

Manipulations élémentaires autour de l'inertie

Analyse statistique des peuplements forestiers du bassin du Congo

EL MAZZOUJI Wahel

GILLET Louison

2024/2025



Table des matières

1	INTRODUCTION	3
2	PARTIE 1	3
2.1	Densités	3
2.1.1	Densité de peuplement	3
2.1.2	Densité centrée-réduite	3
2.2	Barycentre et inertie	4
2.2.1	Barycentre à l'origine	4
2.2.2	Inertie totale	4
2.3	Types forestiers	5
2.3.1	Inertie inter-types	6
2.3.2	Coefficient de détermination R^2	6
2.4	Liens entre espèces et types forestiers	6
3	PARTIE 2	8
3.1	Preliminaires	8
3.2	Programmation et calcul de Π_Y, Π_{x_j} puis de $\text{tr}(\Pi_{x_j} \Pi_Y)$	8
3.2.1	Calcul de Π_Y	8
3.2.2	Calcul de Π_{x_j}	9
3.2.3	Calcul de $\text{tr}(\Pi_{x_j} \Pi_Y)$	9
3.3	Calcul de $\text{tr}(R \Pi_Y)$ et interprétation statistique	9
3.4	Calcul de $\text{tr}(\Pi_{x_j} \Pi_Z), \text{tr}(R \Pi_Z)$ et interprétation statistique	9
4	CONCLUSION	10
5	ANNEXE : CODE R	11

1 INTRODUCTION

Dans le cadre de notre étude, nous avons accès à un jeu de données riche qui examine la diversité de 27 espèces d'arbres au sein de 1000 parcelles forestières. Cette analyse vise à explorer la variabilité des densités de peuplement de ces espèces dans le contexte particulier de la forêt du bassin du Congo. Dans un premier temps, il est essentiel d'examiner attentivement cette base de données afin de prévenir tout biais susceptible de compromettre l'analyse. Nous avons donc exclu la dernière ligne du jeu de données, car elle ne correspond pas à une variable pertinente. Ensuite, nous avons sélectionné les variables d'intérêt pour l'analyse. Le jeu de données final comprend ainsi 30 variables quantitatives, parmi lesquelles figurent principalement le nombre d'individus par espèce, la superficie de chaque parcelle, ainsi que deux variables additionnelles relatives au type de forêt et au type géologique.

Table 1: Extrait du jeu de données trié

gen25	gen26	gen27	surface	forest	geology
32	1	0	5.0	2	3
19	0	2	15.0	7	6
3	21	7	17.5	5	6
34	87	30	20.5	1	5
5	18	8	10.5	1	6
130	1	0	20.0	7	3

2 PARTIE 1

2.1 Densités

2.1.1 Densité de peuplement

Nous cherchons à calculer la densité de peuplement de chaque espèce par unité de surface. Pour chaque parcelle, la densité est donnée par :

$$d_{ij} = \frac{n_{ij}}{S_i}$$

où d_{ij} est la densité pour l'espèce j dans la parcelle i , n_{ij} est le nombre d'individus de l'espèce j dans la parcelle i et S_i est la surface de la parcelle i .

Nous utilisons des densités plutôt que des comptages car cela permet de normaliser les données par rapport à la taille de la parcelle, ce qui rend les comparaisons entre les parcelles équitables.

Table 2: Extrait du tableau des densités de peuplement

gen1	gen2	gen3	gen4	gen5	gen6	gen7
0.0000000	0.0000000	0.0	0	0.0000000	0.4000000	0.0000000
0.6000000	0.0000000	0.2	0	0.1333333	0.1333333	0.0666667
0.5142857	0.0000000	0.0	0	0.0571429	0.0000000	0.1714286
0.0000000	0.1951220	0.0	0	0.4390244	0.0487805	0.5365854
0.0952381	0.0952381	0.0	0	0.0000000	0.0000000	0.3809524

2.1.2 Densité centrée-réduite

Nous devons normaliser et centrer les variables quantitatives afin de faciliter la comparaison de celles qui décrivent les différentes densités. Nous allons utiliser la formule suivante pour le centrage et la réduction :

$$d_{ij} \rightarrow z_{ij} = \frac{d_{ij} - \bar{d}_j}{\sigma_j}$$

où \bar{d}_j est la moyenne des densités pour la variable j et σ_j l'écart-type de la variable quantitative j . Par la suite, nous n'utiliserons plus que les variables centrées-réduites.

