Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n2p170

ClustOfVar-based approach for unsupervised learning: Reading of synthetic variables with sociological data

By Kuentz-Simonet, Lyser, Candau, Deuffic

Published: 14 October 2015

Electronic Journal of Applied Statistical Analysis
Vol. 08, Issue 02, October 2015, 170-197
DOI: 10.1285/i20705948v8n2p170

ClustOfVar-based approach for unsupervised learning: Reading of synthetic variables with sociological data

Vanessa Kuentz-Simonet*, Sandrine Lyser, Jacqueline Candau, and Philippe Deuffic

Irstea, UR ETBX
50 avenue de Verdun Gazinet Cestas, F-33612, France

Published: 14 October 2015

This paper proposes an original data mining method for unsupervised learning, replacing traditional factor analysis with a system of variable clustering. Clustering of variables aims to group together variables that are strongly related to each other, i.e. containing the same information. We recently proposed the *ClustOfVar* method, specifically devoted to variable clustering, regardless of whether the variables are numeric or categorical in nature. It simultaneously provides homogeneous clusters of variables and their corresponding synthetic variables that can be read as a kind of gradient. In this algorithm, the homogeneity criterion of a cluster is defined by the squared Pearson correlation for the numeric variables and by the correlation ratio for the categorical variables. This method was tested on categorical data relating to French farmers and their perception of the environment. The use of synthetic variables provided us with an original approach of identifying the way farmers reconfigured the questions put to them.

keywords: environment, variable clustering, *ClustOfVar*, synthetic variables, typology of farmers.

1 Introduction

Today, the environmental imperative plays an important role in redefining the way farmers see their profession. However, do they see environmental issues as the overriding

*Corresponding author: vanessa.kuentz-simonet@irstea.fr.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

concern, or do other factors exert some influence over this perception? While the implementation of agri-environmental measures (AEMs [1]) can be considered as a binary variable (i.e. the farmer either does or does not sign a contract agreeing to introduce such measures) the same cannot be said of the extent to which farmers actually consider environmental issues in their day-to-day activities. There is no viable way in which to measure such considerations. The aim of this study is to examine how farmers perceive environmental protection relating to their profession. What meanings do they attribute to the environment? Which aspects of their occupation and of their relationship to nature are challenged by the integration of environmental protection into agricultural policy? We address this issue based on a quantitative survey carried out by the National Research Institute of Science and Technology for Environment and Agriculture (Irstea) in 2005 on behalf of the National centre for development and farm structures.

The paper adopts a two-steps exploratory statistical approach, which allows us to identify opinion trends among the farmers interviewed. Specifically, we opt for a variable clustering method, providing synthetic variables (SVs) that can be read as gradients. After finding groups of variables, the second step is to cluster farmers via classic clustering algorithm (Ward's ascendant hierarchical clustering). By this way, our sequential approach is quite similar to "tandem analysis" (Arabie and Hubert, 1994). Note that other alternatives are built on a simultaneous clustering of both observations and variables. We can mention the "factorial k-means" of Vichi and Kiers (2001) or the "disjoint clustering and principal component analysis" of Vichi and Saporta (2009) (see Charrad and Ben Ahmed (2011) for a review of bi-partitioning methods).

In our approach the first step of variable clustering to construct synthetic variables is central to identify elements of interpretation regarding the different types of relationships farmers have with the environment. Variable clustering aims to group together variables that are strongly related to each other, i.e. containing the same information. The idea is to build homogeneous (the meaning of this term will be explained later) clusters of variables. Another objective of variable clustering is to avoid redundancies between variables by reducing the size of datasets. The clustering of variables is thus an alternative to usual factor analysis methods such as Principal Component Analysis (PCA) or Multiple Correspondence Analysis (MCA). In such cases, once the variables are clustered into homogeneous groups, the user may decide to either choose one variable from each group or to construct a synthetic variable. One simple and commonly-used approach to clustering sets of variables is to calculate the dissimilarities between those variables and to apply a traditional cluster analysis method to this dissimilarity matrix. For numeric variables, many measures of dissimilarity involve the coefficient of correlation. For categorical variables, many measures of association can be used, such as: χ^2, Rand, Belson, Jordan, and many others (Abdallah and Saporta, 1998). Besides these traditional methods devoted to the clustering of observation units, other methods exist that are specifically devoted to the clustering of variables. For numeric variables, the best-known method is probably the VARCLUS function of the SAS computer application (SAS Institute Inc., 2013). This procedure divides a set of numeric variables into

[1] AEMs refers to regulatory measures or incentives aiming at protecting the environment.

disjoint or hierarchical clusters. Another approach is to use a clustering algorithm that simultaneously provides clusters of variables and their corresponding synthetic variables. Two similar partitioning algorithms already exist for clustering numeric variables and are based on PCA: the Clustering of variables around Latent Variables (CLV) approach (Vigneau and Qannari, 2003; Vigneau and Chen, 2015) and the Diametrical clustering method (Dhillon et al., 2003). To our knowledge, the clustering of categorical variables has been the subject of less research. It is possible in the CLV approach to integrate categorical variables by their indicator matrix but we get thus clustering of categories. We can mention among others the likelihood linkage analysis method (Lerman, 1990, 1993), which carries out a hierarchical clustering algorithm for numeric or categorical variables. We recently proposed the *ClustOfVar* method, specifically devoted to variable clustering, regardless of whether the variables are numeric or categorical in nature (Chavent et al., 2012a). This approach uses the PCAMIX method (Kiers, 1991), which we re-designed and reprogrammed based on singular value decomposition (Chavent et al., 2012b). More specifically, PCAMIX is a principal component method for a mixture of numeric and categorical variables. It includes the ordinary PCA and MCA as special cases. Two variable clustering algorithms are available in *ClustOfVar*: a hierarchical ascendant algorithm and a k-means type partitioning algorithm. Both algorithms are designed to maximise the same homogeneity criterion, based on the square of the Pearson correlation for numeric variables and on the correlation ratio for categorical variables. By rearranging the variables into homogeneous clusters, the clustering approach simultaneously constructs synthetic variables. Note that regardless of the type of initial data, these synthetic variables are always numeric and can be read as a kind of gradient. Because the application we describe is based on categorical variables, the *ClustOfVar* method for variable clustering is only described for this type of variable. This approach offers more flexibility than the factor analysis because it does not impose orthogonality constraints on the synthetic variables. Each synthetic variable can be read as a sort of gradient. These SVs are easier to interpret and label than the principal components since they (the SVs) only refer to the variables in the corresponding cluster. In this article, we focus on the hierarchical ascendant algorithm, since we have no prior idea of the number of clusters of variables. The proposed algorithms are implemented in the R package *ClustOfVar* available on the CRAN (Comprehensive R Archive Network).

The variable clustering approach is presented in Section 2 and the data are described in Section 3. The methodology was initially proposed to address the complex issue of the extent to which the environment is considered by French farmers. Its application to categorical data is presented in Section 4. Finally, Section 5 discusses the relevance of the approach and provides some conclusions.

2 Variable clustering approach - The case of categorical variables

Let $\{\mathbf{z}_1, \ldots, \mathbf{z}_p\}$ be a set of p categorical variables. We note $\mathbf{Z}$ the corresponding categorical data matrix, of dimension $n \times p$, where n is the number of observation units. For

simplicity, we note $\mathbf{z}_j \in \mathcal{M}_j^n$ the jth column of $\mathbf{Z}$ with $\mathcal{M}_j$ the set of categories of $\mathbf{z}_j$. This section addresses the problem of partitioning a set of p categorical variables into K disjoint clusters. We denote such a partition $P_K = (C_1, \ldots, C_K)$.

