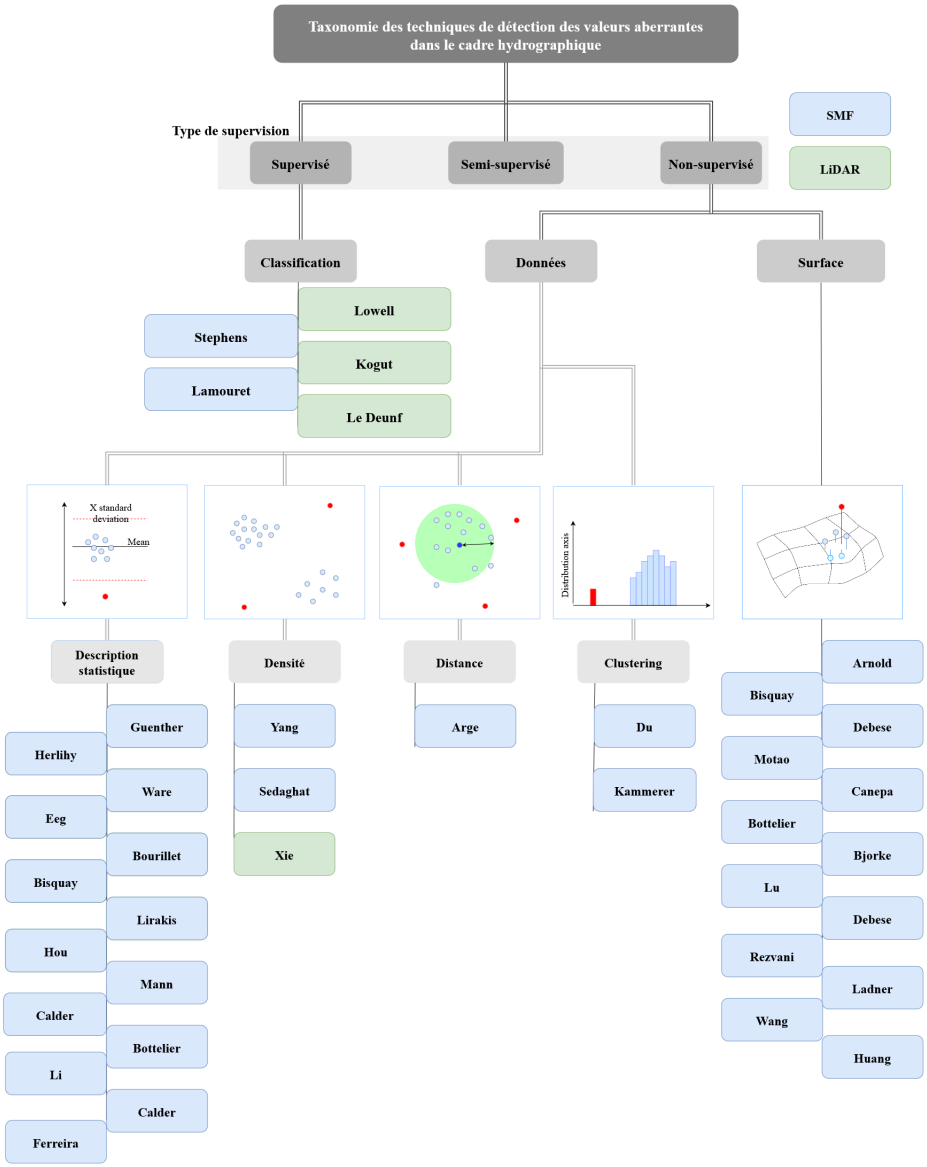


FIGURE 2.4 – Taxonomie exhaustive des méthodes de détections des sondes aberrantes pour le SMF (en bleu) et LiDAR (en vert), issu de [12] et mis à jour à l'été 2022.

### Les algorithmes non-supervisés

Les techniques non-supervisées supposent seulement que les erreurs sont séparées des données "normales" et apparaissent comme des valeurs aberrantes [109]. Les techniques appartenant à cette classe peuvent être regroupées en deux sous-classes : 1) les approches orientées données ou diagnostic qui pointent les aberrations en travaillant directement sur les données, et 2) les approches orientées surface qui construisent de manière robuste un modèle du fond marin à partir de tous les points avant de détecter les aberrations. La figure 2.4 montre la répartition des algorithmes existants dans cette sous-classification.

### Les approches orientées données

Au travers des paradigmes classiques de la science des données [79], les techniques de détection axées sur les données sont globalement classées dans les quatre familles suivantes :

— les approches fondées sur les statistiques généralement divisées en deux classes selon que la distribution des données est supposée connue ou non. Les approches paramétriques sont utilisées pour estimer les paramètres de la distribution supposée, la plupart du temps la moyenne et l'écart-type d'une distribution gaussienne, tandis que les approches non paramétriques estiment la densité de probabilité à partir des données, sans aucune hypothèse sur la forme de la distribution. Dans les deux cas, les valeurs aberrantes sont identifiées comme des points appartenant aux extrémités de la queue de distribution ;

— les approches établis sur la distance, également appelées techniques du plus proche voisin, reposent sur la corrélation spatiale en calculant la distance entre un point donné et son voisinage. Les points ayant une distance plus élevée que les autres points "normaux" sont identifiés comme des points aberrants ;

— les approches construites sur la densité, assez proches des approches établies sur la distance, car la densité de points par unité de surface/volume est inversement liée à la distance du voisinage du point. Les points aberrants sont localisés dans les zones de faible densité tandis que les points normaux sont agrégés ;

— les approches utilisant le clustering, approches classiques de l'apprentissage automatique. Ces approches globales consistent à regrouper des données similaires dans des groupes appelés clusters. Comme les points aberrants sont rares, ils sont soit laissés isolés, soit regroupés dans un cluster très éloigné des autres.

Comme le montre la figure 2.4, les approches statistiques sont de loin les plus nombreuses et aussi les plus anciennes. La première approche, mieux connue sous le nom de programme COP (*Combined Offline Processing*) a été proposée par Guenther en 1982 pour améliorer la sélection des sondes acquis par le système de levé (BS3) [81]. Quatre approches sont fondées sur la mise en cascade de filtres statistiques simples (à l'aide de calculs [110]-[113]). Le nombre de filtres en cascade varie de 3 à 5 selon l'approche. La plupart de ces filtres fonctionnent en mode fauchée sans qu'il soit nécessaire de définir la taille des voisinages, qui sont directement fixés par l'approche. Le seul paramètre à définir par l'opérateur est, si nécessaire, la taille du groupe de pings à traiter. Un algorithme très utilisé de cette catégorie est celui proposé par Calder et communément appelé l'algorithme CUBE [69]. CUBE est un modèle d'erreur basé sur le calcul d'un MNB qui estime les valeurs potentielles de profondeur associées à un intervalle de confiance pour chaque nœud du MNB. L'algorithme fonctionne en trois étapes. La première étape prépare les données pour chaque nœud de la grille, en associant à chaque sonde une estimation de son incertitude totale propagée TPU. La deuxième étape consiste à générer une hypothèse de plus forte probabilité de l'endroit où le fond marin devrait se trouver pour chaque nœud de la grille. La troisième est une phase de désambiguïsation lorsque plusieurs hypothèses peuvent coexister. Dans le cadre de cette étape, l'algorithme présente à l'hydrographe des hypothèses alternatives de fond marin si la dispersion des sondes est trop importante. Cet algorithme est largement utilisé par les SH nationaux telles que la NOAA et le Service hydrographique du Canada (SHC). Cependant, les limites de cet algorithme sont bien connues, comme son mauvais comportement dans le cas de fonds marins chaotiques (zones rocheuses ou obstructions). Dans ce cas, le nombre d'hypothèses augmente significativement nécessitant alors l'intervention de l'hydrographe. Il est fortement recommandé de procéder à un rapide pré-filtrage manuel des données avant d'utiliser cet algorithme. Afin de prendre en compte ces limitations, une version améliorée de CUBE est proposée avec CHRT (CUBE *with Hierarchical*

*Resolution Techniques* présenté dans [83]) incluant la multi-résolution du MNB (adaptée aux zones de plus grande variabilité), le multi-traitement (pour améliorer les performances) et la prise en compte du facteur de qualité [67]. Les derniers algorithmes de cette catégorie sont présentés en détail dans [12].

Deux approches utilisent une approche fondée sur la densité [94], [111]. La première méthode consiste à projeter les données sur la direction transversale et la direction longitudinale respectivement dans le référentiel du SMF. Toutes les données, qui se trouvent maintenant dans le même plan, sont réparties dans des groupes. La plus grande région de données du fond marin est identifiée en recherchant le groupe le plus dense, puis en augmentant la taille de la région avec les groupes adjacents. Ensuite, un algorithme classique d'érosion et de dilatation de l'espace est utilisé pour localiser les valeurs aberrantes qui sont liées à la région de données du fond marin. Tout ce qui se trouve en dehors de cette région de données du fond marin est alors considéré comme une valeur aberrante. Enfin, chaque ping est également filtré par une méthode statistique sur une fenêtre d'étude locale. Cette méthode de densité utilise également des techniques innovantes de traitement d'image qui sont peu vues dans les méthodes étudiées. Segaghat dans [111] utilise une combinaison de deux algorithmes bien connus de la communauté scientifique du traitement des données pour détecter les données aberrantes acquises avec un SBES. Sa technique est établie sur un calcul de densité, utilisant comme algorithme de détection des aberrations le *Local Outlier Factor* (LOF) [114] et le DBSCAN [60] qui sera présenté en détail dans la section 2.4. Ces deux algorithmes sont combinés pour localiser des aberrations spatio-temporelles, temporelles dans le sens où elles restent présentes tout au long du temps (généralement des mois). L'algorithme LOF a été conçu comme une alternative pertinente aux algorithmes [94] car il détecte les outliers en fonction du voisinage local du point observé et donne un score d'anomalie pour chaque sondage. Le paramétrage s'effectue en fixant un seuil sur la valeur maximale autorisée de l'anomalie. Dans son article, Sedaghat a optimisé ce paramétrage afin de minimiser les fausses détections par rapport au jeu de données étudié. Pour le traitement de donnée LiDAR, une méthode basée sur le DBSCAN est proposée par Xie [115]. Cette méthode, présentée dans l'encart 2.4.1, supprime les données aberrantes dans les mesures LiDAR satellitaires d'ICESat-2.

Concernant les approches construites sur la distance, une méthode topologique

impliquant des voisinages non-locaux entre les sommets d'un graphe spécial est proposée dans [104]. Tout d'abord, les points de mesure sont décrits comme l'ensemble des sommets d'un graphe planaire fourni par le squelette de leur triangulation de Delaunay. Ensuite, pour toute paire de triangles adjacents, une arête reliant les deux sommets opposés est ajoutée au graphe initial. Enfin, chaque arête impliquant un changement de profondeur supérieur à un seuil donné est supprimée du graphe. Chaque composante connectée de ce graphe peut alors être identifiée comme un cluster potentiel de valeurs aberrantes. Cette partition permet de discriminer les pointes des structures cohérentes au-dessus du fond marin (par exemple, des tuyaux ou une épave).

Finalement les deux approches proposées par Du dans [89] et Kammerer dans [68] appartiennent à la catégorie des approches par clustering. La technique de Du [89] fonctionne en mode de représentation de la fauchée en mettant en cascade trois filtres. Le premier filtre applique un seuil min/max sur la profondeur à une fenêtre d'étude constituée d'un groupe de pings consécutifs. Une approche *coarse-fine* qui imite la prise de décision d'un opérateur est ensuite appliquée. Elle commence par la construction d'un histogramme de profondeur sur la fenêtre d'étude. L'ensemble des sondes est ensuite partitionné en fonction des modes de l'histogramme de profondeur. Les modes secondaires présentant un écart suffisant par rapport au mode principal sont identifiés comme des valeurs aberrantes. La fenêtre de travail est rétrécie en diminuant le nombre de faisceaux et le processus est répété. Cette approche récursive se termine par l'application d'un test statistique de Dixon [116] à un groupe de six sondes tout en nettoyant un ping. L'approche proposée par Kammerer [68] diffère de toutes les autres approches proposées dans cette étude car elle opère sur les données de rétrodiffusion SMF, le *backscatter*. Quatre algorithmes indépendants sont mis en œuvre pour détecter les zones de forte et de faible rétrodiffusion. L'histogramme, l'entropie et la segmentation K-means fonctionnent dans un mode de représentation ping à ping, tandis que le dernier utilise une analyse de texture d'une mosaïque. Ces algorithmes fournissent des limites de zones dérivées de l'imagerie qui sont visualisées avec les positions des sondes déclarées douteuses par un algorithme de nettoyage des données SMF. De tels outils sont utiles dans un contexte hydrographique où il faut préserver les obstructions pour la sécurité de la navigation.

### Les approches orientées surface

Ces approches sont fondées sur la modélisation du fond marin par un MNB. Les sondes qui s'écartent de ce MNB sont considérées comme des données aberrantes. Cette méthodologie s'appuie sur le choix d' :

1. un modèle mathématique qui décrit un ensemble des caractéristiques les plus représentatives de la morphologie du fond marin ;
2. une approche robuste qui prend en compte la présence de données aberrantes et qui suppose *a priori* un bruit aléatoire tout en estimant les paramètres du modèle ;
3. une stratégie qui identifie les données aberrantes comme un sous-ensemble de sondes distantes du modèle.

La méthodologie utilisée pour calculer le modèle est essentielle dans cette approche surface. Elle doit refléter la variabilité morphologique tout en isolant les sondes considérées comme aberrantes. Les approches proposées dans la littérature se divisent en deux catégories. Les approches globales ajustent une surface de tendance sur l'ensemble de la zone couverte par le levé, alors que les approches locales opèrent par une partition préliminaire de cette zone et ajustent ensuite indépendamment une surface de tendance locale sur chaque élément de la partition. Par conséquent, les approches globales mettent l'accent sur la capacité d'un modèle de tendance flexible capable de rendre compte de la complexité de la morphologie du fond marin. En réduisant la portée géographique de leur modèle, les approches locales peuvent s'appuyer sur des modèles plus simples, à condition que le partitionnement ait été effectué de manière adéquate par rapport aux changements topographiques du fond marin.

Pour créer une surface globale malgré la présence de valeurs aberrantes, différentes approches peuvent être envisagées. Ainsi, Arnold [95] utilise une procédure d'optimisation globale pour trouver une forme de surface stable. La surface bathymétrique est obtenue en minimisant une fonction d'énergie qui combine trois contraintes concurrentes : la régularité de la surface, la continuité et l'adhérence aux sondes. L'algorithme *Graduated Non-convexity* [117] est appliqué pour obtenir la surface d'énergie minimale. Plutôt que de minimiser la fonction d'énergie directement, ce qui est impraticable lorsque le nombre de valeurs aberrantes est élevé, cet algorithme minimise une approximation convexe de la fonction d'énergie. Cette

fonction grossière est raffinée de manière itérative jusqu'à ce qu'elle s'approche de la fonction d'énergie réelle. Une fois la surface d'énergie minimale obtenue, les valeurs aberrantes sont identifiées comme des points de déviation.

Bottelier [103] utilise quant à lui une technique d'interpolation robuste pour produire une surface lisse globale du plancher océanique. La technique de krigeage [118] est utilisée comme méthode d'interpolation. Par hypothèse, la précision de la mesure est la même quelle que soit l'observation. La dépendance spatiale des données est modélisée par une fonction de covariance gaussienne en fixant sa longueur de corrélation à la distance moyenne des points. La robustesse est obtenue en supprimant itérativement l'influence des valeurs aberrantes du système d'équations de krigeage surdéterminé.

Contrairement aux approches globales, les modèles sur lesquels reposent les approches locales peuvent être beaucoup plus simples puisque leur étendue spatiale est limitée. Toutes les approches appartenant à cette catégorie reposent sur des surfaces polynomiales de degré 0 à 3. Elles diffèrent par les techniques utilisées pour définir la taille et la forme du voisinage de travail. Il existe deux groupes d'approches. Les approches ponctuelles définissent un voisinage pour chaque sonde à tester, tandis que les approches surfaciques subdivisent d'abord la zone étudiée en blocs avant d'ajuster une surface de tendance locale sur chaque bloc et de pointer les aberrations.

Ainsi, l'algorithme de prédiction linéaire robuste mis en œuvre par Arnold [95] est une adaptation d'une technique de restauration d'image. Dans sa première forme, l'algorithme examine successivement les sondes d'une fauchée. La profondeur de la sonde est calculée à l'aide d'un M-estimateur [119] défini comme la somme pondérée des sondes appartenant à son voisinage de support non symétrique en demi-plan (x,y). Si la sonde est invalidée, sa profondeur est remplacée par la valeur prédite robuste. Le processus est répété jusqu'à la fin de la fauchée.

Enfin, Bisquay dans [92] applique une technique d'interpolation par krigeage [118] pour modéliser localement la topographie du fond marin en utilisant une fonction aléatoire d'ordre 1. La robustesse est introduite par l'application d'un des deux filtres proposés par la méthode afin d'éliminer les valeurs aberrantes lointaines. Rezvani [120] applique quant à lui une approche similaire pour réduire le volume très important de jeux de données bathymétriques lors de la construction d'un MNB. La profondeur est estimée sur chaque cellule en utilisant un plan

horizontal comme moyenne pondérée des sondes qu'il contient. L'approche a été appliquée en testant quatre M-estimateurs [119] différents, dont les estimateurs bi-poids de Tukey et Huber qui obtiennent les résultats les plus efficaces. Ces deux approches sont établies sur une division régulière de la zone et supposent que le modèle local choisi est statistiquement valide. Pour les cellules dont la taille n'est pas adaptée au modèle choisi, les sondes détectées reflètent l'inadéquation de la surface aux données. Debese [87] propose d'adapter automatiquement la taille des cellules en effectuant une subdivision en quadrillage descendant à l'aide de règles basées sur des inférences statistiques et spatio-temporelles. Cette approche multirésolution fournit un ensemble de valeurs aberrantes ainsi qu'une carte de classification qui indique les zones préoccupantes.

Parce qu'ils offrent un bon compromis entre le niveau de performance et le coût de calcul, les M-estimateurs [119] sont de loin le choix le plus courant lors de l'exécution de régressions robustes. Cependant, d'autres estimateurs sont théoriquement plus robustes, comme par exemple l'estimateur LTS (*Least Trimmed Squares*) utilisé par Lu [101] pour ajuster des données bathymétriques à l'aide d'un arrangement géospatial. Le point de rupture de l'estimateur LTS [121] est déterminé par le réglage de sa constante d'ajustement. Elle définit la fraction des points de données d'entrée (au moins 50%) utilisée pour trouver le sous-ensemble optimal, c'est-à-dire le sous-ensemble fournissant la somme minimale des résidus au carré. Parallèlement, son vaste espace de recherche combinatoire peut nécessiter des heuristiques supplémentaires pour modérer ses exigences informatiques. À cette fin, Lu limite les aménagements aux fenêtres temporelles de fauchées successives.

### Les algorithmes supervisés

La littérature sur la manipulation et le traitement de nuage de points est foisonnante notamment de par les apports récents de l'apprentissage profond comme le montre le papier [108] présentant une nouvelle méthode de segmentation de nuages de points, PointCNN, et ses performances vis-à-vis d'autres méthodes classiques, comme PointNet [122]. Dans le cas des algorithmes supervisés, tous les algorithmes trouvés dans l'état de l'art [54], [72], [106], [107] pour le traitement des données bathymétriques reposent sur des algorithmes de classification (souvent binaire avec une classe bathymétrie et une classe non-bathymétrie).

Pour les capteurs LiDAR, les avancées rapides dans le domaine de l'apprentissage profond sont liées à l'essor des véhicules autonomes et à l'emploi de cette technologie pour leur navigation, comme le montre l'article [108]. Néanmoins, la majorité de ces techniques est difficilement transposable telle quelle au LiDAR bathymétrique. En effet, dans le cadre d'une mesure topographique, l'onde optique ne change pas de milieu (air-eau-air) et par conséquent l'intensité lumineuse retournée permet d'obtenir beaucoup plus d'informations et une densité de points par $m^2$ beaucoup plus importante. Cependant, l'évolution récente des LiDAR bathymétriques, voir [123], améliorant les capacités d'acquisition (en augmentant par exemple la densité à l'aide de nouvelle génération de capteurs) favorise l'emploi de méthodes associées à l'apprentissage machine dans le cadre du traitement de nuages de points bathymétriques, comme nous le montrent les papiers récents [107] et [72]. Ces deux articles partagent le même objectif d'établir une classification d'un nuage de points d'entrée en travaillant directement sur les points et les descripteurs associés.

L'article [107] étudie la classification d'un nuage de points LiDAR provenant d'un AHAB Chiroptera I acquis en septembre 2013. La zone levée est une zone de référence et de test pour le service hydrographique allemand. Elle est donc maîtrisée en termes de connaissance et d'évolution. Quatre groupes d'objets artificiels ont été posés sur le sol dans un souci de capacité de détection (acquisition et traitement). L'objectif de la méthode est de placer chacun des points dans l'une de ces trois classes : fond marin, objets sur le fond et surface d'eau. Les auteurs construisent donc 15 descripteurs bien explicités dans l'article, fondée sur la géométrie du voisinage (de 5 mètres de rayon) du point (différences entre le point testé et le voisinage, planéarité, etc.) ou les caractéristiques du signal optique (amplitude, étalement de l'écho, etc.), afin de nourrir un réseau de neurones convolutif (*CNN* en anglais pour *Convolutional Neural Networks*) pour réaliser la classification. Le résultat issu de cette classification d'objets sur le fond comparé à des méthodes plus classiques d'apprentissage machine (comme le SVM, le Random Forest ou le k-NN) semble bon mais présente des écart-types très importants. Par exemple, 10.09% d'écart-type pour une précision de 72.34% pour la méthode RUSBoosted Trees, mais 3.56% d'écart-type pour une précision de 77.66% pour l'architecture neuronale proposée. Il est intéressant de conserver pour cet article la présentation des descripteurs et notamment l'emploi de descripteurs spatiaux locaux permet-

tant d'avoir une mesure de la morphologie locale. Ces descripteurs semblent être facilement ré-exploitables pour de futures études. Néanmoins, cet article est aussi à nuancer car la donnée LiDAR avant le processus IA semble prétraitée avec très peu de données de mauvaise qualité et les 3 classes déjà bien définies (possiblement traitement manuel pour supprimer les sondes dans la colonne d'eau). Nous sommes donc vraiment face à un problème de classification classique et non un problème de détection des sondes aberrantes. De plus, le volume de données utilisées pour l'entraînement est très faible (6198 points) au regard de la masse habituellement disponible en LiDAR (plusieurs centaines de millions au cours d'un vol quotidien).

Le second article traite un nuage de points LiDAR bathymétrique [72] en se reposant sur les attributs présents dans la donnée brute (tels que le nombre et le numéro de retour du signal, l'incertitude de mesure de position ou d'attitude, l'angle d'incidence, etc.). La donnée de test a été acquise par la NOAA via un RIEGL VQ-880G au large de la Floride en avril 2016 et contrairement au précédent papier, le volume de données est beaucoup plus représentatif (983320 sondes traitées). Cet article indique avoir travaillé avec 3 méthodes de classification supervisée : régression logistique régularisée, perceptrons multicouches et *extrem gradient boosting* régularisée (XGB). En revanche, seuls les résultats de ce dernier modèle sont présentés. La matrice de confusion, voir la table 2.2 qui décrit cette métrique, indique une précision très bonne (à 99.6%) mais qui est à nuancer compte tenu du déséquilibre des classes présentes pour le lot de données d'entrée.

Ces deux articles adoptent le même angle d'attaque : essayer de classifier un nuage de points à partir de descripteurs déjà présents dans la donnée ou construits à partir des informations spatiales ou associées à l'onde optique. La classification se fait également toujours sur le point (pas de rastérisation de la donnée) même si un voisinage peut être construit localement au moment de la génération des descripteurs.

Comme présenté sur la frise chronologique, voir figure 2.5, les méthodes supervisées appliquées à la donnée bathymétrique sont très récentes. Ainsi, pour le SMF, l'article [106] propose de classifier des nuages de points bathymétriques en exploitant une structure de données sous forme de voxel régulier. Ici aussi une classification binaire a été réalisée entre la donnée à conserver et la donnée considérée comme aberrante. Le modèle utilisé pour cette prédiction est adapté à partir d'un *CNN* U-Net [124], [125] classiquement utilisé dans la segmentation d'image-

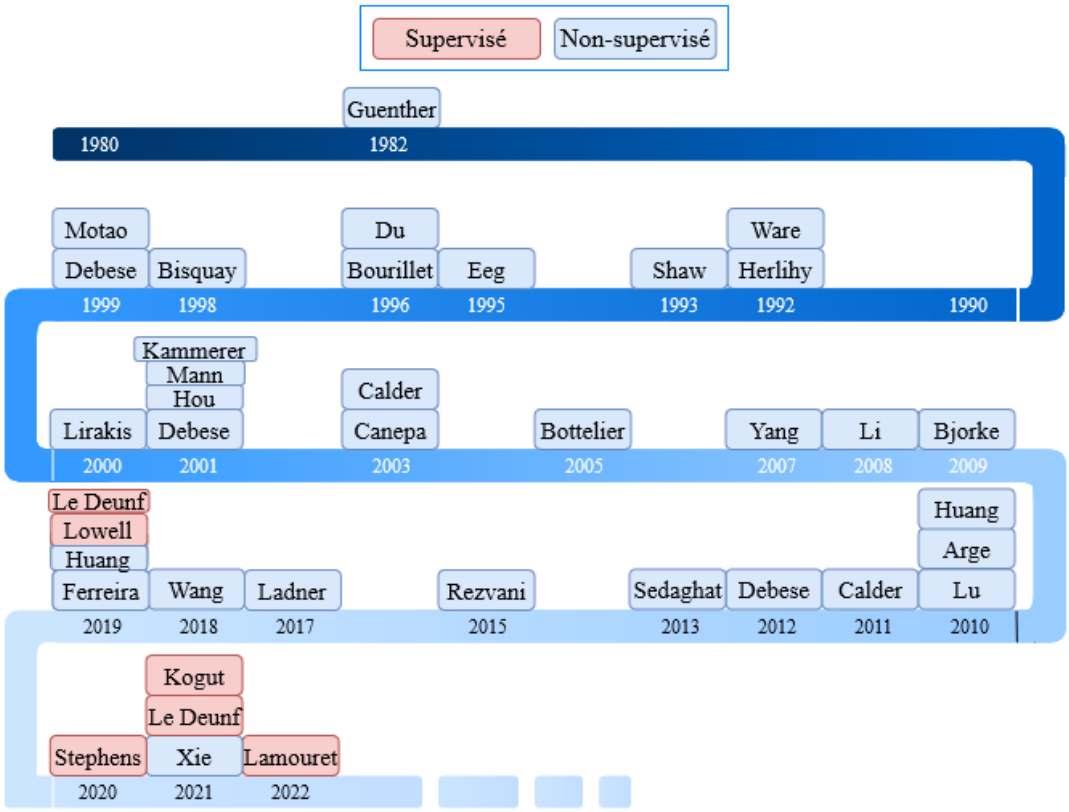


FIGURE 2.5 – Frise chronologique des différents algorithmes de détection des sondes aberrantes, issu de [12].

rie médicale. Ce type d'architecture est également utilisé par l'éditeur de logiciel de traitement de données bathymétrique CARIS et est implémentée dans le logiciel Hips&Sips dans le module CARIS Mira, mais nous n'avons trouvé aucune communication scientifique précisant les bases de leur méthode.

La seconde méthode, proposée pour le traitement de la donnée SMF via une méthode supervisée, a été développée dans le cadre de la thèse de Lamouret [54]. Cette méthode très innovante ne s'intéresse pas aux sondes mais directement à l'échogramme obtenu à partir du sondeur et des données présentes dans la colonne d'eau, voir figure 2.6. Le signal acoustique de l'échogramme est représenté sous la forme de son amplitude et est non complexe.

La méthode de détection du fond repose également sur un *CNN* mais cette fois-ci il s'agit de ResNet50, le réseau ResNet [125] comportant 50 couches. Cette

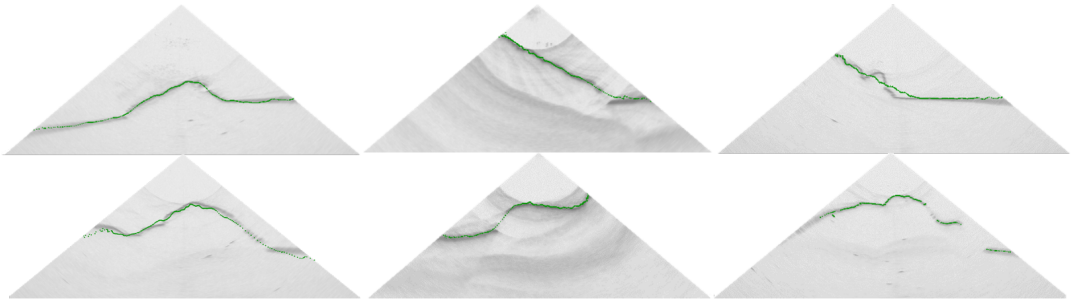

FIGURE 2.6 – Extraction de la bathymétrie (en vert) dans l'imagerie de la colonne d'eau, issu de [54].

méthode permet de travailler directement sur le signal acoustique sans prendre en compte les autres capteurs d'un système de levé hydrographique.

La performance d'un algorithme peut être évaluée selon différents critères tels que la qualité de ses résultats, sa complexité ou son explicabilité. Le nombre de paramètres, leur paramétrage simple ainsi que la possibilité d'un contrôle *a posteriori* du réglage des paramètres doivent être pris en compte dans un contexte opérationnel. Le nombre de paramètres varie fortement d'une méthode à l'autre. Dans certains cas, aucun paramètre n'est nécessaire car l'approche est dédiée à un sondeur spécifique [81]. De plus, ces paramètres peuvent également nécessiter des modifications d'un environnement morphologique à un autre [126]. Cette dernière contrainte opérationnelle rend potentiellement complexe l'utilisation d'algorithmes qui n'arriveraient pas à s'adapter à de nouvelles zones de levés hydro-océanographiques. Or, le Shom, dans le cadre de sa mission, cherche à couvrir une zone d'intérêt de plus de 60 millions de $km^2$ (comprenant la zone économique exclusive et les zones d'intérêt cartographique et de défense). Il est donc indispensable de disposer d'une méthode s'adaptant à toutes ces morphologies mais prenant également en compte les besoins pour la sécurité de la navigation.

### 2.3.3 Le cas du Shom

Le Shom, en tant que plus ancien SH mondial, possède un historique important dans la gestion, le traitement et le contrôle de la donnée bathymétrique issue de capteurs hydrographiques. En 1991, un projet a été mis en place dans le cadre de l'intégration de nouveaux capteurs hydrographiques, démarrant de manière ef-

fective avec l'installation du SMF Thomson Lennermor sur le bâtiment hydrographique Borda et le SMF Simrad EM12 dual sur le bâtiment hydrographique L'Espérance. Ce projet a été maintenu pour l'installation des SMF Kongsberg-Simrad EM1002S sur les bâtiments hydrographiques de seconde classe à partir de 1998 et au-delà avec le navire Pourquoi Pas ? de l'Ifremer (et co-financé par cet institut) pour la partie acquisition de données et également pour assurer le suivi d'études amont, financées par la Direction Générale de l'Armement (DGA) sur le traitement et l'exploitation des données bathymétriques SMF.

Ainsi, dès l'arrivée en 1991 des SMF au Shom, la question du traitement de ces données s'est confirmée. Initialement, la stratégie du Shom était d'implémenter dans sa chaîne de traitement, le démonstrateur CIRCE, développé dans le cadre d'études amont (financées par la DGA), afin d'industrialiser les concepts de traitement validés par le Shom. Ce démonstrateur était développé en priorité pour les SMF grands fonds. Des développements supplémentaires étaient planifiées en 2000 pour poursuivre les travaux de recherche algorithmique, méthodologique et informatique afin d'adapter ces concepts à la volumétrie des SMF très petits fonds. Notamment dans le cadre de l'utilisation de vedettes hydrographiques équipées des SMF Kongsberg-Simrad EM3000 puis EM3002, l'objectif était de réaliser un transfert opérationnel des outils issus de ces études amont.

Suite à la réalisation au début des années 2000 d'un état de l'art des systèmes sur étagère, le Shom a décidé de privilégier la mise en œuvre de HIPS pour le traitement des données de bathymétrie et même dans un second temps la génération des mosaïques de réflectivité *backscatter* (à partir de 2006 suite à l'amélioration des capteurs pour cette fonctionnalité). Après cette décision de basculer vers des logiciels déjà industrialisés, le Shom a suspendu un temps les études amont pour les chaînes de production.

La recherche sur le traitement de données bathymétriques a repris pendant une courte période à la fin des années 2000 avec notamment la publication d'algorithmes novateur sur le sujet comme [127]. Mais ces méthodes ([68], [127], [87]) n'ont jamais été implémentées dans des solutions industrielles et le Shom a très rapidement priorisé les recherches dans d'autres domaines jusque très récemment (2018 et le début de cette thèse).

Aujourd'hui, l'ensemble des données bathymétriques issues des capteurs SMF est traitée manuellement ou semi-automatiquement via l'intégration de l'algo-

rithme CUBE dans le logiciel de traitement de donnée bathymétrique HIPS&SIPS de la société américaine Teledyne Caris (version 11 actuellement en production au Shom).

Pour le capteur LiDAR, le logiciel de traitement de données Qimera de la société néerlandaise QPS a été utilisé jusque 2020. Depuis, le logiciel open-source PFM ABE développé au sein de l'organisme américain NAVal OCEANographic Office (NAVOCEANO) est en production pour l'édition des sondes LiDAR bathymétriques. Pour ces deux logiciels, le traitement est réalisé quasiment intégralement manuellement avec l'utilisation de quelques filtres attributaires en fonction du contexte environnemental. Cette nécessité du traitement manuel (et d'un contrôle poussé) a pour origine le besoin important de sécurité associé aux données exploitées ensuite dans les produits nautiques (comme les CM) mais également pour d'autres produits de la défense. Le rôle de la chaîne de traitement est donc de s'assurer d'un très haut niveau de qualité qui passe par un contrôle exhaustif de la donnée bathymétrique (et particulièrement des points hauts). Cet historique succinct met en lumière les difficultés que peut rencontrer un SH lorsqu'il souhaite mettre en place et maintenir une chaîne de traitement de données interne : conduite du changement, confiance des opérateurs, formation des utilisateurs, mise en place d'une chaîne opérationnelle et maintien en condition de cette chaîne, ainsi que les choix de stratégie de traitement qu'il peut mettre en place pour dépasser des difficultés techniques. Au sein du Shom, une étude en 2019 sur les paramètres de l'algorithme CUBE [69] à employer avec la nouvelle gamme de capteurs SMF, voir [32], a permis de redonner confiance dans l'utilisation de l'algorithme CUBE et de s'intégrer dans les méthodes de traitement facilitant en partie le travail des opérateurs, voir [128] mis à jour à l'été 2022 (première édition été 2021). De plus, intégrer des développements issus de la recherche dans des solutions industrielles est complexe et peu de centres de recherche ont accompli cet objectif, hormis pour le cas de CUBE développé au sein du *Center for Coastal and Ocean Mapping*, qui a pu être utilisé dans HIPS&SIPS dès 2007 puis plus tard dans Qimera, autre logiciel de traitement de données bathymétriques de la société QPS.

## 2.4 Contribution au traitement de la donnée SMF

Une fois cet état de l'art établi, nous avons mis en place une première preuve de concept de traitement de donnée SMF dans un cadre supervisé, qui a fait l'objet de l'article [19]. En effet, aucun article scientifique traitant du sujet du traitement de donnée bathymétrique à l'aide de technique supervisé n'existait en 2019. L'objectif a donc été de déterminer si les techniques récentes de ML pouvaient s'appliquer à ce type de données et par la même occasion de réfléchir aux méthodes à mettre en place pour faciliter le travail des hydrographes. Cette réflexion permet également de planifier les efforts d'infrastructures et de construction de bases d'apprentissage à fournir dans les prochaines années pour satisfaire ce type de méthode et l'augmentation du volume de données engendrée par l'ensemble des SH dont le programme CHOF du Shom.

Pour les premiers tests d'algorithmes ML appliqués aux données bathymétriques, nous avons choisi de tester la perspective de la classification supervisée qui s'apparente en effet à la tâche quotidienne réalisée par les hydrographes des SH via l'étiquetage des données dans deux classes, "acceptées" ou "rejetées". Nous avons ainsi testé trois algorithmes différents (voir les sections suivantes) sur des données acquises avec un sondeur EM2040p de Kongsberg en Nouvelle Calédonie, l'amplitude de profondeur de la zone étudiée allant de 9.93m à 99.25m. La figure 2.7 montre la localisation du levé dans la passe de Koumac.

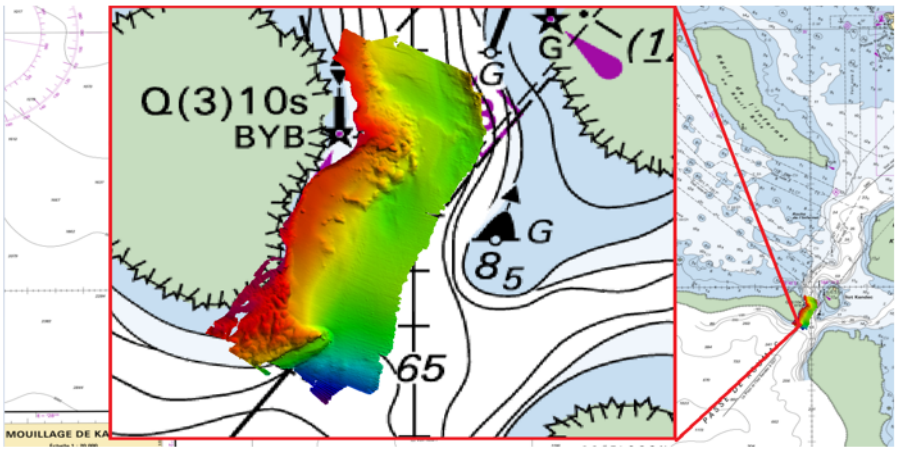


FIGURE 2.7 – Donnée étudiée apposée sur la CM6985 ©Shom.

### 2.4.1 Analyse de la donnée SMF et sélection des descripteurs

Afin de déterminer les descripteurs/caractéristiques pertinents pour l'emploi d'une technique ML, nous avons analysé les données issues de l'acquisition SMF. Ainsi, une fois les données géoréférencées et traitées via le logiciel CARIS HIPS & SIPS, nous avons exporté les données au format .ascii avec le maximum d'attributs possible (comme l'angle d'incidence longitudinale et transversale, la TPU, la force du signal rétrodiffusé, le *timestamp*) et le statut associé à chaque sonde soit :

- acceptée, la sonde est considérée comme valide et sera utilisée dans le produit bathymétrique final ;
- rejetée à la conversion (*conversion - disabled beam* sur la figure 2.8), ce type de sonde a été détecté par l'algorithme de détection de fond du SMF mais filtré par l'algorithme interne car considéré comme un faisceau défectueux (se fondant sûrement sur le *backscatter* mais sans certitude car secret industriel) ;
- rejetée par l'hydrographe (*subset - hydrographer* sur la figure 2.8) (attribut qui permet d'apprendre et de valider).

À partir des données extraites mais également de l'état de l'art présenté dans la section 2.3, nous avons classé les descripteurs possibles en trois familles :

- les descripteurs naturels car présents au moment de l'acquisition ;
- les descripteurs spatiaux (dans le repère géographique des données bathymétriques) car déterminés à partir d'une emprise spatiale autour de la sonde étudiée ;
- les descripteurs séquentiels (dans le repère *beam*/*ping*) car déterminés à partir d'une emprise temporelle autour de la sonde étudiée et donc établie sur le *timestamp*, le numéro de faisceau et la géométrie du sondeur.

Ainsi nous avons mis en place les descripteurs présents dans la table 2.1 :

| Descripteurs naturels | Descripteurs spatiaux | Descripteurs séquentiels |
|---|---|---|
| TPU | DBSCAN ([60]) | score *bad ping* ([93]) |
| Angle d'incidence longitudinal | MAD ([129]) | Test d'échantillons ([93]) |
| Intensité acoustique | LOF ([114]) | Test de variance locale ([93]) |

Table 2.1 – Présentation des descripteurs.

**DBSCAN**

Le DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) est un algorithme de partitionnement de données proposé en 1996 dans l'article [60]. Il s'agit d'un algorithme de clustering, voir 1.3, basé sur la densité : étant donné un ensemble de points dans un espace donné, il regroupe les points qui sont étroitement liés (points ayant de nombreux voisins proches), en marquant comme aberrants les points qui se trouvent seuls dans des régions à faible densité (dont les voisins les plus proches sont trop éloignés). L'algorithme DBSCAN utilise 2 paramètres : la distance $\epsilon$ et le nombre minimum de points $MinPts$ devant se trouver dans un rayon $\epsilon$ pour que ces points soient considérés comme un cluster. L'idée de base de l'algorithme est ensuite, pour un point donné, de récupérer son $\epsilon$-voisinage et de vérifier qu'il contient bien $MinPts$ points ou plus. Ce point est alors considéré comme faisant partie d'un cluster. On parcourt ensuite l'$\epsilon$-voisinage de proche en proche afin de trouver l'ensemble des points du cluster.

Analysons quelques descripteurs en fonction du statut des sondes, l'objectif des algorithmes supervisés étant de classer les sondes soit dans la classe acceptée soit dans la classe rejetée. La première analyse que nous pouvons effectuer est le fait que nous sommes en présence de données fortement déséquilibrées (moins de 0,1% des sondes rejetées par l'opérateur). Sur la figure 2.8, nous poursuivons l'analyse pour les attributs de retour de l'intensité acoustique de l'écho de fond, l'angle d'incidence (*across angle*) et la TPU associée à la sonde.

Dans le cas du retour de l'intensité acoustique, la figure montre clairement que cette caractéristique semble être actuellement utilisée par les fabricants pour rejeter les données à la conversion (en bleu sur la figure) avec un seuil sur cette valeur qui se situerait aux alentours de $50dB$. De plus, le niveau moyen de rétrodiffusion des données rejetées par l'hydrographe (en vert sur la figure) est également inférieur à celui des données acceptées, $-15dB$ contre $-5dB$. Cette caractéristique semble

donc bien discriminante pour les sondes rejetées (que ce soit manuellement ou automatiquement).

Pour la mesure de l'angle d'incidence longitudinale (graphique au centre de la figure 2.8), nous pouvons également remarquer que les trois statuts ont des valeurs moyennes, associées à cette mesure d'angle, différentes, 35° pour le statut accepté, 43° pour les données rejetées à la conversion et 55° pour les données rejetées par l'hydrographe. Cette caractéristique semble également discriminante pour notre problématique.

Enfin, pour la valeur de TPU sur la mesure de profondeur, les valeurs moyennes des trois statuts sont très proches (autour de $0.34m$) mais leur écart-type varie d'avantage allant de $3cm$ pour les sondes rejetées par l'hydrographe à $7cm$ pour les données acceptées. Cette caractéristique semble moins significative que les autres descripteurs.

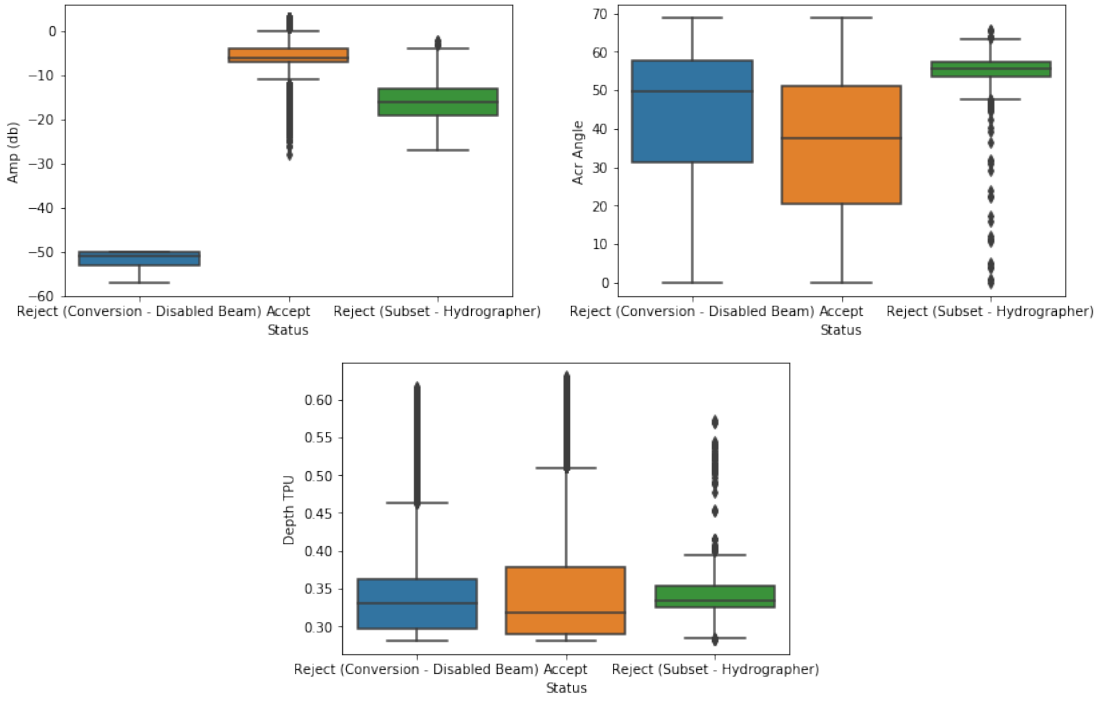


FIGURE 2.8 – Analyse de la donnée bathymétrique selon les statuts acceptés, rejetées à l'acquisition et rejetés par l'hydrographe pour les attributs de retour de l'intensité acoustique pour l'écho de fond (en haut à gauche), l'angle d'incidence (*across angle*, en haut à droite) et la TPU associée à la sonde (en bas).

La figure 2.9 présente l'analyse statistique du descripteur spatial MAD distance

(à gauche) et du descripteur temporel score *bad ping* associé à la sonde (à droite).

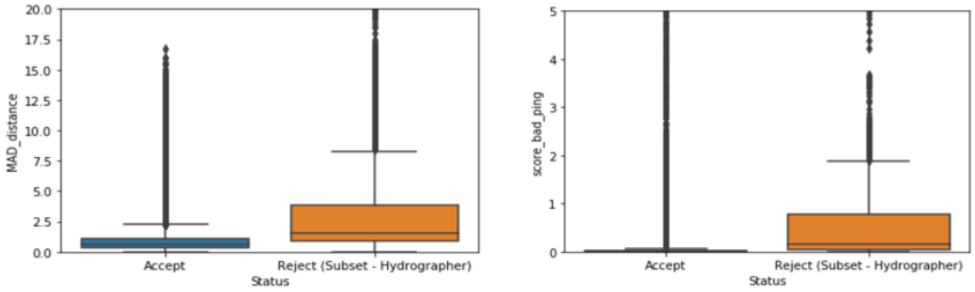


FIGURE 2.9 – Analyse de la donnée bathymétrique selon les statuts acceptés et rejetés par l'hydrographe pour les attributs MAD distance (à gauche) et le score *bad ping* associée à la sonde (à droite).

En ce qui concerne la caractéristique spatiale MAD, nous observons que la MAD-distance est plus grande pour les données rejetées que pour les données acceptées, voir figure 2.9. Ce comportement était attendu car la MAD-distance calcule la dispersion des données. Or, les valeurs aberrantes sont des informations davantage dispersées autour de la médiane. Cette caractéristique est donc bien discriminante pour les sondes rejetées (que ce soit manuellement ou automatiquement). Le calcul du MAD est décrit précisément à la section 3.2 dans le cadre de la régularisation spatiale associée à la prédiction du fond.

Quant au descripteur séquentiel score *bad ping*, nous observons qu'il est plus important pour les données rejetées que pour les données acceptées, voir figure 2.9. Ce comportement est également attendu puisque cette caractéristique est déjà utilisée dans le cadre d'un algorithme classique de détection des valeurs aberrantes et présent dans l'état de l'art 2.3. Cette caractéristique est donc également discriminante pour notre problème.

### 2.4.2 Classifieurs testés et résultats de l'apprentissage

L'objectif de cette première approche est de classer les sondes dans un statut accepté (représentative des fonds marins) ou rejeté (donnée considérée comme aberrante), dans le cadre d'un problème de classification binaire. Ainsi, afin de réaliser cette tâche de manière supervisée, nous avons choisi trois techniques différentes, souvent utilisées dans la littérature sur l'apprentissage automatique, comme

classifieurs potentiels :

— la régression logistique ;
— le RF ;
— le *gradient boosting* (XGBoost).

Ces méthodes ont été exploitées directement à partir de la librairie *Scikit-Learn* [130] qui facilite la mise en oeuvre et la démocratisation de l'approche.

La régression logistique vise à prédire une cible binaire en estimant les paramètres d'un modèle logistique qui sera une combinaison linéaire de nos descripteurs [131]. Cette méthode est simple à mettre en œuvre, efficace peu importe le volume de données, mais le risque de sous-apprentissage, voir 1.3, est important.

Le RF est une méthode ensembliste utilisée pour la classification, qui fonctionne en calculant des combinaisons sur des arbres de décision [132]. Bien que l'algorithme associé soit plus difficile à mettre en œuvre que le modèle de régression logistique, tout en nécessitant beaucoup plus de données, il reste explicable car fondé sur des arbres de décision. Il nécessite un bon équilibre entre les données acceptées et celles rejetées [133]. Pourtant, les données bathymétriques SMF contiennent clairement beaucoup plus de données acceptées que rejetées. Pour tester l'algorithme, nous avons donc sélectionné la zone contenant le plus de données aberrantes.

Le XGBoost est une méthode par descente de gradient (méthode itérative utilisée en optimisation pour trouver le minimum d'une fonction mathématique) combinée à une méthode de *boosting* (algorithme ensembliste d'apprentissage automatique). L'idée est de combiner itérativement des apprenants faibles en un seul apprenant fort [134], mais contrairement au RF où les apprenants sont indépendants dans le cadre du *boosting*, chaque apprenant faible est entraîné pour corriger les erreurs de l'apprenant faible précédent. Cette technique est très rapide et puissante, mais la méthode est par nature moins explicable [56].

Dans le cadre des méthodes ML supervisées de classification, la matrice de confusion est un outil très classique utilisé pour mesurer la qualité d'un classifieur, voir [135]. Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée. Un des intérêts de la matrice de confusion est de montrer rapidement si un système de classification parvient à classifier correctement. La figure 2.2 présente la matrice de confusion théorique, sur la diagonale partant du coin supérieur gauche au coin inférieur droit, les classes prédites par le classifieur

sont cohérentes avec l'étiquetage humain, sur l'autre diagonale les classes sont incohérentes. Plus la matrice de confusion s'approche d'une matrice diagonale et meilleur est le classifieur. Les faux positifs sont des erreurs de type 1 en statistique alors que les faux négatifs sont des erreurs de type 2.

| | | Classe estimée par le classifieur | |
|---|---|---|---|
| | | Classe A | Classe B |
| Classe réelle estimée par un humain | Classe A | Vrais positifs VP | Faux négatifs FN |
| | Classe B | Faux positifs FP | Vrais négatifs VN |

TABLE 2.2 – Présentation de la matrice de confusion.

