ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Agricultural Systems 76 (2003) 141–157

AGRICULTURAL
SYSTEMS

www.elsevier.com/locate/agsy

Typification of farming systems for constructing representative farm models: two illustrations of the application of multi-variate analyses in Chile and Pakistan

C. Köbrich[a], T. Rehman[b,*], M. Khan[c]

[a]Department of Animal Production, University of Chile, Casilla 2 Correo 15, La Granja, Santiago, Chile
[b]Department of Agriculture, The University of Reading, PO Box 237, Earley Gate, Reading RG6 6AR, UK
[c]Agricultural Economics Research Unit, National Agricultural Research Council, Park Road, Islamabad, Pakistan

Accepted 1 March 2002

Abstract

If the fundamental precepts of Farming Systems Research were to be taken literally then it would imply that for each farm 'unique' solutions should be sought. This is an unrealistic expectation, but it has led to the idea of a recommendation domain, implying creating a taxonomy of farms, in order to increase the general applicability of recommendations. Mathematical programming models are an established means of generating recommended solutions, but for such models to be effective they have to be constructed for 'truly' typical or representative situations. The multi-variate statistical techniques provide a means of creating the required typologies, particularly when an exhaustive database is available. This paper illustrates the application of this methodology in two different studies that shared the common purpose of identifying types of farming systems in their respective study areas. The issues related with the use of factor and cluster analyses for farm typification prior to building representative mathematical programming models for Chile and Pakistan are highlighted.

Keywords: Farm types; Multi-variate analysis; Farming systems research; Mathematical programming

* Corresponding author. Tel.: +44-118-931-8480; fax: +44-118-931-6747.
E-mail address: t.u.rehman@reading.ac.uk (T. Rehman).

PII: S0308-521X(02)00013-6

1. Introduction

In the 'systems approach' to agricultural research and extension, a farming system is seen as comprising the totality of production and consumption decisions taken by a farm-household, including the choice of crop, livestock and off-farm enterprises, and food consumed by the household (Byerlee et al., 1980). Nevertheless, no farm-household has the same resources or problems. This implies that every farming system is different, if not unique, facing distinctive decision-making problems, whose solutions could also be unique. Unfortunately it is not feasible in practice, making it necessary to classify or group farms in some way. Such groups constitute the so-called recommendation domains, "a group of roughly homogenous farmers with similar circumstances for whom we can make more or less the same recommendation" (Byerlee et al., 1980). Clearly, the 'best' typology of farms will have to show a maximum amount of heterogeneity between the types, while obtaining maximum homogeneity within particular types or categories, for it to be truly representative of the categories represented. But, the methodologies such as 'sondeo' and 'rapid rural appraisal' used to define domains, which are simple and straightforward, have tended to equate the recommendation domain with a geographic area, when in fact geographic areas exhibit great diversity (Escobar and Berdegué, 1990). The failure to recognise and deal with the heterogeneity of farming systems within a geographic area is one of the main criticisms of Farming Systems Research and Extension.

The ordering of farms, mainly from a geographic point of view, into distinct types has been practised since the beginning of this century. For example, the Commission on Agricultural Typology was established with the task of determining common principles, criteria, methods and techniques for agricultural typification, and to specify the typological and regional classification of world agriculture (Kostrowicki, 1977). This Commission also recognised the importance of typification based on quantitative methods, as qualitative typification based on expert opinion could show different results with time. It was recommended that the typification variables should be limited in number and that they should be synthetic or composite in nature (Kostrowicki, 1977).

Agricultural systems have been classified mostly according to qualitative criteria using subjective assessments based on arbitrary and ad hoc considerations. These classifications also tend to be descriptive rather than explanatory or predictive; two such examples are Spedding's (1988) general classification of agricultural systems and the classification of goat production systems in Peru which is based on environmental, socio-political and economic factors (Perevolotsky, 1990).

The objective methods of classification can provide an exhaustive array of types, allowing condensing of large data sets, and thus helping a researcher to identify those types that are needed for analysis. However, the literature on farming systems research cites only a limited published research on the applications of quantitative typification methods or, indeed, the discussion on the theoretical aspects of typification of farming systems. The examples that can be found include a socio-economic classification of German farm households (Gebauer, 1987), and a cluster analysis of farming systems in Central North China (Hardiman, 1990). However, for Latin-America, the situation is rather better as a set of typification exercises showing a variety of methodological variations have been presented in Escobar and Berdegué (1990). They include the contributions by Berdegué et al. (1990) in Chile, by Landín (1990) in Ecuador, by Duarte (1990) in Colombia, by Martínez et al. (1990) in Guatemala, and by Douglas (1990) in the western Caribbean.

Recently two studies have been carried out at The University of Reading that faced the common problem of creating typical representations of farming systems in two different countries, Chile and Pakistan. In both situations classification schemes were applied to data in order to develop mathematical programming decision-making models for typical farming systems. The purpose of this paper is to present the most important aspects related to the use of multi-variate statistical analysis in the typification of farming systems. The problems encountered in the typification of farming systems are highlighted and the implications for the construction of representative farm models are drawn. The paper begins by detailing the methodology followed by a discussion of results from the two applications before offering some concluding observations on the importance of defining what a "representative farm" is.

