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 (**)

(*) Irstea, UR ADBX
(**) INRIA Bordeaux Sud-Ouest, Équipe CQFD

Introduction

La statistique exploratoire multidimensionnelle permet d'extraire à partir d'un ensemble de données de nombreuses informations synthétiques. Parmi ces méthodes, les techniques de classification automatique sont utilisées pour produire des groupements de lignes ou de colonnes d'une matrice de données. À ce titre, la classification des observations permet de construire des classes d'individus telles que deux individus à l'intérieur d'une même classe soient le plus ressemblant possible et telles que deux individus de classes distinctes soient le plus différent possible. La formation de ces classes homogènes permet ainsi de dégager des profils-types. Dans l'application présentée ici, l'objectif poursuivi par la classification d'observations est de synthétiser l'information et de dégager des tendances chez les agriculteurs concernant leur prise en compte de l'environnement.

Pour des données quantitatives, une stratégie classique consiste à réaliser une Analyse en Composantes Principales (ACP) des données puis à appliquer une méthode de classification sur les scores des individus mesurés sur les composantes principales obtenues. Cette approche est pertinente lorsque le nombre de variables est important car cela permet la réduction du nombre de variables par un ensemble restreint de composantes orthogonales. Elle est aussi utile lorsque l'utilisateur suspecte certaines variables de contribuer très faiblement à la structure de classification cachée dans les données. La première étape d'analyse factorielle permet ainsi d'éliminer le bruit éventuel contenu dans les données et de se concentrer sur l'information essentielle. Concernant les données qualitatives, une Analyse des Correspondances Multiples (ACM) est en général réalisée afin de recoder la matrice initiale en données quantitatives et utiliser ensuite les approches classiques de classification qui prennent en entrée des données quantitatives.

Cependant certains auteurs (voir par exemple DeSarbo *et al.* [1]; De Soete et Carroll [2]; Vichi et Kiers [3]) ont souligné les effets néfastes de cette procédure en deux étapes qu'ils nomment « tandem analysis ». Selon eux, ces deux étapes réalisées indépendamment avec chacune leur propre critère à optimiser, ne peuvent permettre l'identification d'une partition intéressante des observations. Ils précisent qu'il est indispensable de définir une méthodologie qui optimise un seul critère et non deux critères qui peuvent jouer un rôle opposé. Ils illustrent à l'aide de données simulées et réelles le fait que l'ACP identifie parfois des composantes qui contribuent peu à la détection d'une structure dans les observations ou qui au contraire masquent l'information taxinomique.

Différentes alternatives ont donc été proposées. De Soete et Carroll [2] introduisent une méthode appelée « k-means clustering procedure in a reduced space » qui est basée sur le critère défini dans l'algorithme des k-means. L'idée est d'optimiser ce critère sous la contrainte que les centres des classes appartiennent à un sous-espace de l'espace engendré par les colonnes de la matrice des données. Une procédure de moindres carrés alternatifs est utilisée avec comme différence principale par rapport à l'algorithme classique des k-means, l'étape de représentation qui est obtenue grâce à une Décomposition en Valeurs Singulières. Une représentation biplot permet ensuite de représenter simultanément les centres des classes et les variables initiales dans ce même sous-espace. L'avantage principal de cette approche par rapport à un algorithme classique des k-means est sa plus grande généralité : une métrique générale est utilisée au lieu d'une métrique diagonale. Cette méthode fait également preuve de parcimonie puisque la solution obtenue appartient

à un espace de dimension réduite. Selon Vichi et Kiers [3], cette procédure peut se révéler inefficace lorsque les données présentent une variance importante dans des directions orthogonales à celle contenant l'information intéressante pour la classification. En effet ces directions vont être privilégiées puisqu'elles vont contribuer fortement au critère optimisé qui est basé sur la somme des carrés des distances des points aux centres des classes.

Vichi et Kiers [3] proposent une approche nommée « factorial k-means » afin de déterminer un sous-espace de représentation des données tel que les points projetés dans ce sous-espace aient la plus petite distance (au carré) aux centres des classes dans ce même sous-espace. Cette approche combinant algorithme des k-means et ACP est en lien avec la méthode introduite par Diday *et al.* [4]. Elle permet de sélectionner les composantes les plus pertinentes pour la classification en minimisant un même critère. Il s'agit de projeter la matrice des données dans un sous-espace engendré par les colonnes d'une matrice orthonormale. Dans ce sous-espace, avec ces composantes, une partition des objets est cherchée de telle sorte que les objets soient le plus proche possible des centres des classes des objets. Un critère de type moindre carrés est minimisé, il correspond à l'écart entre la matrice de données projetées et la matrice des centres des classes. Le critère minimisé est un critère de déviance intra-classes de la partition des objets décrits par leur projection sur les composantes. Un algorithme des moindres carrés alternatifs est utilisé pour maximiser le critère. Les auteurs précisent qu'un inconvénient de leur approche apparaît lorsque les données présentent des dimensions avec des variances faibles car leur approche va se concentrer en priorité sur ces dimensions car elles contribuent peu à la fonction de perte. Pour éviter ce problème, les auteurs préconisent de supprimer au départ les dimensions triviales présentes dans les données. Citons à ce titre les travaux de De Soete et Carroll [2] qui proposent le même type d'approche intitulée « reduced K-means analysis ». La description de cette méthode ainsi qu'une comparaison avec la méthode « factorial k-means » est réalisée dans Timmerman *et al.* [5].

Différentes méthodes combinant une classification et la recherche d'un sous-espace de représentation ont été proposées. On peut citer les travaux relatifs au « multidimensional scaling » ou « unfolding analysis » (Heiser [6] ; DeSoete et Heiser [7]). Une approche de modèle de mélange reposant sur des hypothèses de normalité a également été proposée par DeSarbo *et al.* [1] pour effectuer simultanément une approche de multidimensional scaling et de classification floue des données.

D'autres travaux ont été récemment proposé à ce sujet : « clustering and disjoint principal component analysis » (Vichi et Saporta [8]), l'approche par modèle de mélange pour la classification croisée de Govaert et Nadif [9] ou encore l'approche de classification simultanée des lignes et des colonnes de Martella *et al.* [10] proposée dans le cadre de l'analyse de puces à ADN.

Cependant la plupart des approches proposées en remplacement de la « tandem analysis » sont dédiées à l'analyse de données quantitatives. À notre connaissance, le cas des données qualitatives a reçu moins d'attention.

Dans ce travail, nous proposons de remplacer la première étape d'ACM (données qualitatives dans notre étude de cas) par une approche de classification de variables. En effet, l'objectif de l'ACM est d'identifier une première composante principale qui explique le plus grand pourcentage de variance initiale contenue dans le nuage de points puis une seconde qui lui est orthogonale et qui explique un grand pourcentage d'inertie et ainsi de suite. Ainsi on peut concevoir que des informations relatives à la structure des observations puissent être masquées par la création de ces composantes non corrélées qui visent seulement à reconstruire au mieux la variance initiale. Au contraire, la classification de variables supprime l'information redondante et la création des variables synthétiques se fait au vu de la réorganisation des variables en classes homogènes. De plus, avec cette approche, les variables synthétiques obtenues ne sont pas nécessairement orthogonales, ce qui offre plus de souplesse dans l'étape suivante de classification des observations à partir de ces variables.

L'article est organisé de la façon suivante. La première section décrit la méthodologie utilisée pour réaliser la typologie des observations, avec une entrée par la classification de variables que nous avons récemment proposée Chavent *et al.* [11]. Dans un second temps, l'étude de cas relative à la perception de l'environnement par les agriculteurs est décrite. Les résultats de la classification de variables sont présentés dans la section 3, avec une description et une interprétation des variables synthétiques de la prise en compte de l'environnement par les agriculteurs. La section 4 est consacrée à l'analyse des profils-types des agriculteurs selon la problématique étudiée. Enfin, quelques remarques et perspectives sont données en conclusion.

1. Une approche par classification de variables

1.1. Objectifs de la classification de variables

L'objectif de la classification de variables est de regrouper les variables liées entre elles afin de construire des classes de variables homogènes. Dans de nombreuses applications, on s'intéresse à la classification des variables et non à celle des individus. C'est le cas par exemple en analyse sensorielle (mise en place de groupes de descripteurs), en biochimie (classification de gènes), en marketing (segmentation d'un panel de consommateurs), en économie (détection de stratégies financières), etc. On peut citer par exemple les travaux de Plasse *et al.* [12] qui utilisent la classification de variables dans la recherche de règles d'association pour une application issue de l'industrie automobile. Gelein et Sautory [13] utilise également ce type d'approche pour analyser la base permanente des équipements des communes produite par l'INSEE. L'idée de la classification de variables est de chercher des groupes de variables liées c'est-à-dire porteuses de la même information. Un autre objectif poursuivi par la classification de variables est la suppression des redondances entre les variables et ainsi la réduction de la dimension du tableau de données. Dans ce cas, il est nécessaire de sélectionner dans chaque classe une variable ou de résumer chaque classe de variables par une variable synthétique.
Une approche simple et courante pour classifier un ensemble de variables est de calculer une matrice de dissimilarités entre les variables et d'appliquer ensuite sur cette matrice une méthode usuelle de classification, dédiée initialement à la classification d'individus. Pour des variables qualitatives, plusieurs critères d'association peuvent être calculés comme par exemple Rand, χ^2 , Belson, Jaccard, Sokal ou Jordan (voir par exemple Abdallah et Saporta [14]). L'inconvénient de ces approches est qu'elles ne fournissent pas directement de variables synthétiques des classes et qu'il est délicat de choisir un critère plutôt qu'un autre.

Parallèlement à cette approche, certaines méthodes ont été proposées spécifiquement pour la classification de variables. La plus célèbre est probablement la procédure Varclus du logiciel SAS. Cette procédure complexe avec peu de justifications des options offertes fournit une hiérarchie ou une partition des variables quantitatives. Une autre approche consiste à utiliser un algorithme de classification qui fournit simultanément des classes de variables et leurs variables synthétiques. Deux algorithmes de partitionnement de ce type existent déjà pour la classification de variables quantitatives et sont basés sur l'ACP: la méthode CLV (Clustering of variables around Latent Variables) proposée par Vigneau et Qannari [15] et la méthode Diametrical Clustering développée par Dhillon *et al.* [16]. Cependant pour des variables qualitatives, peu d'approches ont été proposées. On peut citer par exemple l'Analyse de la Vraisemblance du Lien de Lerman [17 , 18] qui est une méthode de classification hiérarchique de variables quantitatives ou qualitatives.

Nous avons récemment proposé une méthode spécifiquement dédiée à la classification de variables quel que soit leur type, quantitatif, qualitatif ou un mélange des deux. Cette approche utilise la méthode PCAMIX de Kiers [19], proposée sous le nom d'AFDM par Pagès [20], qui est une méthode d'analyse en composantes principales pour un mélange de variables quantitatives et qualitatives, et qui inclue l'ACP et l'ACM comme cas particuliers. Dans nos travaux sur la classification de variables, nous utilisons une approche en termes de Décomposition en Valeurs Singulières de PCAMIX. Deux algorithmes de classification de variables sont proposés : un algorithme hiérarchique ascendant et un algorithme de partitionnement de type k-means. Ces algorithmes visent à maximiser un critère d'homogénéité, basé sur le carré de la corrélation de Pearson pour des variables quantitatives et sur le rapport de corrélation pour des variables qualitatives. Une approche bootstrap est également proposée pour évaluer la stabilité des partitions de variables et ainsi choisir un nombre de classes adéquat. Dans cet article, nous nous concentrerons sur le cas de variables qualitatives et sur l'algorithme de classification ascendante hiérarchique, car nous nous n'avons pas d'idée *a priori* du nombre de classes de variables. Pour plus de détails sur la méthode dans sa généralité, le lecteur peut se référer à Chavent *et al.* [11]. L'implémentation de ces approches est disponible dans le package R nommé *ClustOfVar* (Chavent *et al.* [21]).