2.1 Homogeneity $\mathcal{H}$ of a partition P_K of categorical variables

The homogeneity of a partition P_K of the p categorical variables is defined as the sum of the homogeneity of its clusters:

$$\mathcal{H}(P_K) = \sum_{k=1}^{K} H(C_k), \tag{1}$$

where $H(C_k)$ measures the homogeneity of the cluster C_k of P_K. It is a measure of adequacy between the variables in the cluster and its synthetic numeric variable denoted $\mathbf{y}_k \in \mathcal{R}^n$ (which will be defined in Subsection 2.2) :

$$H(C_k) = \sum_{\mathbf{z}_j \in C_k} \eta^2_{\mathbf{y}_k|\mathbf{z}_j}, \tag{2}$$

where $\eta^2_{\mathbf{y}_k|\mathbf{z}_j} \in [0,1]$ denotes the correlation ratio between $\mathbf{y}_k$ and $\mathbf{z}_j$. It measures the part of the variance of $\mathbf{y}_k$ explained by the categories of $\mathbf{z}_j$:

$$\eta^2_{\mathbf{y}_k|\mathbf{z}_j} = \frac{\sum_{s \in \mathcal{M}_j} n_s (\bar{y}_k^s - \bar{y}_k)^2}{\sum_{i=1}^{n} (y_{k,i} - \bar{y}_k)^2},$$

where n_s is the frequency of category s, $\bar{y}_k^s$ is the mean value of $\mathbf{y}_k$ calculated on the observations belonging to category s, $y_{k,i}$ is the observed value of $\mathbf{y}_k$ for unit i, et $\bar{y}_k$ is the mean of $\mathbf{y}_k$.

In other words $H(C_k)$ measures the link between the categorical variables of C_k and the synthetic numeric variable $\mathbf{y}_k$.

2.2 Definition of the synthetic variable $\mathbf{y}_k$ of a cluster C_k of categorical variables

The synthetic variable $\mathbf{y}_k \in \mathcal{R}^n$ of a cluster C_k is defined as the numeric variable the "most linked" to all the variables in the cluster. It maximises the homogeneity of C_k and is then solution of the following optimisation problem:

$$\mathbf{y}_k = \arg\max_{\mathbf{u} \in \mathcal{R}^n} \left\{ \sum_{\mathbf{z}_j \in C_k} \eta^2_{\mathbf{u}|\mathbf{z}_j} \right\}.$$

We can show that:

- $\mathbf{y}_k$ is the first principal component of MCA applied to $\mathbf{Z}_k$, the matrices made up of the columns of $\mathbf{Z}$ corresponding to the variables in C_k.

- The empirical variance of $\mathbf{y}_k$ is then equal to: $\text{Var}(\mathbf{y}_k) = \sum_{\mathbf{z}_j \in C_k} \eta^2_{\mathbf{y}_k|\mathbf{z}_j} = \lambda^1_k$, the first eigenvalue issued from MCA applied to cluster C_k.

It follows that the homogeneity of a cluster is simply defined by $H(C_k) = \lambda^1_k$. Then the homogeneity of a partition P_K is equal to $\mathcal{H}(P_K) = \lambda^1_1 + \cdots +, \lambda^1_K$.

2.3 Calculation of the synthetic variable of C_k using MCA

We note p_k the number of variables in cluster C_k and m the total number of categories of the variables in C_k. We respectively provide $\mathcal{R}^n$ with the metric $\mathbf{N} = \text{diag}(\frac{1}{n})$ and $\mathcal{R}^m$ with the metric $\mathbf{M} = \text{diag}(\frac{n_s}{np_k})$. We apply MCA to the matrix $\mathbf{Z}_k$. To this end, we carry out the Singular Value Decomposition (SVD) of the matrix $\frac{1}{p_k}\mathbf{JG}$, where $\mathbf{G}$ denotes the indicator matrix of $\mathbf{Z}_k$ and $\mathbf{J}$ is the centering operator: $\mathbf{J} = \mathbf{I}_n - \mathbf{1}\mathbf{1}'/n$ with $\mathbf{I}_n$ the identity matrix of dimension n and $\mathbf{1}$ the row vector with unit entries. We obtain:

$$\mathbf{Z}_k = \mathbf{U}_k \Lambda_k^{1/2} \mathbf{V}'_k,$$

where $\mathbf{U}'_k\mathbf{U}_k = \mathbf{V}'_k\mathbf{V}_k = \mathbf{I}_r$ with r the rank of $\mathbf{Z}_k$ and Λ_k the matrix of the eigenvalues $\lambda^1_k, \ldots, \lambda^r_k$ ranged in decreasing order.

The matrix of the principal component scores of dimension $n \times r$ is given by $\mathbf{U}_k\Lambda_k$. The synthetic variable $\mathbf{y}_k$ corresponds to the first column of this matrix:

$$\mathbf{y}_k = \lambda^1_k \mathbf{u}^1_k, \tag{3}$$

where $\mathbf{u}^1_k$ denotes the first column of $\mathbf{U}_k$.

2.4 The hierarchical ascendant clustering algorithm of categorical variables

The aim is to find a partition of the set of categorical variables in which variables are strongly related to the other variables belonging to their cluster. In other words the objective is to define a partition P_K which maximises the homogeneity criterion $\mathcal{H}$ defined in (1). For this, a hierarchical ascendant clustering algorithm is proposed. It builds a set of p nested partitions of variables in the following way:

1. Step $l = 0$: initialization. Start with the partition into singletons (p clusters).
2. Step $l = 1, \ldots, p-2$: aggregate two clusters of the partition into $p - l + 1$ clusters to get a new partition into $p - l$ clusters. For this, choose clusters A and B with the smallest dissimilarity defined as:

$$d(A, B) = H(A) + H(B) - H(A \cup B) = \lambda^1_A + \lambda^1_B - \lambda^1_{A\cup B}. \tag{4}$$

We can prove that $\lambda^1_{A\cup B} \leq \lambda^1_A + \lambda^1_B$, which implies that the merging of two clusters A and B at each step results in a decrease of criterion $\mathcal{H}$. Then this

dissimilarity measures the loss of homogeneity observed when the two clusters are merged. Therefore the strategy consists in merging the two clusters that result in the smallest decrease in $\mathcal{H}$. Using this aggregation measure the new partition into $p-l$ clusters maximises $\mathcal{H}$ among all the partitions into $p-l$ clusters obtained by amalgamation of two clusters of the partition into $p-l+1$ clusters.

3. Step $l = p-1$: stop. A single cluster consisting of all variables is obtained.

The height of a cluster $C = A \cup B$ in the tree is defined as $h(C) = d(A, B)$. It is easy to verify that $h(C) \geq 0$ but the property "$A \subset B \Rightarrow h(A) \leq h(B)$" has not been proved yet. Nevertheless, inversions in the tree have never been observed in practice neither on simulated data nor on real data sets. This approach provides a tree which enables the user to see the successive aggregations between the variables, and provides a graphical illustration to aid in selecting the number of clusters to be used.

2.5 The reading of the synthetic variable as a sort of gradient

The information carried by the variables of the cluster are summarised by its synthetic variable. Indeed $\mathbf{y}_k$ defined in (3) gives the coordinates of the observation units on the synthetic variable of cluster C_k.

Having applied MCA to calculate the synthetic variable of each cluster of variables C_k, we can also obtain the matrix $\mathbf{A}_k$ of the coordinates of the m_k categories of the p_k variables on the principal components:

$$\mathbf{A}_k = \mathbf{M}^{-1}\mathbf{V}_k\Lambda_k. \tag{5}$$

Then we can calculate the matrix $\mathbf{C}_k = (c_{jl}), j = 1, \ldots, p_k; l = 1, \ldots, r$ which contains the "squared loadings" of the variables. More precisely, $\mathbf{C}_k$ is obtained from the matrix $\mathbf{A}_k$ as follows :

$$c_{jl} = \sum_{s \in I_j} a_{sl}^2, \tag{6}$$

where I_j is the set of row indices of $\mathbf{A}_k$ associated with the categories of the categorical variable j. It is equal to the correlation ratio between the variable j and the lth principal component.

Both matrices $\mathbf{A}_k$ and $\mathbf{C}_k$ play a fundamental role in the interpretation and label of the synthetic variables of the clusters. Indeed they enable to get the same kind of interpretation rules as in MCA. The main difference is that these two matrices are defined inside a cluster. Their dimension is thus lower (in the sense that they only focus on variables of the corresponding cluster and not the whole set of variables), making them easier to read. Because the synthetic variable of a cluster is the first principal component of MCA applied to the cluster, we are only interested in the first column of these matrices. The first column of $\mathbf{C}_k$ is useful to identify those variables with the strongest link to each synthetic variable in terms of correlation ratio. Besides, the first column of $\mathbf{A}_k$ gives the coordinates of the categories of the variables on the synthetic variable.

