

A Quantitative Method for Classifying Farmers Using Socioeconomic Variables

JOSE CROSSA, MAURICIO R. BELLON, AND JORGE FRANCO

Abstract

Small-scale farmers in developing countries are heterogeneous in the resources they control, in the constraints they face, and, hence, in the crop varieties they require. This poses the challenge to participatory plant breeding (PPB) of identifying farmers that experience comparable conditions and needs, and therefore require similar varieties. Practitioners of PPB need methods for classifying farmers into homogenous groups with similar variety demands. This paper presents a statistical method for classifying individuals—farming households in this example—into homogenous but distinct groups. The method allows the use of different types of variables, provides a systematic approach to decide the number of groups present in the data, and assigns a probability that an individual belongs to a group. The method assumes that data have been collected from a sample of farmers in the target socioeconomic or agroecological environments. In the example presented, the method is used to divide a random sample of small-scale maize farmers in Mexico into homogenous groups.

Introduction

Small-scale farmers in developing countries are not always homogenous, even within a community. Ownership of resources such as land, labor, and capital is not equal between households, nor is the sharing of knowledge and information. Consequently, goals and constraints differ between farming households. Variability—spatial and temporal—is another fact of life for every farmer and his/her household. Soils and topography vary and seasons change. All of these factors influence the type of crop varieties that farmers want

and need. In failing to recognize differences between farming households, breeders may overestimate the potential impact of their varieties because they may end up working with a smaller and possibly unrepresentative subset of the farmers they hope to serve, or they may have a very static view of farmers' resources and/or constraints.

Recognizing and addressing the heterogeneity of conditions faced by small-scale farmers and their different needs are key to developing appropriate germplasm through participatory plant breeding (PPB). This creates the

challenge, however, of identifying farmers that experience comparable conditions and needs and, hence, require similar varieties. In other words, the challenge is to identify the recommendation domains for new germplasm.

Practitioners of PPB need methods for classifying farmers into homogenous groups with similar variety demands. There are a number of different approaches to achieving this. A typology can be created based on some *a priori* knowledge of the conditions and needs of farmers. This requires that the researcher decides which variables are important, how they fit together, and what cut-off values in each variable should be used to divide each type. Another approach is to rely on the farmers' own views of their differences and to let them define the groups. A wealth ranking exercise is one example. Key informants in a community identify farmer groups based on wealth, as well as the characteristics defining each group. Then the informants classify farmers into each group (see Bellon, this proceedings). A third approach is to apply geometric or statistical clustering methods. Usually these methods rely on continuous or categorical data, requiring that measurements of farmers' characteristics (attributes) are available. A problem with this method, as with the typology, is that the researcher has to decide which variables should be included. A common mistake is to use all available variables, which makes the interpretation of each group difficult.

In this paper we present a powerful method for classifying "individuals" (e.g., farmers, households, etc.) into homogenous but distinct groups based

on their attributes. The method allows the use of both continuous and categorical variables. This is important because certain attributes are measured as continuous variables (e.g., landholdings), while others are discrete (e.g., gender) or include multiple categories (e.g., membership in different farmer associations). It provides a standardized process for deciding how many groups exist, and, as well as assigning each "farmer" to a group, it gives the probability that each farmer belongs to that group. The method, however, does not solve the problem of deciding which variables to include. This can be addressed by using variables associated with the characteristics identified by a wealth ranking exercise or other participatory methodologies which identify farmer types in an area. This approach relies on variables that are relevant to farmers, but builds on the structure of the actual data to identify the number of groups and to classify each farmer according to these groups. This, in turn, should simplify the interpretation of the meaning of the groups.

The method presented here is illustrated using data from the on-farm conservation project described in Bellon (this proceedings), and involves the use of variables associated with characteristics identified in a wealth ranking exercise. It should be pointed out, however, that the link between the groups formed using the method and the farmers' variety requirements is not automatic. It depends on the variables chosen for the classification and their relationship with the demand for crop traits, and requires the association between the groups and the crop traits to

be tested. This, however, is beyond the scope of this paper. For more information see Bellon (this proceedings).

The paper begins with a brief introduction to classification methods. This is followed by a description of mixture models and the method used to form the initial groups for the classification. The location model, the homogenous conditional mixture model, the independent mixture model, and the canonical variate analysis are then described. Finally, an example is presented of farmer classification using a sequential clustering strategy with categorical and continuous socioeconomic variables.

A Brief Introduction to Classification Methods

Classification methods are grouped in two main categories: cluster analysis and discriminant analysis. Discriminant analysis allocates new individuals to previously defined groups by finding a mathematical function, based on a linear combination of the original variables, that minimizes the chance of misclassification. Cluster analysis is the partition of a heterogeneous population into homogeneous subpopulations using hierarchical or nonhierarchical methods. In hierarchical methods, the individuals or groups are organized in a hierarchy or "tree" and are fused, one by one, to other individuals or groups with the most similar patterns for all attributes. These methods can be used to form a fixed number of groups by truncating the tree at a fixed level. The nonhierarchical clustering methods

guess the number of groups and then use a certain method or algorithm to improve the previous classification by optimizing a particular objective function.

