Jules Malavieille

amU Observatoire des sciences de l'Univers — Institut Pythéas
Aix Marseille Université

Modèle de prévision de probabilités de présence d'espèces de téléostéens marin en Méditerranée et en Atlantiques sur les côtes Françaises et Espagnoles

Projet réalisé à titre personnel

Table des matières

1 Résumé

La prévision de la présence d'espèces marines constitue un défi à la fois écologique et méthodologique, notamment en raison de l'absence de données d'absence fiables dans les jeux de données d'occurrences. Dans ce travail, nous développons un modèle de prévision de probabilité de présence à partir de données d'occurrences issues du Global Biodiversity Information Facility (GBIF), couvrant 24 espèces observées en Atlantique et en Méditerranée entre 2000 et 2025 provenant d'activités de pêche et de plongée. Les occurrences sont associées à des variables environnementales extraites du Copernicus Marine Environment Monitoring Service (CMEMS), décrivant les conditions thermohalines, hydrodynamiques, trophiques et saisonnières.

L'absence d'observations ne pouvant être interprétée comme une absence certaine, des pseudo-absences sont générées pour chaque espèce à partir d'une représentation de la niche écologique réalisée. Une analyse en composantes principales est réalisée par espèce et la niche réalisée est caractérisée dans l'espace de l'ACP. Des points simulés par loi normale multivariée sont ensuite filtrés à l'aide d'un seuil basé sur la distance de Mahalanobis (conservant 75 % des occurrences les plus proches du centre), afin de produire des pseudo-absences statistiquement incompatibles avec la niche observée et respectant des bornes environnementales plausibles.

Un modèle de classification par Random Forest est ensuite entraîné par espèce et évalué à l'aide de la courbe ROC et de l'AUC. Les performances obtenues sont globalement élevées pour la majorité des espèces, tandis que les espèces très peu représentées présentent des performances plus variables. Enfin, une application interactive sous Streamlit permet de visualiser ponctuellement et spatialement les probabilités de présence sous forme de cartes, facilitant l'usage opérationnel des résultats pour la plongée, la pêche et l'aide à l'échantillonnage scientifique.

2 Introduction

Prévoir la présence d'espèces marines reste un défi autant théorique que pratique. En effet, en plus de contraintes écologiques classiques telles que la mobilité des organismes et la variation spatio-temporelle des conditions environnementales. La prévision de présence spécifique fait aussi face à une contrainte conceptuelle : le manque de données d'absence fiables. En effet, prouver qu'un organisme n'est pas présent est impossible théoriquement. Ainsi, la construction d'un jeu de données nécessite la génération de pseudo-absences. La méthode utilisée pour cela constitue toujours une hypothèse qui peut fortement influencer la performance des modèles de prévision. Malgré ces limitations, prévoir la présence d'organismes marins répond aux besoins d'un certain nombre de secteurs tels que la pêche professionnel et le tourisme sub-aquatique avec la plongée sous-marine, l'apnée ou la pêche de loisirs. Ces modèles représentent aussi des outils intéressants pour la recherche scientifique notamment pour orienter les zones d'échantillonnage, analyser la distribution spatiale des espèces ou étudier l'évolution des aires de répartition des espèces marines dans un contexte de changement global.

Dans ce papier on se propose donc de développer un modèle de prévision de probabilité de présence d'espèces marines à partir de de données d'occurrences. Pour cela on mobilise des méthodes d'apprentissage supervisé, entraînées sur des données environnementales afin d'estimer spatialement la probabilité de présence des espèces étudiées.

3 Matériels et méthodes

3.1 Données d'occurrence

Les données d'occurrence représentent des observations d'individus lors d'activités liées à la pêche ou à la plongée. Plus précisément, ces données proviennent d'observateurs de pêche et de plongeurs loisirs. Chaque observation est associé à une latitude, une longitude, une date et une heure. L'ensemble de ces informations est intégré dans une base de données gratuite et libre d'accès : le Global Biodiversity Information Facility (GBIF).

