Chapter XVIII
Multivariate methods for index construction

Savitri Abeyasekera
Statistical Services Centre
University of Reading
Reading, United Kingdom of Great Britain and Northern Ireland

Abstract

Surveys, by their very nature, result in data structures that are multivariate. While recognizing the value of simple approaches to survey data analysis, the present chapter illustrates the benefits of a more in-depth analysis, for selected population subgroups through the application of multivariate techniques. Software packages are now available that make possible the application of these more advanced methods by survey researchers.

This chapter demonstrates a range of situations where multivariate methods have a role to play in index construction and in initial stages of data exploration with specific subsets of the survey data, before further analysis is carried out to address specific survey objectives. The focus is mainly on methods that involve the simultaneous study of several key variables. In this context, multivariate methods allow a deeper exploration into possible patterns that exist in the data, enable complex interrelationships among many variables to be represented graphically, and provide ways of reducing the dimensionality of the data for summary and further analysis. The discussion on index construction uses the broader interpretation of multivariate methods to include regression-type methods.

The emphasis throughout is on providing an overview of multivariate methods so that an appreciation of their value towards index construction can be obtained from a very practical point of view. It is aimed both at those engaged in large-scale household surveys and at survey researchers involved in research and development projects who may have little experience in the application of the analysis approaches described here. The use of these methods is illustrated with suitable examples and a discussion of how the results may be interpreted.

Key terms: Index construction, multivariate methods, principal components, cluster analysis.

A. Introduction

1. In analysing survey data, most survey analysts typically use straightforward statistical approaches. Commonest is the use of one-way, two-way or multi-way tables, and the use of graphical displays such as bar charts, line charts, etc. An overview of these approaches and a good discussion on aspects needing attention during the data analysis process can be found in Wilson and Stern (2001) and chapters XV and XVI of the present publication. In some cases, however, analysis procedures that go beyond simple summaries are desirable. One class of such procedures is discussed in the present chapter.

2. Multivariate methods deal with the simultaneous treatment of several variables (Krzanowski and Marriott, 1994a and 1994b; Sharma, 1996). In a strict statistical sense, they concern the collective study of a group of outcome variables, thus taking account of the correlation structure of variables within the group. Many researchers, however, also use the term "multivariate" in the application of multiple regression techniques because this involves several explanatory (predictor) variables along with the main outcome variable (for example, Ruel, 1999). Once again, the benefit of exploring several variables together is that it allows for intercorrelations. Regression approaches, which essentially involve modelling a key response variable, are discussed more fully in chapter XIX. Here we focus mainly on the joint study of several measurement variables as a preliminary step towards our broader interpretation of multivariate methods in the discussion of index construction.

3. Multivariate techniques are often perceived as "advanced" techniques requiring a high level of statistical knowledge. While it is true that the theoretical aspects of many multivariate procedures and their application can be quite daunting even to statisticians, they do have a useful role in analysing data from developing-country surveys. We first discuss the effective use of such methods: (a) as an exploratory tool with which to investigate patterns in the data; (b) to identify natural groupings of the population for further analysis; and (c) to reduce dimensionality in the number of variables involved. We view these as preliminary steps that lead to the construction of indices from household-level variables, for instance, to create indicators of poverty [see, for example, Sahn and Stifel (2000)].

4. Section B provides a general overview of multivariate techniques as the collective study of a group of outcome variables. It is followed by four sections covering areas of application with a number of illustrative examples. Some conclusions on the value and limitations of these techniques are given in the final section. Technical details have been kept to a minimum and greater emphasis is given to understanding the concepts involved and the interpretation. The reader who wishes to acquire a more in-depth understanding of these techniques should consult Everitt and Dunn (2001); and Chatfield and Collins (1980).

B. Some restrictions on the use of multivariate methods

5. Our emphasis in this chapter is on the use of multivariate approaches as valuable descriptive procedures during the initial stages of data exploration and in index construction. In the application of these methods, however, it is important to stress at the outset that an analysis applied to the full data set from a national household survey is unlikely to produce useful findings owing to the inevitable diversity of households in any country. Valuable information can be lost if an analysis combines urban and rural populations, and different agroecological zones, since the livelihoods of households within these different strata can be quite wide-ranging. The techniques described in this chapter should therefore be used only after a careful examination of the data structure to identify the different sectors or substrata of the population to which the methods can be applied, keeping in mind the main survey objectives.

6. Even within such substrata, or in cases where a whole sample analysis is required, it will be important to pay attention to the sample weights associated with the sampled units. If these vary substantially for the data being analysed, then using a software package that does not have facilities for accounting for sample weights may lead to erroneous conclusions. In such cases, weighting the sample units by the sample weights, using for example the *WEIGHT* statement in SAS (2001) or the *aweight* command in STATA (2003) will tackle this difficulty with respect to methods covered in sections C, D, E and F. Many more software packages will take account of sampling weights with respect to methods described in section G. Where sampling weights are not used, some care is needed in interpreting the results, since they may be subject to some bias.

C. An overview of multivariate methods

7. The basic theme underlying the use of multivariate methods in survey investigations is simplification, for example, reducing a large and possibly complex body of data to a few meaningful summary measures or identifying key features and any interesting patterns in the data. The aim is often exploratory: such methods can help in generating hypotheses of interest to the researcher rather than in testing them. Many of the approaches use distribution-free methods that do not assume an underlying statistical distribution for any of the variables. However, as some care is needed concerning the data types being used (for example, interval-scale, counts, binary), we will refer to this issue where relevant in this chapter.

8. The starting point is a data matrix with rows representing cases (the sample units) and columns representing the variables. Sometimes the rows are of greater interest, for example, if they represent farming households, there may be interest in grouping the households into different wealth categories on the basis of a number of socio-economic criteria represented by some columns of the data matrix. In other cases, columns can be of primary interest themselves, for example, when a set of variables corresponding to a particular theme need to be combined into some form of composite index for further analysis.