1.2. L'algorithme de classification ascendante hiérarchique de variables

1.2.1. Notations

Soit $\{z_1,\ldots,z_p\}$ un ensemble de p 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 $z_j \in M_1\times\ldots\times M_p$ 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.

1.2.2. Variable synthétique d'une classe C_k

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 :

$$y_k = \arg\max_{u\in\mathbb{R}^n} \sum_{z_j\in C_k} \eta^2_{u|z_j} \qquad (1)$$

où $\eta^2_{u|z_j}$ est le rapport de corrélation entre z_j et u. Son expression est donnée par $\eta^2(z_j,u)=\dfrac{\sum_{s\in M_j} n_s(\bar{u}_s-\bar{u})^2}{\sum_{i=1}^{n}(u_i-\bar{u})^2}$, avec n_s l'effectif de la modalité *s* et $\bar{u}_s$ la moyenne de u calculée sur les observations possédant la modalité s. Plus précisément, le rapport de corrélation mesure la part de variance de u expliquée par les modalités de z_j. Cette quantité appartient à [0,1] et évalue le lien entre la variable qualitative z_j et la variable synthétique quantitative u.

Il a été démontré par différents auteurs (Escofier [22]; Saporta [23]; Pagès [20]) 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. Ces auteurs ont également montré que 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 la matrice Z_k. Dans Chavent *et al.* [11], le calcul de la variable synthétique d'une classe est décrit à l'aide d'une présentation en Décomposition en Valeurs Singulières de la méthode PCAMIX. La présentation est uniformisée quel que soit le type de variables (quantitatif et/ou qualitatif). Ainsi dans le cas de variables qualitatives, le calcul de la variable synthétique y_k d'une classe se fait de la façon suivante (correspondant à une ACM) :

- *Recodage de la matrice des données* : $\tilde{Z}_k = JGD^{-1/2}$ est la version standardisée de la matrice des indicatrices G, avec la matrice diagonale D des fréquences des modalités, et $J=I-\mathbf{1}'\mathbf{1}$ l'opérateur de centrage des données (I est la matrice identité et $\mathbf{1}$ le vecteur ligne composé de 1).
- *Décomposition en Valeurs Singulières* de $\tilde{Z}_k = U\Lambda V^t$.
- $\sqrt{n}U\Lambda$ est la matrice des *composantes principales* des individus.
- y_k est la première colonne de cette matrice.
- $V\Lambda$ est la matrice des coordonnées des modalités sur les variables synthétiques.

1.2.3. Homogénéité S d'une classe de variables C_k

Il s'agit d'une mesure d'adéquation entre les variables de la classe et le représentant synthétique quantitatif y_k :

$$S(C_k) = \sum_{z_j \in C_k} \eta^2_{y_k|z_j} = \lambda_1^k \qquad (2)$$

Ainsi l'homogénéité d'une classe est maximale lorsque tous les rapports de corrélation valent 1. Cela signifie alors que les variables de la classe C_k sont très fortement liées entre elles et apportent la même information.

1.2.4. Homogénéité H d'une partition P_k

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

$$H(P_K) = \sum_{k=1}^{K} S(C_k) = \lambda_1^1 + ... + \lambda_1^K \qquad (3)$$

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.

1.2.5. L'algorithme de classification ascendante hiérarchique

Nous avons proposé une approche de classification ascendante hiérarchique basée sur l'optimisation du critère d'homogénéité défini dans (3). L'algorithme part de la partition en singletons puis il procède par agrégations successives de deux classes jusqu'à l'obtention d'une seule classe contenant la totalité des variables. À chaque étape, il s'agit d'agréger les deux classes de la partition qui ont la plus petite dissimilarité d définie de la façon suivante :

$$d(A,B) = H(A) + H(B) - H(A \cup B) = \lambda_1^A + \lambda_1^B - \lambda_1^{A \cup B} \qquad (4)$$

Cette dissimilarité mesure la perte d'homogénéité observée quand deux classes A et B sont agrégées. En utilisant cette mesure d'agrégation, à chaque étape $l = 1, \ldots, p-2$, la nouvelle partition en $p-l$ classes obtenue maximise H parmi l'ensemble des partitions en $p-l$ classes obtenue par agrégation de deux classes de la partition en $p-l-1$ classes. Lors de la dernière étape, la partition en une seule classe est obtenue.

La hauteur d'une classe $C = A \cup B$ dans le dendrogramme est définie par $h(C) = d(A,B)$. On montre facilement que $h(C) \geq 0$ mais la propriété de croissance monotone de l'indice c'est-à-dire « $A \subset B \Rightarrow h(A) \leq h(B)$ » n'a pas encore été démontrée. Notons qu'en pratique sur les jeux de données réelles ou simulées que nous avons utilisés, nous n'avons jamais observé de phénomènes d'inversion.

1.2.6. Stabilité d'une partition de variables

Cette procédure évalue la stabilité des partitions emboîtées qui sont issues du dendrogramme. L'idée est d'utiliser une approche bootstrap afin de perturber légèrement les données et de voir si la partition des variables est stable. Plus précisément, B échantillons boostrap sont tirés à partir des n observations et les dendrogrammes correspondants sont obtenus en appliquant l'algorithme de classification ascendante hiérarchique. Les partitions de ces B dendrogrammes sont alors comparées avec les partitions de la hiérarchie initiale en utilisant le critère de Rand corrigé. Ce critère basé sur le nombre de paires de points classés ensemble dans une même classe mesure le pourcentage d'accord entre les deux partitions. Nous utilisons la version corrigée de ce critère dont

l'espérance vaut 0 lorsque les concordances entre les deux partitions sont dues au hasard, et dont la valeur maximale vaut 1 pour des partitions identiques. Pour plus de précisions sur cet indice, le lecteur peut se référer à Hubert et Arabie [24]. Enfin la stabilité d'une partition est évaluée en calculant la moyenne des B indices de Rand corrigés. Plus cette valeur est proche de 1, plus les partitions se ressemblent, ce qui signifie qu'il y a une structure forte cachée dans les données que l'on retrouve malgré des perturbations sur l'échantillon. Cette approche est une aide dans le choix du nombre de classes, cependant elle ne doit pas se priver de l'interprétation des résultats qui permet également un choix de nombre de classes et de variables synthétiques intéressantes et utiles à l'utilisateur pour la suite de son analyse.

1.2.7. Typologie des individus

Après avoir choisi un nombre de classes de variables, nous pouvons extraire les variables synthétiques issues de la partition des variables choisie. Nous appliquons une méthode de classification des observations sur les coordonnées des individus sur ces variables synthétiques. Plus précisément, nous appliquons une classification ascendante hiérarchique (CAH) avec la distance euclidienne et le critère d'agrégation de Ward.

2. Description des données

En 2005, une équipe de sociologues du Cemagref de Bordeaux a réalisé, via une enquête postale une étude à l'échelle nationale auprès des agriculteurs français[1]. Cette étude se situait dans un contexte d'évolution du monde agricole, qui est passé depuis une trentaine d'années, d'une agriculture qui nourrit les hommes à une agriculture qui en plus de son activité principale, se voit attribuer de nouvelles finalités, comme la protection des ressources naturelles ou encore le lien social dans l'espace rural. C'est donc dans ce contexte de multifonctionnalité de l'agriculture, concept apparu en 1992 au sommet de Rio, que l'enquête portait sur « la prise en compte de l'environnement par les agriculteurs ». Mais cette notion n'est pas une variable binaire, c'est une composante plus complexe, qui évolue en fonction des normes environnementales ou sanitaires, des cahiers des charges des aides publiques, des préoccupations environnementales du monde agricole et non-agricole, etc. Ainsi, s'orienter vers une agriculture environnementale peut être perçu comme un « changement à la fois technique, cognitif et structurel » (Candau *et al.* [25]). En conséquence, l'étude portait sur deux aspects : la perception des professionnels vis-à-vis de l'environnement et les pratiques en faveur de l'environnement. Plus précisément, la prise en compte de l'environnement, le lien entre activité et protection de l'environnement sont abordés au travers d'une centaine de questions fermées, qui concernent les significations données par les agriculteurs à l'environnement, les valeurs et facettes du métier (représentation de la nature, idée de progrès, identité professionnelle, etc.) remises en cause par les orientations environnementales des politiques publiques et les pratiques adoptées pour respecter l'environnement. Des questions relatives aux caractéristiques des agriculteurs enquêtés et de leur exploitation (âge, niveau de formation, taille de l'exploitation, orientation technique, parcours professionnel…) ont été posées en fin de questionnaire pour aider à établir un profil socio-technique des répondants.

Le choix des individus à enquêter s'est fait en fonction de la population ciblée par le commanditaire de l'étude[2], c'est-à-dire les agriculteurs français dont l'exploitation est orientée vers 5 productions spécifiques (l'élevage de montagne, l'élevage intensif, les grandes cultures, les cultures pérennes et la polyculture-élevage). Le choix de la zone d'étude s'est ensuite fait en tenant compte du critère de production (un département par production, où seuls les cantons les plus représentatifs du type de production sont retenus), et de l'existence d'une ou plusieurs problématiques environnementales (ZNIEFF, Natura 2000, SAGE, etc.). Les départements du Puy-de-Dôme, de la Mayenne, de la Seine-et-Marne, de la Gironde et de la Dordogne répondant à ces critères ont donc été sélectionnés.

[1] Sur les 5000 questionnaires envoyés par voie postale début juillet 2005, 1051 ont été reçus dont environ 6 % hors-délai. De plus, près de 11% ont été rejetés car ils ne faisaient finalement pas partie de la population cible ou en raison d'une non-réponse partielle trop importante. Finalement, 879 questionnaires sont exploitables.

[2] L'étude a été commanditée par le Centre national pour l'aménagement des structures des exploitations agricoles (CNASEA).

Département	Type de production	Problématique environnementale	Effectifs
Puy-de-Dôme	Élevage en zone de montagne	Dynamique des paysages et biodiversité	143
Mayenne	Élevage intensif	Qualité de l'eau	181
Seine-et-Marne	Grandes cultures	Qualité de l'eau	179
Gironde	Cultures pérennes	Qualité de l'eau et des sols	192
Dordogne	Polyculture-élevage	Dynamique des paysages	184

Tableau 1 : Terrain d'étude et composition de l'échantillon

Adapté de : Candau et al. [25]

Bien que les méthodes d'analyse de données se prêtent parfaitement au traitement de cette enquête en permettant l'étude simultanée d'un grand nombre de variables, il est nécessaire de porter une attention particulière aux variables à analyser. Ainsi pour améliorer les performances de l'analyse et l'interprétation des résultats à venir, nous avons décidé de ne pas retenir les variables spécifiques à chacune des cinq zones ainsi que celles pour lesquelles la non-réponse est trop importante. L'analyse est donc appliquée à 67 variables[3] relatives à la perception de l'environnement, abordée via la conception du métier, de l'environnement, de la nature et des mesures agro-environnementales (MAE[4]).