The agglomerative hierarchical clustering method starts with an original dissimilarity (distance or dispersion) matrix among all the individuals, and fuses the two individuals which have the smallest dissimilarity between them to form a group with two members. Next, the group-individual dissimilarity between this new group and the remaining individuals is calculated. This set of dissimilarities is added to the matrix of dissimilarities among the remaining individuals to form a new dissimilarity matrix that is one row and column smaller than the original. A new fusion procedure is carried out, and, when two or more groups are present, group-group dissimilarities must be computed. The procedure ends when all of the individuals are in one group. The method used for calculating the group-individual and group-group dissimilarity is called clustering strategy. A number of agglomerative clustering strategies have been proposed such as the single linkage (or nearest neighbor), the maximum linkage (or furthest neighbor), the unweighted pair group arithmetic averaging (UPGMA), and the Ward method (incremental sum of squares). The Ward method uses the within-groups sum of squares as the objective function. It fuses the two groups that increase the within-group sum of squares the least and increase the among-group sum of squares the most, over all of the possible functions.

Classification methods require a measure of association among individuals calculated from measurements of a number of attributes of each individual. The effective use of classification methods requires an understanding of the properties of the forms and type of data collected as well as of the measures of association. Data form consists of a two-way table of n individuals (farmers) and p attributes (or variables), and the type of attribute can be continuous or categorical. Categorical data may be binary or nominal. The two-way table of n individuals and p variables can have one type (only categorical or only continuous) or a mixture of types (continuous and categorical). Classification based on all available information on the individuals is much more reliable and trustworthy than that based on only some attributes.

Franco et al. (1998; 1999) proposed a two-stage sequential strategy for classifying and studying genetic resources. In the first stage, initial groups are formed using an agglomerative hierarchical clustering method such as Ward or UPGMA and includes all of the continuous and categorical variables. Then a statistical method such as the location model (LM) or the modified location model (MLM) for improving the initial groups is used. These statistical models allow the use of continuous and categorical variables. This two-stage sequential clustering strategy is usually called Ward-MLM or UPGMA-MLM.

The objective of the study was to use the two-stage sequential Ward-MLM strategy for classifying 240 farmers using 9 categorical and 22 continuous socioeconomic attributes.

Mixture Models

Agglomerative hierarchical clustering techniques use proximity (distance) matrices for finding groups of objects, and are basically exploratory (or geometrical) methods that do not use any probabilistic density models. Mixture models, on the other hand, cluster the data using a particular probability density function without the need to explicitly use any proximity measurement.

To illustrate the use of mixture models, consider a random sample of farmers, including samples taken from regions where two notoriously different farmer types based on income level can be found, i.e., low income (L) and high income (H) farmers. The attribute, number of hectares, was measured for each randomly selected farmer. Since we know that low income farmers have fewer hectares than high income farmers, the probability density function (pdf) for number of hectares should take this into consideration. If we assume that the attribute is normally distributed with specific mean (μ) and variance (σ^2), and that one farmer from the sample can come from either the low or high income subpopulation with probability α or $(1-\alpha)$, respectively, then the pdf for any farmer can be written as:

$$\text{pdf of number of hectares} = (\alpha)[N(\mu_L, \sigma_L)] + (1-\alpha)[N(\mu_H, \sigma_H)] \quad (1)$$

In equation 1 there are five parameters to be estimated: α , the proportion of low income farmers in the population ($1-\alpha$ is the proportion of high income farmers in the population); and μ_L , σ_L , μ_H , and σ_H , which correspond to the means (μ) and the standard deviations (σ) of the

number of hectares for low income (L) and high income (H) farmers, respectively. N means normal distribution. The function represented by equation 1 is commonly called “finite mixture density”, and in this particular case the distribution of the number of hectares variable results from a weighted mixture of two underlying normal distributions, where the α and $(1-\alpha)$ are called mixing proportions [$\alpha + (1-\alpha) = 1$]. Note that the parameter α is the relative frequency of the underlying distribution $N(\mu_L, \sigma_L)$, and $(1-\alpha)$ is the relative frequency of the other underlying normal distribution $N(\mu_{LH}, \sigma_H)$. Assuming there are $i = 1, 2, \dots, g$ farmer groups, equation 1 can be extended to:

$$\text{pdf of number of hectares} = \sum_{i=1}^g (\alpha_i) [N(\mu_i, \sigma_i)] \quad (2)$$

