LES CAHIERS DE L'ANALYSE DES DONNÉES

F. BENZÉCRI
J. P. BENZÉCRI

# Démonstration de l'équivalence des résultats des algorithmes accélérés à ceux de l'algorithme de base en CAH

*Les cahiers de l'analyse des données*, tome 10, n° 3 (1985), p. 257-271

<http://www.numdam.org/item?id=CAD_1985__10_3_257_0>

*Les Cahiers de l'Analyse des Données*
*Vol. X - 1985 - n° 3 - pp. 257-271*

# DÉMONSTRATION DE L'ÉQUIVALENCE DES RÉSULTATS DES ALGORITHMES ACCÉLÉRÉS A CEUX DE L'ALGORITHME DE BASE EN CAH

## [ALG. & ALG. C.A.H.]

*par F. Benzécri * & J.P. Benzécri ***

On sait que l'algorithme de base de la CAH crée les noeuds un par un dans l'ordre croissant de leurs niveaux en agrégeant à chaque étape une paire de sommets entre lesquels est réalisé le minimum de la distance (ou plus exactement de la fonction utilisée comme critère). Les algorithmes accélérés (cf. de Rham ; [*C.A.H. VOIS. RECIP.*] *in* C.A.D. Vol. V n° 2 1980 ; J. Juan [*PROG. C.A.H. RECIP*] *in* C.A.D. Vol VII n°2 1982) au contraire créent les noeuds plus librement, éventuellement par paquets, sous la seule contrainte de n'agréger en une étape que des paires de sommets qui soient plus proches voisins réciproques l'un de l'autre (i.e. dont chacun réalise le minimum de la distance à l'autre). Il importe cependant de démontrer rigoureusement que toute hiérarchie construite en usant des libertés qu'offre un algorithme accéléré aurait également pu l'être sous les conditions strictes de l'algorithme de base. La présente note propose une démonstration qui sans sortir (du moins le croyons-nous !) des règles mathématiques emprunte les notations du langage ALGOL, pour décrire des structures. Afin de soutenir l'intuition du lecteur, l'exposé abstrait est illustré de planches encadrées.

### 1 *Structure d'arbre indicé*

*1.1 Description d'une structure d'arbre indicé* : Nous énumérons d'abord sous forme de déclaration les éléments constitutifs de cette description ; puis nous donnons les conditions qui doivent être satisfaites pour qu'on ait la description d'une structure d'arbre indicé.

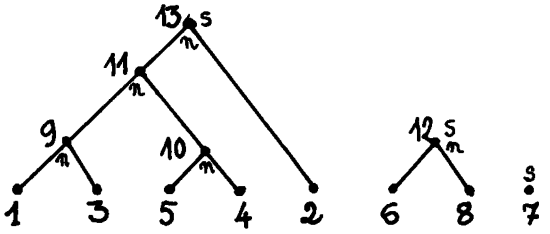
```
entier CARI,CARN;
entier tableau A,B[CARI+1:CARI+CARN];PER,SOM[1:CARI+CARN];
réel tableau D[1:CARI+CARN].
```

*Commentaire :* Pour comprendre ces déclarations, on se souviendra de la représentation graphique usuelle d'un arbre. Soit EI un ensemble dont les éléments sont numérotés de 1 à CARI ; ces éléments sont situés au niveau zéro. D(I)=0 ; au dessus des éléments est placé un ensemble EN de noeuds, numérotés de (CARI+1) à (CARI+CARN). Il est commode d'appeler classe et de désigner par la lettre C aussi bien un élément de EI que de EN : EC = EI ∪ UN. Chaque noeud N est relié à deux classes numérotées A[N] et B[N] (dites aînées et benjamin ; et aussi fils) qui lui sont immédiatement inférieures ; si une classe C n'est aîné ou benjamin d'aucune autre, elle est un sommet, ce qu'on écrit : SOM[C]= 1 ; sinon SOM[C] = 0, et on note PER[C] (père de C) la classe dont C est l'aîné ou le benjamin. Si SOM[C] = 1 on peut donner à PER[C] une valeur arbitraire supérieure à CARI+CARN. Pour que les tableaux de nombres A,B,PER,D correspondent effectivement à la description d'un arbre, il faut donc imposer à leur contenu des conditions telles que celles données ci-dessous.

(* ) *Docteur ès-sciences.*

(**) *Professeur de statistique. Université Pierre et Marie Curie.*

Les cahiers de l'analyse des données - 0339-3097/85/03 257 15/$ 3.50/ 




[ALG % ALG CAH] § 1.1 *ici comme dans la suite, un tableau est représenté par un cadre dont les cases sont numérotées à l'extérieur selon les dimensions déclarées ; le contenu des cases étant les valeurs numériques : e.g. dans la case numérotée 11 du tab. A on lit A[11]=9.*

| | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|
| A | 1 | 5 | 9 | 6 | 11 |
| B | 3 | 4 | 10 | 8 | 2 |

*Conditions*

CARI≥1;CARN≥0;
I∈[1:CARI]⇒D[I]=0;
N∈[CARI+1 : CARI+CARN] ⇒
1≤A[N]<$_{S}$N;1≤B[N]<$_{S}$N;A[N]≠B[N];D[N]≥0.
C∈[1:CARI+CARN]⇒SOM[C]=1-...
Card{N|N∈[CARI+1:CARI+CARN];C∈{A[N],B[N]}};
SOM[C]∈{0,1};
SOM[C]=0⇒PER[C]=2CARI;
C∈{A[N],B[N]}⇒PER[C]=N.

*Remarque 1.* Convenons de dire que N est un père de C si N∈[CARI+1: CARI+CARN] et que (C=A[N])V(C=B[N]);SOM[C] est ici défini comme 1 - (nombre de pères de C) ; la condition SOM[C]∈{0,1} signifie donc que C a au plus un père. (Pour les fonctions A et B, cela implique que la valeur C est prise au plus une fois par les deux fonctions ensemble). Si C a un père unique, le numéro de celui-ci est PER[C] ; sinon, on donne à PER[C] la valeur de 2CARI qui n'est le numéro d'aucun noeud.

