

Représentation des Données et des Connaissances (RDC'05)



Le tableau de données, une structure unique, des réalités multiples

Yves Lechevallier

INRIA-Rocquencourt – AXIS

21 Mars 2005

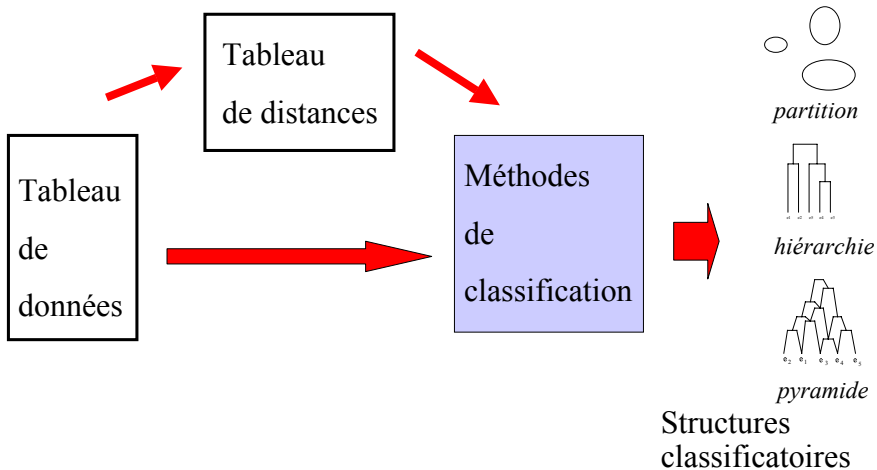
E_mail : Yves.Lechevallier@inria.fr

Plan



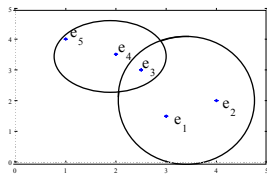
- La classification automatique
- Les concepts fondamentaux
- Le tableau de données
 - Tableau individus x variables
 - Tableau de contingence
 - Les autres formes
- Lien entre base de données relationnelle et tableau de données

La Classification Automatique



Structures classificatoires (1/2)

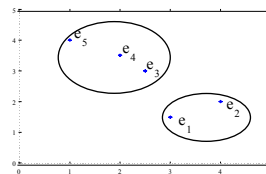
Recouvrement



1) $\forall \ell = 1, \dots, K$ on a $P_\ell \neq \emptyset$

2) $\bigcup_{\ell=1}^K P_\ell = E$

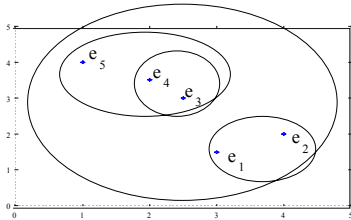
Partition



3) $\forall \ell, m = 1, \dots, K$ et $\ell \neq m$
alors $P_\ell \cap P_m = \emptyset$

Structures classificatoires (2/2)

Hiérarchie

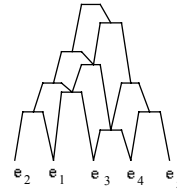


- 1) $E \in H$
- 2) $\forall e \in E$ alors $\{e\} \in H$
- 3) $\forall h, h' \in H$ on a :
 $h \cap h' \neq \emptyset \Rightarrow h \subset h' \text{ ou } h' \subset h$

Yves Lechevallier

RDC'05 ENST-Paris

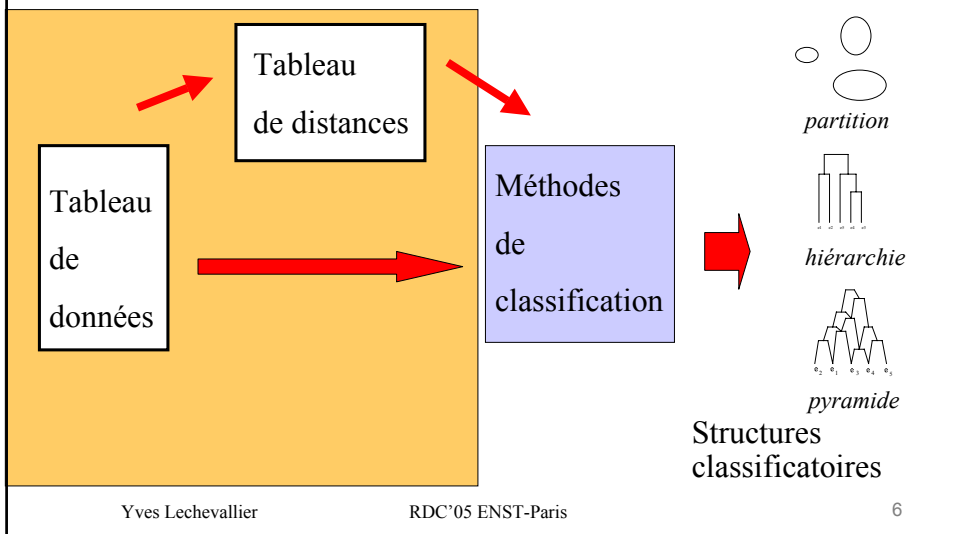
Pyramide



- 3) $\forall h, h' \in H$ on a $h \cap h' = \emptyset$ ou $h \cap h' \in H$
- 4) Il existe un ordre θ tel que
 $\forall h \in H, h$ est un intervalle de θ

5

La Classification Automatique



Yves Lechevallier

RDC'05 ENST-Paris

6

Tableaux de données

Concepts fondamentaux

- Ensemble des **individus**
- Ensemble des **variables** ou **caractères**
- Mise en correspondance
- Codage
- Mesure de **similarité** ou de **dissimilarité**
- Tableaux à N dimensions

Ensemble des individus

C'est l'**unité de base** sur laquelle des mesures vont être réalisées.

