Automatisation de la cartographie d'habitats de récifs coralliens à l'aide d'algorithmes d'apprentissage statistique

Paul Aimé Latsouck Faye, Elodie Brunel, T. Claverie, Solym Manou-Abi, Sophie Dabo-Niang

Automatisation de la cartographie d'habitats de récifs coralliens à l'aide d'algorithmes d'apprentissage statistique

Paul Aimé Latsouck Faye [1], Elodie Brunel-Piccinini [1], Solym Manou-Abi [1,2], Thomas Claverie [2,3] et Sophie Dabo Niang [4]

[1] *IMAG, Université de Montpellier, CNRS, Montpellier, France,*
paul-aime-latsouck.faye@umontpellier.fr, elodie.brunel-piccinini@umontpellier.fr
[2] *CUFR de Mayotte, France,*
solym.manou-abi@univ-mayotte.fr
[3] *MARBEC, Université de Montpellier, CNRS, IRD, Ifremer, France,*
thomas.claverie@univ-mayotte.fr
[4] *Paul Painlevé, CNRS, Université de Lille, INRIA-MODAL, France,*
sophie.dabo@univ-lille.fr

Résumé. Les techniques de télédétection sont les outils les plus couramment utilisés pour la cartographie d'habitats en milieu marin. Basées sur des images satellitaires ou aériennes utilisant divers capteurs, elles sont souvent limitées par des facteurs physiques tels que la profondeur, la turbidité de l'eau, les conditions météorologiques, mais aussi par la nécessité d'une validation manuelle par un expert avec divers relevés de terrain. Elles requièrent également beaucoup de temps et ne sont généralement pas reproductibles car pouvant mener à des résultats différents d'un expert à un autre. Dans ce travail, nous automatisons une approche statistique pour cartographier la structure géormorphologique d'habitats de récifs coralliens. Par une étude de simulation basée sur une cartographie d'experts, nous comparons les résultats obtenus avec deux méthodes d'échantillonnage (l'une spatialement homogène et l'autre dépendante de la complexité), pour différentes tailles d'échantillon et pour différentes résolutions de l'habitat. Nous montrons que la méthodologie proposée permet de reconstruire la carte des experts de manière satisfaisante et mettons en évidence la méthode d'échantillonnage, la taille d'échantillon et la résolution d'habitats optimales.

Mots-clés. Géomorphologie, Modélisation prédictive, Cartographie, Apprentissage automatique, Classification supervisée, Interpolation spatiale, Récifs de corail.

Abstract. Remote sensing techniques are the most commonly used tools for habitat mapping in the marine environment. Based on satellite or aerial images using various sensors, they are often limited by physical factors such as depth, water turbidity, weather conditions, but also by the need for manual validation by an expert with various ground-truthing. They are also time consuming and generally not reproducible as they can lead to different results from one expert to another. In this work, we automate a statistical approach to map the geormorphological structure of coral reef habitats. Through a simulation study based on expert mapping performed in a previous study, we compare the results obtained with two sampling methods (one spatially homogeneous and the other complexity-dependent), for different sample sizes and for different habitat resolutions. We show that the proposed methodology successfully

reconstructs the expert map and highlight the optimal sampling method, sample size and habitat resolution.

Keywords. Geomorphology, Predictive modeling, Mapping, Machine learning, Supervised classification, Spatial interpolation, Coral reefs.