Afin de mieux décrire ces résultats, d'autres métriques associées à la matrice de confusion sont calculées comme la précision et le rappel [135], et sont définies ainsi dans le cas d'une classification binaire.

$$precision = \frac{VP}{VP + FP} \tag{2.1}$$

$$rappel = \frac{VP}{VP + FN} \tag{2.2}$$

Le $F1 - Score$ , voir [135], est la moyenne harmonique de la précision et du rappel (pour une même pondération) et permet ainsi de combiner ces deux grandeurs. Il s'agit donc d'une métrique globale permettant de mesurer les résultats d'un classifieur de manière équilibrée.

$$F_1 = 2 * \frac{precision * rappel}{precision + rappel} \tag{2.3}$$

Les différents algorithmes ont également été comparés en utilisant le $F_1 - Score$. Dans le flux de travail ML, l'objectif est de minimiser les résultats faux négatifs, car il est très important de s'assurer qu'une sonde acceptée par un hydrographe ne sera pas rejetée par la méthode supervisée. Typiquement, nous ne voulons pas filtrer toute information pertinente isolée qui pourrait être une épave ou une obstruction. Pour cette raison, nous voulons que le score de *Traitement manuel Acceptées/Prédiction ML Rejetées* soit le plus faible possible (chiffre en rouge dans les tableaux 2.3, 2.4 et 2.5).

Ces trois méthodes ont été appliquées sur des mêmes sous-échantillons de données issues du levé de la passe de Koumac présenté plus haut. En effet, nous avons construit des zones de 100m par 100m (comportant environ 275 000 sondes) et pour chaque zone mis en place un ratio apprentissage/test de 70%/30%. Nous détaillerons la séparation des jeux de données dans la section 3.2. Nous avons dû réaliser les apprentissages sur ces sous-échantillons car le calcul des descripteurs non-supervisés peut être long, l'algorithme DBSCAN ayant une complexité de $O(n^2)$ avec $n$ le nombre de données à traiter [136]. De plus, cela nous a permis également de sélectionner les sous-échantillons possédant le plus de données aberrantes évitant ainsi d'utiliser des sous-échantillons ne contenant que des données valides afin de limiter le déséquilibre des classes, pour rappel voir la section 1.3. Les résultats des algorithmes suivants ont ainsi été obtenus sur la zone d'étude $Northing \in [7713100 - 7713400]$ et $Easting \in [422100 - 4222200]$.

Le tableau 2.3 présente les métriques et les résultats de la classification pour le modèle de régression logistique. Il montre ainsi que le modèle de régression logistique est peu performant et présente un risque élevé de sous-apprentissage. Ce modèle fonctionne bien lorsqu'il y a très peu de valeurs aberrantes dans les données. Le pourcentage de faux négatifs est supérieur à 2% des données, ce score semble être trop important pour la sécurité de la navigation. Une discussion sur le maximum acceptable de faux négatifs doit être menée dans le cadre de la sécurité de la navigation afin d'évaluer principalement l'impact d'une non-prise en compte de ces données valides dans le produit bathymétrique final. Nous constatons, également, ici la limite de la métrique : si tous les faux positifs se trouvent au-dessus du tapis de sondes, les conséquences sont graves. La métrique ne fournit pas ce niveau de détail.

| | | Prédiction ML | | |
|---|---|---|---|---|
| $F_1$-Score = 0.16 classe rejetée | | Acceptées | Rejetées | Total |
| Traitement manuel | Acceptées | 570 (0.21%) | 17 (0.01%) | 275020 |
| | Rejetées | **5 858** (2.12%) | 269162 (97.66%) | 587 |
| | Total | 269179 | 6428 | 275607 |

TABLE 2.3 – Résultats de classification de la méthode régression logistique.

Le tableau 2.4 montre les métriques et les résultats de la classification pour le modèle RF. Les résultats donnés par cet algorithme sont bien meilleurs que ceux de la régression logistique. Le pourcentage de faux négatifs est d'environ 0,001% des données, ce qui nous donne une plus grande confiance dans la prédiction. Ainsi le $F1 - Score$ est bien supérieur à celui de l'algorithme précédent.

| | | Prédiction ML | | |
|---|---|---|---|---|
| $F_1$-Score = 0.95 classe rejetée | | Acceptées | Rejetées | Total |
| Traitement manuel | Acceptées | 532 (0.19%) | 55 (0.01%) | 275020 |
| | Rejetées | **4** (0.00%) | 275016 (99.79%) | 587 |
| | Total | 275071 | 536 | 275607 |

TABLE 2.4 – Résultats de classification de la méthode RF

Le tableau 2.5 montre les métriques et les résultats de la classification pour le modèle *XGBoost*. Les résultats donnés par cet algorithme sont très proches de l'algorithme RF. Le pourcentage de faux négatifs est également d'environ 0,001% des données. Le $F1 - Score$ est le plus élevé pour cet algorithme mais très proche de la méthode par RF.

| | | Prédiction ML | | |
|---|---|---|---|---|
| $F_1$-Score = 0.97 classe rejetée | | Acceptées | Rejetées | Total |
| Traitement manuel | Acceptées | 552 (0.20%) | 35 (0.01%) | 275020 |
| | Rejetées | **5** (0.00%) | 275015 (99.79%) | 587 |
| | Total | 275050 | 557 | 275607 |

TABLE 2.5 – Résultats de classification de la méthode XGBoost.

Ces résultats nous ont permis de nous assurer de la capacité des méthodes ML à traiter ce type de données bathymétriques. Une étude plus approfondie sur le paramétrage des algorithmes utilisés ainsi que la construction des descripteurs aurait sûrement permis d'améliorer les résultats.

Dans la continuité, l'étude de la donnée SMF a ainsi été proposée comme projet de recherche au département science de donnée de la DGA maîtrise de l'information afin de tester différentes méthodologies.

## 2.5 Conclusion

Ce chapitre nous aura permis de comprendre plus en profondeur la donnée bathymétrique et de formaliser les erreurs présentes dans ce type de donnée et auxquelles les hydrographes sont quotidiennement confrontés. L'état de l'art, voir section 2.3, synthétise 40 ans de recherche et de développement dans le domaine des sciences informatiques appliquées aux mesures bathymétriques, les contraintes rencontrées, et propose une analyse des différents courants associés aux techniques de traitement de données.

Pour ce premier niveau d'échelle, les techniques utilisées se sont focalisées sur la sonde bathymétrique et son voisinage (pour le calcul des descripteurs). L'objectif était, pour la majorité des algorithmes, de classer la sonde comme une donnée acceptée (représentative des fonds marins) ou rejetée (donnée considérée comme aberrante). La section 2.4 a présenté les premières tentatives d'application d'algorithmes ML sur des données bathymétriques. Ces algorithmes ML appliqués aux données bathymétriques et la construction de descripteurs non-supervisés ont généré des résultats prometteurs. Ainsi, la preuve de concept a été validée pour ce type de techniques appliquées à ces données bathymétriques pour la problématique de la détection des données aberrantes. Dans le futur, cette combinaison d'outils fera partie intégrante de la boîte à outils de l'hydrographe, mais il sera nécessaire de s'interroger sur l'impact de ces nouvelles approches sur le quotidien des opérateurs, comme présenté dans le chapitre 5.

Néanmoins, de nombreux axes de recherche doivent encore être investigués. En effet, bien que nos résultats soient acceptables sur de petites zones, ils le sont bien moins sur des zones plus grandes. La structure des données varie beaucoup d'une fenêtre d'étude à l'autre. En effet, les données obtenues sur une zone plane n'ont pas la même disposition que les sondes d'une crevasse. C'est sans compter les conditions de navigation, les spécificités des sondeurs ou la nature des fonds marins. Une piste d'amélioration consisterait à définir des typologies sur ces fenêtres grâce à des descripteurs spécifiques. Cette technique doit être menée conjointement avec des experts métier afin que les résultats correspondent à des types de fonds réels rencontrés en bathymétrie. Chaque typologie de fenêtre permettrait d'entraîner un algorithme de détection des sondes aberrantes. Une fois ces données bathymétriques traitées, elles peuvent être qualifiées, voir le chapitre 4, afin

d'être exploitées par les autres chaînes de production des SH.

La question du passage à l'échelle est également centrale car les temps de calcul des différents descripteurs peuvent être particulièrement longs. C'est pourquoi il nous a semblé essentiel de réfléchir à une structuration plus efficace de la donnée plutôt que de s'intéresser directement aux sondes et ce afin de pouvoir appliquer des algorithmes d'apprentissage à des données beaucoup plus massives comme les données issues des LiDAR bathymétriques. Nous passons donc, de la donnée (la sonde) au niveau micro, à l'information au niveau meso pour un groupe de sondes dans le chapitre 3.

Chapitre 3

# L'échelle meso : structuration de la donnée vers l'information

**Synthèse du chapitre**

L'étude de la donnée à un niveau micro permet d'analyser précisément le comportement d'une donnée aberrante et de construire des descripteurs dans un voisinage spatial ou temporel local. Néanmoins, au changement d'échelle et pour un levé bathymétrique complet, des difficultés apparaissent en matière de temps de calcul des descripteurs pour des volumes de données importants. Il est donc nécessaire de structurer cette donnée afin de bâtir une nouvelle information plus facilement utilisable par des techniques d'apprentissage supervisé ou non-supervisé. Ce chapitre s'intéressera donc à la construction de cette structuration pour le cas applicatif de la donnée LiDAR bathymétrique et présentera les résultats d'algorithmes de régression supervisés avec pour objectif de prédire les fonds marins à partir des sondes bathymétriques brutes. En fin de chapitre, nous présenterons l'intégration de cette méthodologie dans la chaîne de traitement de la donnée LiDAR bathymétrique du Shom et l'avis des opérateurs l'ayant testée.

Suite à la validation de la preuve de concept de l'utilisation de méthodes ML combinées avec des créations de descripteurs fondés sur des méthodes non-supervisées, nous avons orienté (en lien avec le Shom) notre recherche sur la donnée LiDAR bathymétrique afin de soutenir le département Altimétrie Littorale (AL) du Shom. Les raisons de cette orientation sont multiples. En effet, dans l'état de l'art, nous avons pu observer que les publications autour du traitement de données LiDAR bathymétriques se faisaient de plus en plus nombreuses [72], [80], témoignant d'une attention particulière de la recherche pour ce type de donnée. Elle peut s'expliquer par des jeux de données au volume très important pour chaque ligne de vol, au déséquilibre des classes (données bathymétriques ou erreurs) qui penche, contrairement aux données SMF, clairement en faveur de la détection des *outliers*, mais également au format de donnée beaucoup plus ouvert que les données SMF et facilement utilisable dans des scripts informatiques via des librairies ou des utilitaires conçus pour ce type de données. Nous avons également profité d'une conjoncture favorable avec le lancement d'un projet, financé par la Direction interministérielle du numérique dans le cadre d'un appel à manifestation d'intérêt sur l'intelligence artificielle, pour accélérer les développements autour de la donnée LiDAR bathymétrique et ainsi tester plus rapidement auprès des opérateurs nos solutions algorithmiques.

Au vu des volumes de données beaucoup plus importants engendrés par les capteurs LiDAR bathymétriques, plus de 700000 points pour un fichier de vol de 4 secondes couvrant environ $0.04km^2$, nous avons rapidement envisagé d'optimiser la structure de cette information pour faciliter son intégration dans un processus construit avec des algorithmes ML.

De plus, l'état de l'art effectué au chapitre précédent a montré que les solutions proposées aujourd'hui se fondent uniquement sur des méthodes de classification binaire (bathymétrie ou non-bathymétrie) ou multi-classes (fond, surface d'eau, objets sur le fond), voir [72], [107]. Bien que la classification des sondes soit une tâche réalisée quotidiennement par les opérateurs, une classification automatique peut être difficilement explicable pour les opérateurs, notamment en cas d'erreur de la classification sur certains types de zone, ce qui entraînerait potentiellement une perte de confiance dans l'outil de classification automatique dans le cadre de la sécurité de la navigation. Afin de proposer un nouvel outil d'aide à la décision sur la donnée LiDAR bathymétrique et en concertation avec les experts de la donnée

LiDAR, nous avons donc testé une nouvelle approche basée sur une régression afin de prédire le fond plutôt que de classer les données brutes (une comparaison des données brutes à cette prédiction permettant dans un second temps de classer les sondes).

## 3.1 Analyse du capteur LiDAR bathymétrique

La figure 3.1 recense les principaux signaux (comme la surface d'eau) et erreurs présents dans un jeu de données LiDAR topo-bathymétrique classique (à gauche), ainsi que le résultat généré après traitement manuel (à droite). Dans ce contexte, la validation des sondes repose uniquement sur la décision d'un opérateur qualifié [8]. Bien que cette tâche soit soutenue par un logiciel de visualisation (PFM ABE dans le cas du Shom et la NAVO) dédié au traitement et au contrôle de la donnée LiDAR bathymétrique, le processus est fastidieux, chronophage et ne garantit pas la préservation de toutes les caractéristiques bathymétriques (comme les épaves ou obstructions) présentant un risque potentiel pour la navigation. Ce type de traitement est également sujet à erreurs de par la répétitivité de la tâche à accomplir. Ainsi, notre objectif de recherche est d'estimer le fond marin réel à partir de nuages de points LiDAR topo-bathymétriques pour que ce fond prédit soutienne leur méthode actuelle de traitement.

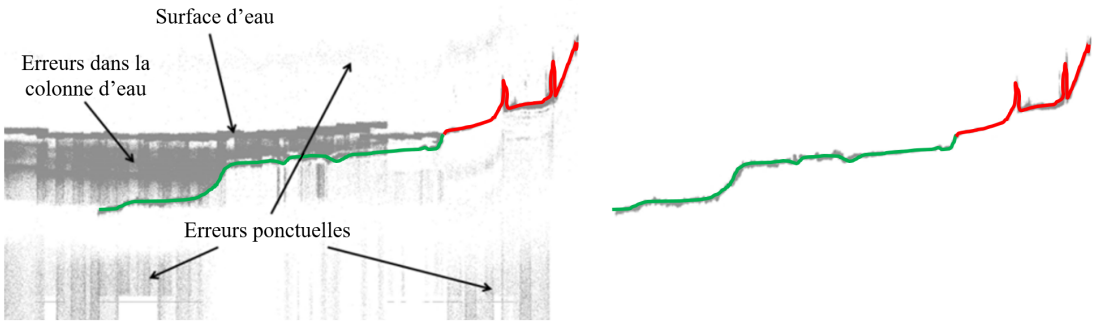


FIGURE 3.1 – Types de signaux et erreurs représentés par une section LiDAR bathymétrique (à gauche), le fond marin (en vert) et la topographie (en rouge) après traitement manuel (à droite).

La donnée LiDAR repose sur le format .las, voir [137], format ouvert à partir duquel nous avons directement travaillé. La figure 3.2 recense les attributs disponibles dans la version du format utilisé actuellement au Shom. Nous remarquons

qu'en plus des données de position (X, Y, Z), la donnée brute contient des descripteurs présents naturellement dans la donnée. Ainsi, nous distinguons les attributs associés directement au spectre lumineux de l'acquisition du signal laser : l'intensité (*intensity* en anglais, sans unité dans le format .las et dépendant du signal lumineux (voir la figure 1.15), le numéro de retour (*return number* en anglais), le nombre de retours dans un même faisceau (*number of returns* en anglais), les canaux RGB associés à la sonde (canaux utilisés seulement pour les lasers topographiques et pour lesquels il y a aussi une caméra qui acquiert dans le spectre du visible dans l'avion) et l'attribut classification qui repose sur le spectre lumineux du faisceau mais dont le fonctionnement est limité en bathymétrie à cause de l'absence de signal IR. Il y a également des attributs liés à la géométrie de l'acquisition : la direction du faisceau (*scan direction flag* en anglais, flag binaire qui indique si le faisceau pointe vers l'avant ou l'arrière), l'information de fin de cycle (*edge of flight line* ce bit de données a une valeur de 1 uniquement lorsque la sonde se trouve à la fin d'un balayage (d'un cycle)) et l'angle de balayage (*scan angle rank* en anglais) qui est l'angle (arrondi à l'entier le plus proche) d'émission du point laser, en prenant en compte le roulis de l'aéronef.

| Item | Format | Size | Required |
|---|---|---|---|
| X | long | 4 bytes | yes |
| Y | long | 4 bytes | yes |
| Z | long | 4 bytes | yes |
| Intensity | unsigned short | 2 bytes | no |
| Return Number | 3 bits (bits 0-2) | 3 bits | yes |
| Number of Returns (Given Pulse) | 3 bits (bits 3-5) | 3 bits | yes |
| Scan Direction Flag | 1 bit (bit 6) | 1 bit | yes |
| Edge of Flight Line | 1 bit (bit 7) | 1 bit | yes |
| Classification | unsigned char | 1 byte | yes |
| Scan Angle Rank (-90 to +90) – Left Side | signed char | 1 byte | yes |
| User Data | unsigned char | 1 byte | no |
| Point Source ID | unsigned short | 2 bytes | yes |
| Red | unsigned short | 2 bytes | yes |
| Green | unsigned short | 2 bytes | yes |
| Blue | unsigned short | 2 bytes | yes |
| *Minimum PDRF Size* | | *26 bytes* | |

FIGURE 3.2 – Attributs disponibles dans la donnée .las et leur type associé (pour le format 7 utilisé au Shom), extrait de [137].

Nous avons alors réalisé des analyses univariées et bivariées sur la donnée acquise avec un LiDAR Leica HawkEye 4X, déployant un triple laser : laser PIR

pour la topographie (1TD), laser vert pour les eaux peu profondes 0-10 mètres (*shallow* - 2HD) et laser vert pour les eaux plus profondes supérieures à 10 mètres (*deep* - 3HD). Cette analyse correspond à la recherche de corrélation linéaire entre deux variables, voir la figure 3.3. Sur la diagonale, nous observons la distribution de la variable d'intérêt, selon la validité de la sonde (considérée comme de la bathymétrie ou comme un *outlier*). Pour l'ensemble ce chapitre, nous nous sommes intéressés uniquement aux lasers portant de l'information bathymétrique (lasers 2HD et 3HD).

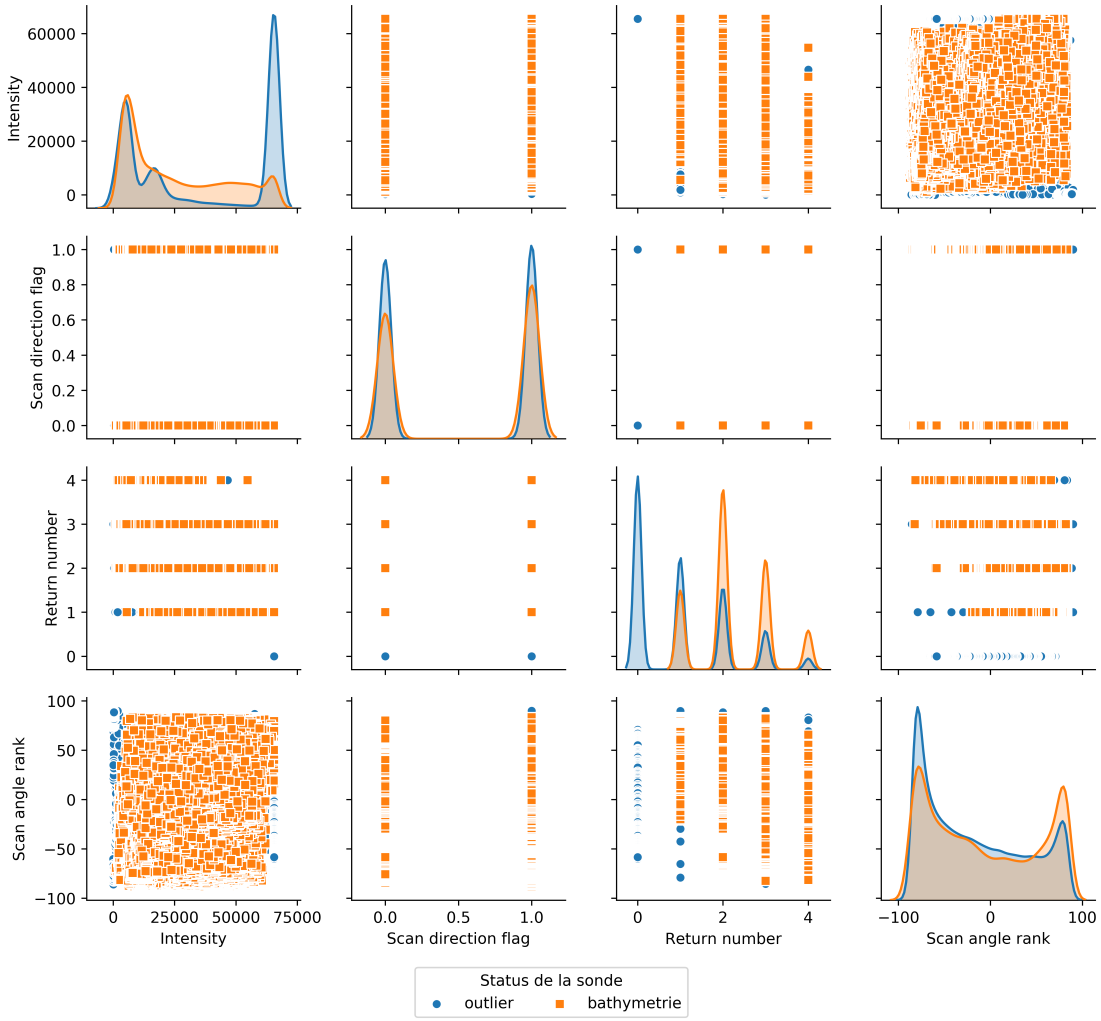


FIGURE 3.3 – Corrélation des variables (*intensity*, *scan direction flag*, *return number* et *scan angle rank*) deux à deux selon la validité de la sonde (bathymétrie ou *outlier*).

La corrélation des variables ne permet pas d'observer de discrimination linéaire

simple par binôme de variables. Néanmoins, nous remarquons que la valeur 0 pour l'attribut *return number* permet de détecter facilement des données aberrantes. Cette valeur correspond au premier retour. Nous l'associons souvent à la surface d'eau. Malheureusement, nous ne pouvons généraliser cet attribut sur l'ensemble d'un levé LiDAR car au niveau des interfaces terre-mer, qui sont les plus complexes à traiter (plage, rocher affleurant...), le premier retour n'est pas nécessairement de la donnée aberrante.

Au vu de la volumétrie importante des données acquises par le capteur LiDAR bathymétrique et de la méthode que nous voulons mettre en place, nous proposons dans la prochaine section une structuration géospatiale permettant de tester des algorithmes ML afin de prédire la bathymétrie.

## 3.2 Méthodologie mise en place pour la détection des *outliers*

Notre approche estime le fond marin en produisant un MNB à partir d'un algorithme de régression. Nous cherchons à produire une surface bathymétrique à partir d'un nuage de points en s'appuyant sur des algorithmes supervisés. Par rapport aux articles présentés dans la section 2.3, nous opérons ce changement de paradigme afin de :

- exploiter les recherches associées au traitement des données matricielles, en s'appuyant sur le traitement classique des images (comme l'article [138]), mais aussi l'apprentissage profond (comme pour les données hyperspectrales dans [139]) qui a considérablement mûri ces dernières années ;
- permettre aux opérateurs de comparer la surface générée avec le nuage de points et utiliser cette surface comme outil de traitement ou de contrôle en fonction de la complexité de la zone considérée, produisant ainsi un outil d'aide à la décision.

La figure 3.4 décrit la démarche méthodologique globale adoptée pour la prédiction d'un MNB et le traitement de la donnée LiDAR bathymétrique mis en place dans le cadre de ce travail de recherche. Le flux proposé présente deux parties : un ensemble d'étapes se réalisant juste après l'acquisition et ne concernant que l'utilisation de processus informatiques (gestion de données, apprentissage si modèle inexistant et prédiction) ; et des étapes où l'humain prend le contrôle du

processus afin d'y apporter son expertise métier.

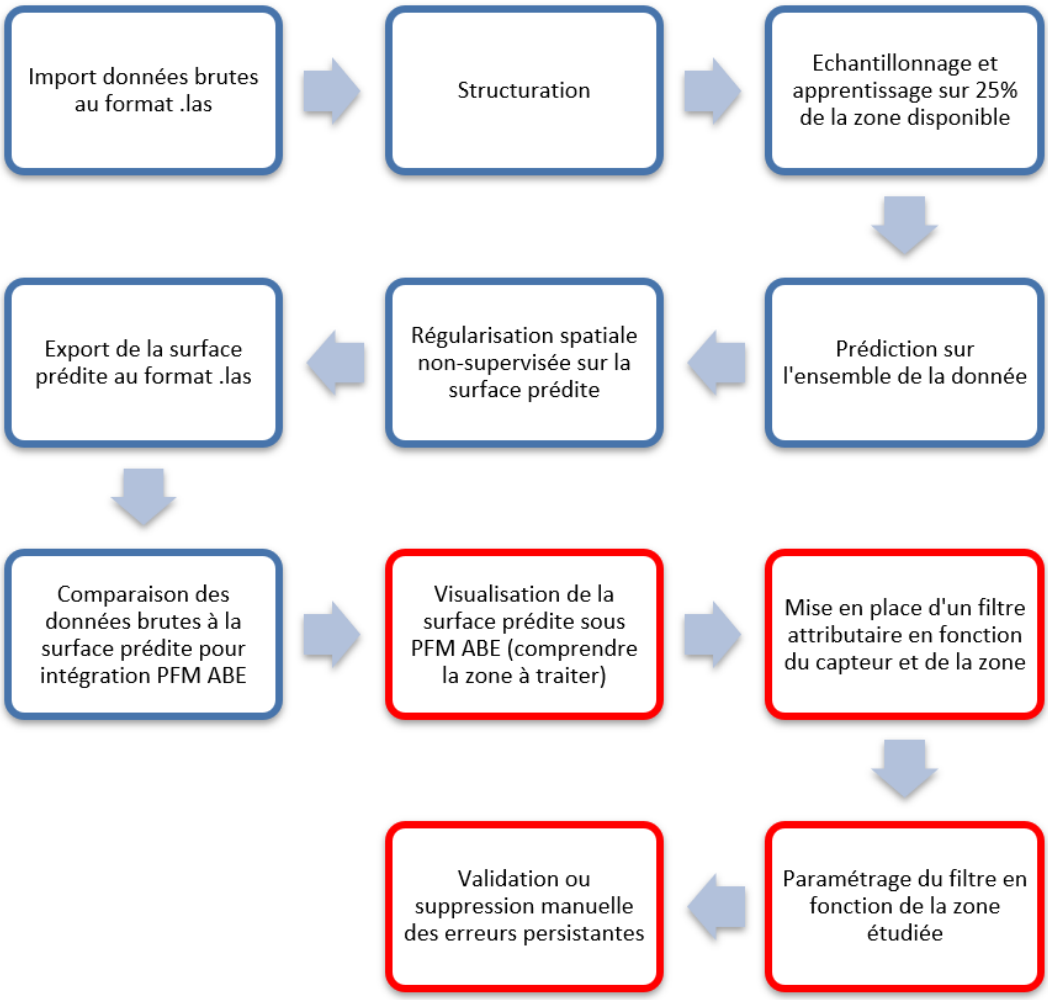


FIGURE 3.4 – Diagramme méthodologique (en bleu les étapes machines, en rouge les étapes opérateurs).

### 3.2.1 Structuration de la donnée et extraction de caractéristiques

Pour effectuer cette régression, notre proposition repose sur une structure de données représentée par la figure 3.5, sous forme d'une grille de voxels en 2D (Nord/Est) + 1D (élévation par rapport à l'ellipsoïde) dans laquelle nous stockons un ensemble de descripteurs. Ce type de structure de données est également présenté dans le cadre de travaux de classification établis sur l'apprentissage profond

à partir de données hyperspectrales, voir [139]. Cette structure permet des résolutions différentes selon les axes verticaux et horizontaux, la densité de points étant très variable dans ces deux dimensions, ce qui donne lieu à une meilleure prise en compte de la variabilité des données d'entrée. Ces deux résolutions sont des hyperparamètres de notre modèle.

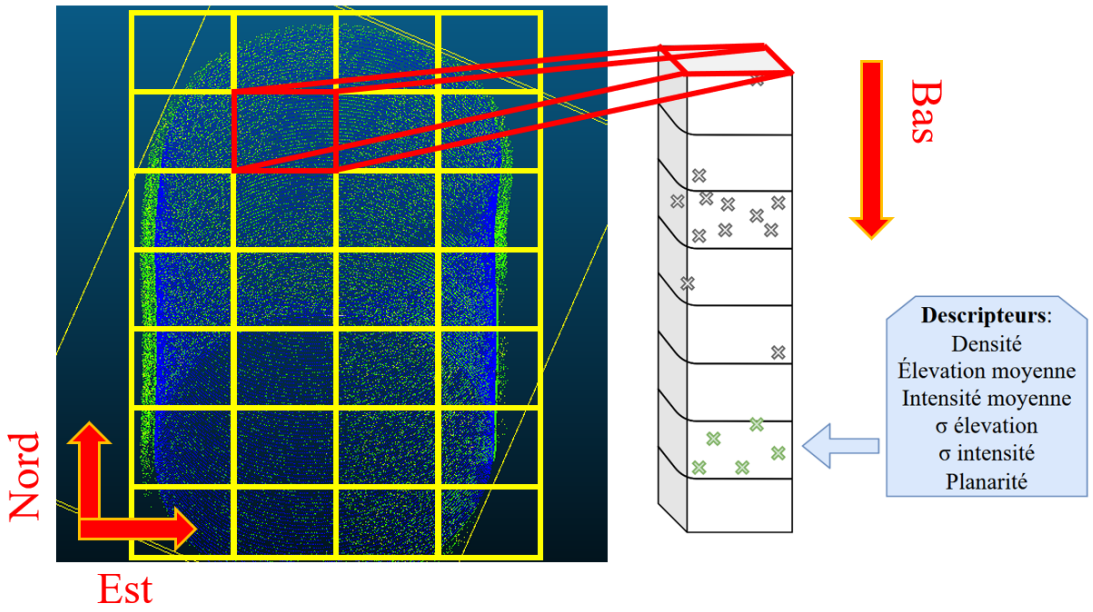


FIGURE 3.5 – Structure de données et descripteurs construits pour la méthode proposée.

Dans le cadre de cette structuration, les descripteurs actuellement considérés et calculés pour chaque voxel composé de $n$ sondes (avec élévation $z$ et intensité $I$) sont les suivants :

— Densité non normalisée $d$ de sondes par voxel :

$$d = n \tag{3.1}$$

— Elévation moyenne $\bar{z}$ :

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i \tag{3.2}$$

— Ecart-type $\sigma_z$ de l'élévation :

$$\sigma_z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3.3}$$

— Moyenne de l'intensité du signal de retour LiDAR $\bar{I}$ :

$$\bar{I} = \frac{1}{n}\sum_{i=1}^{n} I_i \tag{3.4}$$

— Ecart-type $\sigma_I$ de l'intensité du signal de retour LiDAR :

$$\sigma_I = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_i - \bar{I})^2} \tag{3.5}$$

— Aplatissement de l'ellipsoïde associé à la matrice de covariance des sondes $P_\lambda$, comme calculé dans [107] :

$$P_\lambda = \frac{\lambda_2 - \lambda_3}{\lambda_1} \tag{3.6}$$

Où $\lambda_1$, $\lambda_2$ et $\lambda_3$ sont les valeurs propres calculées sur le nuage de points présent dans le voxel étudié (matrice de covariance associée). L'aplatissement capture la forme du nuage de points et permet ainsi de décrire la rugosité du voxel. Cette dernière pourrait être très discriminante entre la surface de l'eau et le fond marin par exemple.

### 3.2.2 Mise en place des régresseurs et premiers apprentissages

Cette structure de données alimente un prédicteur, dont le rôle est de prédire l'élévation du fond marin pour chaque colonne de voxels. Afin de construire notre référence (*baseline* en anglais), nous avons testé différentes méthodes de régression :

— la régression par vecteur de support (SVR) [140] ;

— les forêts aléatoires (RF) [141] ;

— le perceptron multicouche (*MLP* en anglais) [142].

Ces méthodes ont été choisies car elles sont très classiquement utilisées dans le ML. Elles sont donc bien documentées et facilement intégrables dans des chaînes

de traitement. Ce sont de plus des méthodes d'interprétabilité décroissante [143], critère crucial pour faciliter la compréhension et la confiance des opérateurs dans une chaîne de traitement.

La méthode SVR [140] est considérée comme une technique non paramétrique car elle s'appuie sur des fonctions à noyaux (changement d'espace vers une dimension supérieure pour trouver une séparation linéaire). Elle vise à réduire l'erreur en déterminant l'hyperplan (d'équation $< w, x > +b = 0$ avec $w$ un vecteur unitaire dans le cas de la classification) et en minimisant la différence entre les valeurs prédites et observées. Elle essaie d'ajuster la meilleure ligne dans les limites d'une valeur seuil (distance entre l'hyperplan et la ligne frontière). Cela revient à chercher $w \in \mathbb{R}^p$ et $b \in \mathbb{R}$ tels que :

$$|< w, x_i > +b - y_i| \leq \epsilon \tag{3.7}$$

Avec $\epsilon > 0$ petit à choisir par l'utilisateur, on détermine ainsi $(w, b)$ qui minimise $\frac{1}{2}||w||^2$ sous les contraintes $|y_i - w, x_i > -b| \leq \epsilon$, $i = 1, ..., n$.

Pour une régression non linéaire, la fonction à noyau transforme les données en une dimension supérieure et jusqu'à pouvoir effectuer une séparation linéaire. Dans le cadre de nos essais, nous avons utilisé le noyau gaussien $K(x, y)$ (*Radial Basis Function (RBF)* dans [130]), voir l'équation 3.2.2.

$$\begin{aligned} K : X \times X &\to \mathbb{R} \\ K(x, x') = \exp(-\gamma||x - x'||^2) &\text{ avec } \gamma > 0 \end{aligned} \tag{3.8}$$

Les forêts aléatoires (RF) [141] font partie des techniques bien éprouvées en apprentissage supervisé. Cet algorithme combine les concepts de sous-espaces aléatoires et de *bagging* (pour *bootstrap aggregating*). Le *bagging* est une méthode ensembliste conçue pour améliorer la stabilité et la précision des algorithmes d'apprentissage automatique, à travers la génération d'échantillons *bootstrap* depuis l'ensemble d'apprentissage de départ. L'algorithme des forêts d'arbres décisionnels effectue ainsi un apprentissage via de multiples arbres de décision entraînés par l'intermédiaire de sous-ensembles de données légèrement différents. La régres-

sion par forêt aléatoire moyenne ensuite toutes les prédictions pour générer une meilleure estimation de la vérité terrain attendue à partir d'un vote majoritaire. La figure 3.6 présente ce principe de création de multiples arbres de décision pour réaliser une classification ainsi que le processus de vote majoritaire pour prédire une classe finale.

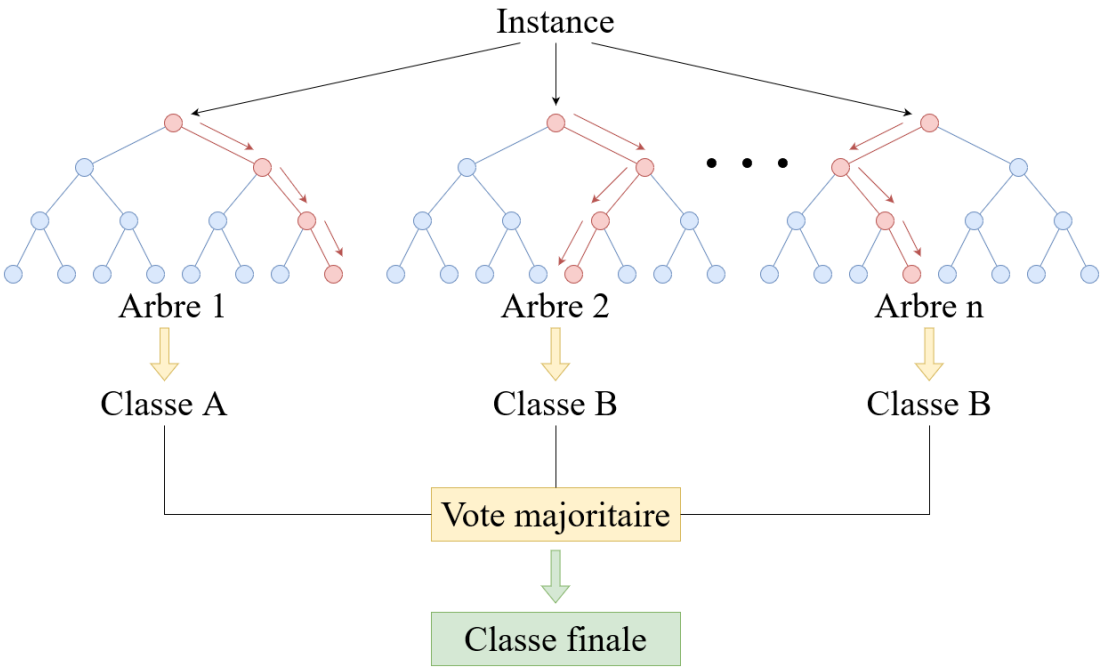


FIGURE 3.6 – Illustration de l'algorithme RF dans le cas d'une classification binaire (classe A ou B).

Le perceptron multicouche (MLP) [142] est un type de réseau neuronal artificiel [144] organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement. Les réseaux de neurones [144] sont des ensembles organisés de neurones artificiels interconnectés qui effectuent des opérations (aussi appelées traitement élémentaires) de combinaisons linéaires afin de résoudre des problèmes complexes à l'aide d'un mode d'apprentissage débouchant sur une forme d'IA dite faible. Cet ensemble de couches, qui constitue le MLP, est appelé architecture. Sa complexité varie en fonction du problème à résoudre et du nombre de descripteurs à construire alors artificiellement. Ainsi, nous parlons d'architecture ou apprentissage profond (Deep Learning (DL)) comportant des couches cachées (*hidden layers* en anglais) pour les réseaux complexes.

La figure 3.7 présente une architecture simple d'un MLP comportant 3 neurones en entrée, 5 neurones dans la couche cachée et un neurone afin de réaliser une prédiction (cette dernière étant dans ce cas une régression, la dernière couche ne comportant qu'un neurone). De façon générale, un réseau totalement interconnecté estime $n \times m$ poids avec $n$ le nombre d'entrées et $m$ le nombre de neurones cachées ; le réseau estimera 15 poids pour le cas de l'architecture de la figure 3.7.

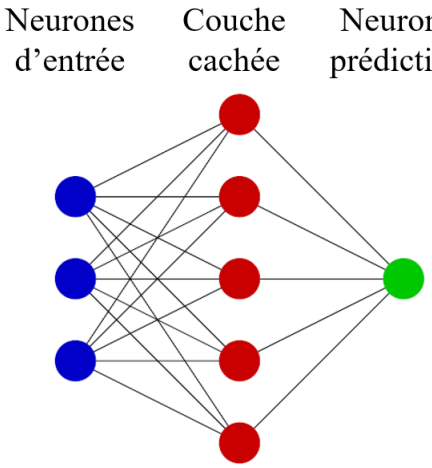


FIGURE 3.7 – Architecture de MLP à trois couches.

Dans le cadre de notre étude, nous avons construit une architecture du MLP simple se composant de quatre couches de neurones, avec des connexions complètes entre les couches. La première couche (entrée) contient 128 neurones, la deuxième couche (intermédiaire) contient 64 neurones, la troisième couche contient 8 neurones et la dernière couche (sortie) contient 1 neurone (la prédiction du fond). Les neurones des trois premières couches ont une fonction d'activation ReLU (*Rectified Linear Unit*) de la forme suivante :

$$f(x) = \begin{cases} 0, \text{ si } x \leq 0 \\ x, \text{ autrement.} \end{cases} \tag{3.9}$$

L'objectif de la fonction d'activation est d'ajouter de la non-linéarité. L'activation est ainsi réalisée après chaque combinaison linéaire. Le neurone de la couche finale possède une fonction d'activation linéaire (fonction identité $f(x) = x$) afin d'effectuer la régression. En utilisant la couche linéaire à la sortie d'un MLP,

le processus de régression permet d'obtenir la valeur prédite. Le modèle a été entraîné à l'aide de l'algorithme de rétropropagation des erreurs avec un taux d'apprentissage de $\rho = 0.01$ (choisi arbitrairement, hyperparamètre à optimiser). La fonction de coût du modèle est l'erreur absolue moyenne (*Mean Absolute Error (MAE)*), métrique communément utilisée en régression [145] et l'optimiseur est l'algorithme d'optimisation d'Adam (raccourci pour *Adaptive moment estimation*). L'algorithme d'optimisation Adam [146] est utilisé pour la formation de modèles d'apprentissage profond. Il s'agit d'une extension de la descente de gradient stochastique [147] utilisée pour calculer les taux d'apprentissage adaptatifs pour chaque paramètre. Classiquement, la fonction de coût (*loss* en anglais) pour la MAE se calcule avec la formule 3.10 suivante, avec $y$ la valeur vraie et $\hat{y}$ la valeur prédite pour $N$ échantillons :

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} |y - \hat{y}_i| \tag{3.10}$$

Pour les premiers essais de prédiction réalisés, nous avons utilisé un jeu de données d'environ 500 000 sondes (environ 0.10 km$^2$) pour lesquelles ont été générées les caractéristiques décrites dans la section précédente. Ce jeu de données a été acquis en Corse en septembre 2018 dans le cadre du projet Shom Litto3D®, voir [148].

Pour notre approche, la vérité terrain a été construite comme l'élévation moyenne des sondes considérées comme valides par un opérateur dans une colonne d'eau. Cette vérité terrain est donc très inégalement distribuée en termes de représentation de l'élévation car elle est fonction de la morphologie des fonds marins. Afin de rendre les données d'entraînement non-biaisées (pour ne pas sur-représenter une gamme de profondeurs par exemple), nous avons mis en place la méthode suivante d'échantillonnage par histogramme.

Supposons qu'il existe $k$ bandes (plage de profondeurs), calculées avec la formule de Doane [149]. Soit $n_i$ le nombre attribué à la $i^{\text{eme}}$ bande, et $a_i$ et $b_i$ ses frontières (gauche et droite). Nous avons calculé la taille minimale ($min_{bin-size}$) et la taille moyenne des bandes ($mean_{bin-size}$) à partir de la fréquence globale des k bandes ; ensuite, $c_i$, qui est le nombre de données échantillonnées au hasard dans chaque bande, est calculé à l'aide de l'équation 3.11 :

$$c_i = \begin{cases} rand(mean_{bin-size_i}), \text{ si } bin_{freq_i} > \min_{bin-size_i}, \\ 0, \text{ si } bin_{freq_i} < 50 \end{cases} \tag{3.11}$$

Pour la modélisation, différentes résolutions horizontales et verticales des voxels (hyperparamètre de la structuration) ont été prises en compte afin de considérer les différentes densités de nuages de points. Le compromis est de viser la résolution la plus détaillée possible (afin de mieux décrire le fond marin en restant cohérent avec l'empreinte au sol du faisceau) tout en conservant suffisamment d'informations pour que la régression soit statistiquement pertinente. Il faut également prendre en considération les temps de calcul et les contraintes de mémoire compatibles avec la puissance de calcul mise à notre disposition dans le cadre du projet.

La figure 3.8 illustre un exemple des essais de paramétrages pour les hyperparamètres suivant la résolution horizontale et verticale. Elle montre ainsi que la MAE du modèle SVR tend à être plus grande par rapport aux modèles RF et MLP pour des résolutions horizontales comprises entre 2 et 20 mètres et pour une résolution verticale de 0.50 mètres. Le SVR n'a pas été calculé pour 2 et 5 mètres de résolution horizontale car étant toujours plus mauvais (avec les jeux de paramètres utilisés au vu de la métrique globale MAE et des jeux de données testés), il a donc rapidement été abandonné. De la même façon, nous remarquons que le modèle MLP est toujours moins bon que le modèle RF pour ces différentes résolutions.

Suite à ces premiers essais sur les trois régresseurs présentés, nous n'avons pas considéré les modèles SVR et MLP en raison de leurs MAE plus élevées, ce qui nous a permis d'économiser du temps de calcul pour le modèle RF qui semblait plus prometteur. Les méthodes SVR et MLP pourront être testées à nouveau (avec plus de caractéristiques ou des noyaux différents, plutôt que le noyau gaussien utilisé pour le SVR par exemple), pour s'assurer qu'elles sont effectivement moins efficaces. Un travail supplémentaire sur l'optimisation des hyperparamètres est à entreprendre, même si les résultats du modèle RF sont déjà prometteurs.

De plus, au cours de l'analyse de ces premiers résultats et des descripteurs associés à la donnée d'entrée, nous avons remarqué que la conservation de toutes les données (et des caractéristiques associées) de tous les capteurs (2HD et 3HD) dans une même structure de données peut rendre confuses les informations enregistrées dans les voxels et finalement rendre certains descripteurs moins discriminants. La

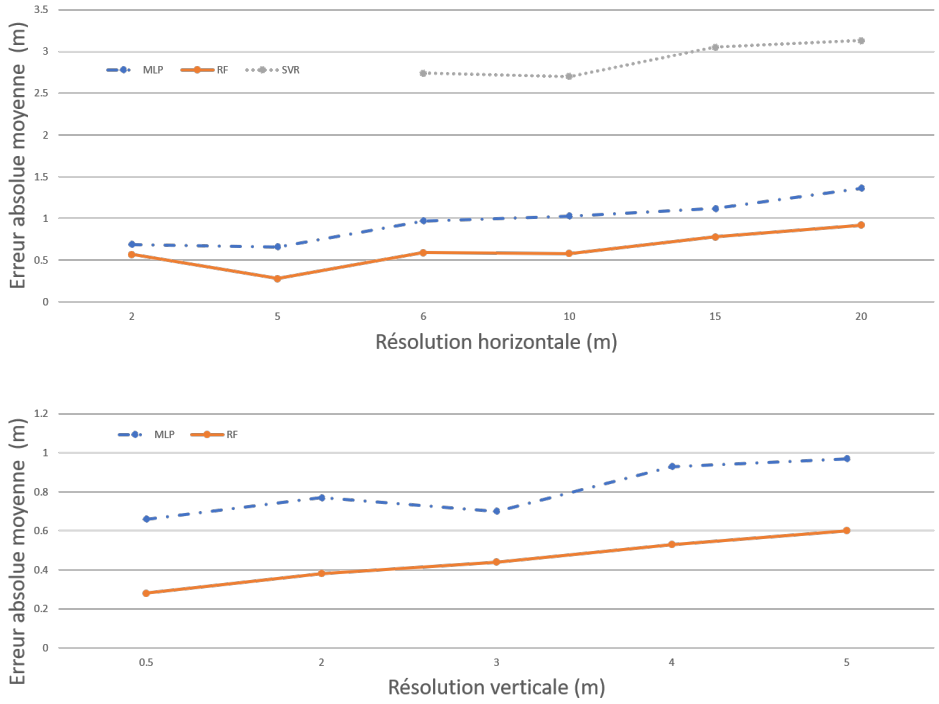


FIGURE 3.8 – Erreurs moyennes absolues en fonction de l'hyperparamètre résolution horizontale (en haut) pour les régresseurs SVR, RF et MLP et en fonction de l'hyperparamètre résolution verticale (en bas) pour les régresseurs RF et MLP.

figure 3.9 montre un exemple de sous-ensemble de données avec le laser 3HD en PIR et le laser 2HD en vert : nous constatons un recouvrement important entre ces deux capteurs.

Ce recouvrement est susceptible de perturber la construction des descripteurs. En effet, la densité de points (2HD et 3HD combinées) dans une zone de recouvrement augmente de manière significative puis diminue drastiquement lorsque le laser 2HD peu profond n'est plus présent. De plus, ce chevauchement peut induire un changement de l'intensité du retour lumineux, comme le montre la figure 3.10. Ce changement d'intensité du signal de retour s'explique par la différence de puissance et d'absorption du signal dans la colonne d'eau pour les lasers profonds et peu profonds, voir [44].

Ainsi, si nous observons d'un point de vue quantitatif les graphiques présentés sur la figure 3.11, nous remarquons que les valeurs d'intensités en fonction de la profondeur on des valeurs différentes pour le capteur 2HD et 3HD. Nous pouvons

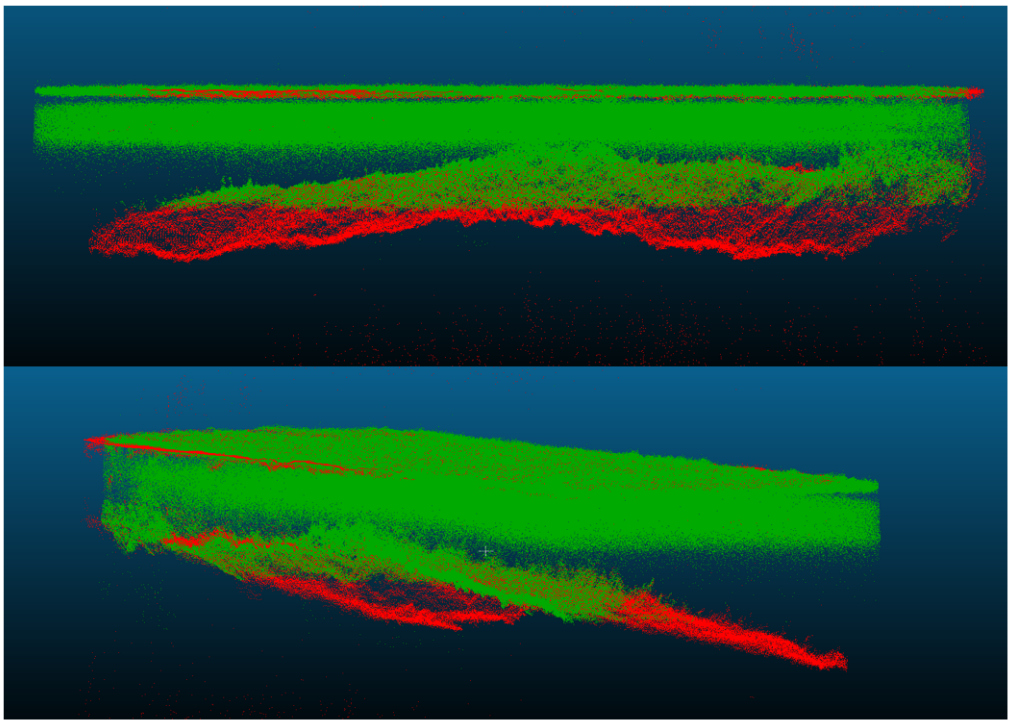

FIGURE 3.9 – Vue de côté et vue en perspective d'un sous-ensemble de données, LiDAR 2HD en vert et LiDAR 3HD en rouge.

également percevoir un phénomène de saturation de la mesure d'intensité qui ne dépasse jamais la valeur de 65530.

Au vu de ces informations sur la densité de points ainsi que les variations d'intensité (liées aux gains à l'émission différents entre les capteurs), nous avons transformé notre structuration pour réaliser une prédiction avec le capteur 2HD et une autre prédiction avec le capteur 3HD. La densité de sondes pour le capteur 3HD étant plus faible que le capteur 2HD, nous avons dù adapter la résolution horizontale de la structuration. Nous avons ainsi formé les couples de résolution horizontale-verticale pour le 2HD (2.0-0.5m) et 3HD (5.0-0.5m). De plus, sur suggestion des opérateurs qui utilisent ce type de procédé manuellement, nous avons séparé la vue avant de la vue arrière (attributs las *scan direction flag*) en considérant des "scènes" d'acquisition (dans le sens où une même zone a été observée à des instants légèrement différés) différentes pendant la trajectoire de vol de l'avion, ce qui permet de combiner ces deux vues pour rendre la prédiction plus robuste (si les deux vues voient une même obstruction par exemple).

