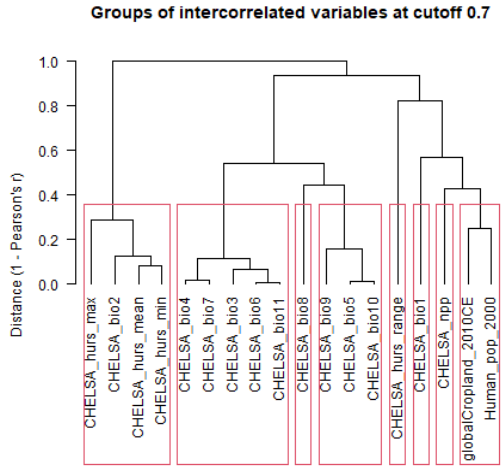
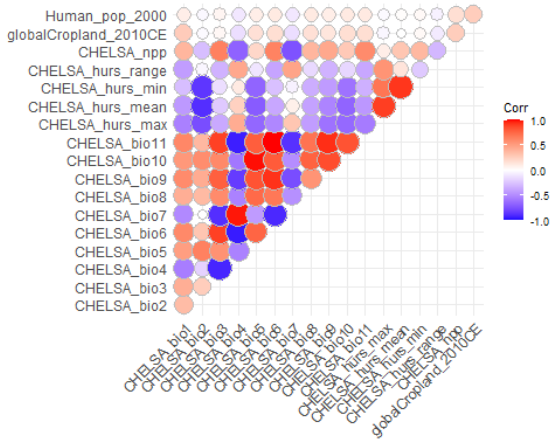
Test-Variables

Eléna

2025-01-20

J'ai appliqué un facteur 15 à tous mes rasters (0.5arcminutes×15=7.5arcminutes) pour que le code tourne plus rapidement.





Colinéarité des variables & Matrice des corrélations

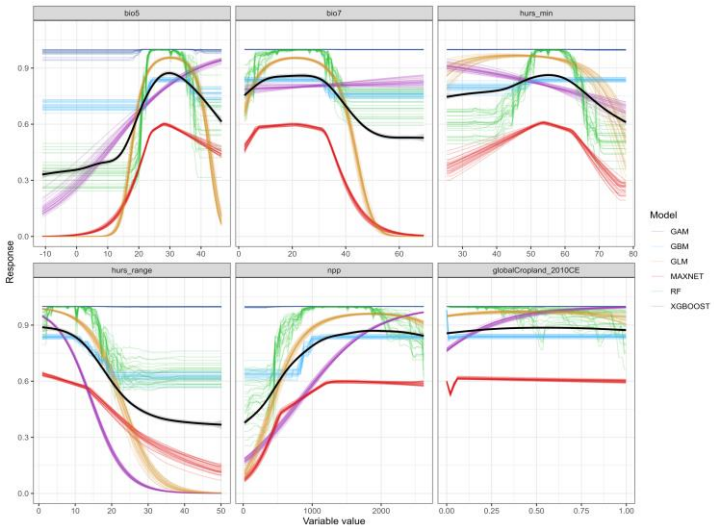
Le choix des variables initiales s'est fait en fonction de la biologie des insectes (leur sensibilité aux conditions climatiques retrouvée dans la littérature) mais aussi en fonction des résultats de l'arbre de décision ci-dessus.

Test initial (0) avec 6 variables (ref document Test-0)

Nous avons fait le choix de ne garder deux variables de températures (bio5 et bio7) et deux variables d'humidité (Hurs min et range). En effet, humidité minimale joue un rôle très important chez les insectes et on considère l'intervalle d'humidité plus important que l'humidité maximale.

Test avec les 6 variables :

- Bio5 (température maximale)
- Bio7 (range température)
- Hurs_min (humidité minimale)
- Hurs_range (range d'humidité)
- NPP (productivité primaire nette)
- globalCropland_2010CE (terres cultivées 2010)

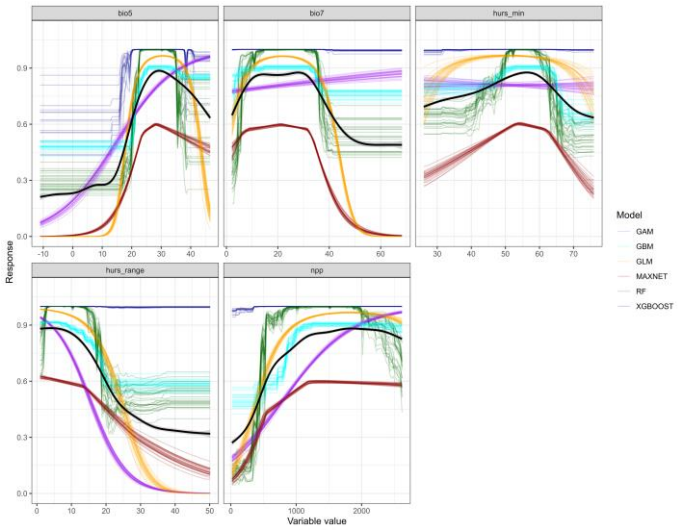


- Absence de réponse de XGBOOST.
- L'algorithme GBM répond très légèrement aux différentes variables.
- Très mauvaise estimation de tous les algorithmes concernant la variable cropland.