Les espèces sélectionnées pour cette étude sont un groupe d'organismes non-ubiquistes, ciblé par la pêche professionnel, ciblé par la pêche de loisir et apprécié par les plongeurs, aussi bien en Atlantique qu'en Méditerranée. Le tableau 1 présente pour chaque espèce retenu, le nombre d'occurrence disponible, le pourcentage de provenance de la pêche et de la plongée ainsi que la zone géographique concernée.
Le tableau 1 met en évidence un fort déséquilibre inter-spécifique en terme de nombre d'observations. A titre d'exemple, le chinchard, le maquereau et la sole comptent respectivement environ 62 000, 25 000 et 17 000 observations. Tandis que le mulet, la liche et le tassergal ne présentent respectivement que 80, 23 et 27 occurrences. Ces déséquilibres reflètent à la fois l'abondance des espèces et l'intérêt qu'elles suscitent chez les pêcheurs ou les plongeurs. Par exemple, la liche est une espèce relativement rare donc peu observée. Alors que le mulet, malgré son abondance est peu ciblé par la pêche et peu recherché par les plongeurs, son nombre est donc largement sous-estimé. Au contraire, le chinchard est un poisson qui vit en banc et qui est ciblé fortement par la pêche, ce qui conduit à un nombre d'occurrence élevé.
Ces déséquilibres influencent directement les performances des modèles de prévision, en effet une classe mieux représentée permet des prédictions plus robustes, du fait d'une meilleure couverture des conditions environnementales associées à la présence ou l'absence de l'espèce. Afin de tenir compte de ce biais d'échantillonnage, un indice de rareté est défini pour chaque espèce selon :

$$N_i = w_{peche} N_{peche} + N_{plongee}$$
$$D_i = \frac{log(N_i)}{max(log(N_j))}$$

ou i désigne l'espèce considérée, j l'ensemble des espèces du jeu de données et $w_{peche} \in [0, 1]$, un poids associé aux observations issues de la pêche. L'indice D_i est normalisé de 0 à 1 ou 0 correspond à la représentation minimum (rareté élevé). Au contraire 1 correspond au maximum de représentation (rareté minimale). On note également que les observations issues de la pêche professionnelle sont pénalisées afin de corriger un biais d'échantillonnage. En effet la pêche cible préférentiellement certaines espèces, ce qui conduit à une augmentation artificielle du nombre d'occurrences des ces espèces. A l'inverse la plongée est considéré comme une activité d'observation plus généraliste car elle n'agit pas sur les individus. Ainsi une espèce fortement ciblé par la pêche peut paraître moins rare qu'une espèce non-ciblé alors que cette différence résulte principalement de l'intensité de l'effort d'observation. La pénalisation des observations de pêche vise donc à atténuer ce biais pour fournir une estimation plus cohérente de la rareté relative des espèces.

3.2 Données environnementales

À partir de la position géographique et de la date associées à chaque observation du jeu de données d'occurrence, des variables environnementales ont été extraites à l'aide du Copernicus Marine Environment Monitoring Services (CMEMS). Ces données proviennent de modèles de réanalyses, fournissant une description cohérente et spatialement continue des conditions physico-chimiques du milieu marin à la date et au lieu considérés.
Les variables sélectionnées répondent à 2 critères principaux : être facilement interprétable par les acteurs susceptibles d'utiliser le modèle de prévision et être écologiquement pertinentes pour expliquer la présence ou l'absence d'organismes marins. Les variables retenues sont ainsi : le jour de l'observation, la température de l'eau, la salinité, la bathymétrie, la hauteur significative des vagues, la composante de courant Est, la composante de courant Nord, la profondeur de la couche de mélange, la production primaire nette et la concentration en oxygène dissous dans l'eau. Ces variables décrivent les conditions thermohalines, hydrodynamique et trophique du milieu ainsi que leur variabilité saisonnière.
A partir des composantes de courant Est (u) et Nord (v), la vitesse du courant et sa direction d'origine (la convention

