



## Prédiction de l'abondance et de la richesse totales des vers de terre

Présenté par : M. Abdou DIALLO

Encadrants: M. Walid HORRIGUE

M. Daniel CLUZEAU

M. Kevin HOEFFNER



TERRITOIRES  
D'INNOVATION



# Prédiction de l'abondance et de la richesse totales des vers de terre

## Sommaire

### 1 Présentation des données

### 2 Analyses exploratoires:

Pré-traitement des données, sélection des variables et partitionnement des données

### 3 Modèles utilisés :

- Modèles linéaires généralisés (GLM)
- Modèles additifs généralisés (GAM)
- Random Forest (RF)
- Gradient Boosting Machine (GBM)
- Réseau de neurones artificiels (ANN)

### 4 Résultats & discussion

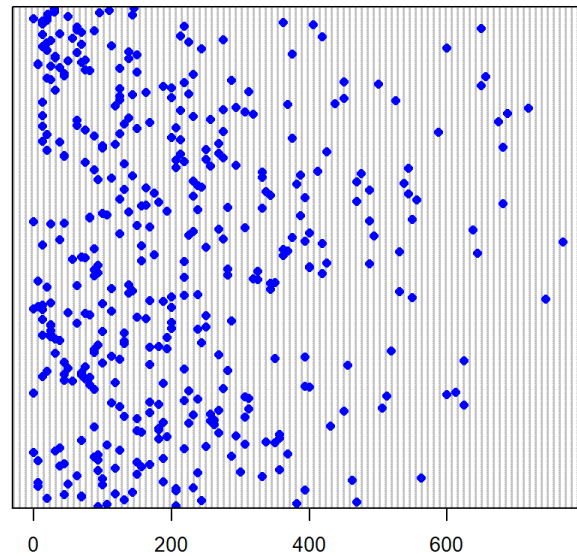




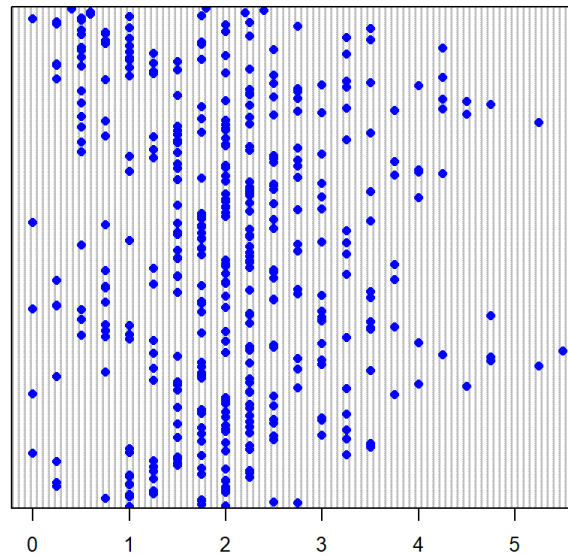
- Les variables à expliquer

Le dataset comportait **386 observations** pour **153 colonnes**

**Abondance totale (ind. /m<sup>2</sup>)**



**Richesse totale (nb. sp. par site)**

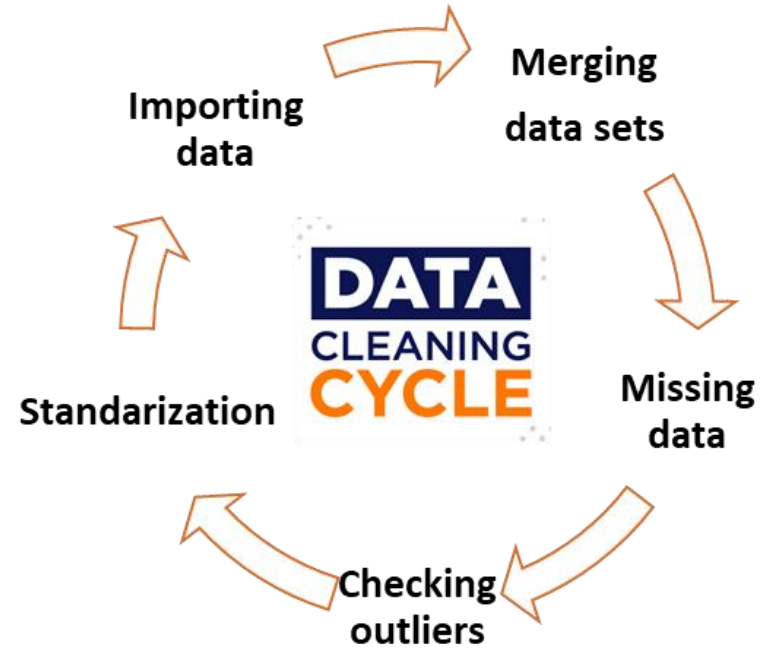


## ■ Les variables explicatives



Variables	Descriptions
GPS_X	Coordonnée GPS X
GPS_Y	Coordonnée GPS Y
pH_eau	pH du sol dans l'eau
SableF	Fraction fine de sable
SableG	Fraction grossière de sable
LimonF	Fraction fine de limon
LimonG	Fraction grossière de limon
Argile	Teneur en argile
Sables	Teneur en sables
Limons	Teneur en limons
MO	Matière organique
C_tot	Carbone total
C_org	Carbone organique
C.N	Ratio Carbone/Azote
N_tot	Azote total
Details_Milieu_Niv3	Occupation du sol

