RAPPORT 8SAE

1. Description des données

La base de données à analyser comporte 200 observations d'oiseaux, un indice de santé et 7 variables descriptives. La variable dépendante *bird_health_index* est numérique et contient des valeurs de -25.36 à 59.44, alors que les variables dépendantes sont un assemblage d'entiers positifs (*feeder_count*, *thrash_can_count*), de nombres positifs (*road_density*, *shrub_density*) de valeurs binaires (*raptor_presence*) et de facteurs imbriqués (*city_id*, *park_id*).

Les facteurs imbriqués *city_id* et *park_id* sont relativement bien distribués, tout comme les NAs qui peuvent être retirés sans débalancer les niveaux de la base de données. La base de données filtrée sur les lignes entières résulte en 183 observations dans 4 villes (n de 39 à 53) distribuées en 12 parcs (n de 9 à 22).

L'examen des distributions et de la matrice de corrélation (Fig. 1) suggère que les distributions de *feeder_count*, *road_density* et *shrub_density* sont des distribution de poisson, que les niveaux de *raptor_presence* sont équitablement distribués entre 0 et 1, que la distribution de *thrash_can_count* est à peu près normale et que la variable dépendante *bird_health_index* est normalement distribuée. L'examen des distributions par ville suggère que *feeder_count* et *thrash_can_count* peuvent être des distributions de poisson ou normale, selon les cas. Finalement, des correlations significatives existent entre *thrash_can_count, feeder_count* et *road_density*, ce qui suggère une potentielle autocorrelation des prédicteurs.

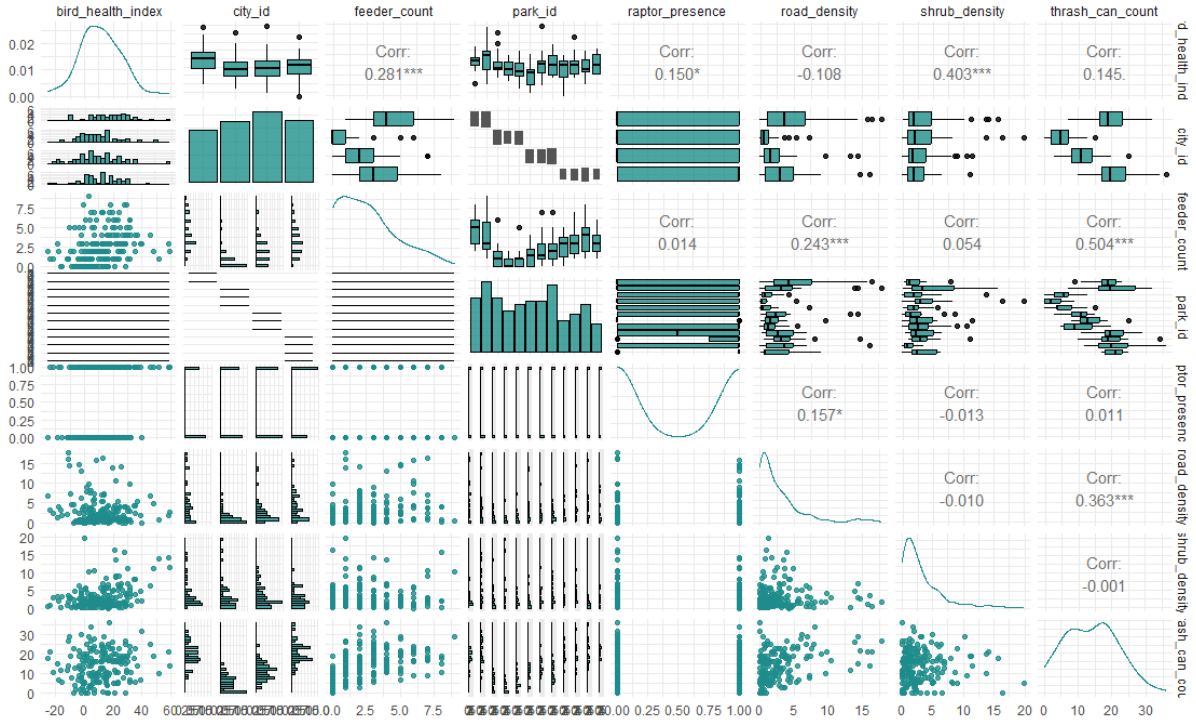

Figure 1. Matrice de corrélations et distributions des variables indépendantes et dépendantes, ainsi que leur distribution par niveau de facteurs. La figure est générée par la fonction mlr3::autoplot().

2. Modélisation de la santé des oiseaux

La modélisation empirique de l'indice de santé des oiseaux *bird_health_index* fut réalisée en deux parties: (a) par modélisation statistique linéaire, linéaire mixte et non-linéaire (modèles additifs généralisés) suivi d'une sélection de prédicteurs et d'une sélection par critère d'information d'Akaike (AIC), et (b) par apprentissage machine via modèles *random forest* de type ranger, avec optimisation des hyper-paramètres par grille automatisée à l'aide de Caret avec une validation croisée pénalisant pour le croisement des facteurs *city_id* et *parks_id*.

a. Modèles classiques

La modélisation classique par lm() excluant les facteurs suggère que la colinéarité des prédicteurs est mineure (VIF < 2) et exclue *thrash_can_count* des prédicteurs significatifs. Le modèle lm2 exclut ce prédicteur. Les modèles mixtes considèrent les villes et les parcs comme des effets aléatoires imbriqués qui résultent à nouveau en l'exclusion de *thrash_can_count* des prédicteurs significatifs. Le modèle mlm2 exclut ce prédicteur. Les modèles gam considèrent les villes et parcs commes des effets aléatoires indépendants, et permettent de tester des *splines* non linéaires sur chaque prédicteur. Le facteur « estimated degrees of freedom » (EDF) permet d'exclure les effets non-linéaires des prédicteurs *raptor_presence* et *road_density*, alors que la significativité des prédicteurs permet d'exclure le prédicteur *thrash_can_count.* La sélection par AIC favorise largement le modèle gam2 non linéaire comme étant le meilleur modèle, avec un R2 de 0.44 et un RMSE 11.09 (Tableau 1).