1 Introduction

Une typologie d'habitats fait intervenir plusieurs niveaux de description hiérarchisés incluant des informations sur la géomorphologie, l'architecture, la couverture benthique et les groupes taxonomiques. D'après Franklin (1995), le développement des techniques de télédétection aérienne et satellitaire au cours des dernières décennies a facilité l'acquisition de données pour les études écologiques à grande échelle. Cela a favorisé l'utilisation de la bathymétrie pour segmenter des zones en communautés reflétant des caractéristiques biologiques distinctes.
La typologie *Millennium Coral Reef Mapping Project* (MCRMP) initiée par l'Institute for Marine Remote Sensing - University of South Florida (IMaRS/USF) et reprise par l'Institut de Recherche pour le Développement (IRD) a permis l'élaboration d'une structure hiérarchique à plusieurs niveaux par Andréfouët et al. (2006). Cette typologie fournit une description géomorphologique encore très utilisée dans de nombreuses études car comme le montrent Andrefouët et Diberg (2006), elle permet de comparer sur une base homogène thématiquement riche, l'ensemble des sites mondiaux.
La cartographie d'experts que nous utilisons dans cette étude s'est intéressée au niveau 4 (N4) de cette typologie qui décrit la structure géomorphologique en distinguant des unités récifales telles que les pentes, les platiers, les passes, les terrasses, les lagons, les chenaux, etc. Dans la section 2, nous présentons les données bathymétriques utilisée dans notre étude. Nous décrivons dans la section 3.1 comment la surface ou zone d'étude est représentée sous forme de Modèles Numériques de Terrain (MNT) de différentes résolutions avec la bathymétrie. Des prédicteurs spatiaux ou attributs de terrain sont ensuite quantifiés sur la base des interactions entre les cellules voisines du MNT, permettant ainsi de créer une grille multi-couches d'attributs de terrain sur toute la surface. Dans la section 3.2, nous décrivons les méthodes d'échantillonnage utilisées pour générer des points de vérification terrain. Ces points sont par la suite labellisés à partir de la carte des experts. Leurs attributs de terrain quant à eux sont déterminés à partir de la grille des prédicteurs spatiaux. Dans la section 3.3, nous présentons un algorithme de classification supervisée basé sur les forêts aléatoires (Random Forest) avec sélection de variables que nous utilisons pour segmenter les données en unités géomorphologiques distinctes. Par prédiction, toutes les cellules du MNT sont labellisées permettant ainsi une cartographie de toute la zone d'étude. Dans la section 4, les résultats obtenus avec les combinaisons de ces différents paramétres (résolutions du MNT, méthode d'échantillonnage et taille d'échantillon) sont ensuite analysées afin de déterminer la meilleure méthode pour cartographier la géomorphologie de niveau 4.

2 Zone d'étude et données

Geyser est un banc récifal situé entre Mayotte et les îles Glorieuses dans le canal du Mozambique et couvrant environ 268 km^2 de surface. Il s'étend sur 21km d'ouest en est entre 46°42'E et 46°62'E et sur 20km de nord en sud entre 12°24'S et 12°37'S. Les données bathymétriques (données géo-référencées de la profondeur du fond marin) utilisées dans cette étude sont acquises par LIDAR et proviennent du Service Hydrographique et Océanographique de la Marine (SHOM). La profondeur mesurée (environ 48 millions observations) varie entre - 50 à 4 mètres, les valeurs positives correspondant aux zones émergées. Certaines zones du banc ont été peu échantillonnées, en raison de l'absence de retour de signal du capteur à partir d'une certaine profondeur.
Lors une précédente campagne de cartographie menée par Roos et al. (2017) sur le banc du Geyser à partir d'images hyperspectrales, les grandes unités géomorphologiques identifiées ont été délimitées et vectorisées à l'aide d'un Système d'Information Géographique (SIG) et de logiciels de traitement d'images. Certaines typologies géomorphologiques telles que la pente externe, la pente interne n'étant pas facilement identifiables, la bathymétrie a ensuite été utilisée pour mettre en évidence la topographie et les zones de relief. Afin de valider cette pré-cartographie et la compléter, une vérification terrain a été réalisée notamment aux zones d'incertitudes. Sur la Figure 1, nous pouvons voir que 11 habitats ont pu être cartographiés. Cette carte d'experts servira de base d'échantillonnage dans notre étude.

