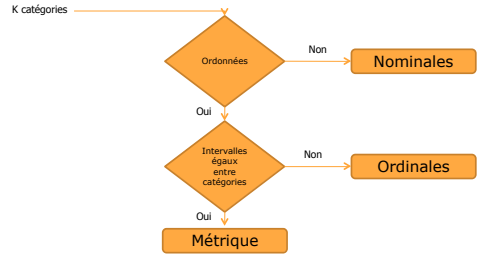
Analyse Typologique

Cluster Analysis

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Le questionnaire : les types de mesures



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Terminologie

Qui ils SONT — **Variables** — **Leurs réponses et comportements**

Caractéristiques C
Variable de regroupement

Réponses X, Y, Z

Individus n

DATA

	A	B	C	D	E	F	G	H
1	ACCTNUM	GENDER	MONEY	RECENCY	FREQUENC	FIRST	NBCHILD	NBYOUTH
2	14086	F	440	14	12	54	3	1
3	31310	F	189	10	11	52	3	1
4	57640	F	249	8	11	40	4	0
5	19155	F	231	32	10	72	2	0
6	50575	F	311	14	11	58	2	1
7	31326	F	433	18	11	60	5	0
8	69486	F	375	10	12	46	6	2
9	41955	F	314	12	11	74	5	0
10	39430	F	234	10	10	60	6	0
11	25520	F	252	12	10	44	3	1
12	20520	F	217	8	10	44	5	2

- Une variable x
- Est mesurée sur un individu i et donne une observation x_i
- Il y a *n* observations (effectifs)

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Codage des données

	Var_1 (âge)	Var_2 (sexe)	Var_3 (satisfaction)	…	Var_V
Sujet 1	25	2	5	…	11
Sujet 2	48	1	2	…	13
…	…	…	…	…	…
Sujet N	37	1	7	…	8

1 : homme 1 : pas du tt s.
2 : femme 2 : très satisfait

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

L'analyse typologique

- L'analyse typologique étudie un ensemble de relations de corrélation
- Elle ne fait pas de distinction entre variables dépendantes et indépendantes
- Son objectif principal est de :
  - Classer des individus dans des groupes relativement homogènes en fonction de l'ensemble des variables considérées
- Les individus d'un même groupe sont similaires selon les critères de ces variables et différents des individus des autres groupes

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Les analyses multivariées

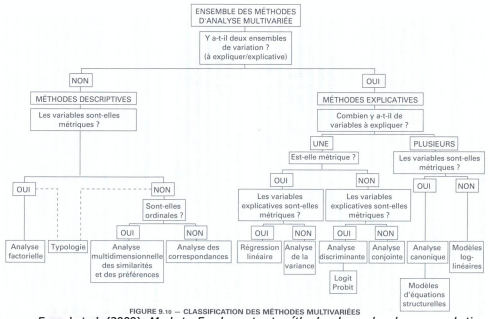


FIGURE 9-10 — CLASSIFICATION DES MÉTHODES MULTIVARIÉES

Source : Evrard et al. (2009), *Market : Fondements et méthodes des recherches en marketing,* Dunod

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

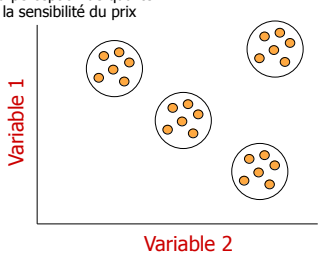
Concepts de base

- L'analyse typologique :
  - est utilisée pour classer des objets ou des individus en ensembles relativement homogènes appelés « groupes » ou « classes »
  - dans lesquels les individus tendent à être semblables entre eux et différents des individus des autres groupes
- Elles sont aussi appelée analyse de classification ou taxinomie
- Les classes ou groupes sont choisis à partir des données, et non définis a priori

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Une situation idéale

Classification idéale : les groupes sont distincts selon deux variables
Chaque consommateur se retrouve dans un groupe et il n'y a pas de recoupement
Variable 1 : la perception de qualité
Variables 2 : la sensibilité du prix



