L'ANALYSE EN COMPOSANTES PRINCIPALES : PRINCIPES ET APPLICATIONS

R. PALM*

RÉSUMÉ

Cette note décrit les principes de l'analyse en composantes principales et donne deux exemples numériques d'applications.

SUMMARY

This note describes the principles of principal component analysis and gives two examples.

1. INTRODUCTION

L'analyse en composantes principales est une technique descriptive permettant d'étudier les relations qui existent entre des variables quantitatives, sans tenir compte, *a priori*, d'une quelconque structure, ni des variables ni des individus.

Les domaines d'application de cette méthode sont très variés et de nombreux exemples sont proposés, notamment, par JACKSON [1991] et PRESS [1972].

L'objectif de cette note est de décrire et d'illustrer les principes de l'analyse en composantes principales. Nous examinerons d'abord comment on détermine les composantes (paragraphe 2) et nous détaillerons leur signification algébrique et géométrique (paragraphe 3). Nous décrirons alors les représentations graphiques (paragraphe 4). Le paragraphe 5 sera consacré à une application concrète et nous terminerons par quelques informations complémentaires (paragraphe 6).

Une présentation de l'analyse en composantes principales peut être trouvée dans la plupart des livres relatifs à l'analyse multivariée et l'ouvrage de JACKSON [1991] est entièrement consacré à ce sujet. Le lecteur y trouvera de nombreuses informations complémentaires ainsi qu'une importante bibliographie.

*Chargé de cours associé à la Faculté universitaire des Sciences agronomiques de Gembloux.

2. DÉFINITION DES COMPOSANTES PRINCIPALES

2.1. Objectifs poursuivis

Le point de départ d'une analyse en composantes principales est un tableau de données quantitatives de n lignes et p colonnes. Les lignes correspondent aux individus et les colonnes correspondent aux variables observées. Un tel tableau peut comporter un très grand nombre de cellules et on va s'efforcer de résumer les données de manière à prendre plus facilement connaissance de l'information qu'elles contiennent.

Le calcul de la moyenne et de l'écart-type donne, pour chaque variable, des informations concernant l'ordre de grandeur et la dispersion des données. De même, le calcul de la matrice de corrélation des variables donne des indications sur l'évolution simultanée des variables prises deux à deux. Ces éléments de statistique descriptive univariée et bivariée ne donnent cependant aucune information sur le problème lorsque les p variables sont considérées simultanément. Cette étude simultanée des variables est précisément le but de l'analyse en composantes principales.

De manière à rendre la présentation aussi concrète que possible, nous allons considérer un exemple relatif à trois variables et 16 individus. Les dimensions réduites de cet exemple n'offriront sans doute pas l'opportunité de mettre clairement en évidence l'utilité de l'analyse en composantes mais elles permettront de bien illustrer les principes de l'analyse notamment par l'établissement de nombreux tableaux. Une application plus pratique, sur un tableau de données plus grand, sera présentée au paragraphe 5.

L'exemple que nous allons examiner concerne les teneurs en protéines, en graisse et en lactose du lait de 16 mammifères. Ces données ont été publiées par HARTIGAN [1975] et sont reprises dans le tableau 1. L'examen de ce tableau montre que, bien que les trois constituants soient exprimés en pour cent, la variabilité des caractéristiques est assez différente. Comme nous le préciserons au paragraphe 6.1, il se justifie de standardiser les trois variables, afin de leur donner la même importance :

$$x_{ij} = (y_{ij} - \bar{y}_j)/\hat{\sigma}_j \quad (i = 1, \ldots, n;\ j = 1, \ldots, p).$$

Dans cette relation, $\bar{y}_j$ et $\hat{\sigma}_j$ sont, respectivement, la moyenne arithmétique et l'écart-type estimé de la colonne j. Les variables ainsi standardisées sont données dans le tableau 2.

Le principe de l'analyse en composantes principales est de définir des indices synthétiques qui résument au mieux l'information contenue dans ce tableau.

Le premier indice synthétique, z_{i1}, est défini de manière à respecter les contraintes suivantes :

– il doit être une combinaison linéaire des variables centrées réduites :

$$z_{i1} = u_{11}\, x_{i1} + u_{21}\, x_{i2} + u_{31}\, x_{i3}\,;$$

Tableau 1 – Teneurs en protéines, graisse et lactose du lait de 16 mammifères : résultats exprimés en pour cent (d'après HARTIGAN, 1975).

Code	Nom	Protéines	Graisse	Lactose
a	ânesse	1,7	1,4	6,2
b	baleine	11,1	21,2	1,6
c	biche	10,4	19,7	2,6
d	brebis	5,6	6,4	4,7
e	buffle	5,9	7,9	4,7
f	chamelle	3,5	3,4	4,8
g	cobaye	7,4	7,2	2,7
h	jument	2,6	1,0	6,9
i	lama	3,9	3,2	5,6
j	lapine	12,3	13,1	1,9
k	mule	2,0	1,8	5,5
l	rate	9,2	12,6	3,3
m	renarde	6,6	5,9	4,9
n	renne	10,7	20,3	2,5
o	truie	7,1	5,1	3,7
p	zèbre	3,0	4,8	5,3
	Moyennes	6,44	8,44	4,18
	Ecarts-types	3,50	6,87	1,60

– les coefficients u_{j1} qui interviennent dans cette combinaison linéaire doivent être tels que :
$$u_{11}^2 + u_{21}^2 + u_{31}^2 = 1\,;$$
– les coefficients u_{j1} doivent, en outre, être tels que la variance des z_{i1} soit maximum.

La seconde contrainte est une forme de normalisation, sans laquelle la troisième contrainte serait dénuée d'intérêt. On pourrait, en effet, toujours augmenter la variance des z_{i1}, en multipliant tous les coefficients u_{j1} par une constante arbitrairement grande.

La solution numérique du problème qui vient d'être posé est la suivante :
$$u_{11} = -0,585\,, \quad u_{21} = -0,569 \quad \text{et} \quad u_{31} = 0,578\,.$$
et nous verrons, au paragraphe 2.2, comment elle est obtenue.

Disposant de ces coefficients, on peut déterminer les valeurs de l'indice pour chacun des individus (tableau 3). Ainsi, par exemple, pour l'ânesse on a :
$$z_{11} = (-0,585)(-1,354) + (-0,569)(-1,024) + (0,578)(1,263) = 2,105\,.$$

Comme nous le préciserons par la suite (paragraphes 3.1 et 3.2), cette combinaison linéaire est celle qui résume au mieux le tableau 2. La variance

Tableau 2 – Teneurs en protéines, graisse et lactose du lait de 16 mammifères : données centrées et réduites.

Code	Nom	x_{i1}	x_{i2}	x_{i3}
a	ânesse	–1,354	–1,024	1,263
b	baleine	1,332	1,858	–1,615
c	biche	1,132	1,639	–0,990
d	brebis	–0,239	–0,297	0,325
e	buffle	–0,154	–0,078	0,325
f	chamelle	–0,839	–0,733	0,387
g	cobaye	0,275	–0,180	–0,927
h	jument	–1,096	–1,083	1,701
i	lama	–0,725	–0,762	0,888
j	lapine	1,675	0,679	–1,428
k	mule	–1,268	–0,966	0,825
l	rate	0,789	0,606	–0,551
m	renarde	0,046	–0,369	0,450
n	renne	1,218	1,727	–1,052
o	truie	0,189	–0,486	–0,301
p	zèbre	–0,982	–0,529	0,700

de cette combinaison linéaire est égale à 2,801. Comme nous le préciserons au paragraphe 3, le rapport entre la variance de l'indice et la somme des variances des trois variables du tableau 2, qui est égale à trois du fait de la standardisation, représente la part de la variance des variables initiales qui est prise en compte par la combinaison linéaire. Cette part est égale à :

$$2,801/3 = 0,93 \text{ ou } 93\ \%\,.$$

Si on souhaite prendre en compte une part plus grande encore de la variabilité des colonnes du tableau 2, il faut définir un deuxième indice, z_{i2}, déterminé de la manière suivante :

– il doit être une combinaison linéaire des variables centrées réduites :

$$z_{i2} = u_{12}\, x_{i1} + u_{22}\, x_{i2} + u_{32}\, x_{i3}\,;$$

– les coefficients u_{j2} qui interviennent dans cette combinaison linéaire doivent être tels que :

$$u_{12}^2 + u_{22}^2 + u_{32}^2 = 1 \quad \text{et} \quad u_{11}\, u_{12} + u_{21}\, u_{22} + u_{31}\, u_{32} = 0\,;$$

– les coefficients u_{j2} doivent être tels que la variance des z_{i2} soit maximum tout en respectant les contraintes ci-dessus.

Par rapport à la définition du premier indice synthétique, une contrainte supplémentaire a été ajoutée : la somme des produits des coefficients des deux

combinaisons linéaires est nulle, ce qui se traduira par la non corrélation entre les valeurs du premier indice et les valeurs du second indice : le second indice contiendra donc une information non redondante par rapport au premier indice.

La solution du problème est la suivante :

$$u_{12} = -0,233\,, \quad u_{22} = 0,801 \quad \text{et} \quad u_{32} = 0,552\,.$$

et on peut, comme ci-dessus, calculer les valeurs de cet indice pour chaque individu. Les résultats sont repris dans le tableau 3. La variance de ce second indice est égale à 0,142 et la part de la variabilité des variables centrées réduites prise en compte par cet indice est de :

$$0,142/3 = 0,047 \quad \text{ou} \quad 5\%\,.$$

La prise en compte simultanée des deux indices permet donc de retrouver 98 % de la variabilité des variables x_1, x_2 et x_3.

On peut encore calculer un troisième indice, z_{i3}, en rajoutant aux contraintes équivalentes à celles imposées lors du calcul de z_{i1} les deux contraintes suivantes :

$$u_{11}\,u_{13} + u_{21}\,u_{23} + u_{31}\,u_{33} = 0 \quad \text{et} \quad u_{12}\,u_{13} + u_{22}\,u_{23} + u_{32}\,u_{33} = 0\,,$$

qui assurent la non-corrélation des z_{i3} et z_{i1} d'une part, et des z_{i3} et z_{i2} d'autre part. La solution obtenue s'écrit :

$$u_{13} = 0,777\,; \quad u_{23} = -0,188 \quad \text{et} \quad u_{33} = 0,601\,.$$

La variance des z_{i3} vaut 0,057 ; elle correspond à 2 % de la variance cumulée des variables x_1, x_2 et x_3. Les valeurs z_{i3} sont données dans le tableau 3.

