

**Université Paris – Dauphine**

**Ecole Doctorale de Gestion**

*M. Gettler – Summa, C. Pardoux*

# **LA CLASSIFICATION AUTOMATIQUE**

# **Une problématique en Gestion - Marketing**

**Découper le marché en sous-ensembles dont les éléments réagissent de façon similaire aux variations des variables d'action du marché.**

# Exemples

**Identifier des groupes d'individus ou de ménages ayant un comportement homogène vis-à-vis de :**

- la consommation de différents produits,
- la consommation de différentes marques ou variétés,
- l'attitude par rapport à un produit,
- ...

**Il s'agit de problèmes souvent traités avec les méthodes de classification automatique.**

# Données

- **$n$  objets (ou individus) caractérisés par  $p$  descripteurs, ou**
- **tableau carré symétrique de ressemblances (similarités, dissimilarités, distances).**

# Indice de dissimilarité

Soit  $E$  l'ensemble des  $n$  objets à classer.

Une dissimilarité  $d$  est une application de  $E \times E$  dans  $\mathbb{R}^+$  telle que :

1.  $d(i, i) = 0 \quad \forall i \in E$
2.  $d(i, i') = d(i', i) \quad \forall i, i' \in E \times E$

Une distance satisfait les propriétés d'un indice de dissimilarité.

# Objectif

**Constituer des groupes d'objets *homogènes et différenciés*, i.e. des groupes d'objets tels que :**

- les objets soient les plus similaires possibles au sein d'un groupe (*critère de compacité*),
- les groupes soient aussi dissemblables que possible (*critère de séparabilité*),

**la ressemblance ou la dissemblance étant mesurée sur l'ensemble des variables descriptives.**

# Hypothèse

On suppose qu'une structure de classes existe au sein de la population étudiée, le but de l'analyse est de la mettre à jour, de l'identifier.

## Exemples

- Classification des consommateurs d'apéritifs,
- Classification de la clientèle d'une banque,
- Classification des 36 000 communes françaises,
- ...

# Tableaux analysés

La classification est réalisée sur :

- un tableau de valeurs numériques,
- un tableau de contingence,
- un tableau de « présence – absence »,

ou

- un tableau carré symétrique de similarités ou de dissimilarités (distances, par ex.).

# Représentation

**La représentation synthétique peut être :**

- une typologie,
- un recouvrement (classes empiétantes),
- une partition,
- une hiérarchie de partitions (arbre hiérarchique),
- une hiérarchie de recouvrements (pyramide).

# Les étapes d'une classification automatique

1. Choix des données.
2. Calcul des dissimilarités entre les  $n$  individus à partir du tableau initial.
3. Choix d'un algorithme de classification et exécution.
4. L'interprétation des résultats :
  - évaluation de la qualité de la classification,
  - description des classes obtenues.

# Une classification : remarque

Une classification automatique obtenue sur un ensemble n'est jamais LA classification de cet ensemble, mais une classification (parmi beaucoup d'autres) établie à partir de variables et de méthodes choisies intentionnellement.

# Etape 1 : choix des données

**La classification obtenue est liée aux variables choisies pour décrire les individus.**

On distingue :

- les variables actives, celles sur lesquelles sera basée la classification des individus,
- les variables illustratives (ou supplémentaires) qui serviront à décrire les classes constituées : variables décrivant les caractéristiques de l'individu (variables sociodémographiques, ...).

## Etape 2 : calcul des ressemblances

Il existe un grand choix de mesures de ressemblances. Le tableau obtenu est un tableau carré de dimension  $n$ .

### ➤ *Variables quantitatives*

La distance euclidienne est une mesure possible de la ressemblance. Dans le cas de variables hétérogènes, il faut travailler sur les données centrées réduites.

### ➤ *Variables qualitatives*

De nombreux indices de ressemblance ont été proposés : dans le cas d'objets décrits par des variables binaires, indice de Jaccard, indice de Russel et Rao, ... (Saporta, 1990).

# Les méthodes de classification hiérarchique

- La classification ascendante hiérarchique (CAH) conduit à la construction d'un *arbre de classification* (ou *dendrogramme*) montrant le passage des  $n$  individus au groupe « total » par une succession de regroupements.
- La classification descendante hiérarchique procède à l'inverse par subdivisions successives de l'ensemble à classer.

On peut obtenir une partition à partir d'une hiérarchie (partitionnement indirect).

# Les méthodes de partitionnement direct

- Opérer au sens d'un critère donné, le « meilleur » regroupement possible des individus en un nombre choisi a priori de classes.
- Méthodes : agrégation autour des centres mobiles, méthode des nuées dynamiques (Lebart et al., 2000).
- Principe de ces méthodes : constitution de  $k$  groupes ( $k$  étant un nombre choisi par l'analyste) à partir des  $n$  individus sur la base d'un algorithme itératif « Recentrage/Réaffectation » en essayant d'optimiser un indice global mesurant la qualité de la classification.