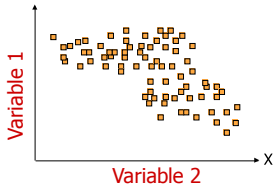
Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Une situation pratique

Cette classification sera plus souvent rencontrée dans la pratique

Les limites de certains groupes ne sont pas clairement découpées

La classification de certains consommateurs n'est pas évidente



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Les notions statistiques associées à la classification

- La plupart des méthodes de classification sont des procédures relativement simples
- Elles ne s'appuient pas sur une longue suite de raisonnements statistiques
- Ce sont des méthodes heuristiques, fondées sur des algorithmes
- La simplicité des méthodes de classification doit être soulignée

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Les notions statistiques associées à la classification

- **La chaîne des agrégations (Agglomeration Schedule)**
  - Elle donne des informations sur les individus combinés à chaque étape du processus de classification hiérarchique

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	1.000	0	0	6
2	6	7	2.000	0	0	7
3	2	13	3.500	0	0	15
4	5	11	5.000	0	0	11
5	3	8	6.500	0	0	16
6	10	14	8.167	0	1	9
7	6	12	10.500	2	0	10
8	9	20	13.000	0	0	11
9	4	10	15.583	0	6	12
10	1	6	18.500	0	7	13
11	5	9	23.000	4	8	15
12	4	19	27.750	9	0	17
13	1	17	33.100	10	0	14
14	1	15	41.333	13	0	16
15	2	5	51.833	3	11	18
16	1	3	64.500	14	5	19
17	4	18	79.667	12	0	18
18	2	4	172.667	15	17	19
19	1	2	328.600	16	18	0

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Les notions statistiques associées à la classification

- **Le centroïde (Cluster centroid)**
  - Également appelé centre de gravité
  - C'est l'ensemble des valeurs moyennes des variables pour tous les individus d'un même groupe
- **Les noyaux (Cluster centers)**
  - Ce sont les points de départ dans la classification non hiérarchique
  - Les groupes sont construits autour de ces centres

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Les notions statistiques associées à la classification

- **L'appartenance à un groupe (Cluster centroid)**
  - Elle indique le groupe auquel chaque individu appartient

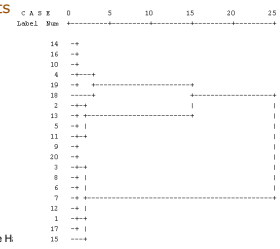
Cluster Membership

Case	6 Clusters	5 Clusters	4 Clusters	3 Clusters	2 Clusters
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	1	1	1
4	4	4	3	3	2
5	5	2	2	2	2
6	1	1	1	1	1
7	1	1	1	1	1
8	3	3	1	1	1
9	5	2	2	2	2
10	4	4	3	3	2
11	5	2	2	2	2
12	1	1	1	1	1
13	2	2	2	2	2
14	4	4	3	3	2
15	1	1	1	1	1
16	4	4	3	3	2
17	1	1	1	1	1
18	6	5	4	3	2
19	4	4	3	3	2
20	5	2	2	2	2

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Les notions statistiques associées à la classification

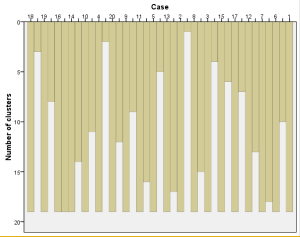
- **Le dendogramme (arbre hiérarchique)**
  - C'est un outil graphique qui permet d'exposer les résultats de la classification
  - Il se lit de gauche à droite
  - Les lignes verticales représentent les groupes qui se rejoignent
  - La position de la ligne sur l'échelle indique les distances auxquelles les groupes sont joints



Copyright © 2010 Pearson Education, Inc. publishing as Prentice H

Les notions statistiques associées à la classification

- **Les distances entre les groupes**
  - Elles montrent à quel point deux groupes sont séparés
  - Les groupes largement séparés sont distincts et, par conséquent, souhaitables
- **Le diagramme en stalactite (icycle plot)**
  - C'est une représentation graphique des résultats de la classification, appelé ainsi car il ressemble à une rangée de stalactites
  - Il se lit de bas en haut
  - Les colonnes correspondent aux individus à classer et les rangs au nombre de groupes



