

TP1 ADM

MARIAC Damien, SCAIA Matteo

October 12, 2024



Contents

1	Introduction	3
2	Partie 1	3
2.1	inertie et barycentre	3
2.2	Autour des types forestiers	5
2.2.1	Calcul des poids des types forestiers	5
2.2.2	Calcul des barycentres des types forestiers	5
2.2.3	Calcul des normes euclidiennes carrées	5
2.2.4	Inertie inter-types et coefficient R^2	5
2.3	Liaisons	6
3	Partie 2	7
3.1	Enoncer	7
3.2	Rappel et calcul de projecteur	7
3.2.1	Rappel	7
3.2.2	Calcul de projecteur	8
4	Conclusion	9

1 Introduction

On dispose d'un jeu de données présentant une étude de 27 espèces d'arbres dans 1000 parcelles de la forêt du Congo. Nous avons enlevé une ligne problématique dans notre jeu de données (ligne 1000 code TGC). Il s'agit d'étudier la variabilité des densités de peuplement d'espèces arborées dans différentes parcelles de la forêt. Nous disposons dans notre jeu de données de 30 variables quantitatives dont : 27 variables de comptage des espèces, une sur la surface de la parcelle, une forestière et une géologique. Et d'une variable qualitative "code". On arrondira les valeurs à 10^{-3} .

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

Table 1: Extrait du jeu de données Datagenus

2 Partie 1

2.1 inertie et barycentre

Afin d'avoir une comparaison plus juste entre chaque parcelles, nous utiliserons la densité de peuplement de ces dernières. C'est à dire 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 poid de nos parcelles sont equipondérée.

Par consequents 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, la matrice transformée X' est définie par :

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}^j}{s_j}$$

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

$$\bar{x}^j = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i^j = \frac{1}{n} \sum_{i=1}^n \frac{x_i^j - \bar{x}^j}{s_j} = \frac{1}{s_j} \left(\frac{1}{n} \sum_{i=1}^n x_i^j \right) - \frac{\bar{x}^j}{s_j} = 0$$

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

Inertie totale égale à 27 : (preuve théorique)

Considérons la meme matrice de données \tilde{X} .

L'inertie de l'ensemble des points \tilde{X} par rapport à leur barycentre \mathbf{y} est définie par:

$$I_{Y,W} = \sum_{i=1}^n w_i \|\tilde{\mathbf{x}}_i - \mathbf{y}\|^2$$

Dans notre cas, le barycentre est nulle. 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:

$$I_{Y,W} = \sum_{i=1}^n w_i \|\tilde{\mathbf{x}}_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|\tilde{\mathbf{x}}_i\|^2$$

Comme chaque $\tilde{\mathbf{x}}_i$ est une observation centrée-réduite et la variance de chaque variable est 1, nous avons (pour i fixé):

$$\|\tilde{\mathbf{x}}_i\|^2 = \sum_{j=1}^p (\tilde{x}_i^j)^2 = p$$

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 :

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

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

Table 2: Tableau des poids forestier

2.2.2 Calcul des barycentres des types forestiers

Pour chaque type forestier i , (avec X la matrice des densités centrés 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:

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

Table 3: normes carrée

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

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

Table 4: Extrait du R^2 des variables

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 :

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

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

[1] 0.06891505
> R2_partition <- sum(poids * rowSums(barycentres_types^2)) / inertie_totale
> print(R2_partition)
[1] 0.06891505
> |
```

3 Partie 2

3.1 Enoncer

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.

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

Tout d'abord, il est important de rappeler que les x^j sont des variables qui sont centrées réduites. De plus, nous rappelons que

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

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

A partir de ces deux résultats, nous pouvons conclure que.

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

Nous venons de montrer que pour tout j ,

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

De plus, nous fixons j , et nous avons l'expression suivante.

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

Or, il est important de préciser que

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

Ici, $\overline{x^j}$ représente la moyenne globale des x^j , $\overline{x^{j^q}}$ représente la moyenne pondérée des x^j pour le type forestier q .

A partir de ces notations, nous avons l'expression suivante.

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

Nous pouvons conclure que $\|\Pi_Y x^j\|_W^2$ s'interprete statistiquement comme étant la mesure des 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 admet l'expression suivante.

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

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

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

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

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

Donc, nous pouvons conclure que $tr(\Pi_{x^j} \Pi_Y)$

4 Conclusion

