École nationale
de la statistique
et de l'analyse
de l'information

Quelles circonstances sont associées à l'expression de signes cliniques respiratoires des porcs en croissance en élevages alternatifs ?

Étudiants :

Antoni Guàrdia Sanz
Clément Yvernault-Collet
Marius Vinclair
Xavier Braquaval

Tutrice :

Christelle Fablet

18 février 2025

Table des matières

1 Introduction

1.1 Introduction du sujet

1.2 Description de la base de données

2 Etude des variables respiratoires

Dans cette section, on se propose d'analyser les relations entre les variables caractérisant la respiration des porcs, en particulier les fréquences d'éternuements et de toux en engraissement et post-sevrage, afin d'identifier les liens existants. L'objectif est de regrouper ces variables et de simplifier l'étude en réduisant leur complexité. Les analyses sont menées en omettant les variables manquantes.

2.1 Présentation et caractérisation des variables respiratoires

Dans cette partie, nous présentons et caractérisons les variables respiratoires mesurées chez les porcs. L'objectif est de mieux comprendre la répartition de ces données et d'identifier d'éventuelles particularités dans le comportement respiratoire des porcs.

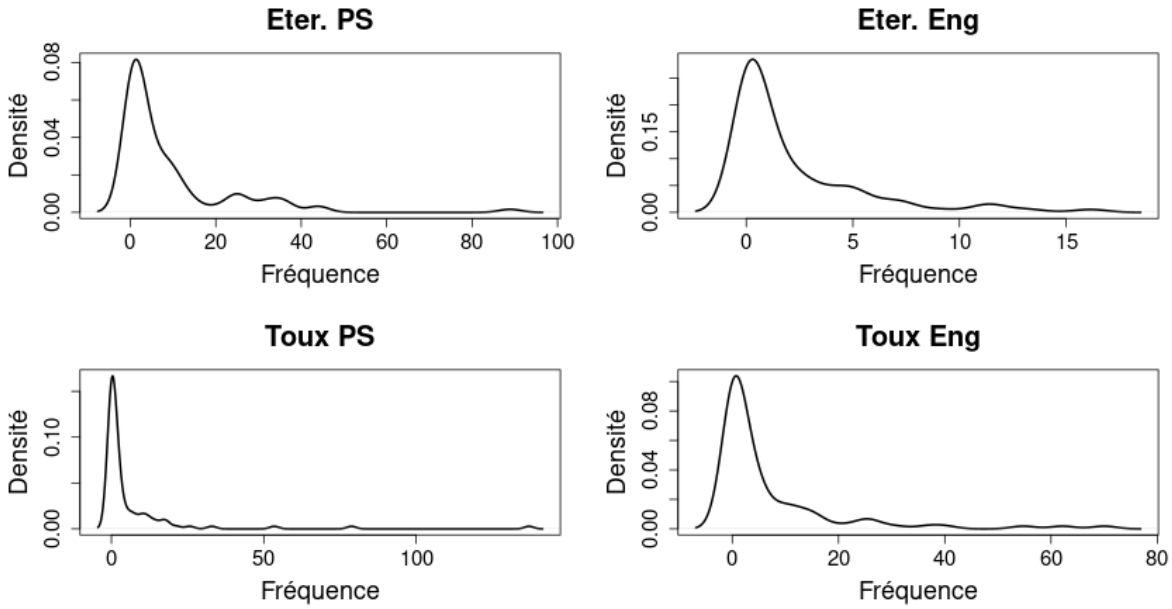


FIGURE 1 – Densité des variables respiratoires (BP [1] : nrd0 [2])

Notons que dans (1) les quatre variables respiratoires présentent une forte concentration de la densité autour de 0, ce qui reflète un bon état respiratoire global chez les porcs étudiés. Cependant, certains élevages montrent des exceptions où cette tendance générale n'est pas observée.