La conception du métier est appréhendée au travers de questions générales sur le métier (r111 à r114) ou par le biais de variables plus spécifiques, qui portent sur les attraits du métier d'agriculteur (q5_51 à q5_56), les finalités poursuivies (q6_61 à q6_67) ou encore les difficultés rencontrées dans l'exercice du métier (r221 à r227 et r331 à r336).

Les variables qui ont trait à l'environnement s'intéressent aux problèmes d'environnement et à l'évaluation de leur gravité (q8_81 à r996), à la relation agriculture-environnement dans les 20 prochaines années (r101 à r104) et au rapport que les agriculteurs entretiennent avec la nature (r121 à r125) et les interrogent en particulier sur les mesures agro-environnementales, leur évaluation et les difficultés de mise en œuvre (r131 à q189).

Ces variables, à deux ou trois modalités, constituent les 4 grandes parties du questionnaire. Des variables supplémentaires (28) viennent enrichir l'analyse. Il s'agit des variables relatives aux caractéristiques socio-économiques des agriculteurs (production principale, effectif dans l'exploitation, exercice d'une activité d'accueil ou de vente directe, pluriactivité du chef d'exploitation, situation familiale, sexe, âge, niveau d'études, parcours professionnel, responsabilités), ainsi que des variables trop éloignées de la thématique étudiée (classement des termes qui représentent le plus ou le moins la nature) ou au contraire trop centrales (définition de l'environnement). Les détails de l'analyse des variables supplémentaires ne seront pas présentés.

Les individus ayant au moins une donnée manquante sur les 67 variables traitées en variables actives pour la classification de variables ont été supprimés, après vérification (statistiques descriptives) que l'échantillon n'est pas ou est peu modifié sur les caractéristiques socio-économiques. Au final, l'échantillon sur lequel porte notre application comporte 544 individus. Malgré la perte importante d'individus (335), nous avons fait le choix de ne pas faire d'imputation pour la non-réponse. En effet l'échantillon étudié reste finalement de taille relativement importante pour une étude exploratoire, dont l'objectif est de décrire la perception de l'environnement par les agriculteurs et de dégager des profils-types.

Classiquement, une analyse factorielle des données (ici ACM, vue la nature exclusivement qualitative des données) ainsi qu'une méthode de classification des individus serait une approche classique permettant de répondre à cette problématique complexe de la prise en compte de l'environnement par les agriculteurs. L'originalité de notre approche réside dans le remplacement de la première étape par une classification de variables, dont les principes et les fondements théoriques ont été décrits dans la première section. Dans la section 4 sur la description de la typologie des agriculteurs, certains résultats de l'approche classique dénommée « tandem analysis » seront brièvement présentés.

[3] Pour la liste complète des variables, se reporter à l'Annexe 1.

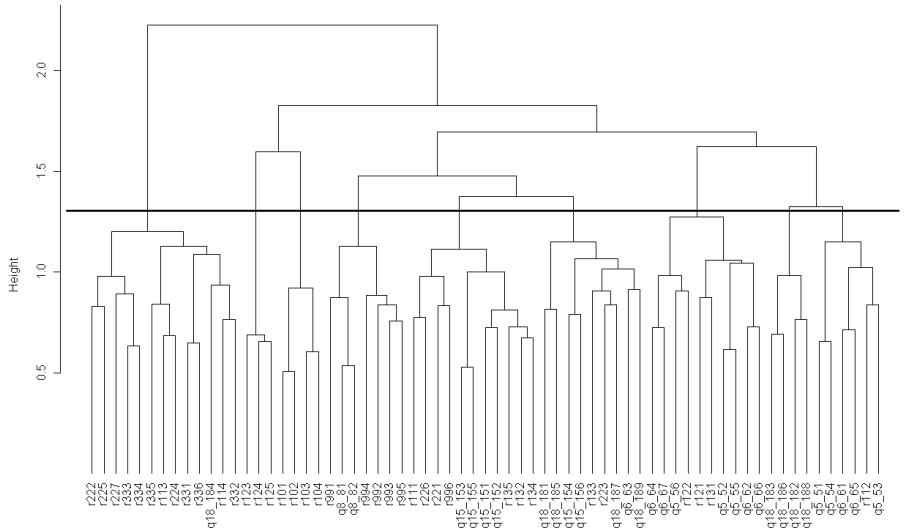
[4] MAE : dans le questionnaire, il s'agit de toute mesure réglementaire ou incitative visant la protection de l'environnement.

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

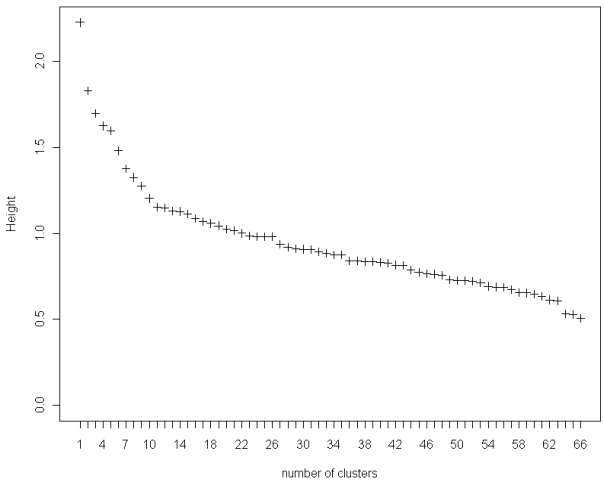
3.1. Dendrogramme et choix du nombre de classes de variables

Le package R intitulé *ClustOfVar* nous permet de réaliser une classification ascendante hiérarchique des variables. Le dendrogramme issu de cette classification est représenté sur le Graphique 1. Il nous permet d'analyser les agrégations successives de l'ensemble des variables et de visualiser ainsi les liaisons entre les variables. Cependant il est difficile de choisir un nombre de classes adéquat en observant seulement l'arbre. Le Graphique 2 représente l'évolution du critère d'agrégation (utilisé pour indicer la hiérarchie) qui peut aider au choix de nombre de classes. À chaque étape de la classification ascendante, ce critère mesure la perte en cohésion lorsqu'on agrège deux classes. Il s'agit donc de détecter un coude dans l'évolution de ce critère, signifiant ainsi qu'on a agrégé des classes trop différentes. Une observation minutieuse de ce graphique semble indiquer qu'un choix de 10 classes est pertinent. Pour finaliser ce choix, le Graphique 3 représente le calcul de la fonction de stabilité présentée dans la section 1. Sur ce graphique, le nombre d'échantillons bootstrap vaut B=100. La valeur de l'indice de Rand corrigé augmente avec le nombre de classes. En regardant les valeurs de la fonction pour un nombre faible de classes, l'objectif étant de réduire la dimension, on observe un léger pic pour 10 classes et donc une stabilité de cette partition de variables. Si on représente la dispersion de la valeur du critère de Rand sur les échantillons bootstrap (Graphique 4), nous remarquons que la valeur de la médiane atteint un pic pour 10 classes (en s'intéressant à un faible nombre de classes) et que sa dispersion est modérée vers des valeurs faibles.
Mais le choix du nombre de variables synthétiques est aussi dicté par l'interprétation que l'on peut faire de ces regroupements de variables. Aussi, en regardant l'interprétation des classes, nous constatons que deux d'entre elles sont composées par des variables qui sont proches dans leur thématique. C'est pourquoi nous décidons de retenir la partition en 9 classes (qui réunit ces deux groupes de variables), d'autant que l'interprétation de ces groupes de variables est relativement aisée.

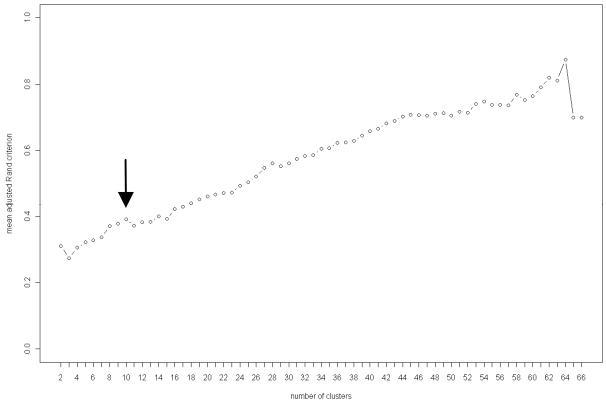
Concernant le choix du nombre de classes, une dernière remarque concerne la structuration du questionnaire. Comme mentionné dans la section précédente, la construction du questionnaire a été organisée autour de 4 thématiques. L'objectif de la classification de variables est de regrouper les variables fortement liées entre elles, c'est-à-dire celles sur lesquelles il y a un lien dans la façon dont les individus ont répondu aux questions. Il est intéressant de remarquer à ce titre que le nombre de 4 classes n'est pas adéquat au vu de ces différents graphiques. D'autre part, la partition des variables en 4 groupes de reprend pas ces 4 grandes parties du questionnaire. Ce premier résultat sur la classification des variables montre que les individus se ressemblent sur les choix qu'ils ont effectués sur les 9 groupes de variables et non uniquement au travers de ces 4 groupes de questions thématiques.



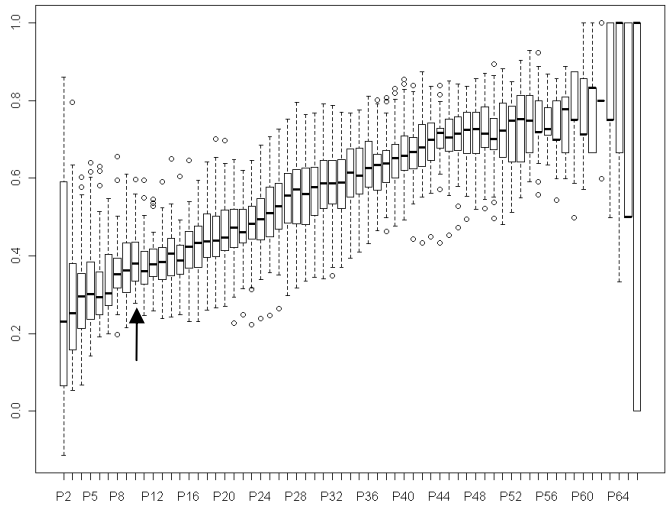
Graphique 1 : Dendrogramme issu de la classification ascendante hiérarchique des 67 variables qualitatives



Graphique 2 : Évolution du critère de classification des 67 variables qualitatives



Graphique 3 : Stabilité des partitions



Graphique 4 : Dispersion du critère de Rand moyen ajusté

3.2. Qualité de la partition de variables

L'homogénéité de la partition des 67 variables en 9 classes définie dans (3) vaut $H\left(P_9\right)=17,3$.

Nous pouvons également évaluer le pourcentage du gain en cohésion de cette partition. Pour cela, définissons tout d'abord l'homogénéité de l'ensemble V des variables à classer : $H(V)=\sum_{j=1}^{p}\eta^2_{y|z_j}$ où y est la variable synthétique globale de V . Notons qu'elle vaut p pour la partition des singletons.

Pour une partition P_K donnée en *K* classes, le pourcentage de gain en cohésion est défini comme le ratio entre le gain obtenu avec cette partition et le gain maximum qui est atteint avec la partition en singletons :

$$E(P_K)=\frac{H(P_K)-H(V)}{p-H(V)} \qquad (5)$$

Cette quantité vaut 0% pour la partition en une seule classe V et 100% pour la partition des singletons. Comme elle croît avec le nombre de classes, elle est surtout utile pour comparer des partitions de même nombre de classes, construites avec des approches différentes. Elle peut être calculée cependant à titre indicatif. Dans notre application, le pourcentage de gain en cohésion avec la construction de 9 variables synthétiques pour 67 variables vaut 21%, ce qui est plutôt satisfaisant.

