PEOPLE AND PLANTS WORKING PAPER - JUNE 1999

6

This series of working papers is intended to provide information and to generate fruitful discussion on key issues in the sustainable and equitable use of plant resources. Please send comments on this paper and suggestions for future issues to

People and Plants

People and Plants Initiative,

Division of Ecological Sciences,

UNESCO, 7 Place de Fontenoy,

Quantitative Ethnobotany

Applications of multivariate and statistical analyses in ethnobotany

M. Höft, S.K. Barik and A.M. Lykke

Authors' addresses:

M. Höft
c/o UNESCO Office Nairobi
P.O. Box 30592
Nairobi
KENYA

S. K. Barik
Centre for Environmental Studies
North-Eastern Hill University
Shillong 793 014
INDIA

A. M. Lykke
Dept. of Systematic Botany
Nordlandsvej 68
8240 Risskov
DENMARK

Photos: all photos by R. Höft except Photo 1 by Y. Morimoto

Cover illustration: M. Höft; Callus formation and bark regeneration of *Rytigynia kiwuensis* (top); fruiting *Rytigynia kigeziensis* (bottom); graph showing relationship between bark weight and diameter of three species of *Rytigynia* (data from Kamatenesi 1997).

Published in 1999 by the
United Nations Educational, Scientific and Cultural Organization
7, place de Fontenoy, 75352 Paris Cedex 07 SP, FRANCE
Printed by UNESCO on chlorine-free recycled paper

Edited by Robert Höft
Design: Ivette Fabbri
Layout: Martina Höft and Robert Höft

SC-99/WS/41

Recommended citation: Höft, M., Barik, S.K. & Lykke, A.M. 1999. *Quantitative ethnobotany. Applications of multivariate and statistical analyses in ethnobotany.* People and Plants working paper 6. UNESCO, Paris.

Quantitative ethnobotany

APPLICATIONS OF MULTIVARIATE AND STATISTICAL ANALYSES IN ETHNOBOTANY

Abstract

Photo 1. In most cases ethnobotanical data collection requires simple tools such as measuring tape or spring balance. This photo shows the Loita Ethnobotany Team quantifying amounts of '*olorien*', *Olea europaea* L. ssp. *africana* (Mill.) P. Green (Oleaceae), used for fuel in Maasai households.

Some wild plant resources are severely threatened by habitat loss and species-selective overexploitation. In addition, indigenous knowledge about the uses of wild plant resources is rapidly disappearing from traditional communities. In the context of conservation and sustainable and equitable use of wild plant resources, quantitative ethnobotany can contribute to the scientific base for management decisions.

In the past, most ethnobotanical studies have recorded vernacular names and uses of plant species with little emphasis on quantitative studies. In this working paper, a selection of multivariate and statistical methods particularly applicable to the analysis of ethnobotanical field data is presented. The working paper aims at assisting researchers and students to recognize the appropriate method to analyse their data and to develop management recommendations from scientifically sound conclusions.

The techniques presented include cluster and principal component analysis, regression analysis, analysis of variance, and log-linear modelling.

Multivariate and statistical analysis requires computerized statistics and graphics programs. Basic technical knowledge to use such tools as well as basic understanding of statistical terms are important requirements to get most benefit from this publication.

Contents

Introduction

In order to enhance the indicative value of ethnobotanical studies, there have been attempts in recent years to improve the traditional compilation-style approach through incorporating suitable quantitative methods of research in ethnobotanical data collection, processing and interpretation. Such quantitative approaches aim to describe the variables quantitatively and analyse the observed patterns in the study, besides testing hypotheses statistically. The concept of quantitative ethnobotany is relatively new and the term itself was coined only in 1987 by Prance and coworkers (Prance, 1991). Quantitative ethnobotany may be defined as "the application of quantitative techniques to the direct analysis of contemporary plant use data" (Phillips & Gentry 1993a and b). Quantification and associated hypothesis-testing help to generate quality information, which in turn contributes substantially to resource conservation and development. Further, the application of quantitative techniques to data analysis necessitates refinement of methodologies for data collection. Close attention to methodological issues not only improves the discipline of ethnobotany but also enhances the image of ethnobotany among other scientists (Phillips & Gentry 1993a and b).

Different approaches are taken to collect and analyse quantitative and qualitative ethnobotanical data. The approaches depend on the objectives of the researcher and the nature of study and aim at the objective evaluation of the reliability of the conclusions based on the data. Multivariate and statistical methods are typically applied to the interpretation of the following types of ethnobotanical data (the list is not exhaustive):

- relative importance of plant taxa and vegetation types to different ethnic, social or gender groups;
- knowledge and uses of plants by different ethnic, social or gender groups;
- preference information on different plant species;
- size class distribution of woody plant species;
- quantitative impact of human uses on growth and regeneration patterns;
- quantitative impact of environmental factors on certain plant traits;
- quantitative impact of agricultural or horticultural techniques on certain plant traits;
- quantitative plant morphological and pharmacological characteristics of useful plants.

The data processing techniques in ethnobotany may range from calculating a simple index to complex computational techniques of multivariate analysis such as classification and ordination. The selection of a particular technique for application to the data is based on the effectiveness of the technique for sound interpretation of the results and identification of the inter-relationships that may exist among the variables studied. In general, statistical applications may be classified into two broad categories:

1. Sets of data where the measurements are taken only on one attribute or response variable and the data so obtained are analysed through a set of techniques called univariate analysis techniques.
2. Sets of data where the measurements are taken simultaneously on more than one variables and the statistical techniques applied to such data sets are called multivariate analysis techniques.

Studies of multivariate nature are more common in ethnobotanical research, and are treated in more detail in this paper.

Dimensions of data

Because of the complexities involved in most ethnobotanical studies, it is common for ethnobotanical researchers to collect observations on many different variables. The need to understand the relationships between many variables makes multivariate analysis mathematically complex and the techniques to analyse such data invariably need a computer. Today a large number of computer packages are available for analysis of multivariate data sets. BMDP (BMDP Statistical Software Inc.), CANOCO (Ter Braak, 1988a and 1988b), NTSYS (Rohlf, 1985), PC-ORD (MjM Software Design), R-Package (Casgrain 1999), SAS (SAS Institute Inc.), SYSTAT (SPSS Inc.), SPSS (SPSS Inc.) and TWINSPAN (Hill, 1979) are some of the popular and powerful software packages widely used for a variety of multivariate and statistical data analyses. Besides their analytical features most of them include graphical functions.

Generally, multivariate and statistical methods aim at making large data sets mentally accessible, structures recognizable and patterns explicable, if not predictable. Johnson and Wichern give five basic applications for these methods (Johnson & Wichern 1988):

1. **Data reduction or structural simplification**: The phenomenon being studied is represented as simply as possible with reduced number of dimensions but without sacrificing valuable information. This makes interpretation easier.
2. **Sorting and grouping:** Groups of similar objects or variables are created.
3. **Examining relationships among variables:** Variables are investigated for mutual interdependency. If interdependencies are found the pattern of dependency is determined.
4. **Prediction:** Relationships between variables are determined for predicting the values of one or more variables on the basis of observations on the other variables.
5. **Testing of hypothesis:** Specific statistical hypotheses formulated in terms of the parameters of multivariate populations are tested. This may be done to validate or reject assumptions.

The different multivariate and statistical analysis techniques, which are available for the above applications are derived from one simple linear mathematical model, the Multivariate General Linear Hypothesis (MGLH). In this paper the following linear models are presented along with their applications:

1. classification and clustering;
2. ordination;
3. analysis of variance:
4. regression analysis;
5. correlation;
6. log-linear modelling;

These techniques will be demonstrated using examples from a 'People and Plants' workshop on species used for woodcarving in Kenya, a Ph.D. study on alkaloid patterns of *Tabernaemontana pachysiphon,* two Ugandan M.Sc. studies, one on *Rytigynia kiwuensis* and one on medicinal plant collection habits of different specialist groups. Before getting to the practical applications, some general remarks regarding types of data, sampling size, sorting and grouping of data are presented.

First of all, the different types of quantitative and qualitative data must be distinguished (see Box 1, page 5). In the majority of cases ethnobotanical data are quantitative on an ordinal scale. Frequency and abundance are key parameters in vegetation analysis and population dynamics, ranking order reveals important information on preferences of user groups and ordered multistate character are data that fall into predefined hierarchical groups. Quantitative data on a ratio or interval scale may be collected to determine growth patterns of individual plant species, to assess the effectiveness of a certain remedy, or to express the impact of human uses.

Qualitative data like 'presence/absence' or 'yes/no' are often recorded during interviews when people's knowledge of certain species or management techniques is assessed or the potential for the acceptance of substitutes for a particular resource is gauged.

Counts are obtained when numbers of people falling in a certain category, or numbers of events taking place in a pre-defined category or time span are recorded. In order to assist in determining relationships among and between variables and how they can be classified and appropriate analysis techniques identified, Box 2 (page 5) lists some common data settings and research questions to which corresponding parametric and non-parametric methods exist. Not all of these techniques are discussed in this paper.

Parametric methods apply to approximately normally distributed data. In a simple linear model

$$Y = a + bX + e$$

Y is the dependant and X the independent variable. Variables are defined as quantities that can vary in the same equation. In contrast, the **parameters** a and b are quantities that are constant in a particular equation, but can be varied in order to produce other equations in the same general family. The parameter a is the value of Y when $X = 0$. This is sometimes called a Y-intercept (where a line intersects the Y-axis in a graph when $X = 0$). The parameter b is the slope of the line, or the number of units Y changes, when X changes by one unit; "e" is referred to as an "error" or residual, which is a departure of an actual Y from what the equation predicts. The sum of all e is zero.

Sampling and organization of data

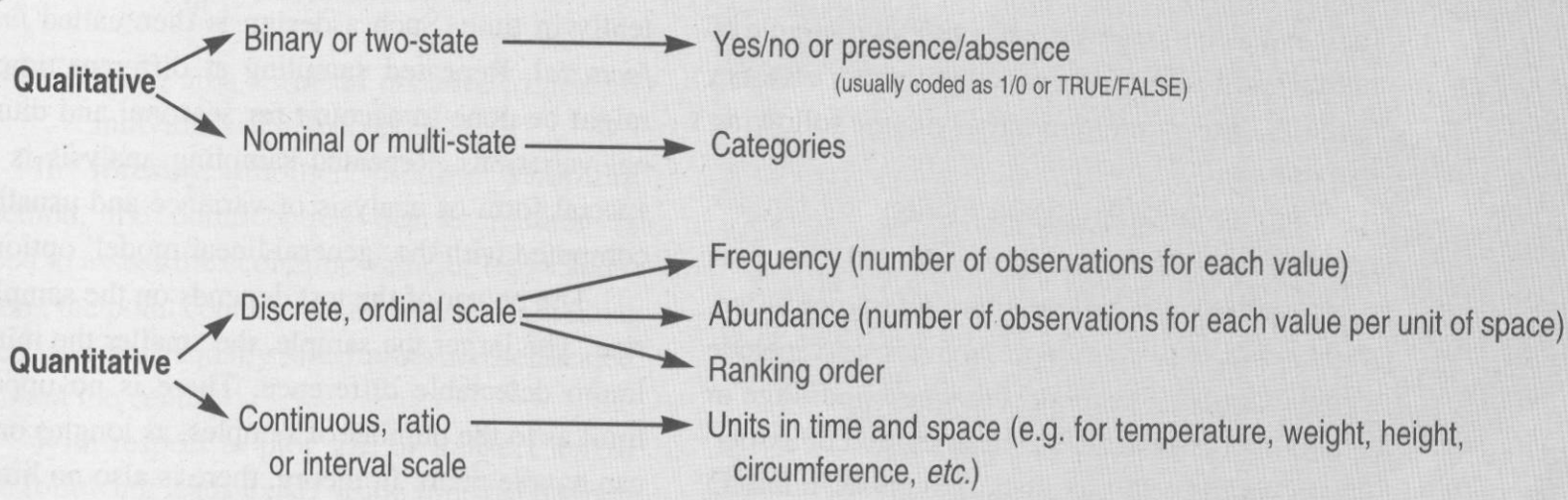
Having developed a well thought out research design before going to the field is likely to:

1. save a lot of time (and money) when analysing data;
2. enhance the expected output in terms of meaningful results;
3. allow more easily for the results being translated into scientifically sound recommendations;
4. leave you and others satisfied with the work.

The following reflections are crucial when planning ethnobotanical research in the field:

- How many samples need to be taken in different categories?

Box 1. Qualitative and quantitative data.



Box 2. Applications of multivariate and statistical analyses techniques based on linear models.

Research interest	Method of analysis
General relationships among variables	⇒ Correlation analysis/analysis of co-variance
Associations between variables	⇒ Detrended correspondence analysis
Quantitative relationships among variables and prediction	⇒ Regression analysis
Similarity/dissimilarity among variables or groups of variables	⇒ Cluster analysis
Variance among variables or subjects in counted observations	⇒ Principal component analysis
Testing of hypothesis regarding factorial effects on variables	⇒ Analysis of variance/Kruskal Wallis test
Relationships among categories in multi-way frequency tables and prediction of cell frequencies based on counts	⇒ Log-linear modelling
Exploring survival rates	⇒ Survival analysis

- How many categories can realistically be studied without cutting down the minimum number of samples to be taken in each category?
- Can equal sampling be assured for each category?
- Which are the categories that would be representative for the research question?
- Have seasonal, diurnal, or circumstantial fluctuations to be accounted for?
- Is repeated sampling necessary? (i.e. same samples studied at different times)
- Are the samples representative for the population?
- If processes are to be documented: can changes realistically be observed within the time-frame of the study? (i.e. growth records of plants)
- What indicators can be used to document processes?

Due to high cost in terms of time and money true random sampling is not practicable in many applied research situations. A stratified or systematic random sampling strategy is, therefore, usually applied to study plant use by people. Charles Peters (1996) provides a good discussion on the different approaches.

Photo 2. Mzee Ali Mwadzpea and Alex Jeremani constructing litter traps. Sixteen traps were randomly set up in a coastal forest in Kenya to study the nutrient dynamics of soil and vegetation.

True random sampling of a finite population would mean the assignment a number to each case and then the random selection of a sample of numbers. In SYSTAT, random numbers between 1 and 73,500 can be generated with the following expression:

1 + INT (73500 * URN)

When people are interviewed and questions asked with respect to some particular knowledge, the sample should be representative and include people from different social backgrounds, age or gender. In order to allow interpretation of associations that may arise from data analysis, it is necessary to record as much information as possible from the interviewees.

If the aim for ethnobotanical applications is to predict one quantitative plant trait from another (usually ready to measure) quantitative plant trait, the sample size must be sufficiently large to include individual variation according to environmental factors at the study site (e.g. altitude, exposition, soil nutrients) and endogenous factors within the species itself (e.g. age, phenological status). While for statistical confidence and accuracy a sampling size of 10% of the total population is desirable, for practical reasons the actual sampling size may not even reach 1% of the total population. An absolute minimum of four individuals within each cell (category) is indispensable for any statistical analysis. Obviously, the predictive value of inference increases with increasing sample size.

The state of environmental factors may be described either quantitatively (e.g. concentration of nitrogen and phosphorus in the soil, average daily light sums, amount of water in the soil) or simply categorically (e.g. 'fertile' soil, 'high' light intensity, 'dry' site). These attributes are subjective depending on the perceptions of the researcher. For quantitative observations of the independent variable on a continuous or ratio scale, regression analysis would apply. However, if the independent variable is categorical, then analysis of variance is applied. In both cases, relationships among variables and significant interactions between the environmental factors may be expected and have to be accounted for in analysis. Sampling schemes in the field should be planned in such manner as to allow separate accounting for environmental effects. In addition to sampling in the field, experimental designs are used to separately test effects of factors and proper planning is imperative in this respect. Linear relationships or 'co-variances' among variables must be tested before applying any of the standard procedures.

Ideally, the sample size in each category has to be equal and samples have to be taken consistently in time. Such a design is then called *full factorial*. Repeated sampling at different times might be done to account for seasonal and diurnal variations. Repeated sampling analysis is a special form of analysis of variance and usually computed with the 'general linear model' option.

The power of the test depends on the sample size. The larger the sample, the smaller the minimum detectable difference. There is no upper limit as to the number of samples, as long as one can handle them. In theory, there is also no limit to the number of factors that might be analysed simultaneously. However, the number of possible interactions becomes unwieldy and interpretation of interactions of more than three or four variables extremely difficult.

Another group of multivariate analysis involves the application of log-linear models, also referred to as discrete analysis of variance. In this case data are counts and are arranged in two- or multi-dimensional contingency tables. Again, the number of observations should ideally be equal in each of the categories.

When preparing the data set it is crucial to clear ambiguous signs (e.g. numbers with question marks) or in-between categories for the final record. Usually a lot of time is wasted in cleaning of data sets from such ambiguous entries. Often the whole entry is lost when the meaning of symbols one used to mark a certain entry at a certain time can not be recalled. It is better to invest more time in the field measuring or inquiring to obtain clear data entries from the beginning.