As previously mentioned, the *ClustOfVar* method simultaneously constructs groups of categorical variables and their synthetic variables that are numeric. Then the detailed reading of the SVs does not appear straightforward at first sight. The idea of the proposed methodology is to read each synthetic variable as a sort of gradient. Obviously this requires that the correlation ratios of the cluster are high, meaning that the initial categorical variables in the cluster are strongly linked. In this case, there are a finite number of possible values for synthetic variables. Therefore by setting up the coordinates of the categories of the categorical variables in the cluster on this synthetic variables, regrouping of categories appear. This enables to interpret the numeric values of the synthetic variable and consequently to label it. This interpretation is simplified by the fact that each SV only relates to variables of the corresponding cluster. Indeed, only the categories of the variables in the cluster have some coordinates on the synthetic variable (contrary to MCA where, for each component, we visualise the categories of all variables). On the other hand, as *ClustOfVar* is a variable clustering method, all categories of each variable have to be in the same group. From this point of view, the method can be seen as "constrained clustering", contrary to another type of approach, such as the CLV method (Vigneau and Qannari, 2003; Vigneau and Chen, 2015), that performs a cluster analysis of the categories, independently of the variable to which the categories belong.

3 Description of the data

In 2005, during a period of reform of common agricultural policy, sociologists from Irstea Bordeaux conducted a nationwide survey using mail-out questionnaires for French farmers. At one time, the role of agriculture was simply to feed the population. However, over the past thirty years, its scope has expanded to include protecting natural resources and preserving the vitality of rural areas. This concept of "multifunctional agriculture" first emerged in 1992 at the Rio Summit. In view of the changing role of agriculture, the Irstea survey was based on a simple question: "Do farmers take the environment into account?" The response to this question cannot be summarised as a simple binary variable. It is a complex issue, which is continually changing as a result of new environmental and health standards, terms and conditions attached to state subsidies, and environmental concerns from both inside and outside the farming community. Therefore, a shift towards environmentally-friendly agriculture can be seen as a "technical, cognitive and structural change" (Candau et al., 2005). Farmers may also be faced with more pressing challenges than those relating to the environment. To reflect the complex nature of the question, the questionnaire was divided into four main sections. Questions related to perception of:

1. The farming profession, described through general questions about their activity (q1_1 to q1_4) or through more specific variables, related to the attractiveness of the job (q5_1 to q5_6), the objectives being pursued (q6_1 to q6_7) or general difficulties in carrying out their work (q2_1 to q2_7 and q3_1 to q3_6).

2. The environment, using questions about environmental problems (q9_1 to q9_6) and the assessment of their severity (q8_1 and q8_2).

3. Nature and the environment, specifically addressing the relationship between farming and the environment over the next 20 years (q10_1 to q10_4) and the relationship that farmers have with nature (q12_1 to q12_5).

4. Agri-environmental measures (AEMs), using questions about farmers' opinions on those measures (q13_1 to q13_5, q15_1 to q15_6) and difficulties in implementing them (q18_1 to q18_9).

The questionnaire ends with questions about farmers' socio-economic characteristics: age, education, visits or direct sales, farm size, multiple activities, professional, community or municipal responsibilities (see Table 1). We chose these characteristics because, while controversial, their relevance has been established in existing literature (Burton, 2014). As most of these previous studies were based on methods of qualitative research (mainly in-depth interviews), our questionnaire survey allowed us to test these characteristics on a larger scale.

In this article, we focused on farmers' perception of the environment, which are contained within a specific dataset. We chose to analyse variables directly related to the protection of the environment. The dataset consists of 67 variables addressing how farmers perceive their activity and the environment (see list in appendix). These variables (categorical, with two or three categories) are used for variable clustering in order to construct synthetic variables. Socio-economic variables are not introduced at this stage but will be used later to characterise specific groups of farmers.

4 Results

4.1 Choosing the number of synthetic variables

The R package called *ClustOfVar* (Chavent et al., 2011), available on the CRAN, was used for the hierarchical ascendant clustering of variables. The tree produced by this clustering (Figure 1) illustrates the successive aggregations of all variables and helps visualise the links between them. Note that the aggregation criterion is used as the relevant node height in the tree. Thus, in addition to observing the tree, the progressively increasing level of aggregation (Figure 2) provides a suitable tool to select the number of clusters when partitioning variables. At each step of the ascendant clustering algorithm, this criterion measures the loss in homogeneity when two clusters of variables are merged. The elbow shape in the curve corresponds to the aggregation of very different clusters. However, choosing a number of clusters of variables from these two plots is difficult. It is not easy to detect either a "jump" in the hierarchical tree (Figure 1) or a clear "break" in Figure 2. However, the choice of the number of synthetic variables is not only based on statistical arguments. In our case study, the emphasis is on understanding the clusters of variables and analysing them in connection with the issue at stake. The purpose of clustering is to group variables that are strongly related, that is to say, those

that are linked by the way people responded to questions. It is interesting to note that despite the questionnaire being organised into four main sections, there is little to be gained from dividing the corresponding variables into four clusters. There was very little similarity in the way individuals answered these four main groups of questions, which is useful in terms of sociological analysis.

Figures 1 and 2 show the relevance of a ten-cluster partition. When examining the composition of clusters, we see that two of them are composed of variables with similar themes. As these two clusters are combined within the next clustering step, we use a partition into nine clusters, which better suits sociological analysis of the farmers' behaviour.

One of the main advantages in the proposed approach of variable clustering is that it simultaneously provides the variable clusters and the synthetic variables (SVs) of these clusters. The interpretation of SVs is a key step in the proposed methodology. To avoid a linear presentation of the nine SVs, these are interpreted according to the sociological results they produce.

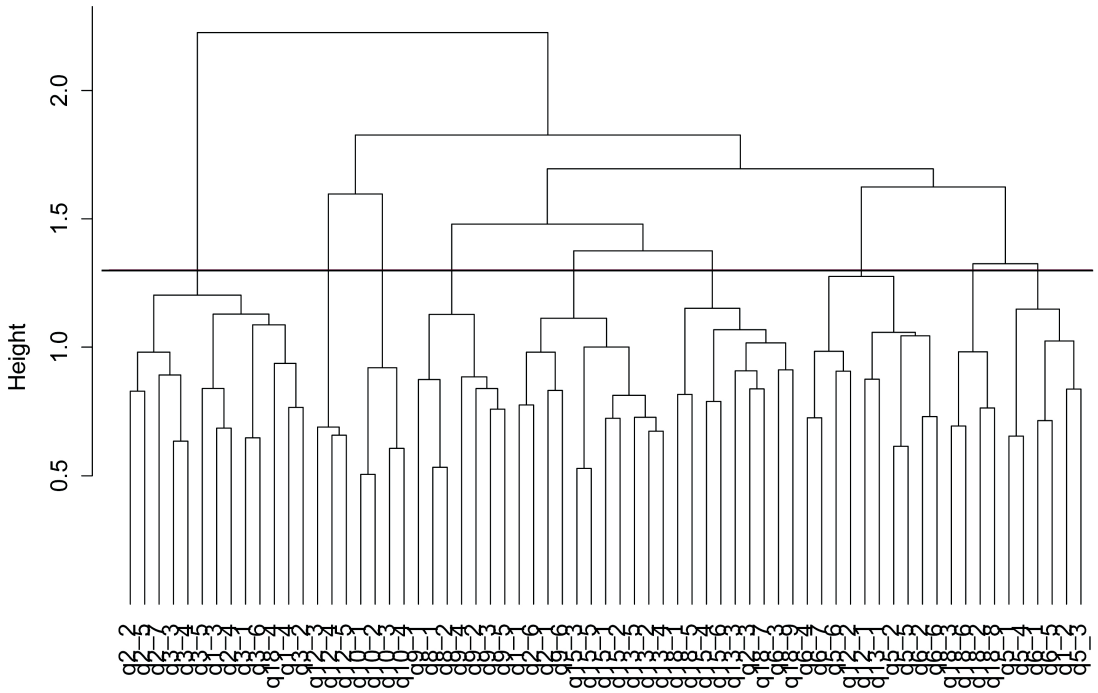

Figure 1: Cluster tree of the 67 categorical variables built with *ClustOfVar*

4.2 First group of synthetic variables: structured according to the questions

Reading each synthetic variable as a gradient. The four synthetic variables described below correspond to the four themes of the questionnaire (see Section 3). Their characteristics are detailed in Table 2: number and list of variables, correlation ratio,

