Quantification de l'occurrence et de l'abondance des plantes sauvages cueillies : un outil pour la gestion de la biodiversité

Chloé Mouillac

L'Institut Agro Rennes-Angers
☒ Site d'Angers ☐ Site de Rennes

Année universitaire : 2022-2023 Spécialité : Horticulture Spécialisation (et option éventuelle) : Génie de l'Environnement (option Préservation et Aménagement des Milieux - Ecologie Quantitative)	**Mémoire de fin d'études** ☒ d'ingénieur de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de master de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de l'Institut Agro Montpellier (étudiant arrivé en M2) ☐ d'un autre établissement (étudiant arrivé en M2)

Quantification de l'occurrence et de l'abondance des plantes sauvages cueillies : un outil pour la gestion de la biodiversité

Par : Chloé MOUILLAC

Garrigue à thym au nord de Montpellier, photo personnelle prise sur le terrain.

Soutenu à Rennes le 22/09/2023 devant le jury composé de :

Président : Loïs MOREL

Maîtres de stage : Guillaume PAPUGA
et Aurélien BESNARD

Enseignant référent : Didier LE COEUR

Autres membres du jury : Simon Dufour
Université de Rennes 2

Remerciements

Tout d'abord, un immense merci à Guillaume pour ton encadrement exceptionnel et la confiance que tu m'as offerte. J'ai beaucoup apprécié travailler à tes côtés, et je suis super heureuse de poursuivre mes 3 ans de thèse avec toi !

Merci aussi à Aurélien Besnard, malgré nos rencontres peu fréquentes, tu m'as toujours donné de précieux conseils et de bonnes pistes de réflexion. J'ai hâte de pouvoir profiter davantage de tes connaissances pour ces 3 années à venir.

Je remercie également Didier Le Coeur pour son suivi en tant qu'enseignant référent durant ce stage. Merci pour vos retours et vos commentaires qui m'ont aidé à avancer !

Merci à mes collègues de bureau : Aurélien, Rémi, Arthur et Clément pour la bonne ambiance, les fous rires et pour avoir su me remonter le moral quand ça n'allait pas (comme quand j'effaçais toutes mes données…).

Un grand merci à Arthur, Guillaume, Aurélien et Hadrien pour votre aide inestimable sur le terrain. Votre bonne humeur a largement compensé nos combats avec les fourrés de chênes verts et les pierriers !

Merci à toute l'équipe Pl@ntNet dont je ne peux citer tous les noms ici. Les discussions que nous avons pu avoir les midis ou aux pauses café m'ont vraiment aidé à prendre du recul sur mon travail, et donné de nouvelles idées.

Merci à Houssem pour les bons moments passés ensemble, et tes irruptions régulières dans mon bureau qui me faisaient prendre des pauses dont j'avais bien besoin.

Un grand merci à tous les autres stagiaires et doctorants que j'ai pu apprendre à connaître et avec qui j'ai pu partager des parties de ping-pong endiablées. Surtout Pablo et Clara, qui sont devenus de très bons amis, merci pour nos moments de folie !

Un merci tout particulier à Noémie pour ton énergie et ton optimisme débordants. Merci encore pour mon fabuleux bananier, qui, je l'espère, survivra à mes 3 ans de thèse avec moi.

Merci à l'équipe bivouac pour les incroyables souvenirs créés ensemble ! Grâce à vous j'étais en vacances tous les week-ends !

Finalement, merci à Max, mon copain, qui a toujours écouté mes découvertes d'une oreille passionnée et qui me soutient dans tous mes projets.

Table des matières

1. Introduction..... 1
2. Matériel et méthodes..... 6
2.1. Aire d'étude et espèce choisie..... 6
2.2. Données d'entrée des modélisations à très haute résolution..... 7
2.2.1. Données de présence..... 7
2.2.2. Variables explicatives..... 8
2.2.2.1. Données climatiques Bioclim..... 8
2.2.2.2. Données sur les propriétés du sol..... 9
2.2.2.3. Calcul d'indice de végétation à partir d'images Sentinel 2..... 9
2.2.2.4. Modèle numérique de terrain et calcul de variables topographiques..... 10
2.2.3. Pseudo-absences..... 10
2.3. Modélisation à très haute résolution de la distribution des espèces..... 11
2.3.1. Calibration et validation des modèles..... 12
2.3.2. Prédiction de l'occurrence du thym grâce aux modèles..... 13
2.3.3. Evaluation des modèles..... 13
2.3.4. Assemblage des modèles pour produire un modèle d'ensemble..... 13
2.4. Caractérisation in-situ de l'occurrence et de l'abondance du thym..... 14
2.4.1. Bilan méthodologique des métriques les plus communes..... 14
2.4.2. Protocole de mesures..... 14
2.4.3. Choix des zones d'échantillonnage..... 15
2.4.4. Validation de la méthode terrain..... 16
2.5. Comparaison des SDM aux données terrain : analyse statistique..... 17
2.5.1. Evaluation de la capacité des modèles à prédire la présence ou absence observée sur le terrain..... 17
2.5.2. Evaluation de la capacité des modèles à prédire l'abondance mesurée sur le terrain..... 18
3. Résultats..... 19
3.1. Résultats des SDMs à très haute résolution : évaluation de leur performance à partir des validations croisées..... 19
3.2. Résultats de l'étude de terrain..... 20
3.2.1. Etude préliminaire : cohérence et répétabilité des métriques choisies..... 21
3.2.1.1. Analyse du biais observateur..... 21
3.2.1.2. Puissance de l'échantillonnage pour les mesures des distances de recouvrement..... 22
3.2.2. Cohérence entre l'abondance mesurée et estimée à vue..... 22
3.3. Mise en regard de résultats des SDM à haute résolution et de l'étude de terrain : évaluation de la capacité des SDM à prédire des métriques de terrain..... 23
3.3.1. Evaluation de la capacité des modèles à prédire la présence observée sur le terrain..... 23
3.3.2. Evaluation de la capacité des modèles à prédire l'abondance mesurée sur le terrain..... 24
4. Discussion et perspectives..... 25
4.1. Variables à haute résolution : suffisantes pour prédire l'occurrence et l'abondance ?..... 25

4.2. Des modélisations réalistes : à quel prix ?.....27
4.3. Quelles perspectives pour une gestion locale de la biodiversité ?..... 28
5. Conclusion..... 30
Annexes..... 31
Bibliographie..... 64
Résumé..... 73

Liste des illustrations et annexes

Figure 1 : Niche climatique du thym (Thymus vulgaris) à l'échelle du Paléarctique-Ouest.. ……7
Figure 2 : Filtrage des données de présence utilisées pour les SDM à très haute résolution du thym.....8
Figure 3 : Méthode générale employée pour réaliser les SDM dans ce travail…..... 11
Table 1 : Tableau récapitulatif des algorithmes utilisés dans le cadre des modélisations à très haute résolution.....12
Table 2 : Récapitulatif des métriques (non destructives) couramment utilisées pour quantifier l'occurrence et l'abondance des plantes..... 14
Table 3 : Récapitulatif des métriques retenues pour l'étude de terrain, et leurs méthodes de mesure associées..... 14
Figure 4 : Schématisation du dispositif expérimental mis en place pour les mesures de thym sur le terrain..... 15
Figure 5 : Carte des zones d'échantillonnage prospectées lors de l'étude de terrain sur le thym..... 16
Figure 6 : Représentation schématique des zones tampon créées pour représenter un quadrat et son environnement.
Figure 7 : Scores TSS calculé par Biomod2 pour chacun des algorithmes utilisés dans le cadre de la modélisation à très haute résolution..... 19
Figure 8 : Cartes produites par les deux approches de modélisation à l'échelle locale, et différences inter-modèles.....20
Figure 9 : Variabilité des valeurs de recouvrement obtenues par 3 observateurs sur 9 quadrats (A a I)..... 21
Figure 10 : Coefficient de variation du recouvrement estimé à vue en fonction du coefficient de variation du recouvrement mesuré à partir des distances..... 21
Figure 11 : Evolution du recouvrement mesuré selon le nombre de cm inclus au sein des transects prospectés pour chaque quadrat (de A à I).....22
Figure 12 : Régression entre les valeurs (en %) de recouvrement estimé à vue et de recouvrement estimé à partir des mesures de distance..... 22
Figure 13 : Régression entre le nombre de touffes de thym touchées sur les 4 axes (Nord, Sud, Est et Ouest) d'un quadrat et l'abondance estimée de ce quadrat (à partir des mesures de distance).....22
Figure 15 : Distribution des points occurrences observées sur le terrain en fonction des prédictions issues du GAM et du modèle d'ensemble..... 23
Figure 16 : Représentation du score AIC en fonction du BIC pour chaque modèle. Les scores sont issus des GLMM réalisés pour évaluer la capacité des modèles à prédire l'occurrence.....24

Figure 17 : R² (coefficients de détermination) calculés à la suite de GLMM réalisés pour évaluer la capacité des modèles à prédire l'occurrence..... 24
Figure 18 : Distribution du recouvrement mesuré sur le terrain en fonction des prédictions issues du GAM et du modèle d'ensemble..... 24
Figure 19 : Représentation du score AIC en fonction du BIC pour chaque modèle..... 25
Figure 20 : R² calculés à la suite de GLMM réalisés pour évaluer la capacité des modèles à prédire l'abondance.....25
Annexe A : Mise en place d'un modèle de distribution du thym (Thymus vulgaris) à l'échelle du Paléarctique Ouest.....31
Figure A.1 : Carte des présence et des pseudo-absences utilisés pour la modélisation réalisée à l'échelle du Paléarctique Ouest.....32
Table A.1 : Tableau récapitulatif des variables utilisées en entrée de la modélisation réalisée à l'échelle du Paléarctique Ouest.....32
Figure A.2 : Matrice de corrélation des variables à sélectionner pour la modélisation réalisée à l'échelle du Paléarctique Ouest.....33
Figure A.3 : Rasters des variables retenues pour la modélisation réalisée à l'échelle du Paléarctique Ouest.....35
Figure A.4 : Probabilité de présence du thym (Thymus vulgaris) dans le Paléarctique Ouest d'après le modèle d'ensemble réalisé avec Biomod2..... 40
Table A.2 : Importance des variables (entre 0 et 100) pour chaque modèle individuel réalisé dans le cadre de la modélisation du thym à l'échelle du Paléarctique Ouest..... 40
Figure A.5.a : Courbes de réponse pour chaque algorithme et "run" (répétition) effectués dans le cadre de la modélisation du thym à l'échelle du Paléarctique Ouest..... 41
Figure A.5.b : Courbes de réponse pour le modèle d'ensemble effectué dans le cadre de la modélisation du thym à l'échelle du Paléarctique Ouest.....41
Table A.3 : Evaluation du modèle d'ensemble réalisé à l'échelle du Paléarctique Ouest, et des modèles individuels le constituant.....42
Figure A.6 : Poids des modèles intégrés dans la modélisation d'ensemble du thym à l'échelle du Paléarctique..... 42
Figure A.7 : Scores TSS calculé par Biomod2 pour chacun des algorithmes utilisés dans le cadre de la modélisation du thym à l'échelle du Paléarctique..... 42
Annexe B : Sélection des variables pour la modélisation du thym (Thymus vulgaris) à très haute résolution..... 44
Table B.1 : Tableau récapitulatif des variables utilisées en entrée de la modélisation réalisée à l'échelle locale..... 44
Figure B.3 : Matrice de corrélation des variables à sélectionner pour la modélisation réalisée à l'échelle locale..... 46
Figure B.4 : Rasters des variables retenues pour la modélisation réalisée à l'échelle locale.....46
Annexe C : Détails sur les modèles de distribution réalisés à l'échelle locale..... 49
Table C.1 : Importance des variables (entre 0 et 100) pour chaque modèle individuel réalisé dans le cadre de la modélisation du thym à l'échelle locale.....49
Table C.2 : Evaluation du modèle d'ensemble réalisé à l'échelle locale, des modèles individuels le constituant, et évaluation du GAM ajusté séparément.....49
Figure C.1 : Poids des modèles intégrés dans la modélisation d'ensemble à très haute résolution du thym.....49
Figure C.2 : Évolution de la variabilité inter-observateur de deux métriques selon le % de recouvrement..... 50
Table C.3 : Evaluation de la capacité des modèles réalisés à l'échelle locale de prédire la

présence du thym observée lors de l'étude de terrain, selon différentes métriques..... 51
Table C.4 : Coefficients issus des GLMM mis en place pour évaluer la capacité des modèles locaux à prédire la présence du thym observée lors de l'étude de terrain..... 51
Figure C.3 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM utilisés pour évaluer la capacité des modèles à prédire l'occurrence du thym..... 52
Figure C.4 : Distribution des points de présence/absence observés sur le terrain en fonction des prédictions issues de chaque modèle individuel réalisé dans le cadre de la modélisation du thym à l'échelle locale.....53
Table C.5 : Evaluation de la capacité des modèles réalisés à l'échelle locale de prédire l'abondance du thym mesurée lors de l'étude de terrain, selon différentes métriques..... 54
Table C.6 : Coefficients issus des GLMM mis en place pour évaluer la capacité des modèles locaux à prédire l'abondance du thym mesurée lors de l'étude de terrain..... 54
Figure C.5 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM utilisés pour évaluer la capacité des modèles à prédire l'abondance du thym..... 55
Figure C.6 : Distribution du recouvrement mesuré sur le terrain en fonction des prédictions pour chaque algorithme utilisé dans le modèle d'ensemble à très haute résolution.....56
Figure C.7.a : Courbes de réponse pour chaque algorithme et "run" (répétition) effectués dans le cadre de la modélisation du thym à l'échelle locale..... 57
Figure C.7.b : Courbes de réponse pour le modèle d'ensemble effectué dans le cadre de la modélisation du thym à l'échelle locale.....57
Annexe D : Choix de lissage pour le modèle GAM à l'échelle locale..... 58
Figure D.1 : Courbes de réponse produites pour le GAM selon trois niveaux de lissage, les coefficient retenu est k=5..... 58
Annexe E : Analyse spatiale de l'effet site dans les prédictions des modélisations à l'échelle locale.....59
Figure E.1 : Effet associé à chaque site d'échantillonnage (Montpellier est localisé par l'étoile rouge).....59
Annexe F : Analyses supplémentaires pour les corrélations des métriques de terrain 61
Table F.1 : Coefficients issus du GLMM mis en place pour évaluer la relation entre la hauteur et la longueur des touffes de thym.....61
Figure F.1 : Représentation de la hauteur de la touffe en fonction de sa longueur.....61
Figure F.4 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM mis en place pour évaluer la relation entre la hauteur et la longueur des touffes de thym..... 61
Table F.2 : Coefficients issus du GLMM mis en place pour évaluer la relation entre la distance moyenne à la première touffe et le recouvrement du thym dans un quadrat..... 62
Figure F.3 : Représentation du recouvrement en fonction de la distance moyenne à la première touffe de thym dans un quadrat..... 62
Figure F.4 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM mis en place pour évaluer la relation entre la distance à la première touffe, et le recouvrement du thym dans un quadrat..... 62
Annexe G : Fiche pour les relevés de terrain.....63

Glossaire

Abréviations et acronymes :

AFC : Association Français des professionnels de la Cueillette

AIC : Akaike Information Criterion

AUC : Area Under the Curve

BDD : Base De Données

BIC : Bayesian Information Criterion

CBN : Conservatoire Botanique National

CMR : Capture-Marquage-Remarquage

CNN : Convolutional Neural Networks

CTA : Classification and Regression Tree Analysis

GBIF : Global Biodiversity Information Facility

GBM : Gradient Boosting Models

GAM : Generalised Additive Models

GLM : Generalised Linear Models

GLMM : Generalised Linear Mixed-Effect Models

GPS : Global Positioning System

IA : Intelligence Artificielle

IGN : Institut Géographique National

IMCISE : IMpact de la CueIllette sur la dynamique des populations et la production primaire des plantes SauvagEs

INPN : Inventaire National du Patrimoine Naturel

logLik : log-likelihood

MARS : Multivariate Adaptive Regression Splines

MAXENT : MAXimum ENTropy

ML : Machine Learning

MNT : Modèle Numérique de Terrain

NDVI : Normalised Difference Vegetation Index

OFB : Organisme Français de la Biodiversité

PA : Pseudo-Absences

PCQM : Point-Centered Quarter Method

PMA : Prélèvement Maximal Autorisé

PNR : Parc Naturel Régional

RF : Random Forest

RMD : Rendement Maximum Durable

ROC : Receiver Operating Characteristic

R2 : R carré

SDM : Species distribution Modelling

TAC : Taux Autorisés de Capture

TPI : Topographic Position Index

TRI : Terrain Ruggedness Index

UICN : Union Internationale pour la Conservation de la Nature

Définitions :

Assemblage de modèles : ou "model averaging", consiste à ajuster plusieurs modèles de distribution d'espèces en utilisant plusieurs méthodes de modélisation et en prenant en compte différentes variables environnementales. Ensuite, au lieu de choisir un seul modèle comme étant le meilleur, l'assemblage permet de pondérer les prédictions de tous les modèles pour obtenir une prédiction globale plus robuste et précise. Il vise à réduire le risque de surajustement (overfitting) en prenant en compte l'incertitude associée au choix d'un modèle particulier.

Transférabilité : se réfère à la capacité d'un modèle préalablement entraîné sur des données de distribution d'espèces dans une région à être utilisé pour prédire la distribution d'espèces dans une autre région, en exploitant les similitudes potentielles entre les deux régions.

Surajustement : ou overfitting, d'un modèle se produit lorsqu'un modèle statistique ou algorithmique s'adapte trop étroitement aux données d'apprentissage, capturant non seulement les motifs réels mais aussi le bruit ou les fluctuations aléatoires présentes dans les données. En conséquence, le modèle peut avoir une performance exceptionnelle sur les données d'apprentissage, mais il généralise mal sur de nouvelles données non vues, car il a mémorisé des détails spécifiques aux données d'entraînement qui ne sont pas applicables plus largement.

Sensibilité : mesure la capacité d'un modèle à identifier correctement les vrais positifs parmi tous les cas positifs réels. En d'autres termes, c'est la proportion de vrais cas positifs que le modèle a correctement détectés par rapport à l'ensemble total des cas positifs. Une sensibilité élevée indique que le modèle a une forte capacité à détecter les cas réellement positifs.

Spécificité : mesure la capacité d'un modèle à identifier correctement les vrais négatifs parmi tous les cas négatifs réels. En d'autres termes, c'est la proportion de vrais cas négatifs que le modèle a correctement identifiés par rapport à l'ensemble total des cas négatifs. Une spécificité élevée indique que le modèle a une forte capacité à exclure les cas réellement négatifs.

Rééchantillonnage : dans le domaine de la cartographie et de la géomatique, processus qui consiste à ajuster la résolution spatiale (taille des pixels) d'une image raster ou d'une couche géospatiale, généralement en la réduisant ou en l'augmentant.

Variable réponse (ou expliquée) : la variable qui est cherchée à être expliquée ou prédite en fonction de variables indépendantes ou prédictives.

Variables prédictives (ou explicatives) : les variables utilisées pour expliquer ou prédire la variation d'une variable réponse. Elles sont utilisées pour établir des relations et des modèles afin de comprendre comment elles influencent la réponse.

Modèle linéaire à effets mixtes (GLMM) : étend le concept du modèle linéaire généralisé (GLM) en y ajoutant des effets aléatoires en plus des effets fixes classiques. Cette approche emprunte également l'idée des modèles linéaires mixtes pour adapter les GLM aux données qui ne suivent pas une distribution normale.

Effets aléatoires : sont des composantes ajoutées à un modèle statistique pour tenir compte de la variabilité non expliquée par les variables indépendantes incluses dans le modèle. Ils sont couramment utilisés dans les modèles à effets mixtes, qui combinent à la fois des **effets fixes** (liés aux variables explicatives) et des effets aléatoires (liés aux unités d'observation ou aux groupes).

Validation croisée : ou cross-validation, est une méthode d'estimation de la performance du modèle en utilisant des données distinctes de celles sur lesquelles le modèle a été ajusté. Cela permet de mesurer à quel point le modèle est généralisable et de prévenir le surajustement. En divisant les données en ensembles d'apprentissage et de test multiples, la validation croisée fournit une estimation plus robuste de la capacité du modèle à fonctionner avec de nouvelles données.

Coefficient de détermination (R^2) : Le coefficient de détermination, souvent désigné par "R carré" (R^2), est une mesure statistique qui évalue à quel point les variations d'une variable dépendante sont expliquées par les variations d'une ou plusieurs variables indépendantes dans un modèle de régression. Il varie entre 0 et 1, où 0 indique que le modèle n'explique aucune variabilité et 1 indique une explication parfaite de la variabilité. En d'autres termes, R^2 mesure l'ajustement du modèle aux données observées.

1. Introduction

La cueillette est une activité humaine qui remonte à la préhistoire. Elle consiste en le prélèvement de parties de plantes sauvages (fruits, fleurs, racines, feuilles, graines…), notamment pour la consommation alimentaire (Cunningham 2001). Elle constitua avec la chasse l'unique source de nourriture jusqu'à la révolution du néolithique (Martin 2014). L'apparition de l'agriculture et l'élevage a peu à peu supplanté la cueillette dans l'alimentation des sociétés de chasseurs-cueilleurs nomades. Cependant, les pratiques de cueillette se sont maintenues, complément essentiel pour maintenir une alimentation équilibrée mais également pour la médication (Cunningham 2001; Martin 2014). Progressivement, la cueillette a perdu de son importance avec le développement de l'agriculture. La plante sauvage développe même une image négative et est vue comme une ressource pour les pauvres et les animaux (Martin 2014). Ce n'est pas avant le XXème siècle, notamment avec le développement des industries pharmaceutiques et cosmétiques, que l'économie de la cueillette se développe (Julliand 2008). Aujourd'hui, la cueillette occupe une place importante dans notre société. Souvent oubliée et peu connue, elle n'en demeure pas moins essentielle pour bon nombre de médicaments, cosmétiques et autres produits que nous utilisons couramment. De cueillettes peu organisées et personnelles, nous observons aujourd'hui une transition vers des modes d'exploitation commerciaux (Labbé 2018). Un rapport très différent aux plantes se développe avec une augmentation des cueillettes productivistes au détriment des savoirs traditionnels (Lieutaghi 1991; Pailhès 2005). Ainsi, la demande en plantes sauvages ne cesse d'augmenter dans le monde entier et pose la question d'une cueillette durable (Jenkins, Timoshyna, et Cornthwaite 2018).

Le rapport 2018 de l'organisation Traffic, illustre bien la situation actuelle, et nous fait part de constats alarmants (Jenkins, Timoshyna, et Cornthwaite 2018). 60 à 90 % des plantes à parfum, aromatiques et médicinales commercialisées dans le monde sont issus de spécimens sauvages. Le commerce mondial déclaré de plantes à usages médicinales a été évalué à plus de 3 milliards de dollars US, soit une multiplication par trois entre 1999 et 2015. De surcroît, parmi les 30 000 plantes documentées pour leurs usages aromatiques et médicinaux, seuls 7% ont fait l'objet d'une évaluation de leur statut UICN à l'échelle globale, parmi lesquelles une sur cinq se trouvait menacée d'extinction à l'état sauvage. Cette pression croissante impacte fortement les populations, parfois même jusqu'à générer des adaptations chez celles-ci. Un exemple parlant est celui de *Fritillaria delavayi*, une plante utilisée traditionnellement en médecine chinoise. En réponse à une cueillette commerciale importante, la couleur de ses feuilles, originellement vertes, a évolué vers le gris/marron pour mieux se camoufler dans son environnement rocailleux (Niu, Stevens, et Sun 2021). Ainsi, la cueillette génère d'importantes pressions localement. Cependant, le manque criant d'information empêche d'évaluer les risques liés à la conservation des espèces touchées (Jenkins, Timoshyna, et Cornthwaite 2018).

La France ne fait pas exception à cette situation. En outre, de par sa position biogéographique au carrefour de l'Europe, elle bénéficie de l'influence des climats méditerannéen, alpin, océanique et continental, avec de nombreuses situations de rareté et de limites d'aire (Sagarin et Gaines 2002; Brown 1984; Lawton 1993). Ainsi, le pays présente une grande diversité de plantes avec environ 7000 taxons, dont au moins 10% sont cueillis. Selon les sources, ce sont entre 700 et plus de 1000 taxons qui feraient l'objet de prélèvements en France, avec 100 plantes cueillies régulièrement (Lescure et al. 2015). Il existe une grande variété de pratiques, et de profils de cueilleurs, allant des traditions locales individuelles (faibles volumes cueillis) à de grandes entreprises pharmaceutiques ou agroalimentaires qui ramassent tout sur leur passage. Les impacts sont donc très variés et peuvent générer des effets catastrophiques sur la flore locale. Dans les années 80, la cueillette abusive de la gentiane jaune par Pernod-Ricard pour l'industrie des apéritifs a complètement éradiqué l'espèce dans certaines zones d'Ariège (Laucoin 2012).

Avec la prise de conscience de la nécessité de conserver la biodiversité, la France a

développé un arsenal de conservation. Ainsi, de nombreux outils de régulations existent (flore protégée, régulations européennes…). Cependant, ces outils ont été le plus souvent pensés pour protéger des espèces rares. Peu d'espèces recueillies bénéficient de ces statuts, l'un des rares exemples de plantes cueillies protégées étant le Génépi laineux (*Artemisia eriantha*, protégé à l'échelle nationale et interdit à la cueillette) (Gargominy et Régnier 2023). Dans le cadre de la cueillette, la plupart des plantes sont communes et généralement jugées peu menacées. Ces groupes sont donc souvent moins étudiés, malgré tout, les CBN (Conservatoires Botaniques Nationaux) font des efforts pour y remédier avec, entre-autres, un groupe de travail inter-CBN sur la cueillette. En France, des arrêtés préfectoraux régulant la cueillette sont mis en place en réponse à des pressions locales importantes. Par exemple, dans les Hautes Alpes il est est interdit en tout temps et sur tout le territoire du département de cueillir plus de 100 brins de génépi jaune/blanc (*Artemisia genipi* Weber, *Artemisia glacialis* L., *Artemisia umbelliformis* Lam.) (Gargominy et Régnier 2023). Il existe aussi des réglementations spécifiques mises en place au sein des Parcs Nationaux pour tenter de s'adapter à la situation locale. Dans le cœur du parc national des Écrins, la cueillette des baies comme les myrtilles (*Vaccinium myrtillus* L.) est notamment limitée à 1 kg par jour et par personne (« Cueillettes et prélèvements » 2014). Cependant, à l'heure actuelle, l'impact des pratiques de cueillette sur cette biodiversité commune est très mal compris ce qui entrave la mise en place de réglementations adaptées.

En effet, à l'échelle internationale, peu d'études portent sur l'impact de la cueillette sur la dynamique des populations. En France, quelques études ont été faites ponctuellement sur certaines espèces cueillies, souvent par les Conservatoires Botaniques ou Parcs Naturels. Deux grandes études ont notamment été réalisées par le Conservatoire Botanique National Pyrénéen (CBNPMP) (Garreta et Morisson 2011), le Conservatoire Botanique National du Massif Central (CBNMC) (Laucoin 2012) et le Parc Naturel Régional des Monts d'Ardèche (Muraz et PNR Monts d'Ardèche 2018). Toutefois, il faut noter que les volumes de cueillette ont été définis seulement au niveau du CBNMC, il n'existe pas d'enquête à l'échelle nationale et surtout, pas de suivi dans le temps mis en place. Par ailleurs, quelques thèses traitant d'espèces cueillies (arnica, génépis notamment) ont vu le jour à Montpellier ces dernières années (Locqueville 2019; Fontaine 2021). Cependant l'approche reste centrée sur une espèce et non transférable à l'ensemble des plantes cueillies, à cause de leur variété de caractéristiques biologiques et de réponses à la cueillette. Nous pouvons noter de nombreuses études ponctuelles qui ont été menées, ainsi qu'un manque de communication entre les personnes qui les ont réalisées, empêchant l'élaboration d'une stratégie nationale (d'après une synthèse bibliographique personnelle).

Par conséquent, nous manquons de stratégie d'anticipation et d'outils de gestion de la ressource. Dans le cas de la pêche et de la chasse, des approches spatialisées visant une gestion durable des ressources ont été développées. Pour la chasse, il y a notamment le dispositif du prélèvement maximal autorisé (PMA) qui permet de limiter le nombre de captures par chasseur et par période (jour, semaine, année) sur un territoire déterminé (« Section 4 : Prélèvement maximal autorisé (Articles R425-18 à R425-20) - Légifrance » s. d.). Pour la pêche, des outils de gestion similaires existent : les TAC (Taux autorisés de capture) qui sont négociés chaque année par le Conseil européen « agriculture et pêche » à Bruxelles pour fixer les quotas de pêche dans les eaux européennes de l'Atlantique et de la mer du Nord. La notion de Rendement Maximum Durable (RMD) a aussi été intégrée, représentant la quantité de poissons qu'il est possible de pêcher à long terme sans perturber le cycle de reproduction (« Quotas de pêche : comment sont-ils fixés ? » 2019). Ces approches sont souvent intégrées dans des systèmes de gestion adaptative qui permettent d'adapter les prélèvements à l'état de la population en temps réel (Marboutin et et al. 2005; Bal et al. 2021). Cependant, encore peu de choses ont été faites pour les plantes, alors même qu'elles sont au centre des activités humaines.

Ainsi, pour répondre aux défis posés par l'augmentation de la pression de cueillette, une

première étape est la quantification de la ressource disponible pour les espèces qui subissent d'importants prélèvements. Cela permettra de mieux rendre compte de la situation, et d'établir des quotas de cueillette localement, afin d'adapter la pression à la ressource disponible. En effet, une même pression de cueillette n'aura pas forcément des impacts identiques sur deux sites différents. La cueillette à la limite de la répartition d'une espèce, où l'abondance est souvent moindre, peut rapidement menacer la ressource même avec une pression de cueillette modérée, tandis que cela ne poserait pas de problème au sein de son aire principale, où la ressource est généralement plus abondante (Brown 1984).

Pour évaluer la ressource, deux principales approches sont possibles : d'un côté, les approches "statiques" qui visent à quantifier de l'abondance d'un taxon sur un territoire à un instant t. Cela peut être fait à partir de modèles statistiques de distributions d'espèces (SDM), ou encore de méthodes de capture-marquage-remarquage pour estimer la taille de population. En parallèle, les approches "dynamiques" de suivi des évolutions temporelles de la ressource (à partir desquels sont établis les quotas de pêche ou chasse notamment) donnent des informations sur l'état de la population et sa dynamique. Cependant, pour les plantes, ces évaluations sont complexes et peu réalisées, en raison de contraintes de temps et d'argent imposées. Les contraintes sont d'autant plus importantes dans le cas des plantes cueillies puisque nous devons faire face à un grand nombre d'espèces aux caractéristiques (écologiques, biologiques, pratiques de cueillette…) très différentes.

En outre, l'un des principaux enjeux dans la quantification des plantes est la notion d'abondance. En effet, elle est protéiforme, et peut signifier plusieurs choses (« Diversité fonctionnelle des plantes » 2023). La première définition est celle du nombre d'individus, classiquement appliquée en écologie animale. Cependant, celle-ci est parfois mal adaptée aux plantes pour lesquelles les individus peuvent être difficiles à délimiter (plantes drageonnantes par exemple). Une autre définition est celle de l'occupation spatiale, qui peut être évaluée par une fréquence d'occurrence ou un recouvrement. Ce dernier est notamment très utilisé en écologie des communautés (Braun-blanquet 1932), mais qui ne donne pas d'information sur la biomasse disponible. Une dernière définition est celle de la biomasse, qui correspond à la masse de tissus produits. Dans le cadre de la cueillette, la mesure de biomasse est un proxy de choix pour estimer l'abondance, toutefois, elle nécessite généralement des méthodologies de terrain destructrices. Malgré tout, même les mesures de biomasse ne permettent pas d'évaluer directement la ressource dans le cas des prélèvements de sève, racines, fruits…

Dans le cadre de l'analyse spatiale des distributions d'espèces, les modèles de distribution d'espèces (SDM) ont gagné en popularité aux cours des dernières années. Il s'agit de méthodes statistiques qui permettent de prédire l'occurrence des organismes en faisant un lien corrélatif avec des variables spatialisées intégrées sous forme de raster. Ces variables peuvent fournir des informations climatiques (précipitations, température…), sur l'élévation, sur les propriétés du sol… Et beaucoup d'autres (Miller 2010). Les SDM se basent à l'origine sur la théorie de la niche écologique synthétisée par l'écologue britannique G.E. Hutchinson dans les années 1957 (Lamotte 1979). Cette théorie est fondamentale en écologie, et permet de comprendre comment les espèces coexistent au sein d'un écosystème complexe. Elle se concentre sur la manière dont les espèces occupent des positions spécifiques, appelées niches, au sein de leur environnement, en tenant compte de leurs besoins, de leurs interactions et des conditions abiotiques. Cette théorie reconnaît que chaque espèce a une série de préférences et de contraintes environnementales qui déterminent sa survie, sa croissance et sa reproduction.

Pour faire mettre en place les SDM, il faut des données d'occurrence. En général les SDM utilisent des données de présence seule, tirées de bases de données opportunistes (telles que les données de sciences participatives comme Pl@ntNet ou iNaturalist, ou alors les bases de données des Conservatoires Botaniques). Cependant, il n'existe le plus souvent pas de données d'absence avérée. Or beaucoup de modèles requièrent des données

d’absence pour fonctionner. Il existe donc multiples méthodologies pour générer ce qu’on appelle des “pseudo-absences” (Barbet-Massin et al. 2012). Les choix pour les générer sont cruciaux pour la représentativité finale des modélisations.

Les SDM sont largement utilisés pour caractériser la distribution d’une grande variété d’espèces, allant de l’évolution de l’aire de distribution des ours polaires face au changement climatique (Durner et al. 2009), à la modélisation de la progression d’espèces envahissantes (Barbet-Massin et al. 2018). Il existe une très large diversité d’approches statistiques, depuis de simples approches corrélatives (GLM ou Generalised Linear Models) aux algorithmes de Machine Learning (type Maxent) qui utilisent de l’intelligence artificielle (IA) et pour certains des réseaux neuronaux convolutifs (CNN). Nous pouvons aussi trouver des algorithmes basés sur des arbres de décision (CTA, BRT, RF…). Une approche populaire en écologie consiste à assembler les modèles pour avoir un résultat supposé plus solide (Hao et al. 2019; Marmion et al. 2009). En effet, au lieu de choisir un seul modèle comme étant le meilleur, l’assemblage permet de pondérer les prédictions de tous les modèles pour obtenir une prédiction globale plus robuste et précise qui prend en compte l'incertitude associée au choix d'un modèle particulier. Il vise à réduire le risque de surajustement* tout en utilisant des modèles plus performants (Hao et al. 2019). Chacune des approches a ses points forts et ses limites. Les modèles plus “simples” type GLM sont plus facilement interprétables d’un point de vue écologique, tandis que ceux utilisant le Machine Learning sont accusés d’être des modèles de type “boîte noire” mais souvent plus performants. Toutefois, la performance des modèles n’est généralement évaluée qu’en théorie, en utilisant le jeu de données de base, mais en n’allant pas sur le terrain pour vérifier la vraisemblance des résultats (Hao et al. 2019). En effet, l’évaluation se fait habituellement en mettant de côté une partie du jeu de données d’entrée pour ensuite voir la capacité du modèle, généré avec le reste données, à prédire les données initialement mises de côté. Des scores peuvent ensuite être attribués grâce à des métriques de type TSS (True Skill Statistic).

Historiquement, les SDM sont plutôt utilisés avec des variables climatiques, de type Bioclim ou Chelsa, à grande échelle (Booth 2018). Cependant, pour le cas des plantes, ces modèles demeurent peu informatifs au niveau local car ils définissent la “macro-niche” de l’espèce sans rendre compte des variations fines définissant les “micro-niches” (Papuga et al. 2018). Au cours des 10 dernières années, le développement de variables à haute résolution (à partir de données de télédétection, de modèles numériques de terrain…) laisse entrevoir des possibilités de prédiction à l'échelle de gestion des habitats naturels (<100m). Cela est actuellement très peu fait. Nous pouvons citer quelques modélisations d’écosystèmes à 5 m de résolution (Singer et al. 2016; Jähnig et al. 2012; Álvarez-Martínez et al. 2014). Il existe aussi quelques applications techniques des SDM pour des animaux et des plantes (Engler et al. 2013; Buckland et al. 2014; Pradervand et al. 2014) dont une seule suggère l’utilisation de ces SDM pour une meilleure gestion de la biodiversité (Pradervand et al. 2014). Ces approches seraient alors extrêmement utiles pour les plantes cueillies, particulièrement dans le cas de la gestion de la ressource. Cela permettrait une prédiction précise de leur occurrence, et éventuellement de leur abondance dans le cadre d’une quantification de la ressource. Cependant beaucoup de questions restent en suspens pour ces SDM à haute résolution.

Ces approches corrélatives ne permettent pas de réellement savoir ce qui est modélisé. Le lien statistique établi entre les variables et la présence de l’espèce peut être influencé par la qualité des données d’entrée : leur résolution, le biais d'échantillonnage des occurrences utilisées. Des questions peuvent également se poser quant à la pertinence biologique des variables et leur capacité à donner une représentation réaliste de la niche de l’espèce. Il peut aussi y avoir un surajustement* des données, c’est-à-dire que le modèle apprend à très bien prédire le jeu de données qui lui est fourni en entrée, mais est incapable de prédire un jeu de données issu d’une autre source (différente origine géographique par exemple). Cela est causé par la création d'information là où il n’y en a pas, générant alors des relations entre les variables prédictives et la réponse qui n’ont pas de sens par rapport à la niche écologique de

l'espèce. Il s'agit de relations "spurieuses" qui impactent l'efficacité des SDM (Garsd 1984; Fourcade, Besnard, et Secondi 2018). Cela limite fortement la validité et la transférabilité* des modèles (Breiner et al. 2015). Il faut alors choisir entre des modèles éventuellement plus grossiers (type GLM), mais plus en phase avec la théorie de la niche attendue, ou des modèles qui s'ajustent mieux aux données (modèles de Machine Learning), au prix du réalisme de la prédiction. Ceux-ci posent souvent des problèmes de transférabilité* dû au surajustement*, et nécessitent par conséquent une validation par un jeu de données indépendant. Malgré tout, si la représentation de la niche est correcte, une grande probabilité de présence traduit des conditions plus favorables à l'espèce. Ainsi, plusieurs auteurs ont émis l'hypothèse que les SDM pourraient non seulement prédire l'occurrence mais aussi l'abondance d'une espèce (Van Couwenberghe et al. 2013; Young et Carr 2015). En effet, la notion de centralité de la niche laisse supposer des zones plus favorables, et potentiellement une plus grande abondance de population (Sagarin et Gaines 2002; Brown 1984; Lawton 1993). Cette hypothèse est l'un des axes abordés lors de cette étude.

