Classification de variables qualitatives pour la compréhension de la prise en compte de l'environnement par les agriculteurs

Vanessa KUENTZ-SIMONET, Sandrine LYSER, Jacqueline CANDAU, Philippe DEUFFIC, Marie CHAVENT, Jérôme SARACCO

XI[e] Journées de Méthodologie Statistique, Paris
24-26 janvier 2012

irstea
www.irstea.fr

Introduction

✓ Classification d'observations : construire des classes homogènes d'individus pour l'élaboration de profils-types.

✓ Etude de cas : dégager des tendances chez les agriculteurs concernant leur perception environnementale.

✓ Stratégie classique : méthode factorielle suivie d'une classification sur les scores des composantes principales des observations.

✓ Effets néfastes de cette procédure « tandem analysis » soulignés par plusieurs auteurs (DeSarbo et al., 1991; De Soete et Carroll, 1994; Vichi et Kiers, 2001).

✓ Différentes alternatives existent dans la littérature :

- k-means clustering procedure in a reduced space (De Soete et Carroll, 1994)
- Factorial k-means (Vichi et Kiers, 2001)
- Clustering and Disjoint PCA (Vichi et Saporta, 2009)
- Multidimensional scaling ou unfolding analysis (Heiser, 1993; De Soete et Heiser, 1993)
- Etc.

✓ Peu d'approches dédiées spécifiquement à l'analyse de données qualitatives.

✓ Notre proposition : remplacer la première étape d'ACM par une approche de classification de variables.

Plan de la présentation

1. Une approche par classification de variables
2. Description des données de l'application
3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs
4. Les profils-types des agriculteurs
5. Conclusion

1. Une approche par classification de variables

a) Objectifs de la classification de variables

✓ Regrouper les variables liées entre elles afin de construire des classes de variables homogènes.

✓ Utilisation dans de nombreuses applications : analyse sensorielle, biochimie, économie, recherche de règles d'association (Plasse et al. 2007, analyse de la base permanente des équipements de l'INSEE (Gelein et Sautory, 2009), etc.

✓ Objectif de sélection de variables ou de réduction de la dimension du tableau de données.

✓ Approche simple et courante : calcul d'une matrice de dissimilarités entre les variables et application d'une méthode usuelle de classification (voir Abdallah et Saporta, 1998 pour la proposition de différents critères d'association pour la classification de variables qualitatives)

1. Une approche par classification de variables

a) Objectifs de la classification de variables

✓ Proposition de méthodes spécifiques pour la classification de variables :

- Procédure *Varclus* du logiciel SAS
- CLV (Vigneau et Qannari, 2003) ; Diametrical clustering (Dhillon et al. 2003)
- Analyse de la Vraisemblance du Lien (Lerman, 1993)

✓ Approche de classification de variables (Chavent et al., 2011) :

- Pour des variables quantitatives, qualitatives ou un mélange des deux
- Deux algorithmes de classification sont proposés : un algorithme hiérarchique ascendant et un algorithme de type k-means

✓ Présentation et utilisation de la version hiérarchique (car pas d'idée a priori sur le nombre de classes de variables) sur des variables qualitatives.

✓ Package R intitulé *ClustOfVar.*

1. Une approche par classification de variables

b) L'algorithme de classification ascendante hiérarchique

Notations :

Soit $\{z_1,\ldots,z_p\}$ un ensemble de variables qualitatives.

Soit Z la matrice de données correspondante de dimensions $n \times p$, où n est le nombre d'observations.

Dans un souci de simplicité, nous notons $z_j \in M_j^n$ la j[e] colonne de Z avec M_j l'ensemble des modalités de z_j.

Notons $P_K = (C_1,\ldots,C_K)$ une partition en K classes des p variables.

Variable synthétique d'une classe :

Dans la classe C_k, la variable synthétique y_k est définie comme la variable quantitative à laquelle les variables (qualitatives) de la classe sont le plus liées : $y_k = \arg\max_{u \in \mathbb{R}^n} \sum_{z_j \in C_k} \eta^2_{u|z_j}$

où $\eta^2_{u|z_j}$ est le rapport de corrélation entre z_j et u. Son expression est donnée par

$$\eta^2_{u|z_j} = \frac{\sum_{s \in M_j} n_s (\bar{u}_s - \bar{u})^2}{\sum_{i=1}^{n} (u_i - \bar{u})^2}$$

avec : n_s effectif de la modalité s, $\bar{u}_s$ moyenne de u calculée sur les observations possédant la modalité s.

- mesure la part de variance de u expliquée par les modalités de z_j (appartient à [0,1]).
- évalue le lien entre la variable qualitative et la variable synthétique quantitative.

