ISSN 1813-3290, http://www.revist.ci

ÉVALUATION DES PERFORMANCES DES ALGORITHMES ET DE LA COMPLEXITÉ DES MODÈLES POUR LA PRÉDICTION DE LA RÉPARTITION BIOGÉOGRAPHIQUE DU GENRE *COCHLOSPERMUM* KUNTH

Y. TOFFA [1] et A. B. FANDOHAN [1,2,3*]

[1] Université Nationale d'Agriculture, Ecole de Foresterie Tropicale, Laboratoire de Sciences Végétales Horticoles et Forestières, Unité de Recherche en Foresterie et Conservation des Bioressources, BP 43, Kétou, Bénin
[2] Université d'Abomey Calavi, Faculté des Sciences Agronomiques, Laboratoire de Biomathématiques et d'Estimations Forestières, 01 BP 526, Cotonou, Bénin
[3] Université d'Abomey Calavi, Faculté des Sciences Agronomiques, Laboratoire d'Ecologie Appliquée, 01 BP 526, Cotonou, Bénin

(reçu le 01 Février 2021 ; accepté le 06 Mai 2021)

* Correspondance, e-mail : *bfandoohan@gmail.com*

RÉSUMÉ

Les algorithmes de modélisation de la distribution des plantes sont devenus des outils très populaires pour estimer l'étendue des aires favorables aux espèces et leur réponse potentielle aux changements environnementaux. Dans cette étude, les performances de six algorithmes en fonction de différents modèles et types de variables de prédiction de la répartition biogéographique du genre *Cochlospermum* ont été évaluées. Les modèles ont été tournés avec 592 observations de présence du genre *Cocholspermum* et six variables bioclimatiques, une variable indicatrice de la variation du rayonnement solaire et deux variables de biais d'échantillonnage. Les performances des modèles ont été évaluées au moyen de la zone sous une courbe de la caractéristique de fonctionnement du récepteur (AUC), la *True Skill Statistic* (TSS) et le *Symmetric Extremal Dependence Index* (SEDI). Les résultats montrent que l'algorithme MAXENT est le plus performant pour modéliser la distribution du Genre *Cochlospermum* et l'impact du changement climatique sur la dynamique spatiotemporelle de ses habitats. De même, la prise en compte de la variable indicatrice de la variation du rayonnement solaire et des variables de biais n'améliore pas les performances des modèles.

Ce travail met en évidence la variation des performances des modèles en fonction des algorithmes utilisés et la nécessité de calibrer des modèles spécifiques pour chaque espèce dans les études d'évaluation de l'impact des changements climatiques sur la distribution des espèces.

Mots-clés : *forêts aléatoires, principe d'entropie maximale, arbre de régression amélioré, machine à vecteurs de support, modèle linéaire généralisé, modèle linéaire généralisé régularisé avec lasso ou elastic-net.*

ABSTRACT

Assessing performances of algorithms and model complexity for predicting the biogeographic distribution of the genus *Cochlospermum* Kunth

Algorithms for modeling plant distribution have become very popular tools for estimating habitat suitability for species and their potential response to environmental changes. In this study, the performance of six algorithms as a function of different models and types of variables predicting the biogeographic distribution of the genus *Cochlospermum* was assessed. The models were run with 592 presence records of the genus *Cocholspermum* and six bioclimatic variables, one indicator variable for variation in solar radiation and two sampling bias variables. The Models' performances were evaluated using the Area Under Receiver Operating Characteristic Curve (AUC), the True Skill Statistic (TSS) and the Symmetric Extremal Dependence Index (SEDI). The results showed that the MAXENT algorithm is the most performant for modelling the distribution of the Genus Cochlospermum and the impact of climate change on the spatiotemporal dynamics of its habitats. Likewise, taking into account the indicator variable of the variation in solar radiation and the bias variables does not improve the performances of the models. This work highlights the variation in model performance depending on the algorithms used and the need to calibrate specific models for individual species in studies assessing the impact of climate change on the distribution of species.

Keywords : *random forest*, *maximum entropy principle*, *boosted regression tree*, *support vector machine*, *generalized linear model, generalized linear model with lasso or elastic-net*.

I - INTRODUCTION

Les modèles de distribution des espèces (SDM) ont été développés et popularisés pour évaluer la distribution des espèces sur le globe terrestre et l'impact du changement climatique sur la dynamique spatiotemporelle potentielle de leurs habitats [1 - 4]. Ils sont également utilisés en combinaison avec d'autres outils pour sélectionner des sites pour la réintroduction d'espèces et établir des priorités de conservation [5]. Pour ce faire, certaines méthodes n'utilisent que des données de présence, e.g. BIOCLIM, DOMAIN et *Limiting Variable and Environmental Suitability* (LIVES) [6, 7], alors que d'autres utilisent à la fois des données de présence et d'absence (eg. modèle linéaire généralisé (GLM), modèle additif généralisé (GAM) [7]). D'autres encore utilisent des données présences et des données pseudo-absence, (eg. analyse des facteurs de niche écologique et Maxent, [8, 9]). En dépit de leur utilité potentielle et parfois avérée, la popularité de ces modèles n'a d'égal que la multitude de leurs faiblesses relevées et discutées. Ces outils ont donc besoin d'être constamment améliorés afin de réduire la probabilité que des recommandations erronées découlent de leurs résultats. L'amélioration de ces outils de modélisation de la distribution des espèces est donc motivée par le besoin croissant de prédire avec le plus de précision possible, les réponses des espèces aux changements environnementaux [10, 11].

A cet effet, des investigations récentes ont mis en évidence, la variation de la performance des différentes techniques de modélisation et de l'influence du nombre et de la répartition spatiale des données de présence d'espèces sur les résultats de modélisation [12]. Pour contourner ce problème, plusieurs études [13 - 17] ont suggéré l'estimation de modèle moyens qui seraient plus robuste que chacun des modèles pris isolement. Cette option n'est pas non plus sans faille majeure. Par exemple, les coefficients moyens de régression estimés sur plusieurs modèles à partir du Critère d'Information Akaike (*AIC*) sont souvent utilisés pour les modèles d'ensemble. Cependant, ils ne sont pas des estimations valides et interprétables des effets partiels des prédicteurs individuels lorsqu'il existe une multi-colinéarité entre les variables prédictives. En effet, la multi-colinéarité suppose que la mise à l'échelle des unités dans les dénominateurs des coefficients de régression peut changer d'un modèle à l'autre de telle sorte que ni les paramètres ni leurs estimations n'ont des échelles communes, donc fondamentalement, leur moyenne n'a aucun sens d'un point de vue statistique [18]. En attendant d'avoir une alternative statistiquement valide, l'une des options pour prendre en compte ces faiblesses est d'imposer aux modélisateurs, non seulement d'intégrer dans leur approche méthodologique, une étude comparative des performances de différents modèles, mais aussi de calibrer des modèles explicitement

