TP1 ADM

MARIAC Damien, SCAIA Mattéo

October 20, 2024

Contents

1 Introduction

Nous disposons d'une base de données comprenant l'analyse de 27 espèces d'arbres réparties dans 1000 parcelles forestières situées dans la forêt du Congo. Il s'agit d'étudier la variabilité des densités de peuplement d'espèces arborées dans différentes parcelles de la forêt. Nous avons enlevé une ligne problématique dans notre jeu de données (ligne 1000 code TGC). Nous disposons dans notre jeu de données de 28 variables quantitatives dont : 27 variables de comptage sur les espèces d'arbres et une sur la surface de la parcelle. De plus, notre base de données intègre deux variables qualitatives : l'une décrivant le type forestier et l'autre le type géologique.
On arrondira nos calculs à 10^{-3}.

Table 1: Extrait du jeu de données Datagenus

code	gen1	gen2	gen3	forest	geology	surface
1299	0	0	0	2	3	5
2644	9	0	3	7	6	15
1838	9	0	0	5	6	17.5
534	0	4	0	1	5	20.5
3213	1	1	0	1	6	10.5
1861	19	3	1	7	3	20

2 Partie 1

2.1 Inertie et barycentre

Afin d'avoir une comparaison plus juste entre chaque parcelle, nous utilisons la densité de peuplement de ces dernières. C'est-à-dire que la densité de peuplement de chaque espèce d'arbre par unité de surface est donnée par :

$$d_i^j = \frac{n_i^j}{S_i}$$

où n_i^j est le nombre d'arbres de l'espèce j présents sur la parcelle i, et S_i représente la surface de la parcelle i.

Nous devons de plus centrer et réduire les variables quantitatives dans le but de mieux comparer celles qui décrivent les différentes densités. Nous allons alors utiliser :

$$(x_i^j)_{1\leq j\leq 27} = \frac{x_i^j - \bar{x}_j}{\sigma_j}$$

avec $\bar{x}_j$ la moyenne pour la j-ème variable et σ_j l'écart-type de la j-ème variable.

Dans la suite, nous ne considérerons plus que les variables centrées-réduites.

De plus, on supposera que le poids de nos parcelles est équipondéré.

Par conséquent, on a :

Barycentre à l'origine : (preuve théorique)

Supposons que nous avons un ensemble de données X composé de n observations et p variables (dans notre cas 1000 et 27). Après le centrage et la réduction, les coefficients de notre matrice transformée X' sont définis par :

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Le barycentre de X est donné par la moyenne de chaque colonne de X. Calculons cette moyenne pour n'importe quelle j :

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

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

Inertie totale égale à 27 : (preuve théorique)

Considérons la même matrice de données X'. L'inertie de l'ensemble des points X par rapport à leur barycentre y est définie par :

$$I_{y,w} = \sum_{i=1}^{n} w_i \|x_i - y\|^2$$

Dans notre cas, le barycentre est nul. De plus, tous les poids w_i sont égaux (par exemple, $w_i = \frac{1}{n}$ ce qui est notre cas), alors l'inertie devient :

$$I_{y,w} = \sum_{i=1}^{n} w_i \|x_i\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|x_i\|^2$$

Comme chaque x_i est une observation centrée-réduite et la variance de chaque variable est 1, nous avons (pour i fixé) :

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

Ainsi, l'inertie totale est :

$$I_{y,w} = \frac{1}{n} \sum_{i=1}^{n} p = p$$

Dans notre cas vaut 27.

2.2 Autour des types forestiers

Dans cette section, nous calculons les barycentres des sept types forestiers présents dans les données, ainsi que l'inertie inter-types et le coefficient R^2 associé à la partition des parcelles selon ces types. Ce calcul nous permet d'évaluer la proportion de la variabilité totale des densités de peuplement.

2.2.1 Calcul des poids des types forestiers

Le poids de chaque type forestier est calculé comme la proportion de parcelles appartenant à ce type par rapport à l'ensemble des 1000 parcelles. C'est à dire, le poids est donné par :

$$p_i = \frac{\text{Nombre de parcelles du type } i}{1000}$$

Table 2: Tableau des poids forestiers

1	2	3	4	5	6	7
0.278	0.105	0.022	0.018	0.169	0.095	0.313

2.2.2 Calcul des barycentres des types forestiers

Pour chaque type forestier i, (avec X la matrice des densités centrées réduites), le barycentre vaut :

$$\bar{X}_i = \frac{1}{n_i} \sum_{j \in \text{Type } i} X^j$$

Avec n_i le nombre de parcelles dans le type forestier i.
Nous obtenons alors une matrice $B \in \mathcal{M}_{7,27}(\mathbb{R})$