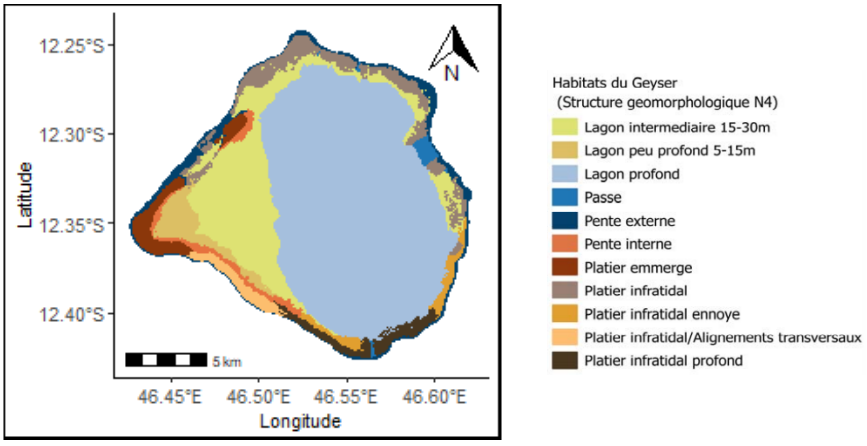

FIGURE 1 – Cartographie de la structure géomorphologique de niveau 4 du banc du Geyser réalisée par Roos et al. (2017).

3 Méthodologie

3.1 Création du MNT et calcul d'attributs de terrain

Le jeu de données bathymétriques brut contient environ 48.10^6 points. Il est sous échantillonné en utilisant une grille régulière de cellules de taille 70 m × 70 m. Dans chaque cellule, nous

avons retenu les points distants d'au moins 5 m. Ce sous échantillonnage a permis de réduire les données de façon homogène en 8.10^6 points environ. A partir de ces données, nous générons un MNT, c'est-à-dire une représentation de la surface sous forme d'une grille régulière de mesures de profondeur d'une résolution donnée. Chaque cellule de cette grille représentée par les coordonnées géographiques (latitude et longitude) de son centre et par la profondeur moyenne des points s'y situant. Trois MNT de résolutions (50 m, 25 m, 5 m) sont générés et analysés dans notre étude.
Aux zones peu échantillonnées, la bathymétrie des cellules vides du MNT a été imputée par interpolation spatiale avec un modèle de krigeage ordinaire (Voir Baillargeon (2005) pour plus de détails sur la méthode).
Une analyse bathymétrique consistant à quantifier les prédicteurs spatiaux de la structure géomorphologique est ensuite réalisée sur la base de la grille matricielle des données de profondeur (MNT) et des interactions algébriques entre les cellules comme décrit dans Wilson et al. (2007). Les attributs de terrain calculés sont listés dans la Table 1 qui suit.

Attributs de terrain	Références
Pente et Orientation	
Pente (*Slope*)	Horn (1981)
Orientation (*Aspect*)	Horn (1981)
Variablilité du terrain	
Rugosité (*Roughness*)	Dartnell (2000)
Indice de rugosité du terrain (*TRI*)	Wilson et al., (2007)
Vecteur de mesure de rugosité (*VRM*)	Sappington et al. (2007)
Courbure et position relative	
Courbure de profile (profc)	Wilson et al. (2007)
Courbure tangentielle (*tanc*)	Wilson et al. (2007)
Indice de position bathymétrique (*BPI*)	Wilson et al. (2007)

TABLE 1 – Attributs de terrain dérivés de la bathymétrie.

3.2 Echantillonnage

3.2.1 Echantillonnage spatialement homogène

Cochran (1977) montre qu'en cartographie, échantillonner le plus uniformément possible augmente la précision. Un échantillonnage spatialement homogéne peut être réalisé avec un algorithme de classification selon Hartigan (1975) en minimisant la trace de la variance intra-groupe. Il a été démontré que dans le cas où tous les centres des classes (les emplacements à partir desquels la distance avec les individus est calculée) coïncident avec les centroïdes des classes (moyenne multivariée des cellules attribuées à une classe à une étape donnée de la procédure de classification), minimiser la trace de la variance intra-groupe revient à minimiser la distance moyenne quadratique $\frac{1}{n}\sum_{i=1}^{n} \underset{k}{min}(D_{ik}^2)$ où D_{ik} est la distance entre le i^{ieme} noeud de la grille et le k^{ieme} point d'échantillonnage et n le nombre total de noeuds de la grille.

