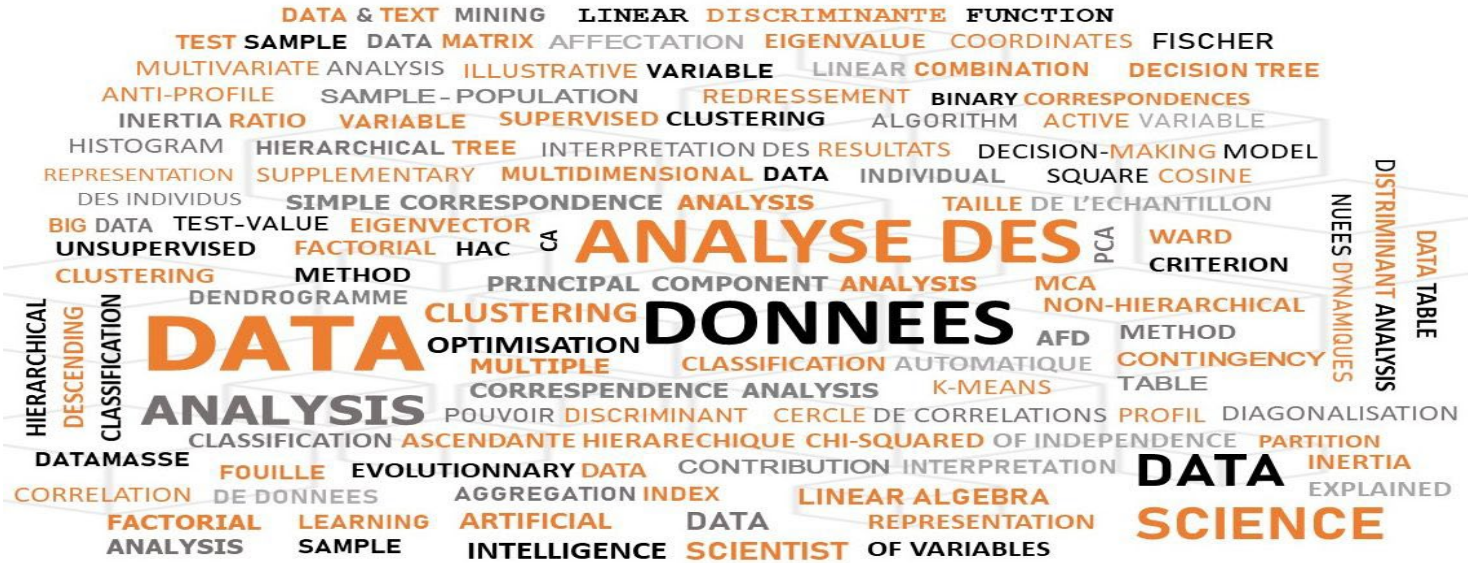
UFR de sciences économiques et de gestion

UNIVERSITÉ DE LYON

M2 – APE – Economic and Statistical Studies



Data Science - Data Analysis
Clustering methods

University year 2025-2026

R. Abdesselam

Courriel ; rafik.abdesselam@univ-lyon2.fr

http://perso.univ-lyon2.fr/~rabdesse/Documents/

COHERIS ANALYTICS SPAD

UFR de Sciences économiques et de Gestion

Université de Lyon

M2 – APE – Chargé d'Etudes Economiques

Science des Données - Analyse des données

Polycopié 2 : Méthodes de classification

Année Universitaire 2025-2026

R. Abdesselam

Courriel ; rafik.abdesselam@univ-lyon2.fr

http://perso.univ-lyon2.fr/~rabdesse/Documents/

COHERIS ANALYTICS SPAD

Objectifs des techniques de classification

- *La classification automatique (clustering) est une méthode mathématique d'analyse de données : pour décrire une population importante d'individus (objets, personnes, entreprises, malades, pays, etc…), on les regroupe en plusieurs classes de telle sorte que les individus d'une même classe soient le plus semblables possible et que les classes soient le plus distinctes possibles.*
- *La classification est une branche de l'analyse des données qui a donné lieu à de nombreuses publications. Suivant le domaine d'application, on la rencontre aussi sous le nom de typologie, de segmentation (en particulier dans le monde du marketing).*
- *Soit à partitionner (classer, regrouper) un ensemble n individus (objets) caractérisés par p variables (attributs).*
- *Les techniques de classification automatique :*
  - *Méthode des nuées dynamiques (ou k-means) (classification directe, non hiérachique)*
  - *Classification Ascendante Hiérarchique (CAH)*

  ont pour objectif de structurer les individus (objets) : faire des regroupements les moins arbitraires possibles des individus à partir de leurs caractères de description.
- *Le but est de découvrir des structures cachées de l'ensemble des individus.*
- *Ces méthodes visent à mettre en évidence des groupes d'individus aussi homogènes que possible, c'est-à-dire que les individus soient très ressemblants entre eux, tandis que deux individus appartenant à des groupes différents doivent être très dissemblants.*
- *Les classes de la classification regroupent des individus ayant des caractéristiques (variables) similaires et séparent les individus ayant des caractéristiques différentes : homogénéité interne et hétérogénéité externe.*
- *Faire une analyse factorielle avant l'analyse de classification pour réduire le nombre de variables et avoir des variables le plus indépendantes possible.*
- *Enfin, ces groupes d'individus sont obtenus au moyen d'algorithmes formalisés.*

Classification par nuées dynamiques

- *La segmentation par nuées dynamiques (ou k-means) est une méthode de classification automatique qui a pour objectif de partitionner l'espace des individus IR^p, muni d'une distance euclidienne appropriée notée M, en un nombre k (fixé - connu) de classes.*

- *A partir d'une partition initiale, on améliore itérativement la partition de l'espace en minimisant la variance et en maximisant l'écart entre les classes.*

- *La solution proposée par cet algorithme dépend de la partition initiale.*

- *Les deux principales étapes de la méthode :*

  - une étape de recentrage : l'individu moyen de chaque groupe (centre de classe) est redéfini en fonction des individus qui lui sont affectés,

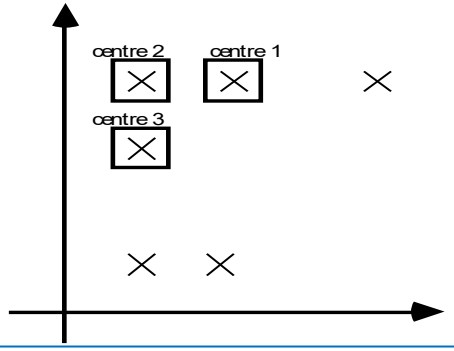
  -une étape d'affectation : chaque individu est placé dans le groupe dont le centre de classe est le plus proche.