Le terme « **individu** » peut désigner un client d'un magasin, un animal, une ville.

L'ensemble des individus observés peut être un **échantillon** d'une population (sondage) ou la **population entière** (l'ensemble des départements français)

Ensemble des variables ou caractères

Sur ces individus on relève un certain nombre de mesures ou d'attributs.

Le choix des variables est étroitement lié au problème posé.

Afin de décrire les individus l'utilisateur va choisir un ensemble de variables.

- Pour une enquête les variables sont les questions.
- Pour la description d'une plante c'est un ensemble de descripteurs

Mise en correspondances

La réalisation d'un **tableau de données** se fait par la mise en correspondance ou la mise en relation de l'ensemble E des individus avec l'ensemble Y des variables

$$T : ExY \rightarrow D \quad T(x_i, Y_j) = Y_j(x_i)$$

Si le recueil des données se fait après la définition des deux ensembles, la construction de ce tableau est souvent facile mais cela devient plus difficile qu'en une analyse secondaire

Un exemple

Année	Partenaires privés	Apprentissage AMI	Commerce et industrie CMI	Transports TMA	Logement et aménagement du territoire LOG	Éducation et culture EDU	Autres ACS	Autres accusés AAS	Accidents commiss. ACC	Défense DEF	Déjà DEF	Librairie LIB	Total
1872	18,0	0,5	0,1	8,7	0,5	2,1	2,0			26,4	41,6	2,1	100
1879	14,1	0,5	0,1	13,3	1,9	3,7	0,5			29,8	31,3	2,5	100
1890	13,6	0,7	0,7	6,8	0,6	7,1	0,7			33,6	34,4	1,7	100
1900	14,3	1,7	1,7	6,9	1,5	7,4	0,8			37,7	36,2	2,2	100
1903	10,3	1,5	0,4	9,3	0,6	8,5	0,9			35,4	27,2	3,0	100
1906	13,4	1,4	0,5	8,1	0,7	8,6	1,8			38,5	25,3	1,9	100
1909	13,5	1,1	0,5	9,0	0,6	9,0	3,4			36,8	23,5	2,6	100
1912	12,9	1,4	0,3	9,4	0,6	9,3	4,3			41,1	19,4	1,3	100
1920	12,3	0,3	0,1	11,9	2,4	3,7	1,7	1,0		42,4	23,1	0,2	100
1923	7,6	1,2	3,2	5,1	0,6	5,6	1,8	10,0		29,0	35,0	6,9	100
1926	10,5	0,5	0,1	4,5	1,8	6,6	2,1	10,1		19,9	41,6	2,3	100
1929	10,9	0,6	0,6	9,0	1,0	8,1	3,2	11,3		28,0	25,8	2,0	100
1932	10,6	0,8	0,3	8,9	3,0	10,0	6,4	13,4		27,4	19,2	0	100
1935	8,8	2,6	1,4	7,8	1,4	12,4	6,2	11,3		29,3	18,5	0,4	100
1938	10,1	1,1	1,2	5,9	1,4	9,5	6,0	5,9		40,7	13,2	0	100
1947	15,6	1,6	10,0	11,4	7,6	8,8	4,8	3,4		32,2	4,6	0	100
1950	11,2	1,3	16,5	12,4	15,8	8,1	4,9	3,4		29,7	4,2	1,5	100
1953	12,9	1,5	7,0	7,0	12,1	8,1	5,3	3,9		36,1	5,2	0	100
1956	10,9	5,3	9,7	7,6	9,6	9,4	8,5	4,6		28,2	6,2	0	100
1959	13,1	4,4	7,3	5,7	8,8	12,5	8,0	5,0		26,7	7,5	0	100
1962	12,8	4,7	7,5	6,6	6,8	15,7	9,7	5,3		21,5	6,4	0,1	100
1965	12,4	4,3	8,4	9,1	6,0	19,5	10,6	4,7		19,8	3,5	1,8	100
1968	11,4	6,0	9,5	5,9	5,0	21,1	10,7	4,2		20,9	4,4	1,8	100
1971	12,8	2,8	7,1	8,5	4,0	23,8	11,3	3,7		18,8	7,3	0	100

Source : C. ANDRÉ et H. DELORME, *L'évolution des dépenses publiques en France (1872-1971)*
rapport CORDÉS, CEPREMAP, 1976.

Les **individus** sont les années, les **variables** représentent les dépenses

Population ou échantillonnage des observations

Notre information est contenue dans un ensemble E d'observations expérimentales. Chaque individu est associé une description qui est un vecteur de dimension p :

Le tableau de données \mathbf{Z} associé à l'ensemble E de N individus est une matrice ayant p colonnes et N lignes

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_N] = \begin{bmatrix} z_1^1 & z_1^j & z_1^p \\ z_i^1 & z_i^j & z_i^p \\ z_N^1 & z_N^j & z_N^p \end{bmatrix}$$

$E = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ est l'ensemble d'**apprentissage**

En statistique on suppose que l'ensemble E est un **échantillon** issu d'une population ayant une distribution inconnue.

