Université de Bretagne Occidentale

# MASTER SML

**SCIENCES DE LA MER ET DU LITTORAL**

**MENTION**

**BIOLOGIE**

**PARCOURS**

**Ecosystèmes**

SANS Mathurin

# Application de méthodes d'apprentissage statistique à la reconnaissance des opérations de pêche des fileyeurs

Mémoire de stage de Master 2

Année Universitaire **2022-2023**

Structure d'accueil : **Ifremer - Plouzané**

Tuteur universitaire : **Olivier Gauthier**

Maître de stage : **Julien Rodriguez**

iuem INSTITUT UNIVERSITAIRE EUROPÉEN DE LA MER

Ifremer

Université de Bretagne Occidentale

L'Université est une chance

# ENGAGEMENT DE NON PLAGIAT

Je soussigné-e ……Mathurin SANS……

Assure avoir pris connaissance de la charte anti-plagiat de l'université de Bretagne occidentale.

Je déclare être pleinement conscient-e que le plagiat total ou partiel de documents publiés sous différentes formes, y compris sur internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

Je m'engage à citer toutes les sources que j'ai utilisées pour rédiger ce travail.

Signature

*Document à insérer en première page de tous les rapports, dossiers et mémoires.*

## Remerciements :

Je tiens tout d'abord à remercier mon maître de stage Julien Rodriguez pour les connaissances et le soutien apportés tout au long de ce projet. Je tiens à remercier toute l'équipe HISSEO et LBH (Ifremer) avec qui j'ai partagé de très bons moments, avec une attention particulière à Charlène Spagnol, Sébastien Demanèche, Jonathan Rault et Emilie Leblond pour les réponses apportées à diverses questions. Je remercie Hervé Barone de m'avoir introduit à Ifremer. Dernièrement, je tiens à remercier mon tuteur pédagogique Olivier Gauthier ainsi que toute l'équipe pédagogique du master SML pour les deux ans passés avec eux.

## Liste des abréviations :

- Acc : Accuracy - Précision
- AIS : Automatic Identification System – système d'identification automatique
- CART : Classification and Regression Tree – arbre de classification et de régression
- ICES – CIEM : Internation Council for the Exploration of the Sea – Conseil International de l'Exploration de la Mer
- RMSE : Root Mean Square Deviation – écart quadratique moyen
- SIH : Système d'Informations Halieutiques
- SSF : Small Scale Fisheries – pêcheries artisanales
- SVM : Support Vector Machine - Séparateur à Vaste Marge
- VMS : Vessel Monitoring System – système de surveillance des navires
- WKSSFGEO : Workshop On Geo-Spatial Data for Small-Scale Fisheries
- XGBoost : eXtreme Gradient Boosting

# Table des matières

# INTRODUCTION :

Depuis 2016 en France, les échouages de petits cétacés présentant des traces de capture ont atteint des niveaux importants sur le littoral Atlantique. Le nombre de dauphins capturés est estimé à plusieurs milliers chaque année le long des côtes atlantiques françaises, essentiellement dans le golfe de Gascogne (Peltier *et al.* 2020b, Peltier *et al.* 2020a, ICES 2022). Ces niveaux de captures accidentelles pourraient remettre en question la viabilité de la population de dauphin commun de l'Atlantique Nord Est (ICES 2022). Sur cette échelle de temps, il n'apparaît pourtant pas de signes d'un changement majeur de l'abondance de la population des dauphins communs (Blanchard *et al.* 2021), ni d'une augmentation de l'effort de pêche dans le golfe de Gascogne.

Le 2 juillet 2020, la Commission européenne met en demeure la France (l'Espagne et la Suède) de mettre en œuvre des mesures afin d'éviter les prises accessoires non durables d'espèces de dauphins et marsouins par les navires de pêche (EU, 2020).

Face à l'absence de mesures ou à des mesures jugées insuffisantes, l'Europe émet un avis motivé le 15 juillet 2022 à l'encontre de la France et de l'Espagne (EU, 2022). Les deux pays disposent alors d'un délai de deux mois pour répondre à ce problème en prenant les mesures nécessaires. A défaut, la Commission pourrait décider de saisir la Cour de justice de l'Union Européenne, avec le risque de lourdes sanctions financières à suivre (LPO, 2022).

En janvier 2023, l'état met en place des mesures pour lutter contre les captures accidentelles de cétacés. Un arrêté oblige les fileyeurs les plus actifs du Golfe de Gascogne à s'équiper de dispositifs répulsifs aussi appelés « pingers », un autre arrêté concernant tous les fileyeurs et chalutiers de plus de 6m les oblige à s'équiper de la VMS (Vessel Monitoring System) d'ici 2030, et enfin les fileyeurs ou chalutiers de plus de 15m sont obligés de participer chaque année à un programme d'observation en mer. A l'échelle nationale, le Conseil d'Etat a ordonné lundi 20 mars au gouvernement de fermer certaines zones de pêche dans le golfe de Gascogne d'ici six mois, ainsi que de prendre des mesures complémentaires pour permettre d'estimer de manière plus précise le nombre de captures annuelles de petits cétacés. L'objectif étant toujours de *"limiter les captures accidentelles de petits cétacés"* et garantir la conservation des dauphins dans la zone (Conseil d'Etat, 2023).

Dans ce contexte, La Rochelle Université-CNRS et l'Institut français de recherche pour l'exploitation de la mer (Ifremer) ont construit en concertation avec l'Office Français de la Biodiversité (OFB), les professionnels de la pêche et l'État, le projet Delmoges (Delphinus Mouvements Gestion). Il vise, d'une part, à combler des lacunes par l'acquisition des nouvelles

données sur les habitats des dauphins communs dans le Golfe de Gascogne, sur leurs interactions trophiques dans l'écosystème et leurs interactions avec les engins de pêche. D'autre part, le projet propose d'intégrer les connaissances sur l'ensemble du socio-écosystème pour envisager une diversité de scénarios de diminution des captures accidentelles incluant des solutions technologiques et, enfin, d'en évaluer les conséquences biologiques et socio-économiques (Delmoges, 2022).

Dans le cadre de l'étude des interactions entre les dauphins et les engins de pêche, des actions sont envisagées pour décrire plus finement l'activité des flottilles de pêche pressenties comme étant les plus concernées par les captures accidentelles de cétacés. À l'heure actuelle, la qualification des trajets de pêche est basée sur des règles de décision simples utilisant des seuils de vitesse. Les navires de pêche européens de plus de 12 mètres sont réglementairement équipés de systèmes de positionnement VMS émettant une position avec une fréquence minimale de deux heures (une heure en France), et ce depuis plusieurs années, ce qui permet d'accéder à des données spatio-temporelles conséquentes.

L'utilisation de données à plus haute résolution temporelle telles que l'AIS (Automatic Identification System) permet d'envisager le développement d'algorithmes plus élaborés, des travaux étant déjà initiés en ce sens en Europe (WKSSFGEO Vol. 4).

Les travaux dans le cadre de ce rapport ont été réalisés à Ifremer, l'institut français de recherche entièrement dédié à la connaissance de l'océan. Par ses recherches scientifiques et technologiques, ses innovations et ses expertises, l'Ifremer contribue à protéger et restaurer l'océan, à gérer durablement les ressources et milieux marins, et à partager des données et informations marines. L'Ifremer s'engage dans des initiatives et programmes scientifiques de portée nationale, européenne et internationale.

Au sein de l'unité HISSEO (Coordination et valorisation de l'Observation Halieutique), qui coordonne le SIH (Système d'Informations Halieutiques), réseau national d'observation des ressources et de la pêche professionnelle ayant un rôle important d'appui aux politiques publiques et à la recherche, la participation au projet Delmoges est centrée sur l'action 3, « interactions dauphins-pêcheries ».

Dans le cadre du programme DELMOGES et du projet précédent IAPESCA, des méthodes d'apprentissage statistique (plus souvent désignées par le terme « machine learning ») ont été testés sur des données de positionnement qualifiées pour certains types d'engins de pêche. La plupart des méthodes développées ont été implémentées dans un package R « iapesca », (nettoyage et manipulation des données, calcul de covariables pertinentes pour décrire les

trajectoires, optimisation de certaines méthodes d'apprentissage statistique et qualification d'opérations de pêche par des méthodes géo-informatiques).

Les méthodes habituelles de qualification des opérations de pêche reposent sur des seuils de vitesse. Ces méthodes, éprouvées pour des navires opérant des engins trainants (chaluts, dragues), s'avèrent peu efficaces pour les navires utilisant des engins de pêche passifs, particulièrement pour les navires artisanaux. Ces bateaux, désignés par l'acronyme « SSF » pour Small Scale Fisheries, ont une longueur totale de moins de 12m (EMFAF, EU 2021) et ne font pas l'objet d'un équipement VMS obligatoire. De plus, les métriques d'effort de pêche classiques, comme le nombre d'heure de pêche, ne sont pas efficaces pour décrire les engins dormants, car l'activité de pêche de l'engin est complètement dissociée des déplacements du bateau. En parallèle, les méthodes d'apprentissage statistique donnent de bons résultats pour la caractérisation des mouvements de différents moyens de transport (Dodge et al., 2009) et les identifications d'engins de pêche à partir des trajectoires (Huang et al., 2019). Ces antécédents montrent que l'apprentissage statistique est un outil propice à l'analyse de données spatialisées. Nous allons donc appliquer des méthodes d'apprentissage statistique à la reconnaissance des opérations de pêche des fileyeurs, sur un jeu de données comprenant aussi bien des bateaux de plus de 12m que des SSF.

Le stage s'inscrit dans la tâche 3.2 du programme DELMOGES : « Caractérisation des pratiques de pêche à haute résolution spatiales (données VMS, AIS) ». De nouvelles sources de données qualifiées étant disponibles avec une plus haute résolution temporelle pour les fileyeurs, l'objectif du stage consiste à optimiser, tester et évaluer différentes méthodes d'apprentissage statistique (CART, KNN, SVM, random-forest, XGBoost…) appliquées à des jeux de données dégradés à différentes résolutions (de 30 secondes à 1 heure). Ce travail permettra :

- D'améliorer les méthodes et modèles élaborés précédemment
- De définir des métriques de mesures de l'effort plus appropriées à l'étude des navires utilisant des engins passifs
- D'établir les résolutions minimales permettant la reconnaissance des filets en fonction de leur longueur et des pratiques de pêche
- D'initier une réflexion sur une intégration opérationnelle de ces méthodes aux algorithmes en place au sein du Système d'Information Halieutiques

# MATERIEL & METHODES :

Dans un premier temps seront décrits succinctement les jeux de données utilisés qui seront ensuite explorés en utilisant des variables permettant une description de l'effort de pêche.

Dans un deuxième temps, les modèles seront élaborés : les covariables utilisées pour créer les modèles d'apprentissage statistique, les hyperparamètres à optimiser pour chaque modèle, et la méthodologie employée pour l'optimisation et l'évaluation des modèles seront définis.

Enfin, des métriques d'effort appropriées aux engins passifs seront calculées à partir des prédictions obtenues en validation et comparées aux observations, afin d'évaluer l'ensemble du processus.

## Description des données :

Les données sont explorées à différentes échelles de résolution : 3600s, 1800s, 900s, 600s, 300s, 120s, 60s, 30s**,** 20s. Ces valeurs correspondent au laps de temps entre deux positions du bateau. Les informations de géolocalisation pour 20 navires (soit 1323 marées) ont été récupérées à raison d'un ping par seconde, c'est la base de données OBSCAMe. Chacun de nos jeux de données est donc le fruit d'un ré-échantillonnage linéaire à partir des données brutes nettoyées pour arriver à la résolution choisie. Nous disposons donc de neuf jeux de données dont le nombre d'observations varie entre 36 000 pour le jeu à 3600s et 6 500 000 pour celui à 20s. Il peut ne pas sembler évident que l'on ait besoin d'autant de résolutions différentes, cependant la distribution des covariables et des règles de décision varie en fonction des résolutions employées (Rodriguez et al., 2023), il s'agira donc de créer des modèles différents pour chacune d'entre elles.

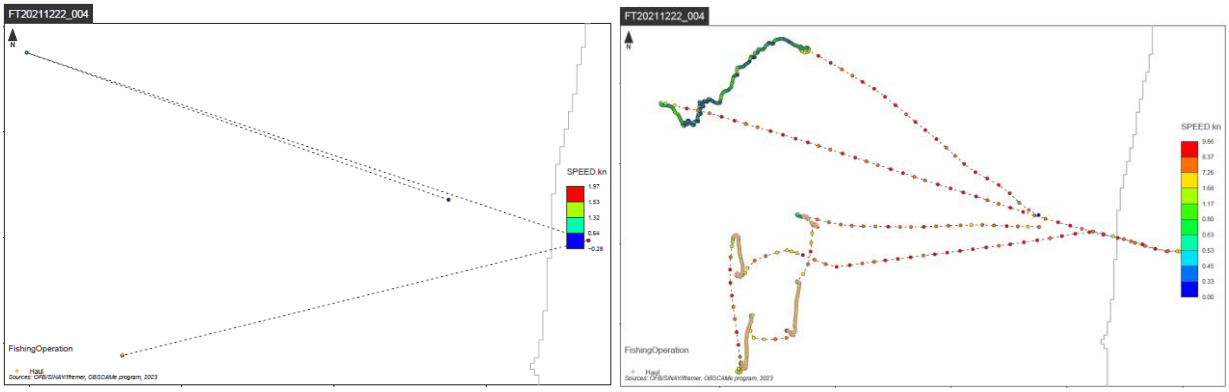


*Figure 1. Marée d'un navire de 12m représentée à 3600s (gauche) et 20s (droite)*

Pour chaque jeu de données, les variables initiales sont les suivantes : Date/Time (instant correspondant à la position), Vessel FK (identifiant navire), Fishing Trip FK (identifiant marée), Fishing operation (NotFishing, Haul, Set), et Gear (engin de pêche utilisé). Les identifiants navires sont anonymisés. Les opérations de pêche sont qualifiées par des observateurs en utilisant les vidéos acquises grâce à des caméras embarquées sur les navires. L'ensemble des covariables qui seront utilisées pour les modèles seront calculées à partir des données de base : temps et positions géographiques.