- *L'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition soit lorsqu'un critère (la variance intra-classe) cesse de décroître de façon sensible soit encore parce qu'un nombre maximal d'itérations a été fixé a priori.*

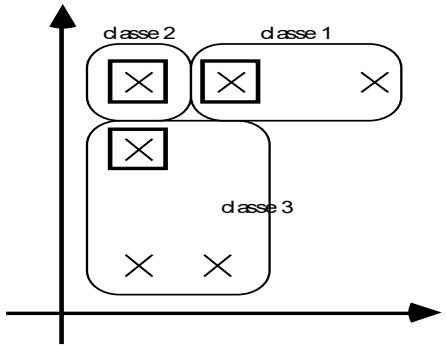
- *Dans tous les cas la partition obtenue dépend du choix initial des centres à l'étape 0*

Etapes de recentrage et d'affectation

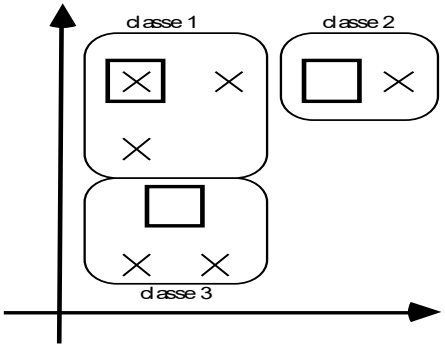
Etape 0 : Initialisation des centres de classes



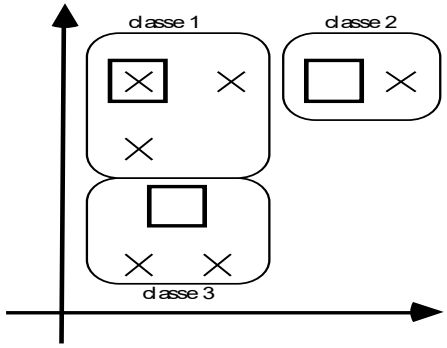
Etape 1 : Association de chaque individu à un centre



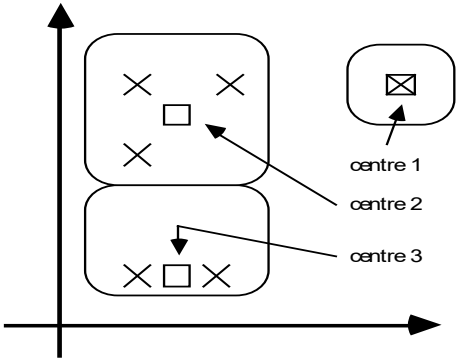
Etape 2 : Calcul de nouveaux centres



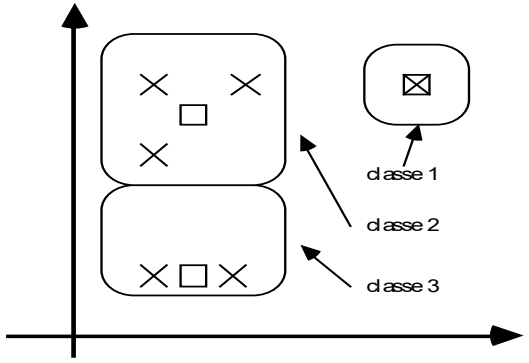
Etape 3 : Réaffectation de chaque individu à un centre



Etape 4 : Calcul de nouveaux centres

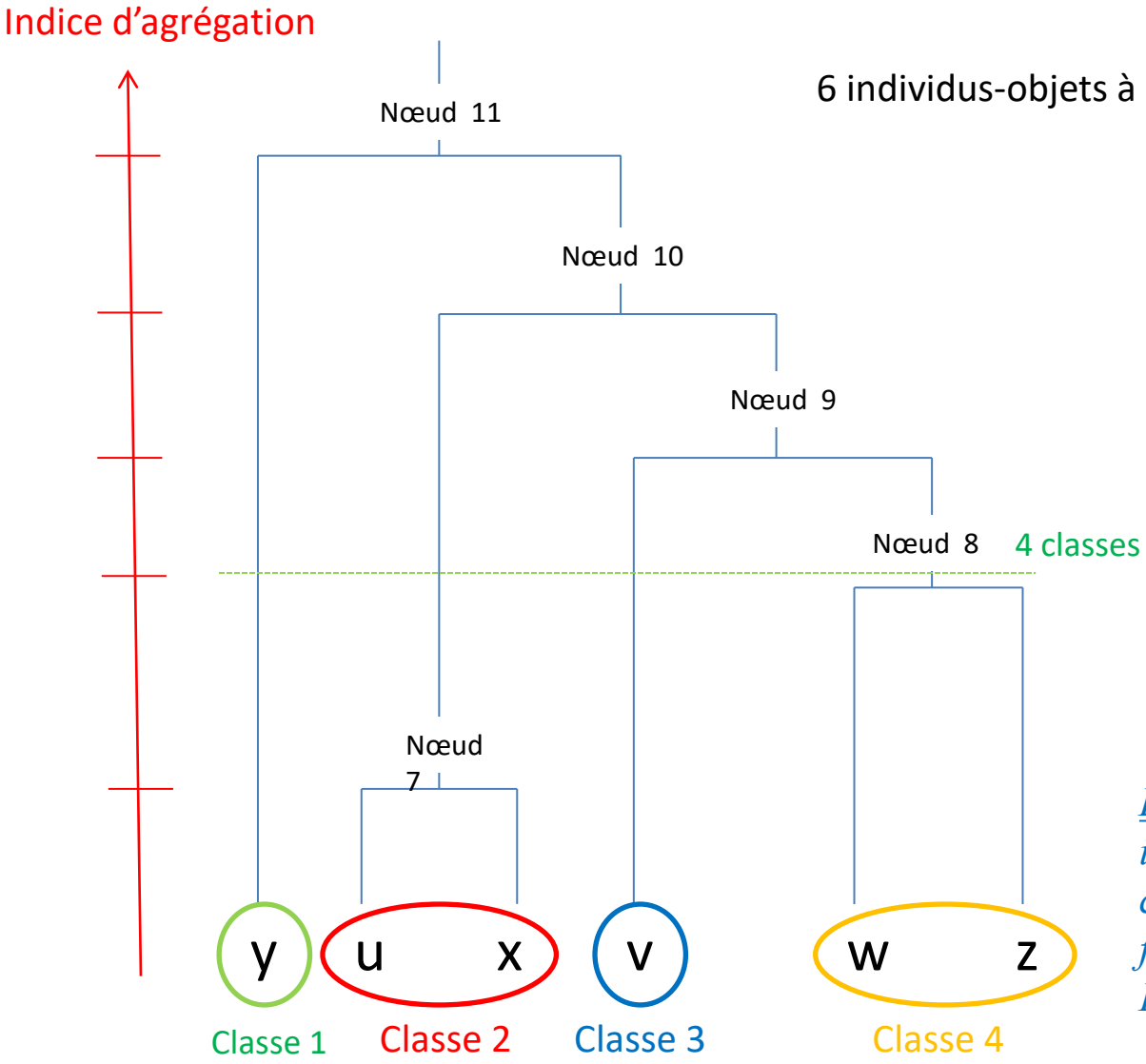


Etape 5 : Réaffectation de chaque individu à un centre



CAH : Classification Ascendante Hiérarchique

Arbre hiérarchique - Dendrogramme



6 individus-objets à classer {u, v, w, x, y, z}

Le critère de Ward :
il consiste à réunir les deux classes dont le regroupement fera le moins baisser l'inertie Interclasse.