Notre étude utilise l'algorithme proposé par Walvoort (2010) adapté pour l'échantillonnage de zones à la forme irrégulière. La zone d'étude est discrétisée en une grille régulière de cellules (environs 2500) qui sont ensuite classées par k-means sur la base des coordonnées géographiques (latitude et longitude) de leurs centres. Il en résulte alors une partition de la grille en clusters, les centres de ces clusters servant de points d'échantillonnage (voir exemple sur la Figure 2).

3.2.2 Echantillonnage complexité-dépendant

L'échantillonnage complexité-dépendant évalue la complexité du milieu dans l'espace bi-dimensionnel de la rugosité (*Roughness*) et de la profondeur par un algorithme de clustering dit CLARA (Clustering LARge Application). L'algorithme crée S sous échantillons de taille η à partir de la grille de n cellules du MNT, $\eta < n$. Sur chaque sous échantillon, il applique un algorithme PAM (Partitionnement Autour de Médoïdes). Un médoïde M est une observation qui minimise sa distance avec l'ensemble des autres observations $M = arg\ \underset{i}{min} \sum_{i=1}^{n} d(i,m)$ avec $m = 1, ..., n.$ et d la distance euclidienne (Voir Kaufman et Rousseeuw (1990)). La qualité de chaque partition obtenue avec un vecteur de k médoïdes est quantifiée par $E = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(i, M_k)$. Les points d'échantillonnage retenus correspondent à la configuration (vecteur de médoïdes) qui minimise E (Voir exemple sur la Figure 2).

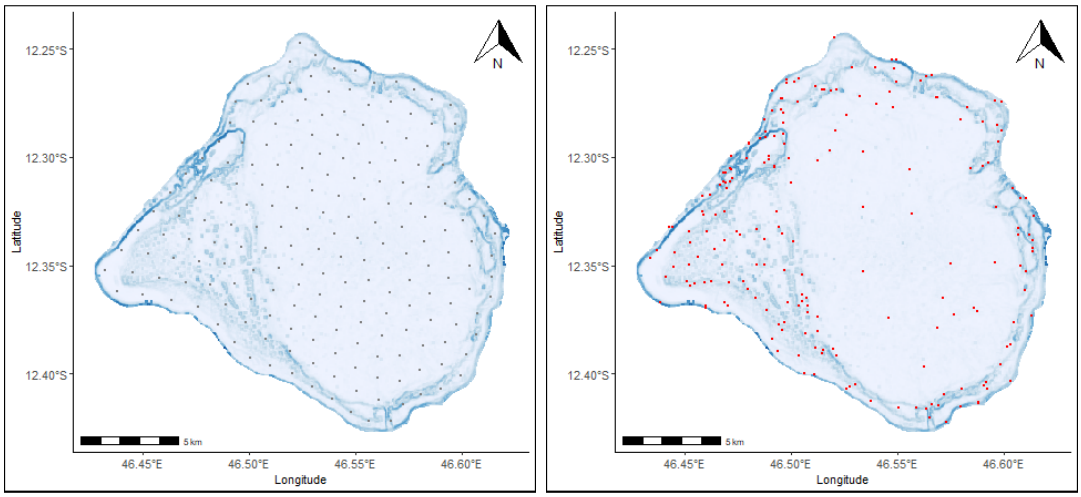

FIGURE 2 – Exemple d'un échantillonnage spatialement homogène (Gauche) et d'un échantillonnage complexité-dépendant (Droite) de 200 points.

3.3 Modélisation avec un algorithme de forêts aléatoires (RF) puis prédiction

Une fois l'emplacement des sites défini, les points échantillonnés sont ensuite projetés sur la carte des experts et sur la grille multi-couches des prédicteurs spatiaux pour déterminer leur label N4 et les attributs de terrain correspondants et serviront d'entrainement au modèle de classification par RF. L'algorithme de RF construit N arbres différents à partir de $a_n \in \{1, ..., n\}$ observations tirées aléatoirement avec remplacement d'un jeu de données de taille n. Pour chaque arbre, chaque noeud est scindé en deux sous-noeuds en maximisant le critère CART sur $T \in \{1, ..., p\}$ directions choisies au hasard (voir Hafezi, Liu et Millward (2018)). La construction des arbres de décision individuels s'arrête lorsque chaque noeud contient moins d'une taille définie $nodesize = \{1, ..., n\}$ d'observations, puis la prédiction est faite par vote majoritaire des arbres. Ambroise et McLachlan (2002) et Stevnik et al. (2004) soulignent que la sélection de variables fait partie du processus de construction de modèles et, en tant que tel, doit être validé de manière externe. Nous utilisons un algorithme de sélection récursive de variables intégré dans une procédure de validation croisée et permettant d'obtenir des mesures de performance qui prennent en compte la variation due à la sélection de variables (Kuhn, M. (2019)). Une sélection implicite de variables qu'on peut visualiser avec l'indice d'importance de Gini, mesure de qualité de la partition à un noeud donné, est utilisée comme indicateur de pertinence des variables tel que proposé par Menze et al. (2009).