Une fois cette nouvelle structuration réalisée, nous avons de nouveau testé

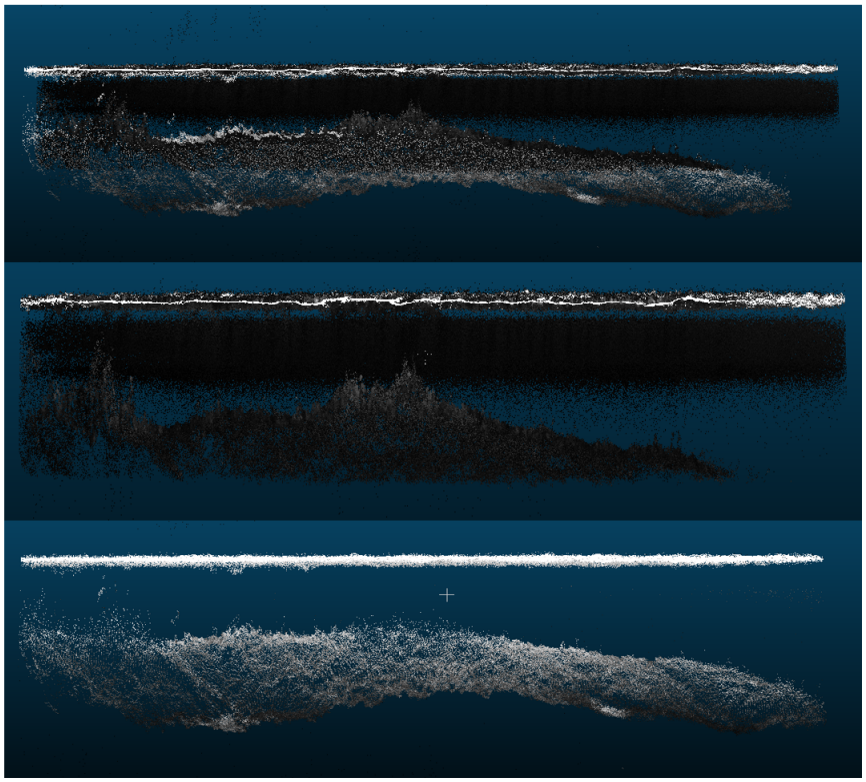

FIGURE 3.10 – En haut, l'intensité pour les capteurs 2HD et 3HD confondus (valeur entière comprise entre 0 et 65530 et propre à l'industriel). Au milieu, l'intensité pour le capteur 2HD seul. En bas, l'intensité pour le capteur 3HD seul. L'échelle du niveau de gris d'intensité est la même pour les deux capteurs.

le jeu de donnée LiDAR bathymétrique Corse. La figure 3.12 montre la vérité terrain (résultat du traitement manuel d'un opérateur en haut à gauche), le jeu de données d'entraînement (en haut à droite) et la différence entre la prédiction et la vérité terrain avec une résolution horizontale de 2m pour le laser vue arrière peu profond (2HD). Dans le cadre de cet apprentissage pour le modèle RF, nous avons réparti la donnée disponible en 70% d'entraînement (*train* en anglais) et 30% test. Pour cette modélisation, les conditions d'apprentissage sont optimales car un échantillonnage aléatoire a été effectué. Il sera intéressant à l'avenir de varier cette stratégie d'apprentissage pour voir comment le modèle est capable de se généraliser. Nous notons que pour le capteur peu profond, la MAE de ce jeu de données est de 0,09m et l'écart-type associé à cette métrique est de 0.26 cm, ce qui est comparable à la sensibilité du capteur LiDAR 2HD (entre 0,10m et 0,30m). La prédiction est donc globalement de bonne qualité au vu de la métrique MAE.

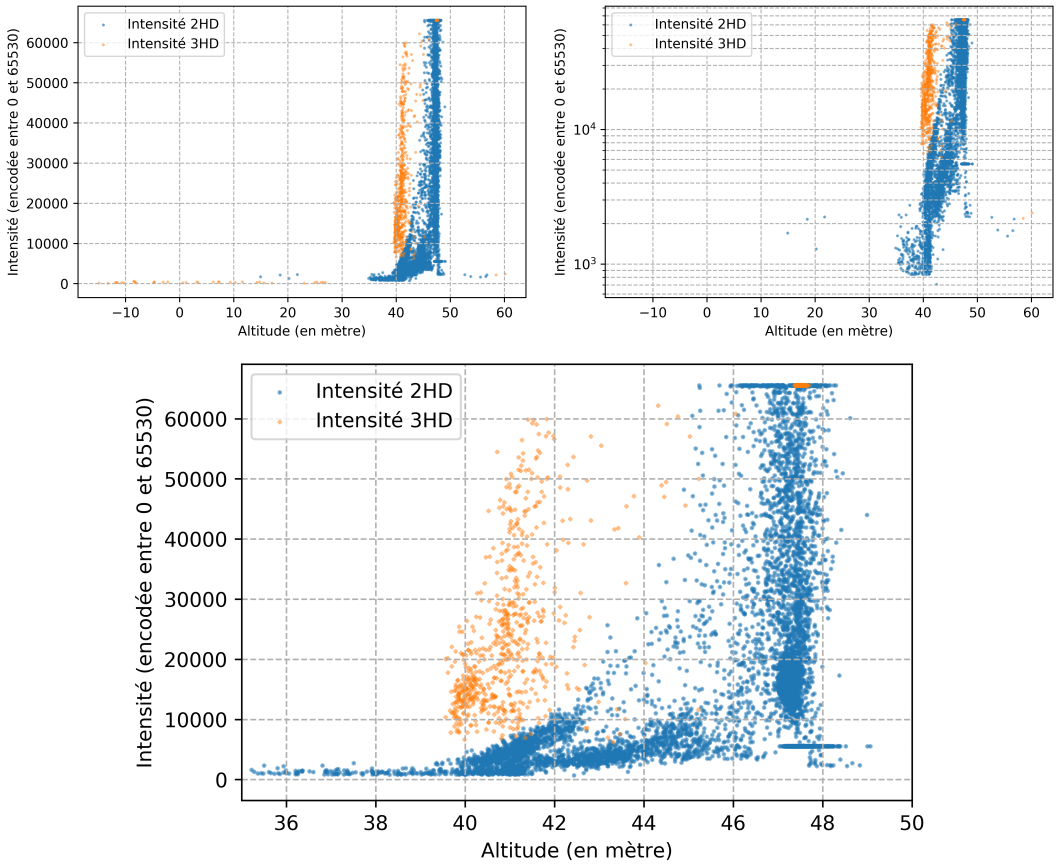


FIGURE 3.11 – Intensité en fonction de l'altitude (référence verticale au moment de l'acquisition, ramené ensuite à une profondeur) pour la zone des figures 3.9 et 3.10 (capteur 2HD en bleu et 3HD en orange), en haut à gauche en échelle linéaire, en haut à droite en échelle logarithmique et en bas zoom sous la surface d'eau.

La figure 3.13 montre la vérité terrain (résultat du traitement manuel d'un opérateur en haut à gauche), le jeu de données d'entraînement (en haut à droite) et la différence entre la prédiction et la vérité terrain avec une résolution horizontale de 2m pour le laser vue arrière profond (3HD). Pour le capteur 3HD, la MAE est de 0,21m et l'écart-type associé à cette métrique est de 1.36m, ce qui est également comparable à la sensibilité du capteur LiDAR 3HD (entre 0,30m et 0,50m). L'écart-type plus élevé que le capteur 2HD s'explique par la présence de différences sur les limites du domaine de modélisation. De plus, il faut prendre en compte un biais dû à l'échantillonnage plus grand de la donnée (résolution horizontale de 5m pour le capteur 3HD contre une résolution horizontale de 2m pour le capteur 2HD). La prédiction est donc globalement de bonne qualité au vu de la métrique

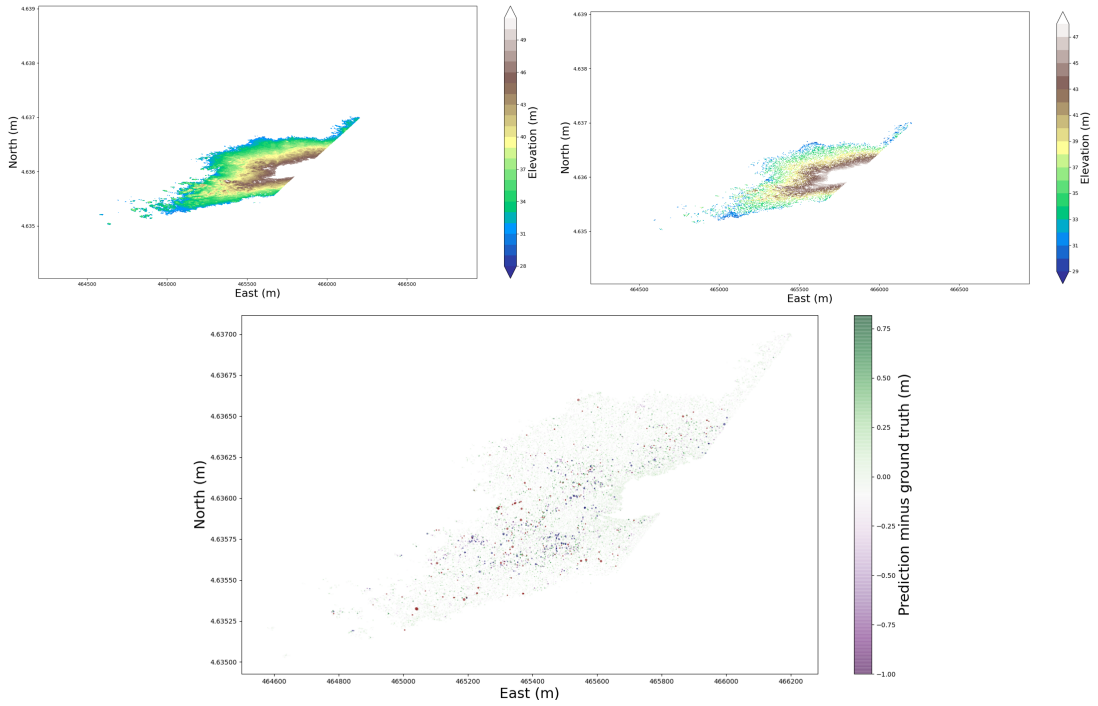


FIGURE 3.12 – *Ground truth* (en haut à gauche), échantillon d'entraînement (en haut à droite) pour une résolution horizontale de 2m pour le LiDAR vue arrière 2HD. Différence entre la vérité terrain et la prédiction (en bas) pour l'ensemble de données sur les échantillons de test.

MAE mais perfectible car il subsiste encore quelques erreurs localisées autour des zones où de petits artefacts morphologiques (comparés à la taille des pixels de la grille) sont présents, ou lorsque peu d'échantillons d'entraînement sont disponibles (typiquement dans les zones les moins profondes), ou près de la limite du domaine (où les caractéristiques calculées sont perturbées par la présence de la limite).

Les figures 3.14 et 3.15 mettent en évidence une corrélation entre le signe des erreurs et la profondeur prédite. En effet, nous observons une corrélation négative entre l'erreur de prédiction signée et l'élévation de la vérité terrain. Ce comportement est en désaccord avec la sécurité de la navigation, pour laquelle il est important de ne pas sous-estimer l'élévation dans les zones peu profondes.

Nous notons ainsi une erreur moyenne maximale en dessous de 3cm pour le capteur 2HD (très faible au vu de la sensibilité du capteur 2HD comprise entre 10cm et 30cm) et une erreur moyenne maximale aux alentours de 65cm pour le capteur 3HD dans la gamme de profondeurs 2-3 mètres (qui s'expliquerait dans ce cas par le faible nombre d'échantillons sur cette gamme de profondeurs). La

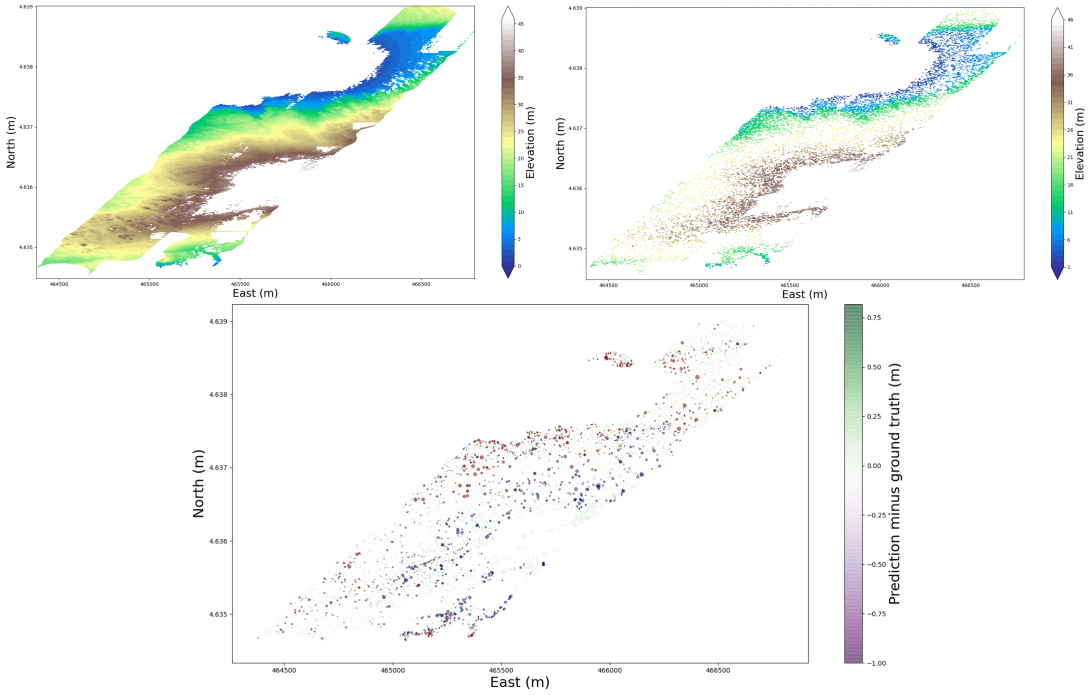


FIGURE 3.13 – *Ground truth* (en haut à gauche), échantillon d'entraînement (en haut à droite) pour une résolution horizontale de 5m pour le LiDAR vue arrière 3HD. Différence entre la vérité terrain et la prédiction (en bas) pour l'ensemble de données sur les échantillons de test.

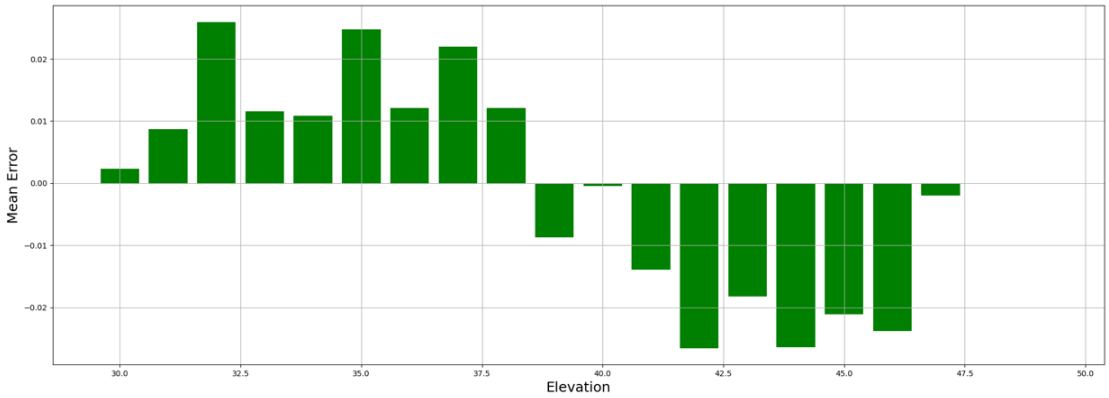


FIGURE 3.14 – Différence moyenne entre la vérité terrain et l'erreur de prédiction par plage d'élévation pour la donnée Corse pour le capteur 2HD.

source de ce biais devra être mis en évidence dans de futures recherches afin de le supprimer et de contraindre le modèle à privilégier des prédictions allant dans le sens de la sécurité de la navigation.

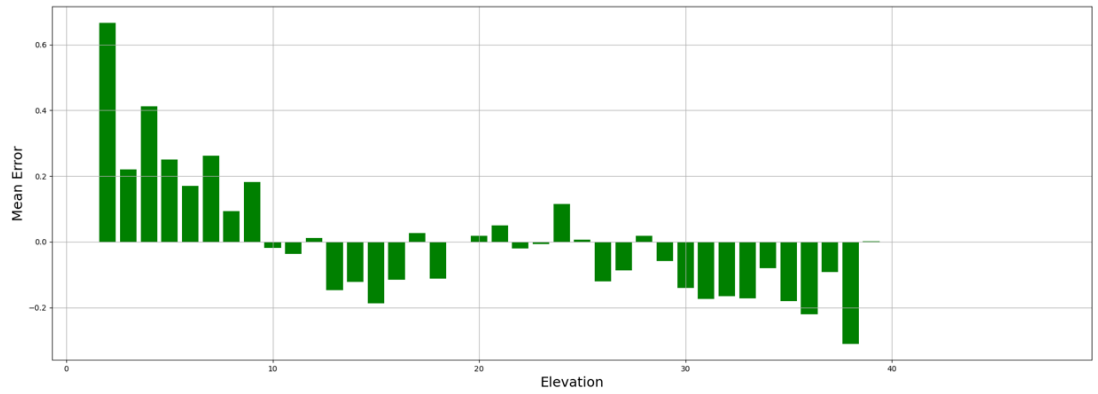


FIGURE 3.15 – Différence moyenne entre la vérité terrain et l'erreur de prédiction par plage d'élévation pour la donnée Corse pour le capteur 3HD.

### 3.2.3 Régularisation spatiale et intégration dans le logiciel de traitement PFM ABE

La figure 3.13 montre encore quelques erreurs localisées notamment en bordure des zones. Ces erreurs sont parfois isolées (une seule cellule concernée). Afin de détecter les cellules prédites du MNB les moins cohérentes avec leurs voisins, nous avons mis en place un calcul du score de l'écart absolu de la médiane (MAD en anglais), voir [129]. Cette méthode statistique réalise une mesure robuste de la variabilité d'un échantillon univarié de données quantitatives.

Pour un ensemble de données univariées $X_1, X_2, ..., X_n$, le MAD est défini comme la médiane des écarts absolus par rapport à la médiane des données $\tilde{X} = median(X)$. Ainsi, cette distance permet de détecter les pixels qui ont une valeur trop différente de leur voisinage local. Dans notre cas, le calcul du MAD se fait sur un ensemble de données d'élévation Z $= (z_1, z_2, ..., z_n)$ dans un rayon (voisinage horizontal) de 5 mètres.

$$MAD_i = 1.4826 * median(|z_i - \tilde{Z}|) \tag{3.12}$$

On calcule ainsi une $MAD_{distance_i}$ pour chaque sonde $z_i$ grâce à 3.13 et qui décrit la différence entre l'échantillon et sa médiane normalisée par le score MAD de l'échantillon :

$$MAD_{distance_i} = \frac{|z_i - \tilde{Z}|}{MAD_i} \tag{3.13}$$

L'équation 3.14 nous permet ainsi de détecter les prédictions aberrantes en fonction d'une distance maximale admissible ($MAD_{max}$) et de les remplacer par la valeur médiane du voisinage.

$$z_i = \begin{cases} z_i, \text{ si MAD\_distance\_i} < \text{MAD\_max}, \\ \tilde{Z}(\text{z\_i exclu}), \text{ sinon.} \end{cases} \tag{3.14}$$

Cette méthode est pertinente pour la détection des pixels isolés mais ne parvient pas à détecter des groupes de pixels incohérents. Un filtre médian devrait fournir le même type de résultats mais n'a pas été testé au cours de cette étude. Une fois détectés, ces pixels jugés comme trop différents de leur voisinage sont remplacés par la valeur médiane de ce voisinage. Des tests supplémentaires sont à réaliser sur le paramétrage, notamment la valeur de seuil et le voisinage d'étude (possibilité de mettre en place une méthode supervisée pour la détermination de ces paramètres) mais aussi sur la manière d'ajuster le pixel considéré comme incohérent (valeur moyenne, médiane, pondérée par la distance...). Dans le cadre de l'expérimentation, cette méthode a donné de très bons résultats (voir la table 3.2 avant et après régularisation) et a donc été proposée aux opérateurs.

Pour réaliser cette régularisation, nous avons également testé un modèle DL basé sur une architecture U-Net [125]. Cette architecture est utilisée dans le cadre de l'estimation de la profondeur bathymétrique à partir de données multispectrales dans l'article [150] avec le réseau BathyNet mais également dans certaines méthodes présentées dans la section 2.3 comme [106]. Afin de tester cette méthode, nous avons transformé les données prédites en images au format *raster* comme présenté sur la figure 3.16.

Le réseau construit a été entraîné et testé sur des imagettes de 256x256 pixels issues des prédictions réalisées par le modèle RF. Malheureusement, ce réseau n'a jamais fourni de résultat convergent pertinent. Il est destructeur en introduisant du bruit après la régression spatiale. La figure 3.17 montre ainsi la différence entre la prédiction réalisée par le réseau BathyNet appliquée à notre donnée et la vérité terrain.

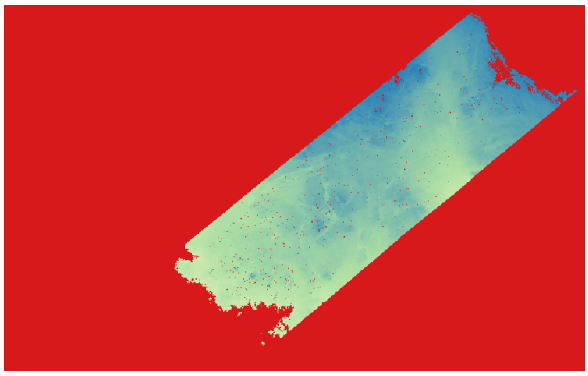

FIGURE 3.16 – Exemple d'image utilisée pour la régularisation spatiale supervisée (en rouge les pixels sans valeur).

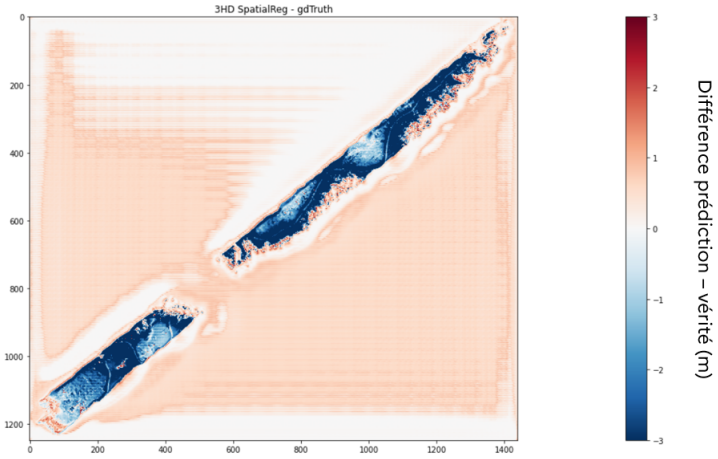


FIGURE 3.17 – Résultat du réseau BathyNet pour la régularisation spatiale suite à une prédiction.

Une des raisons vraisemblables à l'origine de ce phénomène destructeur serait la présence de nombreuses zones sans prédiction lors du découpage en imagettes. Effectivement, assurer pour chaque imagette une couverture de la zone d'au minimum 70% ([150]) avec de la donnée bathymétrique, nous oblige à ne conserver qu'une partie des données disponibles. Un travail de structuration de la donnée et de la génération de données d'entraînement (avec potentiellement de l'augmentation de données à réaliser) permettrait d'améliorer cette méthode et ainsi connaître le nombre d'imagettes nécessaires à une bonne convergence.

Néanmoins, au vu des résultats déjà très intéressants proposés par la régularisation spatiale non-supervisée, voir 3.2, il nous paraît beaucoup plus pertinent de

continuer dans la voie de l'hybridation entre prédiction supervisée et régularisation spatiale non-supervisée.

Une fois la prédiction et la régularisation réalisées, l'objectif est de visualiser cette information dans le logiciel de traitement de données utilisé par le département AL : PFM ABE. La prédiction est donc écrite dans le format .las supporté par PFM ABE. Elle est vue comme une source supplémentaire de données accessibles par l'opérateur et comme un outil de contrôle et d'aide à la décision. Il est conseillé de visualiser la prédiction avant de commencer le traitement pour distinguer les zones où la régression semble pertinente et les zones où un contrôle manuel semble nécessaire. Cette information permet également d'avoir une idée de la morphologie globale de la zone à traiter (zone de dune, roches complexes...).

De plus, afin d'aider les opérateurs sur la phase de traitement des données et de suppression du bruit, un indice d'appartenance à la surface prédite a été calculé pour chaque mesure. Cet indice entier ($In$) est calculé comme une distance verticale entre l'élévation de la sonde ($z$) et la valeur de prédiction la plus proche ($z_{pred}$), cette distance étant normalisée entre 0 et 200. Cet indice centré indique 100 pour si $z = z_{pred}$. La valeur d'indice 100 est ainsi la valeur considérée comme le fond pour le prédicteur (valeur optimale), voir équation 3.15 :

$$\begin{aligned} In &= 100 + (z - z_{pred}) \times 2 \\ In &= \begin{cases} 200, \text{ si } In \geq 200 \\ 0, \text{ si } In \leq 0 \\ In \text{ sinon.} \end{cases} \end{aligned} \tag{3.15}$$

L'indice remplace ensuite un des attributs disponibles dans le format .las (voir la figure 3.2), inutilisé dans la procédure actuelle d'AL. Le choix s'est porté sur l'attribut Red pour la prédiction pour le capteur 2HD, et l'attribut Green pour le capteur 3HD. Dans le futur, une intégration dans un attribut spécifique permettrait d'encoder plus facilement cette valeur (sans contrainte sur le typage et la taille de l'attribut).

Pour utiliser cet indice, les opérateurs appliquent un filtre attributaire sur l'un des deux attributs générés, en fonction du type de capteur à privilégier ou à utiliser de façon séquentielle (Green puis Red par exemple). Ce filtre est nativement intégré dans PFM ABE et permet pour notre utilisation de détecter plus facilement le fond.

Sur la figure 3.18, nous visualisons l'utilisation de ce filtre sur l'attribut Red pour un jeu de données LiDAR dans lequel l'indice 100 est sélectionné.

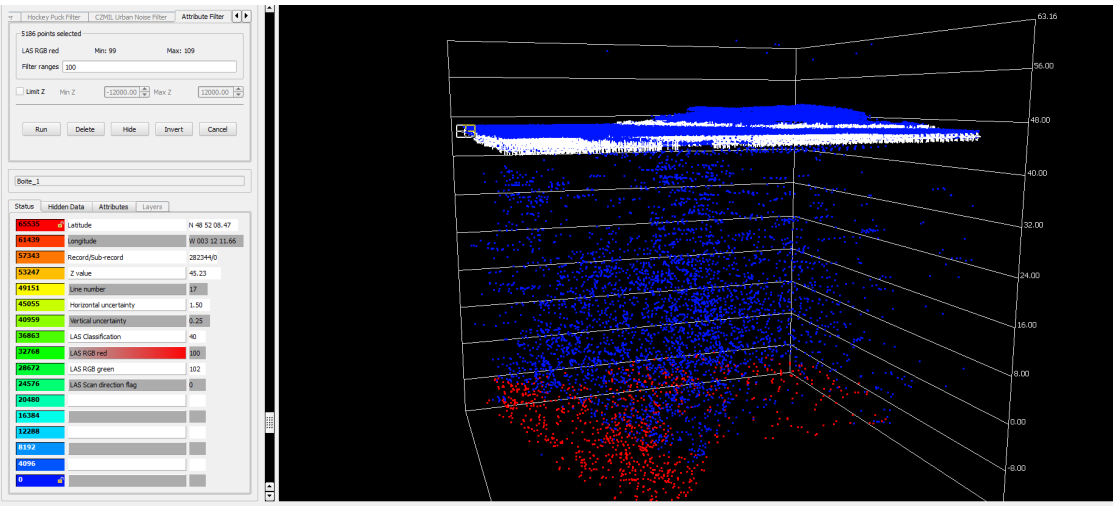


FIGURE 3.18 – Filtre sur l'attribut RGB du format .las dans l'outil de traitement LiDAR PFM ABE (points blancs indice 100, points rouges et bleus les autres indices).

## 3.3 Expérimentation réelle sur la donnée LiDAR

Afin de tester notre approche à plus grande échelle, une expérimentation a été mise en place avec le concours du département AL du Shom. Nous avons eu ainsi la chance de travailler avec des données réelles et des opérateurs dans le but de recueillir leur retour quantitatif et qualitatif sur la méthodologie proposée. L'objectif d'un point de vue opérationnel était de s'intégrer le plus naturellement possible dans la chaîne de traitement des données LiDAR bathymétrique du Shom. Les résultats présentés dans cette partie sont donc issus exclusivement de cette courte expérimentation et mériteraient une consolidation avec des nouveaux essais sur des zones différentes. Les objectifs étaient à la fois techniques et méthodologiques :

- s'assurer de la bonne intégration des fichiers .las (produits par la méthodologie présentée dans la section précédente) dans le logiciel de traitement PFM ABE ;
- prise en main des filtres associés à la méthode ML ;

— comparaison quantitative du traitement manuel classique et semi-automatique ML ;
— comparaison qualitative sur la prise en main et la méthode proposée ;
— propositions et perspectives d'améliorations de la méthodologie.

Dans le cadre de cette expérimentation, notre choix s'est porté sur un jeu de données acquis en 2019 par les opérateurs du département AL car nous souhaitions disposer d'une comparaison récente du traitement manuel avec des données traitées sur le Système d'Information (SI) dédié d'AL accessibles. En effet, pour ce dernier point, travailler avec de la donnée archivée nécessiterait davantage de travail préparatoire sur cette donnée, voire de repasser par un traitement manuel de certaines zones.

### 3.3.1 Présentation de la donnée à traiter

Je me suis intéressé à un jeu de données acquis en Bretagne en novembre 2019 dans le cadre du projet Shom Litto3D® (voir [151]), visant à produire un référentiel terre-mer complet pour la France. Ce projet Litto3D® Bretagne est financé par le Shom, l'IGN, la région Breatgne, l'État et l'Europe, voir [152]. Ces données ont été récoltées à bord d'un avion Cessna 208 B Grand Caravan équipé d'un Leica HawkEye 4X, déployant un triple laser : laser PIR (1064 nm) pour la topographie (1TD), laser vert (532 nm) pour les eaux peu profondes 0-10 mètres (*shallow* - 2HD) et laser vert (532 nm) pour les eaux plus profondes supérieures à 10 mètres (*deep* - 3HD). Les données présentées ici sont issues de cette acquisition. Les conditions environnementales étaient bonnes (peu de turbidité, bonne clarté de l'eau) durant cette acquisition permettant d'atteindre 25 mètres de profondeur.

La figure 3.19 montre l'emplacement des trois zones d'études de l'XP TraitLIA : la côte nord Bretagne dans les Côtes d'Armor, la zone Est Paimpol et Plougrescant. La zone 1 correspond à un environnement rocheux avec de la topographie émergente sur une partie de la zone, morphologie complexe à appréhender. La zone 2 quant à elle correspond à un parc à huîtres avec du sursol complexe à traiter, des zones sableuses et des zones rocheuses. Enfin, la zone 3 présente une bathymétrie plus profonde et des dunes de sables : c'est la zone la plus simple à traiter.

Ces trois zones ont été choisies en raison de leurs caractéristiques différentes comme présentées dans la table 3.1.

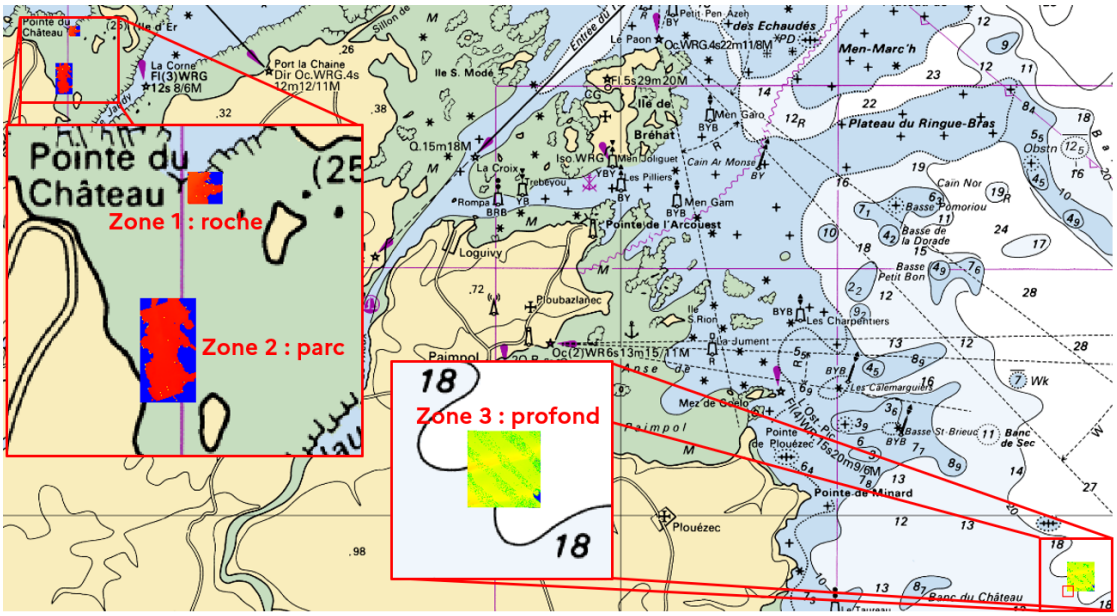


FIGURE 3.19 – Zones d'étude de l'expérimentation, extrait de la CM 6930 du Shom.

TABLE 3.1 – Caractéristiques globales des zones étudiées.

| Caractéristiques | Zone 1 : roche | Zone 2 : parc | Zone 3 : profond |
|---|---|---|---|
| Nbre total de points (2HD – 3HD) | 2 2249 857 | 7 852 417 | 18 129 896 |
| Nbres points validés manuellement | 427 648 | 2 224 351 | 738 276 |
| Ratio donnée valide/total | 19% | 28% | 4% |
| Surface (km$^2$) | 0.07 | 0.38 | 2.25 |

La figure 3.20 présente la distribution statistique des données valides pour les différentes zones. Nous constatons que la Z1 est assez rugueuse (histogramme aplati entre 42 et 47 mètres avec une élévation moyenne de 43.7 mètres et un écart-type de 1.6 mètres) alors que la Z2 est beaucoup plus concentrée (écart-type de 0.45 mètre et une élévation moyenne de 47.1 mètres). La Z3 présente quant à elle un écart-type de 1.2 mètres pour 21.3 mètres d'élévation moyenne.

La figure 3.21 présente quant à elle les histogrammes pour les données invalides sur les différentes zones. Nous remarquons que les données aberrantes sont présentes sur l'ensemble de la colonne d'eau.

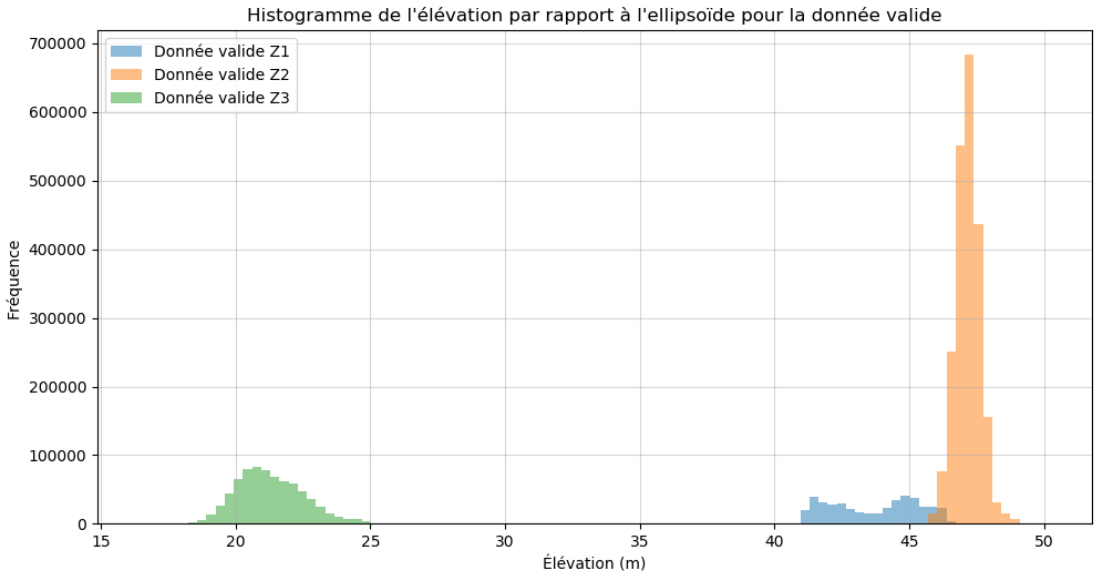


FIGURE 3.20 – Histogramme de l'élévation par rapport à l'ellipsoïde pour la donnée valide sur les 3 zones d'étude.

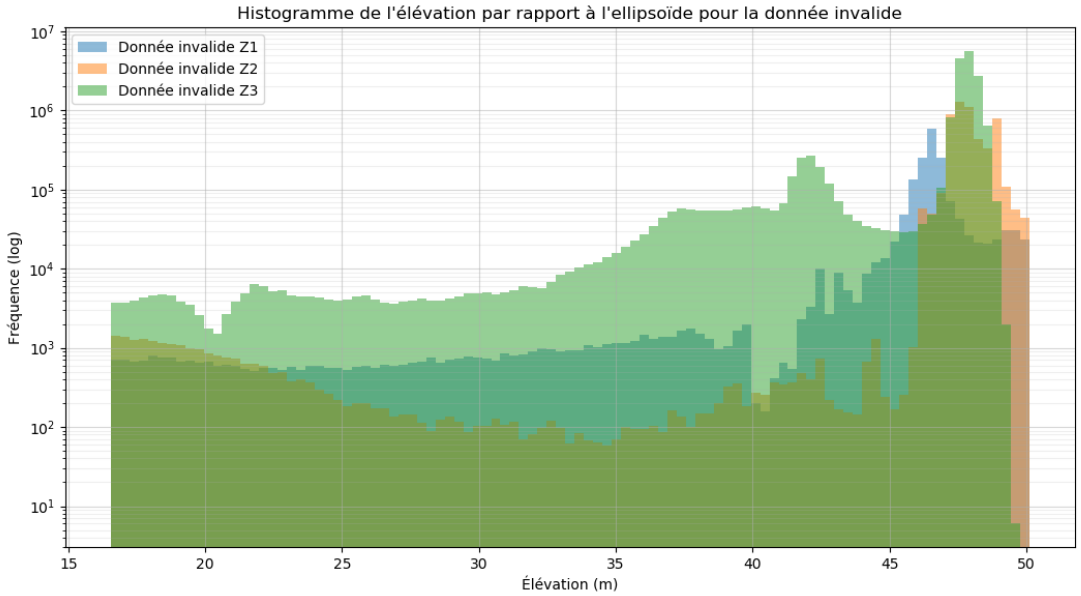


FIGURE 3.21 – Histogramme de l'élévation par rapport à l'ellipsoïde pour la donnée invalide (en échelle logarithmique pour les abscisses) sur les 3 zones d'étude.

Nous observons que la gamme d'élévations (élévations par rapport à l'ellipsoïde IGN69) mesurée est assez proche (dans la plage 40m / 50m) pour la zone 1 et la zone 2, qui sont également assez proches spatialement, alors que, dans la zone 3,

nous mesurons une donnée bathymétrique valide entre 24m et 33m d'élévation (par rapport à l'IGN69). Nous observons également que la zone 1 est la plus petite à traiter (en nombre de points et surface) puis vient la zone 2 et enfin la zone 3 qui est la plus vaste. Nous nous trouvons également devant un déséquilibre important entre la donnée valide et invalide. Le ratio se situe dans la même gamme pour la zone 1 et 2 (ce qui s'explique par une proximité dans le type de fond et l'élévation moyenne). Ce déséquilibre est beaucoup plus important pour la zone 3. En effet, dans cette zone, seul le capteur 3HD fournit de l'information pertinente (le laser vert *shallow* 2HD étant émis avec une puissance moins forte que le 3HD *deep*). Néanmoins, le nombre d'échantillons valides semble être suffisant pour réaliser un apprentissage sur la zone.

### 3.3.2 Résultats de l'apprentissage

Avant la mise à disposition des données prédites aux opérateurs, la structuration de la donnée et l'apprentissage ont été réalisés indépendamment sur les 3 zones en suivant la méthodologie proposée dans la section 3.2. Cette partie présente les résultats statistiques globaux obtenus après cette phase.

Afin d'assurer un apprentissage uniforme et ne pas sur-représenter une certaine gamme d'élévations, nous avons réalisé un échantillonnage basé sur l'histogramme d'élévation, voir 3.11. Seulement 25% de la donnée ont été utilisés pour entraîner le modèle (3050 échantillons pour le train-test sur un total de 12621 échantillons pour la zone 1). Dans ce sous-échantillon d'apprentissage, 70% de la donnée ont été utilisés pour le *train* et 30% pour le test de la prédiction, comme présenté dans la figure 3.22. Les conditions d'apprentissage sont également optimales pour les trois zones car un échantillonnage aléatoire a été effectué, comme le montre la figure 3.22 au centre.

La table 3.2 présente les résultats d'apprentissage sur les trois zones avant et après régularisation spatiale reposant sur le score MAD. Nous avons réalisé la différence entre la prédiction de tous les échantillons (pas uniquement le sous-ensemble d'apprentissage) et la vérité terrain (donnée traitée par un opérateur avec la procédure manuelle classique).

Nous relevons tout d'abord qu'aucune prédiction n'a pu être réalisée pour la zone 2 avec le capteur 3HD. Il y a effectivement trop peu de sondes valides avec

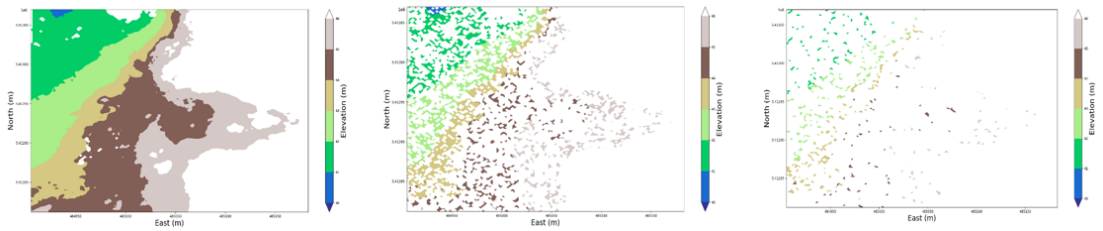

FIGURE 3.22 – Représentation de la vérité terrain disponible (à gauche), donnée utilisée pour l'apprentissage 70% (au centre) et donnée utilisée pour le test 30% (à droite) pour la zone 1.

TABLE 3.2 – Résultats d'apprentissage sur les trois zones avant et après régularisation.

| Résultats | Zone 1 avant | Zone 1 après | Zone 2 avant | Zone 2 après | Zone 3 avant | Zone 3 après |
|---|---|---|---|---|---|---|
| MAE 2HD | 0.08 m | 0.07 m | 0.06 m | 0.06 m | 0.30 m | 0.32 m |
| Écart-type 2HD | 0.12 m | 0.10 m | 0.05 m | 0.05 m | 0.47 m | 0.47 m |
| Écart max 2HD | 3.27m | 2.16 m | 2.03 m | 0.69 m | 4.41 m | 3.30 m |
| Écart min 2HD | -2.20 m | -1.44 m | -0.91 m | -0.36 m | -4.91 m | -4.91 m |
| MAE 3HD | 0.27 m | 0.25 m | - | - | 0.10 m | 0.10 m |
| Écart-type 3HD | 0.34 m | 0.32 m | - | - | 0.15 m | 0.13 m |
| Écart max 3HD | 1.85 m | 1.85 m | - | - | 2.34 m | 1.76 m |
| Écart min 3HD | -2.51 m | -2.51 m | - | - | -2.67 m | -2.18 m |

ce capteur dans cette zone pour que l'apprentissage soit possible. Ensuite, dans l'intégralité des cas, la régularisation spatiale mise en place via le score MAD est meilleure ou au moins aussi bonne (cas Z3 - 2HD) que sans régularisation. Elle a donc une action positive en réduisant notamment les écarts min et max comme pour la zone 2 où nous passons d'une amplitude min/max de presque 3m à 1m. Nous pouvons également remarquer que le capteur 2HD est plus juste pour les zones 1 et 2 alors que, pour la zone 3 le capteur 3HD donne les meilleurs résultats. Ce comportement s'explique par le fait que le capteur 2HD est beaucoup moins bruité et plus dense que le 3HD dans les eaux peu profondes. En revanche, le bruit de mesure devient plus important avec l'augmentation de la profondeur et nous sommes en limite de décrochage du capteur 2HD dans le cas de la zone 3. La prédiction du capteur 3HD sera donc à privilégier dans cette zone (et l'inverse

pour les zones 1 et 2).

La généralisation de la prédiction semble être également très pertinente. Pour rappel, dans les résultats de la table 3.2, 75% de la donnée n'a jamais été utilisées dans le processus d'apprentissage.

La figure 3.23 montre un rendu graphique pour la zone 1 de la vérité terrain générée à partir de la donnée traitée, la prédiction réalisée avec le modèle et cette prédiction régularisée via le score de MAD.

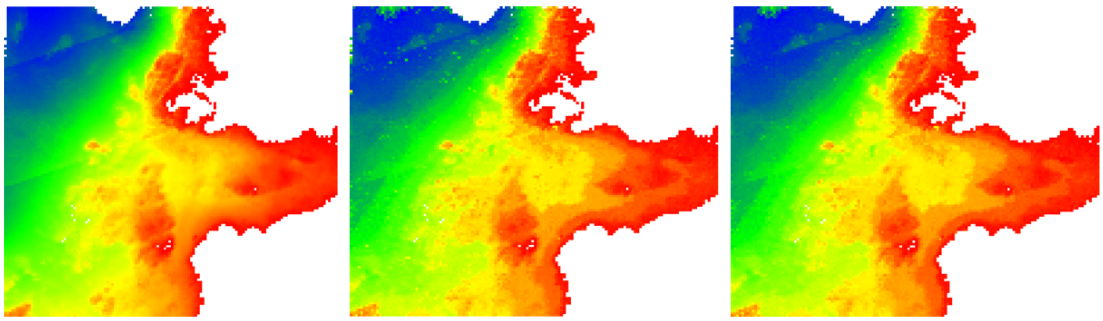

FIGURE 3.23 – MNB à 2m de la vérité terrain (à gauche), de la prédiction en sortie de modèle (au centre) et la prédiction régularisée (à droite) : zone 1 capteur 2HD.

La figure 3.24 présente la carte des différences entre la prédiction régularisée et la vérité terrain. Nous pouvons apercevoir qu'un groupe de pixels au nord-ouest de la carte est encore bien présent. Ce type de groupe de pixels aberrants est la limite principale de notre régularisation spatiale.

### 3.3.3 RETEX des opérateurs

Les trois zones étudiées dans la section 3.3.1 ont ensuite été traitées par 3 opérateurs d'AL (sauf la zone 1 qui a été traitée uniquement par 2 opérateurs par manque de temps). Pour chaque zone, les opérateurs ont passé environ 45 minutes à réaliser un traitement manuel et 45 minutes sur un traitement avec l'aide de la méthodologie ML. Le traitement méthode ML comprend la mise en place du filtre et la vérification manuelle de la zone couverte par l'opérateur (avec suppression ou validation des sondes proposée par le filtre). La table 3.3 montre le pourcentage de zone traitée pour les trois opérateurs sur les trois zones d'études dans le temps imparti.

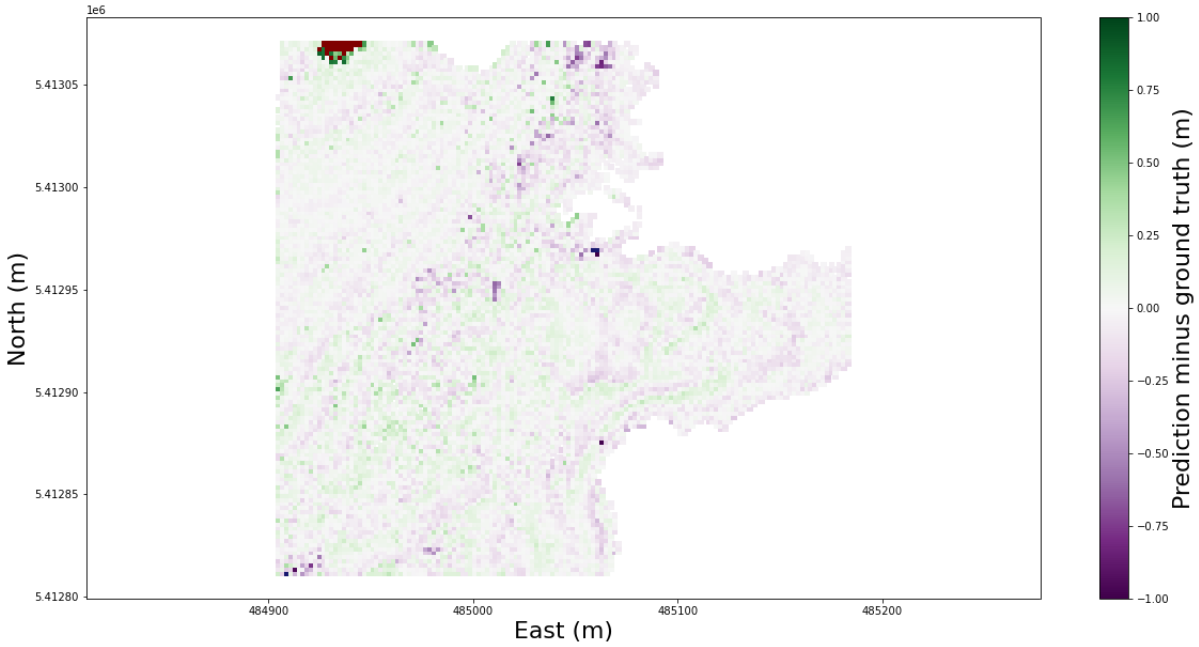


FIGURE 3.24 – Différence entre la prédiction régularisée et la vérité terrain en mètre : zone 1 capteur 2HD)

TABLE 3.3 – Pourcentage de zone traitée (et temps passé) pour la méthode manuelle et ML pour chaque opérateur.

| Traitement manuel / ML | Opérateur 1 | Opérateur 2 | Opérateur 3 |
|---|---|---|---|
| Zone 1 | 80% (45 min) | - | 65% (45 min) |
| Zone 1 | 59% (45 min) | - | 30% (45 min) |
| Zone 2 | 80% (45 min) | 62% (45 min) | 23% (45 min) |
| Zone 2 | 52% (45 min) | 100% (45 min) | 15% (45 min) |
| Zone 3 | 100% (45 min) | 100% (45 min) | 60% (45 min) |
| Zone 3 | 100% (45 min) | 100% (10 min) | 63% (45 min) |

Pour les zones 1 et 2, la méthode de traitement manuel est *a priori* meilleure que la méthode ML dans le pourcentage de zones couvertes par l'opérateur pour un même temps donné (45 minutes). Les résultats sont à mettre en perspective d'une nouvelle méthodologie ML mise en place dans le cadre de cette expérimentation (temps de formation nécessaire à la prise en main) mais également dans le temps très court dédié à cette expérimentation.