1. Une approche par classification de variables

b) L'algorithme de classification ascendante hiérarchique

- ✓ Il a été démontré par différents auteurs (Escofier, 1979; Saporta, 1990; Pagès, 2004) que y_k est la première composante principale issue de l'ACM appliquée à Z_k la matrice formée par les colonnes de Z qui correspondent aux variables de la classe C_k.
- ✓ La variance empirique de y_k est alors égale à $\sum_{z_j \in C_k} \eta^2_{y_k|z_j} = \lambda_1^k$, où λ_1^k est la première valeur propre issue de l'ACM de Z_k.
- ✓ Dans notre approche de classification de variables : calcul de la variable synthétique d'une classe à l'aide d'une présentation en DVS de la méthode PCAMIX.
- ✓ PCAMIX : méthode d'analyse factorielle pour un mélange de données quantitatives et qualitatives.
- ✓ Introduite sous le nom d'AFDM par Pagès (2004) et PCAMIX par Kiers (1991).
- ✓ Inclut comme cas particuliers l'ACP et l'ACM.

1. Une approche par classification de variables

b) L'algorithme de classification ascendante hiérarchique

Homogénéité d'une classe de variables :

Mesure d'adéquation entre les variables de la classe et le représentant synthétique quantitatif : $S(C_k) = \sum_{z_j \in C_k} \eta^2_{y_k|z_j} = \lambda_1^k$

=> Homogénéité maximale lorsque tous les rapports de corrélation valent 1: les variables de la classe sont très fortement liées entre elles et apportent la même information.

Homogénéité d'une partition :

L'homogénéité d'une partition est définie de la façon suivante :

$$H(P_K) = \sum_{k=1}^{K} S(C_k) = \lambda_1^1 + \ldots + \lambda_1^K$$

où $\lambda_1^1 + \ldots + \lambda_1^K$ sont les premières valeurs propres issues des ACM appliquées à chacune des K classes de P_K.

1. Une approche par classification de variables

b) L'algorithme de classification ascendante hiérarchique

Description de l'algorithme :

- ✓ Partition en singletons puis agrégations successives de deux classes jusqu'à l'obtention d'une seule classe contenant la totalité des variables.
- ✓ Mesure d'agrégation entre deux classes A et B :

$$d(A,B) = S(A) + S(B) - S(A \cup B) = \lambda_1^A + \lambda_1^B - \lambda_1^{A \cup B}$$

  - Correspond à la perte d'homogénéité observée quand deux classes A et B sont agrégées.
  - Hauteur d'une classe dans le dendrogramme est définie par $h(C) = d(A,B)$.
  - Indice positif $h(C) \geq 0$ mais la propriété de croissance monotone $A \subset B \Rightarrow h(A) \leq h(B)$ n'a pas encore été démontrée.

1. Une approche par classification de variables

b) L'algorithme de classification ascendante hiérarchique

Choix du nombre de classes de variables :

- ✓ Evaluation de la stabilité des partitions emboîtées issues du dendrogramme.
- ✓ Utilisation d'une approche bootstrap afin de perturber légèrement les données et de voir si la partition des variables est stable.
- ✓ Application de l'algorithme de classification ascendante hiérarchique sur B échantillons bootstrap tirés à partir des n observations.
- ✓ Comparaison des partitions de ces dendrogrammes avec les partitions de la hiérarchie initiale en utilisant le critère de Rand corrigé.
- ✓ Stabilité d'une partition : moyenne des indices de Rand corrigés.
- ✓ Structure forte cachée dans les données retrouvée malgré des perturbations sur l'échantillon ?

Typologie des individus :

- ✓ Extraction des variables synthétiques issues de la partition des variables choisie.
- ✓ Application de la CAH sur les coordonnées des individus sur ces variables (critère de Ward).

2. Description des données de l'application

Le questionnaire :

✓ Enquête postale menée par une équipe de sociologues du Cemagref de Bordeaux à l'échelle nationale auprès des agriculteurs français.

✓ Contexte d'évolution du monde agricole.

✓ Prise en compte de l'environnement par les agriculteurs.

✓ Deux aspects :

- Perception des professionnels vis-à-vis de l'environnement
- Pratiques en faveur de l'environnement

✓ Une centaine de questions fermées relatives à la prise en compte de l'environnement, lien entre activité et protection de l'environnement, valeurs et facettes du métier, politiques publiques, etc. (organisées en 4 grandes parties).

✓ Questions relatives aux caractéristiques des agriculteurs et de leur exploitation afin d'établir un profil socio-technique des répondants.

2. Description des données de l'application

Les individus enquêtés :

✓ Etude commanditée par le Centre National pour l'Aménagement des Structures des Exploitations Agricoles.

✓ Agriculteurs français dont l'exploitation est orientée vers 5 productions spécifiques (élevage de montagne, élevage intensif, grandes cultures, cultures pérennes, polyculture-élevage).

✓ Choix de la zone d'étude en tenant compte du critère de production et de l'existence d'une ou plusieurs problématiques environnementales (Natura 2000, SAGE, etc.).

✓ 5 départements : Puy-de-Dôme, Mayenne, Seine-et-Marne, Gironde, Dordogne.

Les données finales :

✓ Selection de 67 variables relatives à la perception de l'environnement, abordée via la conception du métier, de l'environnement, de la nature et des mesures agro-environnementales.

✓ Variables supplémentaires : caractéristiques socio-économiques, éloignées de la thématique, ou trop centrales.