4 Résultats

Afin de mesurer les performances statistiques des modèles, nous avons calculé la précision $Accuracy = \frac{TP+TN}{TP+FN+FP+FN}$ et la précision tenant compte d'une répartition déséquilibrée des K labels dans l'échantillonnage $Balanced\ Accuracy = \sum_{k=1}^{K}[(\frac{TP_k}{TP_k+TN_k} + \frac{TN_k}{TN_k+FP_k})/2]/K$ à partir de la matrice de confusion avec TP (vrais positifs) et TN (vrais négatifs) les observations correctement prédites, FP (faux positifs) et FN (faux négatifs) les observations mal prédites. Pour comparer la carte prédite à une résolution donnée à celle des experts pixellisée à la même résolution, nous calculons un ratio de correspondance que nous nommons $Match = \frac{N_i}{N}$ avec N_i le nombre de cellules où le label prédit est identique à celui de la carte des experts et N le nombre total de cellules. Le nombre d'habitats non échantillonnés (*Number of missing labels*) a été également calculé. Nous pouvons visualiser sur les Figure 3 et 4 une synthése des indicateurs de performance ($Match$, $Number\ of\ missing\ labels$, $Accuracy$ et $Balanced\ Accuracy$) calculés pour différrentes tailles d'échantillon (50, 100, 200, 500, 700, 1000) et différentes résolutions de MNT (50 m, 25 m, 5 m) en mètres.