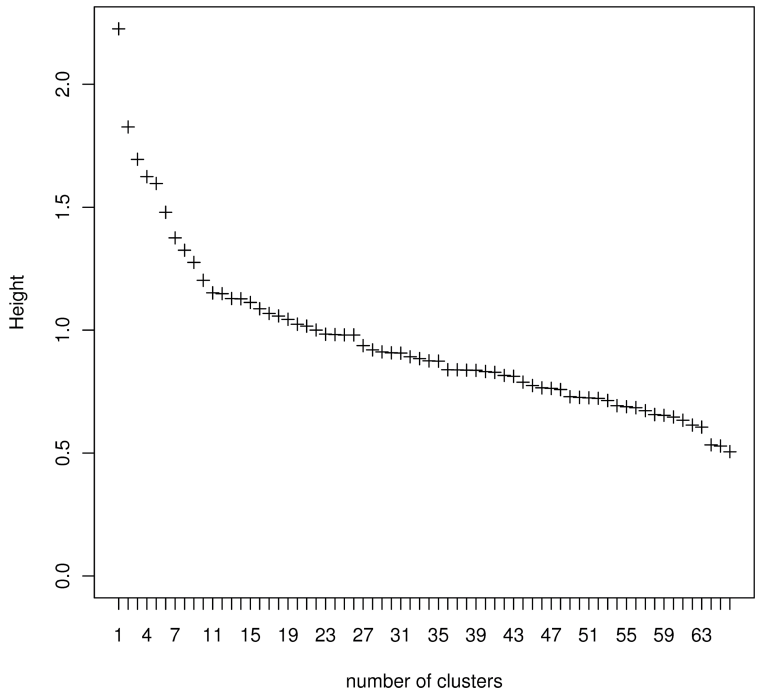

Figure 2: Aggregation level of ascendant hierarchical clustering

homogeneity and percentage of inertia explained. The correlation ratio between the categorical variables and the numeric synthetic variable of the cluster, indicated in brackets, shows that the clusters with smallest number of variables are composed of variables that are most strongly related to the SV. For larger clusters, some values are indeed lower because they include variables covering a wider variety of themes. As explained in Subsection 2.2, the homogeneity of a cluster is defined as the largest eigenvalue of MCA applied to the variables of the cluster, that is to say, the variance of its SV. To compare values to one another, we also calculate the percentage of inertia of the cluster explained by the SV. To do so, we divide the homogeneity of the cluster by the total variance of the cluster, defined by $\frac{m_k}{p_k} - 1$ (it depends on the average number of categories per variable). Cluster 3 shows a lower percentage of inertia, but this needs to be balanced with the fact that it is the cluster containing the most variables. The coordinates of categories on the SVs, defined in (5), allow each variable to be displayed as a sort of gradient. Figure 3 shows that the positive or negative values of each SV (numeric) are associated with a distinct grouping of categories of variables within the corresponding cluster.

Synthetic variable 1: relationship with the non-farming world. This SV almost entirely summarises questions 13 and 15 relating to farmers' opinions of agri-environmental measures (AEMs). The seven variables most associated with the SV come from these two questions. They are associated, but to a lesser extent, with three variables addressing possible problems relating to sharing a particular area with other

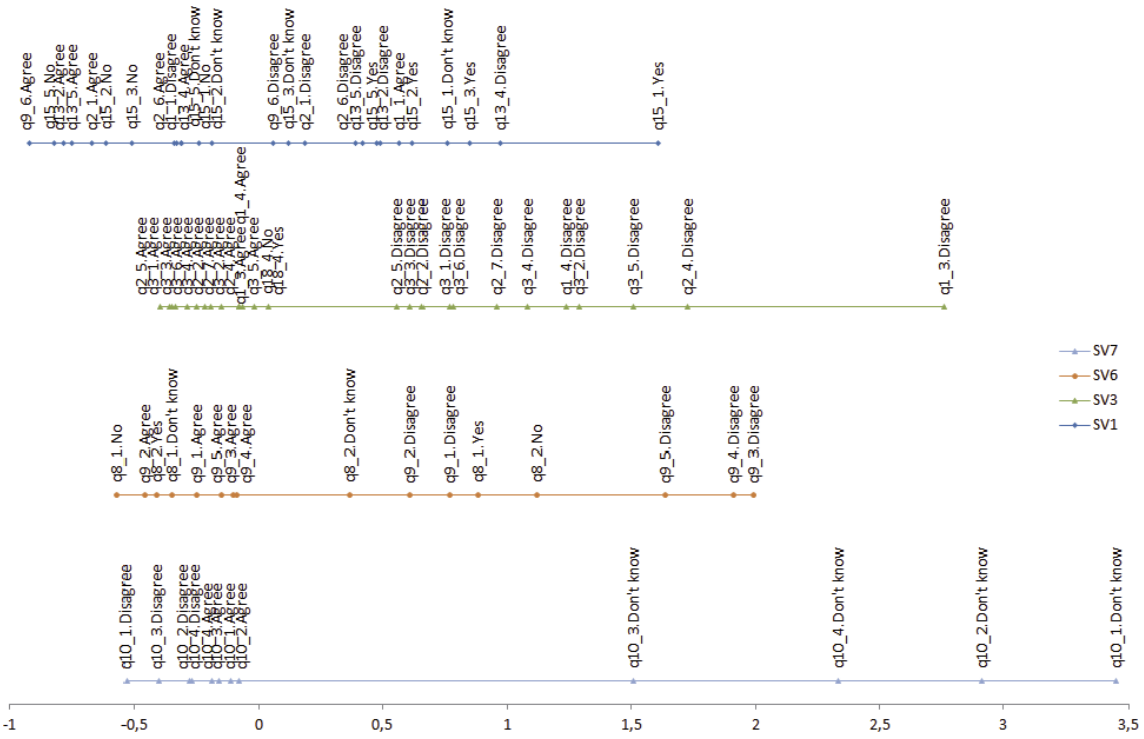

Figure 3: SVs 1, 3, 6 and 7 as gradients

users (tourists, neighbours, etc.). The link between the questions and SVs essentially comes down to the fact AEMs are designed to protect natural resources, a goal which not only benefits farmers, but also the general public. Farmers' relationships with people outside of their profession depend on their (the farmers) views on AEMs. When it comes to perceptions of environmental protection, AEMs, and tourism among farmers, there would appear to be two main camps:

- Those who see them as a constraint placed on their activities (q9_6, q13_2, q2_1, q15_2) or as detrimental to the image of farmers (q15_5).
- Those who see them as an asset to their profession (q15_1), a source of greater solidarity among farmers (q15_3), a way of improving the quality of their produce (q15_2), and a positive influence on the image of farmers in general (q15_5). Positive responses relating to relationships with non-farmers (q1_1, q2_6) also fall into this group.

On this basis, it is possible to easily interpret and label each SV.

Synthetic variable 3: difficulties in exercising the profession. SV3 summarises questions 2 and 3, which addressed the difficulties faced by farmers in exercising their activity. The overlap between the variables making up the SV and the variables describing the theme in two questions is very large. Indeed, 10 out of the 13 variables

that make up this SV come from these two questions. Two others (q1_3 and q1_4) come from the first introductory question, which covered general items related to the job. The variable "your profession is undergoing profound changes" (q1_3) is the most structuring variable. The opinions expressed by farmers fell into two distinct groups: those who consider difficulties are significant and related to structural change, and those who do not have any difficulty in exercising their activity. Unlike SV1, these difficulties are not specific to AEMs but relate to overall agricultural practices (environmental or others).

