

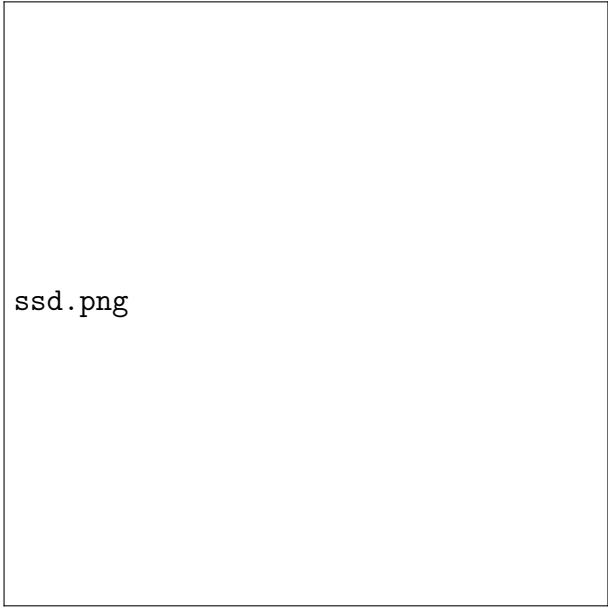
# DEVOIR MAISON ADM

Wahel El Mazzouji

*et*

Louison Gillet

08/10/2024



ssd.png

Table des matières

Table des figures

# 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. Le jeu de données se compose de 30 variables quantitatives, dont les principales incluent le comptage des individus pour chaque espèce, la superficie de chaque parcelle ainsi que deux variables supplémentaires relatives au type forestier et au type géologique. À cela s'ajoute une variable qualitative, identifiée par un "code", qui permet d'apporter des informations contextuelles sur chaque parcelle. Cette étude permettra d'éclairer les dynamiques écologiques en jeu et d'approfondir notre compréhension des interactions entre les espèces arborées et leur environnement.

## 2 PARTIE 1

### 2.1 Question 1

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

$$D_{ij} = \frac{N_{ij}}{S_j}$$

où  $D_{ij}$  est la densité pour l'espèce  $i$  dans la parcelle  $j$ ,  $N_{ij}$  est le nombre d'individus de l'espèce  $i$  dans la parcelle  $j$  et  $S_j$  est la surface de la parcelle  $j$ .

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.

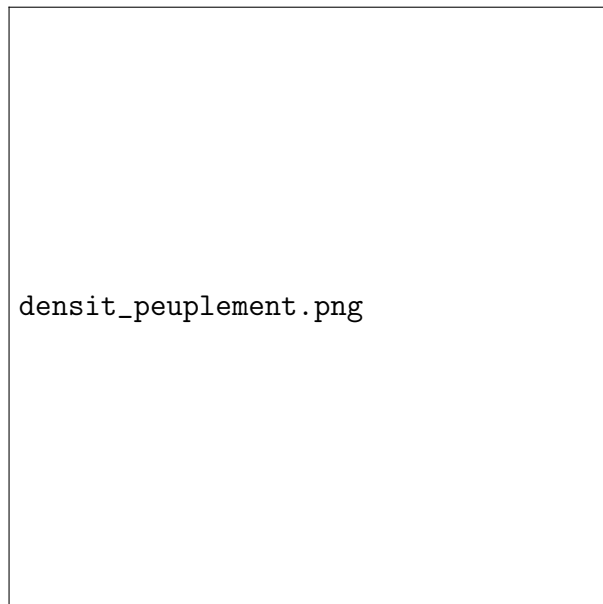


FIGURE 1 – Extrait des densités de peuplement

## 2.2 Question 2

Nous devons centrer et réduire les variables quantitatives pour mieux comparer celles qui décrivent les différentes densités. Nous allons utiliser la formule suivante pour le centrage et la réduction :

$$Z_{ij} = \frac{D_{ij} - \bar{D}_j}{s_j}$$

où  $\bar{D}_j$  est la moyenne pour la variable  $j$  et  $s_j$  l'écart-type de la variable quantitative  $j$ .