Le détail de l'homogénéité de chacune des classes est donné dans le Tableau 2. Elle est définie d'après la formule (2) par la plus grande valeur propre de l'ACM de la classe, c'est-à-dire la variance de sa variable synthétique. Cette valeur, augmentant avec le nombre de variables dans la classe, nous calculons le pourcentage de variance de la classe expliquée par la variable synthétique. Pour cela, nous divisons l'homogénéité de la classe par la variance totale de la classe, qui vaut $\frac{\sum_{j \in C_k} \text{card}(M_j)}{p_k} - 1$ où $\text{card}(M_j)$ désigne le nombre de modalités de la variable j et p_k représente le nombre de variables de la classe C_k.

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

Tableau 2 : Homogénéité des classes et pourcentage de variance expliquée par leur variable synthétique

On voit que les variables synthétiques des classes 8 (zones peu productives) et 9 (MAE) sont celles qui expliquent le plus grand pourcentage de variance. On remarque que ce sont des classes avec peu de variables, dont la thématique est relativement proche. La classe 7, contenant peu de variables, possède un pourcentage d'inertie plus faible. Il s'agit de questions plus « larges » sur les relations agriculture-environnement. La classe 4 est la classe dont le pourcentage est le plus faible, elle regroupe des variables relatives à l'adaptation du métier aux mesures environnementales et à l'aspect économique du métier.

3.3. Constitution des classes

Le Tableau 3 décrit la partition des variables en 9 classes et indique entre parenthèses le rapport de corrélation entre la variable qualitative et le représentant synthétique quantitatif de la classe. Ainsi on peut voir que les classes d'effectif plus faible ont des variables fortement reliées à la variable synthétique. Pour les plus grandes classes, certaines valeurs sont plus faibles car elles regroupent des variables de thématiques plus variées.

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

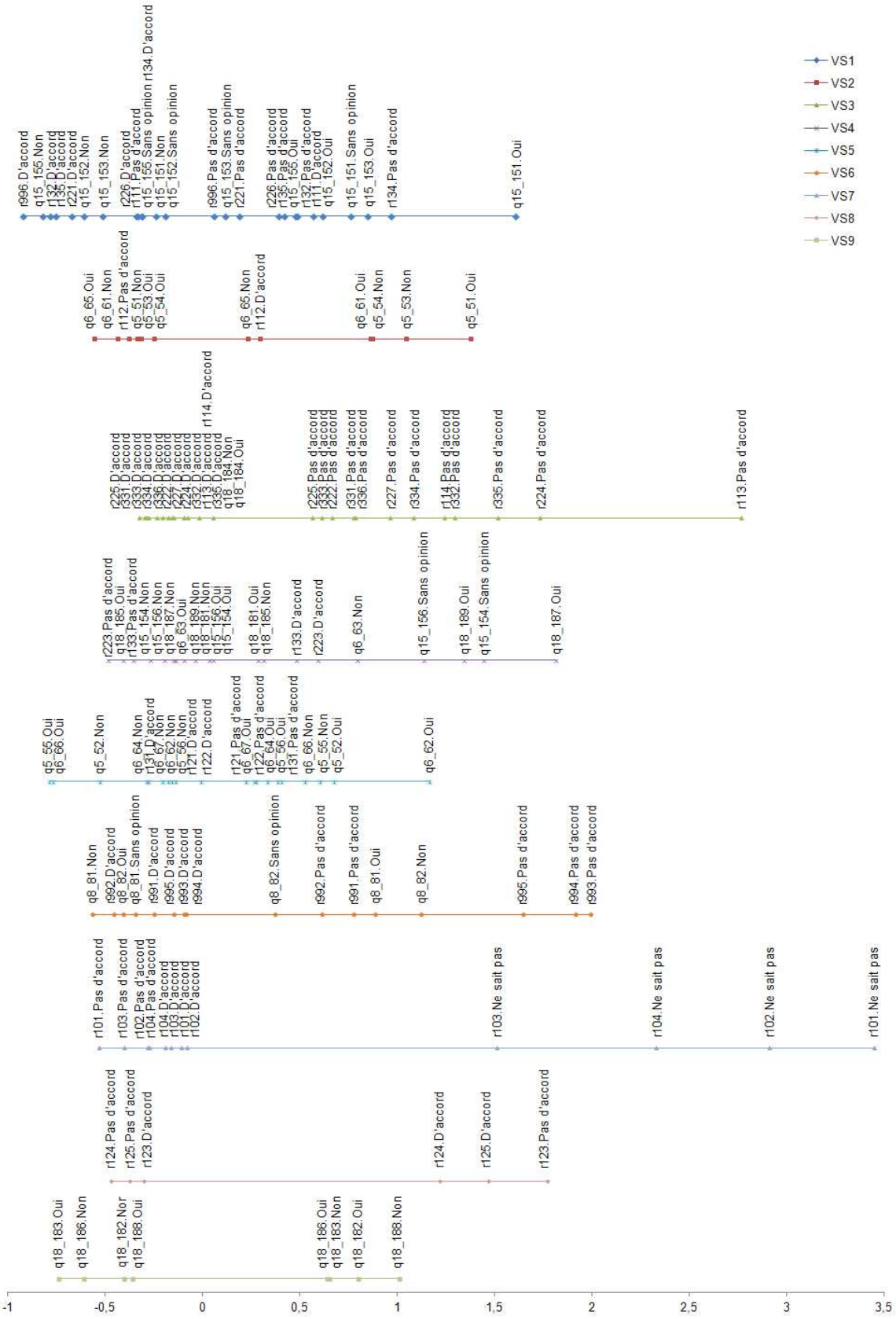
Tableau 3 : Partition des 67 variables qualitatives en 9 variables synthétiques (rapport de corrélation entre la variable et la variable synthétique de la classe)

3.4. Labellisation des variables synthétiques

Les coordonnées des modalités sur les variables synthétiques (définie dans la section 1.2.2) nous permettent de visualiser ces variables comme une sorte de gradient (cf. Graphique 5). En effet, les valeurs négatives ou positives sont associées à un certain ensemble de modalités des variables de la classe. Il est alors possible d'interpréter et de labelliser les variables synthétiques aisément (cf. Tableau 4).

	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

Tableau 4 : Résumé des informations des 9 variables synthétiques



Graphique 5 : Gradients des 9 variables synthétiques

Lecture : La variable synthétique 7 est la variable quantitative à laquelle les variables relatives aux scénarios d'évolution des relations agriculture-environnement dans les 20 prochaines années (r101 à r104) sont le plus liées. Ainsi, les valeurs positives de la VS7 sont associées aux modalités « Ne sait pas » alors que les valeurs négatives sont associées aux avis établis, les valeurs les plus faibles étant les modalités « Pas d'accord », alors que les valeurs les plus proches de zéro sont associées aux modalités « d'accord ».

Il est intéressant d'analyser si les variables synthétiques des classes sont liées entre elles. En effet contrairement à l'ACM, l'algorithme de classification de variables proposé n'impose pas de contraintes d'orthogonalités entre les variables synthétiques. Le Tableau 5 montre que ces variables sont faiblement corrélés, ce qui signifie qu'elles apportent des informations bien distinctes. Seules les variables synthétiques 1 et 6 sont corrélés négativement (-0,33). Un test de corrélation de Pearson confirme la significativité de cette corrélation avec une p-valeur égale à 10^{-5}. Ainsi les agriculteurs qui portent une attention à la situation environnementale ont tendance à penser que les MAE sont bénéfiques pour l'activité et le lien avec le monde non-agricole.

Classe	1	2	3	4	5	6	7	8	9
1	1	0,11	0,14	0,08	-0,11	**-0,33**	-0,03	-0,02	0,18
2		1	0,06	-0,01	-0,12	0,04	-0,12	-0,03	0,15
3			1	-0,01	-0,01	0,09	0,06	0,05	0,05
4				1	0,08	-0,09	0,01	-0,01	0,05
5					1	0,04	0,05	0,01	0,04
6						1	-0,02	-0,03	-0,12
7							1	-0,08	-0,03
8								1	0,07
9									1

Tableau 5 : Corrélation entre les variables synthétiques des classes

Dans Chavent *et al.* [11] la proposition d'un algorithme de partitionnement des variables nécessite la définition d'une mesure de similarité entre deux variables de type quelconque (quantitatif ou qualitatif). Il s'agit d'une sorte de corrélation canonique au carré qui vaut dans le cas particulier de deux variables quantitatives la corrélation de Pearson au carré alors que dans le cas d'une variable qualitative et d'une variable quantitative, cette mesure est égale au rapport de corrélation. Pour deux variables qualitatives, cette similarité ne correspond pas à une mesure d'association connue. Son interprétation est géométrique, plus cette valeur est proche de 1, plus les sous-espaces linéaires engendrés par les indicatrices des variables sont proches. Afin de ne pas surcharger le document, nous présentons seulement les valeurs de la classe 7 dans le Tableau 6. On voit par exemple que les deux variables concernant le lien entre l'agriculture et l'agroalimentaire (r101) et la place de l'environnement au cœur de l'agriculture (r102) ont la plus forte similarité dans la classe.

Variable	r101	r102	r103	r104
r101	1	0,24	0,04	0,11
r102		1	0,06	0,08
r103			1	
r104				1

Tableau 6 : Similarité entre les variables de la classe 7

L'interprétation des variables synthétiques montre que dans chaque groupe de questions, les agriculteurs ont répondu de la même manière, ce qui implique une uniformité dans les réponses. Une classification des individus va à présent apporter davantage de précisions en proposant une typologie des individus.

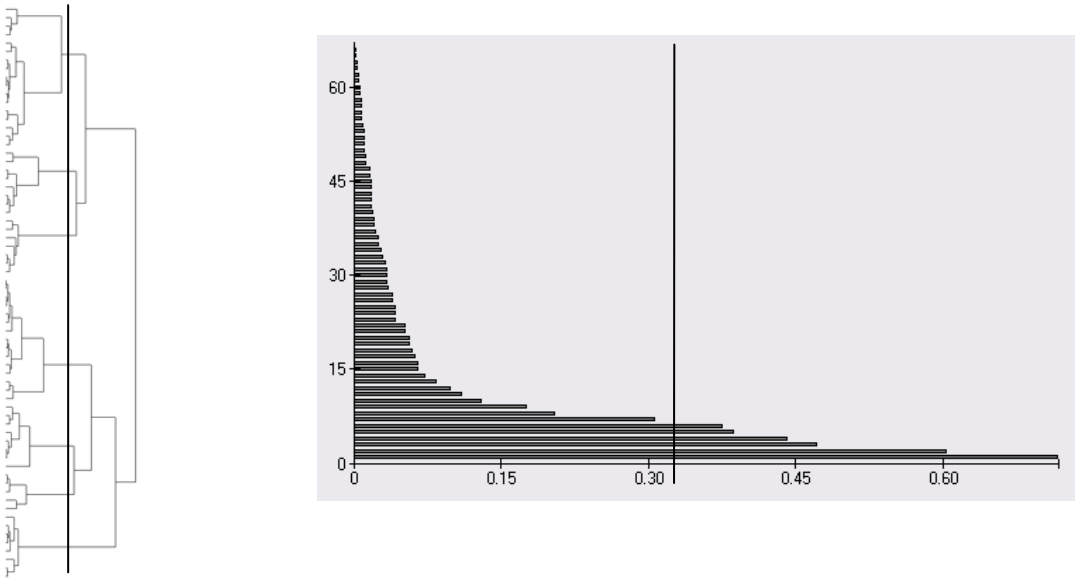
4. Profils-types des agriculteurs

4.1. Une typologie en 7 classes

L'approche par classification de variables, alternative à la stratégie classique utilisant l' ACM, a permis d'identifier 9 variables synthétiques. Toutefois, nous ne disposons pas d'une réelle typologie des agriculteurs pour la compréhension de la prise en compte de l'environnement par cette catégorie de population. Aussi, nous utilisons ces 9 variables quantitatives pour réaliser une classification ascendante hiérarchique avec le critère d'agrégation de Ward, afin d'élaborer une typologie des individus. La classification ascendante hiérarchique aboutit à la construction d'un arbre de classification (dendrogramme) qui montre les différentes étapes d'agrégation des individus, de la première étape où les *n* individus sont séparés (singletons) à la dernière étape où tous les individus sont regroupés en une seule classe. La succession des agrégations s'effectue en choisissant à chaque étape le regroupement d'une classe avec un individu ou un groupe d'individus qui minimise l'inertie intraclasse.