Table 3: Extrait du tableau des densités centrées-réduites

gen1	gen2	gen3	gen4	gen5	gen6	gen7
-0.9525149	-0.4458588	-0.3833563	-0.3454747	-0.4504654	1.6585968	-0.4433017
0.7194413	-0.4458588	0.5368654	-0.3454747	0.4379021	0.1954714	-0.2791772
0.4805904	-0.4458588	-0.3833563	-0.3454747	-0.0697365	-0.5360913	-0.0212673
-0.9525149	0.1536780	-0.3833563	-0.3454747	2.4746469	-0.2684465	0.8777003
-0.6871251	-0.1532277	-0.3833563	-0.3454747	-0.4504654	-0.5360913	0.4945525

2.2 Barycentre et inertie

2.2.1 Barycentre à l'origine

Considérons que nous avons un ensemble de données présenté sous forme d'une matrice $X = (x_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ où x_{ij} représente la densité de l'espèce j dans la parcelle i . Après le centrage et la réduction, la matrice transformée X' est alors définie par :

$$x_{ij} \rightarrow x'_{ij} = z_{ij} = \frac{d_{ij} - \bar{d}_j}{\sigma_j}$$

Le barycentre du nuage des densités représenté par X' est donné par la moyenne de chaque colonne de X' . Calculons cette moyenne $\forall j \in [1, p]$:

$$\bar{x}'_j = \frac{1}{n} \sum_{i=1}^n x'_{ij} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_{ij} - \bar{x}_j}{\sigma_j} \right) = \frac{1}{\sigma_j} \left(\frac{1}{n} \sum_{i=1}^n x_{ij} \right) - \frac{\bar{x}_j}{\sigma_j} = 0$$

Ainsi, comme chaque moyenne est nulle, le barycentre du nuage se trouve à l'origine.

2.2.2 Inertie totale

Considérons à nouveau la matrice de données X' .

L'inertie de l'ensemble des points x'_{ij} , autour de leur barycentre donné par \bar{x}' (vecteur contenant la moyenne des densités pour chaque espèce j), est définie comme la somme des carrés des écarts entre les densités observées et leur barycentre. Ainsi :

$$\text{In}_O\{x_i, w_i\}_i = \sum_{i=1}^n w_i \|x'_i - \bar{x}'\|^2 \text{ où } \bar{x}' = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Nous supposons que le cas présent est équipondéré par conséquent $w_i = \frac{1}{n} \implies \sum_{i=1}^n w_i = 1$. De plus, nous avons vu précédemment que le barycentre du nuage se trouvait à l'origine d'où $\bar{x}' = 0$. On a alors :

$$\text{In}_O\{x_i, w_i\}_i = \sum_{i=1}^n w_i \|x'_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|x'_i\|^2$$

Comme chaque x'_i est une observation centrée-réduite, nous avons :

$$\|x'_i\|^2 = \sum_{j=1}^p (x'_{ij})^2 = p$$

Ainsi, l'inertie totale est :

$$\text{In}_O\{x_i, w_i\}_i = \frac{1}{n} \sum_{i=1}^n p = \frac{1}{n} \times np = p$$

Cela montre que l'inertie totale du nuage des données centrées-réduites est égale au nombre de variables p . En l'occurrence, le nombre de variables p équivaut au nombre d'espèces, soit 27.