In vegetation research extensive relevés are often produced and the crucial problem in the beginning is to decide on the right sampling method. Species distribution can be recorded using transects, whereas species abundance is recorded in plots. In stratified plot sampling, plots are arranged along imaginary lines following environmental gradients. Stratified plot sampling combines two approaches to vegetation analysis and is mainly used for investigating population dynamics. In ecosystems where woody vegetation is sparsely distributed, plotless sampling is most appropriate when wishing to derive estimates of woody species density. The simplest plotless sampling method is the nearest individual method. Random sampling points are determined in the area and the distance to the nearest individual(s) of each tree species is recorded. Successive distance measurements are taken and the procedure is repeated for a number of random points. The density of each species is then derived from the following formula:

$$D_{Sp} = \sqrt{\text{mean area}/2}$$

where: mean area = (mean distance to nearest individual of a species)^2

In forestry, another plotless sampling method, the point-centred quarter method is used to assess the economic value of tree stands. Here, the point centre is marked by an individual tree, and four equally sized plots are delineated around the centre.

With respect to plot size or transect length the following leads exist: inside forest and when dealing with large trees: subplot size should be 20 x 10 m or 20 x 20 m. For the analysis of regeneration patterns, subplot sizes of 10 x 10 m are sufficient to cover total areas between e.g. 0.1 and 1 ha. In grassland and when analysing herbaceous vegetation (including tree seedlings), subplot sizes of 1 x 1 or 5 x 5 m are usually chosen. Transects may have lengths between 100 and 1000 m and are usually between 1 and 5 m wide along each side.

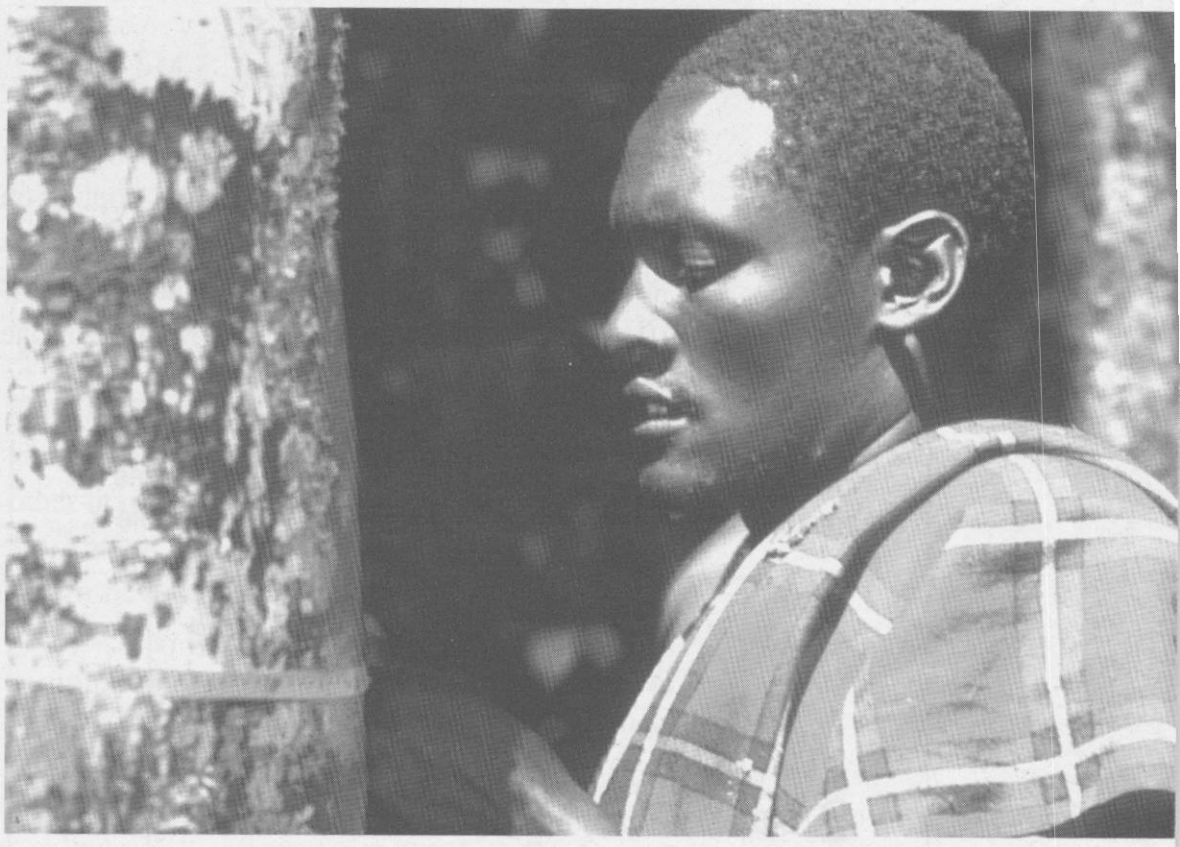


Photo 3. Moses Kipelian of the Loita Ethnobotany Team measuring the DBH of '*oltarakwai*', *Juniperus procera* Endl. (Cupressaceae) trees, which are highly valued for the construction of stockades and fences. The resulting size class distribution curve showed a lack of regeneration which has raised serious concern and has led to the establishment of a tree nursery.

Photo 4. To study the abundance of three species of much sought after medicinal plants, '*nyakibazi*' (*Rytigynia kigeziensis* Verdc., *R. kiwuensis* (K. Krause) Robyns, and *R. bagshawei* (S. Moore) Robyns, Rubiaceae), in Bwindi Impenetrable National Park, Uganda, Maud Kamatenesi set up more than 300 plots of 20 x 20 m, counted the individuals, measured DBH and height and determined the amount of bark used.

Data are entered into the computer with the aid of a spreadsheet (rows and columns), that can later be imported into any statistical package, or simply as ASCII (American Standard Code for Information Interchange) file, where entries are separated by blanks or tabulators. Some statistical programs put limits to the maximum file width (i.e. number of columns) that can be imported without specification of the file width.

Data are entered either as numerical values (SI units) or character values. Character variables are also referred to as 'string' variables and in most programs are marked with the '$' sign after the variable name (i.e. *name*$), while numeric variables have no special sign added (i.e. *length*). In many statistical packages it is not possible to interchange character and string variables by simple editing of the spreadsheet. Instead, a new variable has to be defined, based on the value of the variable that is to be altered. Variable names should be as simple and straightforward as possible.

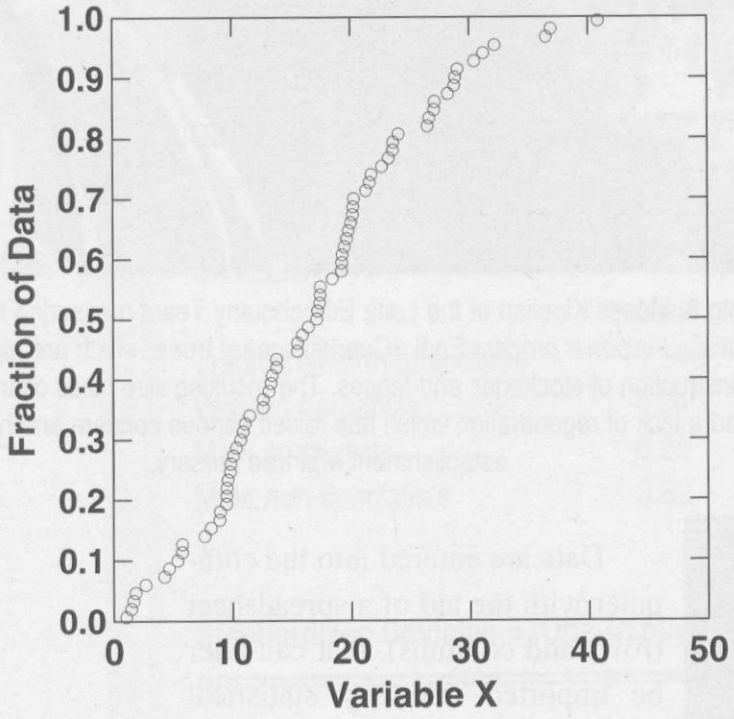
Data standardization and transformation

Classical parametric methods of inference make the assumption that the underlying population from which the sample data are drawn shows a normal distribution. Normal probability plots (Figure 1) help to visualize the distribution of one or more variables. A sample from a normal distribution results in an approximately *S*-shaped curve. A few of these methods, e.g. *t* - test, are robust in the sense that they are not sensitive to modest departure from normality. However, the accuracy of most tests is seriously affected at large deviations from normality. In that case, data are transformed so as to approximate a normal distribution (Berenson *et al.* 1983). In order to meet the conditions of normality, standardization of the basic data matrix is an essential step in most techniques. Besides, standardization in certain multivariate tests (e.g. principal component analysis, factor analysis) is done in order to remove the measurement units from the basic data. Standardization or transformation is achieved by treating the data with one of the transformation functions given in Box 3, where x'_{ij} is the transformed, while x_{ij}, and y are the original data.

Binary/two state character data are not standardized. For combinations of two- and multistate characters ordering should be used. For combinations of qualitative and quantitative data, one of the following options should be followed:

a. ignore the problem;
b. divide the data matrix;
c. convert the quantitative data to qualitative.

Figure 1. A quantile plot showing the standardized values of a variable *Y* (Fraction of Data) as a function of a variable *X*.



Box 3. Data transformations.

Logarithmic transformation:

$x'_{ij} = \log_{10}(x_{ij})$

or

$x'_{ij} = \log_{10}(x_{ij} + 1)$

Square root transformation:

$x'_{ij} = \sqrt{x_{ij}}$

or

$x'_{ij} = \sqrt{x_{ij} + 0.05}$

Divide by standard deviation:

$$x'_{ij} = \frac{x_{ij}}{\delta_1}$$

Standardization:

$$X'_{ij} = \frac{X_{ij} - \bar{X}}{\delta}$$

Proportional function:

$$x'_{ij} = \frac{x_{ij}}{\sum n_{l=1} x_{ij}} \qquad 0.0 \leq x \leq 1.0$$

Divide by the range value:

$$x'_{ij} = \frac{x_{ij}}{x_{max} - x_{min}} \qquad 0.0 \leq x \leq 1.0$$

Ordering

$$x'_{ij} = \frac{x'_{ij}}{x_{max} - x_{min}} \qquad 0.0 \leq x \leq 1.0$$

Linear transformation:

$Y' = (Y - a)/b + c$

SUBTRACTION OPTIONS:

$y - y_{min}$

$y - \bar{y}_i$

DIVIDE OPTIONS:

y / y_{max}

$y / y_{max} - y_{min}$

y / δ

$y / \sqrt{y - \bar{y}}$

$y / \sqrt{\Sigma y}$

$y / \sqrt{\Sigma y^2}$

$y / \Sigma y$

Classification and ordination techniques

In general, multivariate techniques are used to categorize or group the objects or experimental units. The aim of classification or ordination could be:

1. to get an overview of the variance;
2. to compare groups or trends among themselves or with additional data;
3. produce hypotheses to prepare further studies.

Clustering and classification

Classes have boundaries and hence an inner structure and relationships with external objects or other classes. Thus, algorithms have to address the problem of what to include in a particular class and what to exclude. Important criteria for judging, recognizing and testing of classifications and classes are:

- the centres (averages for elements);
- the density of classes;
- the variance of classes;
- the number of members;
- the "distinctness" of delimitation.

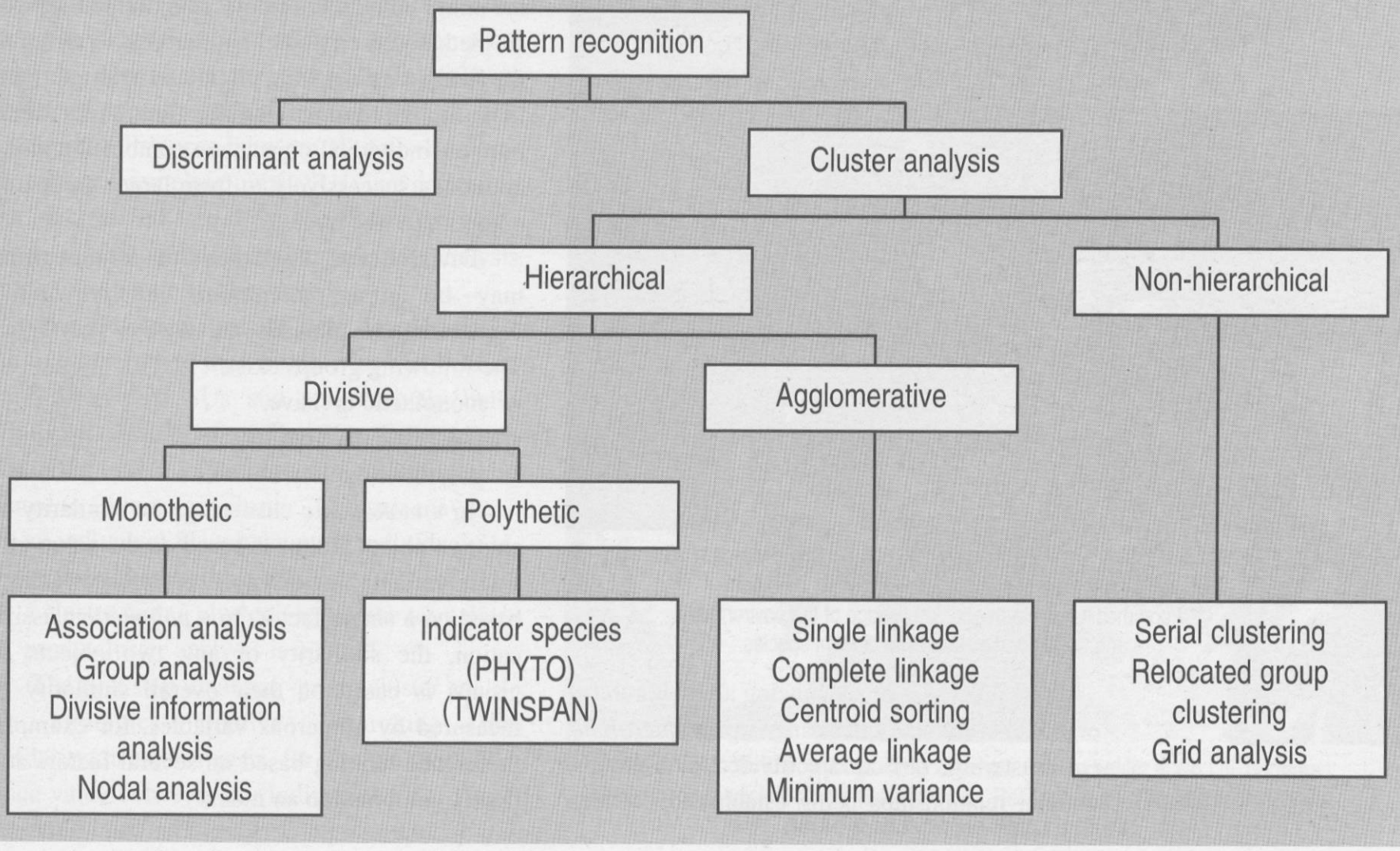
In different methods, different criteria are optimized. The significance of the respective criteria must be seen in relation to the objective of the study. The choice of methods depends on the objectives. Figure 2 gives an overview of the division of classification methods.

There are situations where the categorization is done in terms of groups that are themselves determined from the data. Such exploratory techniques for grouping objects (variables or items) are called 'clustering'. In classification methods other than cluster analysis, the number of groups are known beforehand and the objective is to assign new observations (items) to one of these groups. In cluster analysis, in contrast, no assumptions are made concerning the number of groups. Grouping is done on the basis of similarities or distances. The inputs required are similarity measures or data from which similarities can be computed.

CLUSTER ANALYSIS

Cluster analysis attempts to subdivide or partition a set of heterogeneous objects into relatively homogeneous groups. The objective of cluster analysis is to develop subgroupings such that objects within a particular subgroup are more alike than those in a different subgroup. Thus, the outcome of cluster analysis is a classification scheme that provides the sequence of groupings

Figure 2. Classification of classification methods (after Fischer & Bemmerlein 1986).



by which a set of objects is subdivided. Box 4 lists some examples of data which are suitable for cluster analysis.

Box 4. Examples of data suitable for cluster analysis.

- Similarity/dissimilarity of people's responses to well defined questions.
- Similarity/dissimilarity of plant utilization patterns among different ethnic, social or gender groups.
- Similarity/dissimilarity of species based on people's indication of use values
- Similarity/dissimilarity of phenotypic characteristics (e.g. seeds) in different varieties of food plants.
- Similarity/dissimilarity of the pattern of secondary compounds (e.g. essential oils) in different varieties of medicinal or aromatic plants.

