Test_1

Eléna

2024-12-16

Pour ce test, nous nous concentrons sur la mouche soldat noire (Hermetia illucens) qui compte 17 216 occurrences avant les étapes de filtrage.

Modification des intervalles : 30 valeurs par variables

Nous avons choisi de définir 30 intervalles par variables.

code de modification des intervalles :

intervals <- list(bio5 = seq(min(combinations[, 1]), max(combinations[, 1]), length.out = 30), hurs_min = seq(min(combinations[, 2]), max(combinations[, 2]), length.out = 30), npp = seq(min(combinations[, 3]), max(combinations[, 3]), length.out = 30))

Espace environnemental considéré pour ce test :

- Total number of cells with environmental conditions in the geographical space: 12513421
- Number of duplicated conditions: 121008072
- Number of unique cells (environmental space): 5349

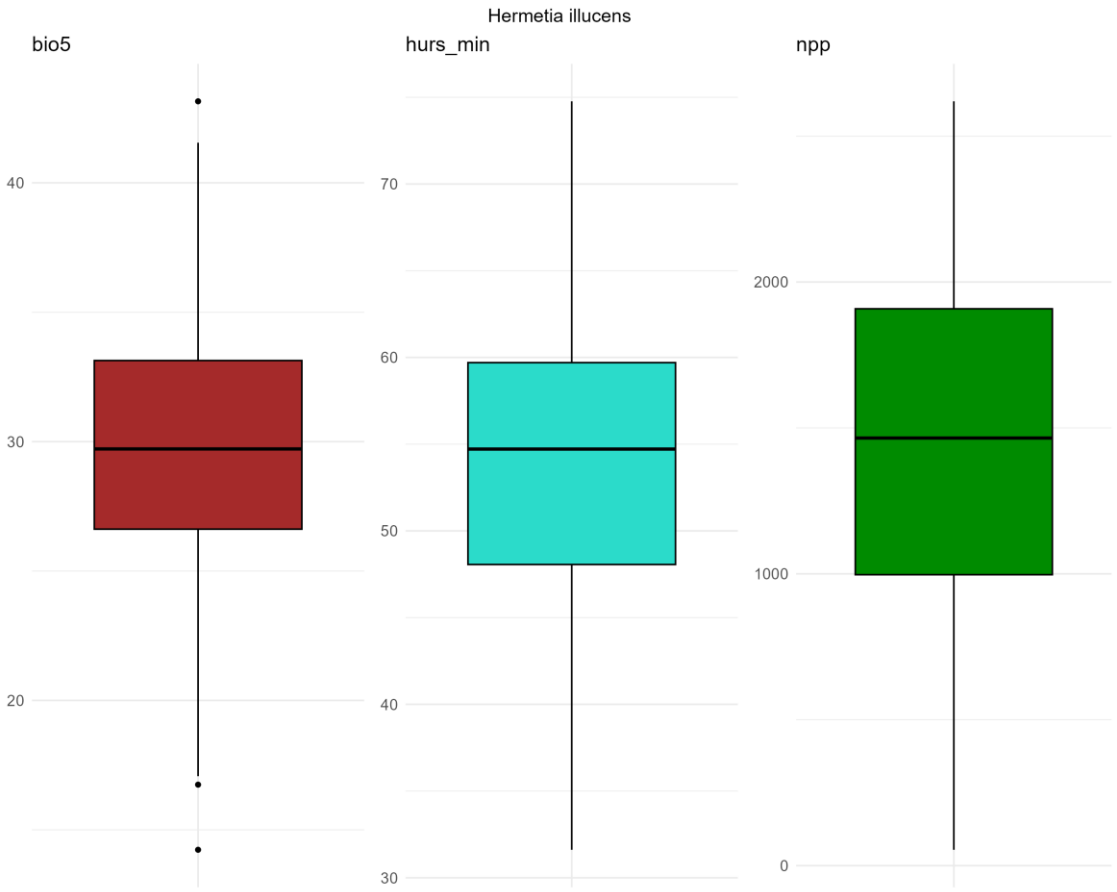
On observe que l'espace environnemental est très restreint, ce qui est dû au fait que les intervalles des variables sont larges, permettant ainsi à ces valeurs d'être facilement retrouvées à la surface de la Terre.