Espèce	Nombre d'observations	Pêche	Plongée	Lieu
Mugil sp. (Mulet)	80	0%	100%	Méditerranée/Atlantique
Symphodus tinca (Crénilabre Paon)	2274	0%	100%	Méditerranée/Atlantique
Sparus aurata (Daurade royale)	1439	3.9%	96.1%	Méditerranée/Atlantique
Dentex dentex (Denti)	1421	0%	100%	Méditerranée
Sphyraena sphyraena (Barracuda)	107	7.1%	92.9%	Méditerranée
Scomber scombrus (Maquereau)	25327	63.5%	36.5%	Méditerranée/Atlantique
Pomatomus saltatrix (Tassergal)	27	44.1%	55.9%	Méditerranée
Lichia amia (Liche amie)	23	0%	100%	Méditerranée
Seriola dumerili (Sériole courronée)	196	0%	100%	Méditerranée
Labrus viridis (Labre vert)	509	0%	100%	Méditerranée/Atlantique
Phycis phycis (Mostelle)	1250	0%	100%	Méditerranée/Atlantique
Diplodus puntazzo (Sar becofino)	1306	0%	100%	Méditerranée/Atlantique
Diplodus sargus (Sar commun)	3401	0%	100%	Méditerranée/Atlantique
Diplodus vulgaris (Sar à tête noir)	4225	0%	100%	Méditerranée/Atlantique
Diplodus cervinus (Sar tambour)	759	0%	100%	Méditerranée/Atlantique
Lithognathus mormyrus (Marbré)	494	8.5%	91.5%	Méditerranée/Atlantique
Sarpa salpa (Saupe)	3333	0%	100%	Méditerranée/Atlantique
Sarda sarda (Bonite)	155	32.6%	67.4%	Méditerranée/Atlantique
Scorpaena scrofa (Rascasse rouge)	2984	6%	94%	Méditerranée/Atlantique
Pagellus erythrinus (Pageot commun)	1926	47.5%	52.5%	Méditerranée/Atlantique
Solea solea (Sole commune)	17170	94.3%	5.7%	Méditerranée/Atlantique
Dicentrarchus labrax (Loup/Bar)	2766	41.1%	58.9%	Méditerranée/Atlantique
Trachurus trachurus (Chinchard)	61864	87.4%	12.6%	Méditerranée/Atlantique
Mullus surmuletus (Rouget de roche)	9130	57.8%	42.2%	Méditerranée/Atlantique

TABLE 1 – Résumé des données d'occurrences spécifique en Méditerranée et en Atlantique, en France et en Espagne, entre 2000 et 2025

météorologique) sont calculées selon :

$$speed = \sqrt{u^2 + v^2}$$
$$direction = (270 - \arctan 2(v, u)) \bmod 360$$

On remplace alors la composante de courant Nord et Est par la vitesse et la direction du courant afin de fournir des variables directement interprétables des conditions hydrodynamiques. Au total, chaque observation est ainsi décrite par dix variables environnementales caractérisant les conditions dans lesquels les espèces étudiées sont observées.

3.3 Pseudo-absence et niche écologique

3.3.1 Problème de l'absence

Pour établir un modèle de prévision de présence, il est nécessaire d'entraîner ce modèle sur des données de présence et d'absence. Cela lui permet d'établir clairement une dichotomie entre les conditions d'existence et d'inexistance de l'espèce. Les données d'occurrences représentent parfaitement des données de présence, puisque cela correspond à l'observation directe d'un organisme, néanmoins les données d'absence sont plus controversées. En effet, une absence certaine n'existe pas. Il est impossible de dire qu'une espèce n'est pas là parce qu'elle n'a pas été observée. C'est une limite théorique et fondamentale de l'observation des organismes en écologie. Il faut donc trouver une méthode pour remplacer ces absences, qui soit réalisable dans le cadre de cette étude.