9. In the sections below, we concentrate on four main approaches to handling multivariate data in developing-country surveys. The first three may be regarded as exploratory techniques leading to index construction. First, we look at graphical procedures and summary measures that will contribute to an understanding of the data. We then look at two popular multivariate

procedures, cluster analysis and principal component analysis (PCA), since these are two of the key procedures that have a useful preliminary role to play in index construction. The latter procedure is discussed more fully in section G along with other ways in which indices can be constructed, taking the broader interpretation of "multivariate" methods as used by many researchers. Throughout, we assume that a suitable subset of the survey data has been selected for analysis and that the aim of subjecting these data to a multivariate procedure is to integrate an exploratory step into an analysis that is attempting to fulfil some broader survey objective.

10. There are of course many other multivariate methods that could be considered in specific situations. Table XVIII.1 shows a range of such methods, together with a brief description of each. This chapter is restricted to just the first three because the aim is to focus on data exploration as a necessary first step for index construction. These three methods are also likely to have the greatest relevance in survey data analysis. Together with the wider application of the term "multivariate" in our discussion on index construction, they form valuable additional methodological tools in survey data analysis. The remaining methods in table XVIII.1 may be useful on specific occasions when relevant to survey objectives. They are, however, beyond the scope of this chapter which proposes to provide only a broad introduction to some of the simpler methods.

Table XVIII.1. Some multivariate techniques and their purpose

	Multivariate technique	Purpose of technique
1.	Descriptive multivariate methods	Data exploration; identifying patterns and relationships
2.	Principal component analysis	Dimension reduction by forming new variables (the principal components) as linear combinations of the variables in the multivariate set
3.	Cluster analysis	Identification of natural groupings among cases or variables
4.	Factor analysis	Modelling the correlation structure among variables in the multivariate response set by relating them to a set of common factors
5.	Multivariate analysis of variance (MANOVA)	Extending the univariate analysis of variance to the simultaneous study of several variates. The aim is to partition the total sum of squares and cross-products matrix among a set of variates according to the experimental design structure
6.	Discriminant analysis	Determining a function that enables two or more groups of individuals to be separated
7	Canonical correlation analysis	Studying the relationship between two groups. It involves forming pairs of linear combinations of the variables in the multivariate set so that each pair in turn produces the highest correlation between individuals in the two groups
8.	Multidimensional scaling	Constructing a "map" showing a spatial relationship between a number of objects, starting from a table of distances between the objects

D. Graphs and summary measures

11. A preliminary understanding of the data is an essential initial stage whenever data analysis is undertaken. A careful look at the data will provide a feel for the meaning and distributional patterns of the data, identify possible outliers (observations not consistent with the pattern of the remaining data), show up data patterns, and provide the user with an idea of whether some variables have greater variability than others [see, for example, Tufte (1983) and Everitt and Dunn (2001)].

12. As in a set of univariate analyses, summary measures such as means and standard deviations for measurement data and frequency tables for binary and categorical data are desirable. Pairs of variables may then be considered in order to identify associations between variables. At this preliminary stage, it would be reasonable to consider data in "bundles", possibly two, one comprising quantitative data (continuous or discrete) and the other comprising qualitative data (categorical and binary). For the former, scatter plots (in pairs) would be meaningful, while for the latter, two-way tables, again in pairs, would be appropriate, possibly combined with some measures of association and the use of a chi-square test statistic. Where relevant, the scatter plots may also be displayed using different symbols to indicate subsets of the data identified by a categorical variable.

13. Most statistics software packages have facilities for matrix plots, for example, the PLOT procedure in SAS (2001), the *Graph/Graphics* menu in SPSS for Windows (SPSS, 2001) and GenStat for Windows (GenStat, 2002). These are graphical displays where scatter plots between all pairs of variables can be shown together, thus providing a quick judgement on how each variable is related to every other variable in the multivariate data set under consideration.

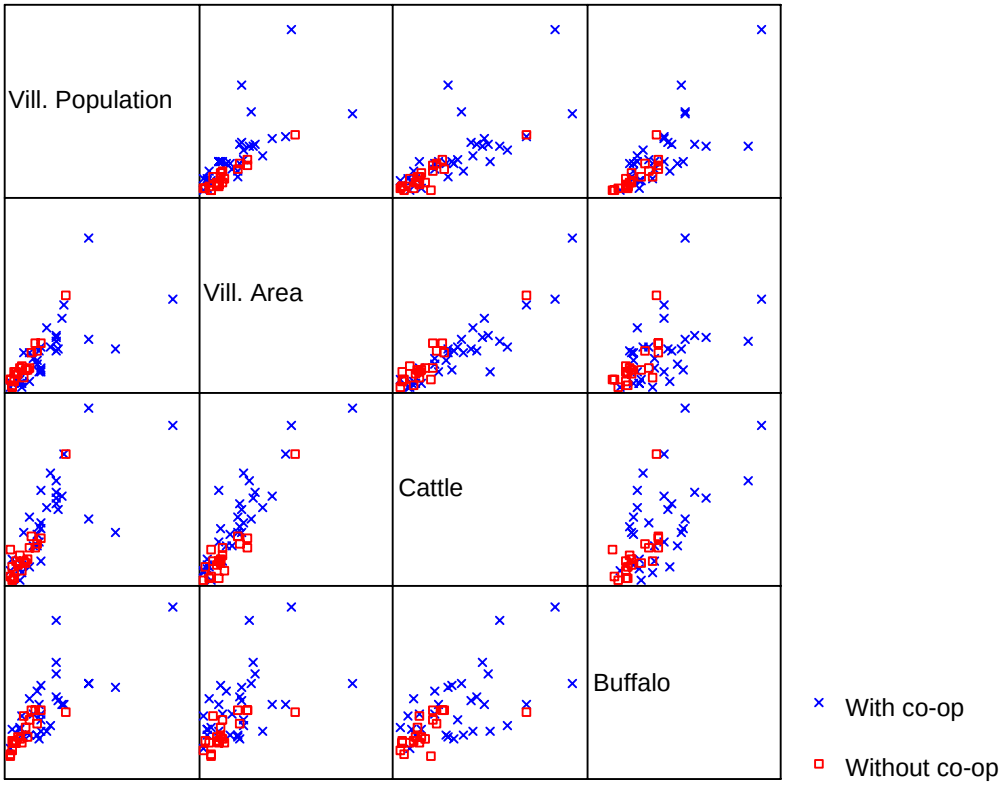
14. As an example, figure XVIII.1 presents a matrix plot, produced from SPSS (2001), that shows the relationships between four variables for 50 villages in Gujarat State in India, according to whether or not they had a dairy cooperative. The variables were: village population, area, and numbers of cattle and buffalo, these being just a few of a larger group of variables. The data come from a baseline study conducted prior to introducing a scheme to promote animal health training. The horizontal and vertical axes for each plot are determined by the axis that runs parallel to the diagonal cells. For example, the three plots in the first row all have village population as their vertical axis and area, cattle and buffalo numbers as their horizontal axes in turn. The same three plots appear in the first column but with their axes reversed. There is possibly one outlier in the data set, clearly seen in the cells in the first row corresponding to a village with a very high population. Some association is observed between all pairs of variables. It is also seen that large values for all variables under consideration are more likely with villages having a dairy cooperative than those without one.