spécifiques à chaque espèce. Ces options pourraient permettre de faire les recommandations à partir des meilleurs modèles. Une autre faiblesse assez discutée notamment au cours de la dernière décennie est celle relative au problème de validité statistique de la spécification des modèles. En effet, dans nombre de cas, la distribution observée des occurrences des espèces est davantage fonction de l'accessibilité de certains sites et des collectes opportunistes de proche en proche, qu'une fonction de la réponse des espèces aux variables environnementales [19, 20]. Ainsi, l'absence de certaines variables indirectes et/ou des variables de biais dans les équations de spécification conduit alors à des modèles statistiquement non valides. En utilisant le genre *Cochlospermum* Kunth comme espèce cible, la présente étude (1) compare les performances de six algorithmes communs à modéliser la distribution de l'espèce, (2) évalue l'effet de l'incorporation des variables indirectes et de biais sur cette performance et (3) évalue la stabilité des contributions des différentes variables prédictives d'un algorithme à un autre. Six algorithmes parmi les plus utilisés ont été ciblés : Forêts Aléatoires (*Random Forest*, RF), Principe d'entropie maximale (MaxEnt), Arbre de Régression Amélioré (*Boosted Regression Tree*, BRT), Machine à vecteurs de support (*Support Vector Machine*, SVM), Modèle Linéaire Généralisé (GLM) et Modèle Linéaire Généralisé régularisé avec la méthode *Lasso* (*Least absolute shrinkage and selection operator*) ou *Elastic-Net* (GLMnet). L'objectif de cette étude est de sélectionner l'algorithme le plus performant pour la modélisation de la distribution et l'impact des changements climatiques sur *Cochlospermum*.

II - MÉTHODE

II-1. Données de présence utilisées

Les données d'occurrence utilisées sont celles du genre *Cochlospermum* Kunth au Benin. Afin d'éviter les problèmes d'incertitudes liées à la qualité des données, sur la base des occurrences disponibles du Genre sur le portail électronique du *Global Biodiversity Information Facility* (GBIF), une exploration profonde du terrain a été faite et a permis la confirmation de 592 présences de l'espèce au moyen d'un *Global Positioning System* (marque Garmin, précision, 10 m) à travers son aire de distribution, telle que décrite par la plateforme GBIF ***(Figure 1)***. Les coordonnées géographiques longitudinales et latitudinaux de ces points ont été extraites du GPS et stockées dans un fichier sous format csv pour une utilisation ultérieure.

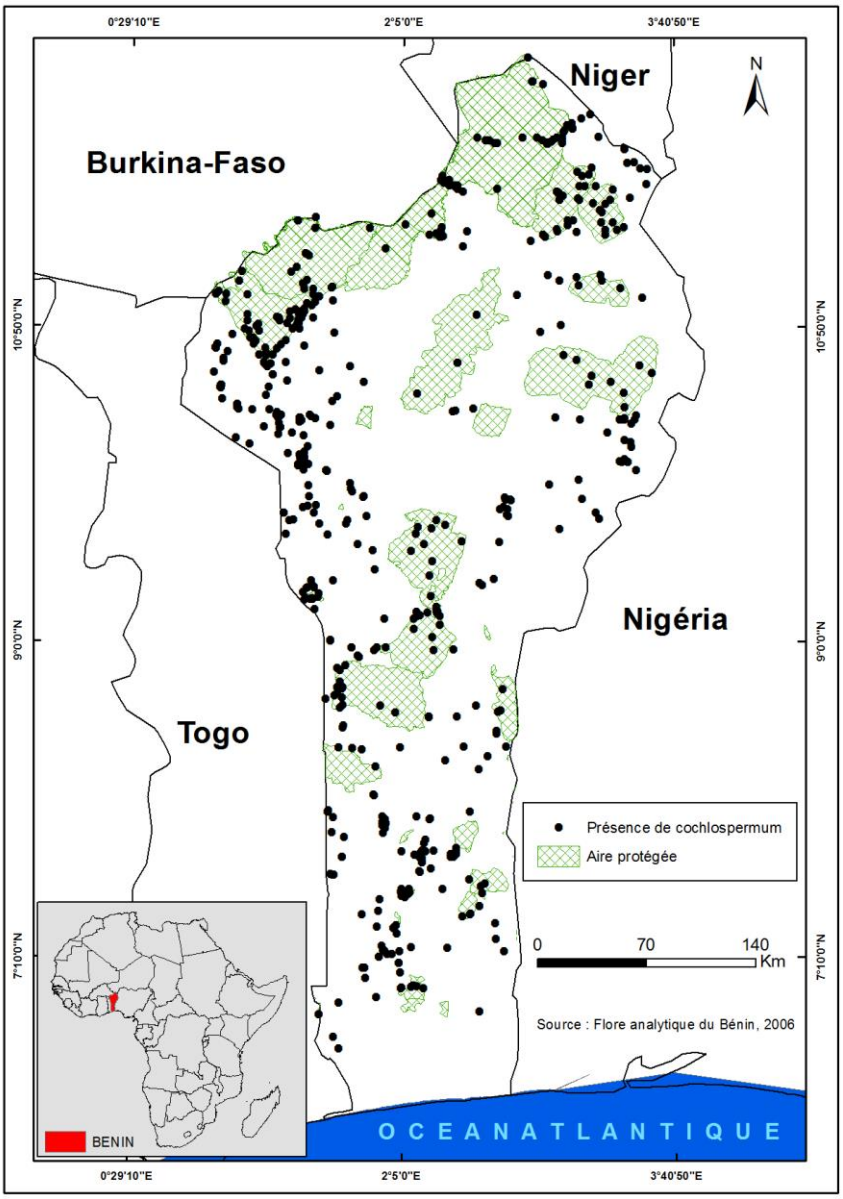

Figure 1 : *Réparation géographiques des points de présence de Cochlospermum au Bénin*

II-2. Variables utilisées

Trois types de variables ont été utilisées dans le cadre de cette étude : les variables bioclimatiques ayant un effet écologique potentiel direct sur la distribution des espèces ; les variables indirectes qui n'ont pas d'effet direct sur la distribution des organismes vivants mais qui ont d'effet sur la variation

des variables à effet direct, les variables de biais qui ont un effet potentiel sur la distribution des points d'occurrences disponibles des espèces, du fait des biais d'échantillonnage fonction de l'accessibilité des zones de collecte, la présence d'agglomération et la géomorphologie du milieu. Une sélection raisonnée des variables bioclimatiques à utiliser a été faite en fonction de la connaissance de l'écologie des plantes dans le milieu d'étude. La distribution des plantes dans le milieu étant une fonction de l'aridité [21], six variables rendant compte du gradient d'aridité et les moins corrélées ont été sélectionnées ($r < 0,7$ test de Pearson) : *arid* (l'inverse de l'indice d'aridité); *bio2* (la variation journalière moyenne des températures); *bio3* (l'iso-thermalité); *bio6* (la température minimale du mois le plus froid), *bio14* (pluviométrie du mois le plus sec); *mi* (l'indice d'humidité annuelle). Les valeurs de ces variables sont disponibles sur la plateforme http://www.york.ac.uk/environment/research/kite/resources/) et représentent les conditions bioclimatiques actuelles du milieu d'étude, calculées sur 50 année [22]. Comme variable indirecte, la variation de la pente (*slope variation*, *sv*) a été utilisée comme indicatrice de la variation du rayonnement solaire. Les variables de biais considérées sont la distance à la plus proche agglomération (*distance to the nearest human settlement*, *dists*) et la distance à l'aire protégée la plus proche (*distance to the nearest protected area*, *distp*). Ces trois dernières variables ont été générées au moyen de l'outil *spatial analyst tool* du logiciel ARCGIS, respectivement à partir des données sur le Modèle Numérique de Terrain (disponible sur le site http://srtm.csi.cgiar.org/), les données sur la carte des aires protégées et les données sur les agglomérations du Benin (disponible sur le site hhttps://www.diva-gis.org/Data).