Copyright © 2010 Pearson Education, Inc. publishing as Prer

Les notions statistiques associées à la classification

- **La matrice des distances ou des similarités**
  - C'est une matrice à base triangulaire contenant les distances, par paires, entre les individus

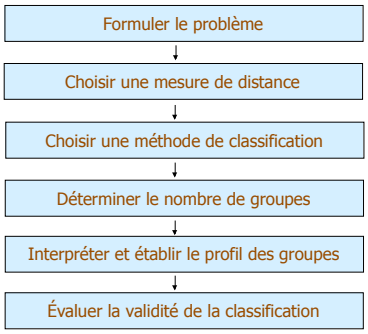
Proximity Matrix

Squared Euclidean Distance

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.000	64.000	8.000	31.000	69.000	3.000	5.000	5.000	48.000	48.000	60.000	7.000	65.000	48.000	13.000	48.000	5.000	56.000	56.000	69.000
2	64.000	.000	68.000	31.000	7.000	47.000	39.000	77.000	8.000	16.000	4.000	35.000	3.000	36.000	49.000	28.000	55.000	24.000	44.000	9.000
3	8.000	68.000	.000	43.000	83.000	11.000	11.000	3.000	84.000	56.000	70.000	11.000	61.000	58.000	19.000	58.000	23.000	78.000	68.000	79.000
4	31.000	31.000	43.000	.000	44.000	20.000	22.000	36.000	31.000	5.000	39.000	12.000	34.000	3.000	22.000	5.000	24.000	17.000	7.000	50.000
5	69.000	7.000	83.000	44.000	.000	52.000	42.000	90.000	5.000	33.000	3.000	46.000	16.000	51.000	58.000	41.000	52.000	41.000	69.000	10.000
6	3.000	47.000	11.000	20.000	52.000	.000	3.000	8.000	35.000	33.000	41.000	4.000	50.000	31.000	10.000	33.000	8.000	45.000	43.000	54.000
7	5.000	39.000	11.000	22.000	42.000	3.000	.000	10.000	29.000	31.000	37.000	4.000	40.000	33.000	14.000	31.000	8.000	47.000	45.000	46.000
8	5.000	77.000	3.000	36.000	90.000	8.000	10.000	.000	89.000	53.000	79.000	10.000	72.000	49.000	18.000	51.000	16.000	71.000	53.000	90.000
9	48.000	8.000	84.000	31.000	5.000	35.000	29.000	89.000	.000	24.000	4.000	31.000	17.000	38.000	45.000	32.000	41.000	24.000	56.000	5.000
10	48.000	16.000	56.000	5.000	33.000	33.000	31.000	53.000	24.000	.000	28.000	21.000	19.000	4.000	39.000	2.000	39.000	14.000	8.000	35.000
11	60.000	4.000	70.000	39.000	3.000	41.000	37.000	79.000	4.000	28.000	.000	37.000	9.000	48.000	51.000	39.000	49.000	39.000	60.000	7.000
12	7.000	35.000	11.000	12.000	46.000	4.000	4.000	10.000	31.000	21.000	37.000	.000	34.000	23.000	8.000	23.000	10.000	31.000	27.000	40.000
13	65.000	3.000	61.000	34.000	16.000	50.000	40.000	72.000	17.000	19.000	9.000	34.000	.000	39.000	52.000	29.000	58.000	29.000	41.000	16.000
14	48.000	36.000	58.000	3.000	51.000	31.000	33.000	49.000	38.000	4.000	48.000	23.000	39.000	.000	39.000	2.000	37.000	21.000	6.000	55.000
15	13.000	49.000	19.000	22.000	58.000	10.000	14.000	18.000	45.000	39.000	51.000	8.000	52.000	39.000	.000	43.000	16.000	43.000	41.000	68.000
16	48.000	28.000	58.000	5.000	41.000	33.000	31.000	51.000	32.000	2.000	36.000	23.000	29.000	2.000	43.000	.000	35.000	24.000	8.000	47.000
17	5.000	55.000	23.000	24.000	52.000	8.000	8.000	16.000	41.000	39.000	49.000	10.000	58.000	37.000	16.000	35.000	.000	49.000	49.000	60.000
18	56.000	24.000	78.000	17.000	41.000	45.000	47.000	71.000	24.000	14.000	30.000	31.000	29.000	21.000	43.000	24.000	59.000	.000	24.000	31.000
19	56.000	44.000	68.000	7.000	69.000	43.000	45.000	53.000	56.000	8.000	60.000	27.000	41.000	6.000	41.000	8.000	49.000	24.000	.000	73.000
20	69.000	9.000	79.000	50.000	10.000	54.000	46.000	90.000	5.000	35.000	7.000	40.000	16.000	55.000	68.000	47.000	60.000	31.000	73.000	.000