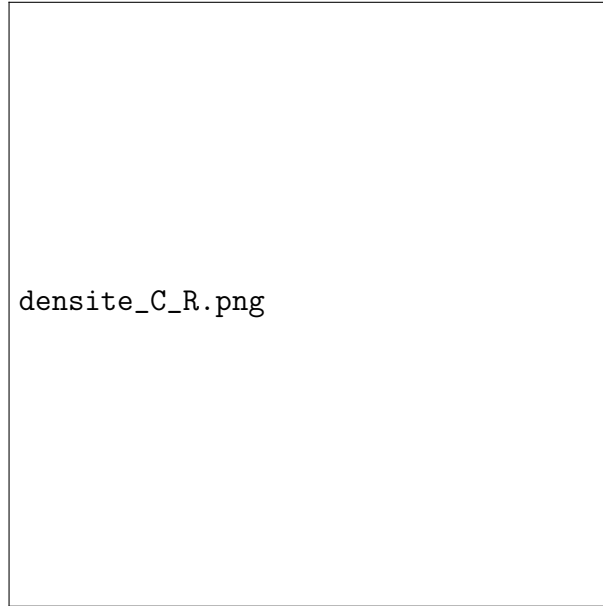


FIGURE 2 – Extrait des densités centrées réduites

## 2.3 Barycentre à l'origine

Considérons que nous avons un ensemble de données  $X$  composé de  $n$  observations et  $p$  variables. Après le centrage et la réduction, la matrice transformée  $X'$  est définie par :

$$X'_{ij} = \frac{D_{ij} - \bar{D}_j}{s_j}$$

Le barycentre de  $X$  est donné par la moyenne de chaque colonne de  $X$ . Calculons cette moyenne pour la variable  $j$  :

$$\bar{D}_j = \frac{1}{n} \sum_{i=1}^n X'_{ij} = \frac{1}{n} \sum_{i=1}^n \left( \frac{D_{ij} - \bar{D}_j}{s_j} \right) = \frac{1}{s_j} \left( \frac{1}{n} \sum_{i=1}^n D_{ij} \right) - \frac{\bar{D}_j}{s_j} = 0$$

Ainsi, le barycentre de chaque variable dans  $X'$  est égal à zéro.

## 2.4 Inertie totale

Considérons à nouveau la matrice de données  $X$ . Après centrage et réduction, chaque élément de la matrice transformée  $X'$  est défini par :

$$X'_{ij} = \frac{D_{ij} - \bar{D}_j}{s_j}$$

L'inertie de l'ensemble des points  $X'$  par rapport à leur barycentre  $y_M$  est définie par :

$$I_{Y,W} = \sum_{i=1}^n w_i \|X_i - y_M\|^2$$

où  $y_M = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$  est le barycentre pondéré. Pour les données centrées-réduites, chaque  $X'_i$  est déjà centré, donc le barycentre  $y_M = 0$ . Par conséquent, la formule de l'inertie se simplifie à :

$$I_{Y,W} = \sum_{i=1}^n w_i \|X'_i\|^2$$

Si tous les poids  $w_i$  sont égaux (par exemple,  $w_i = \frac{1}{n}$ ), alors l'inertie devient :

$$I_{Y,W} = \frac{1}{n} \sum_{i=1}^n \|X'_i\|^2$$

Comme chaque  $X'_i$  est une observation centrée-réduite et que la variance de chaque variable est 1, nous avons :

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

Ainsi, l'inertie totale est :

$$I_{Y,W} = \frac{1}{n} \sum_{i=1}^n p = p$$

Ce qui montre que l'inertie totale du nuage des données centrées-réduites est égale au nombre de variables  $p$ .

## 2.5 Question 3

L'inertie inter-types est calculée par :

$$I_{inter} = \sum_{k=1}^p w_k \cdot n_k^2$$

où  $p$  est le nombre de types forestiers,  $w_k$  est le poids du type forestier  $k$  et  $n_k$  est la norme euclidienne carrée pour ce type.

Le coefficient de détermination  $R^2$  est donné par :

$$R^2 = \frac{I_{inter}}{I_{total}}$$

Ce qui représente la proportion de variance expliquée par les types forestiers dans la variabilité des densités de peuplement.

## 2.6 Question 4

Nous devons vérifier que le  $R^2$  de la partition est égal à la moyenne des  $R^2$  des espèces. Pour cela, nous calculons le  $R^2$  pour chaque espèce  $i$  :

$$R_i^2 = \frac{I_{inter,i}}{I_{total,i}}$$

Puis, nous comparons avec :

$$R_{partition}^2 = \frac{1}{m} \sum_{i=1}^m R_i^2$$

où  $m$  est le nombre d'espèces.

## 3 PARTIE 2

### 3.1 Question 1

### 3.2 Question 1

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  $Y_{\text{centré}}$ .

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.

## 4 CONCLUSION

Cette analyse nous a permis de mieux comprendre les dynamiques de peuplement forestier dans la forêt du bassin du Congo. En intégrant des approches statistiques robustes, nous avons pu quantifier la variabilité des espèces et leur répartition en fonction des types forestiers. Les résultats de cette étude fourniront des bases solides pour des recherches futures et des actions de conservation.