*Remarque 2.* Il est naturel de demander que tout noeud N soit situé à un niveau D[N] supérieur ou égal à celui de ses deux fils A[N] et B[N] ; toutefois il est commode de ne pas imposer cette condition *a priori*, mais de démontrer (cf. § 4.2, corollaire) qu'elle est vérifiée par toute description compatible avec un critère satisfaisant à l'axiome de la médiane (§ 3).

*Remarque 3.* Le nombre des sommets de l'arbre est : CARI-CARN car, en bref, chaque noeud ayant deux fils, la qualité de fils (ou de non-sommet) appartient à 2CARN classes ; restent donc des sommets au nombre de ((CARI+CARN)- 2CARN)=CARI-CARN. Comme il y a toujours au moins un sommet, on a : CARN≤CARI-1.

*Remarque 4.* Le cas D[N] = 0 se rencontre effectivement lorsque A(N) et B(N) sont des individus identiques (quant à leur description) portant des numéros différents dans I.

1.2 *Equivalence de deux descriptions* : Notre but est de démontrer que des algorithmes qui créent les noeuds dans des ordres différents produisent cependant les mêmes arbres. Il faut donc définir ce qu'on entend par "descriptions équivalentes" : cette démarche est d'ailleurs familière aux mathématiciens.

Soit Desc0 et Desc1 deux descriptions d'arbres. (Sur la figure on voit deux descriptions d'un même arbre définies par deux numérotages des noeuds : le numérotage de la description Desc0 est indiqué en caractères gras ; celui de la description Desc1 est indiqué entre parenthèses en caractères maigres). On note :

Desc t ={CARIt, CARNt, At, Bt, PERt, SOMt, Dt}, où t est 0 ou 1. On dit que ces deux descriptions (Desc0 et Desc1) sont équivalentes sous les conditions suivantes :

*Conditions :*

CARI1 = CARI0 ; CARN1 = CARN0 ;

(on notera donc simplement CARI, CARN) ; il existe deux applications Tt de l'intervalle[1:CARI+CARN] dans lui-même telles que (en notant: t = 0,1 ; t' = (1-t)) :

C∈[1:CARI+CARN] ⇒Tt[Tt'[C]]=C ;
I∈[1:CARI]⇒Tt[I] = I ;
N∈[CARI+1:CARI+CARN]⇒
{Tt[At[N]],Tt[Bt[N]]}={At'[Tt[N]],Bt'[Tt[N]]};
SOMt'[TtN]=SOMt[N];Dt'[Tt[N]]=Dt[N];
si SOMt[N]=0 alors PERt'[Tt[N]]= Tt PERt[N]].

*Commentaire :* Les fonctions T0, T1 réalisent une correspondance biunivoque entre les classes, en respectant le numérotage de EI et donnant aux noeuds homologues des paires de descendants homologues. Il importe de noter que dans cette correspondance l'aîné peut être l'homologue du benjamin, le benjamin étant alors l'homologue de l'aîné. C'est pourquoi, dans l'énoncé des conditions, on a fait usage d'accolades pour identifier des paires d'éléments sans spécifier leur ordre. Dans le cas de la figure cet échange se rencontre trois fois, notamment pour l'exemple {13,2} = {2,13}.

1.3 *Description d'arbre tronquée au-dessous du noeud N* : Soit

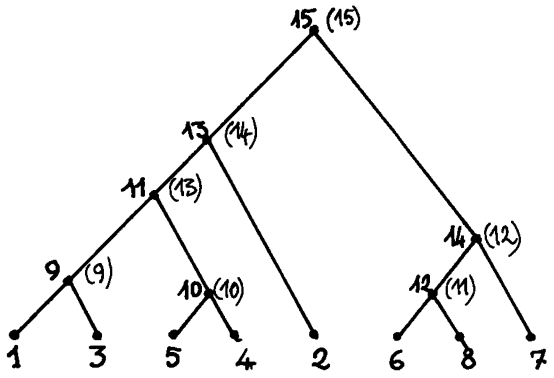
Desc={CARI,CARN,A,B,PER,SOM,D} une description d'arbre et soit N le numéro d'un noeud:N ∈ [CARI+1;CARI+CARN]. On définit la description d'arbre tronquée au-dessous du noeud N :

Desc'={CARI,CARN',A',B',PER',SOM',D'}
en posant :
CARN' = N (CARI + 1),
(i.e., CARI + CARN'=N-1 : le dernier noeud aura pour n° N-1);
A',B', restrictions de A, B à[CARI+1:N-1];
D' restriction de D à [1:N-1];
PER'[C]:=si PER[C]$<_s$N alors PER[C] sinon 2CARI ,

(i.e. une classe n'a de père que si celui-ci est un noeud de numéro inférieur ou égal à N-1);

SOM'[C]:=si PER'[C]=2CARI alors 1 sinon 0

(on redéfinit en conséquence le prédicat SOM' : "être un sommet").




CARI = 8
CARN = 7
CARS = 1

| | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|
| A0 | 1 | 5 | 9 | 6 | 11 | 12 | 13 |
| B0 | 3 | 4 | 10 | 8 | 2 | 7 | 14 |
| A1 | 1 | 4 | 6 | 11 | 9 | 2 | 12 |
| B1 | 3 | 5 | 8 | 7 | 10 | 13 | 14 |
| T0 | 9 | 10 | 13 | 11 | 14 | 12 | 15 |
| T1 | 9 | 10 | 12 | 14 | 11 | 13 | 15 |

[ALG & ALG CAH] §1.2 : Equivalence de deux descriptions DESC0 et DESC1

Exemple :

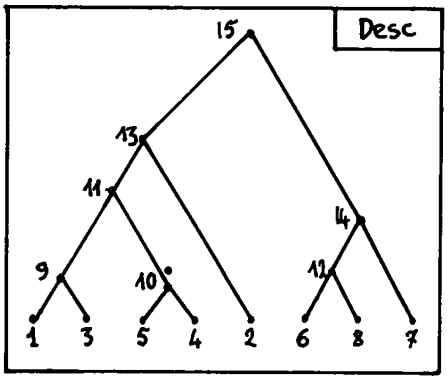
$\{T0[A0[13]], T0[B0[13]]\} =$

$\{T0[11], T0[2]\} = \{13,2\} =$

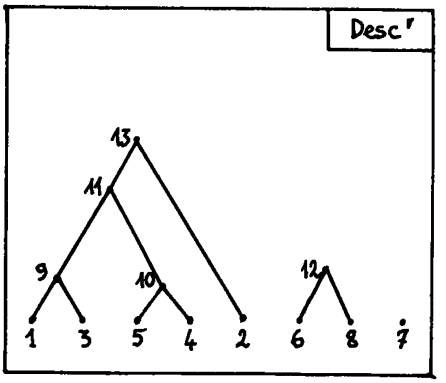
$\{A1[T0[13]], B1[T0[13]]\} =$

$\{A1[14], B1[14]\} = \{2,13\}$

*Note : le numérotage des éléments de EI (de 1 à CARI) étant respecté par les transformations T0 et T1, on a seulement tabulé celles-ci pour N variant de CARI+1 à CARN.*




Desc : description d'arbre ; Desc prolonge Desc'.