2.2.3 Calcul des normes euclidiennes carrées

Une fois les barycentres calculés, nous évaluons la distance de chaque barycentre à l'origine:

Table 3: normes carrées

type forestier	1	2	3	4	5	6	7
distance (en norme 2)	0.772	3.228	4.279	21.210	1.092	1.443	1.628

2.2.4 Inertie inter-types et coefficient R^2

L'inertie inter-types forestiers est calculée en pondérant les normes euclidiennes carrées par les poids des types forestiers. Elle mesure la variabilité des densités de peuplement expliquée par la partition en types forestiers et elle est définie par :

$$\text{Inertie inter-types} = \sum_{i=1}^{7} p_i \|\bar{X}_i\|^2 = 1.860$$

Le coefficient R^2 nous permet d'évaluer dans quelle mesure les types forestiers expliquent la variabilité des densités de peuplement dans les parcelles observées.

$$R^2 = \frac{\text{Inertie inter-types}}{\text{Inertie totale}} = 0.069$$

Cela correspond à une disparité d'environ 7%

2.3 Liaisons

On s'intéresse maintenant à l'interprétation des résultats trouvés. Nous calculons alors le R^2 de chaque variable densité avec $R^2 = \frac{variance_{inter-typesforestiers}}{variancetotale}$. (C'est à dire : $R^2 = \frac{\text{Inertie externe}}{\text{Inertie totale}}$.)

Table 4: Extrait du $\mathbf{R^2}$ des variables

Especes	gen1	gen6	gen17	gen25
$\mathbf{R^2}$	0.072	0.008	0.096	0.184

Nous remarquons numériquement que la variable la plus liée au type est gen25, tandis que la moins liée est gen6.
Montrons que le R^2 de la partition est la moyenne (arithmétique) des R^2 des densités:

$$R^2 = \frac{\text{Inertie externe}}{\text{Inertie totale}} = \frac{1}{27}\sum_{k=1}^{7}\|x_k^j\|^2 = \frac{1}{27}\sum_{k=1}^{7}\sum_{j=1}^{27}(x_k^j)^2 = \frac{1}{27}\sum_{j=1}^{27}\sum_{k=1}^{7}(x_k^j)^2 = \frac{1}{27}\sum_{j=1}^{27} R_j^2$$

On retrouve bien numériquement la même valeur :

Listing 1: Extrait du code R

```
R2_global <- sum(R2_variables) / length(R2_variables)
print(R2_global)

R2_partition <- sum(poids * rowSums(barycentres_types^2)) / 27
print(R2_partition)
```

```
> R2_global <- sum(R2_variables) / length(R2_variables)
> print(R2_global)
[1] 0.06891505
> R2_partition <- sum(poids * rowSums(barycentres_types^2)) / 27
> print(R2_partition)
[1] 0.06891505
```

3 Partie 2

3.1 Enoncé

On notera X la matrice dont les colonnes sont les 27 densités centrées-réduites, Y la matrice dont les colonnes sont les indicatrices de types forestiers, et Z celle dont les colonnes sont les indicatrices de sols (geology).
On notera $W = \frac{1}{n} I_n$ la matrice des poids des individus et $M = \frac{1}{p} I_p$ celle des poids des variables.

3.2 Rappel et calcul de projecteur

3.2.1 Rappel

Montrons l'égalité suivante.

$$\forall j \qquad \Pi_Y x^j = \Pi_{Y^c} x^j$$

Nous rappelons que

$$Y^c = \Pi_{1^\perp} Y$$

$$\Pi_Y = \Pi_{Y^c} + \Pi_1$$

A partir de ces deux résultats, nous avons :

$$\begin{aligned} \Pi_Y x^j &= \Pi_{Y^c} x^j + \Pi_1 x^j \\ &= \Pi_{Y^c} x^j \end{aligned}$$

C'est à dire, pour tout j :

$$\Pi_Y x^j = \Pi_{Y^c} x^j$$

De plus, nous avons l'expression suivante.