Filtration des occurrences :

- Suppression des occurrences correspondant à des données environnementales manquantes.

- Suppression des occurrences dupliquées dans l'espace environnemental.

Ainsi, le nombre d'occurrences utilisables pour calibrer nos modèles est réduit à 1009.



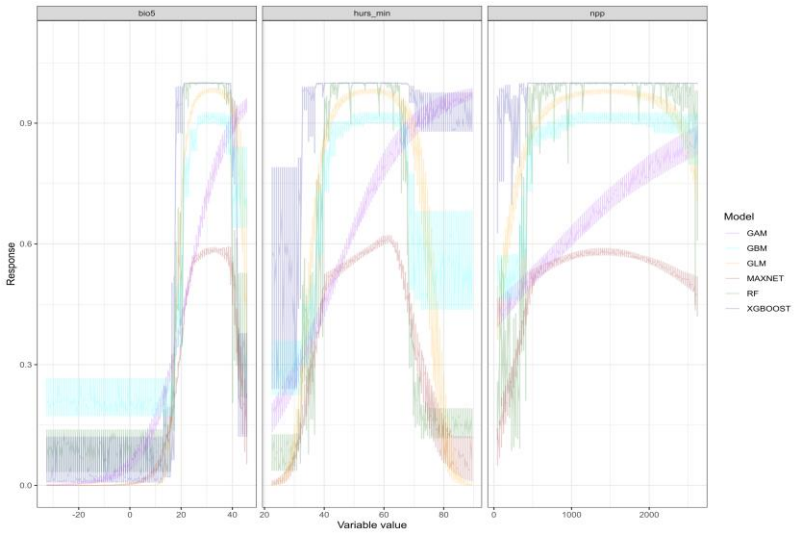
Répartition des occurrences le long des variables

On observe que les occurrences des espèces se situent principalement dans des intervalles assez restreints: Entre 26 et 34 degrés pour la température maximale, 48% et 60% d'humidité minimale, et une productivité primaire nette entre 1000 à 2000. D'après ces données d'occurrences, l'espèce peut vivre jusqu'à maximum 43 degrés avec une humidité minimale de 30% et une productivité primaire nette comprise entre 54 et 2619.

La littérature statistique récente recommande de générer un nombre équivalent de points de pseudo-absences et de présences, ces derniers étant tirés aléatoirement en dehors du convexhull (l'espace environnemental délimité par les points de présence de l'espèce). Cette approche garanti une meilleure représentation des conditions environnementales disponibles pour le modèle.

Pour améliorer la représentation de l'espace environnemental des pseudo-absences, nous procédons à plusieurs runs de génération de pseudo-absences (ici, 5).

Courbes de réponses



Courbes de réponses des variables utilisées pour le test 1

Les courbes de réponses des modèles témoignent toutes d'une réponse forte de la probabilité de présence de l'espèce par rapport aux variables prédictives utilisées. Les courbes de réponses semblent suivre la même trajectoire, c'est à dire en cloche à l'exception des modèles GAM. Bien que légèrement plus large, les modèles prédisent globalement une forte probabilité de présence entre des intervalles similaires que ceux des occurrences de l'espèce.