2.3 Types forestiers

Nous nous penchons ici sur les différents types forestiers du bassin du Congo. Notre objectif est de créer une figure permettant, dans un premier temps, de visualiser la répartition des espèces en fonction des types forestiers.

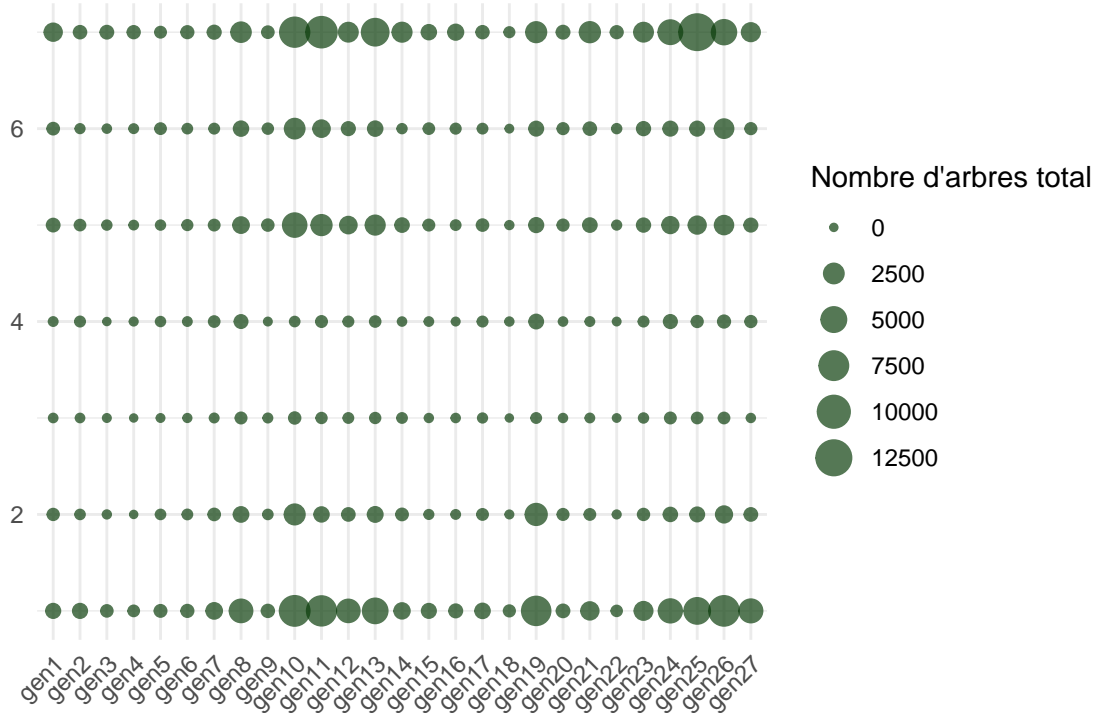


Figure 1: Répartition des espèces d'arbres par type forestier

Tout d'abord, on remarque que les types forestiers 1 et 7 sont ceux ayant le plus d'arbres. Le calcul des poids vient appuyer cette observation :

Table 4: Poids des types forestiers

type1	type2	type3	type4	type5	type6	type7
0.278	0.105	0.022	0.018	0.169	0.095	0.313

Le poids du type forestier k est donné par :

$$p_k = \frac{n_k}{n}$$

où n_k est le nombre de parcelles du type forestier k et n le nombre total de parcelles (1000).

2.3.1 Inertie inter-types

Nous voulons calculer l'inertie inter-types forestiers. Soit :

$$I_{inter-types} = \sum_{k=1}^K p_k \|B_k\|^2$$

où K est le nombre de types forestiers (ici 7) et B_k est le barycentre du type forestier k . Le barycentre du type k $\forall j \in [1, p]$ est obtenu par :

$$B_k^j = \frac{1}{n_k} \sum_{i \in \text{type } k} x_i^j$$

Le calcul de l'inertie nous donne $1.8588457 \approx 1.86$.