Exemple d'application

Dictionnaire des variables - UE-28 en 2014

---

1 . Etat membre (3 MODALITES)
AUTR - Autre ZNEU - Zone Non Euro ZEUR - Zone Euro

2 . Population (%) UE-2000 (CONTINUE)
POPU - Population (%) UE-28

3 . Produit Intérieur Brut (%) UE-28 (CONTINUE)
PIB - Produit Intérieur Brut

Thème : Finances publiques de l'UE-28 en 2014

---

4 . Recettes totales des APU en % du PIB (CONTINUE)
RECE - Recettes totales des

5 . Dépenses totales des APU en % du PIB (CONTINUE)
DEPE - Dépenses totales des

6 . Solde des finances publiques en % du PIB (CONTINUE)
SDFP - Solde des finances p

7 . Dette brute des APU en % du PIB (CONTINUE)
DETB - Dette brute des APU

Thème : Activité-Emploi

---

8 . Taux de chômage en % population active (CONTINUE)
TCHO - Taux de chômage en %

9 . Ventes au détail en volume (%) (CONTINUE)
VENT - Ventes au détail en

10 . Production industrielle hors bâtiment (CONTINUE)
PIND - Production industrie

11 . Taux de croissance en volume du PIB (CONTINUE)
TCRO - Taux de croissance

Dictionnaire des variables

SELECTION DES INDIVIDUS ET DES VARIABLES UTILES

VARIABLES NOMINALES ILLUSTRATIVES

1 VARIABLES 3 MODALITES ASSOCIEES

---

1 . Etat membre (3 MODALITES)

VARIABLES CONTINUES ACTIVES : 4 VARIABLES

---

4 . Recettes totales des APU en % du PIB (CONTINUE)
5 . Dépenses totales des APU en % du PIB (CONTINUE)
6 . Solde des finances publiques en % du PIB (CONTINUE)
7 . Dette brute des APU en % du PIB (CONTINUE)

VARIABLES CONTINUES ILLUSTRATIVES : 4 VARIABLES

---

8 . Taux de chômage en % population active (CONTINUE)
9 . Ventes au détail en volume (%) (CONTINUE)
10 . Production industrielle hors bâtiment (CONTINUE)
11 . Taux de croissance en volume du PIB (CONTINUE)

INDIVIDUS

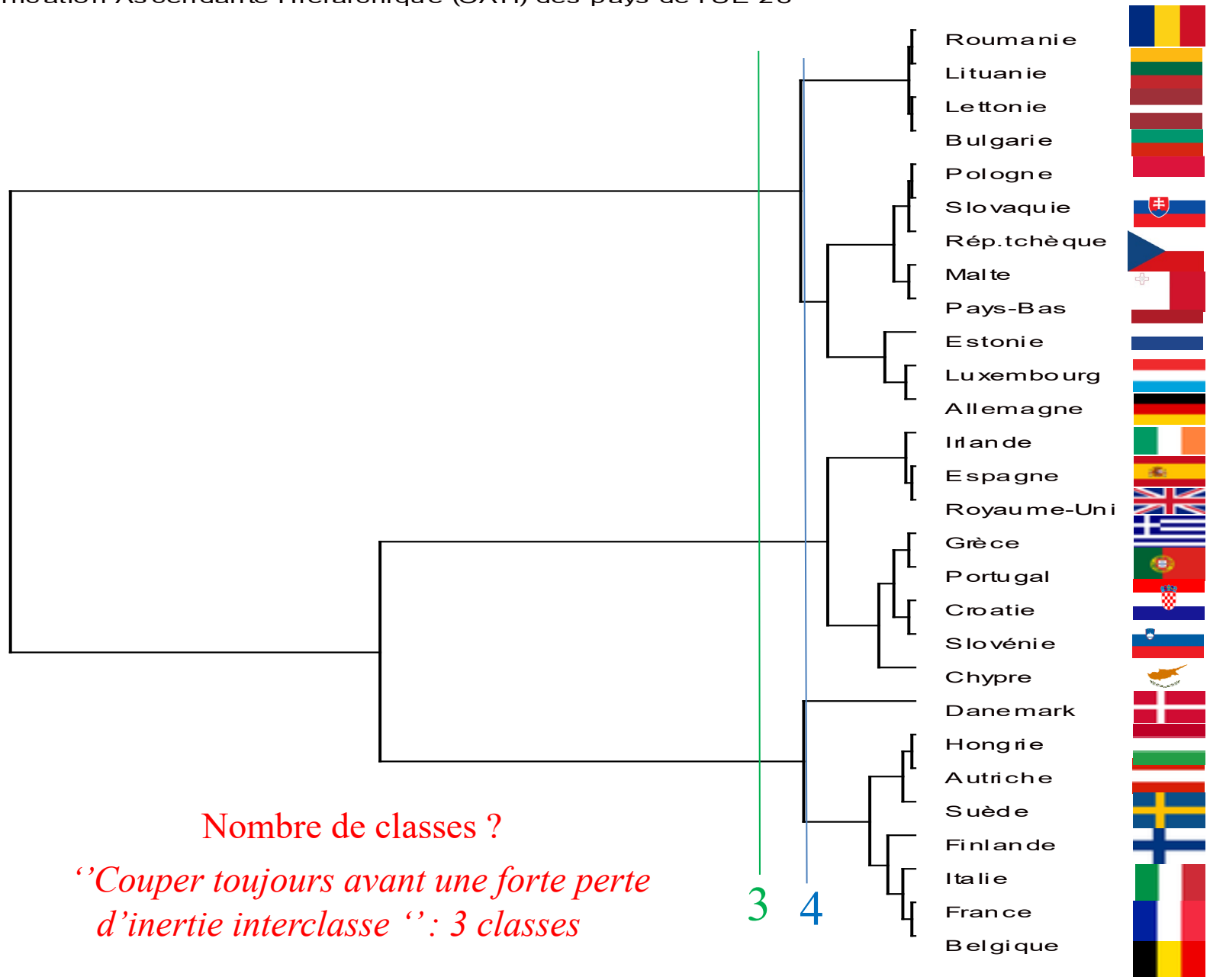
---------- NOMBRE ---------- POIDS ----------

POIDS DES INDIVIDUS: Poids des individus, uniforme egal a 1. UNIF
RETENUS ..... NITOT = 32 PITOT = 32.000
SELECTION APRES FILTRAGE
ACTIFS ..... NIACT = 28 PIACT = 28 .000
SUPPLEMENTAIRES .... NISUP = 4 PISUP = 4.000

Classification Ascendante Hiérarchique

Principaux indicateurs économiques et financiers de UE-28 : 2014

Classification Ascendante Hiérarchique (CAH) des pays de l'UE-28



Nombre de classes ?