II-3. Modélisation et évaluation de la performance des modèles

Le logiciel R (http ://cran.r-project.org/ ; version 4.0.2.) a été utilisé pour la préparation et le traitement des données. Le traitement des données a nécessité les packages *BiodiversityR*, *dismo*, *raster*, *rJava*, *doParallel*, *ggplot2, gridExtra, ggpubr*. Sous chacun des algorithmes évalués, quatre modèles ont été calibrés. Le Modèle 1 inclus les variables bioclimatiques uniquement :

$$Model\ 1 \sim \text{arid} + \text{bio2} + \text{bio3} + \text{bio6} + \text{bio14} + \text{mi} \quad (1)$$

Le modèle 2 en plus des variables bioclimatiques, intègre des approximations des biais potentiels de collecte de présence du Genre *Cochlospermum* dans les aires protégées et autour des zones anthropisées :

$$Model\ 2 \sim \text{arid} + \text{bio2} + \text{bio3} + \text{bio6} + \text{bio14} + \text{mi} + \text{dists} + \text{distp} \quad (2)$$

Le modèle 3 en plus des variables bioclimatiques, intègre une approximation de la radiation solaire :

$$Model\ 3 \sim \text{arid} + \text{bio2} + \text{bio3} + \text{bio6} + \text{bio14} + \text{mi} + \text{sv} \quad (3)$$

Le modèle 4 prend en compte toutes les variables considérées dans les modèles précédents :

$$Model\ 4 \sim \text{arid} + \text{bio2} + \text{bio3} + \text{bio6} + \text{bio14} + \text{mi} + sv + dists + distp \quad (4)$$

Les statistiques suivantes ont été utilisées pour évaluer la performance des modèles en fonction des algorithmes utilisés.

II-3-1. Statistique de précision

Nous utilisons les abréviations suivantes pour les grandeurs empiriques : (*a*), le nombre de vraies présences, (*b*), le nombre de fausses présences (commission), (*c*), le nombre de fausses pseudo-absences (omission) et *d* (le nombre de vraies pseudo-absences). La matrice de confusion ci-dessous est générée à partir de ces grandeurs empiriques ***(Tableau 1)***

Tableau 1 : *Matrice de confusion pour le calcul des statistiques de performance*

	Occurrence Observée	
Occurrence prédite	Présence	Absence
Présence	Vraie (a)	Fausse (b)
Absence	Fausse (c)	Vraie (d)

II-3-2. Zone sous une courbe de la caractéristique de fonctionnement du récepteur (AUC ou AUROC)

Une courbe ROC (courbe de la caractéristique de fonctionnement du récepteur) est un graphe montrant les performances d'un modèle de classification à tous les seuils de classification. Cette courbe trace deux paramètres : Le taux de vrais positifs (vraies présences) et le taux de faux positifs (fausses présences) prédits par le modèle. La zone sous une courbe de la caractéristique de fonctionnement du récepteur, abrégée AUC, est une statistique indépendante de seuil, largement utilisée dans diverses disciplines, y compris l'écologie [23]. C'est une valeur unique de probabilité qui mesure globalement la performance d'un algorithme ou d'un classificateur binaire [24]. Dans une classification ascendante comme une régression logistique, c'est la probabilité qu'un modèle classe un point de présence aléatoirement

choisi plus haut qu'un point d'absence aléatoirement choisi. La valeur de l'AUC est comprise entre 0 et 1. Un modèle dont la valeur de l'AUC est proche de 1 est considérée comme très performant (ayant un taux d'erreur de classification équivalent à zéro). Une valeur inférieure à 0,5 indique un modèle de faible performance.

II-3-3. Erreur de commission

L'erreur de commission ou taux de faux positifs (*False Positive Rate, FPR*) est une mesure seuil-dépendante de précision, définie comme la proportion d'aire faussement classée comme aire de présence de l'espèce modélisée [24]. En général, plus faible est l'erreur de commission plus élevée est la performance d'un modèle. Dans le cas d'utilisation de pseudo-absence, cette mesure doit être interprétée avec caution. Elle est donnée par la ***Formule*** [25] :

$$FPR = b/(b+d) \quad (5) \tag{5}$$

II-3-4. Erreur d'Omission

L'erreur d'omission ou taux de faux négatifs (*False Negative Rate, FNR*) est une mesure de précision définie comme la proportion de zones prédites faussement comme des aires absences pour l'espèce (c.-à-d. proportion de présences d'espèces mal identifié par les modèles). Elle fournit des informations sur la capacité discriminatoire du modèle et sa propension à un sur-ajustement. En général, une erreur d'omission faible indique une performance plus élevée (meilleure discrimination entre les zones favorables à l'espèce et celles non favorables). Les modèles présentant un problème de sous-ajustement ont également des erreurs d'omission élevées. Elle donnée par la ***Formule*** [25] :

$$FNR = c/(c+a) \tag{6}$$

II-3-5. Sensivity

La *Sensivity* encore appelée taux de vraie présence est le rapport des vraies présences à la somme des vraies présences et des fausses absences en ignorant les fausses absences et les vraies absences. C'est la probabilité que le modèle classe correctement une vraie présence.

$$Sensivity = a/(a+c) \tag{7}$$

II-3-6. Specificity

La *Specificity* encore appelé taux de vraie absence est égale à *1 - le taux d'erreurs de commission*. C'est aussi le rapport des vraies absences à la somme des vraies absences et des fausses présences. C'est la probabilité que le modèle classe correctement une vraie absence.

$$Specificity = 1 - b/(b + d) \tag{8}$$

II-3-7. True Skill Statistic

La *True Skill Statistic* est une mesure de précision seuil-dépendant, largement utilisée dans les tests de diagnostic médical. Elle a été introduite dans la littérature des modèles de distribution des espèces par [26]. C'est une mesure de la capacité du modèle à détecter avec précision les vraies présences (sensibilité) et les vraies absences (spécificité). Elle varie entre -1 et 1 avec une valeur approchant 1 indiquant une précision de classification élevée tandis que la valeur au-dessous de 0 indique une classification pas mieux qu'aléatoire. La TSS varie de – 1 à 1 et est donnée par la ***Formule*** :

$$TSS = a/(a + c) - b/(b + d) = Sensitivité + Specificité - 1 \tag{9}$$

Une TSS proche de 1 est indicatrice d'une bonne performance.