2. Methodology

A continuum of six stages of the procedure to establish the farming system (FS) types can be identified (Escobar and Berdegué, 1990): (1) determination of the specific theoretical framework for typification, (2) selection of variables, (3) collection of data, (4) factor analysis, (5) cluster analyses, and (6) validation. The theoretical framework defines the purpose of classification and establishes the hypothesis (es) to guide the process of typification. The inputs required at the beginning are the researchers' previous experience and knowledge of the area, the objectives of the typification exercise and, the quantitative information that is available about the study area's agriculture (Escobar and Berdegué, 1990). The translation of the hypotheses into a set of variables that can be used in the typification scheme is the most critical step, but it is also the least understood. It is necessary to assess the relevance of the variables to the problem being investigated (Aldenderfer and Blashfield, 1984). Although there cannot be a general rule for the selection of variables, the most commonly used ones include farm size, capital, labour, production pattern, soil quality, and managerial ability (Escobar and Berdegué, 1990). Further the identification of types ought to be based on internal and not on external attributes. The use of both types of attributes would presuppose rather than show their impact on the identification of farming systems (Kostrowicki, 1977).

The information collected on the various variables can be processed prior to factor analysis and those variables, which do not show variability, can be discarded. First, any variable that makes little contribution, in terms of its variability, to the measure of distance being used to form clusters can be discarded (Escobar and Berdegué, 1990). Second, some variables may not be relevant to the typification required for the purposes of a particular study and can therefore be discarded, even though the typology obtained initially is consistent with observable farming systems

(Berdegué et al., 1990). Thus a researcher has to assess if the information imparted by a variable is consistent with research objectives. Third, highly correlated variables can be eliminated as an uncritical use of such variables to compute a measure of similarity is essentially an implicit weighting of these variables (Aldenderfer and Blashfield, 1984). For a given variable that is highly correlated to another one, its contribution to the measure of distance is reflected by changes in other variables. Finally, cluster analysis requires that if data are missing the complete observation be discarded or else average values be used. The first option reduces the number of farms, while the second may bias the results. So if possible, the variables should be discarded instead of the observations.

The purpose of the fourth stage (factor analysis) is to reduce the number of variables and thus the 'dimensionality' of the problem. Factor analysis is often used when the variables used in the study are known to be correlated (Aldenderfer and Blashfield, 1984). It is concerned with the internal relationships of a set of variables and is aimed at constructing a set of factors (hypothetical unobserved variables) from a set of observable variables (Lawley and Maxwell, 1971). In other words, observed values (Y) are explained through a linear combination of factors (B) and a residual (E) or $Y = XB + E$. The factors are *common* when they contribute to the variance for at least two observed variables or *unique* when their contribution is only towards one variable. A correlation matrix for a set of observations (R-factor analysis) is prepared or, less frequently, for individuals for a set of variables (Q-factor analysis). Then the initial factors are extracted which can be based on defined factors, principal component analysis (PCA), or on inferred factors, that is common factor analysis (FA). As the exact configuration of the factor structure is not unique, one factor solution can be transformed into another one or rotated to a terminal solution. This can achieve simpler and more meaningful factor patterns, instead of the highly complex extracted factors that are related to many of the variables rather than to just a few (Kim, 1970; Comrey and Lee, 1992).

The main difference between PCA and Common FA is how the communalities are computed, that is the fraction of each variable's variance that is explained by the total of the extracted factors. Communality represents the extent of overlap between the extracted factors and the variable and it equals the sum of squares of the variable's loadings across factors (Comrey and Lee, 1992). As PCA is based on statistical variance, the first chosen factor accounts for most of the variance in the data. The second is chosen in the same way but it has to be orthogonal to the first. The last factor explains all the residual variance (Kim, 1970; Lawley and Maxwell, 1971). The common FA is a covariance or correlation oriented method based on the assumption that each variable is influenced by a set of shared or common factors that determine the correlation between variables. The implied expectation is that the number of common factors will account for all the observed relations and that such factor will be less than the number of variables (Lawley and Maxwell, 1971). In common factor analyses the correlation matrix is transformed before undertaking factor analysis (Kim, 1970).

Determining how many factors should be retained is a problem, as with real data the actual number that merit retention is often considerably smaller than the number of variables. One test searches for a point where there is a break in the Eigenvalues, that is, the variation in the original group of variables, which is accounted for by a particular factor. As factors are extracted from large to small, their Eigenvalues are also decreasing. When they are plotted, a straight line can be drawn through the latter smaller values. The earlier, larger values will fall above the straight line. It is proposed that the number of factors to be retained is at the point where the last small factor is above the line, giving an indication of how many factors there are (Comrey and Lee, 1992). Another test defines a threshold level for the residual correlation, beyond which it would be unnecessary to continue extracting, as any new factor would have very small loadings (Comrey and Lee, 1992). A common rule is to extract all the factors with Eigenvalues of 1.0 or more (Kaiser's rule). Whatever rule is used, it must be kept in mind that it is better to err on the side of extracting too many factors rather than too few, as the idea is to extract enough factors to be relatively certain that no more factors of any importance were discarded (Comrey and Lee, 1992).

Next, cluster analysis is used to classify the observations according to *m*-variables of an *n*-dimensional attribute space, by computing the similarity between any pair of observations though a distance coefficient (Sokal, 1977). The agglomerative hierarchical clustering models form an initial partition of *N* clusters (each object is one cluster) and they proceed, in a stepwise manner, to reduce the number of clusters one at a time until all *N* objects are one cluster. All models can be characterised by a set of *N* partitions and their corresponding criterion values 'α'. The hierarchical methods differ on how 'α' is defined (Mojena, 1977). Once the cluster sequence has been established where the process will be cut and thus how many clusters will be defined can be determined. The heuristic procedures and formal tests are two basic means for determining the number of clusters (Aldenderfer and Blashfield, 1984). Of the two methods the heuristic procedures are used most commonly. The hierarchical tree (dendrogram) can be 'cut' through subjective inspection or, more formally, by plotting the number of clusters against the change in the fusion coefficient (i.e. the difference between the distance coefficient at one clustering stage and the previous one). A flat or even curve suggests that no new information is portrayed by the mergers that follow. Further, when two dissimilar clusters are merged, the slope of the distance coefficient curve gets steeper and 'jumps' can be seen. However, the problem of determining when a 'significant jump' occurs remains (Aldenderfer and Blashfield, 1984).