The model in equation 2 can be extended to multivariate data in such a way that the univariate normal distribution is replaced by the multivariate normal (MVN) density with mean vectors μ_i and variance-covariance (dispersion) matrix Σ_i , such that:

$$\text{pdf of number of hectares} = \sum_{i=1}^n (\alpha_i) [MVN(\mu_i, \Sigma_i)] \quad (3)$$

$$\text{where} = \sum_{i=1}^n \alpha_i = 1$$

Parameter estimation of the mixture models—the maximum likelihood estimation

The parameters of the distribution under a mixture model are estimated by maximum likelihood (ML) procedures, in which case equation 3 gives the likelihood function of the unknown

parameters as a function of the observed values. In ML we consider a $p \times 1$ random vector of observations $x' = x_1, x_2, \dots, x_p$ and ask about the vector of parameters Θ (the true proportion, α ; the mean, μ ; and the dispersion matrix, Σ , under the normal probability density function). The maximum likelihood estimate of an unknown parameter is the linear combination of the observations that maximizes the likelihood of the parameter given the observations. The ML estimates of the unknown parameters Θ , $\hat{\Theta} = (\hat{\alpha}, \hat{\mu}, \hat{\Sigma})$, is the value of $\Theta = (\alpha, \mu, \Sigma)$ corresponding to the maximum of $l(\Theta | x)$. It is usually easier to find the maximum of the logarithm of the maximum likelihood function $L(\Theta | x) = \ln[l(\Theta | x)]$ than to find it from the function itself, because of the mathematical properties of the logarithmic function. For many ML estimation problems, a simple solution for the ML estimator can be obtained by solving the equation $\partial[L(\Theta | x)] / \partial(\Theta) = 0$.

Forming the initial groups

The question of how to form the *a priori* subpopulations used in the mixture models has been examined by Franco et al. (1997a; 1997b) in the context of genetic resource conservation. The initial groups are the starting points of the iterative process by which a solution that corresponds to a global (or local) maximum of the likelihood function is found.

Franco et al. (1997a) compared the performance of several hierarchical and nonhierarchical classification strategies for forming the initial groups and then compared the application of the mixture of normal distributions to these initial groups. The authors found that the

initial groups formed using the Ward clustering method with Gower's distance (so that all continuous and discrete attributes can be used in the classification) recovered a good percentage of the true groups. Furthermore, the authors applied the mixture models to those initial groups and found a great deal of reallocation of individuals among groups and, thus, the formation of more compact, homogeneous, separate, and well characterized groups. They called this a sequential clustering strategy, where the initial groups are formed using the Ward method, and the mixture normal distribution is applied to the groups to improve the classification.

From a statistical perspective the Ward method seems better than other hierarchical clustering strategies. This is because it has an objective function to minimize the within-group sum of variability and therefore to maximize the among-group variability; thus, it gives a natural connection to the analysis of variance. Furthermore, the Ward method is appropriate for multi-normal data distribution. One problem with this sequential clustering strategy, however, is that while all variables, continuous and discrete, are used to form the initial groups using the Ward method, only the continuous variables can be used in the mixture models.

Location Model

In practice there is a mixture of attribute types. Some attributes used for classifying individuals are continuous and others are categorical. A distance measurement such as Gower's distance

can be used for mixed variable types, thus any hierarchical clustering algorithm such as the Ward method could be employed for clustering the individuals and forming the initial groups. While the mixture of normal distributions is appropriate for modeling only continuous variables, the binomial, trinomial, or multinomial distributions should be the natural probability density functions for modeling categorical variables. Therefore, for modeling mixed types of variables, a combination of these probability density functions should be the most appropriate modeling strategy.

A joint distribution of a set of continuous and categorical variables can be written as the product of the marginal distributions of some, and the conditional distribution of others, given the values of the selected variables. For example, for two variables A (continuous) and B (categorical), the joint probability is $P(A \cap B) = P(A|B)P(B)$. Olkin and Tate (1961) proposed a model where the joint distribution of continuous and categorical variables $[P(A, B)]$ is the marginal distribution of the categorical variables $[P(B)]$ multiplied by the conditional distribution of the continuous variables, given the categorical variables $[P(A|B)]$. This is known as the location model (LM) (Krzanowski 1988). The categorical variables are arranged in a contingency table where the table categories follow a multinomial distribution and the continuous variables are assumed to follow a multivariate normal (MVN) distribution. However, the parameters of these MVN distributions depend on their location in the contingency table of the categorical variable.

Homogenous Conditional Mixture Model

Recently Lawrence and Krzanowski (1996) proposed the homogeneous conditional Gaussian mixture (HCM) model which is based on the original location model of Olkin and Tate (1961) for clustering n observations into g underlying subpopulations using a mixture of continuous and categorical variables. The method combines all levels of the categorical variables into one multinomial variable with m multinomial levels (or cells). The algebraic details of this model, named the location model for simplicity, are given by Franco et al. (1998).