5 variables

–> Elimination de la variable CropLand.

- Bio5 (température maximale)
- Bio7 (range température)
- Hurs_min (humidité minimale)
- Hurs_range (range d'humidité)
- NPP (productivité primaire nette)



Courbes de réponses des 5 variables

- Courbe de réponse moyenne des algorithme (en noire) est davantage précise (visible surtout pour la variable npp).
- Toujours mauvaise réponse de l'algorithme XGBOOST bien qu'une légère réponse est observée pour certaines variables).
- Réponse davantage drastique pour GBM.
- les algorithmes GBM et RF ont toujours du mal à trancher (réponse jamais en dessous de 0.5) sur la variable hurs_range.

–> C'est mieux mais toujours de mauvaises estimations.

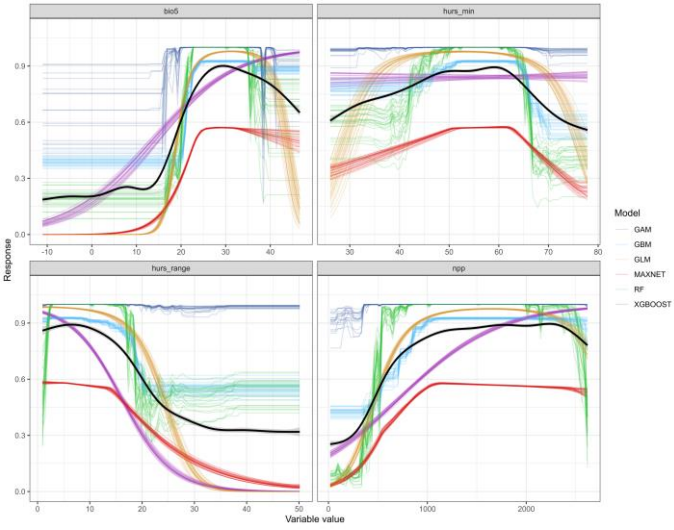
4 variables_Hurs-range

–> Elimination de cropland et de Bio7

- Bio5 (température maximale)
- Hurs_min (humidité minimale)
- NPP (productivité primaire nette)
- Hurs-range

Total number of cells with environmental conditions in the geographical space: 979220 Number of duplicated conditions: 575262 Number of unique cells (environmental space): 403958

Number of unique presences: 5255



- toujours très faible réponse d'XGBOOST.
- les algorithmes GBM et RF ont toujours du mal à trancher (réponse jamais en dessous de 0.5) sur la variable hurs_range.

4 variables_Bio7 (thermal_range)

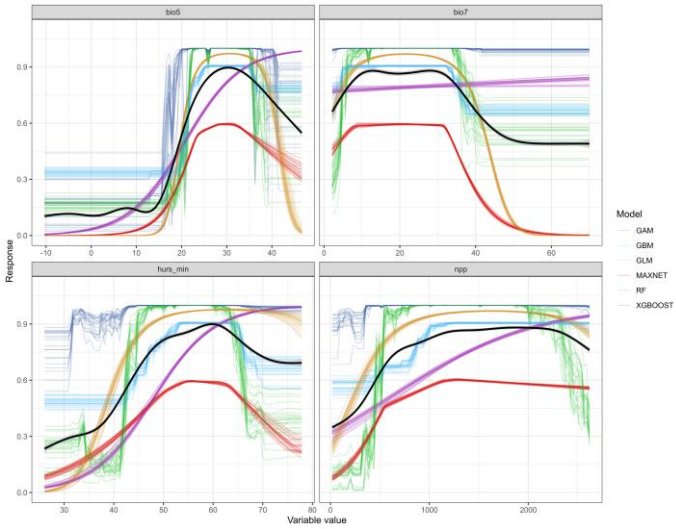
–> - cropland et hurs_range

- Bio5 (température maximale)
- Bio7 (range de température annuel)
- Hurs_min (humidité minimale)
- NPP (productivité primaire nette)

Environmental space :

- Total number of cells with environmental conditions in the geographical space: 979220
- Number of duplicated conditions: 662902
- Number of unique cells (environmental space): 316318

Number of unique presences: 5227



- Meilleure réponse de l'algorithme XGBOOSt bien que toujours pas encore suffisant
- XGBOOST n'arrive pas à utiliser la variable bio7.