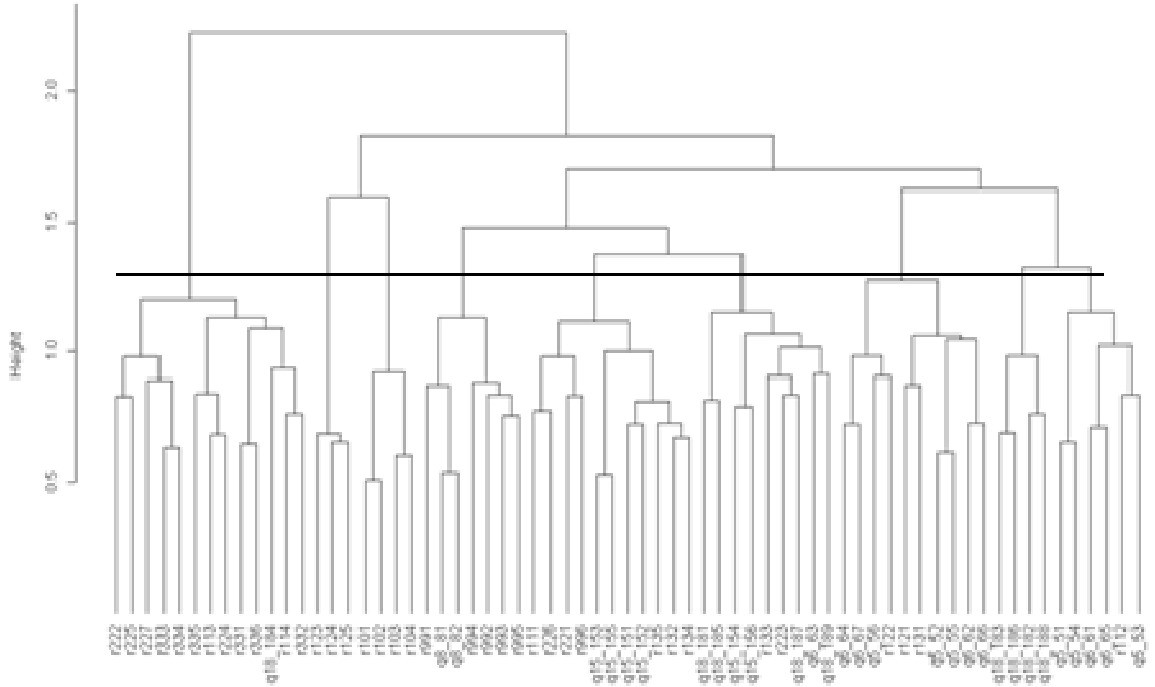
✓ Variables à 2 ou 3 modalités.

✓ Suppression des données manquantes => 544 individus.

3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

✓ Utilisation du package *ClustOfVar*.

✓ Classification ascendante hiérarchique des variables : visualisation de l'arbre pour analyser les agrégations successives des variables.



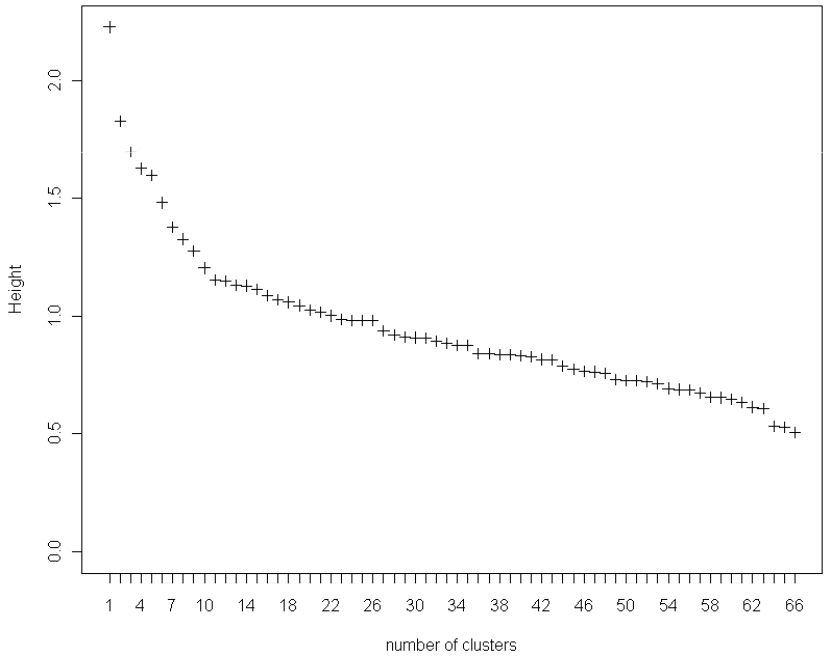
Dendrogramme issu de la classification ascendante hiérarchique des 67 variables qualitatives