Cette étude s'inscrit dans le cadre d'un contrat de recherche et développement financé par l'OFB (Office Français de la Biodiversité) et relatif à la quantification de l'impact de la cueillette sur la dynamique des populations et la production primaire des plantes sauvages (projet IMCISE). L'objectif est d'étudier, en collaboration avec les Conservatoires botaniques nationaux (CBN) et l'Association française des professionnels de la cueillette de plantes sauvages (AFC), la mise en place de protocoles de suivis et de veille pour les espèces actuellement soumises à la cueillette. Ces méthodes permettront de mieux comprendre l'état de conservation des populations de plantes, et d'adapter les programmes de gestion de la ressource ensuite. Ces questionnements seront traités conjointement avec les acteurs de la cueillette et de la conservation de la nature, tels l'OFB, les Parcs nationaux et Parcs Naturels Régionaux. Les principaux objectifs du projet sont 1. d'améliorer la connaissance sur la cueillette de plantes sauvages, à travers un état des lieux des pratiques actuelles de cueillette, 2. de quantifier la ressource au niveau régional grâce à des modèles d'occurrence et d'abondance et 3. de mettre au point des suivis locaux de l'impact de la cueillette. Ce projet doit durer trois ans, les travaux préliminaires produits lors de ce stage seront poursuivis et approfondis lors d'une thèse que je mènerai à partir d'octobre 2023.

Pour le cas d'étude traité dans ce manuscrit, l'aire d'étude a été réduite à une zone autour de Montpellier, située au cœur du bassin Méditerannéen, un lieu de choix pour étudier la cueillette. En effet, celui-ci abrite une riche biodiversité qui inclut une grande variété de plantes cueillies parmi lesquelles le thym, romarin, lavande, ciste, aubépine, le fenouil, l'asperge sauvage… et une multitude d'autres. Certaines des ces espèces subissent d'importantes pressions de cueillette, cependant il n'existe encore aucune réglementation mise en place pour la réguler, mise à part l'interdiction de cueillir les espèces protégées (Gargominy et Régnier 2023). Or cela ne touche qu'un nombre restreint d'espèces dont très peu d'entre elles font l'objet de prélèvements. Le thym fait partie des espèces qui subissent localement des cueillettes excessives, et il est notamment inscrit dans la liste prioritaire des plantes cueillies de l'AFC (Association Française des professionnels de la Cueillette de plantes sauvages 2019). Cette espèce à distribution Méditerranéenne m'a servi de plante modèle pour mon étude. Malgré les constatations concernant des cueillettes excessives, il n'existe pas de données quantifiées à ce sujet. Ce sont souvent des dires de cueilleurs qui voient la ressource décliner localement d'année en année, et qui lancent un signal d'alarme.

Dans ce contexte, les SDM nous ont semblé être une bonne piste pour répondre aux défis posés par le suivi des plantes cueillies. En effet, ils permettent de prédire des distributions d'espèces à partir de peu de données. C'est une solution intéressante qui permettrait de faire des suivis plus généralisés, sur plus d'espèces en dépensant moins. Cette étude vise ainsi à établir une meilleure compréhension des SDM utilisés pour faire des prédictions à haute résolution, en utilisant des variables d'entrée très précises.

La question à laquelle je me suis confrontée lors de cette étude est donc : **Peut-on prédire l'occurrence et l'abondance des plantes sauvages à fine échelle spatiale ?**

Afin d'y répondre, trois grands objectifs ont été définis :

1) **Prédire à fine échelle l'occurrence d'une plante cueillie modèle**, le thym, à travers deux approches : le modèle d'ensemble et le GAM (Generalised Additive Model).
   Nous émettons ici l'hypothèse que les données issues de sciences participatives et des CBN sont suffisantes pour estimer précisément la présence d'une espèce.

2) **Concevoir un protocole d'échantillonnage sur le terrain pour tester la capacité des SDM à prédire différentes métriques d'occurrence et d'abondance.**
   Tester la corrélation entre ces métriques.
   Valider la reproductibilité de protocole, le biais observateur, la stabilité des métriques.

3) **Tester la qualité des modèles face à la réalité terrain** : vérifier si les modèles arrivent à prédire les différentes métriques produites (occurrence et abondance).
   L'intervention des données de terrain permettra de comparer la performance relative des deux approches de modélisation choisies. Par ailleurs, elle nous permettra de voir s'il est possible, à partir de modèles d'occurrence, d'inférer l'abondance d'une espèce en posant l'hypothèse qu'une niche favorable sera corrélée à une plus forte abondance

La suite de ce mémoire sera organisée selon la structure classique d'un article scientifique avec un matériel et méthodes, suivi des résultats, d'une discussion et enfin d'une conclusion. Chaque partie sera sous-divisée selon les objectifs énoncés ci-avant.

2. Matériel et méthodes

2.1. Aire d'étude et espèce choisie

Dans le cadre de ce stage, j'ai décidé de travailler sur le thym qui est une plante déjà bien étudiée et qui subit d'importantes pressions de cueillette. Le thym (*Thymus vulgaris* L.) est une plante herbacée vivace appartenant à la famille des Lamiacées (Tison et Foucault 2014). C'est une plante emblématique des régions méditerranéennes, mais elle a été largement cultivée et naturalisée dans d'autres parties du monde en raison de sa popularité. Cette plante affectionne les sols secs et bien drainés, ainsi que les environnements ensoleillés. Nous la retrouvons fréquemment sur des collines, des pentes rocailleuses et dans les garrigues. Les feuilles du thym sont riches en huiles essentielles, principalement le thymol et le carvacrol, qui lui confèrent son arôme et son goût caractéristiques. En cuisine, le thym est utilisé en tant que condiment pour parfumer divers plats. En plus de son rôle culinaire, le thym possède des propriétés médicinales appréciées depuis l'Antiquité. Il est réputé pour ses effets antiseptiques, antioxydants et anti-inflammatoires. Outre son utilisation en cuisine et en phytothérapie, le thym est également employé dans l'industrie cosmétique et pharmaceutique pour la production d'huiles essentielles, de teintures et d'extraits. Les infusions de thym sont connues pour leurs bienfaits sur les affections respiratoires. Sur le plan écologique, le thym attire une variété d'insectes pollinisateurs, ce qui en fait une plante bénéfique pour la biodiversité locale. Cependant, dans certaines régions, une surexploitation peut avoir des conséquences négatives sur les populations locales de cette plante, soulignant l'importance d'une gestion durable de sa récolte (Agribio 04 et Agribio 05 s. d.). L'Association Française des Cueilleurs l'a notamment placée dans la liste des 55 espèces prioritaires qui font face aux plus grands enjeux de cueillette (Association Française des professionnels de la Cueillette de plantes sauvages 2019). En outre, dans le cadre de l'étude de terrain, ma localisation sur Montpellier était idéale pour couvrir une grande partie du gradient de distribution de cette espèce Méditerranéenne (Fig. 1).

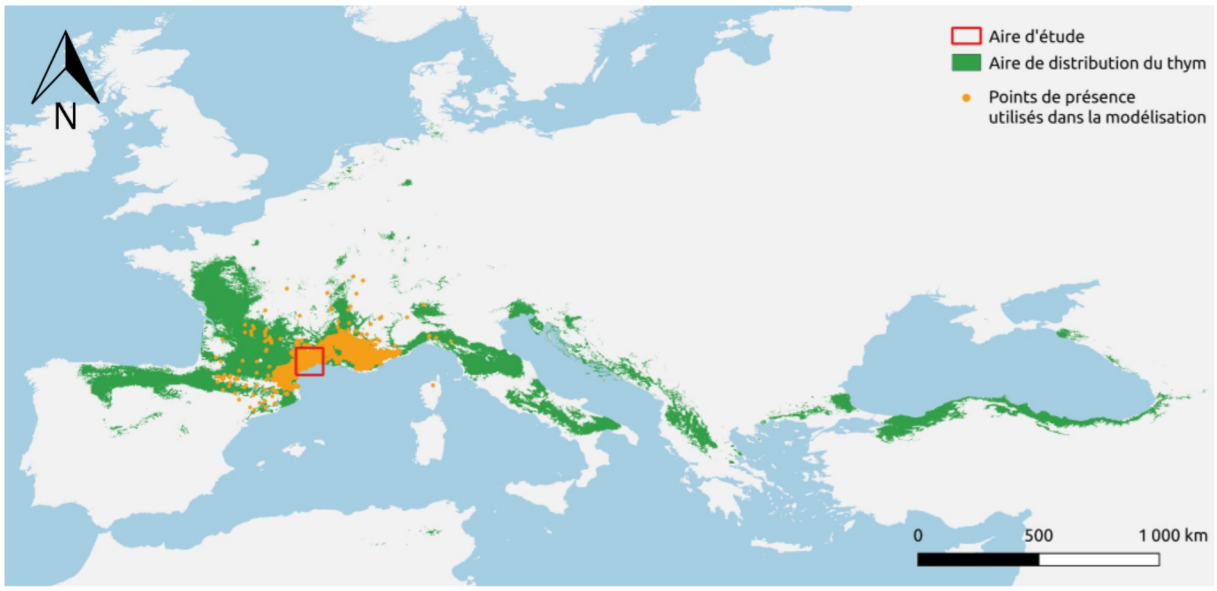


Figure 1 : Niche climatique du thym (*Thymus vulgaris*) à l'échelle du Paléarctique-Ouest. Les résultats sont obtenus à partir d'une modélisation préliminaire de résolution 1km² réalisée avec *Biomod2* (voir Annexe A).

Le travail effectué dans ce stage vise au au développement d'une méthode utilisable à plus large échelle plus tard. L'objectif sera en effet d'avoir une méthodologie applicable à l'ensemble des espèces cueillies afin de quantifier la ressource disponible, ainsi que les enjeux de pression de cueillette à l'échelle locale.

2.2. Données d'entrée des modélisations à très haute résolution

Le premier volet de cette étude s'est focalisé sur la modélisation de la distribution du thym à très haute résolution (10 m). Pour cela, j'ai utilisé les données de présence issues du GBIF (Global Biodiversity Information Facility) ainsi que celles du CBN méditérannéen. Utiliser les sciences participatives (GBIF) est particulièrement intéressant en raison de la quantité d'observations accumulées et de leur évolution constante. Afin d'ajuster les modèles, j'ai obtenu des données climatiques, sur la composition du sol ainsi que des données satellite et topographiques dont je détaillerai l'utilisation ci-après. La résolution de 10 m a été choisie car les variables intégrées dans les SDM ont généralement une résolution maximale de 10 m. Par ailleurs, étant donné l'incertitude liée à la position GPS des occurrences intégrées au modèle (2 à 10 m), il n'aurait pas été cohérent de choisir une résolution plus élevée.

2.2.1. Données de présence

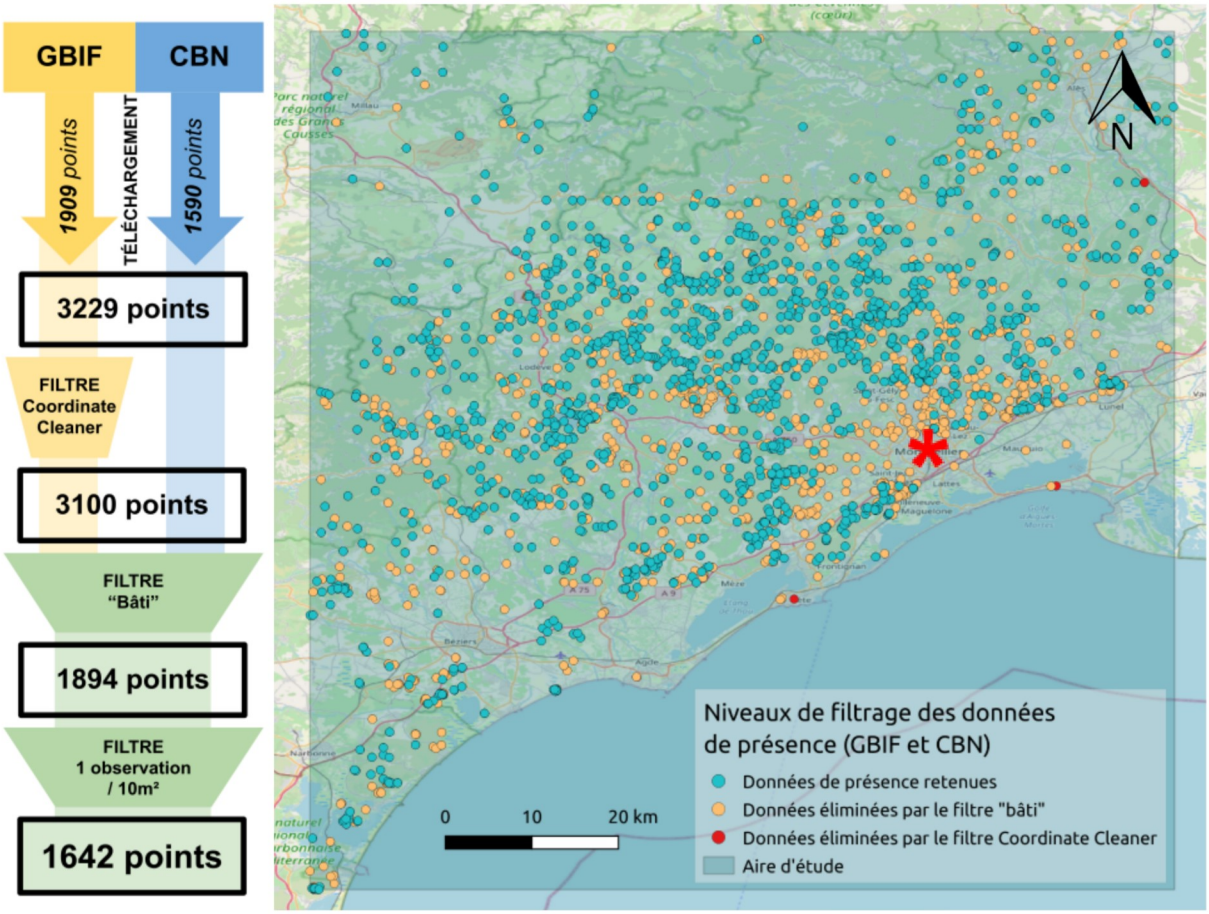
Avant l'intégration des données de présence dans les modèles, j'ai réalisé plusieurs étapes de nettoyage pour améliorer la qualité des données (Fig. 2).

Tout d'abord, lors du téléchargement des données GBIF (grâce au package *rgbif* sur R), seules les données présence ayant une précision inférieure à 10 m ont été gardées, de même pour les observations faites avant l'an 2000 et celles sans coordonnées. De plus, les "flags" du GBIF ont été utilisés pour éliminer toutes les données ayant des "problèmes d'ordre géospatial", soit les observations dont les coordonnées sont invalides, ou arrondies (Buitrago 2020). Les données téléchargées ont ensuite été soumises à un nettoyage permis par le package *CoordinateCleaner* sous R (Zizka et al. 2019). Celui-ci comprend des tests automatisés permettant de repérer facilement (et d'exclure) les enregistrements attribués au centroïde d'un pays ou d'une province, à la mer, au siège du GBIF, ou à l'emplacement des

institutions chargées de la biodiversité (musées, zoos, jardins botaniques, universités). Il identifie en outre, pour chaque espèce, les coordonnées aberrantes, les coordonnées nulles, les latitudes/longitudes identiques et les coordonnées non valides.

Ensuite, pour les données du CBN Méditérannéen, seules les observations faites depuis l'an 2000 avec une précision de niveau GPS ont été gardées (environ 2 à 10 m).

Les données de présence issues du GBIF et du CBN ont alors été réunies dans un même tableau, puis j'ai créé un filtre pour éliminer les observations situées au niveau de zones habitées. Ce filtre vise à éliminer les observations faites dans les jardins ou parcs qui correspondent généralement à des plantes cultivées, et non à la distribution naturelle de l'espèce. L'intégration de ces observations dans les modèles risquerait de biaiser les résultats. Ainsi, le filtre a été créé à partir du Plan Cadastral Informatisé à laquelle ont été rajoutés 200m de zone tampon autour de chaque zone construite. Il faut noter que l'élimination des données situées en zone urbaine est l'une des options du package *CoordinateCleaner* présenté plus tôt. Cependant, ce filtre est bien trop grossier (il ne considère que les grandes villes), c'est pourquoi j'ai décidé d'en faire un qui soit plus précis (Annexe B Fig. B.2). Enfin, j'ai réduit le nombre de points pour n'en garder qu'un seul par



cellule de 10m² (en se basant sur le rasters utilisés dans les modélisations).

Figure 2 : Filtrage des données de présence utilisées pour les SDM à très haute résolution du thym. Montpellier est localisé par l'étoile rouge. *(source personnelle)*

(Remarque : *CoordinateCleaner* n'a pas éliminé beaucoup de points, cela s'explique par le filtrage préalable au niveau du téléchargement des données GBIF).

2.2.2. Variables explicatives

Afin d'ajuster les modèles, différentes variables explicatives ont été utilisées (toutes les cartes associées peuvent être trouvées en Annexe B Fig. B.4.

2.2.2.1. Données climatiques Bioclim

En premier, j'ai téléchargé les données bioclimatiques issues de WorldClim (« Bioclimatic variables - WorldClim 1 documentation » s. d.) en utilisant le package *geodata* sous R (fonction *worldclim_country*) (« R: WorldClim climate data » s. d.). Ces ensembles de données peuvent être utilisés pour expliquer les distributions potentielles des espèces en fournissant des variables biologiquement significatives qui transmettent les conditions climatiques moyennes annuelles et saisonnières ainsi que la saisonnalité intra-annuelle (Hijmans et al. 2005). Il s'agit de 19 variables dérivées des valeurs températures et précipitations mensuelles moyennes (entre 1950 et 2000). Ce sont les variables les plus couramment utilisées pour les modélisations de distribution d'espèces (Booth 2018; Ndlovu et al. 2018). Elles permettent notamment de définir le préférendum des espèces. Ces données existent à différents niveaux de résolution, mais dans le cadre de mon étude, ce sont les données les plus précises que j'ai choisies, soit avec un précision de 30 arcsec (correspondant à environ 1km²). Pour que ces données soient utilisables dans mes modèles, les images ont été rognées pour correspondre à la zone d'étude. Ensuite, j'ai rééchantillonné* les rasters à un taille de pixel de 10 m. Les valeurs utilisées sont donc extrapolées, mais j'ai estimé que la variabilité des données bioclimatiques au sein d'1km² serait négligeable.

2.2.2.2. Données sur les propriétés du sol

Ensuite, j'ai obtenu des données sur les propriétés du sol grâce à SoilGrids, qui fournit une cartographie mondiale, à 250m de précision, des caractéristiques du sol (« SoilGrids » s. d.). Ces cartes sont produites à partir de données environnementales et plus de 230 000 profils de sols qui sont intégrés dans un modèle de machine learning. Les propriétés décrites sont le pH, la masse volumique du sol, la quantité de nitrogène contenue, la CEC (capacité d'échange cationique), la fraction de fragments grossiers (>2 mm, cfvo), la concentration en carbone organique (soc), ainsi que les fractions de sable (sand), argiles (clay) et limons (silt). Cette base de données précise les propriétés du sol sur 6 strates : 0-5 cm de profondeur, 5-15 cm, 15-30 cm, 30-60 cm, 60-120 cm et 120-200 cm. Pour mes modèles, j'ai choisi de prendre les données des 3 premières strates (0 à 30 cm) que j'ai moyennées. Le thym s'enracine souvent assez superficiellement, allant à une profondeur de quelques dizaines de centimètres, c'est pourquoi je n'ai pas intégré les données des strates à plus de 30 cm de profondeur (Al-Ramamneh 2009).

Les propriétés du sol sont des données souvent intégrées dans les SDM car elles permettent, entre autres, de déterminer la disponibilité en eau et en nutriments du sol, qui sont essentielles dans le définition de l'habitat d'une espèce (Coudun et al. 2006; Dubuis et al. 2013).

De même que pour les données Bioclim, pour que ces données soient utilisables dans mes modèles, les images ont été rognées pour correspondre à la zone d'étude. Ensuite, j'ai rééchantillonné* les rasters à un taille de pixel de 10 m.

2.2.2.3. Calcul d'indice de végétation à partir d'images Sentinel 2

Après, j'ai obtenu des images satellite Sentinel 2 qui m'ont permis de calculer le NDVI sur mon aire d'étude. Le NDVI, ou Normalised Difference Vegetation Index, est une métrique largement utilisée pour estimer la couverture végétale d'un sol et sa biomasse (Schwager et Berg 2021; Amaral et al. 2023). Le NDVI est souvent considéré comme un outil essentiel pour évaluer l'activité photosynthétique et la productivité, et pour fournir des informations sur la dynamique de la végétation, ou pour décrire la configuration structurelle de la végétation (Pettorelli et al. 2011). Ainsi, il est particulièrement utile dans les modèles de distribution

d'espèces pour identifier des types de couverts végétaux associés à l'espèce modélisée. L'intégration du NDVI permet aussi d'identifier les surfaces de sol nu, les plans d'eau et les surfaces artificialisées qui ne permettront pas le développement de l'espèce.

En passant par la plateforme Copernicus Open Access Hub (dépendant de l'Agence Spatiale Européenne) (« Open Access Hub » s. d.), j'ai téléchargé toutes les images disponibles entre juin 2022 et juin 2023, sur les tuiles 31TEK, 31TEJ et 31TEH ayant une couverture nuageuse inférieure à 10%, et ayant une précision de 10 m. J'ai sélectionné les images L2A (à 10 m) qui sont déjà pré-traitées utilisant Sen2Cor pour corriger les effets atmosphériques, radiométriques et géométriques, ce qui les rend plus rapidement prêtes à l'emploi. J'ai ensuite assemblé les images en mosaïque sous R, et effectué une moyenne annuelle sur les bandes 4 et 8 nécessaires pour le calcul du NDVI, donné par la formule :

$$NDVI = (NIR - R) / (NIR + R) = (B08 - B04) / (B08 + B04).$$

NIR : near infrared, ou rayonnement infrarouge proche
R : red, ou bande de rayonnement rouge

Enfin, les images ont été rognées pour correspondre à la zone d'étude. Ici, le rééchantillonnage* n'était pas nécessaire.

2.2.2.4. Modèle numérique de terrain et calcul de variables topographiques

Les dernières variables explicatives intégrées aux modèles sont des variables d'élévation et topographiques. A partir de l'IGN, j'ai pu télécharger les données RGE Alti, qui est un MNT (Modèle Numérique de Terrain, autrement dit les valeurs d'élévation) avec une résolution d'1 m (« RGE ALTI® | Géoservices » s. d.). J'ai tout d'abord utilisé le MNT tel quel dans mes modèles, en dégradant les données pour passer d'une résolution d'1 m à 10 m. Puis, j'ai ensuite calculé des variables dérivant de ce MNT grâce à la fonction *terrain* du package *terra* sous R (« Terrain: Terrain Characteristics in Terra: Spatial Data Analysis » s. d.). Ainsi, j'ai pu obtenir : la pente, "l'aspect" du terrain et le TPI (Topographic Position Index). L'aspect correspond à orientation de la pente en degrés, tandis que le TPI est différence entre la valeur d'une cellule et la moyenne des 8 cellules adjacentes. Par ailleurs, il était possible de calculer 2 autres variables : roughness (ou rugosité) et TRI (Terrain Ruggedness Index), mais qui étaient toutes deux corrélées à 99% avec la pente, donc inutiles. Le TRI est défini par la valeur absolue de la différence entre la valeur d'une cellule et la moyenne des 8 cellules adjacentes, et pour Roughness, il s'agit de la différence entre le maximum et le minimum d'une cellule et de ses 8 cellules adjacentes

J'ai sélectionné la variable élévation car il s'agit d'un incontournable des SDM. En effet, elle joue un rôle significatif dans la distribution des espèces à travers son influence sur les gradients de température, le climat et les gradients altitudinaux notamment (Lannuzel et al. 2021). De plus, les variables du terrain influencent le type de sol, l'humidité du sol, l'angle du soleil, les précipitations et par conséquent la distribution de la végétation (Bennie et al. 2006). J'ai alors intégré la variable aspect car elle donne des informations sur l'exposition au soleil et l'hydrologie, des facteurs qui peuvent être essentiels au développement d'une espèce. Par exemple, les pentes orientées vers le sud ont tendance à recevoir plus de lumière du soleil et des températures plus chaudes, tandis que les pentes orientées vers le nord peuvent être plus fraîches et plus humides. Enfin, le TPI m'a semblé une variable pertinente à ajouter à la modélisation car elle mesure les variations locales de topographie et permet ainsi de rendre compte de l'hétérogénéité de l'habitat à échelle très fine. Ces variations peuvent affecter des facteurs tels que la prise au vent, les gradients de température ou encore la structure de la végétation. Intégrer le TPI permet donc de caractériser la réponse des espèces à ces changements subtils (Seif 2014; Weiss 2001).

2.2.3. Pseudo-absences

Peu de SDM reposent uniquement sur des données de présence. Les plus communs sont les modèles de type BIOCLIM et ceux basés sur les distances de Mahalanobis. La difficulté d'acquérir des données d'absence confirmées, nécessitant un échantillonnage plus important pour assurer leur fiabilité par rapport aux données de présence (Mackenzie et Royle 2005), a conduit à l'utilisation fréquente de modèles basés uniquement sur les données de présence (Graham et al. 2004). Cependant, les comparaisons entre différentes méthodes de modélisation montrent que les modèles prenant en compte à la fois la présence et l'absence tendent à présenter de meilleures performances que ceux se basant uniquement sur la présence (Elith* et al. 2006). Par conséquent, les modèles de présence-absence sont de plus en plus préférés lorsque seules les données de présence sont disponibles. Dans ces cas, des données d'absence artificielles sont générées (souvent appelées pseudo-absences ou données de fond) pour enrichir le jeu de données et améliorer la qualité des prédictions des modèles. Ici, j'ai décidé de générer autant de points d'absence (Annexe B. Fig. B.1) que de points de présence, de manière randomisée sur la totalité de l'aire d'étude (Barbet-Massin et al. 2012).

2.3. Modélisation à très haute résolution de la distribution des espèces

Avec l'ensemble des données récoltées, deux approches de modélisation ont été menées, toutes les deux avec le package *Biomod2* sous R (Fig. 3) (Miller 2010). *Biomod2* comprend un ensemble d'outils statistiques et de modélisation pour l'analyse de données de biodiversité et d'écologie. Il permet aux utilisateurs de modéliser la distribution des espèces en fonction de variables environnementales en utilisant différentes méthodes de modélisation. L'une des caractéristiques clés de *Biomod2* est la possibilité de créer des modèles d'ensemble en combinant plusieurs algorithmes de modélisation pour obtenir des prédictions plus robustes et précises. Le package propose également des outils pour sélectionner les variables environnementales les plus pertinentes pour la modélisation. Par ailleurs, il intègre des méthodes de validation croisée* pour évaluer la performance des modèles et évaluer l'apport de chaque variable.

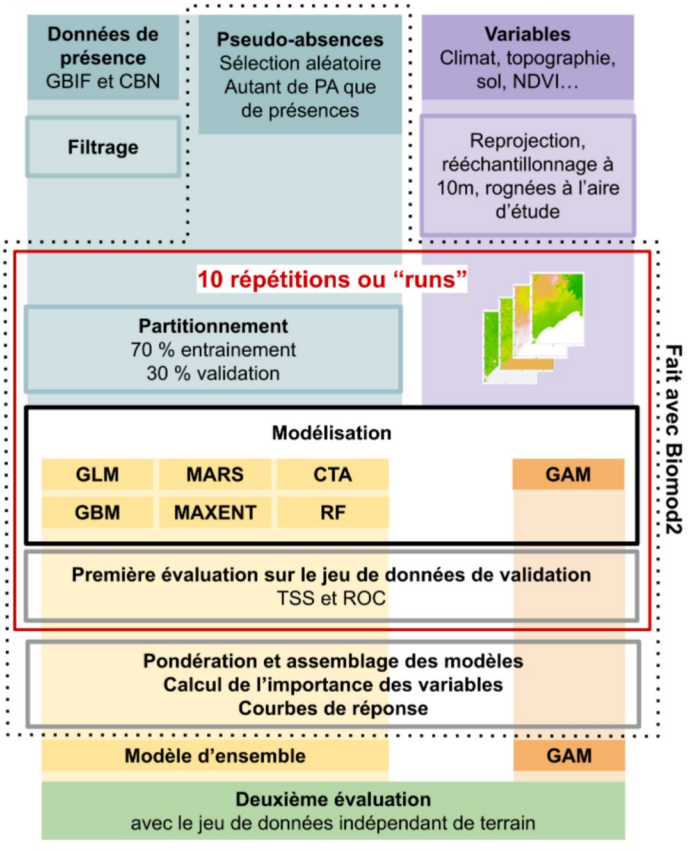


Figure 3 : Méthode générale employée pour réaliser les SDM dans ce travail *(source personnelle)*.

Ainsi, *Biomod2* m'a permis d'un côté d'ajuster des GAM (ou Generalised Additive Models) individuellement, et d'un autre côté, de réaliser un modèle d'ensemble*.
J'ai choisi de confronter mon modèle d'ensemble à un GAM, car c'est un bon intermédiaire entre les modèles trop simples de type GLM, qui ne sont pas capables de capturer l'ensemble des relations entre les prédicteurs et la réponse, et les modèles de machine learning qui produisent des relations difficilement interprétables d'un point de vue écologique.

En effet, les GAM sont utilisés pour modéliser des relations complexes entre des variables indépendantes (ou prédicteurs) et une variable dépendante (ou réponse). Ils sont une extension des modèles linéaires généralisés (GLM) qui permettent de capturer des relations non linéaires entre les prédicteurs et la réponse. Dans un GAM, chaque prédicteur est modélisé individuellement en utilisant des fonctions lisses (ou splines), qui sont des courbes non linéaires flexibles. Ces fonctions lisses sont ensuite combinées de manière additive pour former le modèle global. Les GAM permettent également d'inclure des termes linéaires, ce qui en fait un cadre flexible pour la modélisation de données complexes. C'est pourquoi ils sont particulièrement adaptés pour modéliser les relations écologiques complexes entre les variables. Par ailleurs, ils fournissent généralement une interprétation plus aisée des effets des variables prédictives sur la variable réponse en utilisant des courbes lisses pour illustrer ces effets.
Ainsi, le corps de cette étude est constitué par la mise en place et l'évaluation de modèles à haute résolution (10 m) qui vont faire l'objet des parties suivantes.

2.3.1. Calibration et validation des modèles

Pour mettre en place les modèles, j'ai dans un premier temps sélectionné les variables à maintenir. En effet j'ai décidé d'éliminer toutes les variables corrélées à plus de 70% pour éviter de biaiser le modèle (fonction *select07* du package *mecofun,* matrice de corrélations en annexe B Fig. B.3). En effet, le risque est une mauvaise estimation des coefficients et de leurs erreurs-types associées, qui peut rendre certaines variables insignifiantes ou entraîner un surajustement du modèle (Dormann et al. 2013). A l'issue de cette sélection, 14 variables ont été retenues (bio3, bio8, bio13, bio14, bio15, nitrogen, cfvo, sand, silt, NDVI2022, elev, slope et TPI) (Annexe B Fig. B.4). J'ai ensuite partitionné le jeu de données en deux : 70% pour l'entraînement du modèle et 30% pour sa validation croisée*, et réalisé 10 répétitions, qui sont les recommandations habituelles (Guisan et Thuiller 2005; Araújo et al. 2005). Ensuite, pour mettre en place le modèle d'ensemble, j'ai sélectionné 6 des algorithmes que j'ai pu voir utilisés le plus souvent dans la littérature (Tab. 1) : GLM, MARS, CTA, GBM, RF et MAXENT (Hao et al. 2019).

Table 1 : Tableau récapitulatif des algorithmes utilisés dans le cadre des modélisations à très haute résolution. Tous les algorithmes, sauf le GAM, ont été intégrés dans la modélisation d'ensemble.

Régression	**GLM**	Modèles linéaires généralisés, extension des modèles de régression linéaire, adaptée pour modéliser des relations entre variables en tenant compte de différentes distributions de données et de structures de variance
	GAM	*Generalized Additive Models, sont des techniques statistiques qui permettent de modéliser des relations complexes entre variables en utilisant des fonctions non linéaires additives, offrant ainsi une approche flexible pour l'analyse de données et la prédiction.*
	MARS	Multivariate Adaptive Regression Splines, technique d'analyse statistique qui décompose une relation complexe entre variables en segments plus simples, en utilisant des fonctions linéaires ou spline, pour créer un modèle prédictif flexible et interprétable.
	CTA	Classification and Regression Tree Analysis, également appelés arbres de décision, sont utilisés pour classer ou prédire des valeurs numériques en divisant récursivement les données en sous-groupes en fonction de conditions sur les variables explicatives.
Machine-learning	**GBM**	Gradient Boosting Models, technique d'apprentissage automatique qui combine plusieurs modèles simples en un modèle prédictif puissant en utilisant une approche itérative qui améliore progressivement la prédiction en corrigeant les erreurs précédentes.
	RF	Random Forest, algorithme d'apprentissage automatique qui combine plusieurs arbres de décision pour améliorer la précision prédictive et la stabilité en utilisant des méthodes d'agrégation. Ils sont efficaces pour la classification et la régression, en réduisant le surajustement* et en capturant des relations complexes dans les données.
	MAXENT	Maximum Entropy Models, grâce à l'apprentissage automatique, le modèle cherche à trouver la distribution de probabilité qui est la moins biaisée parmi toutes les distributions possibles, tout en satisfaisant les contraintes spécifiques fournies par les données.

Pour les algorithmes utilisés dans le modèle d'ensemble, j'ai gardé des options de modélisation de base avec :

- GLM : simple (d'ordre 1), sans termes d'interaction, et avec une famille binomiale
- GBM : nombre d'arbres de 1000

Pour les autres algorithmes, ce sont les paramètres par défaut qui ont été gardés.

Pour les GAM, j'ai retenu la famille binomiale car elle est spécifiquement conçue pour modéliser des probabilités binaires de type présence/absence. De plus, j'ai sélectionné la méthode REML (pour l'estimation des paramètres de lissage du modèle) qui est la plus couramment utilisée (Wood 2012). De plus, j'ai précisé le paramètre "k" qui contrôle le degré de lissage des courbes de réponse produites lors de la modélisation. En effet, le "k" correspond au nombre de nœuds de lissage utilisés pour modéliser les composantes lisses dans le modèle. Un "k" plus élevé permet des courbes plus complexes et plus flexibles, tandis qu'un "k" plus faible conduit à des courbes plus lisses et plus simples. En général, choisir le bon paramètre "k" est une question d'équilibre entre la flexibilité du modèle et le risque de surajustement. Un "k" trop élevé peut conduire à un surajustement du modèle aux données d'entraînement, tandis qu'un "k" trop faible peut sous-modéliser les relations sous-jacentes (Pedersen et al. 2019). Ainsi j'ai généré plusieurs modèles avec différents niveaux de lissage, et j'ai choisi celui qui donnait des courbes le plus écologiquement cohérentes (soit à k=5) (Annexe D).

2.3.2. Prédiction de l'occurrence du thym grâce aux modèles

Pour prédire de la probabilité de présence du thym à partir des modèles générés, des fonctions déjà intégrées dans le package *Biomod2* permettent de le faire automatiquement (avec *BIOMOD_projection*). Ainsi, j'ai réalisé des prédictions pour le GAM, le modèle d'ensemble et les modèles le constituant pour voir leur performance individuelle.

2.3.3. Evaluation des modèles

Mesurer les performances d'un modèle est important pour évaluer la précision et la fiabilité de ses résultats. Ici, l'évaluation des modèles est automatiquement faite par le package *Biomod*. L'évaluation des modèles individuels est permise par le partitionnement du jeu de données en deux (70-30%) (Mtengwana et al. 2021). Il suffit juste de sélectionner les métriques. Ainsi, j'ai sélectionné le TSS (True Skill Statistic) et AUC-ROC qui sont les métriques les plus couramment utilisées pour ce type de modélisation.

La "True Skill Statistic" (TSS) est la somme de la sensibilité* (proportion de vrais positifs) et de la spécificité* d'un modèle (proportion de vrais négatifs) moins un. Elle mesure la capacité d'un modèle à distinguer les événements positifs (ceux qui se sont produits) des événements négatifs (ceux qui ne se sont pas produits), en tenant compte à la fois des prédictions correctes et des erreurs. La plage de valeurs du TSS se situe entre -1 et +1, +1 étant le meilleur score, tandis que TSS ≤ 0 indique l'absence d'accord et une faible performance du modèle (Allouche, Tsoar, et Kadmon 2006; Somodi, Lepesi, et Botta-Dukát 2017).

La courbe ROC (Receiver Operating Characteristic) représente la relation entre le taux de vrais positifs (sensibilité*) et le taux de faux positifs (1 - spécificité*) à différents seuils de classification. L'aire sous la courbe ROC (AUC) est souvent utilisée comme mesure pour évaluer la performance d'un modèle de classification binaire. Les valeurs de l'aire sous la courbe (AUC) varient entre 0 et 1, où les modèles peu précis ont des valeurs proches de 0, tandis que les modèles avec une valeur d'AUC ≥ 0,7 démontrent de fortes capacités prédictives (Mtengwana et al. 2021).

2.3.4. Assemblage des modèles pour produire un modèle d'ensemble

Enfin, une partie essentielle du assemblage de modèles* est l'attribution d'un poids à chacun des modèles. Pour cela, j'ai effectué une moyenne pondérée de tous les modèles individuels en fonction de leurs scores TSS obtenus par cross-validation.

2.4. Caractérisation *in-situ* de l'occurrence et de l'abondance du thym

Dans un second volet, j'ai réalisé une étude de terrain pour voir si les prédictions fournies par mes modélisations étaient validées par des observations *in-situ*, et pour vérifier si certains modèles ont éventuellement surajusté les données.

2.4.1. Bilan méthodologique des métriques les plus communes

Tout d'abord, j'ai conduit une recherche méthodologique qui m'a permis de constituer un tableau bilan des métriques régulièrement utilisées pour quantifier l'occurrence et l'abondance d'une espèce (Tab. 2). Je n'ai regardé que les méthodes non destructives qui sont généralement plus faciles à mettre en œuvre, à moindre coût et en minimisant les perturbations sur les espèces observées.