A partir de ces jeux de données, une phase d'exploration sera conduite afin d'évaluer l'effet de la résolution sur la capacité à récupérer des variables d'effort.

Calcul de l'effort de pêche :

A partir de ces jeux de données, la première étape est le calcul des variables d'effort de pêche qui nous semblent être les plus pertinentes, en s'appuyant sur les définitions du WKSSFGEO Vol.4. Nous avons donc choisi de calculer, par bateau et par résolution :

- Le nombre de marées : trajet d'un bateau de pêche durant lequel des activités de pêche ont lieu qui commence lorsque le bateau quitte le port et se termine lorsqu'il y revient. La zone « port » est définie par un buffer d'1km autour du centre du port, et il doit y avoir un intervalle temporel d'au moins une heure entre la sortie du port et le retour au port pour définir une nouvelle marée. (Rodriguez, 2023)
- Le nombre de virages détectés : nombre de séquences continues de positions correspondant à la récupération d'un engin de pêche (filet dans notre cas). (Rodriguez, 2023)
- Le nombre de filages détectés : nombre de séquences continues de positions correspondant au déploiement d'un engin de pêche (filet dans notre cas). (Rodriguez, 2023)
- Le nombre de jours de pêche : « Jours passés en mer avec au moins une opération de pêche » (*WKSSFGEO* Volume 4, 2022)
- Le nombre d'heures de pêche : "Temps total passé à pêcher, soit le nombre d'heures passées en mer moins le temps passé en transit." (*WKSSFGEO* Volume 4, 2022)
- Le nombre de jours en mer : "Toute période continue de 24h durant laquelle un navire est présent dans une zone de pêche et absent du port." (*WKSSFGEO* Volume 4, 2022)

- Le nombre d'heures en mer : "Durée effective en heure qu'un navire a passé en mer en dehors du port." (*WKSSFGEO* Volume 4, 2022)

A partir de ces variables, nous avons pu comparer l'efficacité des différentes résolutions pour mesurer l'effort de pêche, le jeu de données nous servant de référence étant celui le moins dégradé, soit celui à 20 secondes. Pour les variables d'effort de pêche agrégées touchant au nombre de marées ou au temps de pêche, il a été choisi de ne pas présenter de graphiques dans les résultats car ils montrent très peu de variations. Ils sont cependant disponibles en annexe.

Toute l'exploration des données a été réalisée sur R, avec les packages « iapesca » (Rodriguez, 2023), « SIH.Rhelpers » (package interne HISSEO), « mapview » (Appelhans, 2016) … et en utilisant notamment « plotly » pour réaliser les graphs.

Afin de préparer les travaux sur l'apprentissage statistique, des covariables permettant de décrire les caractéristiques d'une trajectoire ont été définies.

Calcul des covariables:

- Vitesse (Speed): La vitesse (nœuds) en un point est définie comme la vitesse entre le point GPS précédent et celui-ci. La distance entre ces points est calculée selon la formule de Vicenty (1975).
- Accélération (Acceleration): Elle s'exprime comme la différence de vitesse moyenne observée entre deux points $P_{n-1}$ et $P_n$ sur la différence de temps $\Delta_t$ entre ces points.
- Index de proximité (Proximity Index): Correspond au décompte d'observations dans une fenêtre spatiotemporelle spécifique. On définit un rayon en fonction de la résolution, il correspond à la distance parcourue dans le temps entre deux pings par un navire à une vitesse constante de 4.5 nœuds (Rodriguez, 2023). Le seuil de 4.5 nœuds est le seuil sous lequel les navires sont considérés en pêche pour les traitements opérationnels français (Ifremer, 2021).
- Jerk : Taux de changement entre l'accélération et la décélération, fréquemment utilisé dans l'étude des transports et d'intérêt majeur pour l'identification des modes de transport (Dabiri et Heaslip, 2018).
- Bearing : mesure d'angle entre une trajectoire et le Nord géographique.
- Bearing rate : valeur absolue de la différence de bearing entre deux points consécutifs $P_n$ et $P_{n+1}$.
- Changement de vitesse (Speed Change): La variation absolue de vitesse entre le point actuel et le point suivant.

- Rectitude : C'est une mesure qui décrit le ratio entre le trajet suivi et la distance euclidienne entre $P_n$ et $P_{n+k}$.
- Sinuosité (Sinuosity): sa mesure entre deux points correspond à un ratio entre la distance à vol d'oiseau et la distance réellement parcoure entre ces deux points. On a donc une valeur entre 0 et 1, une sinuosité de 1 représentant un profil rectiligne.
- Angle de virage (Turning Angle): Angle de chaque segment par rapport au précédent.
- Changement de direction (Direction Change): Variation de trajectoire dans le temps entre deux trajectoires consécutives (Kitamura et Imafuku, 2015).

Pour chaque observation de la base de données, on calcule ces variables dans une fenêtre glissante qui varie avec la résolution. On les prénomme « Variable_previous » et « Variable_next », respectivement au point précédent et au point suivant. Par exemple, à 900s, pour une observation on a « Speed », la vitesse à la position actuelle, « Speed_previous », la vitesse au point précédent (900s avant), et « Speed_next », la vitesse au point suivant (900s après).

Nous pouvons maintenant aborder les 4 modèles qui seront utilisés dans cette étude : SVM (séparateur à vaste marge), CART (arbre de décision), Random Forest (forêt aléatoire), et XGBoost (eXtreme Gradient Boosting). Dans la partie qui suit, le choix de ces modèles sera justifié en se basant sur le travail bibliographique effectué en amont, et les hyperparamètres à optimiser seront décrits.

Présentation des modèles d'apprentissage statistique :

SVM :

Les machines à support vectoriel ou séparateur à vaste marge (SVM) ont fait l'objet de nombreuses études dans le domaine de la classification des modes de transport (Bolbol *et al*., 2012; Jahangiri and Rakha, 2014; Zheng and Xie, 2008). Leur application au domaine maritime et notamment à l'identification des types de navire de pêche (Kim et Lee, 2020; Marzuki *et al*., 2018) a également montré un réel potentiel. Ces algorithmes reposant sur l'apprentissage supervisé sont réputés pour leur performance à résoudre des problématiques de discrimination. Leur fonctionnement repose sur la recherche de l'hyperplan de marge optimal qui permet de classer ou de séparer correctement les données et dont la capacité de généralisation, ou autrement dit la capacité à séparer les données est la plus grande possible (Tong et Chang, 2001). Il existe différents noyaux de SVM : linéaires, polynomial… celui qui a été choisi pour notre

étude est le SVM à noyau radial car lors des essais sur le jeu de données c'est celui qui donnait les meilleures prédictions.

Pour optimiser le modèle SVM radial, le package R « kernlab » a été utilisé (Karatzoglou et al., 2004), les hyperparamètres sont les suivants :

- C : représente le taux de mauvaise classification. Si sa valeur est élevée, l'optimisation choisira un hyperplan avec une plus petite marge si cet hyperplan classifie mieux les points d'entrainement. Au contraire, une valeur de C faible forcera l'optimisation à chercher une marge plus large, même si elle donne une mauvaise classification pour plus de points.
- Sigma : joue sur la linéarité de la marge. Lorsqu'il prend une valeur faible, la frontière de décision ignore les points éloignés, donnant lieu à une marge non linéaire. Dans le cas contraire, ces points ont plus de poids et la marge est plus linéaire.

CART :

Les classifications à arbres de décision (DT) catégorisent les données à chaque étape de la classification. Cela se traduit par la création d'un arbre répondant à un algorithme spécifique aux données explorées et qui sera utilisé pour simplifier le jeu de données. L'arbre est composé de branches et de nœuds. Les nœuds correspondent à des décisions tandis que les branches s'étalant à gauche ou à droite des nœuds représentent des données encore non classées. La simplicité de leur exécution et de l'interprétation des résultats en font un des algorithmes de classification les plus populaires. De ce fait ils ont été largement utilisés dans la reconnaissance des modes de transport (Dabiri et Heaslip, 2018; Xiao et al., 2017; Zheng et Xie, 2008), et c'est pour cette raison que ce modèle a été choisi pour l'étude, bien qu'il présente moins de potentiel en terme de précision que les trois autres, son optimisation dure moins d'une heure pour le jeu de données à 20s, alors qu'elle dure entre 20 et 48h pour les autres. L'utilisation des DT dans ce domaine a montré une performance satisfaisante dans des applications tendant à les identifier comme une piste prometteuse pour la reconnaissance des opérations de pêche à partir du positionnement des navires.

Le package R « rpart » a été utilisé pour travailler sur le CART, les hyperparamètres à optimiser sont :

- Minsplit : taille minimale d'un nœud, soit le nombre d'individus minimum pour séparer les données en deux.

- Max depth : profondeur maximale de l'arbre (nombre de noeuds).

Random Forest (Forêts Aléatoires) :

Les algorithmes de classification à forêts aléatoires (Breimann, 2001) sont composés de plusieurs arbres de décision. Ces multiples arbres de décisions sont obtenus de manière aléatoire lors du processus d'apprentissage qui s'obtient grâce à une méthode dite de « bootstrap ». La constitution aléatoire des arbres de décision permet de supprimer l'effet de surajustement souvent rencontré lors de l'utilisation d'arbres de décision simples (Breiman, 2001). Leur utilisation pour l'identification des modes de transport (Dabiri et Heaslip, 2018; Xiao et al., 2017) et des engins de pêche (Marzuki et al., 2018, 2015) a montré des performances très intéressantes (Tableau.2). De ce fait, ces algorithmes présentent un fort potentiel pour l'exploitation des données issues de la flottille de pêche expérimentale récoltées dans le cadre de ce projet de recherche.

Le package R « ranger » (Wright et al., 2015) a été préféré au plus classique « randomforest » en raison de son temps de calcul plus courts.

- Num trees : Nombre d'arbres de décision créés.
- Mtry : Ce paramètre contrôle le nombre de covariables disponible pour chaque arbre de décision.
- Min Node Size : Taille minimale pour créer un nœud sur les arbres de décision. Equivalent au minsplit du modèle CART.

XGBoost :

Les algorithmes de classification de type eXtreme Gradient Boosting (XGBOOST) sont des algorithmes ensemblistes fonctionnant grâce à une agrégation d'arbres. Ce type d'algorithmes repose sur un apprentissage par l'erreur lors d'un processus itératif. Ainsi, à chaque itération, l'arbre construit apprend des erreurs commises par l'arbre précédemment construit (Chen et Guestrin, 2016). De cette manière, la règle de décision construite résulte de la somme des résultats de chaque arbre permettant d'obtenir un haut niveau de fiabilité. Leur capacité à être exécutés en parallèle rend ces algorithmes particulièrement bien adaptés aux grands volumes de données (Chen et Guestrin, 2016; Huang et al., 2019). Les études menées dans le domaine des transports (Xiao et al., 2015) et dans l'identification des engins de pêche (Huang et al., 2018) ont révélé que ces algorithmes répondaient très bien à ces problématiques et se sont

avérés être les plus performants de ceux testés dans ces travaux. Ces résultats tendent à encourager leur utilisation dans ces domaines et particulièrement à appuyer leur capacité à discriminer les types d'engins utilisés au sein de la flottille de pêche française.

Le booster choisi est le « gbtree », en effet c'est le seul qui se base sur des arbres et non des modèles linéaires, et une réponse non linéaire est fortement attendue. Le package R « xgboost » (Chen et Guestrin, 2016) permet l'optimisation des modèles de XGBoost via les hyperparamètres suivants :

- Eta : Correspond au taux d'apprentissage, après chaque étape de boosting l'eta diminue le poids des covariables pour rendre le procédé plus conservateur.
- Gamma : Spécifie la réduction minimale de la perte nécessaire pour faire un nœud. Plus gamma est grand, plus l'algorithme est conservateur.
- Max depth : Profondeur maximale d'un arbre.
- Subsample : Fraction des observations échantillonnée aléatoirement pour chaque arbre.
- Colsample by tree : Ratio de re-échantillonages à la construction de chaque arbre.
- Objective : Définition de la fonction objectif (loss function en anglais) à minimiser. Pour de la prédiction multiclasse on utilise soit du multi:softmax soit du multi:softprob. Dans cette étude le multi:softprob est utilisé car il donne la probabilité d'un individu d'appartenir à chaque classe.
- Eval metric : Métrique utilisée pour les données de validation, on utilise mlogloss car nous avons plusieurs classes.
- Num_class : Nombre de classes de la variable à prédire (3 dans le cadre des opérations de pêche : virage, filage, et non en pêche).
- Numtrees : Nombre d'arbres créés.

Optimisation des modèles :

Pour optimiser ces modèles, les hyper-paramètres sont sélectionnés par un échantillonnage systématique couvrant une large gamme de possibilités. Ceux-ci constitue une grille multi-dimensionnelle assez large qui peut être réduite dans un second temps par une deuxième étape d'échantillonnage systématique. Cela permet de faire tourner l'algorithme avec chacune des combinaisons d'hyperparamètres, pour ensuite choisir celles qui permet la construction du modèle donnant les meilleures performances en prédiction.

Pour calculer la précision de chaque modèle lors de cette phase d'optimisation la validation croisée est utilisée. La méthode de la validation croisée consiste à diviser le jeu de

données en n sous-ensembles qui joueront tour à tour le rôle d'échantillon d'apprentissage et d'échantillon test. Ainsi, pour une validation croisée à 3 niveaux (3-folds) on divise le jeu en 3, pour à 5 niveaux (5-folds) on le divise en 5, etc…

Pour le processus d'optimisation, la validation croisée 3-folds a été choisie plutôt que la 5-folds car le temps de calcul est plus court. Plutôt que de tester tour à tour chaque échantillon, on sélectionne dès le début un des trois échantillons comme étant l'échantillon test. Cela permet à nouveau de réduire les temps de calcul. Les positions n'étant pas indépendantes, l'objet sur lequel porte la CV est dans ce cas la marée et non pas les index de ligne comme dans la plupart des procédures de CV. Les résultats seront ainsi plus à même de décrire l'objet de la prédiction qui, dans notre cas, sera soit une nouvelle marée soit un nouveau navire.

Une fois que l'on a les résultats pour la première grille de paramètres, on peut l'affiner et sélectionner de nouveaux hyperparamètres afin d'améliorer les résultats si besoin.

