MÉTHODES DE CLASSIFICATION

Pierre-Louis GONZALEZ

MÉTHODES DE CLASSIFICATION

Objet: Opérer des regroupements en classes homogènes d'un ensemble d'individus.

Données: Les données se présentent en général sous la forme d'un tableau individus variables.

1. Ayant défini un **critère de distance** (dissemblance) ou **dissimilarité** (pas nécessairement d'inégalité triangulaire) entre les individus, on procède au regroupement des individus.
2. Ce regroupement nécessite une **stratégie de classification** : critère de classification.

MÉTHODES

NON HIÉRARCHIQUES

Partition en k classes

Exemples : Centres mobiles

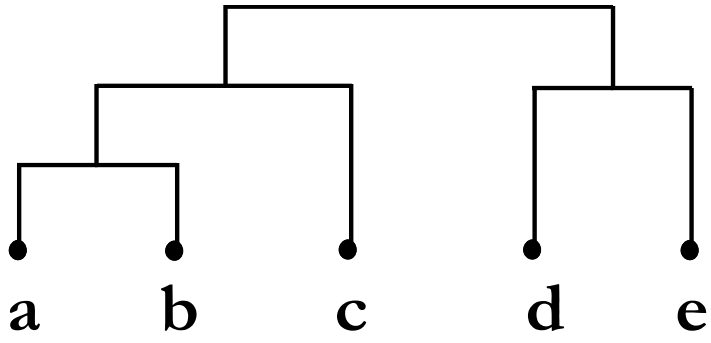
Nuées dynamiques

Avantages : Permettent la classification d'ensembles volumineux.

Inconvénients : On impose au départ le nombre de classes.

MÉTHODES

HIÉRARCHIQUES: suites de partitions emboîtées



OU

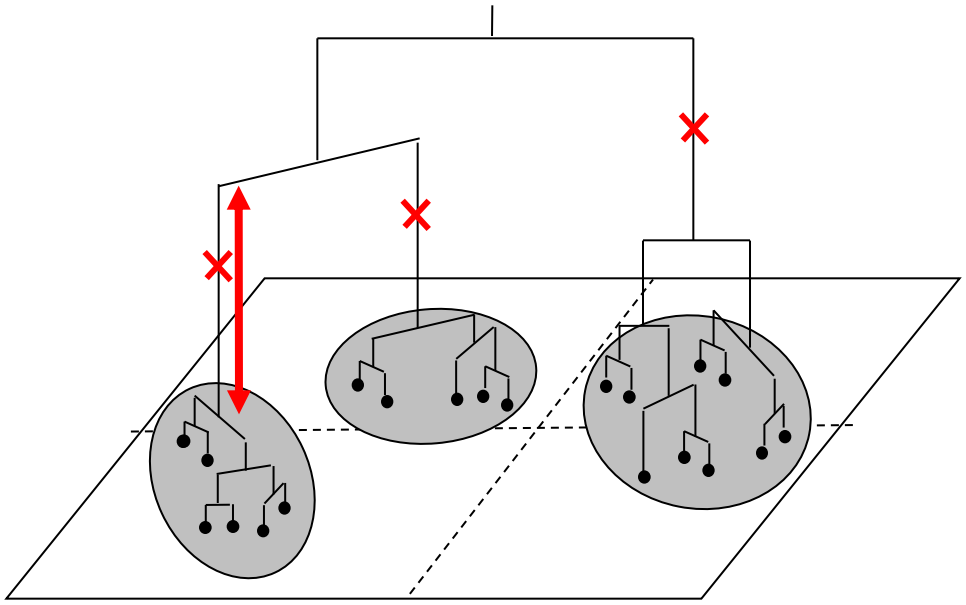
a, b, c, d, e
ab, c, d, e
abc, de
abcde

Avantages : La lecture de l'arbre permet de déterminer le nombre optimal de classes

Inconvénients : Coûteux en temps de calcul.