Synthetic variable 6: severity of and responsibility for environmental problems. SV6 summarises questions 8 and 9: q8_81 and q8_82 addressed the severity of environmental problems and q9_1 to q9_5 related to categories of stakeholders responsible for managing environmental problems. The way in which respondents answered these two groups of questions was very similar.

Synthetic variable 7: prospective scenarios of farming and environment relationships. The interpretation of this SV is simple; it includes the question 10, which suggested four different scenarios describing the relationships between agriculture and the environment over the next 20 years. The four scenarios, which are not mutually exclusive, cover a wide range of possibilities in terms of evolution. Differences in the structure of responses to this SV are mainly between those who vote against or in favour of future scenarios (negative values) and those who have no opinion (positive values).

4.3 Second group of synthetic variables: less clear structured according to the questions

Synthetic variables for clusters 8 and 9 are those that explain the greatest percentage of variance. They contain few variables, relating to themes that are very similar.

Synthetic variable 8: maintaining areas with low levels of production. This SV is made up entirely of variables from question 12, which relate to non-intensive farming practices. The questionnaire suggested three types of practices: the bare minimum of work required to keep farms serviceable, intentionally carrying out no work on a farm in order to let nature take its course, and intentionally carrying out no work on a farm in order to reduce workload. The farmers appear to have interpreted the variables in question 12 differently to ourselves. This is shown by the fact that the results contribute to another SV (SV5). The two variables contributing to SV5 are general (and even abstract) in nature, while the three that contribute to SV8 are directly linked to spatially-anchored farming practices in areas with low levels of production.

Synthetic variable 9: practical difficulties in implementing AEMs. The SV9 includes variables from question 18, which addresses potential difficulties in implementing AEMs. Of the nine challenges listed in the questionnaire, the respondents grouped

together four practical ones (institutional controls to ensure farmers respect the terms and conditions, financial investment required to meet the terms and conditions, paperwork and workload generated). SV9 focuses on the practical difficulties in implementing AEMs. Less concrete and less specific difficulties (lack of training, amount of compensatory aid, lack of solidarity among farmers) were associated with another synthetic variable (SV4).

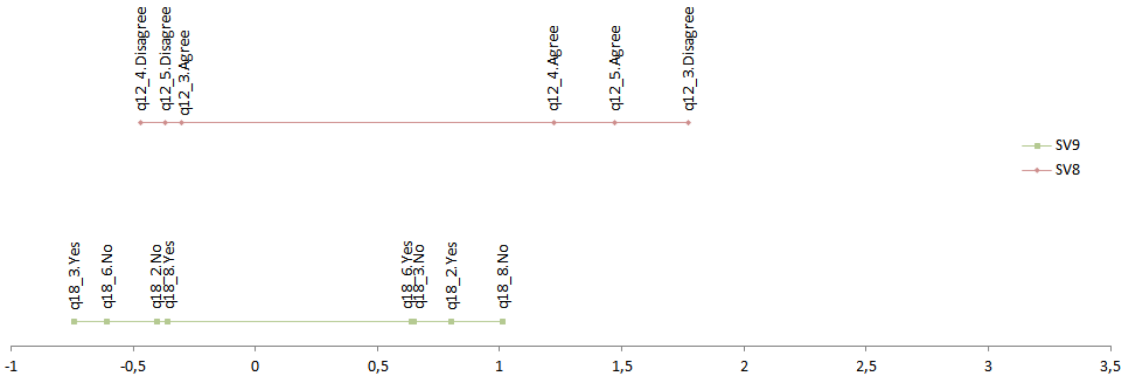

Figure 4: Synthetic variables 8 and 9 as gradients

4.4 Third group of synthetic variables: structured according to several issues

Synthetic variables 2 and 5: attractions and purposes of the job. Both of these SVs are made up of variables from question 5 (advantages of working in agriculture) and question 6 (reasons for working in agriculture). These were associated with one variable from introductory question q1_2 and another from question q13_1. These SVs relate specifically to the farming profession, which the respondents differentiated from the activities of a farmer (i.e. the physical work actually carried out on a farm) as suggested in the questionnaire. But which aspects of the profession appear in which synthetic variable? It would appear that the difference between the two SVs is one of time. SV2 relates to the current vision of agriculture (being at the forefront of technology, adapting to consumer requirements, being close to nature, being independent, feeding human beings, feeling motivated in one's work). On the other hand, SV5 addresses the perception of the evolution of farming over time, by connecting the past and the future (constant change, growing crops while constantly changing, owning your own property, maintaining old buildings, mastering advanced techniques, selling on your farm).

4.5 Particular synthetic variable - SV4

Cluster 4, which has the lowest percentage, consists of eight variables relating to various topics and covering six different questions. SV4 is therefore outside of the intended structure of the questionnaire. It brings together variables from various questions. This

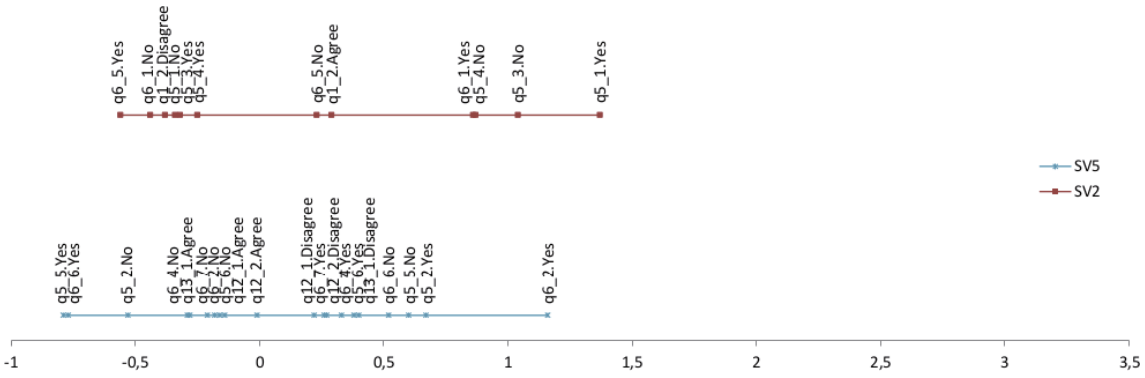

Figure 5: SVs 2 and 5 as gradients

synthetic variable addresses lack of training, effects of AEM, difficulties in implementing AEM, and the purpose of the profession. The uniqueness of this SV makes for a theme that is difficult to characterise. This is made even more complicated by the fact that no groups of categories can be identified (Figure 6).

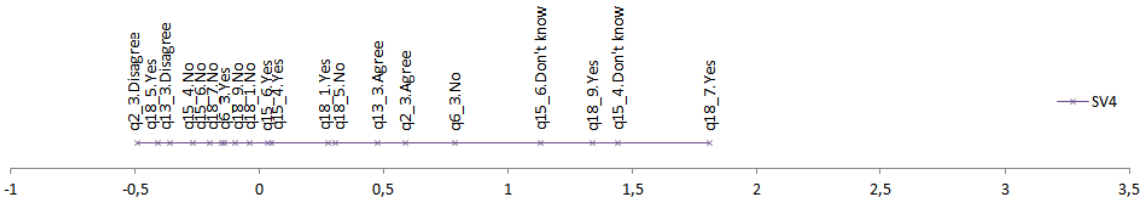

Figure 6: SV4 as gradient

4.6 From labelling of synthetic variables to the typology of individuals

Partitioning farmers into seven clusters. The variable clustering approach, used as an alternative to the traditional strategy based on MCA, identified nine SVs. These SVs show some sets of characteristics that identify certain points of view relating to changes in the farming profession (state of the environment, relationship to the non-agricultural world, trust in the future, etc.). We established a typology of individuals to highlight the profiles of farmers expressing particular points of view. This specifically involved the use of an ascendant hierarchical clustering algorithm (AHC with Ward criterion) on individual scores measured using the nine synthetic variables. It also included a consolidation step using the k-means algorithm in order to stabilise the typology. Analysis of the cluster tree and the histogram showing the AHC level indices indicates a jump for three or seven clusters of individuals.