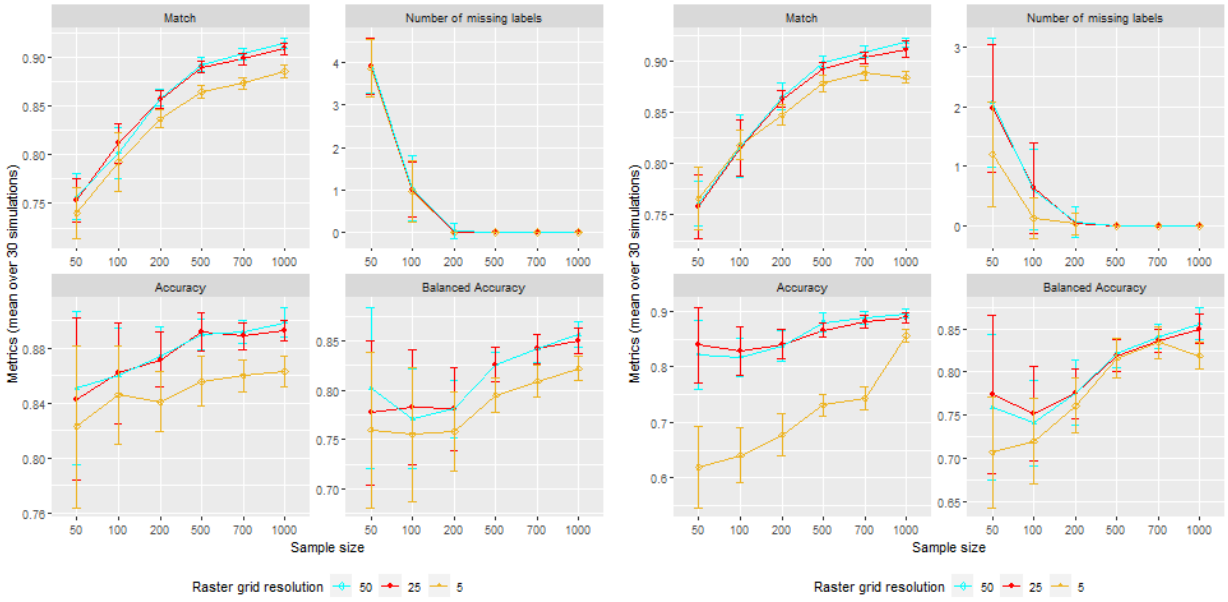

FIGURE 3 – Moyennes (écarts-types) sur 30 simulations des indicateurs de performance obtenues par échantillonnage spatialement homogène (Gauche) et complexité-dépendant (Droite).

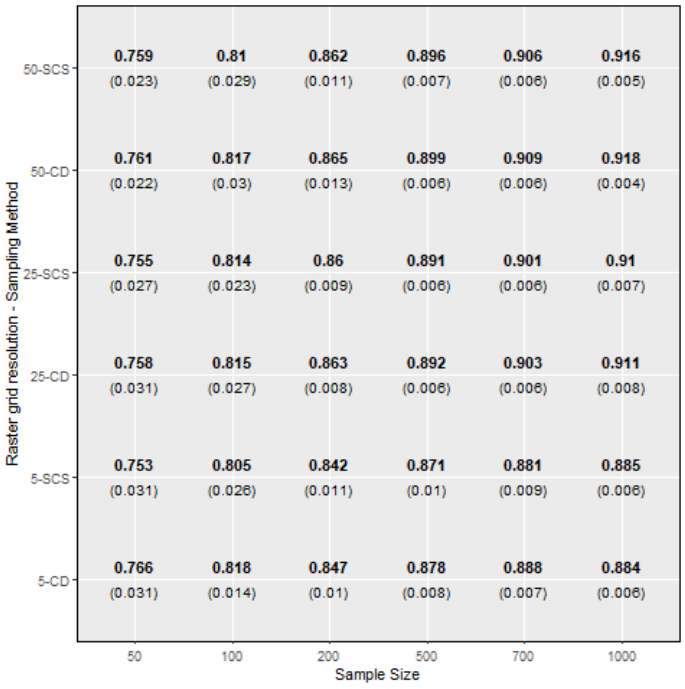

FIGURE 4 – Moyennes (écarts-types) sur 30 simulations de l'indicateur de performance *Match* obtenu par échantillonnage spatialement homogène (SCS) et complexité-dépendant (CD).

En augmentant la taille de l'échantillonnage, nous améliorons nos indicateurs de perfor-

mance par la hausse du *Match*, de l'*Accuracy* et la *Balanced Accuracy* et la baisse de *Number of missing labels* qui devient presque négligeable à partir de 200 points et nulle à 500 points. Au delà de 500 points, nous remarquons que nous ne gagnons plus énormément en termes de prédictions alors que la taille d'échantillon augmente. Par ailleurs, on peut voir plus en détails sur la Figure 3 que les deux méthodes d'échantillonnage donnent sensiblement les mêmes résultats pour toutes les résolutions de MNT. Cependant, des trois résolutions, celle de 50 m donne de meilleures résultats selon le *Match*.