Tableau de données

N individus $E = \{e_1, \dots, e_i, \dots, e_N\}$

p descripteurs $Y = \{Y_1, \dots, Y_j, \dots, Y_p\}$

À chaque individu e_i de E est associé un **vecteur de description** $(x_i^1, \dots, x_i^j, \dots, x_i^p)$ représentant les p mesures

	Y_1	...	Y_j	...	Y_p
e_1	x_1^1	...	x_1^j	...	x_1^p
...		
e_i	x_i^1	...	x_i^j	...	x_i^p
...		
e_N	x_N^1	...	x_N^j	...	x_N^p

À chaque variable ou paramètre Y_j est associé un vecteur $(x_1^j, \dots, x_i^j, \dots, x_N^j)$ représentant l'ensemble des N valeurs observées de E sur Y_j

Représentation tabulaire par un tableur

ID	Type d'iris	longueur du sepale	largeur du sepale	longueur du pétales	largeur du pétales
1	1 Setosa	5.1	3.5	1.4	0.2
2	2 Setosa	4.9	3	1.4	0.2
3	3 Setosa	4.7	3.2	1.3	0.2
4	4 Setosa	4.6	3.1	1.5	0.2
5	5 Setosa	5	3.6	1.4	0.2
6	6 Setosa	5.4	3.9	1.7	0.4
7	7 Setosa	4.6	3.4	1.4	0.3
8	8 Setosa	5	3.4	1.5	0.2
9	9 Setosa	4.4	2.9	1.4	0.2
10	10 Setosa	4.9	3.1	1.5	0.1
11	11 Setosa	5.4	3.7	1.5	0.2
12	12 Setosa	4.8	3.4	1.6	0.2
13	13 Setosa	4.8	3	1.4	0.1
14	14 Setosa	4.3	3	1.1	0.1
15	15 Setosa	5.8	4	1.2	0.2
16	16 Setosa	5.7	4.4	1.5	0.4
17	17 Setosa	5.4	3.9	1.3	0.4
18	18 Setosa	5.1	3.5	1.4	0.3
19	19 Setosa	5.7	3.8	1.7	0.3
20	20 Setosa	5.1	3.8	1.5	0.3
21	21 Setosa	5.4	3.4	1.7	0.2
22	22 Setosa	5.1	3.7	1.5	0.4
23	23 Setosa	4.6	3.6	1	0.2
24	24 Setosa	5.1	3.3	1.7	0.5
25	25 Setosa	4.8	3.4	1.3	0.4
26	26 Setosa	5	3	1.6	0.2
27	27 Setosa	5.1	3.4	1.5	0.4
28	28 Setosa	5	1.1		

Pas de sémantique ou de méta-données

Un autre exemple

Nom	Pays	Type	PG	CA	MG	NA	K	SUL	NO3	HCO3	CL
Evian	F	M	P	78	24	5	1	10	3,8	357	4,5
Montagne des Pyrénées	F	S	P	48	11	34	1	16	4	181	50
Cristaline-St-Cyr	F	S	P	71	5,5	11,2	3,2	5	1	290	20
Filè des Lois	F	S	P	89	31	17	2	47	0	360	28
Vakania	F	S	P	4,1	1,7	2,7	0,9	1,1	0,8	25,8	0,9
Saint Dèry	F	M	G	85	80	385	65	25	1,9	1380	285
Lucifon	F	M	P	26,5	1	0,8	0,2	8,2	1,8	78,1	2,3
Vielje	F	M	P	9,9	6,1	9,4	5,7	6,9	6,3	65,2	8,4
Alpes/Mouettes	F	S	P	63	10,2	1,4	0,4	51,3	2	173,2	1
Orlé du bois	F	M	P	234	70	43	9	635	1	292	62
Aeris	F	M	G	170	92	1,4	0,4	51,3	2	173,2	10
Alpes/Roche des Ecrins	F	S	P	63	10,2	1,4	0,4	51,3	0	2195	387
Oneline	F	S	P	46,1	4,3	6,3	3,5	9	0	163,5	3,5
Thomson	F	M	P	108	14	3	1	27	0,2	341	3
Aix les Bains	F	M	P	84	23	2	1	13	12	350	9
Cortèx	F	M	P	486	84	9,1	3,2	1187	2,7	403	8,6
La Bondoire Saint Hippolite	F	S	P	86	3	17	1	7	19	256	21
Dax	F	M	P	125	30,1	126	19,4	365	0	164,7	156
Quézac	F	M	G	241	95	255	49,7	143	1	1685,4	38
Sulzeat	F	M	G	253	11	7	3	25	1	820	4
Stamina	GRC	M	P	48,1	9,2	12,6	0,4	9,6	0	173,3	21,3
Iohi	GR	M	P	54,1	31,5	8,2	0,8	15	6,2	287,5	13,5
Avea	GR	M	P	110,8	9,9	8,4	0,7	39,7	35,6	308,8	8
Rosvas	GR	M	P	25,7	10,7	8	0,4	9,6	3,1	117,2	12,4
Alvea	IT	M	P	12,3	2,6	2,5	0,6	10,1	2,5	41,6	0,9
San Benedetto	IT	M	P	46	28	6,8	1	5,8	6,6	287	2,4
San Pellegrino	IT	M	G	208	55,9	43,6	2,7	549,7	0,45	219,6	74,3
Levissima	IT	M	P	19,8	1,8	1,7	1,8	14,2	1,5	56,5	0,3
Nera	IT	M	P	36	13	2	0,6	18	3,6	154	2,1
La Française	F	M	P	12,3	6,1	4,9	0,7	1,6	4,3	135,5	1
Saint Bonis	F	S	G	35,4	83	653	22	1055	0	225	982
Flançois	F	M	P	46,1	4,3	6,3	3,5	9	0	163,5	3,5
Saint Alla	F	S	P	36	19	36	6	43	0	195	38
Puits Saint Georges/Casim	F	S	P	10	33	4	20	0,5	84	37	
St-Georges/Casim	F	S	P	33	430	18,5	10	8	1373	39	
Hidron Hénar	GB	M	P	5,2	2,3	14,05	1,15	6	0	30,5	25
Hidron Hénar	GB	M	P	97	1,7	7,7	1	4	26,4	236	16

Les individus sont les marques d'eau en bouteilles

Les variables décrivent l'étiquette de cette eau

Les données sont «hétérogènes »

Chapitre 1 : G. Saporta et N. Niang

Analyse des données

G. Govaert Hermes 2003

M: minérale S: source

P: plate G: gazeuse

ions CA en mg/litre

Yves Lechevallier

RDC'05 ENST-Paris

15

Types de variables

Chaque variable aléatoire Y_j est une fonction mesurable de $\Omega \rightarrow D_j$

Ω est l'ensemble des observables

$$E \subseteq \Omega$$

Y_j est une variable **continue** ou **quantitative** si D_j est \mathbb{R}

X_j est une variable **discrète** ou **qualitative** si D_j est un ensemble fini $\{m_1, \dots, m_j\}$. Les éléments de D_j sont appelés **modalités** de la variable X_j .