Compared to the issue addressed in the sociological analysis, a three-cluster typology seems less relevant than the seven-cluster typology, especially in terms of accounting for the diversity of profiles that emerged during the aforementioned process of variable clustering. However, unlike MCA, the variable clustering algorithm does not impose

orthogonality constraints between SVs. This is an advantage for the search of SVs in terms of flexibility. But to perform and view quality projections in orthogonal planes, we have to conduct a normalized PCA on these nine synthetic variables to have uncorrelated variables. By keeping all the axes, all information is retained. The plot of the partition of farmers into seven clusters on the first factorial plane (Figure 7) shows that the clusters are relatively homogeneous and separated from each other, with a satisfactory quality of projection of individuals.

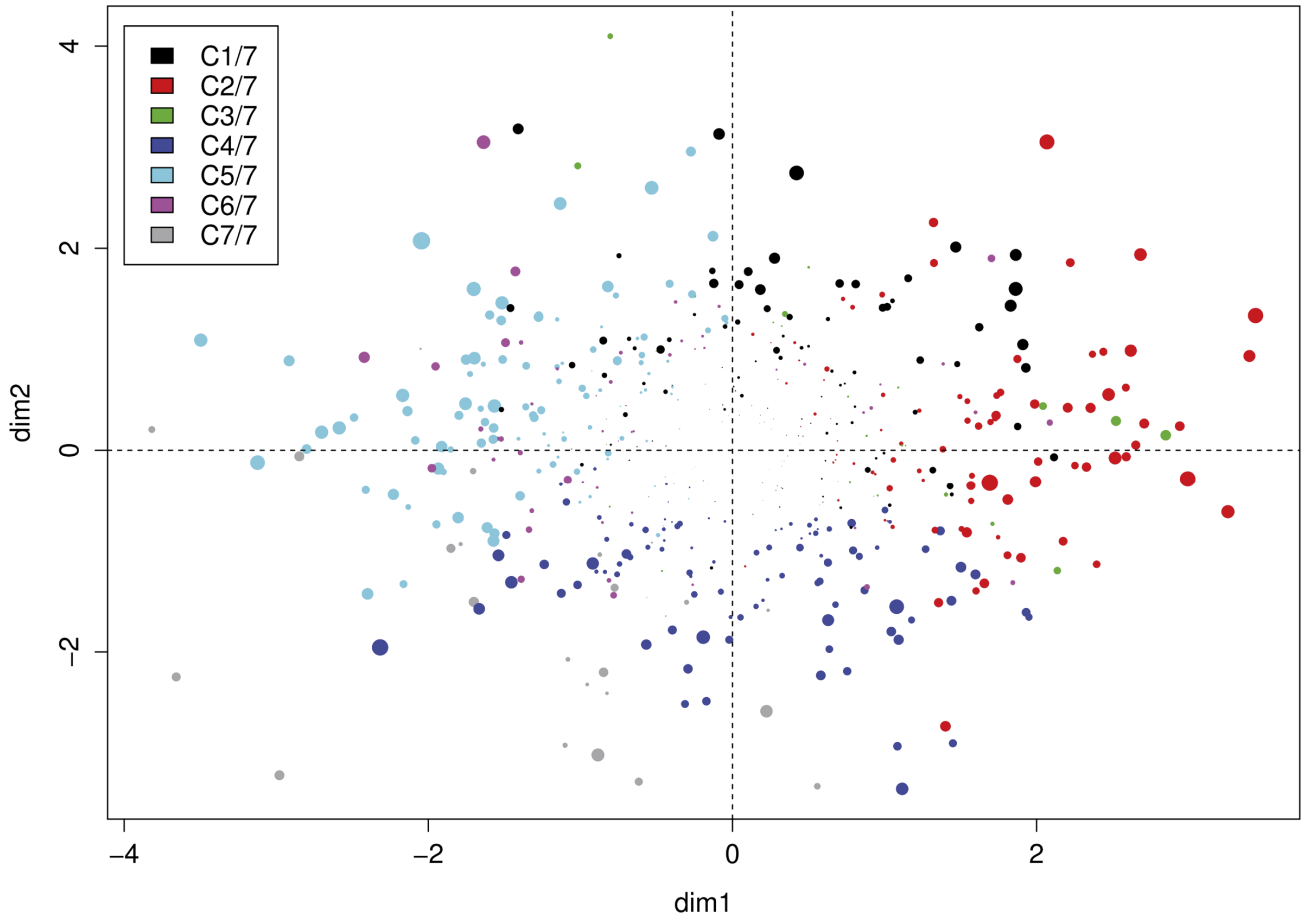

Figure 7: Plot of farmers clustered into seven clusters on the first factorial plane (Dim1 and dim2 are the first two principal components of the normalized PCA on the nine synthetic variables; they correspond in this case to SV1 and SV2. This additional step is necessary for a projection onto an orthogonal plane.)

Interpretation of clusters of individuals using synthetic variables. A partition is interesting when clusters are described by the individuals who compose and/or by the variables that characterise them. Analysing individuals is of little use, due to their anonymity. By reducing the number of variables via *ClustOfVar*, we can interpret the

partition of farmers, not based on the 67 original variables, but by relying on the nine SVs described in the previous section. Table 6 provides the mean value of each SV in the seven clusters of the typology. The parallel with the label of SV makes interpreting results much easier. Specifically, for each SV we compare the means for each cluster of individuals to that of the total sample (null mean because in the variable clustering approach used, SVs are centered). Table 6 shows in bold the negative (positive) means that are significantly lower (superior) to 0 (p-value less than 10^{-3}). However this test has no real statistical value, as SVs were used to create groups of individuals. It is therefore only useful for interpretation, as it indicates which SV characterises different clusters of farmers. Certain clusters (1, 3, 6, 7) may be characterised by a single SV, which facilitates interpretation. For the smaller clusters (3 and 7), we expect a concise characterisation. However, cluster 1, which accounts for the largest number of individuals (almost a quarter of the sample), is also described by a single SV. These results underline both the homogeneity within groups of individuals and the heterogeneity between them. For the remaining three clusters of individuals, the characterisation is less marked. Interpretation through several SVs is necessary. However understanding these groups remains easy given the low number of SVs (previously interpreted and labelled) and is relevant for understanding the problem.

A rich and relevant sociological interpretation. As previously mentioned, the labelling of SVs is an effective aid in the interpretation of clusters of individuals. We will see that the resulting partition is relevant and provides a lot of information for sociological analysis. The introduction of socioeconomic characteristics, as variables illustrating the typology performed, provides an additional aid in interpreting clusters.

Cluster 1 is characterised by a negative mean value of SV5 (-1.471). We can therefore deduce that farmers in this cluster are interested in change, they love their job because it needs to constantly evolve and they consider that AEMs require them to master advanced techniques. They are not particularly sensitive to the area in which farms are located. A large proportion of them (39%) grow crops (cereals, oilseeds and protein crops) and practice, to a lesser extent, intensive farming (29%). On the other hand, few wine growers and mountain stockbreeders appeared willing to share their opinions.

Farmers in Cluster 2 believe environmental problems are a reality and do not feel that they are exaggerated ($\overline{SV6} = 0.747$). Measures in favour of the environment play an important role in the production process ($\overline{SV1} = 1.476$) and administrative procedures related to AEMs are not an issue. On the contrary, they accept government regulation of the agricultural industry. However they remain attached to the market ($\overline{SV2} = 1.176$), which, according to them, also provides guidance. This group of farmers also faces difficulties with the entrepreneurial aspect of their job: they complain about the additional workload and investment required by the implementation of environmental measures ($\overline{SV9} = 1.349$). These are primarily wine growers (59%) and farmers engaging in direct sales (42%). A large proportion of these farmers are under 40 years old (45%) and have an above-average level of education (compared with others surveyed), often 2 or more years study beyond the basic French high-school diploma (27%).

For cluster 3, SV3 has a high positive average value (4.026), suggesting that this cluster is defined by farmers who are confident in the future and seem to exercise their activity without difficulty. They do not share many of the challenges listed in the survey (paperwork, cost of land, labour, etc.). Nearly 33% of individuals in this cluster are female farmers (the only socioeconomic variable characterising them).