3.3.2 Niche écologique

Afin d'avoir un cadre pour répondre à cette problématique, il est nécessaire de revenir à une théorie écologique : la niche de Hutchinson (1957). Cela correspond à l'hyper-volume dans l'espace des paramètres physico-chimiques dans lequel une espèce peut évoluer. Autrement dit, c'est l'ensemble des conditions environnementales compatibles avec la survie de l'espèce. En dehors de cet hyper-volume, l'espèce n'est plus adaptée à son environnement donc elle s'éteint. Dans la théorie de Hutchinson, deux notions complémentaires sont distinguées : la niche fondamentale et la niche réalisée. La niche fondamentale correspond à l'hyper-volume détaillé plus tôt. Tandis que la niche réalisée désigne la partie effectivement occupée de cet espace, contrainte par les interactions biotiques. En effet, lorsque les niches fondamentales de plusieurs espèces se chevauchent, des phénomènes de compétition interspécifique apparaissent. Ces interactions peuvent conduire soit à l'exclusion d'une espèce, soit à une partition de la section communes des niches écologiques entre espèces concurrentes. Dans les deux cas, l'espace réellement occupé par une espèce est restreint par rapport à sa niche fondamentale. La niche réalisée est ainsi, par définition, incluse dans la niche fondamentale et généralement de dimension plus réduite.

3.3.3 Génération des niches écologiques

À partir du tableau de données initial, le jour de l'année (compris entre 1 et 365) et la direction du courant (azimut compris entre O° et 360°) sont transformés en variables périodiques selon les relations suivantes :

$$\theta = \frac{\pi}{180}\,\text{direction} \qquad \text{jour}_{\text{rad}} = \frac{2\pi}{365}\,\text{jour}$$
$$\sin_{\text{direction}} = \sin(\theta) \qquad \sin_{\text{jour}} = \sin(\text{jour}_{\text{rad}})$$
$$\cos_{\text{direction}} = \cos(\theta) \qquad \cos_{\text{jour}} = \cos(\text{jour}_{\text{rad}})$$

Cette transformation permet de prendre en compte la nature périodique de ces variables. Par exemple, les jours 1 et 365 sont très proches dans le cycle annuel, alors qu'ils apparaissent éloignés dans une représentation linéaire classique. L'utilisation des composantes sinus et cosinus permet ainsi de préserver la continuité temporelle et directionnelle. Dans le tableau de données final, la variable *jour* est remplacée par $\sin_{\text{jour}}$ et $\cos_{\text{jour}}$, et la variable *direction* par $\sin_{\text{direction}}$ et $\cos_{\text{direction}}$.

Le jeu de données est ensuite scindé en 24 sous-ensembles, un pour chaque espèce étudiée. Une analyse en composantes principales (ACP) est réalisée indépendamment sur chacun des sous-ensembles, après suppression des variables de longitude et de latitude. Dans l'espace défini par l'ACP, il est alors possible de construire, pour chaque espèce, un hyper-volume représentant sa niche écologique réalisée.

La construction de ces hyper-volumes dans l'espace de l'ACP, plutôt que directement dans l'espace des variables environnementales permet : de pondérer implicitement les variables, de limiter la redondance liée aux corrélations entre variables et de hiérarchiser les gradients environnementaux dominants caractérisant les conditions de présence de chaque espèce.

3.3.4 Génération des pseudo-absences

Dans la mesure ou une absence certaine est impossible en écologie, les points construits dans cette étude sont désignés comme des pseudo-absences. Le principe général consiste à identifier des conditions environnementales statistiquement incompatibles avec la niche réalisée de chaque espèce.