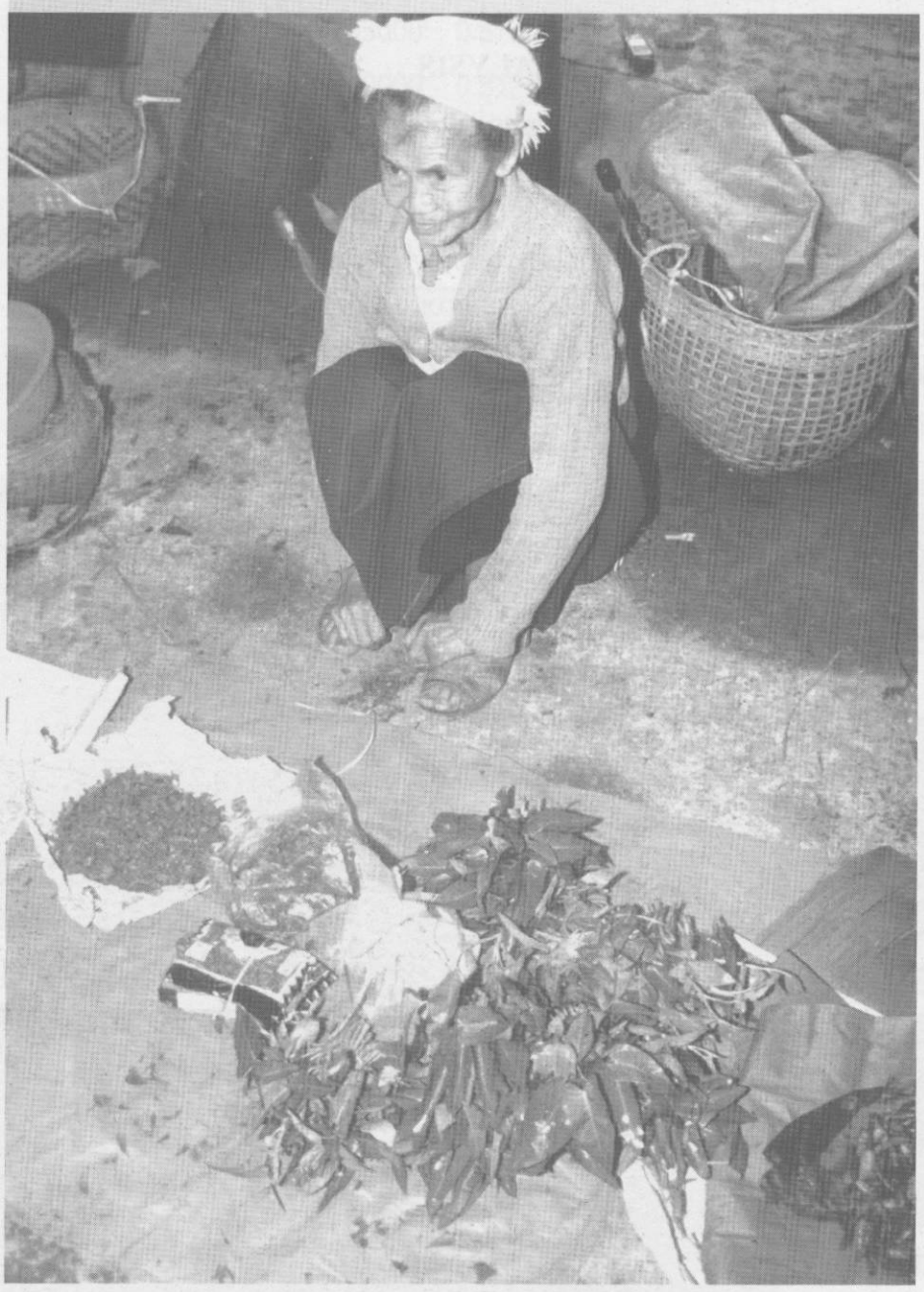


Photo 5. Woman selling herbal medicine at a market in Menglun, Yunnan Province, China. In most cases the older members of a community have a deeper knowledge of the environment and the properties and uses of plant species.

The processes of sequencing are hierarchical or non-hierarchical clustering. In *non-hierarchical* clustering, objects are divided into groups, without relationships being established between them, i.e. no dendrogram can be produced. Non-hierarchical clustering is particularly suitable for large data sets, since no complete similarity matrix must be calculated. All non-hierarchical clusters are calculated in the following way:

1. choice of number and position of initial cluster centres;
2. allocation of all objects to one respective cluster centre;
3. new calculation of cluster centres;
4. re-iteration of steps 2) and 3) until no further changes occur in the structure of clusters;
5. eventually merging of clusters.

The more widely used approach is *hierarchical clustering* arrangement. In this approach, once two objects are linked together at a particular stage, they cannot be separated into different clusters later on. Therefore, clustering decisions at a particular step are conditioned by the arrangement of objects at the previous step. In this approach the number of possible clustering choices decreases at each step. In hierarchical clustering, groups at any lower level of a cluster are exclusive subgroups of those groups at higher levels. In contrast to non-hierarchical clustering, statements on the relationships of classes (but not of the relationships of members in the respective classes) can be made in the hierarchical approach. The results can be depicted in the form of a dendrogram. All methods discussed in the following paragraphs are hierarchical.

Hierarchical clustering may be either *divisive* or *agglomerative*. In a *divisive* cluster analysis, the entire collection of objects is divided and re-divided, based on object similarities, to arrive at the final groupings (i.e., picture an inverted tree). In an *agglomerative* classification, as its name implies, individual objects are combined and re-combined successively to form larger groups of objects, (i.e. the tree).

Divisive and agglomerative arrangements may be either *monothetic* or *polythetic*. Agglomerative methods are always polythetic. The following groups exist:

- monothetic divisive,
- polythetic divisive,
- polythetic agglomerative.

In a *monothetic* clustering, the similarity of any two object groups is based on the value of a single variable, for example, preference ranking based on a single factor. In a *polythetic* classification, the similarity of any two objects or groups is based on their overall similarity as measured by numerous variables, for example, preference ranking based on several factors and finally combined to an index.

Agglomerative clustering procedures begin by considering each object as its own distinct cluster. Then two objects are placed together in a single cluster according to certain optimization criteria while grouping each of the remaining objects separately. In the next step, objects are grouped into either one cluster of three or two clusters of two (with each remaining object grouped separately). This clustering procedure continues sequentially until all objects are merged into one cluster.

Another criterion for defining cluster analyses is related to the measure of distance utilized in linking the objects for cluster formation. Alternative approaches are followed, including complete linkage, single linkage and average linkage. In *complete linkage*, the merger of two subsets of objects is based on the maximum distance between objects. This approach is also called *farthest neighbour* or *diameter method* and produces compact clusters of approximately equal size (unsuitable for ethnobotanical research questions). In *single linkage*, the merger is based on the minimum distance between objects. This approach is alternatively known as *nearest neighbour method* and often produces a single large chain-like cluster and several small clusters during its sequencing process. The *average linkage* approach bases the merger of two subsets of objects on the average distance between objects and is considered to be a way in between the first two approaches.

The general approach to cluster analysis is to compute a normal mode resemblance matrix between the objects (also referred to as sampling units or operational taxonomic units (OTUs)) using appropriate resemblance functions. The similarities/distances between all pairwise combinations of sampling units (SUs) in a collection are summarized into a SU x SU similarity/distance matrix and the various cluster analysis strategies operate on this matrix.

The cluster analysis models described here are agglomerative: they begin with a collection of N individual SUs and progressively build groups or clusters of similar SUs. During each clustering cycle, only one pair of entities may be joined to form a new cluster. This pair may be:

1) an individual SU with another individual SU,
2) an individual with an existing cluster of SUs,
3) a cluster with a cluster. Hence, the term pair-group cluster analysis is applied.

The first step in all pair-group cluster analysis strategies involves searching the similarity/distance matrix for the smallest distance value between two individual SUs. These two individual SUs may be represented by the symbols j and k, respectively. Hence, the first cluster is formed at a distance D(j,k) and this can be diagrammed using a dendogram. The initial collection of N SUs is now reduced to one cluster C1 (= SUs j and k joined) and N – 2 individual SUs. Special equations have been developed to compute the distance between this cluster and each of these N – 2 remaining SUs. A general linear combinatorial equation developed by Lance & Williams (1967) is given below:

$$D(j,k) = \alpha_1 D(j,h) + \alpha_2 D(k,h) + \beta D(j,k)$$

where the distance between the new cluster (j,k) is formed from the *j*th and *k*th SUs. A third *h*th SU or group of SUs can be calculated from the known distances D(j,k) D(j,h) and D(k,h) and the parameters α_1, α_2, and β. The distance between SU 3 and the cluster represented by SUs 1 and 4 is given by:

$$D(1,4)(3) = 1\,D(1,3) + 2\,D(4,3) + D(1,4)$$

The different clustering strategies differ only in their values for α_1, α_2, and β, which are the weights for determining the new distances.

Depending on the weighting scheme used, the resultant cluster formation varies. The group mean clustering strategy (the unweighted pair-group method with arithmetic averages - UPGMA) is most commonly used and it effectively computes the mean of all distances between SUs of one group to the SUs of another and, hence, is unweighted (see Legendre & Legendre 1998 for weighting strategies).

Photo 6. Pramoth Kheowvongsri interviewing a Palong healer in No Lai, northern Thailand, on medicinal plants use and trade. Responses from structured interviews can be analysed using cluster analysis.

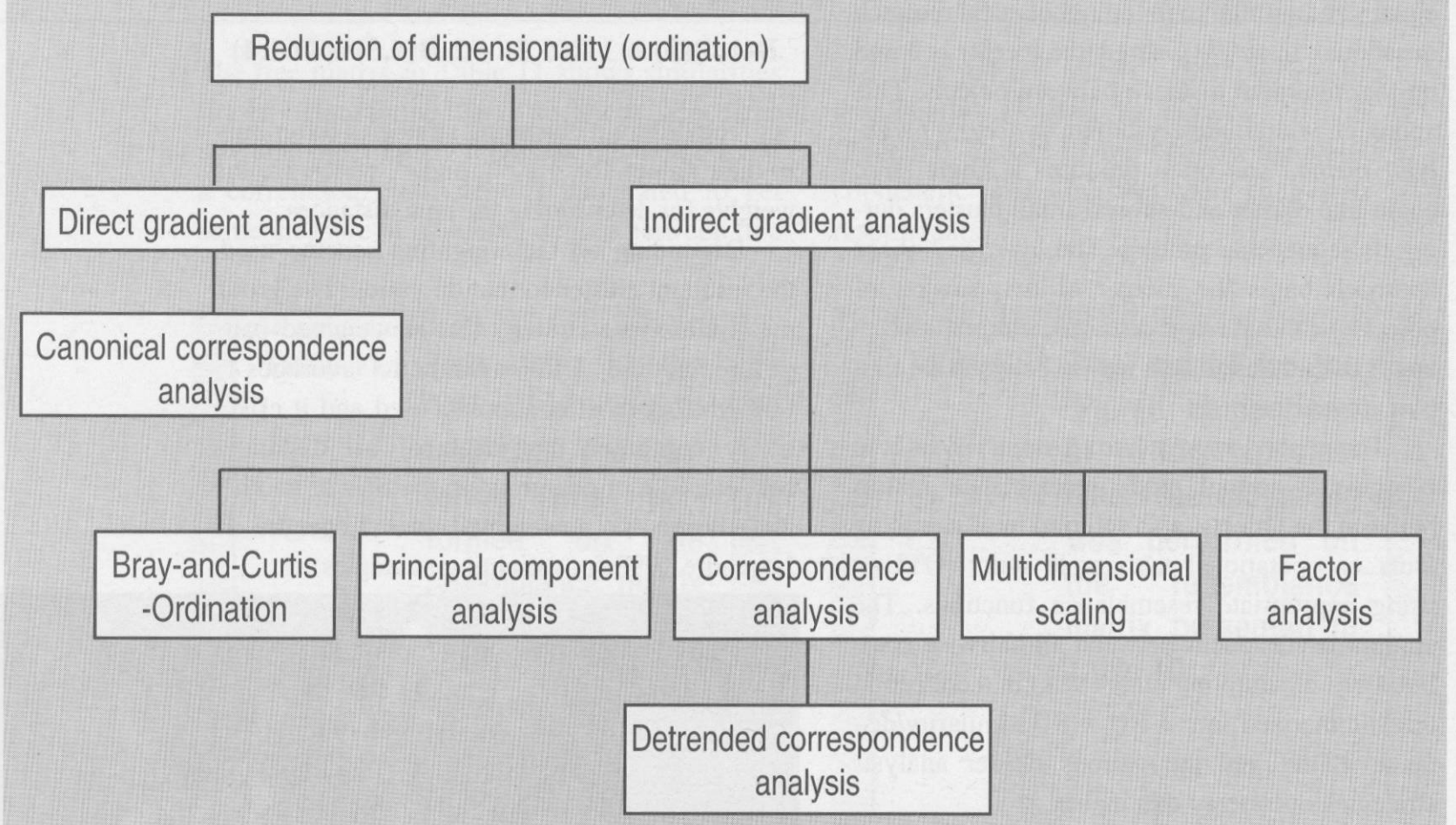
Ordination

Ordination involves reduction of dimensionality. The basic objective of reducing dimensionality in analysing multi-response data is to obtain simplicity for better understanding, visualization and interpretation. While reducing the dimensions, the techniques ensure the retention of sufficient details for adequate representation. Some of the important goals of reducing the dimensionality of multiple response data are as follows (Gnanadesikan 1977):

1. to screen out redundant variables or to find more insightful ones as a preliminary step to further analysis;
2. to stabilize scales of measurement, when a similar property is described by each of several variables. Here the aim is to compound the various measurements into fewer numbers;
3. to help in assessing the significance for testing a null hypothesis by compounding the multiple information. For example, small departures from null conditions may be evidenced on each of several jointly observed responses. It is advisable to integrate these non-centralities into a smaller dimensional space wherein their existence might be more sensitively indicated;
4. to obtain the preliminary specification of a space, which may be used later on in classification and discrimination procedures;
5. to detect the possible functional dependencies among observations in high-dimensional space.

In ordination two distinctly different approaches exist: direct and indirect gradient analysis (Figure 3).

Figure 3. Classification of ordination methods.



Historically, these methods are employed to investigate the relative importance of underlying ecological factors in vegetation analysis. In direct gradient analysis, vegetation relevés are arranged in an ecological space along axes of moisture, nutrients, altitude, *etc.* and the influence of the respective factors on the vegetation is determined. The indirect gradient analysis, in contrast to direct gradient analysis, focuses on the floristic composition. Five methods are distinguished:

- Bray-Curtis-Ordination,
- correspondence analysis,
- multidimensional scaling,
- principal component analysis, and
- factor analysis.

PRINCIPAL COMPONENT AND FACTOR ANALYSIS

The two most widely used classical linear reduction methods are principal component analysis (PCA) and factor analysis. In PCA, a d-dimensional observation (usually with correlated variables) is replaced by a k-linear combination of uncorrelated variables, where k is much smaller than d. Biplots are used to graphically

describe both, relationships among the *d*-dimensional observations x_1, x_2, x_3 ....x_n and relationships among the variables in two dimensions. Underlying assumptions for the data set to be analysed used PCA are:

1) data are normally distributed,
2) linear relationships exist between variables.

Only linear relationships are elaborated through PCA. The method looks at the objects (respondents) as an assembly of dots in a space, who's axes represent the (plant) species in question. The aim of the method is to project the multi-dimensional onto a two-dimensional hyperspace, such that minimum information on the distances between dots is lost. The first axis is laid through the centre of the dot cloud into the direction of largest variance. The second and following axes are perpendicular to the first axes, pointing into the direction of largest rest-variance. PCA is a transformation, in which the origin of the coordinate system is moved to the centre of the dot cloud and the axes are arranged according to variance. The problem of moving axes is mathematically solved through analysis of "Eigen" (German, meaning "self") vectors of the covariance or correlation matrix. Detrended correspondence analysis and reciprocal averaging are forms of PCA which were specifically developed for plant sociological analyses and are not further discussed here.

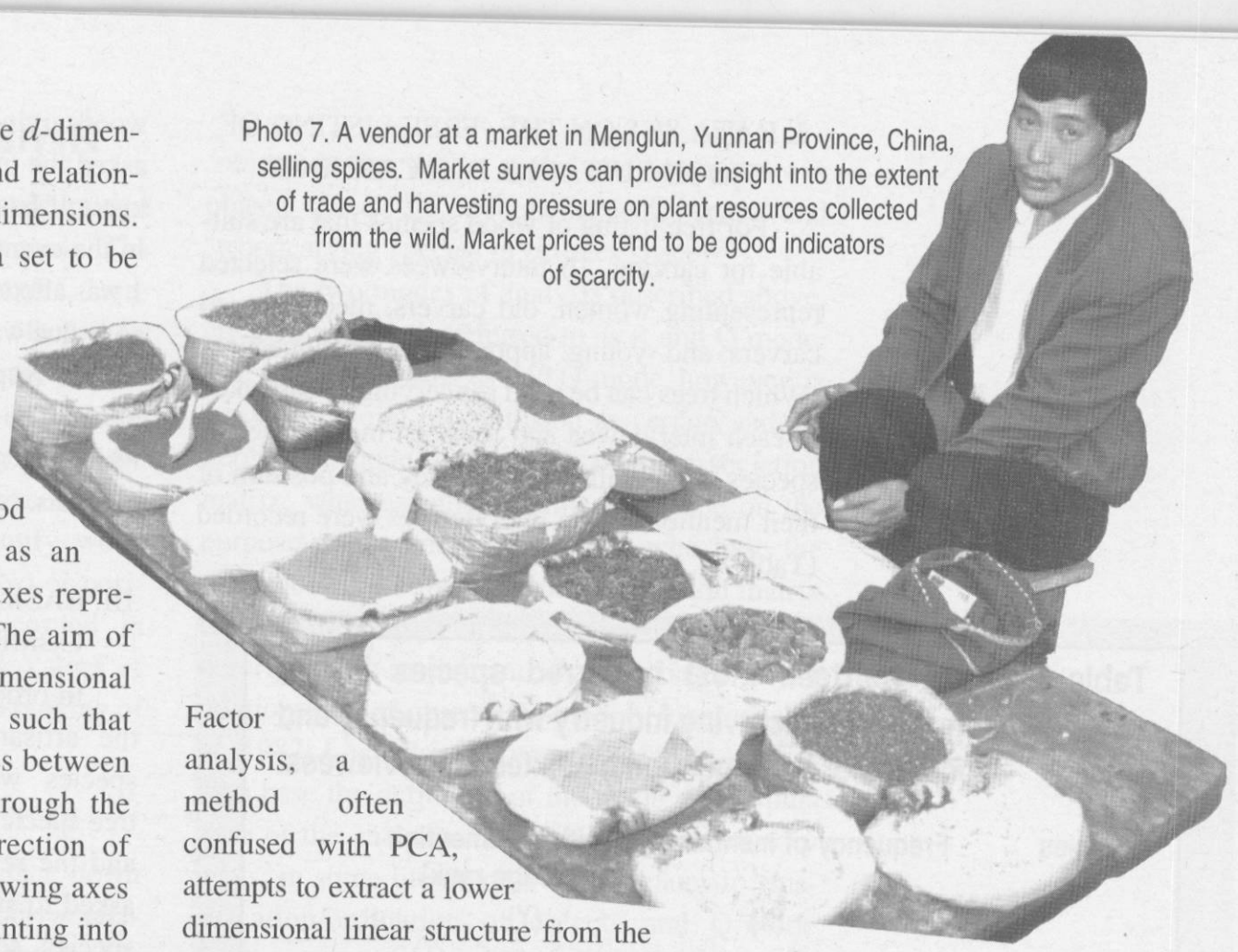


Photo 7. A vendor at a market in Menglun, Yunnan Province, China, selling spices. Market surveys can provide insight into the extent of trade and harvesting pressure on plant resources collected from the wild. Market prices tend to be good indicators of scarcity.