De façon plus générale, pour un tableau de données comportant p colonnes, on pourra définir p indices synthétiques, d'importance décroissante et non corrélés. Ces indices sont appelés scores ou valeurs des composantes principales. La méthode de calcul de ces composantes est précisée au paragraphe 2.2.

En pratique, on ne s'intéressera le plus souvent qu'aux premières composantes principales qui, comme nous venons de le voir, contiennent l'essentiel de l'information. Ce point sera développé aux paragraphes 2.3 et 6.2.

2.2. Valeurs et vecteurs propres de la matrice de corrélation

Au paragraphe précédent, nous avons donné les principes à la base de la détermination des valeurs des composantes principales, dans le cas de trois variables. Nous allons maintenant préciser comment on détermine numériquement les coefficients qui interviennent dans ces combinaisons linéaires et ce, quel que soit le nombre de variables initiales.

Le point de départ est la matrice de corrélation, $\boldsymbol{R}$, des p variables initiales. Cette matrice carrée est de dimensions $p \times p$, de rang r ($r \leq p$) et admet r valeurs propres positives :

$$l_1 \geq l_2 \geq \ldots \geq l_r\,,$$

Tableau 3 – Valeurs des composantes principales.

Code	Nom	z_{i1}	z_{i2}	z_{i3}
a	ânesse	2,105	0,193	–0,100
b	baleine	–2,770	0,285	–0,285
c	biche	–2,167	0,502	–0,023
d	brebis	0,496	–0,002	0,065
e	buffle	0,322	0,152	0,090
f	chamelle	1,132	–0,177	–0,282
g	cobaye	–0,594	–0,720	–0,310
h	jument	2,241	0,328	0,374
i	lama	1,371	0,049	0,113
j	lapine	–2,191	–0,636	0,316
k	mule	1,768	–0,022	–0,308
l	rate	–1,125	–0,004	0,168
m	renarde	0,443	–0,058	0,376
n	renne	–2,303	0,517	–0,010
o	truie	–0,008	–0,599	0,057
p	zèbre	1,280	0,192	–0,243

auxquelles sont associés r vecteurs propres $\boldsymbol{u}_1\,, \boldsymbol{u}_2\,, \ldots, \boldsymbol{u}_r$.

Les valeurs propres sont les variances des valeurs des composantes principales :

$$s^2_{z_1} = l_1\,, \ldots\,, s^2_{z_r} = l_r\,,$$

et, lorsqu'il est normé à l'unité, le vecteur propre $\boldsymbol{u}_j$, relatif à une valeur l_j, a comme éléments les coefficients de la combinaison linéaire de la composante z_j :

$$\mathbf{u}_j = \begin{bmatrix} \mathbf{u}_{1j} \\ \mathbf{u}_{2j} \\ \vdots \\ \mathbf{u}_{pj} \end{bmatrix}\,.$$

Les n valeurs de la composante, z_{ij}, s'obtiennent par le produit matriciel suivant :

$$\boldsymbol{z}_j = \boldsymbol{X}\,\boldsymbol{u}_j\,,$$

$\boldsymbol{X}$ étant la matrice des données centrées et réduites.

En utilisant les matrices partitionnées, on peut aussi écrire :

$$\boldsymbol{Z} = (\boldsymbol{z}_1 \ldots \boldsymbol{z}_r) = \boldsymbol{X}\,(\boldsymbol{u}_1 \ldots \boldsymbol{u}_r) = \boldsymbol{X}\,\boldsymbol{U}\,,$$

où $\boldsymbol{Z}$ est la matrice obtenue par la juxtaposition des r colonnes $\boldsymbol{z}_j$ et $\boldsymbol{U}$ est la matrice obtenue par la juxtaposition des r vecteurs propres $\boldsymbol{u}_j$.

Tableau 4 – Corrélations entre variables initiales et corrélations des variables initiales avec les composantes principales.

Variables	x_1	x_2	x_3	z_1	z_2	z_3
x_1	1,000	0,897	–0,938	–0,979	–0,088	0,186
x_2	0,897	1,000	–0,865	–0,952	0,301	–0,045
x_3	–0,938	–0,865	1,000	0,968	0,208	0,144

Le plus souvent, le rang r de la matrice de corrélation est égal à p. Toutefois si les variables initiales présentent des combinaisons linéaires ou si le nombre de variables est supérieur au nombre d'observations, r sera inférieur à p. Ce sera, par exemple, le cas si une variable est une transformation linéaire d'une autre variable ou, dans le cas de pourcentages (ou de proportions) lorsqu'une variable est obtenue en calculant le complément à 100 (ou à l'unité) de la somme d'une série d'autres variables.

Si on multiplie les éléments d'un vecteur propre $\boldsymbol{u}_j$ par la racine carrée de la valeur propre correspondante, on obtient la corrélation de la composante principale z_j avec chacune des variables initiales. Ces corrélations seront utiles pour préciser la part de la variance d'une variable donnée prise en compte par une composante principale particulière (paragraphe 3.1) et seront utilisées pour les représentations graphiques des variables dans les cercles de corrélation (paragraphe 4.2).

Pour l'exemple considéré, les corrélations de la première composante avec les trois variables initiales sont égales à :

$$r_{z_1x_1} = (-0,585)(\sqrt{2,801}) = -0,979\,; \quad r_{z_1x_2} = (-0,569)(\sqrt{2,801}) = -0,952$$

et
$$r_{z_1x_3} = (0,578)(\sqrt{2,801}) = 0,968\,,$$

et on pourrait, de la même manière, calculer la corrélation de z_2 et de z_3 avec les trois variables initiales. Les résultats sont repris dans le tableau 4, de même que les corrélations entre les variables initiales.

3. SIGNIFICATION ALGÉBRIQUE ET GÉOMÉTRIQUE

3.1. Reconstitution du tableau de départ

Les vecteurs propres u_j sont orthogonaux et de normes unitaires. En effet, le caractère orthogonal se justifie par la nullité du produit scalaire de deux vecteurs propres $\boldsymbol{u}_j$ et $\boldsymbol{u}_{j'}$:

$$u_{1j}\,u_{1j'} + \ldots + u_{pj}\,u_{pj'} = 0$$

et le fait que la norme soit unitaire résulte de la standardisation adoptée :

$$u_{1j}^2 + u_{2j}^2 + \ldots + u_{pj}^2 = 1\,.$$

Si $r = p$, la matrice $\boldsymbol{U}$, obtenue en juxtaposant les p vecteurs propres est dite orthogonale et possède la propriété suivante :

$$\boldsymbol{U}\,\boldsymbol{U}' = \boldsymbol{U}'\,\boldsymbol{U} = \boldsymbol{I}\,,$$

$\boldsymbol{I}$ étant la matrice identité, de dimensions $p \times p$.

En utilisant cette propriété, et en partant de la relation matricielle donnant $\boldsymbol{Z}$ on a, en postmultipliant par $\boldsymbol{U}'$:

$$\boldsymbol{Z}\,\boldsymbol{U}' = \boldsymbol{X}\,\boldsymbol{U}\,\boldsymbol{U}' = \boldsymbol{X}\,.$$

Cette relation montre que les données initiales, $\boldsymbol{X}$, peuvent être retrouvées à partir des vecteurs propres et des valeurs des composantes principales.

En partitionnant les matrices $\boldsymbol{Z}$ et $\boldsymbol{U}'$, on obtient :

$$\boldsymbol{Z}\,\boldsymbol{U}' = (\boldsymbol{z}_1 \ldots \boldsymbol{z}_r) \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_r' \end{pmatrix} = \boldsymbol{z}_1\,\boldsymbol{u}_1' + \ldots + \boldsymbol{z}_r\,\boldsymbol{u}_r' = \boldsymbol{X}_1 + \ldots + \boldsymbol{X}_r = \boldsymbol{X}\,.$$

La matrice $\boldsymbol{X}$ des données initiales peut donc être reconstituée par la somme de r matrices, chacune de celles-ci étant liée à une composante principale.

Le tableau 5 reprend les résultats de la reconstitution des variables initiales, d'une part, à partir de la première composante et, d'autre part, à partir des deux premières composantes principales. La première partie correspond donc aux colonnes de $\boldsymbol{X}_1$ et la seconde partie correspond aux colonnes de $\boldsymbol{X}_1 + \boldsymbol{X}_2$. Le tableau donne également les sommes des carrés des différentes colonnes et on peut vérifier que les moyennes de ces colonnes sont, aux erreurs d'arrondis près, toutes nulles.

Si on exprime les sommes des carrés des variables reconstituées à partir de z_1 en proportion de la somme des carrés des variables initiales, égale à 15, on obtient les valeurs suivantes :

$$14,366/15 = 0,958\,, \quad 13,608/15 = 0,907 \quad \text{et} \quad 14,043/15 = 0,936\,.$$

Ces valeurs sont égales aux carrés des coefficients de corrélation de z_1 et x_1, de z_1 et x_2 et de z_1 et x_3, que nous avons donnés au tableau 4.

Ces valeurs indiquent, pour chacune des variables initiales, la part d'information qui est conservée quand on se limite, dans l'interprétation, à une seule composante, le complément à l'unité étant la part d'information qui est perdue si on néglige les deux dernières composantes. On peut constater que, pour cet exemple, la première composante permet de reconstituer plus de 90 % de l'information contenue dans les variables initiales, cette proportion atteignant même 96 % pour la première variable.

Tableau 5 – Reconstitution de $\mathbf{X}$ à partir de la première composante et à partir des deux premières composantes.