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

Or,

$$\Pi_Y x^j = \sum_{q=1}^{Q} \Pi_{y^q} x^j = \sum_{q=1}^{Q} y^q (\overline{x^j}^q - \overline{x^j})$$

Avec, $\overline{x^j}$ représentant la moyenne globale des x^j, $\overline{x^j}^q$ représentant la moyenne pondérée des x^j pour le type forestier q.
A partir de ces notations, nous avons l'expression suivante.

$$\begin{aligned} \left\|\Pi_Y x^j\right\|_W^2 &= \left\langle \Pi_Y x^j, \Pi_Y x^j \right\rangle_W \\ &= \sum_{q=1}^{Q} (\overline{x^j}^q - \overline{x^j}) \sum_{i=1}^{Q} (\overline{x^j}^i - \overline{x^j}) \left\langle y^q, y^i \right\rangle_W \\ &= \sum_{q=1}^{Q} w^q (\overline{x^j}^q - \overline{x^j})^2 \end{aligned}$$

Ici, $\left\|\Pi_Y x^j\right\|_W^2$ s'interprète statistiquement comme étant une mesure pondérée de la variation des x^j dans un certain type forestier.

3.2.2 Calcul de projecteur

Le but de cette question est de trouver l'expression de Π_Y et de calculer pour tout $j \in [1, 27]$, Π_{x^j} et $tr(\Pi_{x^j}\Pi_Y)$. Nous ferons la démonstration puis nous programmerons le résultat.
Tout d'abord, on rappelle que :

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

Soit $j \in [1, 27]$. Calculons les deux expressions données précédemment.

$$\Pi_{x^j} = x^j(x^{j\prime}Wx^j)^{-1}x^{j\prime}W$$

En utilisant les propriétés de la trace et l'expression de Π_{x^j}, il en suit

$$\begin{aligned} tr(\Pi_{x^j}\Pi_Y) &= tr(x^j(x^{j\prime}Wx^j)^{-1}x^{j\prime}W\Pi_Y) \\ &= tr((x^{j\prime}Wx^j)^{-1}x^{j\prime}W\Pi_Y x^j) \\ &= (x^{j\prime}Wx^j)^{-1}x^{j\prime}W\Pi_Y x^j \\ &= \frac{\langle x^j, \Pi_Y x^j\rangle_W}{\|x^j\|_W^2} \\ &= R^2(x^j \mid Y) \end{aligned}$$

Donc $tr(\Pi_{x^j}\Pi_Y)$ représente le R^2 d'une densité de population sachant son type forestier. C'est-à-dire que nous mesurons la variabilité d'une espèce liée à un type forestier.
Nous pouvons programmer les deux expressions ci dessus.

Listing 2: Extrait du code R

```
P_Y <- Y %*% solve(t(Y)%*% W %*% Y) %*% t(Y) %*% W
P_X <- function(j){
    # Calculer la projection de la colonne j de X
    x_j <- X[, j, drop = FALSE]  # Pour que x_j reste une matrice
    return(x_j %*% solve(t(x_j) %*% W %*% x_j) %*% t(x_j) %*% W)
}
#On calcule la trace du produit matriciel
Tr_1 <- sum(diag(P_X(3) %*% P_Y))
```

Nous effectuons le calcul de la trace pour tout $j \in [1, 27]$. Nous obtenons le tableau suivant.

Table 5: Extrait du R^2 des variables via le calcul de la trace

Especes	gen1	gen6	gen17	gen25
R^2	0.072	0.008	0.096	0.184

3.2.3 Calcul de R et de $tr(R\Pi_Y)$

De la même manière, nous programmons la matrice $R = XMX'W$, ainsi que $tr(R\Pi_Y)$.

Listing 3: Extrait du code R

```
    R <- X %*% M %*% t(X) %*% W
    Tr_2 <- sum(diag(R %*% P_Y))
```

Nous pouvons alors calculer la trace de la matrice demandée et nous obtenons.

$$tr(R\Pi_Y) = 0,069$$

Nous pouvons interpréter $tr(R\Pi_Y)$ comme étant le coefficient R^2 suivant :

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

3.2.4 Interprétation des résultats

Nous avons trouvé les mêmes résultats que dans la partie 1 du TP. Cependant, dans cette partie, nous avons utilisé des calculs différents.
Nous pouvons rapprocher les résultats obtenus dans Table 4 et dans Table 5. En effet, nous avons bien les mêmes valeurs. De plus, les résultats du coefficient R^2 obtenu dans la partie 3.2.3 et celui obtenu dans la partie 2.2.4 sont bien identiques.
En conclusion, avec deux méthodes de calcul différents, nous retrouvons bien les mêmes résultats.

3.3 Rapprochement avec la variable "geology"

Dans cette question, nous allons nous intéresser a la variable Z "geology". Programmons de la même manière, les matrices demandées.