Méthodes de classification

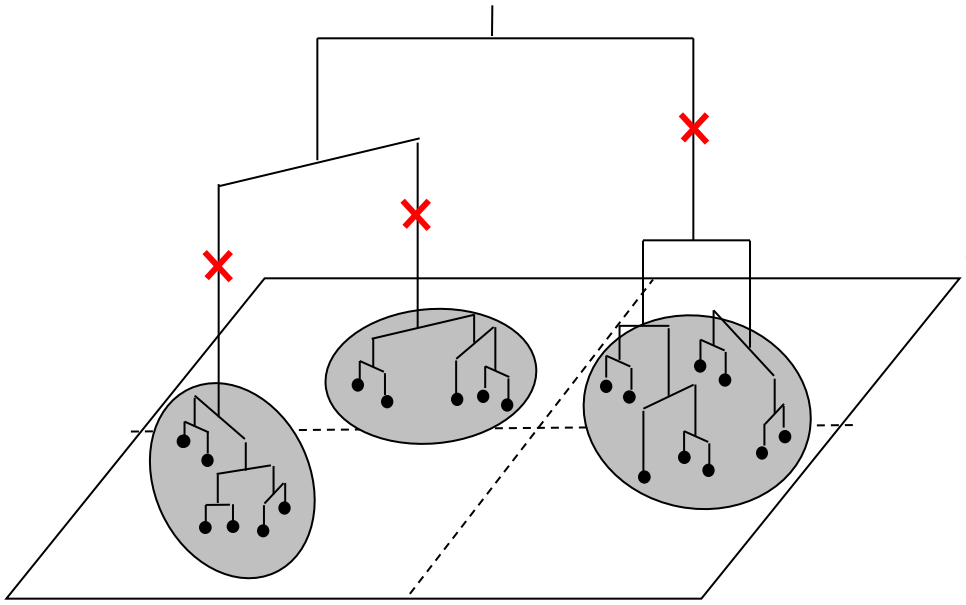
Méthode hiérarchique
de Ward



3 classes

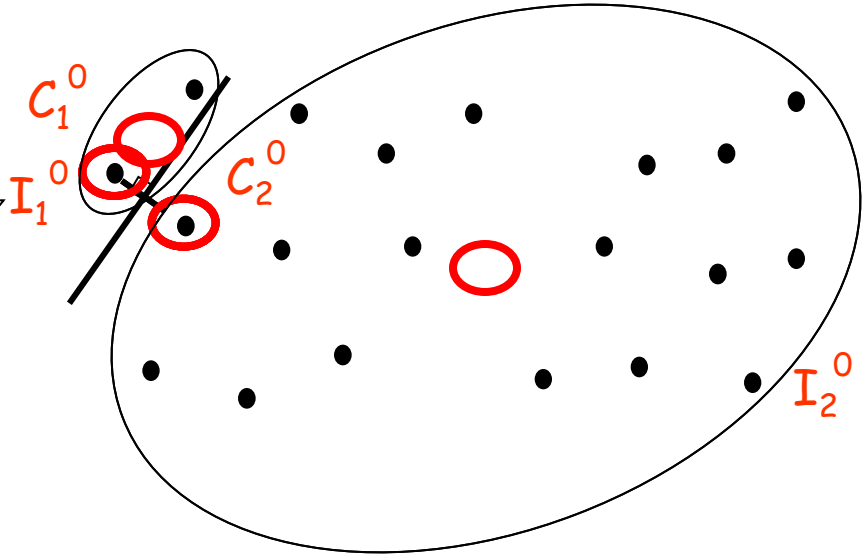
Méthodes de classification

Méthode hiérarchique de Ward



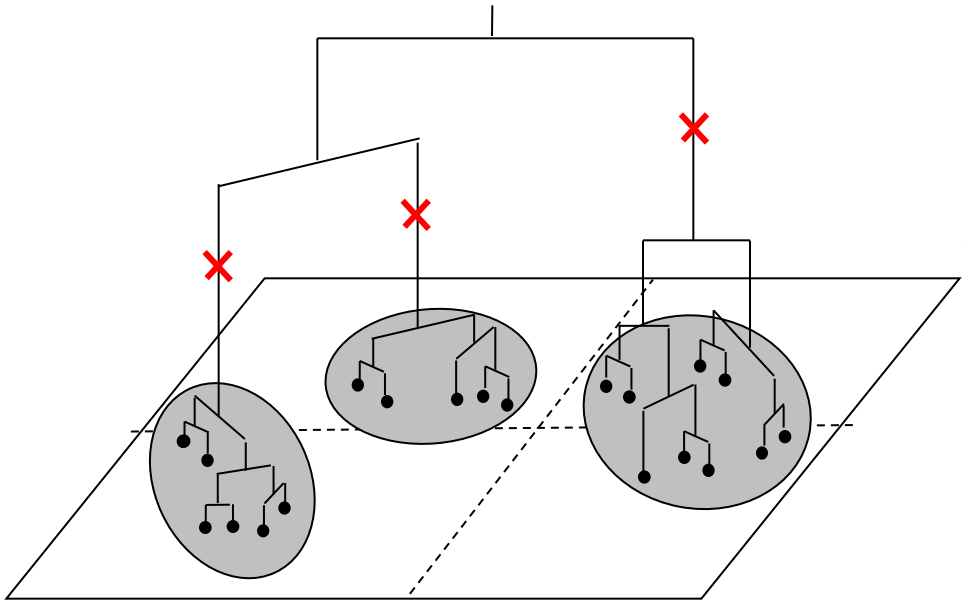
3 classes

Centres mobiles



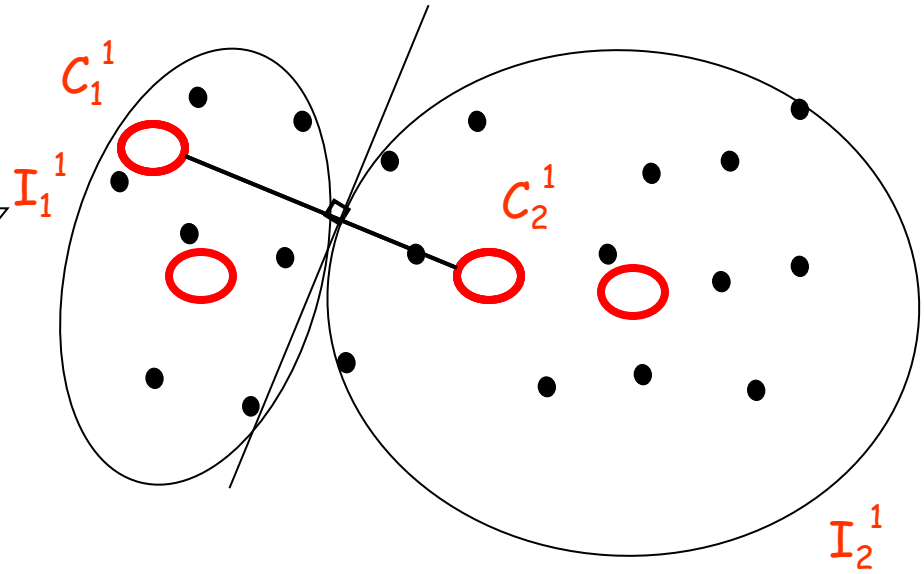
Méthodes de classification

Méthode hiérarchique de Ward



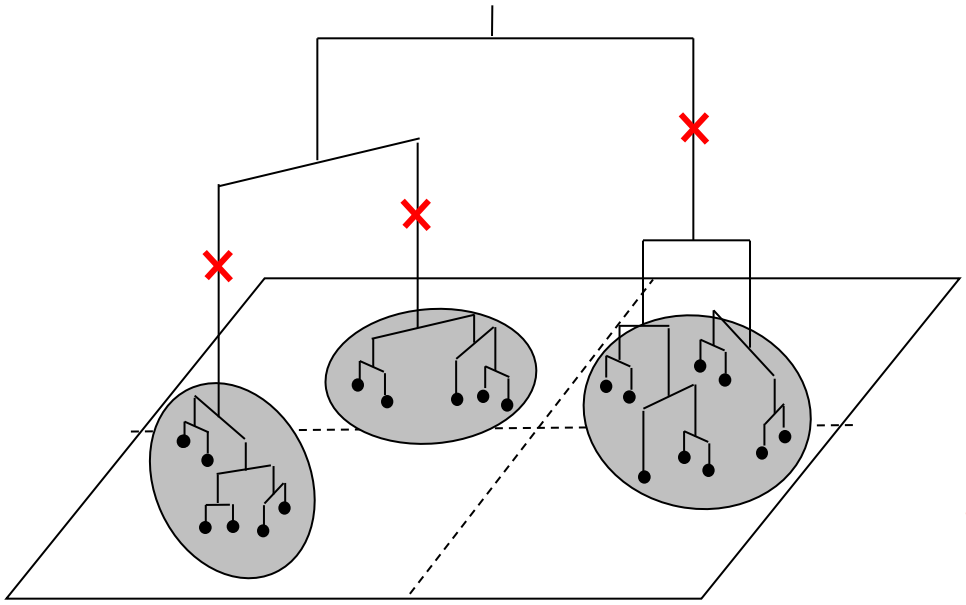
3 classes

Centres mobiles



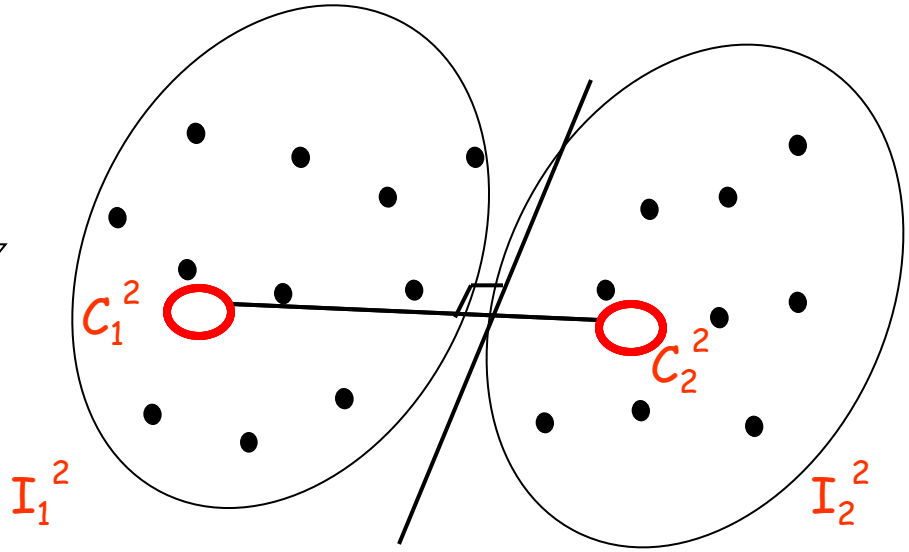