Factor analysis, a method often confused with PCA, attempts to extract a lower dimensional linear structure from the data that explains the correlations between the variables. However, when one subset of variables is compared with the subset of the remaining variables in the set, the method of canonical correlation (not discussed here) is used to find suitable linear combinations within each subset. If any grouping of the observations in a lower dimension is required to be highlighted, then canonical discriminant analysis (discriminant coordinates) can be performed. Linear combinations are then chosen to highlight group separation. In Box 5 some examples for application of principal component analysis are given.

Box 5. Examples of data suitable for principal component analysis.

- People asked to rank or categorize plant use values. PCA can be carried out on the People x Species matrix (with the rank in the cell). The resulting ordination diagram (with people in plants space) will reveal if there are certain groups of people that tend to value the same species in the same way, i.e. gender, ethnic or age groups. The species vectors in the diagram will indicate which species are characteristic for which groups.
- Spot people who respond differently from the majority. If a person just gave random answers or purposely replied incorrectly this person will be seen as an outlier on the ordination diagram, on the condition that there is a pattern in the answers in general.
- People indicating if certain species are useful (or not) for a number of purposes. A Species x Use matrix can be formed (with the number of species indicated in a certain use category). Ordination on these data will group species according to the use values assigned by people, and the vectors will indicate which uses characterize a group of species.
- Characterizing changes in e.g. floristic composition along environmental gradients. The axes would provide information on the most influential factor.

Examples of data matrices

The statistical analysis of the examples provided in this working paper are all based on matrices and matrix algebra. The following examples are drawn from an exercise where sixteen Kenyan woodcarvers were interviewed. During a workshop, three sets of data were collected:

- free listing of wood suitable for carving;
- wood identification task (yes/no; binary or two state character);
- paired comparison of wood species (ordered multistate character);

In the following paragraphs further details are provided on these data sets.

I. DATA SET ON THE 'FREE LISTING OF WOOD SUITABLE FOR CARVING'

For free listing of wood species that are suitable for carving, 16 interviewees were selected representing women, old carvers, medium-aged carvers and young apprentices. The question, 'Which trees can be used for carving?' was asked to each interviewee and fourteen most preferred species along with the frequency and position of their mention by the interviewees were recorded (Table 1).

Table 1. The fourteen most preferred species in the Kenyan woodcarving industry with frequency and position of their mention by sixteen interviewees.

Species	Frequency of mention (X)	Position of mention (average rank) (Y)
Brachylaena huillensis	16	1.4
Dalbergia melanoxylon	16	1.8
Combretum schumannii	16	4.9
Zanthoxylum chalybeum	10	6.2
Azadirachta indica	12	6.3
Sterculia africana	13	8.2
Olea europaea ssp. *africana*	6	9.3
Erythrina sacleuxii	12	9.7
Commiphora baluensis	11	9.7
Mangifera indica	10	10.3
Albizia anthelmintica	8	10.5
Terminalia brownii	9	10.7
Platycelyphium voense	6	12.0
Oldfieldia somalensis	10	12.5

Table 2. A pairwise ranking matrix for five tree species used in woodcarving. *

S_1	S_2	S_3	S_4	S_5
	S_2	S_3	S_4	S_5
		S_2	S_2	S_2
			S_3	S_5
				S_5

	Score	Rank
S_1	0	1
S_2	4	5
S_3	2	3
S_4	1	2
S_5	3	4

*** The table is based on the preferences expressed by one respondent (R_1).**

II. BASIC DATA MATRIX FOR THE 'WOOD IDENTIFICATION TASK'

The data on the 'Wood identification task' were collected on eight species based on the responses of sixteen respondents involved in a woodcarving project in Kenya. Each artisan was asked the question separately for eight species to know if he or she can identify the species or not. In the event of a positive reply ('Yes'), the value 1 was allotted; alternatively, if the reply was 'No', a 0 value was assigned. In this way, the matrix for sixteen respondents and eight species was completed. The species were arranged across the rows, while the respondents were arranged across the columns. (see Table 4, Appendix, p. 36).

III. BASIC DATA MATRIX FOR A 'PAIRED COMPARISON OF WOOD SPECIES'

In order to assess species preference among the artisans, a 'Paired comparison of wood species' was undertaken. For the purpose, five tree species used for woodcarving were selected and the respondents (the sixteen artisans) were asked to state their preference between any two species set or pair combination of the five species. Preferences of each respondent in respect of five such possible species pair combinations (n (n-1) / 2) were tabulated as shown in Table 2. The score is defined by the total number of mentions in the table and the highest rank is assigned to the species with the highest score. Pairwise rank matrices were then prepared in respect of each respondent (R_1.....R_{16}). Finally, the ranks for five species so obtained from the responses were tabulated in matrix form. The rows of the matrix represented the species and the columns were respondents.

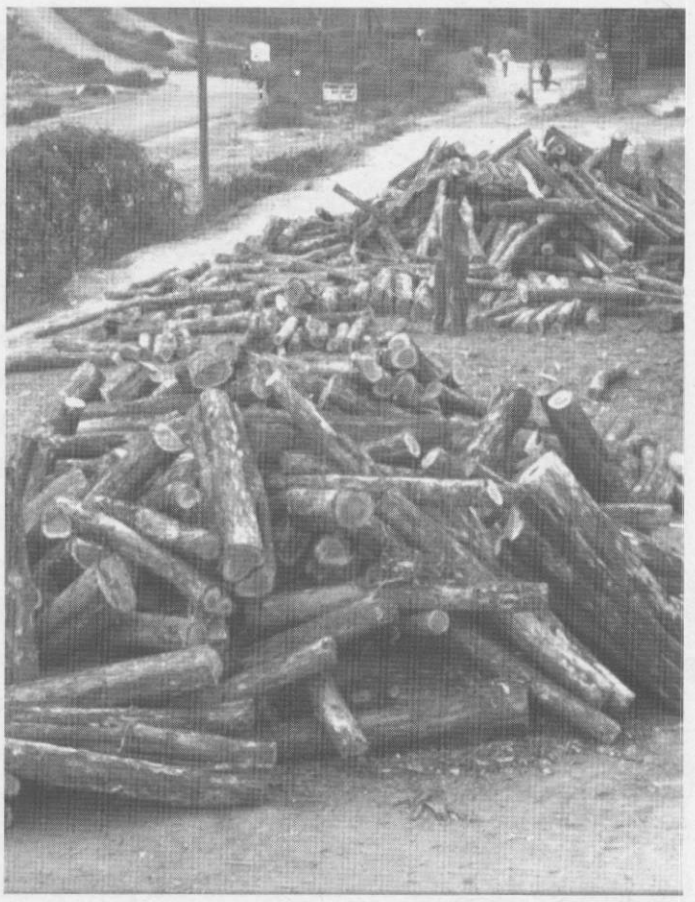


Photo 8. *'Muhuhu'* (*Brachylaena huillensis* O. Hoffm., Asteraceae) logs piled up outside a carving workshop in Wamunyu, Eastern Province, Kenya. Each year 40,000 indigenous trees are felled in Kenya for woodcarving.

Matrix structure and analysis

The term descriptor is used for the attributes that describe or compare the objects of the study. The objects may be the respondents, samples, locations, quadrats, observations or any other sampling units (e.g. operational taxonomic units - OTUs in numerical taxonomy). In our example the respondents (R_1 - R_{16}) were objects and the species were the descriptors, i.e. responses of the artisans (measures of ability to identify wood species used for woodcarving). Yes/No or positive/negative reply values were recorded in 'Wood identification task' while rank values of the species were recorded in 'Pairwise ranking of wood species' data.

NORMAL *VS.* INVERSE ANALYSIS

Data matrices can be viewed either down columns or across rows, i.e. one can look at relations between objects or between descriptors. For instance, one may wish to explore the relationship between respondents/objects to see whether certain groups of people gave similar responses and therefore may have similar attitudes towards carving wood. Or, one may wish to explore relations between descriptors/ rows to highlight for which species people tend to give similar responses. Maximum information can often be obtained by making both modes of analysis. The two modes of analysis require different measures of association as objects are independent of each other (sampling of objects is preferably done in a way to ensure mutual independence of sampling units), whereas descriptors may be dependent. A variety of association measures are available to study the relationship of objects (e.g. Legendre & Legendre, 1998). Different correlation coefficients are applied to the study of relations between descriptors.

If objects are grouped on the basis of the entire set of descriptors, it is sometimes referred to as normal analysis, whereas in an inverse analysis, descriptors are grouped on the basis of their distribution in a series of objects (Kent & Coker 1994). In connection with ordinations the two modes of analysis have been referred to as 'objects in descriptors space' and 'descriptors in objects space', e.g. 'people in species space' and 'woodcarving species in peoples space'.

The two modes of analysis described above, are also frequently referred to as R and Q mode. The use of the terms R- and Q-mode, however, is a possible point of confusion as certain authors define the mode on the basis of the association matrix, whereas others define the mode on the purpose of the analysis. Authors who base the definition of the mode on the association matrix call analyses based on the relationships between descriptors 'R-mode', and analyses based on the relationships of objects 'Q-mode' (Jongmann *et al.* 1987, Legendre & Legendre 1998). Authors who base the definition of the mode on the purpose of the analysis, do so in two contradictory ways: in some literature R-mode relates to classification/ordination of objects and Q-mode relates to species classification/ordination (Pielou 1984; Causton 1988; Kent & Coker 1994). Again, in other literature, Q-mode relates to classification/ordination of objects and R-mode relates to species classification/ordination (Romesburg 1984).

Photo 9. National Museums of Kenya researchers Mohamed Pakia, Raymond Obunga and Hamisi Mududu measuring DBH and basal diameters of standing and cut '*mgurure*', *Combretum schumannii* Engl. (Combretaceae), trees in Dzombo Forest, coastal Kenya.

Object ordinations normally begin with a dispersion/correlation matrix of descriptors (although they can be based on a association matrix of objects). According to Legendre & Legendre (1998), object ordinations can therefore be both, R- and Q-mode. Because of these confusing notations, we prefer to use the terms 'normal' and 'inverse' to describe the purpose of the analysis.

SIMILARITY MEASURES

The analysis is started on a resemblance matrix which is derived either from the original or a transformed/standardized data matrix. These resemblance matrices are called 'similarity matrix' or 'dissimilarity matrix' depending on the way in which resemblance functions are calculated and the matrix is derived. In this section, some resemblance functions that quantify the similarity or dissimilarity between samples are described. The more similar the objects (respondents or samples) are with respect to a particular character (variable), the greater their resemblance and the smaller the distance between them when projected into a geometric space. Resemblance functions quantify the similarity or dissimilarity between two objects (samples) based on observations over a set of descriptors (Sneath & Sokal 1973). To explore the nature of relationships or affinities that exists among the respondents, normal mode analysis is usually applied. Two types of normal mode resemblance functions are distinguished:

1. similarity coefficients and
2. distance coefficients.

Similarity coefficients vary from a minimum of 0 (when a pair of respondents are completely different) to 1 (when the respondents are identical). On the other hand, distance coefficients assume a minimum value of 0 when a pair of respondents are identical and have some maximum value (in some cases infinity) when the pair of respondents are completely different. Hence, distance coefficients are also referred to as dissimilarity coefficients. In fact, a similarity index may always be expressed as a distance just by a simple transformation such as 1 – similarity (Legendre & Legendre 1998). Thus, distance may be thought of as the complement of similarity (Sneath & Sokal 1973).

Similarity coefficients are widely used indices. These indices are based solely on presence/positive reply (indicated with a '1') or absence/negative reply (indicated with a '0') data (see Appendix, Tables 4 to 8 for illustration).

Three indices - Ochiai, Dice and Jaccard - are useful for calculating the similarity index of presence/absence or positive/ negative reply data (qualitative) (see Box 6).

Box 6. Indices for calculating similarity index of presence/absence or positive/negative reply data (qualitative).

Ochiai Index (OI)

$$OI = \frac{a}{\sqrt{a+b}\sqrt{a+c}}$$

In the above example:

$$OI_{R14,R15} = \frac{1}{\sqrt{1}\sqrt{3}} = 0.577$$

Dice Index (DI) (Sorensen Index)

$$DI = \frac{2a}{2a+b+c}$$

In the above example:

$$DI_{R14,R15} = \frac{2}{2+0+2} = 0.5$$

Jaccard Index (JI)

$$JI = \frac{a}{a+b+c}$$

In the above example:

$$JI_{R14,R15} = \frac{1}{1+0+2} = 0.33$$

These indices can be used to measure the degree of association between species (an inverse mode analysis, i.e., across the rows of the data matrix) as well as to compute a normal mode similarity between respondents. It may be mentioned here that these are the only types of functions that are used to measure both normal mode (sample similarity) and inverse mode (species association) resemblance (Ludwig & Reynolds 1988).

DISTANCE COEFFICIENTS

Measures of distance may be categorized into three groups:

1. E-group (the Euclidean distance coefficients);
2. BC-group (the Bray-Curtis dissimilarity index);
3. RE-group (the relative Euclidean distance measures).

The distances computed between all possible pairs of sampling units (SUs) based on any of the above similarity or distance measures are arranged in a SU x SU matrix. Examination of this matrix quickly reveals the distance between any two SUs. It is on this distance matrix that the clustering strategies and ordination techniques such as principal component analysis operate. The distance coefficients are explained in Box 7.

Box 7. Distance coefficients

(after Ludwig & Reynolds 1988).

E-GROUP DISTANCES

EUCLIDEAN DISTANCE (ED)

This measure is the familiar equation for calculating the distance between two points R_j and R_k in Euclidean space:

$$ED_{jk} = \sqrt{\sum^{s}_{i=1}(X_{ij} - X_{ik})^2}$$

The value of ED ranges from zero to infinity, as do all of the E-group measures.

SQUARED EUCLIDEAN DISTANCE (SED)

This measure is the square of ED:

$$SED_{jk} = \sum^{s}_{i=1}(X_{ij} - X_{ik})^2$$

MEAN EUCLIDEAN DISTANCE (MED)

MED is similar to ED, but the final distance is on a smaller scale since the mean difference is used:

$$MED\ jk = \sqrt{\frac{\sum^{s}_{i=1}(X_{ij} - X_{ik})^2}{S}}$$

ABSOLUTE DISTANCE (AD)

This measure is the sum of the absolute differences taken over the S species:

$$AD_{jk} = \sum^{s}_{i=1}(X_{ij} - X_{ik})$$

This distance measure is also known as Manhattan or City block dissimilarity coefficient measure. The AD measure is the character difference in numerical taxonomy (Sneath & Sokal 1973).

MEAN ABSOLUTE DISTANCE (MAD)

The MAD is similar to AD, but a mean distance is used rather than an absolute distance:

$$MAD\ jk = \frac{\sum^{s}_{i=1}(X_{ij} - X_{ik})}{S}$$

MAD is equivalent to the mean character difference used in numerical taxonomy (Sneath & Sokal 1973).

BC-GROUP DISTANCE

This group is represented by a single index first introduced by Bray & Curtis (1957). The step is to compute the percent similarity (PS) between SUs j and k as

$$PS_{jk} = \frac{2W}{A + B}\ 100$$

Where $W = \sum s_{i=1} \min(X_{ij}, X_{ik})$

$A = \sum^{s}_{i=1} X_{ij}$

$B = \sum^{s}_{i=1} X_{ik}$

Percent Dissimilarity (PD):

$$PD = 100 - PS$$

PD may also be computed on a 0 – 1 scale as

$$PD = 1 - [2W/(A+B)]$$

RE-GROUP DISTANCE

This group contains distance indices that are expressed on standardized or relative scales.

RELATIVE EUCLIDEAN DISTANCE (RED)

$$RED_{jk} = \sqrt{\sum^{s}_{i=1}[(X_{ij} / \sum^{s}_{i} X_{ij}) - (X_{ik} / \sum^{s}_{i} X_{ik})]^2}$$

RED ranges from 0 to $\sqrt{2}$.