Desc' : description Desc tronquée au-dessous du noeud 14 .

[ALG & ALG CAH] § 1.3

Réciproquement, on dira que la description Desc *prolonge* la description Desc' si Desc' peut être obtenue en tronquant Desc. Tout algorithme de CAH (cf. § 5) procède en construisant une suite de descriptions dont chacune *prolonge* la précédente. Sur la figure, l'arbre décrit par Desc est représenté tronqué au-dessous du noeud 14 (avec pour sommets : {13,12,7}). La troncature dépend évidemment du numérotage choisi pour les noeuds : elle concerne donc la description particulière choisie. Cependant, par abus de langage, on dit en bref "arbre tronqué" au lieu de "description d'arbre tronquée".

## 2 *Critères d'agrégation*

*2.0* *Représentation spatiale de l'ensemble EI* : Pour construire une classification sur l'ensemble EI il faut, en bref, tenir compte d'une notion de proximité entre les éléments de cet ensemble. Les données utilisées associent par exemple à chaque élément une ligne d'un tableau de correspondance. Le profil de cette ligne est considéré comme un point dans un espace multidimensionnel. Sans entrer dans les détails, nous supposerons toujours qu'à chaque élément de EI est associé un point dans un espace E. Ainsi EI sera identifié à une partie de E, à la réserve près que plusieurs éléments peuvent être décrits par un même point.

*2.1* *Définition d'un critère* : On dit classiquement qu'un ensemble E est muni d'une structure d'espace métrique si est défini sur E×E une fonction réelle d(e,e') appelée distance, positive ou nulle quels que soient e et e', nulle seulement si e = e', et satisfaisant à l'inégalité du triangle :

$$\forall\ e, e', e'' : d(e,e'') \leq d(e,e') + d(e',e'').$$

[ALG & ALG CAH] §2.1

En classification automatique, on doit d'après un critère décider de l'agrégation de deux classes C et C' dont chacune représente une partie finie non vide éventuellement réduite à un élément de l'ensemble EI. Il semble donc qu'un critère doive être une distance sur l'ensemble des parties de EI. En fait (mise à part l'agrégation initiale des éléments de EI représentés par un même point (cf. *supra* § 2.0 et § 1.1 Rem. 4) on n'a jamais à agréger que des parties de EI d'intersection vide, ce qui rend inutile la condition : d(e,e') = 0 implique e = e' ; d'autre part dans la comparaison des distances entre plus de deux éléments, on n'a pas recours à l'inégalité du triangle, mais à un autre axiome que, sous le nom d'axiome de la médiane, nous étudierons ci-dessous (§ 3). Nous poserons donc seulement la définition suivante :

*Définition :* Un critère D sur l'ensemble fini EI est une fonction réelle positive ou nulle, définie et symétrique sur les paires de parties non vides de EI (i.e. si $p \subset EI$ et $p' \subset EI$ ; $p \neq \emptyset$ ; $p' \neq \emptyset$, on a : $D(p,p') = D(p',p) \geq 0$).

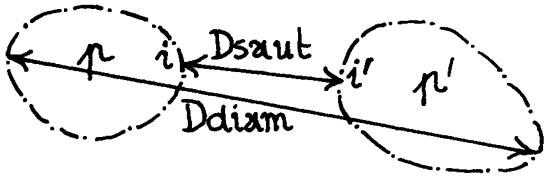
*2.2* *Critères usuels* : Le calcul d'un critère (sorte de distance entre parties de EI) se fonde presque toujours sur une distance usuelle entre éléments de l'ensemble EI, mais utilise aussi dans des cas importants un système de masses positives attribuées aux éléments de EI.

Nous rappelons ici quatre critères classiques satisfaisant à l'axiome de la médiane, (comme on le vérifiera au § 3.2).

2.2.1 *Critère du saut minimum* : EI étant muni d'une distance usuelle notée D, on étend celle-ci en un critère défini pour deux parties non vides p, p' par la formule :

$$D(p,p') = \mathrm{Inf}\{D(i,i') \mid i \in p \ ; \ i' \in p'\} \ ;$$

autrement dit : D(p,p') est défini comme la distance entre les deux points i et i' qui sont le plus proches possible.




[ALG & ALG CAH] §2.2

2.2.2 *Critère du diamètre* : Comme pour le critère du saut, on part d'une distance usuelle D, et l'on pose :

$$D(p,p') = \mathrm{Sup}\{D(i,i') \mid i \in p \ ; \ i' \in p'\} \ ;$$

i.e. : D(p,p') est la distance entre les deux points i et i' qui sont le plus éloignés possible. Le terme de diamètre s'explique parce que le diamètre d'une partie p d'un espace métrique est classiquement le maximum de la distance entre deux points de p.

2.2.3 *Critère de la distance moyenne* : Outre la distance D entre éléments de EI, on utilise un système de masses positives M(i) dont sont munis les éléments de EI. On appelle masse M(p) d'une partie p de EI, la somme des masses des éléments de p ; et on pose :

$$D(p,p') = (M(p)M(p'))^{-1} \ \Sigma\{M(i)M(i')D(i,i') \mid i \in p \ ; \ i' \in p'\} \ ;$$

c'est bien la définition classique d'une moyenne.