Méthodes de classification

Méthode hiérarchique de Ward



3 classes

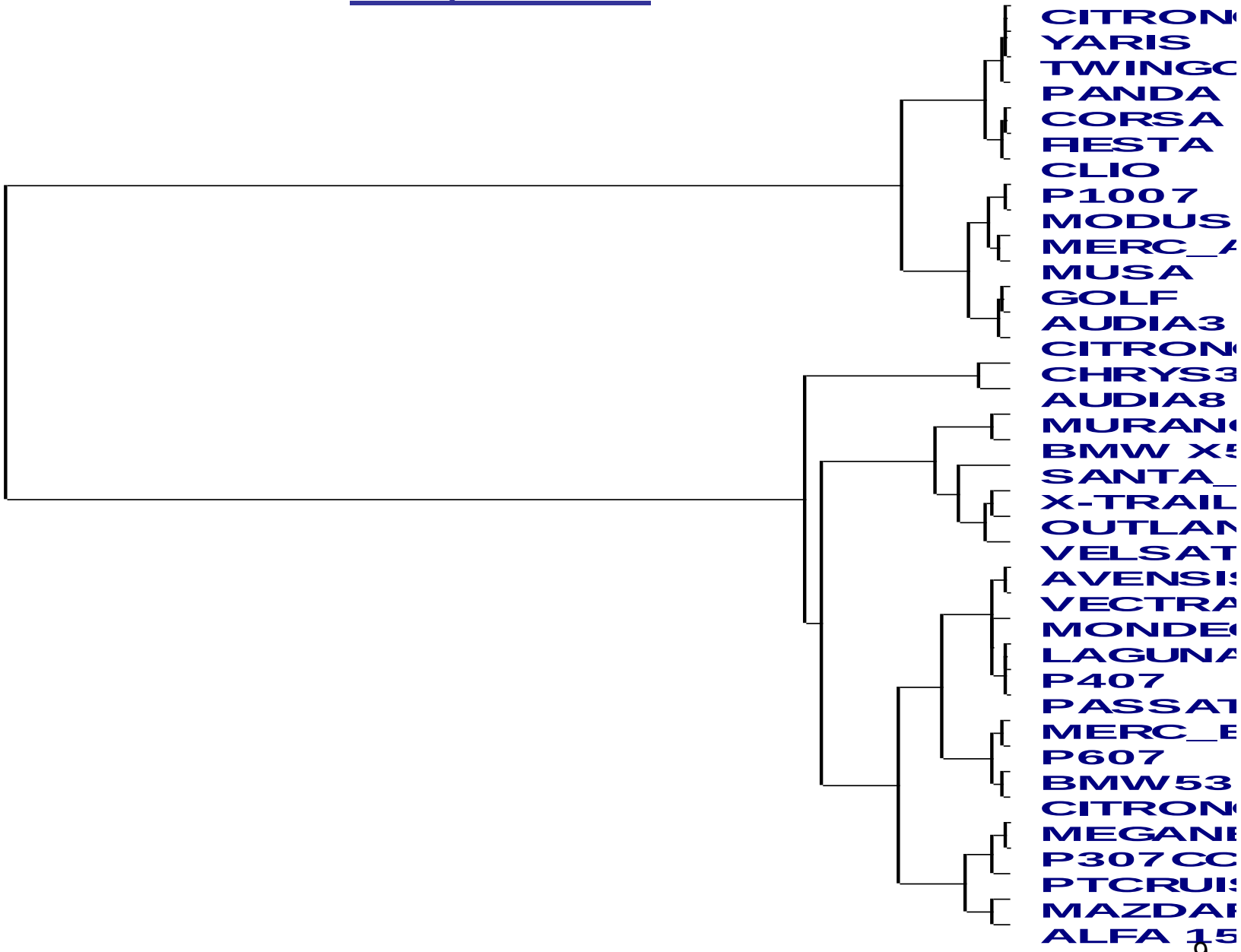
Centres mobiles



2 classes

Classification hiérarchique ascendante

Exemple voitures



Éléments de vocabulaire

- classification automatique
- classification non supervisée
- apprentissage sans professeur

Le terme **« classification »** en anglais fait référence à l'affectation d'un individu à une classe (existant a priori) dans le cadre de l'analyse discriminante. Il se traduit par le terme classement.

