L'analyse des correspondances multiples

$$X(n,p) = \begin{pmatrix} x_1^1 & \dots & \dots & \dots & x_1^p \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ x_n^1 & \dots & \dots & \dots & x_n^p \end{pmatrix}$$

➔ Les p variables sont qualitatives

Etape 1 – détour :
La tableau de contingence

Principe d'un tableau de contingence

- → Il s'agit du tableau résultant de 2 variables qualitatives dans un fichier de données individuel.
- →A partir de 2 variables, on peut demander un tableau croisé des effectifs (commande « tab » dans stata). C'est ce qu'on appelle un tableau de contingence.

- → Les modalités de la première composent les lignes du tableau.
- → Les modalités de la seconde composent les colonnes du tableau

$\mathbf{Y}$ / $\mathbf{X}$	$\mathbf{y}^1$	$\mathbf{y}^2$	…	$\mathbf{y}^j$	…	$\mathbf{y}^J$	Total
$\mathbf{x}^1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
$\mathbf{x}^2$	n_{21}						
…	…						
$\mathbf{x}^i$	n_{i1}			n_{ij}			$n_{i.}$
…							
$\mathbf{x}^I$	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}$

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}$

Effectifs conjoints

Tri à plat pour X

Y / X	$\mathbf{y}^1$	$\mathbf{y}^2$	…	$\mathbf{y}^j$	…	$\mathbf{y}^J$	Total
$\mathbf{x}^1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
$\mathbf{x}^2$	n_{21}						
…	…						
$\mathbf{x}^i$	n_{i1}			n_{ij}			$n_{i.}$
…							
$\mathbf{x}^I$	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}$

Tri à plat pour Y

		Variable 2				
		Modalité 1	Modalité 2	…	Modalité J	Marge 1
Variable 1	Modalité 1	Effectifs conjoints				Effectifs marginaux (tri à plat variable 1)
	Modalité 2					
	…					
	Modalité I					
Marge 2		Effectifs marginaux (tri à plat variable 2)				

Etape 2 -
L’analyse des correspondances (AC) : une méthode d’analyse d’un tableau de contingence

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

➔L'AC analyse un tel tableau et répond aux questions suivantes :

-Y a-t-il des lignes du tableau qui se ressemblent ?

- Y a-t-il des lignes du tableau qui s'opposent ?

- Mêmes questions pour les colonnes

- Y a-t-il des associations de modalités entre X et Y qui s'attirent (effectifs conjoints très élevés) ou qui se repoussent (effectifs conjoints très faibles)

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

➔ Si toutes les lignes et toutes les colonnes du tableaux se ressemblent, à quelle situation cela se rapporte-t-il ?

➔ Egalité des profils lignes ⟺ Egalité des profils colonnes

$$\Leftrightarrow \quad n_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Objectifs de l'AC

- Représenter les écarts à l'indépendance des profils lignes et colonnes

- Faire apparaître les modalités de X qui se ressemblent (et donc à l'inverse qui s'opposent)

- Décrire la distribution de X dans les J sous-populations de Y

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

Fréquences

$$f_{ij} = \frac{n_{ij}}{n} = \frac{Effectif de la cellule (i, j)}{Effectif total}$$

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

➔ **Ces profils lignes peuvent-être appréhendés comme une liste de coordonnées dans un espace à J dimensions**

➔ **On peut alors chercher les directions de plus grande dispersion**

Fréquences lignes (profils lignes)

$$fl_{ij} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}} = \frac{EffectifdelaCellule(i,j)}{EffectifdelaLigne(i)}$$

$$f_{.j} = \frac{n_{.j}}{n} = \frac{EffectifdelaColonne(j)}{EffectifTotal}$$

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

➔ **Ces profils colonnes peuvent-être appréhendés comme une liste de coordonnées dans un espace à I dimensions**

