Sélection de variables dans un modèle de Poisson Log-Normal (PLN)

Application à l'étude des communautés microbiennes dans le processus de production du lait

KIOYE J. Y.[1]
GROLLEMUND P. M.[1,2]
CHASSARD C.[1]; CHAUVET J.[3]

UMRF[1], LMBP[2], LARIS[3]

22 juin 2023

Contexte et Motivations

Comprendre ce qui sous-tend la qualité du lait

- Qualité sensorielle et composition biochimique
- Biodiversité prairiale et pratiques d'élevage
- Lien avec les différentes communautés microbiennes

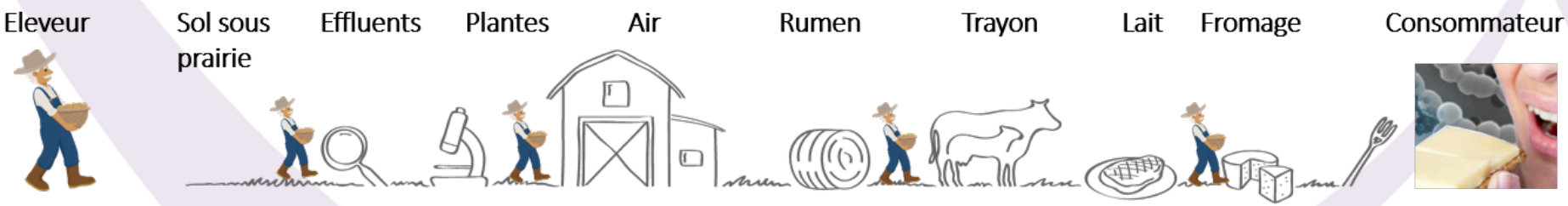
Contexte et Motivations

Comprendre ce qui sous-tend la qualité du lait

- Qualité sensorielle et composition biochimique
- Biodiversité prairiale et pratiques d'élevage
- Lien avec les différentes communautés microbiennes

Amélioration des approches à l'échelle du système agri/agroalimentaire

- Impact des pratiques d'élevage
- Les flux microbiens d'amont en aval
- Identification des facteurs déterministes



Modélisation statistique

- Étudier les abondances conjointes des bactéries
- Évaluer l'intensité des facteurs environementaux
- Identifier des interactions entre bactéries

Modèle de Poisson Log Normal (PLN)

Le modèle PLN [1] : cas particulier de modèle lineaire géneralisé

$$\begin{aligned} \boldsymbol{Y}_i \mid \boldsymbol{Z}_i &\sim \mathcal{P}\big(\exp(\boldsymbol{Z}_i)\big) && \text{(espace observé)} \\ \boldsymbol{Z}_i &\sim N_p\big(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big) && \text{(espace latent)} \end{aligned}$$

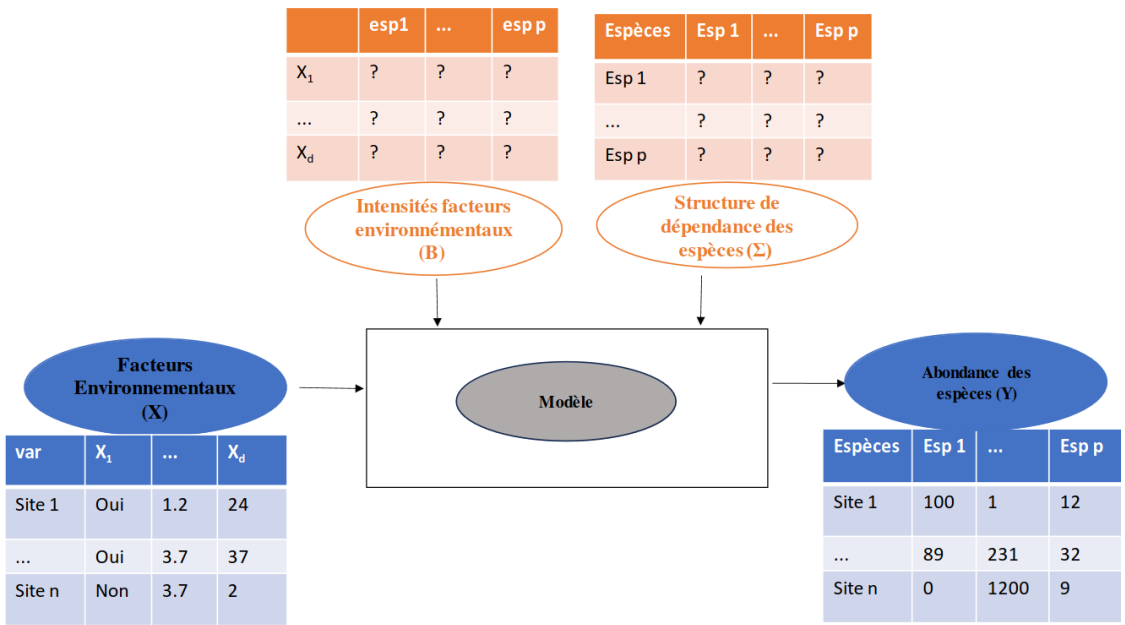
1. Julien Chiquet, Mahendra Mariadassou et Stéphane Robin. « The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances ». In : *Frontiers in Ecology and Evolution* 9 (2021), p. 588292