As a means of solving this problem, 'stopping rules' have been defined to determine which partition best approximates the underlying populations (Mojena, 1977; Milligan and Cooper, 1985). These rules can be based on the distribution of the criterion 'α'. A significant change in 'α' from one stage to the next implies a partition which should not be accepted. One stopping rule is based on the mean and standard deviation of the $N-1$ items in the distribution of 'α' (Mojena, 1977). This means that an optimal partition of a hierarchical clustering solution is selected when $\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}$ is satisfied; where α_{j+1} is the value of the criterion on the stage $j+1$ of the clustering process, k is the standard deviate, and $\bar{\alpha}$ and s_{α} are, respectively, the mean and unbiased standard deviation of the α-distribution. This rule essentially parallels a one-tail confidence interval based on the fusion values. If no such value satisfies the inequality, the solution is one of the three possibilities: (1) one cluster, (2) the stage j for which $j+1$ yields the largest standard deviate, or (3) some other heuristic rule is required (Mojena, 1977). The problem with this approach is the value of the standard deviate. When tested with artificial data sets ('natural clusters'), the best fit between the natural clusters and the clusters established by the stopping rule were found when using values in the range of 1.25–3.00 (Mojena, 1977; Milligan and Cooper, 1985). Such a diverse range for k has a significant effect on the partition selection.

Once all clusters have been established, each representing one of the real farming systems, they still need to be validated. It is important to ensure that these groups are 'real' and not merely imposed on the data by the method being used for classifying (Aldenderfer and Blashfield, 1984). The problem is how to carry out significance, or 'optimality' tests, to validate the classification (Sokal, 1977) as no formal procedure has been developed for this purpose. A good alternative thus is to contrast the FS types with the hypotheses about its structure, as well as with the researcher's perception with regard to the variety of FSs that have been observed empirically (Escobar and Berdegué, 1990). Finally, it must be mentioned that for classes or categories to be meaningful, and useful, they have to be related to the purposes for which they are being created; therefore, the fact that they serve the purposes for which they are intended, provides the most meaningful way of testing their conceptual validity.

3. The Chilean application

This project was concerned with the analysis of the impact of the development policies on peasant farming systems in a micro-region of Chile's coastal mountains (Köbrich, 1997). The hypothesis being tested was that the impact of these policies would depend on the levels of available resources such as labour, land (including the pattern of its use) and capital. Despite the predominant prevalence of wheat–livestock production pattern, great variability in the availability of labour, land and herd size was observed and therefore it was necessary to obtain a typification of farming conditions. The data set for 67 farmers included information on location, productive orientation, household structure, on-farm and off-farm labour use, available land, land use capability, actual land use, and livestock numbers.

The criteria of missing data, variation, relevance, and the presence of correlation were used to determine if a particular variable would be used in typification. First, the three variables related with land use capability (arable, non-arable, and non-agricultural land) as well as the variable, irrigated land, were discarded due to missing data. It was thought that the impact of the elimination of these four variables on the generation of clusters would be small, as irrigated land was scarce and land quality was related to land use. In fact "arable land" and "non-arable land" were highly correlated to other variables. Next, the variables "number of poultry", "number of pigs" and "number of sows" were discarded because no farm possessed them in large numbers, nor were they important from the point of view of the research. Similarly, the county in which the farm was located had no relevance, because the unit of analysis in this study was defined as the micro-region, and there was no reason to justify a typology based on location of farms. If the differences between counties do exist, they should become evident after farm typification. Further, to include the qualitative variable 'county' would require its replacement by a set of variables per county (one less than the number of categories) with values 0 or 1 defining the location of the farm. The third criterion, variability, was evaluated through the coefficient of variation. It was established on an a priori basis that the variables with a coefficient of variation of less than 50% would not be considered. The variables, "manager's age" and "time spent by the manager on farm" did not match these criteria. The correlation coefficients between the remaining variables were also computed to determine which other variables could be discarded. The two variables "land given out" and "area of vineyards and orchards" were not correlated to any other variable and were thus included.