En revanche, nous constatons pour la zone 3 une augmentation de la zone couverte et un traitement plus fin avec la méthode ML comparée à la méthode manuelle classique. Un gain de vitesse de traitement semble donc perceptible pour les zones simples mais reste à quantifier de façon plus précise lors d'une future expérimentation à plus grande échelle.

Pour l'ensemble des zones, il est à noter que les opérateurs traitaient pour la première fois la donnée bathymétrique avec cette méthode ML innovante. Il y a donc un temps d'apprentissage et de familiarisation non négligeable à considérer, même s'il est facilité par l'intégration dans l'outil PFM ABE. De plus, comme indiqué précédemment, la résolution verticale adoptée pour le filtre sur cette première expérimentation était de 50 cm pour les premiers essais. Avec un filtre réglé sur les indices [99,100], les sondes détectées par le filtre se trouvent ainsi à +/- 50cm du fond. Il est probable que les sondes sélectionnées soient trop nombreuses et que les opérateurs n'aient que très peu de latitude sur la configuration du filtre pour une sélection pertinente des sondes à conserver.

Cette dernière justification explique en grande partie les résultats obtenus sur les zones 1 et 2, le tapis de sondes LiDAR bathymétrique étant compris entre 10 et 20cm pour la gamme de profondeurs représentée pour les zones 1 et 2 avec le capteur 2HD. La résolution du filtre proposée sur cette première itération a donc eu tendance à proposer un fond en partie bruité (conservant une partie du bruit présent autour du fond). Pour le capteur 3HD, tout particulièrement adapté à la zone 3, le tapis de sondes est compris verticalement entre 20 et 40cm, ce qui est donc plus proche de la résolution verticale proposée du filtre et ce qui permet de sélectionner les sondes les plus pertinentes et les plus représentatives des fonds marins.

Des essais complémentaires sur la dynamique du filtre ont été réalisés afin d'affiner cette valeur de résolution du filtre en essayant 20cm puis 10cm de résolution, permettant ainsi une meilleure sélection des sondes par les opérateurs. Les premiers retours concernant la détection du fond avec ces paramètres de résolution sont beaucoup plus pertinents pour les zones 1 et 2 et permettent même, pour la zone 3, de réaliser un traitement grossier et de proposer un traitement fin convaincant (pour lequel des reprises manuelles sont toutefois à réaliser). Une seconde campagne d'expérimentation réalisée en 2022 a validé ce nouveau paramètre.

Après la phase d'expérimentation technique, des RETour d'EXpérience (RE-

TEX) ont été réalisés avec les opérateurs participants afin d'évaluer qualitativement cette nouvelle méthodologie. Ils ont été l'occasion d'attribuer une note sur 10 à l'outil sur les thèmes suivants :

- satisfaction de l'ergonomie de l'outil, 8.3/10 ;
- évaluation de l'homogénéisation des résultats grâce à la méthodologie ML, 7.6/10 ;
- satisfaction de la qualité des MNT générées, 7.3/10 ;
- évaluation de la facilitation du travail grâce à la méthode ML, 8/10.

Finalement, sur l'aspect qualitatif, la prise en main de cette nouvelle méthode n'a pas semblé amener de difficulté particulière et l'évolution demandée par le département AL (modification de la résolution du filtre) a permis de faciliter son utilisation. Les opérateurs ont jugé cette nouvelle méthodologie prometteuse et très efficace pour les fonds considérés comme simples. Néanmoins, elle nécessite encore des développements pour faciliter son utilisation comme la présence d'un facteur de confiance/qualité associé à la prédiction. Des essais complémentaires seront également à réaliser sur les zones d'interfaces terre/mer, zones plus complexes à traiter par les opérateurs.

Ces entretiens ont également été l'occasion de réfléchir et d'évaluer les impacts métiers et organisationnels de cette méthodologie dans le cadre du traitement de la donnée LiDAR bathymétrique.

## 3.4 Perspectives

Dans des travaux futurs, nous aimerions mettre en place différents scénarios d'apprentissage afin d'en évaluer leur pertinence : deux scénarios différents seraient considérés en fonction de l'existence des données présentes dans la zone, voir la figure 3.25. Dans le premier scénario, les opérateurs auraient déjà traité une partie des lignes de vol LiDAR (marquée en rouge) que nous utiliserions comme données d'apprentissage. Dans le second scénario, les opérateurs auraient déjà traité la zone (ou une zone voisine), lors d'un levé précédent, et nous utiliserions cet historique comme donnée d'apprentissage.

De plus, nous pensons que de nombreuses perspectives existent sur la construction des descripteurs. De fait, l'ajout d'éléments supplémentaires permettrait de mieux décrire le nuage de points (en prenant exemple sur l'article [107]). Nous

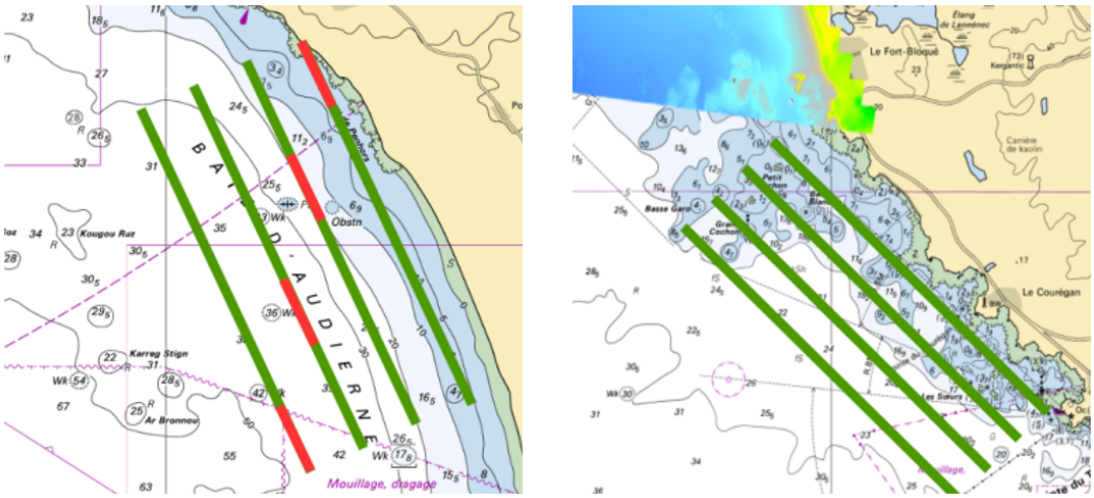


FIGURE 3.25 – Différents scénarios d'apprentissage envisagés. À gauche : apprentissage après une phase de traitement manuel partiel (en rouge). À droite : apprentissage à partir d'une zone déjà traitée (représentée par le MNT en haut de la figure).

pourrions également construire des descripteurs non-supervisée comme avec le clustering de données DBSCAN [60] présenté dans la section 2.4, ou bien l'algorithme de clustering développé par l'Inria, ToMATo [153], déjà testé sur de la donnée bathymétrie SMF dans l'article [17] (algorithme présent dans la librairie [154]).

Enfin, comme vu dans la section 3.3.2, de nombreuses erreurs de prédiction sont localisées sur des zones restreintes voire des pixels isolés. Nous avons ainsi mis en place une régularisation spatiale très simple qui pourrait être complexifiée via l'intégration de la régularisation dans la construction des descripteurs ou l'apprentissage lui-même, et ce afin d'améliorer la prédiction finale en prenant en compte les pixels horizontaux voisins. Nous pourrions également mettre en place des combinaisons de différentes vues et de différents capteurs pour fusionner les données multi-sources et multi-échelles. Néanmoins, il existe également de nombreuses zones qui ne comportent aucune donnée bathymétrique. La figure 3.26 présente ainsi, à gauche, la zone 1 de l'application LiDAR bathymétrique sur laquelle nous voyons en rouge les pixels contenant des données bathymétriques et en bleu les données ne comportant que des erreurs aberrantes (au sens de l'acquisition de donnée LiDAR donc comportant la surface d'eau et les erreurs dans la colonne d'eau). Sur la droite de cette figure, nous remarquons qu'une prédiction a

été réalisée sur les zones ne comportant pas de donnée bathymétrique. La limite principale de cette prédiction est que les données prédites semblent vraisemblables, surtout sur la zone en haut à gauche entourée en bleu. De ce fait, l'ensemble de pixels prédits suite à la régularisation spatiale pourrait être assimilé rapidement à une obstruction par un hydrographe.

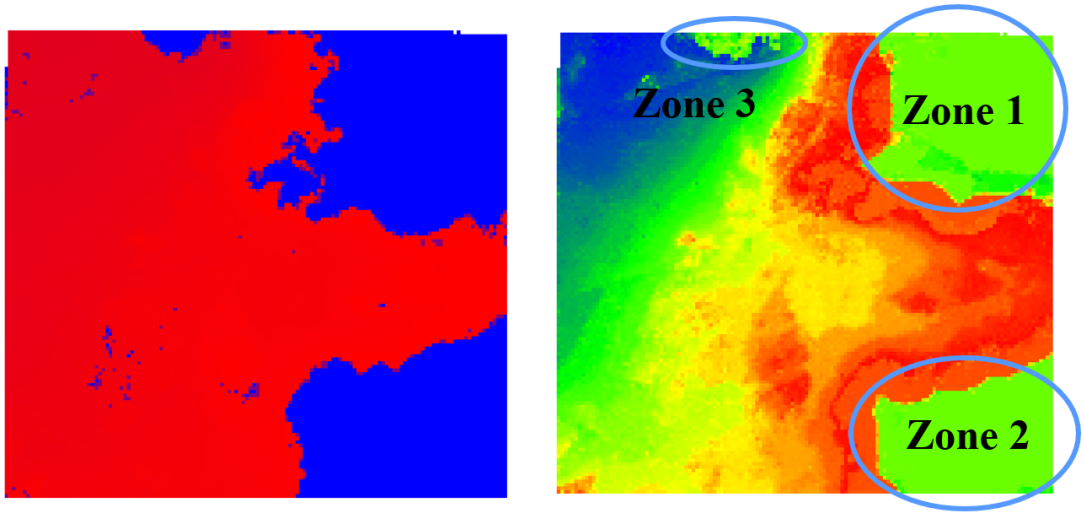


FIGURE 3.26 – À gauche : vérité terrain pour la zone 1, voir section 3.3, en rouge les pixels possédant de la donnée bathymétrique, en bleu les pixels ne contenant que des données aberrantes. À droite : prédiction régularisée de la zone 1, entourées en bleu des zones prédites sans support de données bathymétriques dans la colonne d'eau.

Sur la figure 3.27, cette donnée prédite est représentée en vue 3D isométrique. Nous remarquons que certaines zones sans donnée bathymétrique sont facilement repérables (les zones de plateau en haut et à droite de la figure) mais l'algorithme de régression peut également prédire des données bathymétriques fausses vraisemblables, comme le haut-fond sur la gauche de la figure. De plus, à cause de la régression spatiale, des zones tampons sont générées entre la donnée prédite de bonne qualité et la donnée prédite sans mesure bathymétrique, ce qui peut être perturbant pour les opérateurs.

Il est donc important, dans les perspectives, de réfléchir à la manière de prédire ces zones sans donnée bathymétrique afin de s'assurer qu'elles ne soient pas assimilées à des hauts-fonds et ce afin de ne pas polluer le lot bathymétrique final qui sera intégré en base de données bathymétriques, mais également pour faciliter

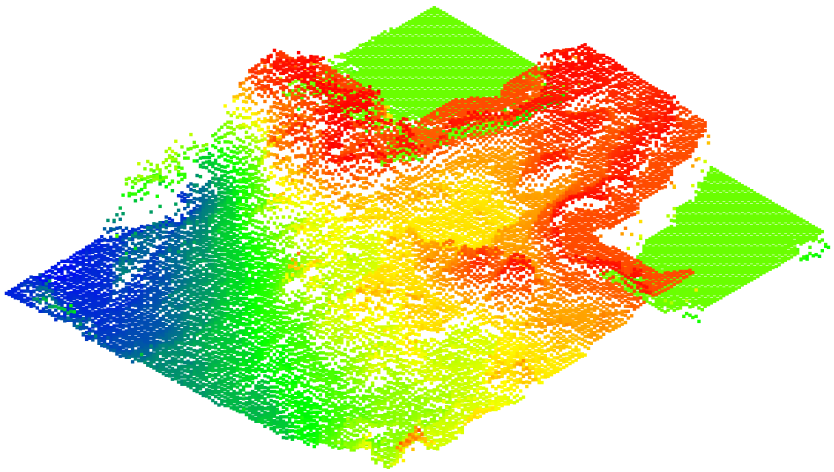

FIGURE 3.27 – Vue 3D isométrique de la prédiction de la zone 1, voir section 3.3.

le travail des opérateurs. Le passage par une couche d'incertitude associée à la prédiction permettrait de faciliter ce travail de détection des zones ne contenant pas de donnée bathymétrique.

## 3.5 Conclusion

Cette proposition méthodologique de traitement de donnée LiDAR bathymétrique via des techniques de ML ne constitue qu'un maillon d'une stratégie globale beaucoup plus importante autour du traitement de données hydrographiques fondé sur des technologies utilisant l'IA (au sens très large avec des techniques ML dans ce chapitre, aide à la décision dans le chapitre 4, et règles expertes dans la section 5). En effet, les premiers résultats de ce travail de recherche sont très prometteurs dans le domaine du traitement semi-automatique des nuages de points bathymétriques, notamment en terme de confiance que portent les opérateurs dans la méthodologie mise en place et l'intégration dans la chaîne de valeur actuelle. Ainsi, comme présenté dans la section 2.3, le Shom s'est mis à jour de l'état de l'art concernant le traitement de données bathymétriques et comme exposé dans ce chapitre, nous avons même contribué à la recherche dans ce domaine, des sciences informatiques appliquées à l'environnement marin [14], en proposant une solution novatrice et unique de régression du fond à partir de techniques ML pour la détection du fond

dans la donnée LiDAR bathymétrique. Cette méthodologie nous servira ainsi de référence pour les développements et projets futurs présentés dans la section 5.

De plus, ces travaux ont abouti à une preuve de concept de gestion de données LiDAR bathymétrique pour la prédiction d'un fond marin, prédiction qui a été intégrée avec succès dans le logiciel de traitement de données bathymétriques d'AL, susceptible d'être généralisé avec PFM ABE. Néanmoins, de nombreux points restent à traiter, comme présenté dans la section 3.4, et l'intégration reste à parfaire même si les objectifs initiaux du projet de recherche ont tous été atteints. Le travail avec des organismes de recherche - Inria Saclay et IMT Atlantique - de premier plan, via l'appel à manifestions d'intérêt sur l'IA remporté par le Shom, a clairement permis à cet axe de recherche de prendre une envergure plus importante. Le partenariat entre le Shom et ces organismes est à poursuivre dans cette même thématique de recherche. Des développements et essais devront être conduits durant l'année 2023 pour améliorer les modèles et continuer à intégrer les utilisateurs finaux dans le processus de développement pour construire un écosystème de confiance autour de la méthode. Ainsi, ces travaux s'inscriront dans la préparation du Projet de Technologies de Défense (PTD) AUTOBATH, soutenu par la DGA, qui s'intéressera à la transition vers la donnée SMF et la construction de nouveaux descripteurs.

Une fois que l'opérateur réalise l'étape de traitement manuel et de contrôle (présenté en rouge sur la figure 3.4), le travail de structuration de la donnée et de prédiction d'un fond marin à partir de données LiDAR bathymétriques permet de finaliser la phase de traitement de la donnée bathymétrique, voir la figure 1.11. Il s'ensuit alors une phase de rédaction et de construction des métadonnées de qualité associées au levé hydro-océanographique. Ainsi, le chapitre 4 suivant étudiera les axes de recherche permettant de faciliter le travail des hydrographes dans l'élaboration des produits bathymétriques finaux tout en prenant en compte les préférences utilisateurs qui diffèrent en fonction du concept d'emploi final de la donnée.

Chapitre 4

# L'échelle macro : la décision appliquée au levé

**Synthèse du chapitre**

Évaluer la qualité d'un levé hydro-océanographique est une tâche complexe et cruciale pour l'aide à la décision et l'exploitation d'un levé hydro-océanographique en fonction du besoin utilisateur final. Ainsi, ce chapitre présente deux contributions principales :

- La détermination de l'emprise d'un levé via un algorithme original QuadSME, première étape primordiale de la localisation spatiale de l'ensemble des critères de qualité pour l'évaluation de la qualité d'un levé ;
- L'utilisation de modèles et de méthodes Multiple-Criteria Decision Analysis (MCDA) pour déterminer la qualité d'un levé hydro-océanographique pour différents concepts d'emploi, en fonction des priorités et des préférences variables des utilisateurs. La prise en compte de tous ces critères de qualité facilitera ensuite la comparaison entre levés pour la prise de décision à un niveau macro, permettant notamment de définir la priorité d'un levé par rapport à l'autre.

L'automatisation de l'évaluation de la qualité de la donnée issue de levés hydrographiques est un problème complexe. En effet, le contexte dans lequel les campagnes multi-capteurs sont réalisées conditionne la qualité de la donnée et de l'information obtenue. Les conséquences peuvent être importantes surtout en fonction de l'usage final associé à la donnée acquise. L'OHI a récemment publié une nouvelle édition de la norme de levés hydrographiques (S-44 édition 6.0.0 - [8]). Cette sixième édition démontre l'importance accordée par le milieu de l'hydrographie à la qualité de la donnée, et tout particulièrement à travers l'ajout d'un nouvel ordre plus strict, « l'ordre exclusif ». Ces ordres (exclusif, spécial, 1A, 1B et 2) requièrent d'appliquer une méthodologie de levé adéquate et l'utilisation de capteurs d'acquisition spécifiques afin de répondre aux normes. De plus, cette nouvelle norme intègre une matrice, voir la figure 4.1, tableau d'exigence issu du §7.6 de [8]. Cette matrice définit des niveaux de critères de qualité de la donnée selon les besoins hydrographiques, ce qui signifie que la qualité d'un levé diffère selon le concept d'emploi défini par le prescripteur du levé ou l'utilisateur final. Aujourd'hui, il est crucial d'avoir des normes plus strictes en terme de qualité de la donnée hydro-océanographique afin de générer des produits nautiques de meilleures résolutions et nécessaires à la navigation autonome de demain, comme les produits bathymétriques de hautes densités S-102 [155]. Cette matrice permet également de qualifier très finement les caractéristiques de qualité bathymétrique d'un levé hydro-océanographique. Nous y retrouvons, par exemple, les informations d'incertitude sur la mesure de profondeur (*Total Vertical Uncertainty (TVU)* en fonction de la profondeur) ou la *Couverture bathymétrique* décrivant le pourcentage de recouvrement entre chaque ligne de levé. Les méthodologies de qualification des données bathymétriques sont rappelées dans la section 2.2.

La qualité de la donnée bathymétrique est nécessaire pour de nombreux besoins, notamment pour la construction d'indicateurs d'incertitude associés à un levé ou un produit bathymétrique (comme EMODnet Bathymetry [156]), ou bien dans l'évaluation des risques hydrographiques [157]. Dans le cas de l'évaluation des risques (*risk assessment* en anglais), de nombreuses méthodes reposent sur l'utilisation de la qualité d'un levé bathymétrique pour produire des cartes de risque comme présenté à la 92$^{e}$ conférence EURO Working Group on Multi-Criteria Decision Aiding [158] ou bien dans le cadre de la méthodologie de l'évaluation des risques hydrographiques définie par le SH néo-zélandais (Land Information New

| | Critères | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | BATHYMÉTRIE | | | | | | | | | | | | | | |
| a | Profondeurs THU [m] | 500 | 200 | 100 | 50 | 20 | 15 | 10 | 5 | 2 | 1 | 0,5 | 0,35 | 0,1 | 0,05 |
| b | Profondeurs THU [% de la profondeur] | 20 | 10 | 5 | 2 | 1 | 0,5 | 0,25 | 0,1 | | | | | | |
| c | Profondeurs TVU « a » [m] | 100 | 50 | 25 | 10 | 5 | 2 | 1 | 0,5 | 0,3 | 0,25 | 0,2 | 0,15 | 0,1 | 0,05 |
| d | Profondeurs TVU « b » Note 1 | 0,20 | 0,10 | 0,05 | 0,023 | 0,02 | 0,013 | 0,01 | 0,0075 | 0,004 | 0,002 | | | | |
| e | Détection d'éléments [m] | 50 | 20 | 10 | 5 | 2 | 1 | 0,75 | 0,7 | 0,5 | 0,3 | 0,25 | 0,2 | 0,1 | 0,05 |
| f | Détection d'éléments [% de la profondeur] | 25 | 20 | 10 | 5 | 3 | 2 | 1 | 0,5 | 0,25 | | | | | |
| g | Recherche d'éléments [%] | 1 | 3 | 5 | 10 | 20 | 30 | 50 | 75 | 100 | 120 | 150 | 200 | 300 | |
| h | Couverture bathymétrique [%] | 1 | 3 | 5 | 10 | 20 | 30 | 50 | 75 | 100 | 120 | 150 | 200 | 300 | |
| | | Ordre 2 | | | Ordre 1b | | | Ordre 1a | | Ordre spécial | | | Ordre exclusif | | |

FIGURE 4.1 – Extraction de la matrice pour la qualification de la donnée bathymétrique et les ordres associés, inspiré de la norme S-44 de l'OHI [8].

Zealand - LINZ), voir [159] et [160].

Les données et informations récoltées lors de levés hydro-océanographiques sont également très importantes dans des processus de décision comme la navigation autonome de porteur en surface ou sous-marins (Autonomous Surface Vehicle (ASV) et AUV) [161] et, tout particulièrement, la donnée bathymétrique dans les applications de recalage de navigation sous-marine et cartographie simultanées [162] (*Simultaneous Localization And Mapping* (SLAM)).

Dans ce chapitre, nous nous intéresserons dans un premier temps à la génération de l'emprise du levé et les problématiques associées à la construction d'une enveloppe géographique pour des jeux de données comportant des millions de sondes et nous proposerons un algorithme itératif basé sur une décomposition Quadtree afin de répondre à ces problématiques. Dans un second temps, nous répondrons à la question suivante : "comment prendre en compte la préférence d'un expert dans la qualification d'un levé hydro-océanographique ?". Enfin, nous présenterons notre méthode basée sur l'aide multicritère à la décision.

## 4.1 L'emprise du levé

Avant de s'intéresser à la qualité d'un levé, la première chose à déterminer est l'emprise géographique de ce levé. En effet, cette emprise géographique définit la géométrie qui sera intégrée dans une BDD géospatiale métier et à laquelle sera rattaché l'ensemble des métadonnées (dont les métadonnées de qualité) du levé hydro-océanographique. De plus, cette géométrie est utilisée dans le processus de déconfliction (comparaison locale des données bathymétriques) afin d'obtenir les données bathymétriques de meilleure qualité sur une zone géographique spécifique. Ce processus de déconfliction constitue l'une des premières étapes de la production d'un MNT, produit pour les forces ou pour une carte nautique. La déconfliction des données hydrographiques, voir [163], permet de définir une surface de référence de données bathymétriques à partir de laquelle l'ensemble des chaînes de traitement réalise des produits (dédensifications de sondes, modélisations bathymétriques ou isobathes). C'est également à partir de cette emprise qu'un ensemble de décisions sera pris pour le levé car elle est l'élément de représentation géospatiale du levé bathymétrique. L'emprise géographique catalyse donc un ensemble de processus clés dans les chaînes de traitement de données des levés hydro-océanographiques.

Ainsi, le multi-polygone, enveloppe concave de l'emprise de chaque levé bathymétrique, est utilisé dans le processus de déconfliction comme enveloppe de découpe (*cookie cutter* en anglais), dans le processus automatique de sélection des meilleures sondes en fonction de la couverture et des métadonnées descriptives du levé source. De manière simplifiée, et à titre d'exemple, le levé le plus récent et de meilleure qualité (établi sur des indicateurs de qualité comme l'incertitude de la mesure verticale et horizontale associée au levé) annule et remplace les levés plus anciens. Dans le cas du Shom, ce sont plus de 300 règles expertes qui régissent le processus de fiabilisation et de déconfliction de la donnée bathymétrique dans le cadre du projet Téthys (présenté dans la section 5). L'emprise géospatiale de chaque jeu de données est appelée SME. La SME d'un lot bathymétrique est l'union des ensembles disjoints dans lesquels s'inscrivent les sondes du lot. Cette union d'ensembles doit décrire la géométrie du lot en minimisant sa surface et en respectant les trois conditions suivantes :

1. Condition d'unicité : toute sonde du lot est incluse dans un unique ensemble.
2. Condition de densité : dans chaque ensemble, les plus proches voisines d'une

sonde sont distantes de strictement moins de 5 fois (norme empirique choisie par le Shom pour limiter les polygones complexes) le maximum de la résolution caractéristique (distance moyenne entre les sondes dépendant de la profondeur des sondes et du capteur utilisé) du lot bathymétrique. Si cette distance entre 2 sondes n'est pas respectée, un nouvel ensemble est créé (voir cas 3.1 de la figure 4.2). Si une sonde ne peut être agrégée avec d'autres, elle est considérée comme isolée (ensemble créé avec une ou deux sondes, voir cas 3.2 de la figure 4.2).

3. Condition de représentativité :

   3.1 Pour un ensemble de sondes agrégées : le contour (interne et externe) de l'ensemble est dilaté (création d'une zone tampon) d'une distance ($D_{tampon}$) choisie comme :

$$D_{tampon} = min(min(POSACC), \frac{max(reslot)}{2})) \tag{4.1}$$

   le $POSACC$ étant l'incertitude horizontale associée aux sondes bathymétriques (et qui peut varier en fonction de la profondeur des sondes mesurées) et le $reslot$ la résolution caractéristique du lot

   3.2 Pour un ensemble de sonde(s) isolée(s) : toute sonde appartient à un ensemble circulaire de rayon minimum ($R_{tampon}$) défini comme :

$$R_{tampon} = min(min(POSACC), \frac{max(reslot)}{2})) \tag{4.2}$$

La figure 4.2 synthétise l'ensemble de ces critères nécessaires à l'élaboration de la SME pour un ensemble de mesures bathymétriques issu d'un levé hydro-océanographique. Les croix représentent les différentes sondes et les aires colorées les emprises composant la SME finale (de géométrie multi-polygone trouée).

Avant 2018, cette emprise géographique était construite totalement manuellement au Shom. Les opérateurs généraient la SME via un outil de tracé en essayant de respecter au mieux les conditions d'unicité, de densité et de représentativité. Cette tâche très fastidieuse dans le cas de lots contenant des millions de points était génératrice d'erreurs, avec notamment de nombreuses sondes hors de la SME. Cette opération manuelle est subjective. Il nous a donc paru important de réfléchir à la manière de générer automatiquement (tout en conservant

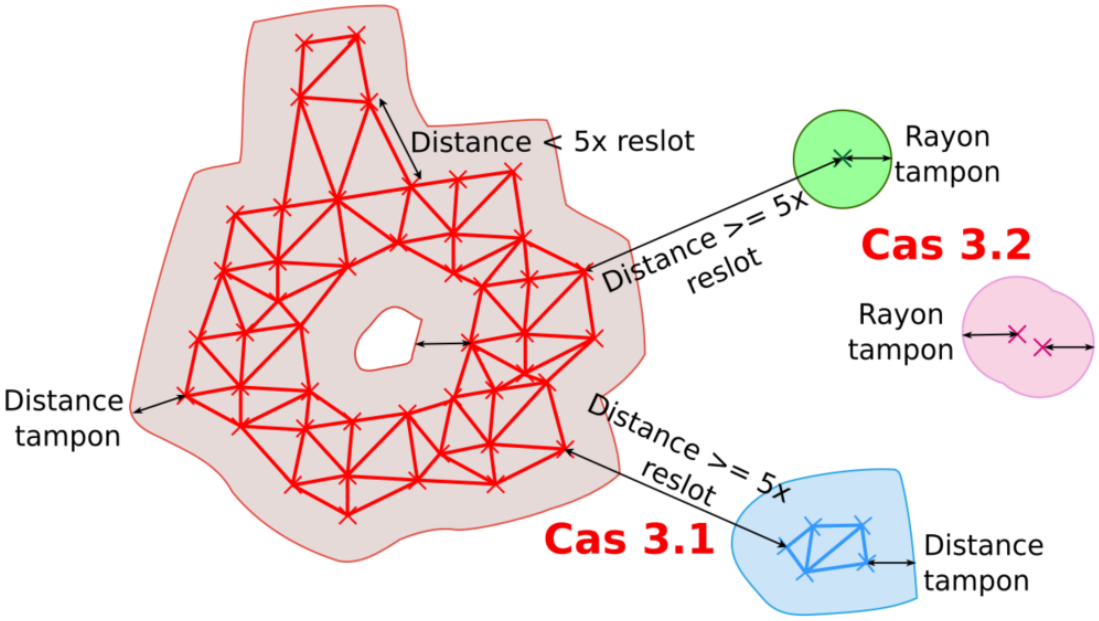


FIGURE 4.2 – Définition de la SME au Shom.

un contrôle manuel) la SME associée à chaque levé bathymétrique, via des méthodes de détection d'enveloppe classique, afin d'améliorer la représentativité des levés hydro-océanographiques et les processus de décision associés dans un objectif d'harmonisation de la SME obtenue.

### 4.1.1 $\alpha$-shape, méthode classique de détermination d'emprise spatiale

#### Présentation de l'algorithme $\alpha$-shape

Afin de déterminer l'emprise spatiale la plus fidèle possible pour un levé bathymétrique, nous avons étudié l'algorithme $\alpha$-shape. Définie par Edelsbrunner *et al.* [164] puis étendue en trois dimensions [165], cette méthode très classique permet de déterminer l'enveloppe d'un groupe de points. L'objectif de l'algorithme est de généraliser la construction d'une enveloppe concave grâce à un algorithme de complexité $\mathcal{O}(n \log n)$, avec $n$ représentant le nombre de points. L'algorithme $\alpha$-shape est utilisé pour obtenir les segments de droites composant le pourtour d'un ensemble de points dans le plan permettant ainsi de construire l'emprise spatiale la plus stricte à partir de ces segments composant la frontière du nuage de points

en entrée. La construction efficace d'enveloppes concaves pour des ensembles finis de points dans le plan est l'un des problèmes les plus étudiés dans le domaine de la géométrie algorithmique. Son champ d'application est encore très riche aujourd'hui notamment dans le domaine de l'extraction et la reconnaissance de forme (avec l'exemple de l'extraction de toit dans de la donnée LiDAR [166]) ou pour quantifier la complexité d'une forme comme pour l'article [167].

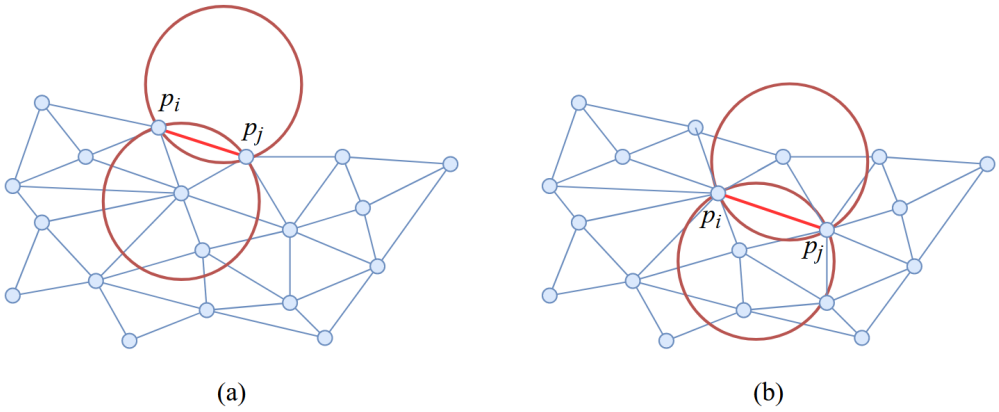


FIGURE 4.3 – Critères de l'algorithme $\alpha$-shape. (a) Les points $p_i$ et $p_j$ forment un segment de frontière, tandis que pour (b) $p_i$ et $p_j$ forment un segment interne.

**$\alpha$-shape algorithme**

En considérant le paramètre $\alpha$ et deux points $p_i$ et $p_j$ qui composent une arête dans le maillage résultant d'une triangulation de Delaunay [168] contrainte, il est possible de définir deux cercles de rayon $\alpha$ passant par $p_i$ et $p_j$, comme le montre la figure 4.3. Si aucun point ne se trouve à l'intérieur d'au moins un des cercles, le bord est traité comme un segment de frontière valide, voir figure 4.3(a) ; sinon, s'il y a au moins un point à l'intérieur de chaque cercle, le segment est considéré comme interne, comme le montre la figure 4.3(b).

Le paramètre clé de cet algorithme est la valeur $\alpha$ qui définit le rayon $r$ des cercles : $r = 1/\alpha$. Dans le cas d'un levé bathymétrique, il est important d'adapter

ce paramètre à la densité du lot pour s'assurer d'avoir la SME la plus représentative du lot de donnée et que l'emprise ne comporte pas de zones sans donnée. La figure 4.4 montre ainsi l'algorithme $\alpha$-shape sur un jeu de données synthétiques pour différentes valeurs de $\alpha$. On voit que, pour cet exemple, les valeurs de $\alpha = [0.5, 2]$ ne sont pas assez restrictives et conservent des zones importantes sans donnée. À l'inverse, pour une valeur de $\alpha = 10$ et pour des valeurs supérieures, l'algorithme va isoler des points. L'ajustement de ce paramètre est critique pour générer l'emprise spatiale la plus fidèle.

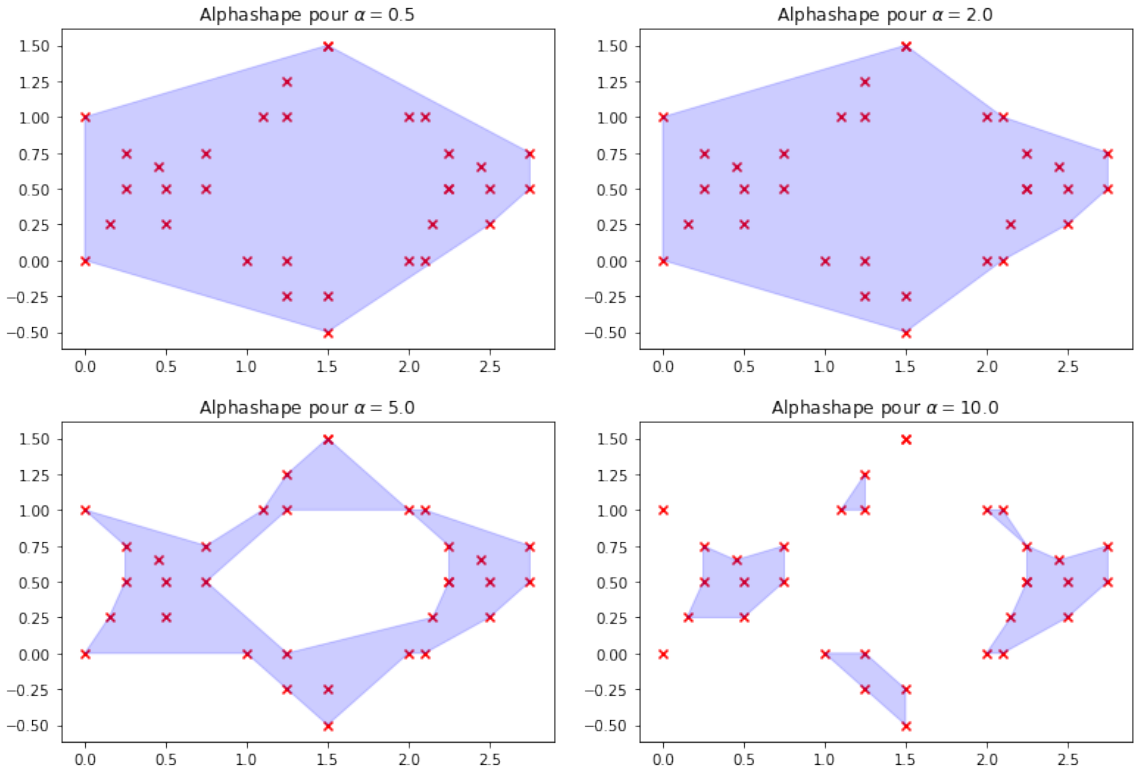


FIGURE 4.4 – Algorithme $\alpha$-shape sur un jeu de données synthétiques pour différentes valeurs d'$\alpha$.

### Application à des levés bathymétriques

Cette méthode classique de construction d'enveloppe spatiale, présente dans de nombreux Système(s) d'Informations Géographiques (SIG), nous a servi de première méthode de référence pour construire des SME pour nos levés bathymétriques. L'objectif est de pouvoir comparer notre méthode de création de SME

avec une méthode reconnue de construction d'enveloppe spatiale dans le plan.

Nous avons appliqué cette méthode à des levés de différentes densités. Ainsi, pour les levés de densité constante comme illustré par la figure 4.5, la méthode fonctionne parfaitement et produit une SME très fidèle comme le met en évidence la distance de Haussdorf-Pompeiu [169] (notre critère quantitatif de comparaison des SME) calculée dans la table 4.1, qui respecte les conditions décrites dans la section 4.1 et correspond également aux attentes des opérateurs en terme d'ergonomie et de représentativité.

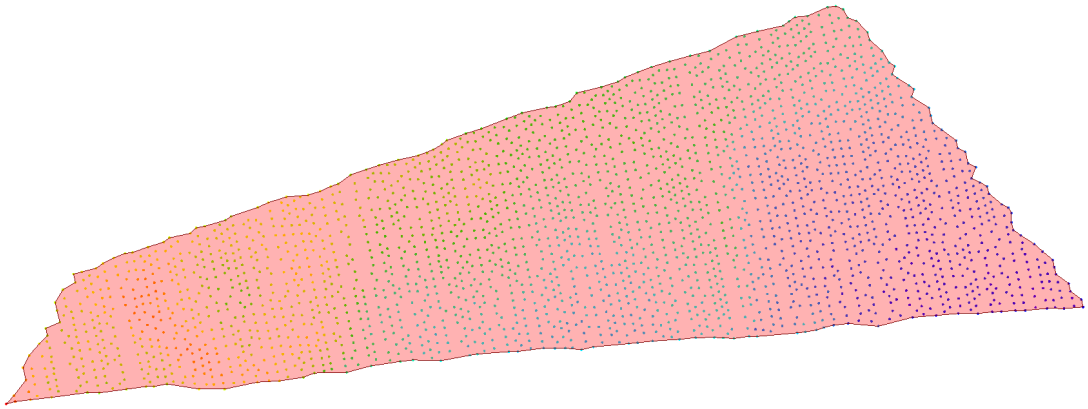

FIGURE 4.5 – Utilisation de l'algorithme $\alpha$-shape sur le lot S202099900-001.

Néanmoins, cette méthode comporte des limites lorsque la densité du nuage de points n'est pas constante pour l'entièreté du jeu de données. La figure 4.6 montre ainsi les limites d'utilisation pour la construction de la SME pour des lots bathymétriques dont la densité varie de façon importante, notamment en fonction de la profondeur mesurée. On voit ainsi sur les zooms de la figure 4.6 que les trous dans la couverture sont parfois mal détectés et représentés par l'algorithme $\alpha$-shape. Pour y remédier, de nombreuses itérations de cet algorithme seraient nécessaires afin de définir le paramètre $\alpha$ optimal mais également de nombreuses reprises manuelles par les hydrographes en charge de la construction de la SME.

Cette limite principale est déjà connue dans la communauté scientifique et des méthodes sont utilisées pour répondre à ces contraintes comme dans l'article [170] qui propose un paramètre $\alpha$ adaptatif local en fonction de la densité du nuage de points. L'article [171] propose également une méthode dont l'objectif est de déterminer le paramètre $\alpha$ optimal à utiliser dans le cadre de la reconnaissance des formes en 2D. Une seconde limitation a également été relevée au cours des premiers

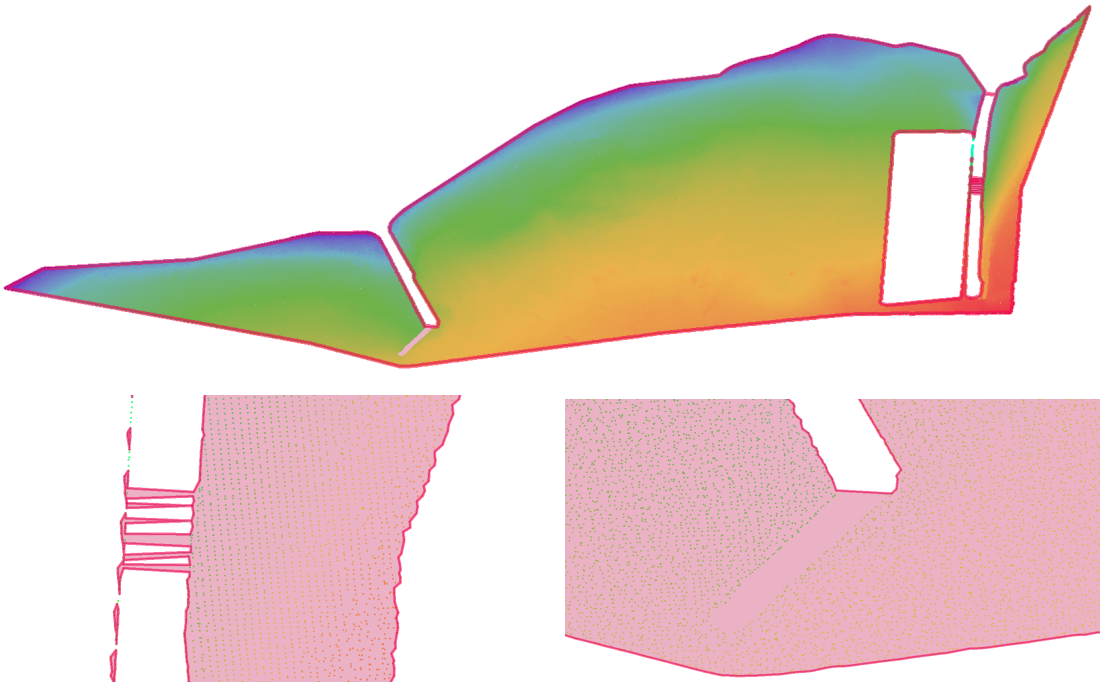

FIGURE 4.6 – Utilisation de l'algorithme $\alpha$-shape sur le lot E201804100-002 dans la partie supérieure et deux zooms sur des zones complexes dans la partie inférieure.

essais : le temps de calcul associé à cette méthode. Le temps de traitement d'un lot de données bathymétriques d'environ 10 millions de sondes met ainsi plus de 2 heures (résultats sur différents volumes de données présentés par la figure 4.14 sur PC windows 10, Intel 2.50GHz et 96 Go RAM).

### 4.1.2 Méthode pragmatique QuadSME

Pour répondre à ces contraintes liées au paramétrage et au temps de traitement important de l'algorithme $\alpha$-shape, nous proposons dans cette sous-section une méthode basée sur deux décompositions quadtree [172] afin, dans un premier temps, de partitionner un jeu de données dans des sous-ensembles homogènes, puis, dans un second temps, d'extraire la SME de ces sous-ensembles via une triangulation de Delaunay [168]. Au sein du Shom, cette méthode est dénommée QuadSME.

**Quadtree**

Le quadtree [172] (ou arbre quaternaire) est une structure de données et une méthode de partitionnement d'un espace plan par subdivision récursive en quatre sous-espaces de même dimension. Cet algorithme construit ainsi un arbre, d'où est tiré son nom, pour lequel chaque nœud possède quatre feuilles. La subdivision se termine une fois un critère d'arrêt atteint (comme la profondeur maximale de l'arbre).

La figure 4.7 présente un exemple de décomposition quadtree en trois niveaux. Les sous-quadrants 2, 3 et 4 n'ont pas été subdivisés car le critère d'arrêt a été atteint, ce qui n'a pas été le cas pour le sous-quadrant 1 qui a lui-même été décomposé en 4 sous-quadrants.

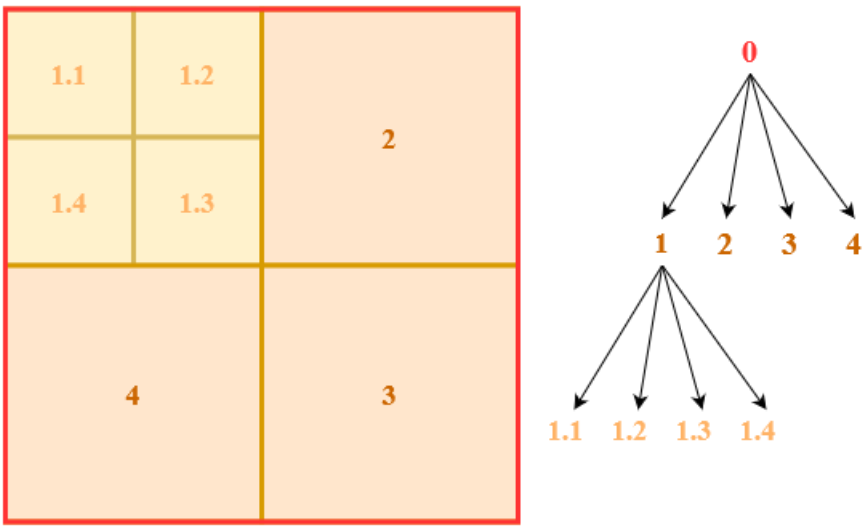


FIGURE 4.7 – Partitionnement quadtree sur 3 niveaux et son arbre associé.

### Présentation de l'approche QuadSME appliquée à la donnée bathymétrique

L'algorithme prend en entrée un nuage de points bathymétriques et l'incertitude horizontale, l'attribut *POSACC*, associée aux données de profondeur (provenant des métadonnées du levé étudié). L'approche fonctionne à la fois pour des

coordonnées géographiques ou bien projetées. La figure 4.8 présente la méthodologie développée pour calculer la SME (représentant l'enveloppe convexe d'un nuage de points bathymétriques). Cette approche a fait l'objet d'une présentation lors d'une conférence internationale [18] et a été abordée dans la publication suivante [15].

1. La première étape de cette méthodologie est l'importation de la donnée sous forme de nuage de points comprenant obligatoirement un triplet (x, y, z) qui donne la position d'un point dans l'espace, et la valeur du *POSACC* (variable en fonction de la profondeur) issue des métadonnées. Les données doivent être encodées en ASCII [173] et différentes extensions du fichier d'entrée sont acceptées (.txt, .ascii, .xyz, .glz/.lgz, .csv). Les coordonnées des sondes peuvent être dans un système projeté (nord/est/profondeur) ou géographique (latitude/longitude/profondeur) mais avec une conversion degré vers métrique pour les calculs de distance en prenant en compte la latitude moyenne des sondes.
2. La seconde macro-étape de la méthodologie consiste en une première indexation quadtree [172]. Via cette première structuration, nous nous assurons de manipuler un nombre acceptable de points dans la suite du processus par la machine informatique utilisée. Le critère d'arrêt du quadtree est donc le nombre maximal de points par quadrant fixé à 5 millions de points : ce paramètre est modifiable en fonction des capacités de la machine qui devra lancer l'algorithme. Cette macro-étape comporte ainsi la création du premier quadrant, le calcul du nombre de points dans ce quadrant et la décomposition quadtree en fonction du critère d'arrêt.
3. La troisième macro-étape de la méthodologie consiste en une seconde indexation quadtree. L'espace est découpé pour ne conserver que des quadrants validant le critère de densité ou bien des quadrants comportant un maximum de 1000 sondes. Cette macro-étape est constituée du calcul du critère de densité dans le quadrant et du processus itératif quadtree (fonction de critère d'arrêt de densité ou du nombre maximal de points). Le critère de densité correspond au nombre de points dans chaque sous-quadrant constituant un quadrant principal. Si ce nombre est identique (jugé grâce à un seuil) pour chaque sous-quadrant alors le quadrant principal est considéré comme homogène en termes de répartition du nombre de sondes et

le critère de densité est validé. Un paramètre de seuil modifiable gère l'homogénéité de ces sous-quadrants (ratio du nombre de sondes dans chaque sous-quadrant et du nombre total de points divisé par quatre).

4. La quatrième macro-étape de la méthodologie consiste en la génération des polygones contenant les sondes. Dans un premier temps, une résolution caractéristique du nuage de points, compris dans le sous-quadrant, est calculée afin de s'adapter aux potentielles différences de densité du nuage de points d'entrée. La résolution correspond à la moyenne de la distance moyenne du point d'étude et de ses 4 plus proches voisins pour les 500 premiers points (ou tous les points si leur nombre est inférieur à 500) constituant le sous-quadrant. Ensuite, un partitionnement et une détection des points isolés sont réalisés via la méthode DBSCAN [60], présentée dans la section 2.3. L'objectif est de construire des enveloppes spécifiques pour les points isolés comme indiqué dans la section 4.1 et de construire des regroupements de points de même densité avant la création des polygones. Enfin, on réalise une triangulation de Delaunay [168] sur les différents regroupements et on en extrait le polygone associé.
5. La cinquième et dernière macro-étape de la méthodologie consiste en une fusion des polygones générés lors des étapes précédentes afin de former la SME finale. Les géométries sont fusionnées via un processus de dilatation/érosion (création d'un *buffer*) des géométries pour supprimer des trous de construction. Cette étape est basée sur la valeur du $POSACC$ ou de la résolution caractéristique du lot global afin de respecter la condition de densité et représentativité.