Table 2 : Récapitulatif des métriques (non destructives) couramment utilisées pour quantifier l'occurrence et l'abondance des plantes. Issu d'une analyse bibliographique.

	Métrique	Méthode(s)	Références
Occurrence	Présence / Absence		
	Time-to-Detection (TTD)	Mesure du temps à la détection du premier individu	(Halstead, Rose, et Kleeman 2021; Bornand et al. 2014; Priyadarshani et al. 2022; Garrard et al. 2013)
Abondance	Recouvrement	Estimation à vue (facteurs de Braun-Blanquet…)	(Wikum et Shanholtzer 1978; Braun-blanquet 1932; Mueller-Dombois et Ellenberg 1974)
	Densité et fréquence d'occurrence	Méthodes d'échantillonnage systématiques (transects, quadrats…)	(Schmutz et al. 1982; Floyd et Anderson 1987; Barkaoui, Bernard-Verdier, et Navas 2013)
		Méthodes d'échantillonnage non systématiques (mesures de distances comme le PCQM, …)	(Dahdouh-Guebas et Koedam 2006; Khan et al. 2016; White et al. 2008; Kumarathunge, R.O.Thattil, et Nissanka 2011)
	Biomasse	Recouvrement mesuré x hauteur, points contacts…	(Barkaoui, Bernard-Verdier, et Navas 2013; Büchi et al. 2016)

2.4.2. Protocole de mesures

Pour l'étude de terrain, j'ai donc choisi de mesurer l'occurrence et l'abondance du thym à travers plusieurs métriques dérivées de ce qui a pu être proposé dans la littérature (Tab. 3 et Fig. 4). Nous avons choisi de réaliser des disques d'échantillonnage d'environ 100 m² (5,6 m de rayon), ce qui correspond à la taille d'une cellule des rasters produits par les SDM. La fiche utilisée sur le terrain pour les mesures peut être trouvée en Annexe G.

Table 3 : Récapitulatif des métriques retenues pour l'étude de terrain, et leurs méthodes de mesure associées.

	Métrique	Ordre	Méthode
Occurrence	**TTD (Time-to-detection)**	**1**	Mesurer le temps à la première détection de l'espèce dans le quadrat. Commencer par le quartier NE, se donner 2 min de prospection, si aucun individu n'a été repéré, passer au prochain quart et faire de même pour tous les autres quarts. Si un individu est repéré, la prospection s'arrête complètement et le temps total de prospection est relevé en secondes.
	Présence / Absence	**-**	Déterminer si l'espèce est présente ou absente (déduit du recouvrement).

Abondance	**Recouvrement estimé à vue**	**2**	Estimer le recouvrement à vue dans chacun des quarts du cercle (1 carré de 50*50cm représente environ 1%). 0% n'est indiqué qu'en cas d'absence absolue, en cas de très faible recouvrement on choisit une valeur légèrement supérieure.
	Recouvrement mesuré à partir de distances de recouvrement	**3**	Sur chacun des axes N, E, S et O, mesurer la distance au centre de chaque touffe rencontrée, en relevant le couple de valeurs représentant la distance au début et à la fin de la touffe observée
	Distance à la première touffe *-indicateur du niveau d'agrégation de la population-*	**4**	Sur les huit axes N, NE, E, SE, S, SO, O et NO, mesurer la distance au premier touffe de thym touchant le mètre ruban en partant du centre.
	Hauteur de la première touffe *-utilisé avec la distance pour une estimation de la biomasse-*	**4**	Sur les huit axes N, NE, E, SE, S, SO, O et NO, mesurer la hauteur maximale de la première touffe touchant le mètre ruban en partant du centre.
Abondance	**Nombre de touches** *-proxy de la fréquence d'occurrence-*	**-**	Déduit à partir des mesures de recouvrement, il s'agit du nombre total de fois que le mètre croise une touffe de thym sur les axes : N, E, S et O.

Le temps à la première détection doit impérativement être effectué en premier pour éviter de biaiser la mesure. L'ordre de relevé des autres métriques correspond à la succession de mesures la plus pratique à réaliser sur le terrain.

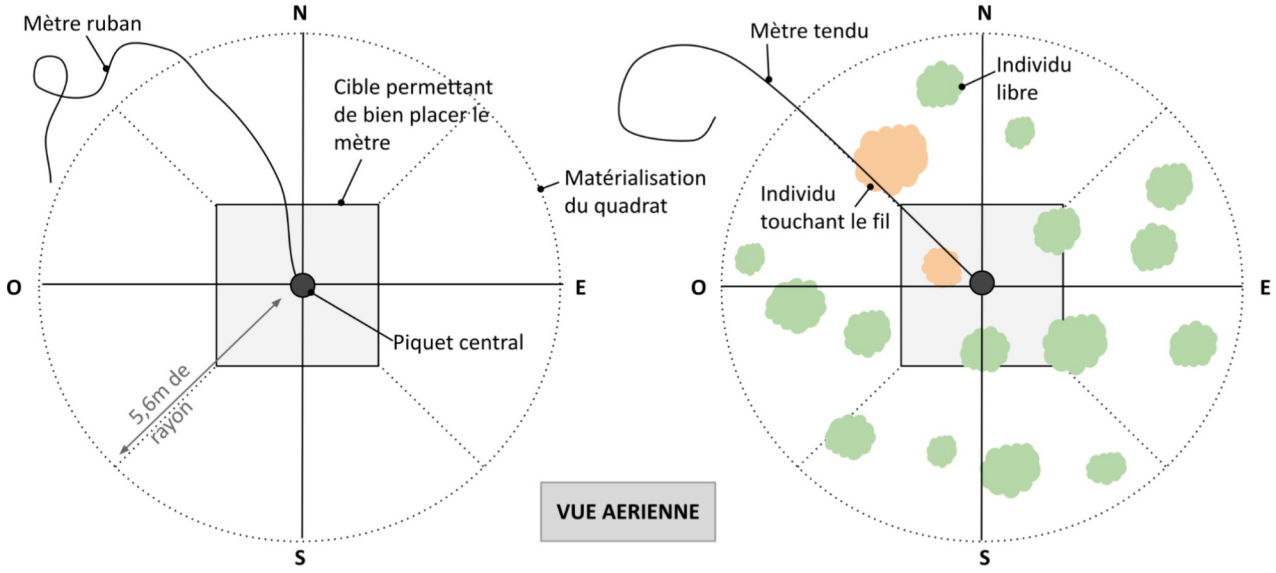


Figure 4 : Schématisation du dispositif expérimental mis en place pour les mesures de thym sur le terrain. *(source personnelle)*

A noter qu'il faut partir à chaque fois du nord pour continuer le long du cercle dans le sens des aiguilles d'une montre. Toutes les mesures sont approchées au centimètre.

2.4.3. Choix des zones d'échantillonnage

Pour l'échantillonnage, j'ai souhaité couvrir un maximum du gradient de distribution du thym. J'ai donc réalisé un transect perpendiculaire à la côte, d'environ 100 km de long qui recoupe la distribution latitudinale de l'espèce. Ensuite, 10 zones d'environ 1 km^2 ont été sélectionnées le long du transect dans lesquelles j'ai tiré 10 points au hasard distancés d'au moins 50 m (Fig. 5). L'homogénéité de la taille de zones a permis de garder une densité d'échantillonnage constante. En plus des 10 points initiaux, j'ai également tiré 20 points au hasard qui me servaient de points de secours si jamais les points prévus n'étaient pas accessibles. Dans ce cas, le point de secours le plus proche était sélectionné. Ce sont donc 110 quadrats qui ont été réalisés sur 10 jours complets de terrain auxquels ont été rajoutés 2 jours pour les expérimentations préliminaires. J'ai eu l'aide de 4 autres personnes pour effectuer ce travail. Nous étions en général 3 par jour de terrain.

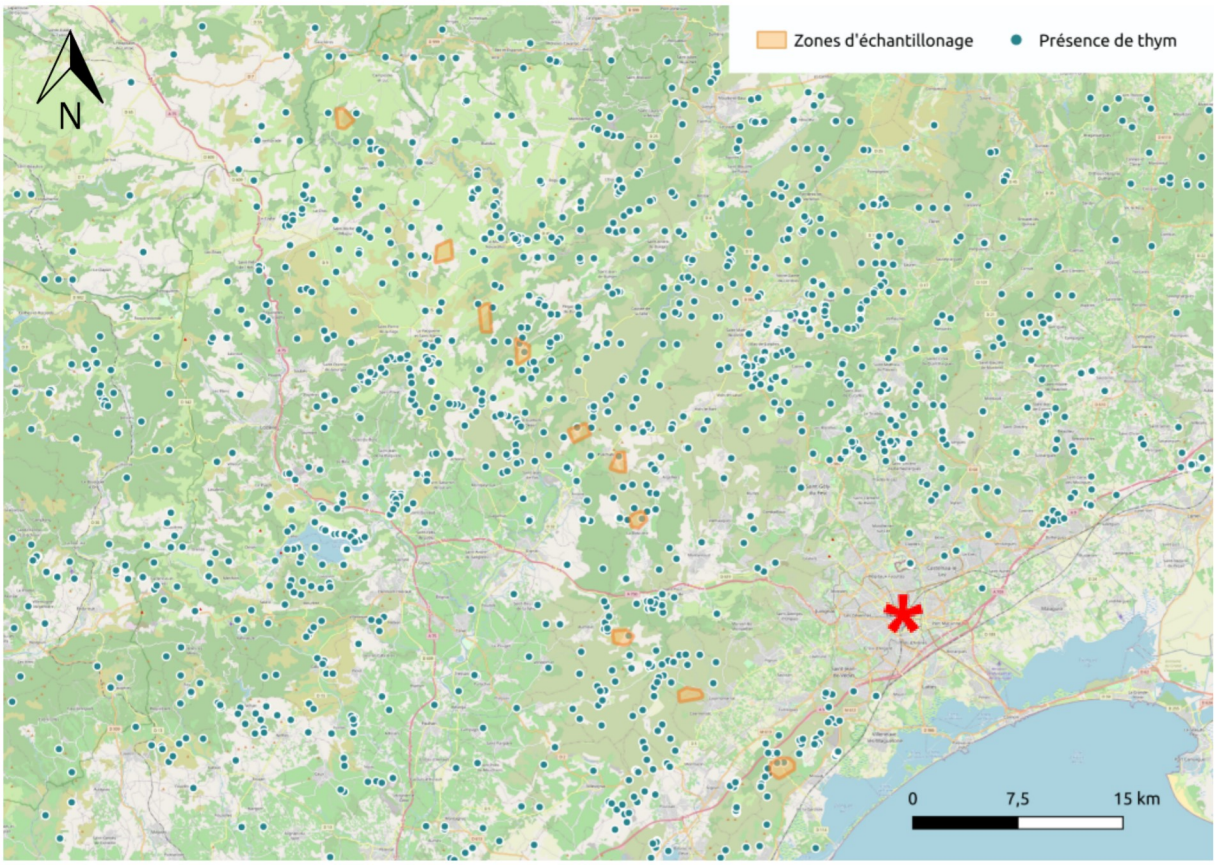


Figure 5 : Carte des zones d'échantillonnage prospectées lors de l'étude de terrain sur le thym. Montpellier est localisé par l'étoile rouge.

2.4.4. Validation de la méthode terrain

Afin de valider la méthodologie de terrain, une première sortie a été organisée pour calibrer les mesures, vérifier la facilité d'utilisation et la réplicabilité des métriques choisies. Pour cela, une zone dans la garrigue au nord de Montpellier a été sélectionnée, dans laquelle 9 quadrats ont été matérialisés. Chaque quadrat a été prospecté avec le même protocole par 3 observateurs différents. Cela m'a notamment permis de rendre compte de la variabilité inter-observateur.

Ensuite, j'ai regardé la corrélation entre les différentes métriques grâce à des régressions linéaires. J'ai donc regardé la relation entre le recouvrement estimé à vue et le recouvrement mesuré, de même pour le nombre de touches et le recouvrement mesuré. J'ai également cherché s'il y avait une relation entre la hauteur moyenne de touffe et sa taille, ainsi qu'entre la distance au premier individu et le recouvrement mesuré.

Par ailleurs, j'ai étudié l'évolution des coefficients de variation inter-observateur pour le recouvrement estimé à vue et le recouvrement mesuré en fonction de l'abondance associée. Le coefficient de variation a été choisi plutôt que l'écart-type. Il permet une mesure relative de la dispersion par rapport à la moyenne qui est comparable même avec des moyennes différentes.

Enfin, pour déterminer si l'échantillonnage réalisé était suffisant. J'ai repris les valeurs de recouvrement mesurées pour les 9 quadrats de l'étude préliminaire. J'ai ensuite cherché à voir si la moyenne évoluait selon la quantité de données intégrée. Pour cela, j'ai regardé l'évolution de la valeur de recouvrement moyenne d'un quadrat selon le nombre de centimètres intégrés. J'ai donc tiré au hasard 1 cm parmi les 2240 cm prospectés (560x4=2240) et regardé si le thym y était présent ou absent. J'ai continué en ajoutant 1 cm

à chaque fois jusqu'à couvrir l'ensemble de 2240 cm. La valeur de recouvrement correspondait alors au *nombre de cm avec présence de thym / nombre de cm tirés.* J'ai répété cette opération 10 fois afin de pouvoir obtenir une moyenne et son écart-type associé.

2.5. Comparaison des SDM aux données terrain : analyse statistique

Afin d'évaluer la capacité des modèles de distribution à haute résolution à prédire l'occurrence et l'abondance du thym, j'ai employé deux approches pour comparer mes modèles et les données de terrain. Tout d'abord, j'ai souhaité évaluer la capacité des modèles à prédire la présence du thym, et dans un second temps leur capacité à prédire son abondance (représentée ici par le recouvrement mesuré). Pour tester le lien entre chaque variable de terrain et nos résultats, j'ai utilisé des modèles linéaires généralisés avec effets aléatoires* sur le site. Cet effet aléatoire permet de modéliser la variabilité qui ne peut pas être expliquée par les variables explicatives fixes du modèle.

2.5.1. Evaluation de la capacité des modèles à prédire la présence ou absence observée sur le terrain

Dans le cas de la prédiction de présence/absence, j'ai réalisé un modèle linéaire généralisé mixte avec une distribution binomiale. Ce modèle a été réalisé grâce au package *lme4* sous R (fonction *glmer*). Voici la formule employée :

glmer(PA ~ PrédictionPrésence + (1|Site), family = "binomial")

PA : présence/absence observée sur le terrain
PrédictionPrésence : probabilité de présence donnée par les modélisations pour le quadrat échantillonné

J'ai retenu la distribution binomiale car elle est spécifiquement conçue pour modéliser des probabilités binaires, de type présence/absence.

Pour obtenir la prédiction de présence associée au quadrat réalisé sur le terrain, j'ai initialement voulu sélectionner la prédiction calculée pour la cellule dans laquelle tombait le point GPS du quadrat. Cependant, ces points GPS correspondent au centre du quadrat réalisé et ne coïncident pas toujours avec le centre d'une cellule. Ainsi, pour avoir une prédiction plus représentative, j'ai créé des zones tampon autour des points GPS (fonction *buffer* du package *terra* sous R). Ces zones ont été matérialisées sous forme de disque de 13,2 m de diamètre (Fig. 6). Les quadrats sur le terrain faisaient environ 11,2 m de diamètre (100 m²), j'ai donc rajouté 2 m pour prendre en compte l'incertitude liée à la position GPS. A partir des disques, j'ai ensuite calculé la moyenne pondérée des prédictions des cellules couvertes (fonction *extract* du package *terra* sous R). Cette pondération a été réalisée en multipliant chaque valeur de prédiction par la fraction de la cellule couverte associée.

Pour comparer la performance des différents modèles, j'ai utilisé les métriques classiques directement calculées lors de la modélisation : AIC, BIC et log-likelihood (Kuha 2004).

L'AIC (Akaike Information Criterion) et le BIC (Bayesian Information Criterion) sont des mesures statistiques utilisées pour comparer et sélectionner des modèles en fonction de leur ajustement aux données tout en pénalisant la complexité des modèles. L'AIC favorise la précision de prédiction, tandis que le BIC favorise la simplicité des modèles. Des valeurs plus basses d'AIC et de BIC indiquent généralement un meilleur ajustement du modèle aux données.

Le log-likelihood (logLik) est une mesure de la vraisemblance logarithmique d'un modèle statistique par rapport aux données observées. Elle indique à quel point les données sont probables sous un modèle spécifique. Une valeur plus élevée indique que le modèle s'ajuste

mieux aux données observées. Le logLik est utilisé pour calculer l'AIC et le BIC. L'AIC est calculé en soustrayant deux fois la valeur de logLik du nombre de paramètres du modèle. Le BIC est calculé en soustrayant la valeur de logLik multipliée par le logarithme du nombre d'observations du modèle

J'ai également calculé le R^2 (coefficient de détermination*) associé au modèle, ainsi que le score TSS. J'ai calculé le TSS pour pouvoir comparer les scores déjà obtenus avec *Biomod2*. Cela m'a notamment permis de voir s'il y avait un surajustement, autrement dit, si en théorie les modélisations semblaient bien fonctionner mais qu'elles étaient en réalité incapables de prédire le jeu de données indépendant obtenu sur le terrain.

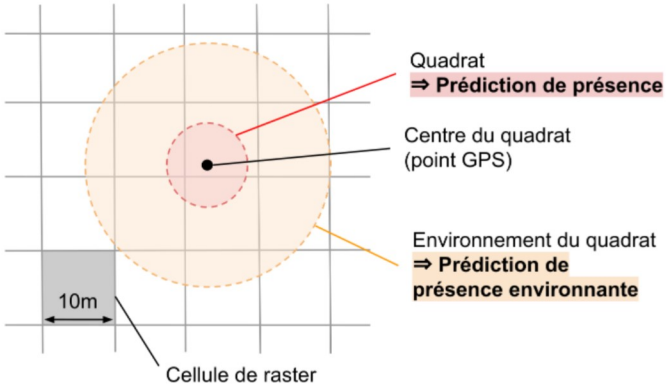


Figure 6 : Représentation schématique des zones tampon créées pour représenter un quadrat et son environnement. Ces zones ont été utilisées pour relier les échantillonnages de terrain aux prédictions de présence générées par les modélisations. Le disque rouge représente le quadrat. L'anneau orange correspond à l'environnement du quadrat, et sa prédiction de présence associée servira à prédire l'abondance de thym dans le quadrat. *(source personnelle)*

2.5.2. Evaluation de la capacité des modèles à prédire l'abondance mesurée sur le terrain

Dans le cas de la prédiction d'abondance, j'ai réalisé un modèle linéaire généralisé mixte avec une distribution bêta. Ce modèle a été réalisé grâce au package *glmmTMB* sous R (le package *lme4* ne supporte pas la distribution bêta). Voici la formule employée :

glmmTMB(AbondanceTerrain ~ PrédictionPrésenceEnvironnante : PrédictionPrésence
+ (1|Site), family = beta_family(link = "logit"))

AbondanceTerrain : recouvrement mesuré sur le terrain
PrédictionPrésence : probabilité de présence donnée par les modélisations pour le quadrat échantillonné
PrédictionPrésenceEnvironnante : prédiction de présence moyenne des cellules situées autour quadrat échantillonné

J'ai choisi d'utiliser une distribution bêta car celle-ci est indiquée pour modéliser des valeurs bornées entre 0 et 1 telles que des proportions ou des taux. Cela est donc parfaitement adapté à la modélisation d'abondance exprimée ici en pourcentages recouvrements. Cependant, les valeurs extrêmes telles que 0 et 1 ne sont pas acceptées, c'est pourquoi j'ai décidé d'effectuer une transformation sur mes données. Smithson et Verkuilen proposent notamment une transformation pratique, donnée par la formule : $\mathbf{(y * (n - 1) + 0.5)/n}$ ou **y** est la variable réponse (PrédictionPrésenceEnvironnante) et n la taille de l'échantillon (Smithson et Verkuilen 2006).

Pour obtenir la prédiction de présence associée au quadrat réalisé sur le terrain, j'ai utilisé la même méthode que celle décrite dans la partie 2.5.2. Ensuite, pour la prédiction de la présence environnante, je suis partie du même principe. J'ai matérialisé des zones tampon circulaires de 34 m de diamètre, pour couvrir environ 900 m² autour du centre du quadrat échantillonné (petit disque représentant le quadrat exclu) (Fig. 6). Cette taille a été choisie suite à mes observations de terrain. J'ai alors émis l'hypothèse que plus la probabilité de présence prédite pour cet anneau serait élevée, plus l'abondance en son centre le serait également (Sagarin et Gaines 2002; Brown 1984; Lawton 1993).

Pour comparer la performance des différents modèles, j'ai utilisé les mêmes métriques que dans la partie 2.5.2 précédente : AIC, BIC, logLik et R^2 .

3. Résultats

3.1. Résultats des SDMs à très haute résolution : évaluation de leur performance à partir des validations croisées

Tout d'abord, j'ai évalué la performance des modèles générés par *Biomod2*. Ci-dessous se trouve une figure montrant les scores TSS pour chacun des algorithmes utilisés (Fig. 7). Nous observons une performance nettement plus élevée pour le modèle d'ensemble (0.629). Les autres algorithmes présentent des scores plus bas avec des valeurs plutôt proches. Seul le GLM semble mal performer (environ 0.3), tandis que le GAM se situe plutôt dans la marge haute de scores (environ 0.46). Les autres métriques produites pour l'évaluation des modèles peuvent être trouvées en Annexe C Tab C.2.

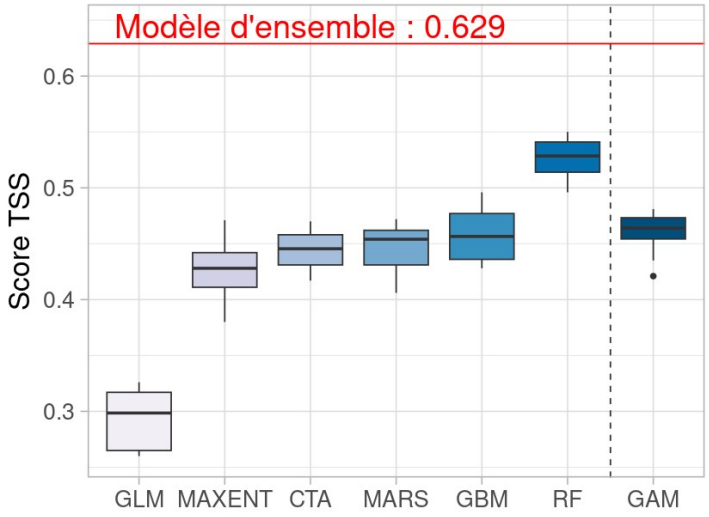


Figure 7 : Scores TSS calculé par *Biomod2* pour chacun des algorithmes utilisés dans le cadre de la modélisation à très haute résolution. Les boîtes à moustache sont basées sur les scores TSS calculés sur le jeu de données de validation (30% restants) de chacun des 10 "runs" effectués pour chaque algorithme. La ligne horizontale centrale indique la médiane. Note : Le "run" 2 de l'algorithme GBM n'a pas abouti, ses évaluations ne sont donc faites que sur 9 répétitions.

Le modèle d'ensemble, ainsi que le GAM fournissent des modélisations assez similaires, sauf en amont de la distribution, sur le causse avec des variations allant globalement de 10 à 30%, et ponctuellement jusqu'à 50 % (Fig. 8). Nous pouvons remarquer que le GAM détecte fortement les surfaces linéaires (fleuves, routes…) qui ne sont pas mises en avant avec le modèle d'ensemble. Nous pouvons également noter une anomalie de prédiction pour le GAM dans l'étang de Thau (flèche noire), avec des probabilités de présence anormalement élevées. De plus, l'effet du filtrage des points de présence et très visible, nous voyons nettement les délimitations des zones urbanisées qui ont de faibles probabilités de présence de thym (flèches rouges). En outre, les scores d'importance attribués à chaque variable montrent que les variables ayant le plus d'influence sont globalement les mêmes dans les deux modèles. Il s'agit d'elev (élévation), bio8 (température moyenne du quart le plus pluvieux), bio14 (précipitations du mois le plus sec) et bio15 (saisonnalité des précipitations) sont celles ayant le plus d'influence dans les deux modèles finaux (Annexe C Tab C.1). Il y a juste bio13 (précipitations du mois le plus pluvieux) pour laquelle une contribution importante est capturée dans le cas du GAM, mais ignorée par le modèle d'ensemble.

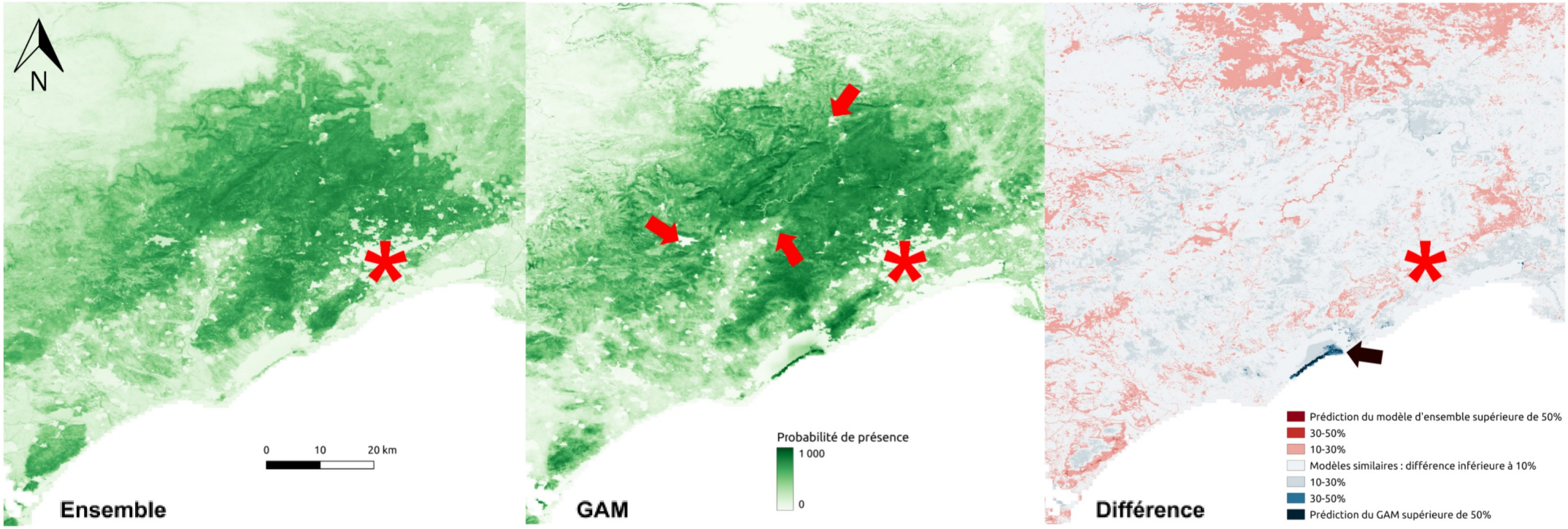


Figure 8 : Cartes produites par les deux approches de modélisation à l'échelle locale, et différences inter-modèles. Montpellier est localisé par l'étoile rouge. Les flèches rouges sont des exemples de zones urbanisées qui affichent une faible probabilité de présence. La flèche noire indique une anomalie de prédiction au niveau de l'étang de Thau.

3.2. Résultats de l'étude de terrain

Ensuite, grâce à l'étude de terrain j'ai pu échantillonner 110 points, parmi lesquels 66 présences et 44 absences. Pour vérifier la pertinence des métriques choisies, une étude préliminaire à été menée dont les résultats sont présentés ci-dessous.

3.2.1. Etude préliminaire : cohérence et répétabilité des métriques choisies

3.2.1.1. Analyse du biais observateur

Pour tester le biais observateur lors du déploiement du protocole, la méthode a été testée sur un même site par différents observateurs. Pour rappel : 9 quadrats ont été matérialisés, allant de A à I, et le protocole a été appliqué par 3 différents observateurs sur chaque quadrat. Les résultats nous ont permis d'observer une faible variabilité entre observateurs (Fig. 9) avec une moyenne de 0.6 % de variabilité pour le recouvrement mesuré et de 2.7 % pour le recouvrement estimé à vue.

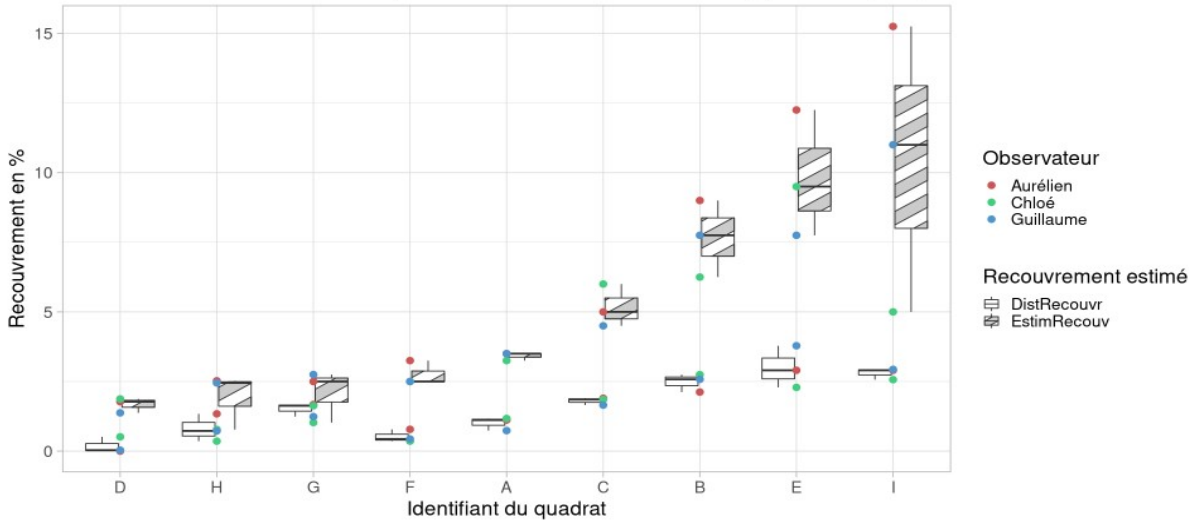


Figure 9 : Variabilité des valeurs de recouvrement obtenues par 3 observateurs sur 9 quadrats (A a I). Les valeurs de recouvrement sont obtenues grâce à deux méthodes : recouvrement estimé à partir de mesures de distances (DistRecouvr) et recouvrement estimé à vue (EstimRecouv).

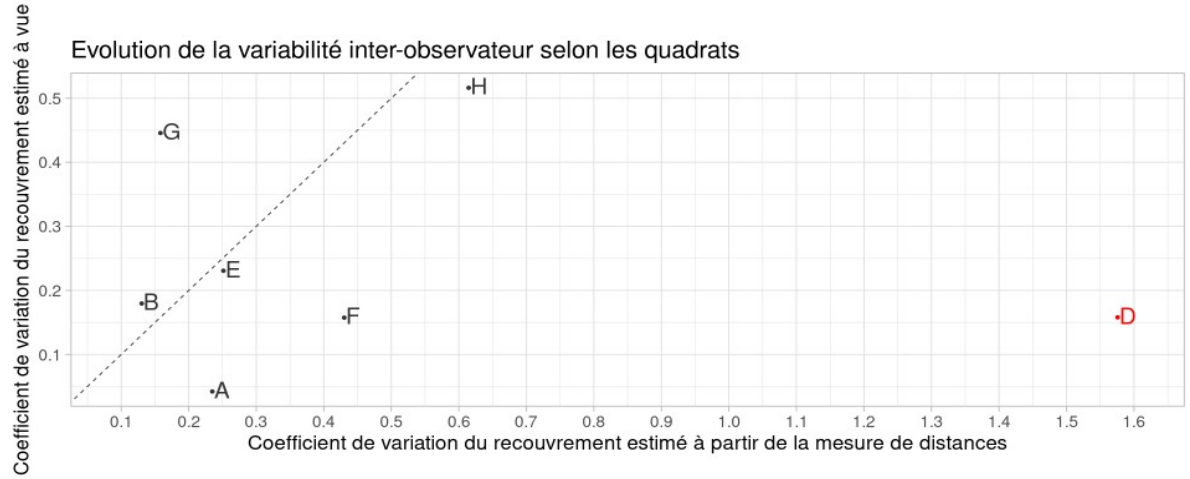


Figure 10 : Coefficient de variation du recouvrement estimé à vue en fonction du coefficient de variation du recouvrement mesuré à partir des distances. Le point D (en rouge) est particulièrement éloigné. Cela est expliqué par le fait que les valeurs de recouvrement mesurées ici à partir des distances sont très faibles, induisant un écart très important entre les relevés des observateurs (cf Fig. 9 ci-dessus).

De plus, l'abondance ne semble pas avoir d'effet déterminant sur la variabilité inter-observateur des mesures (Annexe C Fig C.3). Nous pourrions supposer que plus le recouvrement du quadrat est élevé, plus il est difficile d'avoir une estimation à vue précise, tandis que le recouvrement mesuré à partir des distances devrait maintenir une variabilité constante quelle que soit l'abondance. Ce n'est pas ce que nous observons ici, malgré tout le nombre d'observations est insuffisant pour tirer des conclusions robustes.
En somme, nos deux métriques ont un bon niveau de précision. Les résultats montrent des

performances équivalentes pour les deux méthodes (Fig. 10). Cependant, le biais peut être élevé pour les mesures de distances dans le cas où les recouvrements sont très faibles.

3.2.1.2. Puissance de l'échantillonnage pour les mesures des distances de recouvrement

Par ailleurs, j'ai évalué la puissance de l'échantillonnage effectué pour les mesures des distances de recouvrement (Fig. 11). Nous observons ici une nette stabilisation de la valeur moyenne du recouvrement, ainsi qu'une diminution de l'écart-type associé. Cela indique un échantillonnage suffisant pour obtenir un recouvrement représentatif de la réalité.

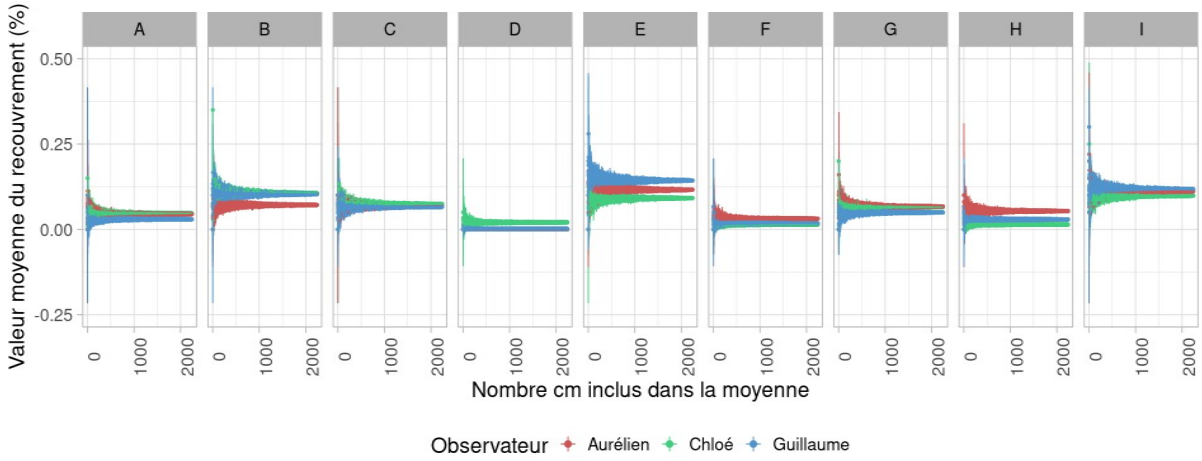


Figure 11 : Evolution du recouvrement mesuré selon le nombre de cm inclus au sein des transects prospectés pour chaque quadrat (de A à I). Les calculs ont été répétés 10 fois pour chaque nombre n de cm entre 1 et 2240 afin d'obtenir une valeur moyenne de recouvrement et son écart-type associé.

3.2.2. Cohérence entre l'abondance mesurée et estimée à vue

Ensuite, pour étudier la cohérence des métriques choisies pour l'étude de terrain. J'ai voulu établir une régression entre le recouvrement estimé à vue et le recouvrement mesuré à partir des distances. Ici, l'ensemble des 110 points échantillonnés sur le terrain ont été utilisés. Les résultats en Fig. 12 et montrent une relation linéaire nette entre les deux méthodes d'estimation du recouvrement. Cependant, nous pouvons noter un facteur 2 entre ces mesures qui supposerait une éventuelle surestimation lors de l'estimation à vue du recouvrement. En outre, nous pouvons observer une très forte corrélation entre le nombre de touffes de thym touchées et l'abondance mesurée du quadrat (Fig. 13). Cela suggère une homogénéité de la taille des touffes. De plus, cela indique que mesurer la taille des touffes apporte peu d'informations supplémentaires.

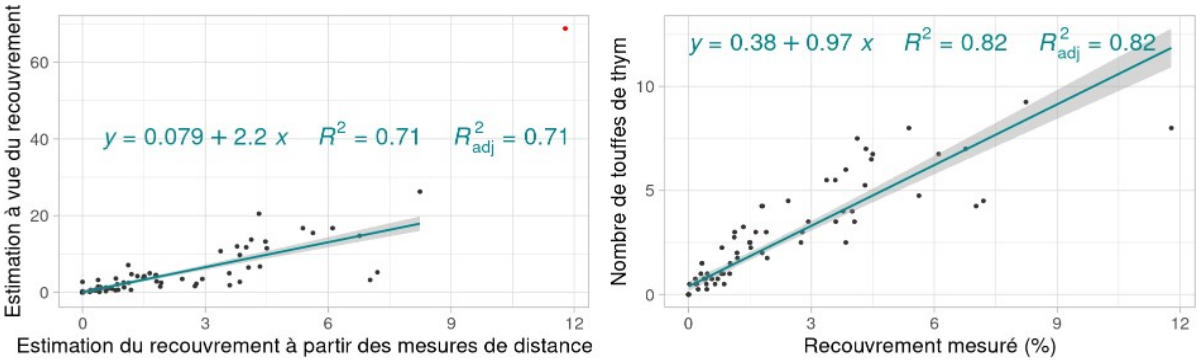


Figure 12 (à gauche) : Régression entre les valeurs (en %) de recouvrement estimé à vue et de recouvrement estimé à partir des mesures de distance.
Figure 13 (à droite) : Régression entre le nombre de touffes de thym touchées sur les 4 axes (Nord, Sud, Est et Ouest) d'un quadrat et l'abondance estimée de ce quadrat (à partir des mesures de distance).