```

1 rm(list = ls())
2
3 tab <- read.csv("C:/Users/damie/Desktop/MASTER/ADM/tp/tp1/
4   Datagenus.csv", sep = ";")
5 data <- tab[1 :1000,]
6 species_columns <- grep("gen", colnames(data), value = TRUE)
7 ##QUESTION 1
8
9 density_data <- data[species_columns] / data$surface
10
11 W <- diag(1/1000, 1000, 1000)
12 I_n <- rep(1, 1000)
13
14 Q <- as.matrix(density_data)
15 x_b <- t(Q) %*% W %*% I_n # moyenne (par projection)
16 x_c <- Q - I_n %*% t(x_b)
17
18 vec_norm <- sqrt(diag(t(x_c) %*% W %*% x_c))
19
20
21 x_cr <- sweep(x_c, 2, vec_norm, FUN = "/")
22 tableau_cr <- as.data.frame(x_cr)
23
24
25 bar <- t(x_cr) %*% W %*% I_n
26 barycentre <- t(bar)
27 print(barycentre)
28
29 inertie_totale <- sum(diag(t(x_cr) %*% W %*% x_cr)) #inertie
30 print(inertie_totale)
31
32 print(tableau_cr)
33
34 ##Question 2
35
36
37 forest_types <- unique(data$forest)
38 n_types <- length(forest_types)
39
40 poids <- numeric(n_types)
41 barycentres_types <- matrix(0, n_types, length(species_
42   columns))
43
44 for (i in 1:n_types) {
45   parcelles_type_i <- which(data$forest == forest_types[i])
46   n_i <- length(parcelles_type_i)
47   poids[i] <- n_i / 1000
48   W_i <- diag(1/n_i, n_i, n_i)

```

```

48   I_i <- rep(1, n_i)
49   Q_i <- as.matrix(x_cr[parcelles_type_i, ])
50   x_b_i <- t(Q_i) %*% W_i %*% I_i
51   barycentres_types[i, ] <- t(x_b_i)
52 }
53
54 normes_carre_barycentres <- rowSums(barycentres_types^2)
55
56 inertie_inter_types <- sum(poids * normes_carre_barycentres)
57 print(inertie_inter_types)
58
59 R2 <- inertie_inter_types / inertie_totale
60 print(R2)
61
62
63
64
65
66 forest_types <- unique(data$forest)
67 n_types <- length(forest_types)
68
69 R2_variables <- numeric(length(species_columns))
70
71 for (j in 1:length(species_columns)) {
72   variance_totale_j <- sum((x_cr[, j] - mean(x_cr[, j]))^2)
73   / 1000
74
75   variance_inter_j <- 0
76   for (i in 1:n_types) {
77     parcelles_type_i <- which(data$forest == forest_types[i])
78     n_i <- length(parcelles_type_i)
79     poids_i <- n_i / 1000
80
81     W_i <- diag(1/n_i, n_i, n_i)
82     I_i <- rep(1, n_i)
83
84     Q_i <- as.matrix(x_cr[parcelles_type_i, j])
85     barycentre_i <- t(Q_i) %*% W_i %*% I_i
86
87     variance_inter_j <- variance_inter_j + poids_i * as.
88       numeric(barycentre_i)^2
89   }
90
91   R2_variables[j] <- variance_inter_j / variance_totale_j
92 }
93
94 print(R2_variables)
95
96 densite_max <- species_columns[which.max(R2_variables)]

```

```

95 | densite_min <- species_columns[which.min(R2_variables)]
96 | print(densite_max)
97 | print(densite_min)
98 |
99 | R2_global <- sum(R2_variables) / length(R2_variables)
100 | print(R2_global)
101 |
102 | R2_partition <- sum(poids * rowSums(barycentres_types^2)) /
    | 27
103 | print(R2_partition)

```