15. If the matrix plot identifies particular pairs of variables that show interesting patterns or outliers, it would be well to repeat these as simple two-way scatter plots, but with attention to the sampling weights associated with each data point. Bubble plots, where each point is represented by a bubble with an area proportional to the sample weight (Korn and Graubard, 1998), are particularly helpful and provide a more meaningful interpretation. For example, an outlier with a large sampling weight will obviously have a greater impact than one with a small sampling

weight. There are a variety of other ways of accounting for the sample design in scatter plots, for example, by subsampling the data with probability proportional to the sample weights and then plotting while ignoring the sample weights, or by applying kernal smoothing methods. The reader is directed to Korn and Graubard (1998) for further details.

16. Many other graphical approaches exist for displaying multivariate data. For example, Manly (1994) shows how several objects, described by several variables, can be drawn in three different ways to show the profile of variable values. Everitt and Dunn (2001) has an excellent chapter on many graphical displays including bivariate boxplots, coplots and trellis graphs, and Jongman, Ter Braak and Van Tongeren (1995) demonstrates the use of biplots. It is not possible to provide further details here but the reader is encouraged to look up the references cited above for further clarification. It is important to note, however, that such graphical procedures are of most value when used with specific subgroups of the population.

Figure XVIII.1. Example of a matrix plot among six variables



E. Cluster analysis

17. Cluster analysis (Everitt, Landau and Leese, 2001) is a data-driven technique, generally aimed at identifying natural groupings among the sampling units (for example, respondents, farms, households) so that units within each group (cluster) are similar to one another while dissimilar units are in different groups. Situations also arise where clustering of variables is relevant, for example, the case where just one or two variables are selected from each cluster so that further analysis could be based on fewer variables. It is thus a useful tool in data exploration and/or data reduction. It can also be used to help in hypothesis generation and in other specific situations.

Example 1

18. As an illustration, consider a study aimed at investigating the effectiveness of a range of low-cost pest management strategies for adoption by resource-poor farming households in a particular region. Suppose that a baseline survey of farmers who may participate in future on-farm trials is conducted with the aim of (a) giving a socio-economic profile of farming households; (b) determining farmers' current pest management practices; and (c) determining farmers' perceptions in respect of pests on the crops they grow. We concentrate here on the first of these three aims and consider how cluster analysis can be used to help determine an effective choice of different groups of farmers for the main study involving on-farm trials.

19. A large number of socio-economic variables were measured during the baseline survey. The aim was to stratify the farming households on the basis of these variables. One approach is to choose, for example, two key variables and form strata defined by combinations of categories associated with the two variables. For example, if the chosen variables were gender of the household head (male/female) and the household's level of food security (low, medium, high), then six strata would result.

20. The disadvantage of this approach is that it ignores other socio-economic characteristics of the households. A multivariate approach allows many variables to be considered simultaneously. Cluster analysis, applied to the farming households on the basis of all relevant socio-economic variables, is a more effective way of stratifying households into a number of clusters so that each cluster represents a distinct socio-economic group of the farming population. This is important inasmuch as recommendations concerning pest management strategies will not necessarily be appropriate for all farming households. An initial classification of farmers into clusters is helpful in providing a basis for choosing different types of farmers to participate in exploring a range of pest management strategies. It also helps in focusing on characteristics specific to the clusters so that interactions between such characteristics and the recommended strategies can be investigated. An illustration is provided in Orr and Jere (1999).

21. To conduct a cluster analysis, two decisions have to be made. First, a measure of similarity (or distance) among the units being clustered must be determined. A similarity measure is one that uses the information from several variables to give a numerical value reflecting the degree of "closeness" between each pair of units. A distance measure is the opposite and reflects how far apart any pair of units is. When all variables are quantitative, or

include at most a few *ordered* categorical variables in addition, the use of a Euclidean[32] distance matrix may be appropriate. Survey data, however, often include binary and non-ordered categorical variables. For such data, various similarity measures have been proposed. For example, if a similarity measure is to be produced between two binary variables, the data may first be cross-tabulated by these two variables to give the 2 × 2 table below.