2.2.4 *Critère de l'inertie* : On suppose ici que l'ensemble EI, est un ensemble fini de points d'un espace euclidien, muni de sa distance usuelle ; et que de plus chaque élément i a une masse M(i). A toute partie p de EI, on associe son centre de gravité, noté également p:

$$p = M(p)^{-1}\Sigma\{M(i)\ i \mid i \in p\} \ ;$$

la distance euclidienne entre les centres de deux parties p et p' est simplement notée $\|p - p'\|$. Ceci dit on pose :

$$D(p,p') = (M(p)\ M(p')/(M(p) + M(p'))) \ \|p - p'\|^2.$$

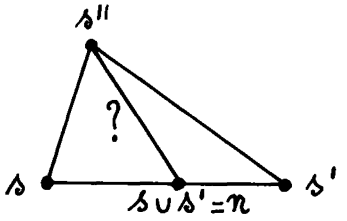
C'est l'inertie d'un nuage réduit aux deux points p et p' munis de leurs masses.

2.3 *Calcul du critère : formule de la médiane* : Lorsqu'une classification hiérarchique est édifiée par voie ascendante, l'arbre s'accroît par création de noeuds N dont chacun a pour descendants A[N] et B[N], des classes qui, dans l'étape antérieure, étaient des sommets. Puisque seuls les sommets sont susceptibles d'être engagés, il suffit de calculer la valeur du critère D pour deux sommets s et s'. Au départ

CARN = 0, l'ensemble des sommets n'est autre que l'ensemble des éléments de EI. Au cours du déroulement de l'algorithme, des sommets disparaissent, d'autres se créent ; un problème de calcul apparaît:

connaissant D(s,s'), D(s,s"), D(s',s"), calculer D(s ∪ s', s").

Il est commode d'énoncer ce problème dans les termes familiers de la géométrie élémentaire en généralisant le sens de ces termes. Nous dirons donc que les trois sommets s,s',s" ou plus généralement trois parties de EI deux à deux d'intersection vide constituent un "triangle", que les trois valeurs D(s,s') etc. du critère pour ces sommets pris deux à deux sont les "côtés" du triangle ; la classe s ∪ s' sera appelée le "milieu" du côté (s,s') (terme justifié si on identifie chaque classe à son centre de gravité et que s et s' ont même masse) ; enfin D(s ∪ s',s") sera la "médiane" relative au sommet s" dans le triangle {s,s',s"}. Ceci posé, le problème devient :




[ALG & ALG CAH] §2.3

calculer la médiane en fonction des côtés : en fait, le calcul n'est possible que pour les deux critères du saut et du diamètre ; pour les deux autres critères cités, il faut de plus connaître les masses dont sont affectés les trois sommets. Voici par exemple trois "formules de la médiane" :

$$Dsaut(s \cup s') = \inf\{D(s,s"), D(s',s")\} ;$$

$$Ddiam(s \cup s',s") = \sup\{D(s,s"), D(s',s")\} ;$$

$$D\,moy(s \cup s',s") = (M(s \cup s'))^{-1}(M(s)D(s,s") + M(s')D(s',s")).$$

La formule pour Dinert est plus compliquée (nous la rencontrerons au § 3.2) ; mais peu importe, voici pourquoi.

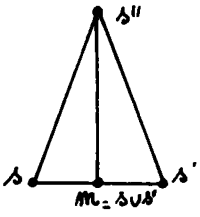
Un algorithme de CAH doit, en bref, véhiculer et mettre à jour des informations relatives aux sommets, afin de décider des agrégations à effectuer. Utiliser une "formule de la médiane" implique de mettre à jour (outre la tableau des masses des sommets) un tableau des distances entre sommets ; tableau dont, au départ la dimension est CARI (CARI-1)/2. Cependant, avec le critère Dinert, une autre voie s'offre : véhiculer avec les masses des classes des sommets, les coordonnées des centres de gravité de celles-ci dans l'espace euclidien ambiant ; et, quand il est nécessaire, calculer le critère D(s,s') en appliquant directement la formule de définition du § 2.2.4.

Or, dans la pratique, la dimension de l'espace ambiant dépasse rarement 50 et peut être réduite à 10 par analyse factorielle ; au contraire, on traite aujourd'hui des ensembles EI de 1000 éléments ou plus. Il est clair que le critère de l'inertie permet de véhiculer des informations beaucoup moins volumineuses que celles requises pour appliquer la formule de la médiane. Cet avantage s'ajoute à d'autres pour faire préférer le critère de l'inertie.

## 3 *Axiome de la médiane et voisins réciproques*

*3.1 Enoncé de l'axiome* : Nous reprenons l'emploi analogique des termes géométriques usuels, et nous énonçons :

Dans un triangle, la médiane opposée au plus petit côté est supérieure ou égale au plus petit des deux autres côtés.




[ALG & ALG CAH] §3.1

Il importe de noter que cet axiome n'est pas vrai en géométrie euclidienne usuelle : mais il l'est pour chacun des quatre critères proposés au § 2.2. Pour plus de précision reformulons l'axiome sans recourir aux termes de la géométrie élémentaire.

Soit s,s',s" trois parties non vides de EI, deux à deux d'intersection vide ;

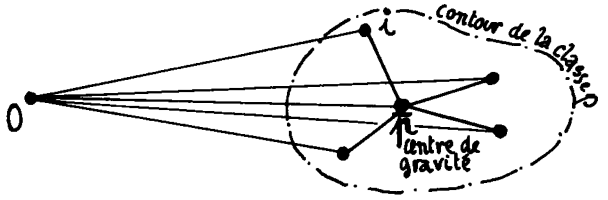
si: $D(s,s') \leq \inf(D(s,s"),D(s',s"))$ ; alors :

$$\inf(D(s,s"),D(s',s")) \leq D(s \cup s',s").$$

(Sour le nom d'axiome de réductibilité, cette formule a été introduite par Bruynooghe : cf. [*CLAS. RAP.*] *in* C.A.D. Vol III (1978) p. 10).