### Application à des levés bathymétriques et comparaison avec la méthode $\alpha$-shape

Nous avons appliqué cette méthode à des levés de différentes densités. Ainsi, pour les levés de densité constante, comme en figure 4.9, l'algorithme $\alpha$-shape produit une SME très fidèle, au vu de la distance de Hausdorff-Pompeiu présentée dans la table 4.1, qui respecte les conditions décrites dans la section 4.1 et correspond également aux attentes des opérateurs. Pour ce lot S202099900-001, aucun quadrant n'a été conservé car le nombre de sondes (2743) est trop faible et

le critère de densité n'a pas été respecté dans le premier quadrant. Le processus QuadSME a donc réalisé la seconde indexation (étape 3 de 4.8) puis a directement procédé à la création des polygones d'emprise spatiale (étape 4 de 4.8) pour les

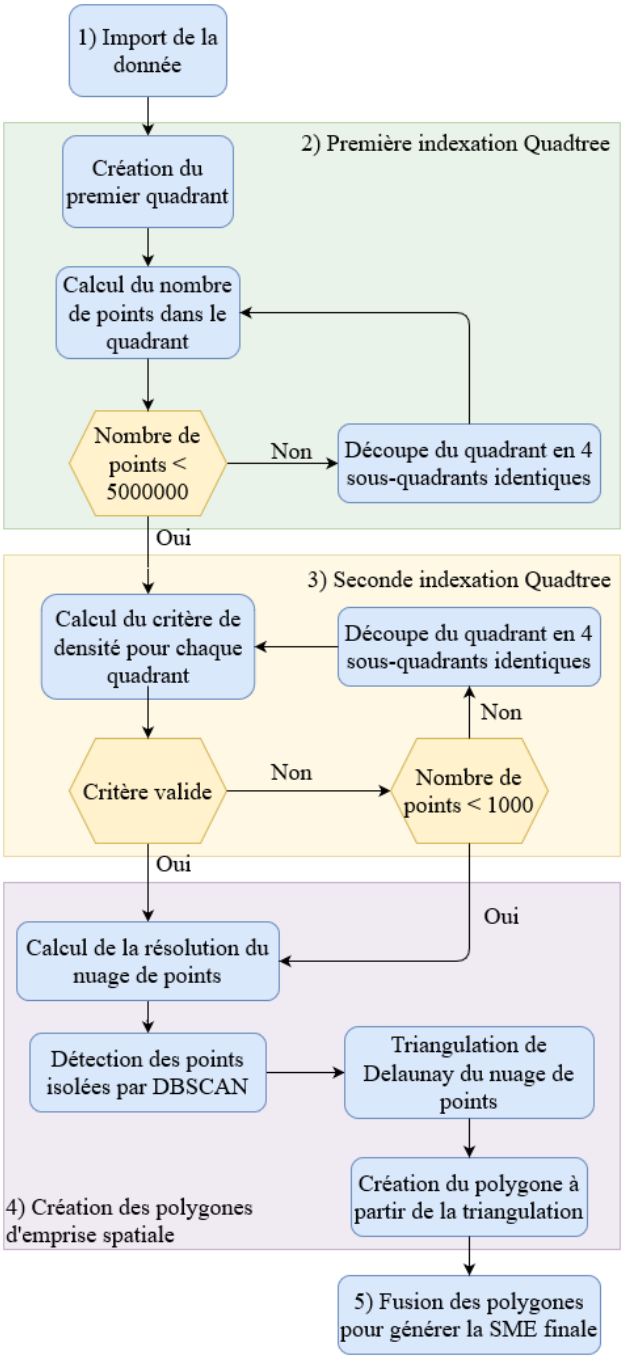


FIGURE 4.8 – Processus global de la méthode QuadSME.

quatre sous-quadrants de niveau 2 (voir 4.7) avant la fusion de ces polygones.

FIGURE 4.9 – Utilisation de la méthode QuadSME pour le lot S202099900-001.

Sur des levés plus importants en nombre de sondes, le résultat est plus performant que la méthode $\alpha$-shape notamment dans la prise en compte des trous, comme le montre la figure 4.10 pour le lot E201804100-002 (128940 sondes). Nous pouvons également observer les quadrants conservés lors de la seconde indexation quadtree (étape 2 de 4.8). Ces images montrent une SME beaucoup plus fidèle que l'algorithme $\alpha$-shape.

La figure 4.11 montre un dernier exemple de la génération d'une SME avec l'algorithme $\alpha$-shape et la méthode QuadSME sur un lot S2001701200-1 très complexe issu d'une acquisition LiDAR. Ce type de levé génère souvent une très forte disparité dans la densité des sondes acquises surtout lorsque les sondes sont intégrées dans un lot unique. Nous voyons ainsi que l'algorithme $\alpha$-shape ne respecte pas les conditions de la SME présentées dans la section 4.1, alors que la méthode QuadSME est beaucoup plus représentative du lot de données bathymétriques. Sur des lots aussi complexes, il reste des reprises manuelles à effectuer par les opérateurs avant que la SME n'intègre la BDBS.

La figure 4.12 montre un zoom sur le lot S200701200-1. Nous pouvons ainsi voir que la détection des trous est beaucoup plus représentative avec la technique QuadSME qu'avec la méthode $\alpha$-shape.

Afin de comparer les SME générées avec une métrique quantitative, nous calculons la distance de Hausdorff-Pompeiu [169] entre les résultats des méthodes proposées ($\alpha$-shape et QuadSME) et une SME validée manuellement par un opérateur

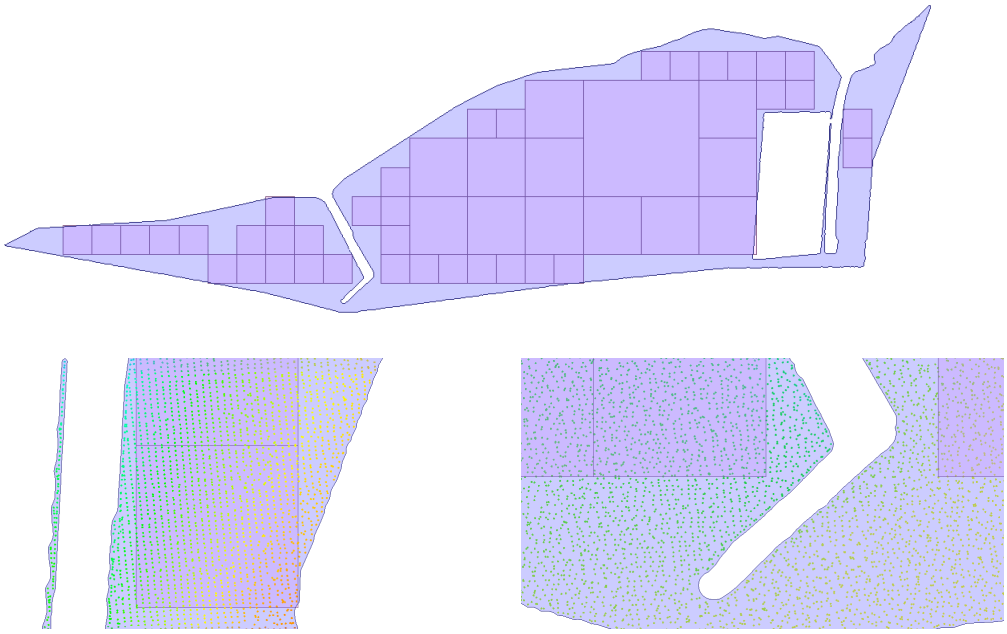

FIGURE 4.10 – Utilisation de la méthode QuadSME pour le lot E201804100-002 sur la partie supérieure et deux zooms sur des zones complexes dans la partie inférieure. Les quadrants représentent les zones vérifiant le critère de densité de l'algorithme QuadSME.

afin de respecter les trois conditions énoncées plus haut. Cette SME opérateur est considérée comme la SME de référence. Nous avons utilisé cette métrique car la distance de Hausdorff-Pompeiu est un outil topologique qui mesure l'éloignement de deux sous-ensembles d'un espace métrique (permet ainsi de vérifier la condition de densité et représentativité). Elle est donc tout à fait indiquée pour comparer l'écart maximal entre deux emprises spatiales. Cette distance permet ainsi de mesurer la dissimilarité de deux formes. Soit $X$ et $Y$ deux sous-ensembles non-vides d'un espace métrique $(E, d)$ où $E$ est un ensemble non vide et $d$ une distance sur $E$ (qui vérifie les propriétés de symétrie, séparation et l'inégalité triangulaire). On définit la distance de Hausdorff-Pompeiu de $X$ et $Y$, notée $d_H(X, Y)$ par :

$$d_H(X,Y) = max\Big\{sup_{x\in X} d(x,Y), sup_{y\in Y} d(X,y)\Big\} \tag{4.3}$$

Avec $d(a, B) = inf_d(a, b)$, quantifiant la distance du point $a$ ($a \in X$) au sous-ensemble $B \subseteq X$, $sup$ représentant le supremum et $inf$ l'infinum. On a la propriété suivante $d_H(X, Y) = 0$ si et seulement si $X$ et $Y$ ont la même emprise spatiale (propriété de séparation de la distance).

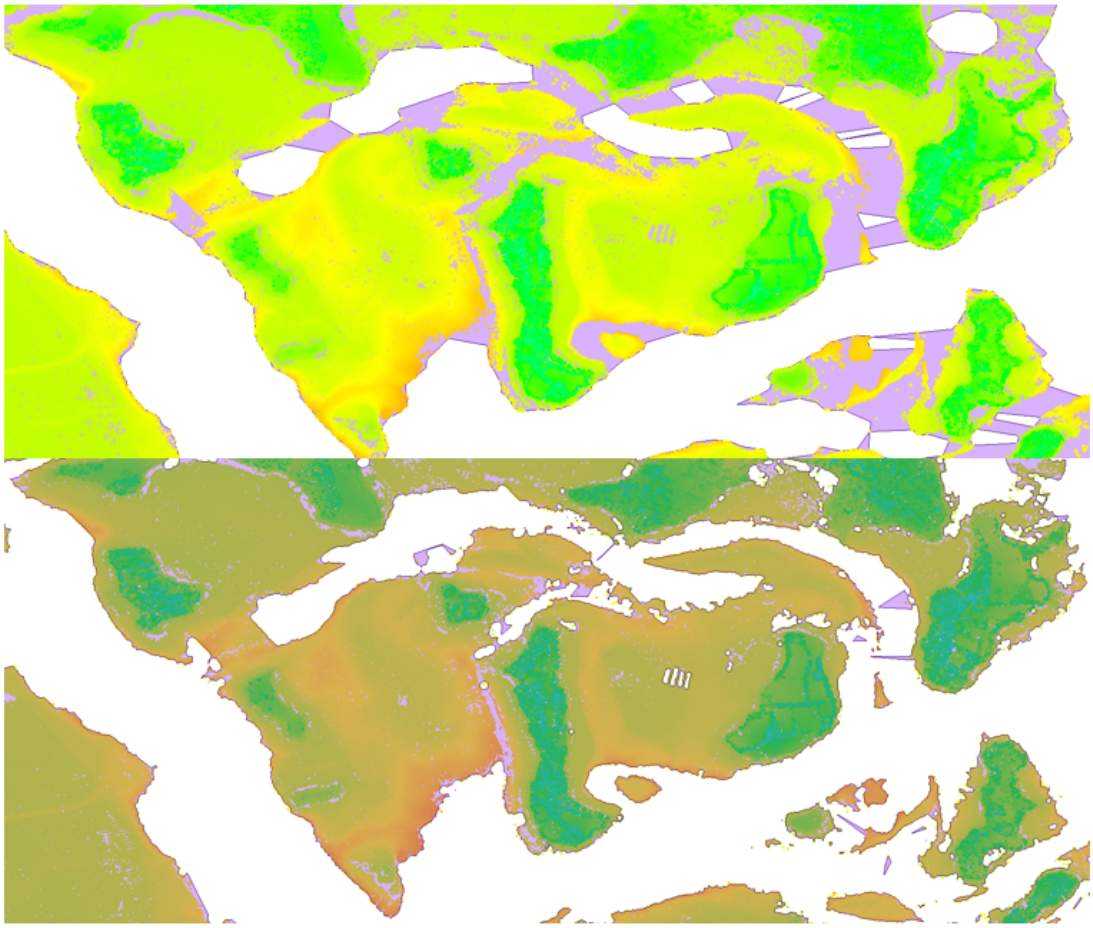

FIGURE 4.11 – Utilisation de l'algorithme $\alpha$-shape (partie supérieure) et de la méthode QuadSME (partie inférieure) sur le lot S200701200-1 représenté par les points jaunes.

La figure 4.13 présente la distance pour deux ensembles $A$ et $B$. $A$ représente le carré vert et $B$ le disque bleu de même surface et de même centre. À l'endroit où les deux figures coïncident, la couleur est bleue. Sinon, elle est verte ou rouge. Les différences entre les deux figures se matérialisent sous la forme de 4 lunules rouges et 4 presque triangles verts. On considère le point du disque le plus éloigné du carré : c'est le sommet de la lunule et sa distance au carré est notée $d_1$. On considère ensuite le point du carré le plus éloigné du disque, qui est un sommet du carré, à une distance $d_2$ du disque. La distance de Hausdorff est la plus grande valeur des deux, $d_H(X, Y) = max(d_1, d_2)$, en l'occurrence $d_2$, pour l'exemple choisi. Les valeurs $d_1$ et $d_2$ sont appelées distance de Hausdorff relative.

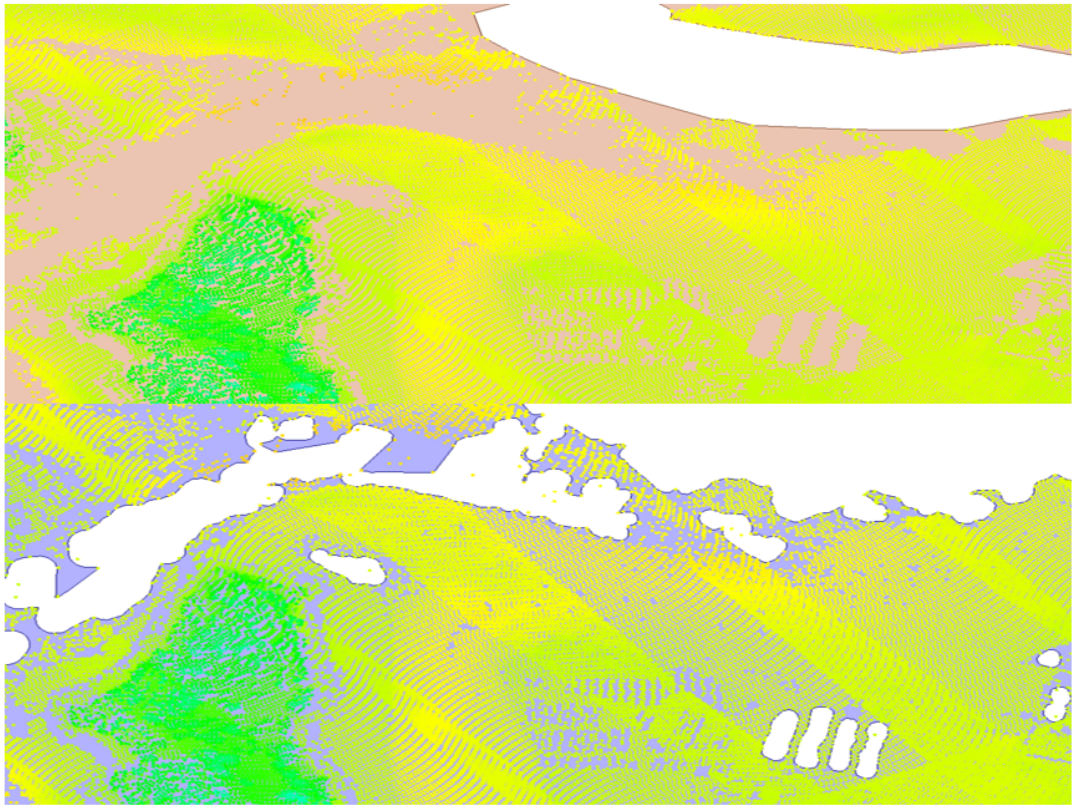

FIGURE 4.12 – Zoom sur le lot S200701200-1, $\alpha$-shape pour la partie supérieure et QuadSME pour la partie inférieure.

TABLE 4.1 – Calcul de la distance (en mètre) de Hausdorff-Pompeiu pour 5 lots de données bathymétriques.

| Nom du lot | Nombre de sondes | Distance $\alpha$-shape | Distance QuadSME |
|---|---|---|---|
| S202099900-001 | 2743 | 165.6 | 141.9 |
| S201207000-5 | 25718 | 86.7 | 76.5 |
| E201804100-002 | 128939 | 16.6 | 0.9 |
| S202102500-001 | 1070131 | 182.9 | 25.1 |
| S200701200-1 | 10829541 | 1340.3 | 118.3 |

La table 4.1 présente les résultats du calcul de cette distance pour 5 levés de densités différentes. On peut observer que la distance de Hausdorff-Pompeiu pour la méthode QuadSME est toujours plus faible et donc plus fidèle à la SME de

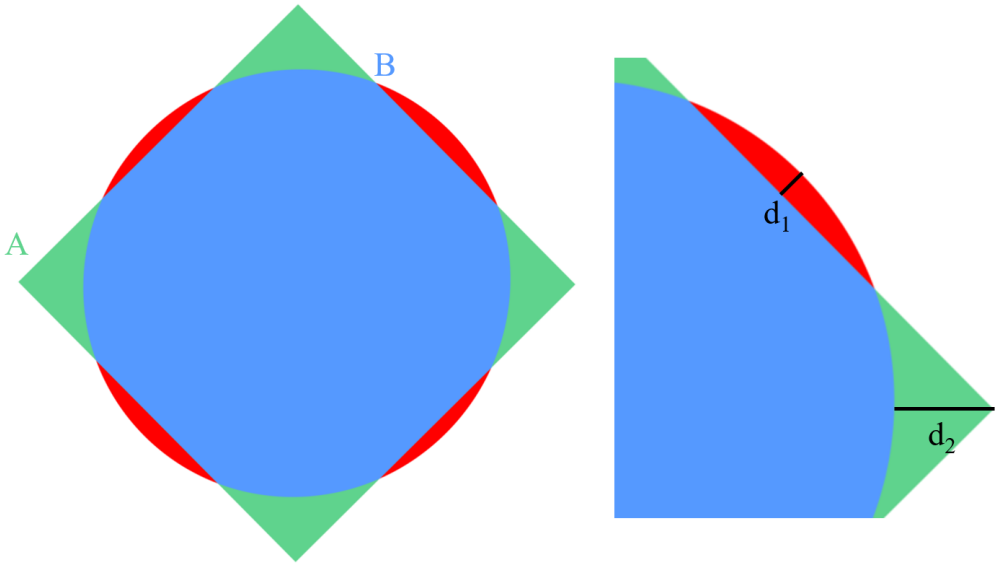


FIGURE 4.13 – Distance de Hausdorff entre deux ensembles $A$ et $B$.

référence, qu'avec la méthode $\alpha$-shape, mais elle reste du même ordre de grandeur pour les lots avec peu de sondes (les deux premiers levés bathymétriques du tableau 4.1). De plus, pour les lots au volume plus important (les trois derniers levés bathymétriques du tableau 4.1, la méthode QuadSME est meilleure d'un ordre de grandeur : la SME est toujours aussi fidèle. Par contre, la SME générée par la méthode $\alpha$-shape est, elle, beaucoup moins fidèle comme ce que nous avions déjà pu observer dans les figures 4.11 et 4.6.

Nous comparons ensuite le temps de calcul associé à chacune des méthodes. La méthode QuadSME, voir figure 4.14, affiche des temps de calcul plus performants que l'algorithme $\alpha$-shape, surtout pour un nombre de points supérieur au million qui correspond au nombre de sondes minimum à traiter pour des lots de données bathymétriques. Nous constatons ainsi que, pour un lot de sondes de l'ordre de grandeur de dix millions de sondes, la méthode QuadSME est quarante fois plus rapide que l'algorithme $\alpha$-shape. Il est possible d'améliorer encore le temps de traitement en réalisant un multiprocessing de la méthode QuadSME. En effet, toutes les opérations réalisées sur les différents quadrants, macro-étapes 2 à 4 de la figure 4.8, sont indépendantes et peuvent donc être réalisées en parallèle. En revanche, le temps de calcul de la méthode QuadSME sera très dépendant de l'homogénéité de la répartition des sondes. Ainsi, si le critère de densité est

rapidement atteint alors le processus de quadtree se stoppera. À l'inverse, si la répartition des sondes n'est pas homogène dans les sous-quadrants ou que la donnée comporte de nombreux trous, alors le temps de calcul sera plus long car il faudra aller jusqu'au bout de la décomposition quadtree.

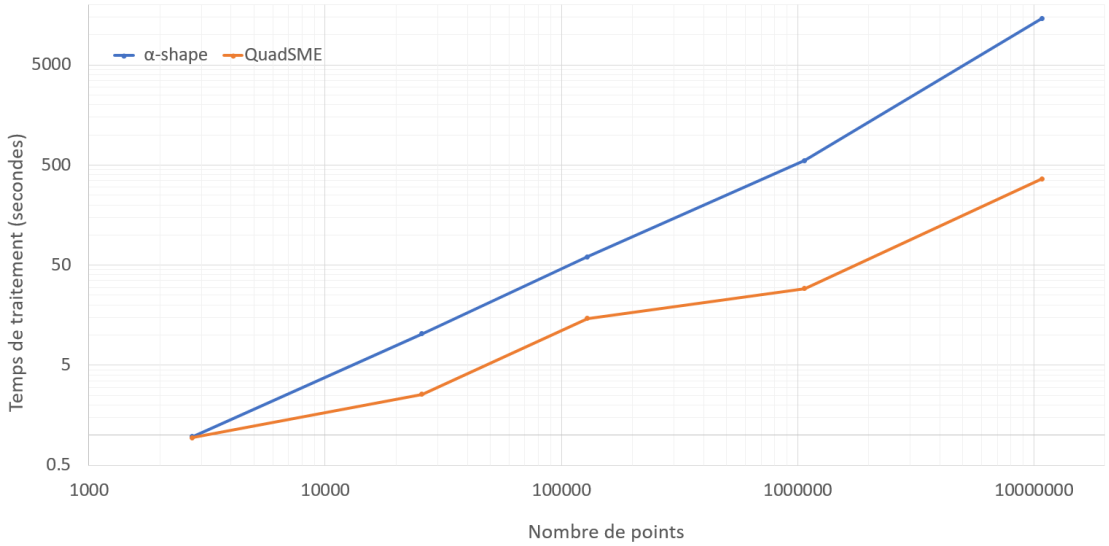


FIGURE 4.14 – Temps de traitement en secondes en fonction du nombre de points et de la méthode utilisée (échelle logarithmique pour l'axe des abscisses et ordonnées).

Cet algorithme a été développé en Python et s'intègre dans un processus opérationnel d'automatisation de traitement de l'information à l'aide de l'Extract-Transform-Load (ETL) FME®, outil de transformations massives d'information d'une source de données vers une autre. Il est important de noter que, bien que ce polygone final simplifie la géométrie de la couverture (à cause du processus de dilution / érosion réalisé à la fusion finale), il est fondamental qu'aucune sonde ne soit filtrée (en dehors de la SME) par ce processus, afin de valider la condition d'unicité décrite plus haut. En effet, la profondeur associée à une sonde filtrée pourrait devoir être affichée sur la carte et être critique pour la sécurité de la navigation.

Une fois la SME générée et l'emprise spatiale la plus fidèle possible d'un levé connue, nous pouvons passer à la seconde étape qui consiste à extraire les paramètres de qualité d'un levé et évaluer sa qualité globale.

## 4.2 Décrire la qualité d'un levé

L'évaluation objective de la qualité d'un levé hydro-océanographique tout en tenant compte des besoins des utilisateurs, ainsi que son automatisation, est un problème complexe et critique pour la bonne maîtrise de la donnée. En parallèle, la qualité de la donnée et des informations est habituellement examinée à travers le spectre de multiples paramètres (ou critères) de qualité. Le choix de ces paramètres, ainsi que leur importance relative, dépend bien souvent de l'utilisation finale des données. Ces paramètres peuvent être également influencés par le contexte ou l'environnement (voir [174]), qui à leur tour peuvent être perçus différemment selon les utilisateurs et, par conséquent, doivent être modélisés pour chaque concept d'emploi particulier. Ces observations ont constitué le fil conducteur de cette section : comment automatiser l'évaluation de la qualité des levés hydro-océanographiques en intégrant à la fois le besoin des utilisateurs finaux ainsi que le contexte de la donnée ? Et comment, dans le même temps, l'outil peut-il conserver un haut niveau de traçabilité du processus d'évaluation et une bonne explicabilité du résultat afin d'aider la prise de décision concernant les levés hydro-océanographiques ?

Pour répondre à ces questions, nous exploitons des techniques d'analyse de la qualité de la donnée afin de déterminer automatiquement, à partir de la donnée et de sa métadonnée associée, différentes mesures de qualité des levés hydro-océanographiques. Puis, nous combinons ces techniques avec des modèles de préférences issus de l'aide multicritère à la décision (MCDA, en anglais) permettant d'intégrer différents profils experts dans la procédure d'évaluation. Nous démontrons, entre autres, que ces multiples critères de qualité, combinés avec différentes visions métiers sur ces données, peuvent conduire à des évaluations globales de la qualité des levés hydro-océanographiques. Pour terminer, nous présentons comment la combinaison de ces outils peut conduire à des informations explicables en sortie de modèle, qui peuvent ensuite être exploitées pour réaliser des évaluations de levés actuels et améliorer la prescription et préparation de futurs levés.

### 4.2.1 La qualité de la donnée pour l'hydro-océanographie

Comme présenté dans la section 1.1, de nombreux capteurs sont présents tout au long de la chaîne d'acquisition de la donnée bathymétrique. Les capteurs, comme des SMF, des sonars, des IMU, des caméras optiques aéroportées et du LiDAR ba-

thymétrique, permettent de détecter, identifier, classifier et cartographier des zones d'intérêt. L'étude de tels phénomènes hydro-océanographiques produit des flux de données importants, bruités, redondants, parfois hétérogènes et contradictoires. Les caractéristiques de ces flux dépendent des caractéristiques du capteur, des paramètres d'acquisition opérationnelle et de phénomènes exogènes non contrôlables. De fait, l'évaluation de la qualité de la donnée est primordiale avant son exploitation pour les différents types de produits au sein du Shom. Des informations de mauvaise qualité peuvent ainsi induire des biais ou des erreurs ayant parfois des impacts critiques pour ces même produits (MNT, cartes nautiques, produits défense...).

À ce titre, l'évaluation de la qualité des levés hydro-océanographiques reste un problème complexe en raison de l'impossibilité de réunir les conditions parfaites d'acquisition mais également l'effort mis en place durant l'acquisition. Cela peut générer de multiples imperfections dans les jeux de données acquis. Il est nécessaire alors d'adapter les critères de la qualité au type de capteur utilisé, aux chaînes de traitement qui varient dans les données manipulées, au concept d'emploi associé (océanographie, cartographie marine, acoustique sous-marine...) et, avant tout, au contexte/environnement d'acquisition de cette donnée. En outre, il faut pouvoir estimer ou calculer la qualité de la donnée à chaque étape de la chaîne de traitement mais également comprendre l'impact de la qualité de la donnée pour son exploitation finale. Généralement, la compatibilité avec le concept d'emploi est une notion largement acceptée dans la caractérisation globale de la qualité de la donnée et des informations. Cette notion englobe la définition des paramètres de la qualité, liés à l'analyse du contexte opérationnel et la typologie de la qualité associée. Toutefois, vu la difficulté associée à la réalisation de mesure automatique de la qualité de la donnée (*i.e.* indisponibilité des valeurs vérité terrain - *ground truth* et des infrastructures nécessaires pour le réaliser en temps réel), une partie de ce travail est réalisée manuellement par les hydrographes qui assignent des estimations globales qualitatives, fondées sur leur expérience métier, des protocoles d'évaluation de la donnée et leurs connaissances des caractéristiques du produit final attendu. Cette évaluation consiste à retirer les erreurs et les anomalies évidentes, en conservant ou rejetant les sondes au lieu de la donnée à qualifier, en évaluant la fiabilité des données à travers des échantillons du jeu de données et en assignant des niveaux estimés de qualité de la donnée selon une référence prédéfinie. De plus, la donnée

peut également contenir des jeux de valeurs imparfaites pouvant être catégorisées dans cette typologie de la qualité :

— erronées : les valeurs associées à la donnée diffèrent de la valeur vraie.
— incomplètes : les données fournies le sont partiellement, il manque des valeurs.
— biaisées : les données manquent de justesse, la valeur moyenne s'éloignent de la valeur vraie.
— incertaines : la donnée ne peut être spécifiée avec une confiance absolue car elle manque de fidélité.
— indisponibles : le système ne peut obtenir certains jeux de données à cause de ses limitations.

Selon le cas d'utilisation, ces imperfections n'auront pas le même impact sur la validation finale de la donnée. Il est donc nécessaire d'adapter des profils particuliers de qualité aux préférences utilisateurs et aux exigences de leur exploitation finale.

### 4.2.2 L'aide multicritère à la décision au service de la qualité

Comme présenté dans la section précédente, l'évaluation de la qualité de l'information ou de la donnée de levés hydro-océanographiques est un problème de nature multi-dimensionnelle et fonction du concept d'emploi final de la donnée. Il est donc pertinent d'utiliser une agrégation de techniques issues du domaine MCDA, permettant de collecter de multiples paramètres, tout en prenant en compte le point de vue (appelé préférences) de l'utilisateur final pour répondre à la problématique de détermination de la qualité d'un levé hydro-océanographique. Le MCDA est une branche de la recherche opérationnelle dont l'objectif est d'aider un ou plusieurs décideurs à préparer et prendre des décisions (ou alternatives) avec un ensemble fini de paramètres, en tenant compte des paramètres contradictoires (ou critères). Ce décideur peut être la personne qui prend la responsabilité de la décision ou bien l'utilisateur expert dont le système de valeurs ou préférences sera pris en compte dans la prescription finale.

Habituellement, trois types de problèmes sont mis en avant dans ce domaine :

— le problème du choix qui vise à recommander le sous-ensemble le plus réduit

possible d'alternatives contenant celles qui sont satisfaisantes (exemple : préconiser des véhicules en fonction de l'usage pour une famille) ;
— le problème du tri qui vise à assigner à chaque alternative une catégorie ou classe prédéfinie (exemple : classer des véhicules en fonction de leur consommation énergétique) ;
— le problème du classement qui vise à assigner un ordre aux alternatives par ordre décroissant de préférences (exemple : trier des véhicules par ordre de préférence en fonction de critères comme le prix, la couleur ou l'empreinte écologique).

L'outil mathématique utilisé en MCDA tire principalement son origine de deux tendances méthodologiques très différentes. D'une part, il y a le paradigme européen qui s'est développé autour du concept de sur-classement des relations, où la recommandation de décision est construite par comparaison en paire d'alternatives. D'autre part, il y a le paradigme anglo-saxon qui se fonde sur le concept d'utilité ou valeur dans la théorie des valeurs à multi-attributs (*Multi-Attribute Value Theory* - MAVT, en anglais) afin d'obtenir, par agrégation, une comparabilité complète des alternatives. Les avantages et des inconvénients de ces deux paradigmes seront comparés avant de choisir le plus pertinent pour notre cas d'étude. La différence principale entre ces deux méthodologies repose sur la manière dont les alternatives sont comparées et sur le type d'informations requises pour le décideur. Les méthodes de sur-classement conviendraient plus à l'évaluation hétérogène d'alternatives de critères, i.e. qualitative et quantitative, et si le décideur souhaiterait inclure des imprécisions quant à ses préférences dans le modèle. Les méthodes fondées sur les valeurs peuvent être préférées si le critère ne repose principalement que sur des critères quantitatifs et si un comportement compensatoire (approche anglo-saxonne) du décideur doit être modélisé.

L'approche mise en place dans le cadre de l'évaluation de la qualité des levés hydro-océanographiques avec la prise en compte des préférences des experts est présentée dans la figure 4.15. Elle a fait l'objet d'une publication [15] et d'une présentation lors d'une conférence internationale [22]. La méthodologie MCDA utilisée dans le cadre de ce processus est présentée dans la sous-section 4.2.2.

Ce processus est destiné aux utilisateurs qui souhaiteraient évaluer la qualité d'un levé hydro-océanographique en fonction d'un besoin spécifique métier, par exemple, des besoins acoustiques, océanographiques ou hydrographiques, domaines

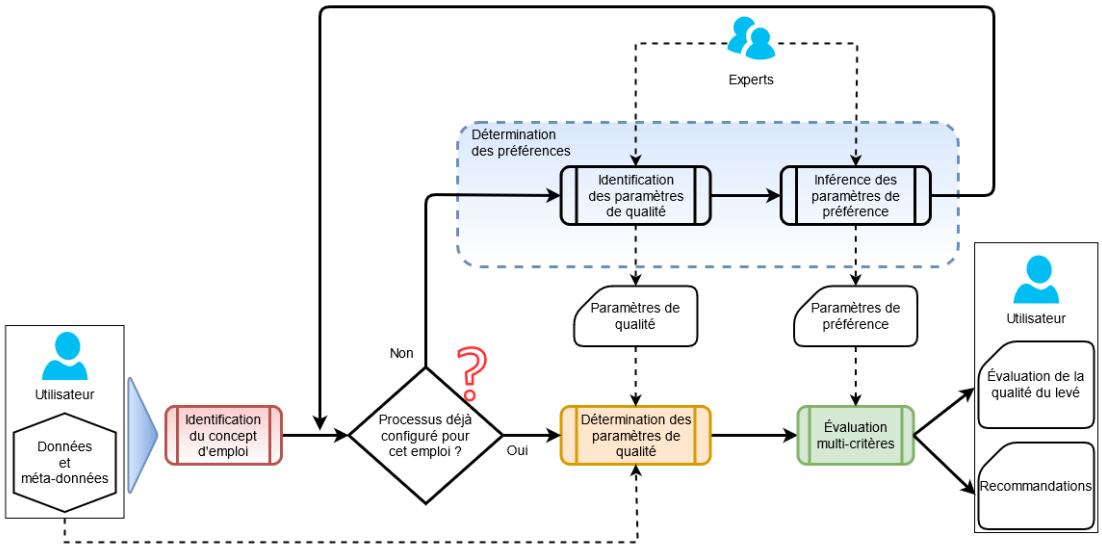


FIGURE 4.15 – Processus d'évaluation de la qualité des levés hydrographiques.

d'expertises qui ont été étudiés dans le cadre de la thèse, voir les tables 4.2, 4.3 et 4.4. La donnée et la métadonnée associée aux levés hydro-océanographiques constituent également les intrants de ce processus. Le prérequis de la chaîne de traitement est qu'elle a besoin d'être configurée en amont par un expert afin de déterminer les préférences associées à son profil d'expert (bloc *détermination des préférences* sur la figure 4.15). Les données de sortie(s) sont l'évaluation partielle et globale des données d'entrée du levé hydrographique, ainsi que les recommandations associées (si nécessaire) pour améliorer la qualité du levé hydro-océanographique évalué mais également pour de futures acquisitions.

Une fois que l'utilisateur a débuté le processus d'évaluation de la qualité du levé, il doit en premier lieu *identifier le concept d'emploi* de ce levé (pour des besoins acoustiques ou hydrographiques par exemple). Si la chaîne de traitement a déjà été configurée (spécifiée en amont par un expert) pour cet objectif spécifique, l'étape *détermination des paramètres de qualité* extrait et calcule les paramètres de qualité depuis les données du levé hydro-océanographique et leurs métadonnées associées, étape présentée en sous-section 4.2.2. Ensuite, l'*évaluation multi-critères* (exploitant le modèle MCDA) est utilisée pour agréger les paramètres de qualité et ainsi réaliser l'évaluation globale de levé tout en prenant en compte le modèle de préférence de l'expert. Dans le cas où le processus n'a pas été configuré en amont pour l'objectif final souhaité par l'utilisateur, un expert métier est interrogé dans le sous-

processus, appelé *détermination des préférences*, afin de configurer tout d'abord l'étape d'*identification des paramètres de qualité* pour choisir les paramètres de qualité à considérer dans le cadre de ce profil. Puis, l'étape d'*inférence des paramètres de préférence* est configurée avec les paramètres de préférence définissant les priorités de l'expert. Une fois le système paramétré, le processus reprend à partir de l'étape *détermination des paramètres de qualité*, l'*évaluation multi-critères* et la fin du processus. Les sous-sections suivantes détaillent chacune des étapes principales de la méthodologie proposée.

### Détermination des préférences

Le sous-processus de détermination des préférences se décompose en deux étapes : l'*identification des paramètres de qualité* et l'*inférence des paramètres de préférence*. Nous rappelons que ce sous-processus ne concerne pas les utilisateurs mais qu'il est seulement adressé à un expert afin de configurer, selon sa connaissance, le processus d'aide multicritère à la décision. Durant la première étape, l'expert doit fournir des paramètres de qualité qu'il utilise dans l'évaluation globale de la qualité du levé hydro-océanographique. La donnée de sortie de cette étape est une liste de ces paramètres de qualité qui seront utilisés par la suite comme données d'entrée pour l'étape *détermination des paramètres de qualité*. La seconde étape est l'*inférence des paramètres de préférence* de la méthode d'évaluation multi-critères. Ces paramètres représentent les priorités de l'expert pour l'évaluation de la qualité des levés hydro-océanographiques.

### Identification des paramètres de qualité

Dans notre contexte des levés hydro-océanographiques, l'extraction des paramètres de qualité est réalisée à partir des données et métadonnées acquises pendant la mission. En effet, durant une campagne en mer, les hydrographes acquièrent un large volume de données dans des domaines d'utilisations divers tels que l'hydrographie, l'océanographie, la sédimentologie, l'acoustique sous-marine, la cartographique nautique, la production de MNT, les produits pour la défense, etc. Les capteurs mis en place sont, de fait, nombreux et dépendent du type de campagne mené et du type de données que nous souhaitons collecter dans la zone d'étude. La matrice présentée dans [8] permet ainsi de spécifier, en fonction d'un usage cible, très précisément le niveau de qualité attendu en fonction de la donnée collectée.

Dans le cas de la bathymétrie, le Shom utilise le SBES, le SMF, le LiDAR bathymétrique ou bien des données satellitaires (SDB) pour les fonds les plus proches des côtes. Comme présenté dans la section 1.1, ces données, acquises sous forme de nuages de points ou données maillées (notamment pour la SDB), permettent d'obtenir une représentation plus ou moins résolue des fonds suivant la technologie utilisée des capteurs et la couverture bathymétrique réalisée lors de l'acquisition. Si nous nous intéressons à la représentation des phénomènes physiques de l'océan, alors il est important de mesurer les courants de surface et les courants de la totalité de la colonne d'eau (via des courantomètres de fond ou via des profileurs de courant à effet doppler par exemple) mais également les variations de hauteurs d'eau (via des marégraphes à capteur de pression, acoustiques ou des bouées GNSS) pour mesurer les séries temporelles de marée à des localisations spécifiques. Dans le cadre des études sédimentologiques du fond marin, la connaissance de la bathymétrie est importante pour la détection locale et à basse résolution des ruptures morphologiques, mais également pour l'accès à la classification des fonds rencontrés (boue, roche, sable ...). Pour établir ce type de classification, beaucoup de capteurs différents sont nécessaires, comme les capteurs acoustiques (SBES et SMF), afin d'obtenir les informations morphologiques sous-marines (voir la figure 4.16), mais également des techniques d'échantillonnage *in situ* avec des bennes de prélèvement (type Shipek ou Van Veen) ou des tubes de carottage pour fournir des informations sur le type de fond marin à des positions spécifiques.

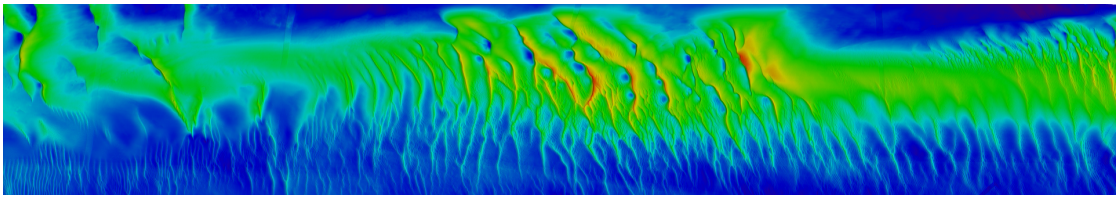

FIGURE 4.16 – MNT montrant des dunes sous-marines ©Shom.

Les hydrographes en charge des campagnes de levés sont familiers des conditions d'acquisition de données. Ils réalisent des calculs et sont donc capables d'assigner au jeu de données des qualifications quantitatives et qualitatives fiables de la qualité des données acquises, voir [70] et [8].

Dans le cadre de ce travail de recherche, nous avons interrogé des experts métiers du Shom afin de mieux comprendre la manière dont est qualifiée la donnée. Nous avons ainsi sélectionné dix paramètres de qualité choisis par les experts. Ces

paramètres sont facilement accessibles dans les levés hydro-océanographiques mais également bien spécifiques aux besoins des experts. Ces paramètres de qualité sont : *CATZOC*, *POSACC*, *SOUACC*, *Couv*, *Capt*, *SEDIM*, *OCEANO*, *Durée* (pour une série temporelle de hauteur d'eau), *Trous* (pour une série temporelle de hauteur d'eau) et *Erreurs* (pour une série temporelle de hauteur d'eau). Le *CATZOC*, le *POSACC* et le *SOUACC* font partie de la norme S-57 [40]. Ces paramètres de qualité sont décrits de la façon suivante :

— *CATZOC* pour la catégorie de zone de confiance (*CATegory of ZOne of Confidence* en anglais). Ce paramètre qualitatif fournit une information sur la confiance globale accordée à la donnée d'un levé bathymétrique dans un cadre d'utilisation cartographique. Les valeurs catégoriques classées de la meilleure à la pire qualification sont : A1, A2, B, C, D et U (pour *Unclassified*) [40]. La typologie (voir la section 4.2.1) de la qualité associée est l'incertitude de la donnée.

— *POSACC* qualifie l'incertitude du positionnement horizontal de la mesure bathymétrique (*POSitional ACCuracy* en anglais). Ce paramètre quantitatif en mètre est seuillé par quatre intervalles de profondeur extrait de la norme [8] : $< 1$m, $< 2$m, $< 5$m + 5% de la profondeur, $< 20$ + 10% de la profondeur. Le paramètre de qualité *POSACC* est traditionnellement obtenu par la propagation de l'incertitude totale prenant en compte l'ensemble de la chaîne d'acquisition bathymétrique, voir [70]. La typologie de la qualité associée est l'incertitude et le biais associé à la donnée.

— *SOUACC* qualifie l'incertitude verticale sur la mesure bathymétrique (*SOUnding ACCuracy* en anglais). Ce paramètre quantitatif en mètre est seuillé par quatre intervalles de profondeur inspirées de la norme [8] : $< 0.25$m, $< 0.50$m, $< 1$m, $< 2$m, $< 10$m, $< 20$m. Tout comme le paramètre de qualité *POSACC*, le *SOUACC* est obtenu par la propagation de l'incertitude totale prenant en compte l'ensemble de la chaîne d'acquisition bathymétrique, voir [70]. La typologie de la qualité associée est l'incertitude et le biais associé à la donnée.

— La *Couv*, pour couverture bathymétrique [8], est une mesure quantitative en pourcentage composée de six valeurs : 300, 200, 150, 100, 80, et 50. Elle correspond au pourcentage de couverture calculé à partir des différentes lignes de levés et de leurs recouvrement respectifs. La typologie de la qualité

associée est la complétude.

— Le *Capt*, pour définir le type de capteur hydrographique, est un paramètre qualitatif représenté par quatre classes de capteurs : SMF, LiDAR bathymétrique, SBES + sonar latéral, et SDB. Le capteur influe sur la densité et la sensibilité des mesures bathymétriques acquises. La typologie de la qualité associée est la complétude, la disponibilité et l'incertitude.

— *SEDIM*, paramètre qualitatif, représente le nombre de capteurs sédimentologiques de 0 à 5 utilisés pour décrire les substrats des fonds marins. Ces substrats dépendent de la réflexion des signaux acoustiques émis. La type de qualité associée est la complétude et l'incertitude.

— *OCEANO* paramètre qualitatif, représente le type de donnée océanographique acquise durant un levé : les mesures de marées, de courant, de météorologie (MTO), de célérité du son dans l'eau (par CTD notamment), et décrite par quatre classes d'évaluations groupées et ordonnancées de la pondération la plus forte à la plus faible : MTO + CTD + marée + courant ; MTO + CTD + marée ; MTO + CTD ; MTO. La typologie de la qualité identifiée est l'incertitude de la donnée associée à la sensibilité des capteurs, ainsi que la complétude de la donnée.

— *Durée* d'acquisition d'une série temporelle marégraphique, paramètre quantitatif (temps en jours) indiquant la durée d'acquisition du capteur marégraphique ancré lors d'un levé hydro-océanographique. La typologie de la qualité associée est la disponibilité.

— *Trous* dans la série temporelle marégraphique, paramètre quantitatif (% de données manquantes) indiquant le pourcentage de périodes manquantes dans l'acquisition sur la durée totale de la série chronologique du marégraphe. La typologie de la qualité associée est la complétude.

— *Erreurs* ponctuelles dans une série temporelle marégraphique, paramètre quantitatif (% de données aberrantes) indiquant le pourcentage de données mauvaises dans l'acquisition sur la durée totale de la série chronologique du marégraphe. La typologie de la qualité associée est la présence de valeur erronée dans la série temporelle.

Ces paramètres de qualité sont directement présents dans les métadonnées des levés (calculés/mesurés par les ingénieurs responsables des levés) ou bien extraits directement depuis la donnée des levés étudiés. Ces valeurs reflètent certains as-

pects de l'évaluation de la qualité du levé, selon le contexte spécifique de l'acquisition.

Afin de savoir si un levé hydro-océanographique est de bonne qualité, en fonction du concept d'emploi associé au levé, nous avons besoin de réaliser l'inférence des paramètres de préférence, étape cruciale qui permet de déterminer la conformité de ces paramètres de qualité avec les préférences de l'expertise métier et l'utilisation envisagée du levé.

### Inférence des paramètres de préférence

Comme indiqué plus haut, cette étape est primordiale pour configurer efficacement le profil de l'expert dans le processus d'aide à la décision multi-critères. Les paramètres de préférence à inférer sont les indications de préférence des échelles de critères (si le critère doit être minimisé ou maximisé), le(s) poids des paramètres (ou critères de qualité) $(p_j, \forall j \in J)$, le seuil indiquant la majorité $\lambda$, le nombre et l'ordre des catégories de sortie(s), les limites des catégories $(b_h, \forall h \in \{1, \dots k-1\})$, et les profils de veto $(b_h^v, \forall h \in \{2, \dots k\})$.

Ces paramètres de préférence peuvent être construits directement ou indirectement. Avec la méthode directe, les valeurs de ces paramètres sont déterminées en interrogeant l'expert, habituellement dans un processus interactif, où les effets des paramètres de préférence sur l'évaluation globale sont présentés à l'expert afin qu'il puisse les ajuster le plus finement possible. Habituellement, les limites de catégorie et les profils de veto peuvent être considérés comme des normes en accord avec le concept d'emploi final du levé, alors que les pondérations des critères sont extraites à partir de questions liées aux coalitions majoritaires de critères. Dans le cas où le paramétrage direct s'avère impossible, une approche indirecte peut être employée, dans laquelle l'expert évalue la qualité globale de quelques levés, appelés exemples d'apprentissage. À partir de ces réponses, des modèles d'optimisation mathématiques déterminent les valeurs des paramètres du modèle *MR-Sort*, compatible avec les évaluations globales données par l'expert (voir [175]-[181]). Dans le cadre de notre recherche, nous avons eu accès à des experts du Shom afin de définir les paramètres de préférence. Nous avons donc utilisé la méthode directe de paramétrage.

Ainsi, comme précisé ci-dessus, les trois experts métiers interrogés présentent des exigences et des priorités différentes quant à la manière de qualifier des le-

vés hydro-océanographiques, ce qui a conduit à trois configurations différentes du processus proposé. La première étape consiste à identifier pour chaque expert les paramètres de qualité qui doivent avoir un rôle dans l'évaluation finale de la qualité. L'hydrographe utilise ainsi les 10 paramètres présentés en section 4.2.2 de l'évaluation de la qualité du levé, tandis que l'acousticien et l'océanographe ne tiennent pas compte du *CATZOC* et n'utilisent donc que 9 d'entre eux. Pour l'acousticien, les deux catégories *Bon (B)* et *Mauvais (M)* sont suffisantes (de façon naturelle dans l'ordre de préférence suivant : *Bon* puis *Mauvais*), tandis que pour l'océanographe et pour l'hydrographe, ils exigent une catégorie intermédiaire *Acceptable (A)* avec l'ordre de préférence : *Bon*, *Acceptable*, *Mauvais*. Puis, nous avons identifié, durant l'échange avec l'hydrographe, que ses paramètres de préférence varient selon la profondeur moyenne du levé. L'influence du contexte du levé sur les paramètres de préférence nous conduit à considérer 3 profils de paramètres pour l'hydrographe. Par conséquent, pour tous les levés qui seront évalués par le profil "utilisation hydrographique", la profondeur moyenne du levé doit d'abord être déterminée, afin de choisir le profil de préférence à configurer pour l'expert hydrographe durant le processus MCDA.

Les trois experts ont défini de façon unanime l'échelle d'évaluation de critères suivante ($\succ$ signifiant : est préféré à) :

- *CATZOC* : (A1 $\succ$ A2 $\succ$ B $\succ$ C $\succ$ D $\succ$ U)
- *POSACC* : ($< 0.5m \succ < 2m \succ < 5m + 5\%$ profondeur $\succ < 20m + 10\%$ profondeur)
- *SOUACC* : ($< 0.25m \succ < 0.5m \succ < 1m \succ < 2m \succ < 10m \succ < 20m$)
- *Couv* : (300% $\succ$ 200% $\succ$ 150% $\succ$ 100% $\succ$ 80% $\succ$ 50%)
- *Capt* : (SMF $\succ$ LiDAR $\succ$ SBES + SonaL $\succ$ SDB)
- *SEDIM* : (5 $\succ$ 4 $\succ$ 3 $\succ$ 2$\succ$ 1$\succ$ 0)
- *OCEANO* : (MTO + CTD + tide + current $\succ$ MTO + CTD + tide $\succ$ MTO + CTD $\succ$ MTO)
- *Durée* : $\geq$ 18 ans $\succ \geq$ 1 an $\succ \geq$ 30 jours $\succ <$ 30 jours $\succ$ pas de marée
- *Trous* : 0% $\succ <$ 10% $\succ \geq$ 10% $\succ$ pas de marée
- *Erreurs* : 0% $\succ <$ 3% $\succ \geq$ 3% $\succ$ pas de marée

Les paramètres de préférence restants ont été déterminés durant les trois entretiens et sont fournis dans les tables 4.2, 4.3 et 4.4. Comme indiqué plus haut, ces paramètres de préférence sont trouvés via une approche directe. La principale

explication pour ce choix réside en la faible disponibilité de paires exemples d'apprentissage/qualification. Ainsi le Shom possède énormément de levés mais aucun label d'expertise n'y est apposé. Par exemple, dans les métadonnées associées au levé nous n'avons pas d'indication de qualité spécifique pour le profil acoustique. De plus, les connaissances avancées des experts simplifient rapidement la tâche car ils ont fourni des profils de séparation, des profils de veto et la pondération associée au profil. Pour chaque expert, nous présentons tout d'abord le(s) profil(s) de séparation, le profil du veto, le critère de pondération et le seuil de majorité. Ces différents profils sont présentés dans la section 4.2.2.

Le profil acousticien, voir la table 4.2, est le plus simple car il ne comporte que deux catégories et qu'un seul contexte d'acquisition. Pour ce profil, c'est la connaissance du type de sédiment dans les fonds marins qui est prépondérant car il permet de modéliser au mieux la rétrodiffusion du signal acoustique dans la colonne d'eau et les sédiments.