X_j est une variable **ordonnée** s'il existe un ordre sur D_j .

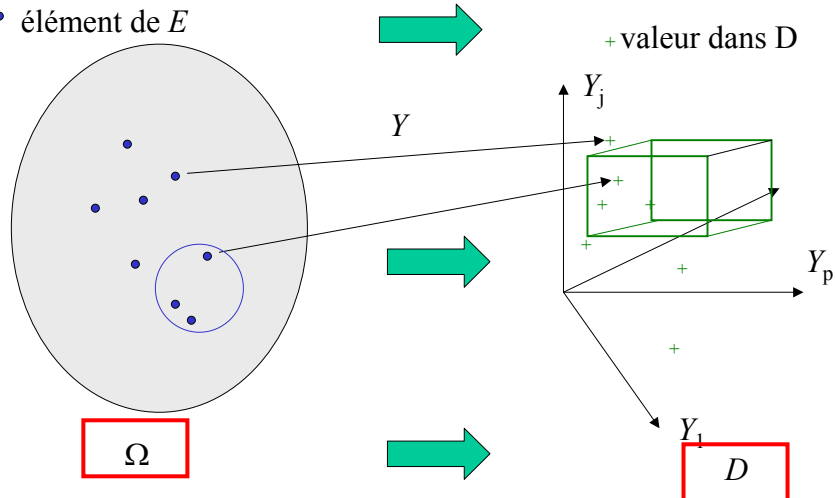
Yves Lechevallier

RDC'05 ENST-Paris

16

Espace de description ou de représentation

- élément de E



Yves Lechevallier

RDC'05 ENST-Paris

17

La première réalité

Beaucoup de tableaux de données ont cette réalité, c'est-à-dire ce sont des tableaux dits « individus x variables »

Chaque individu est souvent représenté par un point appartenant au produit cartésien $D = \prod_{j=1}^p D_j$

C'est dans cet espace que l'on mesure la proximité entre deux individus.

Cette distance est une distance entre les deux points

Yves Lechevallier

RDC'05 ENST-Paris

18

Tableau de données

Représentation d'un ensemble de 5 individus dans un plan.

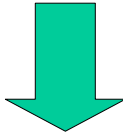
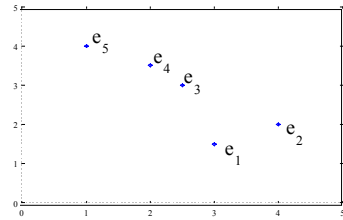


Tableau des distances entre ces 5 individus

	Distance				
e ₁	0				
e ₂	1,1	0			
e ₃	1,6	1,8	0		
e ₄	2,2	2,5	0,7	0	
e ₅	3,2	3,6	1,8	1,1	0

Deux sémantiques...

Variable	Moyenne	Ecart-type	Minimum	Maximum
CA	102.46	118.92	1.20	528.00
MG	25.86	28.05	0.20	95.00
NA	93.85	195.51	0.80	968.00
K	11.09	24.22	0.00	130.00
SUL	135.66	326.31	1.10	1371.00
NO3	3.83	6.61	0.00	35.60
HCO3	442.17	602.94	4.90	3380.51
CL	52.47	141.99	0.30	982.00

Tableau 1.3. Statistiques sommaires des variables continues

Un tableau individu x variable

	CA	MG	NA	K	SUL	NO3	HCO3	CL
CA	1.00							
MG	0.70	1.00						
NA	0.12	0.61	1.00					
K	0.13	0.66	0.84	1.00				
SUL	0.91	0.61	0.06	-0.03	1.00			
NO3	-0.06	-0.21	-0.12	-0.17	-0.16	1.00		
HCO3	0.13	0.62	0.86	0.88	-0.07	-0.06	1.00	
CL	0.28	0.48	0.59	0.40	0.32	-0.12	0.19	1.00

Tableau 1.4. Matrice des corrélations

Un tableau de proximités

✓ Tableau individus x individus

✓ Tableau variables x variables

Mesure de proximité

$E = \{e_1, \dots, e_i, \dots, e_N\}$ ensemble de N individus

(E, d)

d une mesure de proximité entre les individus de E

$$d : E \times E \rightarrow \mathfrak{R}^+$$

Mesure de ressemblance :

Plus deux individus sont proches plus la valeur de la mesure de ressemblance entre ces individus est **élevée**.

Mesure de dissemblance :

Plus deux individus sont proches plus la valeur de la mesure de dissemblance entre ces individus est **petite**.

Distance et similarité

$D = \prod_{j=1}^p D_j$ espace de description de E

Distance d

$$d : D \times D \rightarrow \mathfrak{R}^+$$

(1) $d(\mathbf{x}, \mathbf{y}) = 0$ si et seulement si $\mathbf{x} = \mathbf{y}$

(2) $\forall \mathbf{x}, \mathbf{y} \ d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

(3) $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \ d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$

Similarité s

$$s : D \times D \rightarrow \mathfrak{R}^+$$

(1) $\forall \mathbf{x} \ s(\mathbf{x}, \mathbf{x}) = S$

(2) $\forall \mathbf{x}, \mathbf{y} \ s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$

(3) $\forall \mathbf{x}, \mathbf{y} \ s(\mathbf{x}, \mathbf{y}) \leq s(\mathbf{x}, \mathbf{x}) = S$