Evaluation des modèles :

Suite à l'optimisation des modèles qui permet de choisir les meilleurs hyperparamètres, il faut évaluer les modèles construit avec ceux-ci. Une première évaluation a été effectuée sous la forme d'une validation croisée à 5 niveaux par marée, ceci afin d'évaluer la capacité à prédire une nouvelle marée pour un navire existant dans la base de calibration.

Pour évaluer les modèles, la validation croisée par navire a été choisie car elle se rapproche plus de leur objectif final (les appliquer aux 400 fileyeurs du Golfe de Gascogne) que la validation par marée. Elle permet d'évaluer les capacités de prédiction du modèle sur des bateaux qui ne sont pas présents dans la base de données, ainsi la validation croisée par navire montre dans quelle mesure ces modèles seront applicables à l'ensemble de la flottille des fileyeurs français.

Nous avons aussi effectuée une validation croisée avec omission d'un élément (leave-one-out) par navire (validation croisée avec autant de niveaux qu'il y a de navires dans le jeu de données), ainsi les opérations de pêche de chaque navire sont estimées à partir des données qualifiées de tous les autres navires, mais pas du navire en question. Le choix du leave-one-out pour la validation croisée s'avérera plus sévère qu'en utilisant les marées mais ces résultats correspondront plus à ce qu'il sera possible d'attendre d'une application de ces modèles à de nouveaux bateaux. Ces résultats ont été utilisés pour le calcul des métriques engins.

On calcule la précision (Accuracy) comme suit : $Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$ , avec TP les vrais positifs, TN les vrais négatifs, FN les faux négatif et FP les faux positifs.

Pour comparer les prédictions des différents navires, les résolutions choisies sont 3600s (1h), 900s (15mn) et 60s (1mn). La première a été choisie afin de regarder les données les plus dégradées disponibles, la résolution 15mn correspondra aux fréquences qui seront utilisées pour le déploiement du flux VMS côtier destiné à équiper certains navires de moins de 12 m français, et enfin la résolution à la minute est au cœur des débats dans les travaux sur les SSF menés par l'ICES-CIEM (Conseil International de l'Exploration de la Mer).

Métriques engins :

Utilisation du package R « iapesca » afin de calculer les métriques engins (Rodriguez, 2023). Ce package permet, à partir d'un jeu de données contenant les positions et les opérations de pêche, de créer des objets spatiaux représentants les filets et leurs métriques (longueur, temps d'immersion, identifiant marée du filage et du virage…).

Nous pouvons donc créer les filets « réels » à partir des données qualifiées, et les filets prédits à partir des résultats de la validation croisée leave-one-out par navire, à partir d'un processus implémenté dans le package iapesca. Ce processus va créer un objet spatial à partir d'une opération de virage (séquence continue de points qualifiés en virage), générer un tampon spatial (« buffer ») de 330m et rechercher dans le passé des séquences de points pouvant correspondre au filage par superposition spatiale. La valeur du tampon ayant été définie sur la base de jeux de données qualifiés. Deux méthodes permettent de réaliser une consolidation des filets générés : « BehaviourChange » et « AutoThresholdDetection ».

L'activation de la méthode « BehaviourChange » prend en compte une classification locale des vitesses et changements de direction pour générer des séparations de filets au sein de séquences continues si un changement de comportement est détecté au sein de celle-ci.

L'activation de la méthode « AutoThresholdDetection » détermine sur la base de statistiques sur les filets générés (longueur, durée d'immersion, vitesse de filage…), des valeurs seuils permettant de palier à une expertise sur les engins de pêche. Nous avons choisi d'utiliser l'AutoThresholdDetection mais de désactiver BehaviourChange, car aux résolutions les plus fines cet argument de la fonction générait généralement des filets beaucoup trop nombreux, de nombreux changements de direction apparaissant en affinant la résolution temporelle.

Nous avons décidé d'utiliser les variables suivantes afin de comparer les bateaux et les capacités de prédictions des modèles à différentes résolutions :

- Durée d'immersion médiane par marée
- Longueur totale déployée par marée
- Longueur totale * durée d'immersion

Une fois ces métriques récupérées pour les données qualifiées et prédites nous comparerons les observations aux prédictions. Le $R^2$ sur l'ensemble des marées sera calculé, alors que le RMSE (Root Mean Square Deviation) et le biais seront préférés pour les données par navire.

## RESULTATS :

Dans un premier temps, les variables d'effort de pêche pour les données qualifiées aux différentes résolutions ont été calculées (WKSSFGEO Vol.4).

Par soucis d'anonymisation des navires et de pertinence de l'information, il a été choisi d'indiquer la classe de taille à laquelle un bateau appartient plutôt que son nom pour parler de l'effort de pêche, en ajoutant les identifiants navires (anonymisés) lorsque cela semblait pertinent.

Pour les bateaux d'une taille supérieure à 12m, le nombre de marées détectées est cohérent pour toutes les résolutions (voir annexes). Cependant, pour les navires moins de 10 m, une marée est perdue dès 300 s avec une perte importante d'informations à 1800 s, plus limitée pour les 10-12m. La résolution 900 s permet néanmoins de récupérer une grande majorité des marées avec le paramétrage choisi.

Le décompte des heures de pêche moyennes par marée ne diffère pas d'une résolution à une autre (voir annexes pour les figures).
Ces estimateurs sont maintenant comparés à une échelle plus fine qui est celle de la marée (Fig. 2 et 3).

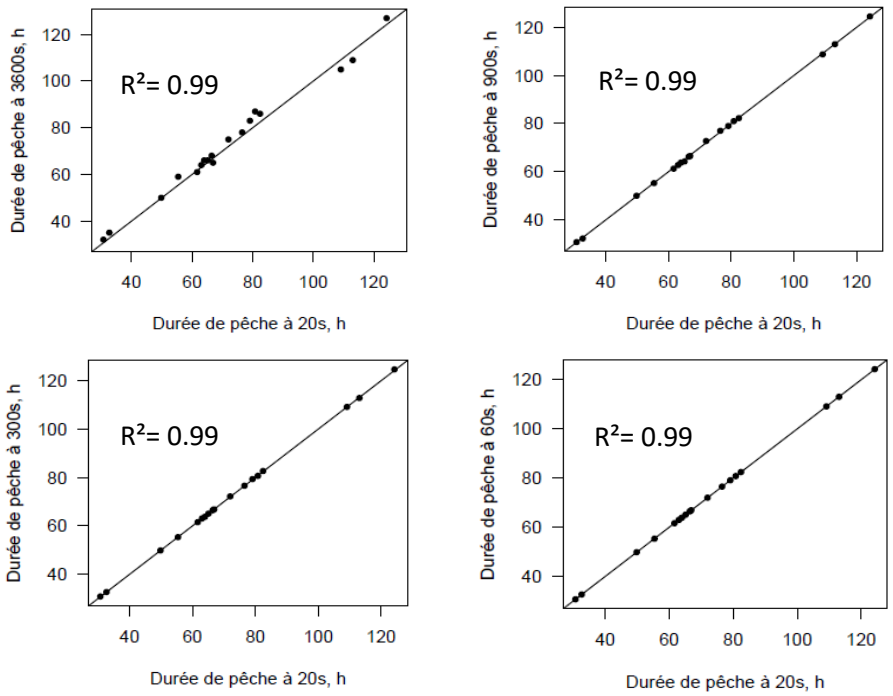


*Figure 2. Durée de pêche d'une marée à différentes résolutions pour un bateau de 24m*

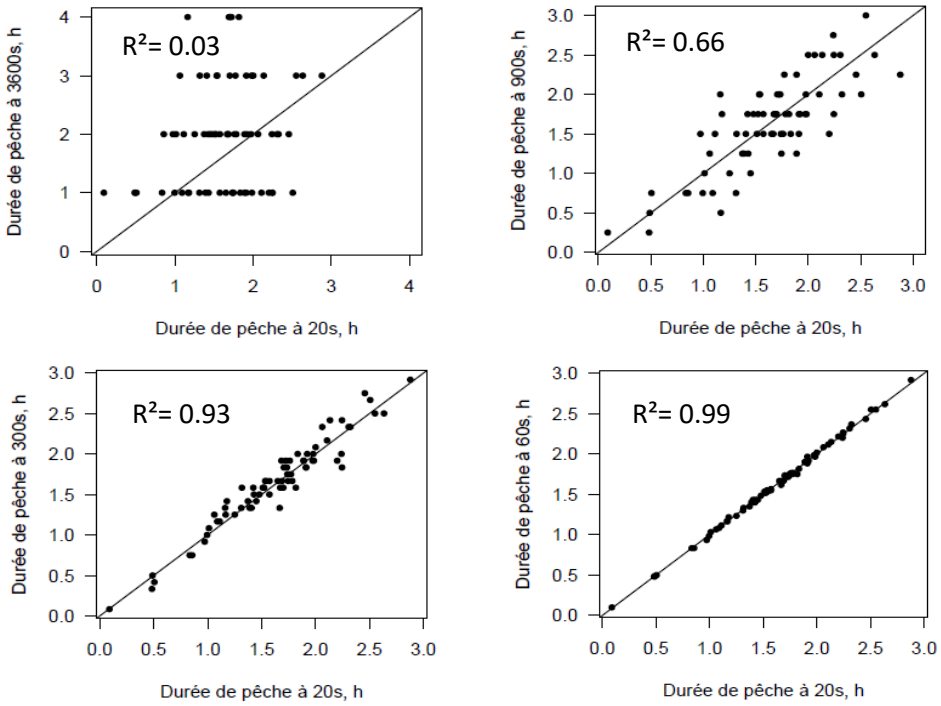


*Figure 3. Durée de pêche d'une marée à différentes résolutions pour un bateau de 10m*

Le calcul de la durée de pêche du bateau de 24m est déjà très fiable à 3600s ($R^2 = 0.99$) (Fig. 2). Pour le bateau de moins de 10m, à 3600s le coefficient de corrélation présente une valeur quasi-nulle et proche de 1 à 60s (Fig. 3). Bien que dégradé à 900 secondes, le calcul est cohérent avec un coefficient de corrélation de 0.66, qui s'améliore très nettement à 300 s.

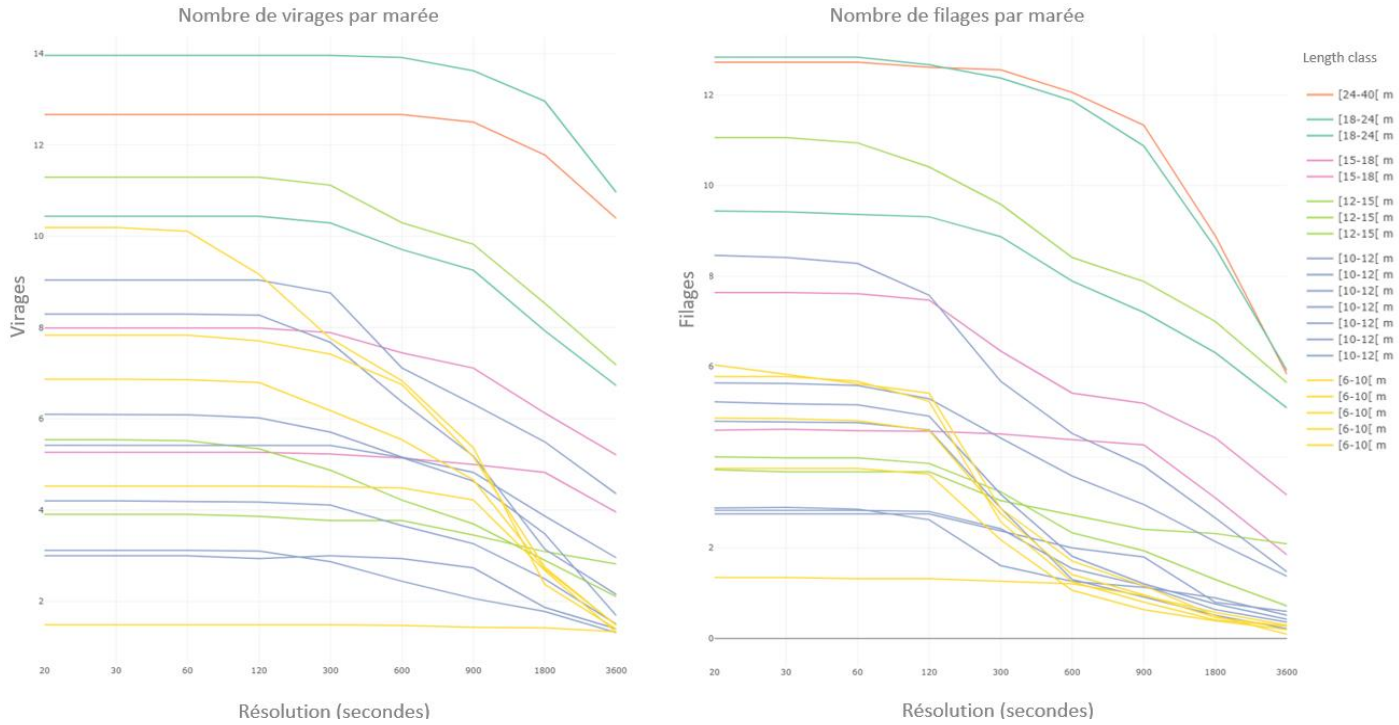


*Figure 4.* Nombres de virages et de filages ramenés à la marée par bateau en fonction de la résolution.

Concernant les opérations de pêche, leur détection se dégrade beaucoup plus rapidement car elles occupent un espace temporel plus fin, les opérations de filage et de virage pouvant avoir des durées de l'ordre de la minute. On peut donc s'attendre à ce que seules les plus fines résolutions permettent une bonne estimation du nombre d'opérations de pêche.

En effet, pour le bateau de plus de 24m il faut prendre tout au plus une position toutes les 600s pour récupérer l'ensemble des opérations de virages. Pour les bateaux entre 15 et 24m, il faut affiner la résolution à 300s pour atteindre ce même objectif, et 120s au minimum pour les 10-15m. Enfin, pour les moins de 12m, 120s s'avèrerait généralement approprié sauf pour deux navires, avec une perte d'informations beaucoup plus brutale aux résolutions plus dégradées (Fig.4) .