Code	Première composante			Première et deuxième composantes		
	x_1	x_2	x_3	x_1	x_2	x_3
a	–1,231	–1,198	1,217	–1,276	–1,043	1,323
b	1,620	1,576	–1,601	1,553	1,804	–1,444
c	1,267	1,233	–1,253	1,150	1,635	–0,976
d	–0,290	–0,282	0,287	–0,290	–0,284	0,286
e	–0,188	–0,183	0,186	–0,224	–0,061	0,270
f	–0,662	–0,644	0,654	–0,620	–0,786	0,556
g	0,347	0,338	–0,343	0,515	–0,238	–0,741
h	–1,310	–1,275	1,295	–1,387	–1,012	1,477
i	–0,802	–0,780	0,793	–0,813	–0,741	0,820
j	1,281	1,247	–1,267	1,430	0,738	–1,617
k	–1,034	–1,006	1,022	–1,029	–1,024	1,010
l	0,658	0,640	–0,650	0,659	0,637	–0,652
m	–0,259	–0,252	0,256	–0,245	–0,299	0,224
n	1,347	1,311	–1,331	1,226	1,725	–1,046
o	0,005	0,005	–0,005	0,145	–0,475	–0,336
p	–0,749	–0,729	0,740	–0,793	–0,575	0,846
Somme des carrés	14,366	13,608	14,043	14,482	14,970	14,690

Si on additionne les sommes des carrés des trois premières colonnes du tableau 5 et qu'on divise le résultat par la somme des sommes des carrés des trois variables initiales on obtient le résultat suivant :

$$(14,366 + 13,608 + 14,043)/45 = (0,958 + 0,907 + 0,936)/3 = 2,801/3 = 0,934\,.$$

En moyenne, pour les trois variables initiales, la part d'information prise en compte par la première composante est égale à 93 %, cette valeur étant égale au rapport de la première valeur propre à la somme des valeurs propres. Cette valeur montre que $\boldsymbol{X}_1$ est une bonne approximation de $\boldsymbol{X}$, puisque $\boldsymbol{X}_1$, contient 93 % de l'information de $\boldsymbol{X}$.

Les calculs ci-dessus peuvent également être effectués pour la deuxième et la troisième composante. On constaterait, par exemple, que la deuxième composante explique 0,8 % de la première variable, et que la troisième composante explique 3,5 % de cette même variable. On constaterait aussi que $\boldsymbol{X}_2$ contient 5 % et que $\boldsymbol{X}_3$ contient 2 % de l'information de $\boldsymbol{X}$, comme nous l'avons déjà signalé au paragraphe 2.1.

Si maintenant on prend en considération les deux premières composantes, l'information prise en compte est égale à la somme de l'information prise en compte par la première composante et par la deuxième composante. Ainsi, pour

la teneur en protéines, les deux composantes expliquent :

$$(-0,979)^2 + (-0,088)^2 = 0,958 + 0,008 = 0,966 \quad \text{soit} \quad 96,6\%\,.$$

Aux erreurs d'arrondis près, cette valeur est égale à la somme des carrés des éléments de la première colonne de $\boldsymbol{X}_1 + \boldsymbol{X}_2$ (tableau 5), divisée par 15 :

$$14,482/15 = 0,965\,.$$

Globalement aussi, les deux premières composantes permettent de retrouver :

$$(2,801 + 0,142)/3 = 0,981 \quad \text{ou} \quad 98,1\%$$

de l'information contenue dans $\boldsymbol{X}$. La somme $\boldsymbol{X}_1 + \boldsymbol{X}_2$ donne donc une très bonne approximation de la matrice $\boldsymbol{X}$.

Pour synthétiser, on peut donc dire que :

- la part de l'information contenue dans une composante z_j et relative à une variable initiale x_k est égale au carré du coefficient de corrélation de z_j et x_k ;

- la part de l'information contenue dans une composante z_j pour l'ensemble des variables est égale à la moyenne des proportions de l'information relative à chacune des variables initiales ou encore à la valeur propre l_j, exprimée en proportion de la somme des valeurs propres.

3.2. Définition d'un nouveau système d'axes

Dans les paragraphes qui précèdent, nous avons examiné l'analyse en composantes principales sous un angle algébrique. Le problème peut cependant aussi être envisagé sous l'aspect géométrique.

Le lait des 16 mammifères peut être représenté dans un espace à trois dimensions, chacune des dimensions représentant une variable initiale. Le nuage des 16 points est centré sur l'origine du fait de la standardisation.

Si on place, dans l'espace à trois dimensions, un nouvel axe gradué passant par l'origine et qu'on calcule la somme des carrés des projections des 16 points sur cette droite, la valeur de cette somme des carrés dépendra de la position de la droite considérée car le nuage des points présente une direction d'allongement nettement privilégiée, du fait des corrélations importantes qui existent entre les variables initiales. Le premier axe principal est précisément l'axe qui maximise la somme des carrés des projections des 16 points sur l'axe et les valeurs des projections sont les valeurs de la première composante principale, z_{i1}. Quant aux coefficients directeurs de l'axe, ils sont égaux aux éléments du premier vecteur propre de la matrice de corrélation.

Le deuxième axe principal est un axe gradué passant par l'origine et perpendiculaire au premier axe principal dont la position est telle que, compte tenu

de la contrainte ci-dessus, la somme des carrés de projections est maximale. Les projections des points sur cet axe sont les valeurs de la deuxième composante, z_{i2}.

Enfin, le troisième axe principal est perpendiculaire au plan formé par les deux premiers axes principaux et les projections des points sur cet axe sont les valeurs de la troisième composante, z_{i3}.

Géométriquement donc, on remplace le système de coordonnées initiales par un nouveau système d'axes, z_1 z_2 et z_3. Cette opération ne modifie pas les distances des points par rapport à l'origine des axes. En effet, pour le premier individu par exemple, le carré de la distance du point à l'origine est égal, pour les axes initiaux, à :

$$x_{11}^2 + x_{12}^2 + x_{13}^2 = (-1,354)^2 + (-1,024)^2 + (1,263)^2 = 4,477\,,$$

et pour les nouveaux axes, à :

$$z_{11}^2 + z_{12}^2 + z_{13}^2 = 2,105^2 + 0,193^2 + (-0,100)^2 = 4,478\,.$$

Aux erreurs d'arrondis près, on trouve bien la même valeur.

L'intérêt de ce changement d'axes est que les nouveaux axes sont d'importance décroissante et que les projections des points sur ces axes sont non corrélées.

Le fait que, pour l'exemple considéré, la première composante représente 93 % de l'information signifie que, dans l'espace à trois dimensions, les points sont très concentrés autour du premier axe principal. Si on prend les deux premières composantes en considération, la qualité de la représentation est de 98 % : les points sont donc approximativement situés dans le sous-espace z_1, z_2, qui est un sous-espace de x_1 x_2 et x_3 ou, ce qui revient au même, de z_1 z_2 et z_3.

L'interprétation géométrique donnée ci-dessus pour trois variables et 16 individus peut évidemment être étendue à un nombre quelconque de variables et d'individus : les composantes principales sont des axes perpendiculaires correspondant aux directions dans lesquelles la variabilité est la plus grande et les projections des individus sur ces axes correspondent aux valeurs des composantes principales.

4. REPRÉSENTATIONS GRAPHIQUES

4.1. Cercles des corrélations

Les cercles des corrélations sont des graphiques visant à représenter géométriquement les variables dans le nouveau système de coordonnées.

Pour l'exemple considéré, seule la représentation des trois variables initiales dans le plan formé par les axes z_1 et z_2 et appelé premier plan factoriel est utile, compte tenu de l'importance de ces deux axes dans la reconstitution des variables (figure 1).

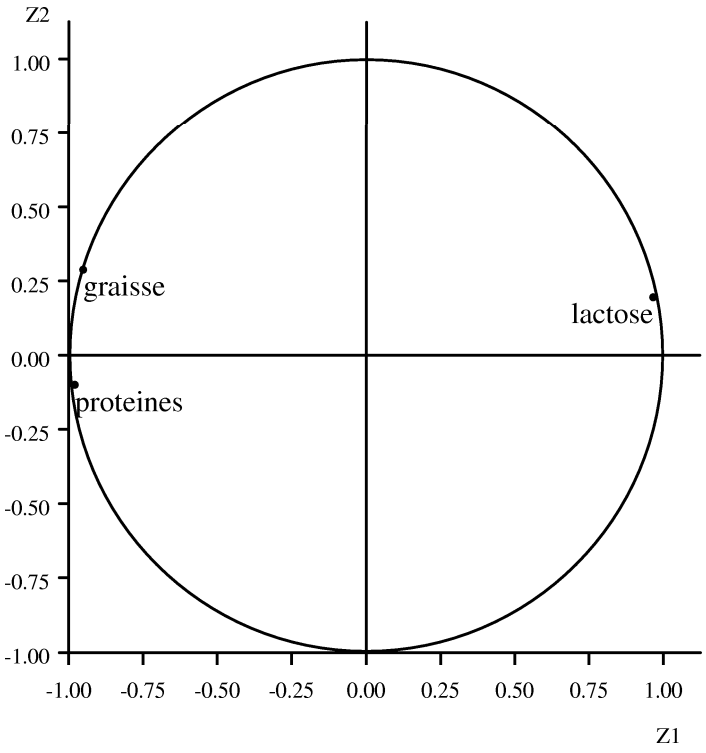

Figure 1 – Cercle de corrélation dans le plan formé par z_1 et z_2.

Les coordonnées des variables initiales sur z_1 sont les corrélations des variables avec z_1 soit, $-0,979$ pour x_1, $-0,952$ pour x_2 et $0,968$ pour x_3. De même, les coordonnées sur z_2 sont les corrélations des variables avec z_2, soit $-0,088$ pour x_1, $0,301$ pour x_2 et 0,208 pour x_3. Ces corrélations ont été données dans le tableau 4.

Il s'agit en réalité d'une représentation déformée, car on a projeté, dans un espace à deux dimensions, trois points qui se situent dans un espace à trois dimensions, la troisième coordonnée étant la corrélation des variables avec z_3. Pour l'exemple considéré, la déformation est cependant très peu importante car les corrélations avec z_3 sont très faibles.