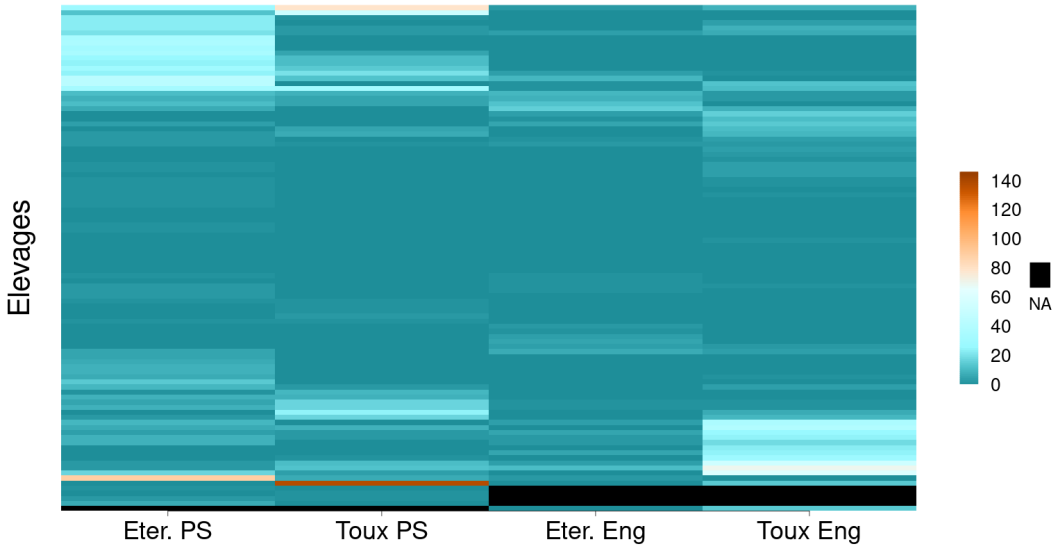


FIGURE 2 – Visualisation des variables respiratoires par élevages

1. Bande passante.
2. La méthode `nrd0` fait référence à la largeur de bande de référence normale.

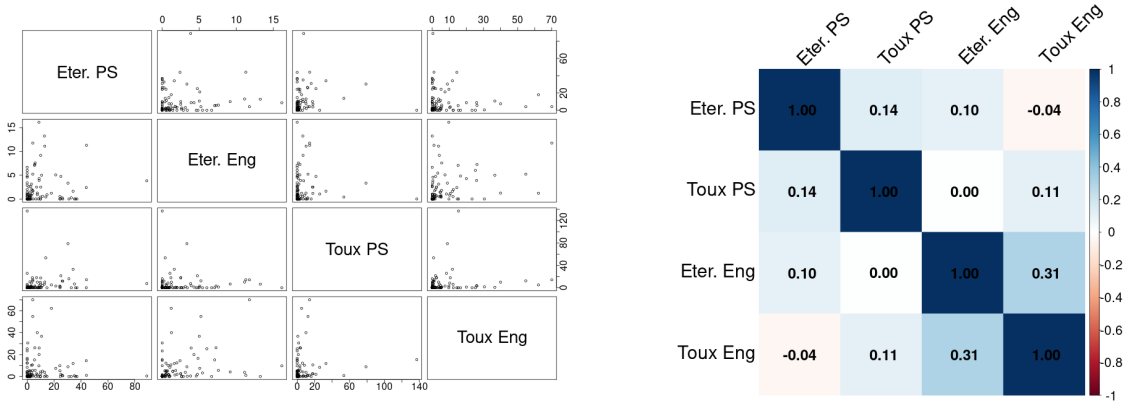
Dans (2), nous constatons une fois de plus que la majorité des individus présentent des valeurs faibles pour l'ensemble des variables respiratoires. Nous relevons également la présence de cinq individus avec des valeurs manquantes. Il apparaît que ces valeurs manquantes concernent soit les variables issues de l'engraissement, soit celles du post-sevrage. Cette observation est cohérente avec le protocole de l'étude, qui prévoyait la mesure simultanée des éternuements et des toux. Des légères corrélations entre la toux et les éternuements sont également observées au sein d'un même lot.

2.2 Exploration des relations entre les variables respiratoires

L'objectif de cette section est d'analyser la force des liens entre les variables respiratoires, afin de déterminer si une régression linéaire pourrait être pertinente pour réduire le nombre de variables à considérer dans les étapes suivantes. De plus, cette exploration permet d'évaluer la pertinence de l'Analyse en Composantes Principales (ACP) pour la réduction de dimensions, afin de juger de son adéquation avec nos données.

