Test-2

Eléna

2024-12-17

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

On reste sur les 3 variables utilisées dans le test 1: bio5, hurs_min et Npp.

Augmentation du nombre de valeurs au sein des intervalles à 80

On augmente à 80 le nombre de valeurs pour chacune des variables.

Code pour le changement d'intervalles:

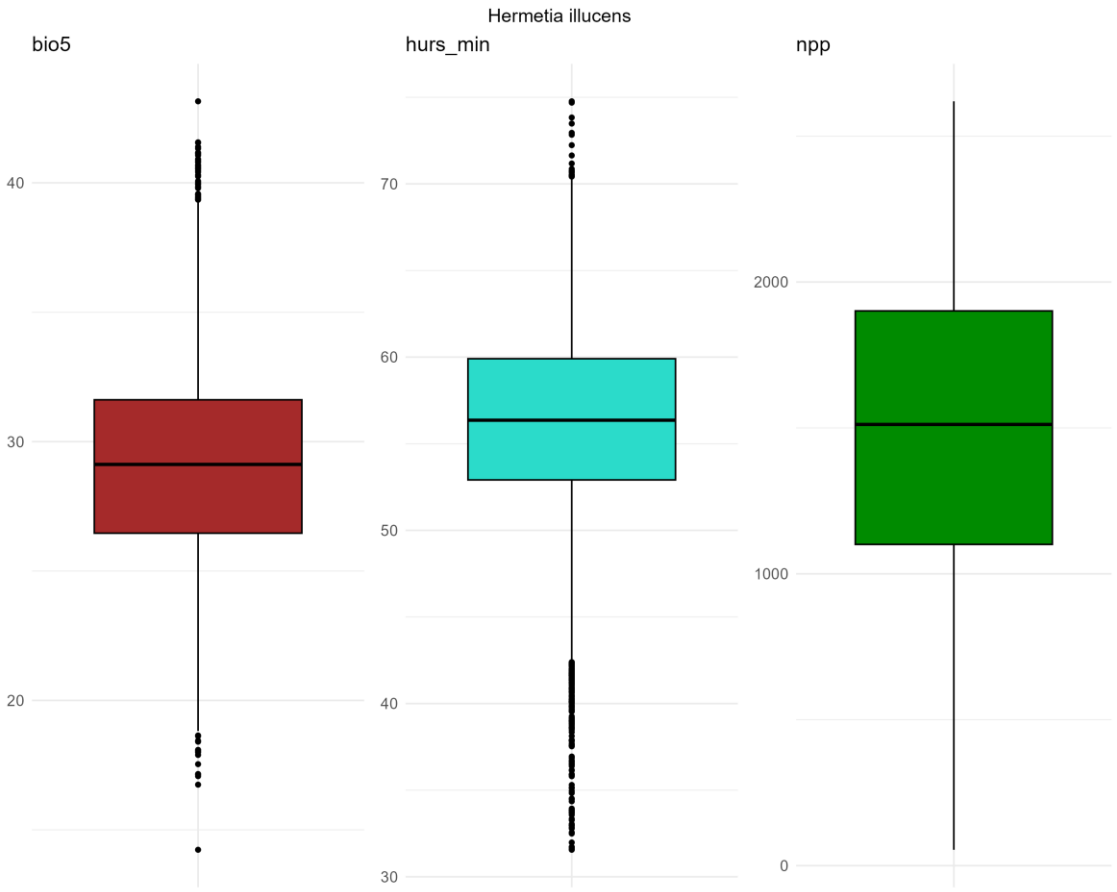
intervals <- list(bio5 = seq(min(combinations[, 1]), max(combinations[, 1]), length.out = 80), hurs_min = seq(min(combinations[, 2]), max(combinations[, 2]), length.out = 80), npp = seq(min(combinations[, 3]), max(combinations[, 3]), length.out = 80))

Espace environnemental considéré pour ce test :

- Total number of cells with environmental conditions in the geographical space: 12513421
- Number of duplicated conditions: 12434039
- Number of unique cells (environmental space): 79382

L'espace environnemental considéré pour ce test est plus large, car les intervalles des variables sont plus serrés, ce qui restreint le nombre de combinaisons similaires.

Après avoir appliqué les étapes de filtration, le nombre d'occurrences utilisables pour calibrer nos modèles est réduit à 4310.