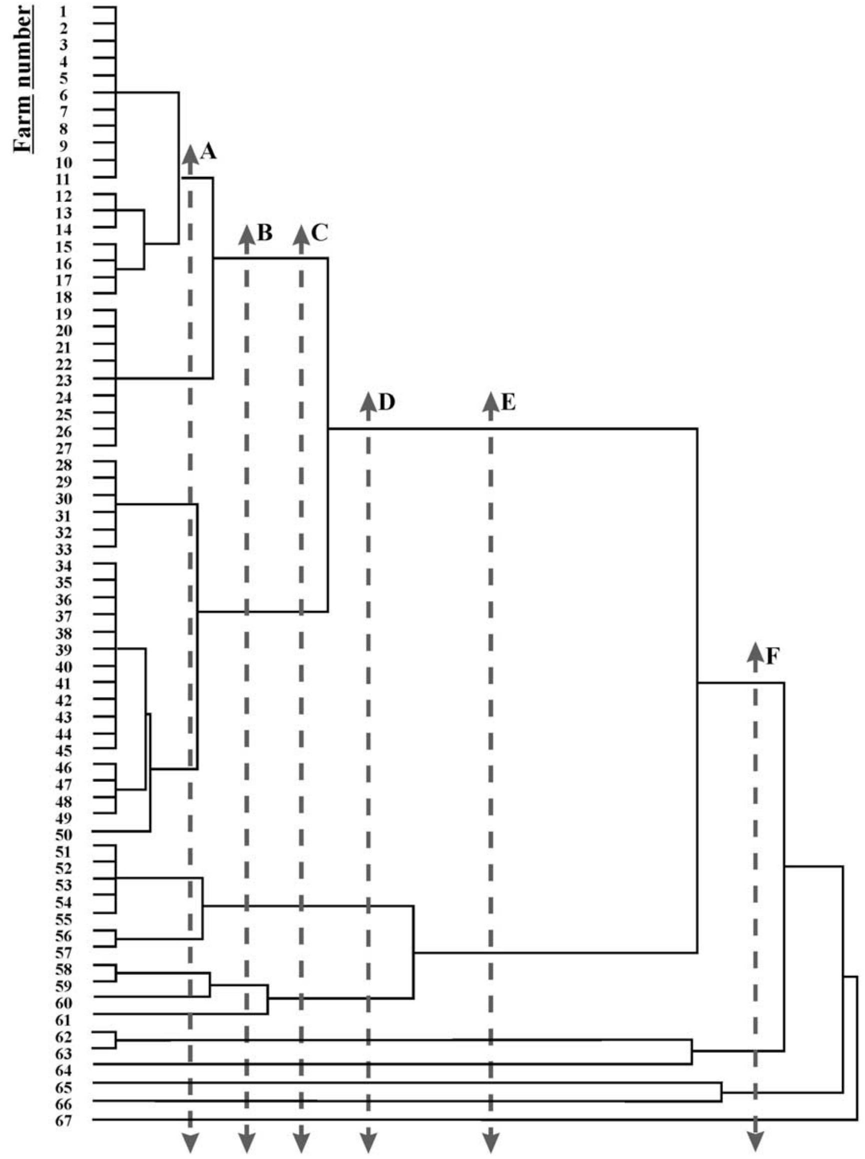
Second, three pairs of highly correlated variables ($R^2 \geqslant 0.90$) were found and one variable from each pair was then discarded. The criteria that determined which variable to keep from each pair were a variable's relevance, the quality of the data obtained from the farm, and the availability of the data on a particular variable. From the pair "total available land" and "natural pastures", the second one was discarded because it is more susceptible to year-by-year change than the total area of available land. Other correlated variables were the two pairs of variables relating the number of livestock to the number of females (i.e. cattle and sheep and goats). The variables "number of cows" and "total number of ewes and does" were kept as data on female livestock has a steadier level and to a great extent determines the total number of livestock. Next, the variables with low correlation coefficients were included in the final data set. These variables were "number of months women spend working on-farm", "area under artificial pastures", and "agricultural land used for other crops". The 10 remaining variables were analysed one by one. The variable "time spent by other male members working on-farm" was included because up to this stage only two variables to which it was correlated were in the data set. The "area of owned land" was excluded because of its high correlation with "total available land" and because it was correlated to three variables already in the data set. The variables "area of land taken-in for sharecropping" and "area under the forage crop" were correlated to only two already selected variables and were thus included. All other variables (arable land, area of woodland or forests, area of bushes, area of non-agricultural land, number of oxen, and number of horses) were discarded because they had at least three correlated variables in the final set. As a result of this process 14 variables were discarded and only 11 were kept for analysis. The high level of correlation between the variables means that a lot of information is redundant, confirming the view that typification surveys should contain relatively few questions but many observations (Escobar and Berdegué, 1990).

The 11 selected variables were then used to construct the factors using principal components analysis. Depending on the criteria used a variable number of factors could be retained. When Kaiser's criterion (Eigenvalue > 1) was used four factors were retained. On the other hand the residual correlation rule suggested extracting eight factors, while the straight-line rule suggested retaining six factors. Considering that a strict selection of correlated variables had been made it was decided that a rather large number of factors should be retained and thus the first seven were extracted, explaining 85.4% of the total observed variation and at least 70% of every original variable's variation.

The seven retained factors were used in for cluster analysis using Ward's minimum variance criterion as the clustering method (SAS, 1985). This method optimises the minimum variance within clusters and it tends to find (or create) clusters of relatively equal sizes and shapes as hyper-spheres. It works by joining those groups or cases that result in the minimum increase in the within-groups sum of squares or the error sum of squares $\mathrm{ESS} = x_i^2 - \frac{1}{n(\Sigma x_i)^2}$ (Ward, 1963; Aldenderfer and Blashfield, 1984). When compared to the other variants of hierarchical agglomerative methods, Ward's method generates clearer solutions (Aldenderfer and Blashfield, 1984). The squared Euclidean distance (the sum of the squared differences in values for each variable) was used to measure the interval between observations. The dendrogram in Fig. 1 shows the sequence by which the observations and clusters were merged. For the study an ideal a priori distribution would have been a small number of similarly sized groups and no clusters formed by single farms. In this sense the cutting line B (Fig. 1) shows a subjective partition which tries to balance the number of clusters with the number of lost observations. That line creates five clusters (farms 1–27, 28–50, 51–57, 58–60 and 62–63) while five farms remain unclassified. Despite forming a cluster, the farms 62 and 63 should probably be discarded as their relevance to the whole sample is rather limited. It can also be seen that if the line is shifted to the left (line A) eight clusters are formed (three with only two observations) with six farms still remaining unclassified. Shifting the line to the right (line C) merges farm 61 with farms 58–60. A further shift to the right (line D) creates one very large cluster comprising 50 farms and two small ones which was not appropriate for the research project.

