. Russel Rao $\quad s(i_1, i_2) = \frac{2}{2+1+0+1} = \frac{2}{6}$.

III / Partition et hiérarchie.

III.1 / Partition

Soit E une population à n individus.

On note par $\mathcal{P}(E)$ l'ensemble de toutes les parties de E.

et soit $P(E) = \{E_1, \dots, E_k\} \subset \mathcal{P}(E)$ tq : . $E_i \neq \phi \quad \forall i = 1, \dots, k$

. $\bigcup_{i=1}^{k} E_i = E$

. $E_i \cap E_j = \phi$ pour $i \neq j$

P(E) est une partition de E.

Exp : $E = \{1, 2, 3, 4\}$. $P(E) = \{\{1,3\}, \{2,4\}\}$ partition de E

III.2 / hiérarchisation de parties d'un ensemble

Def 1 : Soit E un ensemble finie. et soit H un ensemble de parties de E $(H \subset \mathcal{P}(E))$

H est une hiérarchie sur E ssi

. $\forall x_i \in E \quad \{x_i\} \in H$

. $E \in H$

. $\forall h_1$ et $h_2 \in H$ on a soit $h_1 \cap h_2 \neq \phi$ ou bien $(h_1 \subset h_2$ ou $h_2 \subset h_1)$

(si l'intersection est vide l'une est inclue dans de l'autre.

. Toute classe $h \in H$ est la reunion des classes qui sont contenues ds h

Exp. $E = \{1, 2, 3, 4, 5, 6\}$.

$H = \{E, \phi, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}, \{1,2,3\}, \{4,5,6\}\{4,5\}\}$

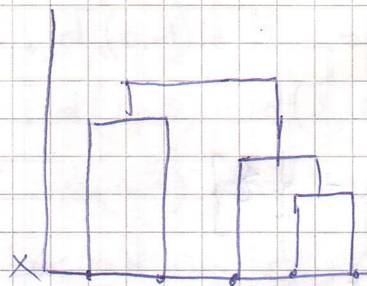
Def 2 : Une hiérarchie indicée est un couple (H, f) ou H est une hiérarchie et $f : H \longrightarrow \mathbb{R}_+$ tq pour $h \in H$

. $f(h) = 0 \iff h$ ne contient qu'un seul elt

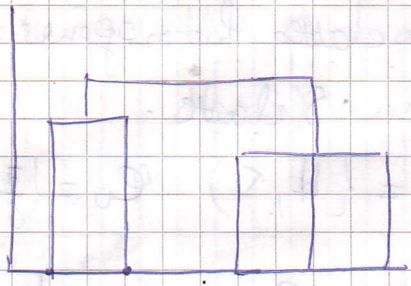
. $\forall h, h' \in H \quad h \subset h' \Rightarrow f(h) < f(h')$ "$f \nearrow$"

Rq : si l'inegalité est large, on parle de l'hierarchie indicée au sens large.

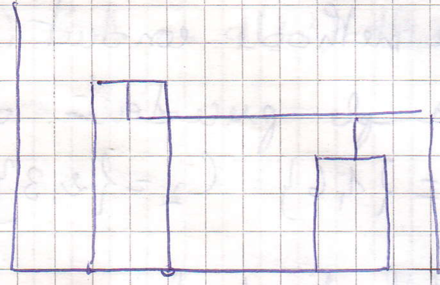
. S'il existe h et h' tq $h \subset h'$ et $f(h) > f(h')$ on dit qu'il y a inversion.



hierarchie indicée



hierarchie indicée au sens large



hierarchie avec inversion

indices sont appelés niveau d'agregation. ** voir page supp.

IV / Methode de classification basée sur la ressemblance.

Soit $E = \{x_1, \ldots, x_n\}$ à n inds à classifier. Partition.

Le but : chercher à représenter ses inds (p^{ts}) par des ensembles de pa ~~hierarchiquement emboités~~

Afin de proceder au regroupement de inds de E, on calcul Les distances (ou dissimilarités) entre les inds de E pris 2 à 2.

Plusieurs ~~tec~~ techniques ont été utilisée, citons les + importante.

a/ Methode de liaison simple : consiste à regrouper les inds les inds d'une m̂ classe ont une distance (ou dissimilarité) 2 à 2 $\leq$ à un seuil fixé δ.

Si par exemple C est une classe $\forall i, i' \in C \Rightarrow d(i,i') \leq \delta$

Rq : La procedure est simple, mais on aboutit generalement à des classes heterogenes ds le sens où chaque ind n'est pas forcement relié à touts les inds de la classes.