The HCM model (1) requires the estimation of a vector of means in each of the $m \times g$ cells (a total of $m \times g \times p$ means), (2) has a likelihood function that compares each observation with the cell mean and not with the subpopulation mean, and (3) estimates the means of cells that may be empty and thus are not represented in the sample.

The Independent Mixture Model

Franco et al. (1998) proposed a model where the means, variances, and covariances depend not on the specific (is)th cell but rather on i^{th} subpopulation. The main difference between the independent mixture (IM) model and the HCM model is that the vector of means and the dispersion matrix of the IM are assumed to be equal for all multinomial cells within a subpopulation, whereas for the HCM model, the vector of means and the dispersion matrix are assumed to be

different in each multinomial cell within subpopulations.

As previously mentioned, each observation (y_{sj}) is compared with the mean of the subpopulation (μ_i) and not with the cell mean (μ_{is}) as for the HCM model.

Canonical Variate Analysis

Canonical variate analysis is an ordination method for graphical display that allows groups on the data matrix and focuses on the separation among groups such that it can be used for discriminant analysis. Assume that p attributes are measured in each of the n individuals (matrix of $n \times p$) [i.e., the p attributes measured on the j^{th} individual are represented by $x_j' = (x_{j1}, x_{j2}, \dots, x_{jp})$] and consider that the n individuals are grouped into g clusters ($i = 1, 2, \dots, g$). One objective is to examine whether there are differences between the g groups of n_j individuals ($j = 1, 2, \dots, g$; where $n = \sum_{i=1}^g n_i$). Also, it is assumed that any direction in the p -dimensional sample space is specified by $a' = (a_1, a_2, \dots, a_p)$, thus we will focus (for the j^{th} individual) on the linear combination $y_j = a_1 x_{j1} + a_2 x_{j2} + \dots + a_p x_{jp}$. The more separate the groups are in the space, the easier it will be to distinguish the various groups. One major aim is to find a low-dimensional representation of the data that will approximate the high dimensional configuration where the various groups are distinguished. Canonical variables attempt to explain complex relationships in terms of a smaller number of attributes, and if the correlation coefficient between the original and the canonical variables can be adequately interpreted, it will help to characterize the various groups in terms of the attributes associated with the canonical variables.

The Sequential Clustering Strategy for Classifying Farmers Using Categorical and Continuous Socioeconomic Variables

To illustrate the methodology, data from the on-farm conservation project described in Bellon (this proceedings) is used. The project included a random survey of a representative sample of farming households in six communities of the Central Valleys of Oaxaca, Mexico. A total of 240 households were surveyed. An exercise to rank the sample farmers according to wealth was carried out with assistance from key informants in each of the communities. The key informants identified characteristics pertaining to well off, intermediate, and poor farmer groups. These characteristics were related to variables in the survey, though this was not possible in all cases. For example, having interest (or motivation) was a characteristic of the well off, according to the key informants, however this is a difficult characteristic to measure. The variables used to create the classification of farmers in the study included age, education, family demographics, landholdings, animal holdings, sources of nonagricultural income, land quality, and ownership of agricultural implements such as plows, trucks, and tractors. Farmers were classified based on 9 categorical and 22 continuous socioeconomic variables (Table 1).

Estimation of the optimal number of initial groups

The two rules described by Franco et al. (1998; 1999) were used: the upper tail approach (Wishart 1986) and the

likelihood profile, associated with the likelihood ratio test. Every hierarchical procedure performs $n-1$ fusions, and it is possible to arrange these values in increasing order. These sets of values have a mean and a standard deviation which are then used for selection of the best partition. The upper tail criterion selects as the best partition that which has a distance of fusion within the interval $(1-\alpha)100\%$ of the distribution of the fusion values. Therefore, a partition with one group less requires a fusion outside the α interval.

The likelihood profile is used as a graphical display for observing the changes to the log-likelihood function in relation to the number of groups. The optimal number of clusters occurs when the log-likelihood function shows its highest increase. Using the Ward method on the 240 farmers, the upper tail approach determined the existence of 7-10 groups, and the changes in the likelihood profile showed that 5 groups is where the highest increase occurs (Figure 1).

The dashed line indicates the value of the log-likelihood for the five groups.

Relevant variables for discriminating among groups

A stepwise discriminant analysis was performed to examine the importance of the 22 continuous variables on the delineation of the 5 groups. Results indicated that the most relevant attributes were:

- C1: Male farmer's age
- C2: Years of education completed by male farmer
- C4: Years of education completed by female farmer
- C5: Family members less than 5 years old
- C7: Family members 16-60 years old

Table 1. Code, type, and description of the attributes used to classify 240 farmers.