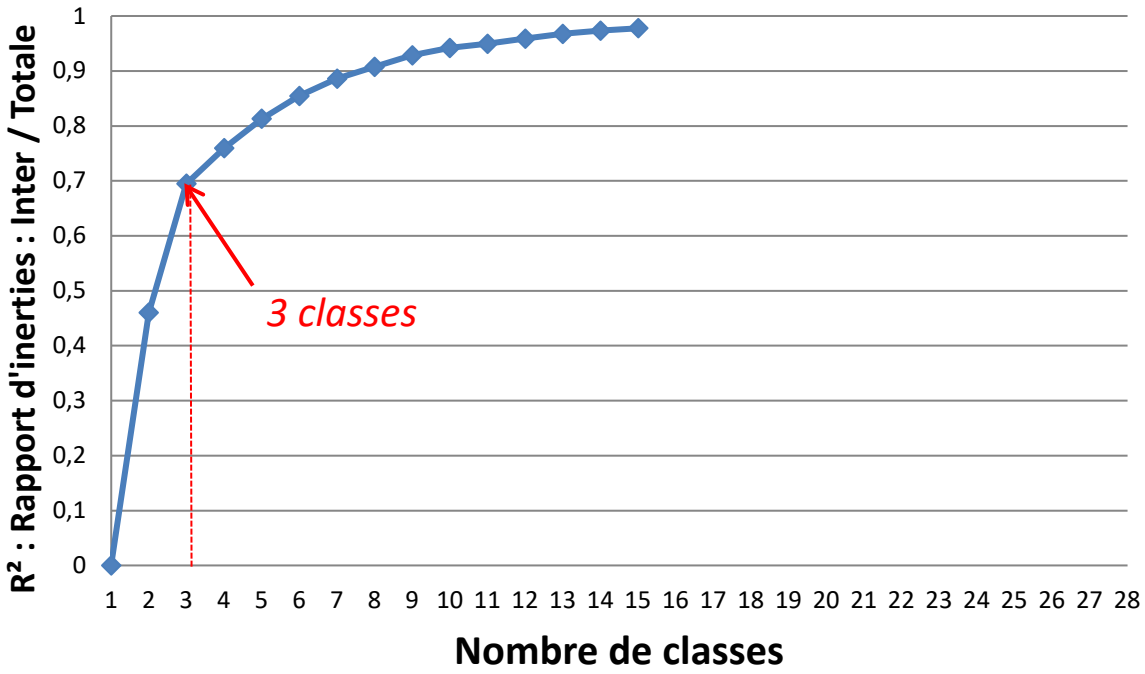
''Couper toujours avant une forte perte d'inertie interclasse '' : 3 classes

Classification Ascendante Hiérarchique

Mesure de la qualité & Choix du nombre de classes

$R^2 = I_{Inter} / I_{Totale}$: Rapport d'inertie expliquée par les classes

- *doit être le plus proche possible de 1 sans avoir trop de classes*
- *s'arrêter après le dernier saut important*



Classes	R^2	Ecart	F
1	0		
2	0,4601	0,4601	22,1571
3	0,69487	0,23477	28,4661
4	0,75933	0,06446	25,2405
5	0,81316	0,05383	25,0250
6	0,85478	0,04162	25,8989
7	0,88615	0,03137	27,2422
8	0,90776	0,02161	28,1180
9	0,9287	0,02094	30,9350
10	0,94195	0,01325	32,4531
11	0,9497	0,00775	32,0972
12	0,95884	0,00914	33,8843
13	0,96781	0,00897	37,5819
14	0,97341	0,0056	39,4241
15	0,97803	0,00462	41,3369
28	1		

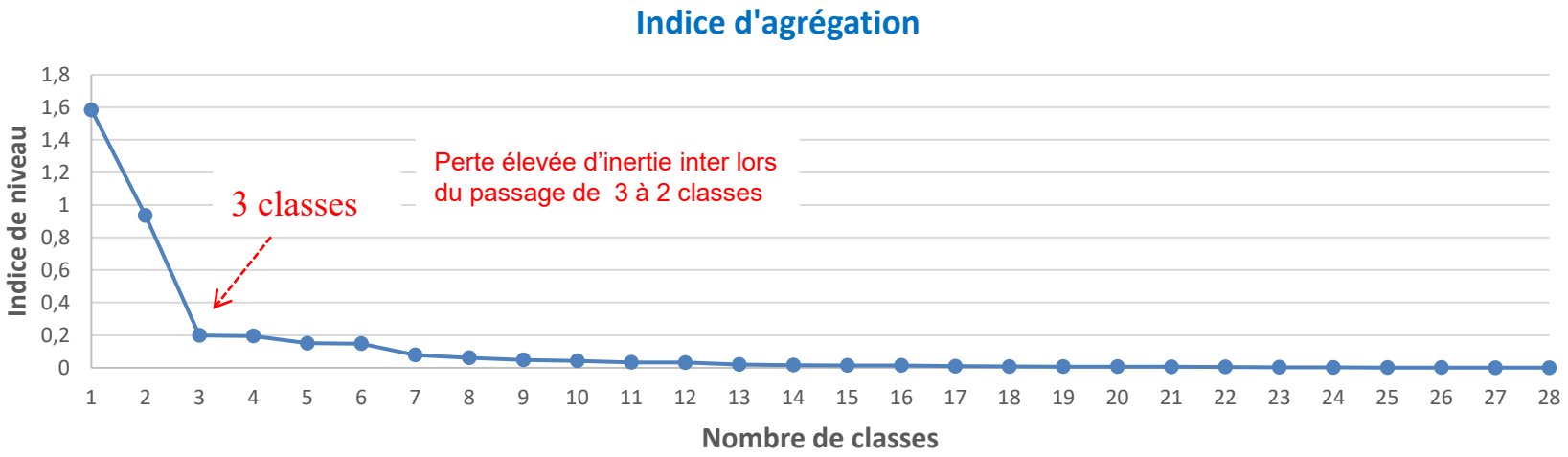
Pseudo F : mesure la séparation entre toutes les classes