Modèle de Poisson Log Normal (PLN)

Le modèle PLN [1] : cas particulier de modèle lineaire géneralisé

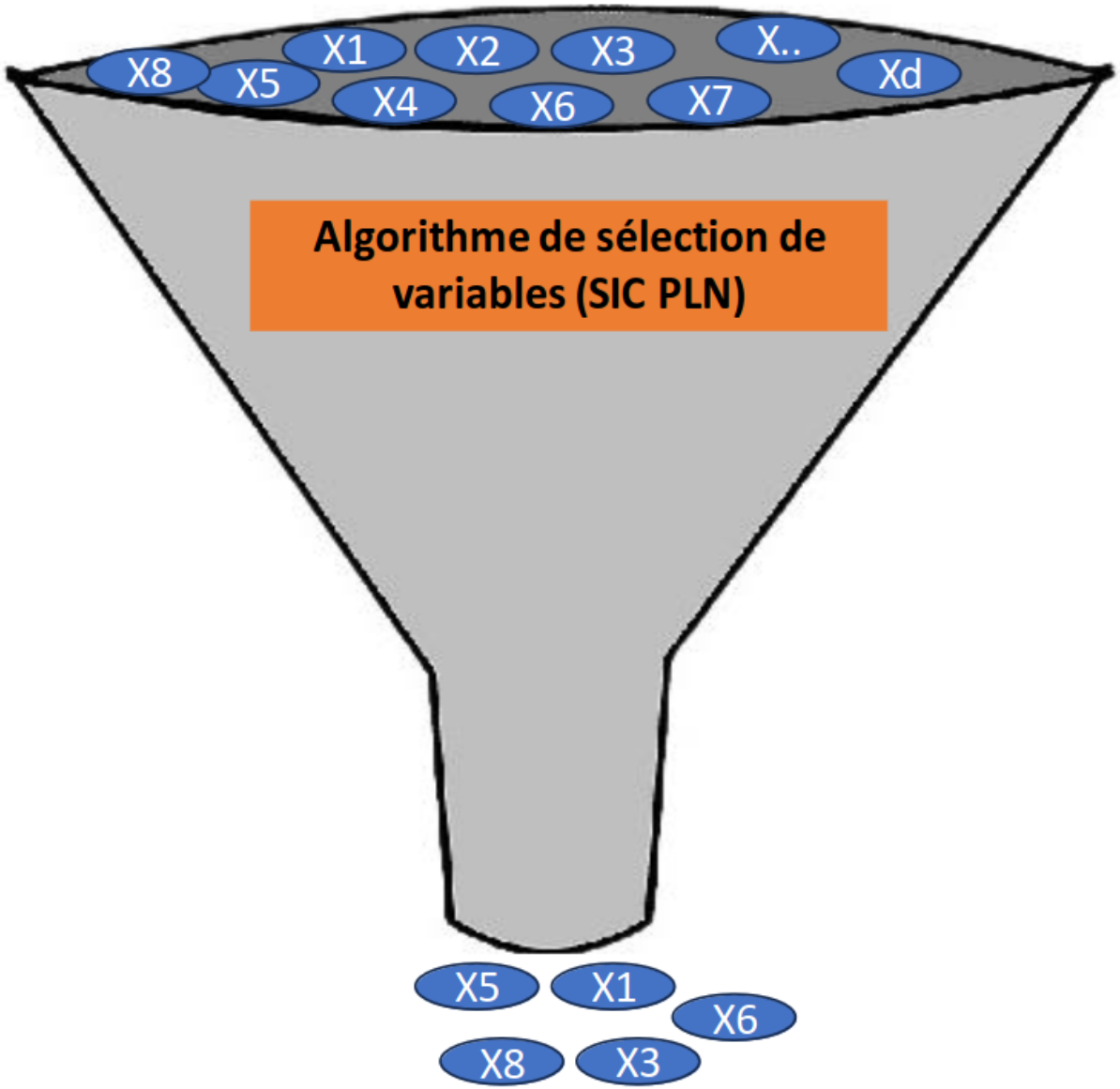
$$\boldsymbol{Y}_i \mid \boldsymbol{Z}_i \sim \mathcal{P}\big(\exp(\boldsymbol{Z}_i)\big) \qquad \text{(espace observé)}$$