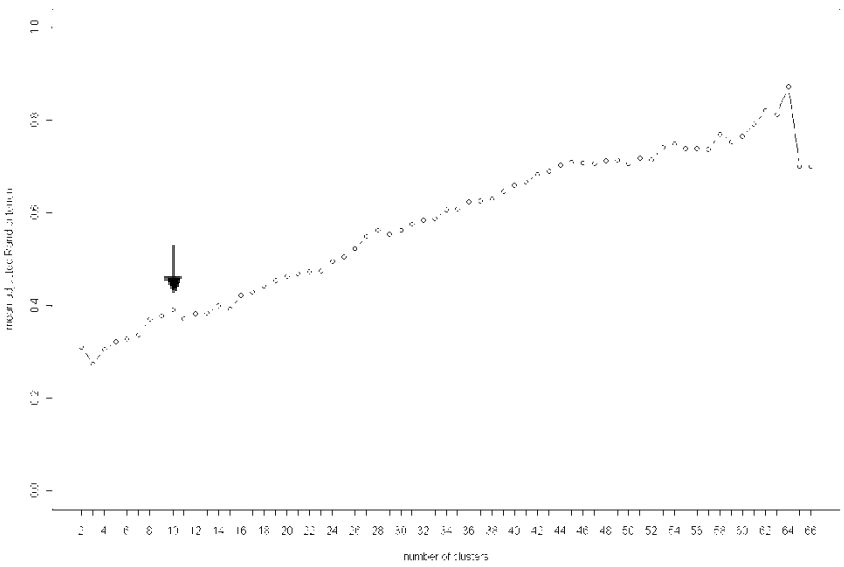
3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

Aide au choix du nombre de classes :

✓ Evolution du critère d'agrégation (utilisé pour indicer la hiérarchie) : perte en cohésion lorsqu'on agrège deux classes.

✓ Calcul de la fonction de stabilité pour 100 échantillons bootstrap.





✓ Au vu de ces graphiques,10 classes semblent être un choix intéressant mais l'interprétation nous guide pour retenir 9 classes.

✓Remarque sur 4 groupes de variables (structure du questionnaire).

3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

Constitution des classes :

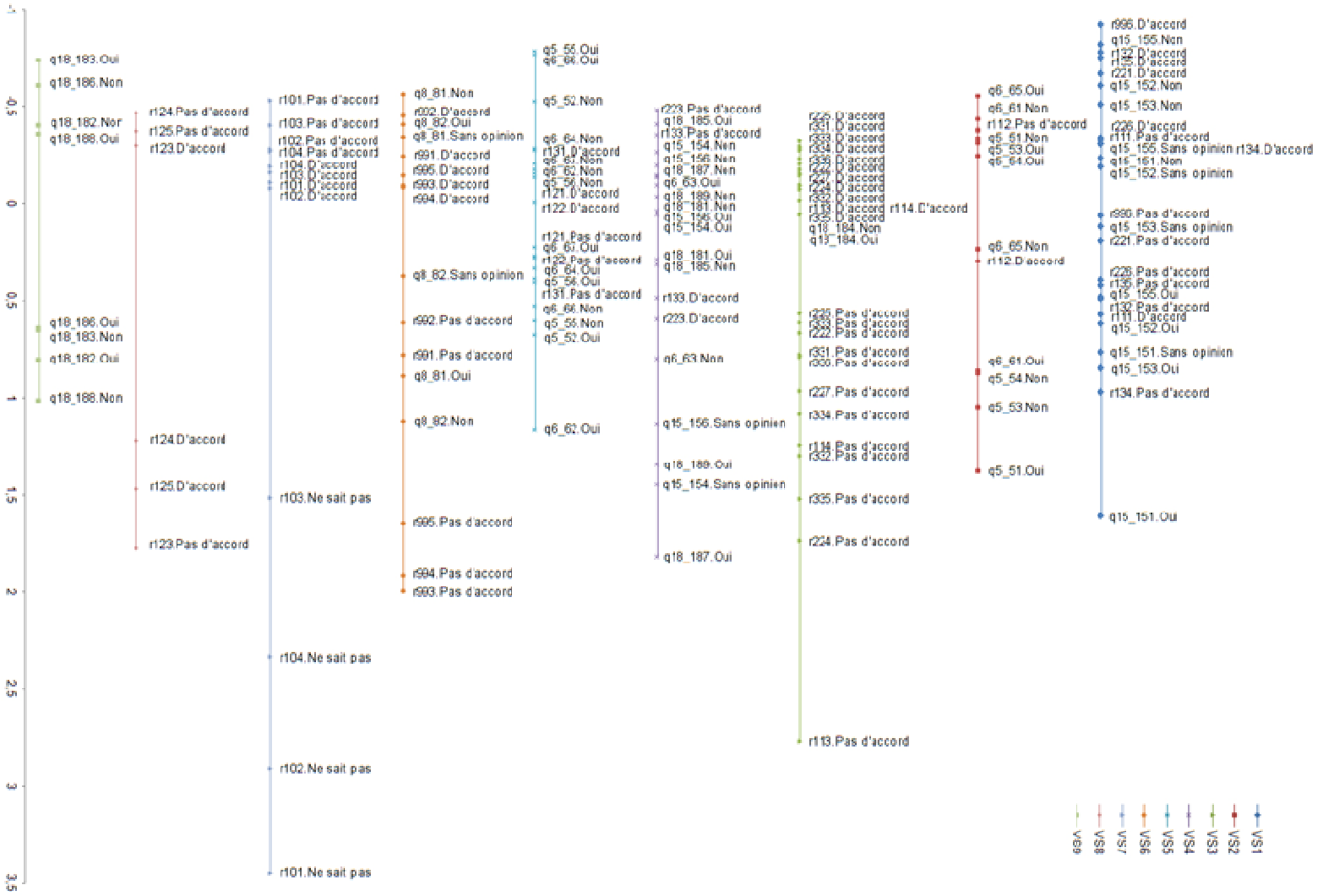
Classe 1 *11 variables*	**Classe 2** *6 variables*	**Classe 3** *13 variables*	**Classe 4** *9 variables*
q15_153 (0,39)	q5_51 (0,46)	r113 (0,28)	r223 (0,29)
r132 (0,38)	q6_61 (0,38)	r334 (0,26)	q18_187 (0,28)
q15_15 (0,35)	q5_53 (0,33)	r224 (0,26)	q15_154 (0,21)
q15_152 (0,34)	q5_54 (0,21)	r336 (0,22)	q15_156 (0,18)
r134 (0,30)	q6_65 (0,13)	r331 (0,23)	r133 (0,17)
r135 (0,31)	r112 (0,11)	r332 (0,21)	q18_185 (0,13)
q15_151 (0,26)		r333 (0,18)	q18_189 (0,13)
r111 (0,19)		r225 (0,18)	q6_63 (0,11)
r221 (0,13)		r227 (0,18)	q18_181 (0,01)
r226 (0,13)		r222 (0,14)	
r996 (0,05)		r114 (0,12)	
		r335 (0,12)	
		q18_184 (0,01)	