Code	Type	Description
Q1	Binary	Male farmer knows how to read and write
Q2	Binary	Female farmer knows how to read and write
Q3	Binary	Has oxen
Q4	Binary	Has tractor
Q5	Binary	Has truck
Q6	Multistate	Importance of agricultural work outside farm
Q7	Multistate	Importance of nonagricultural work outside farm
Q8	Multistate	Importance of money from relatives living in Mexico
Q9	Multistate	Importance of money from relatives living outside Mexico
C1	Continuous	Male farmer's age
C2	Continuous	Years of education completed by male farmer
C3	Continuous	Female farmer's age
C4	Continuous	Years of education completed by female farmer
C5	Continuous	Family members more than 5 years old
C6	Continuous	Family members 5-16 years old
C7	Continuous	Family members 16-60 years old
C8	Continuous	Family members more than 60 years old
C9	Continuous	Number of hectares in <i>ejido</i> [†]
C10	Continuous	Number of hectares in communal lands
C11	Continuous	Number of hectares in small holdings
C12	Continuous	Proportion of irrigated maize
C13	Continuous	Number of cattle
C14	Continuous	Number of horses
C15	Continuous	Number of goats
C16	Continuous	Number of pigs
C17	Continuous	Proportion of land category 1 (good)
C18	Continuous	Proportion of land category 2 (medium)
C19	Continuous	Proportion of land category 3 (regular)
C20	Continuous	Proportion of land category 4 (poor)
C21	Continuous	Proportion of land category 5 (very poor)
C22	Continuous	Proportion of area planted with maize

Note: The variables in bold face are those that had the greatest influence in discriminating between farmers and therefore the most influence for forming the groups using the Ward method.

[†] An *ejido* consists of land distributed to rural communities after the Mexican Revolution in the early part of the 20th century. By law, *ejido* land was held and worked communally. Title did not reside with individual members of the *ejido* (known as *ejidatarios*) but with the *ejido* as a government entity. Constitutional reform in the late 20th century made it possible for individual *ejidatarios* to claim title to their land and dispose of it as they pleased.

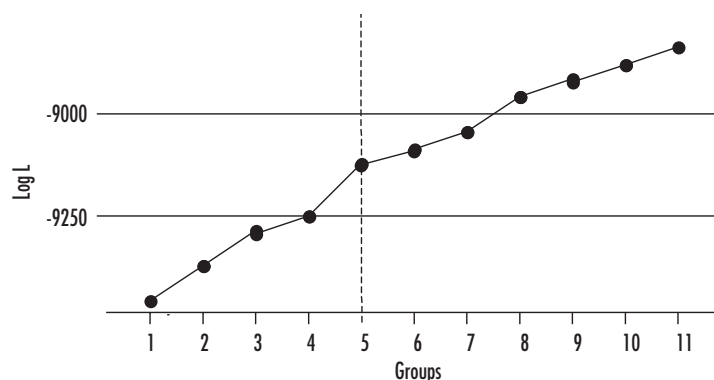


Figure 1. Profile of the log-likelihood function for the number of groups obtained using the Ward method.

- C8: Family members more than 60 years old
- C10: Number of hectares in communal lands
- C11: Number of hectares in small holdings

Years of education, age, and number of hectares seem to be the most important continuous variables for discriminating among the farmers in different groups.

A chi squared test to determine the relevant categorical variables for separating the five groups showed that three binary variables and two multistate variables were the most important discriminators:

- Q1: Male farmer knows how to read and write
- Q2: Female farmer knows how to read and write
- Q3: Has oxen
- Q6: Importance of agricultural work outside farm
- Q7: Importance of nonagricultural work outside farm

Years of education and outside support are the most important categorical variables influencing the groupings.

Ward-MLM

The initial groups formed by the Ward method changed their composition after the MLM method was applied (Table 2). For example, while the initial group of farmers belonging to group G1 comprised 98 farmers, 59 of them remained in G1 after MLM, 16 formed part of final G2, 1 formed part of G3, 18 formed part of final G4, and 4 formed

part of final G5. Similarly, while the initial G2 had 43 farmers, 34 remained in the final G3, but 7 formed part of final G2, and 2 formed the final G5. None of the initial G2 farmers moved to final groups G3 and G4. A total of 32% of farmers were moved from one initial group to another final group. Only three observations were classified in a given group with less than 0.75 probability, i.e., 98.75% of the observations were classified in the 5 groups with at least 0.75 probability.

Characteristics of the five final groups

The 5 final groups, in terms of the 5 binary variables, the 4 multistate variables, and the 22 continuous variables, and after the two-stage sequence clustering strategy Ward-MLM, can be characterized as shown in Table 3.