RELATIVE ABSOLUTE DISTANCE (RAD)

$$RAD_{jk} = \sum^{s}_{i=1}[(X_{ij} / \sum^{s}_{i} X_{ij}) - (X_{ik} / \sum^{s}_{i} Xi_{k})]$$

RAD has a range from 0 to 2.

CHORD DISTANCE (CRD)

This is done by projecting the SUs on to a circle of unit radius through the use of direction cosinuses. The measure is then the chord distance between the two SUs after such a projection.

$$CRD_{jk} = \sqrt{2\,(1 - ccos_{jk})}$$

Where the chord cosinus (ccos) is computed from:

$$ccos\ jk = \frac{\Sigma si = 1\ (X_{ij}, X_{ik})}{\sqrt{\sum^{s}_{i} X^{2}_{ij} \sum^{s}_{i} X^{2}_{ik}}}$$

In case of binary data, this ccos is identical to Ochiai's coefficient. CRD, like RED, ranges from 0 to $\sqrt{2}$.

GEODESIC DISTANCE (GDD)

This measure is the distance along the arc of the unit circle (rather than the chord distance) after projection of the SUs onto a circle of unit radius:

$$GDD_{jk} = \arccos(ccos_{jk})$$

GDD has a range from 0 to $\pi/2$ (i.e. 0 to 1.57).

Applications of cluster and principal component analysis

Cluster analysis of the 'Wood identification' task

The six basic steps involved in cluster analysis are described below taking the data sets from the Kenyan woodcarving project 'Wood identification task' and 'Paired comparison of wood species' as example. The utility of cluster analysis on such data are:

- The responses (objects) can be grouped according to their resemblances, i.e. based on the respondents' ability to identify a particular species used for woodcarving in case of 'Wood identification task' data and on species preferences in the 'Paired comparison of wood species' data. The respondents in each cluster should have a number of common characteristics that set them apart from the respondents of other such clusters.
- The data sets can be reduced to homogeneous groups or clusters. The objective is to demonstrate the relationships of the respondents to each other and to simplify these relationships so that some general statements about the classes of respondents that exist can be made.

Being an ethnobotanical problem, where the interest is to know about the respondents through their view on the individual species, normal mode analysis will be used for both data sets. The procedure is a polythetic, agglomerative classification technique. The results are based on the output of the NTSYS package but the basic steps are similar for any other package.

Box 8. Steps involved in cluster analysis.

STEP 1 **Obtaining the basic data matrix** (see Appendix, Tables 4 and 5, page 36).

STEP 2 **Standardizing the basic data matrix.** The basic data matrix is standardized for following reasons:

- To make the species contribute more equally to the similarity between the respondents.
- To remove all the measuring units (not applicable to the data presented here).

The standardization is performed through a linear transformation of the original values for each character/element of the basic data matrix. Since binary data are not standardized, the basic data matrix for the 'Wood identification task' will be used for further analysis. The basic data matrix for 'Paired comparison of wood species' has been standardized by dividing the matrix elements by the standard deviation (see Appendix, Table 6).

STEP 3 **Computing the resemblance matrix.** The next step in cluster analysis is to compute a normal mode resemblance between the respondents (R_1...R_{16}). Although any of the numerous resemblance functions available could be used, distance measures have been used for multistate character data in the 'Paired comparison of wood species' because of their heuristic value in a cluster analysis (Sneath & Sokal 1973). However, for two state data in the 'Wood identification task', the similarity measure is Jaccard's coefficient. The distances between all pairwise combinations of respondents are summarized into a 16 x 16 distance (D) or resemblance matrix for each data set (see Appendix, Tables 9 and 10, page 37 and 38). The further cluster analysis strategies operate on these resemblance matrices.

STEP 4 **Executing the clustering method and obtaining the tree matrix.** The clustering technique used here is a hierarchical agglomerative procedure based on UPGMA (unweighted pair-group method with arithmetic averages). The clustering was executed on the resemblance matrices (see Appendix, Tables 9 and 10) to yield the tree matrices (see Appendix, Tables 11 and 12).

STEP 5 **Drawing the tree or dendrogram.** The tree matrix derived below produces a tree on scale showing the clustering scheme. The dendrograms or trees for 'Wood identification' and 'Paired comparison of wood species' data is given in Figures 4 and 5, respectively (page19).

STEP 6 **Computing the cophenetic matrix and coefficient, and plotting.** A tree is not exactly like the data matrix it represents. It is necessary to know how well the tree represents the basic data matrix. The cophenetic correlation coefficient measures how well the tree and the resemblance matrix matches. The values that appear in the cophenetic matrix (see Appendix, Tables 13 and 14, page 39) stem from the tree and are compared with those of the basic data matrix either through a matrix plot or Pearson product moment correlation coefficient. Figures 11 and 12 (Appendix, page 39) show the relationships between the cophenetic and resemblance matrices.

The clustering results depicted in the dendrogram (Figure 4) for the 'Wood identification' task exhibit a clear separation of respondents at the eight-cluster level. The two-cluster solution separates the two groups (R_{16} from the rest) which may be different in their socio-economic conditions, age structure, artisan skills or ethnic composition. This may be examined through the already collected data in these respects or a new explorative study may be designed for testing the above hypothesis. Further, the tree shows that R_1, R_2, R_3, R_4, R_5, R_{15} and R_{11} are similar. Based on this, the researcher can treat these respondents as similar in further experiments or in designing new studies. In addition, the factors responsible for such similarity may also be explored, which may have high ethnobotanical relevance.

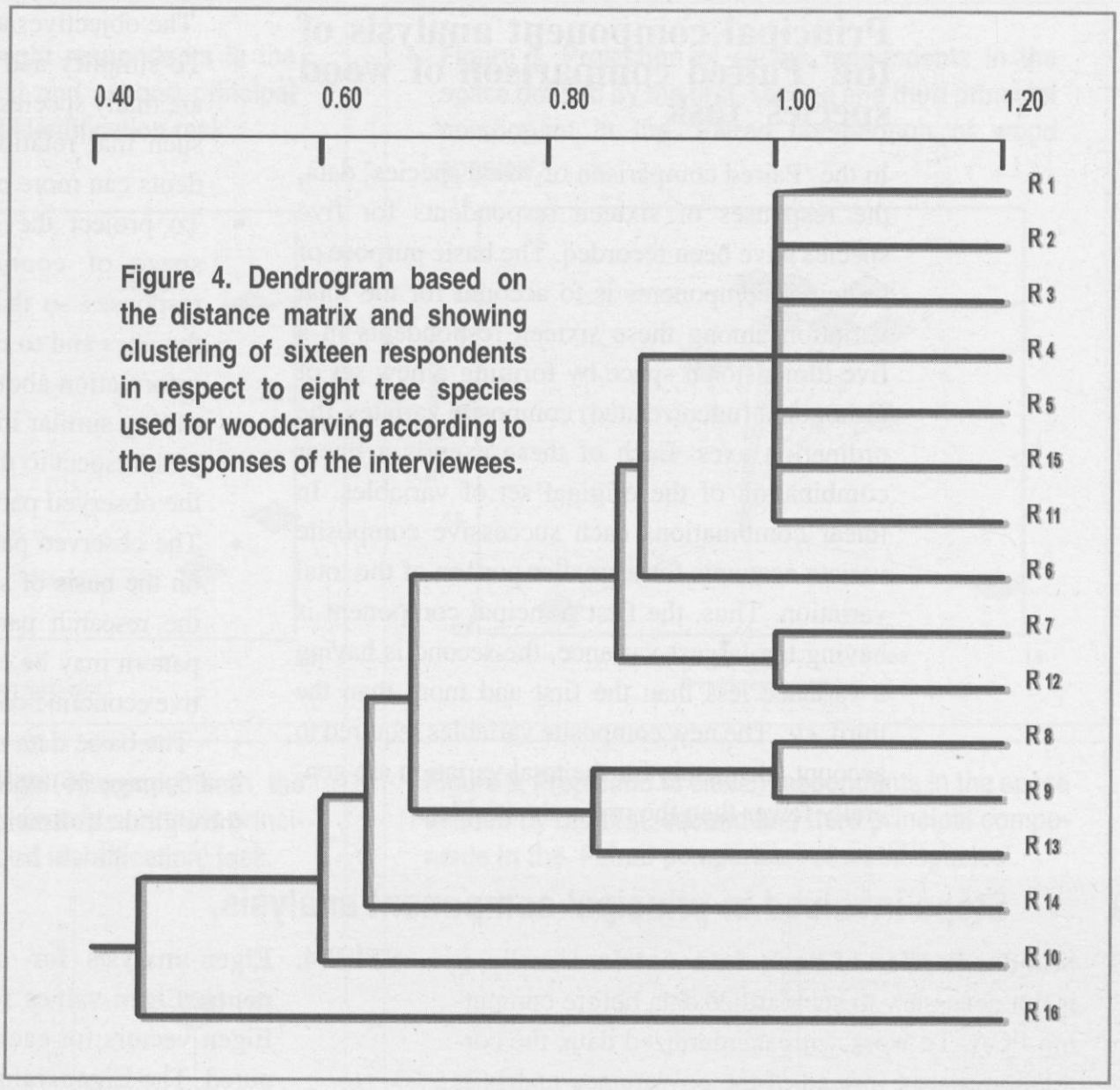


Figure 4. Dendrogram based on the distance matrix and showing clustering of sixteen respondents in respect to eight tree species used for woodcarving according to the responses of the interviewees.

Similarly, the tree in Figure 5 for the 'Paired comparison of wood species', reveals the existence of two groups of respondents (i.e. R_3, R_4, R_{16} and R_7 in one group and the rest in another group). The underlying factors for such grouping pattern may be explored. Further, two distinct groups of respondents exist: one group with R_1, R_2, R_{10} and R_{14} respondents and the other with R_6, R_{13}, R_9, R_{11} and R_{12}. Each group consists of a large number of similar respondents (five and four respectively). The factor(s) behind such similarity may be an interesting ethnobotanical observation.

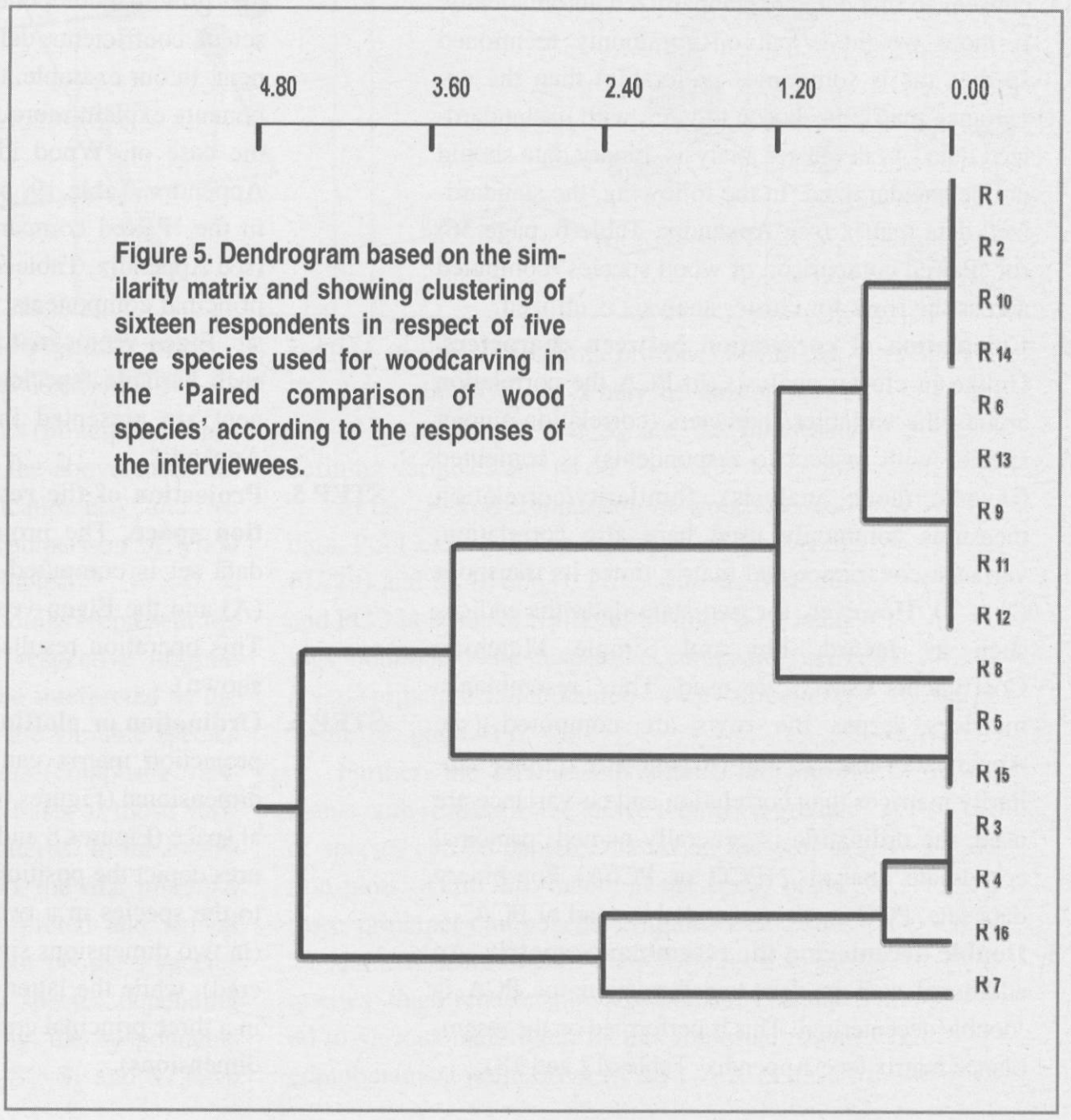


Figure 5. Dendrogram based on the similarity matrix and showing clustering of sixteen respondents in respect of five tree species used for woodcarving in the 'Paired comparison of wood species' according to the responses of the interviewees.

Principal component analysis of the 'Paired comparison of wood species' task

In the 'Paired comparison of wood species' data, the responses of sixteen respondents for five species have been recorded. The basic purpose of principal components is to account for the total variation among these sixteen respondents in a five-dimensional space by forming a new set of orthogonal (uncorrelated) composite variates, the ordination axes. Each of these axes is a linear combination of the original set of variables. In linear combinations each successive composite variate accounts for a smaller portion of the total variation. Thus, the first principal component is having the largest variance, the second is having a variance less than the first and more than the third, *etc.* The new composite variables required to account adequately for the total variation are generally fewer than the original variables.

The objectives are as follows:

- To simplify and condense data sets. If there are many species, dimensions can be reduced, such that relationships between the respondents can more easily be examined.
- To project the research participants in the space of coordinates according to their responses so that their relative positions to the axes and to each other provide maximum information about their similarities. By identifying similar informants from their position with respect to the axes, underlying factors in the observed pattern may be searched.
- The observed patterns may later be explained on the basis of social and cultural features of the research participants. Differences in the pattern may be correlated with ethnicity, relative economic conditions, family structure, etc.

The basic data matrices (Appendix, Tables 4 and 5, page 36) will be used. PCA is carried out through the following six steps (Box 9).

Box 9. Steps involved in principal component analysis.

STEP 1. Standardization of basic data matrix. Usually, it is not necessary to standardize data before computing PCA. To work with standardized data, the correlation matrix instead of the co-variance matrix is chosen, so that data are standardized automatically. If more weight is put on commonly mentioned species (as is sometimes preferable) then the co-variance matrix is chosen to work with unstandardized data. As in cluster analysis, binary data should not be standardized. In the following, the standardized data matrix (see Appendix, Table 6, page 36) for 'Paired comparison of wood species' computed across the rows for cluster analysis is utilized.

STEP 2. Calculation of correlation between characters. Unlike in cluster analysis, in PCA the correlation among the variables/characters (correlation among species with respect to respondents) is computed (inverse mode analysis). Similarity/correlation measures commonly used here are: correlation, variance-covariance and matrix times its transpose (X x T). However, for two-state data, the indices such as Jacard, Phi and Simple Matching Coefficients (SMC) are used. Thus, resemblance matrices across the rows are computed (see Appendix, Tables 15 and 16, page 40). If other similarity matrices than correlation and co-variance are used, the ordination is generally named 'principal coordinate analysis' (PCO or PCoA) For binary data sets, PCO is recommended instead of PCA.

STEP 3. Double decentering the resemblance matrix. An additional step in data transformation for PCA is 'double decentering'. This is performed on the resemblance matrix (see Appendix, Tables 17 and 18).

STEP 4. Eigen-analysis for deriving principal components. Eigen-values for each ordination axis and Eigen-vectors for each variable (species) are computed. The Eigen-value is the variance of a particular principal component while Eigen-vector is the set of coefficients defining the principal component. In our example, the first three principal components explain more than 85% of the variance in the case of 'Wood identification task' data (see Appendix, Table 19, page 41) and more than 89% in the 'Paired comparison of wood species' data (see Appendix, Table 20). Therefore, the first three principal components were used for further analysis. Eigen-vector matrices (U) with the loading of each variable (species) in each principal component are presented in Tables 21 and 22 in the Appendix.

STEP 5. Projection of the respondents into the ordination space. The projection matrix (Y) for each data set is computed from the basic data matrix (A) and the Eigen-vector matrix (U). Y = A x U. This operation results in projection matrices (not shown).

STEP 6. Ordination or plotting of projection matrix. The projection matrix can be plotted in both a two-dimensional (Figures 6 and 7) and three-dimensional space (Figures 8 and 9, page 21). The former figures depict the position of respondents with respect to the species in a two principal component space (in two dimensions since only two PCs are considered), while the latter two arrange the respondents in a three principal components space (thus in three dimensions).