Following a more formal approach, both the distance coefficient and the increase in its size were plotted against the number of clusters. It was seen that until 18 clusters remained, the distance between adjoining clusters was small and fairly constant, without important jumps. Then the increase in the value of coefficient became bigger, but no meaningful jump was observed until 14 clusters were formed. Accordingly, this method suggests that line A (Fig. 1) gives the appropriate number of clusters for this sample. Finally when Mojena's stopping rule was used (setting k at 1.25), two clusters and six lost observations were generated (line E); but when k was set at 2.75 or 3.00 it generated one very large cluster, two very small clusters and one unclassified observation (line F).

To determine which of these cutting lines is more appropriate to determine the farming systems typology, the results have to be examined from the point of view of the research objectives. Mojena's stopping rule is not very strict as it is trying to find differences between individuals of an apparently homogenous population. The Chilean application, however, has contrasting situation as groups with maximum similarities among individuals within a group are being created from a heterogeneous population. Thus the line C in Fig. 1 was defined as the cutting line.



Note: The farms have been recoded to simplify the description of the results

Fig. 1. Dendrogram showing the full history of cluster construction and six possible cutting lines (Chile).

On comparing the clusters it was seen that labour variables were very important in differentiating all clusters (Table 1). The largest single difference between cluster 1 (farms 1–27) and cluster 2 (farms 28–50) was the availability of female labour, while male labour makes a distinctive difference between these two clusters and cluster 3

Table 1
Comparison of selected clusters of farms for the Chilean application

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Farmer on farm	One year	One year	One year	One year	Half a year
Additional labour	One woman	Marginal	Two men	Two women, two men	Marginal
Farm size	Small	Small	Medium	Large	Small
Herd	Small	Small	Small	Large	Large
Arable/available	56.6%	86.5%	67.0%	19.6%	–
Crop/arable	30.9%	17.4%	27.8%	30.0%	–
Sharecropping	Takes-in	Takes-in	Takes-in	–	Gives-out

(farms 51–57) and 4 (farms 58–61). The female labour is also relevant but not unique in distinguishing cluster 4 from 5 (farms 62 and 63). The clusters 1 and 4, specially 4, have less arable land, as they use labour more intensively and their ratio of crop over arable land is almost double that of cluster 2; for cluster 3 better availability of labour allows this ratio to be high. Under a normal rotation for that area, the expected crop/arable ratio would be 20% (1 year fallow, 1 year crop, and 3 years rough grazing). Further the farms of type 2 have over 80% of their area under pastures, while around 18% of farm land of types 4 and 5 are under other use (mainly bushes) or woods.

For the Chilean data the clustering method has been used to define various farming systems; therefore, an important issue is how valid are the resulting farming systems types. As cluster analysis allows grouping any collection of individuals or observations according to any set of variables, it is necessary to determine if the typology that has been generated represents an observable classification and not the one imposed on the data by the cluster analysis itself. In other words the same set of observations in different contexts may lead to distinct typologies, each of them suited for a particular purpose. Thus the usefulness of a typology is generally restricted to the context in which it was constructed. The results of this analysis lead to farming systems that reflect the significance of different resource endowments and would therefore elicit different responses to the development policies from these systems. The distribution of farms across counties also showed a distinctive pattern (Table 2). Although the χ^2 test should not be used to analyse these results, it could be seen that the distribution was not random. The farms located in the counties of Litueche and Marchihue were concentrated in Cluster 2 while 75% of Pumanque's farms belonged to Cluster 1 and none to Cluster 2. The observed distribution in Cluster 1 and Cluster 2 was very different from the expected values of around 35–40% of each county's farms in Cluster 1 and Cluster 2, respectively. It should be mentioned that the four unclassified farms were located in the same county and had either large areas of vineyards, forage crops, artificial pasture, or large sheep and/or goatherds.

Up to this stage no consideration was given to the area currently under a given crop or the farm's production pattern and only the variable area of arable land had been considered. The reason being that crop areas may vary from year to year and

Table 2
Percentage of farms of each county allocated to every cluster: the Chilean application

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Litueche	7.7	69.2	7.7	7.7	7.7
Marchihue	13.3	60.0	10.0	3.3	–
Pumanque	75.0	–	–	20.8	4.2
Expected	34.3	42.9	6.3	11.1	3.2

thus affect the clustering results. It was therefore judged preferable to use a farm's production pattern as a typification criterion. The problem is that the inclusion of one qualitative variable per each production pattern in factor and cluster analysis would imply a large increase in the number of variables. To avoid this it was decided that after clustering a cross-tabulation between clusters and production patterns would put a greater emphasis on present activities. In this way each of these cluster–production pattern pairs was then identified as a farming system (FS). Of a maximum possible of 30 FSs, 16 were not empty. Of these only eight FSs had four or more observations, one had two observations while the other eight had only one farm (Table 3). Thus, even as the observations for each cluster are spread on various production patterns, it can be seen that they were concentrated around one or two production patterns. The distribution of farms in a given cluster across counties as well as across production pattern was not a random one, that is farms belonging to each cluster tended to be located in certain counties and have certain production patterns.