Dans un premier temps, des bornes minimales et maximales sont fixées pour chacune des variables environnementales. Ces bornes correspondent soit aux valeurs observables dans l'environnement considéré, soit à des limites physiquement plausibles. Les valeurs retenues pour chaque variable sont présentées dans le tableau 2.

Variable	Minimum	Maximum
Température (°C)	0	35
Salinité (PSU)	0	42
Bathymétrie (m)	0	5000
Hauteur de vague (m)	0	20
Vitesse du courant (m/s)	0	1
Profondeur de la thermocline (m)	0	500
Production primaire nette ($mg/m^3/jour$)	0	500
Concentration en O_2 ($mmol/m^3$)	100	400
Jour sin (sans unité)	-1	1
Jour cos (sans unité)	-1	1
Direction sin (sans unité)	-1	1
Direction cos (sans unité)	-1	1

TABLE 2 – Valeur minimale et maximale choisi pour chaque variable

La génération des pseudo-absences repose sur l'algorithme suivant :

Etape 1 : Génération du centre de gravité
Dans l'espace de l'ACP ou les niches écologiques ont été construites, on génère le centre de gravité du nuage de point correspondant aux occurrences de l'espèce considéré tel que :

$$\mathbf{g} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

ou $\mathbf{z}_i$ représente le vecteur des coordonnées de l'observation i dans l'espace de l'ACP.

Etape 2 : Calcul de la distance des points au centre de gravité
La distance de Mahalanobis de chaque observation au centre de gravité est ensuite calculé tel quel :

$$\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i - \mathbf{g})(\mathbf{z}_i - \mathbf{g})^\top$$
$$d_M(\mathbf{z}_i, \mathbf{g}) = \sqrt{(\mathbf{z}_i - \mathbf{g})^\top \mathbf{\Sigma}^{-1}(\mathbf{z}_i - \mathbf{g})}$$

ou $\mathbf{\Sigma}$ est la matrice de variance-covariance du jeu de données dans l'espace de l'ACP. La distance de Mahalanobis permet de prendre en compte la variance des axes, les corrélations entre les variables et la forme réelle du nuage de point.

Etape 3 : Conservation de la distance de 75% de la densité
On garde en mémoire la distance de Mahalanobis correspondant à 75% des observations les plus proches du centre de gravité. Le choix de ce seuil vise à définir une frontière souple de la niche écologique réalisée. Cette approche évite l'utilisation d'une frontière catégorique entre présence et absence et introduit une zone d'incertitude résultant de bruit écologique, améliorant ainsi la robustesse et la stabilité du modèle de prévision.

Etape 4 : Génération des pseudo-absences

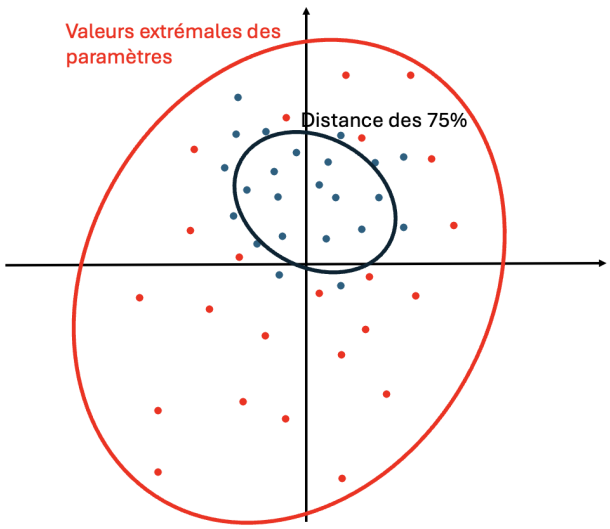

FIGURE 1 – Schématisation de la génération de pseudo-absences. Ce graphe est une simplification en 2 dimensions, représenté dans l'espace de l'ACP. Les points bleus sont des points de présence et les points rouges de pseudo-absence. Le cercle noir représente la distance de Mahalanobis qui inclue 75% des points, le cercle rouge représente les valeurs extrémales des paramètres.