*3.2 Vérification de l'axiome* : La vérification étant très facile pour les critères du saut minimum du diamètre et de la distance moyenne, nous nous bornerons au critère de l'inertie. Pour plus de commodité on notera m = s∪s' (le "milieu" du "côté" s,s') ; il suffira de calculer la médiane D(m,s") en fonction des trois côtés et des masses; i. e. d'appliquer la formule de la médiane (cf. § 2.3) qui s'obtient

[ALG & ALG CAH] § 3.2 : Théorème de HUYGHENS.




$$In(\pi;0) = \sum\{M(i)\,\|i-0\|^2 \mid i \in p\}$$
$$= M(p)\,\|\bar{\pi}-0\|^2 + In(\pi;\bar{\pi}) \;;$$

ici par le théorème de Huyggens. On a par définition de l'inertie de (s,s') par rapport à m :

$$M(s)\,\|s-m\|^2 + M(s')\,\|s'-m\|^2 = D(s,s'),$$

le théorème de Huyggens donne pour l'inertie de la paire (s,s') relativement au point s" :

$$M(s)\|s-s"\|^2 + M(s')\|s'-s"\|^2 = M(m)\|m-s"\|^2 + D(s,s')$$

en remplaçant les distances au carré par des inerties divisées par une masse (suivant la formule $\|a-b\|^2 = D(a,b)(M(a)+M(b))/(M(a)M(b))$), il vient (en notant, en bref $M_a$ au lieu de M(a)) :

$$((Ms+Ms")/Ms")D(s,s") + ((Ms'+Ms")/Ms")D(s',s") =$$

$$((Mm+Ms")/Ms")D(m,s") + D(s,s') \quad ; \quad \text{d'où :}$$

$$D(m,s") = ((Ms+Ms")D(s,s")+(Ms'+Ms")D(s',s")-Ms"D(s,s'))/(Mm+Ms")$$

on a une borne inférieure de D(m,s") en substituant aux deux nombres D(s,s"), D(s',s") le plus petit de ceux-ci (noté Inf) et en remplaçant de même D(s,s') (qui a le signe -) par Inf (qui est supérieur ou égal à D(s,s'). Il en résulte que $D(m,s") \geq Inf$ ; c.q.f.d..(On notera que l'égalité est réalisée si Ms = Ms' = Ms" et que le triangle (s,s',s") est équilatéral).

3.3 *Définition des voisins réciproques* : Soit ES un ensemble de parties non vides de EI, deux à deux d'intersection vide. Soit $s \in ES$ ; on dit qu'un élément s' de ES est un plus proche voisin de s (au sens du critère D) au sein de ES si on a :

$$D(s,s') = Inf\{D(s,s") \mid s" \in ES \ ; \ s" \neq s\},$$

i.e. si s' réalise le minimum de la distance à s parmi les éléments de ES. Il importe de noter que s peut avoir plusieurs plus proches voisins si le minimum est réalisé plusieurs fois.

On dit que deux éléments (s,s') sont des voisins réciproques au sein de ES si chacun d'eux est un plus proche voisin de l'autre au sein de ES.

3.4 *Voisins réciproques et axiome de la médiane* : La proposition suivante permet de relier les diverses notions de compatibilité entre un arbre et un critère, introduites au § 4.1, et donc de démontrer l'équivalence entre les algorithmes accélérés et l'algorithme de base.

*Proposition.* Soit (s1,s1') et (s2,s2') deux paires de voisins réciproques au sein de ES, les quatre éléments s1, s1', s2, s2' étant distincts entre eux. Après agrégation de s1 et s1', (s2,s2') reste une paire de voisins réciproques au sein du nouvel ensemble de sommets ES' :

$$ES' = ES \cup \{s1 \cup s1'\} - \{s1,s1'\}.$$

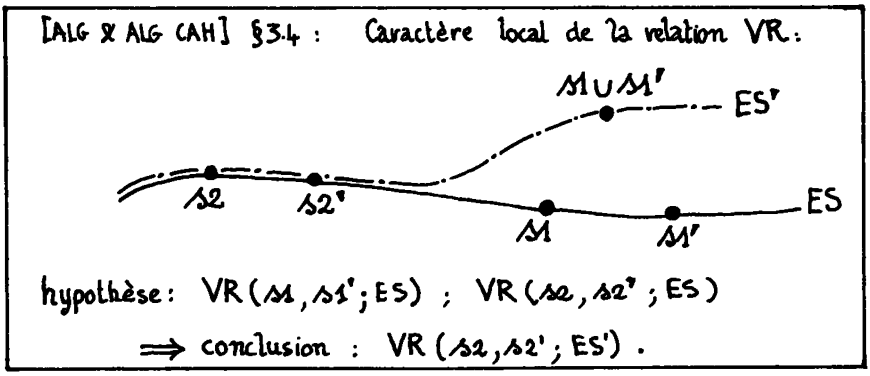
*Preuve.* Il suffit de vérifier que $D(s2,s1 \cup s1') \geq D(s2,s2')$ et de même pour s2'. Or s1 et s1' étant voisins réciproques D(s1,s1') est le plus petit côté du triangle (s1,s1',s2) ; donc d'après l'axiome de la médiane :

$$D(s2,s1 \cup s1') \geq Inf(D(s2,s1),D(s2,s1'))\ ;$$

d'autre part s2 et s2' étant voisins réciproques on a :

$$Inf(D(s2,s1), D(s2,s1')) \geq D(s2,s2').$$

L'inégalité $D(s2,s1 \cup s1')$ D(s2,s2') en résulte. On a de même $D(s2',s1 \cup s1') \geq D(s2,s2')$, ce qui achève la preuve.




## 4 *Arbre indicé compatible avec un critère*

### 4.1 *Notion de compatibilité entre description d'arbre et critère* :

Dans ce § nous reprenons les descriptions d'arbre sous la forme introduite au § 1. Nous écrivons la valeur du critère D(C,C') (en bref, distance entre les deux classes C et C') sans préciser par quelle procédure se calcule cette fonction D (à distinguer du tableau D[C] des niveaux), cf. § 2.3. On définit trois notions de compatibilité entre description d'arbre et critère :

*Définition 1* : Une description d'arbre indicé (CARI,CARN,A,B,PER, SOM,D) est dite *compatible* avec le critère D si :

$N \in [CARI+1 : CARI+CARN] \Rightarrow$

D[N] = D(A(N),B(N)) ; et A[N] et B[N] sont voisins réciproques au sein de l'ensemble des sommets de l'arbre tronqué au-dessous du noeud N (cf. § 1.3).