Figure 6. Projection of eight respondents in the space defined by the first and second principal component in the 'Wood identification task'.

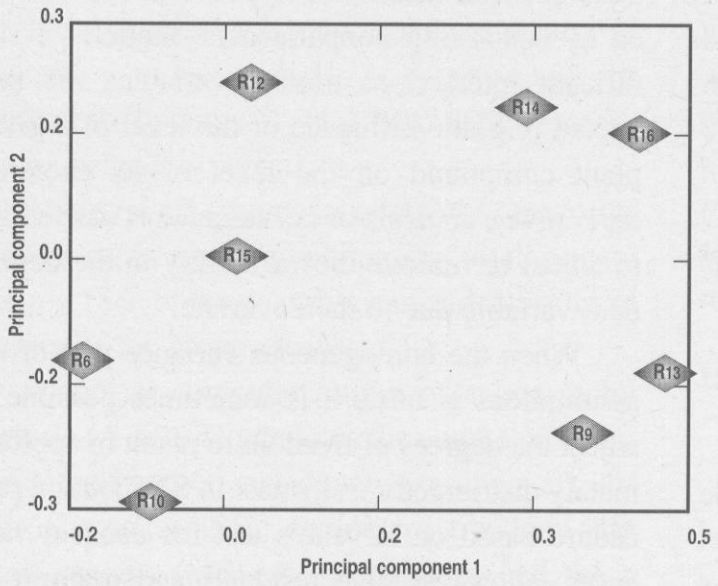


Figure 7. Projection of eleven respondents in the space defined by the first, second and third principal component in the 'Paired comparison of wood species'.

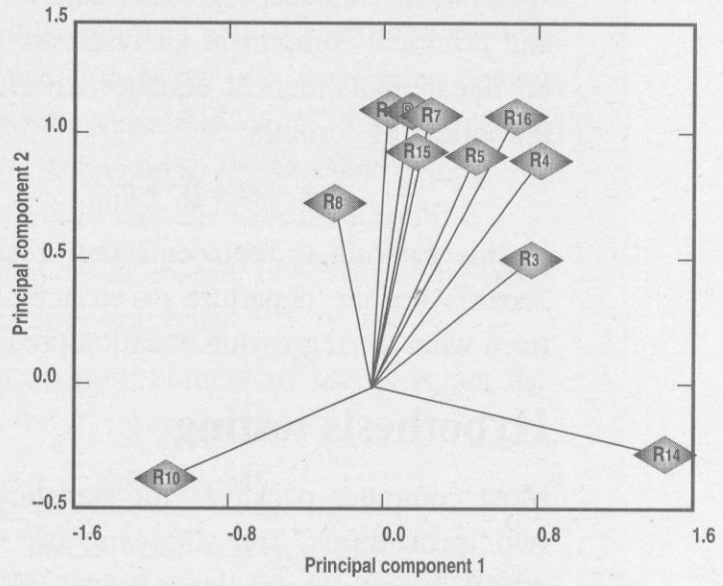


Figure 8. Projection of eight respondents in the space defined by the first, second and third principal components in the 'Wood identification' task.

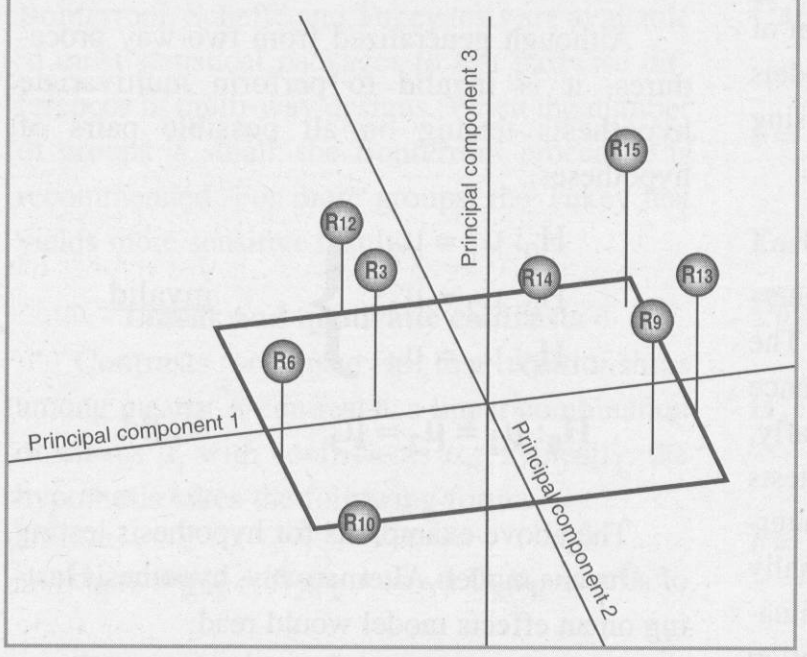
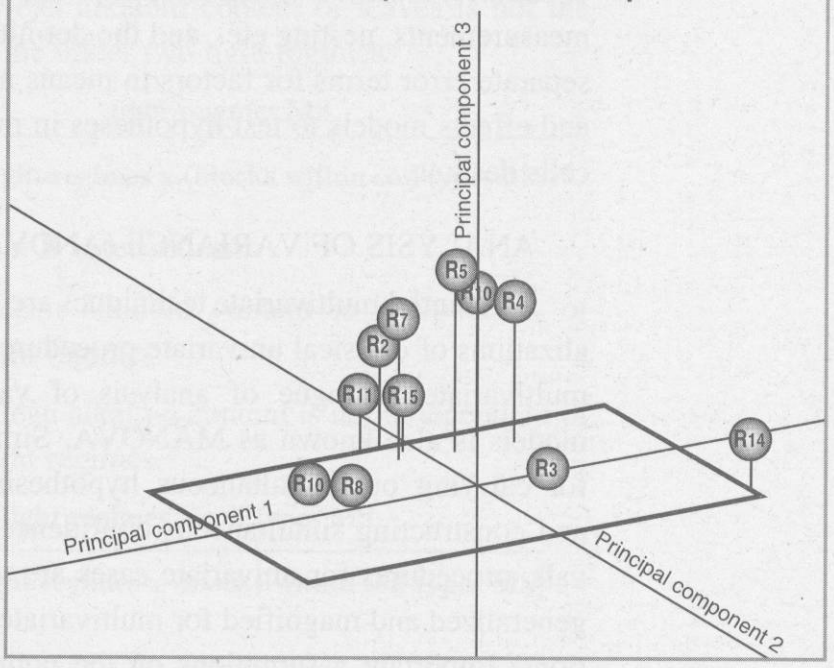


Figure 9. Projection of eleven respondents in the space defined by the first, second and third principal components in the 'Paired comparison of wood species'.



INTERPRETING THE RESULTS OF PRINCIPAL COMPONENT ANALYSIS

Three composite variables (principal components) were derived in both the above cases out of eight in the 'Wood identification task' and five in the case of the 'Paired comparison of wood species' original species variables.

The coefficients (i.e. loadings) represent the correlation of Y_1 with the respective original variable. Thus, 0.2695 can be interpreted as the correlation between Y_1 and the variable species S_1. The principal components (composite variables) are interpreted on the basis of those variables with strong loading patterns. In the example (see Appendix, Table 21), the first principal component Y_1 may be interpreted and named accordingly: for example poor people's species or 'X' ethnic community's species depending upon information concerning the respondents and according to which S_1, S_2, S_4 and S_7 have similar (positive) loading profiles within Y_1. Similarly, the PC 2 (Y_2) may be appropriately named in which S_3 and S_8 are the important defining variables and so on.

In the 'Paired comparison of wood species' data, PC 1 can be defined in terms of species S_1 (0.729) and S_2 (0.6063), PC 2 with S_3 (0.4586) and PC 3 in terms of S_5 (0.5808) since their loadings dominated the respective composite variables (principal components) (see Appendix, Table 22, page 41).

Further, the relationship among the informants with respect to the above identified groups of species can be depicted based on the projection plots of the informants in the space of the three principal components (Figures 6-9, pages 19 and 20) according to their responses to the queries. Such relationships can be later correlated to various underlying factors important from ethnobotanical point of view.

Comparison of several means

Analysis of variance, regression analysis, cluster and principal component analysis, are all based on linear mathematical equations derived from the following formula:

$$y = \alpha + \beta_i + \varepsilon_i$$

In this formula ε_i represents the "residual" or "error", i.e. the departure of an actual y-value from what the regression equation predicts.

Hypothesis testing

Most computer packages for statistics provide two procedures for analysis of variance: ANOVA and general linear model (GLM). The latter is usually less automated and allows the analysis of randomized and incomplete block designs, analysis of co-variance with one or more covariates, crossover designs, split plot, repeated measurements, nesting etc., and the definition of separate error terms for factors in means models and effects models to test hypotheses in missing cells designs.

ANALYSIS OF VARIANCE (ANOVA)

Inferential multivariate techniques are generalizations of classical univariate procedures. The multivariate analogue of analysis of variance models is also known as MANOVA. Similarly, for carrying out simultaneous hypothesis tests and constructing simultaneous confidence intervals, procedures for univariate cases are usually generalized and magnified for multivariate situations. Important assumptions on the population sample for application of ANOVA include the following:

1. data (standardized or otherwise transformed) are normally distributed;
2. distances of variances are equal (condition of homogeneity of variances or homoscedasticity);
3. no significant interactions exist between variables;
4. group means and standard deviations are independent (i.e. the size of the group means is not related to the size of their standard deviations);
5. data contain no gross outliers (outliers may be excluded from analysis upon plausibility checks);
6. number of observations in different categories (cells) are equal (not obligatory).

If after standardization or transformation conditions 1) to 3) are not met, data can be analysed by defining alternative multivariate general linear models or hypotheses can be tested by specifying nonparametric models. If significant interactions among variables are suspected (e.g. the influence of the level of a given plant compound on the level of its chemical derivative), analysis of co-variance is carried out to adjust or remove the variability in the dependent variable due to the covariate.

When the homogeneous variance part of the assumptions is false, it is sometimes possible to adjust the degrees of freedom to result in approximately distributed *F* statistics. In SYSTAT, a procedure based on Levene's test for unequal variances, allows to save residuals and perform an ANOVA on the transformed absolute values of the residuals, merged with the original grouping variables. If the test is significant, separate variance tests in the GLM module can be performed.

Although generalized from two-way procedures, it is invalid to perform multivariate hypothesis testing on all possible pairs of hypotheses.

$$\left.\begin{aligned} H_0&: \mu_1 = \mu_2 \\ H_0&: \mu_2 = \mu_3 \\ H_0&: \mu_1 = \mu_3 \end{aligned}\right\} \quad \underline{\text{invalid}}$$

$$H_0: \mu_1 = \mu_2 = \mu_3 \qquad \textbf{valid}$$

The above example is for hypothesis testing of a means model: Alternatively, hypothesis testing on an effects model would read:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$$

The null hypothesis

The null hypothesis H_0 assumes that there is no difference between population means. When comparing alkaloid contents in leaf samples from two different sites, the null hypothesis assumes that the contents are approximately equal (i.e., not a significantly different). The *F*-test is used to calculate whether the null hypothesis must be accepted or rejected and which confidence level is reached. For example, $p<0.05$ represents a 95% confidence limit.

If the null hypothesis of equal group means is true, then the mean squares (MS) of the groups and the errors of the MS will each be an estimate of the common variance σ^2. However, if the population means are not equal, then the groups' MS in the population will be greater than the populations' error MS. Mean square is the mean

squared deviation from the population mean, and the sum of squares is the summation of these.

$$F = \frac{\textbf{groups' MS}}{\textbf{error MS}}$$

If the calculated *F*-ratio is at least as large as the critical value, then H_0 is rejected and the alternative hypothesis H_A (population means being unequal) accepted.

The critical value is computed from the degrees of freedom involving the total sum of squares (SS) of the overall mean and the sums of squares of the group means.

A typical computer output of a two factor analysis (A and B) tabulates the source of variation (singly: A, B, and interactions: A x B), the sum of squares (SS), degrees of freedom (df), mean squares (MS), the *F* statistics and the *p* value. From this result it can, however, not be determined which groups differ from which other groups. To examine specific pairwise group differences, post hoc testing is used. Bonferroni, Scheffé and Tukey tests are available in most statistical packages to test pairwise differences in multi-way designs. When the number of groups is small, the Bonferroni procedure is recommended. For more groups, the Tukey test yields more sensitive results.

Linear and quadratic contrasts

Contrasts are used to test relationships among means. A contrast is a linear combination of means μ_i with coefficients α_i. Typically, the hypothesis takes the following form:

$$H_0 = \alpha_1\mu_1 + \alpha_2\mu_2 + ...+ \alpha_k\mu_k + = 0$$

The test statistics for a contrast is similar to that for a two-sample *t* - test. The result of the contrast (a certain relation among means) appears in the numerator of the test statistics, and an estimate of within-group variability (the pooled variance estimate or the error term from the ANOVA) is part of the denominator. Specific contrast coefficients can be selected to test for example the following:

- pairwise comparison for testing the difference between two particular means;
- linear combinations of means (e.g. two treatment means *vs.* a control mean); or
- linear or quadratic increases or decreases of a certain quality in response to different categories of treatment.

Block and repeated-measures experimental designs

Imagine we would want to compare the alkaloid accumulation patterns in the leaves of a certain plant species under three different soil conditions (factor A) and two light regimes (factor B). Within each of the three levels of factor A, we sampled seven individuals (or blocks) with an observation for each individual at each of the two levels of factor B. The total variability would be divided into two parts: the variability among blocks and the variability within blocks (due to individual behaviour). Hypotheses testing for **Factor A** would take the following form:

H_0: Mean alkaloid content of leaves is the same for all three soil types.

H_A: Mean alkaloid content of leaves is not the same for all three soil types.

$$F = \frac{\textbf{soil types MS}}{\textbf{blocks within soil types MS}}$$

For **Factor B**:

H_0: Mean alkaloid content of leaves is the same under two light regimes.

H_A: Mean alkaloid content of leaves is not the same under two light regimes.

$$F = \frac{\textbf{light regimes MS}}{\textbf{light regimes x (blocks within soil types MS)}}$$

For A x B interaction:

H_0: Mean alkaloid content is independent of light regimes.

H_A: Mean alkaloid content is not independent of light regimes.

$$F = \frac{\textbf{light regimes x soil types MS}}{\textbf{light regimes x (blocks within soil types MS)}}$$

Repeated measurements may be taken at different time intervals to quantify changes over time. In repeated measures design, the same variable is measured several times for each subject. A paired-comparison *t* - test is the most simple form for this design (e.g. before and after measure). The following steps are involved to manually calculate a *t* statistics:

- For each subject the difference between two measures is computed;
- The average of the differences is calculated;
- The standard deviation of the differences is calculated;
- The test statistics using this mean and standard deviation is calculated as shown below:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}i} - \bar{X}2}$$

Changes are tested within subjects and between subjects. Tests of the within-subjects values are called polynominal tests of order 1,2,...*k*., where *k* is one less than the number of repeated measures. The first polynominal is used to test linear changes (e.g. do the repeated measures increase or decrease around a line with a significant slope?), the second tests if the responses fall along a quadratic curve, etc.

Types of models

1. If the levels of a factor are specifically chosen or predetermined, the design is a *fixed-effects model* or *Model I* Anova.
2. If we are interested in general differences between different categories, and samples are taken truly randomly, then we have *random-effects model* or *Model II* Anova.
3. If we have a factorial design with both, fixed and random effects, the model is called *mixed effects model* or *Model III* Anova.

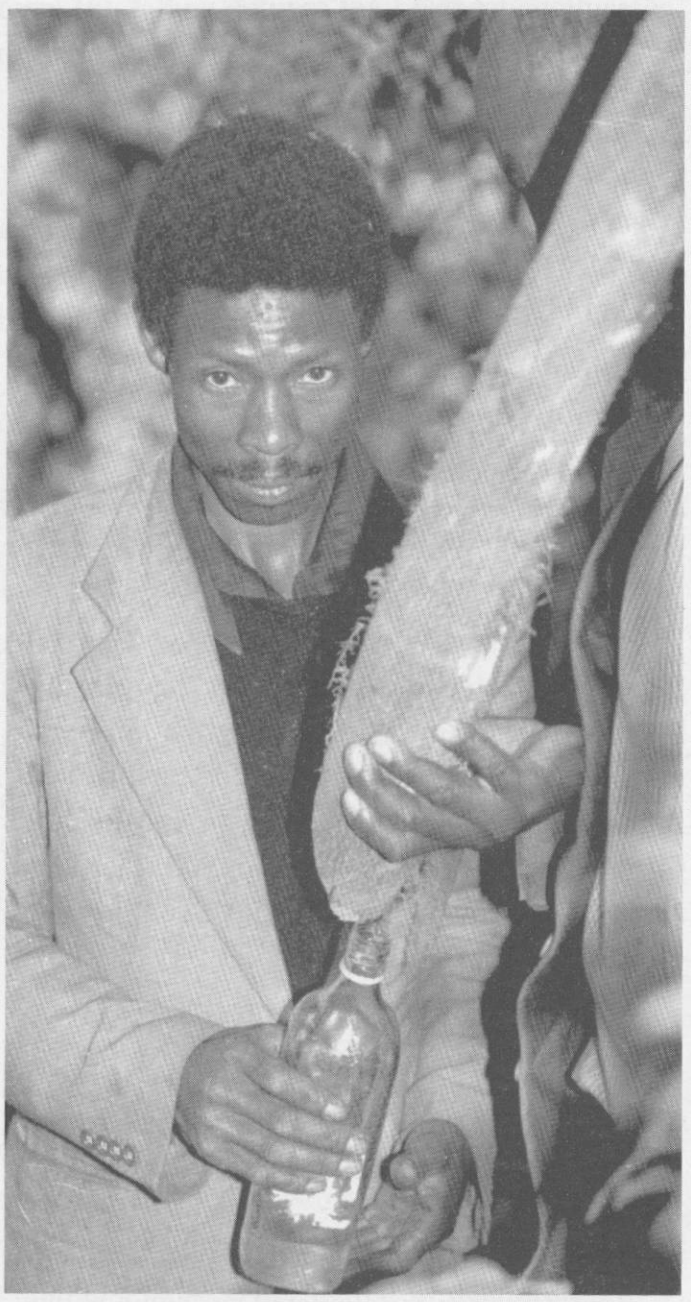


Photo 10. Stephen Weru from Gatei, Kenya, collects the sap of '*mwaritha*', *Dalbergia lactea* Vatke (Fabaceae), used locally to cure hepatitis and other liver ailments. Beyond its medicinal value this liana is also used for the production of durable tea baskets which has led to overharvesting in tea areas.