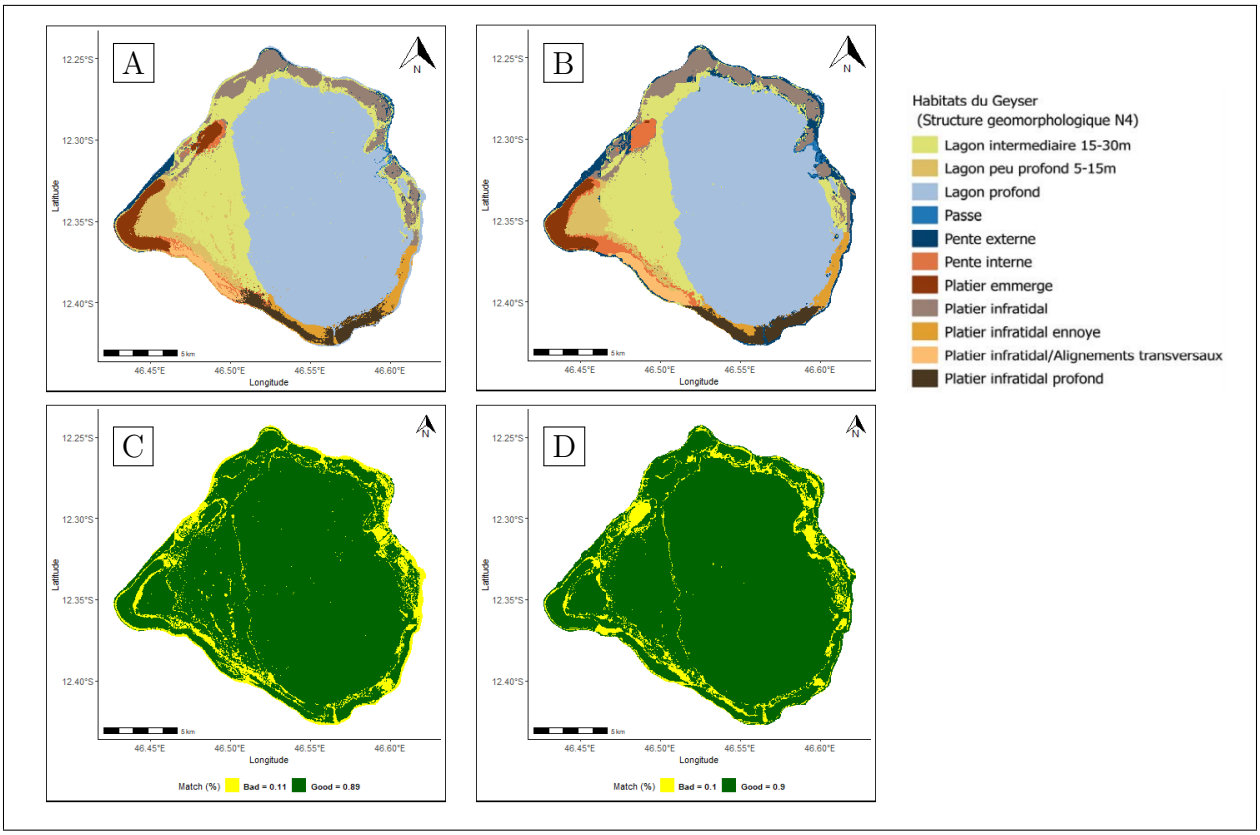

FIGURE 5 – Prédictions (A et B) avec le MNT de résolution 50 m et comparaison (C et D) avec la carte des experts. Les modèles sont entrainés avec 490 observations générés par échantillonnage spatialement homogène (A et C) et par échantillonnage complexité-dépendant (B et D).

Sur la Figure 5, nous montrons des prédictions obtenues avec le MNT de résolution 50 m pour les deux méthodes d'échantillonnage : spatialement homogène (A) et complexité-dépendant (B). Chacune d'elles est comparée à la carte des experts et la proportion des pixels où le label prédit correspond à celui des experts est mesurée (voir *Good* sur les cartes).
Nous pouvons dire que les deux méthodes d'échantillonnage donnent des résultats globalement similaires quand on les compare avec la carte des experts. En effet, les *Match* (ou *Good*) des deux cartes sont presqu'égales (0,89 et 0.9). Nous remarquons également que les cellules où le label prédit ne correspond pas à celui des experts (*Bad*) ne sont pas toujours les mêmes. Cependant, nous pouvons noter que ces cellules sont souvent situées des zones frontières entre

un ou plusieurs habitats. Nous notons également la difficulté de prédire certaines typologies d'habitats telles que la pente et la passe.

5 Conclusion et perspectives

Selon la résolution du MNT et la taille d'échantillon, les prédictions obtenues sont différentes. Une analyse plus approfondie de résultats présentés est envisagée et sera nécessaire pour voir à quel point ceux-ci sont différents. Des travaux sont également en cours pour développer une méthodologie d'échantillonnage permettant à la fois de tenir compte de la complexité du milieu et d'échantillonner toute la zone d'étude.