$$F = [R^2 / (k - 1)] / [(1 - R^2)/(n - k)]$$

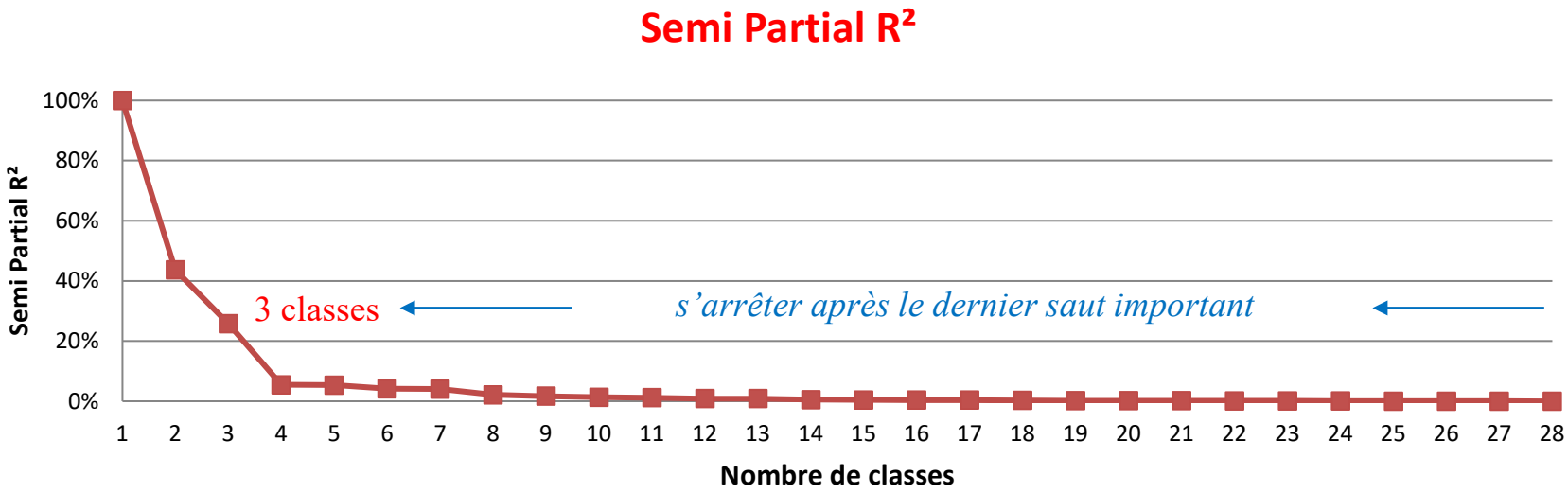
- *n = 28 : nombre d'observations-Pays et k : le nombre de classes*
- *F doit être grand*

Classification Ascendante Hiérarchique

Mesure de la qualité & Choix du nombre de classes



$$\text{Semi Partial } R^2 = \Delta I_{Inter} / I_{Totale} = \text{ Indice d'agrégation} / I_{totale}$$



Classification Ascendante Hiérarchique

CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES) SUR LES 2 PREMIERS AXES FACTORIELS

DESCRIPTION DES NOEUDS

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
29	21	25	2	2.00	0.00011	*
30	9	28	2	2.00	0.00102	*
31	23	15	2	2.00	0.00165	*
32	14	2	2	2.00	0.00215	*
33	10	1	2	2.00	0.00356	*
34	17	20	2	2.00	0.00488	*
35	18	19	2	2.00	0.00582	*
36	11	24	2	2.00	0.00692	*
37	8	22	2	2.00	0.00703	*
38	29	3	3	3.00	0.00774	*
39	31	32	4	4.00	0.01003	*
40	16	5	2	2.00	0.01409	*
41	7	30	3	3.00	0.01470	*
42	12	33	3	3.00	0.01676	*
43	34	27	3	3.00	0.02032	**
44	37	36	4	4.00	0.03255	**
45	38	35	5	5.00	0.03313	**
46	26	42	4	4.00	0.04272	***
47	6	40	3	3.00	0.04805	***
48	44	13	5	5.00	0.06136	****
49	43	46	7	7.00	0.07839	****
50	45	47	8	8.00	0.14829	********
51	41	48	8	8.00	0.15096	********
52	4	49	8	8.00	0.19526	**********
53	39	50	12	12.00	0.19931	**********
54	51	52	16	16.00	0.93633	**
55	53	54	28	28.00	1.58410	***

SOMME DES INDICES DE NIVEAU = 3.62720

Classification Ascendante Hiérarchique

DESCRIPTION DES NOEUDS DE LA HIERACHIE
(INDICES EN POURCENTAGE DE LA SOMME DES INDICES : 3.62720)

NOEUD		SUCCESSEURS				COMPOSITION	
NUMERO	INDICE	AINE	BENJ	EFFECT.	POIDS	PREMIER	DERNIER
29	0.00	24	23	2	2.00	23	24
30	0.03	15	14	2	2.00	14	15
31	0.05	28	27	2	2.00	27	28
32	0.06	26	25	2	2.00	25	26
33	0.10	2	1	2	2.00	1	2
34	0.13	7	6	2	2.00	6	7
35	0.16	21	20	2	2.00	20	21
36	0.19	11	10	2	2.00	10	11
37	0.19	13	12	2	2.00	12	13
38	0.21	29	22	3	3.00	22	24
39	0.28	31	32	4	4.00	25	28
40	0.39	18	17	2	2.00	17	18
41	0.41	16	30	3	3.00	14	16
42	0.46	3	33	3	3.00	1	3
43	0.56	34	5	3	3.00	5	7
44	0.90	37	36	4	4.00	10	13
45	0.91	38	35	5	5.00	20	24
46	1.18	4	42	4	4.00	1	4
47	1.32	19	40	3	3.00	17	19
48	1.69	44	9	5	5.00	9	13
49	2.16	43	46	7	7.00	1	7
50	4.09	45	47	8	8.00	17	24
51	4.16	41	48	8	8.00	9	16
52	5.38	8	49	8	8.00	1	8
53	5.49	39	50	12	12.00	17	28
54	25.81	51	52	16	16.00	1	16
55	43.67	53	54	28	28.00	1	28

Classification Ascendante Hiérarchique

PARTITION PAR COUPURE D'UN ARBRE HIERARCHIQUE
Coupure 'a' de l'arbre en 3 classes
FORMATION DES CLASSES (INDIVIDUS ACTIFS)
DESCRIPTION SOMMAIRE

CLASSE	EFFECTIF	POIDS	CONTENU
aa1a	8	8.00	1 A 8
aa2a	8	8.00	9 A 16
aa3a	12	12.00	17 A 28

COORDONNEES ET VALEURS-TEST AVANT CONSOLIDATION AXES 1 A 2

CLASSES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	0	0	0	1	2	0	0	0	DISTO.
Coupure 'a' de l'arbre en 3 classes													
aa1a - Classe 1 / 3	8	8.00	3.4	-2.5	0.0	0.0	0.0	1.55	-0.92	0.00	0.00	0.00	3.23
aa2a - Classe 2 / 3	8	8.00	1.3	4.0	0.0	0.0	0.0	0.57	1.45	0.00	0.00	0.00	2.42
aa3a - Classe 3 / 3	12	12.00	-4.3	-1.3	0.0	0.0	0.0	-1.41	-0.35	0.00	0.00	0.00	2.11

CONSOLIDATION DE LA PARTITION AUTOUR DES 3 CENTRES DE CLASSES, REALISEE PAR 10 ITERATIONS
A CENTRES MOBILES PROGRESSION DE L'INERTIE INTER-CLASSES

ITERATION	I.TOTALE	I.INTER	QUOTIENT
0	3.62720	2.52043	0.69487
1	3.62720	2.52043	0.69487
2	3.62720	2.52043	0.69487

ARRET APRES L'ITERATION 2 L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES
PAR RAPPORT A L'ITERATION PRECEDENTE N'EST QUE DE 0.000 %.