L'équivalent en anglais de «classification automatique» est **« cluster analysis ».**

Éléments de vocabulaire

- **Dissimilarité**

$$d(i,j) = d(j,i)$$

$$d(i,i) = 0$$

$$d(i,j) \geq 0$$

- **Similarité**

$$s(i,j) = s(j,i)$$

$$s(i,j) \geq 0$$

$$s(i,i) \geq s(i,j)$$

MÉTHODES DE PARTITIONNEMENT

1. Considérations combinatoires

$P_{n,k}$ = nombre de partitions en k classes de n individus

$$P_{n,k} = P_{n-1,k-1} + k\, P_{n-1,k}$$

(récurrence)

(nombre de Stirling de 2ème espèce)

$$P_{12,5} = 1\,379\,400$$

Considérations combinatoires

P_n = nombre total de partitions (nombres de Bell)

Ex : $P_{12} = 4\ 213\ 597$

$\Rightarrow$ **Nécessité d'algorithmes pour trouver une bonne partition.**
Comment définir la qualité d'une partition ?

2. Inertie intra-classe et Inertie inter-classe

n points dans un espace euclidien $d^2(i,i')$ distance euclidienne

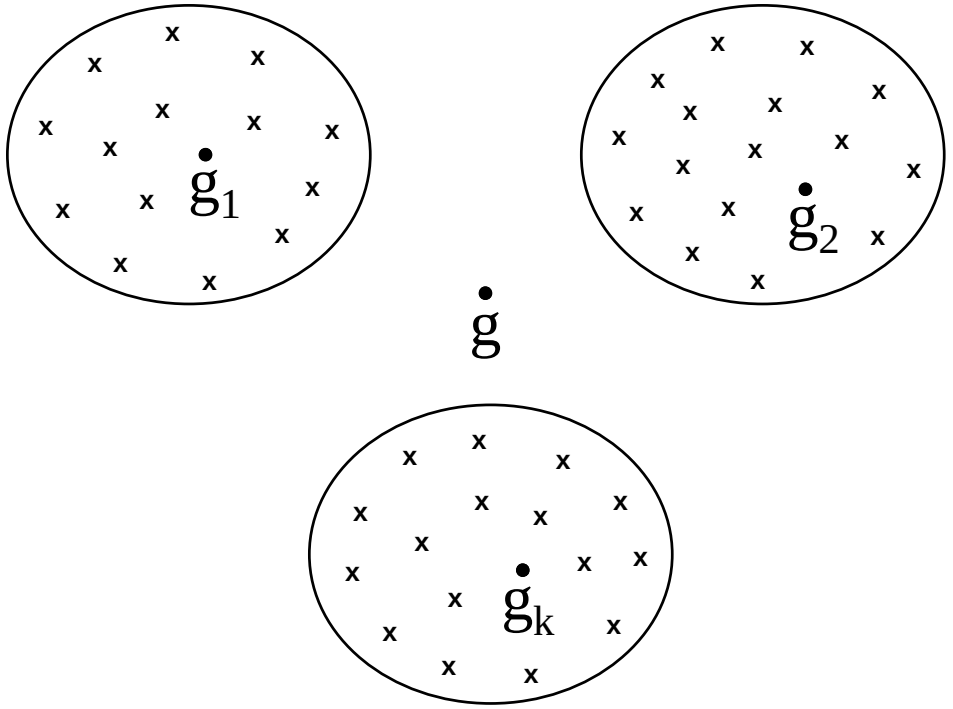
g = centre de gravité des n individus

$g_1, g_2 \dots g_k$ **Centres de gravité des groupes**

$I_1, I_2 \dots I_k$ **Inerties associées**

$$I_W = \sum P_i I_i$$ **Inertie intra-classe**

$$I_B = \sum P_i d^2(g_i, g)$$ **Inertie inter-classe**



$$I_B + I_W = I$$

Inertie-inter + Inertie Intra = Inertie totale

Comparaison de deux partitions en k classes

La meilleure est celle qui a l'inertie-intra la plus faible (ou l'inertie-inter la plus forte).

Remarque : Ce critère ne permet pas de comparer des partitions à nombres différents de classe.

3. Méthode des centres mobiles

Principe:

Etape d'initialisation : choix ou tirage au hasard des centres

Chaque classe est formée de tous les points plus proches de son centre que de tout autre centre.

- **Etape de représentation**

On calcule les représentants de chaque classe : centre de gravité

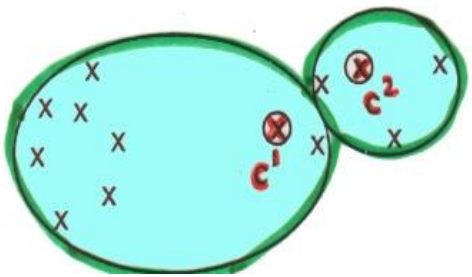
- **Création de nouvelles classes autour des centres de gravité**
- **On procède par itérations successives**

EXEMPLE : Méthode des Centres Mobiles

Etape 0

Choix des centres

c_1 c_2



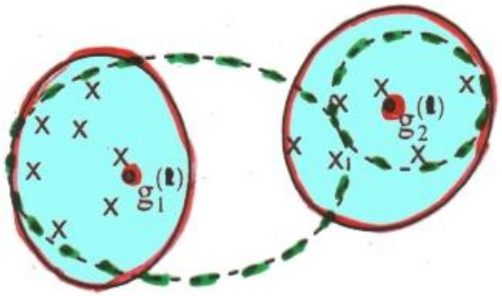
Etape 1

Constitution de classes autour des centres c_1 et c_2

Classe 1 : points plus proches de c_1 que de c_2

Classe 2 : points plus proches de c_2 que de c_1

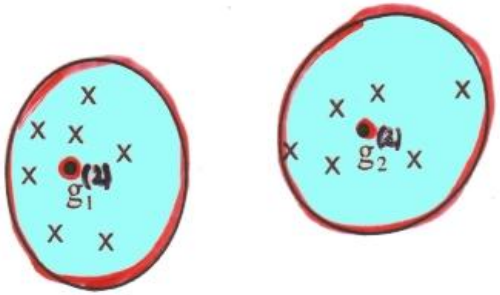
Etape 2



Calcul des centres de gravité
des 2 classes formées à l'étape 1
g_1 g_2

\+ Définition de nouvelles classes
autour des centres de gravité

Etape 3



Calcul des centres de gravité
des classes formées à l'étape 2.
Nouvelle définition des classes
autour de ces centres → STABILITE

⇒ **FIN de l'algorithme**

4. Généralisation : nuées dynamiques

L'idée est d'associer à une classe un **représentant différent de son centre de gravité.**

Par exemple :

- Un ensemble d'individus (noyau formé de q points appelés les étalons)
- Une droite
- Une loi de probabilité

Algorithme - Principe

Il faut faire décroître le critère U mesurant l'adéquation entre les classes et leurs représentants.