L'analyse des corrélations pour les autres métriques peut être trouvée en Annexe F.

3.3. Mise en regard de résultats des SDM à haute résolution et de l'étude de terrain : évaluation de la capacité des SDM à prédire des métriques de terrain

Dans un second temps, j'ai étudié la capacité des SDM à prédire l'occurrence et l'abondance mesurés sur le terrain.

3.3.1. Evaluation de la capacité des modèles à prédire la présence observée sur le terrain

J'ai commencé par confronter les modèles réalisés aux données de présence obtenues sur le terrain. Tout d'abord, j'ai évalué les modèles sans prendre en compte la structure spatiale des données. Pour cela, j'ai calculé les scores TSS associés aux modèles (Fig. 14). Ceux-ci montrent que lorsque les modèles sont confrontés aux données terrain, les TSS diminuent (sauf pour le GLM qui augmente légèrement). Le GAM et le modèle d'ensemble font partie des modèles pour lesquels le TSS a été le plus surévalué. De plus, tous les modèles performant de manière relativement similaire avec des scores moyens.

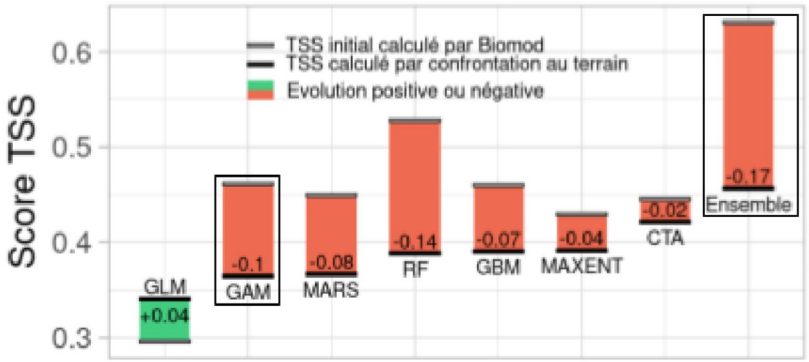


Figure 14 : Comparaison entre les scores TSS calculés par *Biomod2* et les TSS obtenus en confrontant les modèles aux données de présence récoltées sur le terrain. Les TSS calculés par *Biomod2* correspondent aux moyennes des 10 TSS calculés par évaluation croisée pour chaque répétition d'algorithme.

Ensuite, j'ai évalué les modèles en faisant des GLMM qui intègrent l'effet site (Fig. 16). Ceux-ci montrent que seuls les modèles individuels MAXENT (P-valeur=0,035) et RF (P-valeur = 0,018) ont des relations significatives entre la présence observée (réponse) et la présence prédite sur le terrain (prédicteur) (Tab. 4). Ce n'est pas le cas pour le modèle d'ensemble et le GAM, malgré tout le modèle d'ensemble est très proche de la significativité (0,065). Nous pouvons noter que le point de basculement d'absence à présence est très variable selon les algorithmes utilisés : pour le GAM ce seuil se situe à 0.286, tandis que pour le modèle d'ensemble, il est trois fois plus élevé, à 0.675 (Fig. 15). Les sorties complètes des modèles et vérifications d'hypothèses sont en annexe C Tab C.3-4 et Fig C.3.

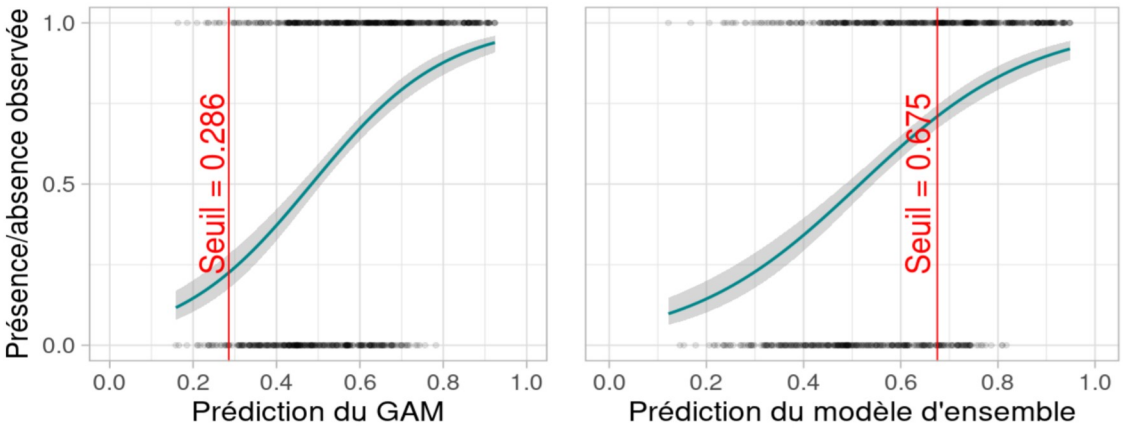


Figure 15 : Distribution des points occurrences observées sur le terrain en fonction des prédictions issues du GAM et du modèle d'ensemble. La courbe bleue correspond à l'ajustement d'un GLMM à distribution binomiale sur les données. Le point de basculement indiquant le passage d'absence à présence est indiqué en rouge (calculé à partir du TSS le plus élevé). Les graphes pour les autres algorithmes sont en annexe C Fig. C.4.

A la suite de ces GLMM, j'ai pu extraire les scores AIC et BIC (Fig. 16) qui montrent une meilleure performance des modèles de ML (RF et Maxent), suivi du modèle d'ensemble et du GAM. Ce classement se retrouve lors du calcul des R^2 (Fig. 17). Cependant, même le modèle le plus performant (RF) n'explique que 38 % de la variabilité des données. En outre, celle-ci est très majoritairement expliquée par l'effet aléatoire du site.

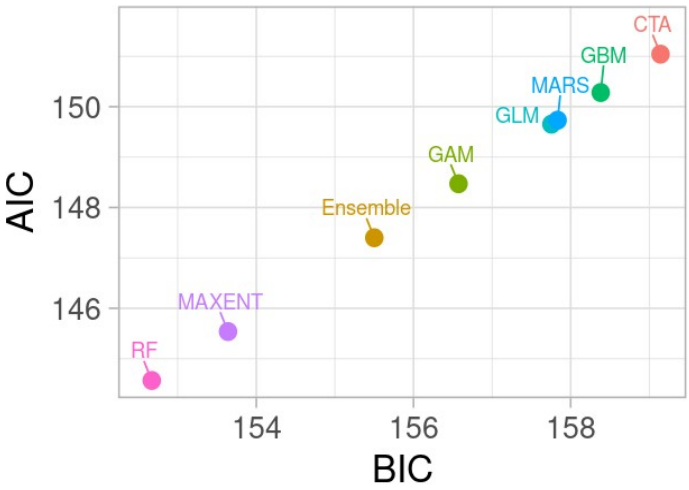


Figure 16 : Représentation du score AIC en fonction du BIC pour chaque modèle. Les scores sont issus des GLMM réalisés pour évaluer la capacité des modèles à prédire l'occurrence. Un score bas suppose un meilleur ajustement.

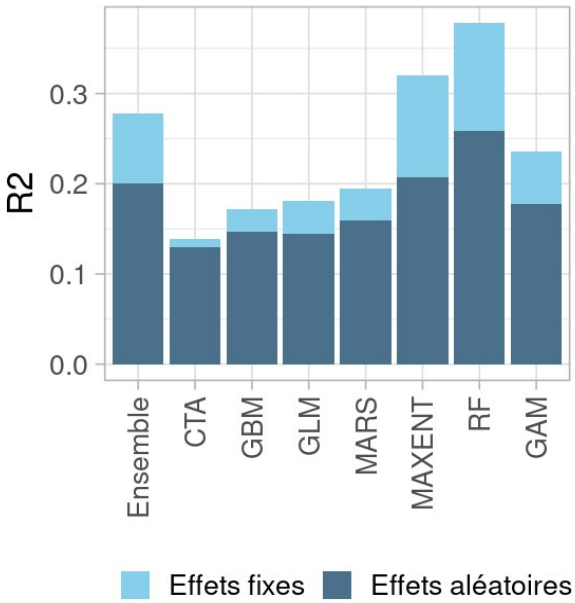


Figure 17 : R^2 (coefficients de détermination) calculés à la suite de GLMM réalisés pour évaluer la capacité des modèles à prédire l'occurrence. La part expliquée par les effets fixes (prédiction de présence) et aléatoires (site) est précisée.

3.3.2. Evaluation de la capacité des modèles à prédire l'abondance mesurée sur le terrain

Ensuite, j'ai voulu savoir si les modèles réalisés permettaient de prédire l'abondance de thym mesurée sur le terrain (Fig. 18). Ici, les GLMM ont démontré une relation significative entre les variables pour CTA (0,039), MAXENT (0,013), RF (0,015) et le modèle d'ensemble (0,023). Les autres restent tout de même très proches de la significativité, mis à part le GLM. Les tableaux avec les sorties complètes des modèles et vérifications d'hypothèses sont en annexe C Tab C.5-6 Fig. C.5.

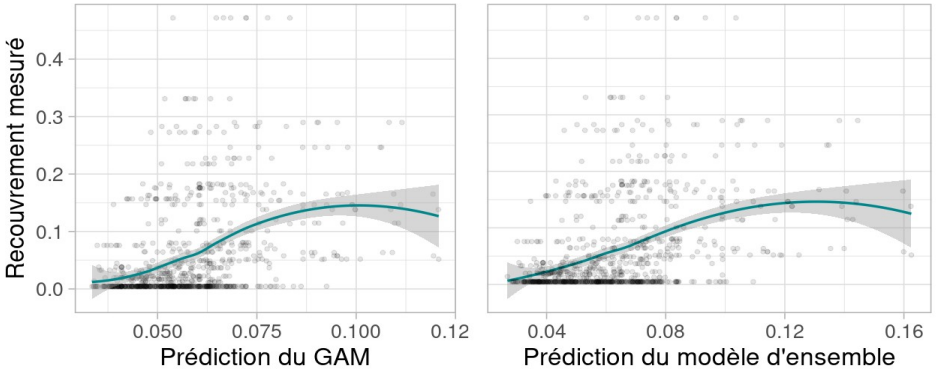


Figure 18 : Distribution du recouvrement mesuré sur le terrain en fonction des prédictions issues du GAM et du modèle d'ensemble. La courbe bleue correspond à l'ajustement d'un GLMM à distribution bêta sur les données. Les graphes pour les autres algorithmes sont en annexe C Fig. C.6.

A la suite de ces GLMM, j'ai pu extraire les scores AIC et BIC (Fig. 19) qui montrent une meilleure performance de Maxent, suivi de RF, et du modèle d'ensemble. Le GAM n'arrive qu'en septième position. Ce classement se retrouve lors du calcul des R^2 (Fig. 20).

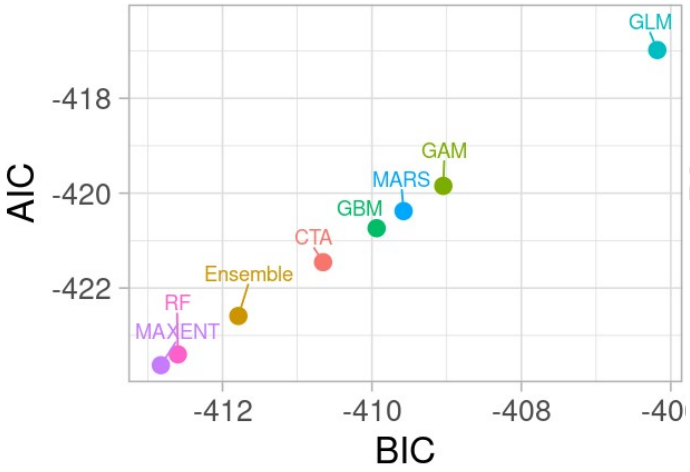


Figure 19 : Représentation du score AIC en fonction du BIC pour chaque modèle. Les scores sont issus des GLMM réalisés pour évaluer la capacité des modèles à prédire l'abondance. Un score bas suppose un meilleur ajustement.

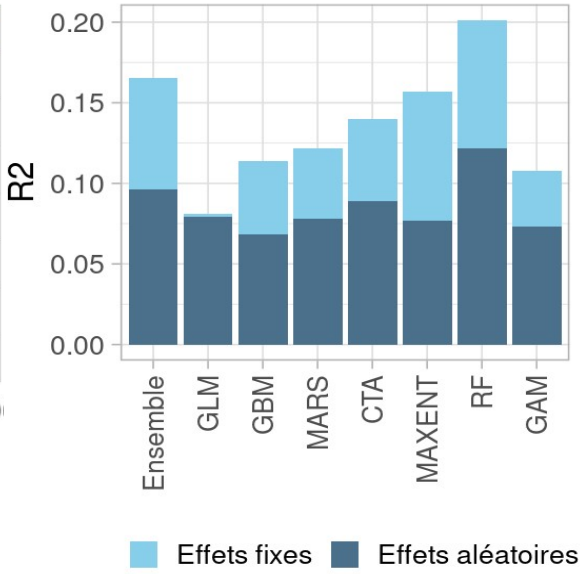


Figure 20 : R² calculés à la suite de GLMM réalisés pour évaluer la capacité des modèles à prédire l'abondance. La part expliquée par les effets fixes et aléatoires (site) est précisée.

Ici, les R² montrent une plus grande implication des effets fixes, avec des effets aléatoires toujours importants. En outre, les modèles ont un pouvoir explicatif légèrement inférieur avec un R² de 0.2 au maximum pour RF, suivi du modèle d'ensemble avec 0.17. Le GAM n'explique que 11% de la variabilité des données.

4. Discussion et perspectives

4.1. Variables à haute résolution : suffisantes pour prédire l'occurrence et l'abondance ?

L'utilisation de variables à haute résolution dans le SDM reste encore limitée, notamment à cause de leur coût et de leur disponibilité. Des questions peuvent alors se poser quant à la pertinence de leur utilisation et le rapport coût-bénéfice associé. Pradervand et al. 2014 ont déjà répondu en partie à ces interrogations. En faisant des modèles à différentes résolutions (2, 5, 10, 25, 50 et 100 m) ils ont mis une en évidence une amélioration globale très faible avec l'affinage des modèles. Cependant, certaines espèces poussant dans des micro-habitats associés avec des variations fines de topographie profitaient grandement des résolutions plus élevées (Pradervand et al. 2014). C'était notamment le cas de *Thymus spp.* qui se développe souvent dans des éboulis ou prairies rocheuses. Ainsi, nos modèles ont probablement bénéficié de ces variables à haute résolution. Toutefois, dans le cas d'une généralisation à d'autres espèces, il convient d'évaluer le bénéfice apporté au cas par cas pour adapter les choix.

Malgré l'utilisation de ces variables précises, nous avons produit des SDM qui ne prédisent que moyennement l'occurrence du thym. Les scores TSS et AUC calculés initialement par *Biomod2* pour évaluer les modèles correspondent aux résultats obtenus par des études similaires utilisant des variables à haute résolution (Engler et al. 2013; Pradervand et al. 2014; Lannuzel et al. 2021; Ndlovu et al. 2018). Cependant, la confrontation de mes résultats aux données terrain montre une baisse significative de ces scores, ainsi qu'une explication limitée de la variabilité des données. En effet, le modèle RF qui présentait le meilleur R² lors de la confrontation aux données terrain, n'a permis d'expliquer correctement que 38% de la variabilité des données (28% pour le modèle d'ensemble et 19% pour le GAM). Cela indique que nos variables restent insuffisantes pour caractériser toutes les composantes de l'environnement qui définissent l'occurrence des plantes à fine échelle.

Peu d'études ont confronté les résultats de leurs modélisations à un jeu de données indépendant, et encore moins un jeu de données spécialement calibré pour l'évaluation. Nous pouvons tout de même citer Engler et al. 2013, qui, dans le cadre d'une cartographie d'arbres individuels à 5 m, ont établi des comparaisons avec des jeux de données indépendants. D'une part, ils ont divisé leur aire d'étude en deux en utilisant les 70 % à l'est pour calibrer le modèle, et les 30% à l'ouest pour son évaluation. L'évaluation n'a montré aucun changement significatif dans la performance des modèles. Dans un second temps, les modèles ont été confrontés à un jeu de données situé dans une autre zone géographique. L'évaluation a alors montré une baisse significative de la performance, du même ordre que celle observée ici. Dans ce travail, le jeu de données indépendant se situait également dans notre aire d'étude. Au vu des résultats obtenus par Engler et al., cela pose donc des questions sur la transférabilité de nos modélisations à d'autres parties de l'aire de distribution du thym. Malgré tout, j'ai réalisé un effort de calibration spatiale pour couvrir un maximum du gradient de distribution du thym. En effet, le jeu de données de terrain a été conçu pour couvrir un grand nombre de situations de présence et d'absence de thym. Nous pouvons donc supposer que même en transférant nos modèles à des zones plus éloignées, leur performance ne diminuerait pas davantage. Dans tous les cas, les modèles développés ici restent un bon point de départ pour modéliser l'occurrence du thym à fine échelle.

Ce travail a également fourni une base intéressante pour la modélisation d'abondance. La méthodologie développée ici n'a pas été employée dans la littérature. En outre, peu d'articles traitent de l'estimation d'abondance à partir de données d'occurrence. Les données d'abondance sont généralement utilisées pour améliorer des modèles d'occurrence, ou générer des modèles d'abondance. Cependant les données d'occurrence sont rarement utilisées pour inférer de l'abondance. Nous avons donc peu d'ordre de comparaison. L'une des principales difficultés réside dans le fait qu'une probabilité de présence élevée n'équivaut pas une abondance élevée (Bradley 2016). Cependant, la théorie macroécologique suggère que l'abondance d'une espèce devrait découler de sa distribution globale (par exemple, avec des abondances élevées se produisant dans le centre géographique et/ou environnemental de la distribution (Brown 1984). Dans le cas du thym, notre capacité à prédire en partie l'abondance semble valider cette hypothèse. Toutefois, nous n'expliquons que 20% de l'abondance dans le meilleur des cas (d'après le R^2). Il nous reste donc à trouver un proxy fiable découlant des prédictions de présence pour prédire l'abondance. L'amélioration des prédictions d'occurrence sera sans doute un premier pas dans l'optimisation des prédictions d'abondance. Cependant, cette méthodologie ne sera pas forcément généralisable à d'autres espèces. En effet, plusieurs études ont montré que l'idée d'une abondance en cœur de distribution n'est pas universelle. Par exemple, Sagarin et Gaines 2002 ont passé en revue des études écologiques comparant l'abondance à l'étendue globale des espèces et ont constaté que seuls 39 % soutiennent l'existence d'un centre abondant.

Par ailleurs, il reste de nombreuses questions concernant la prédiction d'une "abondance" pertinente dans le cas des plantes cueillies. Pour l'instant je me suis focalisée sur le recouvrement comme proxy de l'abondance. Celui-ci est probablement adapté dans le cas du thym. En effet, j'ai observé suite aux mesures de terrain que les tailles de touffe thym étaient en moyenne très similaires, de plus ce sont les parties aériennes qui sont cueillies. Le recouvrement est donc sans doute une bonne estimation de la biomasse de thym disponible. Toutefois, cette métrique risque de ne pas être représentative pour des espèces aux architectures différentes et dont d'autres parties sont prélevées. Il sera alors nécessaire de calibrer les modèles avec d'autres métriques de terrain.

En somme, les modèles produits demeurent limités. L'utilisation de variables à haute résolution est insuffisante mais quelques améliorations techniques pourront probablement améliorer la performance des prédictions.

4.2. Des modélisations réalistes : à quel prix ?

La mise en place de modèles de distribution spatiale (SDM) implique de prendre de nombreuses décisions techniques qui auront un impact significatif sur la performance et la réalisme des modélisations. Dans le cadre de cette étude, nous avons fait le choix de sélectionner des variables à très haute résolution dans le but d'améliorer la précision des prédictions, particulièrement pour une utilisation à l'échelle locale. Cependant, il est essentiel de reconnaître que ces variables présentent encore des limites qui demandent des améliorations.

L'une des principales limites concerne la qualité des données utilisées. Certaines de ces variables à "haute résolution" sont issus d'algorithmes d'IA faisant de l'interpolation (SoilGrids notamment), et la qualité réelle de ces variables est complexe à évaluer. De plus, l'utilisation de variables avec une résolution supérieure à 10 mètres, telles que SoilGrids et Bioclim, soulève des interrogations sur la représentation adéquate des variations microclimatiques et microécologiques fines. Il pourrait être bénéfique d'augmenter la résolution de ces données ou d'exploiter davantage les données de télédétection et les informations d'élévation plus précises pour calculer des variables dérivées.

Par ailleurs, il est à noter que mon NDVI annuel correspond en réalité aux moyennes des NDVI des mois de mars à juin. En effet, pour tous les autres mois, la couverture nuageuse était trop importante pour intégrer les images dans le calcul de NDVI. Il serait intéressant de sélectionner des images sur 2 à 3 ans pour avoir un calcul plus complet, et éventuellement obtenir un NDVI saisonnier, et non plus annuel. Cependant, le temps de téléchargement, de traitement, ainsi que la quantité de stockage nécessaires pour manipuler les images satellite étaient bien trop importants pour que je puisse faire cela dans le cadre d'un stage. Malgré tout, les mois pour lesquels le calcul de NDVI a été possible correspondent au principal pic de biodiversité de l'année (printemps / début été), et donc là où les informations données sont les plus importantes (Eisfelder et al. 2023). En outre, une piste pour améliorer les séries temporelles de NDVI est l'utilisation de PCA (Principal Component Analysis) fonctionnelles (Pesaresi et al. 2020). En effet, celles-ci permettent de considérer le NDVI comme une fonction continue dans le temps en lissant les valeurs. Cela permet de reconstruire des séries temporelles complètes en comblant le manque de données qui peut être dû à la couverture nuageuse par exemple.

Un autre aspect à considérer est que les variables intégrées dans nos modèles représentent un état écologique ou environnemental à un moment donné. Dans ce travail, je n'ai pas étudié l'évolution temporelle de la distribution du thym. Ce type d'approche est classiquement mené dans le cadre du changement climatique pour prédire l'évolution des aires de distribution d'espèces face à différents scénarios (Stanton et al. 2012; Keith et al. 2008; Anderson et al. 2009). Dans le cadre de la cueillette, ajouter des variables temporelles (température, précipitations, empreinte humaine…) serait particulièrement intéressant pour étudier l'évolution de la ressource dans le temps. Cela permettrait de comparer les réactions des populations face à différentes pressions de cueillette. En outre, il a été noté que les SDM bénéficieraient grandement de l'intégration de variables dynamiques pour caractériser les processus démographiques de l'espèce, et non seulement spatiaux (Keith et al. 2008).

La qualité du jeu de données d'occurrence des espèces utilisé revêt également une importance capitale. Dans cette étude, nous avons opté pour des données non-protocolées incluant des observations d'experts (CBN) et des relevés issus de sciences participatives. Cependant, ces données comportent des limites significatives, telles que des erreurs de saisie et la présence de photos de plantes cultivées, de jardins ou parcs, qui ne correspondent pas nécessairement à leur distribution naturelle (Joly et al. 2016). Bien que nous ayons appliqué un filtrage important, notamment en excluant les données liées aux zones bâties, des erreurs résiduelles subsistent inévitablement (Zizka et al. 2019). De plus,

toutes les bases de données issues de sciences participatives sont biaisées spatialement en raison de l'inégalité des efforts d'échantillonnage, de stockage des données et de mobilisation (Beck et al. 2014). Ce biais est particulièrement prononcé pour le GBIF, où les différences nationales de financement et de partage de données entraînent d'énormes disparités dans les contributions. Cela peut impacter les modélisations, malgré tout, ce biais est probablement moins prononcé dans notre aire d'étude. En effet, le thym est une plante qui est très observée, et qui cumule de nombreuses observations sur la plupart de son aire de distribution. De plus, l'intégration des données précises du CBN Méditérannéen m'a permis d'améliorer encore le jeu de données.

Ensuite, un autre problème majeur des SDM réside dans le surajustement. Celui-ci peut avoir de multiples causes. Les modèles plus complexes de type ML y sont souvent sujets à cause de la complexité des relations matérialisées entre variables qui perdent alors en réalité écologique. Toutefois, le travail effectué ici a permis de détecter un éventuel surajustement grâce à la confrontation à un jeu de données indépendant. Celui-ci a permis de déceler un léger surajustement dans tous les algorithmes utilisés, sauf le GLM. Le modèle d'ensemble, RF et le GAM sont ceux qui surajustaient le plus. Cela est surprenant pour le modèle d'ensemble étant donné que plusieurs études ont confirmé sa capacité à limiter le surajustement (Hao et al. 2019; Grenouillet et al. 2011). Pour les GAM, cela laisse entendre que leur lissage doit être retravaillé. En outre, le modèle d'ensemble et le RF sont aussi ceux qui ont détecté les effets site les plus importants (d'après les calculs de R^2). Nous pouvons alors supposer que ces deux algorithmes sont particulièrement sensibles à la structuration spatiale de la présence du thym. Il a été montré dans la littérature que les SDM pouvaient parfois surajuster les dépendance spatiales des processus écologiques impliqués (Roberts et al. 2017). Suite à ces constatations, une analyse spatiale des effets site a été faite en Annexe E qui montre un effet moins important du site en cœur de niche tandis qu'il augmente vers les bords. Cela mène à supposer une spatialisation des erreurs de prédiction avec potentiellement de moins bonnes prédictions en limite d'aire. Nous pourrions en effet émettre l'hypothèse que lorsque nous nous approchons des abords de la niche, cela augmente la quantité de facteurs entrant en compte pour définir la présence du thym. Ainsi cela rend plus difficile la matérialisation et l'explication de la présence du thym à partir de variables écologiques/biologiques/environnementales…

Ainsi, une limite importante de mes modèles réside dans leur sensibilité à l'organisation spatiale de l'occurrence du thym qui entrave leur pouvoir prédictif. L'amélioration des prédictions pourra passer par l'intégration de nouvelles variables pour mieux représenter l'environnement local. Un ajustement plus fin des modèles est aussi à étudier, notamment pour les GAM.

4.3. Quelles perspectives pour une gestion locale de la biodiversité ?

Les SDM réalisés ici ouvrent des perspectives intéressantes pour la gestion des espèces cueillies, mais aussi pour la conservation des plantes en général. Les SDM sont largement décrits comme étant des outils précieux dans le cadre de la conservation (Rodríguez et al. 2007; Franklin 2013; Villero et al. 2017). Il sont notamment utilisés pour évaluer la vulnérabilité d'espèces cibles face au changement climatique (Sinclair, White, et Newell 2010; Durner et al. 2009). Une grande force des SDM est leur potentiel pour la cartographie d'espèces à enjeux qui permet une caractérisation objective de la situation. Ils sont donc particulièrement utiles pour surpasser le "knowing-doing gap" (ou la difficulté à transformer la connaissance obtenue par la recherche appliquée en actions) qui limite souvent la concrétisation d'opérations de conservation (Pfeffer et Sutton 1999). Dans le cadre de la cueillette, nous pourrions déterminer les zones plus vulnérables et organiser des systèmes de rotation pour limiter les impacts sur les zones subissant des prélèvements excessifs. Cela permettrait également d'évaluer la ressource de manière quantitative, pour mettre en place des quotas localement. Par ailleurs, des modélisations précises pourraient

même permettre d'identifier des zones favorables aux réintroductions d'espèces en danger.

Cette étude a permis de répondre à bon nombre de questions d'ordre conceptuel et technique. Toutefois, pour pouvoir réellement utiliser les SDM à des fins de conservation, la méthodologie nécessite d'être améliorée. Pour cela, plusieurs axes de travail à explorer durant la thèse ont été définis. Tout d'abord, il me semble important de mieux quantifier les biais d'échantillonnage des données d'occurrence, particulièrement le biais spatial. Celui-ci pourra éventuellement être matérialisé en créant des cartes d'accessibilité à partir de données de pente, de rugosité du sol, et de proximité aux chemins / villes. La détectabilité de l'espèce sera aussi cruciale à évaluer surtout dans le cas de la généralisation de la méthodologie à d'autres espèces dont l'appareil aérien n'est pas aussi visible que celui du thym. Cela pourra être fait par des expérimentations sur le terrain en utilisant le TTD (temps à la première détection) notamment (Halstead, Rose, et Kleeman 2021; Bornand et al. 2014).

Ensuite, il semble essentiel de parvenir à quantifier la pression de cueillette, afin de pouvoir l'intégrer comme variable dans les modélisations. Ce point comporte beaucoup de difficultés. Plusieurs tentatives ont déjà été effectuées par les CBN, sans grand succès. La méthode la plus commune est la création d'exclos au niveau de zones de cueillette pour comparer l'évolution de la population au sein de l'exclos, et en dehors. Cependant, ceux-ci ne sont jamais réellement respectés par les cueilleurs, empêchant d'obtenir des résultats fiables. Une autre idée est l'utilisation de pièges-photo pour compter le nombre de cueilleurs, mais il faudrait avoir l'accord de toutes les personnes photographiées, ce qui est impossible. La meilleure méthode imaginée pour l'instant est l'utilisation comme point de référence les Parcs et Réserves Naturels dans lesquels la cueillette est interdite, ou alors les zones très peu accessibles. En effet, les zones les plus accessibles sont généralement celles qui subissent le plus de prélèvements. Nous pourrions alors coupler les périmètres des réserves aux cartes d'accessibilité mentionnées avant pour définir des zones de référence. L'objectif serait alors de comparer l'état de la population des ces zones témoin à d'autres populations subissant de la cueillette. L'idéal étant d'avoir des sites aux caractéristiques très similaires, et la cueillette pour seule différence.

Après, la technique des SDM mise en place ici pourrait être améliorée en utilisant des algorithmes plus complexes, adaptés à nos questionnements. Des pistes intéressantes sont les modèles de type "Point-Process", et l'utilisation de réseaux de neurones convolutifs (CNN) dans les DeepSDM. Ces derniers utilisent des techniques d'apprentissage automatique qui sont connues pour leur capacité à capturer des relations complexes. Toutefois, il nécessitent un travail d'entraînement conséquent. Ensuite, les modèles "Point-Process" sont des modèles qui, contrairement aux SDM classiques utilisant des grilles, prédisent des densités de points individuels au sein d'une aire d'étude (Mugumaarhahama, Fandohan, et Glèlè Kakaï 2023). Ces modèles pourraient donc présenter un intérêt important dans le cadre de la quantification de la ressource, d'autant plus en prenant en compte nos résultats qui montrent une relation directe entre le nombre de touffes de thym sur une zone et son recouvrement.

Un autre volet essentiel consiste à améliorer les variables intégrées dans les modélisations. En effet, les résultats ont révélé une large part d'effets inexpliqués dû au site. Il est donc impératif d'accroître la part des effets fixes expliqués, ce qui peut être réalisé en incorporant de nouvelles variables plus représentatives de l'écologie du thym. L'ajout de variables représentatives de la géologie du sol pourrait être particulièrement intéressant. Un bon exemple sont les variables développées par Lehmann et al. (2010) pour la Suisse, cependant nous manquons encore d'une couverture globale. Par ailleurs, les variables issues des données de télédétection sont encore peu utilisées mais pourraient améliorer les modélisations (Zimmermann et al. 2007; Schwager et Berg 2021). Il existe plusieurs variables topographiques calculables à partir d'un modèle numérique de terrain (MNT) qui pourraient être utilisés. Les variables "northness" (exposition au nord) et "eastness"

(exposition à l'est) donnent notamment des informations importantes sur l'exposition au soleil et la température (Zimmermann et al. 2007; Schwager et Berg 2021). Celles-ci peuvent avoir des conséquences déterminantes sur la croissance et la survie des plantes. Ensuite, l'humidité du sol est l'un des facteurs les plus importants influençant la composition et la structure de la végétation (Raduła, Szymura, et Szymura 2018). Celle-ci peut être estimée en utilisant différents indices. Il existe notamment le TWI (Topographic Wetness Index, calculé à partir d'un MNT) qui est un proxy pour l'accumulation d'eau et de nutriments (Riihimäki et al. 2021; Sørensen, Zinko, et Seibert 2006). Il existe aussi le NDMI (Normalised Difference Moisture Index) calculé à partir des bandes spectrales NIR (Near Infra-Red) et SWIR (Short Wave Infra-Red) des images satellite de type Sentinel (Beaury et al. 2023). Celui-ci permet de détecter le niveau d'humidité de la végétation et pourrait éventuellement être couplé avec le TWI pour former un indice composite d'humidité. Par ailleurs, les données LiDAR donnent de bonnes perspectives pour caractériser finement des variables biotiques, souvent manquent dans les SDM. Elles peuvent notamment représenter la végétation avec une très haute résolution (Farrell et al. 2013). En particulier, elles permettent de calculer la taille et la composition de la canopée, ce qui nous permettrait d'obtenir une description verticale précise de l'habitat (He et al. 2015). Nous pouvons aussi en extraire la densité de tiges, qui pourrait être pertinente dans l'estimation de la ressource de certaines espèces cueillies.

5. Conclusion

En conclusion, à travers ce travail, je cherchais à savoir si l'occurrence et l'abondance des plantes sauvages pouvait être prédite à fine échelle spatiale. Les résultats obtenus sont prometteurs. En particulier, cette recherche a démontré la capacité des SDM à haute résolution à relier la présence et l'abondance du thym aux variables intégrées dans les modèles. Cette réussite valide l'hypothèse sous-jacente de la théorie de la niche écologique et d'une abondance accrue en cœur de niche. Cependant, l'étude a également révélé d'importants effets liés au site d'observation et à son histoire, qui demeurent encore largement inexpliqués. Il subsiste de nombreux effets fixes à identifier qui bénéficieront de l'amélioration des variables choisies pour la modélisation.

Les modèles de ML ont montré de bonnes performances globales (RF et MAXENT en particulier). Le modèle d'ensemble et le RF ont montré une tendance à au surajustement, bien que leur performance reste parmi les plus élevées. Le modèle d'ensemble semble notamment bénéficier d'une performance améliorée grâce à la combinaison de modèles. En revanche, le GAM n'a pas apporté de réelle valeur ajoutée par rapport aux autres modèles.

Sur le terrain, les métriques développées ont permis de mesurer efficacement le recouvrement du thym de manière simple et reproductible. Il est également envisageable de simplifier davantage le protocole en mesurant seulement le nombre de touffes de thym dans chaque quadrat, une mesure directement liée au recouvrement. Toutefois, il reste à vérifier si cette approche est adaptable à d'autres espèces présentant des architectures et caractéristiques différentes. Enfin, nous avons révélé une tendance constante à surestimer le recouvrement à vue d'un facteur 2. Cette observation est cruciale pour les futures études écologiques, étant donné l'utilisation généralisée de cette méthode.

Finalement, ce stage s'inscrit dans un projet plus long que je vais poursuivre en thèse. A terme, l'objectif est de généraliser l'étude à toutes les espèces cueillies. Bien entendu, il ne sera pas possible d'effectuer des expérimentations sur l'ensemble des 700 à 1000 espèces concernées. C'est pourquoi j'effectuerai un travail pour former des groupes d'espèces aux caractéristiques similaires pour lesquelles je choisirai des plantes modèles qui me permettront de valider les méthodologies mises en place. En outre, je travaillerai sur la hiérarchisation des priorités de conservation en couplant des approches écologiques et sociologiques pour mieux comprendre les enjeux associés aux plantes cueillies.

Annexes

Annexe A : Mise en place d'un modèle de distribution du thym (*Thymus vulgaris*) à l'échelle du Paléarctique Ouest

1. Introduction

Dans un premier objectif de définition de l'aire de distribution du thym, des SDM ont été réalisés à l'échelle du Paléarctique Ouest avec une résolution de 1km². Le Paléarctique Ouest englobe l'Europe, l'Afrique du Nord et des parties de l'Asie de l'Ouest (Fig. A.1). Pour cette modélisation, j'ai choisi de faire un SDM d'ensemble avec *Biomod2*.

2. Matériel et méthodes

2.1. Nettoyage des données de présence

Nettoyage des données GBIF :

Tout d'abord, lors du téléchargement des données GBIF (grâce au package *rgbif* sur R), seules les données présence ayant une précision inférieure à 1000 m ont été éliminées, de même pour les observations faites avant l'an 2000 et celles sans coordonnées GPS. De plus, le système de "flags" du GBIF a été utilisé pour éliminer toutes les données ayant des "problèmes d'ordre géospatial", soit les observations dont les coordonnées sont invalides, ont été arrondies…

Les données téléchargées ont ensuite été soumises à un nettoyage permis par le package *CoordinateCleaner* sous R. Celui-ci comprend des tests automatisés permettant de repérer facilement (et d'exclure) les enregistrements attribués au centroïde d'un pays ou d'une province, à la mer, au zones densément urbanisées, au siège du GBIF, ou à l'emplacement des institutions chargées de la biodiversité (musées, zoos, jardins botaniques, universités). Il identifie en outre, pour chaque espèce, les coordonnées aberrantes, les coordonnées nulles, les latitudes/longitudes identiques et les coordonnées non valides.

Nettoyage des données CBN :

Ensuite, pour les données du CBN Méditérannéen, seules les observations faites au GPS (2 à 10 m de précision) depuis l'an 2000 ont été retenues.

Dernières étapes de filtrage :

Les données de présence issues du GBIF et du CBN ont alors été réunies dans un même tableau, puis ont été filtrées pour ne garder qu'une seule observation par cellule de 1km² (en se calant sur la géométrie des rasters prédicteurs).

Récapitulatif du filtrage :

- Initialement : 29316 observations
- Filtrage *CoordinateCleaner* (filtre "urban" inclus): 19061 observations (-10255)
- Filtrage pour ne garder qu'1 observation / km² : 1448 observations (-17613)

2.2. Sélection des points de pseudo-absence

Ici, j'ai décidé de générer autant de points d'absence que de points de présence, de manière randomisée sur la totalité de l'aire d'étude (Fig A.1).