Les trois points-variables se trouvent en réalité sur une sphère de rayon unitaire. Nous avons vu, en effet, au paragraphe 3.1, que le carré de la corrélation d'une variable, par exemple x_1, avec z_1 correspond à la proportion de la variance de x_1 prise en considération par z_1. Nous avons vu également que toute la variance de x_1 est restituée si on prend en considération les trois composantes, la somme des carrés des trois corrélations étant égale à l'unité :

$$(-0,979)^2 + (-0,088)^2 + (0,186)^2 = 1\,.$$

Il en va de même pour x_2 et x_3. Les trois points dans l'espace à trois dimensions sont donc à une distance unitaire de l'origine des axes : ils sont donc situés sur une sphère.

Les projections de ces points dans un plan se situent nécessairement sur ou à l'intérieur d'un cercle de rayon unitaire. C'est la raison pour laquelle ce type de représentation graphique est dénommé cercle de corrélation.

D'autre part, plus un point-variable est proche du cercle, plus la qualité de la représentation de la variable est bonne, car, nécessairement, la troisième coordonnée sera faible. Autrement dit aussi, un point proche du cercle correspond à une variable qui est bien reconstituée par les composantes retenues. A l'inverse, une variable plus éloignée du cercle est une variable pour laquelle la troisième composante joue un rôle plus important dans la reconstitution de la variable.

D'une manière plus générale, si le rang de la matrice de corrélation est supérieur à trois, les points-variables se trouvent sur une hypersphère de rayon unitaire et on cherche à se faire une idée de la position relative des points sur cette hypersphère par une série de projections dans des plans factoriels et, lors de l'interprétation de la proximité de deux ou plusieurs points, on tient compte de la qualité des représentations des points dans ces plans : deux points proches dans un plan ne sont pas nécessairement proches sur l'hypersphère, sauf si, dans ce plan, ils sont proches du cercle de rayon unitaire.

Les cercles de corrélation sont des éléments importants pour l'interprétation des données. Ils permettent parfois de donner une interprétation physique à certaines composantes principales. Ainsi, l'examen de la figure 1 fait apparaître une opposition entre la teneur en lactose d'une part (corrélation positive importante avec le premier axe) et la teneur en protéines et la teneur en graisse d'autre part (corrélations négatives importantes avec le premier axe) ; l'axe 1 est donc un axe de richesse en lactose et de pauvreté en protéines et graisse. L'interprétation du deuxième axe est plus difficile et, sans doute sans intérêt, car les corrélations sont nettement moins marquées. On notera simplement que cet axe est lié à la richesse en graisse et en lactose.

4.2. Graphique des individus

Nous avons vu, au paragraphe 3.2, que les points correspondant aux 16 mammifères sont concentrés à proximité du plan formé par les axes z_1 et z_2. On va donc tout naturellement réaliser des représentations graphiques des individus dans ce plan. Les coordonnées de chaque point sont les valeurs de la composante z_{i1} et de la composante z_{i2} (figure 2).

On constate que la dispersion des points selon z_1 est beaucoup plus importante que la dispersion des points selon z_2. En effet, la variance des z_{i1} est de 2,801 alors que la variance des z_{i2} n'est que de 0,142. A ce sujet, on sera attentif aux éventuelles distorsions du graphique qui peuvent résulter de l'utilisation des logiciels, lorsque les procédures standards d'établissement de diagrammes de dispersion prennent en compte des longueurs différentes pour les unités des deux axes.

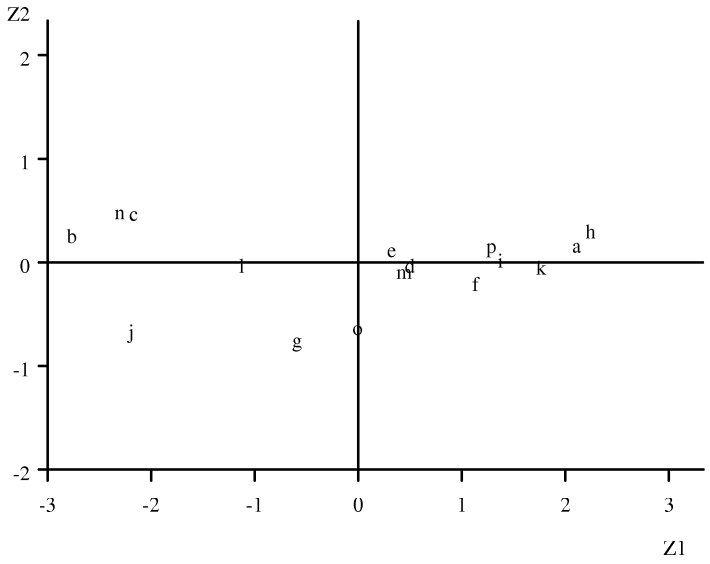

Figure 2 – Représentation des individus dans le premier plan factoriel.

Il ne faut pas perdre de vue que les points sont projetés dans le plan factoriel alors qu'en réalité ils sont situés dans un espace à trois dimensions. Cette troisième dimension est négligeable car, en moyenne (quadratique), la coordonnée des points sur ce troisième axe est égale à la racine carrée de la troisième valeur propre, soit approximativement 0,24. Cette coordonnée moyenne (quadratique) peut cependant cacher des disparités, car, comme le montre le tableau 3, elle varie en fait de 0,010 à 0,376.

En pratique, pour évaluer la qualité de la représentation d'un point projeté dans un sous-espace, ici le sous-espace formé par les axes z_1 et z_2, on détermine le rapport entre les carrés de deux distances. La première distance est la distance euclidienne, par rapport à l'origine, du point projeté dans le sous-espace et la deuxième distance est la distance euclidienne, par rapport à l'origine, du point dans l'espace complet. Ainsi par exemple pour l'ânesse, le carré de la première distance est égal à :

$$d_{12}^{2} = z_{11}^{2} + z_{12}^{2} = 2,105^{2} + 0,193^{2} = 4,468\,,$$

et le carré de la deuxième distance est égal à :

$$d_{123}^{2} = z_{11}^{2} + z_{12}^{2} + z_{13}^{2} = 2,105^{2} + 0,193^{2} + 0,100^{2} = 4,478\,.$$

Le rapport, désigné par le sigle cos^2, est donc égal à :

$$4,468/4,478 = 0,998\,.$$

Tableau 6 – Valeur des cosinus carrés, en %, pour les deux premiers axes et pour le premier plan factoriel.

Code	Nom	Axe 1	Axe 2	Plan 1-2
a	ânesse	98,9	0,8	99,8
b	baleine	97,9	1,0	99,0
c	biche	94,9	5,1	100,0
d	brebis	98,3	0,0	98,3
e	buffle	76,8	17,2	93,9
f	chamelle	92,0	2,3	94,3
g	cobaye	36,5	53,6	90,1
h	jument	95,3	2,0	97,3
i	lama	99,2	0,1	99,3
j	lapine	90,5	7,6	98,1
k	mule	97,0	0,0	97,1
l	rate	97,8	0,0	97,8
m	renarde	57,6	1,0	58,6
n	renne	95,2	4,8	100,0
o	truie	0,0	99,1	99,1
p	zèbre	94,5	2,1	96,6

Il s'agit en fait du cosinus carré de l'angle formé par le segment reliant l'origine des axes au point projeté dans le plan (z_1,z_2) et par le segment reliant l'origine des axes au point dans l'espace complet.

La qualité de la représentation d'un point dans le plan peut d'ailleurs être obtenue en additionnant les cosinus carrés relatifs aux deux axes. Ainsi, pour l'ânesse, le cosinus carré pour le premier axe vaut :

$$z_{11}^2 / \left(z_{11}^2 + z_{12}^2 + z_{13}^2\right) = 2,105^2/4,478 = 0,990\,,$$

et le cosinus carré pour le deuxième axe vaut :

$$z_{12}^2 / \left(z_{11}^2 + z_{12}^2 + z_{13}^2\right) = 0,193^2/4,478 = 0,008\,,$$

et la somme des deux cosinus carrés vaut bien 0,998.

Le tableau 6 donne les cosinus carrés, d'une part, sur chacun des deux premiers axes et, d'autre part, sur le premier plan factoriel. On constate que tous les individus sont très bien représentés dans le premier plan factoriel, à l'exception de la renarde, pour laquelle la troisième composante est non négligeable, relativement aux deux premières, comme le montre le tableau 3, même si, dans l'absolu, cette composante n'est pas très importante puisqu'elle vaut 0,376.

Compte tenu de l'interprétation qui a été faite du premier axe au paragraphe précédent, les animaux sont classés sur l'axe 1 en fonction de la richesse de leur lait en lactose et de la pauvreté en protéines et en graisse. Ainsi, on trouve donc à une extrémité la jument, l'ânesse et la mule dont les laits sont

caractérisés par des teneurs importantes en lactose et des teneurs faibles en protéines et graisse. A l'opposé, le renne, la biche, la baleine et la lapine ont un lait riche en protéines et en graisse mais pauvre en lactose. La différence entre la lapine et les trois autres animaux cités est liée à la plus grande pauvreté en graisse et apparaît lorsqu'on considère le deuxième axe.

Dans le cas d'un problème de dimensions plus importantes et si on souhaite prendre en compte plus de deux composantes principales, on réalise une série de représentations graphiques à deux dimensions, de manière à se faire une idée de la situation des points dans un espace à plus de deux dimensions, comme nous le verrons au paragraphe 5.

4.3. Variables et individus supplémentaires

Il peut arriver qu'on souhaite représenter une ou plusieurs variables dans les cercles de corrélation, alors que ces variables n'ont pas été prises en compte lors du calcul des composantes principales. De telles variables s'appellent variables supplémentaires ou variables passives. Pour positionner une variable supplémentaire dans les cercles de corrélation, il suffit de calculer la corrélation de cette variable avec les composantes principales qui ont été calculées à partir des autres variables, dites actives.

Diverses raisons peuvent justifier l'existence d'une variable supplémentaire. Il peut s'agir d'une variable particulière, de nature un peu différente des autres variables. Ainsi, PHILIPPEAU [1986], dans une étude de variétés de froment, considère le rendement comme variable supplémentaire, les autres variables ayant trait à des caractéristiques relatives au développement de la culture (hauteur de l'épi à une date donnée, date d'épiaison, coefficient de tallage, etc.).