Exp : $E = \{1, 2, 3, 4, 5, 6, 7, 8\}$ et la matrice des distances est

E \ E	1	2	3	4	5	6	7	8
1	0							
2	5	0						
3	3	1	0					
4	2	6	4	0				
5	5	8	2	1	0			
6	1	3	4	3	3	0		
7	6	5	2	2	2	3	0	
8	2	1	5	4	4	3	2	0

Si $\delta = 1 \Rightarrow$ 4 classes

$C_1 = \{1, 6\}$ $C_2 \{2, 3, 8\}$

$C_3 = \{4, 5\}$ $C_4 = \{7\}$

$d(2,3) = d(8,2) = 1 \leq \delta$

mais $d(8,3) = 5 > \delta$

C_2 est donc heterogene

b) Methode de liaison complète : Chaque inds est relié à touts les inds de sa classe à un niveau $\leq \delta$.

Cette methode conduit à des classes homogènes.

Exemple precedent on aura 5 classe.

$C_1 = \{1, 6\}$ $C_2 = \{2, 3\}$ $C_3 = \{4, 5)$ $C_4 = \{7\}$ $C_5 = \{8\}$

c) Methode de liaison moyenne = Regroupement par étapes successives. Chercher la paire de base constituée d'inds les plus proches, calculer les distances moyennes (ou dissimilarité moyennes) entre cette paire et les autres inds. On obtient ainsi une matrice de dimension $(n-1, n-1)$. repérer la distance (ou dissimilarité) la plus faible $\leq$ à un seuil donné.

- si cette distance (ou dissimilarité) correspond à une liaison entre deux inds isolés, ces derniers sont regroupés ds une classe nouvelle matrice.
- Si elle correspond à la liaison entre la paire de base et 1 inds isolé on construit une classe à 3 inds. et ainsi de suite

L'operation s'arrête lorsque les termes de la dernière matrice obtenu soit strictement superieure au seuil fixé.

Exemple : Soit le tableau de distance précédent. On fixe $\delta = 1$

Soit $\{2, 3\}$ la paire de base de depart

	1	{2,3}	4	5	6	7	8
1	0						
{2,3}	$\frac{5+3}{2} = 4$	0					
4	2	$\frac{6+4}{2} = 5$	0				
5	5	$\frac{8+2}{2} = 5$	1	0			
6	1	$\frac{3+4}{2} = 3{,}5$	3	3	0		
7	6	$\frac{5+2}{2} = 3{,}5$	2	2	3	0	
8	2	$\frac{1+5}{2} = 3$	4	4	3	2	0

La distance la plus faible $\leq \delta = 1$ correspond à la liaison entre 5 et 4

$d(5, 4) = 1 \leq 1$

On reuni ds une nouvelle matrice $(6, 6)$

E\E	1	{2,3}	{4,5}	6	7	8
1	0					
{2,3}	4	0				
{4,5}	$\frac{2+5}{2}=3,5$	$\frac{5+5}{2}=5$	0			
6	1	3,5	$\frac{3+3}{2}=3$	0		
7	6	3,5	$\frac{2+2}{2}=2$	3	0	
8	2	3	$\frac{4+4}{2}=4$	3	2	0

$d(1,6) = 1 \leq 1$

on reuni 1 et 6 ds 1 m̂ classe.

on obtient une matrice (5×5)

	{1,6}	{2,3}	{4,5}	7	8	~~8~~
{1,6}	0					
{2,3}	$\frac{3,5+3,5}{2}=3,75$	0				
{4,5}	$\frac{3,5+3}{2}=3,25$	5	0			
~~6~~						
7	$\frac{6+3}{2}=4,5$	3,5	2	0		
8	$\frac{2+3}{2}=2,5$	3	4	2	0	

On arrête puisque toute les distance sont > au seuil $s = 1$

on obtient donc les classes suivantes.

{1,6} ; {2,3} ; {4,5} {7} {8}.

V / Classification hiérarchique ascendante.

V.1/ Hierarchie de partie d'un ensemble.

Le but est de regrouper progressivement les inds d'un ensemble E sur la base de leurs ressemblance. Une hierarchie peut être representée par un arbre de classification.

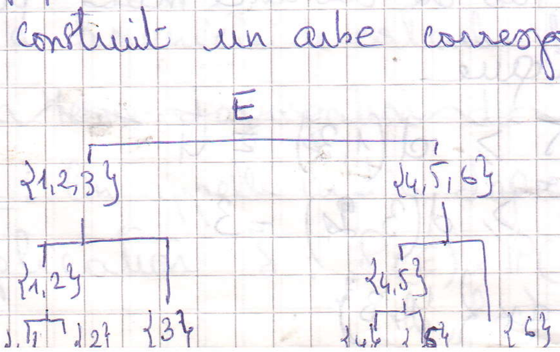
Exemple.

Soit $E = \{1, 2, 3, 4, 5, 6\}$

et $H = \{\phi, \{1\}; \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}; \{1,2,3\}, \{4,5\} \{4,5,6\}, E\}$

exp presédent hierarchie de parties de E.