Initialisation

Deux possibilités :

1. Soit on se donne au départ une fonction d'affectation qui génère une partition $Q = (Q_1 \dots Q_k)$

   sur E. Les noyaux pour chaque classe sont calculés.

2. Soit on se donne k noyaux.

Étape d'affectation

Pour chaque individu, déterminer la classe à laquelle on doit l'affecter (nécessité d'avoir défini une distance entre un point et un noyau, ou un groupe de points).

Étape de représentation

Pour chaque classe définie, calculer le nouveau noyau.

La convergence vers un minimum local est obtenue si chaque étape fait décroître le critère U.

ARRÊT DE L'ALGORITHME quand la décroissance atteint un seuil fixé a priori.

Pratique de la méthode

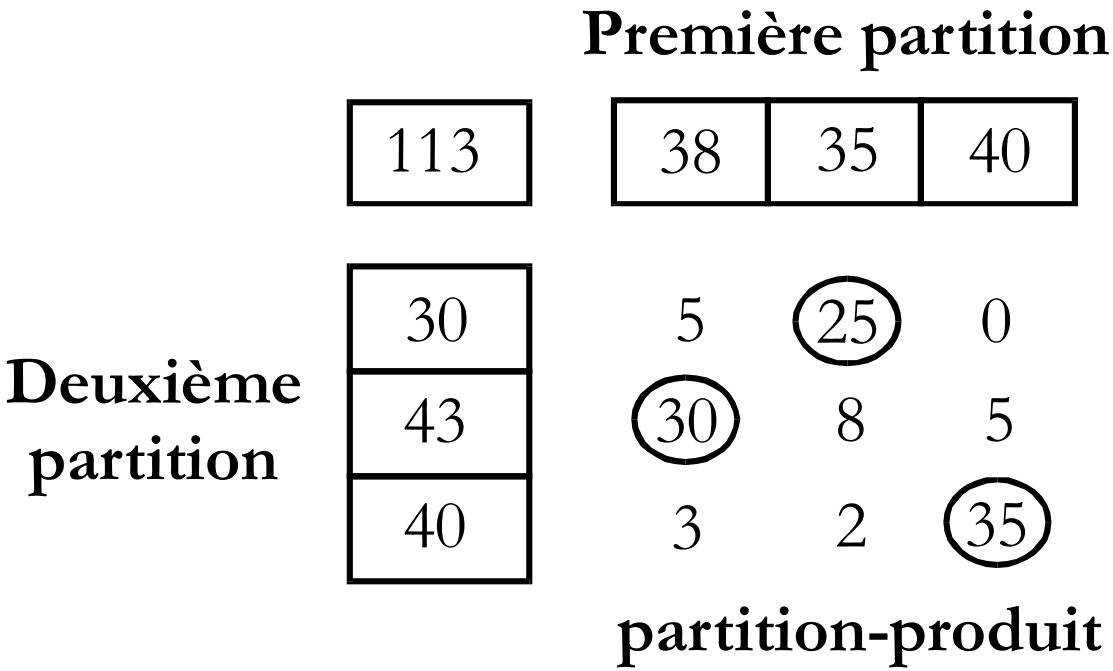
Comme la partition finale peut dépendre de l'initialisation, on recommence s fois (exemple : s tirages aléatoires de noyaux).

Formes fortes

Ensemble d'éléments ayant toujours été regroupés lors de la partition finale.

Exemples :

①


Première partition
113
38
35
40
Deuxième partition
30
43
40
5
25
0
30
8
5
3
2
35
partition-produit

② **1000 individus**

Trois partitions de base en 6 classes :

Partition 1	127	188	229	245	151	60
Partition 2	232	182	213	149	114	110
Partition 3	44	198	325	99	130	204

Ces trois partitions sont ensuite croisées entre elles

6^3 = 216 classes

168	114	110	107	88	83	78 /	26	22	16
15	14	12	12	12	11	10	7	7	7

5. Variantes des méthodes « centres mobiles »

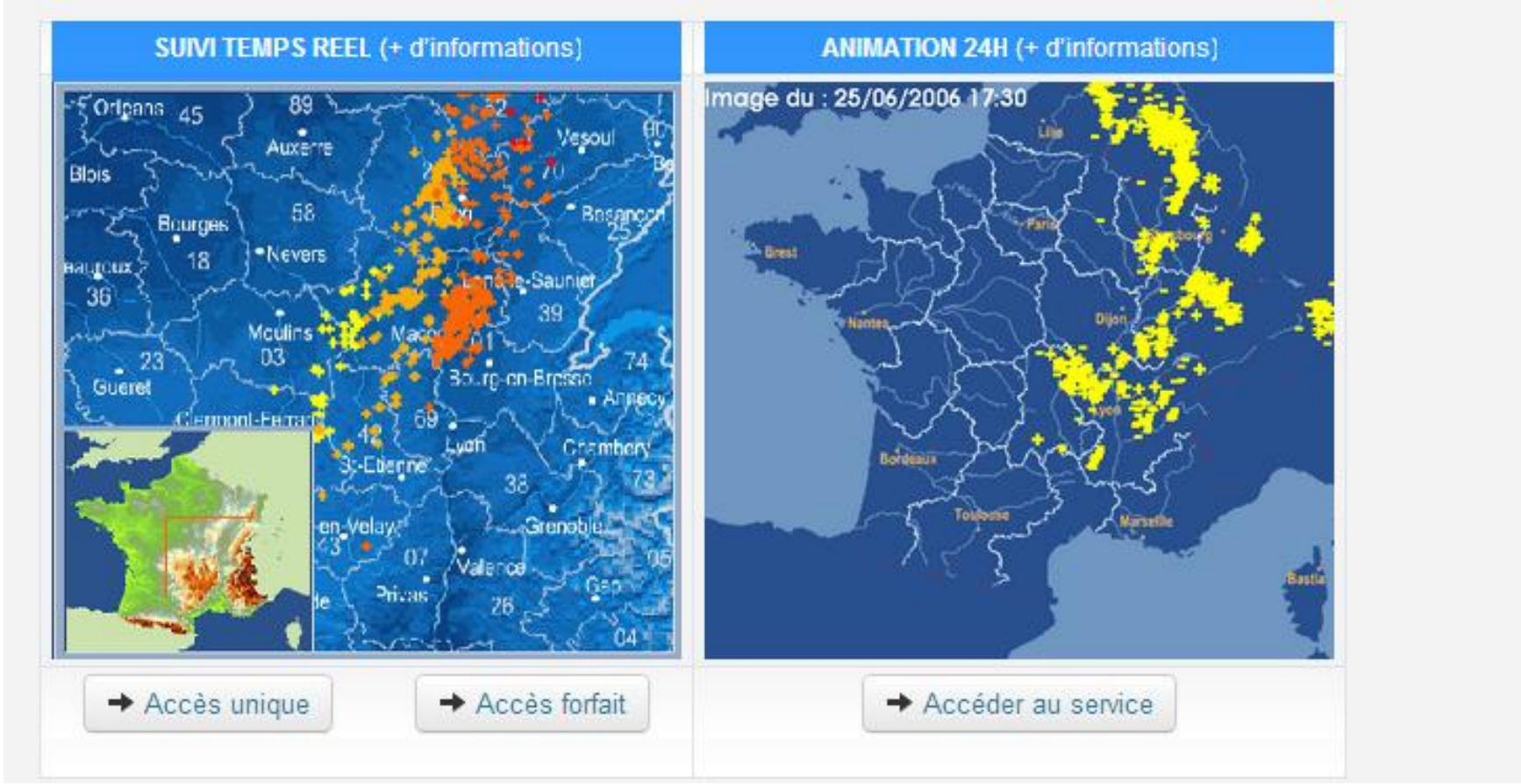
K-means (Mac Queen 1967)