La coupure de l'arbre à un endroit où le palier est important conduit à l'obtention de classes homogènes bien séparées entre elles. Aussi, réaliser une classification des individus s'avère intéressant, en ce sens que cela permet d'avoir une vue concise et structurée des données, qui permettra de montrer des regroupements inattendus ou au contraire qui n'aboutira pas à des regroupements prévus.

La classification des individus est réalisée à l'aide du logiciel Spad. L'examen du dendrogramme et de l'histogramme des indices de niveau (cf. Graphique 6) de la CAH appliquée à nos données, révèle un nombre de classes pertinent quant à la structure des données : un premier saut important est visible entre le deuxième et le troisième niveau et un autre palier assez important apparaît entre le septième et le huitième niveau. Pour notre étude, la typologie en 3 classes ne présente pas un grand intérêt par rapport à la problématique étudiée dans l'analyse sociologique, alors qu'une typologie en 7 classes semble plus pertinente, notamment pour rendre compte de la diversité des profils qui se sont dessinés avec la classification de variables.



Graphique 6 : Dendrogramme issu de la classification ascendante hiérarchique des 544 individus et courbe des indices de niveau

Les 7 classes obtenues sont de tailles sensiblement différentes (cf. Tableau 7). Deux classes sont très petites (la classe 3 qui regroupe 5% des individus et la classe 7 qui en compte 6%), cependant, comme on peut le voir sur le dendrogramme, ces classes se forment assez tôt et semblent donc, pour cette raison, intéressantes à conserver et à analyser. Trois classes sont de taille relativement identique (la classe 1 avec 23% des agriculteurs ainsi que les classes 4 et 5 avec 20% chacune) et regroupent près des $^{2}/_{3}$ de l'échantillon.

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%

Tableau 7 : Résumé des informations des 7 classes d'individus

Une partition s'avère intéressante lorsque l'on décrit les classes par les individus $^{et}/_{ou}$ les variables. Ici, l'analyse par les individus présente peu d'intérêt car ils sont anonymes. Par la réduction du nombre de variables caractérisantes via *ClustOfVar*, nous pouvons interpréter la partition des agriculteurs, non pas au regard des 67 questions initiales, mais en s'appuyant sur les 9 variables synthétiques, décrites dans la section précédente. La valeur moyenne de chaque variable synthétique (VS) dans les 7 classes de la typologie est alors une aide à l'interprétation (cf. Tableau 8).

La classe 1 est caractérisée par une valeur moyenne négative de la VS5 (-1,471), on peut donc en déduire que les agriculteurs de cette classe sont intéressés par le changement, ils aiment leur métier parce que les personnes qui l'exercent doivent évoluer constamment, et ils considèrent que les MAE leur demandent de maîtriser des techniques de pointe.

Pour la classe 3 on note une valeur moyenne positivement élevée de la VS3 (4,026), ce qui nous amène à considérer que cette classe est définie par les agriculteurs qui sont confiants pour l'avenir et qui semblent exercer leur activité sans difficulté. Ils ne partagent pas bon nombre des difficultés proposées par l'enquête (paperasserie, prix des terres, main d'œuvre, etc.).

Les individus de la classe 6 peuvent être jugés comme étant adeptes de la déprise agricole car ils ont une valeur positive pour la VS8 (2,437).

Enfin la classe 7 est associée à une valeur moyenne positive de la VS7 (4,578), autrement dit, ces agriculteurs ne se projettent dans aucun scénario de prospective proposé, sans les rejeter pour autant. L'avenir leur paraît incertain.

Ces classes peuvent être caractérisées par une seule variable synthétique, ce qui rend facile leur interprétation. Certes, on retrouve ici les classes d'effectifs faibles (3 et 7), pour lesquelles on s'attend à une caractérisation concise. Cependant la classe 1, qui comptabilise le plus grand nombre d'individus (près d'un quart de l'échantillon), est également décrite par une seule VS. Ces résultats soulignent l'homogénéité à l'intérieur des classes et l'hétérogénéité entre elles.

Pour les 3 classes restantes, la caractérisation est moins marquée, une interprétation au travers de plusieurs variables synthétiques est nécessaire.

Ainsi, les agriculteurs de la classe 5 rejettent les préoccupations environnementales, pensent que la gravité des problèmes d'environnement est exagérée et que la situation n'est pas inquiétante ($\overline{VS6}$=1,428) et sont en même temps très critiques vis-à-vis des MAE qui sont pour eux un frein à l'activité et leurs projets ($\overline{VS1}$=-1,503).

Les agriculteurs de la classe 2 sont convaincus de la réalité des problèmes d'environnement et estiment qu'ils ne sont pas exagérés ($\overline{VS6}$=-0,747). Les mesures en faveur de l'environnement occupent une place important dans le processus de production ($\overline{VS1}$=1,476) et les démarches administratives liées aux MAE ne posent pas problème. Au contraire, ils acceptent que l'activité agricole soit régulée par les pouvoirs publics et restent néanmoins attachés au marché ($\overline{VS2}$=1,176), qui selon eux donne aussi des orientations à suivre. Cette catégorie d'agriculteurs éprouve conjointement des difficultés avec les dimensions entrepreneuriales de leur activité : ils dénoncent la charge de travail et les investissements nécessaires à la mise en œuvre des mesures en faveur de l'environnement ($\overline{VS9}$=1,349).

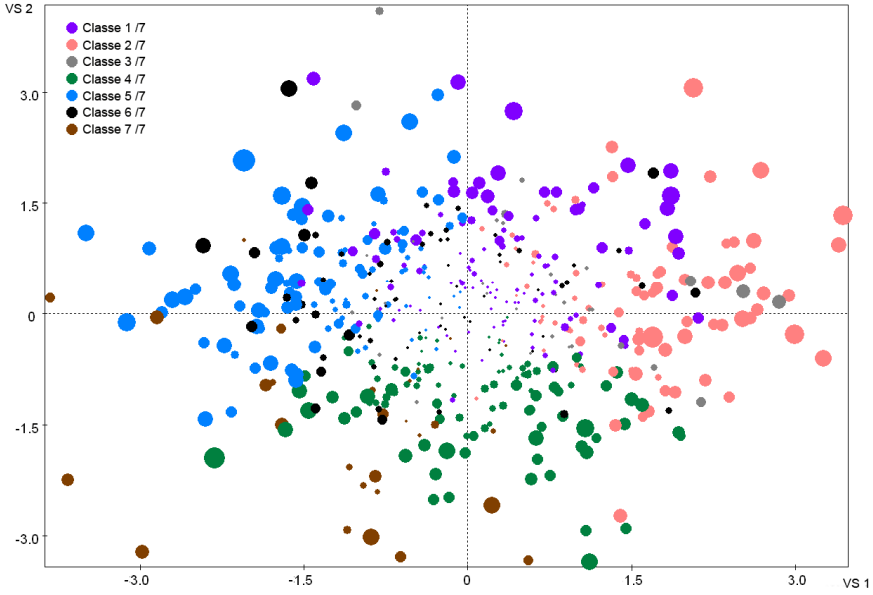
Enfin dans la classe 4, on trouve des agriculteurs particulièrement attentifs à la protection de l'environnement ($\overline{VS6}$=-0,850) qu'ils considèrent difficile à concilier avec le progrès technique ($\overline{VS4}$=0,823). Protéger les ressources naturelles et le paysage est, pour eux, une des premières finalités de leur activité ($\overline{VS5}$=0,826). 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 ($\overline{VS2}$=1,176). Ils déclarent en effet que le changement permanent n'est pas ce qui rend leur métier attrayant. Ils pensent même que les mesures environnementales réactivent des savoir-faire anciens.

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

Tableau 8 : Moyenne des valeurs synthétiques pour les 7 classes d'individus

4.2. Une interprétation simplifiée

La classification effectuée dans notre approche permet de distinguer les catégories d'agriculteurs par rapport à la prise en compte de l'environnement, d'identifier des profils cohérents d'agriculteurs. Elle se démarque des études antérieures sur la problématique, qui étaient basées sur des investigations qualitatives, mais également par la démarche méthodologique employée.
La représentation graphique des individus dans le plan formé par VS1 et VS2, en fonction de leur appartenance aux 7 classes montre des classes qui sont relativement bien homogènes et séparées les unes des autres. De plus, la qualité de projection des individus est satisfaisante (cf. Graphique 7).



Graphique 7 : Nuage des individus dans le plan (VS1, VS2) issu de l'approche par classification de variables

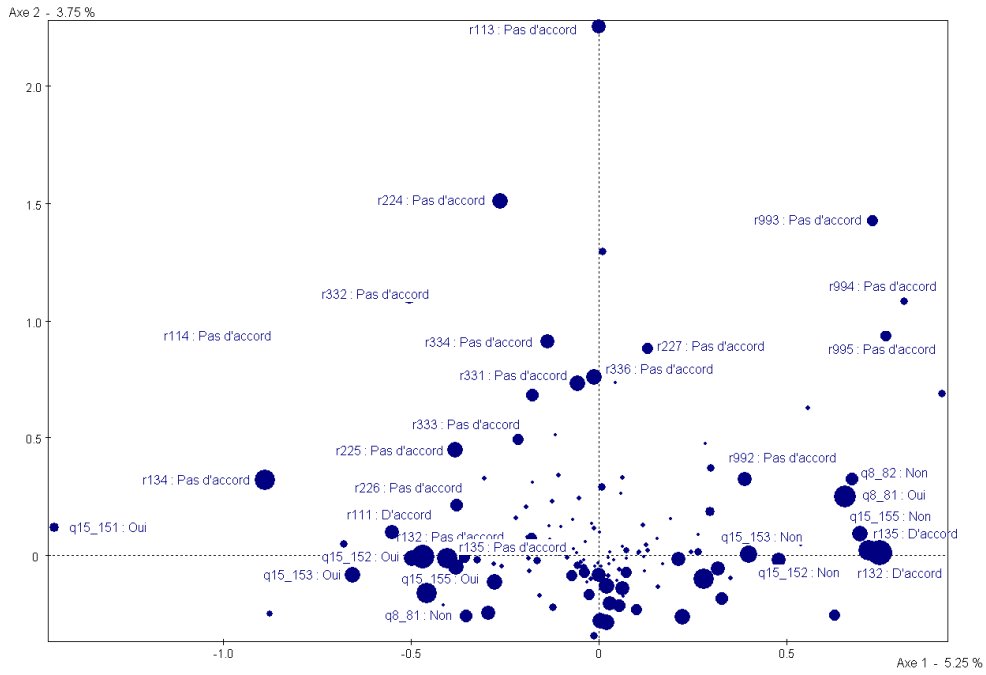
Note : les points sont proportionnels à leur qualité de représentation dans le plan (VS1, VS2)

Comme c'est souvent le cas avec l'ACM, les pourcentages d'inertie portés par les axes sont faibles (5,25% pour F1, 3,75% pour F2 et 3,42% pour F3). Bien que ce faible taux ne soit pas problématique, puisque « les pourcentages d'inertie n'ont qu'un intérêt restreint » (Saporta [26]), cette faiblesse rend le choix du nombre d'axes à conserver difficile (cf. Tableau 9). Ce choix est d'autant plus délicat que les différents éléments mobilisables (pourcentage d'inertie, critère de Kaiser, éboulis des valeurs propres, etc.) ne conduisent pas au même nombre d'axes. Au vu des valeurs propres, on retient 10 axes, ce qui reconstruit près de 30% de l'inertie. Les autres critères conduisent à un nombre beaucoup plus élevé d'axes.