Pour récupérer presque toutes les opérations de filage il est nécessaire d'avoir une position toutes les 60s. Bien que pour certains d'entre eux certaines opérations de filage soient perdues, la perte entre 20 et 60s est minime par rapport à celle qui s'opère après (Fig. 4).

le package R iapesca permet de créer les objets filets de chaque marée à partir des positions et des opérations de pêche du navire. A partir des données qualifiées, nous pouvons donc déjà pour chaque résolution créer les filets, afin de calculer les deux métriques engins qui nous intéressent : la longueur totale déployée par marée et la durée d'immersion médiane des filets par marée.

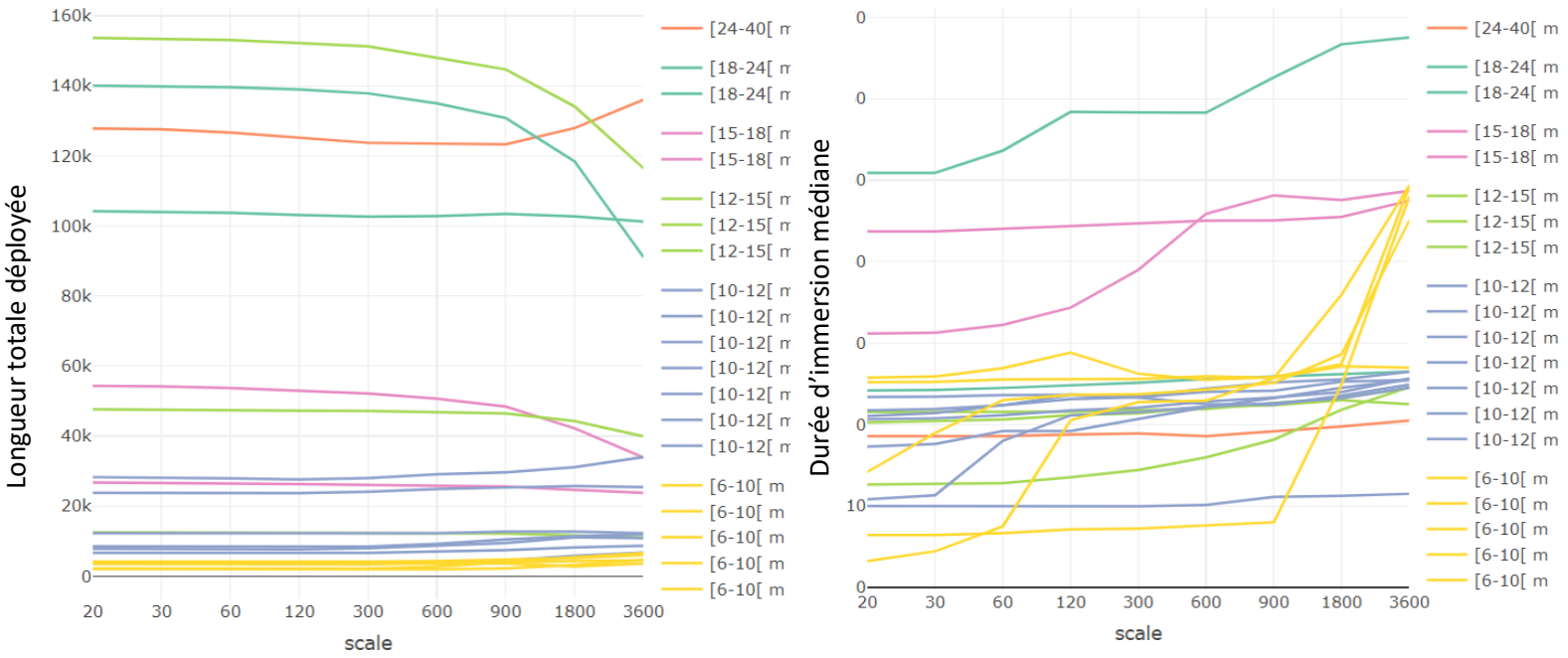


*Figure 5.* Longueur totale déployée et durée d'immersion médiane des filets par marée (moyenne par navire)

En règle générale, il apparaît que plus un bateau est grand, plus la longueur totale de filets qu'il déploie par marée est importante (Fig. 5). On constate une disparité dans l'évolution de ces valeurs en fonction de la résolution, il n'y a pas de tendance globale à la croissance ou la décroissance.

Les durées d'immersion médianes des filets sont surestimées aux résolutions dégradées (Fig. 5). On ne peut pas distinguer de relation entre la taille des bateaux étudiés et les durées d'immersion médianes de leurs filets.

La partie qui suit traite des méthode d'apprentissage statistique utilisées pour prédire les opérations de pêche. Premièrement, des histogrammes ont été réalisés pour la distribution des covariables sur le jeu à 60s, excepté l'angle de virage (Turning Angle) par soucis de présentation. Deuxièmement, les optimisations des modèles seront présentés sous la forme de tableaux contenant les hyperparamètres sélectionnés et les résultats de prédictions. Troisièmement, nous verrons les arbres de décisions CART et les importances des covariables pour le XGBoost. Enfin, les évaluations par marées seront exposées avec les scores de prédiction par navire.

Distribution des covariables :

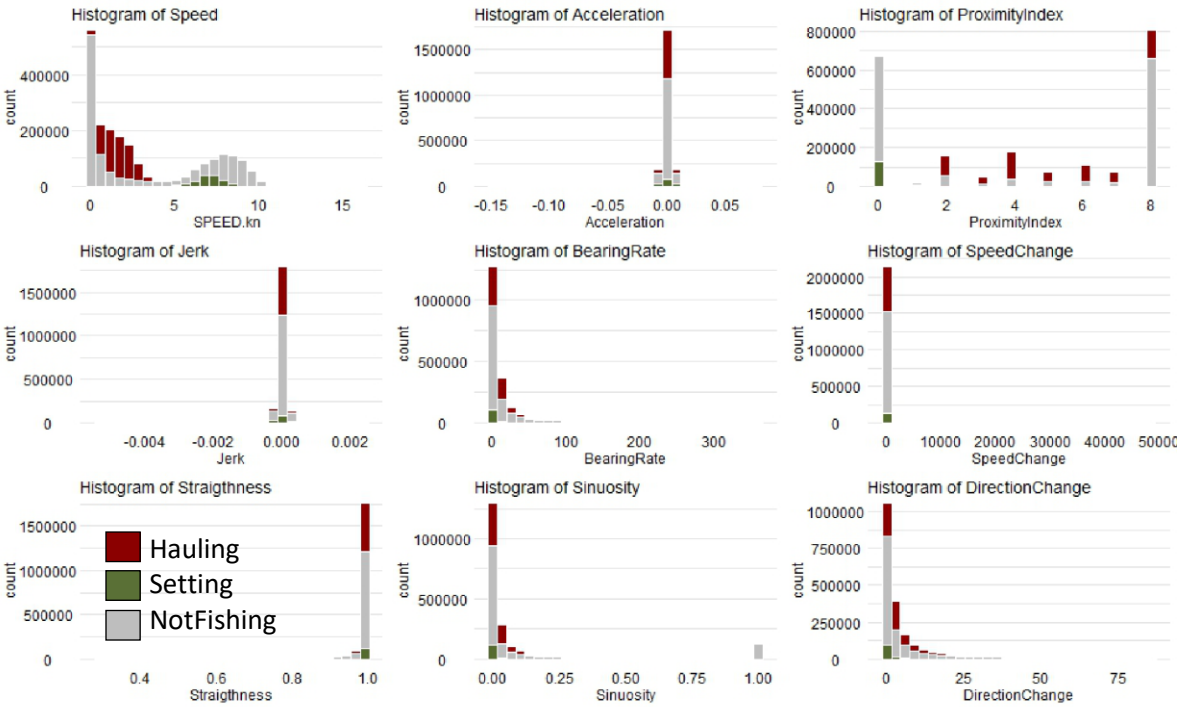


*Figure 6. Histogramme de distribution des covariables à 60s*

A 60s, 98% des évènements de virage ont une vitesse comprise entre 0.15 et 3.43 nœuds, alors que 98% des filages sont compris entre 3.3 et 8.7 nœuds. Les évènements de filage ont un index de proximité souvent nul.

Optimisation des modèles :

Les tableaux ci-dessous présentent, pour chaque modèle, les hyperparamètres retenus à chaque résolution lors de la phase d'optimisation, les couleurs de la colonne « Acc » (précision du modèle) sont le fruit d'un clustering k-means :

| | minsplit | maxdepth | Acc |
|---|---|---|---|
| 3600 | 100 | 8 | 73.14 |
| 1800 | 100 | 10 | 77.04 |
| 900 | 100 | 10 | 80.38 |
| 600 | 100 | 10 | 81.08 |
| 300 | 100 | 10 | 82.84 |
| 120 | 100 | 10 | 83.38 |
| 60 | 100 | 8 | 83.69 |
| 30 | 500 | 5 | 83.28 |

*Table 1.* Hyperparamètres CART retenus aux différentes résolutions

Le choix des hyperparamètres du modèle CART varie très peu d'une résolution à l'autre, mais le score de précision (Acc) augmente de façon presque linéaire avec la résolution jusqu'à 60s .

| | sigma | C | Acc |
|---|---|---|---|
| 3600 | 0.030 | 5 | 75.79 |
| 1800 | 0.025 | 8 | 77.32 |
| 900 | 0.025 | 8 | 81.81 |
| 600 | 0.025 | 5 | 82.85 |
| 300 | 0.040 | 4 | 83.78 |
| 120 | 0.025 | 8 | 83.73 |
| 60 | 0.060 | 40 | 82.09 |
| 30 | 0.050 | 30 | 81.71 |

*Table 2.* Hyperparamètres SVM retenus aux différentes résolutions

Ici il n'y a pas de tendance dans le choix des hyperparamètres. La précision la plus haute est atteinte à 300s, dépassant de quelques centièmes celle à 120s et largement les 60s et 30s.

| | mtry | min.node.size | num.trees | Acc |
|---|---|---|---|---|
| 3600 | 15 | 20 | 500 | 77.90 |
| 1800 | 9 | 10 | 500 | 80.96 |
| 900 | 9 | 10 | 500 | 85.82 |
| 600 | 6 | 15 | 300 | 86.92 |
| 300 | 11 | 10 | 750 | 88.65 |
| 120 | 15 | 10 | 500 | 88.27 |
| 60 | 15 | 20 | 500 | 87.71 |
| 30 | 20 | 30 | 500 | 86.98 |

*Table 3.* Hyperparamètres Random Forest retenus aux différentes résolutions

Même remarque que pour le tableau précédent, on observe une diminution de la précision de plusieurs dixièmes après 300s, puis les scores diminuent jusqu'à 30s.

| | booster | eta | max_depth | gamma | subsample | colsample_bytree | objective | eval_metric | num_class | numtrees | Acc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3600 | gbtree | 0.01 | 15 | 1 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 1250 | 78.78 |
| 1800 | gbtree | 0.01 | 11 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 1200 | 82.34 |
| 900 | gbtree | 0.01 | 11 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 1200 | 86.88 |
| 600 | gbtree | 0.01 | 11 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 1000 | 87.98 |
| 300 | gbtree | 0.01 | 12 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 2000 | 89.66 |
| 120 | gbtree | 0.01 | 14 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 2000 | 89.56 |
| 60 | gbtree | 0.01 | 14 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 2000 | 89.01 |
| 30 | gbtree | 0.01 | 16 | 2 | 0.6 | 1 | multi:softprob | mlogloss | 3 | 2000 | 87.98 |

*Table 4.* Hyperparamètres XGBoost retenus aux différentes résolutions

Encore une fois, la précision augmente jusqu'à 300s puis diminue.

Evaluation des modèles :

| | acc | acc.haul | acc.set | FPR.haul | FPR.set | FNR.haul | FNR.set |
|---|---|---|---|---|---|---|---|
| CART | 73.14 | 88.6 | 0.0 | 40.7 | NA | 11.4 | 100.0 |
| SVM | 75.79 | 89.9 | 14.0 | 37.6 | 48.7 | 10.1 | 86.0 |
| RF | 77.90 | 83.8 | 14.3 | 32.1 | 40.5 | 16.2 | 85.7 |
| XGB | 78.78 | 82.5 | 18.4 | 30.0 | 39.9 | 17.5 | 81.6 |

*Table 5.* Evaluation des modèles en 5fCV par marée à 3600s

| | acc | acc.haul | acc.set | FPR.haul | FPR.set | FNR.haul | FNR.set |
|---|---|---|---|---|---|---|---|
| CART | 80.38 | 83.0 | 0.0 | 26.4 | NA | 17.0 | 100.0 |
| SVM | 81.81 | 88.4 | 45.3 | 28.2 | 36.7 | 11.6 | 54.7 |
| RF | 85.82 | 87.7 | 52.2 | 21.1 | 23.7 | 12.3 | 47.8 |
| XGB | 86.88 | 87.4 | 58.7 | 19.1 | 23.3 | 12.6 | 41.3 |

*Table 6.* Evaluation des modèles en 5fCV par marée à 900s

| | acc | acc.haul | acc.set | FPR.haul | FPR.set | FNR.haul | FNR.set |
|---|---|---|---|---|---|---|---|
| CART | 82.84 | 87.7 | 0.0 | 22.8 | NA | 12.3 | 100.0 |
| SVM | 83.78 | 85.7 | 62.0 | 22.0 | 42.3 | 14.3 | 38.0 |
| RF | 88.65 | 90.5 | 57.6 | 16.1 | 21.3 | 9.5 | 42.4 |
| XGB | 89.66 | 90.3 | 66.0 | 14.4 | 20.6 | 9.7 | 34.0 |

*Table 7.* Evaluation des modèles en 5fCV par marée à 300s

| | acc | acc.haul | acc.set | FPR.haul | FPR.set | FNR.haul | FNR.set |
|---|---|---|---|---|---|---|---|
| CART | 83.69 | 87.7 | 0.0 | 20.9 | NA | 12.3 | 100.0 |
| SVM | 82.09 | 81.9 | 64.6 | 18.2 | 57.6 | 18.1 | 35.4 |
| RF | 87.71 | 91.4 | 40.8 | 15.8 | 34.0 | 8.6 | 59.2 |
| XGB | 89.01 | 91.1 | 57.4 | 14.5 | 28.8 | 8.9 | 42.6 |

*Table 8.* Evaluation des modèles en 5fCV par marée à 60s

Le classement des modèles en terme de précision reste le même pour les différentes résolutions : le XGBoost se place en première position suivi de près par le Random Forest, puis on retrouve le SVM et enfin le CART. La résolution pour laquelle les modèles donnent les meilleurs résultats est 300 secondes excepté pour le CART qui les atteint à 60s (Fig. 11-14).