*Définition 2* : Une description d'arbre indicé est dite *compatible monotone* si elle est compatible et si de plus la suite des niveaux des noeuds est non décroissante (i.e. si :
$CARI <_s N <_s N' \leq CARI+CARN \Rightarrow D[N] \leq D[N']$).

*Définition 3* Une description d'arbre indicé est dite *compatible par segments* s'il existe un entier CQ et une suite strictement croissante d'entier NF[0:CQ] allant de CARI à CARI+CARN (i.e. :
$CARI = NF[0] <_s NF[1] <_s NF[2] \ldots <_s NF[CQ] = CARI + CARN$), avec la condition de compatibilté suivante :

$\forall\, Q \in [1:CQ]\ ;\ \forall\, N \in [NF[Q-1] + 1 : NF[Q]]$ :

D[N] = D(A[N],B[N]) et A[N] et B[N] appartiennent à l'ensemble des sommets de l'arbre tronqué au-dessous du noeud n° (NF[CQ-1] + 1) et sont voisins réciproques au sein de cet ensemble.

*Nota Bene*: La notion de compatibilité par segment est introduite pour décrire la construction ascendante d'un arbre qui passe par une suite d'états numérotés de 0 à CQ. A l'état 0 l'arbre est réduit à l'ensemble EI. A l'état Q, on a l'arbre tronqué au-dessous du noeud NF[Q]+1. Pour passer de l'état (Q-1) à l'état Q, on crée parallèlement tous les noeuds N numérotés de (NF[Q-1]+1) à NF[Q]. Cette création parallèle n'est possible que parce que les deux descendants immédiats A[N] et B[N] existent déjà comme sommets dans l'état (Q-1) (i.e. A[N], B[N] $\leq$ NF[Q-1]). Plus précisément, A[N] et B[N] ne peuvent être agrégés que s'ils sont voisins réciproques au sein de l'ensemble ES[Q-1] des sommets de l'arbre dans son état (Q-1) (i.e. tronqué au-dessous du noeud (NF[Q-1]+1)).

4.2 *Equivalence des notions de compatibilité* : Sous l'hypothèse *admise désormais une fois pour toutes* que le critère D satisfait à l'axiome de la médiane, on démontre les propositions suivantes :

*Proposition 1*:Toute description compatible est équivalente (au sens du § 1.2) à une description compatible monotone.

*Proposition 2*:Toute description compatible par segments est une description compatible (donc, cf. Prop. 1, est équivalente à une description compatible monotone).

*Corollaire* : Dans une description compatible on a, pour tout noeud N les inégalités :

$$D[A[N]] \leq D[N] \; ; \; D[B[N]] \leq D[N],$$

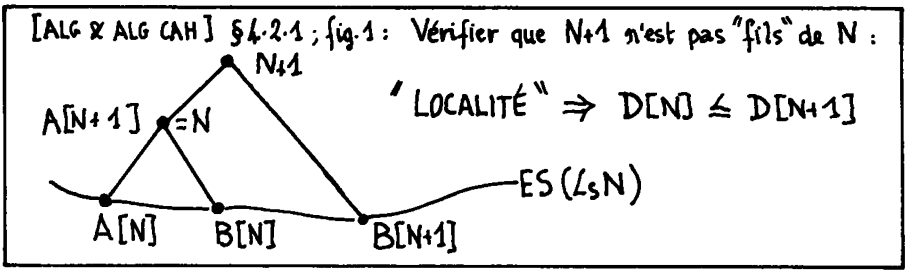
ce qu'on énonce encore en disant que l'arbre ne présente pas d'*inversion* d'indice.

Le corollaire résulte immédiatement des propositions et de ce qu'un arbre compatible monotone ne présente pas d'inversion. Reste à démontrer les propositions.

4.2.1 *Preuve de la proposition 1* : Partons d'une description compatible qui n'est pas monotone ; celle-ci comportera des inversions que nous dénombrerons comme suit :

$$Inv=Card\{(N,N') \mid CARI <_s N <_s N' \leq CARI+CARN \; ; \; D[N'] <_s D[N]\} \; ;$$

en particulier, il existe des inversions relatives à des paires de noeuds (N,N+1) dont les numéros se suivent. Nous montrerons ci-dessous qu'on a une description équivalente à celle donnée en permutant les noeuds N et N+1 (en un sens qui sera précisé) : la description ainsi obtenue comportant une inversion de moins que la description initiale, on aura démontré que toute description est équivalente à une description comportant moins d'inversions qu'elle-même ; et donc, par récurrence, à une description sans inversion, i.e. monotone.




Il importe de vérifier que si $D[N+1] <_s D[N]$, les paires de descendants (A[N]) et B[N]) et (A[N+1], B[N+1]) forment quatre classes distinctes ; autrement dit, que l'on n'a pas N = A[N+1] ou B[N+1]. En effet, supposons e.g. que A[N+1] = N, et tronquons l'arbre au-dessous du noeud N. Alors A[N], B[N] et B[N+1] sont des sommets de l'arbre tronqué ; et puisque la description donnée est compatible avec le critère, A[N] et B[N] sont des voisins réciproques au sein de l'ensemble {A[N], B[N], B[N+1]} d'où il résulte d'après l'axiome de la médiane, que la médiane D[N+1] = D(N,B(N+1)) est supérieure ou égale au plus petit côté D(A[N], B[N]) = D[N], ce qui est contraire à l'hypothèse faite qu'il y a inversion ($D[N+1] <_s D[N]$).

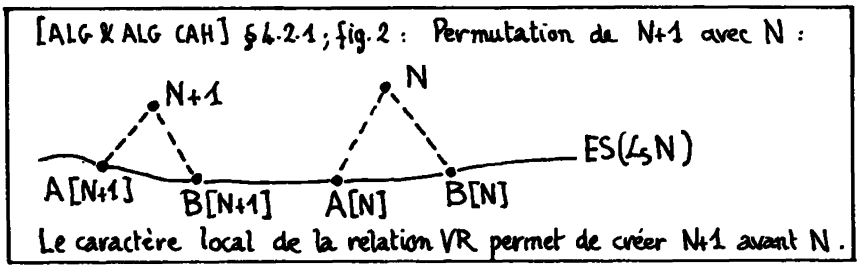
Puisque N n'est pas descendant de N+1, on peut permuter ces noeuds dans la description de l'arbre, i.e. substituer à la

description donnée une autre description qui ne diffère de la première qu'en ce que :

A'[N] = A[N+1]; B'[N] = B[N+1] ; D'[N] = D[N+1];

A'[N+1] = A[N] ; B'[N+1]= B[N] ; D'[N+1] = D[N] .