➔ **On peut alors chercher les directions de plus grande dispersion**

Fréquences colonnes (profils colonnes)

$$fc_{ij} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} = \frac{EffectifdelaCellule(i,j)}{EffectifdelaColonne(j)}$$

$$f_{i.} = \frac{n_{i.}}{n} = \frac{EffectifdelaLigne(i)}{EffectifTotal}$$

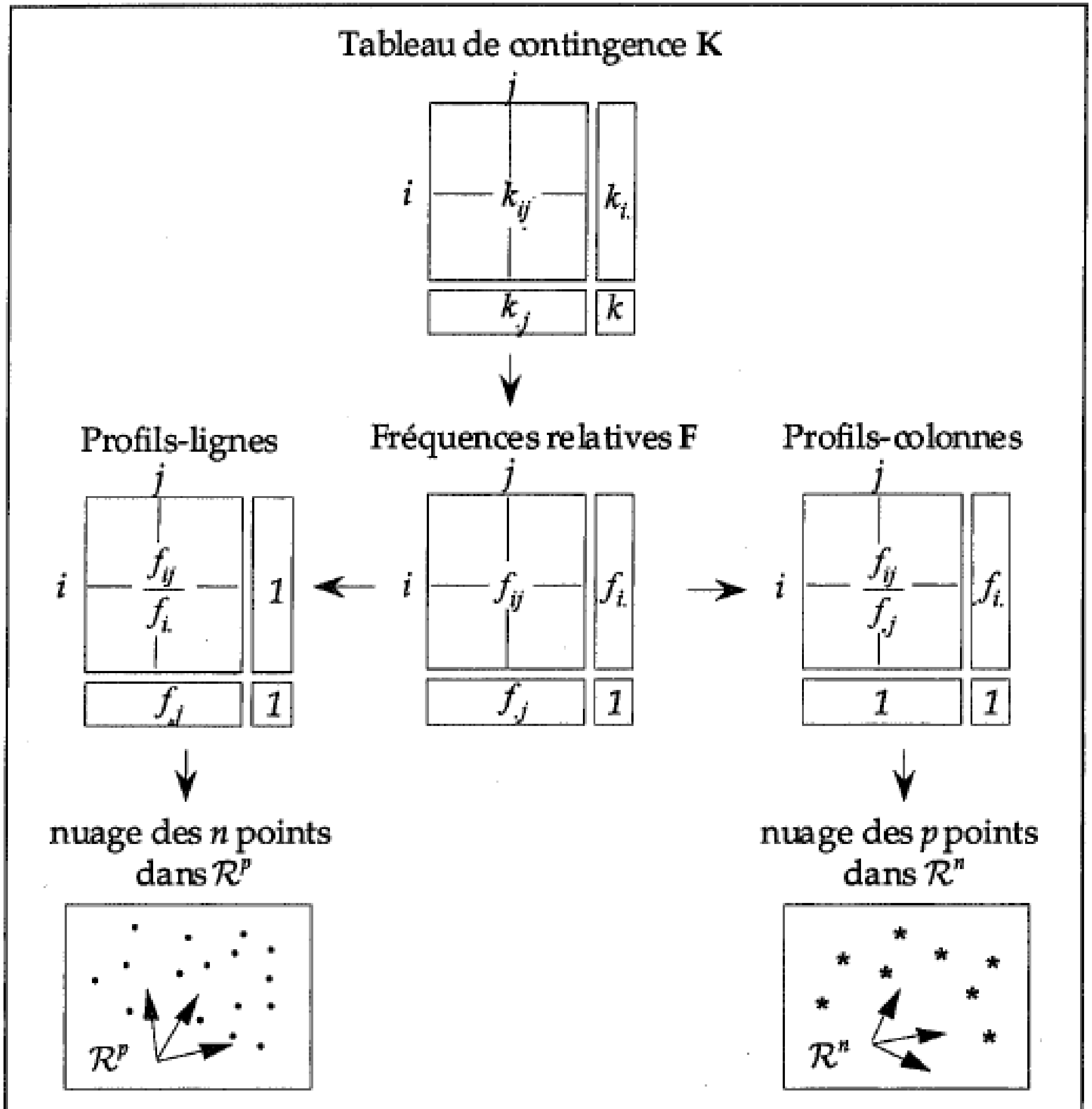

Figure 4.2 – 1. Transformations du tableau de contingence

Source: Lebart, Piron, Morineau, 2006, p. 144

Y X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

Distance entre 2 profils lignes

= Distance entre 2 individus i et i'