Codage / transformation

Trois représentations des données en Analyse des Correspondances Multiples

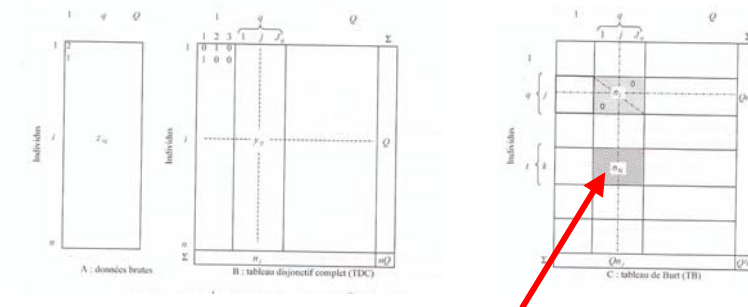
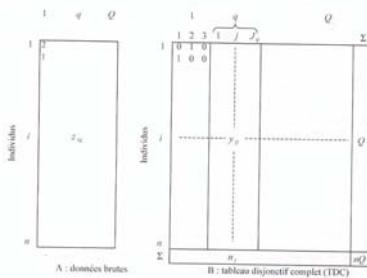


Tableau de contingence entre les variables Y_i et Y_q

Codage



A : Tableau de données

z_{iq} est la valeur de la variable Y_q sur l'individu i

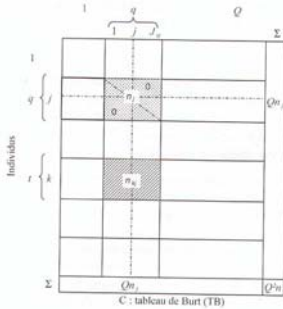
B : Tableau disjonctif complet, chaque colonne représente une modalité d'une variable.

$y_{ij}=1$ si l'individu i possède la modalité j de la variable Y_q

$y_{ij}=0$ sinon

Tableau de contingence multiple

C: Tableau de Burt



Les modalités sont croisées entre elles. Il juxtapose l'ensemble des tableaux de contingence entre toutes les variables prises deux à deux.

n_{kj} est le nombre d'individus possédant à la fois la modalité j de la variable Y_q et la modalité k de la variable Y_t .

Un seul ensemble en correspondance

	1001	1002	1004	1008	1014	1020	1022	1024	1025	1026	1029	1031	1037
1001	0	11	0	15	0	2	0	0	6	0	4	4	0
1002	0	0	1	4	10	0	0	0	0	0	2	3	0
1004	2	4	0	2	0	2	0	0	3	0	11	3	1
1008	11	5	7	35	1	15	1	6	18	5	19	23	0
1014	0	1	0	4	0	0	0	0	1	4	0	0	0
1020	1	2	2	18	0	5	0	1	6	3	5	10	0
1022	0	0	0	1	0	0	0	1	0	0	2	0	0
1024	1	0	2	3	0	3	0	0	2	1	6	7	0
1025	2	4	3	20	0	7	2	6	30	3	25	12	1
1026	0	0	0	9	0	4	0	1	3	0	0	2	0
1029	4	1	2	13	0	7	0	3	30	3	14	13	0
1031	0	0	0	0	0	1	0	1	0	1	11	36	2
1037	0	0	0	0	0	1	0	0	0	0	2	1	0

TAB 1 - Matrice de dénombrements des transitions dans ω

	1001	1002	1004	1008	1014	1020	1022	1024	1025	1026	1029	1031	1037
1001	14,665	0,4078	0	97,657	0	1,8295	0	0	14,665	0	6,5178	6,5178	0
1002	5,305	0,845	0,643	10,327	64,489	0	0	0	0	0	2,58	5,805	0
1004	3,44	13,78	0	3,44	0	3,44	0	0	7,7399	0	0,887	7,7399	0,887
1008	12,829	2,6507	5,1953	129,88	0,109	23,856	0,106	3,817	34,353	2,6507	38,276	58,088	0
1014	0	1,548	0	24,768	0	0	0	0	1,548	24,768	0	0	0
1020	0,2921	1,1963	1,1963	96,632	0	7,2018	0	0,2921	10,519	2,6287	7,2018	20,207	0
1022	0	0	0	3,87	0	0	0	3,87	0	0	15,48	0	0
1024	0,6192	0	2,4768	5,5728	0	5,5728	0	0	2,4768	0,6192	22,291	30,341	0
1025	0,5578	2,3313	1,2551	58,783	0	8,8355	0,5578	5,0205	84,274	1,2551	87,161	20,082	0,1385
1026	0	0	0	65,993	0	13,026	0	0,8147	7,5199	0	0	3,2569	0
1029	2,752	0,172	0,688	29,068	0	8,4279	0	1,548	154,8	1,548	33,712	29,068	0
1031	8,3253	3,2521	0,9203	57,367	0	4,683	0,1301	6,3741	46,96	0,1301	15,74	168,58	0,5203
1037	0	0	0	0	0	3,87	0	0	0	0	15,48	3,87	0

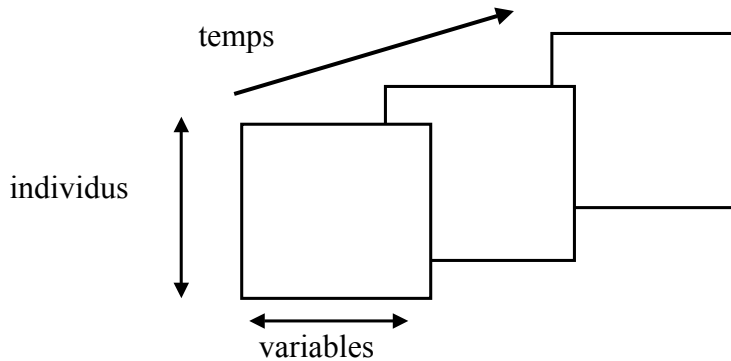
TAB 2 - Matrice BJ de la séquence ω

Tableau de comptage

Tableau de fréquences

Tableaux à plusieurs dimensions

Par exemple un tableau « individus x variables » peut être construit à différents instants.