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0,0620	5,25	5,25
2	0,0442	3,75	9,00
3	0,0403	3,42	12,42
4	0,0354	3,00	15,42
5	0,0336	2,85	18,28
6	0,0299	2,53	20,81
7	0,0286	2,42	23,23
8	0,0267	2,26	25,50
9	0,0247	2,10	27,59
10	0,0236	2,00	29,59
…	…	…	…
41	0,0127	1,07	74,79
…	…	…	…
78	0,0005	0,05	99,97
79	0,0004	0,03	100,00

Tableau 9 : Valeurs propres issues de l'ACM sur les 67 variables

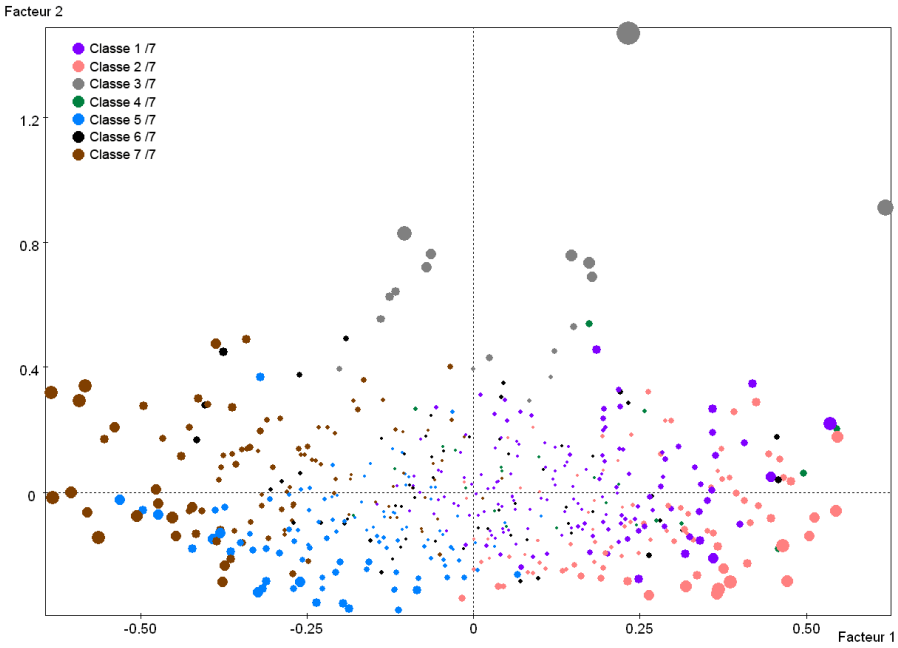
L'interprétation du nuage des modalités dans le premier plan factoriel (cf. Graphique 8) est rendue complexe par le nombre important de modalités, leurs qualités de projection et leurs contributions sur les premiers axes. En effet, les modalités de la totalité des variables sont projetées, contrairement à l'approche *ClustOfVar* où pour chaque variable synthétique, seules les modalités des variables de la classe sont représentées.



Graphique 8 : Nuage des modalités dans le premier plan factoriel issu de l'approche par ACM

Note : les points sont proportionnels à leur qualité de représentation dans le premier plan factoriel. Dans un souci de lisibilité, seules les modalités qui contribuent le plus sont libellées.

Dans un second temps, la classification ascendante hiérarchique qui suit cette analyse factorielle conduit à des classes qui sont moins bien identifiées qu'avec notre approche. L'interprétation de la partition est rendue plus difficile par le nombre de variables utilisées, et ce d'autant plus que les individus sont moins bien projetés sur le premier plan factoriel (cf. Graphique 9).



Graphique 9 : Nuage des individus dans le premier plan factoriel issu de l'approche par ACM

Note : les points sont proportionnels à leur qualité de représentation dans le plan factoriel

Le tableau de contingence du croisement des partitions en 7 classes obtenues respectivement avec l'approche classique « tandem analysis » et avec notre approche montre des ressemblances entre les deux typologies (cf. Tableau 10). Trois classes sont communes : C2_CoV avec C7_TA,C4_CoV avec C5_TA, et C7_CoV avec C4_TA. Les 4 autres classes au contraire sont différentes, à l'instar de la classe C5_CoV ou C6_CoV qui ne sont pas identifiées dans la partition issue de l'approche « tandem analysis ». Or, en termes d'interprétation des résultats, ces classes sont pertinentes et intéressantes pour la compréhension de la perception de l'environnement par les agriculteurs.

	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

Tableau 10 : Croisement des partitions obtenues avec l'approche « tandem analysis » et avec notre approche

À partir du tableau ci-dessus, il est possible de calculer le critère de Rand qui permet de comparer des partitions ayant le même nombre de classes. Notons que Chavent *et al.* [27] ont proposé une version asymétrique de ce critère pour la comparaison de partitions ayant un nombre de classes différent. Comme nous comparons deux partitions sur les mêmes individus, mais issues de méthodes différentes, il est probable que les concordances entre les deux partitions soient dues au hasard. C'est pour cette raison qu'il est plus opportun de retenir dans notre application la version corrigée du critère. Dans notre étude, le pourcentage de concordance entre les deux partitions est relativement faible, puisque le critère de Rand corrigé s'élève à seulement 22%.

Nous poursuivons la comparaison de ces deux partitions en calculant tout d'abord le pourcentage d'inertie expliquée par chacune d'elles. Celui-ci est plus élevé pour l'approche par classification de variables (37%) contre 33% pour l'approche « tandem analysis ». D'autres indices internes de validité sont disponibles (voir par exemple Gordon [28] ou Mirkin [29] pour plus de détails sur ces critères). Pour évaluer ces critères, nous utilisons le package « fpc » proposé par Hennig [30]. Par exemple, la silhouette moyenne de la partition est une mesure populaire qui évalue la correspondance entre une structure de classification et les données à partir desquelles elle a été générée, il s'agit d'un degré de confiance dans l'affectation des observations aux classes. On peut également calculer des mesures de séparation telle que la plus petite distance entre un point de la classe et un point appartenant à une autre classe. Le Tableau 11 présente certains de ces critères qui sont dans l'ensemble peu différents puisque 3 classes parmi 7 sont semblables dans les deux partitions. D'autre part, les classes qui ne sont pas communes sont de qualité similaire. La différence se situe principalement au niveau de l'interprétation.

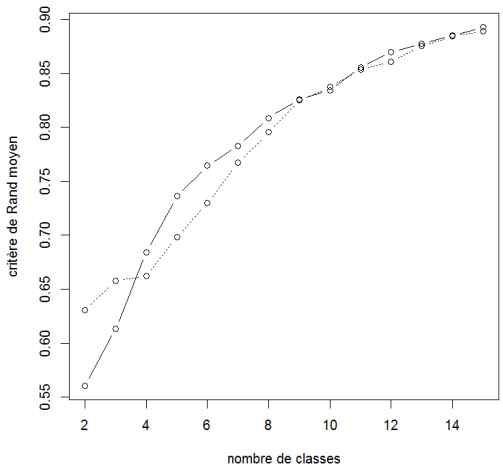
Critères internes de validité	**Approche**	
	« tandem analysis »	***ClustOfVar***
Silhouette moyenne de la partition	0,09	0,11
Maximum des distances entre classes	0,52 0,54 0,71 0,69 0,52 0,64 0,60	0,56 0,59 0,79 0,56 0,61 0,59 0,81

Critères internes de validité	Approche	
	« tandem analysis »	*ClustOfVar*
Plus petite distance entre un point de la classe et un point appartenant à une autre classe	0,09 0,09 0,13 0,12 0,10 0,10 0,09	0,07 0,08 0,16 0,07 0,09 0,11 0,13
Distance moyenne entre un point d'une classe et les points des autres classes	0,37 0,38 0,52 0,48 0,36 0,41 0,40	0,38 0,41 0,48 0,48 0,39 0,41 0,51

Tableau 11 : Critères internes de validité des deux partitions

Afin de tester la stabilité des deux typologies, la méthode du rééchantillonnage par sous-échantillon est utilisée. Selon Mirkin [29] le rééchantillonnage est un moyen de tester la fiabilité d'un résultat de classification et comporte quatre étapes ; générations de copies du jeu de données, lancement de l'algorithme indépendamment sur chaque copie, évaluation des résultats et agrégation. Pour la génération de copies, plusieurs méthodes sont possibles telles que le sous-échantillonnage, la validation croisée, le bootstrap ou encore l'ajout de bruit. Dans cette application, nous décidons de générer les copies grâce à la technique du sous-échantillonnage. Cette méthode, très simple à implémenter, consiste à créer B copies en sélectionnant aléatoirement et sans remise αN entités avec $0 < \alpha < 1$. Dans ce cas, l'algorithme lancé sur chaque copie est le même que celui lancé sur le jeu de données initial. Il s'agit alors de comparer les partitions obtenues sur les sous-échantillons à la partition originale. Cette comparaison est seulement faite sur les αN individus de la typologie initiale (et complète). Pour cela, on peut donc utiliser des indices tels que Rand, Jaccard, Sokal, etc. L'agrégation va consister dans notre cas à calculer l'indice moyen sur l'ensemble des sous-échantillons et sélectionner la typologie la plus stable. L'idée sous-jacente est que si la partition est stable, de légères modifications du jeu de données ne doivent pas fondamentalement modifier les résultats de la classification (structure cachée forte).

Nous appliquons cette approche avec $\alpha = 0,8$ et $B = 100$ c'est-à-dire que l'on tire aléatoirement et sans remise 100 fois 80% de l'échantillon. On obtient donc 100 échantillons de 435 individus chacun à partir desquels on va comparer les deux algorithmes (*ClustOfVar* et « tandem analysis ») pour un nombre de classes allant de 2 à 15 (car nous ne souhaitons pas une partition avec un nombre de classes supérieur). La classification des individus obtenue sur chaque copie est comparée à l'aide du critère de Rand non corrigé, dont les valeurs sont moins pessimistes. L'indice permet de comparer la classification obtenue sur les sous-échantillons à la classification de ces mêmes individus obtenue sur le jeu de données de départ. On obtient donc, pour chaque typologie, et pour chaque nombre de classes, 100 valeurs du critère de Rand que l'on agrège (calcul de la moyenne). Les résultats sont représentés à l'aide du Graphique 10.