Les règles de décision des modèles CART sont représentées sous forme d'arbres (Fig. 7). A 3600s, seule des covariables de vitesse sont utilisées : la vitesse pour les deux premiers noeuds, la vitesse au point précédent pour le troisième, et la vitesse au point suivant pour le dernier. A 900s, les premier, deuxième et quatrième nœuds utilisent la vitesse, le deuxième la vitesse au point précédent, et le dernier l'accéleration. A 60s, on utilise d'abord la vitesse au point suivant, puis deux fois la vitesse au point précédent. En fonction du troisième nœud on prend soit la vitesse au point précédent, soit le Jerk puis le Bearing Rate.

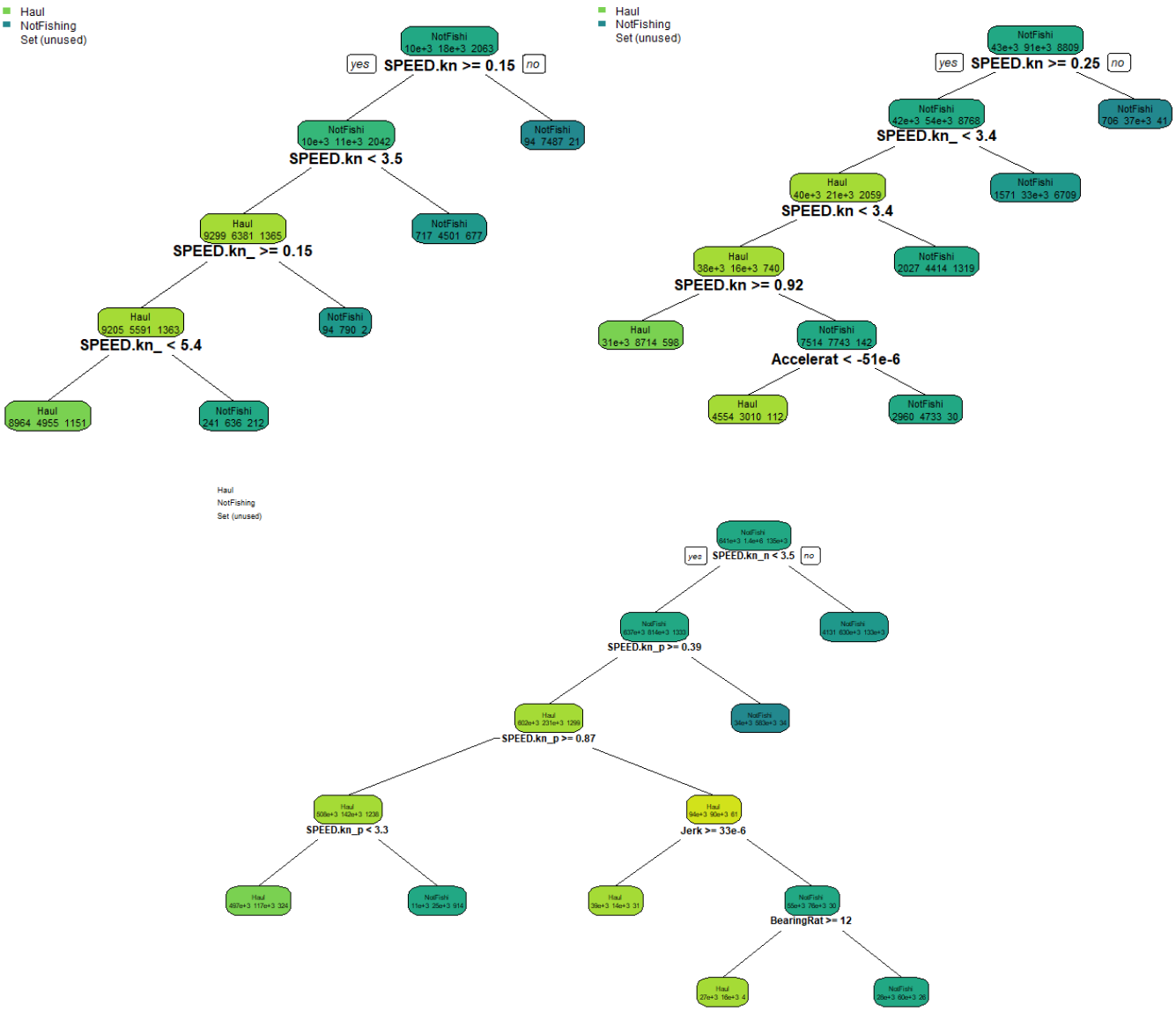




*Figure 7.* Arbres de décision des modèles CART aux résolutions suivantes, en allant de gauche à droite et de haut en bas : 3600s, 900s, 60s

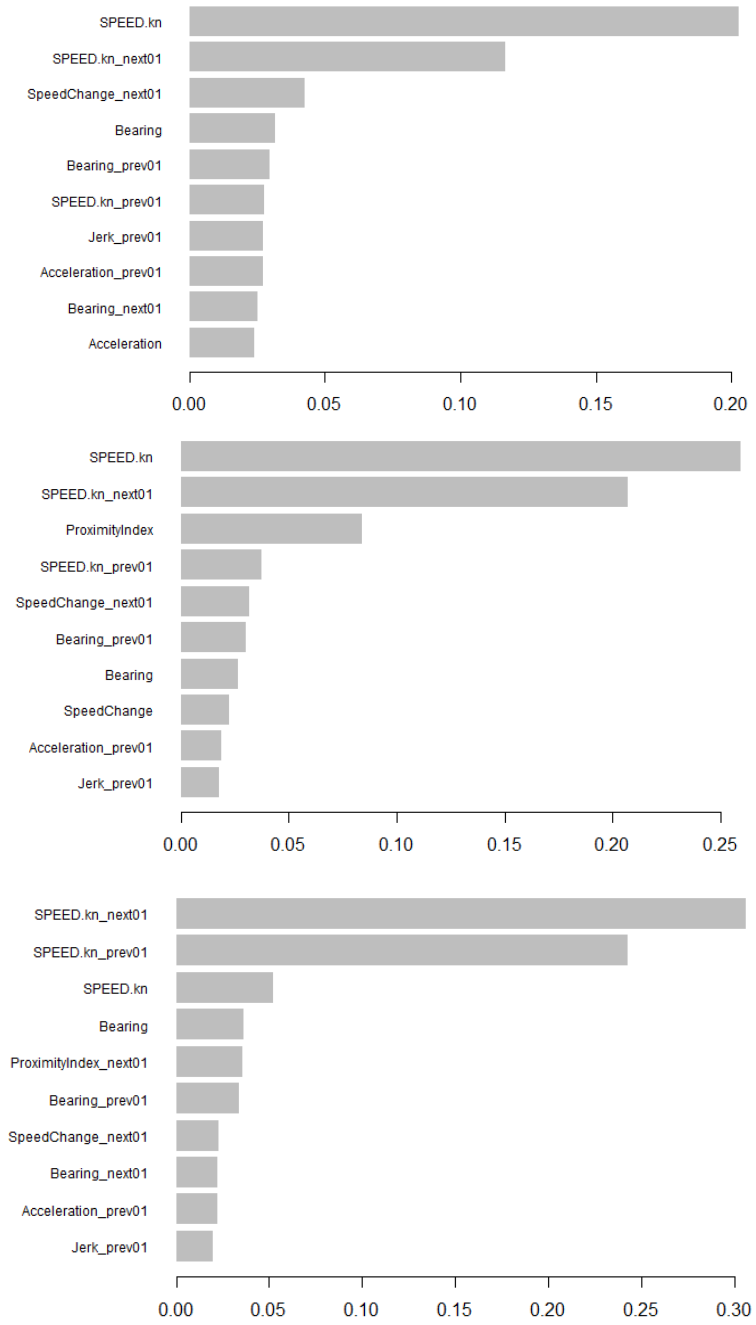


Figure 8. Importance des variables des modèles XGBoost aux résolutions suivantes, en allant de haut en bas : 3600s, 900s, 60s

Pour le XGBoost aussi les covariables de vitesse sont les plus importantes. A 900s, l'index de proximité est la troisième variable la plus importante. A 60s, la vitesse au point actuel est très peu utilisée par rapport aux autres résolutions, et ce sont les vitesses aux points précédent et suivant qui sont les plus importantes.

Validation croisée par navire :

| Identifiant: | Longueur : | Acc 3600s | Acc 900s | Acc 60s |
|---|---|---|---|---|
| NAVIRE_0155 | 24.4 m | 0.59 | 0.6 | 0.71 |
| NAVIRE_0156 | 17.2 m | 0.83 | 0.9 | 0.92 |
| NAVIRE_0157 | 11.9 m | 0.72 | 0.85 | 0.91 |
| NAVIRE_0158 | 11.2 m | 0.81 | 0.85 | 0.87 |
| NAVIRE_0159 | 14.8 m | 0.9 | 0.96 | 0.91 |
| NAVIRE_0160 | 7.7 m | 0.61 | 0.65 | 0.86 |
| NAVIRE_0161 | 12.3 m | 0.72 | 0.81 | 0.86 |
| NAVIRE_0162 | 12 m | 0.75 | 0.84 | 0.87 |
| NAVIRE_0163 | 8.3 m | 0.63 | 0.71 | 0.87 |
| NAVIRE_0164 | 8.6 m | 0.59 | 0.6 | 0.84 |
| NAVIRE_0165 | 12 m | 0.81 | 0.9 | 0.92 |
| NAVIRE_0166 | 15.8 m | 0.8 | 0.9 | 0.93 |
| NAVIRE_0167 | 19.1 m | 0.75 | 0.85 | 0.86 |
| NAVIRE_0168 | 9.5 m | 0.65 | 0.72 | 0.84 |
| NAVIRE_0169 | 14.8 m | 0.88 | 0.95 | 0.94 |
| NAVIRE_0170 | 9.6 m | 0.83 | 0.7 | 0.7 |
| NAVIRE_0171 | 12 m | 0.86 | 0.91 | 0.91 |
| NAVIRE_0172 | 12 m | 0.73 | 0.8 | 0.87 |
| NAVIRE_0173 | 18 m | 0.78 | 0.86 | 0.82 |
| NAVIRE_0174 | 10.2 m | 0.77 | 0.87 | 0.89 |

*Table 9.* Résultats de la validation croisée leave-one-out du XGBoost pour chaque navire

Pour presque tous les bateaux le score de prédiction augmente à mesure que l'on augmente la résolution (Table 9). Cependant, deux bateaux (0159 et 0173) ont un meilleur score à 900s qu'à 60s, et un bateau (0170) a un meilleur score à 3600s qu'à 900 et 60s.

Désormais, le but est de créer des filets à partir des prédictions et de calculer les métriques engins (Longueur Totale déployée et Durée d'Immersion médiane) par marées sur les prédictions. Les fonctions utilisées rencontrent des problèmes pour le jeu à 60s qui ne sont pas réglables dans l'immédiat, les résultats présentés sont donc ceux à 300s.

Calcul des métriques engins sur les prédictions :

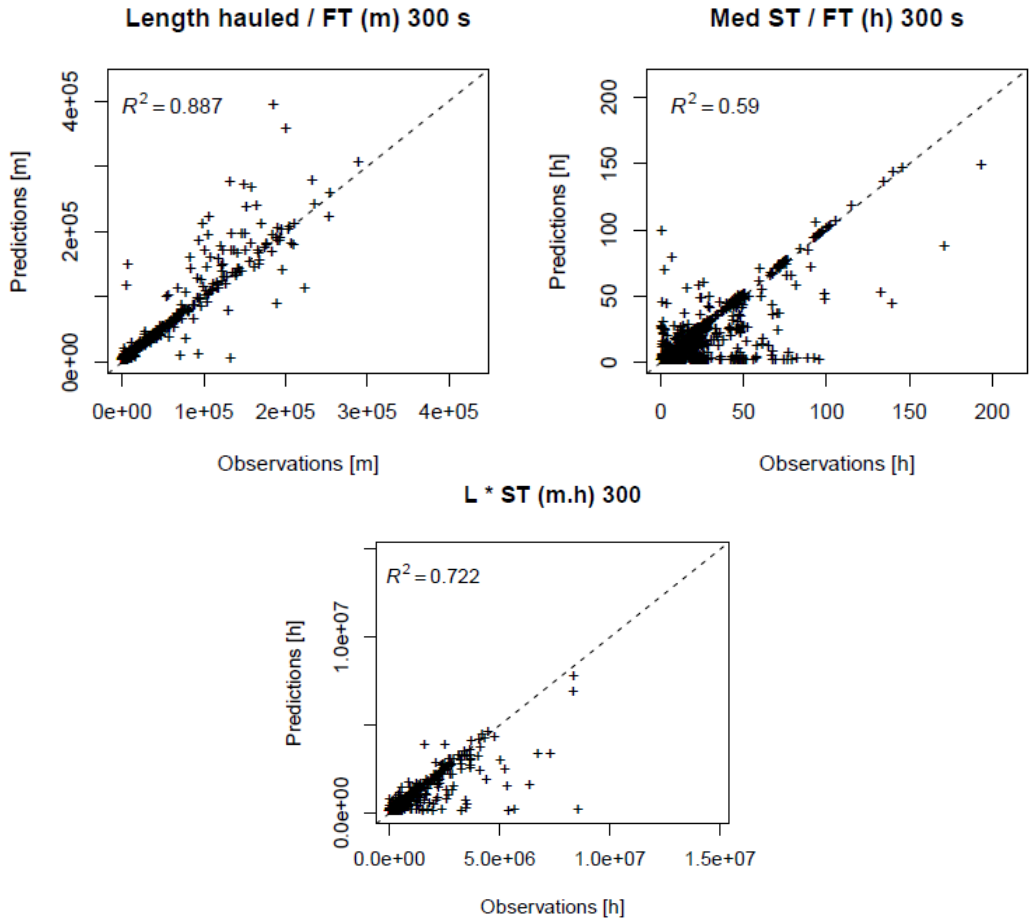


*Figure 9.* Comparaison des métriques prédites avec les métriques observées (Longueur déployée, Durée d'immersion, Longueur déployée * Durée d'immersion)