= distance du Φ² (distance euclidienne inadaptée)

$$d^2_{\Phi^2}(L_i, L_{i'}) = \sum_j \frac{\left(fl_{ij} - fl_{i'j}\right)^2}{f_{.j}}$$

Y / X	y^1	y^2	…	y^j	…	y^J	Total
x^1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1.}$
x^2	n_{21}						
…	…						
x^i	n_{i1}			n_{ij}			$n_{i.}$
…							
x^I	n_{I1}					n_{IJ}	$n_{I.}$
Total	$n_{.1}$			$n_{.j}$		$n_{.J}$	$\sum n_{ij}=n$

Taux de liaison (tableau des taux de liaison soumis à l'analyse)

$$t_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} = \frac{EcartsAlInddépendance}{"PoidsThéorique"} = \frac{n \times n_{ij}}{n_{i.}n_{.j}} - 1$$

Propriétés de la série des taux de liaison

- Moy $(t_{ij}) = 0$

- Var $(t_{ij}) = \Phi^2 = \text{khi2} / n$

$$\sum_i \sum_j f_{i.} \times f_{.j} \times t_{ij}^2 = \sum_i \sum_j \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}} = \sum_i \sum_j \frac{\left(n_{ij}\right)^2}{n_{i.} n_{.j}} - 1 = \frac{\chi^2}{n}$$

- Φ^2 = Somme des valeurs propres = $\sum \lambda_l$

Nuage des modalités de la variable X

- L'AC réalise une analyse factorielle sur le nuage des modalités de la variable X préalablement centré. C'est la représentation qui restitue avec le plus de fidélité les distance du phi2 entre les modalités de X
- Matrice d'inertie, valeurs propres, vecteurs propres, axes et composantes principales
- A chaque modalité on peut associer une coordonnée sur chacun des axes factoriels
- On parle de codage d'une variable qualitative

Nuage des modalités de la variable Y

- La procédure est identique pour la nuage des modalités de la variable Y

Etape 3 -
Généraliser la démarche à plus de 2 variables qualitatives :
L'analyse des correspondances multiples (ACM)

Du tableau de données au tableau disjonctif complet

Exemple

Tableau de données

$$s=3$$

$$\begin{array}{c} 1 \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ n \end{array} \quad \underset{(n,s)}{\mathbf{R}} = \left|\begin{array}{c|c|c} 2 & 2 & 4 \\ 2 & 1 & 3 \\ 3 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \\ 3 & 2 & 2 \\ 1 & 1 & 4 \end{array}\right|$$

Figure 5.1 – 1. Tableau de données sous forme de codage condensé

Source: Lebart, Piron, Morineau, 2006, p. 189

Tableau disjonctif complet

$$\underset{(n,s)}{\mathbf{R}} = \overbrace{\left|\begin{array}{c|c|c} 2 & 2 & 4 \\ 2 & 1 & 3 \\ 3 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \\ 3 & 2 & 2 \\ 1 & 1 & 4 \end{array}\right|}^{s=3} \Rightarrow \underset{(n,p)}{\mathbf{Z}} = \overbrace{\left|\begin{array}{ccc|cc|cccc} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right|}^{p=9}$$

(rows 1 to n)

Figure 5.1 - .2. Construction du tableau disjonctif complet Z

Source: Lebart, Piron, Morineau, 2006, p. 190

Somme des lignes ? Nombre de questions (Q)