Reste à montrer que cette description nouvelle est aussi une description compatible :




Coupons l'arbre au-dessous du noeud N et notons ES l'ensemble des sommets de l'arbre ainsi tronqué ;(A[N], B[N], A[N+1] , B[N+1] ) sont des éléments de ES ; et par hypothèse, A[N] et B[N] sont voisins réciproques au sein de ES ; donc :

$D(A[N+1], A[N]) \geq D(A[N], B[N]) \geq_S D(A[N+1], B[N+1]) = D[N+1]$ ; de même , D(A[N+1], B[N]), D(B[N+1], A[N]), D(B[N+1], B[N]) sont supérieurs à D[N+1] ; d'où il résulte que A[N+1] et B[N+1] qui sont voisins réciproques après agrégation de A[N] et de B[N] (i.e. au sein ES+N-{A[N], B[N]}, le sont aussi au sein de ES. En appliquant la proposition du § 3.4, on voit alors que la description obtenue en permutant N et N+1 reste compatible ; i.e. qu'on peut agréger d'abord A[N+1] et B[N+1], la paire (A[N], B[N]) étant alors une paire de v.r. au sein de ES + {N+1} - {A[N+1], B[N+1]} . Ceci achève la preuve de la proposition 1.

*4.2.2 Preuve de la proposition 2* : On part d'une description compatible par segments ; on veut démontrer que celle-ci est une description compatible. Nous procéderons par récurrence sur la longueur maxima des segments, en appelant longueur du segment Q la différence NF[Q]-NF[Q-1]. Il est d'abord clair que si tous les segments ont longueur 1, cela signifie que CQ = CARN et NF[Q]= CARI+Q ; auquel cas la notion de *"compatible par segments"* coïncide avec celle de *"compatible"*. Supposons maintenant qu'un segment Q ou intervalle [NF[Q-1]+1:NF[Q]] ait une longueur L = NF[Q]-NF[Q-1] supérieure strictement à 1. On peut couper ce segment en deux : d'une part un segment de longueur 1, qui ne comprend que le premier noeud de numéro NF[Q-1]+1, et d'autre part le reste du segment.

Il résulte alors de la proposition du § 3.4 que la description est également compatible avec la nouvelle segmentation : car en bref si pour $N \in [NF[Q-1]+2:NF[Q]]$, A[N] et B[N] sont voisins réciproques au sein de l'ensemble des sommets de l'arbre coupé au-dessous du noeud NF[Q-1], ils le sont aussi après agrégation de A[NF[Q-1]+1] et B[NF]Q-1 +1 , au sein de l'arbre coupé au-dessous du noeud NF Q-1 +1.

Ceci achève la preuve de la proposition 2.

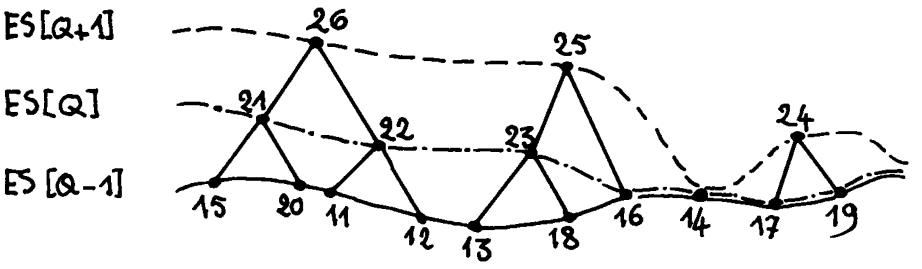
[ALG & ALG CAH] § 4 : Description compatible par segments "DCS" ;

Soit : CQ = Nombre des Segments :
$Q \in [1 : CQ]$ ; $N \in [NF[Q-1]+1 : NF[Q]]$ = "Segment Q" .

$\Rightarrow A[N], B[N] \in ES[Q-1]$ ;
$VR[A[N], B[N] ; ES[Q-1]]$ .

(où $ES[Q-1] = ES(\mathcal{L}_S NF[Q-1]+1)$) .

EXEMPLE




NF[Q-1] = 20 ; ES[Q-1] = {11, 12, 13, 14, 15, 16, 17, 18, 19, 20} ;

NF[Q] = 23 ; ES[Q] = {14, 16, 17, 19, 21, 22, 23} ;

NF[Q+1] = 26 ; ES[Q+1] = {14, 24, 25, 26} .

PROPOSITION : "DC" $\Leftrightarrow$ "DCS" .

Preuve : le caractère local de la relation VR permet de créer les noeuds un par un.

Absence d'inversion : Un arbre compatible ne présente pas d'inversion :

$$D[A[N]] \leq D[N] \; ; \quad D[B[N]] \leq D[N]$$

Preuve : considérer une description compatible monotone.

## 5 *Algorithme de construction d'une description compatible*

En bref, tout algorithme part d'une description d'arbre triviale, i.e. sans noeud : CARN=0, et construit une suite de descriptions emboîtées dont chacune prolonge la précédente (cf. § 4.1 NB) et plus précisément diffère de celle-ci par la création de noeuds dont les descendants (aîné et benjamin) en étaient des sommets : on s'arrête quand il n'y a plus qu'un seul sommet (CARN=CARI-1).

5.1 *Algorithme de base* : On crée les noeuds un par un en agrégeant chaque fois une paire de sommets réalisant le minimum de la distance D(s,s').

*Proposition 1* : L'algorithme de base fournit une description compatible monotone.

*Preuve* : En bref, si N n'est pas un descendant de N+1, alors (A[N],B[N]) et (A[N+1],B[N+1]) appartiennent à l'ensemble des sommets de l'arbre tronqué au dessous du noeud N ; et de par le choix même de l'algorithme D[N]=D(A[N],B[N]) est inférieur ou égal à D[N+1]= D(A[N+1],B[N+1]). Tandis que si, e.g. N=A[N+1], la démonstration du § 4.2.1 prouve que D[N] ≤ D[N+1].