Chaque figure représente toutes les marées de tous les bateaux. On voit que la longueur totale déployée par marée est bien estimée avec un $R^2$ de 0.887 entre les observations et les prédictions. Pour la durée d'immersion ainsi que le produit des deux métriques, les prédictions sont moins précises avec respectivement des coefficients de corrélation de 0.590 et 0.722 (Fig. 9).

| Identifiant | NA Length | RMSE Length | RMSE/mean | Bias Length | NA SoakTime | RMSE SoakTime | RMSE/mean | Bias SoakTime |
|---|---|---|---|---|---|---|---|---|
| NAVIRE_0155 | 0 | 93694.4 | 0.757296345 | 78698.6 | 100 | NA | NA | NA |
| NAVIRE_0156 | 0 | 4840.3 | 0.092881582 | -2330.6 | 8.333333333 | 14 | 0.355681035 | -6.6 |
| NAVIRE_0157 | 0 | 761 | 0.11422688 | -171.9 | 16.93548387 | 14.2 | 0.714129984 | -1.1 |
| NAVIRE_0158 | 0 | 2871.5 | 0.341501646 | 1480.8 | 31.25 | 5.5 | 0.76703645 | -4.9 |
| NAVIRE_0159 | 0 | 4999.4 | 0.106024464 | 1280.4 | 15 | 8.8 | 0.501818599 | 0.6 |
| NAVIRE_0160 | 0 | 1596.2 | 0.755975047 | 919.6 | 39.56043956 | 18 | 0.935359421 | -6.4 |
| NAVIRE_0161 | 0 | 2618.6 | 0.212157545 | -1486.2 | 15.45454545 | 11.8 | 0.823002935 | -3.9 |
| NAVIRE_0162 | 0 | 18097.8 | 1.47898339 | -3442.5 | 9.302325581 | 5.6 | 0.241808991 | -2.1 |
| NAVIRE_0163 | 0 | 1407 | 0.419194459 | -243.8 | 18.70503597 | 15.2 | 0.595924817 | -3.8 |
| NAVIRE_0164 | 0 | 524 | 0.154213408 | -42.5 | 26.47058824 | 12.4 | 0.536845728 | -0.6 |
| NAVIRE_0165 | 0 | 1919.8 | 0.079559289 | 91.7 | 13.79310345 | 12.7 | 0.589668979 | -4.7 |
| NAVIRE_0166 | 0 | 2066.1 | 0.079230591 | -71.9 | 18.07228916 | 11.8 | 0.277612704 | -3.4 |
| NAVIRE_0167 | 0 | 33081.1 | 0.322221198 | 17696.7 | 30.43478261 | 33 | 0.788225059 | -20.4 |
| NAVIRE_0168 | 0 | 1271.7 | 0.695970234 | 573.8 | 11.53846154 | 23.5 | 0.98684173 | -10.3 |
| NAVIRE_0169 | 0 | 7103.2 | 0.046956866 | -4450 | 0 | 2.5 | 0.144088148 | -0.4 |
| NAVIRE_0170 | 0 | 2573.3 | 0.605346243 | -2332.7 | 10.76923077 | 22.1 | 0.854923036 | -17.4 |
| NAVIRE_0171 | 0 | 2965.7 | 0.105836864 | -466.1 | 18.30985915 | 9.5 | 0.531853211 | -3 |
| NAVIRE_0172 | 0 | 1969.8 | 0.246702922 | -867.3 | 19.04761905 | 11.5 | 0.696296018 | -4.3 |
| NAVIRE_0173 | 0 | 43525.3 | 0.31580149 | 4926.2 | 38.46153846 | 28.3 | 1.05032552 | -13.7 |
| NAVIRE_0174 | 0 | 1205 | 0.338569992 | -991.5 | 12.12121212 | 17.6 | 1.191298665 | -8.8 |

*Table 10.* Proportion de valeurs manquantes (%), RMSE, RMSE/moyenne et biais pour les métriques « Longueur totale déployée », « Durée d'immersion médiane », ainsi que leur produit.

Le RMSE est coloré en fonction de sa valeur, plus sa valeur est haute par rapport aux autres plus la couleur de la case tend vers le rouge. Les plus faibles valeurs sont représentées en vert.

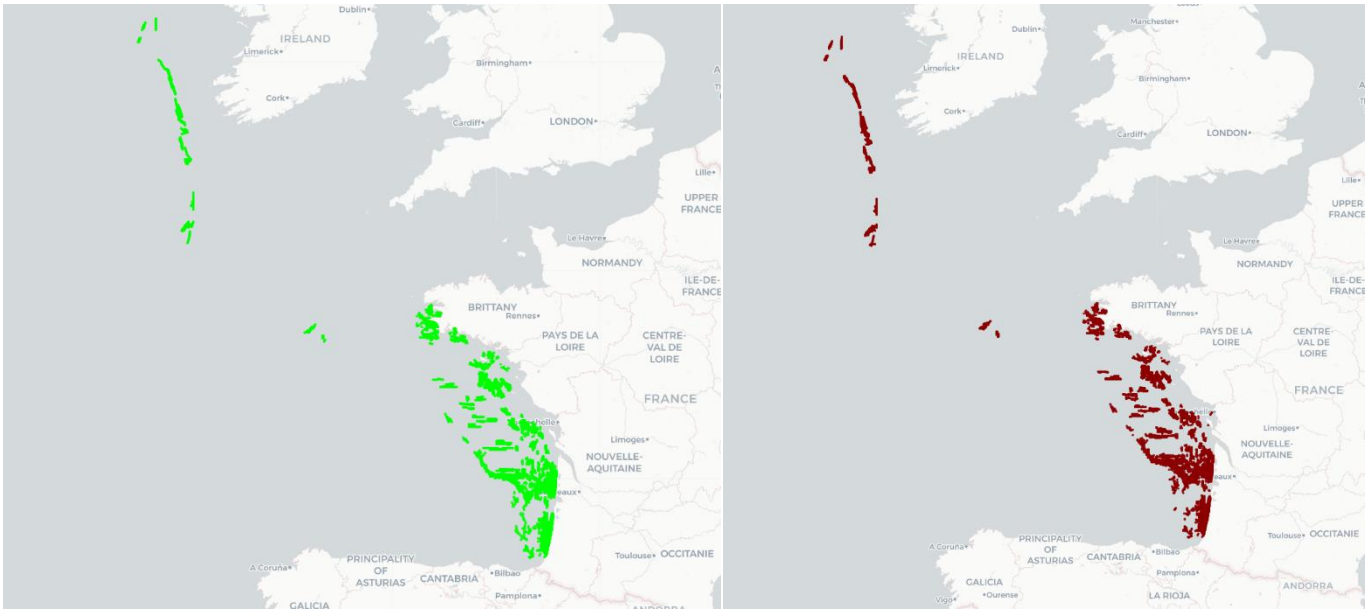


*Figure 10. Comparaison des filets créés à partir des données qualifiées (vert) et prédites (rouge) pour l'ensemble du jeu de données à 300s*

## DISCUSSION :

### Effet de la résolution temporelle sur l'évaluation de l'effort de pêche navire et engin :

En premier lieu nous allons aborder le sujet de l'évaluation de l'effort de pêche aux différentes résolutions. Les premiers résultats décrits seront les comparaisons des métriques d'effort de pêche, calculées à partir des données qualifiées OBSCAMe.
Le calcul de ces variables par bateau pour chaque résolution nous permet d'estimer la capacité d'estimation d'effort de pêche de chacune d'entre elles, et plus précisément de définir la résolution minimale pour atteindre un résultat satisfaisant.
Le nombre de marées détectées par bateau diminue pour les résolutions temporelles de plus de 15min pour les moins de 12m, sans varier pour les navires plus grand. Ces navires opérant près des côtes réalisent des marées courtes, qui ne sont pas détectées à toutes les résolutions car elles ne comportent pas assez de points lorsque l'information est dégradée (Fig. 1).
Le choix d'un tampon de un kilomètre autour du port peut aussi s'avérer parfois insuffisant, en effet si un bateau passe une durée inférieure à deux heures en dehors de cette zone, il risque de n'avoir qu'un point à la résolution la plus dégradée (3600s) et donc son trajet ne sera pas qualifié comme une marée. Une valeur plus grande de ce tampon conduirait cependant à agréger des marées et résulte d'un compromis pour ce jeu de données.
Quant aux temps de pêche moyens par navire agrégés, ils ne varient pas avec la résolution. Ainsi, pour décrire des métriques d'effort de pêche agrégées à l'échelle du navire, une résolution temporelle d'une heure est correcte pour les bateaux de plus de 12m, alors qu'il faut prendre 15mn ou plus précis pour les SSF.

Cependant, pour être en mesure de récupérer les durées de pêche associées à une marée, il peut être nécessaire d'affiner la résolution. En effet, pour le plus grand navire, elles sont bien restituées à toutes les échelles (Fig. 2) mais pour le bateau le plus petit, il est nécessaire de descendre à 300s (5min) pour avoir un coefficient de corrélation élevé (Fig. 3), les données restant cependant cohérentes à 900 s ($R^2 = 0.66$) ce qui n'est plus du tout le cas à 1 heure. S'agissant de navires opérant des engins passifs, l'heure de pêche ne constitue pas une mesure satisfaisante de l'effort de pêche (Mendo et al., 2019). Il apparaît qu'une résolution temporelle beaucoup plus fine est requise pour cet objectif, et ce particulièrement concernant les opérations de filage qui sont beaucoup plus fugaces car réalisées avec une vitesse élevée. L'utilisation de positions ré-échantillonnées à la minute semble être le meilleur compromis qui permettrait potentiellement de récupérer la quasi-totalité des opérations pour les navires de moins de 12m

(Fig. 4). Ce même objectif serait atteint avec une résolution de 2 minutes pour les navires plus grands. Ces premiers résultats ont été présentés au WKSSFGEO2 (Workshop on Geospatial data for Small Scale Fisheries) du CIEM à Faro et corroborent d'autres études portant sur les engins passifs (Portugal, Danemark, Ecosse…), ils ont contribué aux recommandations portant sur la fréquence d'acquisition des données géospatiales pour les navires artisanaux.

Les méthodes du package iapesca permettent de déterminer des métriques d'effort engin agrégées par navire : longueur totale déployée et durées d'immersion. En règle générale, il apparaît que plus un bateau est grand, plus la longueur totale de filets qu'il déploie par marée est importante. Une disparité dans l'évolution de ces valeurs en fonction de la résolution apparaît, il n'y a pas de tendance globale à la croissance ou la décroissance. Il faut noter que si pour les 6-10m (voire pour les 10-12m) la longueur totale déployée par marée ne semble pas beaucoup bouger de 3600s à 20s, c'est uniquement parce que ces valeurs sont minimes comparées à celles de navires plus grands, mais à 3600s certaines estimations sont deux fois plus grandes que celles à 20s (Fig. 5). Cet effet, contraire à l'intuition, est associé aux méthodes de création des filets qui prennent en compte la moitié de la séquence suivant la dernière opération de pêche ce qui peut générer cet effet quand la donnée est trop dégradée.

Quant à la durée d'immersion, elle est surestimé aux résolutions dégradées pour l'ensemble des bateaux (Fig. 5). Cela peut s'expliquer du fait qu'il manque de nombreuses opérations de filage à ces résolutions, et, si un virage se retrouve sans filage associé, la fonction cherche dans les marées précédentes une position qui pourrait correspondre. Cela aide à retrouver certaines opérations de filages en plus, cependant lorsque le filage associé n'est pas le bon, on peut facilement se retrouver avec des durées d'immersion d'une centaine d'heures voire plus. En affinant la résolution, la plupart de ces surestimations s'estompent avec la capacité à détecter les vraies opérations de filage. De plus, la valeur que nous regardons est la médiane des durées d'immersions, or aux plus fines résolutions on a beaucoup plus d'opérations de pêche détectées, et la médiane s'éloigne des valeurs absurdes et devient plus cohérente.

<u>Application des méthodes d'apprentissage statistique :</u>

La suite des travaux a été l'optimisation des modèles de machine learning. Pour l'étude des données de géolocalisation des bateaux et de qualification d'opérations de pêche, le SVM et le Random Forest sont les modèles qui reviennent le plus souvent avec de très bons résultats de précision, entre 90 et 95% (Huang et al., 2018 ; Marzuki et al., 2015 ; Marzuki et al., 2018).

Cependant, une de ces études teste aussi un modèle de XGBoost et obtient un score de 95.5% (Huang et al., 2018).

La phase d'optimisation permet de mettre en évidence deux résultats. Premièrement, le SVM ne donne pas des résultats aussi satisfaisants que le Random Forest et le XGBoost (Fig. 8-10), qui obtiennent tous deux des scores similaires à 1% près en faveur du second. Deuxièmement, la précision n'augmente pas avec la résolution. Ou plutôt, elle augmente jusqu'à 300s, puis diminue légèrement à chaque pas lorsqu'on continue à affiner la résolution, hormis le CART qui ne diminue qu'après 60s (Table 1). De plus, le CART atteint presque le même score que le SVM, bien que le premier mette beaucoup moins de temps à s'exécuter (à 3600s, l'optimisation du CART dure quelques secondes et celle du SVM une heure environ). Pour le SVM et le Random Forest, le choix des hyperparamètres varie beaucoup, ce qui montre l'importance d'optimiser un modèle pour chaque résolution. Cela explique aussi pourquoi le score de précision diminue après 300s, bien que l'on ait une base d'apprentissage plus grande cela ne veut pas dire que le modèle en sera meilleur.