Somme des colonnes ? Effectif de la modalité k (n_k)

Somme totale ? nQ

ACM : notations fondamentales

- $\mathbb{I}$: ensemble des n individus ($i \in \mathbb{I}$) (cardinal : n)
- $\mathbb{Q}$: ensemble des $\mathbb{Q}$ questions actives ($q \in \mathbb{Q}$) (cardinal : $\mathbb{Q}$)
- $\mathbb{K}$: ensemble des modalités des questions actives ($k \in \mathbb{K}$)
- $\mathbb{K}\langle q\rangle$: sous-ensemble des modalités de réponse de la question q

ACM : l'individu i choisit *une et une seule modalité* de réponse pour chaque question
(Variables catégorisées : chaque variable admet un nombre fini de modalités de réponse)

- si $q \neq q'$ alors $\mathbb{K}<q> \cap \mathbb{K}<q'> = \varnothing$

- $\mathbb{K} = \bigcup_{q \in Q} \mathbb{K} < q >$ (ensemble de toutes les modalités de réponse)

- $\mathbb{K}<i>$: patron de réponses de l'individu i

- $\mathbb{I}<k>$: ensemble des individus ayant donné la réponse k à la question q

- $\mathbb{I} = \bigcup_{k \in K<q>} \mathbb{I} < k >$

- pour une question q, si $k \neq k'$ alors $\mathbb{I}<k> \cap \mathbb{I}<k'> = \varnothing$

Principe géométrique de l'ACM :

- Construire un nuage euclidien de points représentant les individus

- Interpréter des axes : déterminer des axes principaux et des variables principales

Etape fondamentale :

- Définir les distances entre individus sur la base de leurs réponses aux questions actives

Le calcul de la distance entre individus dans l'ACM standard

Soit :

- $d_q(i,i')$ la distance entre les individus i et i' à la question q
- n_k le nombre d'individus ayant choisi le modalité k à la question q
- $n_{k'}$ le nombre d'individus ayant choisi la modalité k' à la question q

Propriété fondamentale de l'ACM :

la distance entre deux individus i et i' ne dépend que des *questions de désaccord*.

➔ Par conséquent :

si q est une *question d'accord* entre i et i' : $d_q(i,i') = 0$

Le calcul de la distance entre individus dans l'ACM standard

Si q est une question de désaccord entre i et i', l'un ayant choisi la modalité k et l'autre k' ($\neq k$)

alors

$$\boxed{d_q^2(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}}}$$

Avec $f_k = n_k/n$, et $f_{k'} = n_{k'}/n$

La distance globale $d(i,i')$ entre i et i' est définie par :

$$\boxed{d^2(i,i') = \frac{1}{Q}\sum_{q\in Q} d_q^2(i,i')}$$

(moyenne quadratique des distances des questions)

A partir de ces distances, on définit le nuage euclidien des individus.
Le nombre de dimensions sera au plus égal à ($\mathbb{K}$-$\mathbb{Q}$) (nombre de modalités moins nombre de questions actives) dont on détermine les directions principales.
(on définit les axes principaux en ajustant le nuage par la méthode des moindres carrés orthogonaux)

Propriété :

ACM (sur le tableau $\mathbb{I}$*$\mathbb{Q}$)

= AC du tableau disjonctif complet ($\mathbb{I}$*$\mathbb{K}$)

L'ACM = AFC sur le tableau disjonctif complet Z, bien que ce tableau ne puisse pas être considéré a priori comme un tableau de contingence (puisqu'il est de type : Individus * Variables)

- Même transformation des variables
- Même critère d'ajustement (pondération des points par leurs profils marginaux)
- Même distance (celle du χ^2)