	0	1
0	*a*	*b*
1	*c*	*d*

22. A possible measure of similarity is then $(a+d)/(a+b+c+d)$, which is called the simple matching coefficient. Another is the Jaccard coefficient $d/(b+c+d)$. A range of other measures can be found in Krzanowski and Marriott (1994b). See Gower (1971) for a suitable similarity measure when mixed data types are involved. In practice, if a large number of variables of different types are to be used in the clustering, it may be better to conduct a number of different cluster analyses, considering variables of the same type each time, and then determining whether the different sets of clusters that emerge are similar. This provides a cross-validation of the cluster membership.

23. Once a distance or similarity measure has been determined, a decision has to be made regarding the method of clustering. Again, many options are presented in statistics software. For example, SPSS (2001) offers seven options (for example, between group linkage, within group linkage, nearest neighbour, etc.). Some of these are agglomerative procedures where, initially, the *n* units being clustered form *n* clusters with one member per cluster, and these are then combined sequentially according to their similarity with members of other clusters. The alternative is a divisive process where all *n* units start as a single cluster, which is then divided in a sequential manner until a satisfying solution is obtained. In either case, some care is needed in making the right decision concerning the way in which the clusters are formed. An extensive discussion of these issues can be found in Everitt, Landau and Leese (2001).

Example 2

24. A special case arises when all variables are binary. The procedure can be fairly simple using hierarchic clustering. For purposes of illustration, we will use just a few observations from a small survey involving 74 farmers in an on-farm research programme. Data for a number of variables recorded during farm visits are shown in table XVIII.2 for just eight farmers. The variables correspond to yes (+) and no (-) answers. One aim was to investigate whether the farms can be grouped into a few clusters on the basis of these characteristics.

[32] Euclidean distance can be thought of simply as reflecting the normal meaning of "distance" as applied to a multidimensional space.

25. Again, for purposes of illustration and to keep the construction details simple, consider the formation of a similarity matrix using the number of +'s that any two variables have in common. The results are shown in table XVIII.3. A set of clusters can then be formed by initially regarding the eight farms as constituting eight clusters, and then merging the closest clusters in turn until finally all farms fall within a single cluster.

26. The similarity matrix for the above example is graphically shown in figure XVIII.2. Such a diagram is called a dendogram. It shows how a specified number of clusters can be selected by cutting the "tree" with a horizontal line at any point. For example, a horizontal line placed near the top of the tree will result in three clusters, these being formed from the sets (1), (7) and (2, 3, 4, 5, 6, 8). In most practical situations, subjective judgements are made in determining the number of clusters to be formed from a hierarchic classification. Formal methods addressing this issue are described in Everitt, Landau and Leese (2001).

27. With suitable software, cluster analysis can be performed quite easily but should be undertaken only after paying close attention to the data types being used, the measure of similarity or distance, and the method used to produce the clusters. Special care is needed if the software being used allows only data of one type to be clustered. For example, SPSS (2001) requires all variables used in the clustering to be either continuous, categorical or binary. If a mixture of data types exists, a better option with such software may be to convert all variables to binary scores and use a similarity measure suited to binary variables, while recognizing, however, that this results in some loss of information.

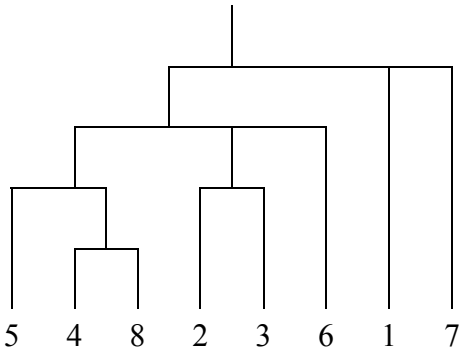
Table XVIII.2 Farm data showing the presence or absence of a range of farm characteristics

	Farm (farmer)							
Characteristics	1	2	3	4	5	6	7	8
Upland (+)/lowland (–)?	–	+	+	+	+	+	–	+
High rainfall?	–	+	+	+	+	–	–	+
High income?	–	+	+	–	–	+	–	–
Large household (>10 members)?	–	+	+	+	–	+	–	+
Access to firewood within 2 km?	+	–	–	+	+	–	+	+
Health facilities within 10 km?	+	–	–	–	–	–	–	–
Female-headed?	+	–	–	–	–	–	–	–
Piped water?	–	–	–	–	–	–	+	–
Latrines present on farm?	+	–	–	–	–	+	–	–
Grows maize?	+	–	–	+	+	–	+	+
Grows pigeon pea?	–	+	+	+	–	+	–	–
Grows beans?	–	–	–	+	+	–	–	+
Grows groundnut?	–	–	–	–	–	–	–	+
Grows sorghum?	+	–	–	–	–	–	–	–
Has livestock?	+	+	–	–	+	–	+	–

Table XVIII.3. Matrix of similarities between eight farms

		Farm							
		1	2	3	4	5	6	7	8
	1	-	1	0	2	3	1	3	2
	2		-	5	4	3	4	1	3
	3			-	4	2	4	0	3
Farm	4				-	5	3	2	6
	5					-	1	3	5
	6						-	0	2
	7							-	2
	8								-

Figure XVIII.2. Dendogram formed by the *between farms* similarity matrix



28. There are two further issues to keep in mind. The first concerns the need to be aware that (as far as the author is aware), the impact of complex sample designs on cluster analysis is unknown. If the survey design involved a cluster sampling procedure, and there were substantial differences between the sampled clusters, a cluster analysis applied to the whole sample data without attention to sampling weights might well generate the survey design clusters themselves. It would therefore be appropriate to consider using a cluster analysis with each of the survey design clusters and study the consistency of the results across these. Again, attention should be paid to differing sampling weights within the survey clusters and results should be interpreted cautiously if the software cannot take weights into account.