Graphique 10 : Critère de Rand moyen pour TA (trait plein) et issue de CoV (pointillés)

On remarque que les deux courbes sont très proches, les deux partitions sont globalement stables puisque les valeurs sont relativement élevées. Plus le nombre de classes augmente, plus le critère se rapproche de la valeur maximale 1 (100% d'accords entre les deux partitions). Ceci peut s'expliquer par le fait que des classes plus petites sont plus homogènes et donc la classification a priori plus stable. La valeur de la stabilité des deux partitions en 7 classes est quasiment identique pour les 2 approches.

4.3. Validation de la partition

Pour valider la partition, nous complétons chacune des approches par une étape explicative, de discrimination, qui permet d'expliquer les règles de construction de ces 7 classes, que ce soit pour « tandem analysis » comme pour notre approche.
La méthode de discrimination employée est celle de la segmentation par arbre. L'utilisation de cette méthode permet de relier une variable à expliquer Y à un ensemble de variables explicatives $X_1, \ldots X_p$, ces variables pouvant être nominales, ordinales ou continues. Ainsi la méthode CART (*Classification And Regression Tree*), développée par Breiman, Friedman, Olsen et Stone en 1984 est analogue à une régression pas à pas, son principe consistant à diviser de façon progressive un échantillon pour obtenir un arbre de classement. Cette méthode comprend deux phases. L'échantillon total est dans un premier temps divisé pour former un échantillon d'apprentissage qui sert à construire l'arbre maximal, par succession de divisions binaires de sorte que par rapport à la variable à expliquer, deux segments descendants soient plus homogènes que le segment parent. L'« arbre est étendu aussi loin que possible sans critère d'arrêt statistique basé sur des niveaux de signification » (Tenenhaus [31]). La deuxième étape consiste à effectuer un élagage de l'arbre maximal à partir de l'échantillon d'apprentissage, pour supprimer les branches de cet arbre qui apportent le moins d'information, c'est-à-dire que cet échantillon-test sélectionne le sous-arbre ayant un coût-complexité et un coût de mauvaise prédiction les plus faibles. Au final cette méthode construit des arbres de décisions binaires ayant de bonnes capacités de prédiction pour l'appartenance à une classe.

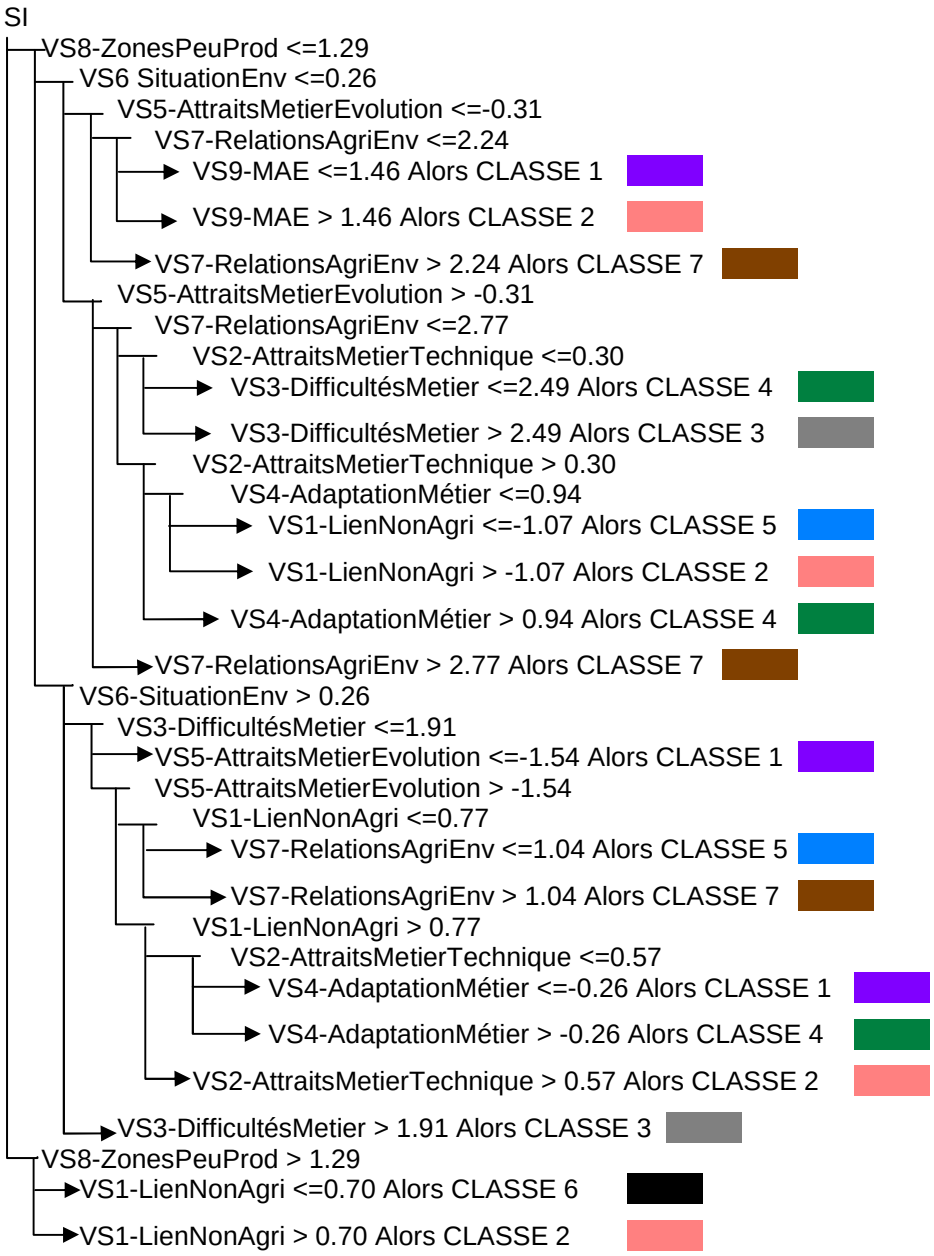
Pour la partition issue de *ClustOfVar*, nous appliquons CART avec comme variables explicatives les variables ayant permis de construire les classes, c'est-à-dire les 9 variables quantitatives synthétiques. Pour comparer les arbres de segmentation issus des deux approches, on utilise pour la partition issue de « tandem analysis » les 10 composantes principales issues de l'ACM. Cependant, l'interprétation de ces composantes principales étant souvent un peu complexe, on utilise en pratique les variables qualitatives initiales. La segmentation est donc également réalisée avec les 67 variables. Pour terminer nous vérifions que le nombre de variables n'impacte pas les résultats de la comparaison en introduisant les 67 variables pour *ClustOfVar* (cf. Tableau 12).

	***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 *<u>remarque</u> : 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

Tableau 12 : Comparaison des règles d'affectation issues de la méthode CART

Nous observons que le nombre de règles avec *ClustOfVar* et les 9 variables synthétiques est faible. Cela signifie que l'affectation des individus est plus simple et que les classes sont donc plus homogènes et distinctes. L'arbre de la segmentation après élagage comporte ainsi peu de feuilles (cf.

Graphique 11). Au contraire, le nombre de règles avec l'approche « tandem analysis » et les 67 variables initiales est plus important. Nous vérifions que le nombre de variables discriminantes n'explique pas ce résultat. Pour cela, nous appliquons *ClustOfVar* avec les 67 variables et nous constatons que le nombre de règles n'a pas augmenté. D'autre part, nous appliquons CART sur « tandem analysis » avec les 10 composantes principales sur lesquelles a été appliquée la classification des observations. Nous observons que le nombre de règles est plus faible et proche de celui de *ClustOfVar* avec les 9 variables synthétiques. La différence est que l'interprétation et la labellisation des composantes principales de l'ACM est plus délicate. Ainsi la lecture des règles d'affectation est plus complexe.



Graphique 11 : Arbre de décision pour la partition issue de l'approche par classification de variables

Lecture : Si la valeur de la variable synthétique 8, qui regroupe les variables relatives aux zones peu productives, est supérieure à 1,29 (pas d'entretien) et si valeur de la variable synthétique 1, qui regroupe les variables relatives aux relations avec le monde agricole, est inférieure à 0,70 (lien difficile), alors l'individu se classe dans la classe 6.

Conclusion

Dans cet article, nous proposons de remplacer la première étape d'ACM précédant en général une typologie des observations par une méthode de classification des variables. Cette approche permet de construire des variables synthétiques qui préservent au mieux les liaisons entre les variables initiales. De plus, cette construction est plus souple qu'en ACM puisqu'on n'impose pas de contraintes d'orthogonalité entre les variables. L'interprétation des variables synthétiques est en outre plus aisée que celle des composantes principales. Nous avons vu qu'elles peuvent être labellisées et lues comme une sorte de gradients.

Sur ce jeu de données, l'utilisation d'une approche de classification de variables permet de simplifier l'interprétation au niveau de l'analyse des relations entre les variables d'une part et les modalités d'autre part. De plus, la compréhension de la formation des groupes d'individus au travers des variables synthétiques construites est facilitée. Par ailleurs, certaines classes formées suite à la classification de variables sont plus intéressantes. Cependant au niveau des critères internes de validité, la comparaison des deux partitions montre des valeurs très proches. Cela signifie que les deux partitions sont « statistiquement » proches. Les différences que l'on observe se situent principalement au niveau de l'interprétation des classes d'agriculteurs.

Concernant les 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. Aussi allons-nous soumettre un autre article afin d'affiner l'interprétation de la typologie succinctement abordée dans celui-ci à dominante statistique.

Pour conclure, la démarche méthodologique proposée peut être considérée comme une « tandem analysis » puisque nous réalisons les deux étapes de réduction de façon indépendante. Une perspective à ce travail consisterait à proposer une méthode qui optimiserait simultanément le critère d'homogénéité proposé dans l'approche de classification de variables et le critère relatif à la typologie des observations.

Bibliographie

[1] DeSarbo W.*, et al.*, «Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups»*, Marketing Letters*, vol 2, n°2, pp 129-146, 1991.

[2] De Soete G., Carroll J.D., «K-means clustering in a low-dimensional Euclidean Space», in *New Approaches in Classification and Data Analysis,* Diday E.*, et al.* (Eds.), 1994

[3] Vichi M., Kiers H.A.L., «Factorial k-means analysis for two-way data»*, Computational Statistics & Data Analysis*, vol 37, n°1, pp 49-64, 2001.

[4] Diday E.*, et al.* (Eds.), «Optimisation en classification automatique», Le Chesnay, 1979

[5] Timmerman M.*, et al.*, «Factorial and reduced K-means reconsidered»*, Computational Statistics & Data Analysis*, vol 54, n°7, pp 1858-1871, 2010.

[6] Heiser W.J., «Information and classification», in *Information and Classification,* Opitz O., Lausen B., Klar R. (Eds.), 1993

[7] DeSoete G., Heiser W.J., «A latent class unfolding model for analyzing single stimulus preference ratings»*, Psychometrika*, vol 58, n°545-565, 1993.

[8] Vichi M., Saporta G., «Clustering and Disjoint Principal Component Analysis»*, Computational Statistics & Data Analysis*, vol 53, n°8, pp 3194-3208, 2009.

[9] Govaert G., Nadif M., «Un modèle de mélange pour la classification croisée d'un tableau de données continues », *CAP'09, 11e conférence sur l'apprentissage artificiel*, Hammamet, Tunisie, 2009

[10] Martella F., Alfò M., Vichi M., «Hierarchical mixture models for biclustering in microarray data 2011»*, Statistical Modelling*, vol 11, n°6, pp 489-505, 2010.

[11] Chavent M.*, et al.*, «Clustering of variables via the PCAMIX method», *International Classification Conference*, St Andrews, Ecosse, 2011

[12] Plasse M.*, et al.*, «Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set»*, Computational Statistics & Data Analysis*, vol 52, n°596-613, 2007.