Listing 4: Extrait du code R

```
P_Z <- Z %*% solve(t(Z)%*% W %*% Z) %*% t(Z) %*% W
Tr_3 <- sum(diag(P_X(3) %*% P_Z))
Tr_4 <- sum(diag(R %*% P_Z))
```

Pour le calcul de $tr(\Pi_{x^j}\Pi_Z)$ nous obtenons les résultats suivants.

Table 6: Extrait du R^2 des variables

Especes	gen1	gen6	gen17	gen25
$\mathbf{R^2}$	0.096	0.009	0.035	0.323

D'après nos calculs numériques, nous observons que la variable la plus liée au type géologique est gen25 ($R^2 = 0,323$), tandis que celle qui est la moins liée est gen20 ($R^2 = 0,004$).

Nous pouvons calculer son expression pour l'interpréter statistiquement.

$$tr(\Pi_{x^j}\Pi_Z) = tr(x^j(x^{j\prime}Wx^j)^{-1}x^{j\prime}W\Pi_Z) = \frac{\langle x^{j\prime}, \Pi_Z x^j\rangle_W}{\|x^j\|_W^2} = R^2(x^j \mid Z)$$

Nous trouvons que $tr(\Pi_{x^j}\Pi_Z)$ est le R^2 d'une densité de population sachant son type géologique. C'est-à-dire que nous mesurons la variabilité d'une espèce liée à un type géologique.

De plus, le calcul de $tr(R\Pi_Z)$ nous donne.

$$tr(R\Pi_Z) = 0,0759$$

Ce dernier représente la variabilité d'une espèce dans un type géologique. C'est-à-dire que la variation d'une espèce s'explique à $7,59\%$ par le type géologique (contre $6,9\%$ pour le type forestier).

4 Conclusion

Au cours de notre analyse des peuplements forestiers du bassin du Congo, nous avons identifié des variations significatives qui peuvent être attribuées aux caractéristiques forestières et géologiques des sites étudiés. Nos résultats suggèrent que le type géologique présente une influence prépondérante sur ces disparités, surpassant l'impact du type forestier.