- 3 conséquences
  - La distance croît avec le nombre de modalité que n'ont pas en commun k et k' et ceux d'autant plus qu'il s'agit de modalités rares
  - Un individu possédant des modalités rares est éloigné de ceux qui n'ont pas cette modalité
  - Deux individus possédant les mêmes modalités pour toutes les variables sont confondus

Les conséquences sont identiques concernant les modalités

Variance du nuage	$$V = \frac{K}{Q} - 1$$
Nombre de valeurs propres (nb de dim°)(non trivialement nulles)	L = K – Q
Moyenne des valeurs propres	$$\bar{\lambda} = \frac{V}{K-Q} = \frac{1}{Q}$$

Variance totale = somme des valeurs propres

Propriétés des coordonnées de l'individu I	$mean\ y_l^I = \sum_{i \in I} \frac{y_l^i}{n} = 0$ $Var\ y_l^I = \sum_{i \in I} \frac{(y_l^i)^2}{n} = \lambda_l$
Propriétés des coordonnées de la modalité k (pour toute question q)	$\sum_{k \in K} \frac{f_k}{Q} y_l^k = 0$
Coordonnée du point moyen modalité k	$\sum_{k \in K} \frac{f_k}{Q} (y_l^k)^2 = \lambda_l$ $\bar{y}_l^k = \sum_{i \in I_k} \frac{y^i}{n_k} = \sqrt{\lambda_l} \times y_l^k$

Taux modifiés de Benzécri

Ces taux permettent d'apprécier l'importance relative des axes principaux.

1. On calcule la moyenne des valeurs propres : $\overline{\lambda} = \frac{1}{Q}$
(Q = nombre de variables actives de l'analyse)

2. On détermine la valeur-propre « seuil » tel que : $\lambda_i \geq \overline{\lambda}$

(Alors $\lambda_{i+1} < \overline{\lambda}$)

3. On pose L tel que ∀ i ≤ L, $\lambda_i \geq \overline{\lambda}$

4. Pour l'axe i, le taux modifié de Benzécri est égal à $\tau_i' = \frac{(\lambda_i - \overline{\lambda})^2}{\sum_{i=1}^{L} (\lambda_i - \overline{\lambda})^2}$

avec i ≤ L.

Variables supplémentaires

- Variable continue: calcul du coef de corrélation entre les var. quantitative et les composantes principales du nuages des individus (impossibilité de la représenter graphiquement)
- Variable qualitative: représentation graphique, relation barycentrique: une modalité supp. est représentée par le point moyen des individus qui prennent cette modalité

Recommandations

- Travailler avec un nombre de modalités comparable entre les variables actives de l'analyse
- Ne pas travailler avec des modalités aux effectifs trop faibles (>5%)

Retour
Tableau disjonctif complet
//
Tableau de Burt

- Exemple de Saporta

  Races de chiens en fonction de différentes caractéristiques

Données fictives (10 premières lignes)

```
##            taille poids velocite intellig affect agress  fonction
## beauceron     T++    P+      V++       I+    Af+    Ag+   Utilite
## basset         T-    P-       V-       I-    Af-    Ag+    Chasse
## ber_allem     T++    P+      V++      I++    Af+    Ag+   Utilite
## boxer          T+    P+       V+       I+    Af+    Ag+ Compagnie
## bull-dog       T-    P-       V-       I+    Af+    Ag- Compagnie
## bull-mass     T++   P++       V-      I++    Af-    Ag+   Utilite
## caniche        T-    P-       V+      I++    Af+    Ag- Compagnie
## chihuahua      T-    P-       V-       I-    Af+    Ag- Compagnie
## cocker         T+    P-       V-       I+    Af+    Ag+ Compagnie
## colley        T++    P+      V++       I+    Af+    Ag- Compagnie
```