29. The second issue concerns the possibility of computational difficulties due to limitations in computing memory. These can arise if cluster analysis is performed using the full survey sample. If consistent with the objectives of performing a cluster analysis, the analysis may be restricted to smaller groups of the surveyed sample to help mitigate this problem.

F. Principal component analysis (PCA)

30. Suppose there are several variables, for instance, 12, which measure facets of one major issue in a survey. For example, in a nutrition survey, the nutrition status of children may be measured in terms of several anthropometric measurements, as well as by variables describing socio-economic characteristics of their families. Such variables are likely to be correlated, and the question then arises whether these variables could be reduced in some fashion to fewer variables that capture as much as possible of the variation in the original data set. Principal component analysis (PCA) aims to do this. The technique is strictly applicable to a set of measurements that are either quantitative or have an ordinal scale. However, as this is largely a descriptive technique, the inclusion of binary variables and/or a small number of nominal categorical variables is unlikely to be of practical consequence.

31. In PCA, a new set of variables is created as linear combinations[33] of the original set. The linear combination that explains the maximum amount of variation is called the first principal component. A second principal component (another linear combination) is then created, independent of the first, that explains, as much as possible, the remaining variability. Further components are then created sequentially, each new component being independent of the previous ones. If the first few components, say, the first 3, explain a substantial amount, say, 90 per cent of the variability among the original set of 12 variables, then essentially, the number of variables to be analysed has been reduced from 12 to 3.

32. It is important to note that the principal component estimators can be severely biased if PCA is applied to the entire survey sample when it is non-self-weighting (Skinner, Holmes and Smith, 1986). As emphasized in section B, PCA is generally recommended in survey data analysis only for smaller subsets of the sample that have (at least approximately) the same sampling weights. If the data subset of interest has substantially differing sampling weights, then some caution should be exercised in interpreting the results.

Example 3

33. Pomeroy and others (1997) applied PCA to data from a survey of 200 households where the respondents were asked to score 10 indicators, on a scale of 1-15, presented to them as rungs of a ladder, to show their perception of the changes that had taken place due to community-based coastal resources management projects in their area. The indicators are listed below, while the PCA results are presented in table XVIII.4.

[33] If $X_1, X_2, \ldots, X_p$ are the original set of p variables, then a variable Y formed from a linear combination of these takes the form $Y = a_1X_1 + a_2X_2 + \ldots + a_pX_p$ where the a_i's (i=1,2,…,p) are numbers, that is to say, the principal component coefficients.

Table XVIII.4. Results of a principal component analysis

Variable	Component		
	PC1	PC2	PC3
1. Overall well-being of household	0.24	0.11	0.90
2. Overall well-being of the fisheries resources	0.39	0.63	0.02
3. Local income	0.34	0.51	0.55
4. Access to fisheries resources	-0.25	0.72	0.17
5. Control of resources	0.57	0.40	0.12
6. Ability to participate in community affairs	0.77	0.13	0.29
7. Ability to influence community affairs	0.75	0.22	0.34
8. Community conflict	0.78	0.03	0.18
9. Community compliance and resource management	0.82	0.12	0.07
10. Amount of traditionally harvested resource in water	0.38	0.66	0.12
Percentage of variance explained	33	19	14

The first principal component is therefore given by:

$$PC1 = 0.24(\text{household}) + 0.39(\text{resource}) \ldots + 0.82(\text{compliance}) + 0.38(\text{harvest}).$$

34. This first component is described by Pomeroy and others (1997) as an indicator dealing with the behaviour of community members, the second component as relating to the fisheries resource, and the third component as an indicator of household well-being. They then use these components as the dependent variables in multiple regression analyses to investigate the effectiveness of a number of explanatory factors in explaining the variability of each indicator.

35. Although the interpretation of the variables is reasonable here, one may question the value of using (say) the first principal component in the form calculated above for further analysis. Only variables 5, 6, 7, 8 and 9 describe the behaviour of the community members and these are the variables that score highly on PC1. Rather than include all 10 variables in the calculation of the first principal component, it would be better to recalculate a new variable as a simple summary of the behaviour variables in the original data set, for example, by taking a simple arithmetical average of variables 5, 6, 7, 8 and 9, or a weighted average of these in which control of resources (variable 5) is given a slightly lower weight relative to the others. Likewise, the resource variables (variables 2, 3, 4 and 10) could be combined to given a simple summary, while variable 1 would stand on its own. Used in this manner, PCA identifies how the 10 indicators may be summarized in a simple way to give a new set of meaningful measures for further analysis, as, for example, Pomeroy and others (1997) have done through regression analysis to explore factors influencing each of their first three principal components.

Example 4

36. The sustainable livelihoods framework adopted by the Department for International Development (DFID) of the Government of the United Kingdom of Great Britain and Northern Ireland provides another practical example. This framework considers five livelihood assets, namely, social capital, human capital, natural capital, physical capital and financial capital. A survey conducted to study household livelihoods would require each of these assets to be measured in terms of a number of subsidiary variables. For example, social capital may be measured in terms of the extent of reliance on networks of support, percentage of household income from remittances, extent of trust in the group, degree of participation in decision-making, etc.; human capital may be measured in terms of the level of education, health status, etc.; and physical capital in terms of ownership of a bicycle or radio, having piped water, electricity, etc.

37. The objective here is to determine a single variable, one for each of the five livelihood assets. This can be done in a straightforward manner for physical assets, for example, by obtaining a simple weighted average of the binary responses corresponding to whether or not items in a given list are owned by a household, using item prices as weights. Social capital, on the other hand, cannot be combined in such a simple way because allocating weights to variables describing social assets is much more difficult. Here we may have to accept data-derived weights via a PCA applied to a set of social variables. The results may be used to produce a suitable overall measure of social capital, again moving towards a simple weighted average after the relative weights of each variable in the first one or two principal components are known.