Remarque :  $(2^{n-1} - 1)$  partitions de  $n$  individus en 2 classes,  
⇒ exploration intelligente, appelée encore *heuristique*.

# Algorithme de la classification ascendante hiérarchique

Phase préalable : Calcul des dissimilarités des objets 2 à 2

Entrées :  $n(n - 1)/2$  dissimilarités

Regroupement des 2 éléments  
les plus proches

jusqu'au regroupement  
de tous les objets  
en un seul groupe :  
( $n - 1$ ) étapes

Calcul d'une nouvelle matrice de  
dissimilarités entre les éléments  
(objets isolés ou groupes) restants

# Stratégie d'agrégation

## 1<sup>ère</sup> étape :

si  $d$  est une dissimilarité, on choisit  $e_i$  et  $e_{i'}$  tels que  $d(e_i, e_{i'})$  minimum  $\Rightarrow G_1 = \{e_i, e_{i'}\}$

## 2<sup>ème</sup> étape :

nouveau tableau de dissimilarités  $(n - 1) \times (n - 1)$

$\Rightarrow$  nécessité de définir une *méthode d'agrégation* entre un individu et un groupe d'individus ou entre deux groupes d'individus.

# Méthodes d'agrégation

*Lien minimum*

$$\delta(A, B) = \min \{d(a, b), a \in A, b \in B\}$$

*Lien maximum*

$$\delta(A, B) = \max \{d(a, b), a \in A, b \in B\}$$

*Distance des centres de gravité*

$$\delta(A, B) = d(g_a, g_b)$$

...

# Exemple

*Agrégation selon le lien minimum*

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	0	23	35	43	50
<i>b</i>	23	0	21	32	45
<i>c</i>	35	21	0	11	25
<i>d</i>	43	32	11	0	17
<i>e</i>	50	45	25	17	0

$$G_1 = \{c, d\} \Rightarrow$$

	<i>a</i>	<i>b</i>	<i>e</i>	$G_1$
<i>a</i>	0	23	50	35
<i>b</i>	23	0	45	21
<i>e</i>	50	45	0	17
$G_1$	35	21	17	0

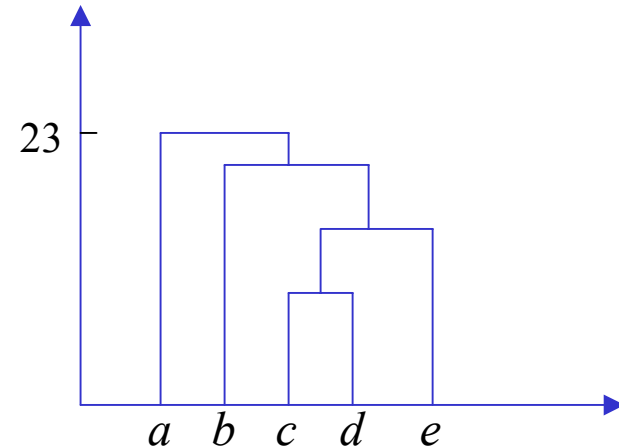
Tableau des dissimilarités

$$G_2 = \{e, G_1\} \Rightarrow$$

	<i>a</i>	<i>b</i>	$G_2$
<i>a</i>	0	23	35
<i>b</i>	23	0	21
$G_2$	35	21	0

$$G_3 = \{b, G_2\} \Rightarrow$$

	<i>a</i>	$G_3$
<i>a</i>	0	23
$G_3$	23	0



# Exemple (suite)

*Agrégation selon le lien maximum*

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	0	23	35	43	50
<i>b</i>	23	0	21	32	45
<i>c</i>	35	21	0	11	25
<i>d</i>	43	32	11	0	17
<i>e</i>	50	45	25	17	0

Tableau des dissimilarités

$$G_1 = \{c, d\}$$

⇒

	<i>a</i>	<i>b</i>	<i>e</i>	$G_1$
<i>a</i>	0	23	50	43
<i>b</i>	23	0	45	32
<i>e</i>	50	45	0	25
$G_1$	43	32	25	0