Des points sont ensuite générés aléatoirement dans l'espace de l'ACP à partir d'une loi normale multivariée tel que :

$$\mathbf{Z_{sim}} \sim \mathcal{N}_d(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma})$$

où $\boldsymbol{\mu}_Z$ représente le vecteur des moyennes des variables dans l'espace de l'ACP. On recalcule alors la distance de Mahalanobis de chaque point simulé, parmi ces points, on ne conserve que ceux dont la distance est strictement supérieur au seuil des 75%. Ce sont les pseudo-absences.

Etape 5 : Retour dans l'espace des variables environnementales
On utilise alors la transformation inverse à l'ACP pour projeter les points simulés dans l'espace des variables environnementales tel que :

$$\mathbf{X_{sim}} = \mathbf{Z_{sim}} A^T \boldsymbol{\Sigma_X} + \boldsymbol{\mu}_X$$

où $\mathbf{A}$ est la matrice des vecteurs propres de l'ACP, $\boldsymbol{\Sigma}_X$ la matrice de variance-covariance des variables environnementales et $\boldsymbol{\mu}_X$ le vecteur des moyennes associées.

Etape 6 : Suppression des valeurs abérantes
Une fois dans l'espace des variables environnementales, les points simulés présentant des valeurs inférieures ou supérieures aux bornes définies dans le tableau 2 sont éliminés. Les étapes 4 à 6 sont répétées jusqu'à l'obtention du nombre souhaité de pseudo-absences.
La figure 1 présente une schématisation de cette méthode dans un espace réduit à 2 dimensions.

3.4 Modélisation

3.4.1 Focus sur les données

Dans notre étude, sont générées autant de pseudo-absences que de présences. Les jeux de données sont donc équilibrés artificiellement afin d'éviter un apprentissage biaisé par une sur-représentation des absences ou des

présences.
Les variables environnementales sont utilisées telles quelles, sans normalisation préalable, les modèles de type Random Forest n'étant pas sensibles à l'échelle des variables. Chaque jeu de données est ensuite scindé en un ensemble d'entraînement et un ensemble de test indépendants comprenant respectivement 90% et 10% des données. Cela permet d'évaluer les performances du modèle sur des données qu'il n'a jamais vu et donc de tester sa généralisation. Cette approche permet d'entraîner un modèle spécifique par espèce, chaque modèle apprenant les relations propres entre conditions environnementales et probabilité de présence.

3.4.2 Random forest

Les modèles de prévision sont basés sur des forêts aléatoires de classification (Random Forest). Cette méthode repose sur l'agrégation d'un grand nombre d'arbres de décision construits à partir de sous-échantillons aléatoires des données et des variables explicatives.
Les Random Forest présentent plusieurs avantages dans le cadre de cette étude : (i) elles permettent de modéliser des relations non linéaires complexes entre variables environnementales et présence d'espèces, (ii) elles sont robustes au bruit et aux corrélations entre variables, (iii) elles offrent de bonnes performances prédictives sans nécessiter d'hypothèses fortes sur la distribution des données.
Chaque modèle est entraîné à prédire une probabilité de présence comprise entre 0 et 1. Les performances sont évaluées sur les jeux de données de test à l'aide de métriques adaptées aux problèmes de classification binaire, telles que la courbe ROC et l'AUC.

3.5 Interface utilisateur

Afin de rendre les résultats exploitables par des utilisateurs non spécialistes, une interface interactive a été développée à l'aide de l'outil Streamlit. Cette application permet de visualiser spatialement ou ponctuellement les probabilités de présence prédites par les modèles, sous forme de cartes de densité par espèce ou de tableau de probabilité de chaque espèce, autour d'un point donnée et pour une date choisie.
L'interface offre la possibilité de sélectionner, une zone géographique, une date et différents paramètres environnementaux. Elle constitue ainsi un outil d'exploration et d'aide à la décision, permettant de relier directement les résultats du modèle à des usages concrets, tels que la plongée, la pêche ou l'exploration scientifique.