Réciproquement, on a la proposition suivante :

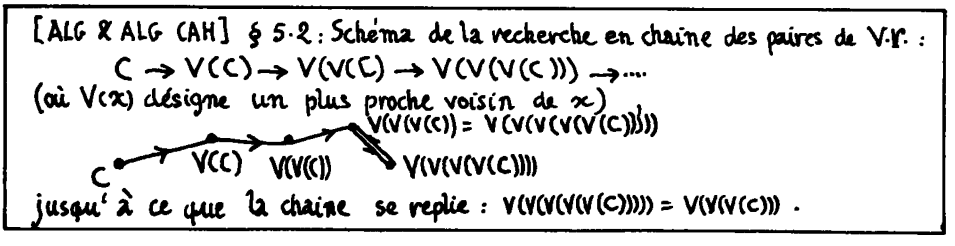
*Proposition 2* : Toute description compatible monotone n'ayant qu'un seul sommet (i.e. cf. § 1.1 Remarque, telle que CARN=CARI-1) peut être obtenue par l'algorithme de base (l'unicité de la description produite par l'algorithme de base n'étant assurée que s'il ne se rencontre jamais plus d'une paire réalisant le minimum du critère).

*Preuve* : On procède par l'absurde. Supposons que dans une description monotone le noeud N soit tel qu'il existe au sein de l'ensemble Es des sommets de l'arbre tronqué au-dessous du noeud N une paire (s,s') telle que $D(s,s') <_S D[N]$. Nécessairement s et s' devront être agrégés soit entre eux soit avec d'autres puisqu'on a supposé que la description compatible considérée n'a qu'un seul sommet ; or cette agrégation (de s avec s', ou celui de ces deux qui est agrégé le premier avec un tiers) se fera à un niveau inférieur ou égal à D(s,s') puisqu'on n'agrège que des paires de v.r. ; et par conséquent strictement au-dessous du niveau D[N] ce qui est contraire à l'hypothèse que la description est monotone.

5.2 *Algorithmes accélérés* : Ces algorithmes construisent une suite de descriptions compatibles emboîtées : on passe d'une description à la suivante en créant un ou plusieurs noeuds par agrégation de paires de sommets de voisins réciproques. Ainsi la description construite est compatible par segment avec le critère. Elle est donc compatible; et il suffit de numéroter à nouveau les noeuds pour avoir une description compatible monotone, donc (cf. § 5.1) un résultat équivalent à celui produit par l'algorithme de base (lequel est lent). Le renumérotage existe de par la proposition 1 du § 4.2 ; et il est unique sauf s'il y a des noeuds N, N' situés au même niveau : D[N] = D[N']. Pratiquement, on procède en ne permutant N et N' que si N < N' et $D[N'] <_S D[N]$; on se garde de permuter des noeuds de niveau égal, ce qui pourrait donner à N un numéro inférieur à celui de A[N] si D[N] = D[A[N]]; éventualité qui ne peut être exclue (cf. § 3.2, cas du triangle équilatéral).

La conception d'un algorithme accéléré se ramène donc à la recherche de paires de voisins réciproques au sein de l'ensemble des sommets d'un arbre. On peut (cf. Juan [*PROG. C.A.H. RECIP.*] *in* C.A.D. Vol. VII n° 2 1982) à agréger à chaque étape toutes les paires de v.r. .

On peut encore cf. [*CAH CHAINE RECIP.*] *in* C.A.D. *ibid)*, chercher seulement une paire de v.r. à l'extrémité d'une chaîne de classes dont chacune est plus proche voisin de la précédente ;



l'intérêt de la méthode étant que, s'il faut une longue chaîne pour aboutir à une paire de v.r., la chaîne subsiste après agrégation de cette paire et sert à la recherche de la paire suivante. Il faut enfin rappeler que l'algorithme de Bruynooghe (cf. [*CLASS. RAP.*] *in* C.A.D. Vol III n° 1, 1978 utilisé notamment par M. Jambu accélère la recherche d'une paire de v.r. en limitant par un seuil de distance la liste des paires susceptibles d'être agrégées. Présentement, on retiendra que [*PROG. C.A.H. RECIP.*] donne l'algorithme le plus rapide en moyenne ; et [*C.A.H. CHAINE RECIP.*] donne un algorithme très simple qui n'est jamais beaucoup moins rapide que le précédent ; et est le seul à notre connaissance qui, quelles que soient les données, achève de classer n individus en un temps de l'ordre de $n^2$.

5.3 *Améliorations ultérieures : accélération de la recherche du plus proche voisin d'un point* : Comme le note F. Murtagh (cf. Information Processing Letters 16 (1983) ; pp 237-241), l'algorithme de CAH fondé sur la recherche en chaîne des v.r. requiert que toutes les distances soient examinées au moins une fois : et c'est pourquoi le temps ne peut être inférieur à $n^2$. Cependant, il n'est manifestement pas nécessaire de calculer toutes les distances à i pour trouver le point i' le plus proche de i. Si l'ensemble EI des individus à classer (puis, dans la suite l'ensemble ES des sommets est muni d'une partition en boules de rayon borné (cf. C.A.D. Vol IX n° 1 (1984) ; pp 119-122) ou mieux encore muni d'une hiérarchie de boules (cf. K. Bensalem ; ibid pp 123-124) il peut suffire de calculer la distance de i au centre d'une boule, pour être assuré que celle-ci ne contient pas le plus proche voisin de i ; traitant deux cas où EI est un ensemble de points du plan, F. Murtagh *(loc. cit.)* utilise tout simplement la partition suivant un quadrillage pour accélérer grandement la recherche. Une autre manière, moins ambitieuse, d'accélérer la recherche du plus proche voisin, consiste à ne visiter dans la recherche en chaîne que les points qui lors de la construction d'un précédent maillon ont été marqués pour être assez proches : car, e.g. l'écart de V(V(V(C))) à son plus proche voisin étant inférieur à l'écart de C à V(C) , l'écart de C à ce plus proche voisin cherché peut également être borné supérieurement. Il faut toutefois prendre garde que la mise en oeuvre de tels perfectionnements est rendue plus ardue par le fait que l'écart n'est pas à proprement parler une distance (satisfaisant à l'inégalité du triangle : cf. § 2.1).