$$G_2 = \{a, b\}$$

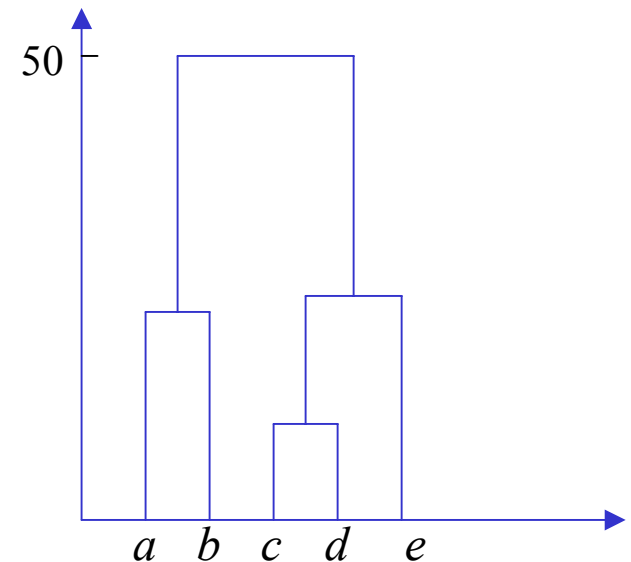
⇒

	<i>e</i>	$G_1$	$G_2$
<i>e</i>	0	25	50
$G_1$	25	0	43
$G_2$	50	43	0

$$G_3 = \{e, G_1\}$$

⇒

	$G_2$	$G_3$
$G_2$	0	50
$G_3$	50	0



# Classification dans un espace euclidien

## Inerties interclasse et intraclasse

Soit une classification en  $k$  groupes d'effectifs  $n_1, \dots, n_k$ , les individus étant des points d'un espace euclidien. Notons les groupes  $G_1, \dots, G_k$ , et  $g_1, \dots, g_k$  leurs centres de gravité ( $g$  est le centre de gravité du nuage).

$$\textit{Inertie totale} : \quad I_{tot} = \frac{1}{n} \sum_{i=1}^n d^2(e_i, g)$$

$$\textit{Inertie interclasse} : \quad I_{inter} = \frac{1}{n} \sum_{i=1}^k n_i \cdot d^2(g_i, g)$$

$$\textit{Inertie intraclasse} : \quad I_{intra} = \frac{1}{n} \sum_{i=1}^k \sum_{e \in G_i} d^2(e, g_i)$$

# Critère d'agrégation selon l'inertie

**Théorème de Huygens :**

**Inertie totale = Inertie inter-classe + Inertie intra-classe**

Au fur et à mesure que les regroupements sont effectués, l'inertie intra-classe augmente et l'inertie interclasse diminue, car leur somme est une constante liée aux données analysées.

# La méthode de Ward

Lorsqu'on remplace deux classes  $A$  et  $B$  par leur réunion, on montre que la diminution de l'inertie interclasse (et donc l'augmentation de l'inertie intraclasse) est égale à :

$$\frac{n_A \cdot n_B}{n \cdot (n_A + n_B)} \cdot d^2(g_A, g_B)$$

**La méthode de Ward consiste à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'inertie intraclasse, utilisée comme *indice de niveau*, soit minimum.**

# Intérêt de la méthode de Ward

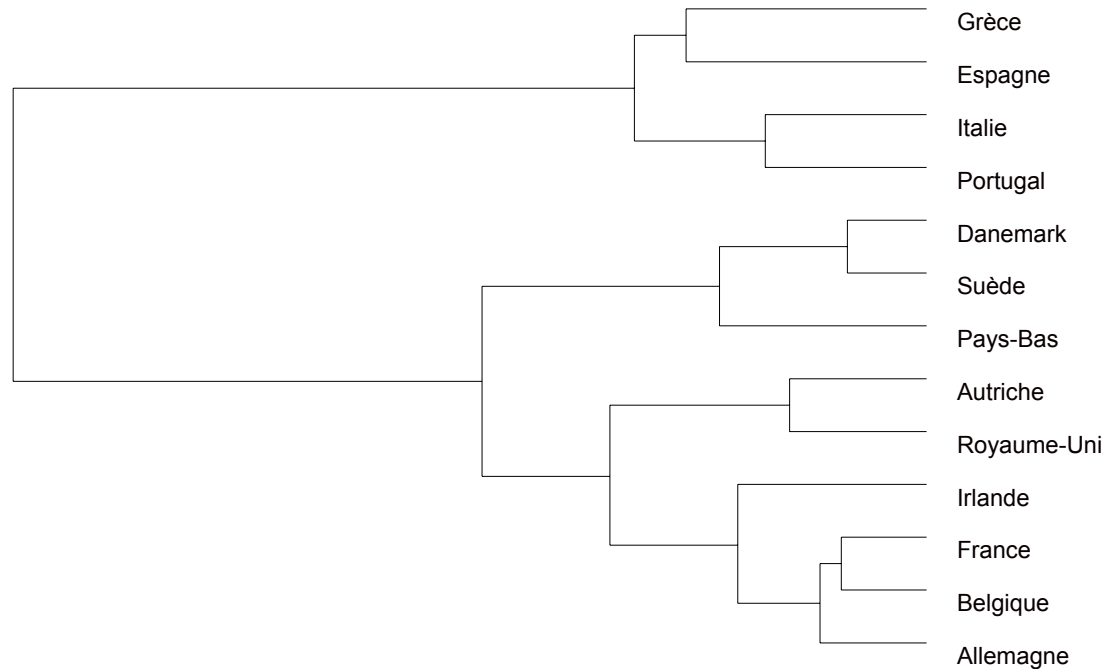
L'agrégation selon le lien minimum a l'inconvénient d'induire des « effets de chaîne » (les objets s'agrègent un par un au groupe déjà constitué), mais déforme peu si on reconstitue les dissimilarités à partir de l'arbre.