Final group G1 is characterized by low values for C5 variables; high values for C9, C15, and C18 variables; and a high proportion of YES for the binary variables Q1, Q2, and Q3 (Table 3). This means that households in this group have few very small children. On average they own the highest amount of *ejido*¹ land, most of it of

Table 2. Number of farmers that moved from the initial groups formed by the Ward methods to the final groups obtained after MLM analysis.

Initial groups	Final groups					Total
	G1	G2	G3	G4	G5	
G1	59	16	1	18	4	98
G2	7	34	0	0	2	43
G3	3	1	19	0	0	23
G4	2	0	0	27	4	33
G5	1	0	0	17	25	43
Total	72	51	20	62	35	

Note: Numbers on the diagonal are the farmers that remained in the same group after the modified location model (MLM) analysis.

¹ An *ejido* consists of land distributed to rural communities after the Mexican Revolution in the early part of the 20th century. By law, *ejido* land was held and worked communally. Title did not reside with individual members of the *ejido* (known as *ejidatarios*) but with the *ejido* as a government entity. Constitutional reform in the late 20th century made it possible for individual *ejidatarios* to claim title to their land and dispose of it as they pleased.

good to very good quality. Many households in this group have oxen and the average number of goats. Most of the male and female heads know how to read and write. These families seem to be in the middle of the demographic cycle and have good access to agricultural assets.

Final group G2 has high values for variables C6, C7, C10, C13, and C19; low values for variables C9 and Q6; and a high proportion of YES for variables Q1, Q2, and Q3. These families have access to family labor, given that most of their members are in the age groups of 5-16 years and 16-60 years (some have small children, but most are teenagers and above). They own, on average, the highest number of cattle and largest landholdings in communal areas, and may depend on cattle farming, though on a small scale. Their land is distributed among all land quality types. Most of the

male and female heads know how to read and write. Most have oxen. Off farm labor and, to a lesser extent, non farm labor are important sources of income for these households. The availability of family labor and cattle seem to be important components of their livelihoods.

Final group G3 showed high values for variables C2, C4, C5, C12, and C16; low values for C1, C3, C8, C14, C15, C18, and C19; low values for Q7; high values for Q6, Q8, and Q9; and the highest proportion of YES for Q1, Q2, and the lowest for Q3. These families are the youngest and the best educated of the sample. Of all groups they rely the most on non farm labor. They own the highest proportion of irrigated land and high quality land; however, on average, they own the smallest landholdings. They also own, on average, the highest number of pigs, but no oxen. These farmers are

Table 3. Mean value of the 5 final groups for 22 continuous variables (C1-C22), 4 multistate variables (Q6-Q9), and the proportion of 5 binary variables (Q1-Q5) for each case.

Group	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
G1	55.2	3.5	47.8	3.6	0.3	1.2	3.1	0.3	2.7	0.5	0.4	0.2	1.7
G2	41.4	5.0	34.9	5.1	0.7	2.1	2.7	0.1	0.7	2.4	0.4	0.1	1.7
G3	38.1	8.9	32.6	9.0	1.1	1.8	2.3	0.0	0.9	0.9	0.6	0.3	1.1
G4	68.2	1.4	62.5	0.8	0.3	0.5	1.3	1.7	1.6	0.4	0.8	0.1	1.0
G5	55.3	1.6	51.1	0.2	0.3	1.0	4.6	0.1	0.9	0.1	2.4	0.1	0.4
Group	C14	C15	C16	C17	C18	C19	C20	C21	C22	Q6	Q7	Q8	Q9
G1	1.4	4.1	1.1	0.4	0.1	0.1	0.1	0.2	0.8	2.8	2.4	2.7	2.5
G2	1.8	3.3	1.2	0.4	0.1	0.2	0.2	0.1	0.9	1.9	2.1	2.6	2.9
G3	0.6	1.9	5.1	0.6	0.0	0.0	0.2	0.1	0.7	3.0	1.3	2.9	3.0
G4	1.5	1.9	1.2	0.4	0.1	0.1	0.3	0.2	0.2	2.8	2.7	2.5	2.6
G5	1.2	2.7	1.5	0.7	0.1	0.1	0.1	0.0	0.9	2.5	2.6	2.5	2.5
Group	Q1		Q2		Q3		Q4		Q5				
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes			
G1	6	94	4	96	33	67	96	4	90	10			
G2	4	96	2	98	27	73	98	2	94	6			
G3	0	100	0	100	100	0	95	5	50	50			
G4	38	62	68	32	53	47	100	0	92	8			
G5	37	63	89	11	17	83	100	0	88	12			

Note: Values in bold represent the highest or the lowest values for each variable.

probably the better off families, with strong links to the non farm economy, practicing a more suburban type of agriculture.