2.3.2 Coefficient de détermination R^2

Le coefficient de détermination R^2 est donné par :

$$R^2 = \frac{I_{inter-types}}{I_{total}} = \frac{\sum_{k=1}^K p_k \|B_k\|^2}{p}$$

Ce qui représente la proportion de variance expliquée par les types forestiers dans la variabilité des densités de peuplement. Le résultat obtenu est $0.0688461 \approx 0.07$. Cela signifie qu'environ 7% de la variabilité des densités de peuplement est expliquée par les types forestiers.

2.4 Liens entre espèces et types forestiers

Nous explorons à présent les liens entre les espèces et les types forestiers.

La figure 1 permet de formuler des hypothèses sur ces relations. À première vue, les espèces 10, 11, 25 et 26 semblent avoir des affinités marquées avec certains types forestiers. En revanche, les espèces 2 à 6 présentent une répartition relativement homogène parmi les différents types, suggérant une moindre influence du type forestier sur leur distribution. Toutefois, cette observation ne prend pas en compte l'ensemble des variables disponibles. Une analyse plus approfondie et des calculs complémentaires pourraient révéler que nos intuitions initiales méritent d'être nuancées.

Nous voulons donc calculer le R^2 de chaque variable densité de peuplement où

$$R^2 = \frac{\text{variance inter-types forestiers}}{\text{variance totale}}$$

Après calcul, nous isolons les 5 espèces les plus (respectivement moins) liées aux types :

Table 5: R^2 des 5 espèces les plus liées au type

gen25	gen11	gen2	gen16	gen17
0.1847	0.1439	0.1325	0.1181	0.0964

Table 6: R^2 des 5 espèces les moins liées au type

gen6	gen20	gen14	gen18	gen13
0.0084	0.0139	0.0191	0.0251	0.0303

Nous réalisons que nos premières intuitions étaient en partie correctes, mais nécessitent bien des ajustements. En effet, les espèces 25 et 11 se révèlent être les plus fortement associées au type forestier, tandis que l'espèce 6 montre la plus faible corrélation. À l'inverse de notre hypothèse initiale, l'espèce 2 n'est pas sans lien avec le type forestier. La table 3 indique clairement qu'elle est, en fait, l'une des espèces les plus liées à ce facteur, avec environ 13 % de la variabilité de sa densité de peuplement expliquée par le type forestier.

Nous devons vérifier que le R^2 de la partition est égal à la moyenne des R^2 des espèces. Soit :

$$\begin{aligned}
 R^2 &= \frac{I_{inter-types}}{I_{total}} = \frac{\sum_{k=1}^K p_k \|B_k\|^2}{p} \\
 &= \frac{1}{p} \sum_{k=1}^K p_k \sum_{j=1}^p (B_k^j)^2 \\
 &= \frac{1}{p} \sum_{j=1}^p \sum_{k=1}^K p_k (B_k^j)^2 \\
 &= \frac{1}{p} \sum_{j=1}^p \frac{\sum_{k=1}^K p_k (B_k^j)^2}{1} \\
 &= \frac{1}{p} \sum_{j=1}^p \frac{I_{inter-types,j}}{I_{total,j}} \\
 &= \frac{1}{p} \sum_{j=1}^p R_j^2
 \end{aligned}$$

Nous retrouvons bien informatiquement cette égalité comme le montre l'extrait de code suivant :

```
print(R2)
```

```
[1] 0.06884614
```

```
print(moyenne_R2_especes)
```

```
[1] 0.06884614
```

(le détail des calculs est fourni en Annexe)

3 PARTIE 2

3.1 Préliminaires

Nous savons que l'espace Y peut être décomposé de la manière suivante :

$$\langle Y \rangle = \langle 1 \rangle + \langle Y_{\text{centré}} \rangle$$

où $\langle 1 \rangle$ est le sous-espace vectoriel engendré par le vecteur constant $\mathbf{1}$, et $\langle Y_{\text{centré}} \rangle$ représente le sous-espace vectoriel engendré par Y une fois centré.