L'agrégation selon le lien maximum a, par contre, l'inconvénient de déformer beaucoup.

La méthode de Ward, aisée à mettre en œuvre lorsque la classification est effectuée après une analyse factorielle (les objets à classer étant repérés par leurs coordonnées sur les premiers axes factoriels), constitue une très bonne méthode de classification ascendante hiérarchique sur données euclidiennes.

# Exemple d'arbre hiérarchique (ou dendrogramme)

Classification hiérarchique directe



L'utilisateur peut choisir de former une partition avec un nombre de classes arbitraire : il coupe l'arbre pour obtenir des classes les plus homogènes possibles.

# Aides à l'interprétation d'une partition

Une partition est considérablement enrichie par une description des classes à l'aide des individus et des variables.

# Interprétation par les individus

Pour chaque classe, on examine :

- son effectif,
- son diamètre (distance entre les 2 points les plus éloignés),
- la séparation (distance minimum entre la classe considérée et la classe la plus proche) et le numéro de la classe la plus proche,
- les identités des individus les plus proches du centre de gravité de la classe ou « parangons »,
- les identités des l'individus les plus éloignés du centre de gravité de la classe ou « extrêmes ».

# Interprétation par les variables : une par une

On calcule un critère mesurant la pertinence de chaque variable de façon isolée pour interpréter la classe.

Exemple :

- prix ..... **critère fort** pour cette classe
- âge : [18 ; 25 ans] ..... **critère faible** pour cette classe

Est-ce que tous les éléments de la classe ont certaine(s) valeur(s) de cette variable (condition nécessaire d'appartenance à la classe) ?

Est-ce que certaine(s) valeur(s) de cette variable ne se rencontrent que dans cette classe (condition suffisante d'appartenance à la classe) ? ...

# Interprétation par des groupes de variables

Méthode de Marquage de données qui lie par des conjonctions des plages de valeurs de diverses variables entre elles, caractéristiques de la classe.

Exemple :

Âge : [18 ; 25 ans] **ET** distribution : « grande surface » **ET** achat : VTT

On calcule de même un critère mesurant la pertinence de chaque groupe de variables pour interpréter la classe (Gettler-Summa, 2000).

# Interprétation par les variables continues

Comparaison de la moyenne  $\bar{x}_k$  et de l'écart-type  $s_k$  d'une variable  $X$  dans la classe  $k$  à la moyenne générale et à l'écart-type général.

# Interprétation par les variables nominales

	Classe $k$	Autres classes	Population
Modalité $j$	$n_{kj}$	*	$n_j$
Autres modalités	*	*	*
Population	$n_k$	*	$n$

Pourcentage global  $\Rightarrow n_j / n$

Pourcentage « mod/clas »  $\Rightarrow n_{kj} / n_k$

Pourcentage « cla / mod »  $\Rightarrow n_{kj} / n_j$

# Un exemple de critère : la valeur-test

**Ces statistiques sur les variables peuvent être converties en un critère appelé « valeur-test ».**

La valeur-test permet de sélectionner les variables continues ou les modalités des variables nominales les plus caractéristiques de chaque classe.

C'est un critère de pertinence qui s'applique **aussi bien** dans l'interprétation d'une classe :

- par chaque variable une par une,
- que par Marquage, pour un groupe de variables.

# Valeur-test pour les variables continues

La valeur-test est égale à l'écart entre la moyenne dans la classe et la moyenne générale exprimée en nombre d'écart-types :

$$\text{v-test} = \frac{\overline{x_k} - \overline{x}}{s_k(X)}$$

$$\text{avec : } s_k^2(X) = \frac{n - n_k}{n - 1} \cdot \frac{s^2(X)}{n_k}$$

# Valeur-test pour les variables nominales

Valeur-test de la modalité  $j$  dans la classe  $k$  :

$$\text{v-test} = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n - 1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

# Interprétation de la valeur-test

**Si  $|v\text{-test}| > 2$ , la moyenne ou la proportion dans la population globale diffère significativement de celle dans la classe.**

Cette interprétation n'a de sens que pour les variables supplémentaires n'ayant pas participé à la construction des classes : il n'y a pas d'indépendance entre les classes d'une partition et une des variables ayant servi à définir la partition.

Pour les variables actives, les valeurs-test constituent de simples mesures de similarité entre variables et classes.

# Pratique de la classification

Pour une classification ascendante hiérarchique, on coupe l'arbre hiérarchique de façon à avoir des classes les plus homogènes possibles tout en étant bien séparées entre elles en se référant à l'histogramme des indices de niveau (cf. exemple).

La stratégie « Analyse factorielle + Classification » permet d'éliminer les fluctuations aléatoires et d'obtenir des classes plus stables, les axes factoriels étant très stables relativement à l'échantillonnage.