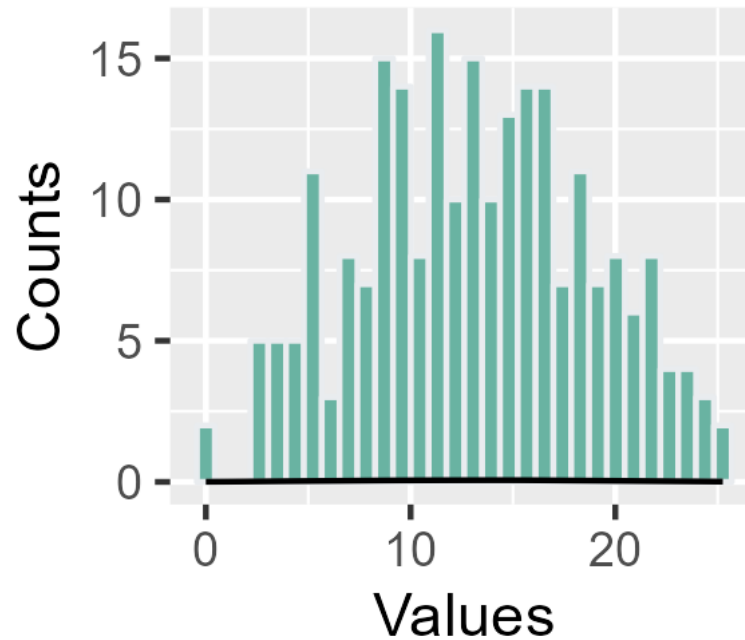




- Train data (80 %)  
307 observations



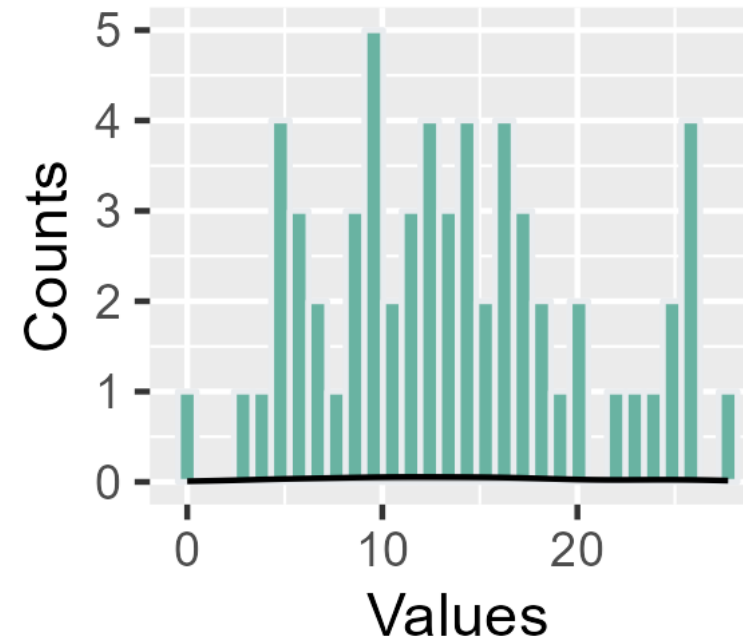
**(a)** Abundance: Train



- Test data (20 %)  
79 observations



**(b)** Abundance: Test

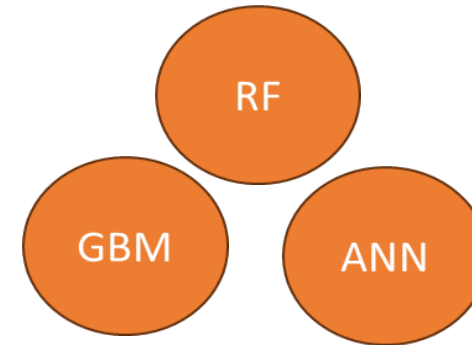


## Variables considérées dans les modèles de prédiction

Variables	Descriptions
GPS_X	Coordonnée GPS X
GPS_Y	Coordonnée GPS Y
pH_eau	pH du sol dans l'eau
SableF	Fraction fine de sable
SableG	Fraction grossière de sable
LimonF	Fraction fine de limon
LimonG	Fraction grossière de limon
Argile	Teneur en argile
C_org	Carbone organique
C.N	Ratio Carbone/Azote
Details_Milieu_Niv3	Occupation du sol
Sables	Teneur en sables
Limons	Teneur en limons
MO	Matière organique
C_tot	Carbone total
N_tot	Azote total