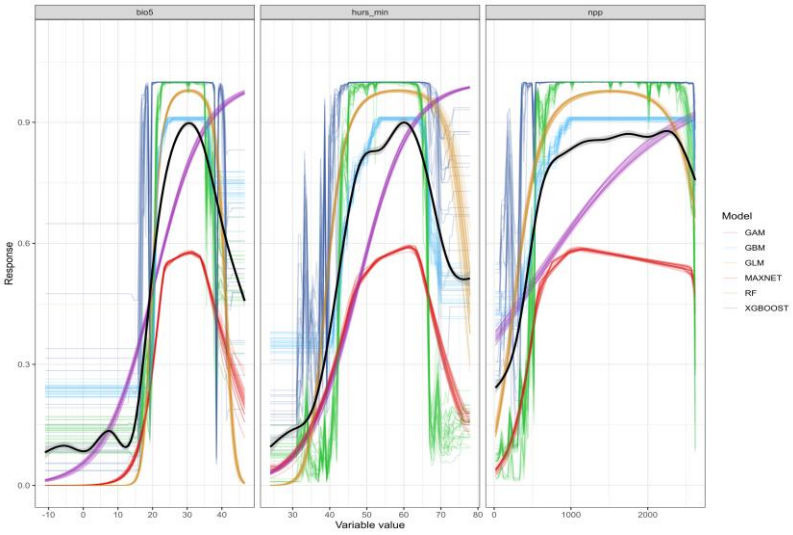


Repartition des occurences le long des variables

Cette figure met en évidence un plus grand nombre de valeurs d'occurrences aux extrêmes, ainsi qu'un intervalle interquartile plus réduit par rapport au test 1. Pour rappel, une seule valeur de présence est conservée par intervalle pour nos modèles, bien qu'une grande diversité de valeurs existe au sein de ces intervalles, particulièrement lorsque ceux-ci sont plus larges (comme observé dans le test 1). Dans ce test, les occurrences sont i) plus nombreuses et ii) associées à des intervalles plus restreints, ce qui permet une meilleure représentation de la niche climatique de présence de l'espèce.

De manière similaire au Test 1, nous procédons à 5 séries de tirages de 4310 pseudo-absences, sélectionnées aléatoirement en dehors du convex hull de l'espèce. Cela nous donne un total de plus de 21 550 tirages de pseudo-absences.

Courbes de réponses des modèles



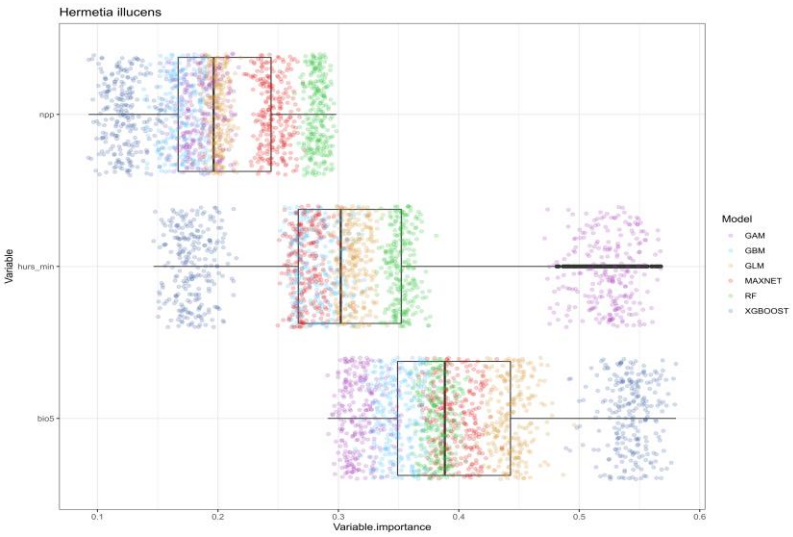
Courbes de réponses des variables utilisées pour le test 2

Sur cette figure, on remarque, par rapport au test 1 :

- Une même réponse générale des modèles (en cloche)
- Présence moins forte d'inflexions, à l'exception de XGBOOST qui en présente davantage.

Puisqu'on observe une réponse en cloche pour la majorité des modèles, je propose d'enlever l'algorithme GAM pour la modélisation finale.