La projection $\Pi_Y x^j$ peut donc être décomposée comme suit :

$$\Pi_Y x^j = \Pi_1 x^j + \Pi_{Y_{\text{centré}}} x^j$$

où $\Pi_1 x^j$ est la projection de x^j sur le vecteur constant, et $\Pi_{Y_{\text{centré}}} x^j$ est la projection de x^j sur l'espace engendré par les colonnes de $Y_{\text{centré}}$.

Cependant, la projection sur le vecteur constant, $\Pi_1 x^j$, est égale à zéro, car x^j est une variable centrée. Autrement dit, lorsque nous projetons x^j sur le vecteur constant, nous obtenons :

$$\Pi_1 x^j = 0$$

En substituant cette relation dans l'équation précédente, nous obtenons :

$$\Pi_Y x^j = 0 + \Pi_{Y_{\text{centré}}} x^j$$

d'où il s'ensuit que :

$$\Pi_Y x^j = \Pi_{Y_{\text{centré}}} x^j$$

Ainsi, nous avons montré que, pour tout j , la projection de x^j sur Y est égale à la projection de x^j sur $Y_{\text{centré}}$.

Le fait que ces deux projections soient égales signifie que la projection sur cet espace (lié aux types forestiers) est la même qu'elle soit centrée ou non, car les types forestiers sont des variables qualitatives représentées par des indicatrices. Le centrage ne change donc pas le résultat.

On a :

$$\|\Pi_Y x^j\|_W^2 = \langle \Pi_Y x^j, \Pi_Y x^j \rangle_W$$

La projection de x^j sur Y est donnée par :

$$\Pi_Y x^j = \sum_{q=1}^Q w^q (\bar{x}^q - \bar{x})$$

où \bar{x}^q est la moyenne pondérée de x^j pour le type forestier q , et \bar{x} est la moyenne globale.

Statistiquement, $\|\Pi_Y x^j\|_W^2$ représente la variance inter-groupes de x^j , c'est-à-dire la part de la variance totale de x^j qui est expliquée par les types forestiers :

$$\|\Pi_Y x^j\|_W^2 = \sum_{q=1}^Q w^q (\bar{x}^q - \bar{x})^2$$

Cette expression mesure la part de la variabilité de x^j expliquée par la partition en types forestiers.

3.2 Programmation et calcul de Π_Y , Π_{x_j} puis de $\text{tr}(\Pi_{x_j} \Pi_Y)$

3.2.1 Calcul de Π_Y

Le projecteur Π_Y est défini comme le projecteur sur l'espace généré par les colonnes de la matrice Y . Mathématiquement, il est formulé comme suit :

$$\Pi_Y = Y (Y' W Y)^{-1} Y' W$$

3.2.2 Calcul de Π_{x_j}

De la même manière, nous avons :

$$\Pi_{x_j} = x_j (x_j' W x_j)^{-1} x_j' W$$

3.2.3 Calcul de $\text{tr}(\Pi_{x_j} \Pi_Y)$

Pour chaque vecteur x_j , le calcul de $\text{tr}(\Pi_{x_j} \Pi_Y)$ implique la projection de x_j dans l'espace défini par Y . Cette opération nous permet d'évaluer la quantité de variabilité dans x_j qui est expliquée par la classification en types forestiers. Ainsi,

$$\begin{aligned} \text{tr}(\Pi_{x_j} \Pi_Y) &= \text{tr}(x_j (x_j' W x_j)^{-1} x_j' W \Pi_Y) \\ &= (x_j' W x_j)^{-1} \text{tr}(x_j x_j' W \Pi_Y) \\ &= \frac{1}{\|x_j\|^2} \text{tr}(x_j' W \Pi_Y x_j) \\ &= \frac{1}{\|x_j\|^2} \langle x_j, \Pi_Y x_j \rangle_W \\ &= \frac{\|\Pi_Y x_j\|_W^2}{\|x_j\|^2} \\ &= R^2(x_j | Y) \end{aligned}$$

où $R^2(x_j | Y)$ représente la part de variabilité de x_j qui est due à Y . En somme, $\text{tr}(\Pi_{x_j} \Pi_Y)$ mesure la part de variabilité des densités d'une espèce due au type forestier.