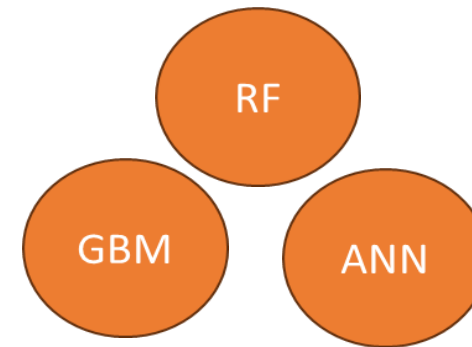
## Algorithmes d'apprentissage automatique



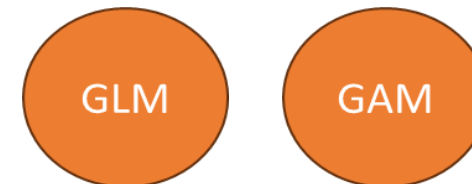
Variables	Descriptions
GPS_X	Coordonnée GPS X
GPS_Y	Coordonnée GPS Y
pH_eau	pH du sol dans l'eau
SableF	Fraction fine de sable
SableG	Fraction grossière de sable
LimonF	Fraction fine de limon
LimonG	Fraction grossière de limon
Argile	Teneur en argile
C_org	Carbone organique
C.N	Ratio Carbone/Azote
Details_Milieu_Niv3	Occupation du sol
Sables	Teneur en sables
Limons	Teneur en limons
MO	Matière organique
C_tot	Carbone total
N_tot	Azote total



## Algorithmes d'apprentissage automatique



## Algorithmes de régression traditionnels





## ☐ Abondance totale

$$\overline{m}_{test} = 218 \text{ ind./m}^2$$

Modèles	R2_adj_train	R2_test	RMSE
Modèles linéaires généralisés	0.09	0.01	166
Modèles additifs généralisés	0.26	0.08	155
<b>Random Forest</b>	<b>0.4</b>	<b>0.11</b>	<b>152</b>
Gradient Boosting Machine	0.29	0.10	156
Réseau de neurones artificiels	0.00	0.08	248



- $R^2$  ajusté du modèle
- $R^2$  (Coefficient de détermination d'une régression)
- RMSE (Erreur Quadratique Moyenne)



## ☐ Richesse totale

$$\overline{m}_{test} = 2.1 \text{ sp./site}$$

Modèles	R2_adj_train	R2_test	RMSE
Modèles linéaires généralisés	0.18	0.11	0.92
Modèles additifs généralisés	0.32	0.14	0.93
<b>Random Forest</b>	<b>0.33</b>	<b>0.18</b>	<b>0.91</b>
Gradient Boosting Machine	0.35	0.14	0.92
Réseau de neurones artificiels	-0.10	0.16	0.96



- $R^2$  ajusté du modèle
- $R^2$  (Coefficient de détermination d'une régression)
- RMSE (Erreur Quadratique Moyenne)





## ☐ Abondance totale

$$\overline{m}_{test} = 14.76 \text{ ind./m}^2$$

Modèles	R2_adj_train	R2_test	RMSE
Modèles linéaires généralisés	0.12	0.07	6.75
Modèles additifs généralisés	0.21	0.06	6.77
<b>Random Forest</b>	<b>0.52</b>	<b>0.19</b>	<b>6.33</b>
Gradient Boosting Machine	0.29	0.16	6.4
Réseau de neurones artificiels	-0.07	0.04	7.16



- $R^2$  ajusté du modèle
- $R^2$  (Coefficient de détermination d'une régression)
- RMSE (Erreur Quadratique Moyenne)