On effectue un recentrage dès qu'un objet change de classe.

Isodata (Ball et Hall 1965)

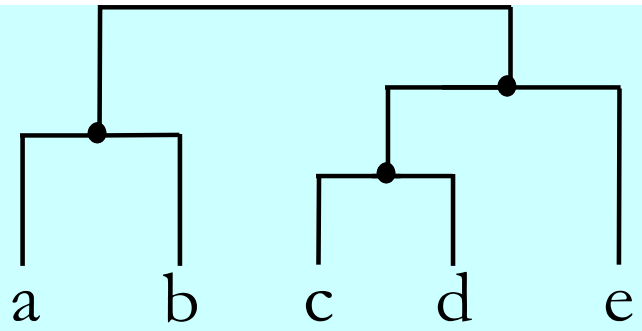
Un certain nombre de contraintes sont imposées pour empêcher la formation de classes d'effectifs trop faibles ou de diamètre trop grand.

Caractérisation de cellules orageuses – Société Météorage



II. LA CLASSIFICATION HIÉRARCHIQUE

Elle consiste à fournir un **ensemble de partitions de E en classes de moins en moins fines** obtenues par regroupements successifs de parties.



Arbre de classification ou dendrogramme

1.Démarche :

Cet arbre est obtenu dans la plupart des méthodes de manière **ascendante** :

- On regroupe d'abord les deux individus les plus proches qui forment un « sommet »
- Il ne reste plus que (n-1) objets et on itère le processus jusqu'à un regroupement complet.

Un des problèmes consiste à définir **une mesure de dissimilarité entre classes.**

Remarque : Les méthodes descendantes (ou algorithmes divisifs) sont pratiquement inutilisées dans le cadre de la classification non supervisée.

2. Aspect formel:

Hiérarchie d'un ensemble de parties de E

Une famille H de parties de E (appelées classes) est une hiérarchie si:

• E et les parties à un élément appartiennent à H.

•Dans une hiérarchie deux classes sont soit disjointes, soit l'une est contenue dans l'autre.

• Toute classe est la réunion des classes qui sont incluses en elles.

Une partition de E compatible avec H est une partition dont les classes sont des éléments de H.

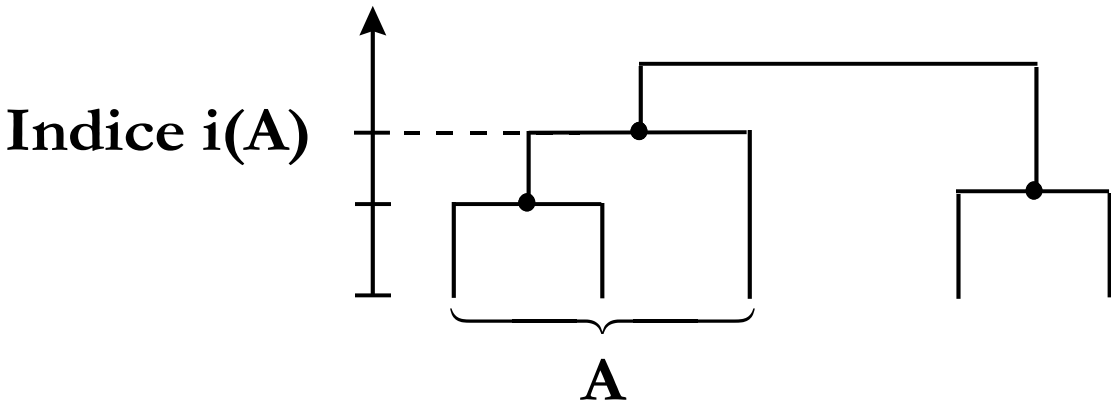
D'une manière imagée, c'est une partition obtenue en coupant l'arbre selon une horizontale

Une hiérarchie est indicée s'il existe une application i de H dans R+ croissante, c'est-à-dire telle que:

si $A \subset B$ alors $i(A) \leq i(B)$

Les indices sont aussi appelés niveaux d'agrégation.

L'indice i(A) est le niveau auquel on trouve agrégés pour la première fois tous les constituants de A



Distances ultramétriques

A toute hiérarchie indicée H correspond un indice de distance entre éléments de H

d(A,B) est le niveau d'agrégation de A et B, c'est-à-dire l'indice de la plus petite partie de H, contenant à la fois A et B.