Importance des variables



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

- L’ordre d’importance des variables inchangé par rapport au Test 1.
- Les valeurs d’importance sont plus équilibrées dans ce test.

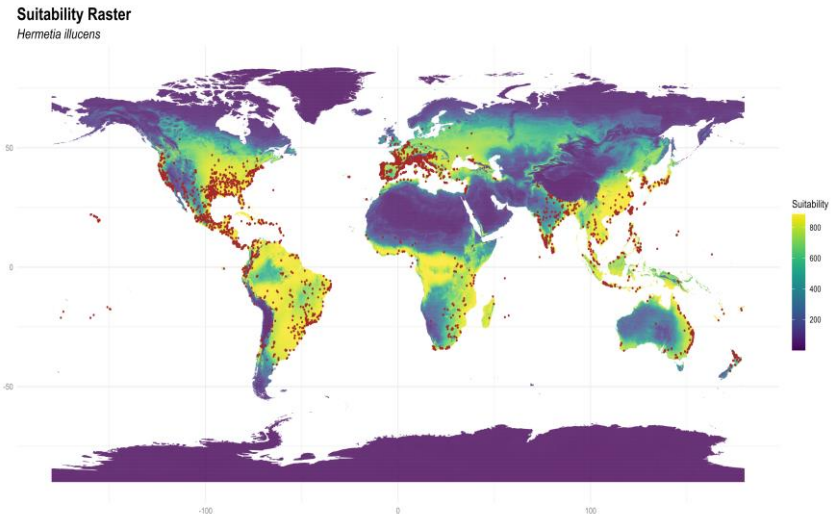
L’algorithme XGBOOST accorde moins d’importance à la variable bio5 par rapport au test précédent (environ 0,6 ici contre 0,8). Il continue toutefois à se démarquer des autres modèles en attribuant des valeurs extrêmes. L’algorithme GAM, quant à lui, montre une distinction notable pour la variable hurs_min, une différence qui n’était pas observée dans le Test 1.

Avec un plus grand nombre de points à analyser, les algorithmes révèlent davantage leurs spécificités. En effet, chaque méthode utilise des approches de prédiction différentes (par exemple, relations linéaires, non linéaires ou quadratiques), ce qui semble accentuer leurs divergences dans les résultats.

Compte tenu des différences observées dans les courbes de réponse et les variables d’importance, il semble pertinent de retirer le modèle GAM. Il pourrait également être envisagé d’exclure XGBOOST, dont les courbes de réponse présentent des inflexions très marquées et étranges (peut-être dû à une mauvaise paramétrisation), et qui fournit des résultats très divergents concernant l’importance des variables par rapport aux autres modèles.

Si nous décidons finalement de conserver XGBOOST, je propose de retravailler sa paramétrisation afin d’obtenir des résultats plus cohérents avec les autres modèles.

Carte d’indice de favorabilité



Indice de favorabilité final

On observe des zones à forte suitabilité globalement similaires au raster de suitability du test 1. Cependant, on observe davantage de zones à plus faibles probabilité d’établissement. En effet, L’Australie présente désormais davantage de zones à

favorabilité plus faible que pour le test 1. L'estimation des zones de favorabilité des modèles est plus précis pour ce test. Ceci peut être expliqué par le nombre d'occurences considérées dans le test 2, plus important que dans le test 1 ce qui permet aux modèles d'avoir une meilleure assurance sur leur prédiction.

Carte incertitude

Logiquement, l'incertitude devrait être moins élevée si on suit la logique du raster de suitabilité puisque les modèles sont plus précis.

Conclusion

Je trouve que la modélisation incluant davantage d'intervalles est plus pertinente car elle permet d'avoir une meilleure représentation de la niche climatique de présence de l'espèce. Cette paramétrisation semble permettre aux modèles d'avoir une prédiction plus précise. Concernant les modèles utilisés, je pense que nous pouvons retirer le modèle GAM qui présente des formes de courbes de réponses totalement différentes des autres modèles. Concernant le modèle XGBOOST, je propose de soit le retiré, soit le reparamétriser. Enfin, je pense qu'il serait pertinent d'inclure une variables d'habitat et ou socioéconomique qui influe sur nos populations (e.g la population humaine mondiale).

Note : Il faut que je réduise mon extent de latitude pour ne plus inclure l'antarctique et l'arctique de mes modèles.