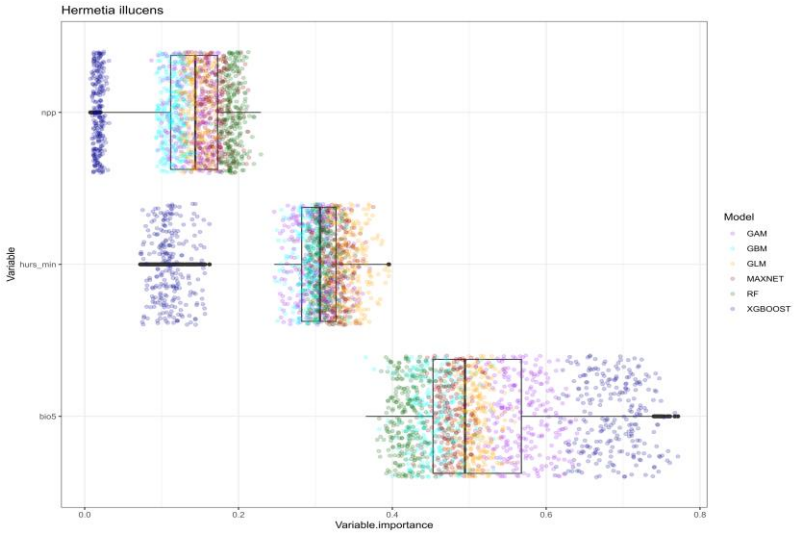
Cependant on observe d'importantes et fréquentes inflexions, en particulier pour les algorithmes random forest XGBoost et GBM, qui suggèrent que les modèles semblent constamment se réajuster.

Hypothèses liées à ces inflexions :

- Est-ce que ces inflexions seraient dues au choix des intervalles des variables qui ne sont pas assez nombreux ? En effet nous avons environ 30 valeurs par variables ce qui donne 5349 pixels (espace environnemental) considérés pour la modélisation ce qui doit être surement trop peu pour pouvoir faire de bonnes estimations. Nous n'avons peut-être pas assez de vraies occurrences pour pouvoir faire de bonnes estimations de probabilité d'occurrences.

- Est-ce qu'il y a de l'overfitting des modèles ? En effet, on a l'impression d'observer aussi un surajustement des modèles.

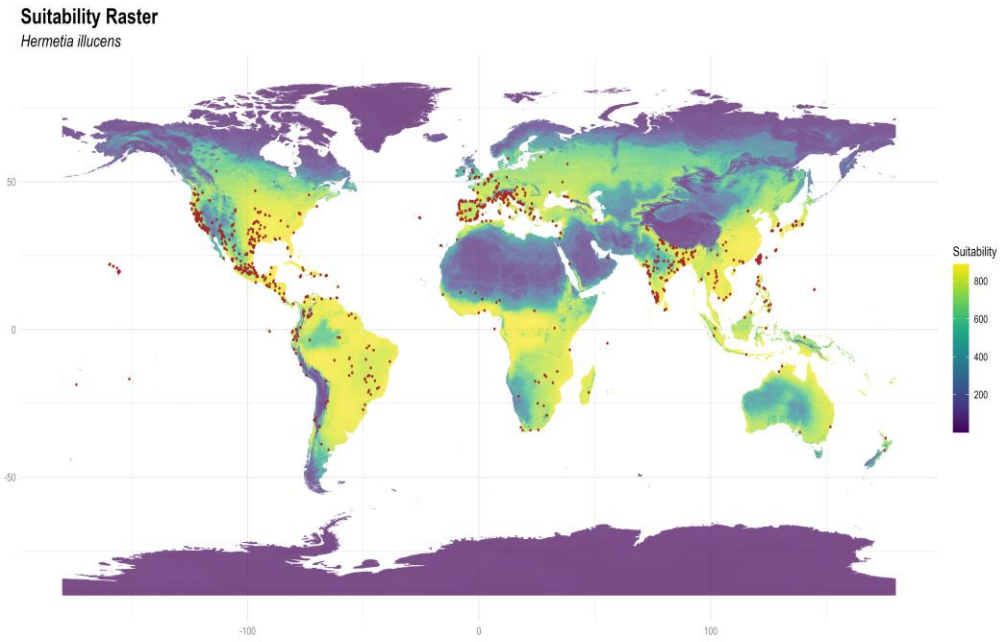
Importance des variables



Importance des variables dans la prédiction de l'établissement de l'espèce par les modèles