Classification Ascendante Hiérarchique

DECOMPOSITION DE L'INERTIE CALCULEE SUR 2 AXES.

INERTIES	INERTIES AVANT	INERTIES APRES	EFFECTIFS AVANT	EFFECTIFS APRES	POIDS AVANT	POIDS APRES	DISTANCES AVANT	DISTANCES APRES
INTER-CLASSES	2.5204	2.5204						
INTRA-CLASSE								
CLASSE 1 / 3	0.3619	0.3619	8	8	8.00	8.00	3.2294	3.2294
CLASSE 2 / 3	0.2745	0.2745	8	8	8.00	8.00	2.4239	2.4239
CLASSE 3 / 3	0.4704	0.4704	12	12	12.00	12.00	2.1121	2.1121
TOTALE	3.6272	3.6272						

QUOTIENT (INERTIE INTER / INERTIE TOTALE) : AVANT ... 0.6949
APRES ... 0.6949

COORDONNEES ET VALEURS-TEST APRES CONSOLIDATION AXES 1 A 2

CLASSES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	0	0	0	1	2	0	0	0	DISTO.
Coupure 'a' de l'arbre en 3 classes													
aa1a - Classe 1 / 3	8	8.00	3.4	-2.5	0.0	0.0	0.0	1.55	-0.92	0.00	0.00	0.00	3.23
aa2a - Classe 2 / 3	8	8.00	1.3	4.0	0.0	0.0	0.0	0.57	1.45	0.00	0.00	0.00	2.42
aa3a - Classe 3 / 3	12	12.00	-4.3	-1.3	0.0	0.0	0.0	-1.41	-0.35	0.00	0.00	0.00	2.11

COMPOSITION DE : Coupure de l'arbre en 3 classes

Classe 1 / 3 : Belgique Danemark France Italie Hongrie Autriche Finlande Suède

Classe 2 / 3 : Irlande Grèce Espagne Croatie Chypre Portugal Slovénie Royaume-Uni

Classe 3 / 3 : Bulgarie Rép.Tchèque Allemagne Estonie Lettonie Lituanie Pays-bas Pologne Roumanie Slovaquie Luxembourg Malte

Classification Ascendante Hiérarchique

APPARTENANCE DE CHAQUE INDIVIDU : Coupure 'a' de l'arbre en 3 classes

Belgique : 1 Bulgarie : 3 Rép.Tchèque : 3 Danemark : 1 Allemagne : 3 Estonie : 3
Irlande : 2 Grèce : 2 Espagne : 2 France : 1 Croatie : 2 Italie : 1 Chypre : 2
Lettonie : 3 Lituanie : 3 Luxembourg : 3 Hongrie : 1 Malte : 3 Pays-bas : 3
Autriche : 1 Pologne : 3 Portugal : 2 Roumanie : 3 Slovénie : 2
Slovaquie : 3 Finlande : 1 Suède : 1 Royaume-Uni : 2

MATRICE DES DISTANCES ENTRE CLASSES

	1	2	3
1	0.000		
2	2.560	0.000	
3	3.008	2.677	0.000
	1	2	3

INDIVIDUS ILLUSTRATIFS AFFECTATION DANS LES CLASSES

CLASSE	EFFECTIF	POIDS
CLASSE 1	0	0.00
CLASSE 2	1	1.00
CLASSE 3	3	3.00

COMPOSITION : Coupure 'a' de l'arbre en 3 classes

Classe 1 / 3 :

Classe 2 / 3 : Serbie

Classe 3 / 3 : Monténégro Albanie Turquie

APPARTENANCE DE CHAQUE INDIVIDU :

Albanie : 3 Monténégro : 3 Serbie : 2 Turquie: 3

Classification Ascendante Hiérarchique

CARACTERISATION PAR LES CONTINUES DES CLASSES OU MODALITES DE Coupure 'a' de l'arbre en 3 classes

Classe 1 / 3

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		Classe 1 / 3	(POIDS = 8.00		EFFECTIF = 8)		aa1a
		Belgique Danemark France Italie Hongrie Autriche Finlande Suède					
4.26	0.000	51.88	43.44	3.47	6.51	6.Recettes Publiques	Rece
4.00	0.000	54.24	46.24	2.95	6.57	5.Dépenses Publiques	Dépe

Classe 2 / 3

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		Classe 2 / 3	(POIDS = 8.00		EFFECTIF = 8)		aa2a
		Irlande Grèce Espagne Croatie Chypre Portugal Slovénie Royaume-Uni					
3.24	**0.001**	**15.17**	**10.10**	**6.55**	**5.15**	**7.Taux Chômage**	**Taux**
3.21	0.001	109.69	73.61	29.50	36.97	4.Dette Publique	Dett
-3.82	0.000	-5.38	-2.83	1.51	2.20	3.Solde Public	Sold

Classe 3 / 3

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		Classe 3 / 3	(POIDS = 12.00		EFFECTIF = 12)		aa3a
		Bulgarie Rép.Tchèque Allemagne Estonie Lettonie Lituanie Pays-bas Pologne Roumanie Slovaquie Luxembourg Malte					
2.89	0.002	-1.42	-2.83	1.35	2.20	3.Solde Public	Sold
2.11	**0.018**	**2.14**	**1.47**	**0.77**	**1.43**	**8.Taux Croissance**	**Taux**
-1.92	**0.027**	**7.90**	**10.10**	**2.39**	**5.15**	**7.Taux Chômage**	**Taux**
-2.84	0.002	39.33	43.44	3.79	6.51	6.Recettes Publiques	Rece
-3.48	0.000	45.03	73.61	18.48	36.97	4.Dette Publique	Dett
-3.78	0.000	40.72	46.24	3.62	6.57	5.Dépenses Publiques	Dépe

Classification Ascendante Hiérarchique

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE Coupure 'a' de l'arbre en 3 classes

Classe 1 / 3

V.TEST	PROBA	---- POURCENTAGES ---- CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				28.57	Classe 1 / 3		aa1a	8