II-3-8. Indice de Dépendance Extrême Symétrique (Symmetric Extremal Dependence Index, SEDI)

L'Indice de Dépendance Extrême Symétrique est une mesure de précision seuil-dépendante récemment jugée meilleure à la *True Skill Statistics*, notamment lorsque le nombre de points de pseudo-absence est très élevées [27]. C'est généralement le cas dans les modèles tournés avec des variables dont les données sont de résolution élevée (eg., données issues de modèles régionaux de circulations constituant les plus utilisées présentement, dans le cadre de la modélisation de la distribution des espèces). Son domaine de définition est [−1, 1] et il est donné par la ***Formule*** :

$$SEDI = \frac{log\left(\frac{b}{(b+d)}\right) - log\left(\frac{a}{(a+c)}\right) - log\left(1 - \frac{b}{(b+d)}\right) + log\left(1 - \frac{a}{(a+c)}\right)}{log\left(\frac{b}{(b+d)}\right) + log\left(\frac{a}{(a+c)}\right) + log\left(1 - \frac{b}{(b+d)}\right) + log\left(1 - \frac{a}{(a+c)}\right)} \tag{10}$$

Un SEDI proche de 1 est indicatrice d'une bonne performance.

II-3-9. Permutation importance

La *Permutation importance* est utilisée pour mesurer l'impact d'une variable donnée sur un modèle. Elle estime plus précisément la perte en performance subit par un modèle lorsqu'une variable donnée est retirée de ce dernier. Dans cette étude, la permutation importance basée sur l'AUC (*AUC-based permutation importance*) a été utilisée [28]. Elle mesure donc la baisse en grandeur AUC subit par un modèle lorsqu'une variable donnée est soustraite du modèle. Pour ce faire, on simule la suppression de chaque variable du modèle en mélangeant ses valeurs entre elles de sorte à éliminer du modèle les informations qu'elle lui apporte. On obtient la réduction en performance (en fraction de l'AUC) que subit le modèle du fait de cette opération. Cette réduction est considérée comme estimateur de l'importance de la variable pour le modèle. Cette opération est effectuée pour chacune des variables prises en compte dans le modèle. Contrairement à d'autres statistiques de mesure d'impact comme le *Gini importance*, la permutation importance est moins sensible à une grande variation des magnitudes/dimensions des variables prédictives utilisées dans les modèles [29] et est aussi robuste lorsque le nombre de pseudo-absences est très hautement plus élevés que le nombre de présences [28]. Compte tenu de leurs meilleures propriétés statistiques (plus puissantes), les valeurs brutes de la *permutation importance* ont été préférées à celles standardisées au *z-score*, souvent utilisées [30].

III-RÉSULTATS

III-1. Performance des modèles en fonction de l'algorithme et de la spécification du modèle

III-1-1. Modèle 1

L'analyse de la performance des modèles tourner avec les variables bioclimatiques seules suggère que les algorithmes *Random Forest* (RF) et *Booted Regression Tree* (BRT) présentent les meilleures performances et GLMnet la plus faible performance, sur la base de l'AUC et de la TSS, ***(Figure 2a)***. La statistique SEDI indique par contre les meilleures performances pour MAXENT et GLM, et la plus faible pour GLMnet et *Support Vector Machine* (SVM). On note également que les valeurs de la statistique TSS indiquent globalement des pouvoirs de prédiction relativement aléatoires (inférieur à 0,5) alors que l'AUC et le SEDI indique des pouvoirs relativement bons.

III-1-2. Modèle 2

Lorsque les variables de biais sont ajoutées à celles bioclimatiques, RF, MAXENT et BRT présentent les meilleures performances avec des pouvoirs prédictifs relativement bons, et GLM et GLMnet les plus faibles performances (AUC, TSS ; ***Figure 2b***). La statistique SEDI suggère par contre SVM comme l'algorithme le plus performant et GLMnet, le moins performant. A l'exception du cas de RF, les valeurs de la TSS indiquent également des pouvoirs prédictifs relativement aléatoire, contrairement à l'AUC et au SEDI.

III-1-3. Modèle 3

Pour les modèles tournés avec la variable indirecte variation de la pente combinée avec les variables bioclimatiques, RF présente les meilleures performances pour l'AUC et la TSS et SVM et GLMnet les plus faibles ***(Figure 2c)***. MAXENT présente cependant la meilleure performance du point de vue de la statistique SEDI, et l'algorithme SVM la plus faible. Quel que soit l'algorithme, les valeurs de la TSS présentent aussi un pouvoir prédictif relativement aléatoire contrairement aux autres statistiques de performance.

III-1-4. Modèle 4

Pour le modèle intégrant toutes les variables, RF, MAXENT et BRT présentent les meilleures performances et SVM la plus faible (AUC, TSS ; ***Figure 2d***). Par contre, d'après la statistique SEDI, MAXENT et BRT présentent les meilleures performances et SVM la plus faible. Les valeurs de la TSS indiquent aussi un pouvoir prédictif relativement aléatoire contrairement à la SEDI et à l'AUC. Globalement, pour l'algorithme MAXENT, l'intégration séparée de la variable indirecte et des variables de biais entraine une amélioration du score de l'AUC et du TSS mais une réduction du score du SEDI. Le modèle global présente toutefois des scores plus élevés pour toutes les statistiques de performance. L'algorithme BRT présente des tendances similaires à MAXENT pour l'AUC et la TSS. Quant au SEDI, son score est amélioré par l'intégration des variables de biais et le modèle global mais réduit pas l'intégration de la variable indirecte. L'intégration des variables de biais et de la variable indirecte améliore les scores de l'AUC mais réduit ceux du SEDI, y compris au niveau du modèle global pour l'algorithme RF. Le Modèle intégrant les variables de biais en plus des variables bioclimatiques présente le meilleur score pour l'algorithme SVM. La prise en compte des variables de biais et de la variable indirecte a un effet positif sur le score de l'AUC et de la TSS pour les algorithmes GLM et GLMnet. Au niveau du SEDI toutefois, alors que cela a un effet positif sur le score du SEDI pour le GLMnet, le contraire est observé au niveau du GLM.