2.2.1 Relations linéaires entre les variables respiratoires

Examinons les corrélations ainsi que la répartition des élevages selon les variables respiratoires.



(a) Répartition des élevages

(b) Matrice des correlations

FIGURE 3 – Illustration étude linéaire

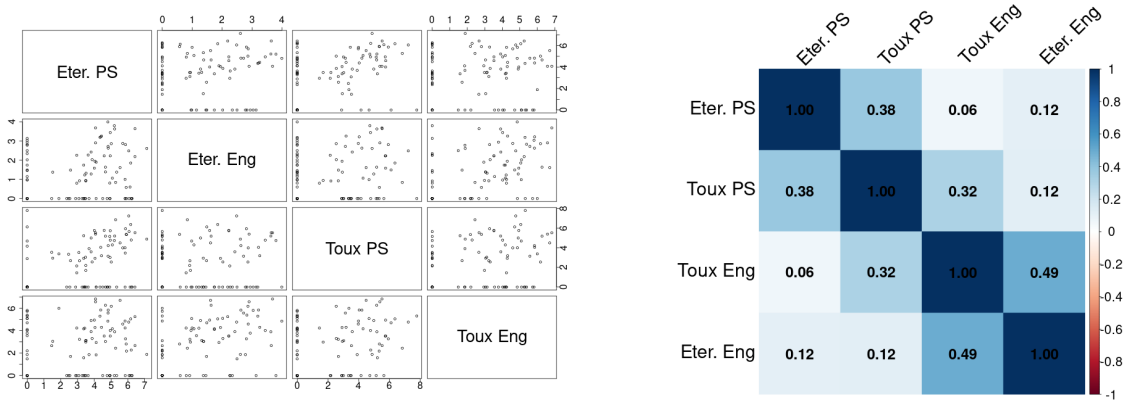
Dans la figure (3), aucune relation linéaire forte n'apparaît entre les différentes variables. Ce constat est particulièrement évident dans la figure (3a), où les nuages de points ne présentent pas la structure triangulaire typique d'une dispersion linéaire. Cependant, la matrice de corrélation (3b) met en évidence des faibles corrélations entre la plupart des variables, à l'exception notable des éternuements en engraissement et des toux en post-sevrage. On observe en effet une corrélation marquée entre les toux et les éternuements, qui appartiennent au même groupe (soit engraissement, soit post-sevrage), cette corrélation étant particulièrement prononcée chez les porcs en engraissement. Le test de Pearson confirme ces observations (Cf. A.1) : la seule corrélation statistiquement significative concerne les variables toux et éternuements en engraissement.

Les analyses graphiques et le test de Pearson révèlent une absence de relations linéaires globales entre les variables, à l'exception de la corrélation entre toux et éternuements en engraissement. Par conséquent, l'ACP, qui repose sur des hypothèses de linéarité, ne semble pas être la méthode d'analyse la mieux adaptée à ce jeu de données.

2.2.2 Relations logarithmiques entre les variables respiratoires

Nous appliquons la transformation $x \longmapsto \ln(x \cdot \hat{\sigma_x} + 1)$ [3] Examinons maintenant les corrélations ainsi que la répartition des élevages selon les nouvelles données.

3. Nous utilisons la transformation $x \longmapsto \ln(x \cdot \hat{\sigma_x} + 1)$ plutôt que $x \longmapsto \ln(x) \cdot \hat{\sigma_x}$ afin de tenir compte de la présence de zéros dans nos données, tout en conservant une fonction bien définie. Le facteur $\hat{\sigma_x}$ (écart-type) est ajouté pour normaliser les données, en atténuant l'effet des différences de dispersion entre les variables, ce qui permet de stabiliser la variance et de rendre les données plus comparables.