Classe 5 *10 variables*	**Classe 6** *7 variables*	**Classe 7** *4 variables*	**Classe 8** *3 variables*	**Classe 9** *4 variables*
q5_55 (0,47)	q8_81 (0,48)	r101 (0,54)	r124 (0,58)	q18_183 (0,48)
q6_66 (0,40)	q8_82 (0,41)	r102 (0,54)	r125 (0,54)	q18_186 (0,39)
q5_52 (0,35)	r992 (0,28)	r104 (0,52)	r123 (0,53)	q18_188 (0,37)
q6_62 (0,21)	r995 (0,25)	r103 (0,37)		q18_182 (0,32)
r131 (0,11)	r993 (0,20)			
q6_64 (0,10)	r991 (0,19)			
q5_56 (0,06)	r994 (0,18)			
q6_67 (0,05)				
r121 (0,03)				
r122 (0,01)				

Partition des 67 variables qualitatives en 9 variables synthétiques (rapport de corrélation entre la variable et la variable synthétique de la classe)

3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

Les variables synthétiques vues comme « gradients » :



3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

Labellisation des variables synthétiques :

	Label	Valeurs négatives	Valeurs positives
VS1	Lien, relation avec le monde non-agricole	Lien difficile avec le monde non-agricole, MAE semblent être un frein à l'activité, problèmes d'environnement ignorés	Mesures environnementales bénéfiques pour l'activité et le lien avec le monde non-agricole
VS2	Attraits du métier	Indépendance, contact avec la nature, nourrir les hommes	Adaptation au marché, technique de pointe, activité motivante
VS3	Difficultés du métier, de son exercice	Difficultés nombreuses, de plusieurs ordres	Confiance en l'avenir, pas de difficulté
VS4	Adaptation du métier aux mesures environnementales et aspect économique du métier	Préoccupations économiques pour l'application des MAE et la finalité du métier	Difficulté d'adaptation du métier aux mesures en faveur de l'environnement et ses applications. Les mesures en faveur de l'environnement véhiculent une image ancienne de l'agriculture, incitent à revenir à des savoir-faire anciens
VS5	Finalité du métier	Adaptation, évolution	Protection, histoire familiale, patrimoine
VS6	Situation de l'environnement	Inquiétude, attention portée à la situation environnementale	Pas d'inquiétude, rejet de la situation environnementale
VS7	Relation agriculture-environnement dans 20 ans	Avis tranché (d'accord ou pas)	Indécis
VS8	Zones peu productives	Entretien de ces zones	Pas d'entretien de ces zones
VS9	MAE	Difficultés d'ordre administratif	Difficultés d'ordre économique, de travail

Résumé des informations des 9 variables synthétiques

3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

Quelques quantités statistiques :

✓ Qualité de la partition des variables : $H(P_9) = 17,3$

✓ Homogénéité des classes :

- Homogénéité d'une classe : plus grande valeur propre de l'ACM de la classe
- Calcul du pourcentage de variance de la classe expliquée par la variable synthétique

Classe	1	2	3	4	5	6	7	8	9
Homogénéité de la classe	2,8	1,6	2,4	1,5	1,8	2,0	2,0	1,7	1,6
% de variance expliquée	18,9	27,1	18,3	13,7	17,9	22,0	24,6	55,2	39,1

✓ Pourcentage de gain en cohésion :

- Ratio entre le gain obtenu avec cette partition et le gain maximum qui est atteint avec la partition en singletons :

$$E(P_K) = \frac{H(P_K) - H(V)}{p - H(V)}$$

où $H(V) = \sum_{j=1}^{p} \eta^2_{y|z_j}$ est l'homogénéité de l'ensemble des variables V à classer, avec y la variable synthétique globale de V .