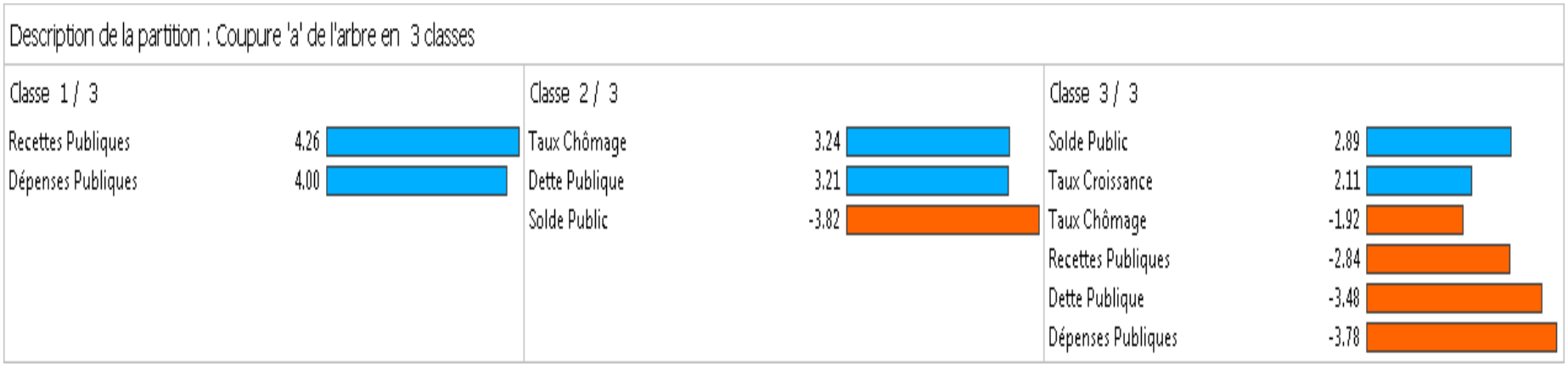
Classe 2 / 3

V.TEST	PROBA	---- POURCENTAGES ---- CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
28.57	Classe 2 / 3						aa2a	8

Classe 3 / 3

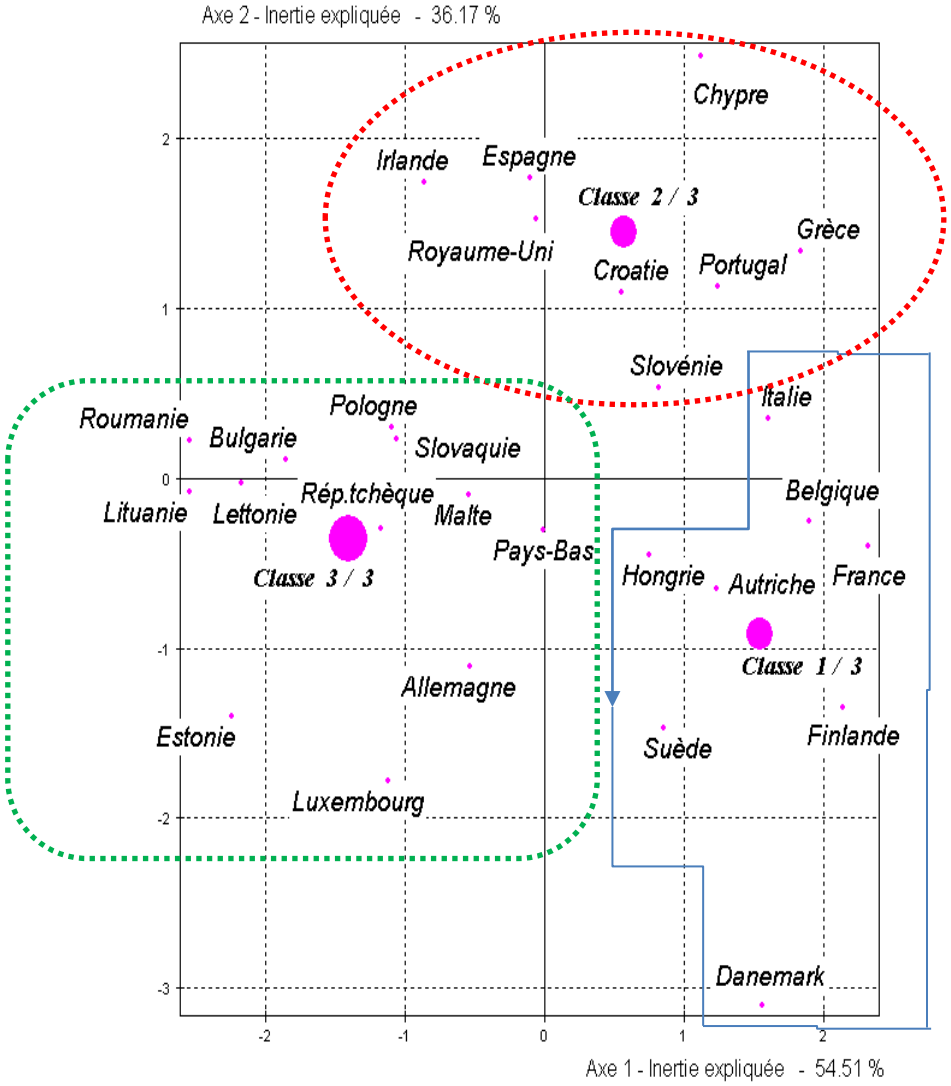
V.TEST	PROBA	---- POURCENTAGES ---- CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				42.86	Classe 3 / 3		aa3a	12