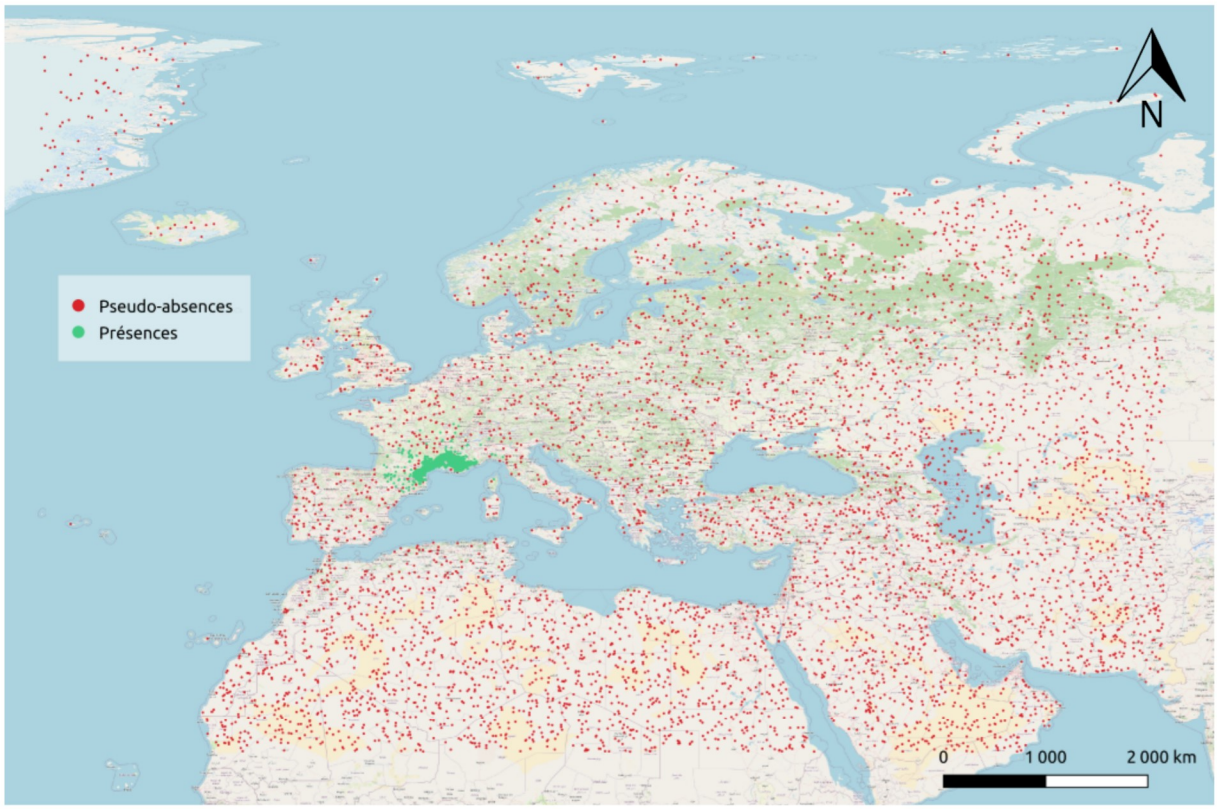


Figure A.1 : Carte des présence et des pseudo-absences utilisés pour la modélisation réalisée à l'échelle du Paléarctique Ouest.

2.3. Choix des variables prédictives utilisées en entrée du modèle

Pour ajuster le modèle, j'ai décidé d'intégrer trois types de variables : des variables bioclimatiques, un MNT ainsi que les propriétés du sol (Tab A.1).

Table A.1 : Tableau récapitulatif des variables utilisées en entrée de la modélisation réalisée à l'échelle du Paléarctique Ouest.

Nom de la variable	Courte description	Source	Résolution
Bioclim	Il s'agit de 19 variables bioclimatiques dérivées des valeurs températures et précipitations mensuelles moyennes.	Bioclim	1 km (30 arcsec)
Modèle numérique de terrain	Les valeurs d'élévation.	RGE Alti	1 m
SoilGrids	Cartographie mondiale des caractéristiques du sol entre 0 et 200 cm de profondeur. Les propriétés décrites sont le pH, la masse volumique du sol, la quantité de nitrogène contenue, la CEC (capacité d'échange cationique), la fraction de fragments grossiers (>2 mm), la concentration en carbone organique, ainsi que les fractions de sable, argiles et limons.	SoilGrids	250 m

Pour voir les définitions des variables individuelles, se reporter à l'annexe B Tab. B.1.

Tous les rasters de prédicteurs ont été alignés sur la géométrie des rasters Bioclim. Autrement dit, une reprojection à été faite lorsque cela était nécessaire. De plus, les rasters ont été rognés au niveau de l'aire d'étude, et les données SoilGrids de précision 250m ont été rééchantillonnées* à 1 km.

Ensuite, j'ai décidé d'éliminer toutes les variables corrélées à plus de 70%, pour éviter de biaiser le modèle (Fig A.2). Cela a été fait avec la fonction *select07* du package *mecofun* sous R.

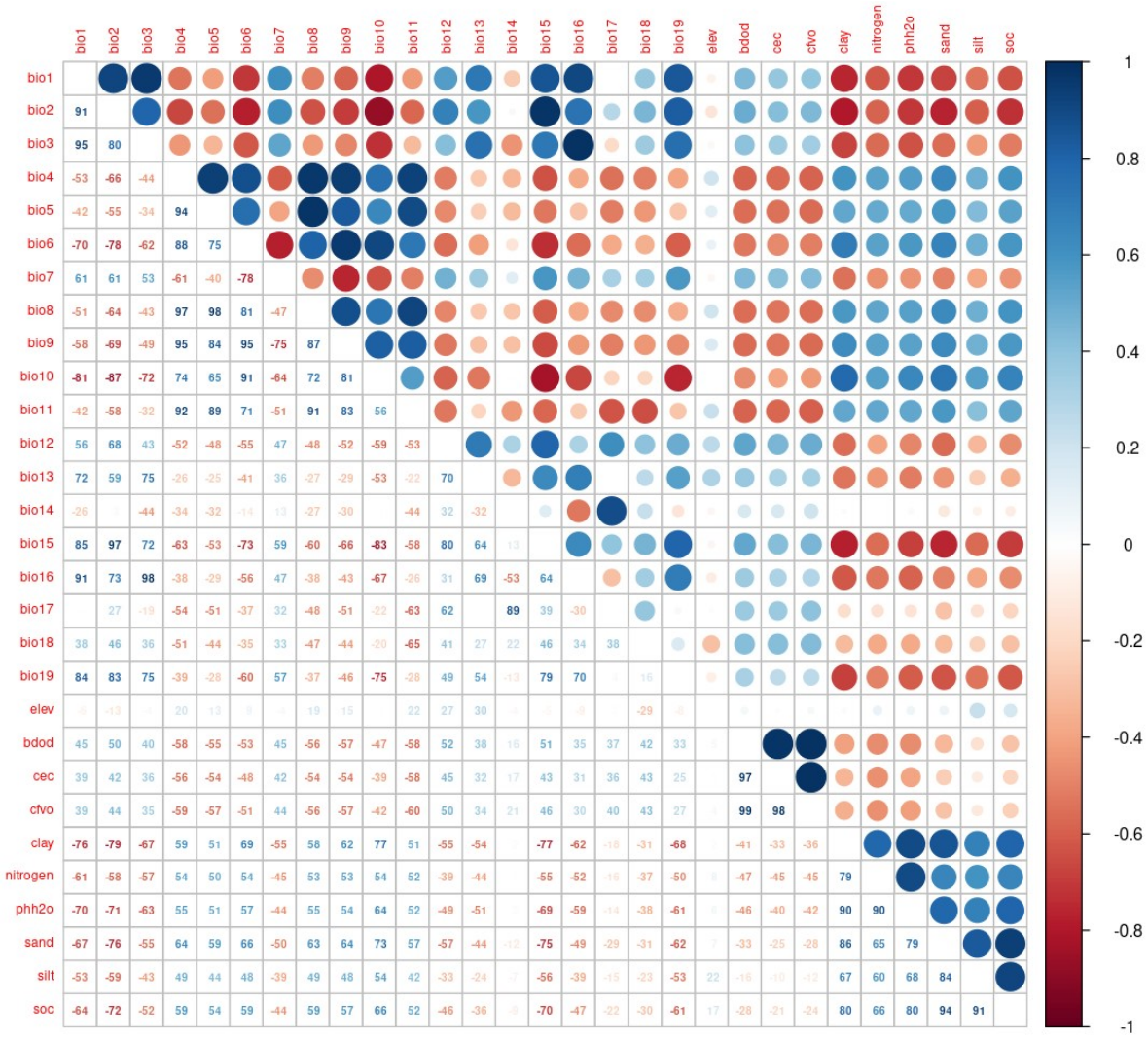


Figure A.2 : Matrice de corrélation des variables à sélectionner pour la modélisation réalisée à l'échelle du Paléarctique Ouest (package *corrplot*, fonction *corrplot.mixed*).

A l'issue de cette sélection, 10 variables ont été retenues : bio3, bio7, bio11, bio12, bio17, bio18, cfvo, sand, nitrogen et elev (Fig A.3).

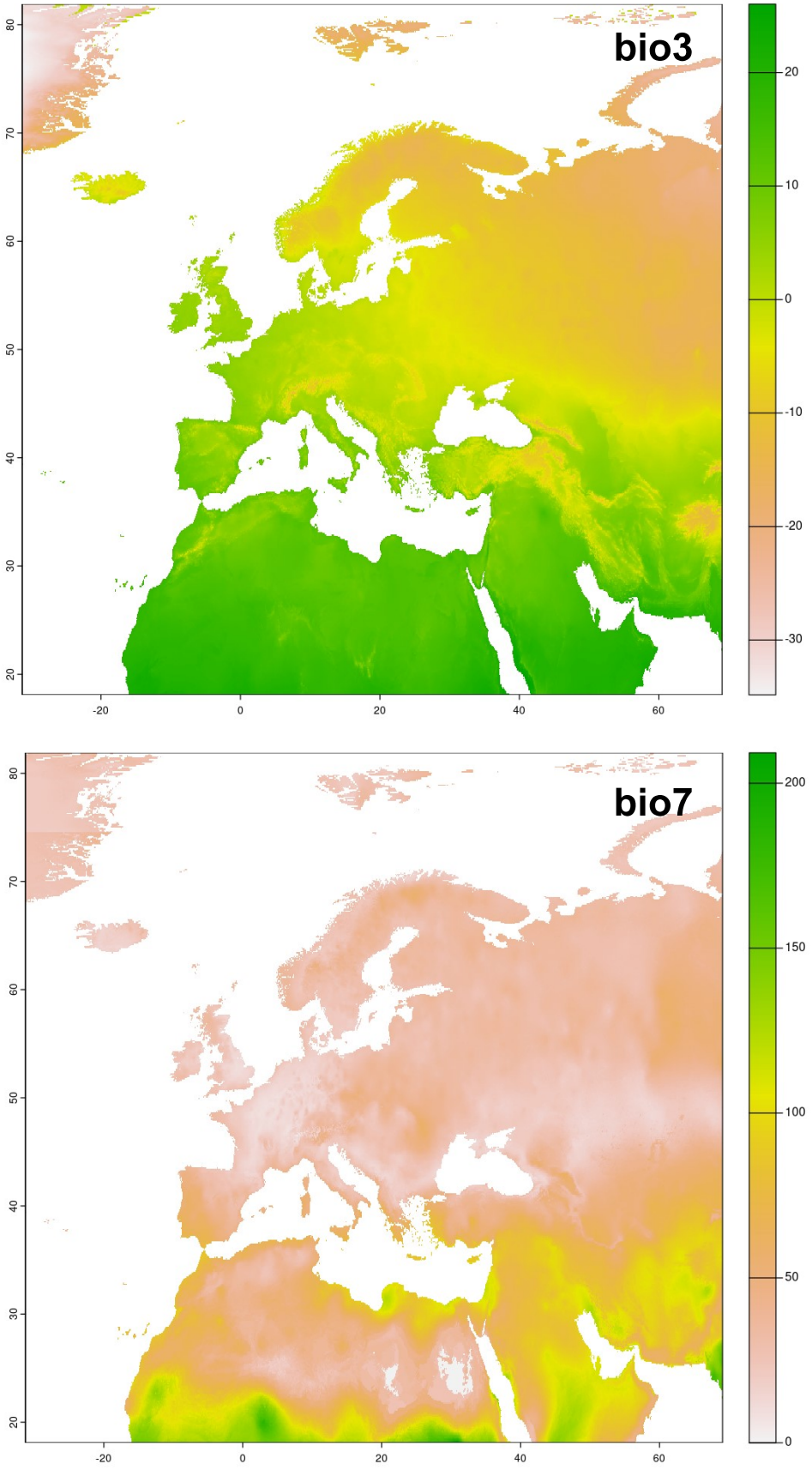
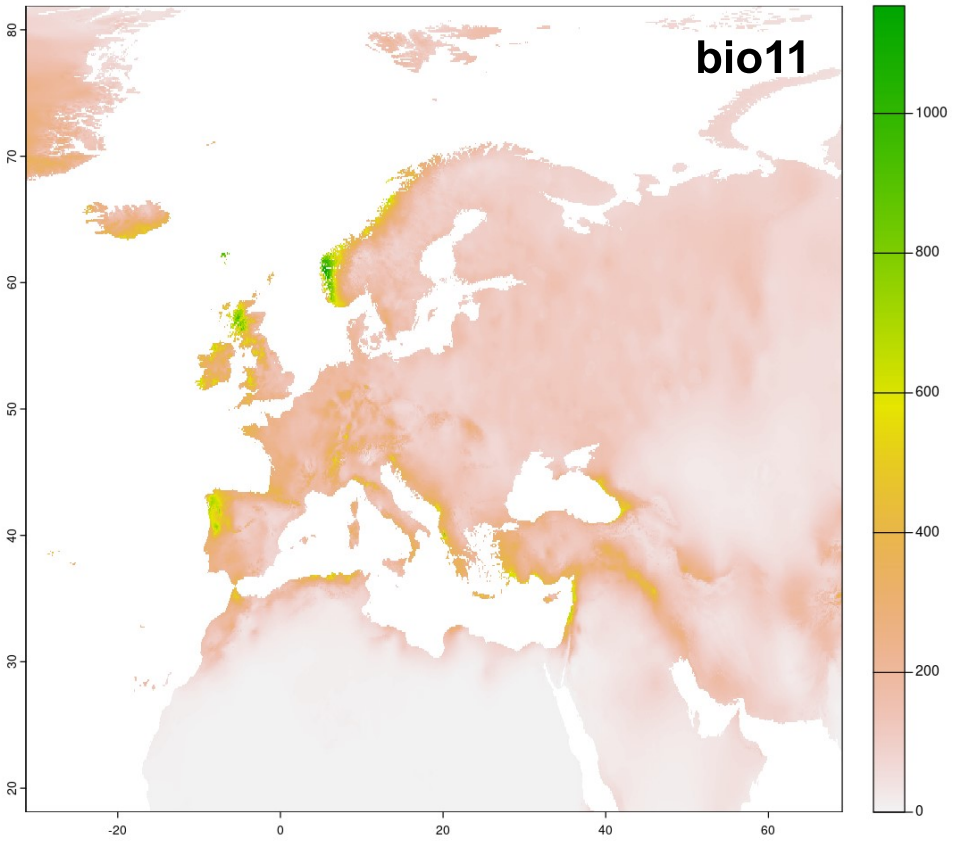


Figure A.3 : Rasters des variables retenues pour la modélisation réalisée à l'échelle du Paléarctique Ouest.



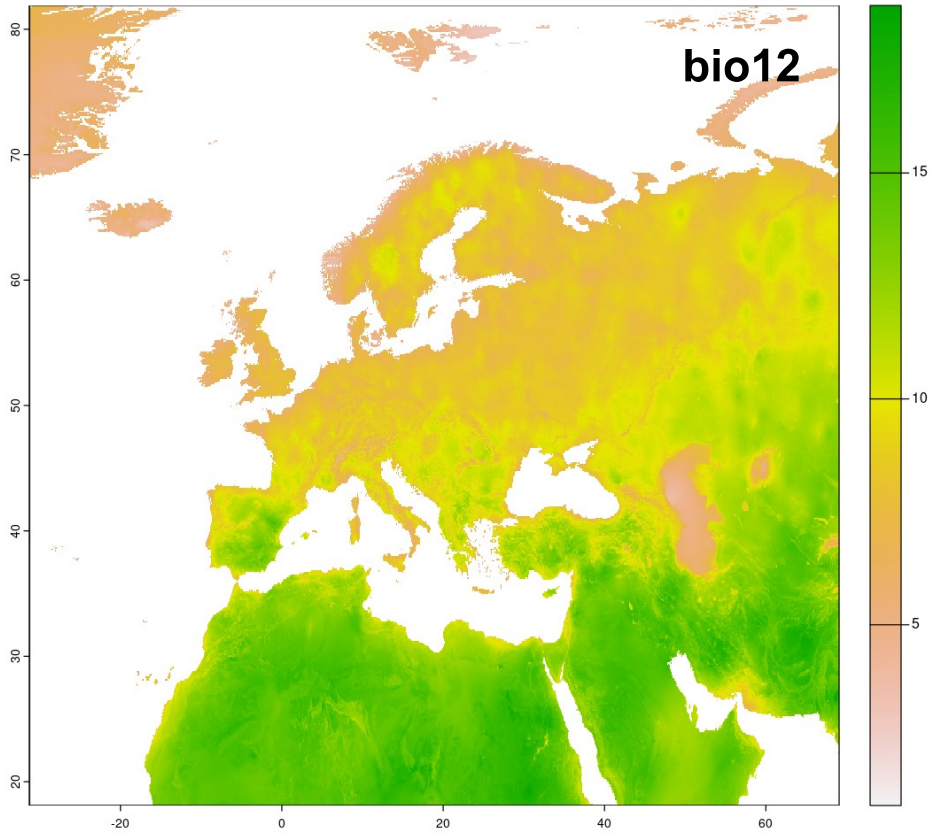


Figure A.3 (suite) : Rasters des variables retenues pour la modélisation réalisée à l'échelle du Paléarctique Ouest.

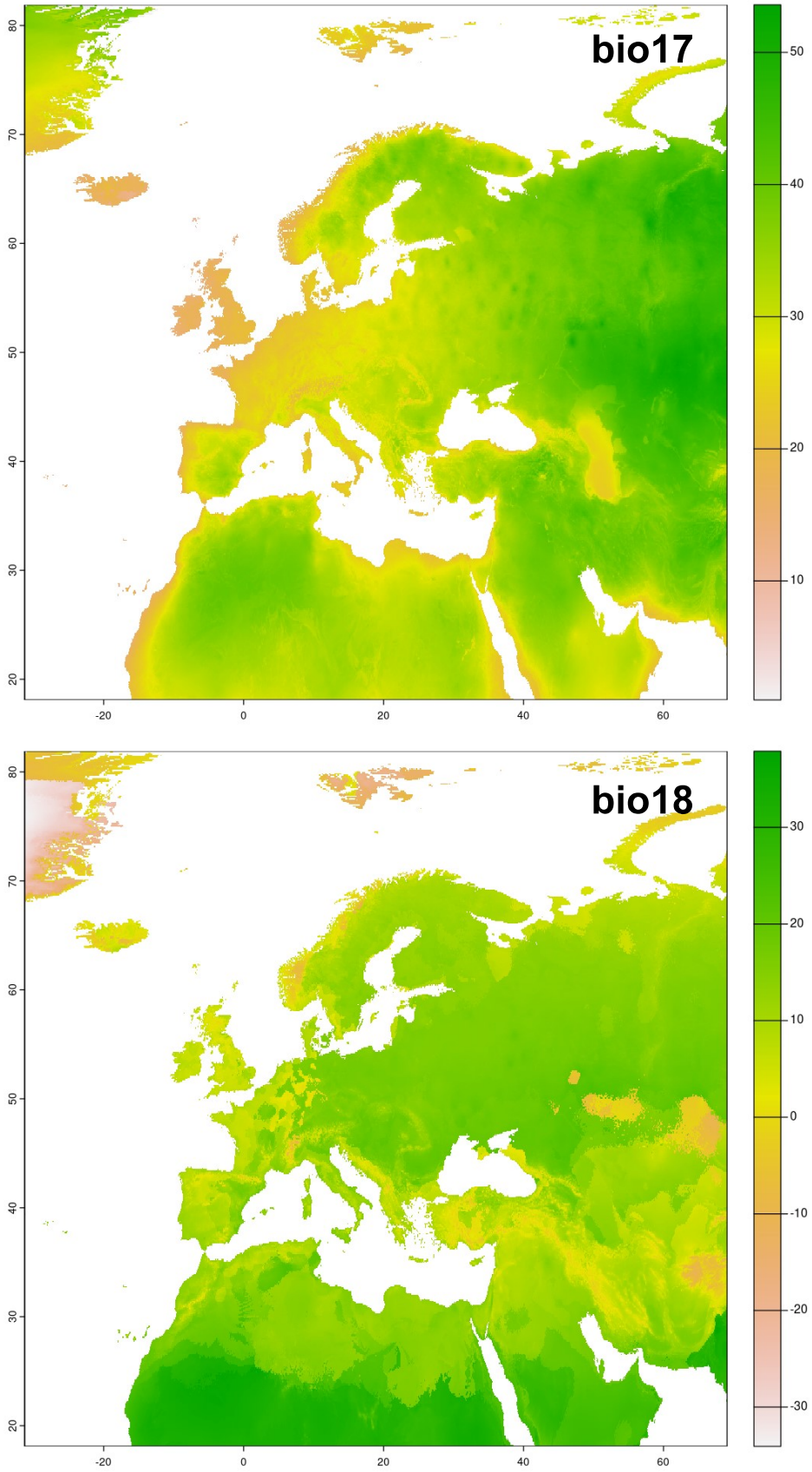


Figure A.3 (suite) : Rasters des variables retenues pour la modélisation réalisée à l'échelle du Paléarctique Ouest.

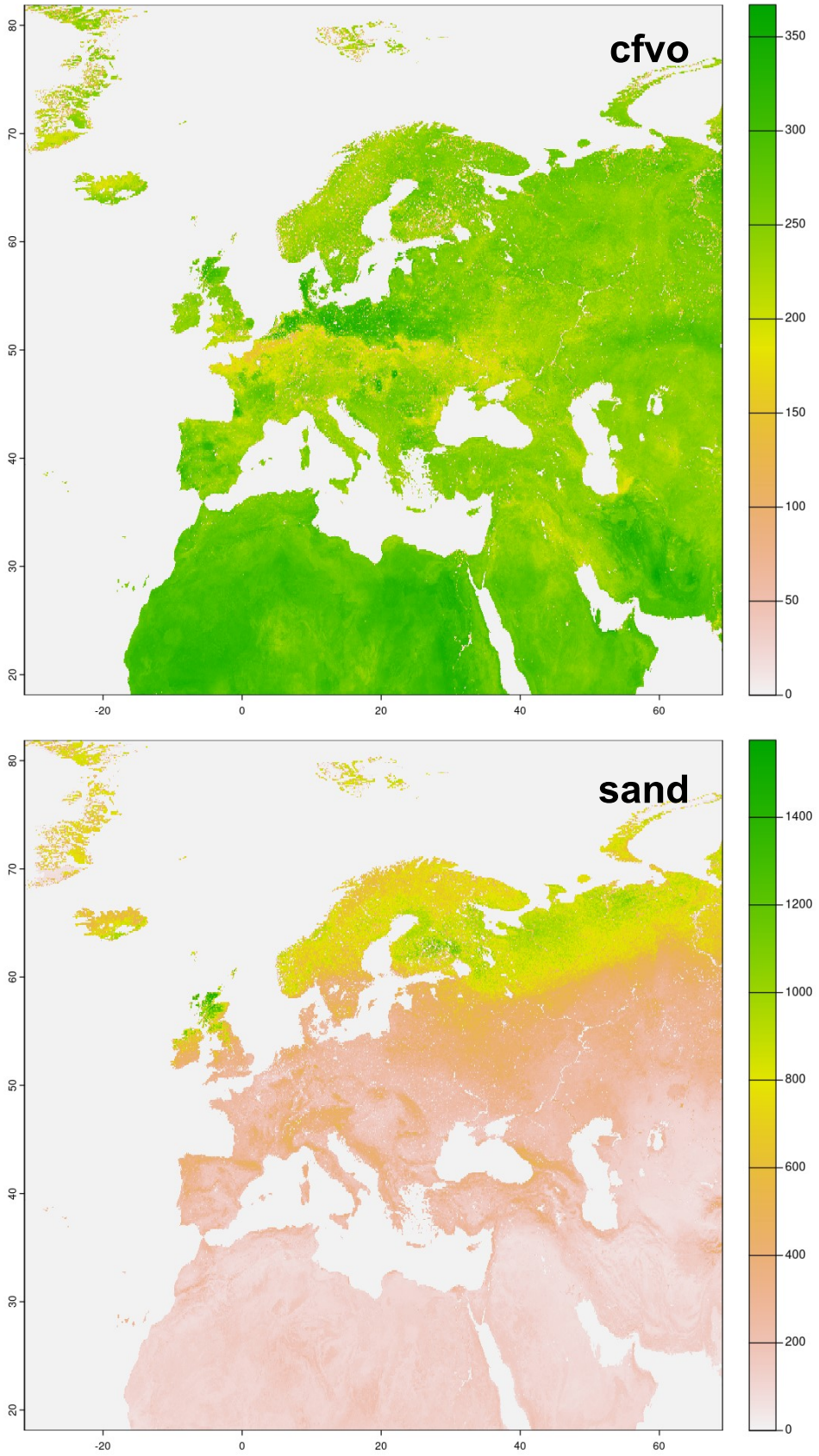


Figure A.3 (suite) : Rasters des variables retenues pour la modélisation réalisée à l'échelle du Paléarctique Ouest.

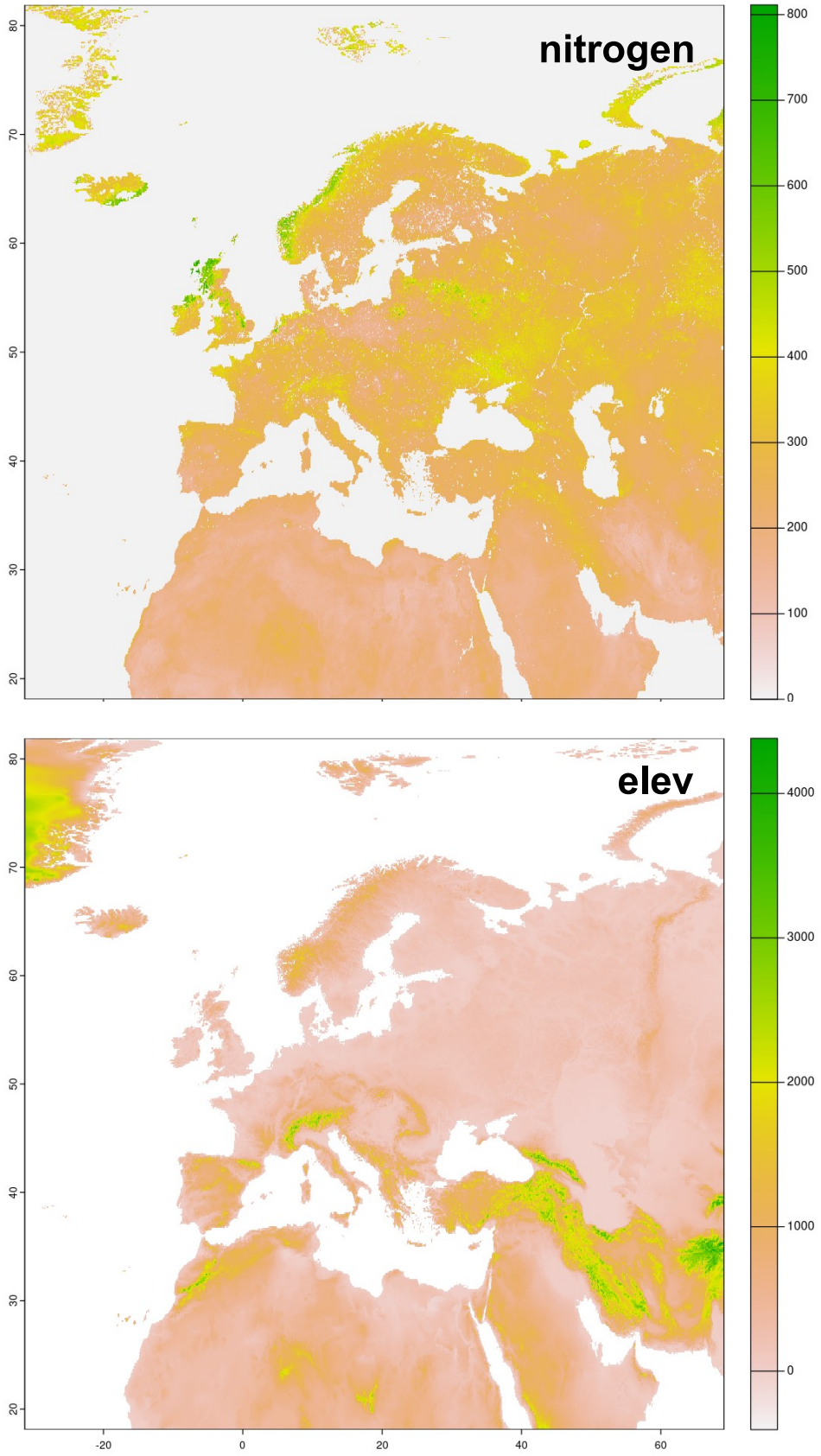


Figure A.3 (suite) : Rasters des variables retenues pour la modélisation réalisée à l'échelle du Paléarctique Ouest.

2.4. Mise en place du modèle

Les algorithmes choisis pour cette modélisation sont GLM, GAM, MARS, CTA, GBM, RF et MAXENT.

J'ai gardé des options de modélisation de base avec :

- GLM : simple (d'ordre 1), sans interaction, et avec une famille binomiale
- GBM : nombre d'arbres de 1000
- GAM : algorithme gam_mgcv, type s_smoother, sans interaction, famille binomiale et method REML
- Pour les autres algorithmes, ce sont les paramètres par défaut de *Biomod2* qui ont été gardés.

J'ai ensuite partitionné le jeu de données en deux : 70% pour l'entraînement du modèle et 30% pour sa validation croisée, et réalisé 10 répétitions *.

2.5. Assemblage des modèles

Pour l'étape d'assemblage de modèles*, j'ai effectué une moyenne pondérée de tous les modèles individuels issus des différents algorithmes en fonction de leurs scores TSS.

2.6. Prédictions à partir du modèle

J'ai pu prédire la probabilité de présence du thym grâce aux fonctions déjà intégrées dans *Biomod2* (fonction *BIOMOD_projection*). Ainsi, j'ai réalisé des prédictions pour le modèle d'ensemble, mais aussi des modèles le constituant pour voir leur performance individuelle.

2.7. Evaluation du modèle d'ensemble et des modèles individuels le constituant

Ici, l'évaluation des modèles est automatiquement faite par *Biomod2*. L'évaluation avec ces métriques est permise par le partitionnement en deux du jeu de données (70-30%). Il suffit juste de sélectionner les métriques. Ainsi, j'ai sélectionné le TSS et AUC-ROC qui sont les métriques les plus couramment utilisées pour ce type de modélisation.

3. Résultats de la modélisation

Grâce à cette modélisation, j'ai pu créer la carte Fig. 1 présentée en début du Matériel et Méthodes. Cette carte a été réalisée, à partir du modèle d'ensemble complet pour lequel j'ai calculé le point de basculement correspondant au score TSS le plus élevé. Ce point de basculement (ou cutoff) est en fait un seuil qui se réfère à la valeur limite utilisée pour délimiter la présence ou l'absence d'une espèce. J'ai donc utilisé ce seuil pour transformer les valeurs continues de mes prédictions en valeurs binaires de présence/absence (0/1). En d'autres termes, si la valeur prédite dépasse le cutoff, j'ai considéré que l'espèce était présente, sinon elle était considérée comme absente.
Ci-dessous se trouve la carte de distribution du thym, avec les valeurs de prédiction brutes, non transformées en présences/absences (Fig. A.4).

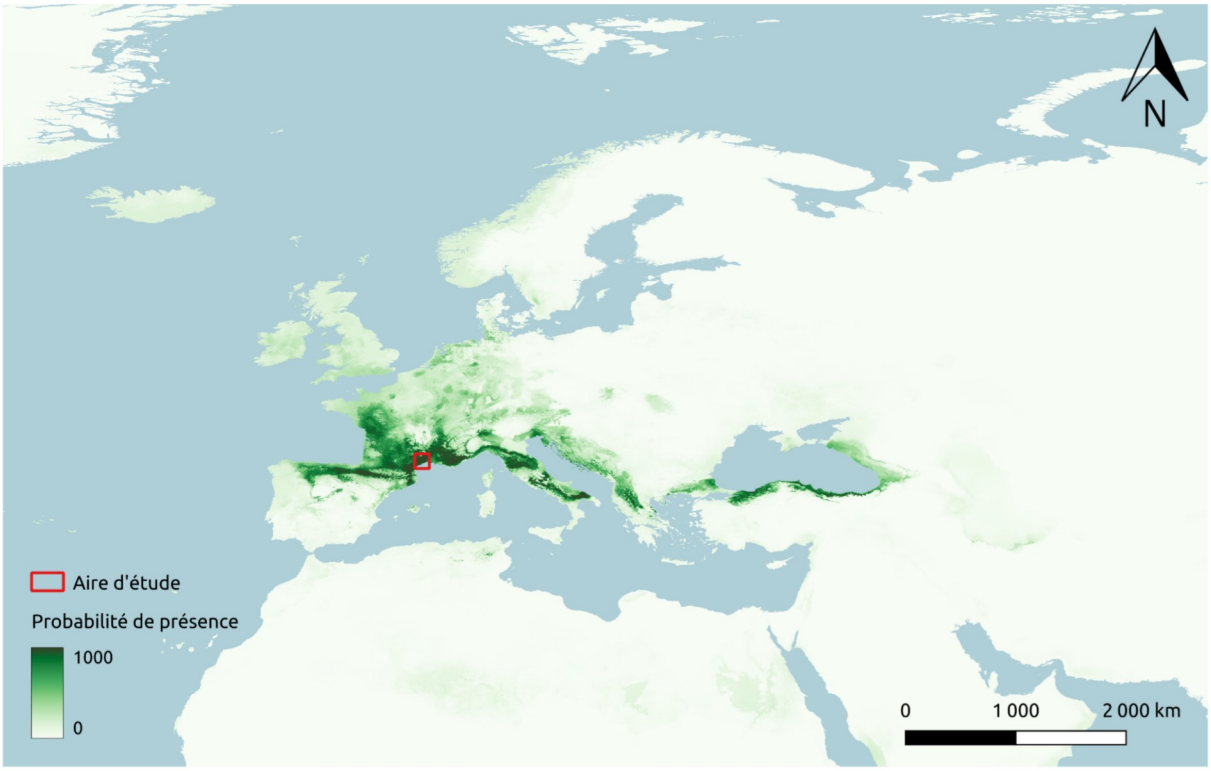


Figure A.4 : Probabilité de présence du thym (*Thymus vulgaris*) dans le Paléarctique Ouest d'après le modèle d'ensemble réalisé avec *Biomod2*.

Lors de l'ajustement des modèles avec *Biomod2*, une importance est attribuée à chaque variable selon la valeur ajoutée qu'elle apporte au modèle final. Ces importances sont présentées en Tab A.2.

Table A.2 : Importance des variables (entre 0 et 100) pour chaque modèle individuel réalisé dans le cadre de la modélisation du thym à l'échelle du Paléarctique Ouest. L'importance des variables mesure la contribution de chaque variable à la prédiction de la distribution de l'espèce. Elle est calculée ici en faisant 3 permutations.

algorithme	bio11	bio12	bio17	bio18	bio3	bio7	cfvo	elev	nitrogen	sand
CTA	**41.4**	0.45	12.98	8.84	20.39	**28.61**	10.01	**0.09**	**0.02**	**0**
GBM	**41.07**	**0.01**	3.2	5.53	16.92	**28.33**	2.38	**0.01**	**0**	**0**
GLM	11.05	1.19	**0.06**	2.78	**42.52**	**40.29**	4.81	**0**	10.02	**0.11**
MARS	**36.85**	**1.83**	**0**	**0**	**33.76**	25.91	**2.36**	**0**	**0**	**0**
MAXENT	**21.24**	4.82	9.8	4.82	**23.87**	19.65	5.63	**4**	**2**	14.01
RF	21.57	2.56	**21.75**	14.34	20.83	**22.82**	8.62	0.79	**0.37**	**0.67**
GAM	18.89	19.34	28.32	7.74	**34.35**	**27.06**	6.38	3.73	**0.13**	**1.38**
Ensemble	**21.03**	1.11	4.84	2.68	21.00	**23.12**	2.92	**0.30**	**0.20**	0.35

Les deux plus hautes contributions sont précisées en orange, et les deux plus basses en bleu. Les variables qui ont été éliminées sont sur fond gris.
Nous pouvons observer que la présence de thym est particulièrement dépendante des variables bio3 (isothermalité), bio11 (température du quart le plus froid) et bio7 (amplitude de température annuelle). Ensuite, nous pouvons également obtenir les courbes de réponse pour chaque variable, qui sont présentées en Fig A.5.