$$\boldsymbol{Z}_i \sim N_p\big(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big) \qquad \text{(espace latent)}$$



1. Julien Chiquet, Mahendra Mariadassou et Stéphane Robin. « The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances ». In : *Frontiers in Ecology and Evolution* 9 (2021), p. 588292

Quelles sont les variables environnementales pertinentes ?



Sélection de variables

Plusieurs méthodes existent :

- Sélection du meilleur sous-ensemble : forward-backward, setpwise, etc.
- Couteux sur le plan calculatoire
- Sélection de modèle : AIC, BIC, etc.

2. Meadhbh O'Neill et Kevin Burke. « Variable selection using a smooth information criterion for distributional regression models ». In : *Statistics and Computing* 33.3 (2023), p. 71

Sélection de variables

Plusieurs méthodes existent :

- Sélection du meilleur sous-ensemble : forward-backward, setpwise, etc.
- Couteux sur le plan calculatoire
- Sélection de modèle : AIC, BIC, etc.

Question pas simple : approches modernes

- Méthode de régularisation : optimisation sous contrainte
- Contraintes relaxées : lasso, ridge, elastic-net, etc.
- Calibrage d'un paramètre de régularisation

2. Meadhbh O'Neill et Kevin Burke. ≪ Variable selection using a smooth information criterion for distributional regression models ≫. In : *Statistics and Computing* 33.3 (2023), p. 71

Sélection de variables

Plusieurs méthodes existent :

- Sélection du meilleur sous-ensemble : forward-backward, setpwise, etc.
- Couteux sur le plan calculatoire
- Sélection de modèle : AIC, BIC, etc.

Question pas simple : approches modernes

- Méthode de régularisation : optimisation sous contrainte
- Contraintes relaxées : lasso, ridge, elastic-net, etc.
- Calibrage d'un paramètre de régularisation
- Récente contribution : smooth information criterion (SIC) [2]

2. Meadhbh O'Neill et Kevin Burke. « Variable selection using a smooth information criterion for distributional regression models ». In : *Statistics and Computing* 33.3 (2023), p. 71

Smooth Information Criterion (SIC)

- Approxime le vrai problème
- Évite de calibrer un paramètre de régularisation par validation croisée
- Ne nécessitent pas l'ajustement d'un très grand nombre de modèles
- Computationnellement moins couteux

Algorithme proposé : Couplage SIC et PLN

Algorithme 1 : Poisson Log Normal (PLN)

1: Initialisation des paramètres
2: **répéter**
    Optimisation des paramètres
    jusqu'à convergence;
3: Mise à jour des paramètres
4: Retourne les meilleurs paramètres

Algorithme proposé : Couplage SIC et PLN

Algorithme 1 : Poisson Log Normal (PLN)

1: Initialisation des paramètres
2: **répéter**
    Optimisation des paramètres
    jusqu'à convergence;
3: Mise à jour des paramètres
4: Retourne les meilleurs paramètres

Algorithme 2 : Smooth Information Criterion (SIC)

1: Initialisation : objectif, paramètres
2: Définir une séquence de valeurs
3: **Pour** *chaque valeur de la séquence*
    Optimisation sous contrainte
4: Retourne les paramètres filtrés