Final group G4 had high values for variables C1, C3, C8, and C9; low values for C2, C4, C6, and C7; a high value for variable Q7; and high proportions of NO for Q4 and Q5. Families of G4 can be seen as the opposite of those in G3. They are the oldest and least educated and have the highest number of members in the oldest age group. They plant the smallest area to maize and have, on average, the second smallest landholdings and little irrigated land. These families are probably the worst off: they are old, have little available labor, and few assets. Remittances do not seem to be an important source of income either.

Final group G5 showed high values for variables C7, C11, and C22; low values for variables C4, C10, C12, C13, C20, and C21; a high value for Q7; a high proportion of YES for Q3; and high proportion of NO for Q4 and Q5. These households have the highest average number of members in the most productive age group (16-60 years), they own the largest area of private land, but one of the lowest in communal lands. They have the lowest cattle ownership, but most have oxen. Most of their land is planted to maize. While they own a small proportion of irrigated land, most of it is of very high quality. This group may be the most agriculturally oriented, particularly with respect to maize—in general they have good labor availability, a team of oxen, and private land that can be used as collateral. They may have the highest agricultural productivity potential given the high quality of their land.

The classification method has created a set of groups with similar resources and constraints. It reflects the importance of

the household demographic structure, education, and access to agricultural assets. Rather than interpreting this classification exclusively in terms of poverty or wealth—although these were the basis for eliciting the variables used, and the patterns of wealth and poverty in it are obvious—the classification can be better interpreted in terms of a diversity of livelihood strategies that respond to the control of different assets. In any case, it is clear from the data that the resources controlled by the sample are relatively modest, even for the better off farmer group.

Canonical analysis and canonical plot

The canonical analysis involves only the 22 continuous variables. The first 2 canonical variables explained almost 90% of the variability existing in the entire data set. The pair-wise squared distances between the 5 final groups are shown in Table 4.

Clearly the final groups that are farthest apart are G3 and G4, followed by G3 and G5. The groups that are closest, with a large overlap between farmers, are groups G1 and G2, closely followed by G1 and G5, and G2 and G3. Group G4 seems to be fairly compact and well separated from the others.

The canonical variables are shown in Table 5. The first canonical variable is positively and highly correlated with

Table 4. Pair-wise squared distance between the five (G1-G5) final groups after the Ward-MLM[†] strategy.

	G1	G2	G3	G4	G5
G1	0.00	11.24	24.81	21.14	12.11
G2	-	0.00	14.23	45.14	25.93
G3	-	-	0.00	71.20	54.95
G4	-	-	-	0.00	28.10
G5	-	-	-	-	0.00

[†] MLM = modified location model.

continuous variables C1, C3, and C8, and negatively and highly correlated with C2 and C4. This indicates that the first canonical variable is associated with the age and education of the male and female household heads and the age group of household members in the oldest category. The second canonical variable is positively correlated with C4 and C8 and negatively correlated with C7. This shows that this canonical variable is associated with the education level of the female household head and the demographic composition of the household (particularly the members in the most productive and the oldest age groups). The third canonical variable is negatively correlated with variable C10, which is the area owned in communal lands. The demographic structure of the household and education of household heads are fundamental components of the classification.

Table 5. Canonical variables for each of the 22 continuous variables.

Continuous variable	Canonical variables		
	Can1	Can2	Can3
C1	0.458224	0.029365	0.221170
C2	-0.389385	0.294195	0.168153
C3	0.437052	0.000497	0.173032
C4	-0.444267	0.391052	0.209704
C5	-0.094468	0.093268	-0.085827
C6	-0.150245	-0.001007	-0.164493
C7	-0.090059	-0.407822	0.123471
C8	0.486446	0.471022	-0.219938
C9	0.048433	0.025511	0.326002
C10	-0.130322	0.099257	-0.435796
C11	0.051708	-0.234342	-0.017993
C12	-0.042150	0.119581	0.224169
C13	-0.034151	0.043551	-0.006289
C14	0.012748	0.004001	-0.194377
C15	-0.021852	-0.031214	0.031700
C16	-0.052913	0.065072	0.113935
C17	-0.021000	-0.129863	0.114719
C18	0.034930	-0.014545	0.032909
C19	-0.025951	-0.047472	-0.119569
C20	-0.002225	0.124978	-0.164192
C21	0.023587	0.073985	0.078142
C22	-0.059653	-0.145481	-0.176298

Note: Values represent correlations between canonical variables and the original variables. Values in bold represent the highest or the lowest values for each variable.

Figure 2 shows the plot of the first canonical variable against the second. This graphical representation is useful for visualizing the relationship between groups. It is clear that G4—older households with a low education level—forms a very compact group, well separated from the others, as well as G3—younger households with the highest education level—and G5. Groups G1 and G2 represent intermediate groups with several overlapping observations in terms of the two canonical variables.