# Pratique de la classification (suite)

La stratégie « Classification mixte », consistant à pratiquer une classification ascendante hiérarchique sur quelques dizaines de groupes homogènes obtenus par un algorithme d'agrégation autour de centres mobiles, est bien adaptée au partitionnement d'un ensemble comprenant un grand nombre d'individus (des milliers, voire des dizaines de milliers).

L'homogénéité des classes obtenues peut être optimisée par une procédure de consolidation qui consiste à utiliser de nouveau une procédure d'agrégation autour des centres mobiles.

# Pratique de la classification (suite)

La méthode de Ward s'allie efficacement avec les constructions de partition du type « Réaffectation / Recentrage » en fournissant une partition initiale de bonne qualité. L'exigence de variables quantitatives pour cette méthode peut être satisfaite grâce à un traitement préalable par analyse factorielle.

# Conclusion

**Complémentarité entre analyse factorielle et classification :**

**la classification (dans l'espace entier) permet de « voir » au-delà du plan factoriel.**

# Intérêts de la classification

- Les classes obtenues assurent une vue concise et structurée des données.
- Des regroupements inattendus apparaissent.
- Des regroupements attendus n'existent pas.
- Les classes significatives entraînent la définition de fonctions de décision permettant d'attribuer un nouvel individu à la classe dont il est le plus proche.

# Classification sur variables

La classification sur individus, afin de les regrouper en un nombre restreint de classes représentatives, est la plus utilisée,

mais on peut aussi faire, après avoir transposé le fichier de données, une classification sur variables afin de réduire leur nombre et éventuellement, étudier leurs redondances.

# Classification sur les données utilisées pour l'exemple traité en ACP

## Economie et Statistique n°332-333, 2000, Insee

La classification a été réalisée à l'aide du logiciel SPAD sur les 10 premiers facteurs de l'ACP qui rendent compte de 98,9% de l'inertie totale.

CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES)

SUR LES 10 PREMIERS AXES FACTORIELS

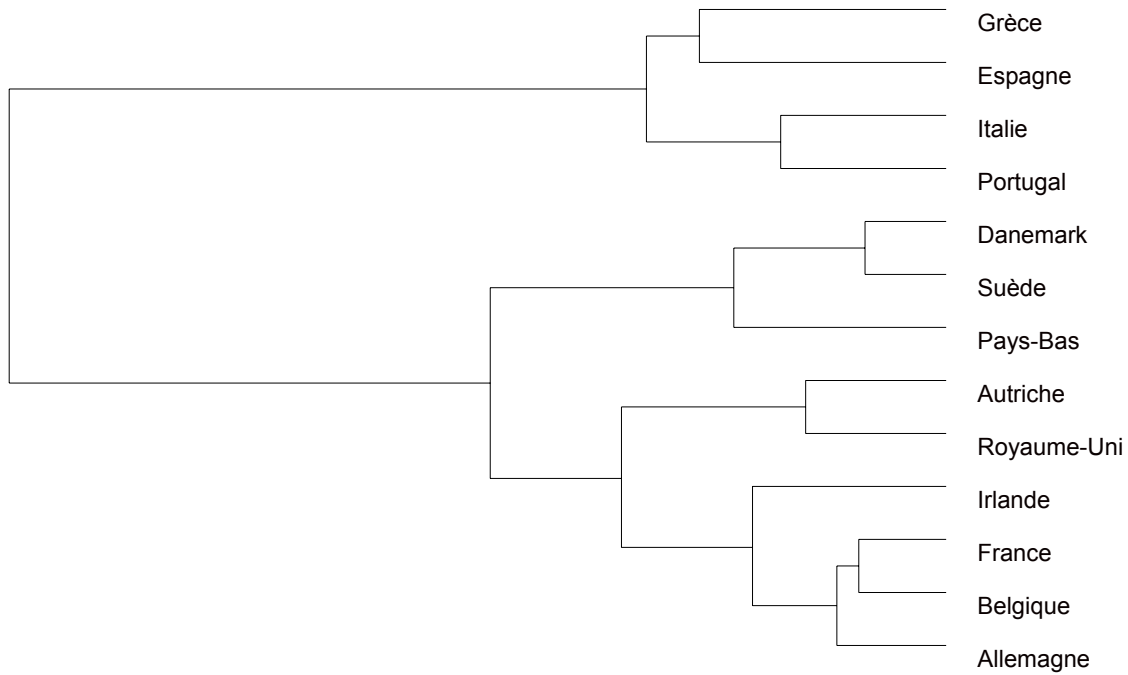
DESCRIPTION DES NOEUDS

NUM.	AIN	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
14	4	13	2	2.00	0.52631	*****
15	6	3	2	2.00	0.57602	*****
16	15	1	3	3.00	0.71932	*****
17	2	12	2	2.00	0.93005	*****
18	9	11	2	2.00	1.10279	*****
19	8	16	4	4.00	1.30177	*****
20	14	10	3	3.00	1.41883	*****
21	7	5	2	2.00	1.64450	*****
22	21	18	4	4.00	2.00764	*****
23	17	19	6	6.00	2.18091	*****
24	20	23	9	9.00	3.04725	*****
25	22	24	13	13.00	6.30253	*****
SOMME DES INDICES DE NIVEAU =						21.75792

D'après l'histogramme des indices de niveau, on peut envisager 2, 3 ou 5 classes.