## ☐ Richesse totale

$$\overline{m}_{test} = 2.1 \text{ sp./site}$$

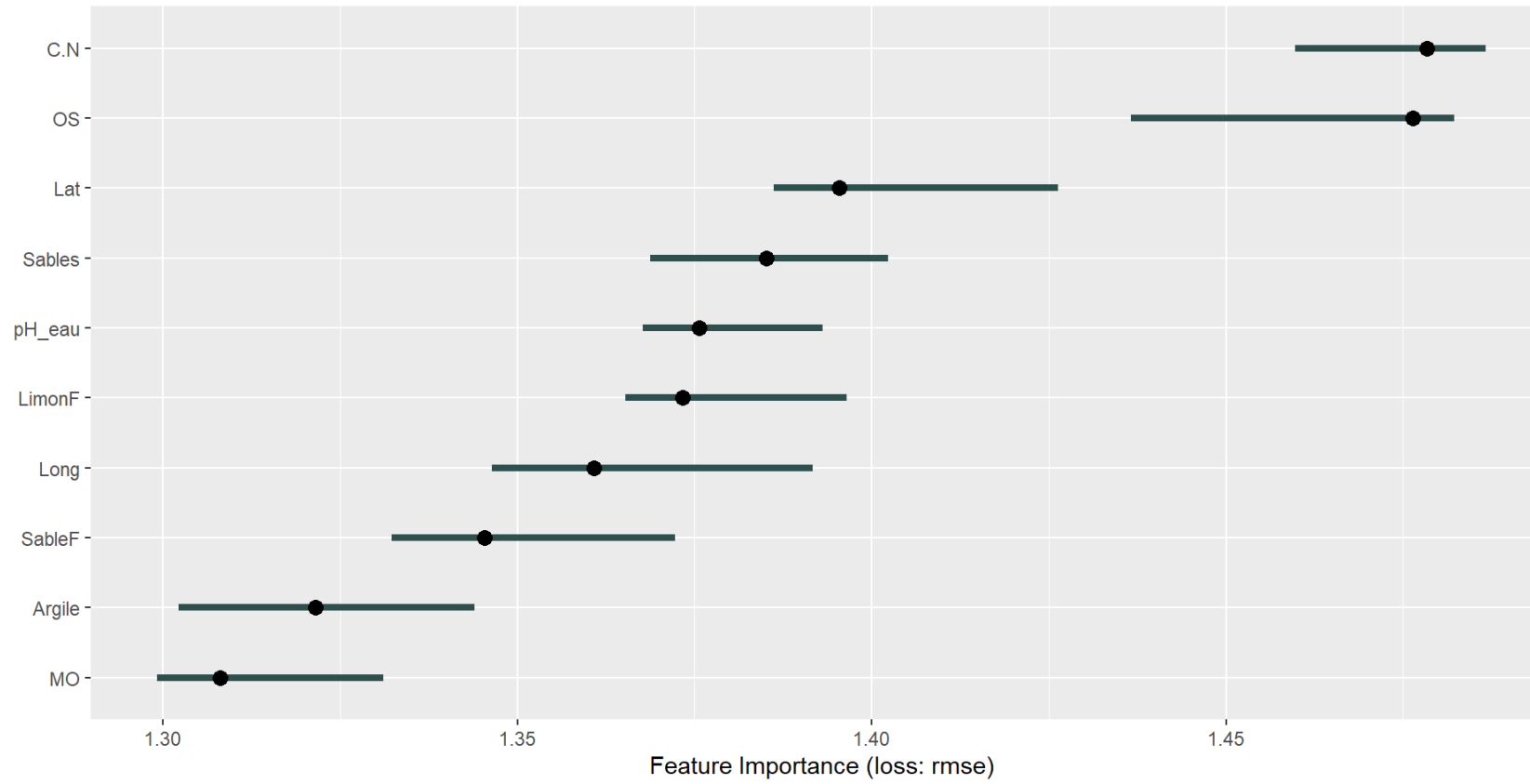
Modèles	R2_adj_train	R2_test	RMSE
Modèles linéaires généralisés	0.13	0.10	0.95
Modèles additifs généralisés	0.21	0.18	0.83
<b>Random Forest</b>	<b>0.53</b>	<b>0.26</b>	<b>0.82</b>
Gradient Boosting Machine	0.30	0.24	0.82
Réseau de neurones artificiels	-0.13	0.13	0.96



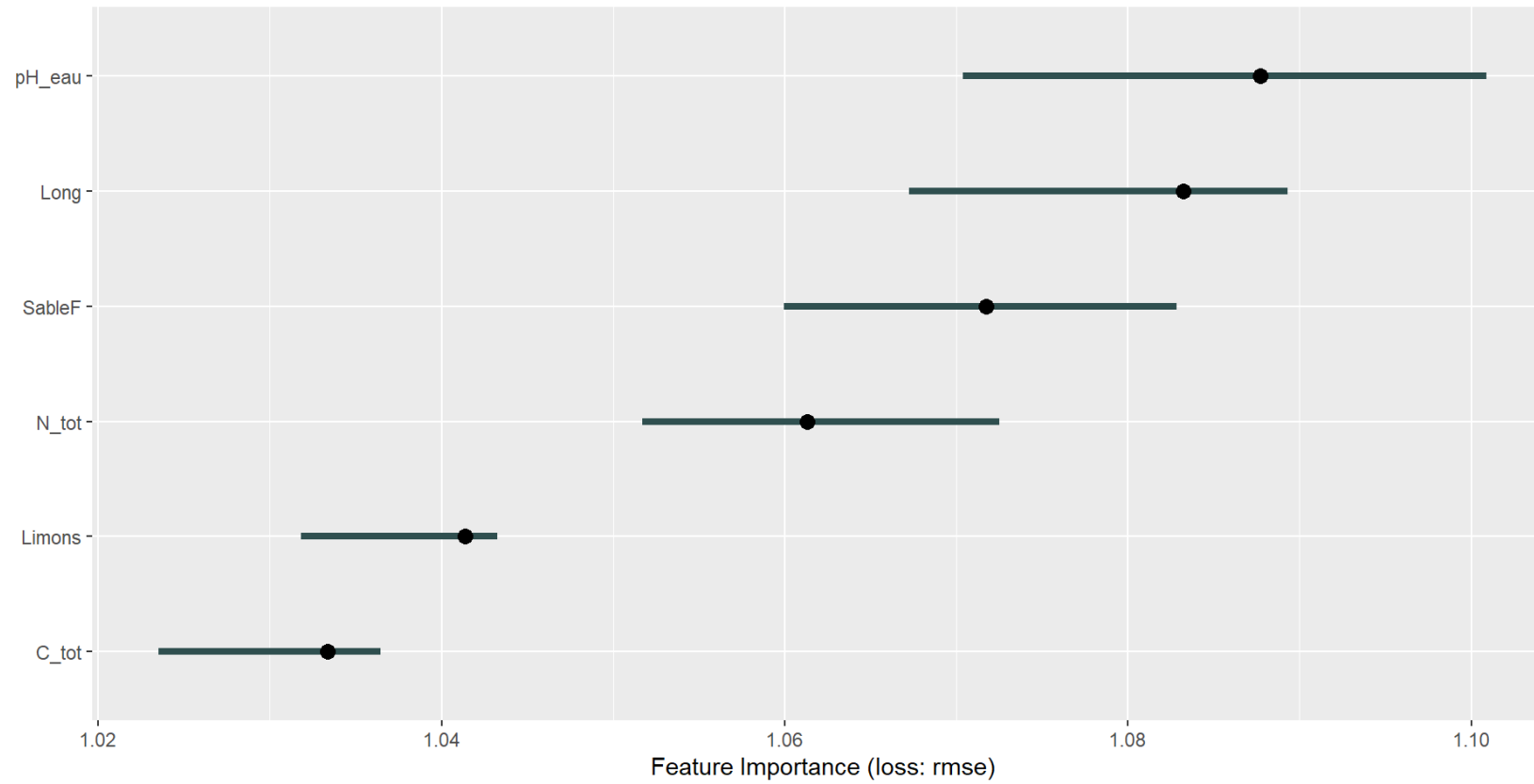
- $R^2$  ajusté du modèle
- $R^2$  (Coefficient de détermination d'une régression)
- RMSE (Erreur Quadratique Moyenne)