It can therefore be argued that the typification process was able to determine some underlying structure, as county and production pattern pairs had not been considered in the set of clustering variables. Such a non-random distribution strongly suggests that a valid typology was developed. Such a typology deals with the differences in resource endowment (mainly natural environment) and can therefore be used for the assessment of the impact of development policies on the sustainability of the farming systems. Finally, a multiobjective programming model was constructed for eight of these farms, each representing one of the more relevant types. As different development policies were evaluated using these farm models, it was seen that the impact showed significant differences between farms. These results underscore the validity of the typology obtained further.

4. The Pakistan application

The farming systems in Pakistan have been classified into ten broad crop agro-ecological zones (PARC, 1980) which were identified, in the main, on the basis of the variation in agricultural, climatic and soil conditions. The socio-economic features of the farming systems were hardly taken into account in the delineation of these zones. The resulting classification is, therefore, very general and has a limited usefulness for

Table 3
Cross-tabulation of farms according to cluster and production pattern: the Chilean application

Production pattern	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Wheat–sheep	1	4			
Wheat–legume–sheep	7	11	1	1	1
Wheat–maize–orchard		1			1
Wheat–pasture–cattle	11			5	
Wheat–pasture–cattle–legume–maize		4	1	1	
Wheat–vineyard–sheep	4	7	2		

defining representative farming systems for modelling purposes. The inclusion of the socio-economic variables in the classification scheme is critical to the identification of representative farming system situations. In presenting the results from the Pakistan data, the methodological nuances are not discussed as most of these have been covered in the previous two sections.

A representative sample of 72 farmers from the rice–wheat zone of the irrigated Punjab of Pakistan were interviewed in 1995. This study was concerned with examining the influence of multiple objectives on introducing 'new' crops to the existing cropping systems in the area. As part of the research strategy, the sample was partitioned into representative farming systems for which different mathematical programming models were to be built (Khan, 1998). Cross-sectional data were collected on five sets of qualitative and quantitative variables. These variables were related to: (1) characteristics of the farm and the household; (2) productive resources and related factors such as land ownership, land type, irrigation sources and fragmented nature of land parcels; (3) farm labour (family labour and permanent hired labour) and inventory of farm machinery and equipment; (4) combination of farm enterprises (crops grown currently, new crops and livestock numbers); and (5) farm and off-farm income sources.

Information on 40 quantitative and qualitative variables, in all, was collected. Before carrying out principal component analysis (PCA) variables with smaller than 25% variation coefficient and high correlation coefficients were excluded from analysis. Nineteen variables were excluded because they were either highly correlated or they had low variation, leading to a considerable reduction in the dimensionality of the typification problem. PCA was used to obtain a linear transformation of the remaining set of variables into factors representing most of the information contained in the original set of variables (Dunteman, 1989). Principal components with Eigenvalues less than one were dropped, as with each succeeding component a smaller amount of variance is explained and the remaining components are less likely to be not correlated. The first factor explained greater proportion of variance (14%) as compared to the following one. In total five factors are required to explain over half of the total variance. Eight factors were retained which accounted for 67% of the variation.

Ward's clustering method was used to construct the typology of farms. The nested tree structure of the dendrogram suggests that many different groups are present

(Fig. 2). Obviously, where to cut the tree in order to obtain an optimal number of groups is an issue. Using the heuristic procedures, based on an subjective judgement, the hierarchical tree could rationally be cut into six clusters at the cutting line C (Fig. 2). The farms 22 and 62 could not be considered as a separate type as the

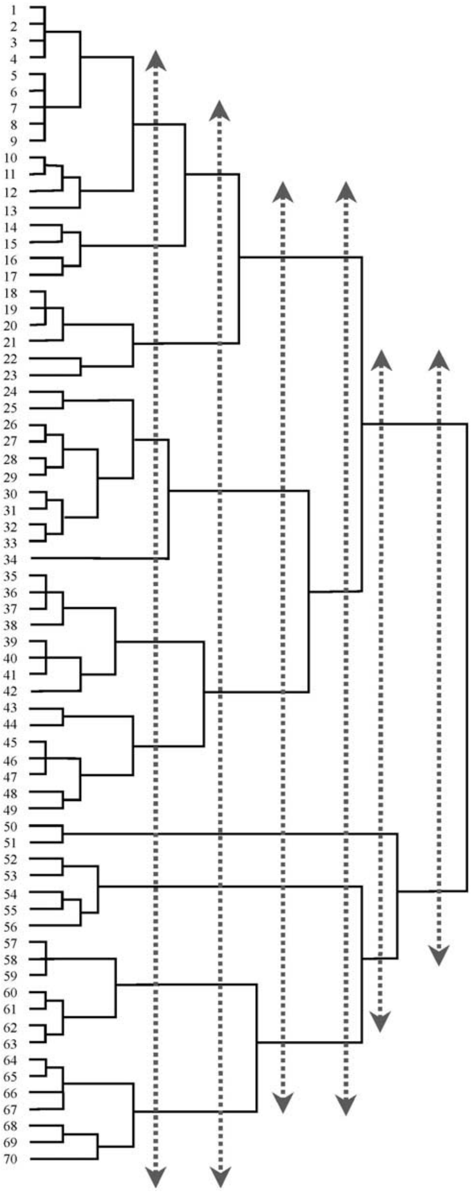


Fig. 2. Dendrogram showing the full history of cluster construction and six possible cutting lines (Pakistan).

number of farms is insufficient to consider it a relevant group. The cutting line C therefore creates five clear clusters with sufficient number of farms in each cluster. Shifting the cutting line B to its left (lines A and B) increases the number of clusters, whereas on shifting it to its right (lines D, E, and F) that number is reduced.