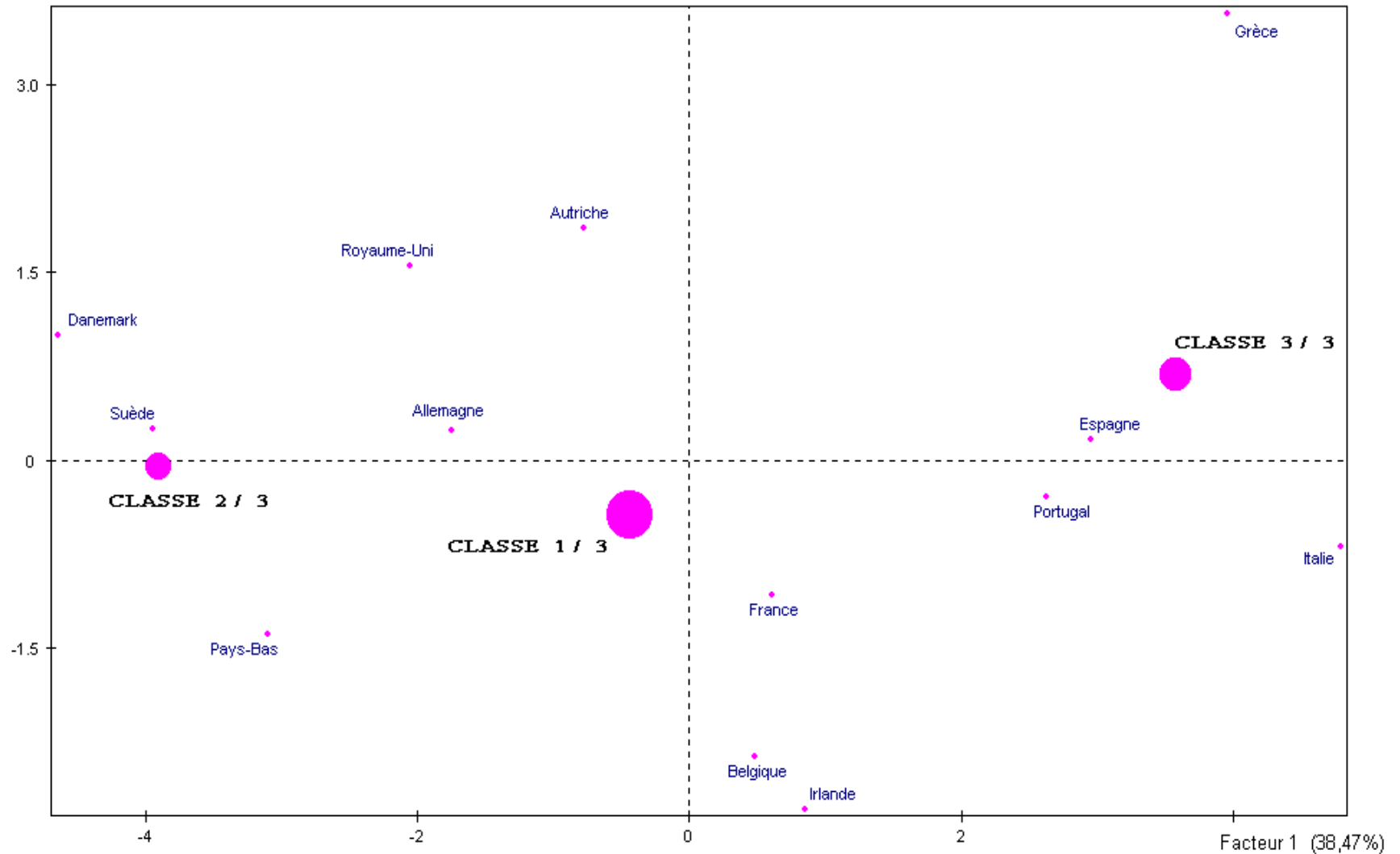
# Dendrogramme

Classification hiérarchique directe



# Représentation de la partition en 3 classes dans le 1<sup>er</sup> plan principal de l'ACP

Facteur 2 (12,82%)