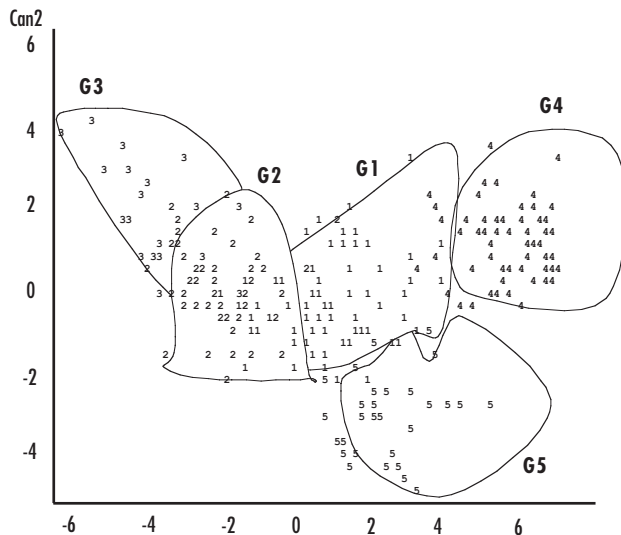


Figure 2. Plot of the first two canonical variables for 240 farmers from the canonical analysis using 22 continuous variables.
Final groups after the Ward-MLM clustering strategy are G1-G5.

Conclusions

This paper presents a method for classifying individuals—farming households in this example—into homogenous but distinct groups. The method allows the use of different types of variables, provides a systematic approach to decide the number of groups present in the data, and assigns a probability that an individual belongs to a group. The value of this method to practitioners of PPB is that it should allow them to group farmers into homogeneous groups, hopefully with similar variety requirements. It should be pointed out, however, that the link between the groups formed and the variety requirements is not automatic. It depends on the choice of variables used in the classification and their relationship with the demand for crop traits. It requires testing of the association between the groups and the crop traits and, hence, the varieties demanded.

References

- Franco, J.E., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1997a. Classifying Mexican maize accessions using hierarchical and density search methods. *Crop Science* 37:972-980.
- Franco, J.E., J. Crossa, J. Díaz, T. Taba, J. Villaseñor, and S.A. Eberhart. 1997b. A sequential clustering strategy for classifying gene bank accessions. *Crop Science* 37:1656-1662.
- Franco, J.E., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1998. Classifying genetic resources by categorical and continuous variables. *Crop Science* 38:1688-1696.
- Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1999. A two-stage, three-way method for classifying genetic resources in multiple environments. *Crop Science* 39:259-267.
- Krzanowski, W.J. 1988. *Principles of multivariate analysis. A user's perspective*. Oxford, U.K.: Oxford University Press.
- Lawrence, C.J., and W.J. Krzanowski. 1996. Mixture separation for mixed-mode data. *Statistics and Computing* 6:85-92.
- Olkin, I., and R.F. Tate. 1961. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* 32:448-465 (correction in 36:343-344).
- Wishart, D. 1986. Hierarchical cluster analysis with messy data. In W. Gaul and M. Schader (eds.), *Classification as a tool of research*. New York: Elsevier Science Publishers BV.

Discussion Summary

The discussion centered on how to identify and characterize socioeconomic environments that are appropriate for participatory breeding. We do not have enough understanding of the way different socioeconomic variables determine or influence the demand for varietal traits by farmers. This understanding should be the basis for selecting variables to be included in a classification exercise, using the methodology described here, to identify the “social environments” that the breeding should target. It was suggested that one way of looking at the data is to set up hypotheses proposing that certain socioeconomic variables (e.g., animal ownership) would affect cultivar adoption. By generating relevant hypotheses from current varietal adoption patterns, multivariate data (variables) for effective classification can be identified. There is a lot of literature on the factors affecting adoption, but it is still not clear how to make these data relevant to participatory breeding. Another option is to use participatory diagnostics to determine the parameters to use.

There was also discussion on the advantages and disadvantages of looking at the impact of individual factors on the adoption or demand of traits versus using multivariate groupings of factors. It was pointed out that there is no inherent contradiction between both approaches. Clearly in the former approach, the impact of one factor at a time can be tested (keeping the others constant), while the latter does not allow this. However, in many cases factors are highly correlated (i.e., there is high multicollinearity in the data), hence, testing individual factors is difficult. No matter which approach is chosen, it is important to make clear hypotheses at the beginning.

It was emphasized that the most important objective of this presentation was to make participants aware of a powerful methodology to generate homogenous groupings. At the end of the day, any participatory breeding effort has to deal not only with biophysical heterogeneity, but also with socioeconomic and even cultural variability. For practical reasons it is important to segregate this variability into units that can be identified, characterized, and targeted. Any participatory plant breeding effort should target certain “recommendation domains” and therefore should have tools to accomplish this.