To examine this issue more formally the distance coefficients and changes in these coefficients were plotted against the number of clusters. The graphical presentation highlights the marked "flattening" and "quantum jump" in the value of distance coefficients (Fig. 2). A marked "flattening" in the squared distances begins at the sixth cluster solution and becomes flatter with the decreasing number of cluster solutions. Similarly, a quantum jump between the five and four cluster solutions was found which indicates that two dissimilar clusters have been merged. Both of these procedures suggest that a five or six cluster solution is appropriate. Finally Mojena's stopping rule was used by setting k at 1.25 and eight clusters were generated (line B in Fig. 2), but when k was set at 2.75 or 3.0 it generated one very large one and another relatively small cluster (line F).

The procedures used above to define the number of clusters have shown different preferences for a few to a large number of cluster solutions. The line C was selected as the best solution considering the needs of this research project and the overall structure of data shown in the dendrogram (Fig. 2). Each cluster displays some unique characteristics, which could be used to establish typologies of the main farming systems. The major characteristics are: (1) operational size of holding; (2) land types; (3) family and hired labour profiles; (4) level of mechanisation; (5) tenancy status; (6) full-time or part-time managers; (7) livestock numbers; (8) household farm and off-farm income sources; (9) distance of farms from roads and market; (10) land allocated to different crops; and (11) joint or single family farms. Although many similar characteristics resolve the farm typologies across the zone, but still a significant difference was accounted for in their frequency as well as their potency as shown in Table 4.

The clusters that have been identified are differentiated primarily in terms of the physical and socio-economic characteristics of farms. The differences in the size of holdings, tenancy status, land types, and mechanization level are recognised as physical factors in the formation of clusters. For instance cluster 1 contains large and mechanized farms of the rice–wheat farming system. On the contrary cluster 2 consists of farms with smallest land resources that are also dependent on hired labour and land is cultivated using hired farm machinery. The farms in other clusters have 'moderate' to 'difficult' or easy access to these resources. The structure of farm families in terms of their social status, off-farm employment, tenancy status and managerial control (sole or jointly managed farms under extended family arrangements) are other criteria used in the formation of these clusters. For instance in cluster 4 all farms are tenant operated farms with no high social profile, which are also operated as single family farms with no off-farm employment opportunities. In clear contrast to other clusters, all the farmers in cluster 4 hold some social status among the farming community in the village.

The above differences in the types of farming systems also reflect some significant differences in farmers' attitudes towards planning or choice of production activities

Table 4
Comparison of selected clusters of farms: the Pakistan application

Characteristics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of farms per cluster	14	23	26	5	2
Farming experience (years)	19.6	27.4	31.2	26.2	30.0
Family education index	0.35	0.24	0.16	0.28	0.12
Distance from village (furlongs)	0.67	1.1	3.2	0.90	2.25
Distance from road (km)	0.71	1.26	1.11	0.90	0.50
Operational holding (acres)	28.5	7.5	12.0	21.6	15.2
Total parcels of land (Nos.)	1.9	1.7	1.7	2.4	1.0
Family labour (h/day)	19.4	12.5	19.4	13.6	18.0
Family labour employed off-farm	37.3	27.7	46.2	41.3	0
Loamy area (%)	35.0	26.0	61.0	30.4	22.2
Value of farm equipment (Rs)	5185	985	1271	3800	400
Area under new crops (%)	16.8	16.2	11.7	23.0	27.0
Livestock income (Rs/year)	43	214	143	104	39
Off-farm income (Rs/year)	1086	457	1114	1385	0
Livestock unit per farm	11.3	7.2	10.9	10.4	4.4
Farmers with some social status (%)	0	0	4.0	100	0
Full-time farmers (%)	71.0	87.0	96.0	60.0	100
Single family farms (%)	43.0	87.0	27.0	27.0	100
Owner operator farmers (%)	57.0	87.0	65.0	20.0	0
Tenant farmers (%)	0	0	0	0	100
Irrigation well owners (%)	64.0	35.0	65.0	80.0	50.0

under similar circumstances. The clusters 4 and 5 also indicate that where farmers depend solely on agriculture as a prime source of livelihood, they tend to diversify as shown by the existence of a large number of enterprises present on the farm. This observation could imply that farms with above attributes would be more willing to introduce ‘new’ crops into the existing cropping systems. The farming types identified through cluster analysis show logical arrangement of complex farm situations into representative types of the farming systems in the area. Such an outcome lends further credence to the argument in favour of using numerical clustering procedures for a farm typology from which relevant farm types can be selected for developing representative farm models.

5. Concluding observations

In building models for portraying farm decision-making situations, classifying or typifying farming systems is a fundamental step. The use of multi-variate statistical techniques, such as cluster analysis, for identifying farm types is not new. However, in studies that involve the construction of representative mathematical programming models, there is a tendency to ‘gloss over’ the problem of representative farming systems or farm types. The two studies reported here have dealt with the issue of establishing typical farming systems when empirical information on farm and farmer characteristics exists, and, there is no justification for preferring qualitative or non-quantitative methods over the identification of typing farming systems quantitatively. The approach advocated here can be particularly useful in studies, especially those in less developed countries, where farm typology has to be derived from scratch. The models based on classification schemes established through rigorous analysis of both qualitative and quantitative data should prove to be reliable tools for generating recommendation domains in farming systems research.

Acknowledgements

The authors are grateful to the following organizations for providing financial support for this research: Fundación FIA of the Chilean Ministry of Agriculture and Pakistan Agricultural Research Council, Islamabad, Pakistan.