Cluster 4 is made up of farmers who are particularly careful about protecting the environment ($\overline{SV6} = -0.850$), which they find difficult to reconcile with technical progress ($\overline{SV4} = 0.823$). One of the first goals of their activity is to protect natural resources and landscape ($\overline{SV5} = 0.826$). This environmental concern suggests that these farmers are likely to question some of their own practices. However they are not really interested in the constant evolution of their activity ($\overline{SV2} = -0.695$). Indeed, they explain that incessant change is not what makes their profession attractive. They believe that environmental measures are reactivating old expertise. These farmers are primarily mountain stockbreeders, rather young and with relatively low education.

Cluster 5 farmers reject environmental concerns, believe that the severity of environmental problems is exaggerated, consider that the situation is not alarming ($\overline{SV6} = 1.428$). At the same time they are also very critical towards AEMs, which they believe slow down their activity and projects ($\overline{SV1} = -1.503$). These are mostly farmers (dairy and/or meat) based in Mayenne or Dordogne, with often sizeable farms employing three to four people, without visits or on-farm sales (84.26%). These farmers are acutely aware of social relationships in exercising their job: according to them, relationships with non-farmers are a source of tensions.

Individuals in Cluster 6 may be considered as contributing to agricultural abandonment as they have a positive SV8 value (2.437). Almost 40% of them produce predominantly crops.

Finally, Cluster 7 is associated with a positive average value of SV7 (4.578), i.e. these farmers do not see themselves being involved in any of the projected scenarios proposed, but do not reject any of them either. The future seems uncertain. These respondents have been exercising their job for a long time (57% of them began work before 1984) and do not have any responsibility in any farming trade bodies (87%).

Upon closer examination of these seven clusters of farmers, it is clear that each is characterised by a number of factors, many of which are far from relating solely to the environment : the business-oriented farmers (Cluster 1), those who accept state regulation of their activities (Cluster 2), those who are confident (Cluster 3), those who take an active interest in environmental protection (Cluster 4), those who are against environmental protection and who defend their local area (Cluster 5), those engaged in agricultural abandonment (Cluster 6), and those to whom the future appears uncertain (Cluster 7).

5 Conclusion

This article proposed an original approach to analyse the extent to which French farmers consider environmental issues when carrying out their activities. It introduced new

advances in the fields of sociology and statistics. The innovation of this study lies in the use of variable clustering in the place of the more traditional approach of factor analysis to create a typology of farmers. The *ClustOfVar* method generates groups of categorical variables, according to the similarities in the way respondents answered the questionnaire. The simultaneous construction of SVs does not follow the construction of the questionnaire, but highlights the main trends in farmers' opinions. The SVs are displayed as a sort of gradient, whose values correspond to distinct associations in the categories of variables. As a result, the SVs are easier to label than traditional principal components obtained through factor analysis. The final clustering of individuals using the scores on these SVs enables typical profiles of farmers to be identified, which can then be interpreted with relative ease.

This article also contributes to sociological research, since it helps to identify how farmers perceive environmental protection relating to their activity. The results show that farmers take on board environmental issues in a variety of ways. It is not simply a question of those who are pro-environment and those who are not. The plurality in the perception of farmers is not distributed around classic variables: some are influential (production, education level, and gender in one case), but not always. The social and technical characteristics of the respondents do not seem to form variables that are representative of their relationship with the environment. This is consistent with the findings of another recent literature review (Burton, 2014). The diverse nature of these results is partly due to the continuous reflection of farmers against a backdrop of constant change. In farming, as well the broader ecological movement, environmental concerns and willingness to defend a particular geographical area are not always one and the same. While some farmers exhibit a genuine willingness to protect their local ecosystems, "business-oriented" farmers want to protect their local area simply because they see it as a place to live. In addition, farmers who are sceptical about environmental issues tend to be very attached to the idea of farming being intrinsically linked with a particular area.

References

Abdallah, H. and Saporta, G. (1998). Classification d'un ensemble de variables qualitatives. Revue de Statistique Appliquée, 46(4):5–26.

Arabie, P. and Hubert, L. (1994). Cluster analysis in marketing research. In Bagozzi, R. P., editor, Advanced methods of marketing research, pages 160–189. Blackwell, Cambridge, MA.

Burton, R. J. F. (2014). The influence of farmer demographic characteristics on environmental behaviour: A review. Journal of Environmental Management, 135:19–26.

Candau, J., Deuffic, P., Ginelli, L., Lewis, N., and Lyser, S. (2005). La prise en compte de l'environnement par les agriculteurs. Résultats d'enquête. Rapport d'étude, Cemagref.

Charrad, M. and Ben Ahmed, M. (2011). Simultaneous Clustering: A Survey. In Pattern Recognition and Machine Intelligence. Springer Berlin / Heidelberg.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2011). ClustOfVar: An R Package for the Clustering of Variables. In The R User Conference.

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012a). ClustOfVar: An R Package for the Clustering of Variables. Journal of Statistical Software, 50(13):1–16.

Chavent, M., Kuentz-Simonet, V., and Saracco, J. (2012b). Orthogonal rotation in PCAMIX. Advances in Data Analysis and Classification.

Dhillon, I., Marcotte, E., and Roshan, U. (2003). Diametrical Clustering for Identifying Anticorrelated Gene Clusters. Bioinformatics, 19(13):1612–1619.

Kiers, H. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika, 56(2):197–212.

Lerman, I. (1990). Foundations of the likelihood linkage analysis classification method. Applied Stochastics Models and Data Analysis, 7(1):63–76.

Lerman, I. (1993). Likelihood linkage analysis classification method : An example treated by hand. Biochimie, 75(5):379–397.

SAS Institute Inc. (2013). The varclus procedure. In SAS/STAT ® 13.1 User's Guide. SAS Institute Inc., Cary, NC.

Vichi, M. and Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis, 37(1):49–64.

Vichi, M. and Saporta, G. (2009). Clustering and Disjoint Principal Component Analysis. Computational Statistics & Data Analysis, 53(8):3194–3208.

Vigneau, E. and Chen, M. (2015). ClustVarLV: Clustering of Variables Around Latent Variables. R package version 1.3.2.

Vigneau, E. and Qannari, E. (2003). Clustering of variables around latent components. Communications in statistics Simulation and Computation, 32(4):1131–1150.

Appendix: The 67 categorical variables used in the variable clustering step

Question	Variable	Description	N	Percentage
1		**Do you think:**		
	q1_1	Your activity is perceived positively by non-farmers	199	36.58
	q1_2	Your profession is motivating	310	56.99
	q1_3	Your profession is undergoing profound changes	525	96.51
	q1_4	Your are worried about the future of your activity	504	92.65
2		**What seems difficult today in your job in general?**		
	q2_1	Increasing tourist visits	120	22.06
	q2_2	Decreasing number of farms	414	76.10
	q2_3	Professional training needs	249	45.77
	q2_4	Paperwork	501	92.10
	q2_5	Protection of the environment	344	63.24
	q2_6	Relationship with non-farmers neighbours	291	53.49
	q2_7	Selling on your farm	457	84.01
3		**What do you find difficult about doing your job today?**		
	q3_1	Workforce	392	72.06
	q3_2	Conforming to industry standards	483	88.79
	q3_3	Need to expand	370	68.01
	q3_4	Land prices	444	81.62
	q3_5	Price and sales of products	516	94.85
	q3_6	Working hours	395	73.16
5		**What is attractive about your job?**		
	q5_1	Being at the forefront of technology	108	19.85
	q5_2	Owning your own property	240	44.12
	q5_3	Being close to nature	417	76.65
	q5_4	Being independent	424	77.94
	q5_5	Constant change	235	43.20
	q5_6	Carrying on the family history locally	163	29.96