(a) Répartition des élevages

(b) Matrice des correlations

FIGURE 4 – Illustration étude logarithmique

Cette fois-ci, nous observons qu'il pourrait exister une relation linéaire entre les logarithmes des variables respiratoires (4a). Notamment, cette relation linéaire est particulièrement visible pour les variables toux post-sevrage et éternuements post-sevrage. Observons également que dans (4a), trois groupes de points se distinguent clairement. En considérant les deux variables précédentes, ces groupes peuvent être interprétés comme suit :

— Absence d'éternuements en post-sevrage
— Absence de toux en post-sevrage
— Présence simultanée de toux et d'éternuements en post-sevrage

C'est dans ce dernier groupe que l'on observe une relation plus ou moins linéaire entre ces deux variables. Dans (4b), cette relation est à nouveau perceptible entre les variables d'engraissement et celles du post-sevrage. De plus, une corrélation relativement forte est mise en évidence entre la toux en post-sevrage et l'engraissement. Ces observations graphiques sont confirmées par les tests de Pearson (A.2).

Cependant, il paraît difficile d'expliquer une variable à partir des trois autres, même au sein du sous-groupe d'élevages qui semble le mieux s'adapter à un modèle linéaire, en raison de la forte amplitude du bruit [4]. De plus, les autres sous-groupes d'élevages ne semblent pas non plus adaptés à ce type d'approche.

2.3 Simplification des variables respiratoires

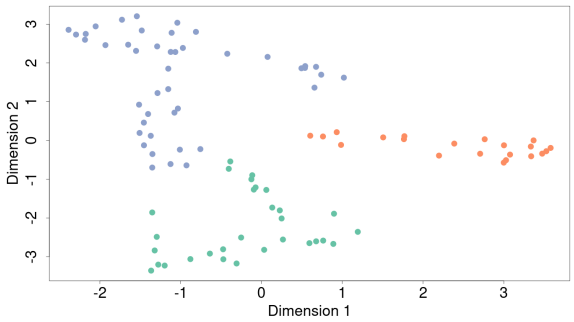
À cette étape, nous proposons de regrouper les élevages en différentes classes afin de réduire le nombre de variables explicatives en une seule. Pour ce faire, nous utilisons la méthode de réduction de dimension UMAP [5], qui s'adapte bien aux structures non linéaires de nos variables. Ensuite, nous appliquons une classification hiérarchique ascendante afin de constituer des groupes d'élevages, qui seront ensuite analysés et interprétés.

2.3.1 Réduction des dimensions

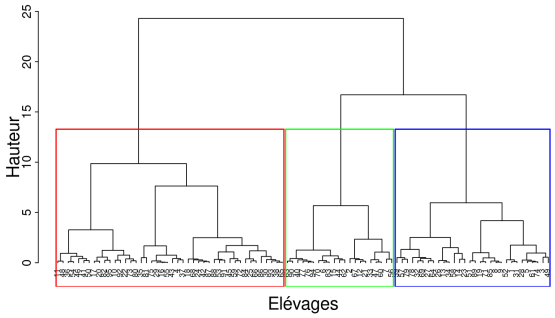
On fait le choix arbitraire de réduire les variables en deux dimensions avec UMAP et on apliqque la classification hiérarchique ascendante (CHA) on obtient alors les résultat suivant :

4. En (A.3), une régression linéaire est détaillée afin de modéliser le logarithme des toux en post-sevrage. Les résultats obtenus montrent toutefois une capacité explicative limitée du modèle, avec un coefficient de détermination de $R^2 = 0.4$

5. *Uniform Manifold Approximation and Projection* en anglais. C'est une technique de réduction de dimension qui permet de représenter des données complexes dans un espace de dimension inférieure tout en conservant au mieux leur structure.



(a) Répartition des élevages dans le plan UMAP par cluster



(b) Dendogramme

FIGURE 5 – Résultats UMAP et CHA

On à fait le choix de garder trois clusters suite au résultats du dendogramme (5b). Passons à l'interpretation des clusters :