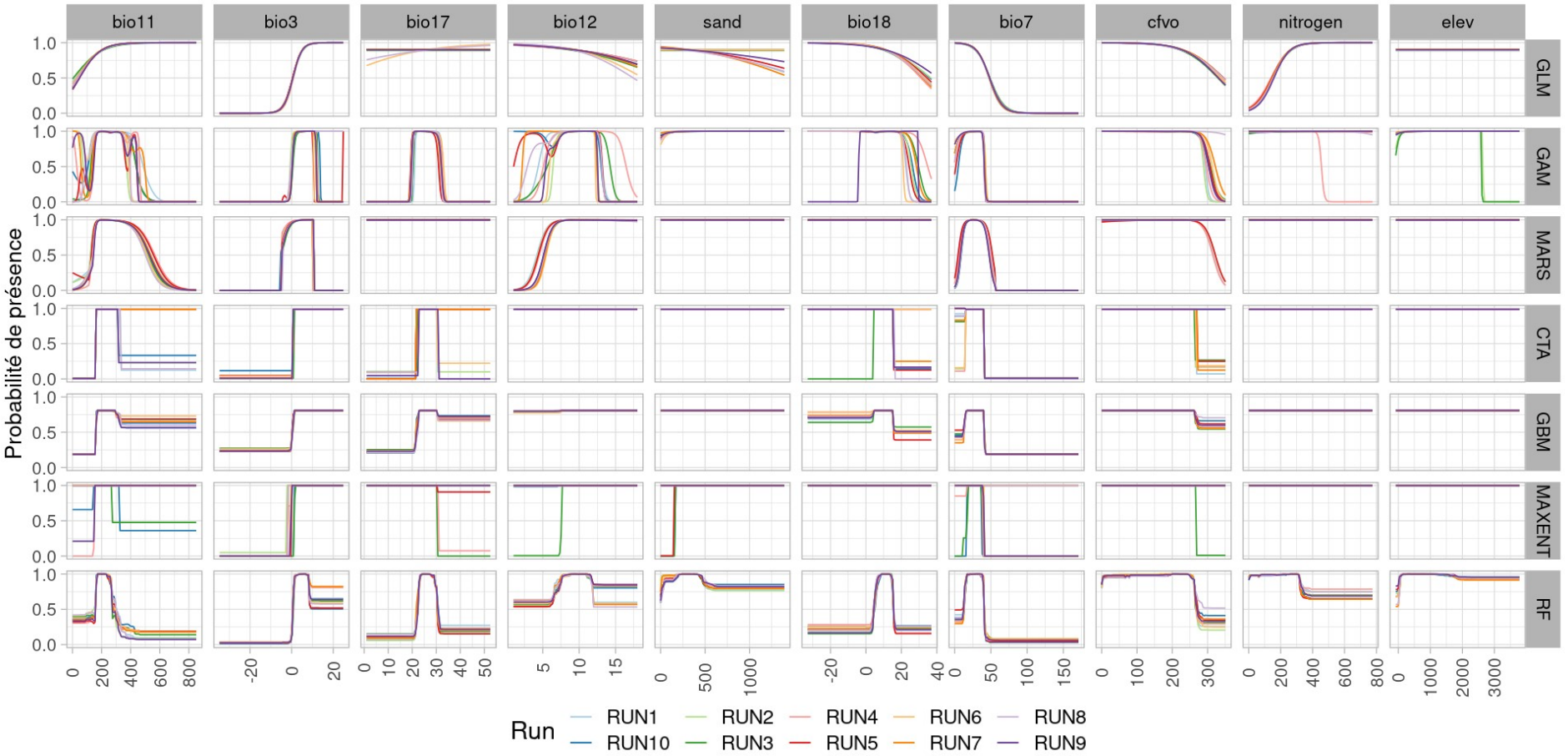


Figure A.5.a : Courbes de réponse pour chaque algorithme et "run" (répétition) effectués dans le cadre de la modélisation du thym à l'échelle du Paléarctique Ouest.

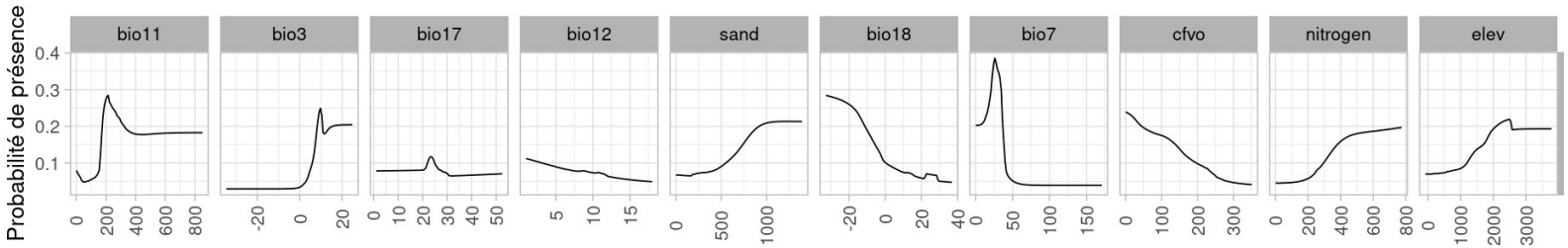


Figure A.5.b : Courbes de réponse pour le modèle d'ensemble effectué dans le cadre de la modélisation du thym à l'échelle du Paléarctique Ouest.

Enfin, j'ai souhaité évaluer les performances du modèle d'ensemble réalisé, ainsi que des modèles individuels le constituant (Tab A.3 et Fig A.7). La performance des modèles individuels détermine directement leur poids dans le modèle d'ensemble final (Fig. A.6).

Table A.3 : Evaluation du modèle d'ensemble réalisé à l'échelle du Paléarctique Ouest, et des modèles individuels le constituant.

algorithme	métrique	seuil	sensitivité	specificité	score
GLM	TSS	515	99.147	90.111	0.893
GLM	AUC	508.5	99.221	90.037	0.947
GBM	TSS	457	99.055	98.111	0.972
GBM	AUC	454.5	99.055	98.111	0.995
CTA	TSS	599	98.999	98.389	0.975
CTA	AUC	604	98.962	98.5	0.992
MARS	TSS	556	98.813	97.37	0.962
MARS	AUC	556.5	98.813	97.37	0.995
RF	TSS	495	99.703	99.333	0.99
RF	AUC	519.5	99.666	99.407	1
MAXENT	TSS	166	98.572	98.5	0.971
MAXENT	AUC	191.5	98.498	98.648	0.996
GAM	TSS	677	99.277	98.926	0.982
GAM	AUC	663	99.314	98.907	0.998
Ensemble	TSS	523*	99.277	98.704	0.98
Ensemble	AUC	516.5	99.351	98.685	1

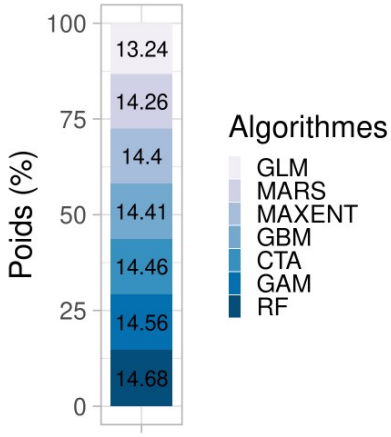


Figure A.6 : Poids des modèles intégrés dans la modélisation d'ensemble du thym à l'échelle du Paléarctique.

*Seuil utilisé pour la création de la carte de présence absence du thym (Fig. 1).

La colonne "score" correspond au score TSS ou ROC moyen calculé pour chaque algorithme.

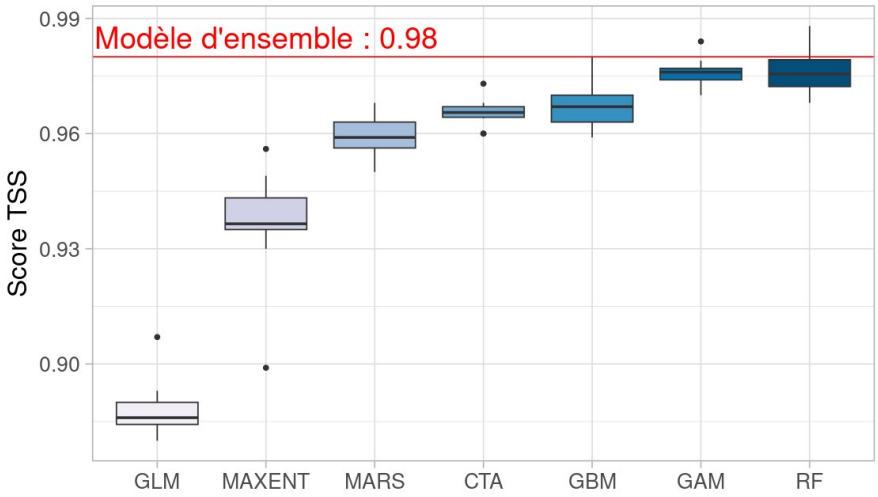


Figure A.7 : Scores TSS calculé par *Biomod2* pour chacun des algorithmes utilisés dans le cadre de la modélisation du thym à l'échelle du Paléarctique. Les boîtes à moustache sont basées sur les scores TSS calculés sur le jeu de données de validation (30% restants) de chacun des 10 "runs" effectués pour chaque algorithme. La ligne horizontale centrale indique la médiane.

4. Discussion et conclusions

Cette première modélisation nous donne une bonne idée de l'aire complète de distribution de thym. Les résultats montrent que sa distribution est particulièrement liée à la température. Cette espèce a une distribution méditerranéenne très marquée dont les principales caractéristiques sont un été chaud et sec suivi d'un hiver doux et humide. Il n'est donc pas surprenant de voir que l'aire générale de distribution du thym dépend directement de la température. Par ailleurs, les scores TSS et ROC montrent une très bonne performance des modèles, tous au-dessus de 0,89. Malgré tout, la question d'un éventuel surajustement se pose étant donné que les modèles n'ont pas été validés avec un jeu de données indépendant. Finalement, la principale limite de ces modèles de distribution à grande échelle est le filtrage des données de présence utilisées. En effet, la plupart des données intégrées proviennent des sciences participatives et incluent de nombreuses erreurs. Un filtrage a été mis en place, mais ne semble pas tout à fait suffisant. En effet, en comparant les cartes obtenues avec la distribution connue du thym, nous pouvons remarquer que les prédictions indiquent une présence un peu trop au nord de sa distribution réelle. De plus, une analyse plus fine des points de présence utilisés montre que tous les points situés au latitudes les plus hautes sont situés dans des zones habitées. Il s'agit donc très probablement d'observations de thym cultivé faites dans les jardins qui biaisent le modèle. Nous devons donc envisager la mise en place d'un meilleur filtrage, comme celui qui a été fait à l'échelle locale (Fig. 2). En outre, il sera intéressant de comparer quelles variables ressortent dans les SDMs à large et fine échelle pour mieux comprendre comment s'organise la macro- et micro-niche du thym.

Annexe B : Sélection des variables pour la modélisation du thym (*Thymus vulgaris*) à très haute résolution

Table B.1 : Tableau récapitulatif des variables utilisées en entrée de la modélisation réalisée à l'échelle locale. En **bleu** les 14 variables retenues pour les modélisations.

Variable	Courte description	Source	Résolution
bio1	température annuelle moyenne	Bioclim	1 km
bio2	moyenne des amplitudes de température journalières (température maximale - minimale)	Bioclim	1 km
bio3	isothermalité (bio2/bio7) (×100)	Bioclim	1 km
bio4	saisonnalité des températures (écart-type x 100)	Bioclim	1 km
bio5	température maximale du mois le plus chaud	Bioclim	1 km
bio6	température minimale du mois le plus froid	Bioclim	1 km
bio7	amplitude de température annuelle (bio5 - bio6)	Bioclim	1 km
bio8	température moyenne du quart le plus pluvieux	Bioclim	1 km
bio9	température moyenne du quart le plus sec	Bioclim	1 km
bio10	température moyenne du quart le plus chaud	Bioclim	1 km
bio11	température moyenne du quart le plus froid	Bioclim	1 km
bio12	précipitations annuelles	Bioclim	1 km
bio13	précipitations du mois le plus pluvieux	Bioclim	1 km
bio14	précipitations du mois le plus sec	Bioclim	1 km
bio15	saisonnalité des précipitations (coefficient de variation)	Bioclim	1 km
bio16	précipitations du quart le plus pluvieux	Bioclim	1 km
bio17	précipitations du quart le plus sec	Bioclim	1 km
bio18	précipitations du quart le plus chaud	Bioclim	1 km
bio19	précipitations du quart le plus froid	Bioclim	1 km
phh2o	pH de l'eau (pH *10)	SoilGrids	250 m
bod	masse volumique de la fraction fine du sol (cg/cm3)	SoilGrids	250 m
nitrogen	quantité de nitrogène contenue en cg/kg	SoilGrids	250 m
cec	Capacité d'Échange Cationique à ph 7 (mmol(c)/kg)	SoilGrids	250 m
cfvo	fraction de fragments grossiers (>2 mm)	SoilGrids	250 m
soc	concentration en carbone organique (dg/kg)	SoilGrids	250 m
sand	fraction de sable en g/kg	SoilGrids	250 m
clay	fraction d'argiles en g/kg	SoilGrids	250 m
silt	fraction de limons en g/kg	SoilGrids	250 m
NDVI2022	Normalised Difference Vegetation Index pour 2022	dérivé des images Sentinel 2 (Copernicus Open Access Hub)	10 m
elev	valeurs d'élévation	RGE Alti	1 m
slope	pente	dérivé de elev	1 m
aspect	orientation de la pente, en degrés	dérivé de elev	1 m
TPI	Topographic Position Index, différence entre la valeur d'une cellule et la moyenne des 8 cellules adjacentes	dérivé de elev	1 m

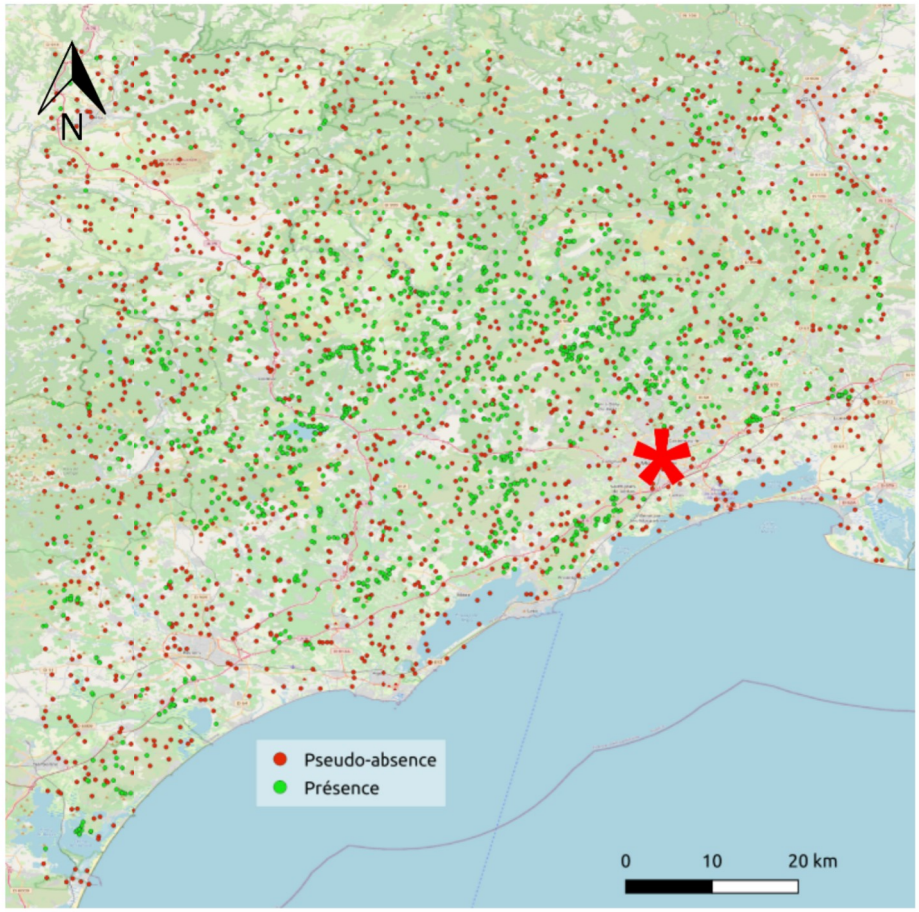


Figure B.1 (à gauche) : Carte des présence et des pseudo-absences utilisés pour la modélisation réalisée à l'échelle locale. Montpellier est localisé par l'étoile rouge.

Figure B.2 (à droite) : Délimitation du bâti dans la zone d'étude, avec 200m de zone tampon, couche utilisée pour filtrer les données de présence de thym utilisées pour les SDM à très haute résolution. Créé à partir du Plan Cadastral Informatisé (https://cadastre.data.gouv.fr/datasets/cadastre-etalab). Montpellier est localisé par l'étoile rouge.

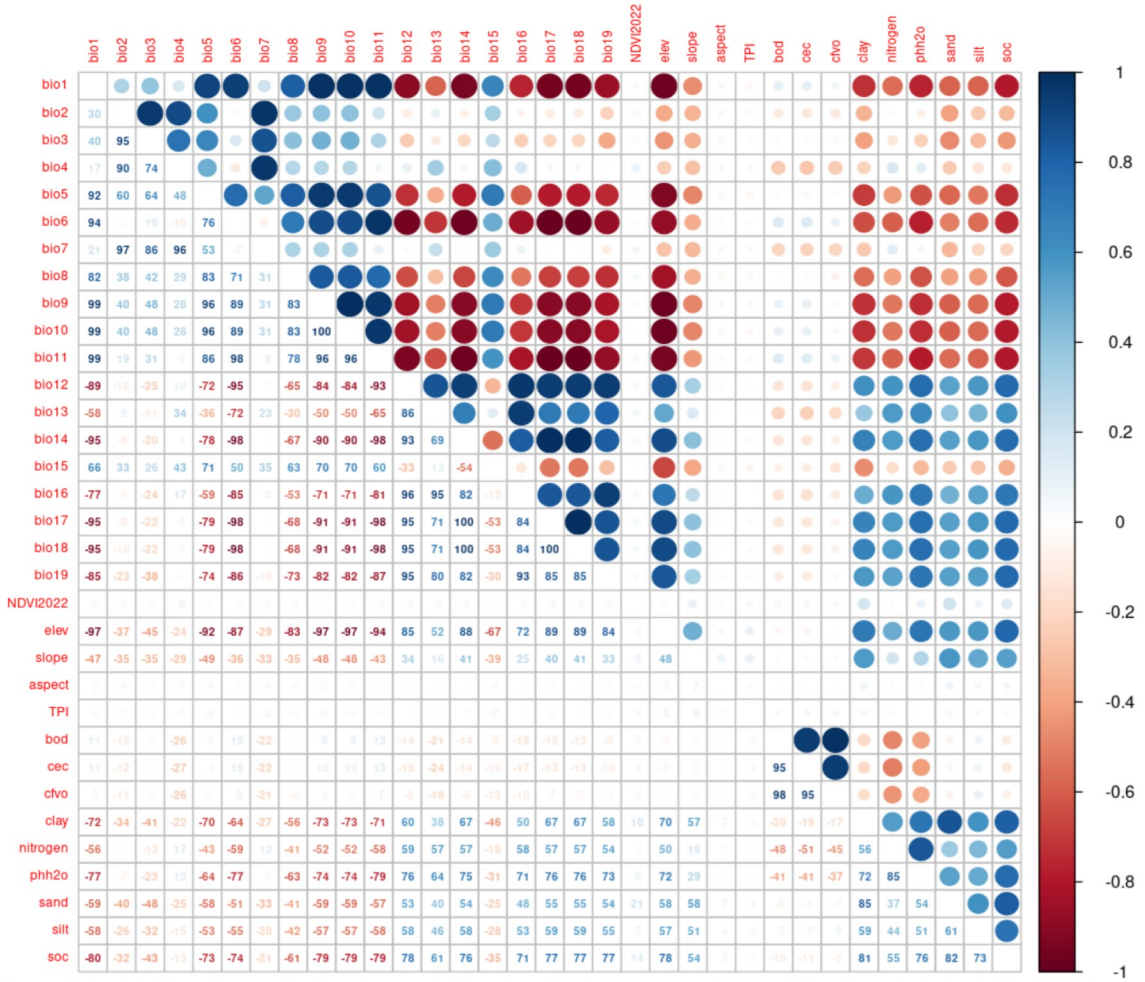


Figure B.3 : Matrice de corrélation des variables à sélectionner pour la modélisation réalisée à l'échelle locale (package *corrplot*, fonction *corrplot.mixed*).

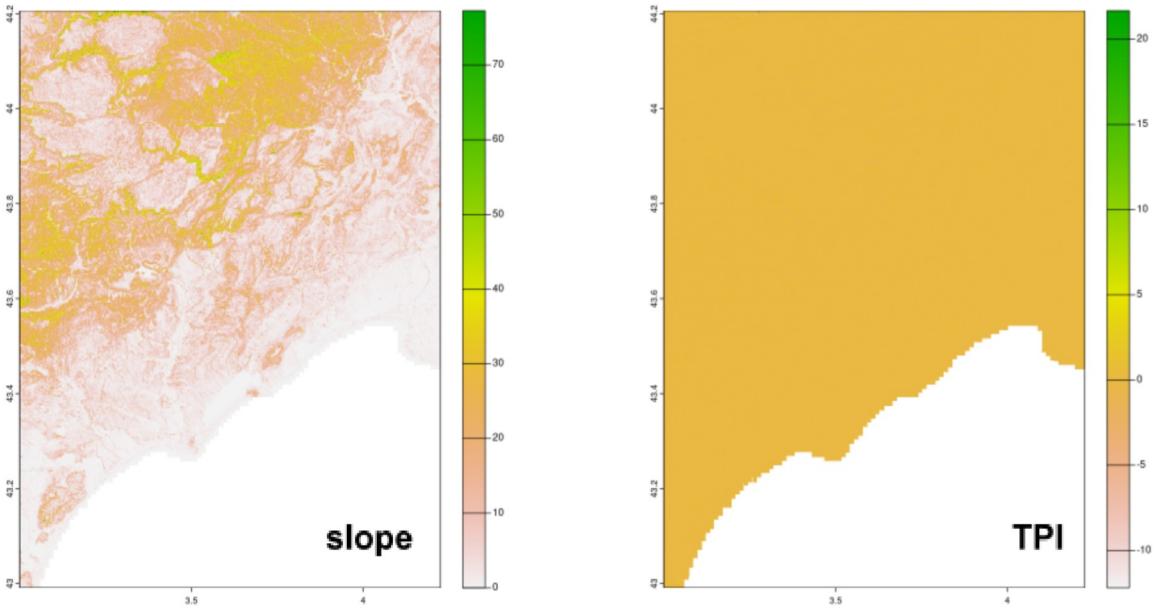


Figure B.4 : Rasters des variables retenues pour la modélisation réalisée à l'échelle locale.

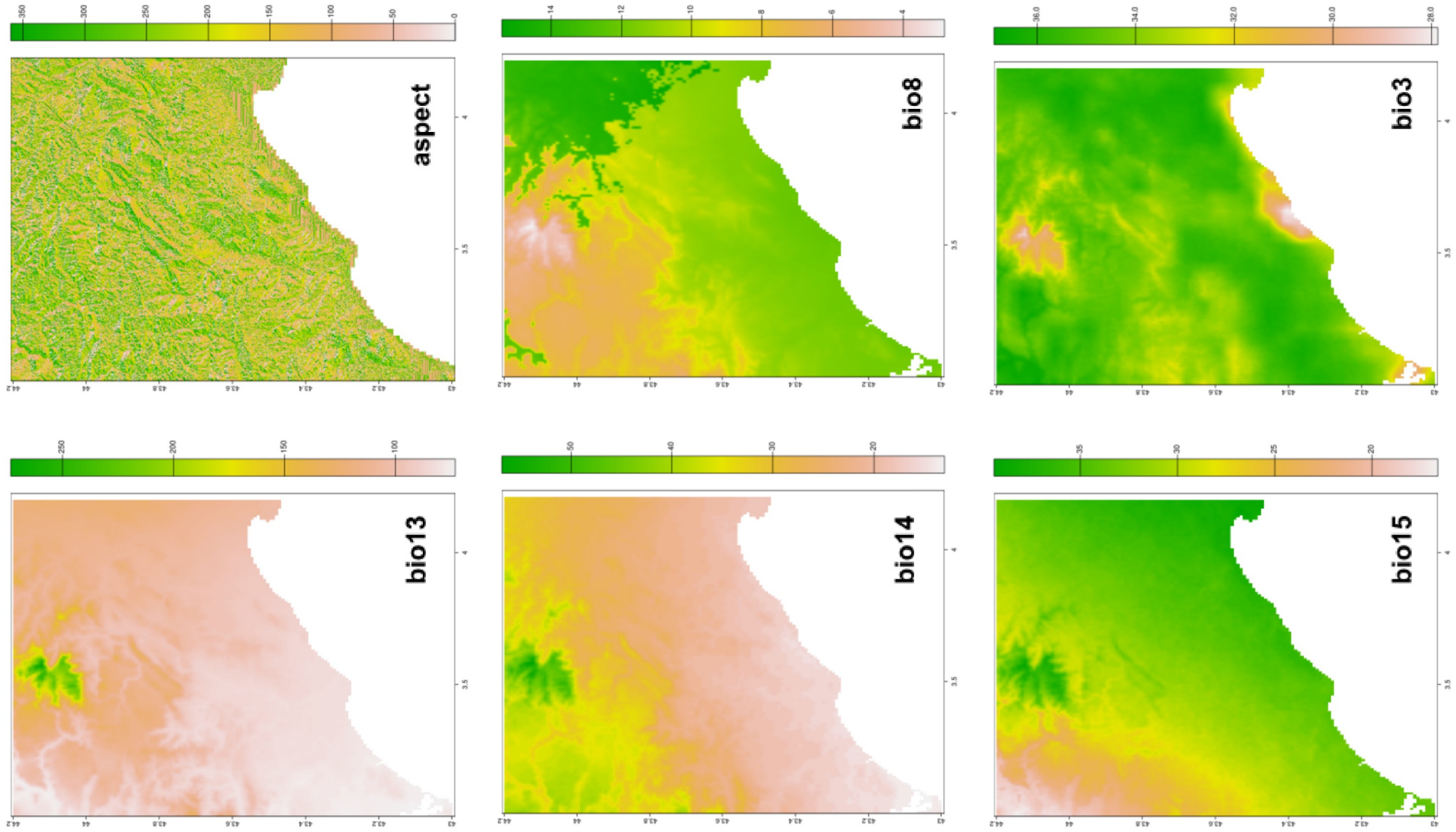


Figure B.4 (suite) : Rasters des variables retenues pour la modélisation réalisée à l'échelle locale.

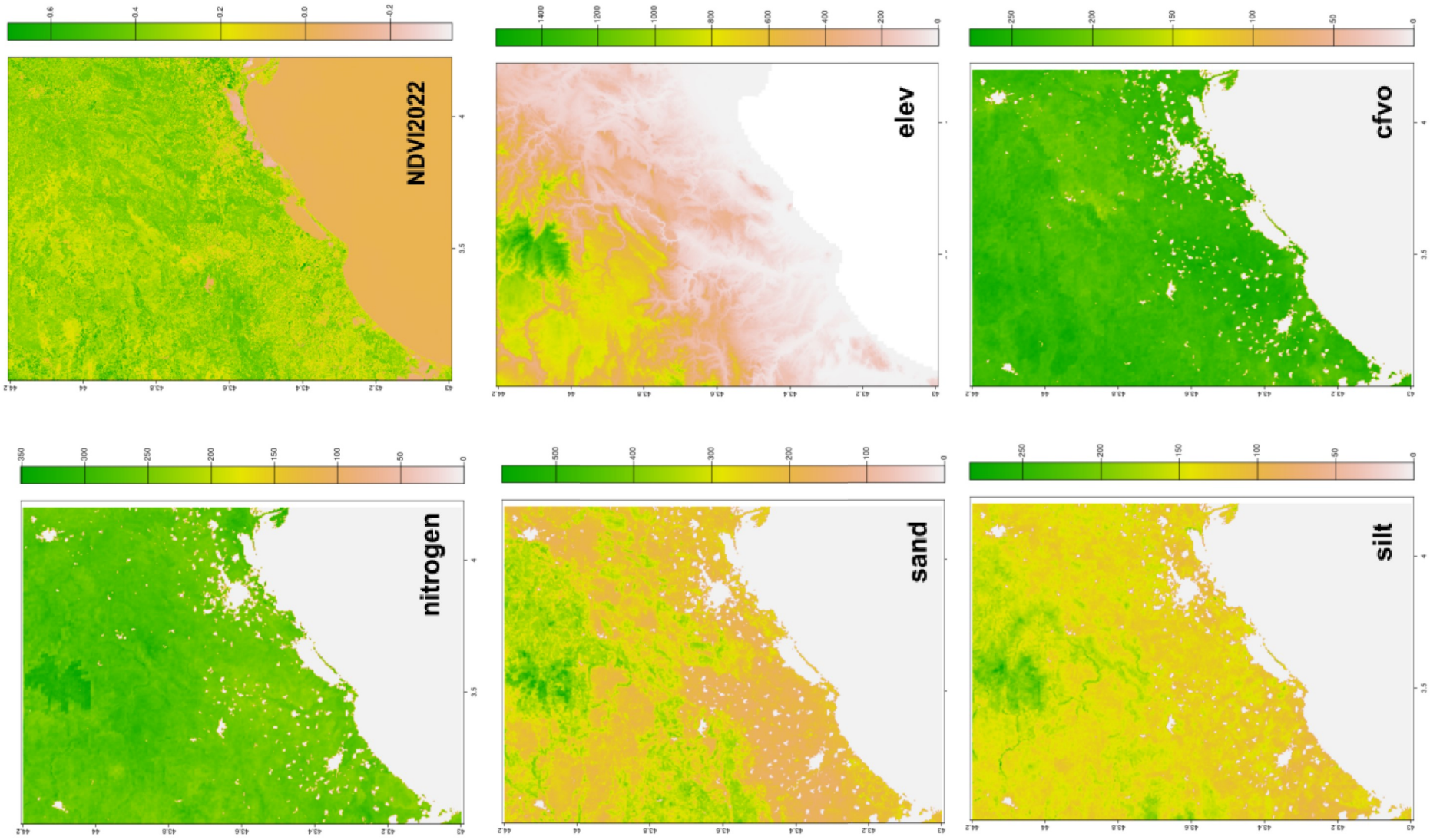


Figure B.4 (suite) : Rasters des variables retenues pour la modélisation réalisée à l'échelle locale.

Annexe C : Détails sur les modèles de distribution réalisés à l'échelle locale

Table C.1 : Importance des variables (entre 0 et 100) pour chaque modèle individuel réalisé dans le cadre de la modélisation du thym à l'échelle locale. L'importance des variables mesure la contribution de chaque variable à la prédiction de la distribution de l'espèce. Elle est calculée ici en faisant 3 permutations.

algorithme	aspect	bio13	bio14	bio15	bio3	bio8	cfvo	elev	NDVI2022	nitrogen	sand	silt	slope	TPI
CTA	0.9	3.8	47.5	18.0	1.2	46.0	1.0	1.7	6.4	0.9	11.0	3.7	15.9	0.7
GBM	0	0.1	45.7	7.0	0.2	26.3	0.1	1.5	2.5	0	2.8	1.0	4.6	0.1
GLM	1.8	51.7	38.3	64.7	1.2	4.7	14.4	34.6	0.1	43.2	0.3	26.3	2.3	0
MAXENT	2.5	6.0	29.2	19.4	3.3	13.6	5.2	16.2	6	1.8	19	5.7	3.6	8.4
RF	0.9	2.8	16.3	6.7	2.4	10.3	1.1	4.2	4.6	1.1	3.4	2.7	3.7	1.2
MARS	0	30.6	45.8	43.8	0.9	9.3	1.9	49.1	3.5	0	15.0	7.8	3.3	5.8
Ensemble	0.3	3.9	36.4	18.9	0.4	11.1	0.7	13.7	2.9	0.8	4.6	2.7	3.0	1.8
GAM	0.5	40.6	35.3	60.7	2.1	6.0	4.9	52.7	3.5	2.0	22.1	14.9	0.8	4.1

Les deux plus hautes contributions sont précisées en orange, et les deux plus basses en bleu. Les variables qui ont été éliminées sont sur fond gris.

Table C.2 : Evaluation du modèle d'ensemble réalisé à l'échelle locale, des modèles individuels le constituant, et évaluation du GAM ajusté séparément.

algorithme	métrique	seuil	sensitivité	spécificité	score
GLM	TSS	456	80.113	51.218	0.314
GLM	AUC	469.5	77.604	53.898	0.714
GBM	TSS	502	78.733	72.229	0.513
GBM	AUC	502.5	78.67	72.655	0.829
CTA	TSS	417	86.386	71.437	0.58
CTA	AUC	421.5	86.261	71.924	0.854
MARS	TSS	474	81.368	67.174	0.486
MARS	AUC	471.5	81.681	66.931	0.818
RF	TSS	539	95.797	96.529	0.923
RF	AUC	538.5	95.797	96.529	0.994
MAXENT	TSS	353	80.176	71.864	0.52
MAXENT	AUC	349.5	80.489	71.559	0.846
ensemble	TSS	491	83.752	79.111	0.629
ensemble	AUC	503.5	82.811	80.146	0.909
GAM	TSS	464	82.497	66.991	0.494
GAM	AUC	470.5	82.058	67.600	0.817

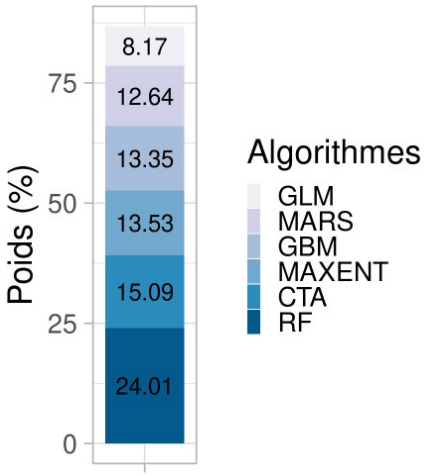


Figure C.1 : Poids des modèles intégrés dans la modélisation d'ensemble à très haute résolution du thym.

La colonne "score" correspond au score TSS ou ROC moyen calculé pour chaque algorithme.

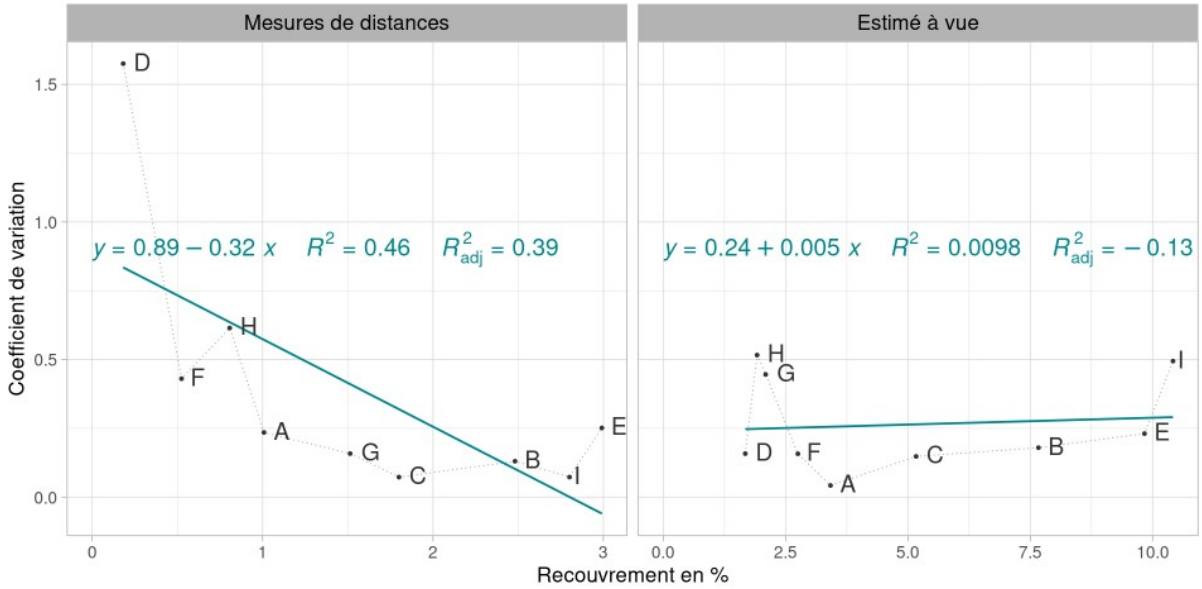


Figure C.2 : Évolution de la variabilité inter-observateur de deux métriques selon le % de recouvrement. Les lettres correspondent aux identifiants des 9 quadrats prospectés. Chaque quadrat a été prospecté par 3 observateurs.
A gauche : coefficient de variation inter-observateur du recouvrement estimé à vue, en fonction du recouvrement estimé moyen pour chaque quadrat.
A droite : Coefficient de variation inter-observateur du recouvrement mesuré à partir des distances, en fonction du recouvrement mesuré moyen pour chaque quadrat.
Attention, les échelles des deux figures ne sont pas identiques.

Evaluation de la capacité des modèles à prédire l'occurrence observée sur le terrain :

Rappel formule : PA ~ PrédictionPrésence + (1|Site)

PA : présence/absence observée sur le terrain
PrédictionPrésence : probabilité de présence donnée par les modélisations

Table C.3 : Evaluation de la capacité des modèles réalisés à l'échelle locale de prédire la présence du thym observée lors de l'étude de terrain, selon différentes métriques.

algorithme	AIC	BIC	logLik	deviance	df.resid	R2_fixe	R2_total	TSS	seuil
CTA	151,04	159,14	-72,52	145,04	107	0,009	0,138	0,420	0,430
GBM	150,28	158,38	-72,14	144,28	107	0,025	0,172	0,389	0,815
GLM	149,65	157,76	-71,83	143,65	107	0,036	0,181	0,339	0,609
MARS	149,73	157,83	-71,87	143,73	107	0,036	0,195	0,365	0,514
MAXENT	145,54	153,64	-69,77	139,54	107	0,113	0,321	0,390	0,407
RF	144,57	152,67	-69,28	138,57	107	0,120	0,378	0,387	0,529
ensemble	147,40	155,50	-70,70	141,40	107	0,078	0,278	0,455	0,675
GAM	148,47	156,57	-71,24	142,47	107	0,058	0,236	0,363	0,286

R2_fixe correspond au R^2 qui ne considère que les effets fixes dans le modèle.
R2_total correspond au R^2 qui considère les effets fixes et aléatoires*.

	algorithme	Coeff	Erreur	Z-valeur	P-valeur (<0.05)
(Intercept)	**CTA**	-0,27	0,98	-0,27	0,786
Prediction	**CTA**	1,13	1,44	0,78	0,434
(Intercept)	**GBM**	-1,75	2,00	-0,88	0,381
Prediction	**GBM**	3,76	3,36	1,12	0,264
(Intercept)	**GLM**	-1,83	1,75	-1,05	0,295
Prediction	**GLM**	4,05	3,05	1,33	0,185
(Intercept)	**MARS**	-1,36	1,44	-0,94	0,347
Prediction	**MARS**	2,67	2,08	1,29	0,198
(Intercept)	**MAXENT**	-1,65	1,06	-1,55	0,120
Prediction	**MAXENT**	3,15	1,49	2,11	**0,035**
(Intercept)	**RF**	-2,55	1,33	-1,92	**0,055**
Prediction	**RF**	4,64	1,96	2,37	**0,018**
(Intercept)	**ensemble**	-2,35	1,56	-1,51	0,131
Prediction	**ensemble**	4,41	2,39	1,85	0,065
(Intercept)	**GAM**	-2,12	1,62	-1,31	0,190
Prediction	**GAM**	3,76	2,31	1,625	0,104

Table C.4 : Coefficients issus des GLMM mis en place pour évaluer la capacité des modèles locaux à prédire la présence du thym observée lors de l'étude de terrain.

Le TSS est calculé grâce à la fonction performance du package *ROCR* qui permet de déterminer le taux de vrai-positifs (tpr) et de faux-positifs (fpr). Le TSS est alors déduit par la formule **TSS=tpr-fpr**. Le seuil indiqué dans la Table C.3 correspond au seuil associé au plus haut TSS.