TDC

```
##           T- T+ T++ P- P+ P++ V- V+ V++ I- I+ I++ Af- Af+ Ag- Ag+
## beauceron  0  0   1  0  1   0  0  0   1  0  1   0   0   1   0   1
## basset     1  0   0  1  0   0  1  0   0  1  0   0   1   0   0   1
## ber_allem  0  0   1  0  1   0  0  0   1  0  0   1   0   1   0   1
## boxer      0  1   0  0  1   0  0  1   0  0  1   0   0   1   0   1
## bull-dog   1  0   0  1  0   0  1  0   0  0  1   0   0   1   1   0
## bull-mass  0  0   1  0  0   1  1  0   0  0  0   1   1   0   0   1
## caniche    1  0   0  1  0   0  0  1   0  0  0   1   0   1   1   0
## chihuahua  1  0   0  1  0   0  1  0   0  1  0   0   0   1   1   0
## cocker     0  1   0  1  0   0  1  0   0  0  1   0   0   1   0   1
## colley     0  0   1  0  1   0  0  0   1  0  1   0   0   1   1   0
## dalmatien  0  1   0  0  1   0  0  1   0  0  1   0   0   1   1   0
## dobermann  0  0   1  0  1   0  0  0   1  0  0   1   1   0   0   1
## dogue_all  0  0   1  0  0   1  0  0   1  1  0   0   1   0   0   1
## epagn_bre  0  1   0  0  1   0  0  1   0  0  0   1   0   1   1   0
## epagn_fra  0  0   1  0  1   0  0  1   0  0  1   0   1   0   1   0
## fox_hound  0  0   1  0  1   0  0  0   1  1  0   0   1   0   0   1
## fox_terri  1  0   0  1  0   0  0  1   0  0  1   0   0   1   0   1
## grand_ble  0  0   1  0  1   0  0  1   0  1  0   0   1   0   0   1
## labrador   0  1   0  0  1   0  0  1   0  0  1   0   0   1   1   0
## levrier    0  0   1  0  1   0  0  0   1  1  0   0   1   0   1   0
## mastiff    0  0   1  0  0   1  1  0   0  1  0   0   1   0   0   1
## pekinois   1  0   0  1  0   0  1  0   0  1  0   0   0   1   1   0
## pointer    0  0   1  0  1   0  0  0   1  0  0   1   1   0   1   0
```

Tableau de Burt

```
##      T- T+ T++ P- P+ P++ V- V+ V++ I- I+ I++ Af- Af+ Ag- Ag+
## T-    7  0   0  7  0   0  5  2   0  3  3   1   1   6   5   2
## T+    0  5   0  1  4   0  1  4   0  0  4   1   0   5   3   2
## T++   0  0  15  0 10   5  4  2   9  5  6   4  12   3   6   9
## P-    7  1   0  8  0   0  6  2   0  3  4   1   1   7   5   3
## P+    0  4  10  0 14   0  0  6   8  3  7   4   7   7   8   6
## P++   0  0   5  0  0   5  4  0   1  2  2   1   5   0   1   4
## V-    5  1   4  6  0   4 10  0   0  4  5   1   5   5   5   5
## V+    2  4   2  2  6   0  0  8   0  1  5   2   2   6   5   3
## V++   0  0   9  0  8   1  0  0   9  3  3   3   6   3   4   5
## I-    3  0   5  3  3   2  4  1   3  8  0   0   6   2   3   5
## I+    3  4   6  4  7   2  5  5   3  0 13   0   4   9   8   5
## I++   1  1   4  1  4   1  1  2   3  0  0   6   3   3   3   3
## Af-   1  0  12  1  7   5  5  2   6  6  4   3  13   0   5   8
## Af+   6  5   3  7  7   0  5  6   3  2  9   3   0  14   9   5
## Ag-   5  3   6  5  8   1  5  5   4  3  8   3   5   9  14   0
## Ag+   2  2   9  3  6   4  5  3   5  5  5   3   8   5   0  13
```

- On peut réaliser une AC du tableau de Burt (pratique anglo-saxonne), mais dans ce cas on ne peut pas procéder à l'analyse des individus