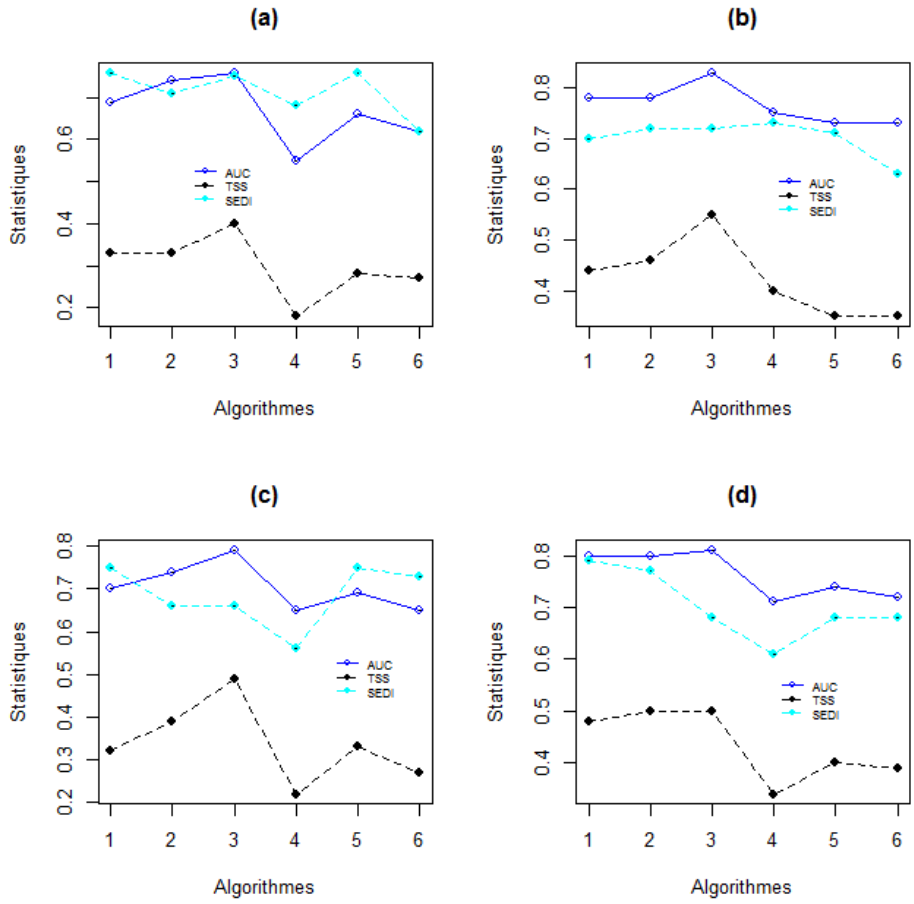

Figure 2 : *Variation des performances des algorithmes en fonction de différentes statistiques et différent modèles spécifiés*

Légende : (a) modèle 1, (b) modèle 2, (c) modèle 3, (d) modèle 4 ; 1 = MAXENT ; 2 = BRT, 3 = RF, 4 = SVM, 5 = GLM, 6 = GLMnet

III-2. Impact (importance) des variables sur les modèles en fonction de l'algorithme et de la spécification du modèle

III-2-1. Modèle 1

Dans le modèle n'incluant que les variables bioclimatiques, le retrait des variables *bio2* et *bio6* entraine une plus grande réduction de performance lorsque l'algorithme MAXENT est utilisé ***(Figure 3)***. Lorsque le modèle est tourné avec BRT ou RF, *arid* et *bio3* sont les variables ayant le plus d'impact. Les tendances pour les autres algorithmes indiquent *bio3* et *mi* pour SVM, *arid* et *bio6* pour GLM et *arid* et *bio2* pour GLMnet.

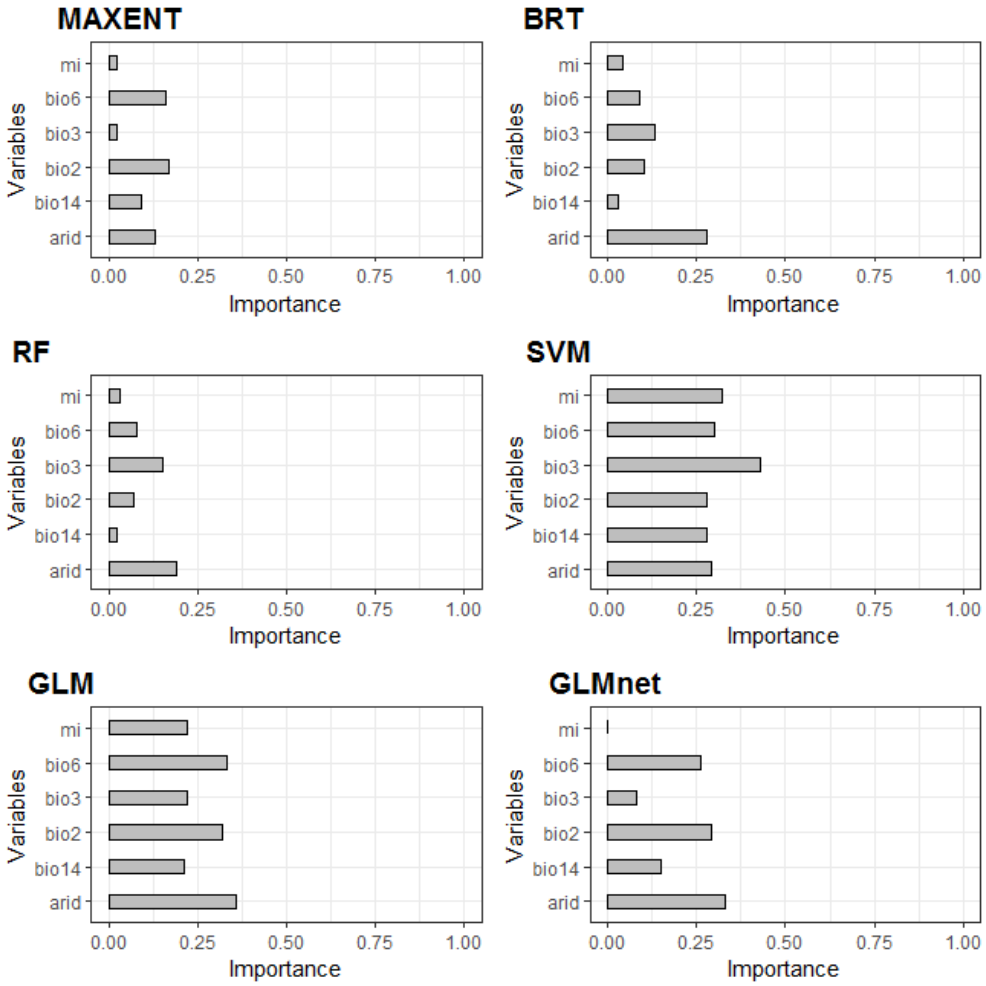

Figure 3 : *Variation de l'impact des variables sur la performance du modèle en fonction des algorithmes, lorsque seules les variables bioclimatiques sont prises*

III-2-2. Modèle 2

Dans le modèle ajoutant les variables de biais à celles bioclimatiques, les variables de biais deviennent les plus importantes au niveau des algorithmes MAXENT, BRT et RF ***(Figure 4)***. Les tendances indiquent toutefois *arid* et *bio3* pour SVM et *arid* et *bio3* pour GLM et GLMnet.