- La température maximale (bio5) = variable qui joue le plus sur la présence d'Hermetia (0.8/1 à son maximum).
- XGBOOST est plus radical que les autres algorithmes, soit en donnant plus d'importance à la variable que les autres (bio5), soit moins (npp et hurs_min).
- XGBOOST considère que bio5 est presque uniquement la seule variable qui joue sur la présence de notre espèces tandis que Random Forest considère les 3.

Carte d’indice de favorabilité

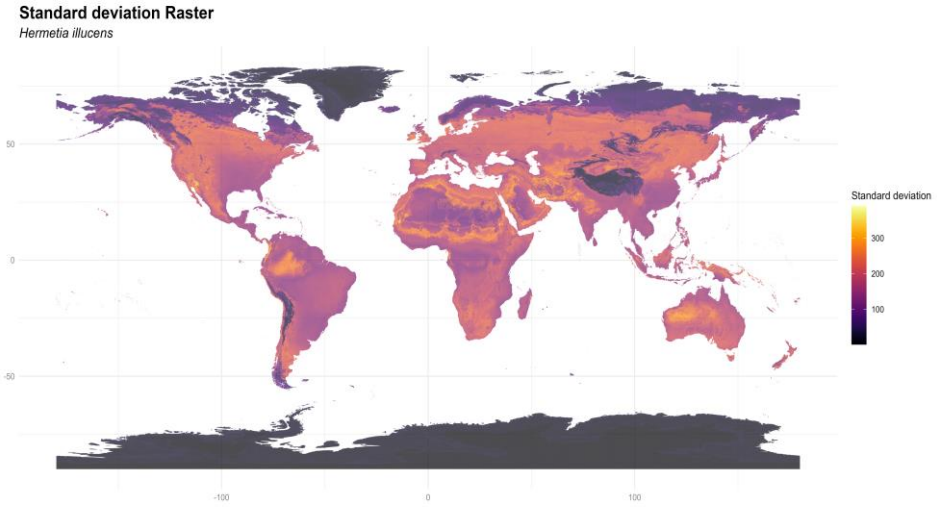


Indice de favorabilité final

La carte de l’indice de favorabilité final représente la moyenne des modèles ayant satisfait les critères de qualité (c’est-à-dire un indice de Boyce supérieur à 0,7). Elle met en évidence le gradient de favorabilité de l’habitat à l’échelle mondiale, avec une zone de forte favorabilité globalement marquée par des valeurs très élevées. On observe également des zones spécifiques où la favorabilité est faible ou inconnue, principalement au-delà de 55 degrés de latitude. Ces zones incluent la région désertique de l’Afrique du Nord-Centrale, la côte ouest de l’Amérique du Sud, ainsi qu’une zone particulière en Asie centrale.

- Prédiction d’établissement fort plus large que dans le test-0.
- Prédiction d’établissement est globalement importante. Les algorythmes ont du mal à trancher sur la probabilité de présence de l’espèce.

Carte incertitude



Incertitude des modèles

La carte d'incertitude correspond à l'écart-type des indices des modèles ayant passé les critères de qualité. Elle complète l'analyse de la carte de l'indice de favorabilité finale. On remarque que les écart types sont élevés principalement dans les zones à forte favorabilité à l'exception de la bio région Saharo-Arabian où l'écart-type entre les modèles est très important.

- Ecart-types globalement plus faibles que dans le test 0 : moins de contradiction entre les modèles.
- Ecart-types moyens présents presque partout, même dans les régions avec beaucoup d'occurrences (contrairement au Test-0). Seulement à certains endroits les algorithmes sont tous en accord (e.g cordillère des Andes).