## □ Abondance: importances des variables



## ☐ Richesse: importances des variables



- ❖ Peu de variabilité dans les données
- ❖ Amélioration des modèles
- ❖ Références & diagnostics



# Merci !

ProDij.  
mieux manger, mieux produire



[metropole-dijon.fr](http://metropole-dijon.fr)







## Prédiction de l'abondance et de la richesse totales des vers de terre

Présenté par : M. Abdou DIALLO

Encadrants: M. Walid HORRIGUE

M. Daniel CLUZEAU

M. Kevin HOFFNER



TERRITOIRES  
D'INNOVATION



[https://rpubs.com/Abdou diallo/12026](https://rpubs.com/Abdou_diallo/12026)  
37

## The permutation feature importance algorithm based on Fisher, Rudin, and Dominici (2018):

Input: Trained model  $\hat{f}$ , feature matrix  $X$ , target vector  $y$ , error measure  $L(y, \hat{f})$ .


1. Estimate the original model error  $e_{orig} = L(y, \hat{f}(X))$  (e.g. mean squared error)
2. For each feature  $j \in \{1, \dots, p\}$  do:
  - Generate feature matrix  $X_{perm}$  by permuting feature  $j$  in the data  $X$ . This breaks the association between feature  $j$  and true outcome  $y$ .
  - Estimate error  $e_{perm} = L(Y, \hat{f}(X_{perm}))$  based on the predictions of the permuted data.
  - Calculate permutation feature importance as quotient  $FI_j = e_{perm}/e_{orig}$  or difference

$$FI_j = e_{perm} - e_{orig}$$

3. Sort features by descending FI.

Show  entries

Search:

Numbers 

111_Forêt de feuillus	75
210_Prairie agricole permanente	55
214_Culture annuelle	230
218_Vignes et autres Cultures pérennes	56