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique

Formuler le problème

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Formuler le problème

- **La partie la plus délicate de la formulation du problème est peut être le choix des variables**
- L'inclusion ne serait ce que d'une ou de deux variables non pertinentes peut fausser les résultats d'une classification
- L'ensemble des variables choisies doit décrire les similitudes entre les individus selon des critères appropriées
- Les variables doivent être choisies à partir de recherches antérieures, d'éléments de théorie…

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique

Choisir une mesure de distance ou de similarité

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une mesure de distance ou de similarité

- L'objectif de la classification est de regrouper des individus similaires
- **Il faut évaluer le degré de similarité ou de différence**
- L'approche la plus fréquente consiste à mesurer la similarité en fonction de la **distance entre les paires d'individus**
  - Plus la distance est petite entre les individus, plus ils sont similaires
- Il existe plusieurs méthodes pour calculer la distance entre deux individus

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une mesure de distance ou de similarité

- La mesure de similarité la plus utilisée est:
  - **LA DISTANCE EUCLIDIENNE OU SON CARRE**
- La **Distance euclidienne** est :
  - La racine carrée de la somme des carrés des différences entre valeurs pour chaque variable

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- D'autres mesures sont aussi disponibles
  - La distance de **city-block ou Manhattan** entre deux objets est la somme des différences absolues des valeurs pour chaque variable

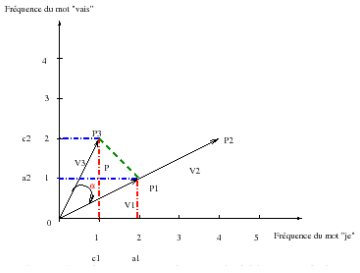
$$d = \sum_{i=1}^{n} |x_i - y_i|$$

  - La distance de **Chebychev** distance entre deux individus est la différence absolue maximale des valeurs pour toute variable

$$\mathbf{Max}_i\ |X_i - Y_i|$$

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une mesure de distance ou de similarité



$$D_e(P1P3) = \sqrt{(a_1 - c_1)^2 + (a_2 - c_2)^2}$$

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une mesure de distance ou de similarité

- Si les variables sont mesurées avec des unités très différentes, la solution de classification sera influencée
- Ex :
  - Dans une étude sur les achats en supermarché, les variables d'attitude peuvent être mesurées sur une échelle de Lickert en 9 points :
    - Comportements en termes de fréquence de visite par mois, de montant dépensé en euros
    - La fidélité à la marque, en termes de pourcentage de dépenses alimentaires réalisées…
- **Avant de classer les individus, on doit standardiser ou réduire les données** en transformant chaque variable afin qu'elle ait une moyenne de 0 et un écart-type de 1

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une mesure de distance ou de similarité

- L'utilisation de mesures de distances différentes peut conduire à des résultats de classification différents
- **Il est recommandé d'utiliser différentes mesures de classification différents et d'en comparer les résultats**
- Après avoir choisi une mesure de distance ou de similarité, il faut ensuite choisir une méthode de classification

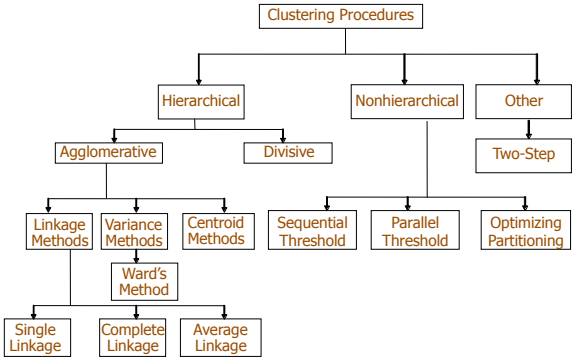
Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique

Choisir une méthode de classification

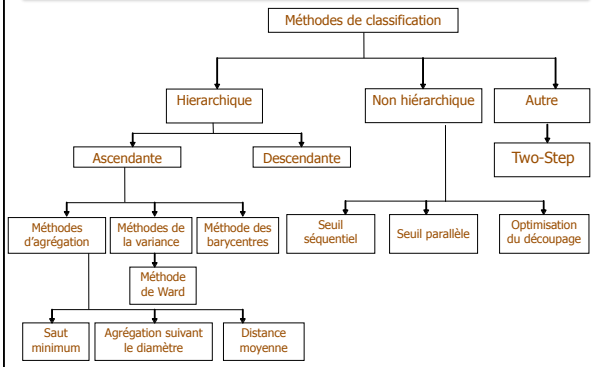
Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

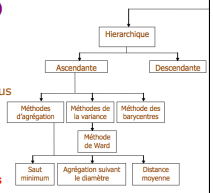
Choisir une méthode de classification



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification hiérarchique

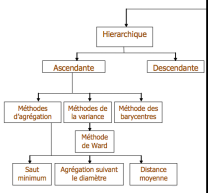
- **Classification hiérarchique**
  - Caractérisée par l'établissement d'une hiérarchie ou structure arborescente
  - Elle peut être ascendante ou descendante
- **Classification ascendante (utilisées en mkt)**
  - **Elle commence avec chaque individu dans un groupe différent**
  - Les groupes sont ensuite formés en agglomérant les individus
  - Ce processus continue avec des groupes de plus en plus gros jusqu'à ce que tous les individus appartiennent à un seul groupe
- **Classification descendante**
  - **Elle commence avec tous les individus regroupés dans un seul groupe**
  - celui-ci est ensuite divisé ou éclaté jusqu'à ce que chaque individu se retrouve dans un groupe séparé



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification hiérarchique

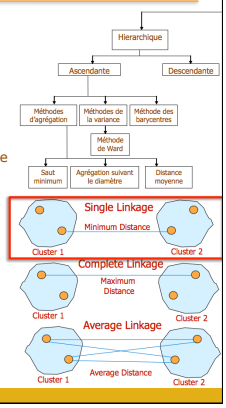
- **Méthodes ascendantes**
  - Elles regroupent des méthodes :
    - **D'agrégation**
      - Saut minimum
      - Distance du diamètre
      - Distance moyenne
    - **De sommes des carrés des erreurs ou de variance**
      - Méthode de Ward
    - **Des barycentres**



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification ascendante

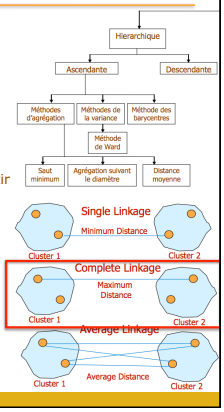
- La méthode du **saut minimum**
  - **Repose sur la distance minimum ou la règle du plus proche voisin**
  - Les 2 premiers individus classés sont ceux qui ont la plus petite distance entre eux
  - La distance suivante la plus petite est identifiée
  - Le 3ème individu est classé avec les 2 premiers, soit un nouveau groupe est créé



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification ascendante

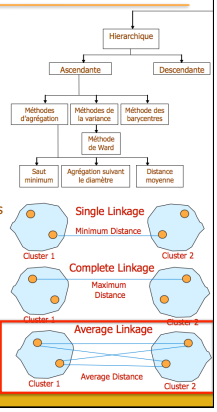
- La méthode de **l'agrégation suivant le diamètre**
  - Est comparable à la précédente
  - Toutefois, **elle repose sur la distance maximum ou la règle du voisin le plus éloigné**
  - La distance entre 2 groupes est calculée à partir de la distance entre leurs 2 points les plus éloignés



