A) Random forest pour trouver ce qui détermine bird_health_index

Afin de trouver les indices de santé manquants, nous avons tout d'abord généré un code qui nous permettrait de prédire l'indice de santé en fonction des autres variables de la matrice de données. Nous avons défini les variables feeder_count, raptor_presence, road_density, shrub_density et thrash_can_count comme des prédicteurs fixes; nous avons considéré les variables city_id et park_id comme des effets aléatoires. Étant donné que les variables city_id et park_id sont connectées, nous les avons regroupées en une seule variablec(id_merf) avec la fonction *interaction*. Nous avons aussi décidé de ne pas inclure les rangées contenant des valeurs manquantes dans nos prédictions.

Nous avons par la suite fait une forêt d'arbres décisionnels à effets mixtes avec les valeurs suivantes :

- X : feeder_count, raptor_presence, road_density, shrub_density et thrash_can_count.
- Y : bird_health_index
- Z : 1 (pour considérer les effets aléatoires)
- time : 1 (puisqu'il n'y a pas de mesures de temps)
- id : id_merf

Les résultats de la forêt ont indiqué que l'importance relative des cinq variables fixes pour prédire l'indice de santé (fig. 1). La racine de l'erreur quadratique moyenne (RMSE) était 11,09 et le R^2 de validation croisée était 0,4320.

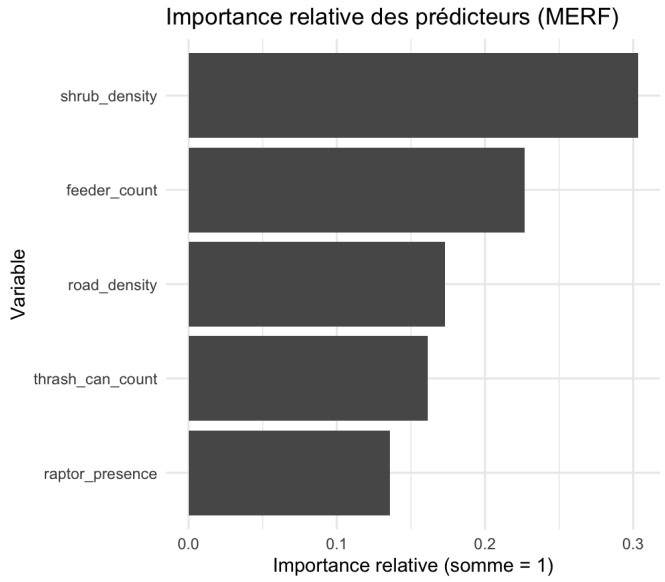

Figure 1. Importance relative des prédicteurs fixes pour déterminer la valeur de l'indice de santé des oiseaux.

B) Utilisation de la random forest pour déterminer les indices de santé manquants

Nous avons alors utilisé la forêt entraînée pour prédire les indices de santé des dix inconnus en fonction des autres variables. Les valeurs estimées des indices sont les suivantes :

1. 9.574333
2. 2.308461
3. 16.784649
4. 2.883590
5. 2.541459
6. 10.980779
7. 17.954286
8. 9.616491
9. 16.169152
10. 16.479205

C) Comparaison des indices observés et prédits

Enfin, nous avons construit un nuage de points représentant les valeurs observées en abscisse et les valeurs prédites en ordonnée, afin d'évaluer le pouvoir prédictif de notre modèle de forêt (figure 2). Le R^2 de la corrélation entre les deux valeurs était 0,8126.

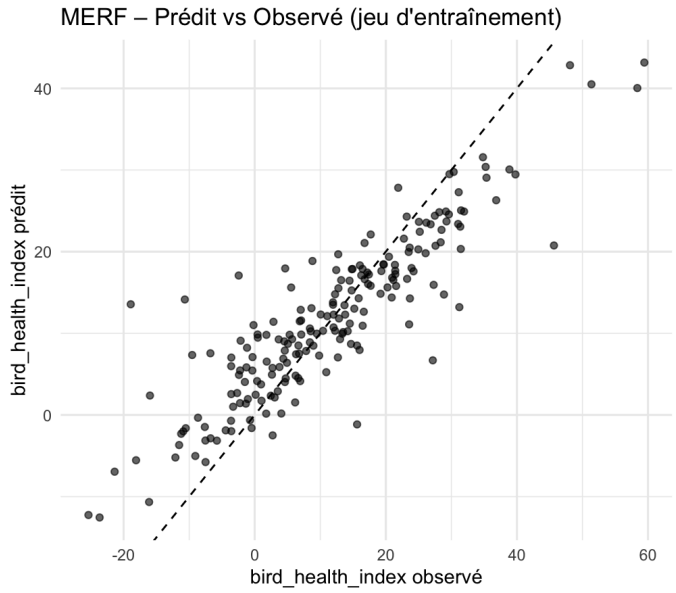

Figure 2. Valeurs des indices de santé des oiseaux prédites en fonction des valeurs observées.