Showing 1 to 4 of 4 entries

Previous

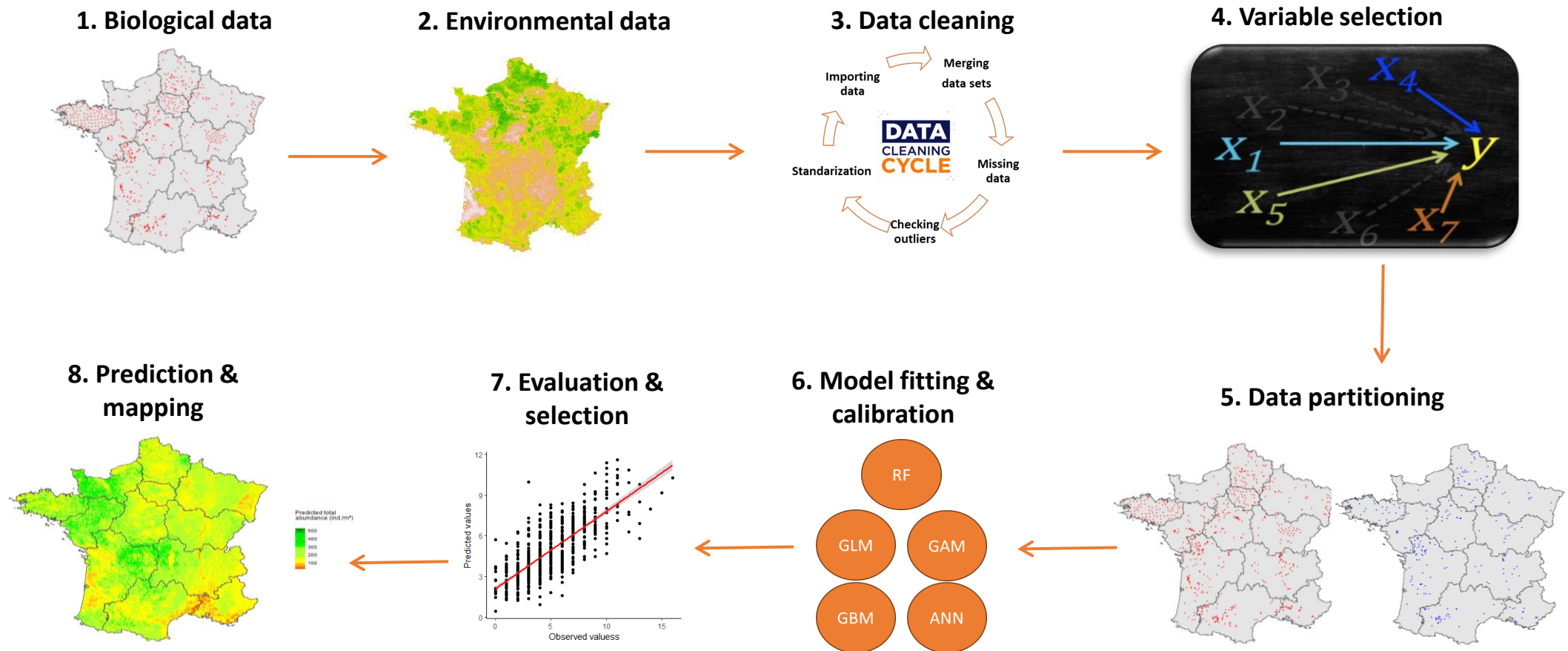
1

Next

# Materials and methods

## Modeling strategy

**ODMAP protocol:** Overview, Data, Model, Assessment, Prediction (zurell et al., 2020)



# Results and discussions

## Superiority of ensemble methods in predicting earthworm communities compared to traditional regression models

Algorithms	Response variables	R <sup>2</sup>	RMSE
GLM	Total abundance	0.22	34.57
GAM		0.26	33.06
<b>RF</b>		<b>0.43</b>	<b>25.20</b>
GBM		0.43	25.30
ANN		0.35	28.94
GLM	Total biomass	0.23	10.69
GAM		0.24	10.50
<b>RF</b>		<b>0.35</b>	<b>8.76</b>
GBM		0.32	9.30
ANN		0.27	10.50
GLM	Total taxonomic richness	0.36	2.18
GAM		0.44	2.04
<b>RF</b>		<b>0.59</b>	<b>1.75</b>
<b>GBM</b>		<b>0.59</b>	<b>1.75</b>
ANN		0.40	2.16

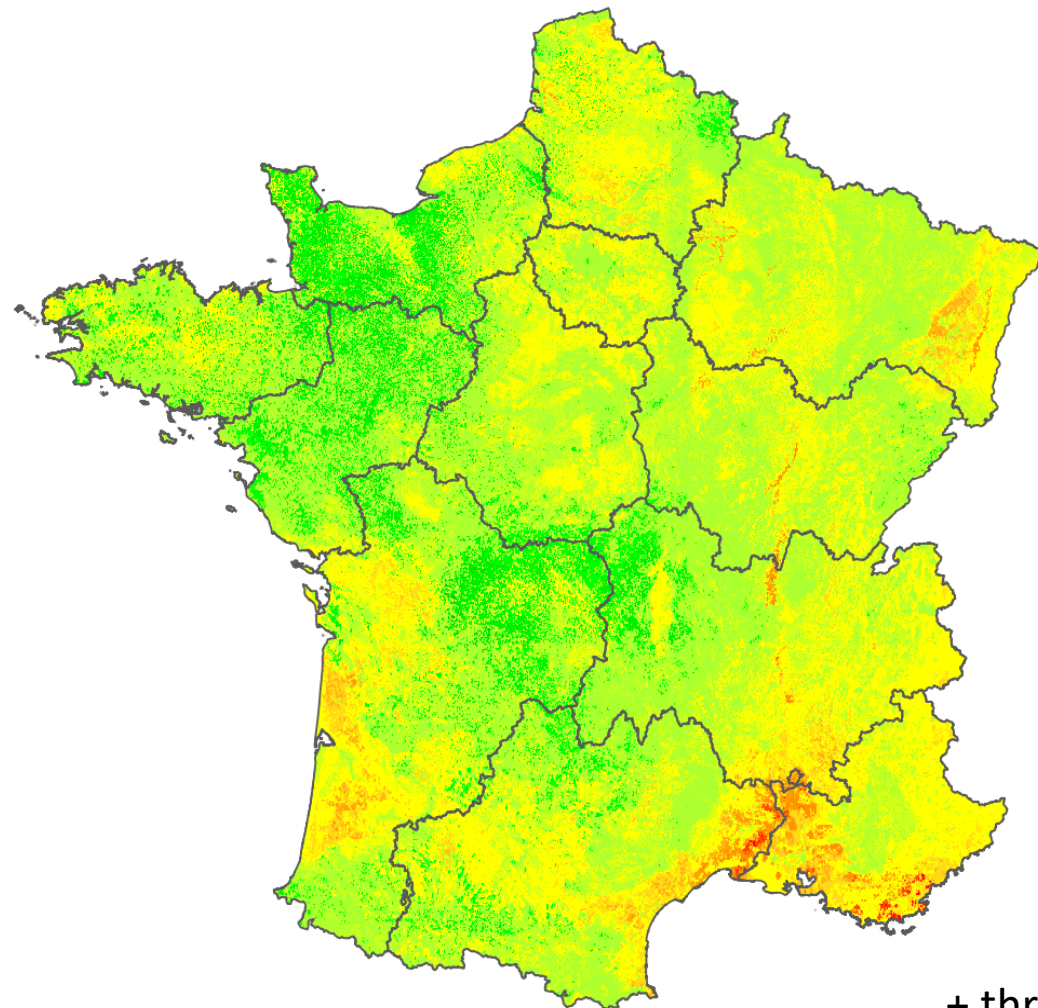
- Ensemble methods (Breiman, 2001; Li & Wang, 2013)
- Better captures nonlinear relationships (Breiman, 2001)