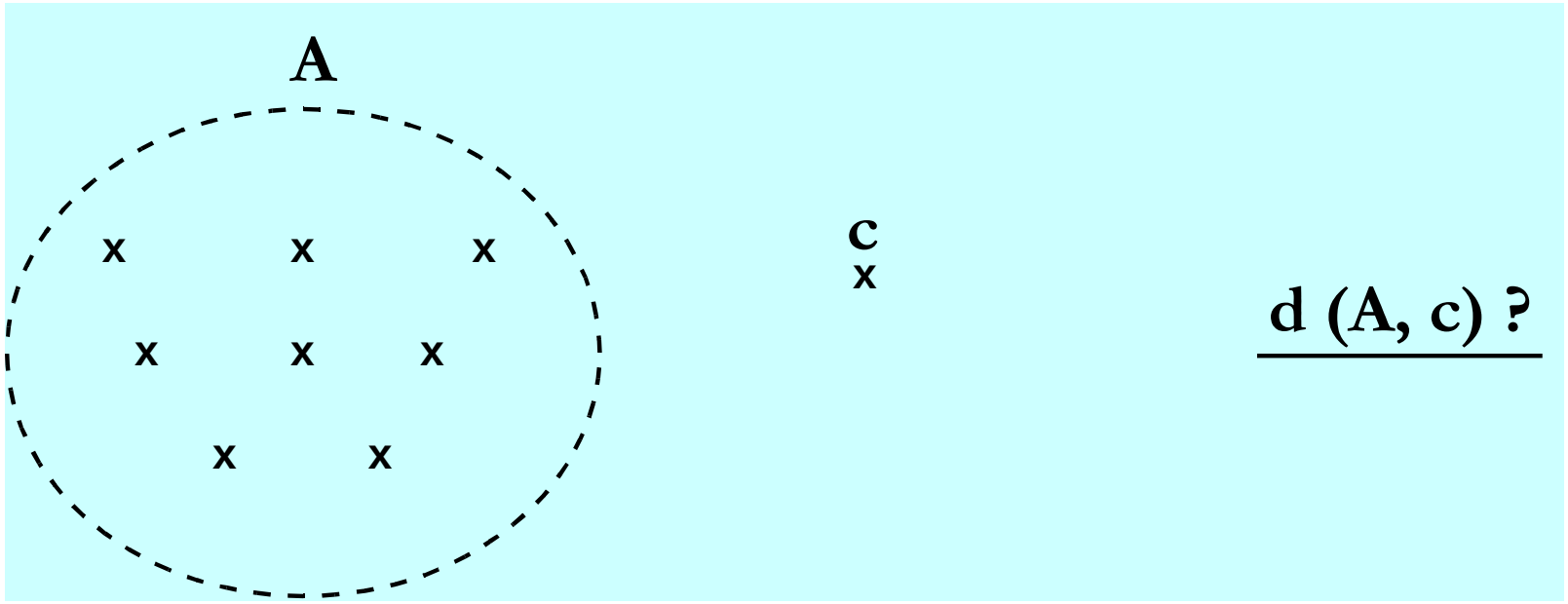
Cette indice de distance possède la propriété suivante dite propriété **ultramétrique**

$$\forall A, B, C \qquad d(A,B) \leq \text{Sup}(d(A,C), d(B,C))$$

3. Stratégies d'agrégation sur dissimilarités

Le problème est de définir la dissimilarité entre la réunion de deux éléments et un troisième :

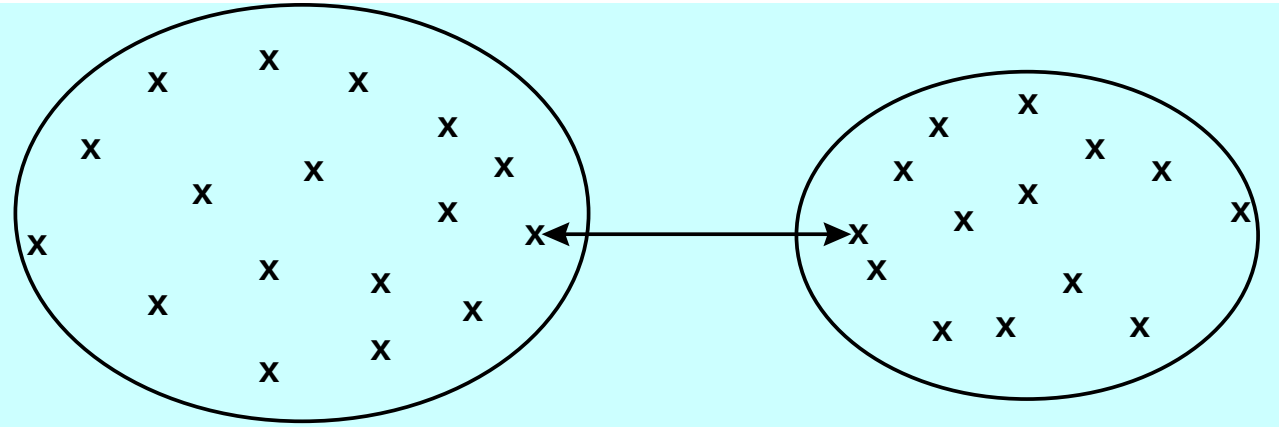
A chaque solution correspond une ultramétrique différente.



a. Le saut minimum

Cette méthode (connue sous le nom de « single linkage » en anglais ») consiste à écrire que :

$$d(a-b,c) = \inf \left\{ d(a,c) ; d(b,c) \right\}$$

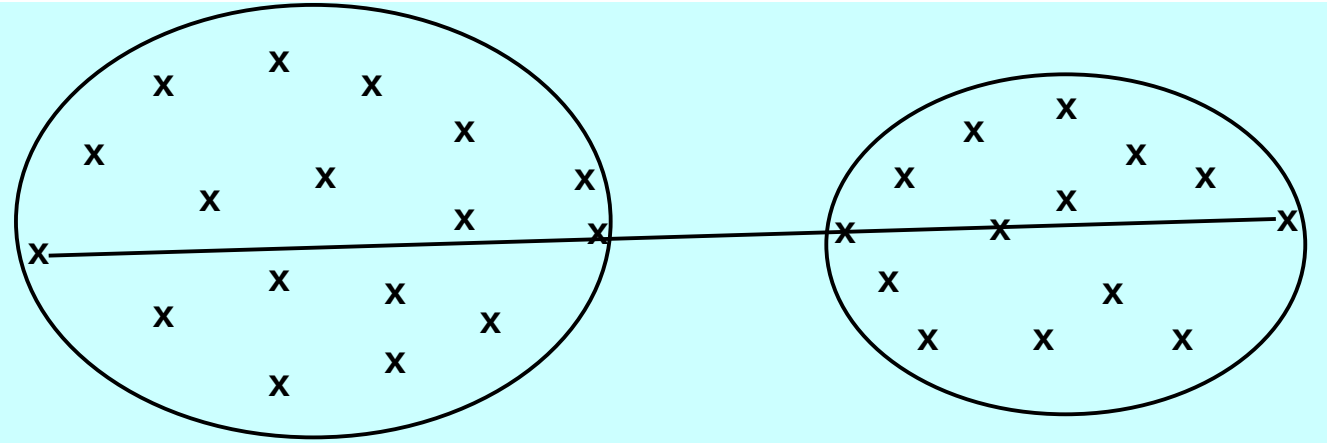


La distance entre parties est donc la plus petite distance entre éléments des deux parties.

b. Le diamètre (« complete linkage »)

On prend ici comme distance entre parties la plus grande distance entre deux éléments.

$$d\left[(a,b)\ ;\ c\right] = \sup\left[d\,(a,c),\ d\,(b,c)\right]$$



c. Stratégies diverses

- saut minimum (plus proche)
- diamètre
- moyenne des distances
- médiane des distances
- distance au centre de gravité.