G. Multivariate methods in index construction

38. Index construction can have several different meanings. In a health study, for example, the nutritional status of children is typically measured by creating indices from anthropometric measurements, for example, weight-for-age, height-for-age and weight-for-height, these representing underweight, stunting and wasting, respectively.

39. In a more complex example, responses to items on breastfeeding, use of baby bottles, dietary diversity, the number of days the child receives selected food groups in past seven days, and feeding frequency, may be summed to create a child feeding index (Ruel and Menon, 2002). This is a second type of index where the researcher decides on the specific scores to be allocated, ensuring that the ordinal scale for each variable is such that high values always represent either "good" or "bad". When binary variables are involved, as, for example, in ownership of a number of assets, the price of the asset could be used to give different weights to each item, as shown in example 4 (sect. F) above.

40. Another type of index can arise in the case where a survey involves determining attitudes or views, say, of the quality of access to health services. Here several questions may be asked, requiring answers on a scoring scale of 1-5 with 1 being "very poor" and 5 being "very good". Again, the resulting scores could be summed across all relevant questions to provide an index reflecting householders' views of the value of health services.

41. Our discussion here goes further to include situations where the data determine the form of the index by use of a multivariate procedure. This still retains the common interpretation of an index as being a single value that captures the information from several variables in one composite measure, typically taking the form:

$$\text{Index} = a_1X_1 + a_2X_2 + a_3X_3 + \ldots\ldots + a_pX_p$$

where the a_i terms are weights to be determined from the data and the X_i terms are an appropriate subset of p variables measured in the survey. We illustrate two ways in which the weights a_i can be determined from the data (see below). Which one is more appropriate will usually depend on the objectives underlying index construction.

42. The first is based on a regression modelling approach; the second, on an application of PCA. These are discussed in relation to indices used for measuring proxy indicators of household wealth or socio-economic status in developing countries. There is a vast literature on this topic and a comprehensive overview can be found in Davis (2002). See also chapter XVII of the present publication which provides a useful discussion on the use of household survey data to understand poverty.

1. Modelling consumption expenditure to construct a proxy for income

43. An approach for modelling consumption expenditure as a proxy for income has been developed by Henstchel and others (2000) and Elbers, Lanjouw and Lanjouw (2001). It involves using data from a detailed household budget survey to identify variables indicative of poverty. This is done by using consumption expenditure as the dependent variable in a multiple linear regression model and a series of household-level variables (for example, assets owned by the household, quality of housing, access to facilities, etc.) as potential explanatory (predictor) variables in the model. The best small subset of the explanatory variables that explains maximum variation in the response (dependent) variable is used to predict consumption expenditure. If the explanatory variables have been collected in a population census, the resulting model equation can then be applied to census data to predict consumption expenditure for each census household. These can then be used to construct poverty maps on a national scale. If the household budget survey is conducted well before the expected date of the census, the appropriate set of predictor variables can be identified from the budget survey data and included in the census questionnaire. We present an example directly below to illustrate this approach.

Example 5

44. The National Bureau of Statistics in the United Republic of Tanzania undertook a National Household Budget Survey (HBS) in 2000-2001 covering approximately 22,000 households. On the basis of details collected on household expenditure over a 28-day period, the total 28-day consumption expenditure per adult equivalent was calculated for each household. Regression modelling with preliminary data available from the HBS identified a series of potential household-level variables (separate sets for urban and rural areas) that explained a high proportion of the variability in consumption expenditure. These variables were included in a

questionnaire administered to a census of households at three sentinel surveillance sites under study by the Adult Morbidity and Mortality Project (AMMP) team based in Dar es Salaam. The aim was to develop an index reflecting consumption expenditure using HBS data for each AMMP site, and to apply the index to households covered by the AMMP at each site.

45. Full details of the modelling approaches and an evaluation of the effectiveness of the models can be found in Abeyasekera and Ward (2002). Here we present a summary of the results for one rural region (see table XVIII.5) to show the variables that entered the model equation and the weights (regression coefficients) used in computing an index of consumption expenditure.

46. From the results of table XVIII.5, the index predicting consumption expenditure for households in Kilimanjaro region in the United Republic of Tanzania is the following:

Index of consumption expenditure =

9.79388+(0.11043*lamp)+(0.19950*sofa)+(0.12870*bicycle)+(0.11858*seed)

+(0.16254*fertiliser)+(0.025824*landarea)+(0.088769*meat)+(0.076132*income4)

+(0.13451*income3)+(0.098303*income2)+(0.27985*edu4)+(0.15878*edu3)

-(0.0091977*edu2) - (0.0022552*age)+(0.010456*hhsize2)-(0.23902*hhsize)

47. The model explained 65 per cent of the variability in consumption expenditure. This is a significantly high figure given the complexity of livelihoods among rural households. The quality of this index at its development stage was judged by (a) comparing it with the true values of consumption expenditure; and (b) considering its ability to identify the true proportion of households below the basic needs poverty line of the United Republic of Tanzania. Method (a), utilized by graphing the index versus true values, showed a very good correspondence. It performed less well when the population of true values and the population of predicted values were categorized into five wealth quintiles, and tabulated against each other. Only 46 per cent of households were classified into the correct quintile. The classification by poverty line was better, with 87 per cent classified correctly as being above or below the poverty line.