- Quantité vaut 0% pour la partition en une seule classe et 100% pour la partition des singletons.
- Pourcentage de gain en cohésion avec la construction de 9 variables synthétiques pour 67 variables : 21%

3. Les variables synthétiques de la prise en compte de l'environnement par les agriculteurs

✓ Corrélation entre les variables synthétiques :

- Pas de contraintes d'orthogonalités imposées dans la construction des variables synthétiques.
- Application : deux VS sont significativement corrélées (négativement).

  => Lien entre l'attention portée à la situation environnementale et l'opinion vis-à-vis du caractère bénéfique ou non des MAE pour l'activité et le lien avec le monde non-agricole.

✓ Similarités entre les variables d'une classe :

- Définition d'une mesure de similarité entre deux variables de type quelconque (quantitatif ou qualitatif) dans Chavent et al. (2011) pour la version de l'algorithme de partitionnement des variables.
- Sorte de corrélation canonique au carré.
- Pour deux variables qualitatives : interprétation géométrique

  => Plus cette valeur est proche de 1, plus les sous-espaces linéaires engendrés par les indicatrices des variables sont proches.

4. Les profils-types des agriculteurs

✓ CAH sur les coordonnées des individus sur les 9 variables synthétiques (VS).

✓ Identification de 7 classes d'individus au vu des résultats statistiques et de l'interprétation sociologique.

✓ 2 classes d'effectifs faibles mais intéressantes car il s'agit d'individus dont la perception environnementale est bien distincte.

Classe	Effectif	Pourcentage
1	127	23,35%
2	77	14,15%
3	28	5,15%
4	110	20,22%
5	109	20,04%
6	58	10,66%
7	35	6,43%

Composition de la typologie des agriculteurs

✓ Avantage de l'approche *ClustOfVar* : caractérisation des classes d'individus par les VS et non les 67 variables qualitatives initiales.

4. Les profils-types des agriculteurs

Variable synthétique	Classe d'individus						
	1	2	3	4	5	6	7
1	0,517	**1,476**	0,668	0,418	**-1,503**	-0,886	-0,909
2	0,175	**1,176**	-0,160	**-0,695**	-0,022	-0,372	-0,499
3	-0,101	0,162	**4,026**	-0,523	-0,653	-0,322	0,251
4	-0,548	0,241	0,356	**0,823**	-0,278	-0,272	-0,185
5	**-1,471**	0,287	0,343	**0,826**	0,403	0,025	0,092
6	-0,289	**-0,747**	0,691	**-0,850**	**1,428**	0,039	0,130
7	-0,398	-0,200	-0,036	-0,022	-0,458	-0,313	**4,578**
8	-0,432	0,079	0,030	-0,513	-0,573	**2,437**	-0,381
9	-0,388	**1,349**	0,316	-0,416	-0,179	-0,079	-0,434

Moyenne des VS pour les 7 classes d'agriculteurs

Classe 1 : valeur moyenne négative de VS5 => agriculteurs intéressés par le changement, ils aiment leur métier car cela demande d'évoluer constamment.

Classe 3 : agriculteurs confiants pour l'avenir, exercent leur métier sans difficultés (VS3).

Classe 5 : rejettent les préoccupations environnementales, pensent que la gravités des pbs de l'environnement est exagérée (VS6), et sont très critiques vis-à-vis des MAE (VS1).

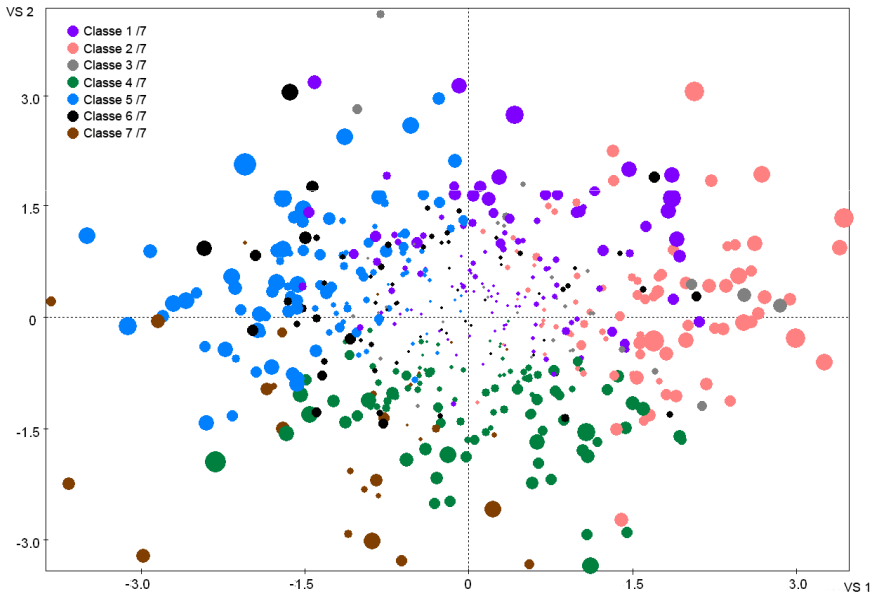
Classe 4 : agriculteurs attentifs à la protection de l'environnement (VS6) qu'ils considèrent difficile à concilier avec le progrès technique (VS4). Protéger les ressources naturelles et le paysage est, pour eux, une des premières finalités de leur activité (VS5). Si cette préoccupation environnementale laisse entrevoir des individus en questionnement et propices à remettre en cause certaines pratiques, l'évolution constante de leur activité ne les intéresse pas plus que cela (VS2).