Une variable peut aussi être traitée en variable supplémentaire parce qu'elle présente des données manquantes pour un grand nombre d'individus. En effet, les logiciels éliminent systématiquement de l'analyse les individus présentant une ou plusieurs données manquantes. Si, pour une variable, de nombreuses données sont manquantes, alors que les autres variables sont à peu près complètes, la prise en compte de la variable en question pour la définition des composantes imposerait la suppression de tous les individus pour lesquels les données sont manquantes, ce qui conduirait à un appauvrissement injustifié des données. Cette façon de procéder permet de conserver tous les individus ayant des données complètes pour les autres variables, sans toutefois éliminer tout à fait la variable incomplète.

Pour l'exemple relatif à la composition du lait de 16 mammifères, nous allons considérer comme variable supplémentaire la teneur en eau du lait. Cette variable est donnée par HARTIGAN [1975], en même temps que les trois autres variables qui ont été analysées. Ce n'est que pour simplifier la présentation de l'analyse en composantes principales que la teneur en eau a été négligée jusqu'à présent. Les données relatives à cette variable ne sont pas reprises ici, mais nous avons calculé les coefficients de corrélation de cette variable avec z_1 et z_2. Nous avons obtenu 0,972 pour z_1 et $-0,218$ pour z_2.

Le point relatif à cette variable serait donc situé légèrement en-dessous du point relatif au lactose, puisque sur l'axe 1 les coordonnées sont pratiquement identiques tandis que sur l'axe 2, elles sont du même ordre de grandeur, mais de signe opposé.

De façon analogue, on peut positionner, dans le graphique des individus, un ou plusieurs individus supplémentaires. Ici aussi, différentes raisons peuvent justifier la présence de ces individus supplémentaires. Un exemple typique est le cas où les individus se répartissent en groupes, en fonction d'un critère qualitatif qui n'a pas été pris en compte dans l'analyse. Pour mieux apprécier l'effet de ce critère, on peut reporter sur le graphique des individus, des individus artificiels qui représentent les individus moyens des groupes.

On peut aussi porter en individus supplémentaires des individus différents de l'ensemble de manière à ce que la position des axes ne soit pas influencée par ces individus atypiques. Si ceux-ci sont trop différents des autres, leur représentation graphique risque cependant d'être sans intérêt, dans la mesure où ils se situeront nettement en dehors du nuage de points.

Pour positionner un individu supplémentaire, on calcule la valeur des composantes principales pour l'individu en question par la relation suivante :

$$z_{sj} = \boldsymbol{x}_s \, \boldsymbol{u}_j \, ,$$

$\boldsymbol{x}_s$ étant le vecteur contenant les observations (centrées et réduites) de l'individu s et $\boldsymbol{u}_j$ étant le $j^{ième}$ vecteur propre.

A titre d'illustration, considérons un individu supplémentaire dont le lait serait caractérisé par une teneur en protéines de 2,1 %, une teneur en graisse de 1,4 % et une teneur en lactose de 6,2 %.

Les valeurs des variables centrées réduites sont donc :

$$x_{s1} = (2,1 - 6,44)/3,50 = -1,239 \, ,$$
$$x_{s2} = (1,4 - 8,44)/6,87 = -1,024 \, ,$$
et $$x_{s3} = (6,2 - 4,18)/1,60 = 1,263 \, ,$$

et les valeurs des composantes seraient :

$$z_{s1} = (-0,585)(-1,239) + (-0,569)(-1,024) + (0,578)(1,263) = 2,038 \, ,$$
$$z_{s2} = (-0,233)(-1,239) + (0,801)(-1,024) + (0,552)(1,263) = 0,167 \, ,$$
et $$z_{s3} = (0,777)(-1,239) + (-0,188)(-1,024) + (0,601)(1,263) = -0,011 \, .$$

Dans le premier plan factoriel, l'individu supplémentaire se situerait à proximité du point représentant l'ânesse.

5. ANALYSE D'UN PROBLÈME PRATIQUE

5.1. Nature des données

Dans le cadre d'une étude sur la qualité du bois de hêtre, LECLERCQ [1979] a sélectionné 68 arbres dans lesquels il a débité des morceaux de bois, appelés

éprouvettes, afin de déterminer, selon des normes bien particulières, une série de propriétés physiques et mécaniques. Nous avons retenu les neuf caractéristiques suivantes :

D	:	masse volumique (kg/m^3),
RVT	:	retrait volumétrique total (%),
N	:	dureté,
F	:	module de flexion statique (kg/cm^2),
K	:	coefficient de résilience (kgm/cm^3),
C	:	résistance unitaire en compression (kg/cm^2),
T	:	résistance unitaire en traction (kg/cm^2),
FD	:	résistance unitaire en fendage (kg/cm),
CS	:	résistance unitaire en cisaillement (kg/cm^2).

La masse volumique est le rapport de la masse de l'éprouvette à son volume. Le retrait volumétrique total traduit la variation du volume d'une éprouvette de bois de l'état saturé d'eau à l'état anhydre. La dureté est une caractéristique mi-physique, mi-mécanique, qui traduit la résistance du bois à la pénétration d'un cylindre d'acier appliqué sur la face radiale de l'éprouvette.

Les trois caractéristiques suivantes mesurent la cohésion axiale du bois. La flexion statique mesure la charge nécessaire à la rupture d'une éprouvette placée sur deux appuis, la charge étant appliquée progressivement. Le coefficient de résilience, appelé aussi module de flexion dynamique, est lié à la charge provoquant la rupture par un choc. La résistance unitaire en compression est lié à la charge nécessaire à la rupture de l'éprouvette, cette charge s'exerçant dans le sens de l'axe de l'arbre.

Enfin, les trois dernières propriétés ont trait à la cohésion transversale. La résistance unitaire en traction est la résistance du bois à une traction perpendiculaire aux fibres du bois dans la direction radiale et exercée aux deux extrémités de l'éprouvette. La résistance unitaire en fendage est liée à la charge provoquant la rupture de l'éprouvette par un effet de traction exercé, dans la direction radiale, à une extrémité de l'éprouvette. La résistance unitaire au cisaillement est liée à la charge provoquant la rupture de l'éprouvette par cisaillement longitudinal radial sous un effort de compression. Pour cet essai, l'éprouvette prend appui sur la moitié d'une section et la charge est appliquée sur la moitié de la section opposée.

Les données, reprises en annexe, concernent les valeurs moyennes observées sur les 68 arbres, plusieurs essais ayant été réalisés par arbre.

5.2. Interprétation des résultats

L'analyse en composantes principales a été réalisée avec le logiciel SAS, en utilisant les procédures PRINCOMP et FACTOR. Les figures 3 et 4, extraites

	D	RVT	N	F	K	C	T	FD	CS
D	1.0000	0.3757	0.8077	0.6669	0.6481	0.5054	0.7176	0.6870	0.6637
RVT	0.3757	1.0000	0.2914	0.3161	0.4862	0.2053	0.0343	0.0670	0.2132
N	0.8077	0.2914	1.0000	0.5420	0.4705	0.5198	0.5320	0.5797	0.7189
F	0.6669	0.3161	0.5420	1.0000	0.4865	0.6961	0.5360	0.3273	0.3253
K	0.6481	0.4862	0.4705	0.4865	1.0000	0.3960	0.3592	0.3213	0.4641
C	0.5054	0.2053	0.5198	0.6961	0.3960	1.0000	0.3324	0.1386	0.2994
T	0.7176	0.0343	0.5320	0.5360	0.3592	0.3324	1.0000	0.7769	0.3713
FD	0.6870	0.0670	0.5797	0.3273	0.3213	0.1386	0.7769	1.0000	0.4480
CS	0.6637	0.2132	0.7189	0.3253	0.4641	0.2994	0.3713	0.4480	1.0000

Figure 3 – Matrice de corrélation des caractéristiques technologiques.

des documents de sorties produits par ces procédures, reprennent la matrice de corrélation des variables initiales, les informations relatives aux valeurs propres de la matrice de corrélation et les coefficients de corrélation des variables initiales avec les premières composantes principales.

L'examen de la matrice de corrélation (figure 3) montre que toutes les corrélations sont positives et que, pour certains couples de variables, elles sont assez élevées. La valeur la plus grande est égale à 0,81 et, si on élimine le retrait volumétrique, la valeur la plus faible est de 0,14. Le retrait volumétrique, RVT, est, dans l'ensemble, moins corrélé aux autres caractéristiques, la valeur la plus grande n'étant que de 0,49. Cette variable semble donc se distinguer des autres.

La figure 4 donne les valeurs propres de la matrice de corrélation, les différences entre les valeurs propres successives, les proportions et les proportions cumulées de la variance expliquée par les composantes. Elle donne également les coefficients de corrélation des variables initiales avec les trois premières composantes.

La première composante principale prend en compte 53 % de la variabilité (figure 4). Elle est, de loin, la plus importante, puisque les deux composantes suivantes n'expliquent, respectivement, que 15 et 11 %. A partir de la quatrième, les composantes sont nettement moins utiles et correspondent à des valeurs propres inférieures à l'unité. Nous limiterons notre analyse aux trois premières composantes qui, ensemble, expliquent 79 % de la variabilité.

L'examen des corrélations des variables avec la première composante montre que toutes les variables sont corrélées positivement avec le premier axe. Sur les deux cercles de corrélation (figure 5), tous les points variables sont donc situés dans la partie droite du graphique. Cet axe peut s'interpréter comme un axe de qualité globale du bois : les arbres qui ont des valeurs élevées pour les différentes caractéristiques ont une valeur de la première composante qui est élevée également et sont des arbres à bonnes propriétés technologiques. Inversement, les arbres qui ont des valeurs faibles pour les différentes caractéristiques ont une valeur faible de la première composante.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
Z1	4.78860	3.45846	0.532066	0.53207
Z2	1.33014	0.35858	0.147793	0.67986
Z3	0.97155	0.24236	0.107951	0.78781
Z4	0.72920	0.27446	0.081022	0.86883
Z5	0.45474	0.18734	0.050526	0.91936
Z6	0.26740	0.06770	0.029711	0.94907
Z7	0.19970	0.04431	0.022189	0.97126
Z8	0.15539	0.05211	0.017266	0.98852
Z9	0.10329	.	0.011476	1.00000

Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
D	0.94980	-0.06294	0.07269
RVT	0.40564	0.64611	0.43421
N	0.85903	-0.06079	0.07913
F	0.75446	0.25593	-0.45102
K	0.69942	0.35339	0.24591
C	0.62483	0.37055	-0.57589
T	0.74375	-0.47423	-0.17809
FD	0.69433	-0.58727	0.14149
CS	0.70706	-0.08685	0.35240

Figure 4 – Informations relatives aux valeurs propres et corrélations des variables initiales avec les trois premières composantes principales.