Table 7: R^2 des espèces selon le type forestier par calcul de la trace

gen2	gen6	gen11	gen14	gen20	gen25
0.1327	0.0085	0.1441	0.0191	0.0139	0.1849

Nous retrouvons alors les mêmes R^2 que dans la première partie. (les minimales différences sont dues aux arrondis)

3.3 Calcul de $\text{tr}(R \Pi_Y)$ et interprétation statistique

On note

$$R = X M X' W$$

.

Alors $\text{tr}(R \Pi_Y)$ représente la somme des variances expliquées par la partition en types forestiers, comme représenté par Y , sur l'ensemble des variables contenues dans la matrice X . Cette valeur constitue une mesure de l'inertie inter-types forestiers et fournit une indication globale de la quantité d'information expliquée par la classification des types forestiers sur l'ensemble des variables analysées. Après calcul, nous obtenons $\text{tr}(R \Pi_Y) = 0.0688461$. Nous retrouvons alors la proportion de variance expliquée par les types forestiers dans la variabilité des densités de peuplement obtenue dans la première partie.

3.4 Calcul de $\text{tr}(\Pi_{x_j} \Pi_Z)$, $\text{tr}(R \Pi_Z)$ et interprétation statistique

Le projecteur Π_Z est défini sur l'espace des indicatrices de sols, représentées par la matrice Z . Le calcul de $\text{tr}(\Pi_{x_j} \Pi_Z)$ permet d'évaluer combien de la variabilité de x_j est expliquée par la partition en types de sols. La formule correspondante est :

$$\text{tr}(\Pi_{x_j} \Pi_Z) = \text{tr}(x_j (x_j' W x_j)^{-1} x_j' W \Pi_Z) = R^2(x_j | Z)$$

Table 8: R^2 des espèces selon le type de sol par calcul de la trace

gen2	gen6	gen11	gen14	gen20	gen25
0.0175	0.0088	0.0673	0.0317	0.0044	0.3233

Cet extrait nous montre par exemple qu'environ 32% de la variabilité des densités de peuplement de l'espèce 25 est expliquée par le type de sol.

En parallèle, $\text{tr}(R\Pi_Z)$ quantifie la somme des variances expliquées par la partition en types de sols sur toutes les variables dans X . Cette valeur constitue une mesure de l'inertie inter-sols et fournit une indication globale de la quantité d'information expliquée par la classification des sols sur l'ensemble des variables analysées. Le calcul de $\text{tr}(R\Pi_Z)$ nous donne 0.0758867 soit environ 7,6%, ce qui est légèrement plus que l'inertie inter-types forestiers. Ainsi, la classification des sols sur l'ensemble des variables analysées nous fournit un peu plus d'informations que la classification des types forestiers.

4 CONCLUSION

En somme, cette analyse a approfondi notre compréhension des dynamiques de peuplement forestier dans le bassin du Congo. Nous avons comparé la densité de peuplement de 27 espèces d'arbres sur 1000 parcelles, en tenant compte de la taille des parcelles. La standardisation des densités a révélé que les types forestiers expliquent une part significative de la variabilité des densités de peuplement. L'analyse des R^2 a mis en évidence les espèces les plus influencées par ces types, et les projections ont clarifié l'impact des types forestiers et géologiques sur la répartition des espèces. Ces résultats sont essentiels pour la gestion et la conservation des forêts du bassin du Congo.