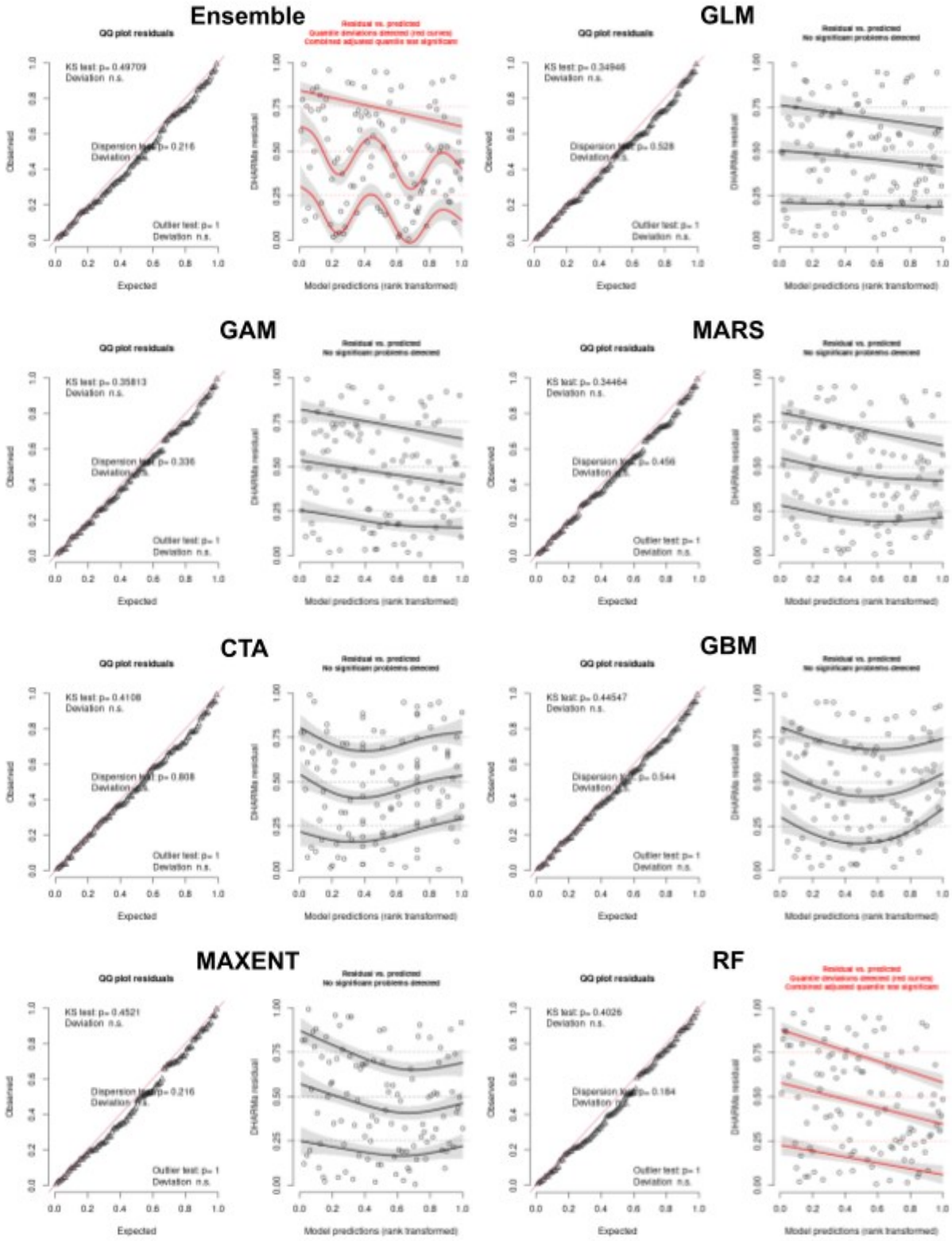


Figure C.3 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM utilisés pour évaluer la capacité des modèles à prédire l'occurrence du thym (package *DHARMa*, fonction *simulateResiduals* sous R). Malgré quelques problèmes mineurs détectés, les hypothèses sont globalement toutes validées.

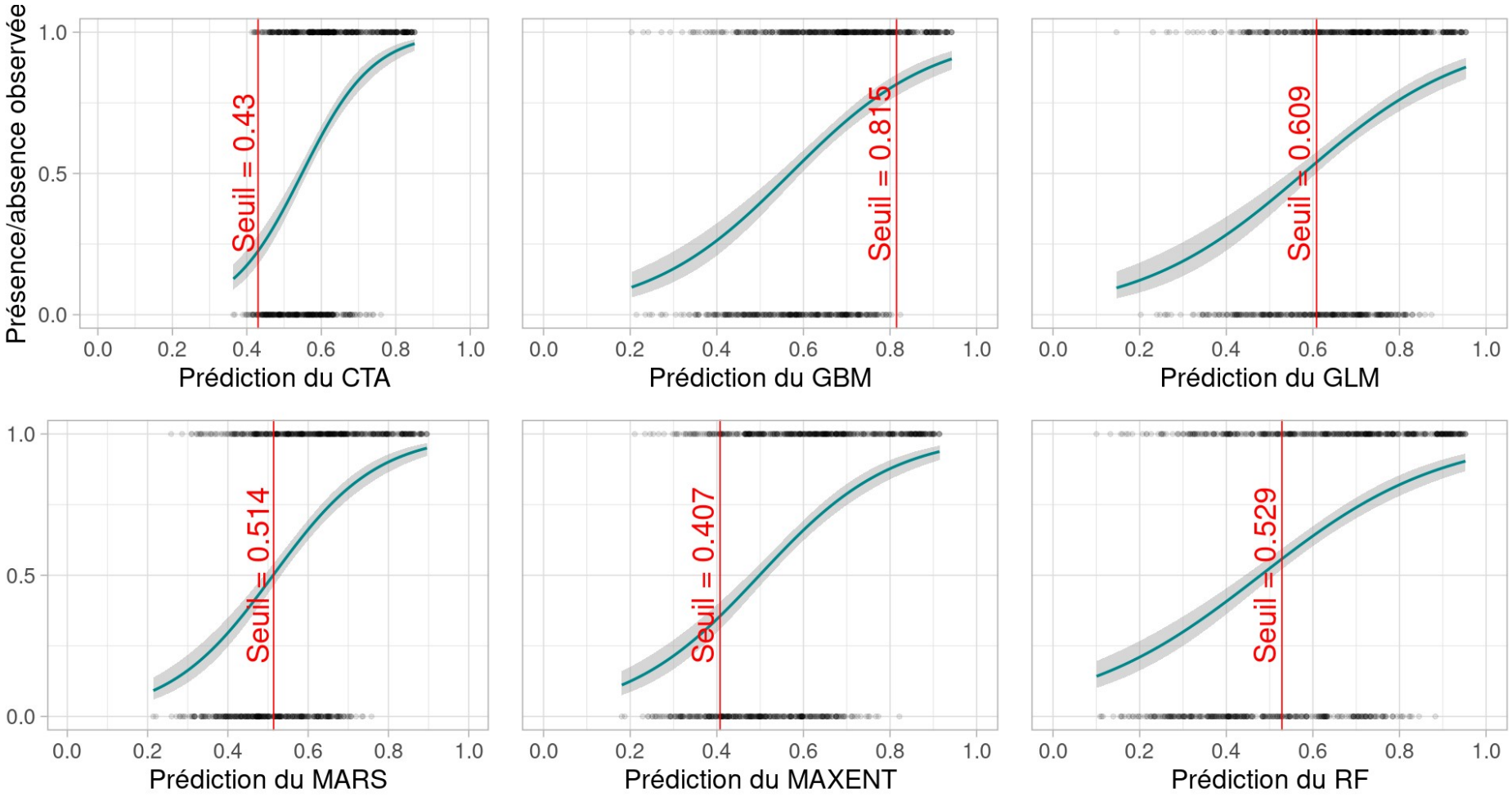


Figure C.4 : Distribution des points de présence/absence observés sur le terrain en fonction des prédictions issues de chaque modèle individuel réalisé dans le cadre de la modélisation du thym à l'échelle locale. La courbe bleue correspond à l'ajustement d'un GLMM à distribution binomiale sur ces données. Le seuil correspondant au meilleur TSS est indiqué en rouge.

Evaluation de la capacité des modèles à prédire l'abondance mesurée sur le terrain :

Rappel formule : AbondanceTerrain ~ PrédictionAbondance : PrésenceEnvironnante + (1|Site)

AbondanceTerrain : recouvrement mesuré sur le terrain
PrédictionPrésence : probabilité de présence donnée par les modélisations
PrédictionPrésenceEnvironnante : probabilité de présence donnée par les modélisations pour la zone environnant le quadrat.

Table C.5 : Evaluation de la capacité des modèles réalisés à l'échelle locale de prédire l'abondance du thym mesurée lors de l'étude de terrain, selon différentes métriques.

algorithme	AIC	BIC	logLik	deviance	df.resid	R2_fixe	R2_total
CTA	-421,46	-410,65	214,73	-429,46	106	0,051	0,140
GBM	-420,74	-409,94	214,37	-428,74	106	0,045	0,114
GLM	-416,98	-406,18	212,49	-424,98	106	0,002	0,081
MARS	-420,38	-409,58	214,19	-428,38	106	0,044	0,122
MAXENT	-423,63	-412,83	215,81	-431,63	106	0,080	0,157
RF	-423,40	-412,60	215,70	-431,40	106	0,079	0,201
ensemble	-422,59	-411,79	215,29	-430,59	106	0,069	0,166
GAM	-419,85	-409,04	213,92	-427,85	106	0,035	0,108

R2_fixe correspond au R^2 qui ne considère que les effets fixes dans le modèle.
R2_total correspond au R^2 qui considère les effets fixes et aléatoires*.

Table C.6 : Coefficients issus des GLMM mis en place pour évaluer la capacité des modèles locaux à prédire l'abondance du thym mesurée lors de l'étude de terrain.

	algorithme	Coeff	Erreur	Z-valeur	P-valeur (<0.05)
(Intercept)	**CTA**	-3,37	0,34	-9,78	**1,36E-22**
Pred_presence : Pred_pres_env	**CTA**	1,30	0,63	2,06	**0,039**
(Intercept)	**GBM**	-3,60	0,48	-7,54	**4,64E-14**
Pred_presence : Pred_pres_env	**GBM**	2,31	1,21	1,91	0,057
(Intercept)	**GLM**	-2,88	0,38	-7,65	**1,99E-14**
Pred_presence : Pred_pres_env	**GLM**	0,39	1,01	0,39	0,697
(Intercept)	**MARS**	-3,34	0,37	-9,13	**6,80E-20**
Pred_presence : Pred_pres_env	**MARS**	1,15	0,64	1,81	0,071
(Intercept)	**MAXENT**	-3,32	0,28	-11,83	**2,88E-32**
Pred_presence : Pred_pres_env	**MAXENT**	1,01	0,41	2,47	**0,013**
(Intercept)	**RF**	-3,43	0,33	-10,39	**2,82E-25**
Pred_presence : Pred_pres_env	**RF**	1,33	0,55	2,43	**0,015**
(Intercept)	**ensemble**	-3,50	0,37	-9,36	**8,18E-21**
Pred_presence : Pred_pres_env	**ensemble**	1,64	0,72	2,27	**0,023**
(Intercept)	**GAM**	-3,30	0,36	-9,06	**1,30E-19**
Pred_presence : Pred_pres_env	**GAM**	1,07	0,63	1,69	0,09

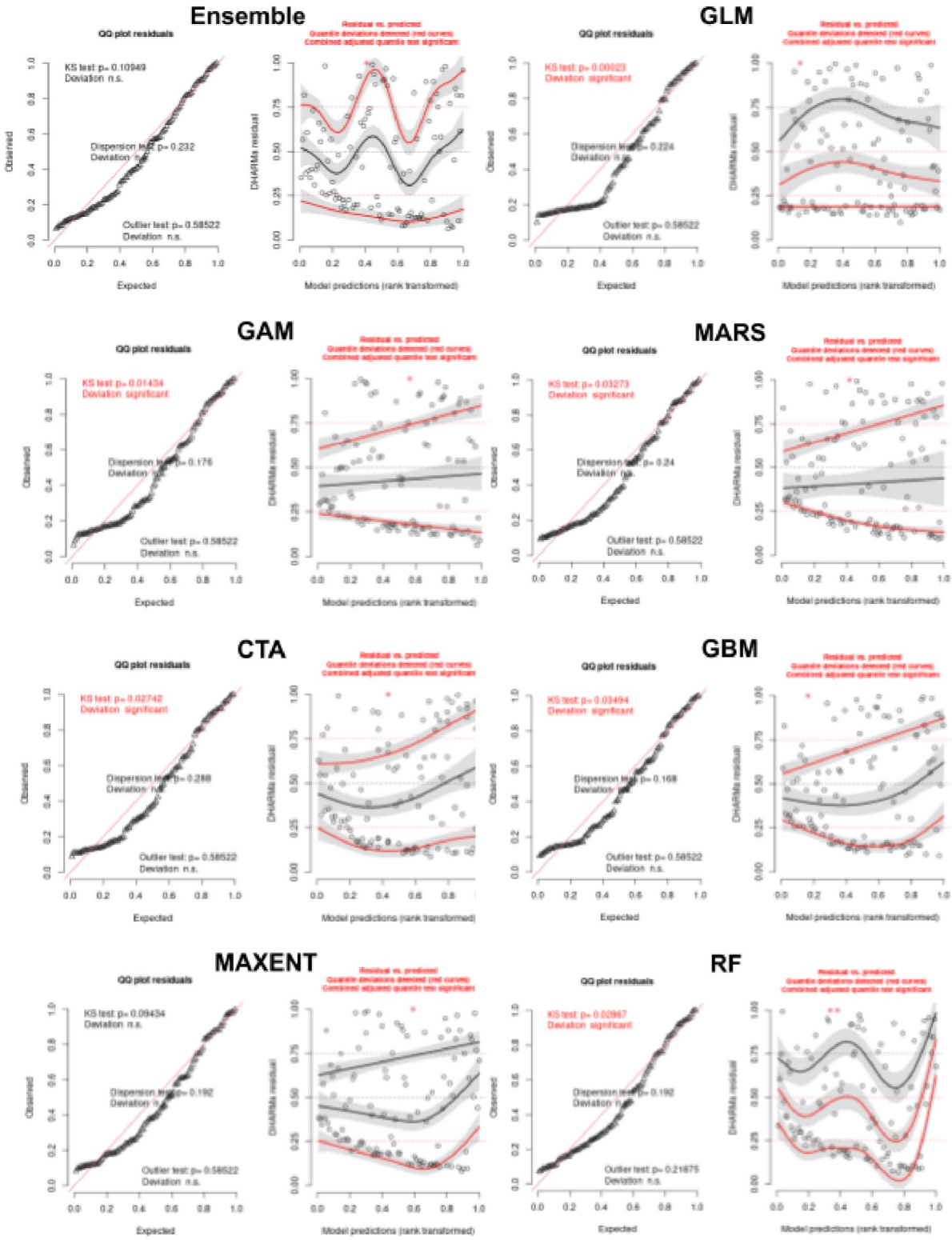


Figure C.5 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM utilisés pour évaluer la capacité des modèles à prédire l'abondance du thym (package *DHARMa*, fonction *simulateResiduals* sous R). Malgré quelques problèmes mineurs détectés, les hypothèses sont globalement toutes validées.

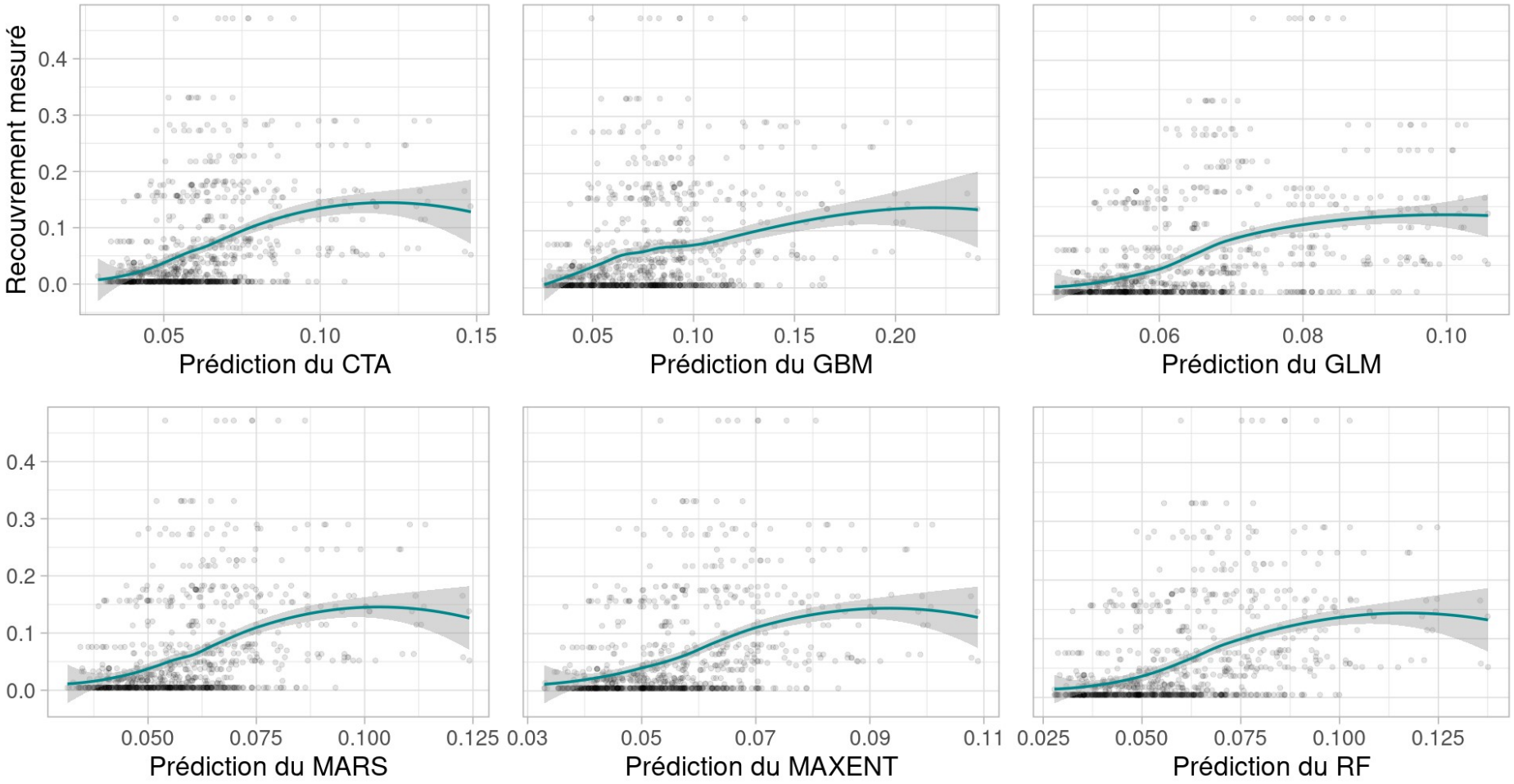


Figure C.6 : Distribution du recouvrement mesuré sur le terrain en fonction des prédictions pour chaque algorithme utilisé dans le modèle d'ensemble à très haute résolution. La courbe bleue correspond à l'ajustement d'un GLMM à distribution bêta sur ces données.

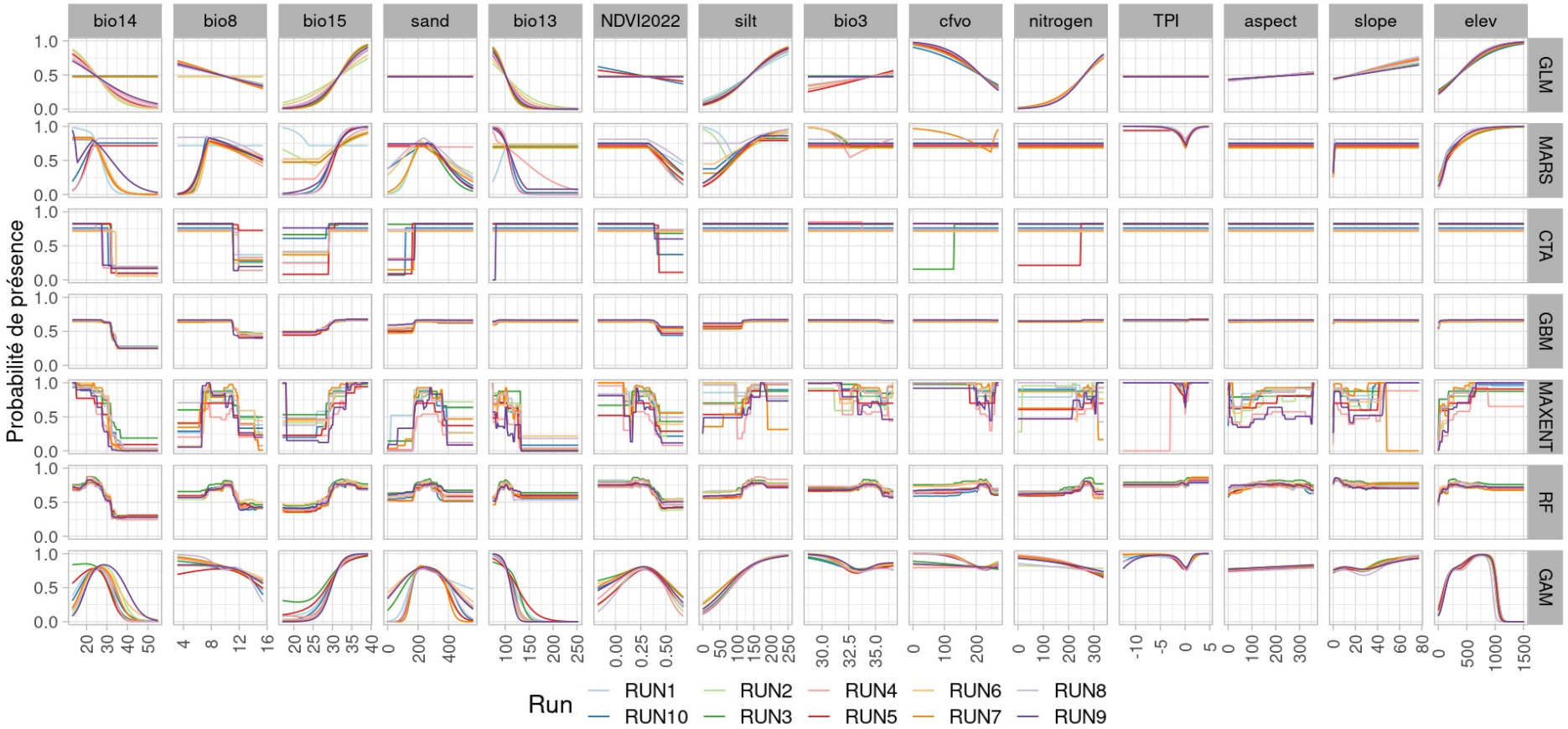


Figure C.7.a : Courbes de réponse pour chaque algorithme et “run” (répétition) effectués dans le cadre de la modélisation du thym à l’échelle locale.

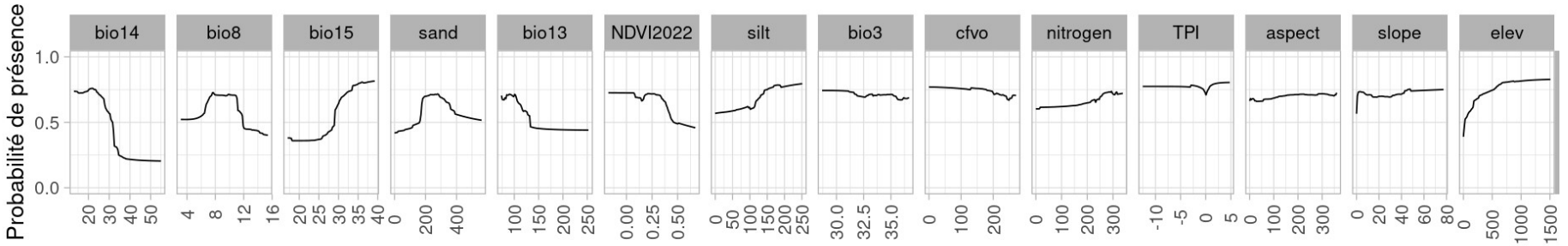


Figure C.7.b : Courbes de réponse pour le modèle d’ensemble effectué dans le cadre de la modélisation du thym à l’échelle locale.

Annexe D : Choix de lissage pour le modèle GAM à l'échelle locale

Ci-dessous sont 3 exemples de coefficients de lissage pour les GAM. J'ai retenu k=5 (au centre) car il permettait de capturer les variations globales, sans générer trop de bruit qui ferait perdre en réalisme écologique. Un bon exemple d'un surajustement est la courbe de réponse de bio3 pour k=10 (encadré en noir), qui crée beaucoup trop de vagues.

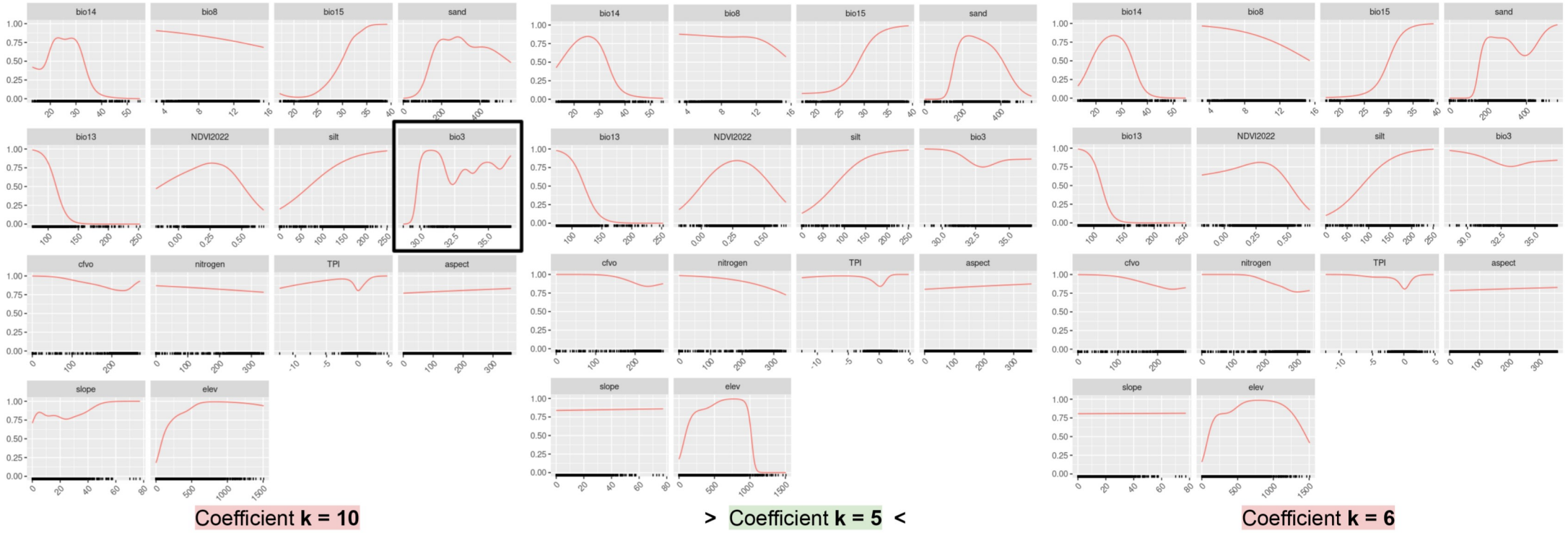


Figure D.1 : Courbes de réponse produites pour le GAM selon trois niveaux de lissage, les coefficient retenu est k=5.

Ici, un seul coefficient a été appliqué à toutes les variables, cependant, selon le package utilisé, il est possible d'ajuster individuellement le nombre de nœuds (ou inflexions) de la courbe, ainsi que leurs positions. Cela n'était pas permis par *Biomod2,* mais constitue une piste de réflexion que je souhaite explorer.

Annexe E : Analyse spatiale de l'effet site dans les prédictions des modélisations à l'échelle locale

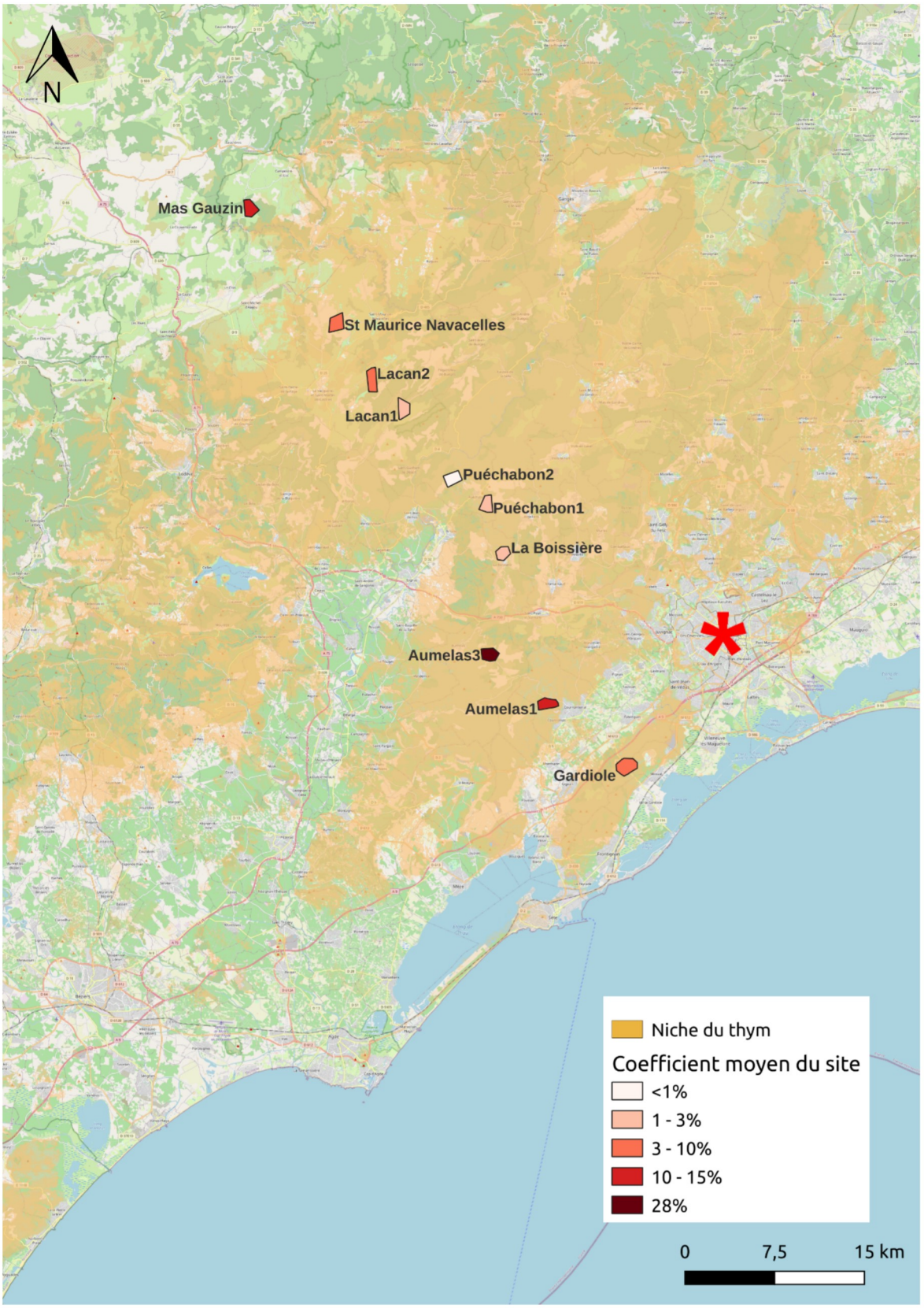


Figure E.1 : Effet associé à chaque site d'échantillonnage (Montpellier est localisé par l'étoile rouge). Les coefficients sont issus de GLM qui confrontent les modèles locaux à l'échantillonnage de terrain :
glm(PrésenceObservée ~ PrésencePrédite + EffetDuSite, family="binomial")
Un coefficient plus faible indique un effet du site moindre, et inversement.
La niche du thym représentée ici correspond à la présence moyenne prédite par les modèles locaux (une superposition de cartes de présence-absence générées en utilisant le seuil calculé en Tab C.2 a été faite).

La Fig E. 1 nous permet d'observer un effet moins important du site en cœur de niche tandis qu'il augmente vers les bords. Cela mène à supposer à une spatialisation des erreurs de prédiction avec potentiellement de moins bonnes prédictions en limite d'aire. Nous pourrions en effet émettre l'hypothèse que lorsque nous nous approchons des abords de la niche, cela augmente la quantité de facteurs entrant en compte pour définir la présence du thym. Ainsi cela rend plus difficile la matérialisation et l'explication de la présence du thym à partir de variables écologiques/biologiques/environnementales… Ceci est une piste qui mérite approfondissement.

Annexe F : Analyses supplémentaires pour les corrélations des métriques de terrain

1. Etude de la relation entre la hauteur et la longueur des touffes de thym

Formule : lmer(Hauteur ~ Longueur + (1|Site), family=gaussian())

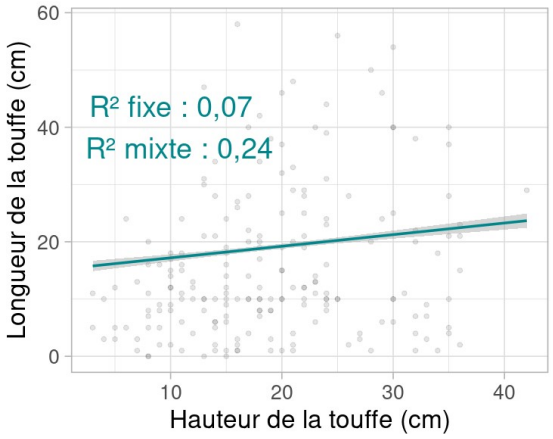


Figure F.1 : Représentation de la hauteur de la touffe en fonction de sa longueur. La droite bleue correspond à l'ajustement d'un GLMM gaussien sur ces données.

Table F.1 : Coefficients issus du GLMM mis en place pour évaluer la relation entre la hauteur et la longueur des touffes de thym.

	Coeff	Erreur	T-valeur	P-valeur (<0.05)
(Intercept)	16,23	1,42	11,41	**0,00**
Longueur	0,18	0,04	4,12	**0,00**

R² fixe correspond au R² associé aux effets fixes (soit la longueur).
R² mixte correspond au R² associé aux effets fixes et mixtes (site).

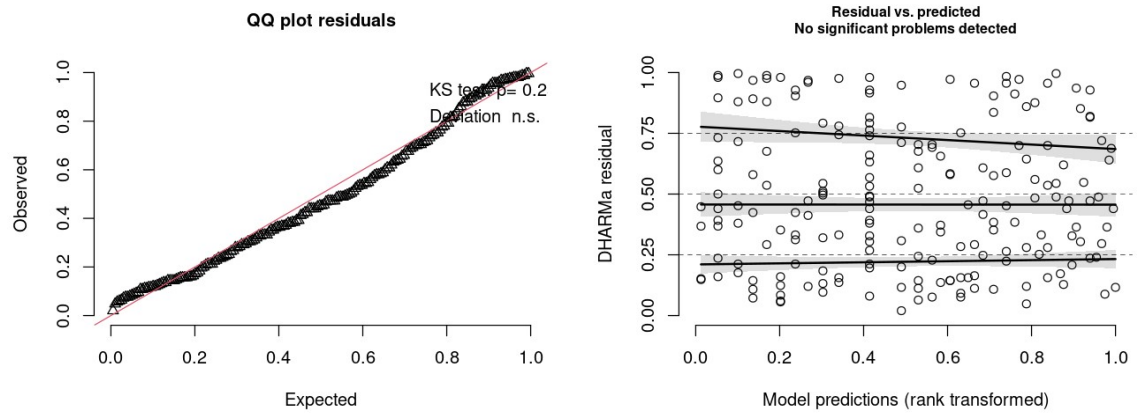


Figure F.4 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM mis en place pour évaluer la relation entre la hauteur et la longueur des touffes de thym (package *DHARMa*, fonction *simulateResiduals* sous R). Toutes les hypothèses sont validées

Conclusions :

Le GLMM met en évidence une relation significative entre la hauteur et la longueur d'une touffe de thym : la longueur augmente légèrement avec la hauteur. Toutefois, le R² associé aux effets fixes indique que la longueur d'une touffe de thym ne permet d'expliquer que 7% de sa hauteur. En outre, il y a un fort effet site. Ainsi, la hauteur d'une touffe de thym dépend de beaucoup d'autres variables.

2. Etude de la relation entre la distance moyenne à la première touffe, et le recouvrement dans un quadrat

La distance moyenne à la première touffe correspond à la distance mesurée en partant du centre de quadrat.

Formule : lmer(Recouvrement ~ Distance à la première touffe + (1|Site), family=gaussian())

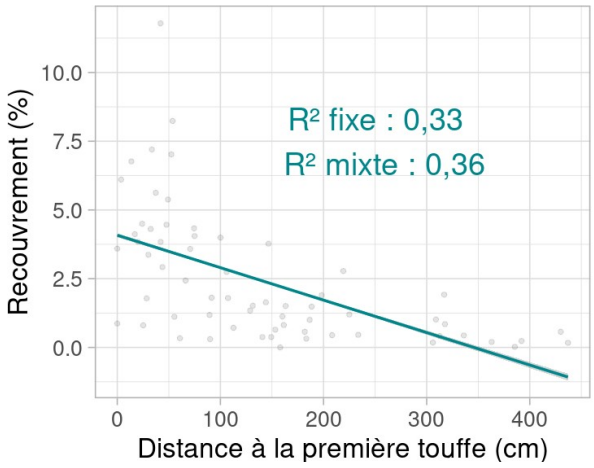


Figure F.3 : Représentation du recouvrement en fonction de la distance moyenne à la première touffe de thym dans un quadrat. La droite bleue correspond à l'ajustement d'un GLMM gaussien sur ces données.

Table F.2 : Coefficients issus du GLMM mis en place pour évaluer la relation entre la distance moyenne à la première touffe et le recouvrement du thym dans un quadrat.

	Coeff	Erreur	T-valeur	P-valeur (<0.05)
(Intercept)	4,07	0,41	9,91	**0,00**
Distance	-0,01	0,00	-5,46	**0,00**

R² fixe correspond au R² associé aux effets fixes (soit la distance à la première touffe).
R² mixte correspond au R² associé aux effets fixes et mixtes (site).