Algorithme proposé : Couplage SIC et PLN

Algorithme 1 : Poisson Log Normal (PLN)

1: Initialisation des paramètres
2: **répéter**
    Optimisation des paramètres
    jusqu'à convergence;
3: Mise à jour des paramètres
4: Retourne les meilleurs paramètres

Algorithme 2 : Smooth Information Criterion (SIC)

1: Initialisation : objectif, paramètres
2: Définir une séquence de valeurs
3: **Pour** *chaque valeur de la séquence*
    Optimisation sous contrainte
4: Retourne les paramètres filtrés

Algorithme 3 : Couplage SIC et PLN

1: Initialisation : fonction objective, paramètres
2: Définir une séquence de valeurs
3: **Pour** *chaque valeur de la séquence*
    Resoudre un problème PLN complexe
4: Retourne les paramètres filtrés

Expérimentation sur des données simulées

Processus de géneration :

- Générer des variables environnementales ($n = 10000$, $d = 6$)
- Considérer des intensités nulles (0) pour certaines variables
- Considérer des intensités moyennes (0.5) pour certaines variables
- Considérer des intensités fortes (1) pour certaines variables
- Générer des données de comptage suivant le modèle PLN ($(n = 10000,\ p = 4)$

Expérimentation sur des données simulées

Processus de géneration :

- Générer des variables environnementales ($n = 10000$, $d = 6$)
- Considérer des intensités nulles (0) pour certaines variables
- Considérer des intensités moyennes (0.5) pour certaines variables
- Considérer des intensités fortes (1) pour certaines variables
- Générer des données de comptage suivant le modèle PLN (($n = 10000$, $p = 4$)

Objectif :

- Mettre l'intensité des variables non active à zéro
- Minimisé les erreurs des intensités estimées

Résultat sur des données simulées

Table – Vraies intensités (estimées avec PLN)

	espèce 1	espèce 2	espèce 3	espèce 4
x_1	0 (0.159)	0.5 (0.546)	1 (1.120)	1 (1.048)
x_2	1 (1.107)	0 (0.161)	0.5 (0.559)	1 (1.007)
x_3	1 (1.143)	0 (0.089)	0.5 (0.649)	0 (0.026)
x_4	1 (1.148)	1 (1.037)	1 (1.111)	0 (0.098)
x_5	1 (1.136)	1 (1.034)	1 (1.127)	0.5 (0.571)
x_6	0 (0.098)	0 (0.096)	0 (0.090)	0 (0.095)

Résultat sur des données simulées

Table – Vraies intensités (estimées avec PLN)

	espèce 1	espèce 2	espèce 3	espèce 4
x_1	0 (0.159)	0.5 (0.546)	1 (1.120)	1 (1.048)
x_2	1 (1.107)	0 (0.161)	0.5 (0.559)	1 (1.007)
x_3	1 (1.143)	0 (0.089)	0.5 (0.649)	0 (0.026)
x_4	1 (1.148)	1 (1.037)	1 (1.111)	0 (0.098)
x_5	1 (1.136)	1 (1.034)	1 (1.127)	0.5 (0.571)
x_6	0 (0.098)	0 (0.096)	0 (0.090)	0 (0.095)

Table – Vraies intensités (intensités estimées avec SIC PLN)

	espèce 1	espèce 2	espèce 3	espèce 4
x_1	0 (0.059)	0.5 (0.446)	1 (1.020)	1 (0.948)
x_2	1 (1.006)	0 (0.061)	0.5 (0.459)	1 (0.907)
x_3	1 (1.043)	0 (0)	0.5 (0.549)	0 (0)
x_4	1 (1.048)	1 (0.937)	1 (1.011)	0 (0)
x_5	1 (1.036)	1 (0.934)	1 (1.027)	0.5 (0.471)
x_6	0 (0)	0 (0)	0 (0)	0 (0)

Erreur d'estimation

- Erreur d'estimation des intensités avec PLN $\widehat{\boldsymbol{B}}$

$$\frac{\|\boldsymbol{B} - \widehat{\boldsymbol{B}}\|_F}{\|\boldsymbol{B}\|_F} = 0.136$$