TABLE 4.2 – Paramètres de préférence pour l'usage acoustique.

| Acousticien | *POSACC* | *SOUACC* | *Couv* | *Capt* | *SEDIM* | *OCEAN* | *Durée* | *Trou* | *Erreurs* |
|---|---|---|---|---|---|---|---|---|---|
| B / M ($b_1$) | $< 5m + 5\%$ | $< 1m$ | 150% | SMF | 4 | MTO + CTD + marée + courant | $< 30$ jours | $\geq 10\%$ | $\geq 3\%$ |
| Veto (B) ($b_1^v$) | | · | · | · | 1 | · | · | · | · |
| Poids ($p_j$) | 2/46 | 1/46 | 7/46 | 9/46 | 9/46 | 6/46 | 5/46 | 4/46 | 3/46 |
| Seuil de majorité ($\lambda$) | 23/46 | | | | | | | | |

Le profil océanographe, voir la table 4.3, est plus complexe car il compte trois catégories et deux contextes de 10/100m et 1000m. En effet, par plus de 200m de profondeur, l'impact de la marée est marginal et il est donc moins pertinent pour un océanographe de l'étudier. Ces deux contextes entraînent également deux veto et des poids différents. Pour ce profil, c'est naturellement toute la partie océanographique et mesure de la hauteur d'eau qui est importante.

Le profil hydrographique, voir la table 4.4, est le plus complexe car il cumule trois catégories et trois contextes de 10m, 100m et 1000m. En effet, les critères de qualité d'un hydrographe sont beaucoup plus variables en fonction de la profondeur

TABLE 4.3 – Paramètres de préférence pour l'usage océanographique.

| Océanographe | *POSACC* | *SOUACC* | *Couv* | *Capt* | *SEDIM* | *OCEAN* | *Durée* | *Trou* | *Erreurs* |
|---|---|---|---|---|---|---|---|---|---|
| B / A (10/100m) ($b_{2,10/100m}$) | $< 2m$ | $< 0.5m$ | 200% | SMF | 4 | MTO + CTD + marée + courant | $\geq$ 1 année | 0% | 0% |
| A / M (10/100m) ($b_{1,10/100m}$) | $< 5m + 5\%$ | $< 2m$ | 100% | SBES + SonaL | 2 | MTO + CTD + marée | $\geq$ 30 jours | $< 10\%$ | $< 3\%$ |
| B / A (1000m) ($b_{2,1000m}$) | $< 2m$ | $< 0.5m$ | 200% | SMF | 4 | MTO + CTD + courant | . | . | . |
| A / M (1000m) ($b_{1,1000m}$) | $< 5m + 5\%$ | $< 2m$ | 100% | SBES + SonaL | 2 | MTO + CTD | . | . | . |
| Veto (A et B) ($b_{1,10/100m^v}$ et $b^v_{2,10/100m}$) | | . | . | . | . | < MTO + CTD | $\leq< 30jours$ | . | . |
| Veto (A et B) ($b^v_{1,1000m}$ et $b^v_{2,1000m}$) | | . | . | . | . | < MTO + CTD | . | . | . |
| Poids 10/100m ($p_j$) | 3/46 | 2/46 | 4/46 | 1/46 | 5/46 | 8/46 | 9/46 | 7/46 | 7/46 |
| Poids 1000m ($p_j$) | 3/21 | 2/21 | 4/21 | 1/21 | 5/21 | 6/21 | . | . | . |
| Seuil de majorité 10/100m ($\lambda$) | 23/46 | | | | | | | | |
| Seuil de majorité 1000m ($\lambda$) | 10.5/21 | | | | | | | | |

du levé, l'hydrographe se concentrant sur la bathymétrie. Pour ce profil, c'est toute la partie bathymétrique qui est la plus importante. Les poids sont donc plus important pour les critères suivants : le *CATZOC*, la *Couv* et le *Capt*.

TABLE 4.4 – Paramètres de préférence pour l'usage hydrographique.

| Hydrographe | *CATZOC* | *POSACC* | *SOUACC* | *Couv* | *Capt* | *SEDIM* | *OCEAN* | *Durée* | *Trou* | *Erreurs* |
|---|---|---|---|---|---|---|---|---|---|---|
| B / A (10m) ($b_{2,10m}$) | A1 | $< 0.5m$ | $< 0.25m$ | 200% | SMF | 2 | MTO + CTD + marée + courant | $\geq$ 30 jours | 0 | 0 |
| A / M (10m) ($b_{1,10m}$) | B | $< 5m + 5\%$ | $< 2m$ | 100% | SBES + SonaL | 2 | MTO + CTD + marée | < 30 jours | $< 10\%$ | $< 3\%$ |
| B / A (100m) ($b_{2,100m}$) | A2 | $< 5m + 5\%$ | $< 1m$ | 150% | SMF | 2 | MTO + CTD + marée + courant | $\geq$ 30 jours | $< 10\%$ | $< 3\%$ |
| A / M (100m) ($b_{1,100m}$) | D | $< 20m + 10\%$ | $< 2m$ | 100% | SBES + SonaL | 2 | MTO + CTD + marée | < 30 jours | $< 10\%$ | $\geq 3\%$ |
| B / A (1000m) ($b_{2,1000m}$) | B | $< 20m + 10\%$ | $< 20m$ | 100% | SMF | 2 | MTO + CTD | Pas de marée | Pas de marée | Pas de marée |
| A / M (1000m) ($b_{1,1000m}$) | D | $< 20m + 10\%$ | $< 20m$ | 100% | SBES + SonaL | 2 | MTO + CTD | Pas de marée | Pas de marée | Pas de marée |
| Veto (A et B) ($b^v_1$ et $b^v_2$) | . | . | . | $< 100\%$ | . | . | . | . | . | . |
| Poids ($p_j$) | 10/55 | 6/55 | 7/55 | 9/55 | 8/55 | 3/55 | 5/55 | 4/55 | 2/55 | 1/55 |
| Seuil de majorité ($\lambda$) | 27.5/55 | | | | | | | | | |

### Détermination des paramètres de qualité

Comme indiqué dans la section 4.2.2, les paramètres de qualité utilisés dans le cadre de ce processus sont directement prélevés depuis les métadonnées des levés (calculés/mesurés par les ingénieurs responsables des levés) ou bien extraits depuis la donnée des levés étudiés. Nous observons ainsi sur l'extraction d'un fichier de métadonnées du levé S2015001 après la rédaction de ce levé, voir listing 4.1, quelques paramètres de qualité choisis par les experts du Shom pour évaluer la qualité globale d'un levé hydro-océanographique.

Listing 4.1 – Extract of the metadata file (.xml format) of the S2015001 survey.