De nombreuses autres méthodes de calcul de distances entre parties ont été proposées, qui sont toutes des cas particuliers de la formule de Lance et Williams généralisée par Jambu

$$d((a,b),c) = a_1\, d(a,c) + a_2\, d(b,c) + a_3\, d(a,b) + a_4\, i(a) + a_5\, i(b) + a_6\, i(c) + a_7\, |\, d(a,c) - d(b,c)\, |$$

Avec les conditions:

$$a_1 + a_2 + a_3 \geq 1$$

$$a_1, a_2, a_4, a_5, a_6 \geq 0$$

$$a_7 \geq -\min(a_1, a_2)$$

4. La méthode de Ward pour distance Euclidienne

Si on peut considérer E comme un nuage d'un espace $\mathbf{R}^p$, on agrège les individus qui font le moins varier l'inertie intra-classe.

A chaque pas, on cherche à obtenir un **minimum local de l'inertie intra-classe ou un maximum de l'inertie inter-classe**.

L'indice de dissimilarité entre deux classes (ou niveau d'agrégation de ces deux classes) est alors égal à **la perte d'inertie inter-classe résultant de leur regroupement**.

Calculons cette **perte d'inertie** :

g_A centre de gravité de la classe A (poids p_A)
g_B centre de gravité de la classe B (poids p_B)
g_{AB} centre de gravité de leur réunion

$$g_{AB} = \frac{p_A g_A + p_B g_B}{p_A + p_B}$$

L'intertie inter-classe étant la moyenne des carrés des distances des centres de gravité des classes au centre de gravité total, la variation d'inertie inter-classe, lors du regroupement de A et B est égale à :

$$p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_A + p_B)\, d^2(g_{AB}, g)$$

Elle vaut:

$$\delta(A,B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$$

Remarque : Cette méthode entre dans le cadre de la formule de Lance et Williams généralisée :

$$\delta[(A,B);C] = \frac{(p_A + p_C)\,\delta(A,C) + (p_B + p_C)\,\delta(B,C) - p_C\delta(A,B)}{p_A + p_B + p_C}$$

On peut donc utiliser l'algorithme général.

On notera que la somme des niveaux d'agrégation des différents nœuds de l'arbre doit être égale à l'inertie totale du nuage, puisque la somme des pertes d'inertie est égale à l'inertie totale.

Cette méthode est donc **complémentaire de l'analyse en composantes principales** et repose sur un critère d'optimisation assez naturel.

Elle constitue à notre avis la **meilleure méthode de classification hiérarchique sur données euclidiennes.**

Il ne faut pas oublier cependant que le choix de la métrique dans l'espace des individus conditionne également les résultats.

5. Complexité des algorithmes de classification hiérarchique

L'algorithme général consiste à chaque étape:

- « Balayer » un tableau de n(n-1)/2 distances ou dissimilarités, afin de rechercher l'élément de valeur minimale

- Réunir les deux éléments correspondants

- Mettre à jour les distances

- Recommencer avec un élément de moins

La complexité d'un tel algorithme est en n^3 (ordre du nombre d'opérations à effectuer)

La méthode des voisins réciproques (Mac Quitty et Jean-Paul Benzecri)

Elle consiste à réunir simultanément plusieurs paires d'individus, à chaque lecture du tableau des distances.

La complexité de l'algorithme est alors en n^2.

La recherche des voisins réciproques s'effectue en chaîne:

• On part d'un objet quelconque et on cherche son plus proche voisin, puis le plus proche voisin de celui-ci…jusqu'à aboutir à un élément dont le plus proche voisin est son prédécesseur dans la liste.

• On réunit ces deux éléments, et on recommence à partir du nœud créé, ou de l'avant dernier élément de la liste jusqu'à la création de tous les nœuds.

III. LA PRATIQUE DE LA CLASSIFICATION

1. Les méthodes mixtes

En présence d'un grand nombre d'individus (>10^3), il est impossible d'utiliser directement les méthodes de classification hiérarchique.

On combine les techniques non hiérarchiques et hiérarchiques.

Étape 1

Méthode **« centres mobiles »** ou **« nuées dynamiques ».**
On forme par exemple 50 classes.

Étape 2

Construction d'un arbre à partir des k classes formées à l'étape 1. **Coupure de l'arbre en un nombre judicieux de classes.**

Étape 3

Consolidation de la partition obtenue à l'étape 2
(méthode de type « centres mobiles »).

2-1. Utilisation des outils de base de la statistique

Pour chaque variable :

Calcul de paramètres caractéristiques de chaque classe **(moyenne, écart-type, min, max...)**

Représentations graphiques : **boîtes à moustaches**, intervalle de confiance pour les moyennes.

Analyse de la variance à un facteur pour chaque variable (on peut ainsi « classer » les variables par ordre de contribution à la création des classes).

2-2. En liaison avec une analyse factorielle (A.C.P. dans le cas de variables quantitatives)

On peut **repérer les classes formées dans le plan des individus**.

Projeter les points moyens représentant chaque classe

Utiliser les valeurs-tests pour chaque classe sur les axes interprétés.

2-3. Les deux approches sont complémentaires

La première approche peut être longue à mettre en oeuvre si le nombre de variables est élevé.

3. Valeurs tests pour les variables continues

On compare la moyenne $\bar{X}_k$ d'une variable X dans la classe k, à la moyenne générale $\bar{X}$ et on évalue l'écart en tenant compte de la variance $s_k^2(\bar{X})$

Valeur-test: $$t_k(X) = \frac{\bar{X}_k - \bar{X}}{s_k(\bar{X})}$$

n_k = effectif de la classe k et $$s_k^2(\bar{X}) = \frac{n - n_k}{n - 1} \frac{s^2(X)}{n_k}$$

La valeur test t_k suit approximativement une loi de Laplace-Gauss

Repérer les valeurs-tests t.q. | V.T.| > 2