5 ANNEXE : CODE R

```
# En tête ----

# Desc : DM1 ADM Date: 02/10/2024 Auteur : EL MAZZOUJI Wahel & GILLET Louison

# Dataframe ----

Datagenus <- read.csv("data/Datagenus.csv", sep = ";")
data <- Datagenus[1:1000, ] # On ne prend pas la ligne 1001
especes <- paste0("gen", 1:27)
colonnes_selectionnees <- c(especes, "surface", "forest", "geology")
data <- data[, colonnes_selectionnees]

# Partie 1 ---- 1.1 Calcul de la densité de peuplement pour chaque espèce (gen1
# à gen27) ####

densite_peuplement <- as.matrix(data[especes]/data$surface) # Conversion en matrice

#### 1.2 Centrage et réduction avec des opérations matricielles ####

### Calcul des moyennes pour chaque espèce (colonne)

moyennes_especes <- (colMeans(densite_peuplement))

### Calcul des écarts-types pour chaque espèce (colonne)
n <- nrow(densite_peuplement)
p <- ncol(densite_peuplement)
mat_moyenne <- matrix(moyennes_especes, nrow = n, ncol = p, byrow = TRUE)
# remplit chaque ligne avec la densité de la colonne

sd_especes <- sqrt(colSums((densite_peuplement - mat_moyenne)^2)/(n - 1)) #racine de la variance sans-bias

### Centrage et réduction : on soustrait les moyennes et on divise par
### l'écart-type

mat_sd <- matrix(sd_especes, nrow = n, ncol = p, byrow = TRUE)
# idem avec l'écart-type

densite_centree_reduite <- (densite_peuplement - mat_moyenne)/mat_sd

#### 1.3 Barycentre et inertie ####

### 1.3.1 Barycentre à l'origine (moyennes des colonnes proches de 0)
### summary(densite_centree_reduite)
moyennes_apres_centrage <- colMeans(densite_centree_reduite)
test_barycentre_a_lorigine <- all(abs(moyennes_apres_centrage) < 1e-10)

### 1.3.2 Inertie totale (variances des colonnes proches de 1) Variance par
### colonne
variances_apres_centrage <- colSums(densite_centree_reduite^2)/(n - 1)

# Inertie totale = somme des variances
inertie_totale <- sum(variances_apres_centrage)
```

```

#### 2.1 Calcul des poids, barycentres des types forestiers et normes
#### euclidiennes de ces barycentres ####

### Identification des types forestiers
types_forestiers <- sort(unique(data$forest))

### Création d'une matrice pour les poids, barycentres et normes euclidiennes
### carrées
d <- length(types_forestiers)
poids_forestiers <- numeric(d)
barycentres_forestiers <- matrix(0, nrow = d, ncol = p) #p=ncol(densite_centree_reduite)
normes_euclidiennes_carre <- numeric(d)

### Calcul par opérations matricielles
for (i in 1:d) {
  # Filtrer les parcelles appartenant à chaque type forestier
  parcelles_type_forestier <- densite_centree_reduite[data$forest == types_forestiers[i],
    ]

  # Calculer le poids : proportion des parcelles de ce type
  poids_forestiers[i] <- nrow(parcelles_type_forestier)/nrow(densite_centree_reduite)

  # Calcul du barycentre pour chaque type forestier (moyenne par espèce)
  barycentres_forestiers[i, ] <- colMeans(parcelles_type_forestier)

  # Calcul de la norme euclidienne carrée pour chaque type forestier
  normes_euclidiennes_carre[i] <- sum(barycentres_forestiers[i, ]^2)
}

#### 2.2 Calcul de l'inertie inter-types et du R2 (coefficient de
#### détermination) ####

### Inertie inter-types
inertie_inter_types <- sum(poids_forestiers * normes_euclidiennes_carre)

### Coefficient de détermination R2
R2 <- inertie_inter_types/inertie_totale

#### 2.3 Pourcentage d'information (variabilité du peuplement) ####
pourcentage_information <- R2 * 100

#### 3.1 Calcul de la variance totale et de la variance inter-types pour chaque
#### espèce ####

### Variance totale
variance_totale_par_espece <- colSums(densite_centree_reduite^2)/(n - 1)

### Variance inter-types
variance_inter_types_par_espece <- numeric(ncol(densite_centree_reduite))

# Boucle pour chaque espèce
for (j in 1:ncol(densite_centree_reduite)) {