Erreur d'estimation

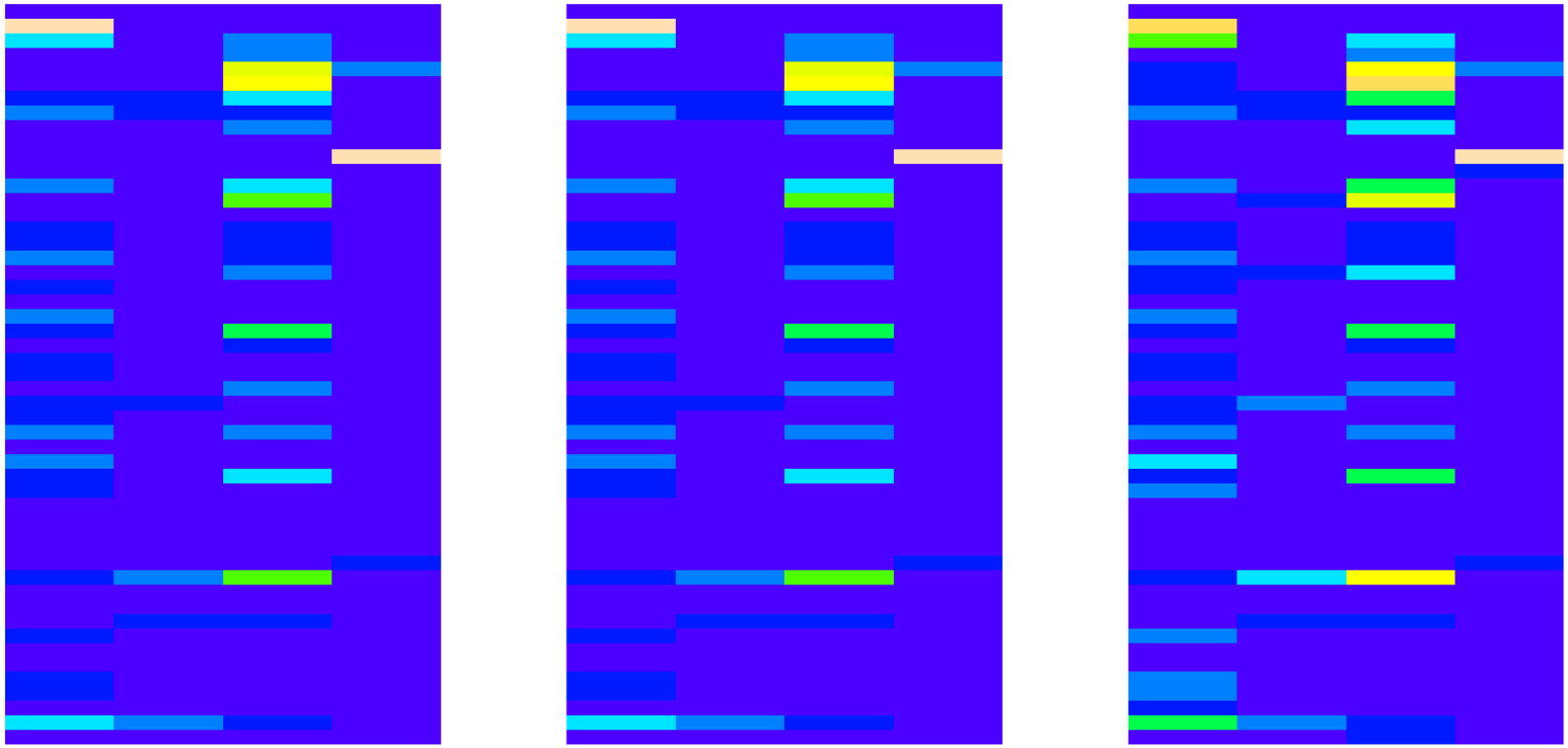
- Erreur d'estimation des intensités avec PLN $\widehat{\boldsymbol{B}}$

$$\frac{\|\boldsymbol{B} - \widehat{\boldsymbol{B}}\|_F}{\|\boldsymbol{B}\|_F} = 0.136$$

- Erreur d'estimation des intensités avec SIC PLN $\widehat{\boldsymbol{B}}$

$$\frac{\|\boldsymbol{B} - \widehat{\boldsymbol{B}}\|_F}{\|\boldsymbol{B}\|_F} = 0.058$$