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<BDB_Simple_Attributes>
  <Attribute name="CATZOC">
    <Value>zone of confidence B</Value>
  </Attribute>
  <Attribute name="OBJNAM">
    <Value>S201500100-1</Value>
  </Attribute>
  <Attribute name="POSACC">
    <Value>2.1</Value>
  </Attribute>
  <Attribute name="RECDAT">
    <Value>20170216</Value>
  </Attribute>
  <Attribute name="SOUACC">
    <Value>0.43</Value>
</BDB_Simple_Attributes>
```

Nous récupèrons ainsi directement à partir de ce type de fichier de métadonnées, le *CATZOC*, le *POSACC*, le *SOUACC* ou le *Capt.* D'autres éléments de qualité sont extraits à partir du rapport particulier (rapport de la campagne hydro-océanographique) comme les paramètres *Couv*, *SEDIM* et *OCEANO*. Enfin, les paramètres de qualités restants, *Durée*, *Trous* et *Erreurs*, peuvent être directement calculés à partir des données brutes de hauteurs d'eau extraites des capteurs marégraphiques mouillés lors des campagnes hydro-océanographiques. Nous remarquons que pour le paramètre *Couv*, l'algorithme QuadSME proposé dans la section 4.1.2 pourrait être utilisé pour déterminer l'emprise spatiale de chaque ligne de levé afin de calculer la couverture associée au levé bathymétrique. Cette technique est encore en développement et n'a pas été systématisée dans le cadre de ce processus de recherche. Il s'agit néanmoins d'un axe très clair d'automatisation et d'amélioration qui permettrait également d'avoir une approche beaucoup plus locale sur ce paramètre de qualité (qualifier certaines zones du levé et non le levé dans son intégralité).

Comme exemple pour l'extraction automatique des éléments de qualité, nous nous sommes intéressé à la qualification de la série temporelle du marégraphe de la Pointe-des-Galets sur l'île de la Réunion (020° 55' 00.0" S, 055° 17' 00.0" E). Cette mesure de marée a été acquise lors d'un tsunami le 13 août 2021 : elle est présentée

sur la figure 4.17. Ce signal est perturbé par le transit de la vague (au début de la série temporelle) mais nous avons tout de même pu extraire automatiquement les paramètres de qualité des mesures de hauteurs d'eau présentées dans la sous-section 4.2.2.

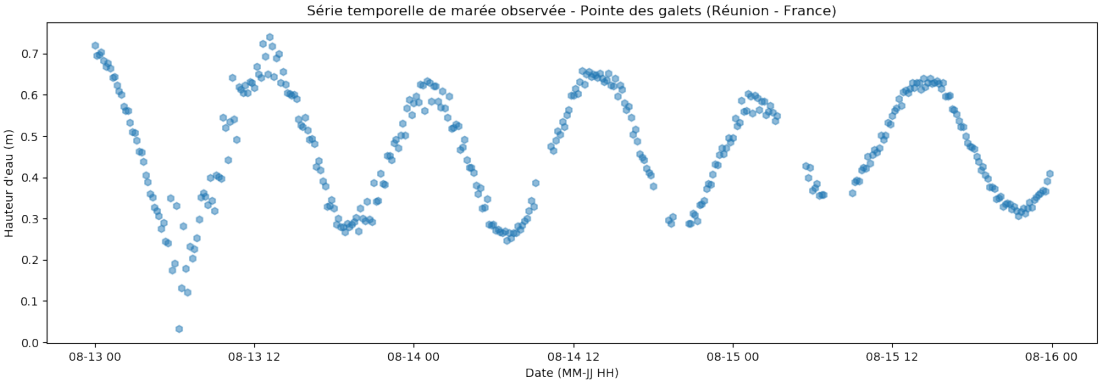


FIGURE 4.17 – Série temporelle de la marée mesurée à la Pointe-des-Galets (Réunion - France) pendant un événement extrême.

Le premier paramètre de qualité à extraire est la durée de la série temporelle $\Delta T$. Cette quantité, en jours, est obtenue par la différence entre la date de fin d'acquisition $T_{end}$ et la date de début d'acquisition $T_{start}$ ($\Delta T = T_{end} - T_{start}$). Pour notre exemple, cette valeur est de 3 jours.

Le deuxième paramètre de qualité associé aux données de hauteurs d'eau est le pourcentage de périodes manquantes dans la série temporelle. La période d'intégration du marégraphe ($\Delta_{integration}$) (période pendant laquelle le marégraphe va moyenner plusieurs mesures pour ne garder qu'une seule valeur) étant de 10 minutes, nous considérons qu'une période manque dès que la différence temporelle entre 2 mesures consécutives $t_i$ est strictement supérieure à 10 minutes. L'algorithme 1 permet ainsi de détecter le nombre de trous.

La figure 4.18 montre ce paramètre de qualité appliqué aux données du marégraphe de la Pointe-des-Galets. La valeur 1 indique la présence de l'information et la valeur 0 l'absence de donnée sur la série temporelle. Dans ce cas, 5 périodes manquantes sont à noter pour un total de 11% de données manquantes sur la durée totale de la série temporelle. Ce paramètre de qualité est très pertinent pour détecter une éventuelle défaillance du capteur ou un problème de liaison avec les serveurs de stockage des données de la série temporelle.

Enfin, le dernier paramètre de qualité extrait automatiquement est la présence

**Algorithm 1** Calcul $N_{Gap}$

$N_{Gap} \leftarrow 0$
**for** $i \leftarrow 1$ to $N-1$ **do**
  **if** $t_{i+1} - t_i > \Delta_{integration}$ **then**
    $N_{Gap}$ += 1
  **end if**
**end for**
**return** $N_{Gap}$

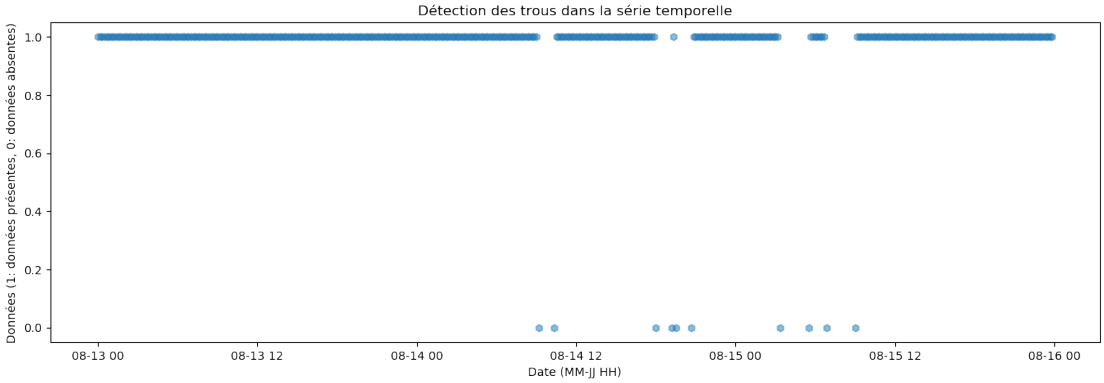


FIGURE 4.18 – Données de marée manquantes pour la marée de la Pointe-des-Galets (Réunion - France).

de valeurs aberrantes ou de fortes variations dans le signal qui pourraient perturber une future modélisation. Ce paramètre est crucial pour détecter les événements extrêmes tels que les tsunamis ou les tremblements de terre. Il garantit également la qualité des informations qui seront utilisées pour modéliser le phénomène de marée nécessitant de longues périodes de mesures. Pour que cette modélisation soit la plus précise possible, il peut être nécessaire d'éliminer ces événements exceptionnels qui ne correspondent pas à une situation normale de marée.

Pour calculer la présence d'une valeur aberrante, nous exploitons la différence entre la valeur absolue de la variation locale de la mesure et la valeur absolue de la variation locale de la marée prédite. Si cette différence est supérieure à un certain seuil, la mesure de hauteur d'eau est considérée comme aberrante. Le seuil est défini par l'incertitude liée à la mesure du marégraphe (5cm) et par une variation maximale acceptable de la pression atmosphérique de 10hPa sur une heure (soit une variation de 10cm par heure) qui correspond à des événements météorologiques fortement dépressionnaires ; donc un seuil à 7cm dans le cas de notre $\Delta_{integration}$.

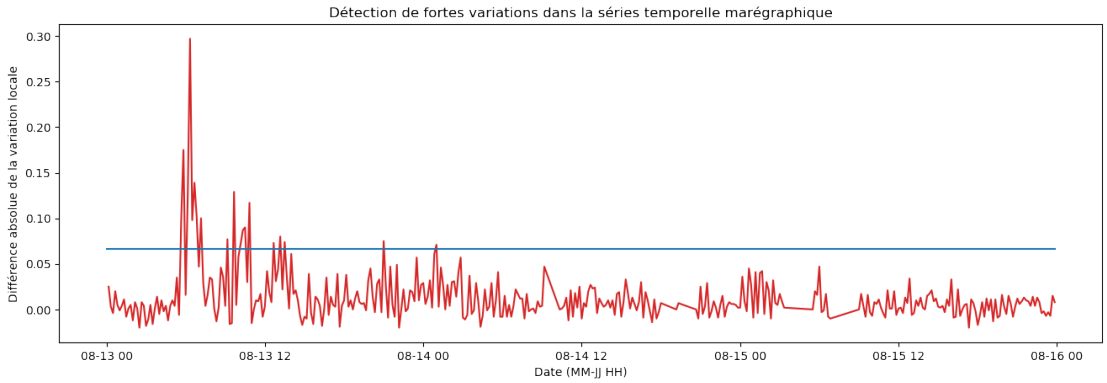


FIGURE 4.19 – Détection de fortes variations pour la série temporelle de marées de la Pointe-des-Galets (Réunion - France).

La figure 4.19 montre ce paramètre de qualité appliqué aux données du marégraphe de la Pointe-des-Galets ainsi que la représentation du seuil par la ligne bleue. Dans ce cas, 9% des données sont considérées comme aberrantes.

### Évaluation multicritère : choix du modèle

Les paramètres de qualité extraits et l'inférence des paramètres de préférence nourrissent le processus MCDA. Dans cette section, nous motiverons notre choix de modèle de préférence spécifique MCDA, avant de détailler sa formulation mathématique. Comme vu dans la section 4.2.2, l'échelle d'évaluation des paramètres de qualité est assez hétérogène. En effet, certains d'entre eux sont qualitatifs ou catégoriques (*CATZOC*, *Capt*, *SEDIM*, *OCEAN*), alors que d'autres sont clairement quantitatifs (*POSACC*, *SOUACC*, *Couv*, *Durée*, *Trous*, *Erreurs*). Ce constat intervient en faveur du paradigme européen et des méthodes de sur-classement, mentionnées en section 4.2.2 dans le choix de la méthode MCDA, qui gère bien par nature cette diversité d'échelles. Ensuite, afin de faciliter l'appropriation de ce processus par les utilisateurs, nous visons à proposer une solution dont chaque étape est facilement explicable et compréhensible et où l'évaluation finale peut être aisément interprétée, ceci afin de générer des recommandations pour améliorer la qualité de futurs levés. Cette interprétabilité peut être réalisée en résolvant des problèmes de tri provenant du paradigme de sur-classement. En effet, les alternatives de tri dans le cadre MCDA reposent sur la création d'une échelle d'évaluation globale qualitative ou catégorique afin d'agréger des critères multiples. Cela se

fait généralement par le biais de limites ou de profils de catégories qui peuvent être considérés comme des frontières, par rapport auxquelles les alternatives sont comparées afin de décider à quelle catégorie elles appartiennent. Parmi tous les algorithmes de tri disponibles dans le paradigme du sur-classement, nous proposons d'utiliser la méthode *MR-Sort*, caractérisée dans [182], [183], qui a l'avantage de générer des résultats hautement interprétables, tout en demeurant un modèle très expressif. Dans la suite, nous introduirons le formalisme de la méthode *MR-Sort*.

Considérons un ensemble fini d'alternatives $A$ (les campagnes hydrographiques dans notre cas applicatif décrit dans la section 4.2.3) et un ensemble fini de critères $J$ (correspondant aux paramètres de qualité de la section 4.2.2). Pour chaque critère, les évaluations possibles sont ordonnées selon les préférences de l'expert, ce qui définit les sens de préférence des critères (une valeur inférieure est préférée à une valeur supérieure, ou vice-versa). L'évaluation globale est catégorique, soit $\{c_1, \ldots, c_k\}$ les $k$ catégories de sortie (représentant les $k$ niveaux de l'échelle d'évaluation ordinale de sortie), ordonnées par leur désirabilité, de $c_1$ étant la pire catégorie à $c_k$ étant la meilleure : $c_k \succ \ldots \succ c_1$. Dans notre contexte d'évaluation de la qualité, ces catégories pourraient par exemple être {très haute qualité $\succ \ldots \succ$ moyenne qualité $\succ$ qualité faible}. Ces catégories sont caractérisées par un ensemble de profil(s) de séparation $B = \{b_1, \ldots b_{k-1}\}$. Chaque catégorie $c_h$ est ainsi définie par sa limite supérieure $b_h$ et sa limite inférieure $b_{h-1}$, à l'exception des catégories pire et meilleure, qui n'ont qu'une seule limite. Chaque alternative et chaque limite de catégorie peuvent être représentées par un vecteur d'évaluation par rapport aux critères. L'évaluation par rapport au critère $j$ peut être vue comme une fonction $g_j : A \cup B \to \mathbb{R}$, où $g_j(a)$ désigne l'évaluation de l'alternative $a \in A$ sur le critère $j$ et $g_j(b_h)$ désigne l'évaluation de la limite de catégorie $b_h, \forall h \in \{1, \ldots, k-1\}$, sur le critère $j$. Dans cette présentation de MR-Sort, nous supposons sans perte de généralité, que les performances sont telles qu'une valeur plus élevée dénote une meilleure performance. Il est évident que ce n'est pas nécessairement le cas dans une application réelle. De plus, les performances des limites de la catégorie sont croissantes, c'est-à-dire que $\forall j \in J, 1 < h < k$ : $g_j(b_{h-1}) \leq g_j(b_h)$.

La méthode MR-Sort utilise deux *règles d'affectation* pour placer les alternatives dans les catégories : les règles d'affectation pessimiste et optimiste [184], [185]. La règle pessimiste affecte une alternative $a$ à la catégorie la plus élevée

possible $c_h$ de sorte que $a$ surclasse la frontière inférieure de la catégorie $b_{h-1}$. La règle optimiste assigne $a$ à la catégorie la plus basse possible $c_h$ de sorte que la frontière supérieure de la catégorie $b_h$ surclasse $a$. La règle pessimiste est la plus communément utilisée en pratique, car elle génère des recommandations plus sécuritaires. On dit d'une alternative $a$ qu'elle surclasse une frontière $b_{h-1}$ si et seulement s'il existe une coalition suffisante de critères soutenant l'assertion : $a$ est au moins aussi bonne que $b_{h-1}$ et aucun critère ne s'oppose fortement (veto) à cette assertion. Pour mesurer si une coalition suffisante de critères considère que $a$ est au moins aussi bon que $b_{h-1}$, nous définissons d'abord pour chaque critère $j$ une fonction $C_j : A \times B \to \{0, 1\}$ qui évalue si le critère $j$ soutient cette affirmation ou non :

$$\forall j \in J, a \in A, 1 \leq h \leq k,$$
$$C_j(a, b_{h-1}) = \begin{cases} 1, \text{ si } g_j(a) \geq g_j(b_{h-1}), \\ 0, \text{ autrement.} \end{cases} \tag{4.4}$$

Pour évaluer si une coalition de critères est en faveur du surclassement ou non, $\forall a \in A, 1 \leq h \leq k$, nous définissons d'abord la concordance globale comme :

$$C(a, b_{h-1}) = \sum_{j \in J} p_j C_j(a, b_{h-1}), \tag{4.5}$$

où $p_j$ est le poids du critère $j$ et $C_j(a, b_{h-1}) \in \{0, 1\}$ mesure si $a$ est au moins aussi bon que $b_{h-1}$ du point de vue du critère $j$ ou non : $C_j(a, b_{h-1}) = 1 \Leftrightarrow g_j(a) \geq g_j(b_{h-1})$, 0 sinon.

Les poids sont définis de façon à ce qu'ils soient positifs ($p_j \geqslant 0, \forall j \in J$) et que leur somme soit égale à 1 ($\sum_{j \in J} w_j = 1$). Cette concordance globale est ensuite comparée à un seuil de majorité $\lambda \in [0.5, 1]$ extrait des préférences du décideur avec les poids (dans notre cas ce sera l'expert de l'exploitation finale du levé).

Même quand la coalition est suffisamment forte, le critère peut poser un veto à la situation de surclassement. Dans un tel cas, on dit que $a$ surclasse une frontière $b_{h-1}$ si et seulement s'il existe une coalition suffisante de critères soutenant l'assertion "$a$ est au moins aussi bon que $b_{h-1}$" et qu'aucun critère ne s'oppose fortement (veto) à cette assertion. Une alternative $a$ est donc dans une relation de veto (notée V) avec un profil $b_{h-1}$ lorsque :

$$a \,\mathrm{V}\, b_{h-1} \iff \exists j \in J \textit{ tel que } g_j(a) < g_j(b^v_{h-1}) \tag{4.6}$$

Le profil de veto $b^v_{h-1}$ représente le niveau minimum de performance qu'une alternative doit avoir pour être autorisée dans la catégorie $c_h$ via la coalition pondérée de critères en faveur de cette affectation. Si pour un critère quelconque, une alternative a une performance inférieure au profil de veto de $c_h$, alors il lui est interdit d'être assignée à $c_h$ ou au-dessus.

La contrainte suivante sur les évaluations d'un profil de veto $b^v_{h-1}$ doit être respectée : $g_j(b^v_{h-1}) \leqslant g_j(b_{h-1}), \forall j \in J, h \in 1..k$. Les performances sur les profils de veto sont non décroissantes, c'est-à-dire que $\forall j \in J, h \in 1..k+1, g_j(b^v_{h-1}) \leq g_j(b^v_h)$. De plus, les performances du profil de veto le plus élevé seront fixées pour être égales à celles du profil de catégorie le plus élevé, de sorte qu'aucune situation de veto ne peut être déclenchée dans ce cas. Ces valeurs peuvent être définies directement par le décideur et sont utilisées pour former ce que nous appelons des profils de veto pour chaque limite inférieure d'une catégorie, où $b^v_{h-1} = \{g_j(b_{h-1}) - v_j^{h-1} \; \forall j \in J\}, 2 \geqslant h \geqslant k$. L'opérateur de la relation de surclassement entre deux alternatives est noté S.

Pour résumer, l'alternative $a$ surclasse la frontière $b_{h-1}$ (et est donc affectée au moins à la catégorie $c_h$) si et seulement si :

$$a \,\mathrm{S}\, b_{h-1} \iff C(a, b_{h-1}) > \lambda \text{ et pas } (a \,\mathrm{V}\, b_{h-1}) \tag{4.7}$$

Pour illustrer l'utilisation de la méthode *MR-Sort*, nous considérons un problème de décision où différentes campagnes hydro-océanographiques doivent être qualifiées dans l'une des deux catégories *Bon* ($B$) ou *Mauvais* ($M$) selon les préférences d'un expert. Considérons trois levés $a_1$, $a_2$ et $a_3$ évalués sur trois critères *CATZOC* ($C$) ($A_1 \succ_\mathrm{C} A_2 \succ_\mathrm{C} B \succ_\mathrm{C} C \succ_\mathrm{C} D \succ_\mathrm{C} U$) ($\succ_\mathrm{C}$ est la relation de préférence stricte définissant l'échelle d'évaluation du critère *CATZOC*), *Couv* ($H$) (plus la couverture est importante, mieux c'est), et *SEDIM* ($Se$) (plus il est élevé, mieux c'est également). L'évaluation de ces alternatives selon ces 3 critères est présentée dans le tableau 4.5 ainsi que les paramètres du modèle MR-Sort (qui ont été déterminés au préalable à partir des préférences de l'expert via les procédures présentées dans la section 4.2.2). Le profil de séparation des catégories $b_1$ délimite les deux catégories $B$ et $M$, par des évaluations à $B$ pour le *CATZOC*, de 100

pour le *Couv hydrographique* et de 3 pour le *SEDIM*. Le profil de veto $b_1^v$ est défini par la pire évaluation de *CATZOC* ($U$) et de *Couverture hydrographique* (50), et un nombre de capteurs pour *SEDIM* supérieur strict à 1. Cela implique que pour le dernier critère, la méthode déclenchera un veto lorsque le nombre de capteurs sédimentologiques sera égal à zéro ou un.

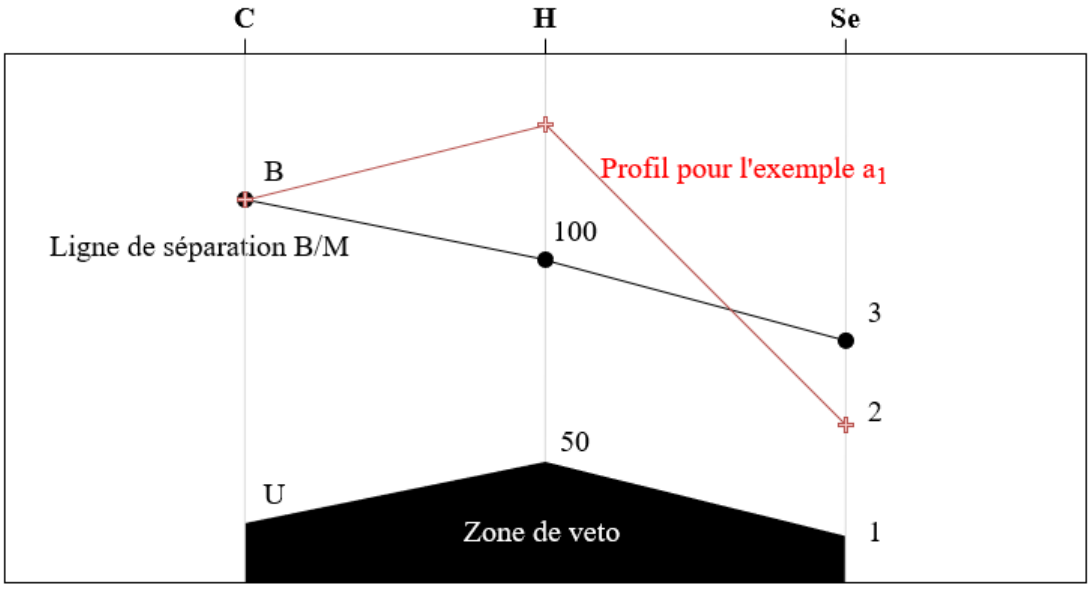


FIGURE 4.20 – Représentation graphique de la ligne de séparation $B/M$ pour le levé $a_1$ dans le cadre de notre exemple d'illustration.

TABLE 4.5 – Exemple illustratif pour la méthode MR-Sort.

Paramètres du modèle

| | *CATZOC* | *Couv* | *SEDIM* | $p_C$ | $p_H$ | $p_S$ | $\lambda$ |
|---|---|---|---|---|---|---|---|
| $b_1$ | $B$ | 100 | 3 | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{2}$ |
| $b_1^v$ | $U$ | 50 | 1 | | | | |

Qualification

| | *CATZOC* | *Couv* | *SEDIM* | $C(a,b_1)$ | $C(a,b_1) \geqslant \lambda$ | $a \,\mathrm{V}\, b_1$ | $a \,\mathrm{S}\, b_1$ | qualification |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $B$ | 150 | 2 | $\frac{1}{3}+\frac{1}{3}+0=\frac{2}{3}$ | ✓ | x | ✓ | *Bon* |
| $a_2$ | $A_1$ | 300 | 0 | $\frac{1}{3}+\frac{1}{3}+0=\frac{2}{3}$ | ✓ | x | x | *Mauvais* |
| $a_3$ | $C$ | 80 | 4 | $0+0+\frac{1}{3}=\frac{1}{3}$ | x | ✓ | x | *Mauvais* |

Comme le premier levé $a_1$ est au moins aussi bon que $b_1$ sur les deux critères $C$ et $H$, il possède une coalition suffisante de critères soutenant son affectation dans la catégorie $B$. De plus, il possède 2 capteurs pour le critère *SEDIM-Se* : il

n'est donc pas en situation de veto. Le deuxième levé $a_2$ a également une coalition suffisante de critères en faveur de son affectation dans la catégorie $B$ mais elle n'a pas de capteur pour le critère *SEDIM*. Par conséquent, elle soulève un veto qui invalide la relation de sur-classement entre $a_2$ et $b_1$. Par conséquent, elle est affectée à la catégorie $M$. La dernière campagne $a_3$ est au moins aussi bonne que $b_1$ sur un seul critère, $S$, donc elle ne surclasse pas $b_1$, et elle est donc assignée à la catégorie $M$.

### 4.2.3 Application aux levés hydro-océanographiques

Afin de démontrer la pertinence du processus d'évaluation que nous proposons, nous présentons dans cette sous-section une étude de cas sur des levés hydro-océanographiques réels acquis par le Shom. Nous avons considéré un utilisateur ingénieur hydrographe qui souhaiterait évaluer la qualité globale de huit campagnes à la mer. Il débute le processus de la figure 4.15 avec, pour première étape, l'identification du concept d'emploi de ces levés. Il souhaite évaluer la qualité de ces levés pour satisfaire des besoins pour l'acoustique sous-marine, puis océanographie et enfin hydrographie pure. Initialement, ce processus n'avait pas été configuré pour ces trois utilisations spécifiques. De fait, il faut procéder à l'étape de détermination des préférences comme indiqué dans la figure 4.15. Trois experts : un acousticien, un océanographe et un hydrographe ont été interrogés dans le cadre de ce sous-processus. Chacun de ces 3 entretiens a abouti à des configurations différentes du modèle de préférence (différents paramètres de qualité et différents paramètres de préférence). Les trois exploitations finales possibles des levés sont en effet très différentes mais tout comme leur métier et leur domaine d'expertise. Voici une présentation des profils que nous avons consulté :

— L'acousticien.ne modélise la propagation du son dans l'eau. Les sources acoustiques sous-marines sont de natures différentes : biologique, géophonique ou anthropique (et souvent des combinaisons des différents types). La compréhension du paysage sonore sous-marin demande une maîtrise de toutes ces sources acoustiques, il est de fait important de pouvoir modéliser la totalité de la propagation des ondes acoustiques dans la colonne d'eau mais également dans leur interaction et la réflexion avec les fonds marins. Pour ce profil, la qualité de l'information doit être bonne pour le paramètre *SEDIM*, *OCEANO* et *couverture bathymétrique*.

— L'océanographe étudie les activités en lien avec la compréhension et la modélisation des paramètres physiques de la colonne d'eau (température, salinité, transparence ...) et leur évolution au cours du temps. Pour ce domaine, il est important que l'expert dispose d'informations de bonne qualité pour le paramètre *OCEANO*, les paramètres associés à la marée (*durée*, présence de *trous*, présence d'*erreurs aberrantes*) et dans une moindre mesure pour le paramètre *SEDIM*. Comme il bâtit des modèles physiques à partir d'un volume d'eau, il est aussi important de posséder une connaissance fidèle du fond bathymétrique. En particulier, la présence ou l'absence d'artefacts sur le fond qui perturberaient le modèle physique pour les forçages (action qui agit sur un système dynamique) océanographiques comme les épaves ou les obstructions sous-marines qualifiées par les paramètres de qualités *POSACC* et le *SOUACC*.

— L'hydrographe : a pour domaine d'expertise la mesure du fond marin, des courants et des marées d'assurer la sécurité à la navigation des produits nautiques créés à partir des données bathymétriques et, tout particulièrement, les cartes nautiques. Il est primordial que les informations bathymétriques (comme le *CATZOC*, la *Couv* et le *Capt* hydrographique) soient de la meilleure qualité possible car ce sont ces informations qui garantissent des cartes nautiques fidèles et sécuritaires pour le navigateur. Aussi, selon la profondeur moyenne du levé, le paramètre *OCEANO* sera plus ou moins pertinent (le phénomène de marée ayant un impact plus important sur les zones côtières que sur les zones hauturières).

Les huit levés hydrographiques pour lesquels nous avons testé notre chaîne de traitement sont présentés dans les figures 4.21, 4.22, 4.23, 4.24, 4.25, 4.26, 4.27 et 4.28. Ils possèdent des caractéristiques et des emplacements géographiques très différents. Ils sont ainsi représentatifs des types de levés acquis au Shom et des types de données que les experts en données hydro-océaniques peuvent rencontrer.

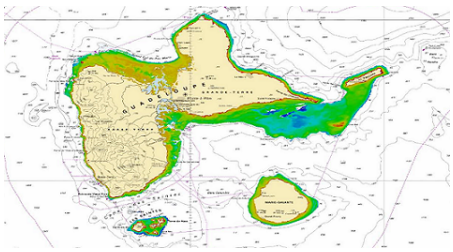

FIGURE 4.21 – Levé S2012056.

La campagne S2012056 est un levé topo-bathymétrique LiDAR aéroporté de l'île de la Guadeloupe. La profondeur moyenne est d'environ 15 mètres. Ce levé a été spécifié et acquis pour établir une référence terre-mer et possède très peu de mesures océanographiques (pas de courantomètre ni de marégraphe mouillé).

La campagne S2015009 est un levé d'exploration, voir section 1.1, au SBES, réalisé sur l'îlot Clipperton dans des conditions environnementales très complexes (logistique faible sur l'île, conditions météorologiques très mauvaises). La profondeur moyenne relevée est de 20 mètres. Le but de ce levé était de mesurer le chenal d'accès à l'îlot de Clipperton pour permettre un atterrissage et un embarquement en toute sécurité de l'île.

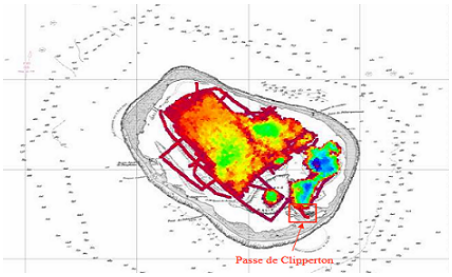


FIGURE 4.22 – Levé S2015009.

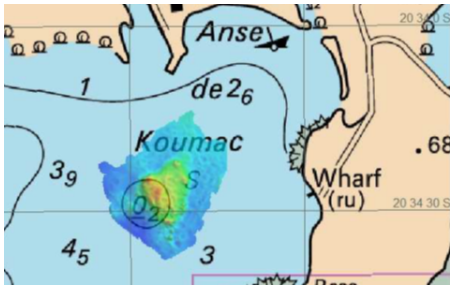


FIGURE 4.23 – Levé S2015001.

La campagne S2015001 est également un levé d'exploration au SBES avec SonaL afin d'effectuer une détection d'obstruction avec une couverture complète. Le levé a été réalisé dans l'anse de Koumac (Nouvelle-Calédonie). La profondeur moyenne relevée est de 5 mètres. Le but de ce levé était de mesurer un haut-fond dans l'anse représentant un danger potentiel pour la navigation.

La campagne S2017026 est un levé hydrographique classique, voir section 1.1, au SMF, qui vise à améliorer la connaissance bathymétrique du dispositif de séparation du trafic (DST) au large de l'île d'Ouessant. Avec une profondeur moyenne de 100 mètres, ce levé a permis la mise à jour d'ouvrages nautiques et plus particulièrement des CM.

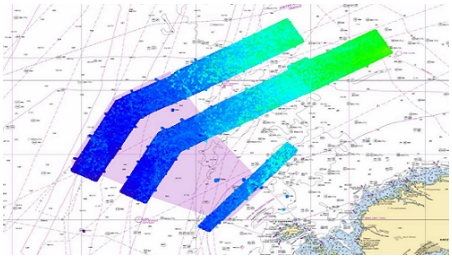

FIGURE 4.24 – Levé S2017026.

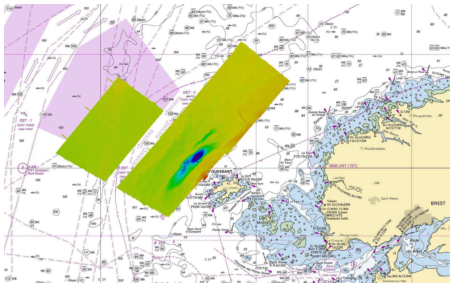

FIGURE 4.25 – Levé S2018057.

La campagne S2018057 a été acquise avec des besoins quasiment identiques au levé S2017026 mais sur une emprise géographique différente. Avec une profondeur moyenne de 100 mètres, ce levé a également permis de mettre à jour les ouvrages nautiques et plus particulièrement les CM mais aussi de mesurer fidèlement la *Fosse d'Ouessant*.

La campagne S2019011 est également un levé topo-bathymétrique LiDAR aéroporté des îles Saint-Martin et Saint-Barthélemy. La profondeur moyenne est d'environ 20 mètres. Ce levé a été spécifié et acquis pour fournir une nouvelle référence terre-mer après l'ouragan Irma en 2017.

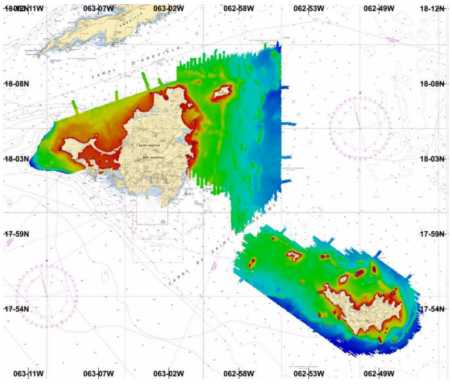


FIGURE 4.26 – Levé S2019011.

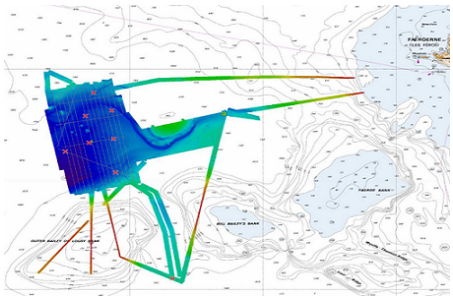

FIGURE 4.27 – Levé S2019026.

La campagne S2019026 est un levé SMF hauturier, voir section 1.1, qui a été réalisé dans l'océan Atlantique au large des côtes des îles Féroé. Sa profondeur moyenne est de 1400 mètres. Ce levé a été réalisé afin de connaître le plus précisément possible l'environnement acoustique. Ainsi, de nombreuses mesures océanographiques et sédimentologiques ont été mises en place.

La campagne S2019029 est un levé de reconnaissance dont l'objectif est de suivre la dynamique sédimentaire en Manche et ainsi mesurer le déplacement des dunes qui présenteraient des dangers éventuels pour la navigation côtière au large de Dunkerque mais aussi de mieux connaître les forçages sédimentaires de cette région. La profondeur moyenne du levé est de 30 mètres.

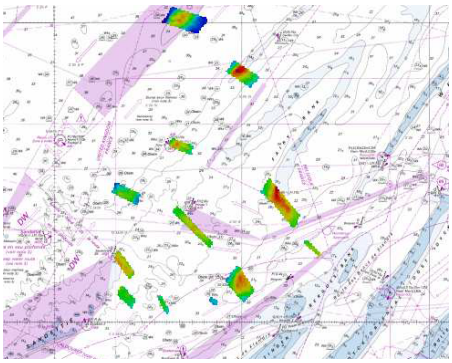

FIGURE 4.28 – Levé S2019029.

L'étape de détermination des paramètres de qualité pour nos levés génère ensuite les données qui figurent dans la table 4.6. Nous observons ainsi une grande diversité de levés étudiés avec des niveaux de qualité disparates au sein d'un même contexte de levé (les levés dans la même gamme de profondeur) mais également entre les contextes de levés acquis avec des objectifs similaires.

TABLE 4.6 – Table de performance : sortie de l'étape *détermination des paramètres de qualité* pour nos levés étudiés.

| Levés (profondeur) | *CATZOC* | *POSACC* | *SOUACC* | *Couv* | *Capt* | *SEDIM* | *OCEAN* | *Durée* | *Trou* | *Erreurs* |
|---|---|---|---|---|---|---|---|---|---|---|
| S2012056 (10m) | B/D | $< 5m + 5\%$ | $< 1m$ | 50% | LiDAR | 0 | MTO | Pas de marée | Pas de marée | Pas de marée |
| S2015009 (10m) | C | $< 5m + 5\%$ | $< 1m$ | 50% | SBES + SonaL | 0 | MTO + CTD | Pas de marée | Pas de marée | Pas de marée |
| S2015001 (10m) | B | $< 2m$ | $< 0.5m$ | 100% | SBES + SonaL | 0 | MTO + CTD + marée | $< 30$ jours | 0% | 0% |
| S2017026 (100m) | A1/B | $< 5m + 5\%$ | $< 1m$ | 100% | SMF | 2 | MTO + CTD + marée | $\geq 30$ jours | 0% | 0% |
| S2018057 (100m) | B | $< 5m + 5\%$ | $< 0.5m$ | 200% | SMF | 4 | MTO + CTD + marée + courant | $\geq 30$ jours | 0% | 0% |
| S2019011 (10m) | B | $< 5m + 5\%$ | $< 0.5m$ | 200% | LiDAR | 0 | MTO | Pas de marée | Pas de marée | Pas de marée |
| S2019026 (1000m) | B/C | $< 20m + 10\%$ | $< 20m$ | 200% | SMF | 5 | MTO + CTD | Pas de marée | Pas de marée | Pas de marée |
| S2019029 (10m) | B | $< 2m$ | $< 0.5m$ | 50% | SMF | 3 | MTO + CTD + marée | $\geq 30$ jours | 0% | 0% |

Dans ce tableau, nous observons que le *CATZOC*, attribué par les hydrographes grâce aux métadonnées, n'est pas nécessairement associée à une valeur unique. Pour le levé S2019026, nous trouvons par exemple deux valeurs de *CATZOC* (B et C) alors que, pour le levé S2017026, le *CATZOC* vaut soit A1 soit B en fonction de la zone étudiée dans le levé. Cette valeur *CATZOC* peut évoluer selon la zone et la fourchette de profondeurs acquises. Durant l'étape MCDA, la valeur la plus pessimiste de *CATZOC* est conservée afin de rester le plus conservateur possible pour la sécurité de la navigation. Néanmoins, cela montre qu'il peut être parfois

pertinent de descendre d'un niveau d'échelle géographique lorsque nous nous intéressons à la qualité d'un levé hydro-océanographique. En effet, la qualité peut varier au sein d'un même levé en fonction de la zone prise en compte. Ainsi, les mesures de *POSACC* et de *SOUACC* au Shom s'appuient sur des calculs de la TPU prenant en compte l'ensemble de la chaîne d'acquisition bathymétrique décrite dans la section 1.1.3. Ici encore, le majorant de la TPU (une pour la TPU vertical et une pour la TPU horizontal) est conservé pour l'ensemble du levé et c'est cette valeur qui est utilisée pour étudier la qualité des levés bathymétriques via le processus MCDA.

Une fois les paramètres de préférence établis (voir table 4.2, 4.3 et 4.4) et les paramètres de qualité extraits des levés (voir table 4.6), le processus MCDA décrit dans la section 4.2.3 peut être mis en oeuvre. La méthode d'évaluation globale *MR-sort* génère pour chaque utilisation finale (acoustique, hydrographique, océanographique) une évaluation globale de la qualité. La donnée de sortie de cette étape d'évaluation est inscrite dans la table 4.7, les niveaux bon/acceptable/mauvais étant définis dans la section 4.2.2.

TABLE 4.7 – Évaluations finales : résultat de l'étape *évaluation multi-critères*.

| Levés (profondeur) | Besoin acoustique | Besoin océanographique | Besoin hydrographique |
|---|---|---|---|
| S2012056 (10m) | Mauvais | Mauvais | Mauvais |
| S2015009 (10m) | Mauvais | Mauvais | Mauvais |
| S2015001 (10m) | Mauvais | Mauvais | Acceptable |
| S2017026 (100m) | Bon | Bon | Bon |
| S2018057 (100m) | Bon | Bon | Bon |
| S2019011 (10m) | Mauvais | Mauvais | Acceptable |
| S2019026 (1000m) | Bon | Bon | Bon |
| S2019029 (10m) | Bon | Acceptable | Mauvais |

Comme indiqué plus haut, l'utilisation d'un modèle *MR-Sort* permet d'obtenir un résultat complètement explicable pour l'évaluation proposée par la table 4.7. Par exemple, pour l'utilisation hydrographique, le levé S2019029 à 10 mètres de profondeur moyenne est classé *Mauvais* malgré un très bon niveau de qualité (comme sur les paramètres Capteurs ou SOUACC) car son paramètre de qualité de

couverture hydrographique atteint le veto du profil hydrographique pour ce levé. De même, le levé S2015001, pour une utilisation acoustique, est classé à un niveau de qualité *Mauvais* car sa valeur Sédimentologie atteint également une valeur de veto pour ce profil. En revanche, le levé S2019026, toujours pour le profil acoustique, a un niveau de qualité évalué à *Bon* car ses paramètres de qualité avec des poids importants (poids 9 pour la sédimentologie, 9 pour les capteurs et 7 pour la couverture) sont supérieurs au profil de séparation *Bon-Mauvais*. Uniquement avec ces trois paramètres de qualité, le levé atteint déjà un score de 25 dépassant ainsi le seuil de majorité fixé à 23 pour le profil acoustique.

De plus, comme mentionné précédemment, ce processus permet également d'émettre des recommandations afin d'améliorer la qualité des futures acquisitions de données. Par exemple, nous constatons que pour le levé S2019029, il suffirait d'ajouter un système d'acquisition de données sédimentologiques (le critère sédimentologie passant alors à 5) ou de mesurer le courant (le critère données océaniques étant alors à son maximum) pour que la coalition des critères de qualité dépasse le seuil majoritaire pour un cas d'usage océanographique. Nous constatons la force de l'interprétabilité de ce type de méthode qui permet un accès rapide aux recommandations pour les décideurs.

### 4.2.4 Discussions et perspectives du processus MCDA pour décrire la qualité d'un levé

La méthodologie proposée présente au moins trois avantages :

— Premièrement, comme vu dans la table 4.7, l'évaluation finale d'un levé n'est pas nécessairement identique pour les trois concepts d'emploi proposés. En effet, elle dépend des préférences spécifiées par l'expert métier et du contexte du levé (pour notre cas la profondeur moyenne). Cela peut être observé plus spécifiquement sur les levés S2017026 et S2018057 qui sont évalués différemment pour chacune des trois utilisations finales.

— Deuxièmement, un levé peut conduire à une même évaluation globale pour deux exploitations finales mais pour des raisons différentes. Par exemple, le levé S2015009 est évalué comme *Mauvais* pour une utilisation finale à la fois hydrographique, acoustique et océanographique. Cette évaluation négative pour une utilisation acoustique et océanographique s'explique par le

fait que même si le levé est assez bon au niveau des paramètres de qualité *POSACC* et *SOUACC*, la pondération ajoutée de ces paramètres (3/46) demeure loin de la coalition majoritaire (23/46). Pour une utilisation hydrographique, l'évaluation négative provient simplement du fait que toutes les évaluations s'avèrent en dessous du niveau d'exigence minimal requis pour un levé correct.

— Troisièmement, l'étape d'évaluation multicritère permet d'émettre des recommandations pour de futurs levés en fournissant, si nécessaire, des explications précises et explicites pour l'utilisateur final de la donnée hydro-océanographique. Si nous prenons en exemple le levé S2019011, il est considéré comme *Mauvais* par le profil acousticien car les trois critères *POSACC*, *SOUACC* et *types de capteurs* considérés comme *Bon*, ne le sont pas suffisamment pour que l'estimation soit globalement bonne. Si toutefois, pour un futur levé, les critères *OCEANO* et *couverture bathymétrique* sont améliorés ne serait-ce que d'un niveau, l'évaluation de la qualité du levé passerait alors à *Bon*. Les paramètres de qualités associés à la donnée d'un levé représentent un choix spécifique de caractérisation, en lien avec les entretiens menés dans le cadre de ce travail de recherche. Selon les données et métadonnées disponibles, il est possible d'affiner la représentation de la qualité, permettant ainsi d'inclure des préférences plus spécifiques de l'expert, sans modifier le processus d'évaluation de la qualité proposé dans ce manuscrit pour les levés hydrographiques.

## 4.3 Conclusion

Comme présenté dans cette section 4.1, la construction de la SME est une étape essentielle lors de la rédaction et le contrôle d'un levé hydro-océanographique. Cette emprise spatiale est également le support de nombreuses décisions appliquées aux levés hydro-océanographiques, notamment dans le cadre de la génération de la surface de référence bathymétrique mais également afin d'affecter une emprise spatiale cohérente avec les informations de qualité présentées dans la section 4.2.2. Cette surface est essentielle pour la quasi-totalité des produits à base de données bathymétriques, comme le projet Téthys présenté dans la section 5.

De plus, la quantité de données bathymétriques augmente avec l'amélioration

de la résolution des capteurs et l'utilisation généralisée de nouvelles technologies comme le LiDAR. Ainsi, QuadSME, voir section 4.1.2, fournit aux hydrographes un nouvel outil efficace d'automatisation et de standardisation d'une tâche de numérisation manuelle auparavant fastidieuse et potentiellement sujette à la subjectivité et aux erreurs. C'est à partir de cette SME que de nombreuses décisions concernant un levé hydro-océanographique sont ensuite prises.

Cette méthode a ainsi été validée par les hydrographes du Shom avec de la donnée synthétique et de la donnée réelle. En effet, elle respecte les contraintes associées à la définition de la SME, les contraintes de performances associées, aux lots ayant un volume important, et les contraintes d'ergonomie et d'intégration dans le SI du Shom. Aussi, cette méthode est meilleure que la méthode classique de détermination (voir table 4.1). Elle est actuellement utilisée en production au sein du Shom comme méthode unique et novatrice de génération de la SME et permet un gain de temps important pour cette tâche autrefois fastidieuse. Des retouches manuelles, à la marge, peuvent toujours être effectuées dans le cas de lots de données bathymétriques particulièrement complexes et dans un souci de généralisation de la SME. De plus, afin de faciliter l'expérience utilisateur, cette méthode a directement été intégrée sur un serveur de calcul délocalisé et relié à une interface web en production dans le cadre du projet Téthys (section 5), voir la figure 4.29. Réussir à automatiser complètement cette tâche nécessiterait sûrement d'adapter les réglages en fonction des types de capteurs bathymétriques utilisés. Cela pourrait être étudié dans le futur pour éviter les reprises manuelles sur les lots issues de l'acquisition LiDAR par exemple.

Enfin, des améliorations de la méthode proposée sont à mettre en place notamment par rapport à l'optimisation du temps de calcul. Ce dernier peut être grandement amélioré par la mise en place du multiprocessing et en traitant chaque sous-quadrant de façon indépendante.

Une fois l'emprise spatiale générée, la seconde phase dans la rédaction et le contrôle de la donnée consiste à associer à cette emprise des paramètres de qualité et les métadonnées correspondantes à ce levé. Or, le problème de la validation de la qualité des données hydro-océanographiques dépend de multiples facteurs. Il devient alors difficile de vérifier si un jeu de données est compatible avec un concept d'emploi spécifique (défini par les préférences d'un expert). Les travaux de la section 4.2.2 ont défini et validé une approche qui combine l'utilisation de paramètres

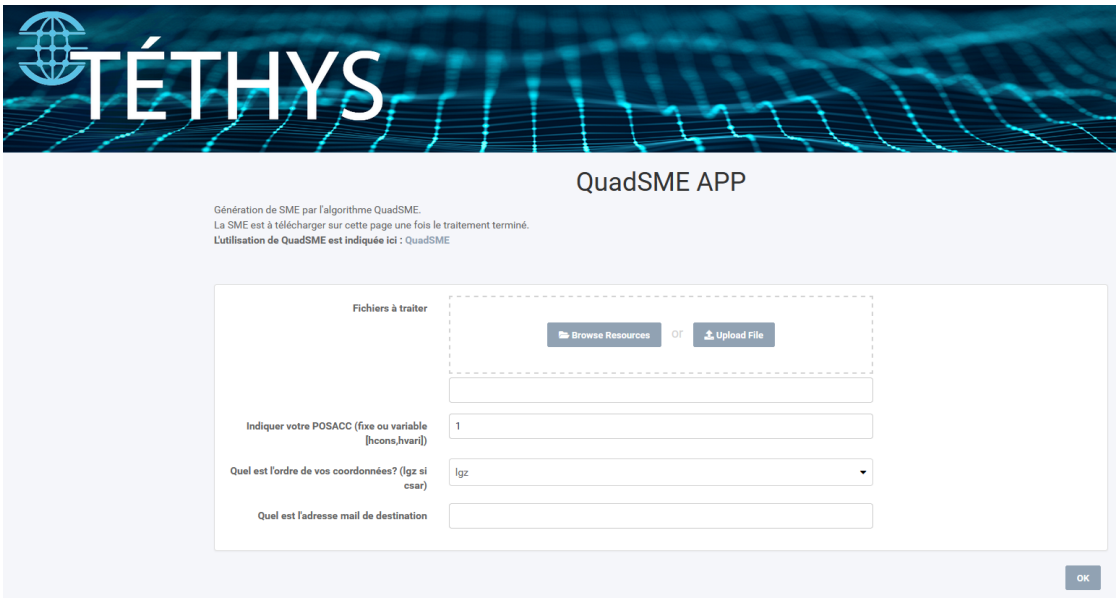


FIGURE 4.29 – Interface web de l'utilisation de la méthode QuadSME via l'outil d'ordonnancement FME Server®.

de qualité de la donnée et d'aide multicritère à la décision pour résoudre ce problème. Toutefois, la métadonnée bathymétrique contient des paramètres de qualité (sous forme de paires clé-valeur). Ces valeurs ne fournissent pas d'information sur l'utilisation qui peut être faite de ce jeu de données et ne sont pas utiles pour tous les utilisateurs (comme montré dans la sous-section précédente). L'intérêt principal de la méthode présentée est de rendre l'analyse de la qualité explicable via l'approche MCDA en ce qui concerne les profils des experts, en sélectionnant les paramètres de qualité et en indiquant une préférence sur ces paramètres. Des paramètres supplémentaires de qualité de la donnée peuvent être étudiés et comparés avec les estimations réalisées par les experts dans le but de définir un indicateur global de confiance des données acquises. Le travail n'a été réalisé, pour l'instant, que sur un ensemble restreint de paramètres de qualité dont une majorité issue directement des métadonnées. Il serait pertinent de poursuivre la réflexion sur un ensemble plus important de paramètres de qualité pour généraliser la chaîne de traitement proposée et prendre en compte tous les paramètres de la qualité au Shom et pour d'autres SH.

Dans l'optique de démontrer l'intérêt de l'approche proposée, nous prévoyons également de déployer cette méthode sur des levés en contexte réel. Plutôt qu'utili-

ser une approche d'extraction de la préférence directe (via les entretiens par profil d'expertise) pour déterminer les paramètres de préférence de chaque expert, une approche indirecte serait employée, dans laquelle l'expert serait confronté à des levés passés qu'il doit évaluer. Cette information serait ensuite utilisée par un algorithme d'apprentissage machine qui déterminerait automatiquement les paramètres du modèle de préférence. Enfin, point primordial, l'explication de l'évaluation finale est actuellement réalisée à la main. L'automatisation de cette tâche via des algorithmes de génération de règles permettrait de produire facilement un rapport de qualité exploitable par les utilisateurs finaux. Cela faciliterait l'utilisation du processus mis en place.

La section 4.2.3 démontre clairement l'intérêt de notre méthode d'aide à la décision proposée pour un usage à destination du Shom. En effet, à partir d'une campagne hydro-océanographique (données et métadonnées associées), nous évaluons le niveau global de qualité mais également l'utilité de la donnée et de l'information pour différents concepts d'emploi finaux : hydrographique, acoustique ou océanographique. Elle peut donc être utilisée *a posteriori* pour évaluer l'intérêt d'un levé après son acquisition mais pourquoi pas également en temps réel pour surveiller la qualité de la donnée acquise quotidiennement lors d'une campagne, [8].

Pour finir, notre proposition peut également être utilisée *a priori* pour planifier de futurs levés hydrographiques selon les besoins exprimés par le prescripteur du levé. Le levé répondrait ainsi strictement aux exigences du Programme National d'Hydrographie (PNH) [186] et empêcherait la surqualification, c'est-à-dire la réalisation d'un levé avec un niveau de qualité plus important que demandé par le prescripteur. Cela permettrait également de s'assurer du bon emploi des moyens coûteux à l'usage (les coûts d'un levé hydro-océanographique étant d'environ 40k€/jour pour le bâtiment hydrographique et océanographique Beautemps-Beaupré) et de fusionner éventuellement des campagnes ayant des objectifs différents, tout en s'assurant du respect de la qualité pour chaque concept d'emploi voulu. Cette méthodologie s'avérerait donc un outil très pertinent d'aide à la décision et de fusion d'information pour la préparation aux campagnes hydro-océanographiques.

Chapitre 5

# Discussion

## Contributions de ce travail de recherche

Dans le cadre de ce travail de recherche, trois contributions majeures dans le domaine des sciences des données et de la décision ont été présentées :

- une étude détaillée des algorithmes de détection des données aberrantes et une taxonomie associée dans le contexte des données nuages de points bathymétriques, voir 2.3 et [12] ;
- la mise en place d'une structure de donnée innovante permettant l'usage d'algorithmes de régression prometteurs pour traiter la problématique de la détection des données aberrantes dans les nuages de points bathymétriques, voir 3.2 et [14] ;
- la mise en place d'un processus basé sur le MCDA pour déterminer la qualité d'une donnée tout en prenant en compte les préférences de l'utilisateur, voir 4.2.2 et [15].

Dans le cadre de l'état de l'art exhaustif, voir 2.3, plus de 80 articles ont été étudiés pour n'en conserver que 39 dans la taxonomie. Nous avons ainsi pu mettre en avant les différentes techniques de détection des données aberrantes bathymétriques tout en notant comment ces techniques se sont adaptées au fur et à mesure du temps, voir la figure 2.5, aux nouvelles capacités des capteurs évoluant ainsi d'une approche centrée sur la sonde à une approche beaucoup plus structurée centrée sur le produit final. Avec l'arrivée massive des algorithmes d'apprentissage machine, nous pouvons également constater ces dernières années une nouvelle bascule dans les algorithmes passant d'un paradigme non-supervisé à des techniques supervisées. Pour pousser la réflexion plus loin, il serait intéressant d'étudier les impacts des différents articles sur la communauté scientifique et notamment les interconnexions qui existent par le biais des citations. La figure 5.1 présente une amorce de cette approche en recensant le nombre de citations pour chaque article.

Nous pouvons ainsi noter clairement que l'algorithme CUBE [69] est la technique la plus utilisée et la plus citée. Ceci s'explique par le fait que cette technique a été implémentée dans de nombreux logiciels commerciaux, favorisant ainsi son utilisation par les principales organisations hydrographiques nationales.

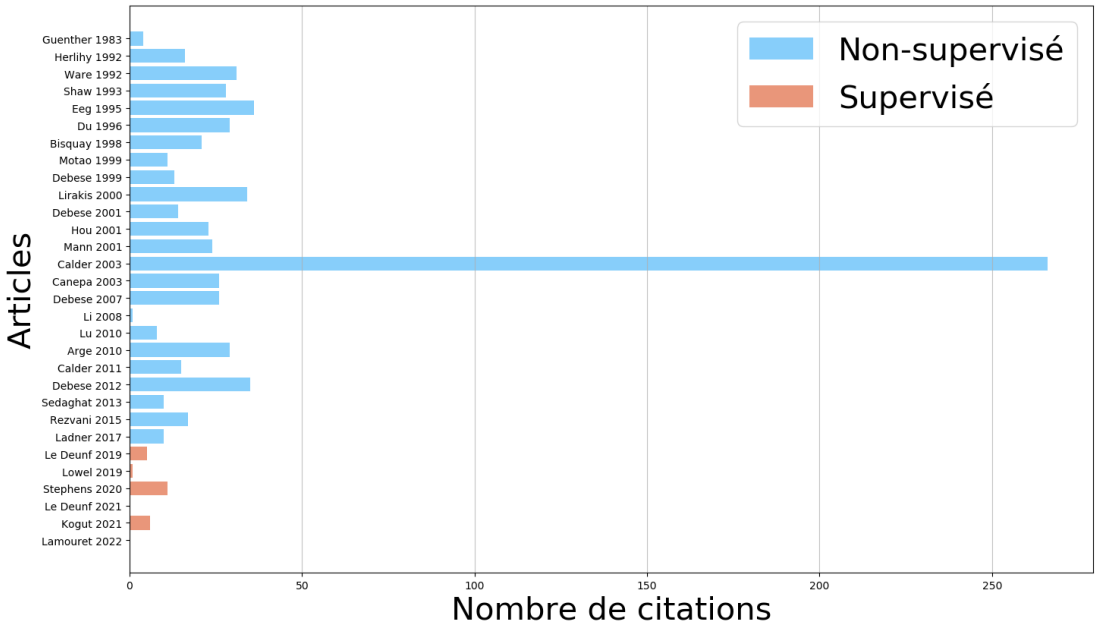


FIGURE 5.1 – Nombre de citations par algorithme (extraites de Google Scholar le 12/10/2022).

Concernant la mise en place d'une structure de donnée spécifique pour la détection des données aberrantes dans les nuages de points bathymétriques, contrairement aux recherches précédentes sur le LiDAR bathymétrique (voir [72], [107]) qui se concentrent sur la classification de nuages de points LiDAR bathymétriques, nous avons formalisé notre question comme un problème de régression. En effet, dans le cadre de l'exploitation des nuages de points LiDAR bathymétriques pour des produits nautiques, et notamment pour la sécurité de la navigation, il est intéressant de construire un modèle établi sur une régression et non sur une classification binaire stricte. Nous notons que ces approches régressives ne semblent pas avoir été utilisées dans les études précédentes. De plus, la compréhension de la sortie du processus prédictif par les opérateurs est largement facilitée par la génération d'un MNT, en termes de contrôle et de qualité des données, plutôt qu'apposer une étiquette (accepté/rejeté) sur les sondes. Notre approche apporte

ainsi les contributions suivantes :

- combinaison d'une approche orientée surface avec des méthodes de régression supervisée ;
- comparaison des performances de différents régresseurs intégrés à notre approche ;
- création d'une nouvelle structure de données 2D + 1D (en s'inspirant des études sur la donnée hyperspectrale, voir [139]) pour la donnée LiDAR bathymétrique ;
- utilisation de la complémentarité des capteurs 2HD et 3HD ;
- mise à l'échelle avec de grands ensembles de données pour l'apprentissage et les tests (plus de 40 millions de sondes par ensemble de données) ;
- technique hybride supervisée et non-supervisée.

Au vu des résultats des expérimentations, il est nécessaire de consolider la généralisation de la méthode aux différentes zones mais également de s'assurer d'un maximum de fluidité dans le parcours utilisateur pour l'usage de ces algorithmes. Ainsi, il faut travailler sur les sorties de l'algorithme pour améliorer son ergonomie sans détériorer les résultats. Enfin, après avoir testé la méthode sur de nombreux jeux de données LiDAR, il faudrait tester un apprentissage sur la donnée SMF et étudier les résultats obtenus pour faciliter le traitement des opérateurs pour cette donnée.

Enfin, nous avons cherché dans notre dernière contribution à répondre à l'interrogation suivante : "comment les modèles de préférence et les méthodes MCDA peuvent-ils être utilisés pour déterminer la qualité d'un ensemble de données pour différentes utilisations, en fonction des priorités et des préférences variables des utilisateurs ?" Cette question n'avait jamais été abordée auparavant dans la littérature et nous proposons une approche permettant d'initier une réponse généralisable à de nombreux domaines. Dans notre cas, les entrées de l'évaluation MCDA sont des paramètres de qualité estimés de flux de données brutes provenant de capteurs, dans des conditions d'acquisition spécifiques liées à l'environnement. Notre travail propose un processus d'évaluation de la qualité des données fonction des besoins de l'utilisateur, en sélectionnant des paramètres de qualité prédéfinis par un expert. Ce processus utilise ces paramètres comme entrée d'une méthode MCDA et évalue les profils de qualité d'un ensemble de données au travers de modèles de préférence, tout en conservant un très haut niveau d'explicabilité et en utilisant une méthode

complètement interprétable qui facilite l'intégration dans les processus d'un SH quand il s'agit d'exprimer la qualité d'une information.

Les limitations rencontrées aujourd'hui sont le nombre assez faible de levés testés et d'entretiens menés avec les experts. Il serait intéressant dans le futur de consolider l'inférence des paramètres tout en testant dans le même temps davantage de levés pour vérifier si les résultats restent cohérents. De plus, la détermination automatique des paramètres de qualité doit être approfondie afin de faciliter le travail des opérateurs dans la phase de rédaction des levés hydro-océanographiques, ce qui permettra aussi d'ajouter des paramètres de qualité au modèle MCDA.

## De la sonde au jumeau numérique bathymétrique : le projet Téthys

Comme présenté dans l'introduction, les données bathymétriques du Shom sont actuellement archivées dans la BDBS et gérées comme une pile de levés susceptibles de se superposer (plus de 11 400 levés en base). Les données issues de ces levés proviennent de différents types de catpeurs/systèmes comme présentés dans la section 1.1.2. Ainsi, la qualité de la donnée stockée dans cette base est variable de par l'évolution des procédures d'acquisition dans le temps. Jusqu'en 2021, chaque cartographe générait une surface de référence bathymétrique via un processus laborieux de sélection de l'information bathymétrique dans cette base, la surface de référence faisant office de socle pour toute action sur la donnée bathymétrique apposée sur une CM. Depuis, le Shom a lancé le projet Téthys qui vise à constituer cette surface de référence bathymétrique et ainsi proposer la meilleure connaissance bathymétrique actualisée pour tous types de besoin (défense, génération de MNT et cartographie nautique). Cette surface de référence ne conservera que les données bathymétriques jugés comme meilleure (processus de fiabilisation de la donnée) en les comparant localement (processus de déconfliction). La Téthys utilise comme donnée source toutes les informations présentes dans la BDBS. La Téthys repose ainsi sur des éléments présentés dans ce manuscrit. La fusion de données bathymétriques afin de fournir une compilation cohérente de données appartient à un domaine de recherche actuellement actif, voir [163], [187]-[190]. Par exemple, la NOAA a lancé en 2021 son projet BlueTopo™, voir [191], qui vise à produire une référence bathymétrique dans la Zone Économique Exclusive (ZEE) américaine.

La génération de cette surface de référence permettra d'accélérer la production par le Shom des CM et des MNT bathymétriques en capitalisant les efforts sur la sélection de données, tout en renforçant la gestion et la filiation de l'information originelle. Cette démarche répond à l'objectif suivant : une acquisition unique pour une utilisation multiple. La construction de cette surface de référence se fait via le processus suivant :

1. à partir des différents levés sources, une vérification de toutes les données et du contenu des métadonnées est effectuée ;
2. pour chaque intersection de jeux de données, des règles de priorité sont définies entre ces jeux de données de qualités et d'âges divers, en fonction des éléments d'analyse de qualité décrits dans les métadonnées ;
3. cette donnée est compilée en respectant les priorités définies précédemment entre les jeux de données.

Le projet Thétys, qui a fait l'objet d'une publication [16] et de présentations à deux conférences internationales [21], [25], vise à automatiser les processus de gestion de la donnée destinée à la production d'information bathymétrique opérationnelle. Il permet ainsi de :

— mettre rapidement à disposition (dans un premier temps au sein du Shom uniquement) la connaissance bathymétrique à jour et ainsi accélérer considérablement la vitesse de transmission des informations mises à jour qui seront ensuite transformées en produits à destination des utilisateurs de données bathymétriques. Aujourd'hui, plusieurs années peuvent s'écouler avant qu'un levé bathymétrique ne soit pris en compte dans son intégralité sur une CM ;
— améliorer la qualité et la cohérence des données de référence ;
— renforcer la cohérence des données et des produits du Shom en maintenant le cap sur la sécurité des utilisateurs marins ;
— anticiper l'inflation du volume de données bathymétriques provenant des nouveaux capteurs et nouvelles technologies, notamment dans le cadre du programme CHOF. En effet, le projet Téthys fournira de la donnée bathymétrique facile d'accès à destination d'un large panel d'opérateurs ;
— les prescripteurs de levé auront accès à la synthèse de la meilleure connaissance des informations bathymétriques actuelles, ce qui les aidera à planifier

plus pertinemment les levés dans le cadre du PNH (connaissance bathymétrique à interroger géographiquement au regard de la stratégie ou de critères commerciaux) ;

— les producteurs de données bathymétriques pourront utiliser la Téthys pour les aider à qualifier leurs nouveaux levés. Cette qualification sera réalisée en analysant ce nouveau levé au regard de la connaissance du moment ;

— pour la génération de produits bathymétriques, les cartographes ou les producteurs de MNT seront en capacité de visualiser et d'extraire seulement les données nécessaires pour leur zone de travail, en opérant une sélection entre la donnée existante en cours de réalisation et de contrôle. Le Shom prévoit que la Téthys lui permettra de gagner au moins 10% du temps nécessaire aux cartographes pour qu'ils puissent compiler les cartes nautiques.

La figure 5.2 présente les macro-tâches composant la chaîne de traitement du processus permettant de générer la BDD Téthys. L'algorithme QuadSME présenté dans la section 4.1.2 est utilisé dans la macro-tâche de fiabilisation de la donnée et plus spécifiquement pour le contrôle de la SME, voir 4.1. En effet, pour les levés SMF les plus anciens, la génération de la SME était réalisée manuellement et pouvait être peu fidèle vis-à-vis de l'emprise de la donnée bathymétrique (sondes hors de la SME ou SME trop lâches, contenant des trous de couverture importants).

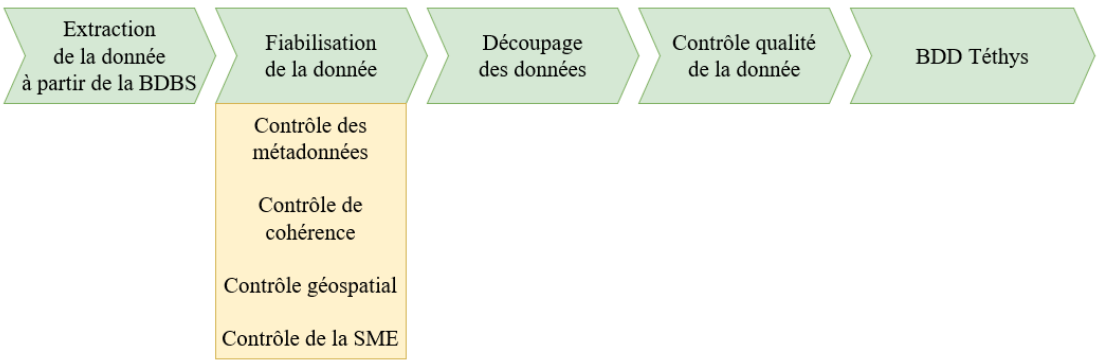


FIGURE 5.2 – Chaîne de traitement du processus menant à la génération de la BDD Téthys.

La première déconfliction effectuée est présentée sur la figure 5.3 qui distingue les étapes avant et après ce processus. Sur cette première dalle (appelée 145_81, nommage inspiré du carroyage MARSDEN), ce sont 115 levés qui ont été utilisés en donnée d'entrée et 441 418 088 sondes associées à traiter. À la fin de la chaîne

de traitement, seulement 96 levés ont finalement été retenus et 310 970 981 sondes intégrées dans la Téthys. Parmi ces sondes, plus de 6 000 sondes ont été numérisées à partir d'anciennes CM [192]. La zone couverte s'étend du port de Saint-Malo à l'ouest, à la baie du Mont Saint-Michel à l'est, et du sud de la Rance à la ville de Coutance au nord. Le résultat de ce processus de déconfliction a soulevé 44 167 conflits entre des sources de données qui s'intersectent. Le contrôle qualité de la dalle est effectué en comparant, entre autres, les produits de navigation générés précédemment, tels que les Electronic Navigational Chart (ENC) officielles.

Ce premier travail a profité récemment aux cartographes qui ont publié la carte nautique couvrant les îles Chausey et la production du MNT topo-bathymétrique qui couvre les approches de Saint-Malo, voir [193].

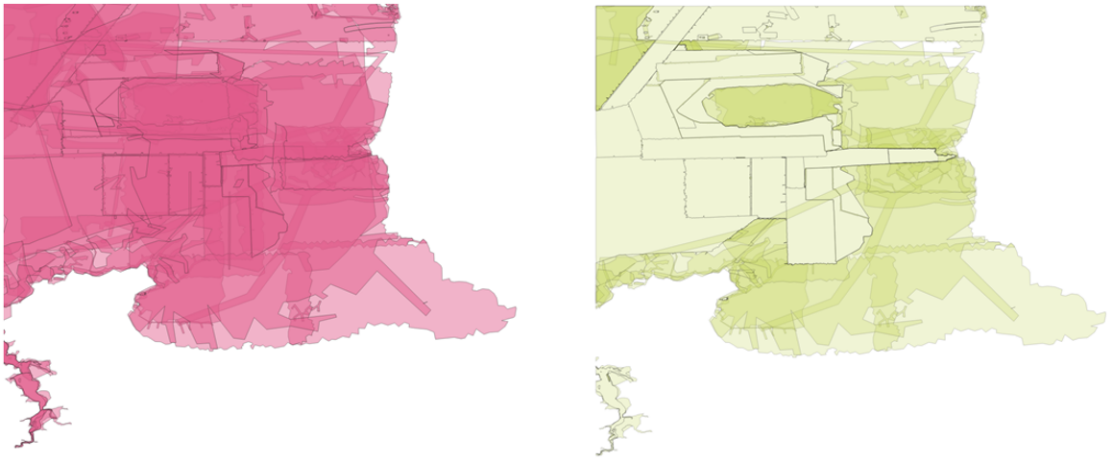

FIGURE 5.3 – Première tuile du projet Téthys, à gauche l'ensemble des données étudiées pour la déconfliction, à droite les levés coupés et conservés après le processus de déconfliction.

Afin que l'accès à la donnée se fasse le plus facilement possible, un portail web, voir 5.4, a été construit dans le cadre de ce projet. Le portail de données Téthys est un site interne dont l'interface est basée sur un SIG et qui fournit un moyen d'accès rapide et facile à la surface de référence bathymétrique ainsi que les informations qui y sont associées. Cette interface utilisateur web évite toute installation d'un SIG sur le poste client et améliore donc l'accessibilité à la donnée. Trois composants de la surface de référence bathymétrique peuvent actuellement être consultés à tout moment :

— la couverture spatiale de chaque levé présent dans Téthys, représenté par

une couche de polygones vectoriels ;

— les sondes, représentées par une palette de couleurs configurable. La fluidité de l'affichage est assurée par une décimation des sondes à la volée selon le niveau de zoom ;

— la métadonnée associée aux levés bathymétriques sélectionnés (auteur, incertitude de position verticale et horizontale, technique de sondage, restrictions d'accès ...) est disponible dans le panneau "attributs". Les liens aux ressources associées sont fournis pour quiconque rechercherait des informations complémentaires (exemple : rapports de levé hydrographique, minutes bathymétriques).

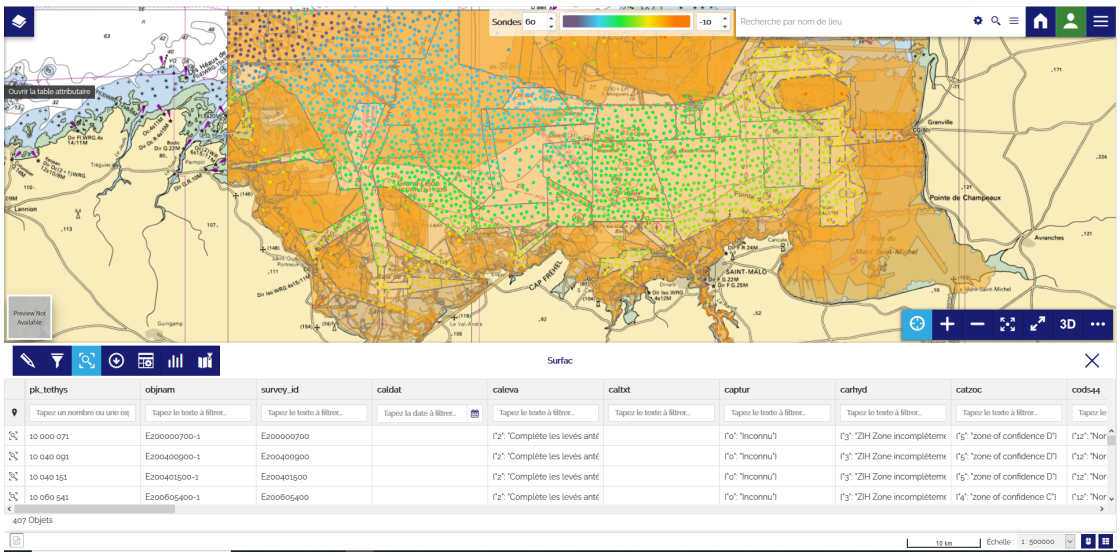


FIGURE 5.4 – Page principale du portail de données Téthys.

Ce portail permettra dans le futur d'accéder à des informations de référence fournies par le PNH, voir [186], avec la connaissance bathymétrique actuelle des eaux françaises. Le PNH est un document stratégique qui fixe l'objectif de connaissance bathymétrique sur une plage de 4 ans, en fonction de différents critères, tels que le trafic, les risques, les itinéraires recommandés, l'environnement ...L'accès direct à la connaissance permet de suivre de manière plus précise et plus détaillée la mise en œuvre du PNH. Ainsi, ce suivi garantit une meilleure visibilité et une priorisation des nouvelles zones à lever pour les personnes en charge de la planification des levés.

Cependant, le concept d'emploi actuel de la Téthys est orienté sur une utili-

sation cartographique, qui se traduit par la mise en place de plus de 300 règles expertes visant à assurer le contrôle et la déconfliction des levés bathymétriques en BDBS. Or, des concepts d'emploi différents (utilisation océanographique ou acoustique) pourraient exiger la mise en place et le respect de règles différentes, de la même façon que ce qui a été présenté dans la section 4.2.2.

Il serait intéressant de coupler les approches autour de la MCDA, proposées dans ce manuscrit, avec les approches développées dans le projet Téthys, afin d'identifier les règles communes (comme la construction de la SME la plus fidèle) et les règles particulières (comme la comparaison de paramètres de qualité spécifiques à un besoin océanographique). Ce couplage permettrait de généraliser le concept d'emploi de la Téthys à d'autres besoins autour de la donnée bathymétrique. Nous gagnerions également en rapidité dans la fourniture de réponses à des finalités plus spécifiques, comme la construction de surfaces de référence bathymétrique pour les reconstructions morphologiques historiques, destinées à l'étude de l'impact de l'érosion du littoral, et la montée des eaux au cours des siècles.

Dans le cadre de ce projet, l'automatisation de certaines tâches de faible niveau technique pose la question de l'impact métier de ce type de méthodologie qui vise à faciliter le travail quotidien des opérateurs.

# Impact métier et méthodologique de l'Intelligence Artificielle dans la chaîne de valeur bathymétrique

À partir des années 80, les SH, dont le Shom, ont commencé à opérer une première transformation numérique en utilisant des sondeurs numériques, voir [3]. Ainsi, dès cette période, des impacts méthodologiques et métiers se sont fait ressentir : l'hydrographe s'est éloigné progressivement de la sonde analogique pour manipuler la donnée numérisée. Aujourd'hui, nous avons constaté que le même phénomène se reproduisait au Shom qui est amené à utiliser de nouvelles technologiques de l'information fondées sur des processus d'IA. Le Shom n'est pas le seul SH national à se questionner sur l'impact au sein de son environnement de travail de ces évolutions associées aux nouvelles technologies qui peuplent aujourd'hui notre quotidien. Le Service Hydrographique du Canada (SHC) a par exemple tenu un atelier sur "l'hydrographe du futur" au mois d'avril 2022. Il en est notamment ressorti que nous ne pouvions plus identifier de *"profil unique d'hydrographe"*,

qu'il était dorénavant nécessaire de mettre en place des techniques automatisées ou issues de l'IA pour analyser les données en mer, et qu'il fallait intégrer des formations supplémentaires dans les domaines suivants : *"mathématiques, données géospatiales et développement informatique"*.

De façon générale, l'impact de l'IA au sein du ministère des Armées est particulièrement à l'étude comme le montre le rapport [194]. En effet, le ministère vise à construire des IA de confiance, robustes et non vulnérables aux attaques ciblées (empoisonnement des données, leurrage, brouillage, porte dérobée, etc.).

Ainsi, dans le cadre de ces travaux, et plus particulièrement des processus associés à une prise de décision tels que le traitement de données bathymétriques et la détermination de la qualité d'un levé hydro-océanographique, nous avons cherché à construire des processus d'IA interprétables qui reposent sur des expertises des agents du Shom. Pour compléter cette approche, il faudrait prendre en compte les besoins spécifiques opérationnels d'un SH comme la vélocité d'une solution technique, la possibilité d'embarquer la solution à la mer et les notions de protections du secret.

Par construction, l'intégration des modèles de qualité de l'information et de préférence d'évaluation de la qualité des levés hydro-océanographiques permet à l'utilisateur d'appréhender parfaitement le résultat fourni. Celui-ci peut facilement identifier des pistes d'amélioration de la qualité d'un levé hydro-océanographique en suivant les recommandations fournies par la méthode. Le paramétrage est aujourd'hui réalisé par des experts mais, dans le futur, nous pourrions tester l'apprentissage de nouveaux paramètres à partir des données déjà qualifiées par les experts. Cet apprentissage permettrait de capter des éléments plus complexes, non fournis actuellement par les experts ou de faire ressortir de nouvelles manières de décrire la donnée afin de nourrir le processus MCDA.

Concernant l'outil de traitement de données bathymétriques via une approche ML, des impacts métiers et organisationnels ont pu être évalués durant les différentes phases d'expérimentation de l'outil mais aussi via des entretiens avec les opérateurs et l'encadrement du département AL du Shom, voir la section 3.3.

Au sujet des impacts sur le processus métier, un gain de temps est pressenti, notamment sur les zones jugées comme simples par les opérateurs (zones sableuses et par plus de 15 mètres de fonds). Un passage à l'échelle pourrait permettre

de conclure sur cette intuition. Ce temps libéré permettrait aux opérateurs de se focaliser sur l'analyse et le traitement des zones plus complexes améliorant globalement la qualité des données traitées et permettant une homogénéisation du traitement de la donnée.

Concernant les impacts pressentis sur l'organisation, l'outil est facile à intégrer dans les processus actuels. La partie ML n'impacte pas la préparation des données avant traitement et ne comporte que du temps machine, comme présenté dans la figure 3.4. De plus, l'outil permettrait une revalorisation du métier des opérateurs en le transformant et en le recentrant sur des activités à plus forte valeur ajoutée, actionnant ainsi un levier de motivation et donnant un nouveau sens à leur métier pour en augmenter son attractivité aujourd'hui mise à mal comme en témoignent des taux importants de renouvellement de personnels.

Toutefois, l'outil nécessitera un temps de montée en compétence pour permettre aux opérateurs expérimentés d'adapter leurs méthodologies de travail à ce nouvel outil, ce qui constitue un enjeu important de la conduite du changement. Il sera donc nécessaire de prendre en compte une phase de formation pour que les opérateurs comprennent la méthodologie globale et surtout pour développer leur sens critique. Effectivement, nous ne souhaitons pas que l'opérateur s'appuie trop rapidement et aveuglement sur l'outil, et oublie de conserver cet oeil critique sur les résultats, démarche qui est le coeur du métier de l'hydrographe. Enfin, il reste aujourd'hui un effet d'encodage qui perturbe la compréhension de l'indice associé à chaque sonde par l'utilisateur et ce car l'indice est encodé sur les attributs RGB du format .las (seuls attributs disponibles). Une amélioration du logiciel PFM ABE permettrait de réduire ce risque en utilisant un attribut dédié au degré d'appartenance au fond pour chaque point.