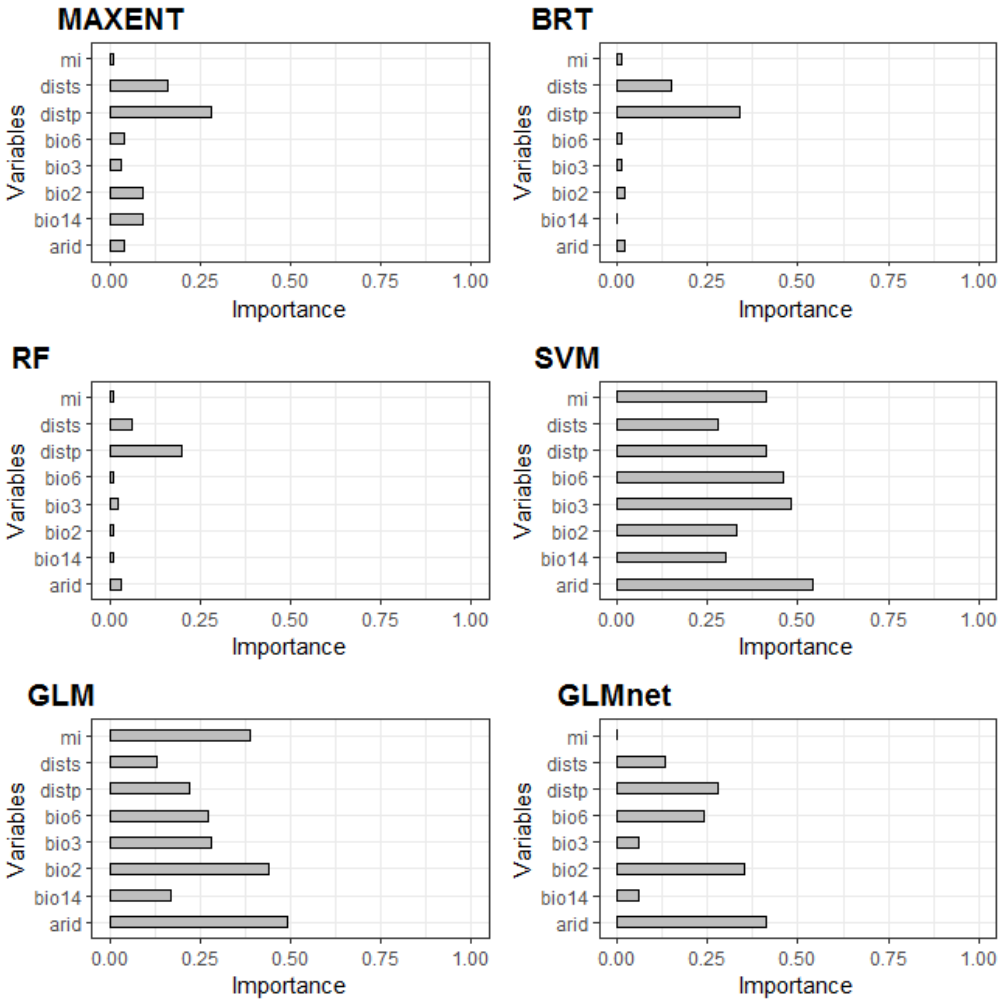

Figure 4 : *Variation de l'impact des variables sur la performance du modèle en fonction des algorithmes, lorsque les variables bioclimatiques sont prises avec les variables de biais*

III-2-3. Modèles 3

Pour l'algorithme MAXENT, les variables *bio2* et *bio6* demeurent les plus importantes lorsque la variable indirecte de variation de pente (*sv*) est ajoutée aux variables bioclimatiques. Cependant, pour BRT, *arid* et *sv* deviennent les variables ayant le plus grand impact. *arid* et *bio3* ont les plus grands impacts pour RF, *bio6* et *arid* pour SVM *arid* et *bio2* pour GLM et GLMnet ***(Figure 5)***.

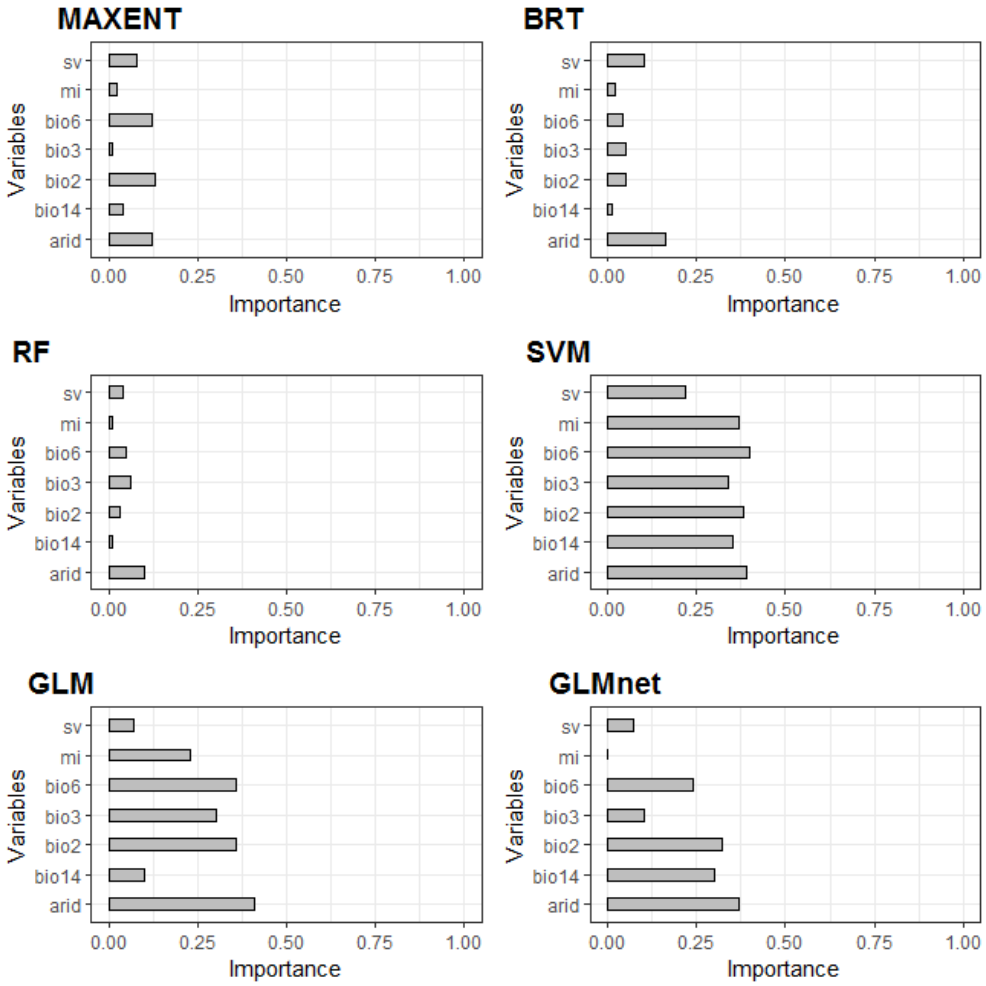

Figure 5 : *Variation de l'impact des variables sur la performance du modèle en fonction des algorithmes, lorsque les variables bioclimatiques sont prises avec la variable indirecte de variation de la pente*

III-2-4. Modèle 4

Pour le modèle global intégrant toutes les variables, la variation de pente (*sv*) et la variable de biais *distp* ont le plus grand impact pour l'algorithme MAXENT ***(Figure 6)***. Les deux variables de biais (*distp* et *dists*) sont les plus importantes lorsque le modèle est tourné avec BRT, *distp* et *bio6* pour RF, *bio6* et *bio2* pour SVM, *arid* et *bio2* pour GLM et GLMnet.