48. Further examples of the modelling approach are presented in the final sections of chapter XIX.

Table XVIII.5. Variables used and their corresponding weights in the construction of a predictive index of consumption expenditure for the Kilimanjaro region in the United Republic of Tanzania

Predictor variable	Significance probability	Weight (model coefficient) (STATA estimate)
Household size	0.000	–0.239
Square of household size	0.000	0.0104
Age of household head (years)	0.038	–0.00226
Education of household head a/	0.000	0, -0.00920, 0.159, 0.280
Main source of income b/	0.017	0, .0983, 0.1345, 0.0761
Days meat eaten in past week	0.000	0.0888
Area of land owned by household	0.000	0.0258
Fertilizer c/	0.000	0.1625
Seeds c/	0.004	0.1186
Ownership of bicycle	0.000	0.1287
Ownership of sofa	0.000	0.1995
Ownership of lamp	0.001	0.1104
Constant in model equation	0.000	9.794

Sample size = 1,026 $R^2 = 0.651$ Adjusted $R^2 = 0.646$

a/ None; primary; secondary; tertiary and above.

b/ Sale of crops; sale of livestock; business/wages/salaries; other sources.

c/ If bought in past 12 months.

2. Principal components analysis (PCA) used to construct a "wealth" index

49. The methodology discussed in section G.1 above can be applied only if reliable data on consumption expenditure – the dependent variable - are available from a previous survey. The difficulty of collecting reliable information on consumption expenditure, combined with the high costs of data collection, has prompted some researchers to recommend the use of an asset-based poverty index, derived from conducting a PCA. The first principal component is used as an index of socio-economic status following previous research that has suggested that the asset-consumption relationship is a quite close one (Filmer and Pritchett, 1998). However, some caution must be exercised in interpreting the asset index as a poverty measure, since its effectiveness will depend on the choice of assets used and the particular set of data to which the PCA is applied. As an example of this approach, Gwatkin and others (2000) illustrate the PCA methodology for determining wealth quintiles in the United Republic of Tanzania, using the following set of mixed asset based variables and health-related:

- Whether the household has electricity, a radio, television, refrigerator, bicycle, motorcycle, car (each coded as 1 = yes, 0 = no)
- Number of persons per sleeping room (a quantitative response)
- Principal household sources of drinking water (seven categories)
- Principal type of toilet facility used by members of the household (five categories)
- Principal type of flooring material in the household (six categories)

50. The data they used come from information gathered through the Demographic and Health Survey (DHS) questionnaire. Appropriate sampling weights were used in the analysis.

51. The authors emphasized that theirs was an initial effort applied to a whole country sample, but that future attempts to examine population differences by socio-economic class would produce different results. They suggested that this might happen as a result of the use of some basis other than assets for defining socio-economic status, or as a result of sampling errors, etc. A more obvious reason would be wealth differentials across sites. Indeed, there was evidence of differences in wealth quintile cut-offs when their methodology was applied to three subpopulations in the United Republic of Tanzania, namely, the three regions referred to in section G.2, using data from the national Household Budget Survey (table XVIII.6). It is therefore advisable not to regard PCA results as being portable even within a single country over time or when applied to different strata of the population.

52. Researchers have also used the first principal component of a principal component analysis as a summary index for further analysis of the data. Ruel and Menon (2002), for example, constructed a socio-economic index from DHS data sets in order to categorize households into terciles for the purpose of controlling for socio-economic status in a multiple regression analysis carried out to determine factors affecting child nutritional status. They undertook separate analyses for urban and rural populations using seven data sets from five countries in Latin America. The variables used were water source, sanitation, housing materials (floor, wall, roof) and ownership of a list of assets. The values of these variables were ranked in ascending order (from worst to best) before subjecting them to a principal component analysis. Only variables with principal component coefficients greater than 0.5 were retained in the final index. The approach here was reasonable, the primary objective having been the construction of an index to correct for socio-economic differentials in a subsequent analysis.

Table XVIII.6. Cut-off points for separating population into five wealth quintiles

Wealth quintile	Dar es Salaam a/ (HBS)	Kilimanjaro a/ (HBS)	Morogoro a/ (HBS)	All United Republic of Tanzania a/ (HBS)	All United Republic of Tanzania b/ (DHS)
20th percentile	-1.2993	-0.8452	-0.9190	-1.0317	-0.5854
40th percentile	-0.7709	-0.6289	-0.6180	-0.5704	-0.5043
60th percentile	-0.1054	-0.2459	-0.3645	-0.3051	-0.3329
80th percentile	1.1603	0.3239	0.4586	0.4609	0.3761

a/ Household Budget Survey 2000-2001.
b/ Demographic and Health Survey, 1996.

H. Conclusions

53. Our aim in this chapter has been to demonstrate the use of multivariate methods in index construction, with an emphasis on the need for multivariate exploratory tools as a first stage in the analysis. The application of these methods, however, requires careful thought, with due attention to their meaning and their limitations. The success of PCA for variable reduction, for example, depends on being able to summarize a substantial proportion of the variation in the data by means of just a few component indices, and being able to give a meaningful interpretation to each of these. One is also well advised to think carefully about the effectiveness of the PCA procedure if only a small part of the variation in the complete set of variables is accounted for by the first principal component. Sufficient attention should also be given to the appropriateness of the variables included in the calculation of the index in relation to the objectives of the analysis.

54. Cluster analysis suffers from difficulties associated with identifying a suitable similarity or distance measure and with decisions concerning the method of clustering to be used. A variety of factors must be considered here, including the types of data being used, computational aspects and the robustness of the procedure to small changes in the data.

55. It is also necessary to stress once more that methods described in this chapter are best applied to appropriate subsets of the population when there is a clear structure into which the population may be divided. This is particularly true if the data for analysis come from a national survey. Decisions regarding the choice of subsets to be used must then be made, with appropriate justification. One consequence is that different indices may be produced for different subsets. This in itself, however, will be a useful finding, suggesting that further analysis would be more meaningful within the population subsets under consideration.