Require large amounts of data (Yiu, 2021)

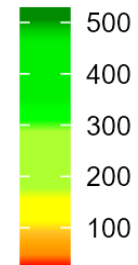
Poor interpretability

# Results and discussions

## Predicted spatial distribution of earthworm total abundance



Predicted total  
abundance (ind./m<sup>2</sup>)



Earthworm community was more abundant  
in the northwest and center of France

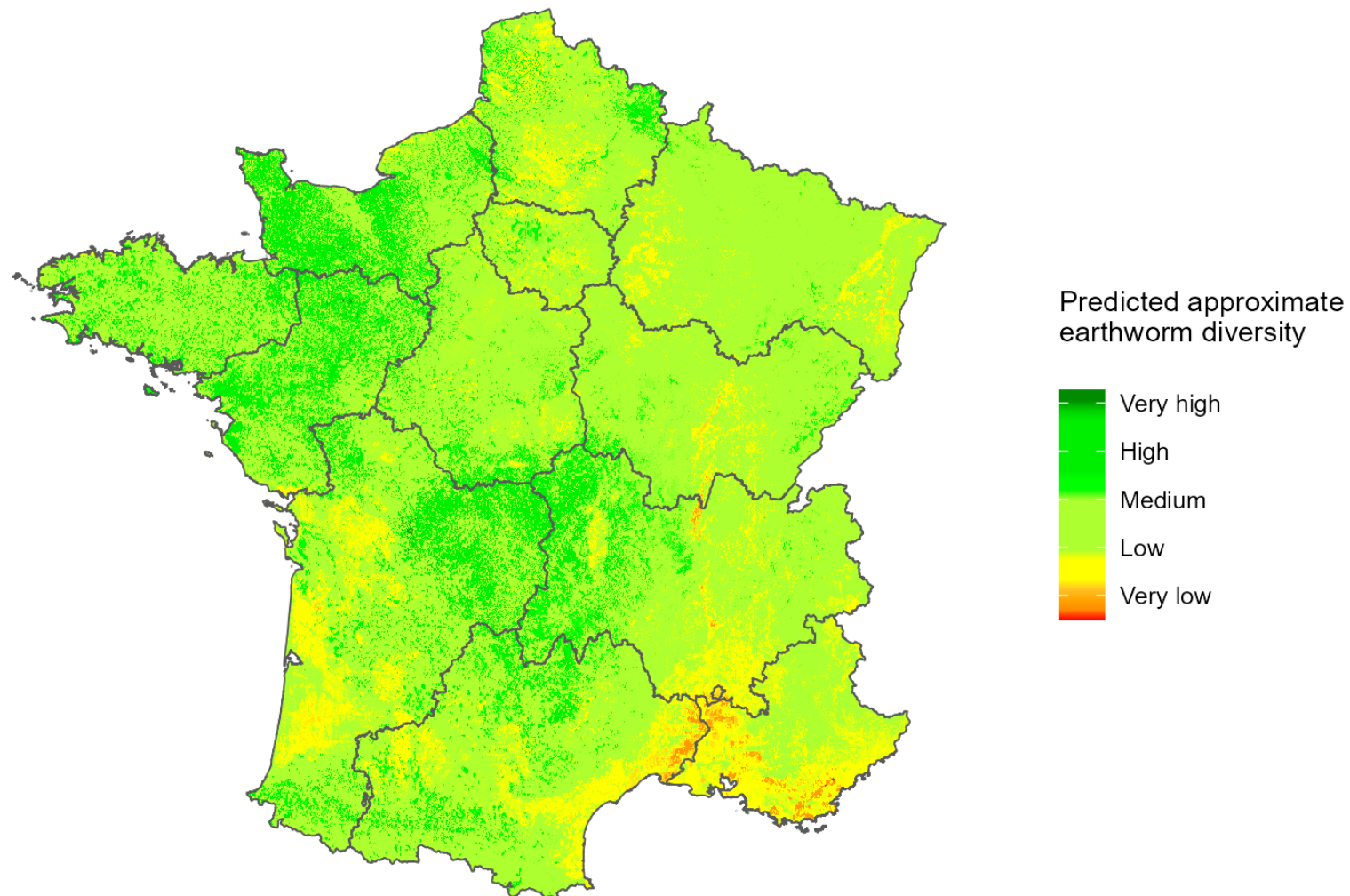
- Land use
- Climate
- Soil

+ three more maps!