Classification Ascendante Hiérarchique



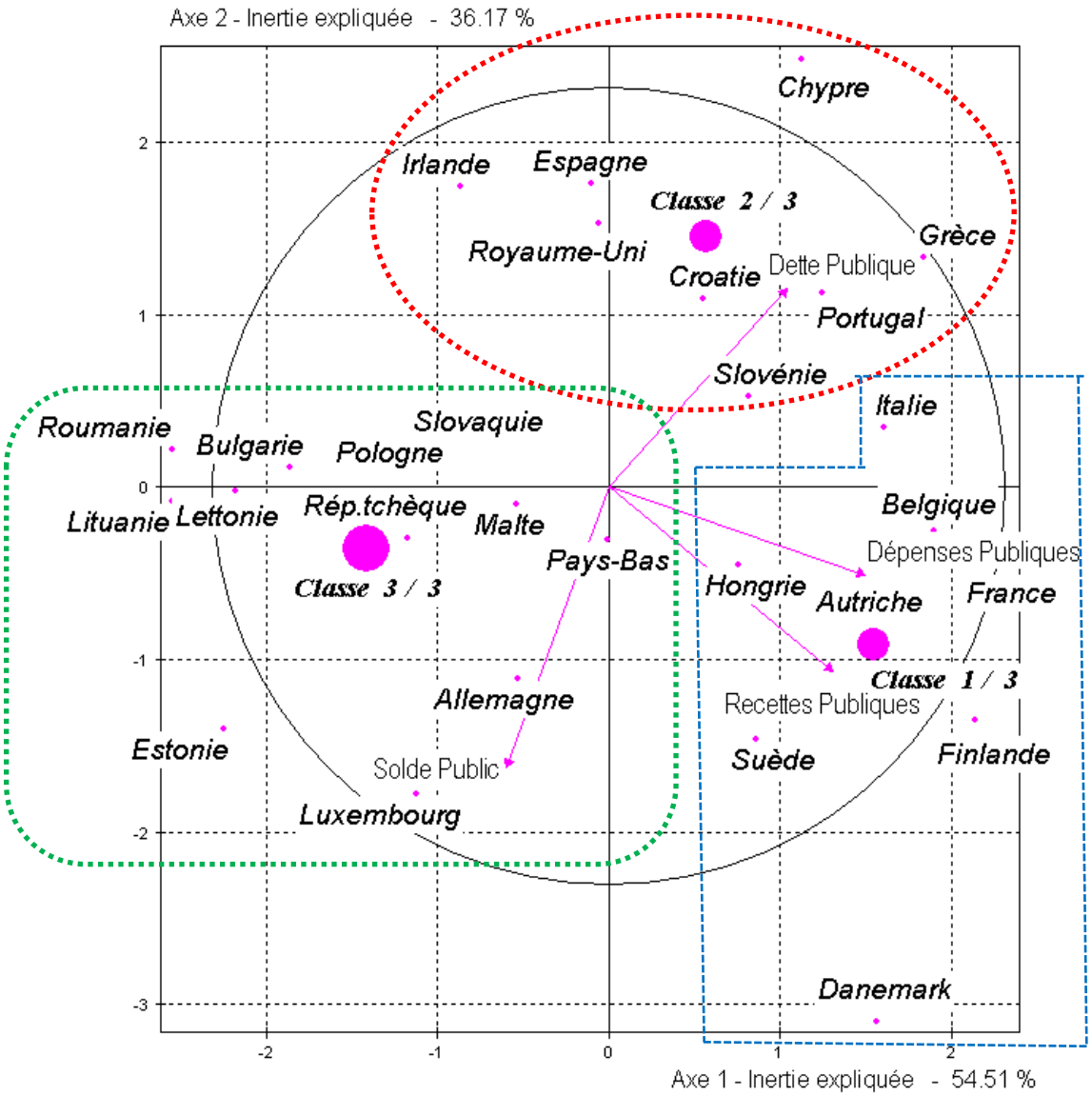
Classification Ascendante Hiérarchique

Représentation des Centres de classes et des Individus actifs sur le premier plan principal



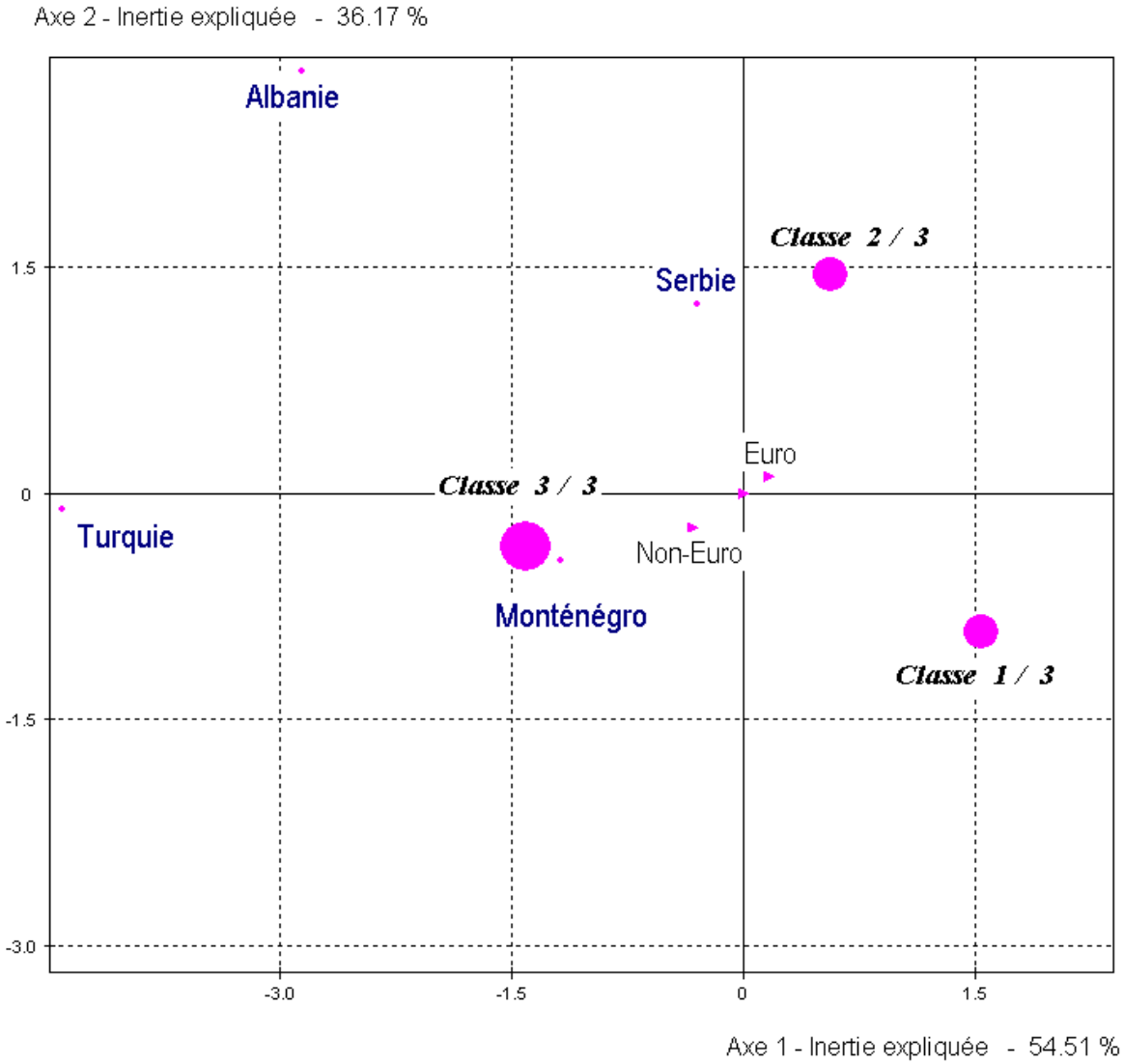
Classification Ascendante Hiérarchique

Représentation des Centres de classes et des Individus actifs sur le premier plan principal



Classification Ascendante Hiérarchique

Représentation des individus illustratifs et des modalités illustratives sur le premier plan principal



Analyse des données

Stratégie «Thèmascope»
Méthode factorielle & Classification