6		**Today the aim of your work is to**		
	q6_1	Adapting to consumer requirements	184	33.82
	q6_2	Maintain old buildings	73	13.42
	q6_3	Support your family	461	84.74
	q6_4	Maintain and pass on your farm	258	47.43
	q6_5	Feed human beings	158	29.04
	q6_6	Produce while adapting to the expectations of society	221	40.62
	q6_7	Protect natural resources and landscapes	245	45.04
8		**Do you think...**		
	q8_1	The severity of environmental problems is exaggerated.	208	38.24
	q8_2	The environment situation is worrying.	379	69.67
9		**In your opinion, environmental problems are the concern of**		
	q9_1	Farmers	412	75.74
	q9_2	Environmental protection associations	310	56.99
	q9_3	Each consumer	518	95.22
	q9_4	Industrialists	519	95.40
	q9_5	Authorities	498	91.54
	q9_6	Nobody, because there is no problem	32	5.88
10		**Relationships between agriculture and environment are evolving. In the next 20 years, which scenarios seem the most likely to you?**		
	q10_1	Agriculture will be more linked to the food-processing industry and have to respect quality standards.	464	85.29
	q10_2	The environment will be at the heart of agriculture with systems close to organic farming.	248	45.59
	q10_3	Europe will give the general framework of production and the environment and the region will manage more specific objectives	310	56.99

	q10_4	There will be both intensive areas dedicated to production areas earmarked for preservation	288	52.94
12		**Do you agree with the following?**		
	q12_1	I have to control nature for my activity	339.	62.32
	q12_2	I have to adapt to nature	522	95.96
	q12_3	I have to keep the areas of low production on my farm serviecable	465	85.48
	q12_4	I try not to maintain the areas of low production on my farm in order to let nature take its course	152	27.94
	q12_5	I do not maintain areas of low production in order to decrease workload	109	20.04
13		**As a farmer, you are invited (or required) more frequently to respond to measures for the protection of the environment. For your activity, such measures ...**		
	q13_1	Require a thorough knowledge of advanced techniques	319	58.64
	q13_2	Impede progress	209	38.48
	q13_3	Encourage you to use traditional farming know-how	233	42.83
	q13_4	Limit your freedom of action	414	76.10
	q13_5	Affect domains that are your concern	195	35.85
15		**Do you consider that measures for the environment**		
	q15_1	Enable young people to settle	28	5.15
	q15_2	Improve product quality	252	46.32
	q15_3	Strengthen solidarity within the sector	183	33.64
	q15_4	Are a good way to limit production	224	41.18
	q15_5	Promote the image of agriculture	320	58.82
	q15_6	Convey an old-fashioned image of agriculture	117	21.51
18		**In the application of AEM, what seems the most difficult to you?**		
	q18_1	Technical changes proposed	63	11.58
	q18_2	Workload generated	182	33.46

q18_3	Institutional controls	256	47.06
q18_4	Effectiveness of measures	134	24.63
q18_5	Weak amount of compensatory aid	232	42.65
q18_6	Financial investment	265	48.71
q18_7	Lack of training	42	7.72
q18_8	Paperwork	400	73.53
q18_9	Solidarity among farmers	37	6.80

Table 1: Socio-economic and demographic characteristics of the survey.

Variable	N	%
Type of production	**544**	**100.0**
Mixed farming	95	17.5
Crops	132	24.3
Intensive farming	114	21.0
Farming in mountain areas	88	16.2
Perennial crops	115	21.1
Farm reception or on-farm sales	**544**	**100.0**
Yes	130	23.9
No	413	75.9
NA	1	0.2
Number of employees	**544**	**100.0**
1 employee	187	34.4
2 employees	207	38.1
3-4 employees	108	19.9
5 employees and above	40	7.4
NA	2	0.4
Other activity	**544**	**100.0**
Yes	60	11.0
No	480	88.2
NA	4	0.7
Gender	**544**	**100.0**
Female	80	14.7
Male	463	85.1
NA	1	0.2
Family situation	**544**	**100.0**
Single	112	20.6
With a partner	428	78.7
NA	4	0.7
Age	**544**	**100.0**
Under 40 years	187	34.4
Between 40 et 49 years	203	37.3
Between 50 et 59 years	126	23.2
60 years and above	23	4.2
NA	5	0.9

Variable	N	%
Levels of education	**544**	**100.0**
Vocational diplomas	243	44.7
High school diploma	123	22.6
Undergraduate	113	20.8
Degree or master's degree	62	11.4
NA	3	0.6
Responsibilities	**544**	**100.0**
Yes	319	58.6
No	225	41.4
Professional responsibilities	**544**	**100.0**
Yes	178	32.7
No	361	66.4
NA	5	0.9
Political mandate	**544**	**100.0**
Yes	101	18.6
No	435	803.8
NA	8	1.5
Member of an association	**544**	**100.0**
Yes	207	38.1
No	331	60.8
NA	6	1.1
Parents farmers	**544**	**100.0**
Yes	476	87.5
No	68	12.5
Professional experience	**544**	**100.0**
Prior 1984	192	35.3
Between 1984 and 1991	147	27.0
Since 1992	196	36.0
NA	9	1.7
Other previous professional activity	**544**	**100.0**
Yes	233	42.8
No	302	55.5
NA	9	1.7

Table 2: Description of the synthetic variables 1, 3, 6 and 7.

Cluster	**1**		**3**		**6**		**7**	
Number of variables	*11*		*13*		*7*		*4*	
Variables (correlation ratio)	q15_3	(0,39)	q1_3	(0,28)	q8_1	(0,48)	q10_1	(0,54)
	q13_2	(0,38)	q3_4	(0,26)	q8_2	(0,41)	q10_2	(0,54)
	q15_5	(0,35)	q2_4	(0,26)	q9_2	(0,28)	q10_4	(0,52)
	q15_2	(0,34)	q3_1	(0,23)	q9_5	(0,25)	q10_3	(0,37)
	q13_4	(0,30)	q3_6	(0,22)	q9_3	(0,20)		
	q13_5	(0,31)	q3_2	(0,21)	q9_1	(0,19)		
	q15_1	(0,26)	q3_3	(0,18)	q9_4	(0,18)		
	q1_1	(0,19)	q2_5	(0,18)				
	q2_1	(0,13)	q2_7	(0,18)				
	q2_6	(0,13)	q2_2	(0,14)				
	q9_6	(0,05)	q1_4	(0,12)				
			q3_5	(0,12)				
			q18_4	(0,01)				
Homogeneity of the cluster	*2,8*		*2,4*		*2,0*		*2,0*	
Percentage of inertia explained	*18,9*		*18,3*		*22,0*		*24,6*	

Table 3: Description of the synthetic variables 8 and 9.

Cluster	**8**		**9**	
Number of variables	*3*		*4*	
Variables (correlation ratio)	q12_4	(0,58)	q18_3	(0,48)
	q12_5	(0,54)	q18_6	(0,39)
	q12_3	(0,53)	q18_8	(0,37)
			q18_2	(0,32)
Homogeneity of the cluster	*1,7*		*1,6*	
Percentage of inertia explained	*55,2*		*39,1*	

Table 4: Description of the synthetic variables 2 and 5.

Cluster	**2**		**5**	
Number of variables	*6*		*10*	
Variables (correlation ratio)	q5_1	(0,46)	q5_5	(0,47)
	q6_1	(0,38)	q6_6	(0,40)
	q5_3	(0,33)	q5_2	(0,35)
	q5_4	(0,21)	q6_2	(0,21)
	q6_5	(0,13)	q6_2	(0,11)
	q1_2	(0,11)	q6_2	(0,10)
			q13_1	(0,11)
			q6_4	(0,10)
			q5_6	(0,06)
			q6_7	(0,05)
			q12_1	(0,03)
			q12_2	(0,01)
Homogeneity of the cluster	*1,6*		*1,8*	
Percentage of inertia explained	*27,1*		*17,9*	

Table 5: Description of the synthetic variable 4.

Cluster	**4**	
Number of variables	*9*	
Variables (correlation ratio)	q2_3	(0,29)
	q18_7	(0,28)
	q15_4	(0,21)
	q15_6	(0,18)
	q13_3	(0,17)
	q18_5	(0,13)
	q18_9	(0,13)
	q6_3	(0,11)
	q18_1	(0,01)
Homogeneity of the cluster	*1,5*	
Percentage of inertia explained	*13,7*	

Table 6: Mean value of each SV in the seven clusters of farmers (in bold: significantly lower or higher than the mean of the SV in the entire sample equal to 0, p-value less than 10^{-3}).

Synthetic variable	**Cluster of farmers**						
	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
N	115	85	34	103	108	69	30
Percentage	21,1	15,6	6,3	18,9	19,9	12,7	5,5