4 Résultats

4.1 Performances du modèle par espèces

Les performances des modèles varient selon les espèces considérées tel que présentées dans le tableau 3. Globalement on observe que les espèces disposant d'un grand nombre d'occurrences présentent des meilleures performances prédictives, traduisant une meilleure couverture des conditions environnementales associées à leur présence. À l'inverse, les espèces rares ou faiblement observées présentent des performances plus variables, en lien avec une incertitude plus élevée sur leur niche écologique réalisée. On considérera que les performances des modèles du tassergal (0.875) et de la liche (0.5) ne sont pas assez bon, ils sont donc retiré de l'étude.
Globalement, les modèles obtiennent des scores satisfaisants, indiquant une bonne capacité à discriminer les conditions favorables et défavorables à la présence des espèces étudiées.

Espèce	Nombre d'observations	Performance AUC
Mugil sp. (Mulet)	80	0.984
Symphodus tinca (Crénilabre Paon)	2274	0.999
Sparus aurata (Daurade royale)	1439	0.999
Dentex dentex (Denti)	1421	0.999
Sphyraena sphyraena (Barracuda)	107	0.941
Scomber scombrus (Maquereau)	25327	0.999
Pomatomus saltatrix (Tassergal)	27	0.875
Lichia amia (Liche amie)	23	0.5
Seriola dumerili (Sériole courronée)	196	1
Labrus viridis (Labre vert)	509	1
Phycis phycis (Mostelle)	1250	0.999
Diplodus puntazzo (Sar becofino)	1306	0.999
Diplodus sargus (Sar commun)	3401	0.999
Diplodus vulgaris (Sar à tête noir)	4225	0.998
Diplodus cervinus (Sar tambour)	759	0.997
Lithognathus mormyrus (Marbré)	494	0.997
Sarpa salpa (Saupe)	3333	0.999
Sarda sarda (Bonite)	155	1
Scorpaena scrofa (Rascasse rouge)	2984	0.999
Pagellus erythrinus (Pageot commun)	1926	0.999
Solea solea (Sole commune)	17170	0.999
Dicentrarchus labrax (Loup/Bar)	2766	0.999
Trachurus trachurus (Chinchard)	61864	0.999
Mullus surmuletus (Rouget de roche)	9130	0.999

TABLE 3 – Performance de l'AUC par modèle de chaque espèces avec le nombre d'observation de chaque espèce

4.2 Probabilités ponctuelles

Les probabilités de présence prédites ponctuellement montrent une forte dépendance aux conditions environnementales locales. Certaines espèces présentent des probabilités élevées uniquement dans des conditions bien spécifiques, tandis que d'autres affichent une réponse plus large, traduisant une niche écologique plus étendue. Ces résultats sont cohérents avec les connaissances écologiques générales sur les espèces étudiées et illustrent la capacité du modèle à capturer des gradients environnementaux complexes.

4.3 Cartes de probabilité de présence

Les cartes de probabilité de présence mettent en évidence des structures spatiales cohérentes, avec des zones de forte probabilité correspondant à des habitats connus comme favorables. La pénalisation des biais d'échantillonnage liés à la pêche conduit à des cartes plus prudentes pour certaines espèces fortement ciblées, sans modifier radicalement la hiérarchie spatiale des zones favorables.
Ces cartes constituent un support visuel pertinent pour l'interprétation des résultats et pour leur utilisation opérationnelle. Néanmoins elles sont contraintes par l'échelle spatiale des données fourni par le CMEMS qui est parfois grande et provoquent des manques de précision sur les cartes de densité de probabilité.