Tableau 1. Performance des modèles classiques

Name	Model	AIC (weights)	RMSE	R2
df.lm	lm	1453.5 (<.001)	12.355	0.298
df.lm2	lm	1453.0 (<.001)	12.407	0.292
df.mlm	lmer	1455.7 (<.001)	11.808	0.335
df.mlm2	lmer	1483.9 (<.001)	11.722	0.336
df.gam	gam	1426.2 (0.301)	11.121	0.435
df.gam2	gam	1424.5 (0.699)	11.093	0.437

b. Modèle ML

Le modèle ranger a été calibré selon un *tuneLength* de 3, selon une validation croisée à 5 niveaux tenant en compte les villes et parcs avec 1000 arbres par niveau. L'importance des prédicteurs est estimée par permutation. Bien que le R2 (OOB) globale du modèle soit de 0.35, avec un root mean squared error (RMSE) *out-of-bag* de 11.88, cette performance diminue entre 0.15 et 0.19 et 13.42 et 14.27 selon la validation croisée et la calibration

lorsque la ville et le parc sont considérés, ce qui suggère que ces facteurs sont importants à la prédiction. En ordre décroissant, les prédicteurs les plus importants sont *shrub_density*, *feeder_count*, *raptor_presence* et *road_density*, alors que *thrash_can_count* ne contribue pas au modèle.

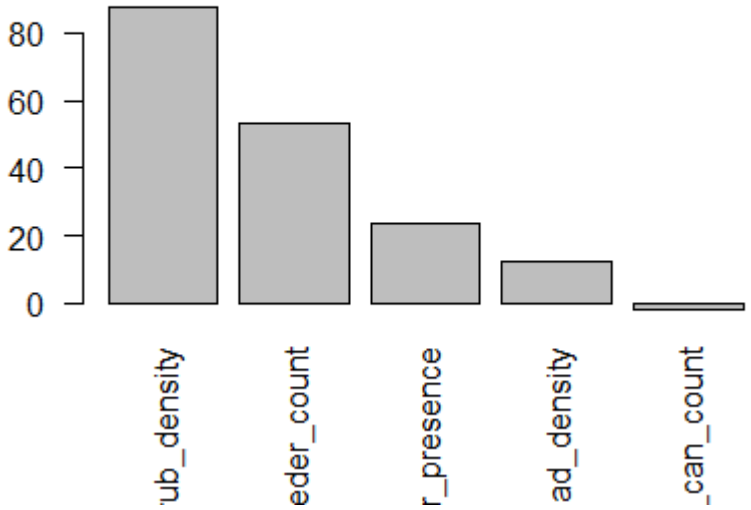

Figure 2. Importance des prédicteurs selon le modèle *random forest* ranger.

c. Prédictions

La figure 4 présente les prédictions de chaque modèle pour les 10 individus inconnus.

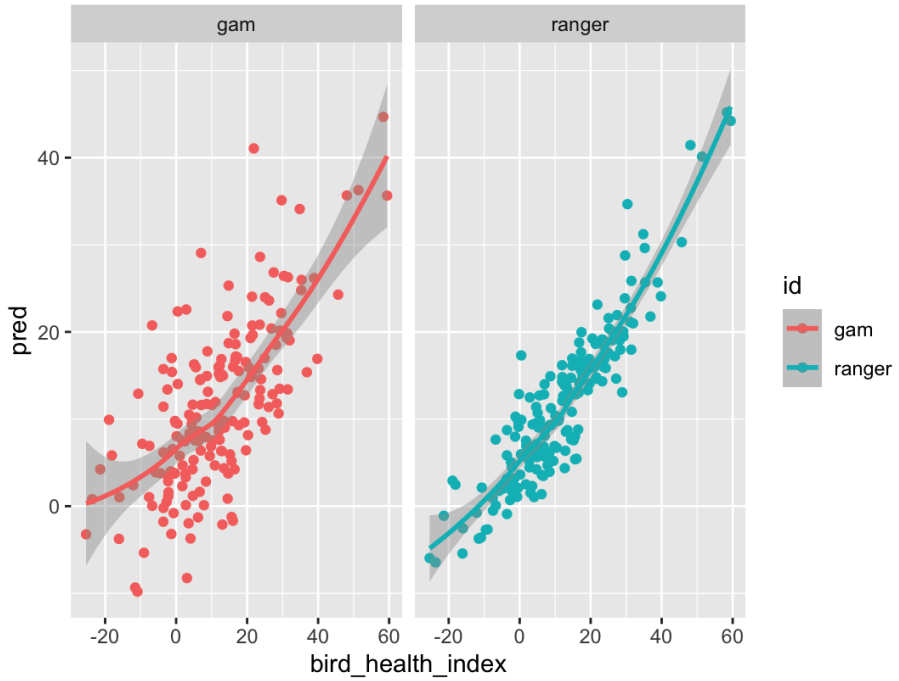

Figure 3. Valeurs prédites vs. observées de *bird_health_index* par (a) le modèle gam2 et (b) le modèle ranger.