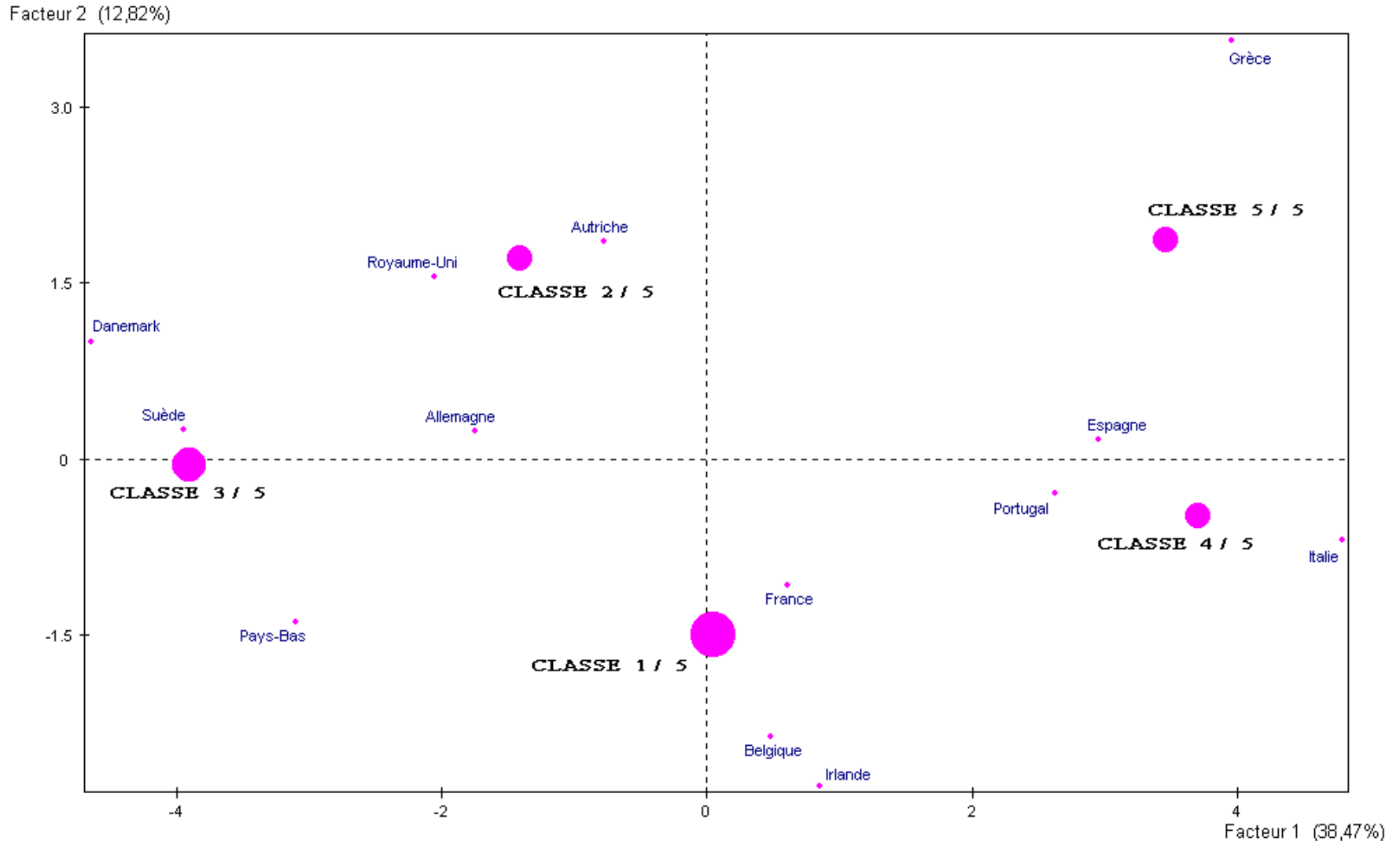
# Description de la partition en 3 classes

{Autriche, Royaume-Uni, France, Irlande, Belgique, Allemagne}  
 {Danemark, Suède, Pays-Bas}      {Grèce, Espagne, Italie, Portugal}

Caractérisation par les variables continues des classes de la partition						
CLASSE 2 / 3 (Poids = 3.00 Effectif = 3)						
Variables caractéristiques	Moyenne dans la classe	Moyenne générale	Ecart-type dans la classe	Ecart-type général	Valeur-Test	Probabilité
TP	27,200	15,785	6,632	8,512	2,54	0,005
DEP	4,970	3,106	0,638	1,464	2,42	0,008
Hres	33,467	38,231	0,899	3,052	-2,96	0,002
CLASSE 3 / 3 (Poids = 4.00 Effectif = 4)						
Variables caractéristiques	Moyenne dans la classe	Moyenne générale	Ecart-type dans la classe	Ecart-type général	Valeur-Test	Probabilité
same	38,850	28,615	4,053	7,765	3,04	0,001
<b>infl</b>	8,525	4,831	3,798	3,346	2,55	0,005
<b>mDP</b>	7,838	5,046	2,777	2,624	2,46	0,007
TP	7,075	15,785	1,266	8,512	-2,36	0,009
<b>fé95</b>	1,268	1,507	0,097	0,229	-2,41	0,008
<b>mDE</b>	-6,475	0,158	5,395	6,207	-2,47	0,007
<b>seul</b>	17,975	24,982	4,391	6,569	-2,55	0,005
<b>div</b>	0,900	1,975	0,274	0,863	-2,92	0,002
2nd	31,375	55,508	8,287	19,000	-2,93	0,002
sala	67,975	80,585	8,018	9,923	-2,93	0,002

Aucune variable ne caractérise la classe 1. Les variables illustratives sont en gras.