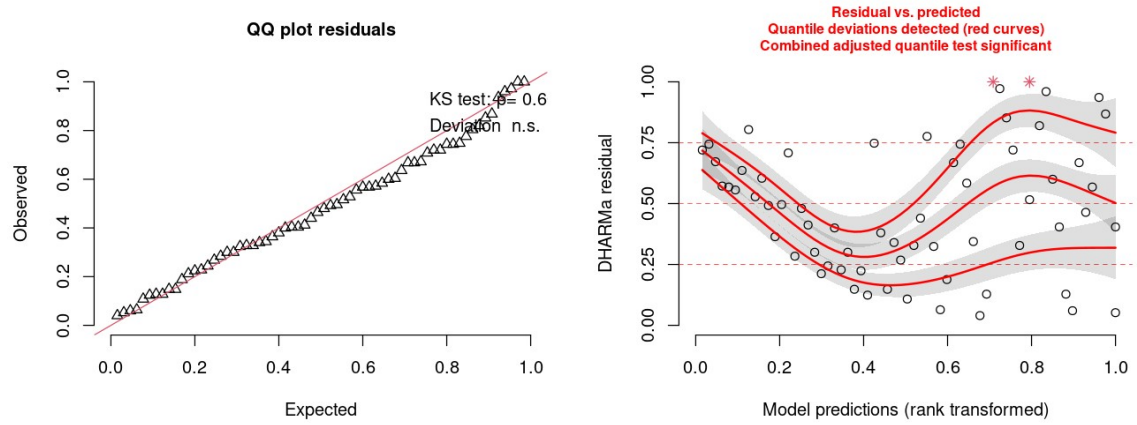


Figure F.4 : Vérification des hypothèses de normalité et d'homoscédasticité pour les GLMM mis en place pour évaluer la relation entre la distance à la première touffe, et le recouvrement du thym dans un quadrat (package *DHARMa*, fonction *simulateResiduals* sous R). L'hypothèse de normalité est validée, cependant il y a des problèmes au niveau de l'homoscédasticité, les résultats doivent donc être interprétés avec caution.

Conclusions :

Le GLMM met en évidence une relation significative entre la distance moyenne à la première touffe et le recouvrement du thym dans un quadrat : le recouvrement diminue avec la distance à la première touffe. Toutefois, le R² associé aux effets fixes indique que cette distance ne permet d'expliquer que 33% du recouvrement. En outre, nous notons un faible effet site. Ainsi, la distance à la première touffe ne suffit pas à expliquer le recouvrement d'un quadrat.

Annexe G : Fiche pour les relevés de terrain

Identifiant du relevé : **Date :**

Point GPS : **Observateur :**

Description (localisation, type d'habitat, accessibilité...) :

Temps à la 1ère détection (2 minutes max par section, temps en secondes)							
	Section NE		Section SE		Section SO		Section NO
	s		s		s		s

Estimation à vue du recouvrement (1% environ égal à 50*50cm)							
	Section NE		Section SE		Section SO		Section NO
	%		%		%		%

Distance au 1er individu et hauteur (distance/hauteur)							
Transect Nord	Transect NE	Transect Est	Transect SE	Transect Sud	Transect SO	Transect Ouest	Transect NO
cm / cm	cm / cm	cm / cm	cm / cm	cm / cm	cm / cm	cm / cm	cm / cm

Mesure du recouvrement (à partir du centre, jusqu'à 5m60)							
Transect Nord		Transect Est		Transect Sud		Transect Ouest	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	
- cm		- cm		- cm		- cm	

Commentaires :

Bibliographie

Agribio 04, et Agribio 05. s. d. « Enquête sur la cueillette en milieu sauvage de PPAM dans le 04/05 ». Consulté le 1 septembre 2023. https://www.bio-provence.org/IMG/pdf/synthese_enquete_cueillette_2022_ab0405.pdf.

Allouche, Omri, Asaf Tsoar, et Ronen Kadmon. 2006. « Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS) ». *Journal of Applied Ecology* 43 (6): 1223‑32. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Al-Ramamneh, Ezz AL-Dein Muhammed. 2009. « Plant growth strategies of Thymus vulgaris L. in response to population density ». *Industrial Crops and Products* 30 (3): 389‑94. https://doi.org/10.1016/j.indcrop.2009.07.008.

Álvarez-Martínez, Jose Manuel, Susana Suárez-Seoane, Jetse J. Stoorvogel, et Estanislao de Luis Calabuig. 2014. « Influence of Land Use and Climate on Recent Forest Expansion: A Case Study in the Eurosiberian–Mediterranean Limit of North-West Spain ». *Journal of Ecology* 102 (4): 905‑19. https://doi.org/10.1111/1365-2745.12257.

Amaral, Silvana, Cristina Bestetti, Costa Camilo, et Camilo Rennó. 2023. « Normalized Difference Vegetation Index (NDVI) improving species distribution models: an example with the neotropical genus coccocypselum (Rubiaceae) ».

Anderson, B.j, H.r Akçakaya, M.b Araújo, D.a Fordham, E Martinez-Meyer, W Thuiller, et B.w Brook. 2009. « Dynamics of range margins for metapopulations under climate change ». *Proceedings of the Royal Society B: Biological Sciences* 276 (1661): 1415‑20. https://doi.org/10.1098/rspb.2008.1681.

Araújo, Miguel B., Robert J. Whittaker, Richard J. Ladle, et Markus Erhard. 2005. « Reducing Uncertainty in Projections of Extinction Risk from Climate Change ». *Global Ecology and Biogeography* 14 (6): 529‑38. https://doi.org/10.1111/j.1466-822X.2005.00182.x.

Association Française des professionnels de la Cueillette de plantes sauvages. 2019. « Guide de bonnes pratiques. Une cueillette durable de plantes sauvages. Liste de plantes prioritaires pour l'établissement de livrets techniques. »

Bal, Guillaume, Aurélien Besnard, Léo Bacon, Emmanuel Menoni, Clément Calenge, et Alexandre Millon. 2021. *Modélisation de la dynamique du grand tétras des Pyrénées françaises pour sa gestion adaptative Conservation biology of threatened plants from the Balearic Islands View project Long-Term Evaluation of the Success of a Reintroduction Program of the European Pond Turtle View project Modélisation de la dynamique du grand tétras des Pyrénées françaises pour sa gestion adaptative*.

Barbet-Massin, Morgane, Frédéric Jiguet, Cécile Albert, et Wilfried Thuiller. 2012. « Selecting Pseudo-Absences for Species Distribution Models: How, Where and How Many? » *Methods in Ecology and Evolution* 3 (avril): 327‑38. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Barbet-Massin, Morgane, Quentin Rome, Claire Villemant, et Franck Courchamp. 2018. « Can Species Distribution Models Really Predict the Expansion of Invasive Species? » *PLOS ONE* 13 (3): e0193085. https://doi.org/10.1371/journal.pone.0193085.

Barkaoui, Karim, Maud Bernard-Verdier, et Marie-Laure Navas. 2013. « Questioning the Reliability of the Point Intercept Method for Assessing Community Functional Structure in Low-Productive and Highly Diverse Mediterranean Grasslands ». *Folia Geobotanica* 48 (octobre): 393‑414. https://doi.org/10.1007/s12224-013-9172-2.

Beaury, Evelyn M., Catherine S. Jarnevich, Ian Pearse, Annette E. Evans, Nathan Teich, Peder Engelstad, Jillian LaRoe, et Bethany A. Bradley. 2023. « Modeling Habitat Suitability across Different Levels of Invasive Plant Abundance ». *Biological Invasions*, juin. https://doi.org/10.1007/s10530-023-03118-z.

Beck, Jan, Marianne Böller, Andreas Erhardt, et Wolfgang Schwanghart. 2014. « Spatial bias in the GBIF database and its effect on modeling species' geographic distributions ». *Ecological Informatics* 19 (janvier): 10‑15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Bennie, Jonathan, Mark O. Hill, Robert Baxter, et Brian Huntley. 2006. « Influence of Slope and Aspect on Long-Term Vegetation Change in British Chalk Grasslands ». *Journal of Ecology* 94 (2): 355‑68. https://doi.org/10.1111/j.1365-2745.2006.01104.x.

« Bioclimatic variables — WorldClim 1 documentation ». s. d. Consulté le 1 septembre 2023. https://www.worldclim.org/data/bioclim.html.

Booth, Trevor H. 2018. « Why Understanding the Pioneering and Continuing Contributions of BIOCLIM to Species Distribution Modelling Is Important ». *Austral Ecology* 43 (8): 852‑60. https://doi.org/10.1111/aec.12628.

Bornand, Christophe N., Marc Kéry, Lena Bueche, et Markus Fischer. 2014. « Hide-and-Seek in Vegetation: Time-to-Detection Is an Efficient Design for Estimating Detectability and Occurrence ». *Methods in Ecology and Evolution* 5 (5): 433‑42. https://doi.org/10.1111/2041-210X.12171.

Bradley, Bethany A. 2016. « Predicting Abundance with Presence-Only Models ». *Landscape Ecology* 31 (1): 19‑30. https://doi.org/10.1007/s10980-015-0303-4.

Braun-blanquet, J. 1932. « Plant Sociology. The Study of Plant Communities. First Ed. » *Plant Sociology. The Study of Plant Communities. First Ed.* https://www.cabdirect.org/cabdirect/abstract/19331600801.

Breiner, Frank T., Antoine Guisan, Ariel Bergamini, et Michael P. Nobis. 2015. « Overcoming Limitations of Modelling Rare Species by Using Ensembles of Small Models ». *Methods in Ecology and Evolution* 6 (10): 1210‑18. https://doi.org/10.1111/2041-210X.12403.

Brown, James H. 1984. « On the Relationship between Abundance and Distribution of Species ». *The American Naturalist* 124 (2): 255‑79.

Büchi, Lucie, Pauline Mouly, Camille Amossé, et Cindy Bally. 2016. « Méthode non destructive d'estimation de la biomasse de couverts végétaux ». *Recherche Agronomique Suisse* 7 (3): 136‑43.

Buckland, Steeves, Nik C. Cole, Jesús Aguirre-Gutiérrez, Laura E. Gallagher, Sion M. Henshaw, Aurélien Besnard, Rachel M. Tucker, Vishnu Bachraz, Kevin Ruhomaun, et Stephen Harris. 2014. « Ecological Effects of the Invasive Giant Madagascar Day Gecko on Endemic Mauritian Geckos: Applications of Binomial-Mixture and Species Distribution Models ». *PLOS ONE* 9 (4): e88798. https://doi.org/10.1371/journal.pone.0088798.

Buitrago, Leonardo. 2020. « GBIF Issues & Flags ». 27 octobre 2020. https://data-blog.gbif.org/post/issues-and-flags/.

« Cadastre Etalab ». s. d. Cadastre Etalab. Consulté le 1 septembre 2023. https://cadastre.data.gouv.fr/.

Coudun, Christophe, Jean-Claude Gégout, Christian Piedallu, et Jean-Claude Rameau. 2006. « Soil Nutritional Factors Improve Models of Plant Species Distribution: An Illustration with Acer Campestre (L.) in France ». *Journal of Biogeography* 33 (10): 1750‑63. https://doi.org/10.1111/j.1365-2699.2005.01443.x.

« Cueillettes et prélèvements ». 2014. Parc national des Ecrins. 12 août 2014. https://www.ecrins-parcnational.fr/cueillettes-et-prelevements.

Cunningham, A. B. 2001. *Applied Ethnobotany: People, Wild Plant Use, and Conservation*. « People and Plants » Conservation Manuals. London: Earthscan.

Dahdouh-Guebas, Farid, et Nico Koedam. 2006. « Empirical estimate of the reliability of the use of the Point-Centred Quarter Method (PCQM): Solutions to ambiguous field situations and description of the PCQM+ protocol ». *Forest Ecology and Management* 228 (juin): 1‑18. https://doi.org/10.1016/j.foreco.2005.10.076.

« Diversité fonctionnelle des plantes ». 2023. De Boeck Supérieur. 12 mai 2023. https://www.deboecksuperieur.com/ouvrage/9782804175627-diversite-fonctionnelle-des-plantes.

Dormann, C. F., J. Elith, S. Bacher, G. C. G. Carré, J. R. García Márquez, B. Gruber, B. Lafourcade, et al. 2013. « Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance : Open Access ». *Ecography* 36 (1): 27‑46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Dubuis, Anne, Sara Giovanettina, Loïc Pellissier, Julien Pottier, Pascal Vittoz, et Antoine Guisan. 2013. « Improving the Prediction of Plant Species Distribution and Community Composition by Adding Edaphic to Topo-Climatic Variables ». *Journal of Vegetation Science* 24 (4): 593‑606. https://doi.org/10.1111/jvs.12002.

Durner, George M., David C. Douglas, Ryan M. Nielson, Steven C. Amstrup, Trent L. McDonald, Ian Stirling, Mette Mauritzen, et al. 2009. « Predicting 21st-Century Polar Bear Habitat Distribution from Global Climate Models ». *Ecological Monographs* 79 (1): 25‑58. https://doi.org/10.1890/07-2089.1.

Eisfelder, Christina, Sarah Asam, Andreas Hirner, Philipp Reiners, Stefanie Holzwarth, Martin Bachmann, Ursula Gessner, et al. 2023. « Seasonal Vegetation Trends for Europe over 30 Years from a Novel Normalised Difference Vegetation Index (NDVI) Time-Series—The TIMELINE NDVI Product ». *Remote Sensing* 15 (juillet): 3616. https://doi.org/10.3390/rs15143616.

Elith*, Jane, Catherine H. Graham*, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, et al. 2006. « Novel Methods Improve Prediction of Species' Distributions from Occurrence Data ». *Ecography* 29 (2): 129‑51. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Engler, Robin, Lars T. Waser, Niklaus E. Zimmermann, Marcus Schaub, Savvas Berdos, Christian Ginzler, et Achilleas Psomas. 2013. « Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution ». *Forest Ecology and Management* 310 (décembre): 64-73. https://doi.org/10.1016/j.foreco.2013.07.059.

Farrell, S. L., B. A. Collier, K. L. Skow, A. M. Long, A. J. Campomizzi, M. L. Morrison, K. B. Hays, et R. N. Wilkins. 2013. « Using LiDAR-Derived Vegetation Metrics for High-Resolution, Species Distribution Models for Conservation Planning ». *Ecosphere* 4 (3): art42. https://doi.org/10.1890/ES12-000352.1.

Floyd, Donald A., et Jay E. Anderson. 1987. « A Comparison of Three Methods for Estimating Plant Cover ». *Journal of Ecology* 75 (1): 221-28. https://doi.org/10.2307/2260547.

Fontaine, Ninon. 2021. « Les génépis : dynamiques d'espèces de haute montagne dans leur socio-écosystème ». These en préparation, Université de Montpellier (2022-….). https://www.theses.fr/s297914.

Fourcade, Yoan, Aurélien Besnard, et Jean Secondi. 2018. « Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics ». *Global Ecology and Biogeography* 27 (février): 245-56. https://doi.org/10.1111/geb.12684.

Franklin. 2013. « Species distribution models in conservation biogeography: developments and challenges ». *Diversity and Distributions*. https://onlinelibrary.wiley.com/doi/10.1111/ddi.12125.

Gargominy, O., et C. Régnier. 2023. « Base de connaissance « Statuts » des espèces en France. Version pour TAXREF v16.0. » PatriNat (OFB-MNHN-CNRS-IRD). https://inpn.mnhn.fr/telechargement/referentielEspece/bdc-statuts-especes.

Garrard, Georgia E., Michael A. McCarthy, Nicholas S. G. Williams, Sarah A. Bekessy, et Brendan A. Wintle. 2013. « A General Model of Detectability Using Species Traits ». *Methods in Ecology and Evolution* 4 (1): 45-52. https://doi.org/10.1111/j.2041-210x.2012.00257.x.

Garreta, Raphaële, et Béatrice Morisson. 2011. « La cueillette des plantes sauvages en Pyrénées et Midi-Pyrénées. Phase 1, état des lieux (2010-2011). » Conservatoire Botanique Pyrénéen.

Garsd, Armando. 1984. « Spurious correlation in ecological modelling ». *Ecological Modelling* 23 (3): 191-201. https://doi.org/10.1016/0304-3800(84)90100-5.

« GBIF ». s. d. Consulté le 1 septembre 2023. https://www.gbif.org/.

Graham, Catherine H., Simon Ferrier, Falk Huettman, Craig Moritz, et A. Townsend Peterson. 2004. « New developments in museum-based informatics and applications in biodiversity analysis ». *Trends in Ecology & Evolution* 19 (9): 497-503. https://doi.org/10.1016/j.tree.2004.07.006.

Grenouillet, Gaël, Laetitia Buisson, Nicolas Casajus, et Sovan Lek. 2011. « Ensemble modelling of species distribution: The effects of geographical and environmental ranges ». *Ecography* 34 (février): 9-17. https://doi.org/10.1111/j.1600-0587.2010.06152.x.

Guisan, Antoine, et Wilfried Thuiller. 2005. « Predicting Species Distribution: Offering More than Simple Habitat Models ». *Ecology Letters* 8 (9): 993‑1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Halstead, Brian J., Jonathan P. Rose, et Patrick M. Kleeman. 2021. « Time‐to‐detection occupancy methods: performance and utility for improving efficiency of surveys ». *Ecological Applications* 31 (3): e2267. https://doi.org/10.1002/eap.2267.

Hao, Tianxiao, Jane Elith, Gurutzeta Guillera-Arroita, et José J. Lahoz-Monfort. 2019. « A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD - Hao - 2019 - Diversity and Distributions - Wiley Online Library ». *Diversity and Distributions* 25 (5): 839‑52. https://doi.org/10.1111/ddi.12892.

He, Kate S., Bethany A. Bradley, Anna F. Cord, Duccio Rocchini, Mao-Ning Tuanmu, Sebastian Schmidtlein, Woody Turner, Martin Wegmann, et Nathalie Pettorelli. 2015. « Will Remote Sensing Shape the next Generation of Species Distribution Models? » *Remote Sensing in Ecology and Conservation* 1 (1): 4‑18. https://doi.org/10.1002/rse2.7.

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, et Andy Jarvis. 2005. « Very High Resolution Interpolated Climate Surfaces for Global Land Areas ». *International Journal of Climatology* 25 (15): 1965‑78. https://doi.org/10.1002/joc.1276.

Jähnig, Sonja C., Mathias Kuemmerlen, Jens Kiesel, Sami Domisch, Qinghua Cai, Britta Schmalz, et Nicola Fohrer. 2012. « Modelling of Riverine Ecosystems by Integrating Models: Conceptual Approach, a Case Study and Research Agenda ». *Journal of Biogeography* 39 (12): 2253‑63. https://doi.org/10.1111/jbi.12009.

Jenkins, M, A Timoshyna, et M Cornthwaite. 2018. « Wild at Home: Exploring the global harvest, trade and use of wild plant ingredients ». TRAFFIC.

Joly, Alexis, Pierre Bonnet, Hervé Goëau, Julien Barbe, Souheil Selmi, Julien Champ, Samuel Dufour-Kowalski, et al. 2016. « A Look inside the Pl@ntNet Experience ». *Multimedia Systems* 22 (6): 751‑66. https://doi.org/10.1007/s00530-015-0462-9.

Julliand, Claire. 2008. *Chapitre 20 : Itinéraires de cueillette. Dans Aux origines des plantes : Tome 2, Des plantes et des Hommes, de Francis Hallé, p 502-529.* Fayard.

Keith, David A, H. Resit Akçakaya, Wilfried Thuiller, Guy F Midgley, Richard G Pearson, Steven J Phillips, Helen M Regan, Miguel B Araújo, et Tony G Rebelo. 2008. « Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models ». *Biology Letters* 4 (5): 560‑63. https://doi.org/10.1098/rsbl.2008.0049.

Khan, Md Nabiul, Renske Hijbeek, Uta Berger, Nico Koedam, Uwe Grüters, S Islam, Md Hasan, et Farid Dahdouh-Guebas. 2016. « An Evaluation of the Plant Density Estimator the Point-Centred Quarter Method (PCQM) Using Monte Carlo Simulation ». *PloS one* 11 (juin): e0157985. https://doi.org/10.1371/journal.pone.0157985.

Kuha, Jouni. 2004. « AIC and BIC: Comparisons of Assumptions and Performance ». *Sociological Methods & Research* 33 (2): 188‑229. https://doi.org/10.1177/0049124103262065.

Kumarathunge, Dushan, R.O.Thattil, et S. Nissanka. 2011. « Evaluation of the plotless sampling method to estimate aboveground biomass and other stand parameters in tropical rain forests ». *Applied Ecology and Environmental Research* 9 (janvier): 425‑31.

Labbé, Joël. 2018. « Les plantes médicinales et l'herboristerie : à la croisée de savoirs ancestraux et d'enjeux d'avenir ». Rapport d'information 727. MI Développement de l'herboristerie. https://www.senat.fr/rap/r17-727/r17-727.html.

Lamotte, M. 1979. « La niche écologique, des concepts théoriques aux utilisations pratiques ». *Revue d'Écologie*, n° 3: 509‑20.

Lannuzel, Guillaume, Joan Balmot, Nicolas Dubos, Martin Thibault, et Bruno Fogliani. 2021. « High-Resolution Topographic Variables Accurately Predict the Distribution of Rare Plant Species for Conservation Area Selection in a Narrow-Endemism Hotspot in New Caledonia ». *Biodiversity and Conservation* 30 (4): 963‑90. https://doi.org/10.1007/s10531-021-02126-6.

Laucoin, Violaine. 2012. « La cueillette des plantes sauvages sur le territoire d'agrément du CBN Massif central : état des lieux et perspectives ». Conservatoire Botanique Massif Central.

Lawton, John H. 1993. « Range, population abundance and conservation ». *Trends in Ecology & Evolution* 8 (11): 409‑13. https://doi.org/10.1016/0169-5347(93)90043-O.

Lehmann, Anthony, Karin Allenbach, Ramona Maggini, Jean-Philippe Richard, Jean-Michel Jaquet, et Hy Dao. 2010. « Swiss Environmental Domains. A new spatial framework for reporting on the environment. »

Lescure, Jean-Paul, Thierry Thevenin, Raphaële Garreta, et Béatrice Morisson. 2015. « Les plantes faisant l'objet de cueillettes commerciales sur le territoire métropolitain. Une liste commentée. » *Le Monde des Plantes*, n° 517: 19‑39.

Lieutaghi, Pierre. 1991. *La plante compagne: pratique et imaginaire de la flore sauvage en Europe occidentale*. Conservatoire et jardin botaniques de Genève.

Locqueville, Jonathan. 2019. « Les pratiques de gestion des plantes entre culture et cueillette : apports croisés de l'ethnoécologie et de l'écologie fonctionnelle ». These en préparation, Université de Montpellier (2022-....). https://www.theses.fr/s226003.

Mackenzie, Darryl I., et J. Andrew Royle. 2005. « Designing Occupancy Studies: General Advice and Allocating Survey Effort ». *Journal of Applied Ecology* 42 (6): 1105‑14. https://doi.org/10.1111/j.1365-2664.2005.01098.x.

Marboutin, Eric, et et al. 2005. « Gestion adaptative de la population de loups en France : du monitoring à l'évaluation des possibilités de prélèvements ». Scientifique. ONCFS. https://www.parcs-naturels-regionaux.fr/mediatheque/ressources/gestion-adaptative-de-la-population-de-loups-en-france-du-monitoring.

Marmion, Mathieu, Miia Parviainen, Miska Luoto, Risto K. Heikkinen, et Wilfried Thuiller. 2009. « Evaluation of Consensus Methods in Predictive Species Distribution Modelling ». *Diversity and Distributions* 15 (1): 59‑69. https://doi.org/10.1111/j.1472-4642.2008.00491.x.

Martin, Lucie. 2014. *Premiers paysans des Alpes. Alimentation végétale et agriculture au Néolithique*. https://doi.org/10.4000/books.pufr.24722.

Miller, Jennifer. 2010. « Species Distribution Modeling ». *Geography Compass* 4 (6): 490-509. https://doi.org/10.1111/j.1749-8198.2010.00351.x.

Mtengwana, Bhongolethu, Timothy Dube, Bester Tawona Mudereri, et Cletah Shoko. 2021. « Modeling the geographic spread and proliferation of invasive alien plants (IAPs) into new ecosystems using multi-source data and multiple predictive models in the Heuningnes catchment, South Africa ». *GIScience & Remote Sensing* 58 (4): 483-500. https://doi.org/10.1080/15481603.2021.1903281.

Mueller-Dombois, Dieter, et Heinz Ellenberg. 1974. *Aims and Methods of Vegetation Ecology*. *John Wiley and Sons*. https://doi.org/10.2307/213332.

Mugumaarhahama, Yannick, Adandé Belarmain Fandohan, et Romain L. Glèlè Kakaï. 2023. « Performance of Inhomogeneous Poisson Point Process Models under Different Scenarios of Uncertainty in Species Presence-Only Data ». *Environmental Systems Research* 12 (1): 27. https://doi.org/10.1186/s40068-023-00312-9.

Muraz, Maëliss, et PNR Monts d'Ardèche. 2018. « La cueillette commerciale de plantes sauvages sur les Monts d'Ardèche. Mémoire ingénieur agronome. »

Ndlovu, Phindile, Onisimo Mutanga, Mbulisi Sibanda, John Odindi, et Ian Rushworth. 2018. « Modelling potential distribution of bramble (rubus cuneifolius) using topographic, bioclimatic and remotely sensed data in the KwaZulu-Natal Drakensberg, South Africa ». *Applied Geography* 99 (octobre): 54-62. https://doi.org/10.1016/j.apgeog.2018.07.025.

Niu, Yang, Martin Stevens, et Hang Sun. 2021. « Commercial Harvesting Has Driven the Evolution of Camouflage in an Alpine Plant ». *Current Biology* 31 (2). https://doi.org/10.1016/j.cub.2020.10.078.

« Open Access Hub ». s. d. Consulté le 1 septembre 2023. https://scihub.copernicus.eu/.

Pailhès, Catherine. 2005. « Les pratiques de cueillette de fleurs sauvages dans les Pyrénées centrales - Mémoire de maîtrise d'ethnologie ». CBNPMP et Université de Toulouse le Mirail.

Papuga, G., P. Gauthier, V. Pons, E. Farris, et J. D. Thompson. 2018. « Ecological Niche Differentiation in Peripheral Populations: A Comparative Analysis of Eleven Mediterranean Plant Species ». *Ecography* 41 (10): 1650-64. https://doi.org/10.1111/ecog.03331.

Pedersen, Eric J., David L. Miller, Gavin L. Simpson, et Noam Ross. 2019. « Hierarchical Generalized Additive Models in Ecology: An Introduction with Mgcv ». *PeerJ* 7 (mai): e6876. https://doi.org/10.7717/peerj.6876.

Pesaresi, Simone, Adriano Mancini, Giacomo Quattrini, et Simona Casavecchia. 2020. « Mapping Mediterranean Forest Plant Associations and Habitats with Functional Principal Component Analysis Using Landsat 8 NDVI Time Series ». *Remote Sensing* 12 (7): 1132. https://doi.org/10.3390/rs12071132.

Pettorelli, Nathalie, Sadie Ryan, Thomas Mueller, Nils Bunnefeld, Bogumila Jędrzejewska, Mauricio Lima, et Kyrre Kausrud. 2011. « The Normalized Difference Vegetation Index (NDVI): Unforeseen Successes in Animal Ecology ». *Climate Research* 46 (1): 15-27. https://doi.org/10.3354/cr00936.

Pfeffer, Jeffrey, et Robert I. Sutton. 1999. « Knowing “What” to Do Is Not Enough: Turning Knowledge into Action ». *California Management Review* 42 (1): 83‑108. https://doi.org/10.1177/000812569904200101.

Pradervand, Jean-Nicolas, Anne Dubuis, Loïc Pellissier, Antoine Guisan, et Christophe Randin. 2014. « Very High Resolution Environmental Predictors in Species Distribution Models: Moving beyond Topography? » *Progress in Physical Geography: Earth and Environment* 38 (1): 79‑96. https://doi.org/10.1177/0309133313512667.

Priyadarshani, Dinusha, Res Altwegg, Alan T. K. Lee, et Wen-Han Hwang. 2022. « What Can Occupancy Models Gain from Time-to-Detection Data? » *Ecology* 103 (12): e3832. https://doi.org/10.1002/ecy.3832.

« Quotas de pêche : comment sont-ils fixés ? » 2019. Ministère de l’Agriculture et de la Souveraineté alimentaire. 16 décembre 2019. https://agriculture.gouv.fr/quotas-de-peche-comment-sont-ils-fixes.

« R: WorldClim climate data ». s. d. Consulté le 1 septembre 2023. https://search.r-project.org/CRAN/refmans/geodata/html/worldclim.html.

Raduła, Małgorzata W., Tomasz H. Szymura, et Magdalena Szymura. 2018. « Topographic wetness index explains soil moisture better than bioindication with Ellenberg’s indicator values ». *Ecological Indicators* 85 (février): 172‑79. https://doi.org/10.1016/j.ecolind.2017.10.011.

« RGE ALTI® | Géoservices ». s. d. Consulté le 1 septembre 2023. https://geoservices.ign.fr/rgealti.

Riihimäki, H., J. Kemppinen, M. Kopecký, et M. Luoto. 2021. « Topographic Wetness Index as a Proxy for Soil Moisture: The Importance of Flow-Routing Algorithm and Grid Resolution ». *Water Resources Research* 57 (10): e2021WR029871. https://doi.org/10.1029/2021WR029871.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2017. « Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure ». *Ecography* 40 (8): 913‑29. https://doi.org/10.1111/ecog.02881.

Rodríguez, Jon Paul, Lluís Brotons, Javier Bustamante, et Javier Seoane. 2007. « The Application of Predictive Modelling of Species Distribution to Biodiversity Conservation ». *Diversity and Distributions* 13 (3): 243‑51.

Sagarin, Raphael D., et Steven D. Gaines. 2002. « The ‘Abundant Centre’ Distribution: To What Extent Is It a Biogeographical Rule? » *Ecology Letters* 5 (1): 137‑47. https://doi.org/10.1046/j.1461-0248.2002.00297.x.

Schmutz, Ervin M., Michael E. Reese, Barry N. Freeman, et Larrye Chris Weaver. 1982. « Metric belt transect system for measuring cover, composition, and production of plants Vegetation ». *Rangelands* 4 (4). https://www.semanticscholar.org/paper/Metric-belt-transect-system-for-measuring-cover%2C-of-Schmutz-Reese/073625cd6ad9ac371cc392bacb1dc851ef42a0ba.

Schwager, Patrick, et Christian Berg. 2021. « Remote sensing variables improve species distribution models for alpine plant species ». *Basic and Applied Ecology* 54 (août): 1‑13. https://doi.org/10.1016/j.baae.2021.04.002.

« Section 4 : Prélèvement maximal autorisé (Articles R425-18 à R425-20) - Légifrance ». s. d. Consulté le 1 septembre 2023. https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006074220/LEGISCTA000006176910/.

Seif, Abdollah. 2014. « Using Topography Position Index for Landform Classification (Case Study: Grain Mountain) ». *Bulletin of Environment, Pharmacology and Life Sciences* 3 (11): 33-39.

Sinclair, Steve J., Matthew D. White, et Graeme R. Newell. 2010. « How Useful Are Species Distribution Models for Managing Biodiversity under Future Climates? » *Ecology and Society* 15 (1). https://www.jstor.org/stable/26268111.

Singer, A., U. Schückel, M. Beck, O. Bleich, H.-J. Brumsack, H. Freund, C. Geimecke, et al. 2016. « Small-Scale Benthos Distribution Modelling in a North Sea Tidal Basin in Response to Climatic and Environmental Changes (1970s-2009) ». *Marine Ecology Progress Series* 551 (juin): 13-30. https://doi.org/10.3354/meps11756.

Smithson, Michael, et Jay Verkuilen. 2006. « A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables ». *Psychological methods* 11 (avril): 54-71. https://doi.org/10.1037/1082-989X.11.1.54.

« SoilGrids ». s. d. Consulté le 1 septembre 2023. https://www.isric.org/explore/soilgrids.

Somodi, Imelda, Nikolett Lepesi, et Zoltán Botta-Dukát. 2017. « Prevalence Dependence in Model Goodness Measures with Special Emphasis on True Skill Statistics ». *Ecology and Evolution* 7 (3): 863-72. https://doi.org/10.1002/ece3.2654.

Sørensen, R., U. Zinko, et J. Seibert. 2006. « On the Calculation of the Topographic Wetness Index: Evaluation of Different Methods Based on Field Observations ». *Hydrology and Earth System Sciences* 10 (1): 101-12. https://doi.org/10.5194/hess-10-101-2006.

Stanton, Jessica C., Richard G. Pearson, Ned Horning, Peter Ersts, et H. Reşit Akçakaya. 2012. « Combining Static and Dynamic Variables in Species Distribution Models under Climate Change ». *Methods in Ecology and Evolution* 3 (2): 349-57. https://doi.org/10.1111/j.2041-210X.2011.00157.x.

« Terrain: Terrain Characteristics in Terra: Spatial Data Analysis ». s. d. Consulté le 1 septembre 2023. https://rdrr.io/cran/terra/man/terrain.html.

Tison, Jean-Marc, et Bruno de Foucault, éd. 2014. *Flora Gallica: flore de France*. Mèze, Hérault: Biotope éditions.

Van Couwenberghe, Rosalinde, Catherine Collet, Jean-Claude Pierrat, Kris Verheyen, et Jean-Claude Gégout. 2013. « Can Species Distribution Models Be Used to Describe Plant Abundance Patterns? » *Ecography* 36 (6): 665-74. https://doi.org/10.1111/j.1600-0587.2012.07362.x.

Villero, Dani, Magda Pla, David Camps, Jordi Ruiz-Olmo, et Lluís Brotons. 2017. « Integrating Species Distribution Modelling into Decision-Making to Inform Conservation Actions ». *Biodiversity and Conservation* 26 (2): 251-71. https://doi.org/10.1007/s10531-016-1243-2.

Weiss, Andrew D. 2001. « Topographic Position and Landform Analysis (Poster) ». The Nature Conservancy. http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf.

White, Neil A., Richard M. Engeman, Robert T. Sugihara, et Heather W. Krupa. 2008. « A comparison of plotless density estimators using Monte Carlo simulation on totally enumerated field data sets ». *BMC Ecology* 8 (1): 6. https://doi.org/10.1186/1472-6785-8-6.

Wikum, Douglas A., et G. Frederick Shanholtzer. 1978. « Application of the Braun-Blanquet Cover-Abundance Scale for Vegetation Analysis in Land Development Studies ». *Environmental Management* 2 (4): 323-29. https://doi.org/10.1007/BF01866672.

Wood, Simon. 2012. « mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation », http://cran.r-project.org/web/packages/mgcv/index.html.

Young, Mary, et Mark H. Carr. 2015. « Application of Species Distribution Models to Explain and Predict the Distribution, Abundance and Assemblage Structure of Nearshore Temperate Reef Fishes ». *Diversity and Distributions* 21 (12): 1428-40. https://doi.org/10.1111/ddi.12378.

Zimmermann, N. E., T. C. Edwards, G. G. Moisen, T. S. Frescino, et J. A. Blackard. 2007. « Remote Sensing-Based Predictors Improve Distribution Models of Rare, Early Successional and Broadleaf Tree Species in Utah ». *The Journal of Applied Ecology* 44 (5): 1057-67. https://doi.org/10.1111/j.1365-2664.2007.01348.x.

Zizka, Alexander, Daniele Silvestro, Tobias Andermann, Josué Azevedo, Camila Duarte Ritter, Daniel Edler, Harith Farooq, et al. 2019. « CoordinateCleaner: Standardized Cleaning of Occurrence Records from Biological Collection Databases ». *Methods in Ecology and Evolution* 10 (5): 744-51. https://doi.org/10.1111/2041-210X.13152.

Résumé / Abstract

La cueillette commerciale des plantes sauvages est une pratique ancestrale dont les modalités varient avec les cultures, les époques, et les caractéristiques biologiques des plantes prélevées. Face à un regain de la popularité des plantes sauvages, de nombreuses espèces font face à d'importantes pressions de cueillette, localement très problématiques. Or il y existe un important manque de données à leur sujet, qui empêche la mise en place d'actions et de régulations adaptées. Dans ce contexte, les SDM (Species Distribution Models) semblent une piste intéressante pour développer nos connaissances sur un grand nombre d'espèces avec un coût et une quantité de données limitées.

Ainsi, cette étude s'est concentrée sur la capacité des SDM à prédire l'occurrence et l'abondance des plantes sauvages à fine échelle spatiale. Tout d'abord, j'ai cherché à prédire l'occurrence à fine échelle d'une plante cueillie modèle, le thym (*Thymus vulgaris* L.). Ensuite, j'ai conçu un protocole d'échantillonnage sur le terrain pour tester la capacité des SDM à prédire différentes métriques d'occurrence et d'abondance. J'ai ainsi pu tester la qualité des modèles face à un jeu de données indépendant conçu spécialement à cet effet.

Les modèles développés ont montré une capacité moyenne à prédire l'occurrence et l'abondance du thym. Il reste beaucoup de variables à identifier, et de points techniques à affiner. Cependant, les résultats restent prometteurs et offrent de belles perspectives pour l'utilisation des SDM pour une meilleure gestion locale de la cueillette.

Mots clés : cueillette sauvage, modèles de distribution d'espèces, SDM, conservation, distribution, haute résolution, modèle d'ensemble, Biomod

—

The commercial harvesting of wild plants is an ancient practice, the methods of which vary with cultures, eras, and the biological characteristics of the plants collected. Faced with a resurgence in the popularity of wild plants, many species are experiencing significant harvesting pressures, which can be locally problematic. However, there is a significant lack of data on these species, hindering the implementation of appropriate actions and regulations. In this context, Species Distribution Models (SDM) appear to be an interesting avenue to develop our knowledge of a wide range of species with limited data and cost constraints.

Thus, this study focused on the capacity of SDM to predict the occurrence and abundance of wild plants at a fine spatial scale. First, I sought to predict the fine-scale occurrence of a model harvested plant, thyme (*Thymus vulgaris* L.). Then, I designed a field sampling protocol to test the ability of SDMs to predict different occurrence and abundance metrics. This allowed me to assess the quality of the models against an independent dataset specifically created for this purpose.

The developed models demonstrated a moderate ability to predict the occurrence and abundance of thyme. There are still many variables to identify and technical aspects to refine. However, the results remain promising and offer great prospects for the use of SDM in the improved local management of harvesting practices.

Key words : wild-plant harvesting, species distribution modelling, SDM, conservation, distribution, high resolution, ensemble models, Biomod