On construit un arbe correspondant H de la manière suivante.



C'est un arbre ascendant.

On regroupe d'abord des ~~objets~~ inds les plus proches qui forme un noeud ou sommet, il ne reste donc q que (n-1) inds et on répète l'opération jusqu'à regroupement complet.

I.2/ Compatibilité d'une hiérarchie avec une partition.

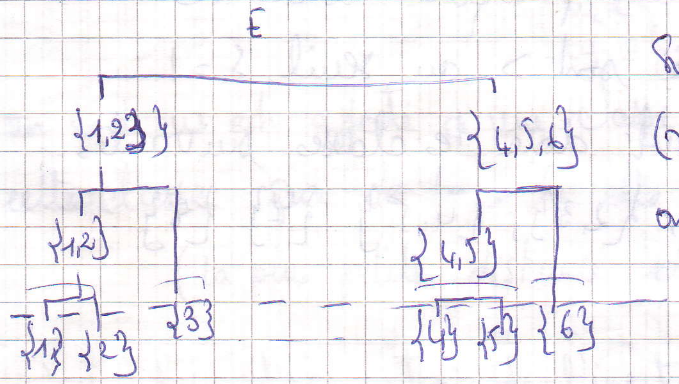
Def : une hiérarchie H est compatible à une partition d'un ensemble E si toute classe de la partition est ds H.

Exemple :

Soit la partition de $E = \{1, 2, 3, 4, 5, 6\}$.

$C_1 = \{1, 2\}$ $C_2 = \{3\}$ $C_3 = \{4, 5\}$ $C_4 = \{6\}$.

ds l'exp précédent C_1, C_2, C_3 et C_4 appartiennent à H



Si on coupe l'arbre par une horizontale (regardez les pointillés) on obtient la partition juste sur l'horizontal

I.3/ Hiérarchie indicée et ultramétrique.

Définition : Une ultra métrique est une application $\delta : E \times E \longrightarrow \mathbb{R}_+$ tq

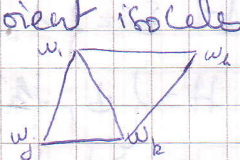
1 - $\forall w_i, w_j$ $\delta(w_i, w_j) \geq 0$

2 - $\delta(w_i, w_j) = 0 \Rightarrow w_i = w_j$

3 - $\delta(w_i, w_j) = \delta(w_j, w_i)$

4 - $\forall i, j, k$ $\delta(w_i, w_j) \leq \max\{\delta(w_i, w_k), \delta(w_k, w_j)\}$

Proposition : Une condition suffisante et nécessaire pour que une distance soit une ultramétrique est que toutes les triangles de E soient isocèles avec la base inférieure aux côtés



Exemple.

E\E	1	2	3	4
1	0	5	4	5
2	5	0	3	1
3	4	3	0	2
4	5	1	2	0

c'est un tab de distance mais pas ultrametrique

$d(1,2) = 5 \geq d(1,3) = 4$

$\geq d(3,2) = 3$

$\Rightarrow d(1,2) \geq \max\{4, 3\}$.

Le tab suivant est un tab d'ultrametrique

E \ E	1	2	3	4
1	0	6	0	6
2	6	0	5	2
3	0	5	0	5
4	6	2	5	0

Theorème : Il existe une bijection entre une hiérarchie indicée et ultrametrique.

En d'autres termes, toute hiérarchie indicée sur E correspond une ultrametrique et inversement.

On propose deux familles de méthodes pour obtenir une hiérarchie indicée : - Les regroupements progressifs

- Les passages directes à l'ultrametrique

V. 4 / Les regroupements progressifs.

On a plusieurs étapes successives.

Sur le tableau des distances ou dissimilarité, on repère la plus petite valeur

Soit A la classe $\{(w_i ; w_{i'})\}$ tq $d(w_i, w_{i'})$ = plus petite valeur.

On calcule la distance ~~ou~~ dissimilarité de chaque ind à A, on obtient un tableau de dim $(n-1, n-1)$

On repère du nouveau tableau la dist ou diss la plus petite

On répète la procedure jusqu'à obtenir un regroupement complet

Rq si la + petite valeur correspond à une dist ou dissi entre deux autres ind il seront regroupés ds une m̃ classe, si cela correspond à dist ou dissi entre un ind et A on forme une classe B de 3 inds.

On a 3 manière de calculer les distances par utilisation d'indice ~~au Mesure d'agrégation du lien minimum~~. d'agrégation

Déf : On appelle indice d'agrégation entre deux groupes d'individus une application $\delta : \mathbb{P}(E) \times \mathbb{P}(E) \longrightarrow \mathbb{R}+$

$(h_1, h_2) \longmapsto \delta(h_1, h_2)$ tq $\delta(h_1, h_2) = \delta(h_2, h_1)$