Bibliographie

Ambroise, C. and McLachlan, G. J. (2002), *Selection bias in gene extraction on the basis of microarray gene-expression data*, Proceedings of the national academy of sciences, 99(10), 6562-6566.

Andréfouët, S. and Dirberg, G. (2006). *Cartographie et inventaire du système récifal de Wallis, Futuna et Alofi par imagerie satellitaire Landsat 7 ETM+ et orthophotographies aériennes à haute résolution spatiale*. Conventions Sciences de la Mer, Biologie Marine.

Andréfouët, S., Muller-Karger, F. E., Robinson, J. A., Kranenburg, C. J., Torres-Pulliza, D., Spraggins, S. A., and Murch, B. (2006), *Global assessment of modern coral reef extent and diversity for regional science and management applications : a view from space*, In Proceedings of the 10th international coral reef symposium 2, 1732-1745. Okinawa : Japanese Coral Reef Society.

Baillargeon, S. (2005), *Le krigeage : revue de la théorie et application*, Mémoire, Université de Laval.

Cochran, W. G. (1977), *Sampling techniques*, John Wiley & Sons.

Dartnell, P., (2000), *Applying Remote Sensing Techniques to map Seafloor Geology/Habitat Relationships*, Masters Thesis, San Francisco State University.

Franklin, J. (1995). *Predictive vegetation mapping : geographic modelling of biospatial patterns in relation to environmental gradients*. Progress in physical geography, 19(4), 474-499.

Hafezi, M. H., Liu, L., and Millward, H. (2018), *Learning daily activity sequences of population groups using random forest theory* Transportation research record, 2672(47), 194-207.

Hartigan, J. A. (1975), *Clustering algorithms*, John Wiley & Sons, Inc.

Horn, B. (1981), *Hill shading and the reflectance map.*, Proceedings of the IEEE, 69, 14–47.

Kaufman, L., and Rousseeuw, P. J. (1990), *Partitioning around medoids (program pam)*, Finding groups in data : an introduction to cluster analysis, 344, 68-125.

Menze, B.H., Kelm, B.M., Masuch, R. et al. (2009), *A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data.* BMC Bioinformatics 10, 213 . https ://doi.org/10.1186/1471-2105-10-213

Roos, D., Dupon, P., Gaboriau, M., Bigot, L., Durville, P., Mulochau, T., Pinault, M., Wickel, J., Urbina-Barreto, I., Mouquet, P., Maurel, L., Cantou, M., Fallourd, S., Guilbert, A., Hoarau, J., Aumond, Y., Huet, J., Evano, H., Sabathe, Y., Giannasi, P., Adami, P., Mercky, Y., Jac, C., Sucre, E., Pelletier, D. and Claverie, T., (2017)*Projet EPICURE : Étude des Peuplements Ichtyologiques et des CommUnautés RécifalEs à partir d'indicateurs spatiaux et de l'approche fonctionnelle, des bancs du Geyser, de la Zélée et de l'Iris. Programme du Xème FED régional « Gestion durable du patrimoine naturel de Mayotte et des îles Eparses ».* Rapport de contrat no 15/1212185. RST/RBE-DOI/2017-07. https ://doi.org/10.13155/54549

Sappington, J. M., Longshore, K. M. and Thompson, D. B. (2007), *Quantifying Landscape Ruggedness for Animal Habitat Analysis : A Case Study Using Bighorn Sheep in the Mojave Desert.*, Journal of Wildlife Management, 71(5), 1419-1426.

Svetnik, V., Liaw, A., Tong, C., and Wang, T. (2004), *Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules*, In Multiple Classifier Systems : 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004. Proceedings 5 (pp. 334-343). Springer Berlin Heidelberg.

Walvoort, D. J., Brus, D. J., and De Gruijter, J. J. (2010). *An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means*, Computers & geosciences, 36(10), 1261-1267.

Wilson, M. F., O'Connell, B., Brown, C., Guinan, J. C., and Grehan, A. J. (2007), *Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope*, Marine Geodesy, 30(1-2), 3-35.

Kuhn, M., (2019-03-27), The caret package.
`http://topepo.github.io/caret/recursive-feature-elimination.html`