GENERAL LINEAR MODEL (GLM)

Specific general linear models (*means* or *effects* models) can be defined using the general linear model option, available in standard statistical computer packages, if the data design is not fully factorial, i.e. if numbers of observations are unequal in categories or 'missing cells' occur. The analysis is not robust to violation of normal distribution and equal distances of variance. The latter can be tested using e.g. Levene's Test. The means model takes the following form:

Y = constant + A + B + C + A x B x C

In a means model, predictors are coded as cell means, while in the classic effects model effects are coded as differences from the grand mean. Box 10 lists some examples of data suitable for analysis of variance.

KRUSKAL-WALLIS TEST

The multivariate analogue to the univariate Mann-Whitney Rank Sum test is the Kruskal Wallis Rank Sum Test. The Kruskal Wallis test is also referred to as "analysis of variance by ranks" and is applied when data do not meet any of the six conditions listed on page 22.

Box 10. Examples of data suitable for analysis of variance and Kruskal-Wallis Test.

- Quantifiable effects of herbal medicines as dependant on site, growing season, prepatation procedure, etc.
- Bioassay testing of ethno-medical recipes.
- Validating and quantifying the described effect of an ethno-medical recipe in different user groups.
- Assessment of quantifiable ethnobotanical knowledge as dependent on age, gender, ethnicity or other social factors.
- Storability of grains and grain quality as dependant on the quality of granaries.
- Life span of beehives as dependent on storage conditions during rainy season.
- Effect of specific agricultural methods (e.g. soil working methods, burning, mechanical treatment of fruit trees) on yields.
- Effect of reduced harvesting schemes on the regeneration potential of wild plant populations.
- Effect of harvesting season for raw material on quality of baskets or other household items.

As in parametric analysis of variance it can, however, not be concluded which groups differ from which other groups. The only inference to be drawn is that at least one difference among the groups exists. The test is called nonparametric because no population parameters are used in the statement of hypotheses, and neither parameters nor sample statistics are used in the test calculations. Examples of data suitable for analysis of variance and Kruskal-Wallis test are listed in Box 10 (page 24).

Prediction

As for analysis of variance, two underlying assumptions with respect to the distribution of values must be true for regression analysis:

1) data must come from an approximately normal distributed population;
2) variances must be equal.

MULTIVARIATE REGRESSION ANALYSIS

The relationship between two variables may be one of *functional* dependence of one variable on the other. The magnitude of one variable may thus be a function of the magnitude of the second variable, whereas the reverse is not true.

Regression analysis is a statistical method for predicting values of one or more response (i.e. dependent) variables from a collection of predictor or explanatory (i.e. independent) variable values (Poole, 1974; Zar, 1996). In a simple linear regression analysis, a linear model is developed from which the values of a dependent (i.e. response) variable can be predicted based on particular values of a single independent variable. The **population** regression model is expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

where:

β_0 = is the true intercept, a constant factor in the regression model representing the expected or fitted value of Y when $X = 0$;

β_1 = the true slope representing the amount that Y changes (either positively or negatively) per unit change in X;

ε_i = the random error or residual in Y for observation i.

Since the entire population can not be measured, it is not possible to compute the parameters β_0 and β_1 and obtain the population regression model. Therefore, the approximations b_0 (for β_0) and b_1 (for β_1) are generally estimated from the sample using the methods of least squares. With this method the statistics b_0 and b_1 are computed from the sample in such a manner that the best possible fit within the constraints of the least squares model is achieved. Thus, the following **sample** regression equation is obtained, in which the residual does not figure:

$$Y = b_0 + b_1 X_1$$

Multivariate models from samples can be considered as the extensions of univariate model described above. In multiple regression at least two independent variables (X_1 , X_2) are used to predict the value of a dependent variable (Y). As in the case of simple linear regression model, when sample data are analysed, the sample regression coefficients (b_0 , b_1 and b_2) are used as estimates of the true parameters (β_0 , β_1 and β_2). Thus the **sample** regression equation for the multiple linear regression model with two independent variables would be:

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

Both the models described above focus on linear relationships between the variables. However, in nature nonlinear relationships are quite often encountered. One of the common nonlinear relationships between two variables is a curvilinear polynomial wherein Y increases (or decreases) at a changing rate for various values of X. The second degree polynomial relationship between X and Y may be expressed by the following model:

$$Y = b_0 + b_1 X_1 + b_2 X^2_2$$

where:

b_0 = Y intercept;

b_1 = linear effect on Y;

b_2 = curvilinear effect on Y.

In addition to the above, interaction terms involving the product of independent variables also contribute to the multiple regression model. When two such independent variables are involved, the model is:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$$

Regression models are also developed by transforming the values of the independent variables, the dependent variable or both, depending upon the situation.

If a reciprocal transformation is applied to the values of each of the independent variables, the multiple regression model would be:

$$Y = b_0 + b_1 (1/X_1) + b_2 (1/X_2)$$

A logarithmic transformation would result in a model:

$$Y = b_0 + b_1 \ln X_1 + b_2 \ln X_2$$

On square root transformation, the model would be:

$$Y = b_0 + b_1 \sqrt{X_1} + b_2 \sqrt{X_2}$$

In some situations, the transformations can be applied to change what appear to be nonlinear models into linear models. For example, the multiplicative model :

$$Y = b_0 X_1^{\,b1} X_2^{\,b2}$$

can be transformed (by taking natural logarithms of both sides of the equation) to:

$$ln\ Y = ln\ b_0 + b_1 \ln X_1 + b_2\ ln\ X_2$$

which is linear in logarithms. Similarly, the exponential model:

$$Y = \exp [b_0 + b_1 X_1 + b_2 X_2]$$

can also be transformed into one of linear form (by taking natural logarithms of both sides of the equation):

$$ln\ Y = b_0 + b_1 X_1 + b_2 X_2$$

For detailed methods on multivariate regression analysis, Johnson & Wichern (1988) may be consulted. Box 11 lists some examples of ethnobotanical data suitable for regression analysis.

Box 11. Examples of data suitable for regression analysis.

- Relationship between fuelwood consumption and household size;
- Relationship between demand and harvesting activities for wild plant species;
- Relationship between tree diameter and bark yield;
- Relationship between distance of residence from forest and amount of forest products collected per unit of time;
- Relationship of distance to nearest health service and percentage of reliance on traditional medical practitioners.

Linear correlation

Correlations are calculated when variables can not be designated as being either X (independent) or Y (dependant). Generally, correlations are computed between properties or quantifiable acts of the same individual which are not connected by cause and effect (see Box 12 for examples). An ethnobotanical example would be to analyse the relationship between age and daily amounts of firewood collected by women in a given area.

The simple correlation coefficient is calculated as:

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

A positive correlation implies that an increase in the value of one variable is accompanied by an increase in the second variable, while a negative correlation implies that a decrease in one variable is accompanied by an increase in the second variable. The coefficient of determination or 'correlation index' r^2 may be described as a measure of how much the variability of one variable is accounted for by correlating it to the second variable. R^2 and not r may be considered as a measure of strength of the straight line relationship. Correlation indices below 0.4 or larger than –0.4 may be regarded as indices of weak correlations.

No statistical assumptions need to be satisfied in order to compute correlation coefficients. However, X and Y values are assumed to stem from a bivariate normal population. Sometimes only one of two variables come from a bivariate normal population and data may be transformed to alter this situation. If, like in most ethnobotanical applications neither variable comes from a normal population, rank correlations come into play. Two widely known methods have been proposed by Spearman and Kendall, respectively.

SPEARMAN'S RANK CORRELATION COEFFICIENT

Instead of the actual values, the ranks of the measurements of each variable are used in computing Spearman's rank correlation coefficient. The correlation index is also referred to as Spearman's ρ :

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

where:

$$d_i = \text{rank of } X_i - \text{rank of } Y_i.$$

The value r_s may range from -1 to $+1$ and has no units. Instead of using differences between ranks of pairs, the sums of the ranks of each pair can be used:

$$r_s = \frac{6 \sum S_i^2}{n^3 - n} - \frac{7n - 5}{n - 1}$$

where S_i = rank of X_i + rank of Y_i.

If identical values repeatedly occur for measurements of X or Y variables, then the variables are *tied*, Goodman-Kruskal's γ, Kendall's τ-*b* and Stuart's τ-*c* are used to calculate correlation coefficients for tied data. The methods differ only in the way ties are treated. "Tied" means that the same value for an observation occurs more than once in a column. Tied values must be corrected for.

KENDALL'S TAU-*B* RANK CORRELATION COEFFICIENT

According to Kendall, the correction for tied X and Y data, respectively, is computed:

$$\Sigma\tau_{X \text{ or } Y} = \frac{\Sigma\,(\tau_i^{\,3} - \tau_i)}{12}$$

where τ_i is the number of tied variables (X or Y).

If $\Sigma\tau_X$ and $\Sigma\tau_Y$ are zero, then the equation is identical as for Spearman's rank correlation. The performances of the Spearman and Kendall coefficients are very similar. However, Spearman is recommended when n becomes large.

ANALYSIS OF VARIANCE OF MULTIPLE CORRELATION

A situation where the Y variable is associated with more than one X variable calls for multiple correlation analysis. As in multiple regression analysis, the hypothesis that no interrelationships exist among the variables is tested by:

$$F = \frac{\textbf{regression MS}}{\textbf{residual MS}}$$

$$H_0 = \beta_1 = \beta_2 = ...\beta_k = 0$$

The coefficient of multiple determination is:

$$r^2 = 1 - \frac{\textbf{regression SS}}{\textbf{total SS}}$$

In the case of correlation, r^2 may be considered to be the amount of variability in any of the variables that is accounted for by correlating it with another variable. A measure of goodness of fit is the adjusted coefficient of determination:

$$r_a^{\,2} = 1 - \frac{\textbf{residual MS}}{\textbf{total MS}}$$

or

$$r_a^{\,2} = 1 - \frac{\mathbf{n-1}}{\mathbf{n-m-1}}\,(1-R^2)$$

In contrast to r^2 which increases at each X_j added to the model, $r_a^{\,2}$ will increase only if an added X_j results in an improved fit of the regression equation. The square root of the multiple determination coefficient is referred to as the multiple correlation coefficient. R is equal to the Pearson correlation coefficient, r.

$$r = \sqrt{r^2}$$

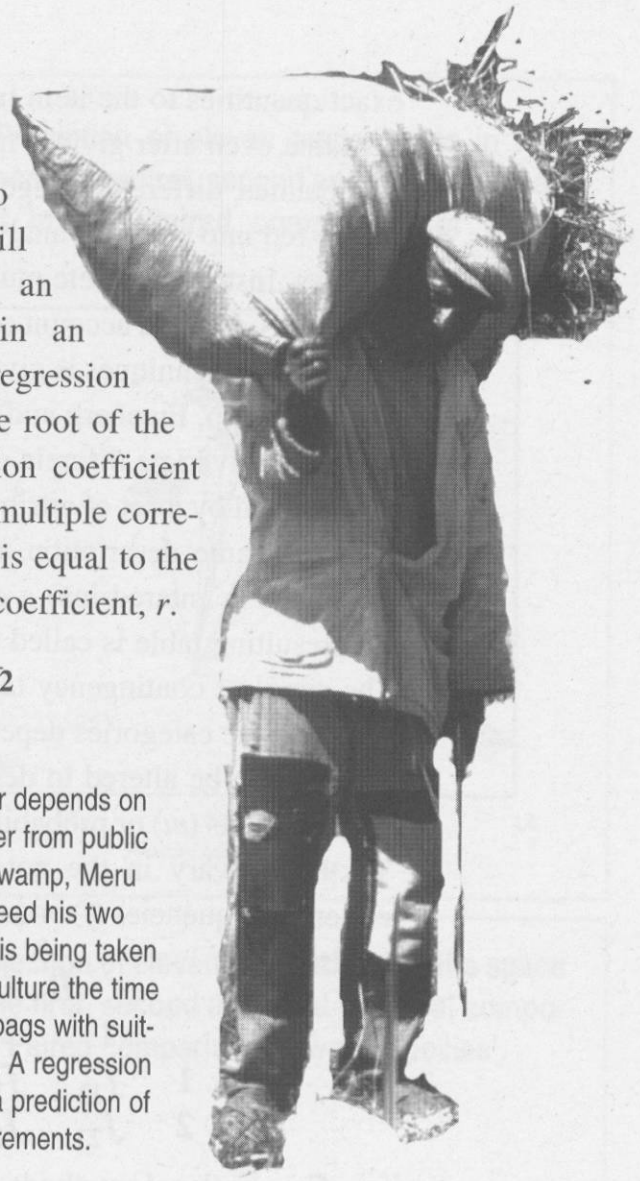


Photo 11. This farmer depends on the collection of fodder from public land around Rurie Swamp, Meru District, Kenya, to feed his two cows. As public lands is being taken by individuals for agriculture the time he requires to fill two bags with suitable fodder increases. A regression analysis would allow a prediction of future time requirements.

Box 12. Examples of data suitable for correlation analysis.

- Correlation between seed size and awn length in sorghum plants.
- Correlation between degree of sclerenchymatization (formation of sclerenchyme cells) and drought resistance in wild cereals.
- Correlation between age and knowledge.

Cross-tabulation

When variables are categorical, frequency tables (cross-tabulations) provide useful summaries. Categories may be unordered (e.g. forest, fallow, garden), ordered (high, medium, low) or formed by defining intervals on a continuous variable such as age (e.g. child, teen, adult and elderly). Such tables can be exploited in three ways:

1. purely descriptive, e.g. calculating percentages of cases falling in specified categories of cross-classifications;
2. test of independence or measure of association between two categorical variables;
3. model relationships among two or more categorical variables by fitting a log-linear model to the cell frequencies.

In a number of cases, typically arising from questionnaires, people may reply by categorizing dependant variables rather than by assigning

exact quantities to the item in question. It is evident that even after giving 'numbers' later to the so obtained different categories, such data can not be fed into standard analysis of variance procedures. Instead, discrete multivariate analysis is applied. A detailed account of discrete multivariate analysis techniques is given in Agresti (1996) and in Bishop, Fienberg and Holland (1995).

In multivariate discrete data, each individual is described by a set of attributes. All individuals with the same description are enumerated, and this count is entered into a cell of a cross-table. The resulting table is called a contingency table. The simplest contingency table is based on four cells and the categories depend on two variables. Models can be altered to describe either expected cell counts (m) or probabilities in each cell (p). Textbooks vary in the notation of cell counts either as frequencies (f) or (m). In this paper f will be used:

		B	
		1	**2**
A	**1**	f_{11}	f_{12}
	2	f_{21}	f_{22}

Conventional methods for analysing relationships between two (or more) discrete variables are χ^2 (Pearson chi-square) and G^2 (Goodman or likelihood-ratio chi-square) tests. The traditional χ^2 test has been used since the beginning of the twentieth century and is described in detail by Zar (1996). For ethnobotanical data, however, the G^2 test is recommended because of better additive properties, i.e. the results of several G^2 tests can be summed. (Sokal & Rohlf 1995; Tabachnick & Fidell 1996). The general structure of a contingency table is depicted in Box 13.

Box 13. General appearance of contingency tables

Variable A	Variable B			Totals
	category 1	category 2	category n	
category 1				
category 2				
category 3				
category n				
Totals				**Grand total**

Different tests and measures exist for different table structures and also depend on whether or not the categories of the variables are ordered. The Pearson and likelihood-ratio chi-square statistics apply to larger than 2 x 2, i.e. r x c tables and categories need not be ordered. Other tests include the following:

- **McNemar's test of symmetry:** applies for square tables where the number of rows equals the number of columns. This structure arises when the same subject is measured twice (as in a paired comparison t test) or when subjects are paired or matched (e.g. cases and controls). In such a design, the categories of rows and columns are the same, but they are measured at different times or under different circumstances or for different groups of subjects. This test ignores the the counts along the diagonal of the table and tests whether the counts above the diagonal differ from those below the diagonal. A significant result indicates a greater change in one direction than another.
- **Phi, Cramer's V, and contingency:** like Pearson's chi-square, these are measures for testing independence of variables in a table. They are applied to tables with unequal sample size. The three measures are scaled differently but test the same null hypothesis. For tables with two rows and two columns, phi and Cramer's V are the same.
- **Goodman-Kruskal's gamma, Kendall's tau-*b*, Stuarts's tau-*c*, Spearman's rho and Somer's *d*:** these are appropriate when both row and column variables have ordered categories. The first three differ only in how ties are treated, the fourth is like the usual Pearson correlation except that the the rank order of each value is used. Somer's d is an asymmetric measure (in SYSTAT, the column variable is considered 'dependent').
- **Fisher's exact test and Yates's corrected chi-square:** these are used specifically in the analysis of 2 x 2 contingency tables with small sample sizes.
- **Yule's Q and Yule's Y:** these measure dominance in a 2 x 2 table. If either cell off the diagonal is 0, both statistics are equal (otherwise they are less than 1). These statistics are 0 only if the chi-square statistics is 0.