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

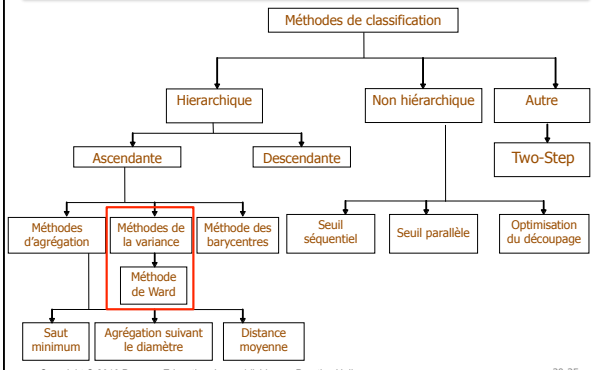
Choisir une méthode de classification ascendante

- La méthode de **distance de moyenne entre classes**
  - Elle fonctionne de façon similaire
  - Cependant, **la distance entre 2 groupes est alors définie comme la moyenne des distances entre toutes les paires d'individus**, avec, pour chaque paire, un membre de chaque groupe
  - Cette méthode utilise l'information de toutes les paires de distances et pas simplement les distances minimum ou maximum
  - C'est pour cette raison qu'on la préfère souvent aux autres



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

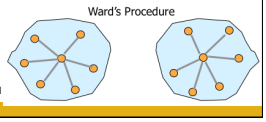
Choisir une méthode de classification



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification ascendante

- Les **méthodes de variance**
  - Tentent de générer des groupes afin de minimiser la variance à l'intérieur des groupes
- L'une d'entre elles, fréquemment utilisée, s'appelle la **méthode de Ward**
  - Pour chaque groupe, **les moyennes pour toutes les variables sont calculées**
  - Ensuite, **pour chaque individu, le carré de la distance euclidienne au centre de la classe est calculé**
  - **Ces distances sont additionnées pour tous les individus**
  - À chaque étape, **les 2 groupes ayant la plus petite augmentation dans la somme globale des carrés des distances à l'intérieure des groupes sont réunis**



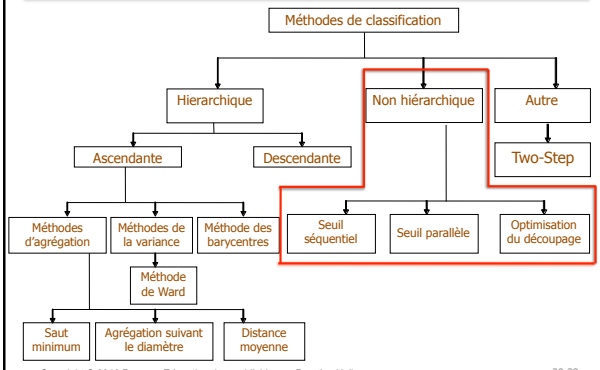
Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une méthode de classification ascendante

- La **méthode des barycentres**
  - **La distance entre 2 groupes correspond à la distance entre leurs barycentres** (moyenne pour toutes les variables)
  - Chaque fois que des individus sont regroupés, un nouveau barycventre est calculé
  - De toutes les méthodes, les méthodes de Ward et de la distance euclidienne moyenne se sont révélées être les plus performantes

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

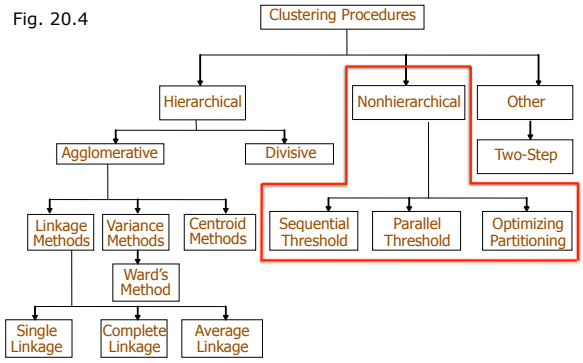
Choisir une méthode de classification



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

A Classification of Clustering Procedures

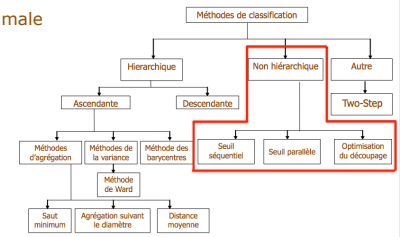
Fig. 20.4



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une procédure de classification – Non hiérarchique