Certaines classes sont caractérisées par une seule VS (pas seulement celles d'effectif faible, on a aussi les classes 1 et 3)
=> Homogénéité à l'intérieur des classes et hétérogénéité entre elles.

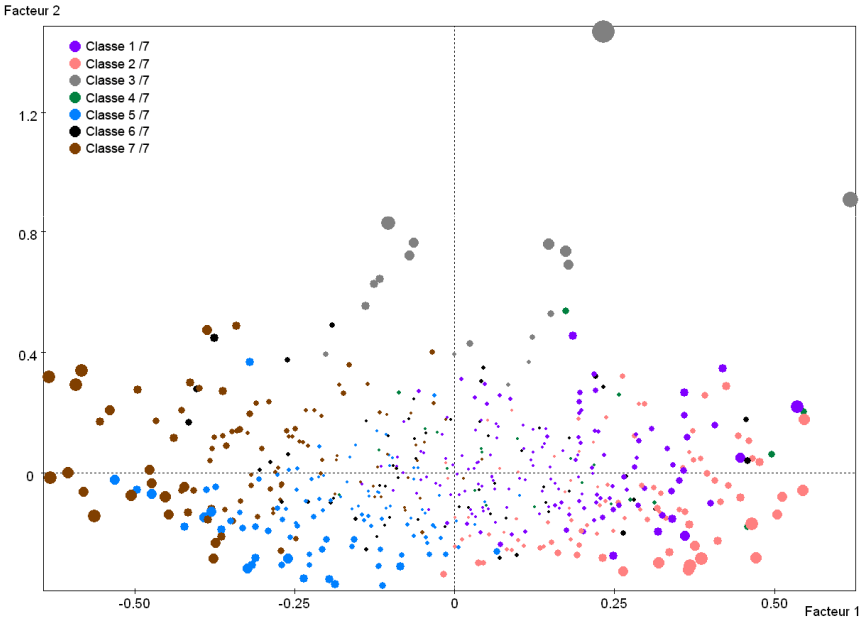
4. Les profils-types des agriculteurs

Une interprétation simplifiée :

- ✓ Difficulté du choix du nombre d'axes en ACM (et pourcentages d'inertie faibles…).
- ✓ Classes plus homogènes et séparées entre elles, et mieux projetées avec ClustOfVar qu'avec « tandem analysis » (TA).



CoV : nuage des individus dans (VS1, VS2)



TA: nuage des individus dans le plan (F1, F2)

- ✓ **Interprétation simplifiée** avec les VS comme gradients plutôt que toutes les modalités des variables avec TA.

4. Les profils-types des agriculteurs

Une interprétation simplifiée :

✓ Critère de Rand corrigé : concordance entre les 2 partitions égale à 22%.

	Partition en 7 classes obtenue avec l'approche classique « tandem analysis » (TA)							
Partition en 7 classes obtenue avec l'approche *ClustOfVar* (CoV)	C1_TA	C2_TA	C3_TA	C4_TA	C5_TA	C6_TA	C7_TA	Total
C1_CoV	**54**	1	0	0	**36**	11	13	115
C2_CoV	6	2	0	0	6	9	**62**	85
C3_CoV	0	1	**17**	0	0	6	**10**	34
C4_CoV	7	19	0	0	**51**	20	6	103
C5_CoV	**49**	**50**	0	2	0	7	0	108
C6_CoV	**28**	**21**	0	0	5	3	12	69
C7_CoV	1	2	1	**21**	2	3	0	30
Total	145	96	18	23	100	59	103	544

Croisement des partitions obtenues avec CoV et TA

✓ Pourcentage d'inertie légèrement plus élevé avec ClustOfVar (37%) qu'avec TA (33%).

✓ Calculs d'indices internes de validité (Gordon, 1999 ou Mirkin, 2005) : silhouette de la partition, distance maximum entre classes, plus petite distance entre des points de classes différentes, etc.

=> Valeurs proches pour les 2 partitions.

✓ Principales différences dans les résultats « sociologiques » (disparition de certaines classes, ou interprétation légèrement différente).

4. Les profils-types des agriculteurs

Validation de la typologie :

✓ Etape explicative de discrimination pour comprendre les règles de construction des classes : segmentation par arbre avec la méthode CART.