# Results and discussions

## Predicted spatial distribution of earthworm diversity

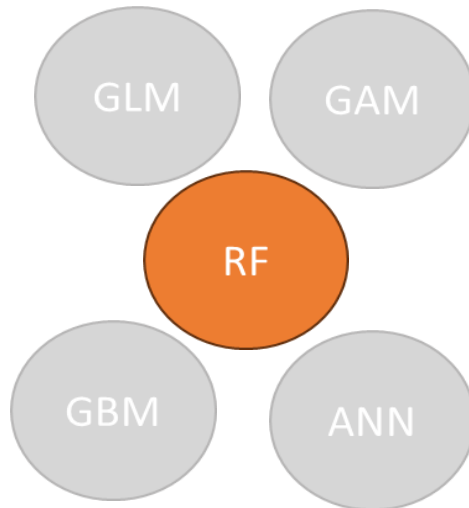




## 6. Model fitting and calibration

Default model

### Machine learning algorithms



### Random Forests (RF)

```
Y3 = randomForest (data[-rep.var], data[[rep.var]],  
                    importance = TRUE)
```

RF model tuning by grid

- *ntree* = 100 to 2000 in increments of 200
- *mtry* = 2 to 10 in increments of 1
- *maxnodes* = NULL and 2 to 15 in increments of 1

```
Y3 = randomForest (data[-rep.var], data[[rep.var]], mtry = 3,  
                    ntree = 500, maxnodes = NULL, importance = TRUE)
```

## 6. Model fitting and calibration

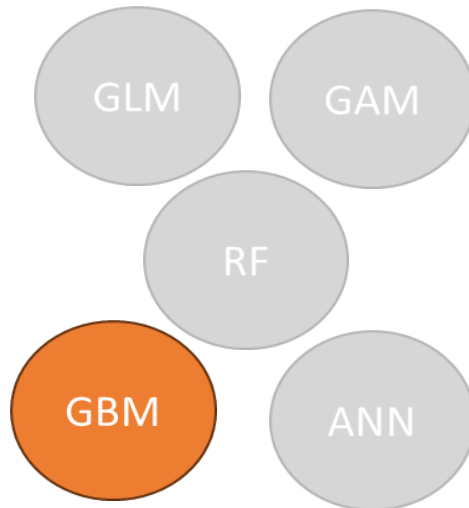
Default model

$$Y4 = gbm(y \sim ., data = data, distribution = 'gaussian')$$

GBM model tuning by grid

- $n.trees = 500$  to  $2000$  in increments of  $100$
- $shrinkage = 0.01, 0.02, 0.05, 0.001, 0.002$  and  $0.005$
- $interaction.depth = 1, 3, 5, 6, 8$  and  $10$
- $n.minobsinnode = 2, 5, 10, 20, 30$  and  $50$

### Machine learning algorithms

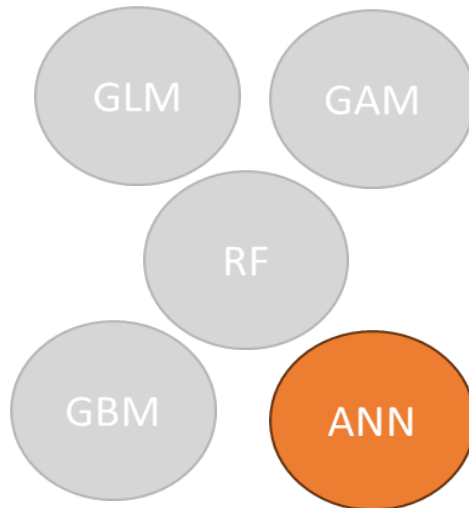


### Generalized Boosted Models (GBM)

$$Y4 = gbm(y \sim ., data = data, distribution = 'gaussian', n.trees = 1000, shrinkage = 0.01, interaction.depth = 5, n.minobsinnode = 10)$$

## 6. Model fitting and calibration

### Machine learning algorithms



### Artificial Neural Networks (ANN)

### Tunning

```
runs = tuning_run("Experiment.R", flags = list(dense_units1 = c(64, 32),  
dense_units2 = c(16, 32),  
dense_units3 = c(8, 16),  
dense_units4 = c(4, 8),  
dropout1 = c(0.4, 0.5),  
dropout2 = c(0.3, 0.4),  
dropout3 = c(0.2, 0.3),  
dropout4 = c(0.1, 0.2),  
batch_size = c(32, 64)))
```