Il s'agit là d'une interprétation très globale, qui doit être légèrement nuancée car un retrait volumétrique élevé n'est pas un facteur de qualité du bois, contrairement aux autres caractéristiques, mais on constate que, précisément, ce retrait volumétrique est la variable, de loin, la moins corrélée au premier axe.

La figure 6 donne une représentation graphique des arbres dans le premier plan factoriel. Selon l'axe z_1, les arbres s'ordonnent d'après leur qualité générale. Pour illustrer cette interprétation, nous avons calculé la moyenne des diverses caractéristiques pour les dix arbres qui ont les valeurs les plus faibles de z_1 et pour les dix arbres qui ont les valeurs les plus élevées de z_1. Les résultats sont repris dans le tableau 7, qui donne aussi les moyennes des caractéristiques pour l'ensemble des arbres. Il apparaît très clairement que les arbres avec les valeurs petites de z_1 ont, en moyenne, des valeurs plus faibles des diverses caractéristiques que les arbres avec les valeurs élevées de z_1.

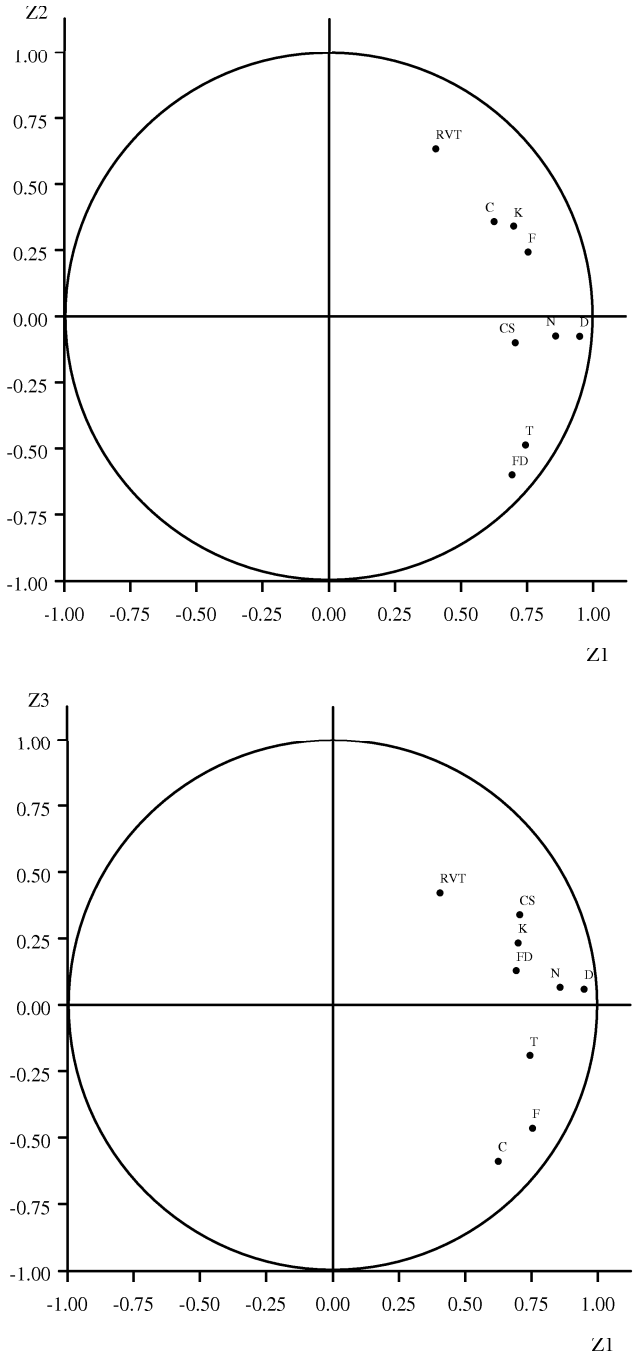

Figure 5 – Cercles de corrélation dans les deux premiers plans factoriels.

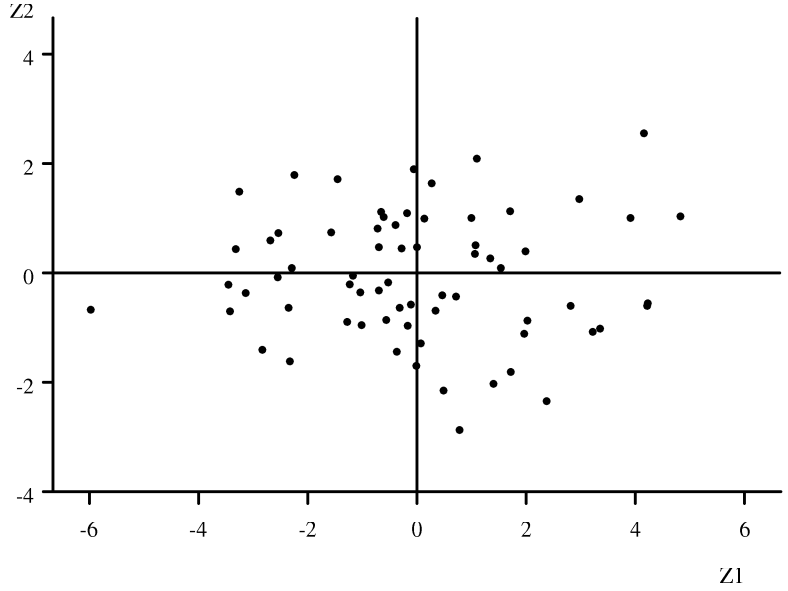

Figure 6 – Représentation des arbres dans le premier plan factoriel.

Pour cet exemple, l'examen détaillé de la figure 6 ne présente guère d'intérêt, dans la mesure où nous n'avons pas donné d'informations concernant les arbres. Ceux-ci n'ont pas d'intérêt par eux-mêmes et il nous est donc assez indifférent de connaître la position dans le plan factoriel d'un arbre particulier. Un examen rapide du graphique se justifie cependant pour vérifier s'il n'y a pas un ou éventuellement plusieurs arbres tout à fait particuliers, qui seraient situés en dehors du nuage de points. Une telle situation pourrait se présenter, par exemple, du fait d'erreurs dans les données.

La deuxième composante principale est corrélée positivement avec le retrait volumétrique total et, dans une mesure moindre, avec les trois caractéristiques liées à la cohésion transversale (F, K, et C). Elle est corrélée négativement avec deux des trois caractéristiques liées à la cohésion radiale (FD et T). Sur le graphique des individus, les arbres caractérisés par des valeurs faibles de z_2 sont donc des arbres à résistances au fendage et en traction (FD et T) supérieures à la moyenne et à retrait et cohésion axiale (F, K, et C) inférieurs à la moyenne, tandis que les arbres avec des valeurs élevées de z_2 correspondent à des arbres à retrait et cohésion axiale supérieurs à la moyenne et à résistances en fendage et en traction inférieures à la moyenne. Le tableau 7, qui donne les moyennes des caractéristiques pour les dix arbres les plus extrêmes en ce qui concerne z_2, confirme bien cette interprétation.

Ce deuxième axe complète donc l'information apportée par le premier axe.

Tableau 7 – Valeurs moyennes des caractéristiques pour l'ensemble des arbres et pour les dix arbres dont les valeurs des composantes sont les plus petites et les plus grandes.

Variables	Ensemble	z_1		z_2		z_3	
		petit	grand	petit	grand	petit	grand
D	717	663	772	723	716	710	727
RVT	24,2	23,3	25,1	22,7	25,1	23,6	25,6
N	3,09	2,41	3,73	3,08	3,10	2,93	3,29
F	1163	1037	1281	1156	1227	1213	1109
K	0,423	0,352	0,533	0,384	0,464	0,392	0,455
C	531	499	577	513	562	566	504
T	44,6	39,8	49,7	47,8	41,8	45,0	44,3
FD	28,6	25,1	33,0	31,5	25,3	27,3	30,4
CS	157	138	174	160	161	144	168

En effet, à égalité de qualité globale, c'est-à-dire pour une valeur fixée de z_1, ce deuxième axe permet de différencier les arbres plutôt meilleurs en cohésion axiale et moins bons en cohésion transversale (FD et T) des arbres qui présentent la situation inverse.

Quant à la troisième composante, elle oppose le retrait volumétrique, RVT, et la résistance au cisaillement, CS, à la résistance en compression, C, et au module de flexion statique, F. Les deux premières caractéristiques citées sont corrélées positivement tandis que les deux dernières sont corrélées négativement avec z_3. Les arbres caractérisés par une valeur faible de z_3 ont donc dans l'ensemble un retrait volumétrique et une résistance au cisaillement plus faibles que la moyenne, un module de flexion statique et une résistance en compression plus importants que la moyenne. La situation inverse s'observe pour les arbres à valeur élevée de z_3 (tableau 7). Cette troisième composante dissocie donc les trois variables liées à la cohésion axiale, qui, dans le premier plan factoriel, étaient relativement groupées. Au-delà des constatations résumées ci-dessus, il est difficile de donner une interprétation pratique et concrète de ce troisième axe, qui, rappelons-le, n'explique que 11 % de la variabilité.

6. QUELQUES INFORMATIONS COMPLÉMENTAIRES

6.1. Transformation des variables

Dans les deux exemples qui ont été examinés, nous avons calculé les valeurs et les vecteurs propres de la matrice de corrélation des variables et nous avons vu, au paragraphe 3.1, que les premières composantes principales obtenues permettaient de reconstituer au mieux le tableau des données centrées réduites.

Les valeurs et vecteurs propres de la matrice de corrélation sont aussi, à une

constante près, les valeurs et les vecteurs propres de la matrice $\boldsymbol{X}'\boldsymbol{X}$, puisque, si $\boldsymbol{X}$ est la matrice des données centrées réduites, la matrice de corrélation est, à une constante près, égale à $\boldsymbol{X}'\boldsymbol{X}$.