	ClustOfVar **avec les 9 variables synthétiques**	**« tandem analysis » avec les 10 composantes principales issues de l'ACM**	**« tandem analysis » avec les 67 variables initiales**	***ClustofVar*** **avec les 67 variables initiales**
Nombre de règles d'affectation pour les 7 classes *remarque : il n'y a pas de correspondance entre les numéros de classe des différentes méthodes*	3 4 2 3 2 1 3	3 4 4 4 3 3 3	13 12 5 2 8 6 12	3 4 0 3 1 2 2
Coût de mauvais classement				
- Échantillon d'apprentissage	0,14	0,12	0,13	0,34
- Échantillon test	0,33	0,33	0,45	0,50

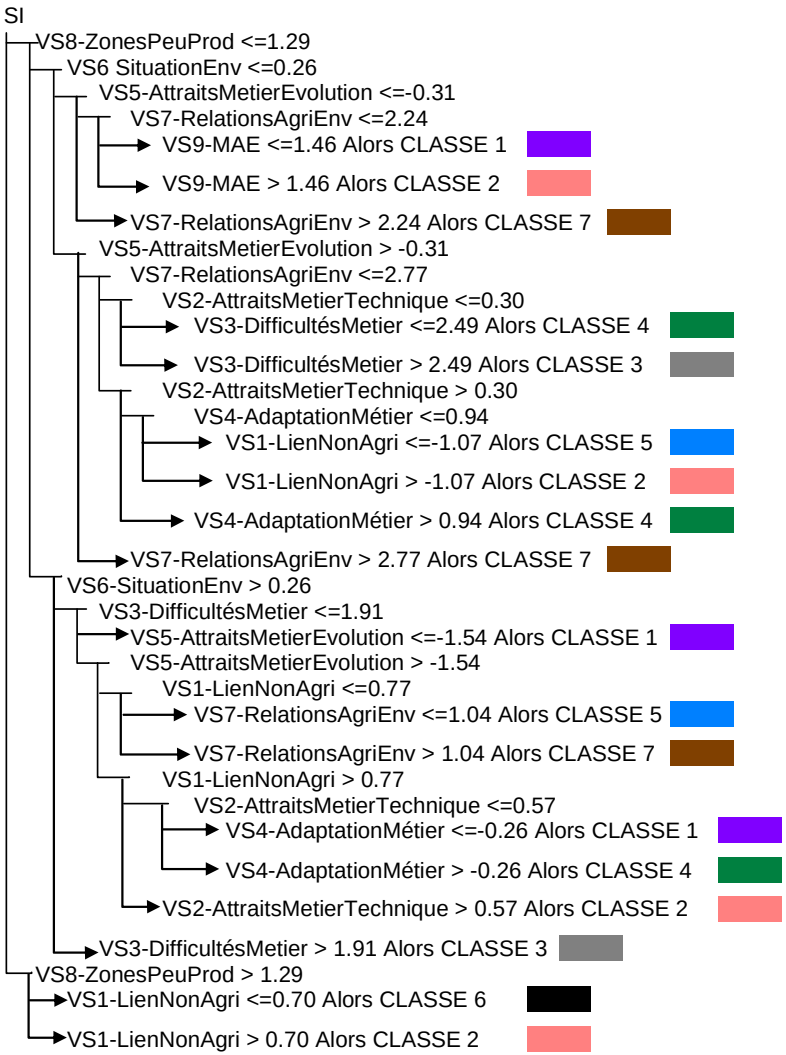
Comparaison des règles d'affectation issues de la méthode CART

✓ Nombre de règles plus faible avec ClustOfVar et les 9 VS.

=> Affectation des individus plus simple, classes plus homogènes et distinctes.

4. Les profils-types des agriculteurs

Validation de la typologie :



Exemple de lecture :

Si VS8 (zones peu productives) est supérieure à 1.29 (pas d'entretien) et si VS1 (relations avec le monde agricole) est inférieure à 0.70 (lien difficile), alors l'individu se classe dans la classe 6 (agriculteurs adeptes de la déprise agricole).

5. Conclusion

✓ Proposition d'utiliser une approche de classification de variables en remplacement de l'ACM, préalable à une classification des observations.

✓ Construction plus souple de variables synthétiques (pas de contraintes d'orthogonalités) et qui préserve les liaisons entre les variables initiales.

✓ Interprétation des VS plus simples que celle des composantes principales (sorte de gradients facile à labelliser).

✓ Compréhension des classes d'individus au travers des VS simplifiée.

✓ Application relative à la perception environnementale des agriculteurs : typologie plus intéressante en termes d'interprétation qu'avec la stratégie classique, mais « statistiquement » proches.

✓ Travaux en sociologie : l'utilisation d'une méthode d'analyse quantitative pour établir une typologie des agriculteurs sur la perception environnementale n'a pas été proposée dans la littérature française.

✓ Perspective : proposition d'une méthode qui optimiserait simultanément le critère d'homogénéité de la classification de variables et le critère relatif à la typologie des individus.