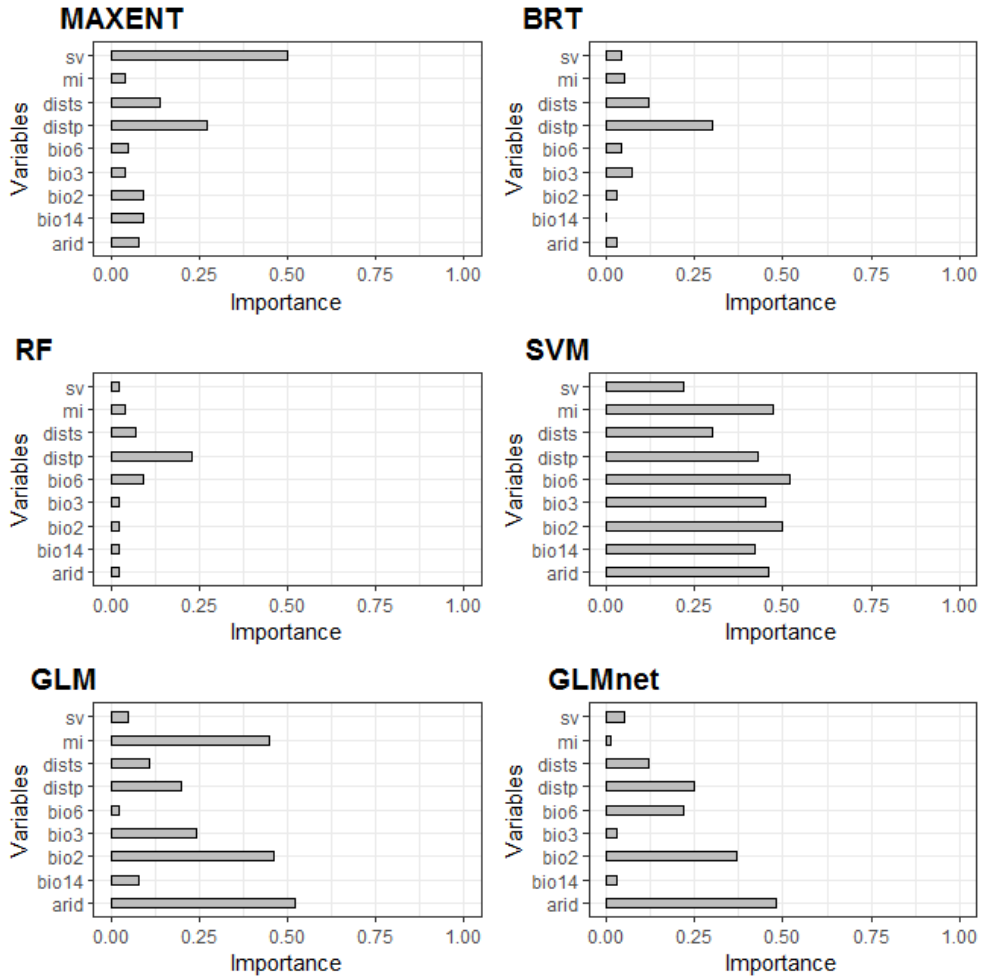

Figure 6 : *Variation de l'impact des variables sur la performance du modèle en fonction des algorithmes dans le modèle global incluant toutes les variables considérées*

IV- DISCUSSION

L'évaluation des performances des modèles distribution des espèces en fonction des variables et des algorithmes constituent un outil puissant de décision. Elle permet de calibrer des modèles spécifiques à chaque espèce et de réduire les incertitudes dues aux faiblesses inhérentes aux modèles. Elle permet également de mettre à la disposition des décideurs politiques des recommandations basées sur meilleurs modèles. Dans l'évaluation du pouvoir prédictif des algorithmes sur la répartition des espèces, des aires de distributions potentielles sont prédites, mais la performance des algorithmes

n'est mesurée qu'en utilisant des données de distribution réelles. Une telle procédure de validation peut paraître absurde en ce sens que ce qui est prédit n'est pas ce qui est évalué dans le processus de validation [15, 31]. Pourtant, une telle évaluation est utile puisqu'une similitude entre les schémas de distribution prévus et observés peut indiquer que les variables prédictives sont les principaux facteurs limitant la distribution des espèces concernées. De même, une déviation de la distribution à partir de celle prévue peut être interprétée comme une preuve de l'effet d'autres facteurs non pris en compte, sur la distribution réelle du taxon étudié. Avec la mise en évidence de l'effet des caractéristiques spécifiques des espèces sur les résultats de modélisation (eg., distribution géographiques, prévalence) [15, 31], l'évaluation et la comparaison des performances des différents algorithmes doivent être intégrées dans les routines de modélisation de chaque taxon ciblé. Dans le contexte de ce travail, il a évalué, l'effet de différents modèles et types de variables sur les performances de différents algorithmes de prédiction de la répartition biogéographique de *Cochlospermum*. Globalement, on note une très grande variabilité de la structuration des scores de performance des algorithmes d'un modèle à l'autre pour l'AUC et le TSS. La statistique *Symmetric Extremal Dependence Index* (SEDI) a récemment été introduite comme une plus fiable statistique d'évaluation de la performance des modèles, notamment lorsque le nombre de pseudo-absences est très hautement plus élevés que le nombre de présences ; ce qui est le cas dans les modèles de distribution des espèces (eg. plus de 30000 pseudo-absences pour 592 présences) [27].

Sur la base la statistique SEDI, l'algorithme MAXENT présente la meilleure performance pour trois modèles sur quatre et la troisième meilleure performance pour le dernier modèle. Il présente également une meilleure stabilité de rang parmi tous les algorithmes quelle que soit la statistique utilisée (meilleure, deuxième ou troisième performance). De même, pour l'algorithme MAXENT, la statistique SEDI a la valeur la plus élevée pour le modèle global intégrant toutes les variables et celui intégrant uniquement les variables bioclimatiques. On note aussi une importante variabilité de l'impact des variables sur les performances des modèles en fonction de l'algorithme utilisés. Les statistiques d'importance présentent l'utilité d'une variable dans un modèle mais pas sa valeur explicative absolue relative à la distribution du taxon étudié. Cette grande instabilité de l'importance des variables à travers différents modèles et différents algorithmes confirme la nécessité d'interpréter l'effet isolé de variables sur la distribution des taxons avec beaucoup de cautions. On remarque par ailleurs une réduction notoire en grandeur de l'importance de la plupart des variables bioclimatiques lorsque la variable indirecte de variation de la pente et/ou les variables de biais sont ajoutées dans la spécification du modèle, exception faite des modèles tournés avec l'algorithme SVM. Les modèles de distributions des espèces sont prédictifs et non explicatifs [32]. Toutefois, des modèles où les variables

ayant un effet direct sur les espèces (i.e., variables bioclimatiques) ont une grande importance pourraient être d'une plus grande utilité écologique que des modèles statistiquement plus robustes mais calibrés avec des variables de biais et/ou des variables ayant un effet indirect. Aussi, contrairement aux observations antérieures sur plusieurs groupes taxonomiques [33], la prise en compte de variables indirectes et/ou de biais n'a pas eu un effet positif marqué sur la performance des modèles de distribution du genre *Cochlospermum* dans le milieu d'étude. Ceci suggère de faibles biais d'échantillonnage dans la collecte des données d'occurrence et une faible importance de la géomorphologie dans la distribution observée du taxon étudié.