L'évaluation des modèles met bien évidence la difficulté principale du sujet : classifier les opérations de filage. En effet, la précision des modèles pour les filages ne dépasse jamais 66%, ce qui est très faible quand on la compare aux précisions de 90% plus fréquemment atteintes pour les virages (Tables 5-8). En effet, les virages sont des mouvements plus lents avec un comportement bien particulier, alors que les filages sont des opérations fugaces peu caractérisable même aux faibles résolutions.

Le XGBoost donnant les meilleurs performances et le CART étant le plus facile à mettre en œuvre, nous avons donc sélectionné ces deux modèles pour la suite de l'étude.

Pour chaque modèle CART nous pouvons regarder l'arbre de décision utilisé (Fig. 7). On remarque que bien que le 3600s n'utilise que des seuils de vitesse, le modèle à 300s implémente l'accéleration, alors que le modèle à 60s ajoute un seuil de Jerk et de Bearing Rate. L'ajout du BearingRate permet de prendre en compte les changements de direction, une valeur faible indique un trajet plutôt rectiligne et une valeur forte un trajet sinueux. Il est effectivement apparu sur les cartes à haute résolution que certains bateaux semblent avoir une trajectoire sinueuse lors du virage de leurs filets, cependant c'est un résultat qui devra être validé avec les pêcheurs par la suite.

Pour le XGBoost, la vitesse et la vitesse au point suivant sont les variables les plus importantes à 3600 et 900s, alors que la vitesse au point précédent et la vitesse au point suivant

sont les plus importantes à 60s (5-6 fois plus importantes que la vitesse au point actuelle) (Fig. 8). Ces trois variables étant très corrélées aux résolutions fines, le graphique d'importance est biaisé. En effet, entre la vitesse au point précédent et celle au point actuel, on a un coefficient $R^2$ de 0.97 à 60s, contre 0.53 à 900s et 0.17 à 3600s. Cela peut expliquer pourquoi les performances des modèles se dégradent après 300s, les voisins sont trop proches pour utiliser les vitesse sur la fenêtre glissante comme des variables indépendantes.

La validation croisée (CV) par navire en leave-one-out donne des résultats un peu moins bons que la CV 5-folds par marée, à 300s pour le CART on obtient 82.84% de précision en validation par marée contre 81.2% par navire, et à 300s pour le XGBoost on obtient 89.66% par marée contre 81.2% par navire. Cela découle du fait qu'il y ait plus de variabilité inter-navires qu'inter-marées. Dans le cas de la CV leave-one-out par navire, on ne rencontre pas ce biais, mais il en ressort que les bateaux « atypiques » seront moins bien prédits et font chuter légèrement les scores.

Puisque l'on parle de bateaux « atypiques » regardons les résultats des prédictions bateau par bateau (Table 9). Le bateau le moins bien prédit est celui de plus de 24m, ce bateau est le seul fileyeur du large à merlu dominant de la base de données OBSCAMe sur la nouvelle typologie des fileyeurs français (Demaneche et al., 2021). La difficulté à prédire ce navire pourrait venir d'un comportement différent directement lié à une méthode de pêche unique dans le jeu de données. C'est un obstacle important qui devra être pris en compte pour pouvoir créer des modèles applicables à tous les fileyeurs français.

Nos derniers résultats porte sur les nouvelles métriques d'effort de pêche : la longueur déployée par marée, la durée d'immersion médiane par marée, et le produit des deux. Malheureusement, les résultats à 60s ne sont pas disponibles pour des raisons techniques, cependant les résultats à 300s sont prometteurs. En effet, le coefficient de corrélation entre les prédictions des longueurs déployées et les observations est de 0.89, celui des durées d'immersion de 0.59 et le $R^2$ du produit est 0.72. Le résultat des durées d'immersion peut sembler faible, cependant il y a de nombreuses valeurs manquantes pour cette métrique, du aux éléments de filage perdus. Le tableau du RMSE et du biais montre une grand hétérogénéité entre les navires, ce qui indique que nos modèles nécessitent encore du travail avant de pouvoir prédire correctement l'effort de pêche pour les 400 fileyeurs du Golfe de Gascogne.

<u>Perspectives :</u>

Le résultat le plus surprenant au premier abord est la diminution de la précision après

300s. Cela vient surement du choix des covariables. En effet, il a été choisi pour chaque ligne d'inclure le point suivant et le précédent dans les covariables. Seulement, si on est à 300s cela veut dire que l'on étudie une fenêtre de 10mn, contre seulement 3 à 60s. Ainsi, ce résultat nous apprend qu'il y a plus d'informations sur les opérations de pêche du bateau avec une fenêtre glissante de 5mn qu'avec 60s. On aurait pu prendre plus de voisins pour les covariables, cependant cela aurait beaucoup augmenter la taille de stockage ainsi que les temps de calculs, ce qui aurait rendu la tâche plus compliquée. De plus, avoir plus de voisins permet de prendre partiellement en compte l'autocorrelation des données géospatiales.
Une autre solution, et un indice pour les prochains travaux, serait de garder la même fenêtre à toutes les résolutions, par exemple en utilisant le jeu de données à 60s, mais en prenant les voisins 5mn avant et 5mn après. Ainsi on aurait peut-être des covariables plus parlantes, sans pour autant augmenter le stockage et le temps de calcul.

Les modèles Hidden Markov Chains auraient aussi pu être utilisés, étant donné qu'ils prennent en compte l'autocorrelation et ont déjà été utilisés pour calculer l'effort de pêche à partir de données VMS (Vermard et al., 2010), mais ces modèles ne permettent généralement pas de reproduire correctement le comportement des navires opérant des engins passifs (WKSSFGEO Vol.4, 2022). Les réseaux de neurones donnent de très bons résultats dans les applications à la reconnaissance des modes de transport et des engins de pêche (Dabiri et Heaslip, 2018 ; Kim et Lee, 2020) et pourraient se prêter à la qualification des opérations de pêche. C'est une autre piste qui pourra être explorée dans de futures études.

Enfin, le modèle CART pourra être utilisé pour remplacer les seuils de vitesse actuellement mis en œuvre par ALGOPESCA (algorithme d'exploitation des données de géolocalisation du SIH), car les règles de décision sont simples et permettraient une meilleure description de l'effort de pêche navire des fileyeurs. L'effet de cette implémentation est à évaluer.

Pour le projet Delmoges, la prochaine étape est la création de d'effort agrégées à fine résolution spatiale, afin de cartographier l'effort de pêche dans le Golfe de Gascogne. Pour l'instant les cartes d'effort de pêche utilisent des carrés beaucoup trop larges, avec des métriques peu pertinentes pour décrire les fileyeurs. Ces travaux seront donc d'une importance cruciale pour pouvoir modéliser le risque de captures accidentelles de dauphins.

## Bibliographie :

- Appelhans, T., Detsch, F., Reudenbach, C. and Woellauer, S., 2016, April. mapview-Interactive viewing of spatial data in R. In *EGU General Assembly Conference Abstracts* (pp. EPSC2016-1832).
- Blanchard, A., Dorémus, G., Laran, S., Nivière, M., Sanchez, T., Spitz, J. and Van Canneyt, O., 2021. Distribution et abondance de la mégafaune marine en France métropolitaine. Rapport de campagne SAMM II Atlantique-Manche - Hiver 2021. Observatoire Pelagis (UMS 3462, La Rochelle Université / CNRS) pour la Direction de l'Eau et de la Biodiversité et L'Office Français de la Biodiversité**:** 103-pp.
- Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J., 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. Computers, Environment and Urban Systems 36, 526–537.
- Breiman, L., 2001. Random forests. *Machine learning*, *45*, pp.5-32.
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794.
- Conseil d'Etat 2023 : : https://www.conseil-etat.fr/actualites/captures-accidentelles-de-dauphins-et-marsouins-le-gouvernement-doit-agir-sous-6-mois-pour-garantir-leur-survie-dans-le-golfe-de-gascogne)
- Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transportation Research Part C: Emerging Technologies 86, 360–371.
- Delmoges rapport intermédiaire, 2022
- Demaneche S., Berthou P., Biseau A., Begot E., Leblond S., Leblond E., 2021. Caractérisation et typologie des engins et de l'effort de pêche dans le Golfe de Gascogne en période d'échouages des petits cétacés. Rapport d'expertise Ifremer.
- Dodge, S., Weibel, R. and Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, *33*(6), pp.419-434.
- EMFAF 2021/1139. Regulation (EU) 2021/1139 of the European Parliament and of the council of 7 July 2021 es-tablishing the European Maritime, Fisheries and Aquaculture Found and amending Regulation (EU) 2017/1004.
- EU 2020 : https://ec.europa.eu/commission/presscorner/detail/fr/inf_20_1212
- EU 2022 : https://ec.europa.eu/commission/presscorner/detail/fr/inf_22_3768
- Huang, H., Hong, F., Liu, J., Liu, C., Feng, Y. and Guo, Z., 2019. FVID: fishing vessel type identification based on VMS trajectories. *Journal of Ocean University of China*, *18*, pp.403-412.
- ICES 2022. Working Group on Bycatch of Protected Species (WGBYC). ICES Scientific Reports. 4**:** 265.
- Ifremer. Système d'Informations Halieutiques (2021). Algorithme de traitement de données de géolocalisation ALGOPESCA. Note synthétique.
- Jahangiri, A., Rakha, H., 2014. Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data [WWW Document].
- Karatzoglou, A., Hornik, K., Smola, A. and Zeileis, A., 2004. kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, *11*(9).
- Kim, K., Lee, K.M., 2020. Convolutional Neural Network-Based Gear Type Identification from Automatic Identification System Trajectory Data. Applied Sciences 10, 4010.

- Kitamura, T., Imafuku, M., 2015. Behavioural mimicry in flight path of Batesian intraspecific polymorphic butterfly Papilio polytes. Proceedings of the Royal Society B: Biological Sciences 282.
- LPO 2022 : lpo_l'europe_intime_a_la_france_de_proteger_les_dauphins
- Marzuki, M.I., Garello, R., Fablet, R., Kerbaol, V., Gaspar, P., 2015. Fishing gear recognition from VMS data to identify illegal fishing activities in Indonesia, in: OCEANS 2015 - Genova. Presented at the OCEANS 2015 - Genova, pp. 1–5.
- Marzuki, M.I., Gaspar, P., Garello, R., Kerbaol, V., Fablet, R., 2018. Fishing Gear Identification From Vessel-Monitoring-System-Based Fishing Vessel Trajectories. IEEE J. Oceanic Eng. 43, 689–699.
- Mendo, T., Smout, S., Photopoulou, T. and James, M., 2019. Identifying fishing grounds from vessel tracks: model-based inference for small scale fisheries. *Royal Society Open Science*, *6*(10), p.191161.
- Peltier, H., Authier, M., Caurant, F., Dabin, W., Daniel, P., Dars, C., Demaret, F., Meheust, E., Ridoux, V., Van Canneyt, O. and Spitz, J., 2020a. Bilan 2020 des évènements d'échouages de l'hiver et de l'été, cartographie des mortalités et corrélation spatiale avec les pêcheries. Observatoire Pelagis (UMS 3462, La Rochelle Université / CNRS)**:** 12-p.
- Peltier, H., Authier, M., Caurant, F., Dabin, W., Daniel, P., Dars, C., Demaret, F., Meheust, E., Ridoux, V., Van Canneyt, O. and Spitz, J., 2020b. Identifier la co-occurrence spatio-temporelle des captures accidentelles de dauphins communs et des pêcheries dans le golfe de Gascogne de 2010 à 2019. Observatoire Pelagis (UMS 3462, La Rochelle Université / CNRS)**:** 25-p.
- Rodriguez, J., 2023. Iapesca, a R-package for manipulating and interpretating high resolution geospatial data from fishing vessels.
- Rodriguez Julien, Sans Mathurin, Demaneche Sebastien (2023). Computation of net fishing effort by combining machine-learning and geocomputing methods available in the R-package "iapesca". ICES WKSSFGEO2. 13-16 March 2023, Faro - Journées DELMOGES. 29 March 2023, La Rochelle.
- Tong, S., Chang, E., 2001. Support vector machine active learning for image retrieval, in: Proceedings of the Ninth ACM International Conference on Multimedia, MULTIMEDIA '01. Association for Computing Machinery, New York, NY, USA, pp. 107–118.
- Vincenty, T., 1975. Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. Survey Review 23, 88–93.
- Vermard, Y., Rivot, E., Mahevas, S., Marchal, P. and Gascuel, D., 2010. Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models. Ecological Modelling 221(15), 1757-1769.
- WKSSFGEO (Workshop on Geospatial data for Small Scale Fisheries) Vol. 4. 2022. ICES Scientific Reports. 4:10.
- Wright, M.N. and Ziegler, A., 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409.*
- Xiao, G., Juan, Z., Zhang, C., 2015. Travel mode detection based on GPS track data and Bayesian networks. Computers, Environment and Urban Systems 54, 14–22.
- Xiao, Z., Wang, Y., Fu, K., Wu, F., 2017. Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. IJGI 6, 57.
- Zheng, Y., Xie, X., 2008. Learning Transportation Mode from Raw GPS Data for Geographic Application on the Web, in: Proceedings of the 17th World Wide Web Conference.

