# Prédiction de l’indice de santé chez les Charles-Martin pêcheur en fonction de variables environnementales

Rive hacking 2026

Groupe : V7ZL

Université du Québec à Trois-Rivières

## Introduction

L'objectif du projet est de prédire la santé de 10 individus de Charles-Martin pêcheur où les valeurs de l'indice de santé n'ont pas été correctement notées sur le terrain, en se basant sur les autres variables environnementales.

Nous disposons de deux jeux de données. Le premier "dataset.csv" contient les informations de 200 individus en incluant leur indice de santé. Le deuxième jeu de données "to_predict.csv", quant à lui, regroupe les 10 individus dont les données de santé sont manquantes.

## Méthode

Préparation des données

Les variables catégorielles (city_id et park_id) ont été transformées en facteurs.

L'ensemble des variables numériques ont été normalisées.

Les variables explicatives ont été sélectionnées par une analyse de corrélation entre les variables numériques.

Analyse des données

Différents modèles linéaires généralisés à effet mixtes (GLMM) ont été réalisés. Le meilleur modèle a été sélectionné à l'aide d'une comparaison du Critère d'Information d'Akaike (AIC) et par principe de parcimonie. Les indices de santé des 10 individus de Charles-Martin pêcheur ont été prédits à partir de ce modèle. L'intervalle de significativité est indiqué par ces symboles correspondant aux valeurs de p : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

## Résultats et discussion

L'analyse de corrélations nous montre trois variables corrélant significativement avec l'indice de santé des oiseaux (Fig. 1). Ces variables sont, par ordre d'importance, la densité de buissons (Cor = **0.405*****), le nombre de mangeoires (Cor = **0.281*****), et la présence des rapaces (Cor = **0.150***). Le nombre de poubelles a été testé dans les modèles au vu de sa significativité marginale (Cor = 0.145.). La distribution des indices de santé des oiseaux est gaussienne et centrée, tandis que les autres variables, à l'exception de la présence des rapaces (binomiale) présentent une kurtose plus (densité de routes, densité de buissons, nombre de mangeoires) ou moins forte (nombre de poubelles) à gauche de la distribution.

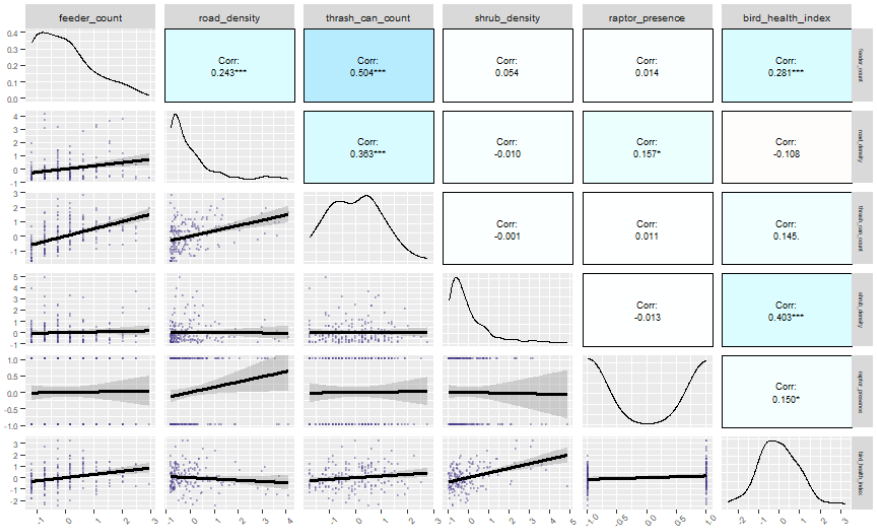


*Figure 1 : table des coefficients de corrélation entre chaque variable numérique du jeu de données (en haut à droite) ; régressions linéaires et distribution des points pour chaque variable 2 à 2 (en bas et à gauche) ; courbe de fréquence des données pour chaque variable (en diagonale).*

L'indice de santé des oiseaux semble dépendre de la densité de buissons (est.= **1.67*****, se = 0.28), le nombre de mangeoires (est.= **1.64*****, se = 0.48) et la présence de rapaces (est.= **5.13****, se = 1.9). L'effet aléatoire des parcs a été préféré à celui des villes par comparaison de la variance expliquée (Int.=7.62, se = 2.76 contre Int.=0.24, se = 0.49). Le facteur du nombre de poubelles a été rejeté par parcimonie, étant donné qu'il n'était pas significatif.

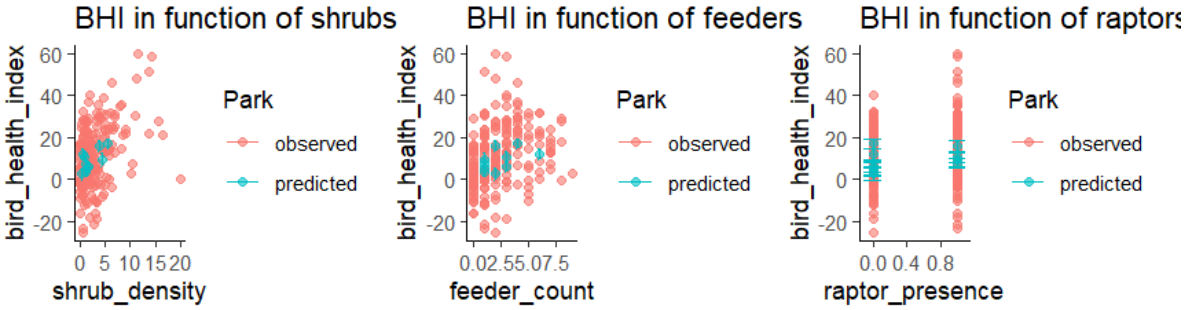


*Figure 2 : Indice de la santé des oiseaux en fonction de la densité de buissons (à gauche), du nombre de mangeoires (au centre) et de la présence des rapaces (à droite). Les valeurs observées sont en rouge, tandis que les valeurs prédites sont en bleu.*

Les valeurs prédites avec le modèle GLMM retenu sont parfaitement intégrées au nuage de données pour chacune des variables explicatives (Fig. 2). Les valeurs prédites sont concentrées au centre du nuage de points dans tous les cas, ce qui semble conservateur, car aucune des données prédites ne frôle la marge du nuage de points, mais aussi rassurant car autour de la moyenne des individus.

Les valeurs de santé prédites pour les individus sont comprises entre 2.51 et 17.11 (Tab.1).

*Table 1 : Valeurs de santé prédites pour chaque individu*

| Individu | Valeur de santé |
|---|---|
| 1 | 9.38 |
| 2 | 5.15 |
| 3 | **15.92** |
| 4 | 7.6 |
| 5 | 3.66 |
| 6 | 2.51 |
| 7 | **17.11** |
| 8 | 10.52 |
| 9 | 6.1 |
| 10 | 12.03 |