```

```

    # On calcule la variance inter-types pour l'espèce j
    variance_inter_types_par_espece[j] <- sum(poids_forestiers * (barycentres_forestiers[,
        j]^2))
}

#### 3.2 Calcul du R2 pour chaque espèce ####
R2_par_espece <- variance_inter_types_par_espece/variance_totale_par_espece

#### 3.3 Identification des espèces les plus et les moins liées au type
#### forestier ####
especes_most_liees <- (sort(R2_par_espece, decreasing = TRUE))[1:5] # Les 5 espèces les plus liées
especes_least_liees <- (sort(R2_par_espece, decreasing = FALSE))[1:5] # Les 5 espèces les moins liées

#### 3.4 Calcul de la moyenne arithmétique des R2 des espèces ####
moyenne_R2_especes <- mean(R2_par_espece)

#### 3.5 Vérification que le R2 de la partition est égal à la moyenne des R2
#### des variables ####
verification_R2 <- R2 == moyenne_R2_especes

# Partie 2 ---- 1.1 Calcul Projection Y ####

### 1.1.1 Préliminaires : création des matrices
X <- densite_centree_reduite
dim(X)
Y <- model.matrix(~as.factor(forest) - 1, data = data)
colnames(Y) <- paste("type", seq_along(levels(as.factor(data$forest))))
W <- diag(1/n, n, n) # Matrice de poids équipondérés
M <- diag(1/p, p, p) # Matrice de poids pour les variables

### 1.1.2 Calcul
Pi_Y <- Y %%% solve(t(Y) %%% W %%% Y) %%% t(Y) %%% W #solve donne l'inverse de Y'WY

#### 1.2 Calcul Pi_xj et tr(Pi_Y*Pi_xj) ####
tr_Pi_xj_PiY <- numeric(p)

for (j in 1:p) {
    x_j <- X[, j]
    Pi_xj <- x_j %%% solve(t(x_j) %%% W %%% x_j) %%% t(x_j) %%% W
    tr_Pi_xj_PiY[j] <- sum(diag(Pi_xj %%% Pi_Y))
}

#### 1.3 Calcul de tr(RPI_Y)
R <- X %%% M %%% t(X) %%% W
trace_R_Pi_Y <- sum(diag(R %%% Pi_Y))

#### 2.1 Calcul de tr(Pi_xj*Pi_Z) ####

### 2.1.1 Création matrice Z
Z <- model.matrix(~as.factor(geology) - 1, data = data)
colnames(Z) <- paste0("geology", setdiff(1:6, 4)) #il n'y a pas de 4 pour geology

### 2.1.2 Calcul Pi_Z
Pi_Z <- Z %%% solve(t(Z) %%% W %%% Z) %%% t(Z) %%% W

```

```
### 2.1.3 Calcul
tr_Pi_xj_PiZ <- numeric(p)

for (j in 1:p) {
  x_j <- X[, j]
  Pi_x_j <- x_j %*% solve(t(x_j) %*% W %*% x_j) %*% t(x_j) %*% W
  tr_Pi_xj_PiZ[j] <- sum(diag(Pi_x_j %*% Pi_Z)) # trace(Pi_x_j * Pi_Z)
}

#### 2.2 Calcul de tr(RPi_Z) ####
tr_R_Pi_Z <- sum(diag(R %*% Pi_Z))

# Sauvegarde pour Rmd ----
save.image(file = "ressources/prepa.RData")
```