# Représentation de la partition en 5 classes dans le 1<sup>er</sup> plan principal de l'ACP



# Description de la partition en 5 classes

{ France, Belgique, Allemagne, Irlande}    {Autriche, Royaume-Uni}  
 {Danemark, Suède, Pays-Bas}    { Italie, Portugal}    {Grèce, Espagne}

<b>Caractérisation par les variables continues des classes de la partition</b>						
<b>CLASSE 3 / 5 (Poids = 3.00 Effectif = 3 )</b>						
<b>Variables caractéristiques</b>	<b>Moyenne dans la classe</b>	<b>Moyenne générale</b>	<b>Ecart-type dans la classe</b>	<b>Ecart-type général</b>	<b>Valeur-Test</b>	<b>Probabilité</b>
TP	27,200	15,785	6,632	8,512	2,54	0,005
DEP	4,970	3,106	0,638	1,464	2,42	0,008
Hres	33,467	38,231	0,899	3,052	-2,96	0,002
<b>CLASSE 4 / 5 (Poids = 2.00 Effectif = 2 )</b>						
<b>Variables caractéristiques</b>	<b>Moyenne dans la classe</b>	<b>Moyenne générale</b>	<b>Ecart-type dans la classe</b>	<b>Ecart-type général</b>	<b>Valeur-Test</b>	<b>Probabilité</b>
Djeu	0,380	0,139	0,040	0,125	2,84	0,002
<b>CLASSE 5 / 5 (Poids = 2.00 Effectif = 2 )</b>						
<b>Variables caractéristiques</b>	<b>Moyenne dans la classe</b>	<b>Moyenne générale</b>	<b>Ecart-type dans la classe</b>	<b>Ecart-type général</b>	<b>Valeur-Test</b>	<b>Probabilité</b>
CDD	22,300	11,597	11,300	6,666	2,37	0,009
infl	10,100	4,831	4,600	3,346	2,33	0,010
sala	64,550	80,585	10,250	9,923	-2,39	0,008

Aucune variable ne caractérise les classes 1 et 2.

# Comparaison de partitions obtenues sur le même ensemble d'individus

Une classification des 13 pays peut aussi être effectuée en utilisant les descripteurs d'un autre groupe de variables.

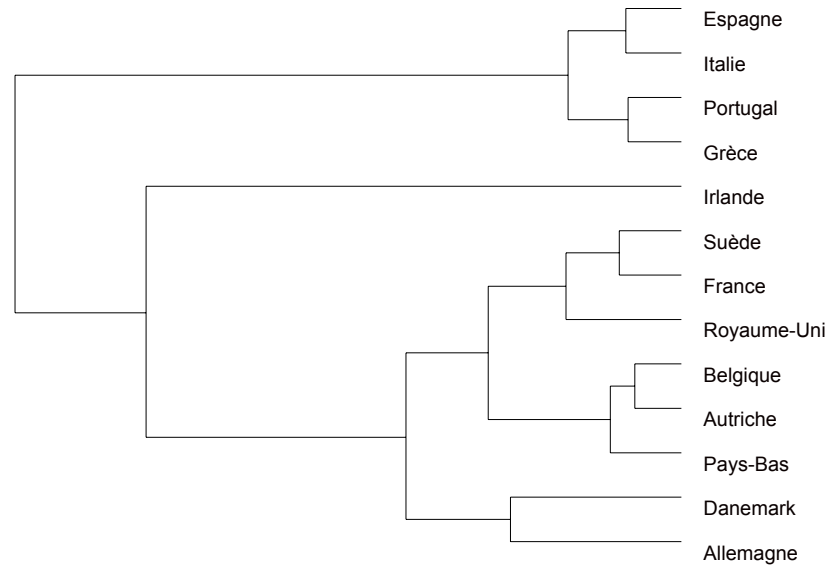
Une partition obtenue avec un groupe de variables peut être archivée et utilisée en variable supplémentaire pour expliquer des partitions opérées avec un autre groupe, ce qui peut mettre en évidence des liaisons entre groupes de variables.

On peut aussi construire le tableau de contingence qui croise deux partitions obtenues avec deux groupes de variables.

Le logiciel SPAD permet d'enregistrer des partitions : « Archivages, exportations », puis « Archivages des coordonnées factorielles et partitions ».

# Dendrogramme de la classification sur les variables du 1<sup>er</sup> groupe « Structures familiales et démographie »

Classification hierarchique directe



Trois partitions peuvent être envisagées : 3 classes, 4 classes, 6 classes.

# Représentation de la partition en 3 classes

avec projection de la partition en 3 classes obtenue avec le groupe « Marché du travail »