Qualité de prédiction



Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

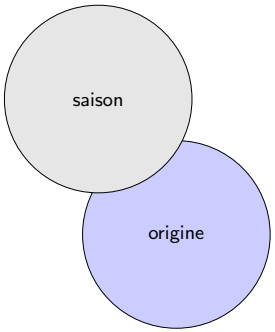
- 12 variables catégorielles
- Les plus pertinentes

Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

- 12 variables catégorielles
- Les plus pertinentes

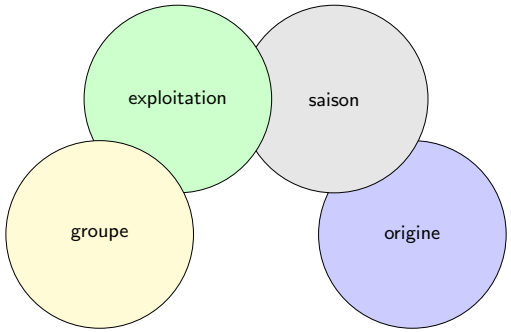


Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

- 12 variables catégorielles
- Les plus pertinentes

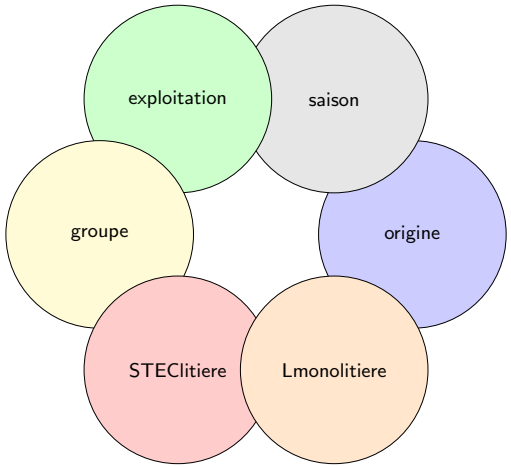


Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

- 12 variables catégorielles
- Les plus pertinentes

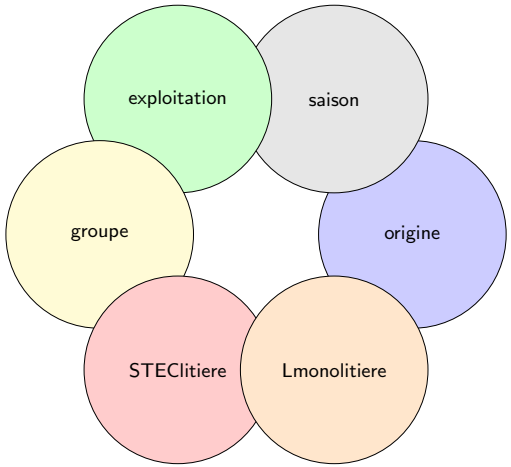


Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

- 12 variables catégorielles
- Les plus pertinentes



Autres données

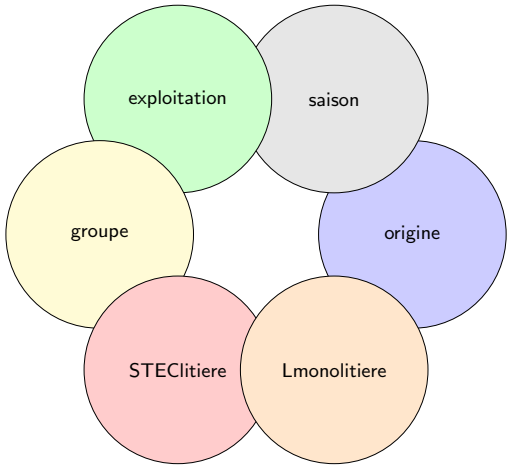
- Projet MINDS : diversité botanique
- Projet TANDEM : pratiques agricoles

Données de microbiologie (UMRF) en cours

Projet Amont Saint-Nectaire

- 536 observations de l'abondance de 1458 bactéries
- Abondances : entre 0 et 39671

- 12 variables catégorielles
- Les plus pertinentes



Autres données

- Projet MINDS : diversité botanique
- Projet TANDEM : pratiques agricoles

Quelles sont les variables environnementales et les pratiques agricoles qui expliquent les abondances ?

Merci pour votre attention ! ! !

"Le choix des variables est l'essence même de l'art de la modélisation."
George E. P. Box