5 ANNEXE

```
rm(list = ls())

tab <- read.csv("C:/Users/damie/Desktop/MASTER/ADM/tp/tp1/Datagenus.csv", sep =
    ";")
data <- tab[1 :1000,]
species_columns <- grep("gen", colnames(data), value = TRUE)

##QUESTION 1

density_data <- data[species_columns] / data$surface

W <- diag(1/1000, 1000, 1000)
I_n <- rep(1, 1000)

Q <- as.matrix(density_data)
x_b <- t(Q) %*% W %*% I_n # moyenne (par projection)
x_c <- Q - I_n %*% t(x_b)

vec_norm <- sqrt(diag(t(x_c) %*% W %*% x_c))


x_cr <- sweep(x_c, 2, vec_norm, FUN = "/")
tableau_cr <- as.data.frame(x_cr)


bar <- t(x_cr) %*% W %*% I_n
barycentre <- t(bar)
print(barycentre)

inertie_totale <- sum(diag(t(x_cr) %*% W %*% x_cr)) #inertie
print(inertie_totale)

print(tableau_cr)

##Question 2


forest_types <- unique(data$forest)
n_types <- length(forest_types)

poids <- numeric(n_types)
barycentres_types <- matrix(0, n_types, length(species_columns))

for (i in 1:n_types) {
  parcelles_type_i <- which(data$forest == forest_types[i])
  n_i <- length(parcelles_type_i)
  poids[i] <- n_i / 1000
  W_i <- diag(1/n_i, n_i, n_i)
  I_i <- rep(1, n_i)
  Q_i <- as.matrix(x_cr[parcelles_type_i, ])
  x_b_i <- t(Q_i) %*% W_i %*% I_i
  barycentres_types[i, ] <- t(x_b_i)
}

normes_carre_barycentres <- rowSums(barycentres_types^2)

inertie_inter_types <- sum(poids * normes_carre_barycentres)
print(inertie_inter_types)

R2 <- inertie_inter_types / inertie_totale
print(R2)
```

```r





forest_types <- unique(data$forest)
n_types <- length(forest_types)

R2_variables <- numeric(length(species_columns))

for (j in 1:length(species_columns)) {
  variance_totale_j <- sum((x_cr[, j] - mean(x_cr[, j]))^2) / 1000

  variance_inter_j <- 0
  for (i in 1:n_types) {
    parcelles_type_i <- which(data$forest == forest_types[i])
    n_i <- length(parcelles_type_i)
    poids_i <- n_i / 1000

    W_i <- diag(1/n_i, n_i, n_i)
    I_i <- rep(1, n_i)

    Q_i <- as.matrix(x_cr[parcelles_type_i, j])
    barycentre_i <- t(Q_i) %*% W_i %*% I_i

    variance_inter_j <- variance_inter_j + poids_i * as.numeric(barycentre_i)^2
  }

  R2_variables[j] <- variance_inter_j / variance_totale_j
}

print(R2_variables)

densite_max <- species_columns[which.max(R2_variables)]
densite_min <- species_columns[which.min(R2_variables)]
print(densite_max)
print(densite_min)

R2_global <- sum(R2_variables) / length(R2_variables)
print(R2_global)

R2_partition <- sum(poids * rowSums(barycentres_types^2)) / 27
print(R2_partition)
```

```r

# Partie 2 :
rm(list = ls())
data <- read.csv("./Datagenus.csv", sep=";")
data2 <- data[1 :1000,] #On enlève la ligne qui pose problème

##### Rappel des résulats obtenu dans la partie 1 ####

# Sélectionner les colonnes des espèces
espece <- paste0("gen", 1:27)

# Calculer la densité de peuplement pour chaque espèce
densité <- data2[espece] / data2$surface
densité <- as.matrix(densité) # Convertir la dataframe densité en matrice
    densité
# Calcul de la moyenne des densités pour chaque espèce
moyenne_densité <- colMeans(densité)
#Centrons nos densité :
densités_centrées <- densité - matrix(moyenne_densité, nrow = nrow(densité),
```

```
    ncol = ncol(densité), byrow = TRUE)
#On calcul les écarts-types
ecarts_type_densité <- apply(densités_centrées, 2, sd)
#Enfin on standarise nos densités
densités_standardisées <- densités_centrées / matrix(ecarts_type_densité, nrow
    = nrow(densité), ncol = ncol(densité), byrow = TRUE)

##### Commencement partie 2 #####

# On créer les matrices du sujet

X <- as.matrix(densités_standardisées)

Y <- matrix(0, nrow = 1000, ncol = 7)
# Remplir la matrice Y
for (i in 1:nrow(Y)) {
  type <- data2$forest[i]  # Récupérer le type de la parcelle i
  if (type >= 1 && type <= 7) {  # Assurer que le type est entre 1 et 7
    Y[i, type] <- 1  # Mettre 1 dans la colonne correspondante
  }
}

Z <- matrix(0, nrow = 1000, ncol = 5)
# Remplir la matrice Z
for (i in 1:nrow(Y)) {
  type <- data2$geology[i]  # Récupérer le type de la parcelle i
  if (type==5){
    type =4
    Z[i, type] <- 1  # Mettre 1 dans la colonne correspondante
  }
  if (type==6){
    type = 5
    Z[i, type] <- 1  # Mettre 1 dans la colonne correspondante
  }
  else{
    Z[i, type] <- 1  # Mettre 1 dans la colonne correspondante
  }
}

W <- diag(1/1000, 1000)
M <- diag(1/27,27)

#Question 1 b)
# On programme la projection de Y et celle de x^j
P_Y <- Y %*% solve(t(Y)%*% W %*% Y) %*% t(Y) %*% W
P_X <- function(j){
  # Calculer la projection de la colonne j de X
  x_j <- X[, j, drop = FALSE]  # S'assurer que x_j reste une matrice
  return(x_j %*% solve(t(x_j) %*% W %*% x_j) %*% t(x_j) %*% W)
}
#On calcule la trace du produit matriciel avec j=25 par exemple.
Tr_1 <- sum(diag(P_X(25) %*% P_Y))


# Question 1 c)
#On pose R et on calcul la trace demandée
R <- X %*% M %*% t(X) %*% W
Tr_2 <- sum(diag(R %*% P_Y))

# Question 2
# On programme Z et on calcul les traces demandées
P_Z <- Z %*% solve(t(Z)%*% W %*% Z) %*% t(Z) %*% W
Tr_3 <- sum(diag(P_X(25) %*% P_Z))
```

```r
Tr_4 <- sum(diag(R %*% P_Z))
# Calcul de Tr_3 avec tout les j
L <- list()
for(i in 1:27){
  Tr_3 <- sum(diag(P_X(i) %*% P_Z))
  L[i] <- Tr_3
}
```