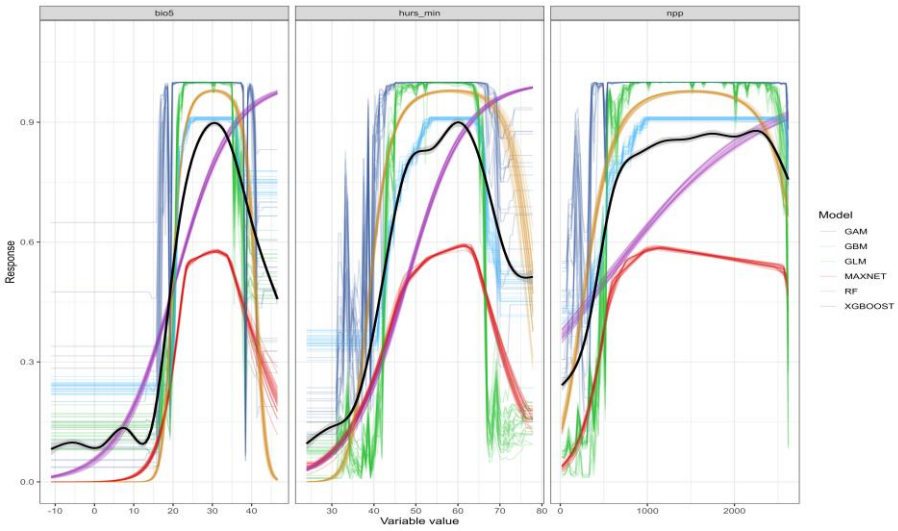
3 variables (ref document Test-2)

–> - cropland, hurs_range et bio7

- Bio5 (température maximale)
- Hurs_min (humidité minimale)
- NPP (productivité primaire nette)

Total number of cells with environmental conditions in the geographical space: 979220 Number of duplicated conditions: 893303 Number of unique cells (environmental space): 85917

Number of unique presences: 4077



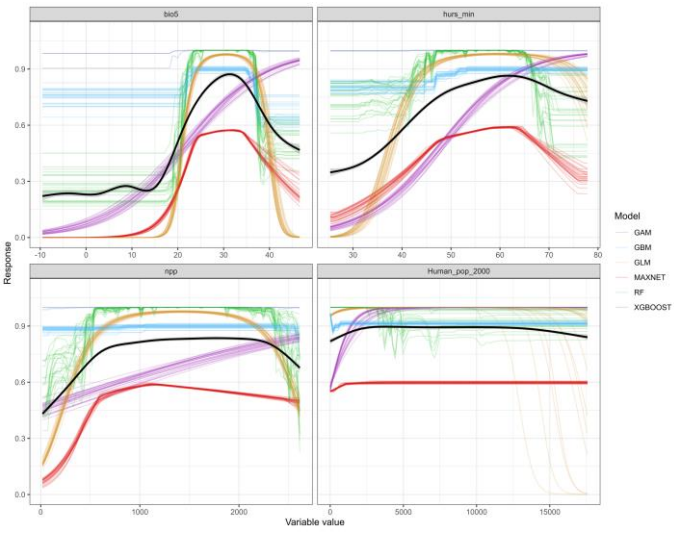
Courbes de réponses des 3 variables

- Meilleure réponse de tous les algorithmes sur toutes les variables.
- XGBOOST est davantage sensible aux variables.

–> Meilleur modèle pour le moment.

3 variables + Human population

Rajout de la variable densité humaine pour changer de cropland (les deux sont corrélées)



- XGBOOST et GBM ne fonctionnent plus lorsqu'il y a rajoute de la variable de densité de population humaine.
- La réponse des algorithmes à la population humaine est globalement mauvaise.

-- > Nous n'utiliserons pas cette variable dans nos modèles.

Conclusion

Les algorithmes rencontrent des difficultés, voire ne répondent pas, à estimer correctement les probabilités de présence de notre espèce (Hermetia illucens) en fonction des variables liées aux terres cultivées, à la densité de population humaine, et aux intervalles de conditions climatiques (hurs_range et bio7). Par conséquent, nous avons décidé d'utiliser uniquement trois variables pour la modélisation et les prédictions.

- bio2 (annual maximum temperature)
- hurs_min (annual min humidity)
- npp (productivité primaire nette)

Je propose également de retirer l'algorithme GAM, car ses résultats semblent trop simples et très différents de ce que nous observons aux niveau des données d'occurrences ainsi que par rapport aux autres algorithmes.

Enfin, je suggère de chercher à optimiser davantage l'algorithme XGBoost pour améliorer ses performances.