# ANNEXES :

## Formules covariables :

- vitesse : La vitesse en un point est définie comme la vitesse entre le point GPS précédent et celui-ci. La distance entre ces points est calculée selon la formule de Vicenty (1975) et nommée $d_{(P_{n-1},P_n)}$ . Ensuite, la différence de temps $\Delta_t$ entre ces points est utilisée pour calculer la vitesse moyenne $S_{p_n} = \frac{d_{(P_{n-1},P_n)}}{\Delta_t}$ .
- accélération : Elle s'exprime comme la différence de vitesse moyenne observée entre deux points $P_{n-1}$ et $P_n$ sur la différence de temps $\Delta_t$ entre ces points : $A_{P_n} = \frac{S_{P_{n+1}} - S_{P_n}}{\Delta_t}$ .
- Index de proximité : Correspond au décompte d'observations dans une fenêtre spatiotemporelle spécifique. On définit un rayon en fonction de la résolution, il correspond à la distance parcourue dans le temps entre deux pings par un navire à une vitesse constante de 4.5 nœuds. [Iapesca git] Le seuil de 4.5 nœuds est le seuil sous lequel les navires sont considérés en pêche pour les traitements opérationnels français (Ifremer, 2021). Cette variable est calculée ainsi : $Proximity\ Index = \sum_{i=t-1}^{t+1} \sum_{k=0}^{2} n_{i,k}$ avec $n_{i,k}$ le décompte d'observations situées temporellement entre $t - res * 4$ et $t + res * 4$ comprises dans le rayon.
- jerk : Taux de changement entre l'accélération et la décélération, fréquemment utilisé dans l'étude des transports et d'intérêt majeur pour l'identification des modes de transport (Dabiri et Heaslip, 2018), sa formule est la suivante : $J_{p_n} = \frac{A_{P_{n+1}} - A_{P_n}}{\Delta_t}$ .
- Bearing : mesure d'angle entre une trajectoire et le Nord géographique. En un point, la trajectoire et calculée avec le point suivant. On calcule d'abord $y = \sin[P_{n+1}(long) - P_n(long)] * \cos[P_{n+1}(lat)]$ ainsi que $x = \cos[P_n(lat)] * \sin[P_{n+1}(lat)] - \sin[P_n(lat)] * \cos[P_{n+1}(lat)] * \cos[P_{n+1}(long) - P_n(long)]$ , pour enfin avoir $Bearing_{(Pn)} = arctan2\ (x, y)$ .
- Bearing rate : valeur absolue de la différence de bearing entre deux points consécutifs $P_n$ et $P_{n+1}$ .
- Changement de vitesse : La variation absolue de vitesse entre le point actuel et le point suivant.
- Rectitude : C'est une mesure qui décrit la différence entre le trajet suivi et la distance euclidienne entre $P_n$ et $P_{n+k}$. Sa formulation générale est la suivante : $Rectitude = \frac{\Delta D_{(n,k)}}{L_{(n,k)}}$

où $\Delta D_{(n,k)}$ représente la distance en ligne droite entre les points $P_n$ et $P_{n+k}$ et où $L_{(n,k)}$ correspond à la distance réellement parcourue entre $P_n$ et $P_{n+k}$ et avec k = 1.

- Sinuosité : sa mesure entre deux points correspond à un ratio entre la distance à vol d'oiseau et la distance réellement parcoure entre ces deux points. On a donc une valeur entre 0 et 1, une sinuosité de 1 représentant un profil rectiligne. Ainsi, $Sinuosité_{P,k} = \frac{\sum_{i=p-k}^{i=p+k-1} di,i+1}{d_{p-k.p+k}}$ , avec $Sinuosité_{p,k}$ la sinuosité entre le point P et le point $P_k$, k le paramètre de lag et d la distance mesurée. Dans le cadre de cette étude le calcul de sinuosité sera réalisé avec k = 1.
- Angle de virage : Calcule les angles de pas (en radians) de chaque segment par rapport au segment précédent, selon la méthode de Baschelet (1981).
- Changement de direction : Variation de trajectoire dans le temps entre deux trajectoires consécutives (Kitamura et Imafuku, 2015). Il se calcule de la manière suivante : $DC_{\propto 2} = \left(180 - \frac{180}{\pi} arccos\left(\frac{a^2+b^2+c^2}{2ab}\right)\right) * \frac{1}{t}$ , où *arccos* est la fonction trigonométrique, *t* le temps écoulé entre le premier et le troisième point, *a* la longueur du segment entre le premier et le second point, *b* la longueur du segment entre le second et le troisième point, et *c* la longueur du segment entre le premier et le troisième point (Fig. 1).

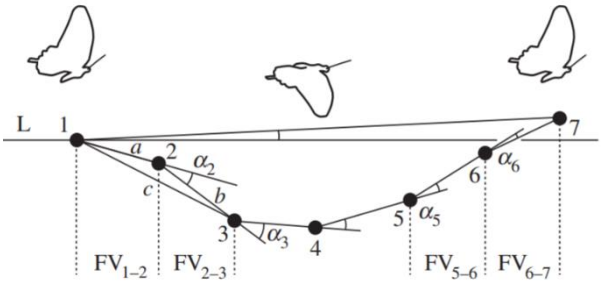


**Figure 1**. Schéma représentant la trajectoire de vol d'un papillon (Kitamura et Imafuku, 2015).a, b et c indique la distance entre les coordonnées 1-2, 2-3 et 1-3 respectivement. $\alpha_2$ indique l'angle extérieur au niveau de la coordonnée 2 en considérant le triangle formé par les positions 1, 2 et 3.

Graphiques d'effort de pêche agrégé :

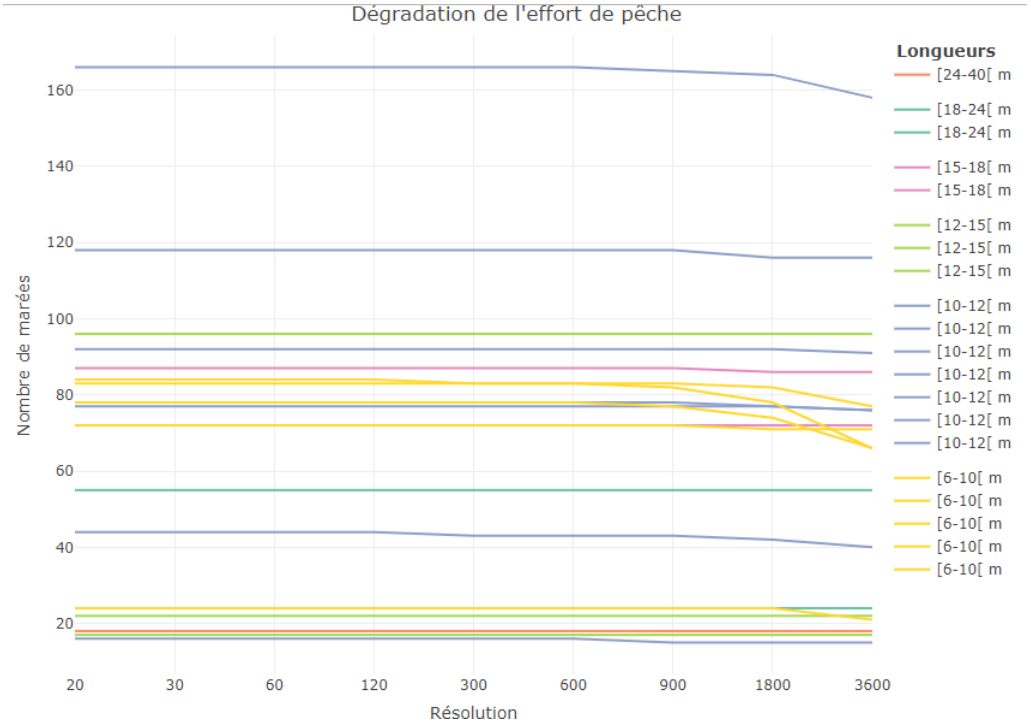


*Figure 2. Nombre de marées agrégé par navire à différentes résolutions*

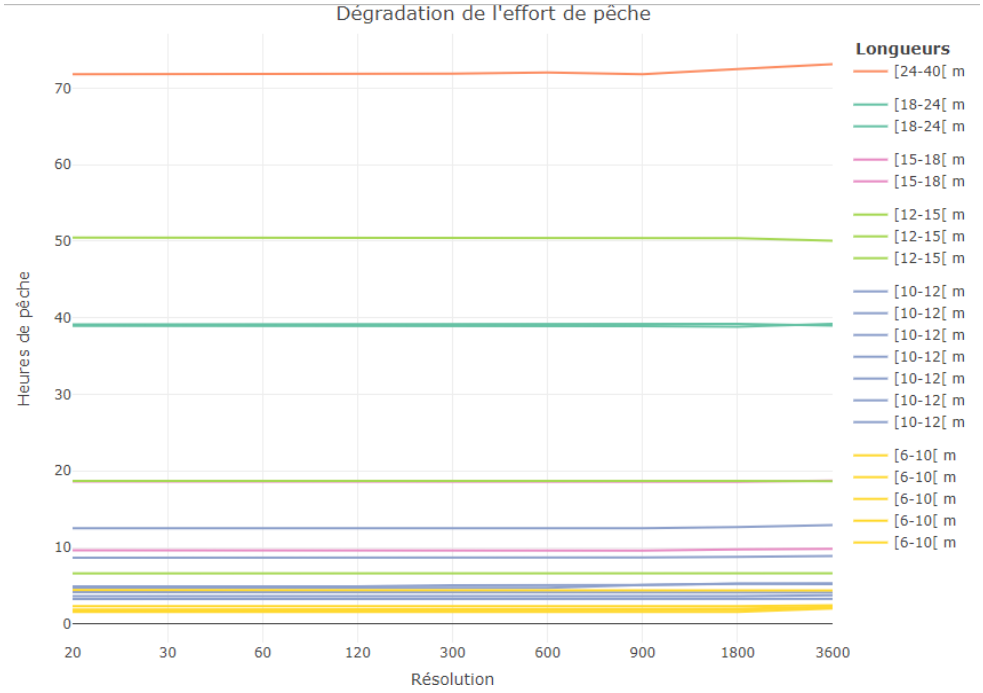


*Figure 3. Durée de pêche agrégée par navire à différentes résolutions*

Evaluation par navire :

| | acc | acc.haul | acc.set | FPR.haul | FPR.set | FNR.haul | FNR.set |
|---|---|---|---|---|---|---|---|
| 3600 | 72.64 | 85.6 | 0 | 40.6 | NA | 14.4 | 100 |
| 1800 | 74.81 | 74.1 | 0 | 33.7 | NA | 25.9 | 100 |
| 900 | 77.62 | 76.8 | 0 | 29.9 | NA | 23.2 | 100 |
| 600 | 78.11 | 79.1 | 0 | 28.2 | NA | 20.9 | 100 |
| 300 | 81.20 | 84.8 | 0 | 25.2 | NA | 15.2 | 100 |
| 120 | 81.70 | 85.4 | 0 | 24.5 | NA | 14.6 | 100 |
| 60 | 81.80 | 82.7 | 0 | 23.0 | NA | 17.3 | 100 |
| 30 | 81.71 | 81.8 | 0 | 22.9 | NA | 18.2 | 100 |

*Figure 4. Résultats de la validation croisée par navire du modèle CART à différentes résolutions*

| | acc | acc.haul | acc.set | FPR.haul | FPR.set | FNR.haul | FNR.set |
|---|---|---|---|---|---|---|---|
| 3600 | 75.93 | 78.8 | 12.9 | 33.4 | 49.3 | 21.2 | 87.1 |
| 1800 | 78.93 | 78.6 | 29.8 | 28.6 | 40.5 | 21.4 | 70.2 |
| 900 | 83.68 | 81.4 | 49.8 | 22.3 | 26.5 | 18.6 | 50.2 |
| 600 | 84.99 | 83.4 | 54.3 | 19.4 | 24.5 | 16.6 | 45.7 |
| 300 | 87.34 | 86.8 | 56.1 | 16.6 | 24.8 | 13.2 | 43.9 |
| 120 | 87.19 | 88.6 | 47.3 | 16.0 | 33.8 | 11.4 | 52.7 |
| 60 | 86.33 | 89.0 | 39.8 | 16.7 | 41.5 | 11.0 | 60.2 |
| 30 | 85.43 | 88.0 | 35.5 | 17.7 | 44.9 | 12.0 | 64.5 |

*Figure 5. Résultats de la validation croisée par navire du modèle XGBoost à différentes résolutions*

## Résumé / Abstract :

Depuis 2016, les échouages de petits cétacés présentant des traces de capture ont atteint des niveaux importants, qui pourraient remettre en question la viabilité de la population de dauphins communs de l'Atlantique Nord (ICES 2022). C'est dans ce contexte que le CNRS et Ifremer en concertation avec l'OFB ont co-construit le programme Delmoges (Delphinus Mouvement Gestion) qui développe une approche scientifique multidisciplinaire visant à une meilleure compréhension des mécanismes de captures accidentelles de dauphins. Des développements sont notamment proposés pour décrire plus finement l'activité des flottilles de pêche pour lequel le risque de capture accidentelle de cétacés est le plus élevé. La qualification des trajets de pêche est basée à l'heure actuelle sur des règles de décision simples utilisant des seuils de vitesse, cette étude se concentre donc sur d'autres méthodes de qualification qui se sont révélées très efficace pour d'autre travaux : l'apprentissage statistique. En utilisant comme jeu d'apprentissage les positions et opérations de pêche qualifiées des 20 fileyeurs de la base OBSCAME, nous avons entraîné différents modèles (CART, Random Forest, XGBoost) à prédire les opérations de pêche d'un fileyeur à partir de ses positions. Les modèles de XGBoost donnent les meilleurs résultats, jusqu'à près de 90% de précision en validation croisée par marée. La dernière étape, pour l'instant encore incomplète, consiste à modéliser les filets des navires du jeu de données à partir des prédictions obtenues pour les navires, dans le but d'évaluer l'effort de pêche des fileyeurs du Golfe de Gascogne en utilisant de nouvelles métriques engins plus à même de décrire l'effort des engins de pêche passifs.

Since 2016, strandings of small cetaceans showing signs of capture have reached significant levels, which could question the viability of the North Atlantic common dolphin population (ICES 2022). It is against this backdrop that CNRS and Ifremer, in conjunction with the OFB, have co-constructed the Delmoges program (Delphinus Mouvement Gestion), which is developing a multidisciplinary scientific approach aimed at gaining a better understanding of the mechanisms involved in the accidental capture of dolphins. In particular, developments are being proposed to provide a more detailed description of the activities of fishing fleets deemed to be most at risk of bycatching cetaceans. The qualification of fishing routes is currently based on simple decision rules using speed thresholds, so this study focuses on other qualification methods that have proved highly effective in other work: machine learning. Using the qualified positions and fishing operations of the 20 gillnetters in the OBSCAME database as a training set, we trained different models (CART, Random Forest, XGBoost) to predict a gillnetter's fishing operations based on its positions. The XGBoost models give the best results, with almost 90% accuracy using fishing trips based cross-validation. The final step, which is still incomplete at the moment, is to model the nets of the vessels in the dataset based on the predictions, with the aim of assessing the fishing effort of Bay of Biscay gillnetters using new gear metrics.