[13] Gelein B., Sautory O., «Classification de variables - Application à la base permanente des équipements», *Xe Journées de Méthodologie statistique* Paris, 2009

[14] Abdallah H., Saporta G., «Classication d'un ensemble de variables qualitatives», *Revue de Statistique Appliquée*, vol 46, n°4, pp 5-26, 1998.
[15] Vigneau E., Qannari E.M., «Clustering of variables around latent components», *Communications in statistics Simulation and Computation*, vol 32, n°4, pp 1131-1150, 2003.
[16] Dhillon I., Marcotte E., Roshan U., «Diametrical Clustering for Identifying Anticorrelated Gene Clusters», *Bioinformatics*, vol 19, n°13, pp 1612-1619, 2003.
[17] Lerman I.C., «Foundations of the likelihood linkage analysis classification method», *Applied Stochastics Models and Data Analysis*, vol 7, n°1, pp 63-76, 1990.
[18] Lerman I.C., «Likelihood linkage analysis classification method : An example treated by hand», *Biochimie*, vol 75, n°5, pp 379-397, 1993.
[19] Kiers H., «Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables», *Psychometrika*, vol 56, n°2, pp 197-212, 1991.
[20] Pagès O., «Analyse Factorielle de Données», *Revue de Statistique Appliquée*, vol 52, n°4, pp 93-111, 2004.
[21] Chavent M.*, et al.*, «ClustOfVar: An R Package for the Clustering of Variables», *The R User Conference*, University of Warwick, Coventry, UK, 2011
[22] Escofier B., «Traitement simultané de variables qualitatives et quantitatives en analyse Factorielle», *Les cahiers de l'analyse des données*, vol 4, n°2, pp 137-146, 1979.
[23] Saporta G., «Simultaneous treatment of quantitative and qualitative data», *Atti 35° Riunione Scientifica della Societa Italiana di Statistica*, 1990
[24] Hubert L., Arabie P., «Comparing partitions», *Journal of Classification*, vol 2, n°1, pp 193-218, 1985.
[25] Candau J.*, et al.*, «La prise en compte de l'environnement par les agriculteurs. Résultats d'enquête», City, 83, 2005
[26] Saporta G., «Probabilités, analyse des données et statistique», Editions Technip, Paris, 1990
[27] Chavent M., Lacomblez C., Patouille B., «Critère de Rand asymétrique - Application en chimie organique», *8èmes Rencontres de la Société Francophone de Classification (SFC01)*, Point à Pitre, 2001
[28] Gordon A.D., «Classification», Chapman & Hall, 1999
[29] Mirkin B., «Clustering for Data Mining: A Data Recovery Approach», Chapman and Hall/CRC, 2005
[30] Hennig C., «Package R 'fpc' Flexible procedures for clustering», 2012
[31] Tenenhaus M., «Statistique : Méthodes pour décrire, expliquer et prévoir», DUNOD, 2007

Annexes

Les annexes figurent dans les pages suivantes.

- Annexe 1 : Liste des variables de l'enquête sur la prise en compte de l'environnement par les agriculteurs

- Annexe 2 : Arbre de décision pour la partition issue de l'approche par ACM

Annexe 1 : Liste des variables de l'enquête sur la prise en compte de l'environnement par les agriculteurs

Variable	Libellé court	Libellé	Nombre de modalités
r111	avis_perception_non_agri	Votre activité est perçue positivement par les non agriculteurs	2
r112	avis_activite	Votre activité est motivante	2
r113	avis_changement_metier	Votre métier connaît de profonds changements	2
r114	avis_avenir_activite	Vous êtes inquiet pour l'avenir de votre activité	2
r221	difficulte_hausse_tourisme	Aujourd'hui l'augmentation de la fréquentation touristique est une difficulté	2
r222	difficulte_baisse_nb_agri	Aujourd'hui la baisse du nombre d'agriculteurs est une difficulté	2
r223	difficulte_formation	Aujourd'hui les besoins en formation professionnelle sont une difficulté	2
r224	difficulte_paperasserie	Aujourd'hui la paperasserie est une difficulté	2
r225	difficulte_environnement	Aujourd'hui les contraintes environnementales sont une difficulté	2
r226	difficulte_voisinage	Aujourd'hui les relations avec les voisins non agriculteurs sont une difficulté	2
r227	difficulte_transmission	Aujourd'hui la transmission des exploitations est une difficulté	2
r331	difficulte_main_oeuvre	Il existe des difficultés liées à la main d'œuvre	2
r332	difficulte_normes	Il existe des difficultés liées à la mise aux normes	2
r333	difficulte_agrandir	Il existe des difficultés liées à la nécessité de s'agrandir	2
r334	difficulte_prix_terres	Il existe des difficultés liées au prix des terres	2
r335	difficulte_prix_produits	Il existe des difficultés liées au prix et à la commercialisation des produits	2
r336	difficulte_temps_travail	Il existe des difficultés liées au temps de travail	2
q5_51	attrait_metier_pointe_technique	Être à la pointe de la technique est l'un des trois attraits principaux du métier	2
q5_52	attrait_metier_patrimoine	Être détenteur d'un patrimoine est l'un des trois attraits principaux du métier	2
q5_53	attrait_metier_nature	Être en contact avec la nature est l'un des trois attraits principaux du métier	2
q5_54	attrait_metier_independance	Être indépendant est l'un des trois attraits principaux du métier	2
q5_55	attrait_metier_evolution	Évoluer constamment est l'un des trois attraits principaux du métier	2
q5_56	attrait_metier_histoire_familiale	Perpétuer l'histoire familiale sur la région est l'un des trois attraits principaux du métier	2
q6_61	finalite_metier_entreprendre	Entreprendre en s'adaptant au marché est l'une des trois plus importantes finalités du métier	2
q6_62	finalite_metier_entretien_batiment	Entretenir les bâtiments anciens est l'une des trois plus importantes finalités du métier	2
q6_63	finalite_metier_subvenir_famille	Faire vivre la famille est l'une des trois plus importantes finalités du métier	2
q6_64	finalite_metier_trasmission	Maintenir et transmettre l'exploitation est l'une des trois plus importantes finalités du métier	2
q6_65	finalite_metier_nourrir_hommes	Nourrir les hommes est l'une des trois plus importantes finalités du métier	2
q6_66	finalite_metier_adaptation_attentes	Produire en s'adaptant aux attentes de la société est l'une des trois plus importantes finalités du métier	2
q6_67	finalite_metier_protection	Protéger les ressources naturelles et le paysage est l'une des trois plus importantes finalités du métier	2
q8_81	environnement_exageration_problemes	On exagère la gravité des problèmes de l'environnement	3
q8_82	environnement_inquietant	La situation de l'environnement inquiétante	3

Variable	Libellé court	Libellé	Nombre de modalités
r991	environnement_problemes_agriculteurs	Les problèmes d'environnement sont l'affaire des agriculteurs	2
r992	environnement_problemes_associations	Les problèmes d'environnement sont l'affaire des associations de protection	2
r993	environnement_problemes_consommateur	Les problèmes d'environnement sont l'affaire de chaque consommateur	2
r994	environnement_problemes_industriels	Les problèmes d'environnement sont l'affaire des industriels	2
r995	environnement_problemes_pouvoirs_publics	Les problèmes d'environnement sont l'affaire des pouvoirs publics	2
r996	environnement_problemes_personne	Les problèmes d'environnement ne sont l'affaire de personne car il n'y a pas de problème	2
r101	20ans_lien_agroalimentaire	Dans 20 ans, l'agriculture sera plus liée à l'agroalimentaire et devra respecter des normes de qualité	3
r102	20ans_proche_agriculture_bio	Dans 20 ans, l'environnement sera au cœur de l'agriculture avec des systèmes proches de l'agriculture biologique	3
r103	20ans_gestion_europe_et_region	Dans 20 ans, l'Europe donnera le cadre général de la production et de l'environnement et la Région gèrera les objectifs plus précis	3
r104	20ans_intensive_et_preservation	Dans 20 ans, il y aura d'un côté des zones intensives vouées à la production et de l'autre des zones vouées à la préservation	3
r121	accord_maitrise_nature	Je dois maîtriser la nature pour mon activité	2
r122	accord_adaptation_nature	Je dois m'adapter à la nature	2
r123	accord_entretien	Je dois entretenir les parties peu productives de mon exploitation pour qu'elles soient propres	2
r124	accord_developpement_nature	J'évite d'intervenir sur les parties peu productives de mon exploitation pour laisser se développer la nature	2
r125	accord_diminution_travail	Je n'interviens pas sur les parties peu productives de mon exploitation pour diminuer mon travail	2
r131	mesures_activite_techniques_pointe	Les mesures pour la protection de l'environnement demandent de maîtriser des techniques de pointe	2
r132	mesures_activite_empeche_progression	Les mesures pour la protection de l'environnement empêchent de progresser	2
r133	mesures_activite_retour_savoirfaire	Les mesures pour la protection de l'environnement incitent à revenir à des savoir-faire anciens	2
r134	mesures_activite_limite_action	Les mesures pour la protection de l'environnement limitent la liberté d'action	2
r135	mesures_activite_domaines_particuliers	Les mesures pour la protection de l'environnement touchent des domaines qui ne regardent que vous	2
q15_151	mesures_environnement_incitation_jeunes	Les mesures en faveur de la protection de l'environnement incitent les jeunes à s'installer en agriculture	3
q15_152	mesures_environnement_amelioration_qualite	Les mesures en faveur de la protection de l'environnement permettent d'améliorer la qualité des produits	3
q15_153	mesures_environnement_renforcement_solidarite	Les mesures en faveur de la protection de l'environnement renforcent la solidarité entre milieux agricole et non-agricoles	3
q15_154	mesures_environnement_limite_production	Les mesures en faveur de la protection de l'environnement sont un bon moyen de limiter la production	3
q15_155	mesures_environnement_valorisation_agriculture	Les mesures en faveur de la protection de l'environnement valorisent l'image de	3

Variable	Libellé court	Libellé	Nombre de modalités
		l'agriculture	
q15_156	mesures_environnement_image_ancienne	Les mesures en faveur de la protection de l'environnement véhiculent une image ancienne de l'agriculture	3
q18_181	mesures_difficulte_changement_technique	Les changements techniques sont l'une des trois difficultés principales dans l'application des mesures agri	2
q18_182	mesures_difficulte_charge_travail	La charge de travail est l'une des trois difficultés principales dans l'application des mesures agri	2
q18_183	mesures_difficulte_controle	Les contrôles sont l'une des trois difficultés principales dans l'application des mesures agri	2
q18_184	mesures_difficulte_efficacite	L'efficacité des mesures est l'une des trois difficultés principales dans l'application des mesures agri	2
q18_185	mesures_difficulte_montant_aide	Le faible montant de l'aide est l'une des trois difficultés principales dans l'application des mesures agri	2
q18_186	mesures_difficulte_investissement	L'investissement financier est l'une des trois difficultés principales dans l'application des mesures agri	2
q18_187	mesures_difficulte_manque_formation	Le manque de formation adaptée est l'une des trois difficultés principales dans l'application des mesures agri	2
q18_188	mesures_difficulte_paperasserie	La paperasserie est l'une des trois difficultés principales dans l'application des mesures agri	2
q18_189	mesures_difficulte_solidarite_agriculteurs	La solidarité entre agriculteurs est l'une des trois difficultés principales dans l'application des mesures agri	2

Annexe 2 : Arbre de décision pour la partition issue de l'approche par ACM