Conclusions

Il faut des informations supplémentaires (meta-données) pour pouvoir analyser un tableau de données.

La représentation tabulaire (Excel) ne suffit pas.

Créer un langage de représentation d'un tableau ?

Latex, XML ...

Une solution ?

passage du monde des arbres au monde des tableaux

Article	Valeur de pH
Champignon d'ormeau	5.00
Crabe	6.60

FIG. 2 – pH approximatif des produits alimentaires

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<tableau><titre> <titre-tableau>pH approximatif des produits alimentaires</titre-tableau>
<tbody><tr><td>Article</td><td>Valeur de pH</td></tr><tr><td>Champignon d'ormeau</td><td>5.00</td></tr>
<tr><td>Crabe</td><td>6.60</td></tr></tbody></tableau>
```

FIG. 3 – Représentation d'un tableau en XTab

Produit	Qte	Lipides	Nombre de calories
merlan au citron	100 g	7,8 g	92 kcal
crabe de terre	150 g	11,25 g	192 kcal
poulet	250 g	18,75 g	312 kcal

FIG. 5 – Compositions nutritionnelles des aliments

VOIR SANSIMPLE TABLE, 8 REGLES SUR LES SITES 4 ET 5.10.10.10.

```
<tableau> <titre><titre-tableau>pH approximatif des produits alimentaires</titre-tableau>
<tbody><tr><td>Article</td><td>Valeur de pH</td></tr><tr><td>Champignon d'ormeau</td><td>5.00</td></tr>
<tr><td>Crabe</td><td>6.60</td></tr></tbody></tableau>
```

FIG. 4 – Représentation simplifiée du tableau de la figure 3 en SML

Xtab

SML

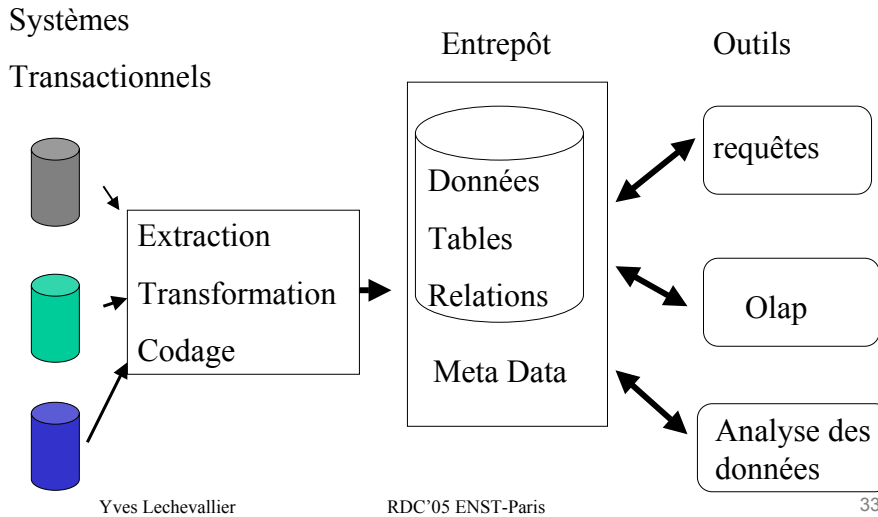
où est la sémantique ????

Construction d'un tableau de données

Analyse secondaire:

- les données sont dans une base de données relationnelle
- Couple (attribut, valeur)
- Schéma entité-relation
- Entrepôt de données

Architecture d'un entrepôt de données



Représentation d'une relation

Schéma de la relation

Req_simple(id, Region, superficie, nb_caisses)

id	Region	superficie	nb_caisses
Bassicaia	etranger		
Haut-Brion	graves	40	12000
Mission Haut-Brion	graves	20	8000
Latour	haut medoc	45	15000
Sociando Mallet	haut medoc	30	15000
Margaux	margaux	66	25000
Lafite-Rothschild	pauillac	90	35000
Lynch-Bages	pauillac	80	35000
Mouton-Rothschild	pauillac	75	20000
Pichon C. de Lalande	pauillac	60	28000
Pichon Longueville	pauillac	50	15000
Lafleur	pomerol	4,5	1600

Tableau de données ?

Schéma de relations

- **Approche base de données** : organiser les données de manière à éviter la redondance de l'information mais sans perte d'information
- **Approche statistique** : organiser les tables en fonction des unités statistiques que l'on désire analyser

Individus / Variables

- Recherche des **ensembles d'individus** qui doivent être analysés.
- Chaque ensemble d'individus est une **clé primaire** de la base.
- Associer à chaque ensemble d'individus les **variables / champs** les décrivant.
- Définition des « unités élémentaires » et des variables associées.

SGDB-R et tableau de données

La structure de données d'entrée d'un logiciel statistique est le **tableau de données** qui met en correspondance un ensemble d'**individus** avec un ensemble de **variables**.

C'est une structure très simple :

☞ A chaque **colonne** est associé une **variable**

☞ Chaque **ligne** contient la **description d'un individu**

Exemple de base de données relationnelle

Dans la base *wine.mdb* se trouve 4 tables :

appellation, gouteur, chateaux et deguste

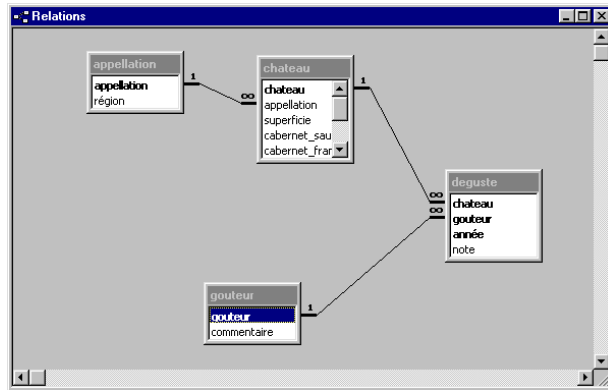
et 3 populations :

- *châteaux* 23 unités

- *années* 3 {1983, 1985 et 1990} unités

- *gôûteurs* 21 unités

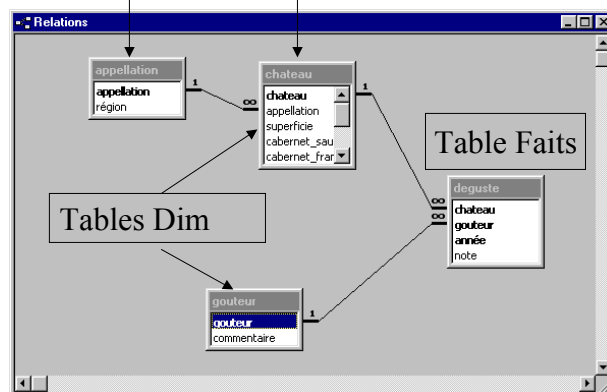