References

Aldenderfer, M.F., Blashfield, R.K., 1984. Cluster Analysis. The International Professional Publishers, Beverly Hills, USA.

Berdegué, J., Sotomayor, O., Zilleruelo, C., 1990. Metodología de tipificación y clasificación de sistemas de producción campesinos de la Provincia de Ñuble, Chile. In: Escobar, G., Berdegué, J. (Eds.), Tipificación de Sistemas de Producción Agrícola. Red Internacional de Metodologías de Investigación en Sistemas de Producción, Santiago, Chile, pp. 85–117.

Byerlee, D., Collinson, M., Perrin, R., Winkelmann, D., Biggs, S., Moscardi, E., Martinez, J.C., Harrington, L., Benjamin, A., 1980. Planning Technologies Appropriate to Farmers—Concepts and Procedures. Centro Internacional de Mejoramiento de Maiz y Trigo, Mexico.

Comrey, A.L., Lee, H.B., 1992. A First Course in Factor Analysis. Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, USA.

Douglas, C., 1990. Clasificación de sistemas de fincas en el Caribe Oriental. In: Escobar, G., Berdegué, J. (Eds.), Tipificación de Sistemas de Producción Agrícola. Red Internacional de Metodologías de Investigación en Sistemas de Producción, Santiago, Chile, pp. 233–248.

Duarte, O.A., 1990. Tipificación de fincas en la comarca de San Gil, Colombia, con base en una encuesta dinámica. In: Escobar, G., Berdegué, J. (Eds.), Tipificación de Sistemas de Producción Agrícola. Red Internacional de Metodologías de Investigación en Sistemas de Producción, Santiago, Chile, pp. 201–220.

Dunteman, G.H., 1989. Principal Component Analysis (Sage University Paper series on Quantitative Application in the Social Sciences). Sage, Beverly Hills, USA.

Escobar, G., Berdegué, J., 1990. Conceptos y metodología para la tipificación de sistemas de fincas: la experiencia de RIMISP. In: Escobar, G., Berdegué, J. (Eds.), Tipificación de Sistemas de Producción Agrícola. Red Internacional de Metodologías de Investigación en Sistemas de Producción, Santiago, Chile, pp. 13–43.

Escobar, G., Berdegué, J., 1990. Tipificación de Sistemas de Producción Agrícola, 1990. Red Internacional de Metodologías de Investigación en Sitemas de Produción, Santiago, Chile.

Gebauer, R.H., 1987. Socio-economic classification of farm households—conceptual, methodical and empirical considerations. European Review of Agricultural Economics 14, 261–283.

Hardiman, R.T., 1990. Use of cluster analysis for identification and classification of farming systems in Qingyang County, Central North China. Agricultural Systems 33, 115–125.

Khan, M., 1998. Farmers' Objectives and the Choice of New Crops in the Irrigated Farming Systems of Pakistan's Punjab. PhD thesis, Reading, UK, Department of Agriculture, The University of Reading.

Kim, J., 1970. Factor analysis. In: Nie, N.H., Hull, C.H., Jenkins, J.G., Steinberger, K., Bent, D.H. (Eds.), Statistical Package for the Social Sciences. McGraw Hill Book Company, New York, USA, pp. 468–473.

Köbrich, C., 1997. The Construction and Use of Compromise Programming Models to Measure the Impact of Development Policies on the Sustainability of Peasant Farming Systems in Central Chile. PhD thesis, Reading, UK, Department of Agriculture, The University of Reading.

Kostrowicki, J., 1977. Agricultural typology concept and method. Agricultural Systems 2, 33–45.

Landín, R., 1990. Tipificación de cuencas lecheras en Ecuador. In: Escobar, G., Berdegué, J. (Eds.), Tipificación de Sistemas de Producción Agrícola. Red Internacional de Metodologías de Investigación en Sistemas de Producción, Santiago, Chile, pp. 167–179.

Lawley, D.N., Maxwell, A.E., 1971. The scope of factor analysis. In: Factor Analysis as a Statistical Method. Butterworths, London, UK, pp. 1–5.

Martínez, E., Ortíz, A., Reyes, L., 1990. Caracterización de los sistemas de producción minifundistas de la parte alta de la cuanca del río Achiguate, Guatemala. In: Escobar, G., Berdegué, J. (Eds.), Tipificación de Sistemas de Producción Agrícola. Red Internacional de Metodologías de Investigación en Sistemas de Producción, Santiago, Chile, pp. 221–231.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Mojena, R., 1977. Hierarchical grouping methods and stopping rules: an evaluation. The Computer Journal 20, 359–363.

PARC, 1980. Agro-ecological Regions of Pakistan. Islamabad, Social Sciences and Plant Sciences Divisions. Pakistan Agricultural Research Council of Pakistan.

Perevolotsky, A., 1990. Goat production systems in Piura, Peru: a multidisciplinary analysis. Agricultural Systems 32, 55–81.

SAS, 1985. SAS User's Guide: Statistics. Version 5 Edition. Statistical Analysis System Inc, Cary, USA.

Sokal, R.R., 1977. Clustering and classification: background and current directions. In: van Ryzin, J. (Ed.), Classification and Clustering. Proceedings of an Advanced Seminar Conducted by the Mathematics Research Centre, The University of Wisconsin, Madison, 1976. pp. 1–15.

Spedding, C.R.W., 1988. An Introduction to Agricultural Systems. Elsevier, London, UK.

Ward, J., 1963. Hierarchical grouping to optimise an objective function. Journal of American Statistical Association 58, 236–244.