4.4 Vérification et cohérence

Il est vérifier le modèle en le confrontant à la réalité pour les mêmes raison qu'il n'est pas possible d'obtenir des absences certaines. Néanmoins les résultats obtenus sont quand même confrontés à des connaissances empiriques issues de l'observation de terrain. Une dizaine de sortie ont été réalisé dans le cadre de sorties de chasse

sous marine sur les côtes marseillaises. On observe généralement une cohérence entre les observations de terrain et les prédictions du modèle. De plus, il est intéressant de voir que si certaines espèces apparaissent très communes à l'échelle locale (notamment pour les plongeurs), leur probabilité de présence modérée à l'échelle spatiale est cohérente avec une distribution écologiquement contrainte autour des côtes, par exemple pour les espèces de sars. Le modèle est donc capable de distinguer l'abondance locale de la distribution spatiale globale, ce qui constitue un point fort de l'approche proposée.

5 Discussion

5.1 Mots sur la cohérence du modèle

Dans l'ensemble, le modèle présente une cohérence écologique satisfaisante. Les relations apprises entre variables environnementales et présence des espèces sont compatibles avec les connaissances existantes et les prédictions spatiales évitent les extrapolations excessives.
La prise en compte explicite de l'incertitude, notamment via la génération de pseudo-absences et l'utilisation de frontières souples de niche, contribue à renforcer la robustesse des résultats.

5.2 Générations des absences

Néanmoins, la génération des pseudo-absences repose sur des hypothèses fortes, notamment : il n'y a pas d'organismes dans des conditions hors de leur niche écologique, la niche écologique est stable dans le temps, les données d'occurrences représentent correctement la niche écologique et la forme de la niche écologique réalisée dans l'espace de l'ACP est approximativement gaussienne en plus des hypothèses plus classique comme l'indépendance ou l'équité des poids de chaque observation. Bien que ces hypothèses constituent une simplification, elles permettent de construire un cadre méthodologique cohérent et reproductible.
De plus, le choix d'un seuil à 75% comme frontière de la niche réalisée représente un compromis entre conservatisme et pouvoir discriminant, évitant une séparation trop stricte entre présence et absence.

5.3 Biais d'échantillonnages

Les données d'occurrences utilisées sont fortement influencées par les pratiques humaines, en particulier la pêche et la plongée. La pénalisation des observations issues de la pêche et l'indice de rareté d'une espèce permet de limiter l'impact de ces biais, mais ne les supprime pas totalement. De plus, l'outil développé dans le cadre de cette étude ne prédit pas une prévision d'occurrence mais de présence seulement. En effet l'observabilité d'une espèce est influencé par d'autre facteurs clés tel que la visibilité de l'eau, le déplacement local des organismes, l'activité humaine et même la performance de l'utilisateur dans la pratique de son activité (aquacité, pêche, plongée). Ces biais doivent être pris en compte dans l'interprétation des résultats, notamment pour les espèces très ciblées par la pêche ou très peu observées.

5.4 Perspectives

Plusieurs pistes d'amélioration peuvent être envisagées, telles que l'intégration de variables d'habitat plus fines, l'utilisation de modèles dynamiques ou l'évaluation de la sensibilité des résultats aux choix méthodologiques (seuil de densité, nombre de pseudo-absences). Une extension à d'autres zones géographiques ou à d'autres groupes taxonomiques constitue également une perspective intéressante.

6 Conclusion

Ce travail propose une approche permettant de prévoir la probabilité de présence d'espèces marines à partir de données d'occurrences, en combinant théorie de la niche écologique de Hutchinson, méthodes statistiques (ACP) et apprentissage automatique par Random Forest. Malgré les limites inhérentes aux données disponibles, les résultats semblent écologiquement plausibles et exploitables.

L'ensemble de la méthodologie développée constitue une base pour des applications opérationnelles pour la pêche, la plongée ou les observations scientifiques et pour des développements futurs dans le domaine de la modélisation de la biodiversité marine.