- Les **méthodes de classification non hiérarchiques** :
  - Le seuil séquentiel
  - Le seuil parallèle
  - La partition optimale



Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une procédure de classification – Non hiérarchique

- Dans la **méthode du seuil séquentiel :**
  - **Un centre de groupe est choisi**
  - **Tous les individus dont la distance au centre est inférieure à une valeur seuil prédéfinie sont regroupés**
  - Ensuite, un autre centre de groupe est choisi et le processus est répété pour les points qui ne sont pas encore classés dans un groupe
  - Une fois qu'un individu est classé, il n'est plus pris en compte pour la classification avec les autres centres suivants

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une procédure de classification – Non hiérarchique

- Le **seuil parallèle** s'applique de la même manière
  - Toutefois, **plusieurs centres de groupes sont choisis simultanément**
  - Les individus au-dessous de ce seuil sont regroupés avec le centre le plus proche

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une procédure de classification – Non hiérarchique

- La méthode de **partition optimisée :**
  - A la différence des précédentes, **les individus peuvent être réaffectés à d'autres groupes** pour optimiser un critère global, tel que la distance moyenne intragroupe pour un nombre de groupes donné

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Choisir une procédure de classification – Non hiérarchique

- Les 2 inconvénients majeurs des méthodes non hiérarchiques sont que :
  - **Le nombre de groupes doit être spécifié avant**
  - **Le choix des centres de classe est arbitraire**
    - En général, les programmes sélectionnent les k premiers individus (k=nb de groupes) sans valeurs manquantes comme les centres des groupes de départ
    - Les résultats dépendent de l'autre des observations dans les données

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique

Décider le nombre de groupes

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Décider le nombre de groupes

- Bien qu'il n'y ait pas de règles strictes, on peut suivre quelques indications
  - **Théoriques, conceptuelles, pratiques…**
  - **Pour la classification hiérarchique :**
    - Les distances auxquelles les groupes sont agrégées peuvent servir de critère
    - Cette information peut être obtenue à partir de la chaîne d'agrégation ou du dendogramme
  - **Pour la classification non hiérarchique :**
    - Le rapport de la variance intragroupe totale sur la variance intergroupe peut être représenté graphiquement en fonction du nombre de groupes
    - Le point auquel se forme un coude ou une courbure brutale indique le nombre approprié de groupes
  - **Les tailles relatives des groupes doivent être sensées**

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique

Interpréter le profil des groupes

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Interpréter le profil des groupes

- Interpréter et établir le profil des groupes passe par l'examen des centres de groupes
- Ces centres permettent de décrire chaque groupe en lui attribuant un nom
- Il est souvent utile de décrire les groupes par des variables qui n'ont pas été utilisées pour l'analyse typologique :
  - Variables démographiques, psychométriques, utilisation de produit, …

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Mener une analyse typologique

Évaluer la fiabilité et la validité

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall

Évaluer la fiabilité et la validité

1. Réaliser l'analyse typologique sur les mêmes données en utilisant des mesures de distances différentes. Comparer les résultats afin de déterminer la stabilité des solutions.
2. Utiliser différentes méthodes de classification et comparer les résultats
3. Diviser les données en 2 moitiés de façon aléatoire. Réaliser la classification séparément sur chaque moitié. Comparer les centres de groupes pour les 2 sous-échantillons.
4. Supprimer aléatoirement des variables. Réaliser la classification sur les variables conservées. Comparer les résultats avec ceux obtenus à partir de l'ensemble des données
5. Dans la classification non hiérarchique, la solution peut dépendre de l'ordre des individus dans l'ensemble des données. Réaliser l'analyse plusieurs fois en utilisant dans des ordres différents jusqu'à ce que la solution se stabilise

Copyright © 2010 Pearson Education, Inc. publishing as Prentice Hall