V - CONCLUSION

Cette étude a testé l'effet de l'intégration d'une variable indirecte et de deux variables de biais sur les performances des modèles, et selon l'algorithme de modélisation utilisé. Les résultats illustrent la variation des performances des modèles en fonction des algorithmes utilisés et la nécessité de calibrer des modèles spécifiques pour chaque espèce dans les études d'évaluation de l'impact des changements climatiques sur la distribution des espèces. Les résultats montrent également que l'algorithme MAXENT est le plus performant pour modéliser la distribution du Genre *Cochlospermum* et l'impact du changement climatique sur la dynamique spatiotemporelle de ses habitats. En fin, en l'absence d'amélioration importante des performances des modèles par la prise en compte de la variable indirecte et des variables de biais, on peut conclure que le modèle ne prenant en compte que les variables bioclimatiques est suffisant pour modéliser la distribution du groupe taxonomique objet de l'étude.

Remerciements

Nous remercions le Gouvernement de la République du Bénin et le Ministère de l'Enseignement Supérieur et de la Recherche Scientifique pour avoir financé ce travail à travers une Bourse d'Etude Doctorale octroyée à Y. TOFFA. Une partie du financement a été couverte par le second auteur du travail, A. B. FANDOHAN. Nous sommes reconnaissants à Robert AGBOVOEDO et Fréderic A. ABIODOUN pour leur assistance lors de l'apprentissage pour le développement des scripts R et ARCGIS en vue de tourner les modèles.

RÉFÉRENCES

[1] - J. ELITH, H. C. GRAHAM, R. P. ANDERSON, M. DUDÍK, S. FERRIER, A. GUISAN, R. J. HIJMANS *et al.*, *Ecography*, 29 (2) (2006) 129 - 151

[2] - D. I. WARTON et L. C. SHEPHERD, *Annals of Applied Statistics*, 4 (3) (2010) 1383 - 1402

[3] - A. CHAKRABORTY, A. E. GELFAND, A. M. WILSON, A. M. LATIMER et J. A. SILANDER, *Applied Statistics*, 60 (5) (2011) 757 - 776

[4] - G. AARTS, J. FIEBERG et J. MATTHIOPOULOS, Methods in Ecology and Evolution, 3 (1) (2012) 177 - 187

[5] - G. N. GOUWAKINNOU, *West African Research and Development - Sub-Saharan Africa*, 6 (2013) 1 - 8

[6] - J. LI et D. W. HILBERT, *Biodiversity and Conservation*, 17 (13) (2008) 3079 - 3095

[7] - M. P. ROBERTSON, N. CAITHNESS et M. H. VILLET A., *Diversity and Distributions*, 7 (1-2) (2001) 15 - 27

[8] - A. A. H. HIRZEL, J. HAUSSER, D. CHESSEL et N. PERRIN, *Ecology*, 83 (7) (2012) 2027 - 2036

[9] - S. J. PHILLIPS, R. P. ANDERSON, M. DUDÍK, R. E. SCHAPIRE et M. E. BLAIR, *Ecography*, 40 (7) (2017) 887 - 893

[10] - K. PACIFICI, B. J. REICH, D. A. MILLER et B. S. PEASE, *Ecology*, 100 (6) (2019) 1 - 15

[11] - D. P. WILKINSON, N. GOLDING, G. GUILLERA-ARROITA, R. TINGLEY, et M. A. MCCARTHY, *Methods in Ecology and Evolution*, 10 (2) (2019) 198 - 211

[12] - G. GRENOUILLET, L. BUISSON, N. CASAJUS et S. LEK, *Ecography*, 34 (1) (2011) 9 - 17

[13] - W. THUILLER, *Global Change Biology*, 10 (12) (2004) 2020 - 2027

[14] - M. B. ARAÚJO, D. ALAGADOR, M. CABEZA, D. NOGUÉS-BRAVO et W. THUILLER, *Ecology Letters*, 14 (5) (2011) 484 - 492

[15] - M. MARMION, M. LUOTO, R. K. HEIKKINEN et W. THUILLER, *Ecological Modelling*, 220 (24) (2009) 3512 - 3520

[16] - W. THUILLER, B. LAFOURCADE, R. ENGLER et M. B. ARAÚJO, *Ecography*, 32 (2) (2009) 369 - 373

[17] - L. COMTE et G. GRENOUILLET, *Diversity and Distributions*, 19 (8) (2013) 996 - 1007

[18] - B. S. CADE, *Ecology*, 96 (9) (2015) 2370 - 2382

[19] - V. KOSHKINA, Y. WANG, A. GORDON, R. M. DORAZIO, M. WHITE, L. STONE, *Methods in Ecology and Evolution*, 8 (4) (2017) 420 - 430

[20] - Y. MUGUMAARHAHAMA, A. B. FANDOHAN, A. C. MUSHAGALUSA, I. A. SODE et R. GLELE KAKAÏ, Preprints, (2021), 2021040400 doi: 10.20944/preprints202104.0400.v1

[21] - A. C. ADOMOU, B. SINSIN et L. J. G. VAN DER MAESEN, *Systematics and Geography of Plants*, 76 (2) (2006) 155 - 178

[22] - P. J. PLATTS, P. A. OMENY et R. MARCHANT, *African Journal of Ecology*, 53 (2015) 103 - 108

[23] - J. M. LOBO, A. JIMÉNEZ-VALVERDE et R. REAL, *Global Ecology and Biogeography*, 17 (2008) 145 - 151
[24] - J. A. HANLEY et B. J. MCNEIL, *Radiology*, 143 (1982) 29 - 36
[25] - A. H. FIELDING et J. F. BELL, *Environmental Conservation*, 24 (1997) 38 - 49
[26] - O. ALLOUCHE, A. TSOAR et R. KADMON, *Journal of Applied Ecology*, 43 (6) (2006) 1223 - 1232
[27] - R. F. WUNDERLICH, Y-P. LIN, J. ANTHONY et J. R. PETWAY, *Nature Conservation*, 35 (2019) 97 - 116
[28] - S. JANITZA, C. STROBL et A-L. BOULESTEIX, *BMC Bioinformatics*, 14 (119) (2013) 10.1186/1471-2105-14-119
[29] - C. STROBL, A. L. BOULESTEIX, A. ZEILEIS et T. HOTHORN, *BMC Bioinformatics*, 8 (25) (2007) 10.1186/1471-2105-8-25
[30] - C. STROBL et A. ZEILEIS, in "*COMPSTAT 2008 – Proceedings of the 18th International Conference on Computational Statistics*", Ed. Brito P., Physica Verlag, Volume II, Heidelberg, (2008) 59 - 66
[31] - R. KADMON, O. FARBER et A. DANIN, *Ecological Applications*, 13 (3) 13 (2003) 853 - 867
[32] - G. SHMUELI, *Statistical Science*, 25 (3) (2010) 289 - 310
[33] - M. M. SYFERT, M. J. SMITH et D. A. COOMES, *PlosOne*, 8 (7) (2013) 10.1371/annotation/35be5dff-7709-4029-8cfa-f1357e5001f5