Le schéma des relations entre les tables



On choisit le champ *château* comme population

Modélisation orientée sujet

Niveau de détail (grain)



Une table de dimension

id	Region	superficie	nb_caisnes
Bassicala	etranger		
Haut-Brion	graves	40	12000
Mission Haut-Brion	graves	20	8000
Latour	haut medoc	45	15000
Sociando Mallet	haut medoc	30	15000
Margaux	margaux	66	25000
Lafite-Rothschild	pauillac	90	35000
Lynch-Bages	pauillac	80	35000
Mouton-Rothschild	pauillac	75	20000
Pichon C. de Lalande	pauillac	60	28000
Pichon Longueville	pauillac	50	15000
Lafleur	pomerol	4,5	1600

Liens entre le tableau de données et la base de données

Comment passer de la structure « base de données » à la structure du « tableau de données » ?

Problème très difficile et ouvert

➤ **On transforme le résultat d'une requête (Vue) en un tableau de données**

- ❑ Construire des **résumés** ou des **données agrégées**
- ❑ Utiliser des **opérateurs de généralisation**

Requête de synthèse/ Résumé

Individus	Groupe	X1	X2
I1	G1	3	1
I2	G1	6	1
I3	G1	1	6
I4	G1	5	1
I5	G2	6	3
I6	G2	0	2
I7	G2	3	4

Requête de synthèse

	X1	X2
G1	3,25	2,5
G2	3	3

Opérateurs de généralisation

La généralisation de chaque groupe est un processus qui utilise une approche non supervisée

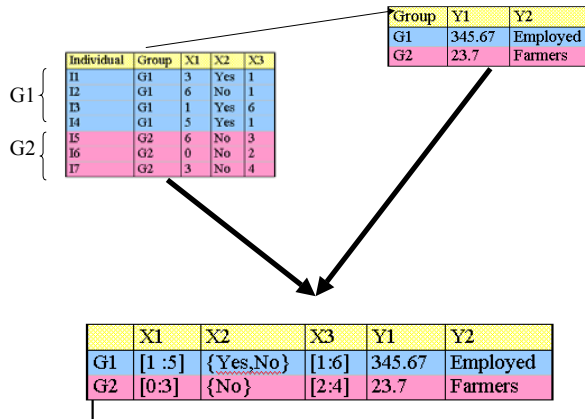
La description est obtenue à partir des conjonctions des descriptions de chaque individu du groupe

$$d_j(G_i) = Y_j(\omega_1) \oplus \dots \oplus Y_j(\omega_{n_i}) \quad \begin{array}{l} \forall j=1, \dots, p \\ n_i = |C_i| \end{array}$$

d_j : opérateur de généralisation appliqué sur la variable Y_j

Opérateurs de généralisation

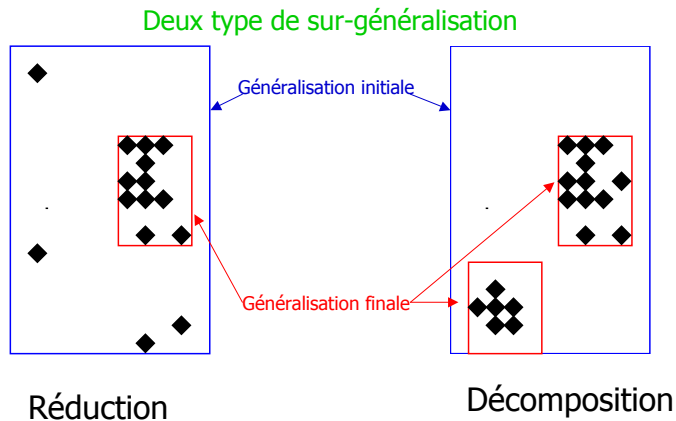
A partir d'un groupe d'individus directement



Méthode de généralisation

- **Approche non supervisée** : Extraire les caractéristiques des individus du groupe G.
- **Approche supervisée** : Définir les spécificités des éléments de G par rapport aux éléments n'appartenant pas à G. (exemples / contre-exemples)
- **Adéquation** de la généralisation en fonction des données.

Sur-généralisation



Yves Lechevallier

RDC'05 ENST-Paris

47

Enquête ASU

- Décomposer l'information en respectant la structure du questionnaire en différents thèmes.
- individus:

Utilisateurs

numero

Logiciels

logiciel

Méthodes

methode

Machines

machine

BDR

sgbd

Yves Lechevallier

RDC'05 ENST-Paris

48

Construction d'un tableau de données

- Les lignes = individus
- Les colonnes = variables

Deux solutions

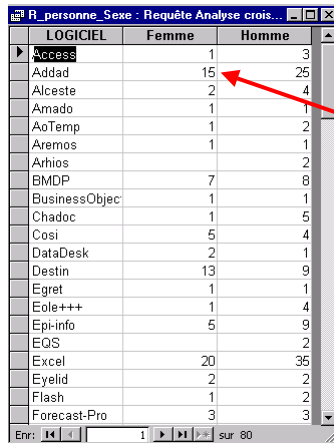
- une table de contingence / croisement
- juxtaposition de tables de contingences

Chapitre 5; du livre
« Statistique textuelle »
L. Lebart et A. Salem,
DUNOD

Table de contingence croisement simple

- Lignes = les logiciels
- Colonnes = ensemble des modalités de la variable « sexe »
(*table* : utilisateur, *champ* : sexe)
- Vue : table de contingence

Résultat de la requête



LOGICIEL	Femme	Homme
Access	1	3
Addad	15	25
Alceste	2	4
Arnado	1	1
AoTemp	1	2
Aremos	1	1
Arhios		2
BMDP	7	8
BusinessObjec	1	1
Chadoc	1	5
Cosi	5	4
DataDesk	2	1
Destin	13	9
Egret	1	1
Eole+++	1	4
Epi-info	5	9
EQS		2
Excel	20	35
Eyelid	2	2
Flash	1	2
Forecast-Pro	3	3

Le nombre 15 correspond aux nombre de femmes utilisant le logiciel Addad

Table de contingence croisement double

- Lignes = les logiciels
- Colonnes = produit cartésien des deux ensemble des modalités des variables « sexe » et « secteur »

table : utilisateur, *champ* : sexe

table : entreprise, *champ* : secteur