Ces impacts doivent encore être analysés plus finement au cours d'autres expérimentations pour assurer la meilleure conduite du changement possible, en associant le plus possible d'opérateurs au processus.

Cette crainte d'une méthode type "boîte noire" est également partagée par les autres acteurs du monde de l'hydrographie. Dans ce même atelier sur "l'hydrographe du futur" tenu par le SHC, les participants relèvent un risque de dépendance jugée excessive à l'égard de l'IA et rappellent la nécessité de *"conserver le sens commun/expertise"* des données de la chaîne de valeur bathymétrique. Ainsi, la pédagogie et la formation semblent être des éléments clés de l'adhésion de ces

nouvelles technologies afin de les rendre accessibles à tous. De plus, avec l'usage toujours plus important des réseaux adverses génératifs [195] (*generative adversarial networks* ou *GANs* en anglais), de plus en plus de chercheurs s'interrogent sur les capacités de ces algorithmes à "tricher" (sans contrôle de l'humain) afin d'être récompensés plus rapidement, comme présenté dans l'article [196].

Enfin, les SH nationaux, producteurs des produits nautiques, sont responsables légalement de la sécurité de la navigation portée par ces produits. Au sein du Shom, la chaîne de responsabilité est ainsi clairement identifiée (via des matrices de responsabilités par exemple) pour l'acquisition de la donnée, les processus de contrôles, l'exploitation finale et la production d'ouvrages nautiques. Avec l'arrivée de processus exploitant l'IA, il est nécessaire de réfléchir aux impacts légaux associés à la production de la donnée. La question se pose ainsi d'identifier clairement les responsabilités lors de l'utilisation de l'IA : serait-ce l'utilisateur du processus, celui qui a construit la base d'apprentissage, le concepteur de la chaîne de traitement ou les trois à la fois ? Dans le cadre de données personnelles, le règlement général sur la protection des données (RGPD), voir [197], indique le cadre à suivre concernant la collecte et l'exploitation des données personnelles. Toutefois, pour ce qui est des données géospatiales qui ne présentent pas les mêmes enjeux de sécurité, de souveraineté, de pouvoir de nuisance et de discrimination, le cadre légal semble pour l'instant encore très ouvert.

## Généralisation de la méthode aux autres producteurs de données bathymétriques

Dans le cadre d'un travail de recherche et développement, la généralisation des méthodes développées est essentielle de manière à fédérer une communauté autour de pratiques communes permettant l'échange de données et d'informations, comme l'OHI le réalise actuellement dans le cadre de la mise en place des nouvelles normes des produits de navigation S-1XX, présenté dans la vidéo [198].

Les outils générés dans le cadre de cette thèse présentent différents niveaux de généralisation vers les autres producteurs de données bathymétriques. Ils sont classés du plus facilement généralisable au plus complexe :

1. Intégration des modèles de qualité de l'information et de préférence pour

évaluer la qualité des levés hydro-océanographique, voir section 4.2.2 ;

2. QuadSME, voir section 4.1.2 ;
3. Traitement bathymétrique via ML (TraitLIA), voir section 3.2.

La méthode MCDA utilisée pour la qualité des levés hydro-océanographiques est par construction facilement généralisable. En effet, dans le processus de détermination de la qualité d'un levé hydro-océanographique, il est prévu le cas où le concept d'emploi n'est pas connu par la méthode et qu'il est donc nécessaire qu'un expert paramètre un nouveau type d'utilisation. L'algorithme peut alors être paramétré à destination d'autres usages ou d'autres besoins. De fait, la méthode peut se généraliser à d'autres domaines que les géosciences marines, dès l'instant où la qualité d'une information dépend d'une préférence utilisateur, d'un concept d'emploi ou d'un contexte environnemental. La méthode proposée peut être utilisée pour déterminer la qualité de cette information en fonction des éléments cités précédemment. Ainsi, tous les SH (portuaires ou nationaux) qui cherchent à évaluer la qualité d'une campagne hydro-océanographique peuvent s'appuyer sur ce processus pour une évaluation adaptée à leur préférence, avec une compréhension claire des paramètres critiques pour la détermination de la qualité finale de la campagne.

La méthode QuadSME, quant à elle, voir 4.1.2, prend pour unique entrée un nuage de points bathymétriques d'un levé hydro-océanographique et détermine l'emprise spatiale (2D dans le plan longitude / latitude) associée à cette donnée d'entrée. L'algorithme est donc exploitable par toute personne possédant ce type de jeux de données et souhaitant déterminer son emprise en suivant les mêmes règles que présentées dans la section 4.1. En revanche, si un producteur de données souhaite modifier ces règles, en changeant par exemple les facteurs de dilution / érosion ou bien la détermination des sondes isolées ne participant pas à un regroupement de sondes, il serait nécessaire d'adapter la solution en modifiant le code source de l'algorithme. L'ensemble de l'algorithme a été développé sur Python et l'ETL FME®, la diffusion des scripts est donc simple mais nécessite une licence propriétaire pour faire fonctionner l'intégralité de la solution. Nous pourrions basculer l'entièreté du processus en Python pur mais nous perdrions alors un peu en vélocité, tant qu'une optimisation de type *multiprocessing* n'aura pas été réalisée.

Qiant à la méthode de traitement bathymétrique via ML (TraitLIA), elle, prend

en entrée le format open source .las, voir [137] et repose sur un code source uniquement en Python. Cette méthode est donc très ouverte (en termes d'outils mis en oeuvre) et facilement généralisable vers d'autres producteurs de données bathymétriques qui souhaiteraient l'utiliser dans leur chaîne de valeur bathymétrique. Néanmoins, l'apprentissage associé aujourd'hui dans le modèle de régression RF est basé sur un traitement manuel réalisé par des opérateurs du département AL du Shom. L'intelligence dans la base d'apprentissage est donc directement corrélée à la méthodologie mise en place dans le traitement manuel de ces opérateurs. Il n'est donc pas impossible que des producteurs de données avec une autre méthodologie ne soient pas satisfaits du type de sortie de l'algorithme, ce qui nécessiterait un nouvel apprentissage à partir de leurs données étiquetées. De plus, comme présenté dans la section 3.3.3, les .las sont transformés de manière à être facilement manipulés dans le logiciel open source PFM ABE. Donc, si un producteur de données utilise un autre type de solution de visualisation / traitement de données LiDAR, alors il faudrait qu'il s'assure tout d'abord d'avoir accès aux attributs modifiés (pour rappel les attributs RGB). Enfin, l'algorithme a aujourd'hui été entraîné sur de la donnée LiDAR. Il serait intéressant de tester la même approche sur la donnée SMF ainsi que d'étudier les capacités de généralisation sur une autre source de données.

Ces trois outils sont donc parfaitement utilisables par d'autres producteurs (au sens large allant de l'acquisition, le traitement, la qualification et l'exploitation) de données bathymétriques avec plus ou moins de modifications en fonction de l'adéquation qu'ils veulent mettre en place pour leurs méthodologies actuelles.

## Interconnexion des différents niveaux d'échelle

Quelle interdépendance existe entre les différents niveaux de zoom proposés dans le cadre de ce travail de recherche (le niveau micro sur la donnée, le niveau meso sur l'information et le niveau macro sur la décision) ?

Dans ce manuscrit, nous avons en effet pu mettre en avant un niveau évident de connexions, qui suit le chapitrage de ce document, avec une vision ascendante comme présentée dans la figure 5.5. Ces connexions ascendantes sont les plus naturelles car chaque niveau N+1 a besoin du niveau N pour exister. Enlever un niveau de la pyramide a alors un effet sur l'entièreté de celle-ci. Ainsi, l'existence

et la capacité de traiter la donnée (niveau micro) conditionne l'information qui est la donnée structurée et analysée (niveau meso). De la même façon, la prise de décision (niveau macro) repose majoritairement sur de l'information afin de prendre une décision éclairée et reposant sur de multiples arguments. Finalement, la prise de décision sert un besoin spécifique et répond souvent à une question posée par un utilisateur final (*end-user* en anglais).

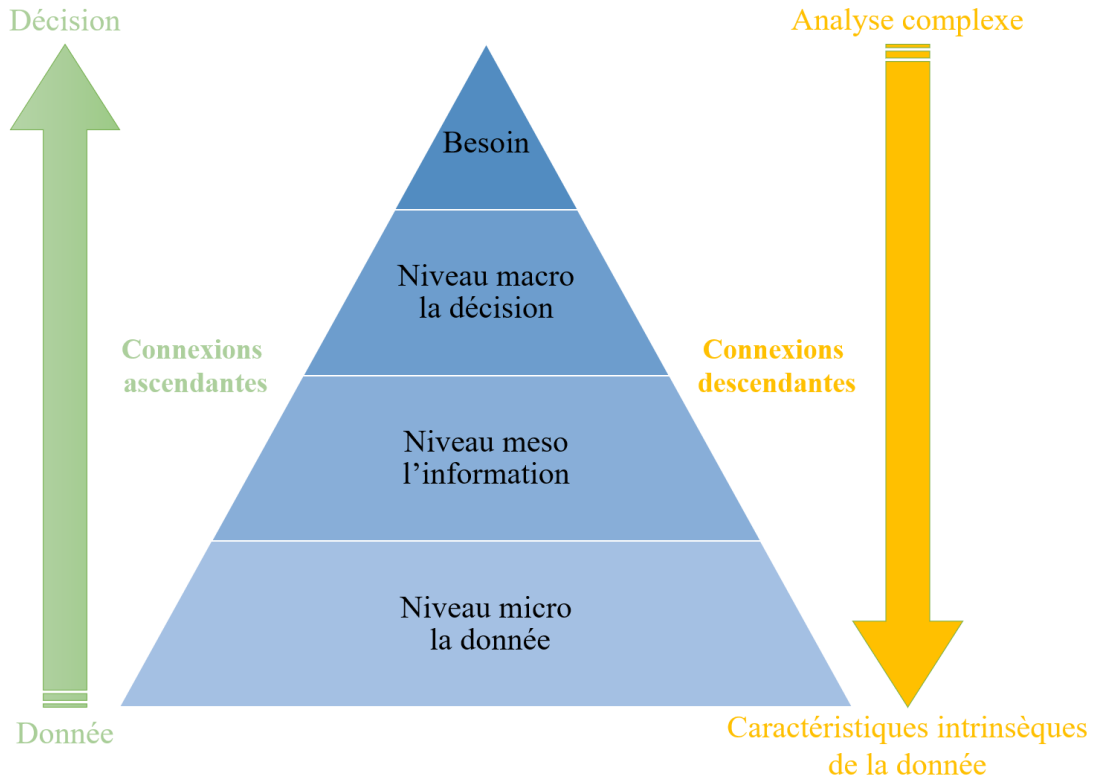


FIGURE 5.5 – Interconnexion à travers les niveaux d'échelle présentés dans le manuscrit.

Dans le contexte des systèmes de levé hydrographique, nous avons pu constater l'importance de s'appuyer sur différentes sources de données pour construire un premier niveau d'information avec les sondes bathymétriques géoréférencées, puis un second niveau avec la structuration mise en place et les descripteurs associés afin d'augmenter le niveau de connaissance des données acoustiques brutes discrètes. Cette interdépendance ascendante permet aussi de créer une valeur ajoutée et enrichir la donnée brute en la transformant en une information beaucoup plus complexe. De la même façon, l'information obtenue à partir des données bathy-

métriques traitées et qualifiées permet la mise en place d'un algorithme de prise de décision comme présenté dans la section 4.2.2 ou des processus de déconfliction comme présenté dans la section 5.

Dans le cadre de la prescription des levés et la préparation des levés, nous nous plaçons dans une interconnexion inverse - descendante, voir la figure 5.5. En effet, pour ce type de contexte, le premier élément de réflexion consiste à comprendre le besoin associé à l'acquisition sur une emprise spatiale (macro). À partir de la définition du besoin, le prescripteur de la campagne peut décider du meilleur moyen d'y répondre et fournir les informations pertinentes (meso) qui vont permettre l'acquisition de la donnée (micro). Cette interdépendance descendante permet de ce fait de toujours conserver en ligne de mire le concept d'emploi et le plus juste besoin utilisateur associé à l'ensemble de la chaîne de valeur proposée par les levés hydro-océanographiques.

Ces interconnexions ascendantes et descendantes forment la majorité des interdépendances classiques associées aux processus qui s'intéressent à la donnée. Néanmoins, dans le cadre hydrographique, cette interconnexion se joue aussi sur des niveaux d'échelle au sens cartographique [3]. En effet, au sens strictement cartographique, l'échelle désigne le rapport entre une distance réelle (mesurée dans l'espace terrestre) et sa représentation sur une carte. L'échelle cartographique permet donc de définir un ordre de grandeur dans lequel nous étudions un espace géographique. Puisqu'il s'agit d'un rapport, l'échelle est petite lorsque le dénominateur est grand et donc que la zone géographique d'intérêt est très étendue (échelle mondiale, ou continentale en topographie terrestre). Inversement, l'échelle est grande lorsque le dénominateur est petit (du *1 :1 000 000e* au *1 :25 000e* par exemple) et que la zone géographique étudiée est beaucoup plus restreinte (échelle locale ou régionale en topographie terrestre). Néanmoins, l'échelle *1 :1* est pratiquement et théoriquement impossible à atteindre comme le présente de manière satyrique Umberto Eco dans [199]. Il est donc nécessaire de passer par ces différents niveaux d'échelle pour modéliser l'espace réel afin de mieux connaître son environnement.

Nous constatons donc assez naturellement que dans un environnement marin au niveau microscopique, nous engageons une réflexion à grande échelle cartographique car les sondes sont des mesures ponctuelles très localisées spécifiquement dans l'espace et sans emprise étendue spatiale. Au niveau meso, ces mêmes sondes

sont traitées et structurées sur une emprise spatiale plus importante, notamment dans le cadre de création de MNT, par exemple, afin de générer de l'information. Enfin, à l'échelle macro, l'emprise spatiale est beaucoup plus importante et s'intéresse souvent à l'échelle d'un levé voire à une échelle plus importante, notamment dans le cadre des dalles Téthys (1° par 1°). Les niveaux micro, meso et macro sont donc interconnectés spatialement dans le cadre de ce travail de recherche.

# Conclusion et perspectives pour le Shom

Devant la nécessité de garantir un très haut niveau de qualité des produits générés tout en assurant davantage de missions, les SH ont constamment recherché à se réinventer tout au long de leur histoire pour se placer en permanence à la pointe de la technique, tout en conservant un très haut niveau de qualité. Les évolutions ont été non seulement d'ordres méthodologiques mais également technologiques. Celles-ci se sont même accélérées au cours des dernières années avec l'avènement du numérique et la multiplication des capacités de stockage et de calcul. Dans le cas du Shom, la révolution numérique a commencé en 1983 [3] avec le remplacement des minutes de bathymétrie par des fichiers numériques. La navigation par satellite est, quant à elle, arrivée en 1987 [3] avec l'utilisation du GPS (Global Positionning System), qui est le premier système GNSS ouvert au public, ce qui a permis de grandement faciliter le positionnement des navires. C'est en 1999 que le Shom, par le biais de l'association PRIMAR, diffuse les premières cartes marines électroniques - ENC en anglais. Celles-ci sont diffusées comme produits de navigation officiels et validées par l'OHI et l'Organisation Maritime Internationale (OMI).

Depuis quelques années, les SH et l'OHI s'intéressent de plus en plus à la SDB, présentée dans la section 1.1. Cette technique passive d'acquisition de donnée bathymétrique permet de couvrir de très grandes surfaces avec un coût plus faible que la mobilisation classique lors de campagnes hydro-océanographiques à la mer.

C'est dans ce contexte d'innovation perpétuelle que la DGA finance des projets de recherches et de développements dans le cadre des PTDs, afin de soutenir le Shom dans ses innovations technologiques et le prototypage de nouvelles solutions techniques. Les présents travaux de recherche s'inscrivent dans la préparation du PTD AUTOBATH (traitement AUTOmatique de la BATHymétrie) qui s'intéresse au traitement de la donnée bathymétrique multi-capteurs via des méthodes ML. En ce sens, les perspectives issues de la thèse sont nombreuses.

Comme présenté dans la section 5, les enjeux associés à ces transformations

numériques sont nombreux et de plus en plus de questions autour du parcours utilisateur et la confiance apportée dans les méthodes issues de l'IA se posent. Il est donc important de mener de front les innovations techniques mais également les innovations sociétales qui sont sous-jacentes à la transformation du métier d'hydrographe. L'accompagnement des nouveaux métiers en émergence et les parcours de formation associés se construisent en parallèle des développements scientifiques réalisés pour assurer une conduite de changement la plus fluide possible.

**Soutien aux hydrographes**

Comme présenté tout au long de ce manuscrit, ce travail s'inscrit dans un contexte très opérationnel qui est celui d'un SH national dont le volume des missions est croissant malgré des ressources humaines disponibles constantes voire déclinantes dans certains métiers (dans l'acquisition et le traitement des données bathymétriques par exemple). De fait, il est crucial que l'expertise des hydrographes soit utilisée sur les tâches ayant le plus de valeur pour les SH. Ainsi, l'objectif principal de ce travail de recherche est d'étudier et d'implémenter des outils numériques et informatiques modernes pour soutenir au quotidien les opérateurs du Shom tout en respectant un paradigme fondamental : conserver l'humain au centre du processus décisionnel.

# Perspectives de recherches et développement

Les applications développées dans le cadre de cette thèse ne sont pas toutes au même stade de maturité. Ainsi, de nombreuses évolutions restent à prévoir pour les prochaines années.

## Planifications à court terme

Sur le court terme, il est important de poursuivre les efforts de développement sur la chaîne de traitement LiDAR bathymétrique afin de préparer une opérationnalisation du processus et, dans le même temps, capitaliser sur les expérimentations

afin d'adapter la méthodologie associée à ce nouveau type de traitement. Des actions de formations doivent également être entreprises afin de prendre en main et présenter la méthode à l'ensemble du département AL du Shom (à l'été 2022, la moitié du département a pu tester l'outil et proposer des améliorations).

## Prévisions à moyen terme

Le présent travail de recherche ne s'est pas intéressé aux infrastructures (comme les BDD ou les moyens de calculs) à mettre en place pour la généralisation des méthodes proposées dans ce manuscrit, que ce soit sur les capacités de stockage ou sur les capacités de calcul. Un travail d'étude important reste donc à effectuer sur l'optimisation de la gestion de la donnée massive LiDAR et SMF, avec une réflexion particulière à mener sur l'intégration de la donnée dans des bases de traitement spécifiques afin d'améliorer la vélocité d'accès et de manipulation de la donnée ainsi que la construction de base de donnée d'apprentissage utile pour les mécanismes d'apprentissage. De plus, après une phase de développement assez intense en 2021 et 2022, il faut à présent procéder à une phase de documentation et de généralisation d'une partie du code afin de garantir sa maintenabilité, sa clarté, sa robustesse pour toute reprise future du projet dans un autre environnement de développement et sa réutilisabilité dans le cadre d'autres projets.

Enfin, et comme présenté dans la section 1.3.2, d'autres techniques d'apprentissage supervisé seraient à étudier comme l'apprentissage par renforcement. En effet, les opérateurs du Shom réalisent aujourd'hui encore beaucoup d'actions manuelles dans leur planification quotidienne des levés. Des apprentissages par renforcement permettraient d'automatiser ces planifications en ajustant leur politique d'apprentissage par l'adaptation aux conditions environnementales durant l'acquisition, voir 1.3.2. En couplant l'apprentissage par renforcement avec les techniques d'apprentissage supervisé plus classiques, un pré-traitement rapide serait réalisé et la politique de l'apprentissage adaptée en fonction des résultat du traitement, ce qui permettrait de revenir sur une zone comportant un trou de donnée, à cause d'une crevasse non-présente sur les CM par exemple, ou insonifier davantage une potentielle épave.

## Prospections à long terme

Sur le long terme, une veille des méthodes de gestion des données doit être régulièrement assurée afin de n'en utiliser que les plus performantes. Il est primordial de garantir pour l'hydrographe une compréhension du processus global mis en branle lors du traitement des données issues des campagnes hydro-océanographiques. Enfin, la mise à disposition de ces données constitueraient une des mailles essentielles dans le travail de préservation des écosystèmes à travers la maîtrise des risques littoraux comme dans le cadre de la prévention des risques de vagues submersion (en lien avec Météo France) [3].

Finalement, je considère la conclusion de ce manuscrit de thèse non comme une fin mais comme l'amorce de progrès dans le domaine du traitement des données bathymétriques via l'apprentissage machine et de la qualification des campagnes hydrographiques. Cette thèse a été l'opportunité de collaborer avec de nombreuses personnes mais également d'initier des projets qui s'inscrivent sur la durée avec de grands instituts de recherche français. Un travail important demeure pour améliorer le quotidien des opérateurs du Shom et leur proposer des outils toujours plus performants. Ainsi, j'espère que les résultats issus de ce travail de recherche y contribueront et que de nombreux développements et innovations seront produits dans le futur.

# BIBLIOGRAPHIE


[1] R. Bellman, *Adaptive Control Processes : A Guided Tour*. Princeton University Press, 1961, isbn : 978-0-691-07901-1. adresse : http://www.jstor.org/stable/j.ctt183ph6v (visité le 08/08/2022).

[2] J. Bourgoin, « The French State and Nautical Cartography », *The International Hydrographic Review*, t. 65, *2*, mai 2015. adresse : https://journals.lib.unb.ca/index.php/ihr/article/view/23360.

[3] O. Chapuis, P. Souquiere et G. Bessero, *300 ans de cartes marines autour du monde*, française, Gallimard Loisirs. Gallimard, 2011, isbn : 978-2-7424-6199-8.

[4] G. Bessero et H. Richard, *300 ans d'hydrographie française*, française, Shom. Locus Solus, 2020, isbn : 978-2-36833-285-6.

[5] C. Beautemps-Beaupré, R. Copeland, A. Dalrymple et M. Flinders, *An Introduction to the Practice of Nautical Surveying, and the Construction of Sea-charts : Illustrated by Thirty-four Plates*. R.H. Laurie, 1823. adresse : https://books.google.fr/books?id=q14OAAAAYAAJ.

[6] International Maritime Organization (IMO), *International Convention for the Safety of Life At Sea*, English, nov. 1974. adresse : https://www.refworld.org/docid/46920bf32.html.

[7] X. Lurton, *An Introduction to Underwater Acoustics : Principles and Applications*, sér. Geophysical Sciences Series. Springer, 2002, isbn : 978-3-540-42967-8. adresse : https://books.google.fr/books?id=VTNRh3pyCyMC.

[8] International Hydrographic Organization (IHO), « S-44 Standards for Hydrographic Surveys », Monaco, rapp. tech. Ed. 6.0.0, sept. 2020.

[9] B. Calder et S. Smith, « A time/effort comparison of automatic and manual bathymetric processing in real-time mode », *International Hydrographic Review*, t. 5, p. 10-23, 2004.

[10] L. Mayer, M. Jakobsson, G. Allen et al., « The Nippon Foundation—GEBCO Seabed 2030 Project : The Quest to See the World's Oceans Completely Mapped by 2030 », *Geosciences*, t. 8, p. 63, fév. 2018. doi : 10.3390/geosciences8020063.

[11] F. Parly, *Stratégie ministérielle de maîtrise des fonds marins*, française, Paris, France, 2022. adresse : https://www.defense.gouv.fr/sites/default/files/ministere-armees/20220211_GT%20MAITRISE%20FONDS%20MARINS_dossier%20de%20presse.pdf.

[12] J. Le Deunf, N. Debèse, T. Schmitt et R. Billot, « A Review of Data Cleaning Approaches in a Hydrographic Framework with a Focus on Bathymetric Multibeam Echosounder Datasets », *Geosciences*, t. 10, *7*, 2020, issn : 2076-3263. doi : 10.3390/geosciences10070254. adresse : https://www.mdpi.com/2076-3263/10/7/254.

[13] M. Malik, A. C. G. Schimel, G. Masetti et al., « Results from the First Phase of the Seafloor Backscatter Processing Software Inter-Comparison Project », *Geosciences*, t. 9, *12*, 2019, issn : 2076-3263. doi : 10.3390/geosciences9120516. adresse : https://www.mdpi.com/2076-3263/9/12/516.

[14] J. Le Deunf, R. Mishra, Y. Pastol, R. Billot et S. Oudot, « Seabed prediction from airborne topo-bathymetric lidar point cloud using machine learning approaches », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-9. doi : 10.23919/OCEANS44145.2021.9706113.

[15] J. Le Deunf, A. Khannoussi, L. Lecornu, P. Meyer et J. Puentes, « Automatic Data Quality Assessment of Hydrographic Surveys Taking Into Account Experts' Preferences », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-10. doi : 10.23919/OCEANS44145.2021.9705772.

[16] J. Le Deunf, R. Jarno, Y. Keramoal et al., « Téthys : automating a data workflow compiling over 300 years of bathymetric information », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-6. doi : 10.23919/OCEANS44145.2021.9705727.

[17] M. Michel, J. L. Deunf, N. Debese, L. Bazinet et L. Dejoie, « Multibeam outlier detection by clustering and topological persistence approach, ToMATo algorithm », in *OCEANS 2021 : San Diego – Porto*, 2021, p. 1-8. doi : 10.23919/OCEANS44145.2021.9705930.

[18] J. Le Deunf, T. Schmitt et Y. Keramoal, *A pragmatical algorithm to compute the convex envelope of bathymetric surveys at variable resolutions*, International Cartographic Conference 2021 (ICC 2021), Florence, 2021.

[19] J. Le Deunf, N. Debèse, T. Schmitt et al., « Outlier detection for Multibeam echo sounder (MBES) data : from past to present », in *OCEANS 2019 - Marseille*, 2019, p. 1-10. doi : 10.1109/OCEANSE.2019.8867321.

[20] J. Le Deunf, A. Ben-ammar, Y. Pastol, R. Billot et S. Oudot, *Cleaning Outliers Bathymetric Lidar Data Using Machine Learning Techniques : Shom's Feedback*, Canadian Hydrographic Conference 2022 (CHC 2022), Ottawa, 2022.

[21] T. Schmitt, J. Le Deunf, M. Fally, R. Jarno et Y. Kearmoal, *Automating the generation of the bathymetric surface reference through expert rules and ETL : the Tethys project*, Canadian Hydrographic Conference 2022 (CHC 2022), Ottawa, 2022.

[22] A. Khannoussi, J. Le Deunf, P. Meyer, L. Lecornu et J. Puentes, *Integrating user preferences in the automatic quality assessment of hydrographic surveys*, 92nd Meeting of EURO Working Group on Multicriteria Decision Aiding "Multi-Criteria Decision Analysis as a transdisciplinary science", Cracow, sept. 2021.

[23] J. Le Deunf, *Traiter la bathymétrie avec l'aide de l'intelligence artificielle*, Seat Tech Week 2020, Virtuelle, oct. 2020.

[24] J. Le Deunf, J. Eddadsi, R. Billot, N. Debèse et T. Schmitt, *Automated detection of bathymetric outliers*, Canadian Hydrographic Conference 2020 (CHC 2020), Québec, 2020.

[25] J. Le Deunf, *Prospects for innovation at Shom*, Vecteur 2019, Rimouski, 2019.

[26] ——, *Traitement lidar bathymétrique utilisant des techniques d'apprentissage automatique pour une prédiction automatique du fond marin : application aux données de la Bretagne*, Journée de l'information scientifique et technique du Shom, Brest, juin 2022.

[27] T. Schmitt, J. Le Deunf, Y. Keramoal et al., *L'automatisation des processus de traitement et l'intelligence artificielle au service de la Bathymétrie*, Journées techniques de l'Afhy 2022, Lyon, 2022.

[28] Y. Pastol, J. Le Deunf, C. Salvaterra et C. Vrignaud, *Levés par lidar bathymétrique aéroporté, dans le cadre du projet Litto3D®, et traitement des données par IA au Shom*, merIGeo 2020, Virtuelle, nov. 2020.

[29] J. Le Deunf, *Détection automatisée de sondes aberrantes, l'approche machine learning pour le traitement de la bathymétrie*, 29eme Journées de la Recherche de l'IGN, Virtuelle, oct. 2020.

[30] J. Le Deunf, N. Debèse, T. Schmitt et B. Romain, *Détection automatisée de sondes aberrantes*, Journée de l'information scientifique et technique du Shom, Brest, mars 2019.

[31] J. Le Deunf, « Rapport du projet TraitLIA : Traitement semi-automatique de nuages de points issus de lidar bathymétrique aéroporté. », Shom, rapp. tech. n°18 DOPS / STM / BATHY / NP, avr. 2022.

[32] ——, « Utilisation de l'algorithme CUBE au Shom », Shom, Rapport technique n°27 DOPS / MIP / BATHY / NP, mai 2019.

[33] J. Le Deunf, M. Legris et J. McManus, « Automatic Calibration for MBES Offsets », *Hydro International*, t. 25, *3*, p. 40, 42, oct. 2020.

[34] R. Legouge, G. André, J. Le Deunf, A. Missault et S. Branchu, « Utilisation d'infrastructures géodésiques mondiales pour la réalisation nationale », *Revue XYZ*, t. 158, mars 2019, Publisher : Association Française de Topographie. adresse : https://hal.archives-ouvertes.fr/hal-02317630.

[35] Y. Xie, W. Shen, J. Han et X. Deng, « Determination of the height of Mount Everest using the shallow layer method », *Geodesy and Geodynamics*, t. 12, *4*, p. 258-265, 2021, issn : 1674-9847. doi : https://doi.org/

10.1016/j.geog.2021.04.002. adresse : https://www.sciencedirect.com/science/article/pii/S1674984721000343.

[36] J. V. Gardner, A. A. Armstrong, B. R. Calder et J. Beaudoin, « So, How Deep Is the Mariana Trench ? », *Marine Geodesy*, t. 37, *1*, p. 1-13, 2014, Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/01490419.2013.837849. doi : 10.1080/01490419.2013.837849. adresse : https://doi.org/10.1080/01490419.2013.837849.

[37] Shom, *Paré pour la sonde*, française, Documentaire, oct. 2020. adresse : https://www.youtube.com/watch?v=8UMIA7CP7bE.

[38] International Hydrographic Organization (IHO), « C-13 Manual on Hydrography », Monaco, rapp. tech. Ed. 1.0.0, fév. 2011.

[39] N. Debese, *Bathymétrie - Sondeurs, Traitement des Données Modèles Numériques de Terrain - Cours Exercices Corrigés*. Ellipses, juin 2013. adresse : https://hal.archives-ouvertes.fr/hal-00843130.

[40] International Hydrographic Organization (IHO), « S-57 Transfer standard for digital hydrographic data », IHO, rapp. tech., 2000.

[41] Y. Bidel, « Gravimètre à atomes froids embarquable », Habilitation à diriger des recherches, UNIVERSITE PARIS-SACLAY, oct. 2020. adresse : https://hal.archives-ouvertes.fr/tel-03111583.

[42] J.-F. Oehler et M.-F. Lequentrec-Lalancette, « The contribution of marine magnetics in the Gulf of Saint-Malo (Brittany, France) to the understanding of the geology of the North Armorican Cadomian belt », *Comptes Rendus Géoscience*, t. 6420, *1*, p. 1-58, 2018, issn : 1631-0713. doi : http://dx.doi.org/10.1016/j.crte.2018.10.002. adresse : http://www.sciencedirect.com/science/article/pii/S1631-0713(18)30134-2.

[43] Shom, *Profil au top*, française, Documentaire, jan. 2018. adresse : https://www.youtube.com/watch?v=omwlyh_6TqU.

[44] W. Philpot, G. Guenther, J. Lilycrop et al., *BBII Airborne Laser Hydrography*. oct. 2019. doi : 10.7298/4sb5-8434.

[45] D. R. Lyzenga, « Passive remote sensing techniques for mapping water depth and bottom features », *Appl. Opt.*, t. 17, *3*, p. 379-383, fév. 1978, Publisher : Optica Publishing Group. doi : 10.1364/AO.17.000379. adresse : http://opg.optica.org/ao/abstract.cfm?URI=ao-17-3-379.

[46] B. Simon, *La marée océanique côtière*, sér. Collection "Synthèses" vol. 1. Institut océanographique, 2007, isbn : 978-2-903581-32-9. adresse : https://books.google.fr/books?id=FKBJPQAACAAJ.

[47] C.-T. Chen et F. J. Millero, « Speed of sound in seawater at high pressures », *The Journal of the Acoustical Society of America*, t. 62, *5*, p. 1129-1135, 1977, _eprint : https ://doi.org/10.1121/1.381646. doi : 10.1121/1.381646. adresse : https://doi.org/10.1121/1.381646.

[48] G. L. Pickard et W. J. Emery, « CHAPTER 4 - Typical Distributions of Water Characteristics in the Oceans », in *Descriptive Physical Oceanography (Fifth Edition)*, G. L. Pickard et W. J. Emery, éd., Fifth Edition, Amsterdam : Pergamon, 1990, p. 34-63, isbn : 978-0-08-037952-4. doi : https://doi.org/10.1016/B978-0-08-037952-4.50009-0. adresse : https://www.sciencedirect.com/science/article/pii/B9780080379524500090.

[49] R. C. Selley, *Applied sedimentology*. Elsevier, 2000.

[50] R. Keyetieu Nlowe, « Calibration of Multi-Beam Echo Sounder systems by inverse methods », Anglais, thèse de doct., ESNTA Bretagne, Brest, France, nov. 2018.

[51] N. Seube, *Geo-location equations and a priori total propagated error. Hydrography course module (UV 3.6) of ENSTA Bretagne.* Anglais, 2013.

[52] M. Legris, *Systèmes sonars de bathymétrie et d'imagerie. Cours d'hydrographie de l'ENSTA Bretagne.* française, 2013.

[53] R. Urick, *Principles of Underwater Sound*. McGraw-Hill, 1983, isbn : 978-0-07-066087-8. adresse : https://books.google.fr/books?id=hfxQAAAAMAAJ.

[54] M. Lamouret, « Traitements automatisés des données acoustiques issues de sondeurs multifaisceaux pour la cartographie des fonds marins. », française, thèse de doct., Université de Toulon, Ecole doctorale 548 Mer et Sciences, mars 2022.

[55] A. Euzen, C. Jeandel et R. Mosseri, *L'eau à découvert*. CNRS éditions, 2015. adresse : https://hal-enpc.archives-ouvertes.fr/hal-01234616.

[56] A. C. Müller et S. Guido, *Introduction to Machine Learning with Python*. O'Reilly, 2016. adresse : https://learning.oreilly.com/library/view/introduction-to-machine/9781449369880/.

[57] T. Hastie, R. Tibshirani et J. Friedman, *The Elements of Statistical Learning*, sér. Springer Series in Statistics. New York, NY, USA : Springer New York Inc., 2001.

[58] Y. Sola, G. L. Chenadec et B. Clement, « Simultaneous Control and Guidance of an AUV Based on Soft Actor–Critic », *Sensors*, t. 22, *16*, p. 6072, août 2022, Publisher : MDPI. doi : 10.3390/s22166072. adresse : https://hal-ensta-bretagne.archives-ouvertes.fr/hal-03752054.

[59] S. Lloyd, « Least squares quantization in PCM », *IEEE Transactions on Information Theory*, t. 28, *2*, p. 129-137, 1982. doi : 10.1109/TIT.1982.1056489.

[60] M. Ester, H.-P. Kriegel, J. Sander et X. Xu, « A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise », in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, sér. KDD'96, event-place : Portland, Oregon, AAAI Press, 1996, p. 226-231.

[61] F. Chazal, L. Guibas, S. Oudot et P. Skraba, « Persistence-Based Clustering in Riemannian Manifolds », *Journal of the ACM*, t. 60, juin 2011. doi : 10.1145/1998196.1998212.

[62] K. Pearson, « LIII. On lines and planes of closest fit to systems of points in space », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, t. 2, *11*, p. 559-572, 1901, Publisher : Taylor & Francis. doi : 10.1080/14786440109462720.

[63] C. M. Bishop et N. M. Nasrabadi, *Pattern recognition and machine learning*, 4. Springer, 2006, t. 4.

[64] T. K. Ho, « Random decision forests », in *Proceedings of 3rd international conference on document analysis and recognition*, t. 1, IEEE, 1995, p. 278-282.

[65] C. Cortes et V. Vapnik, « Support-vector networks », *Machine learning*, t. 20, *3*, p. 273-297, 1995, Publisher : Springer.

[66] H. K. Jabbar et R. Z. Khan, « Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study) », 2014.

[67] A. J.-M. Lurton Xavier Ladroit Yoann, « A quality estimator of acoustic sounding detection », *International Hydrographic Review*, *4*, p. 35-45, 2012. adresse : https://archimer.ifremer.fr/doc/00115/22601/.

[68] E. Kammerer, D. Charlot, S. Guillaudeux et P. Michaux, « Comparative study of shallow water multibeam imagery for cleaning bathymetry sounding errors », t. 4, fév. 2001, 2124-2128 vol.4, isbn : 0-933957-28-9. doi : 10.1109/OCEANS.2001.968327.

[69] B. Calder et L. Mayer, « Automatic processing of high-rate, high-density multibeam echosounder data », *Geochemistry, Geophysics, Geosystems*, t. 4, juin 2003. doi : 10.1029/2002GC000486.

[70] R. Hare, « Depth and Position Error Budgets for Mulitbeam Echosounding », *The International Hydrographic Review*, t. 72, *2*, mai 2015. adresse : https://journals.lib.unb.ca/index.php/ihr/article/view/23178.

[71] R. Ladner, P. Elmore, A. Perkins, B. Bourgeois et W. Avera, « Automated cleaning and uncertainty attribution of archival bathymetry based on a priori knowledge », *Marine Geophysical Research*, t. 38, p. 1-11, sept. 2017. doi : 10.1007/s11001-017-9304-9.

[72] K. Lowell et B. Calder, « Machine Learning Strategies for Enhancing Bathymetry Extraction from Imbalanced Lidar Point Clouds », in *OCEANS 2019 MTS/IEEE SEATTLE*, ISSN : 0197-7385, oct. 2019, p. 1-8. doi : 10.23919/OCEANS40490.2019.8962400.

[73] P. De Bièvre, « The 2012 International Vocabulary of Metrology : "VIM" », *Accreditation and Quality Assurance*, t. 17, *2*, p. 231-232, avr. 2012, ISSN : 1432-0517. DOI : 10.1007/s00769-012-0885-3. adresse : https://doi.org/10.1007/s00769-012-0885-3.

[74] J. E. Hughes Clarke, « The Impact of Acoustic Imaging Geometry on the Fidelity of Seabed Bathymetric Models », *Geosciences*, t. 8, *4*, 2018, ISSN : 2076-3263. DOI : 10.3390/geosciences8040109. adresse : https://www.mdpi.com/2076-3263/8/4/109.

[75] D. M. Hawkins, *Identification of outliers*, sér. Monographs on applied probability and statistics. London [u.a.] : Chapman et Hall, 1980, ISBN : 0-412-21900-X. adresse : http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+02435757X&sourceid=fbw_bibsonomy.

[76] L. Davies et U. Gather, « The Identification of Multiple Outliers », *Journal of the American Statistical Association*, t. 88, *423*, p. 782-792, 1993, Publisher : Taylor & Francis _eprint : https ://www.tandfonline.com/doi/pdf/10.1080/01621459.199 DOI : 10.1080/01621459.1993.10476339. adresse : https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476339.

[77] F. Hampel, E. Ronchetti, P. J. Rousseeuw et W. A. Stahel, « Robust statistics : the approach based on influence functions », 1986.

[78] V. Planchon, « Traitement des valeurs aberrantes : concepts actuels et tendances générales », *Biotechnologie, Agronomie, Société et Environnement*, t. 9, jan. 2005.

[79] C. C. Aggarwal, *Outlier Analysis*, en, 2e éd. Springer International Publishing, 2017, ISBN : 978-3-319-47577-6. adresse : https://www.springer.com/gp/book/9783319475776 (visité le 23/10/2019).

[80] V. Kotu et B. Deshpande, « Chapter 13 - Anomaly Detection », in *Data Science (Second Edition)*, V. Kotu et B. Deshpande, éd., Morgan Kaufmann, jan. 2019, p. 447-465, ISBN : 978-0-12-814761-0. DOI : 10.1016/B978-0-12-814761-0.00013-7. adresse : http://www.sciencedirect.com/science/article/pii/B9780128147610000137.

[81] G. C. Guenther, « Improved Depth Selection in the Bathymetric Swath Survey System (BS3) Combined Offline Processing (COP) Program », NOAA technical report OTES ; 10, OTES (Ocean Technology and Engineering Services), éd., 1982. adresse : https://repository.library.noaa.gov/view/noaa/23480.

[82] N. Debese et H. Bisquay, « Automatic Detection of Punctual Errors in Multibeam Data Using a Robust Estimator », *International Hydrographic Review*, t. 76, 1999.

[83] B. R. Calder et G. Rice, « Design and Implementation of an Extensible Variable Resolution Bathymetric Estimator », 2011.

[84] P. Weatherall, K. M. Marks, M. Jakobsson et al., « A new digital bathymetric model of the world's oceans », *Earth and Space Science*, t. 2, *8*, p. 331-345, 2015. doi : https://doi.org/10.1002/2015EA000107.

[85] V. Chandola, « Outlier Detection : A Survey », 2007.

[86] R. G. Burke, S. R. Forbes et K. White, « Processing 'Large' Data Sets From 100% Bottom Coverage 'Shallow' Water Sweep Surveys A New Challenge for the Canadian Hydrographic Service », *International Hydrographic Review*, t. 65, 1988.

[87] N. Debese, R. Moitié et N. Seube, « Multibeam echosounder data cleaning through a hierarchic adaptive and robust local surfacing », *Comput. Geosci.*, t. 46, p. 330-339, 2012.

[88] S. Wang, P. Zhou, Z. Wu, J. Li et Y. Wei, « Detection and Elimination of Bathymetric Outliers in Multibeam Echosounder System Based on Robust Multi - quadric Method and Median Parameter Model », *Journal of Engineering Science and Technology Review*, t. 11, p. 70-78, 2018.

[89] Z. Du, D. Wells et L. Mayer, « An approach to automatic detection of outliers in multibeam echo sounding data », *Oceanographic Literature Review*, t. 7, *43*, p. 737, 1996.

[90] G. Canepa, O. Bergem et N. G. Pace, « A new algorithm for automatic processing of bathymetric data », *IEEE Journal of Oceanic Engineering*, t. 28, p. 62-77, 2003.

[91] X. Huang, C. Huang, G. Zhai et al., « Data Processing Method of Multibeam Bathymetry Based on Sparse Weighted LS-SVM Machine Algorithm », *IEEE Journal of Oceanic Engineering*, t. 45, p. 1538-1551, 2020.

[92] H. Bisquay, X. Freulon, C. d. Fouquet et C. Lajaunie, « Multibeam data cleaning for hydrography using geostatistics », *IEEE Oceanic Engineering Society. OCEANS'98. Conference Proceedings (Cat. No.98CH36259)*, t. 2, 1135-1143 vol.2, 1998.

[93] T. Hou, L. Huff et L. A. Mayer, « Automatic Detection of Outliers in Multibeam Echo Sounding Data », 1996.

[94] F. Yang, J. Li, F.-y. Chu et Z. Wu, « Automatic Detecting Outliers in Multibeam Sonar Based on Density of Points », *OCEANS 2007 - Europe*, p. 1-4, 2007.

[95] J. F. Arnold et S. Shaw, « A surface weaving approach to multibeam depth estimation », *Proceedings of OCEANS '93*, II/95-II/99 vol.2, 1993.

[96] J. Eeg, « On the Identification of Spikes in Soundings », *International Hydrographic Review*, t. 72, 2015.

[97] Í. O. Ferreira, A. d. P. d. Santos, J. C. d. Oliveira, N. d. G. Medeiros et P. C. Emiliano, « Robust methodology for detection of spikes in multibeam echo sounder data », *Boletim de Ciências Geodésicas*, 2019.

[98] J. T. Bjørke et S. Nilsen, « Fast trend extraction and identification of spikes in bathymetric data », *Comput. Geosci.*, t. 35, p. 1061-1071, 2009.

[99] M. Mann, P. Agathoklis et A. Antoniou, « Automatic outlier detection in multibeam data using median filtering », *2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE Cat. No.01CH37233)*, t. 2, 690-693 vol.2, 2001.

[100] M. Li, Y. C. Liu, Z. Lv et J. Bao, « Order Statistics Filtering for Detecting Outliers in Depth Data along a Sounding Line », *Hydrographic Surveying and Charting*, p. 258-262, 2008.

[101] D. Lu, H. S. Li, Y. Wei et T. Zhou, « Automatic outlier detection in multibeam bathymetric data using robust LTS estimation », *2010 3rd International Congress on Image and Signal Processing*, t. 9, p. 4032-4036, 2010.

[102] H. Motao, Z. Guojun, W. Rui, O. Yongzhong et G. Zheng, « Robust Method for the Detection of Abnormal Data in Hydrography », *International Hydrographic Review*, t. 76, 2015.

[103] P. Bottelier, C. Briese, N. Hennis, R. C. Lindenbergh et N. Pfeifer, « Distinguishing features from outliers in automatic Kriging-based filtering of MBES data : a comparative study », 2005.

[104] L. Arge, K. G. Larsen, T. Mølhave et F. v. Walderveen, « Cleaning massive sonar point clouds », in *GIS '10*, 2010.

[105] B. R. Calder, « Multibeam Swath Consistency Detection and Downhill Filtering from Alaska to Hawaii », 2005.

[106] D. A. Stephens, A. D. Smith, T. Redfern, A. Talbot, A. Lessnoff et K. Dempsey, « Using three dimensional convolutional neural networks for denoising echosounder point cloud data », 2020.

[107] T. Kogut et A. Słowik, « Classification of Airborne Laser Bathymetry Data Using Artificial Neural Networks », *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, t. 14, p. 1959-1966, 2021.

[108] Y. Li, L. Ma, Z. Zhong et al., « Deep Learning for LiDAR Point Clouds in Autonomous Driving : A Review », *IEEE Transactions on Neural Networks and Learning Systems*, t. 32, p. 3412-3432, 2021.

[109] V. J. Hodge et J. Austin, « A Survey of Outlier Detection Methodologies », *Artificial Intelligence Review*, t. 22, p. 85-126, 2004.

[110] C. B. Lirakis et K. P. Bongiovanni, « Automated multibeam data cleaning and target detection », *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No.00CH37158)*, t. 1, 719-723 vol.1, 2000.

[111] L. Sedaghat, J. Hersey et M. P. McGuire, « Detecting spatio-temporal outliers in crowdsourced bathymetry data », in *GEOCROWD '13*, 2013.

[112] D. R. Herlihy, T. N. Stepka et T. D. Rulon, « Filtering Erroneous Soundings from Multibeam Survey Data », *International Hydrographic Review*, t. 69, 1992.

[113] J. F. Bourillet, C. Edy, F. A. Rambert, C. Satra et B. Loubrieu, « Swath mapping system processing : Bathymetry and cartography », *Marine Geophysical Researches*, t. 18, p. 487-506, 1996.

[114] M. M. Breunig, H.-P. Kriegel, R. T. Ng et J. Sander, « LOF : identifying density-based local outliers », in *SIGMOD '00*, 2000.

[115] C. Xie, P. Chen, D. Pan, C. Zhong et Z. Zhang, « Improved Filtering of ICESat-2 Lidar Data for Nearshore Bathymetry Estimation Using Sentinel-2 Imagery », *Remote Sensing*, t. 13, *21*, 2021, issn : 2072-4292. doi : 10.3390/rs13214303. adresse : https://www.mdpi.com/2072-4292/13/21/4303.

[116] D. B. Rorabacher, « Statistical treatment for rejection of deviant values : critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level », *Analytical Chemistry*, t. 63, p. 139-146, 1991.

[117] A. Blake et A. Zisserman, *Visual Reconstruction*. jan. 1987, isbn : 978-0-262-25576-9. doi : 10.7551/mitpress/7132.001.0001.

[118] P. Chauvet, *Aide-mémoire de géostatistique linéaire*, sér. Les Cours - École des mines de Paris. les Presses de l'École des Mines, 1999, isbn : 978-2-911762-16-1. adresse : https://books.google.fr/books?id=EmTUjxwGCk0C.

[119] D. Q. F. d. Menezes, D. M. Prata, A. R. Secchi et J. C. Pinto, « A review on robust M-estimators for regression analysis », *Comput. Chem. Eng.*, t. 147, p. 107 254, 2021.

[120] M.-H. Rezvani, A. G. Sabbagh et A. A. Ardalan, « Robust Automatic Reduction of Multibeam Bathymetric Data Based on M-estimators », *Marine Geodesy*, t. 38, p. 327-344, 2015.

[121] P. J. Rousseeuw, « Least Median of Squares Regression », *Journal of the American Statistical Association*, t. 79, p. 871-880, 1984.

[122] C. R. Qi, H. Su, K. Mo et L. J. Guibas, « PointNet : Deep Learning on Point Sets for 3D Classification and Segmentation », *CoRR*, t. abs/1612.00593, 2016, arXiv : 1612.00593. adresse : http://arxiv.org/abs/1612.00593.

[123] N. Quadros et J. Keysers, « Emerging Trends in Bathymetric Lidar Technology », Anglais, *Hydro International*, 2019.

[124] X.-J. Mao, C. Shen et Y. Yang, « Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections », *ArXiv*, t. abs/1606.08921, 2016.

[125] O. Ronneberger, P. Fischer et T. Brox, « U-Net : Convolutional Networks for Biomedical Image Segmentation », *ArXiv*, t. abs/1505.04597, 2015.

[126] M. E. Vásquez, « Tuning the CARIS implementation of CUBE for Patagonian Waters. », 2007.

[127] N. Debese, « Multibeam Echosounder Data Cleaning Through an Adaptive Surface-based Approach », jan. 2007.

[128] GHOA/HYDRO, « PS2022-018 Processus de traitement de la bathymétrie », française, rapp. tech., juill. 2022.

[129] C. Leys, C. Ley, O. Klein, P. Bernard et L. Licata, « Detecting outliers : Do not use standard deviation around the mean, use absolute deviation around the median », *Journal of Experimental Social Psychology*, t. 49, *4*, p. 764-766, 2013, issn : 0022-1031. doi : https://doi.org/10.1016/j.jesp.2013.03.013. adresse : https://www.sciencedirect.com/science/article/pii/S0022103113000668.

[130] F. Pedregosa, G. Varoquaux, A. Gramfort et al., « Scikit-learn : Machine Learning in Python », *J. Mach. Learn. Res.*, t. 12, p. 2825-2830, 2011.

[131] D. W. Hosmer et S. Lemeshow, « Applied Logistic Regression », 1989.

[132] T. K. Ho, « The Random Subspace Method for Constructing Decision Forests », *IEEE Trans. Pattern Anal. Mach. Intell.*, t. 20, p. 832-844, 1998.

[133] C. Drummond et R. C. Holte, « C4.5, Class Imbalance, and Cost Sensitivity : Why Under-Sampling beats Over-Sampling », 2003.

[134] J. H. Friedman, « Greedy function approximation : A gradient boosting machine. », *Annals of Statistics*, t. 29, p. 1189-1232, 2001.

[135] D. M. W. Powers, « Evaluation : from precision, recall and F-measure to ROC, informedness, markedness and correlation », *ArXiv*, t. abs/2010.16061, 2020.

[136] J. Gan et Y. Tao, « DBSCAN Revisited : Mis-Claim, Un-Fixability, and Approximation », *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015.

[137] E. Silvia, H. Butler et K. Damkjet, *LAS Specification*, 2019. adresse : https://github.com/ASPRSorg/LAS.

[138] K. He, J. Sun et X. Tang, « Guided Image Filtering », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 35, p. 1397-1409, 2013.

[139] N. Audebert, B. Le Saux et S. Lefevre, « Deep Learning for Classification of Hyperspectral Data : A Comparative Review », *IEEE Geoscience and Remote Sensing Magazine*, t. 7, *2*, p. 159-173, juin 2019, Publisher : Institute of Electrical and Electronics Engineers (IEEE), issn : 2373-7468. doi : 10.1109/mgrs.2019.2912563. adresse : http://dx.doi.org/10.1109/MGRS.2019.2912563.

[140] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola et V. N. Vapnik, « Support Vector Regression Machines », in *NIPS*, 1996.

[141] L. Breiman, « Random Forests », *Machine Learning*, t. 45, p. 5-32, 2004.

[142] S. Haykin, « Neural Networks : A Comprehensive Foundation », 1998.

[143] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter et L. Kagal, « Explaining Explanations : An Overview of Interpretability of Machine Learning », *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, p. 80-89, 2018.

[144] W. S. McCulloch et W. Pitts, « A logical calculus of the ideas immanent in nervous activity », *The bulletin of mathematical biophysics*, t. 5, *4*, p. 115-133, 1943, Publisher : Springer.

[145] C. J. Willmott et K. Matsuura, « Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance », *Climate Research*, t. 30, p. 79-82, 2005.

[146] D. P. Kingma et J. Ba, « Adam : A Method for Stochastic Optimization », *CoRR*, t. abs/1412.6980, 2015.

[147] H. E. Robbins, « A Stochastic Approximation Method », *Annals of Mathematical Statistics*, t. 22, p. 400-407, sept. 1951.

[148] Shom - IGN, *Litto3D Corse*, Type : dataset Licence ouverte / Open Data, 2020. doi : 10.17183/L3D_MAR_CORSE_2017_2018.

[149] D. P. Doane, « Aesthetic Frequency Classifications », *The American Statistician*, t. 30, p. 181-183, 1976.

[150] G. Mandlburger, M. Kölle, H. Nübel et U. Soergel, « BathyNet : A Deep Neural Network for Water Depth Mapping from Multispectral Aerial Images », *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, t. 89, p. 71-89, 2021.

[151] Y. Pastol, « Use of Airborne LIDAR Bathymetry for Coastal Hydrographic Surveying : The French Experience », *Journal of Coastal Research*, t. 62, p. 6-18, avr. 2011. doi : 10.2112/SI_62_2.

[152] Shom - IGN, *Litto3D Bretagne*, Type : dataset Licence ouverte / Open Data, 2022. doi : 10.17183/L3D_MAR_BZH_2018_2021.

[153] F. Chazal, L. J. Guibas, S. Y. Oudot et P. Skraba, « Persistence-Based Clustering in Riemannian Manifolds », *Journal of the ACM (JACM)*, t. 60, *6*, p. 38, nov. 2013, Publisher : Association for Computing Machinery. adresse : https://hal.archives-ouvertes.fr/hal-00923563.

[154] C. Maria, J.-D. Boissonnat, M. Glisse et M. Yvinec, « The Gudhi Library : Simplicial Complexes and Persistent Homology », in *ICMS*, 2014.

[155] International Hydrographic Organization (IHO), « S-102 Bathymetric Surface Product Specification », Monaco, rapp. tech. 2.1.0, mai 2022.

[156] EMODnet Bathymetry Consortium, *EMODnet Digital Bathymetry (DTM 2020)*, Anglais, europe, déc. 2020. adresse : https://doi.org/10.12770/bb6a87dd-e579-4036-abe1-e649cea9881a.

[157] C. Bongiovanni, A. Armstrong, B. Calder et T. Lippmann, « Identifying future hydrographic survey priorities : A quantitative uncertainty based approach », *International Hydrographic review*, t. 25, 2021.

[158] A. Mallégol, P. Meyer et S. Loyer, « A hydrographic risk assessment methodology integrating the user's preferences : application to a use-case in the English Channel », in *92nd meeting of the EURO Working Group on Multi-Criteria Decision Aiding (EWG-MCDA 92)*, Krakow, Poland, sept.

2021. adresse : https://hal-imt-atlantique.archives-ouvertes.fr/hal-03376337.

[159] J. Riding et I. Weeb, « Risk methodology : South West Pacifig regional hydrography programme », Anglais, Land Information New Zealand (LINZ), Wellington, New Zealand, Risk methodology : South West Pacifig regional hydrography programme Marico Marine Report n°12NZ246, fév. 2013, p. 55.

[160] J. Riding et A. Rawson, « LINZ Hydrography Risk Assessment Methodology Update », Anglais, Land Information New Zealand (LINZ), Wellington, New Zealand, rapp. tech. Marico Marine Report n°15NZ322, août 2015, p. 76.

[161] L. Paull, S. Saeedi, M. Seto et H. Li, « AUV Navigation and Localization : A Review », *IEEE Journal of Oceanic Engineering*, t. 39, *1*, p. 131-149, 2014. doi : 10.1109/JOE.2013.2278891.

[162] G. Donovan, « Position Error Correction for an Autonomous Underwater Vehicle Inertial Navigation System (INS) Using a Particle Filter », *Oceanic Engineering, IEEE Journal of*, t. 37, p. 431-445, juill. 2012. doi : 10.1109/JOE.2012.2190810.

[163] P. Elmore, C. A. Steed et al., « Algorithm design study for bathymetry fusion-review of current state-of-the-art and recommended design approach », *Naval research lab Stennis space center MS marine geosciences DIV*, 2008.

[164] H. Edelsbrunner, D. Kirkpatrick et R. Seidel, « On the shape of a set of points in the plane », *IEEE Transactions on Information Theory*, t. 29, *4*, p. 551-559, 1983. doi : 10.1109/TIT.1983.1056714.

[165] H. Edelsbrunner et E. P. Mücke, « Three-Dimensional Alpha Shapes », *ACM Trans. Graph.*, t. 13, *1*, p. 43-72, jan. 1994, Place : New York, NY, USA Publisher : Association for Computing Machinery, issn : 0730-0301. doi : 10.1145/174462.156635. adresse : https://doi.org/10.1145/174462.156635.

[166] R. C. dos Santos, M. Galo et A. C. Carrilho, « Extraction of Building Roof Boundaries From LiDAR Data Using an Adaptive Alpha-Shape Algorithm », *IEEE Geoscience and Remote Sensing Letters*, t. 16, *8*, p. 1289-1293, 2019. doi : 10.1109/LGRS.2019.2894098.

[167] J. D. Gardiner, J. Behnsen et C. A. Brassey, « Alpha shapes : determining 3D shape complexity across morphologically diverse structures. », eng, *BMC evolutionary biology*, t. 18, *1*, p. 184, déc. 2018, issn : 1471-2148. doi : 10.1186/s12862-018-1305-z.

[168] B. Delaunay, « Sur la sphère vide », *Bulletin de l'Académie des Sciences de l'URSS. VII. Série*, t. 1934, *6*, p. 793-800, 1934, Publisher : Academy of Sciences of the Union of Soviet Socialist Republics - USSR (Akademiya Nauk SSSR).

[169] R. Rockafellar, M. Wets et R. Wets, *Variational Analysis*, sér. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009, isbn : 978-3-540-62772-2. adresse : https://books.google.fr/books?id=w-NdOE5fD8AC.

[170] Y. Maillot, B. Adam et M. Melkemi, « Shape Reconstruction from Unorganized Set of Points », in *Image Analysis and Recognition*, A. Campilho et M. Kamel, éd., Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, p. 274-283, isbn : 978-3-642-13772-3.

[171] B. Presles, J. Debayle, Y. Maillot et J.-C. Pinoli, « Automatic Recognition of 2D Shapes from a Set of Points », in *Image Analysis and Recognition*, M. Kamel et A. Campilho, éd., Berlin, Heidelberg : Springer Berlin Heidelberg, 2011, p. 183-192, isbn : 978-3-642-21593-3.

[172] H. Samet et R. E. Webber, « Storing a Collection of Polygons Using Quadtrees », *ACM Trans. Graph.*, t. 4, *3*, p. 182-222, juill. 1985, Place : New York, NY, USA Publisher : Association for Computing Machinery, issn : 0730-0301. doi : 10.1145/282957.282966. adresse : https://doi.org/10.1145/282957.282966.

[173] M. H. Weik, « American Standard Code for Information Interchange », in *Computer Science and Communications Dictionary*, Boston, MA : Springer US, 2001, p. 43-43, isbn : 978-1-4020-0613-5. doi : 10.1007/1-4020-0613-6_580. adresse : https://doi.org/10.1007/1-4020-0613-6_580.

[174] P. Merino Laso, D. Brosset et J. Puentes, « Monitoring Approach of Cyber-physical Systems by Quality Measures », in *S-CUBE 2016 : 7th International Conference on Sensor Systems and Software*, t. 205, Nice, France : Springer, déc. 2016, p. 105-117. doi : 10.1007/978-3-319-61563-99. adresse : https://hal.archives-ouvertes.fr/hal-01609035.

[175] A. Leroy, V. Mousseau et M. Pirlot, « Learning the Parameters of a Multiple Criteria Sorting Method », in *Algorithmic Decision Theory*, R. Brafman, F. Roberts et A. Tsoukiàs, éd., t. 6992, Springer, 2011, p. 219-233.

[176] V. Mousseau, J. Figueira et J. Naux, « Using assignment examples to infer weights for ELECTRE TRI method : Some experimental results », *European Journal of Operational Research*, t. 130, *2*, p. 263-275, avr. 2001.

[177] V. Mousseau et R. S\lowiński, « Inferring an ELECTRE TRI model from assignment examples », *Journal of Global Optimization*, t. 12, *2*, p. 157-174, 1998.

[178] A. Ngo The et V. Mousseau, « Using Assignment Examples to Infer Category Limits for the ELECTRE TRI Method », *JMCDA*, t. 11, *1*, p. 29-43, nov. 2002.

[179] A.-L. Olteanu et P. Meyer, « Inferring the parameters of a majority rule sorting model with vetoes on large datasets », in *DA2PL 2014 : From multiple criteria Decision Aid to Preference Learning*, Ecole Centrale Paris et Université de Mons, 2014, p. 87-94.

[180] O. Sobrie, V. Mousseau et M. Pirlot, « Learning a Majority Rule Model from Large Sets of Assignment Examples », in *Algorithmic Decision Theory*, Springer Berlin Heidelberg, 2013, p. 336-350.

[181] ——, « A Population-Based Algorithm for Learning a Majority Rule Sorting Model with Coalitional Veto », in *Evolutionary Multi-Criterion Optimization*, H. Trautmann, G. Rudolph, K. Klamroth et al., éd., Cham : Springer International Publishing, 2017, p. 575-589.

[182] D. Bouyssou et T. Marchant, « An axiomatic approach to noncompensatory sorting methods in MCDM, I : The case of two categories »,

*European Journal of Operational Research*, t. 178, *1*, p. 217-245, avr. 2007, ISSN : 0377-2217.

[183] ——, « An axiomatic approach to noncompensatory sorting methods in MCDM, II : More than two categories », *European Journal of Operational Research*, t. 178, *1*, avr. 2007, ISSN : 0377-2217.

[184] B. Roy, *Multicriteria Methodology for Decision Aiding*. Dordrecht : Kluwer Academic, 1996.

[185] D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs et P. Vincke, *Evaluation and decision models with multiple criteria : Stepping stones for the analyst*, 1st, sér. International Series in Operations Research and Management Science, Volume 86. Boston : Springer, 2006, ISBN : 0-387-31098-3.

[186] Shom, « Programme national d'hydrographie 2021-2024 », Shom, rapp. tech., 2021.

[187] B. Calder, « On the Uncertainty of Archive Hydrographic Data Sets », *IEEE Journal of Oceanic Engineering*, t. 31, *2*, p. 249-265, 2006. DOI : 10.1109/JOE.2006.872215.

[188] S. M. Smith, « The navigation surface a multipurpose bathymetric database », mém. de mast., University of New Hampshire, 2003.

[189] L. Gomes et P. Oliveira, « Bathymetric data fusion : PCA based Interpolation and regularization, sea tests, and implementation », in *OCEANS 2008*, 2008, p. 1-8. DOI : 10.1109/OCEANS.2008.5151973.

[190] A. Stateczny et I. Bodus-Olkowska, « Sensor data fusion techniques for environment modelling », in *2015 16th International Radar Symposium (IRS)*, 2015, p. 1123-1128. DOI : 10.1109/IRS.2015.7226263.

[191] National Oceanic and Atmospheric Administration (NOAA), *What is BlueTopo™ ?*, Anglais, institutionnel. adresse : https://nauticalcharts.noaa.gov/data/bluetopo.html (visité le 28/08/2022).

[192] R. Créach, S. Bosh, L. Boutry, G. Genevier Jonathan, P. Claverie et B. Alexandre, « ScanBathy : A new solution to digitize depth data from historic survey sheets », in *11th Annual GEBCO Bathymetric Science Day (poster presentation)*, Valparaiso, Chile, oct. 2016. adresse : https :

//www.gebco.net/about_us/gebco_symposium/documents/gebco_sd_2016_ts2.pdf.

[193] L. Biscara, C. Salvaterra, J. Rosetto et al., « Production de MNT topo-bathymétriques pour l'amélioration de la gestion du risque de submersion marine : cas du PAPI Saint-Malo », in *XVI Journées Nationales Génie Côtier*, jan. 2020, p. 727-734. doi : 10.5150/jngcgc.2020.081.

[194] F. Parly, « L'intelligence artificielle au service de la défense », française, Ministère des armées, Paris, France, Rapport de la task force IA, sept. 2019, p. 20.

[195] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., « Generative adversarial nets », in *Advances in neural information processing systems*, 2014, p. 2672-2680.

[196] M. K. Cohen, M. Hutter et M. A. Osborne, « Advanced artificial agents intervene in the provision of reward », *AI Magazine*, t. n/a, *n/a*, août 2022, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12064. doi : https://doi.org/10.1002/aaai.12064. adresse : https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12064.

[197] E. Commission, D.-G. f. Justice et Consumers, *Le RGPD : nouvelles opportunités, nouvelles obligations : tout ce que les entreprises doivent savoir à propos du règlement général européen sur la protection des données*. Publications Office, 2018. doi : doi/10.2838/975187.

[198] International Hydrographic Organization (IHO), *An all embracing data model the S 100*, Anglais, Monaco, sept. 2021. adresse : https://www.youtube.com/watch?v=IfKqA7ZkN1w (visité le 31/08/2022).

[199] U. Eco, *Comment voyager avec un saumon*, française, Le Livre de Poche. jan. 2000, isbn : 978-2-253-14792-3.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

**Titre :** Apprentissage automatique de données massives bathymétriques pour l'optimisation de systèmes de levé hydrographique

**Mot clés :** Apprentissage machine ; Détection d'erreurs ; Donnée bathymétrique ; Aide à la décision multicritère

**Résumé :** Les services hydrographiques ont pour pour mission de connaître et décrire l'environnement physique marin dans ses relations avec l'atmosphère, les fonds marins et les zones littorales, d'en prévoir l'évolution et d'assurer la diffusion des informations correspondantes. Dans le cadre de ces missions, dont la sécurité de la navigation est la mission fondamentale, ils réalisent des campagnes à la mer afin d'acquérir le maximum d'informations bathymétriques et océanographiques sur une zone précise. Dans ce manuscrit, nous proposons différentes méthodes visant à faciliter le travail quotidien des opérateurs en considérant différents niveaux d'échelles : micro, meso et macro. Le niveau micro touche à la donnée et donc, dans notre cas, à la sonde bathymétrique afin d'en extraire le maximum de valeur ajoutée possible. Le niveau meso construit à partir de cette donnée bathymétrique les bonnes informations permettant de détecter des données aberrantes via des méthodes d'apprentissage machine dans les lots de données bathymétriques. Enfin, le niveau macro s'attache à la qualification du levé dans son intégralité tout en prenant en compte les préférences de l'utilisateur final via des méthodes d'aide à la décision multicritère.

**Title:** Machine learning on massive bathymetric data for the optimization of hydrographic survey systems

**Keywords:** Machine Learning; Outlier detection; Bathymetric data; Multi-criteria decision analysis

**Abstract:** The mission of hydrographic services are to know and describe the physical marine environment in its interactions with the atmosphere, the seabed and coastal areas, to forecast its evolution and to ensure the dissemination of the corresponding informations. Within the framework of these missions, of which the safety of navigation is the key mission, they carry out campaigns at sea in order to acquire the maximum of bathymetric and oceanographic informations on a precise zone. In this manuscript, we propose different methods to facilitate the daily work of operators, by studying different levels of scales: micro, meso and macro. The micro level focuses on the data, and thus for our case the bathymetric soundings, in order to extract the maximum added value possible. The meso level will build from this bathymetric data the right information to detect outliers data via machine learning methods in bathymetric data. Finally, the macro level concentrates in the qualification of the survey in its entirety while taking into account the preferences of the end user via multiple-criteria decision analysis method.