56. This chapter has offered an assessment of the value of multivariate techniques, as an exploratory tool and, more specifically, for their use in index construction. Facilities are now available in general-purpose statistical software [for example, SPSS (2001), STATA (2003)] to enable such analyses to be performed relatively easily. Researchers are therefore encouraged to consider their use during survey data analysis with a view to extracting as much information as possible from the data and contributing usefully to the survey objectives.

Acknowledgements

I wish to express my sincere thanks to my colleague Ian Wilson and to two anonymous referees for their valuable comments on initial drafts of this chapter. The National Bureau of Statistics of the United Republic of Tanzania is also thanked for allowing access to their data for some of the examples used in this paper, and I am grateful to the Department for International Development (DFID) of the Government of the United Kingdom of Great Britain and Northern Ireland for providing ideas for this chapter through its funding of many interesting projects involving surveys in the developing world. The material in this chapter, however, remains the sole responsibility of the author and does not imply the expression of any opinion whatsoever on the part of DFID.

References

Abeyasekera, S., and P. Ward (2002). *Models for Predicting Expenditure per Adult Equivalent for AMMP Sentinel Surveillance Sites*. Dar es Salaam: Adult Morbidity and Mortality, Ministry of Health of the United Republic of Tanzania. Available from www.ncl.ac.uk/ammp/tools_methods/socio.html.

Chatfield, C., and A.J. Collins (1980). *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Davis, B. (2002). Is it possible to avoid a lemon? Reflections on choosing a poverty mapping method. Available from http://www.povertymap.net/pub/Pov_mapping_methods_18-9-02.pdf.

Elbers, C., J. Lanjouw and P. Lanjouw (2001). Welfare in villages and towns: micro-level estimation of poverty and inequality. Mimeo.Vrije Universiteit, Yale University and World Bank.

Everitt, B.S., and G. Dunn (2001). *Applied Multivariate Data Analysis*. London: Arnold.

Everitt, B.S., S., Landau and M. Leese (2001). *Cluster Analysis*. London: Arnold.

Filmer, D., and L. Pritchett (1998). *Estimating Wealth Effects without Expenditure Data–or Tears: An Application to Educational Enrolments in States of India*. Washington, D.C.: World Bank Policy Research Working Paper, No. 1994.

GenStat (2002). *GenStat for Windows,* 6th Ed. Oxford, United Kingdom: VSN International, Ltd.

Gower , J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics,* vol. 27, pp. 857-872.

Gwatkin, D.R., and others (2000). *Socio-economic Differences in Health, Nutrition and Population in Tanzania*. Washington, D.C.: Thematic Group on Health, Population, Nutrition and Poverty of the World Bank. Available from http://www.worldbank.org/poverty/health/data/tranzania/tanzania.pdf (accessed 30 June 2004).

Hentschel, J., and others (2000). Combining census and survey data to trace spatial dimensions of poverty: a case study of Ecuador. *The World Bank Economic Review*, vol. 14, No. 1, pp. 147-165.

Jongman, R.H.G., C.J.F. Ter Braak and O.F.R. Van Tongeren (1995). *Data Analysis in Community and Landscape Ecology*. Cambridge, United Kingdom: Cambridge University Press.

Korn, E.L., and B.I. Graubard (1998). Scatterplots with survey data. *The American Statistician*, vol. 52, No. 1.

Krzanowski, W.J., and F.H.C. Marriott (1994a). *Multivariate Analysis,* Part 1. *Distributions, Ordination and Inference*. London: Arnold.

_____ (1994b). *Multivariate Analysis,* Part 2. *Classification, covariance structures and repeated measurements*. London: Arnold.

Manly, B.F.J. (1994). *Multivariate Statistical Methods*: *A Primer*. 2nd ed. London: Chapman and Hall.

Orr, A., and P. Jere (1999). Identifying smallholder target groups for IPM in southern Malawi. *International Journal of Pest Management*, vol. 45, No. 3, pp. 179-187.

Pomeroy, R.S., and others (1997). Evaluating factors contributing to the success of community-based coastal resource management: the Central Visayas Region Project-1, Philippines. *Ocean and Coastal Management,* vol. 36, Nos. 1-3, p. 24.

Ruel, M.T., and others (1999). *Good Care Practices Can Mitigate the Negative Effects of Poverty and Low Maternal Schooling on Children's Nutritional Status: Evidence from Accra*. Food Consumption and Nutrition Division Discussion Paper, No. 62, Washington, D.C.: International Food Policy Research Institute.

Ruel, M.T., and P. Menon (2002). *Creating a Child Feeding Index Using the Demographic and Health Surveys: an Example from Latin America*. Food Consumption and Nutrition Division Discussion Paper, No. 130, Washington, D.C.: International Food Policy Research Institute.

Sahn, D.E., and D. Stifel (2000). Assets as a measure of household welfare in developing countries. Working Paper 00-11. St. Louis, Missouri: Washington University, Center for Social Development.

SAS (2001). *SAS Release 8.2.* Cary, North Carolina: SAS Institute, Inc., SAS Publishing.

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: Wiley and Sons, Inc.

Skinner, C.J., D.J. Holmes and T.M.F. Smith (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, vol. 81, Issue 395, pp. 789-798.

SPSS (2001). *SPSS for Windows. Release 11.0.* Chicago, Illinois: LEAD Technologies, Inc.

STATA (2003). *Intercooled Stata 8.0 for Windows*. College Station, Texas: Stata Corporation.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Chesire, Connecticut: Graphics Press.

Wilson, I.M., and R.D. Stern (2001). *Approaches to the Analysis of Survey Data*. Statistical Guideline Series supporting DFID Natural Resources Projects. Reading, United Kingdom: Statistical Services Centre, University of Reading. Available from http://www.reading.ac.uk/ssc (accessed 25 June 2004).