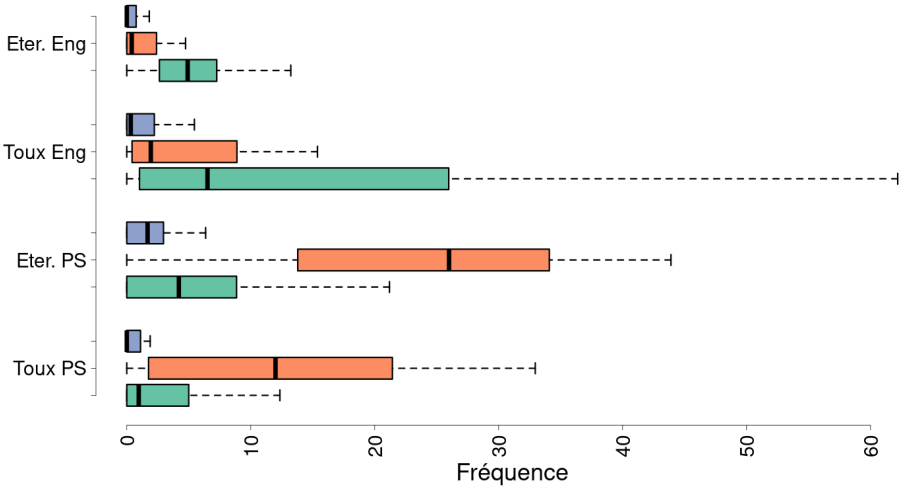


FIGURE 6 – Répartition des clusters par variable respiratoire

On constate à travers (6) la caractérisation suivante des clusters :

— Cluster orange : Contient majoritairement les individus malades en post-sevrage.
— Cluster vert : Contient les individus malades en engraissement, bien que cet effet soit moins marqué.
— Cluster bleu : Représente les individus en bonne santé.

Qu'on nomera dans la suite respectivement : malades post-sevrage, malades-engraissement et sains. Pour une justification plus rigureuse, cf. (A.4)

3 Imputation des variables manquantes

A Annexe : compléments de l'analyse sur les variables respiratoires

A.1 Test de pearson sur les corrélations des variables réspiratoires

Variable 1	Variable 2	Coefficient de corrélation	p-valeur	Significatif
Eter. PS	Eter. Eng	0.104	0.317	Non
Eter. PS	Toux PS	0.135	0.191	Non
Eter. PS	Toux Eng	-0.041	0.696	Non
Eter. Eng	Toux PS	0.003	0.979	Non
Eter. Eng	Toux Eng	0.305	0.003	Oui
Toux PS	Toux Eng	0.114	0.273	Non

Table 1 – Résultats des tests de corrélation de Pearson entre les variables respiratoires.

Constatons que la seule corrélation significative est celle des variables de toux et éternouments en engraissement.

A.2 Test de pearson sur les corrélations du logarithme des variables réspiratoires

Variable 1	Variable 2	Coefficient de corrélation	p-valeur	Significatif
Eter. PS	Eter. Eng	0.119	0.251	Non
Eter. PS	Toux PS	0.377	0.000	Oui
Eter. PS	Toux Eng	0.061	0.557	Non
Eter. Eng	Toux PS	0.122	0.240	Non
Eter. Eng	Toux Eng	0.491	0.000	Oui
Toux PS	Toux Eng	0.315	0.002	Oui

Table 2 – Résultats des tests de corrélation de Pearson entre les variables respiratoires.

TODO

A.3 Résultats régression linéaire entre les variables respiratoires et leur logarithme

TODO GLS with log transform

A.4 Résultats classification

Variable	v.test	$\bar{x}$ (var)	$\bar{x}$ (total)	σ_x (var)
ENG_Eter_freq	6.73	5.55	2.20	3.94
ENG_Tx_freq	4.67	16.31	7.04	19.46
p.value	1.75e-11			

Table 3 – ENG_malade_var

Variable	**v.test**	$\bar{x}$ **(var)**	$\bar{x}$ **(total)**	σ_x **(var)**
PS_Eter_freq	6.35	26.39	9.27	18.71
PS_Tx_freq	4.58	21.98	6.50	32.00
p.value	2.21e-10			

TABLE 4 – PS_malade_var

Variable	**v.test**	$\bar{x}$ **(var)**	$\bar{x}$ **(total)**	σ_x **(var)**
PS_Tx_freq	-2.66	1.34	6.50	2.55
ENG_Tx_freq	-3.53	1.92	7.04	3.15
PS_Eter_freq	-4.19	2.78	9.27	3.59
ENG_Eter_freq	-4.92	0.41	2.20	0.53
p.value	7.75e-03			

TABLE 5 – Sain_var