Le principe de la reconstitution d'une matrice $\boldsymbol{X}$ n'est, en fait, pas limité au cas où $\boldsymbol{X}$ correspond aux variables centrées réduites, mais peut s'envisager quelle que soit $\boldsymbol{X}$. L'analyse en composantes principales telle qu'elle a été présentée, n'est donc qu'un cas particulier d'une analyse plus générale. Des informations à ce sujet sont données par PALM [1994], notamment.

La standardisation a comme conséquence de donner à chaque variable un poids identique dans l'analyse. Elle se justifie pratiquement toujours quand les variables initiales sont exprimées dans des unités différentes, comme dans le cas de l'étude des propriétés du bois de hêtre. Dans cette situation, les résultats obtenus en l'absence de standardisation seraient fonction des unités dans lesquelles sont exprimées les caractéristiques initiales.

Il peut arriver cependant, et notamment quand les variables sont exprimées dans des unités identiques, qu'on souhaite accorder aux variables un poids fonction de leur variance. On procède alors au calcul des valeurs propres et des vecteurs propres de la matrice des variances et covariances. La somme de toutes les valeurs propres est, dans ce cas, égale à la somme des variances des variables initiales.

Pour l'exemple de la composition du lait des mammifères, les trois variables sont exprimées en pour cent et on aurait pu envisager de ne pas standardiser les variables. L'analyse de la matrice des variances et covariances aurait, dans ce cas, attribué un poids beaucoup plus important à la teneur en graisse qu'à la teneur en protéines et surtout qu'à la teneur en lactose puisque l'écart-type de la teneur en graisse est plus grand que les deux autres écarts-types (tableau 1).

Pour certaines applications, un compromis peut être trouvé en procédant à une transformation de variables et en réalisant une analyse sur la matrice des variances et covariances des données transformées. Un exemple est donné par NAIK et KHATTREE [1996], qui analysent les résultats obtenus par une cinquantaine de pays lors d'épreuves olympiques de courses (100 m, 200 m, ..., marathon). Ces auteurs ont trouvé que le fait de centrer et de réduire les temps obtenus aux différentes épreuves ne permettait pas une bonne interpréation des résultats. Ils ont proposé de définir de nouvelles variables, qui sont les distances couvertes par unité de temps dans chaque épreuve, et de procéder à l'analyse de la matrice de variances et covariances de ces distances.

Indépendamment du problème de la standardisation, des transformations de variables peuvent être utiles en présence de non-normalité très accusée. En effet, comme le soulignent LEBART *et al.* [1995], le critère des moindres carrés est particulièrement bien adapté au cas de variables normales mais peut donner un poids excessif aux observations extrêmes dans le cas de distributions nettement non normales. De même, les transformations peuvent améliorer la linéarité des relations existant entre les variables. Parmi les transformations classiques, on pensera notamment à la transformation logarithmique, à la transformation racine carrée et, dans le cas de données de rangs, au calcul des scores normaux.

Enfin, dans certaines situations, on peut souhaiter éliminer, avant l'analyse, l'effet de certaines variables. On réalise alors une analyse en composantes principales sur des résidus de régression [LEBART *et al.*, 1995].

6.2. Nombre de composantes à retenir

Aucune règle rigoureuse ne peut être donnée pour le choix du nombre de composantes à prendre en considération lors d'une analyse en composantes principales.

Si les observations constituent un échantillon aléatoire et simple d'individus prélevés dans une population normale à p dimensions, on peut tester l'égalité des $p-q$ dernières valeurs propres [DAGNELIE, 1975 ; JACKSON, 1991 ; SAPORTA, 1990]. Si l'hypothèse est acceptée, on conserve les q premiers axes et on néglige les $p-q$ derniers axes. L'utilisation de ce test conduit cependant souvent à considérer un nombre élevé de composantes, dont certaines risquent de ne présenter aucun intérêt pratique.

Des règles empiriques peuvent également guider l'utilisateur. Une de ces règles consiste à ne prendre en considération que les composantes pour lesquelles la valeur propre est supérieure à la moyenne arithmétique de toutes les valeurs propres. En particulier, si on travaille sur les données centrées réduites, cela revient à négliger les composantes dont la variance est inférieure à l'unité.

L'examen de la décroissance des valeurs propres successives peut également donner des indications quant aux composantes à retenir. Le graphique des valeurs propres en fonction de leur rang présente souvent l'allure d'un éboulis au pied d'un escarpement[1], ce qui justifie le nom anglais de ce graphique. On essaye, sur ce graphique, de détecter l'existence d'un coude, c'est-à-dire d'une réduction assez brutale de la pente du graphique et on néglige les composantes correspondant aux valeurs propres situées après ce coude.

De nombreuses autres règles sont encore proposées dans la littérature pour déterminer le nombre de composantes à retenir. Une synthèse est donnée par JACKSON [1991].

Dans la pratique, le nombre de composantes retenues est également largement conditionné par la possibilité d'interpréter les composantes, du moins lorsque l'analyse est réalisée dans un but purement descriptif.

Indépendamment de la détermination du nombre de composantes à retenir, l'examen du graphique de l'ensemble des valeurs propres permet de tirer des informations sur les variables soumises à l'analyse. Ainsi, si ces variables présentent q relations linéaires exactes, les q dernières valeurs propres seront nulles. A l'opposé, si toutes les variables sont parfaitement non corrélées et si les données sont centrées et réduites, toutes les valeurs propres seront égales à l'unité. La première situation se rencontre de façon systématique si le nombre d'individus est inférieur ou égal au nombre de variables. Si ce nombre est supérieur au nombre de variables, la colinéarité exacte est rare en pratique. Par contre, une

[1] En anglais : *scree plot*.

colinéarité approximative est assez courante ; la ou les dernières valeurs propres sont alors presque nulles. De même, la non corrélation parfaite ne s'observe pratiquement jamais dans les situations réelles car, même si les variables sont indépendantes, de faibles corrélations existeront par le fait du hasard et les valeurs propres présenteront une faible décroissance. A titre d'illustration, nous avons simulé 68 observations pour 9 variables normales indépendantes et calculé les valeurs propres de la matrice de corrélation correspondante. Les calculs ont été répétés 1000 fois et les moyennes de la plus grande valeur propre et de la plus petite valeur propre sont respectivement 1,60 et 0,51. La comparaison de ces deux valeurs aux valeurs correspondantes de la figure 4 montre bien que, pour les données relatives au hêtre, il existe effectivement des directions privilégiées dans le nuage des points.

6.3. Interprétation des composantes

L'étude des corrélations des variables initiales et des différentes composantes, ou l'examen des cercles de corrélation qui donnent des représentations graphiques de ces corrélations, est un élément important de l'analyse des résultats.

Ces corrélations permettent parfois de donner une signification concrète aux axes. Cette signification dépend évidemment de la nature des données. Une situation que l'on peut cependant rencontrer pour des problèmes variés est la présence d'un "facteur de taille". Ce facteur se rencontre quand toute les variables présentent des corrélations positives. Dans ce cas, tous les éléments du premier vecteur propre sont de même signe. Puisque ce signe est arbitraire et qu'on peut toujours multiplier un vecteur propre par -1, nous allons considérer le cas où il est positif. La première composante est alors corrélée positivement à chacune des variables et, dans le cercle de corrélation, toutes les variables initiales se situent d'un même côté de l'axe. Si on classe les individus dans l'ordre croissant des projections sur le premier axe, ceux-ci sont aussi rangés par valeurs croissantes de l'ensemble des variables : les individus ayant des valeurs négatives pour z_1 ont des valeurs plus faibles que la moyenne pour l'ensemble des variables et les individus ayant des valeurs positives pour z_1 ont des valeurs plus élevées que la moyenne pour l'ensemble des variables.

Ainsi, dans une étude non publiée relative à la morphologie des moules, au cours de laquelle une dizaine de mesures de longueurs et de poids avaient été réalisées sur près de 800 moules, la première composante permettait d'expliquer 93 % de la variabilité totale et traduisait simplement le fait que l'essentiel de la variabilité des observations était lié à la présence dans l'échantillon de moules de tailles fort variables. Ce premier axe était donc un axe de taille, au sens premier du terme.

Pour cet exemple, l'importance du facteur taille masquait d'ailleurs l'intérêt des autres composantes, auxquelles n'étaient associées que des valeurs propres faibles. Ainsi, la part de la variance expliquée par la seconde composante n'était que de 3 %, soit beaucoup moins que la proportion moyenne (10 % lorsqu'il y a dix variables). La prise en compte de cette seconde composante se justifie cepen-

dant pleinement car, en définitive, c'est elle qui contient le plus d'information utile, la première composante ne faisant que traduire une évidence. Dans un cas comme celui qui est évoqué ici, on peut d'ailleurs très facilement éliminer l'effet de taille en réalisant l'analyse en composantes sur des valeurs relatives. Il suffit, pour cela, de diviser, par exemple, toutes les mesures de longueur par la longueur totale de l'individu et toutes les mesures de poids par le poids total de l'individu. De cette façon, non seulement on élimine l'axe de taille, mais on réduit de deux unités le nombre de variables, puisqu'on ne considère plus la taille totale et le poids total des individus.

Pour les données relatives aux caractéristiques mécaniques du bois de hêtre, nous avons vu, au paragraphe 5.2, que la matrice de corrélation ne contenait que des valeurs positives et le premier axe a été interprété comme un axe de qualité globale. Il s'agit en fait, ici aussi, d'un axe de taille.

En complément aux cercles des corrélations, les graphiques des individus peuvent également être utiles pour l'interprétation des composantes principales, du moins lorsque les individus peuvent être facilement identifiés. L'examen des caractéristiques des individus les plus extrêmes pour une composante donnée peut, en effet, fournir des renseignements sur la signification des axes.

Il faut cependant noter que les composantes principales sont des variables artificielles qui n'ont pas toujours une signification physique réelle, ce qui rend évidemment leur interprétation nettement moins aisée [KSHIRSAGAR, 1972]. On se limitera, par exemple, à constater qu'un axe oppose plus ou moins nettement une caractéristique à une autre ou à un groupe d'autres caractéristiques ou qu'un axe est lié à telle ou telle variable. Ainsi, dans l'exemple relatif aux caractéristiques du bois de hêtre, aucune signification physique nette n'a pu être donnée au deuxième et au troisième axe.