- Vue : table de contingence

Résultat de la requête

LOGICIEL	Femme Admini	Femme Collectif	Femme Privé	Femme Pu	Homme Admi
Addad	12	1	2		17
Alceste	1				2
Arnado	1				
AoTemp		1			1
Aremos					
Arhios					
BMDP	6		1		5
BusinessObjec			1		1
Chadoc	1				3
Cosi	1		3	1	
DataDesk	1		1		1
Destin	1	1	8	3	1
Egret	1				1
Eole+++	1				
Epi-info	4		1		5
EQS					2
Excel	5		13	2	6
Eyelid	1		1		2
Flash	1				
Forecast-Pro	2		1		2
Gauss	2				2
Genstat	1				2
GLIM					2
Harry					

Le nombre de femmes utilisant le logiciel Addad est ventilé en fonction du secteur d'activité de ces femmes.

Juxtaposition de tables de contingences

- Lignes = les logiciels
- Colonnes = ensemble des champs de la table « utilisateur_logiciel_methode »
- Vue : plusieurs lignes par logiciel

Résultat de la requête

Expr1	nom_logiciel	Analyse de la variabilité	Analyse discriminatoire	Analyses de survie	Analyses factorielles
10002	Addad	Tressouvent	Ponctuellement	Jamais	Tressouvent
104002	Addad	Jamais	Ponctuellement	Jamais	Tressouvent
149002	Addad	Ponctuellement			Ponctuellement
167002	Addad	Ponctuellement	Ponctuellement		Tressouvent
169002	Addad	Tressouvent	Tressouvent	Jamais	Tressouvent
170002	Addad	Souvent			Tressouvent
181002	Addad	Jamais	Souvent	Jamais	Tressouvent
22002	Addad	Jamais	Jamais	Jamais	Tressouvent
227002	Addad	Souvent	Ponctuellement	Jamais	Tressouvent
249002	Addad	Souvent	Souvent	Ponctuellement	Souvent
250002	Addad	Ponctuellement	Ponctuellement	Jamais	Souvent
256002	Addad	Tressouvent	Ponctuellement	Jamais	Jamais
267002	Addad	Ponctuellement	Ponctuellement	Jamais	Souvent
270002	Addad	Souvent	Ponctuellement	Jamais	Tressouvent
27002	Addad	Jamais	Ponctuellement	Jamais	Souvent
3002	Addad	Ponctuellement	Jamais	Ponctuellement	Tressouvent
31002	Addad	Ponctuellement	Ponctuellement	Jamais	Souvent
317002	Addad	Souvent	Tressouvent	Ponctuellement	Tressouvent
32002	Addad	Souvent	Jamais	Jamais	Souvent

Comment regrouper ou résumer ou agréger ces lignes ?
Choix de l'opérateur de généralisation

Yves Lechevallier

RDC'05 ENST-Paris

57

Conclusions

- Le tableaux de données est la structure de données essentielle dans le domaine de la classification.
- Il y a beaucoup d'informations implicites dans la représentation d'un tableau de données.
- La partie « sémantique » d'un tableau de données est souvent cachée.
- Le modèle relationnel permet une bonne représentation tabulaire des données, le modèle XML pourra t'il intégrer correctement l'aspect «sémantique » du tableau de données ?

Yves Lechevallier

RDC'05 ENST-Paris

58

Bibliographie



- H.-H. Bock, E. Diday, *Analysis of Symbolic Data*, Springer Verlag, 2000.
- J-M Bouroche, G. Saporta, *L'analyse des données*, PUF, Que sais-je?, 1980
- G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, H.Ralambondrainy, *Classification automatique des données; environnement statistique et informatique*. Dunod, Paris, 1989.
- G. Govaert, *Analyse des données*, Hermes, 2003
- L. Lebart, A. Salem, *Statistique textuelle*, Dunod, 1994.
- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*, Dunod, 2000.
- F. Sermier, Bases de données et logiciels statistiques, Journées du groupe logiciel de la SfDS, Juin 1999 (et XML et statistique en juin 2004)
- F. Salis, H. Gagliardi, O. Haemmerlé, N. Pernelle, Enrichissement sémantique de documents XML, représentant des tableaux, EGC 2005, Paris pp 407-418.
- Saporta G., *Probabilités, analyse des données et statistique*, Technip, 1990.