6.4. Quelques utilisations de l'analyse en composantes principales

Dans les paragraphes qui précèdent, l'analyse en composantes principales a essentiellement été présentée comme une méthode descriptive, visant à explorer la structure d'un tableau de données par la prise en considération des liaisons qui existent entre les variables. Elle conduit à une vision synthétique du tableau de données et peut faire apparaître une éventuelle structure dans les données qui n'était peut-être pas soupçonnée au départ.

Elle peut par exemple révéler l'existence de groupes de variables ou de groupes d'individus. Elle peut aussi mettre en évidence la présence d'individus aberrants, dont le comportement s'écarte du comportement global de l'ensemble des individus. Sur les graphiques des individus, ceux-ci seront situés en dehors du nuage de points, soit pour un axe, soit pour plusieurs axes.

Dans de telles situations, il peut être utile de recommencer l'analyse après suppression de ces individus aberrants. En effet, leur élimination peut provoquer des modifications, parfois importantes, des composantes principales. A ce sujet, on notera que la suppression de ces individus ne signifie pas nécessairement que ceux-ci sont définitivement éliminés de l'interprétation ultérieure des résultats.

Ils sont éliminés de l'analyse en composantes principales de manière à permettre à celle-ci de révéler des informations initialement masquées.

L'analyse en composantes principales peut aussi être utilisée pour visualiser et caractériser des groupes d'individus identifiés *a priori* ou résultant, par exemple, d'une classification numérique. Un tel exemple est présenté dans une autre note [PALM, 1996].

Une autre application importante de l'analyse en composantes principales est son utilisation comme méthode de réduction de la dimension d'un problème. L'idée est de remplacer les p variables initiales par q composantes qui sont utilisées pour des analyses ultérieures. Cette substitution permet non seulement de réduire la taille du problème, car q est inférieur à p, mais peut simplifier les calculs puisque les nouvelles variables sont non corrélées. Cette dernière propriété est notamment exploitée dans la technique de régression orthogonale. Des informations à ce sujet sont données par PALM et IEMMA [1995].

En relation avec les problèmes de régression, l'analyse en composantes principales permet aussi d'identifier les variables impliquées dans les phénomènes de colinéarité et, par conséquent, de fournir des informations quant aux variables à exclure des modèles de régression. Des illustrations concrètes de cette application sont données par PALM [1988] et TOMASSONE *et al.* [1983], notamment.

BIBLIOGRAPHIE

DAGNELIE P. [1975]. *Analyse statistique à plusieurs variables.* Gembloux, Presses agronomiques, 362 p.

HARTIGAN J.A. [1975]. *Clustering algorithms.* New York, Wiley, 351 p.

JACKSON J.E. [1991]. *A user's guide to principal components.* New York, Wiley, 569 p.

KSHIRSAGAR A.M. [1972]. *Multivariate analysis.* New York, Dekker, 534 p.

LEBART L., MORINEAU A., PIRON M. [1995]. *Statistique exploratoire multidimensionnelle.* Paris, Dunod, 439 p.

LECLERCQ A. [1979]. *Influence du milieu et du traitement sur la qualité du bois de hêtre.* Gembloux, Faculté des Sciences agronomiques, 339 p.

NAIK D.N., KHATTREE R. [1996]. Revisiting olympic track records : some practical considerations in the principal component analysis. *Amer. Statist.* 50, 140-144.

PALM R. [1988]. Les critères de validation des équations de régression linéaire. *Notes Stat. Inform.* (Gembloux) 88/1, 27 p.

PALM R. [1994]. Les méthodes d'analyse factorielle : principes et applications. *Biom. Praxim.* 34, 35-80.

PALM R. [1996]. La classification numérique : principes et application. *Notes Stat. Inform.* (Gembloux) 96/1, 28 p.

PALM R., IEMMA A.F. [1995]. Quelques alternatives à la régression classique dans le cas de la colinéarité. *Rev. Stat. Appl.* 43, 5-33.

PHILIPPEAU G. [1986]. *Comment interpréter les résultats d'une analyse en composantes principales ?* Paris, Institut technique des Céréales et des Fourrages, 63 p.

PRESS J. [1972]. *Applied multivariate analysis.* New York, Holt, Rinehart and Winston, 521 p.

SAPORTA G. [1990]. *Probabilités, analyse des données et statistique.* Paris, Technip, 493 p.

TOMASSONE R., LESQUOY E. et MILLIER C. [1983]. *La régression : nouveaux regards sur une ancienne méthode statistique.* Paris, Masson, 180 p.

ANNEXE

Données relatives aux caractéristiques du bois de hêtre (d'après LECLERCQ, 1979).

OBS	D	RVT	N	F	K	C	T	FD	CS
1	739	24.4	3.38	1204	0.480	548	45.0	29.9	177
2	705	22.6	3.53	1096	0.364	525	40.7	28.2	175
3	746	25.4	3.94	1317	0.532	580	45.2	28.5	181
4	709	26.3	2.62	1008	0.530	470	44.1	29.8	150
5	758	25.7	3.95	1166	0.366	512	47.7	32.5	176
6	700	23.8	2.98	1205	0.472	549	41.0	25.4	148
7	694	24.3	3.27	1055	0.344	495	42.0	26.7	178
8	706	23.4	3.42	1159	0.435	547	40.2	26.6	163
9	710	23.7	2.66	1184	0.421	533	42.6	24.9	123
10	672	22.7	2.10	1050	0.424	466	38.9	26.2	153
11	672	25.2	2.34	1130	0.417	517	39.1	22.9	154
12	752	25.9	3.18	1216	0.412	542	45.8	26.4	165
13	671	24.8	2.44	1072	0.351	487	39.8	28.2	157
14	684	23.5	2.48	962	0.399	495	41.8	27.7	154
15	715	26.5	2.83	1153	0.528	518	42.3	26.5	152
16	672	24.4	2.43	1109	0.345	503	39.5	24.2	157
17	697	23.3	2.72	1005	0.365	487	35.8	23.7	136
18	761	24.6	3.54	1219	0.483	558	46.1	28.3	175
19	666	24.0	2.51	951	0.328	464	42.0	26.6	136
20	712	25.0	3.29	1221	0.368	588	42.7	28.0	127
21	613	22.1	2.19	951	0.295	451	36.8	24.1	110
22	708	24.5	3.08	1143	0.302	546	44.4	29.3	144
23	731	23.6	3.70	1239	0.434	584	45.8	27.2	165
24	665	22.5	2.54	1006	0.323	483	41.9	27.7	158
25	680	23.7	3.30	1061	0.368	522	42.4	26.3	164
26	721	25.6	3.84	1268	0.431	593	39.5	26.5	160
27	685	24.3	3.38	1057	0.332	482	44.1	28.6	168
28	653	24.4	2.43	1068	0.372	532	42.1	25.9	131
29	724	24.5	3.23	1263	0.401	576	44.4	26.5	133
30	708	24.9	3.04	1121	0.502	505	43.0	25.0	152
31	729	24.6	3.06	1105	0.395	497	42.8	28.4	157
32	745	25.0	3.35	1109	0.609	477	42.3	30.2	171
33	705	25.9	2.98	1164	0.411	499	43.3	27.6	142
34	732	25.3	3.34	1090	0.405	489	45.4	31.5	160

Données relatives aux caractéristiques du bois de hêtre (d'après LECLERCQ, 1979) (suite).

OBS	D	RVT	N	F	K	C	T	FD	CS
35	713	25.9	2.70	1440	0.373	487	46.7	30.4	122
36	686	25.1	2.52	1215	0.366	574	40.7	25.2	136
37	642	22.6	2.61	1031	0.337	520	39.9	25.0	137
38	698	24.0	2.45	1194	0.378	524	45.4	27.3	144
39	685	23.0	2.33	1070	0.400	522	41.2	26.6	132
40	789	27.2	4.33	1208	0.581	584	45.7	34.5	196
41	733	25.2	3.29	1182	0.415	564	39.8	24.9	171
42	724	23.3	3.04	1294	0.502	582	42.7	29.8	162
43	769	24.8	3.62	1225	0.489	534	43.1	24.6	189
44	687	23.6	2.86	1066	0.304	548	44.9	29.2	153
45	758	25.8	3.44	1451	0.654	622	46.7	28.1	184
46	660	23.6	2.24	1127	0.334	565	40.1	20.6	132
47	737	22.9	3.07	1141	0.412	489	47.2	33.3	168
48	715	22.7	2.95	1149	0.381	539	51.2	29.9	147
49	710	23.4	3.02	1132	0.438	519	44.6	29.4	156
50	741	22.9	3.21	1275	0.449	508	50.2	32.5	177
51	788	24.6	3.93	1378	0.495	585	50.4	35.3	166
52	761	22.9	3.33	1145	0.468	524	50.7	33.6	148
53	715	22.2	3.11	1171	0.381	563	48.1	29.1	154
54	767	25.4	3.81	1288	0.587	567	54.5	35.9	158
55	713	23.5	3.08	1076	0.483	505	46.9	30.0	151
56	703	25.2	2.59	1148	0.437	551	45.0	27.0	155
57	772	26.1	3.43	1174	0.503	528	49.6	35.5	167
58	779	25.2	3.50	1368	0.593	595	51.6	29.2	170
59	768	24.3	3.74	1187	0.528	571	50.2	35.0	176
60	734	22.3	3.52	1254	0.302	483	50.3	32.3	169
61	713	23.4	3.38	1187	0.466	530	48.0	27.4	156
62	740	24.0	2.85	1188	0.435	554	46.8	30.8	155
63	686	23.1	2.36	1073	0.360	457	45.6	29.4	140
64	709	22.5	2.87	1235	0.402	559	48.5	26.5	153
65	778	23.2	3.50	1191	0.394	539	50.6	35.2	174
66	698	22.0	3.24	1154	0.369	547	42.4	31.8	161
67	772	23.6	3.69	1244	0.458	599	52.0	32.7	172
68	747	23.5	3.38	1270	0.365	581	48.9	32.2	173