Because of their wide range of applicability, Pearson chi-square and likelihood-ratio chi-square are presented in more detail in the following paragraphs.

CHI-SQUARE ANALYSIS OF CONTINGENCY TABLES

With two variables (A, B) under consideration, the observed frequency is denoted as f_{ij}, whereby i refers to rows and j to columns of the

contingency table. For the computation of chi-square, only the expected frequencies are used, never proportions or percentages.

The total number of observations in all cells of the table, the grand total is $\Sigma^{row}_{i=1}\Sigma^{column}_{j=1} f_{ij}$ or simply $\Sigma\Sigma\, f_{ij}$. For chi-square analysis of contingency tables, the following formula is used:

$$\chi^2 = \Sigma\Sigma \frac{(\hat{f}_{ij} - f_{ij})^2}{\hat{f}_{ij}}$$

or (mathematically equivalent):

$$\chi^2 = \Sigma\Sigma \frac{f_{ij}^{\,2}}{\hat{f}_{ij}} - 1$$

$$\chi^2 = n\left(\Sigma\Sigma \frac{f_{ij}^{\,2}}{R_i C_j} - 1\right)$$

In the above formulas, $\hat{f}_{ij}$ refers to the frequency expected in a row i column j if the null hypothesis is true. To calculate expected values in row i and column j, the Pearson χ^2 statistics is used:

$$\frac{\textbf{row } i \textbf{ total} * \textbf{column } j \textbf{ total}}{\textbf{grand total}}$$

or

$$\hat{f}_{ij} = \frac{(R_i)\,(C_j)}{n}$$

The χ^2 test is intended to be used when the values of observations are small enough for sampling variation to leave some doubt as to the interpretation of data (Box 14). The lower limit to the sample size for which the method is sufficiently reliable is 5 for the expected value, although the actually observed value may be lower. If some cells in the table show expected values smaller than 5, the respective rows or columns may be merged into one combining two features.

Apart from demonstrating a significant association among variables by a contingency table, one may want to assess the strength of that association. A measure of association is:

$$\chi^2 / n$$

where n is the total number of observations.

LIKELIHOOD-RATIO CHI-SQUARE OR GOODMAN SQUARE ANALYSIS OF CONTINGENCY TABLES

The likelihood-ratio chi-square or Goodman square test statistics is additive for nested models. Two models are nested if all the effects of the first are a subset of the second. The likelihood-ratio chi-square is additive because the statistics for the second model can be subtracted from that of the first. The difference provides a test of the additional effects, i.e. the difference in the two statistics has an asymptotic chi-square distribution with degrees of freedom equal to the difference between those for the two model chi-squares (or the difference between the number of effects in the two models). This property does not hold for the Pearson chi-square. Goodman chi-square statistics are calculated as following:

$$G^2 = 2\ln L$$
$$= 2\,\Sigma\,(\text{observed})\,\ln(\text{observed} / \text{expected})$$

The summation is done over all cells in the table. Both chi-square and Goodman test statistics are used to examine interactions among variables by testing for goodness of fit for the observed frequency distribution to the expected frequency distribution representing the H_0. The greater the departure of the actual value from the expected value, the greater are the values of G^2 and χ^2.

Both test statistics are distributed as χ^2. If the χ^2 or G^2 value is greater than the tabulated χ^2 at the desired α-level, then the null hypothesis of no interaction between variables is rejected.

Frequency tests are non-parametric as there are no assumptions made on the underlying population distribution. There are, however, requirements for independent samples, adequate sample size, and the minimum size of expected frequency in each cell. Problems may occur if there are too few cases relative to the number of categories and the following guidelines should be borne in mind (after Tabachnick & Fidell 1996):

- the number of cases should be at least five times the number of cells;
- the expected cell frequency should be >1 in all cells and <5 in no more than 20% of the cells;.
- if the contingency table contains smaller expected frequencies, Fishers exact test is recommended.

Box 14. Examples of data suitable for Pearson chi-square or Goodman square analysis.

- Different user groups using different plant parts.
- Different user groups gaining different amounts of money from the sale of their products.
- Effectiveness of a drug (plant species) to combat a certain ailment.
- Number of incidents in which a plant is mentioned to cure a certain ailment.

Applications of general linear models

The theory of the previous chapter is exemplified in four case studies:

- The effects of soil nutrient levels and light supply on alkaloid accumulation of *Tabernaemontana pachysiphon* Stapf (Apocynaceae) were studied under experimental conditions;
- Data on diameter, bark weight and bark thickness were sampled from *Rytigynia* trees, valued for their anthelmintic properties around Bwindi Impenetrable Forest;
- A correlation was calculated to examine whether a relationship exists between the age of women collecting firewood from Ragati forest, Mt. Kenya, and the average weight of their headloads;
- The preferences of women and men specialists and non-specialists for harvesting of medicinal plants in five different habitat types were investigated around Bwindi Impenetrable National Park.

Analysis of variance

An example for analysis of variance using the general linear model option in SYSTAT has been chosen to demonstrate the effects of three levels of fertilizer and two light intensities on the accumulation of the alkaloid apparicine in leaves of *Tabernaemontana pachysiphon*. The output of a full factorial (all interactions included and tested automatically) analysis of variance model is shown in Box 15. Alternatively, the general linear model option in SYSTAT and SAS would allow to manually specify the interactions one wants to test. The means model is also used to test hypotheses when missing cells are encountered or to test hypothesis about specific cell means. No significant interactions between fertilizer and light intensity affect the accumulation of apparicine ($p = 0.48$). Thus, the effect of either factor can be interpreted independently. The Kolmogorov-Smirnov one-sample test is used to compare the shape and location of a sample distribution with a uniform, normal or chi-square distribution. The Lilliefors test uses standardized variables (mean of zero and standard deviation of 1) and tests whether data are normally distributed. A probability of $p > 0.05$ indicates normal distribution of standardized data.

Box 15. Analysis of variance computation using the general linear model option (data from Höft, 1995).

Effects coding used for categorical variables in model.

Categorical values encountered during processing are:
F(ertilizer)$ (3 levels)
F0, F1, F2
L(ight)$ (2 levels)
L-, L+

Dep Var: Apparicine N: 36
Multiple R: 0.70 Squared multiple R: 0.49

Analysis of Variance

Source	SS	df	MS	F-ratio	P
F$	426.40	2	213.20	4.49	0.02
L$	1047.85	1	1047.85	22.09	0.00
F$*L$	71.94	2	35.97	0.76	0.48
Error	1423.09	30	47.44		

Durbin-Watson D Statistics 1.971
First Order Autocorrelation 0.009

Means Model

Dep Var: Apparicine N: 36
Multiple R: 0.70 Squared multiple R: 0.49

H_0: All means equal.

Unweighted Means Model
Analysis of Variance

Source	SS	df	MS	F-ratio	P
Model	1394.75	5	278.95	5.88	0.00
Error	1423.09	30	47.44		

Durbin-Watson D Statistics 1.971
First Order Autocorrelation 0.009

COL/ROW	F$	L$
1	F0	L-
2	F0	L+
3	F1	L-
4	F1	L+
5	F2	L-
6	F2	L+

Box 15. continued

Using unweighted means.
Post Hoc test of Apparicine

Using model MSE of 47.436 with 30 df.
Matrix of pairwise mean differences:

	1	2	3	4	5	6
1	0.00					
2	-7.39	0.00				
3	10.08	17.48	0.00			
4	-1.23	6.16	-11.31	0.00		
5	8.15	15.54	-1.94	9.38	0.00	
6	-6.24	1.15	-16.32	-5.01	-14.39	0.00

Tukey HSD Multiple Comparisons.
Matrix of pairwise comparison probabilities:

	1	2	3	4	5	6
1	1.00					
2	0.36	1.00				
3	0.16	0.00	1.00			
4	1.00	0.56	0.08	1.00		
5	0.43	0.01	1.00	0.28	1.00	
6	0.59	1.00	0.01	0.78	0.03	1.00

Kolmogorov-Smirnov One Sample Test using Normal (0.00,1.00) distribution

Variable	N-of-Cases	MaxDif	Lilliefors Probability (2-tail)
Apparicine	36.00	0.10	0.43

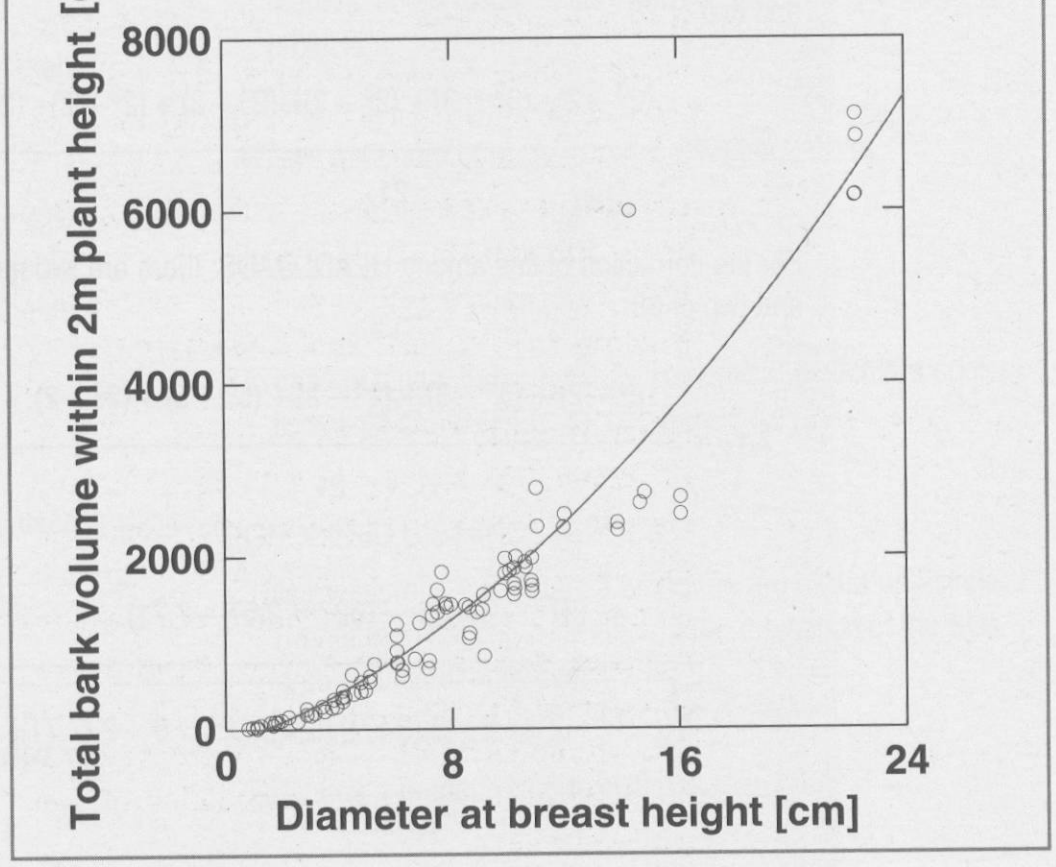
Regression analysis

Data on bark volume and diameter of *Rytigynia kiwuensis* (K. Krause) Robyns (Rubiaceae) were evaluated using regression analysis. A quadratic model was fitted to predict the total bark volume of the bottom 2 metres of *Rytigynia* plants from diameter measurements between 1 and 24 cm (from Kamatenesi 1997). The relationship is depicted in Figure 10. Bark volumes within reach for harvesting from a standing tree (i.e. up to 2 m height) can be predicted with sufficient accuracy from diameter measurements according to the following formula:

$$Y = -105.75 + 99.11\ X + 8.88\ X^2$$
$$(r^2 = 0.795)$$

Likewise, a quadratic regression model can be fitted from the trees' diameters to predict the total bark dry weight of the bottom 2 metres of the plant (see graph on cover page).

Fig. 10. Relationship between total bark volume ot the bottom 2 metres of the plant and diameter at breast height in *Rytigynia kiwuensis*, Rubiaceae (data from Kamatenesi 1997).



Correlation

In the following example it was tested whether there is a significant correlation between the age (X) and the amount of firewood (Y) collected by 21 women in Ragati forest, Mt. Kenya. Since data are not normally distributed Spearman's rank correlation coefficient (r_s) wss calculated. The null hypothesis of no correlation takes the form H_0: $\rho = 0$.

Box 16. Computation of Spearman's rank correlation coefficient (r_s) using age and fire wood collection data [kg/d] of women in Ragati Forest, Mt. Kenya (data from Wanja Waiganjo, 1999).

Age (X)	Headload(Y)	rank of X	rank of Y	d_i	d_i^2
40	53	8	18.5	-10.5	110.25
41	51	9	16	-7	49
50	51	13.5	16	-2.5	6.25
44	71	10.5	21	-10.5	110.25
38	51	6.5	16	-9.5	90.25
28	40	2.5	7	-4.5	20.25
50	44	13.5	11	2.5	6.25
55	41	18	10	8	64
30	37	5	2	3	9
51	45	16	12.5	3.5	12.25
38	40	6.5	7	-0.5	0.25
44	38	10.5	3.5	7	49
28	53	2.5	18.5	-16	256
29	40	4	7	-3	9
45	60	12	20	-8	64
70	50	20.5	14	6.5	42.25
65	40	19	7	12	144
51	38	16	3.5	12.5	156.25
70	35	20.5	1	19.5	380.25
51	45	16	12.5	3.5	12.25
23	40	1	7	-6	36

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

$$r_s = 1 - \frac{6\,(1627)}{9240}$$

$$r_s = -0.056$$

$\bar{n} = 21$

$\sum d_i^2 = 1627$

To test H_0: $(r_s)_{0.05,\,21} = 0.370$ (critical value)

$r_s < 0.370$, therefore do not reject H_0 ($p = 0.05$)

For the correction of ties among AGE: there are two measurements of 70, three of 51, two of 50, two of 44, two of 38 and two of 28

$$\Sigma\tau_X = \frac{(2^3-2)+(3^3-3)+(2^3-2)+(2^3-2)+(2^3-2)+(2^3-2)}{21} = 2.57$$

For the correction of ties among HEADLOADS: there are two measurements of 53, three of 51, two of 45 five of 40 and two of 38.

$$\Sigma\tau_Y = \frac{(2^3-2)+(3^3-3)+(2^3-2)+(5^3-5)+(2^3-2)}{21} = 7.71$$

$$(r_s)c = \frac{(21^3-21)/6 - 1627 - 2.57 - 7.71)}{\sqrt{[(21^3-21)/6 - 2\,(2.57)]\,[(21^3-21)/6 - 2\,(7.71)]}} = \frac{-5.66}{3246.36} = -0.0017$$

To test H_0: $(r_s)_{0.05,21} = 0.370$ (critical value)

$(r_s)c < 0.370$, therefore do not reject H_0 ($p = 0.05$)

The Greek rho represents the correlation coefficient. Box 16 (page 32) details major steps involved in hypothesis testing and special attention is paid to the correction of data which appear in a column more than once ("tied data"). The "critical value" can be obtained from statistical tables. Standard computer packages calculate these steps automatically. If the calculated correlation coefficient of the actual data, r_s is smaller than the critical value the null hypothesis of no correlation must be accepted. In this example there is thus no significant correlation between age and average amount of firewood collected.

Chi-square analysis of contingency tables

Chi-square analysis was applied to test whether or not male and female herbalists prefer different habitat types for the collection of medicinal plants (see Table 3). The frequency of mention is shown in a two-dimensional contingency table and the expected frequencies are indicated in brackets below the actual observations. A typical computer output (SYSTAT) is given in Box 17 (see Appendix, Table 23, page 42 for a more detailed output of analysis results). Log-linear modelling is applied to predict cell frequencies in multi-way tables. Some theoretical background on the analysis of multi-dimensional contingency tables and the prediction of cell frequencies is given in the Appendix (page 44.)

Alternatively, Table 3 can be analysed in a different way. The values for female and male professionals and for female and male non-professionals can be combined to yield two groups: female and male users. These groups can be tested on paired combinations of habitats. Furthermore, pairwise comparisons can be made between female and male professionals and female and male non-professionals for each possible pair of habitats.

Table 3. Two-dimensional contingency table with two variables (habitats and professional category) showing the frequency of mention of the habitat preference for the collection of medicinal plants around Bwindi Impenetrable Forest (data from Kyoshabire 1998).

	Habitat type					
Professional category	Garden	Early fallow	Old fallow	Bushland	Forest	**N**
Traditional birth attendants (f)	17	42	46	39	40	**184**
(expected)	(15.5)	(38.1)	(40.5)	(42.3)	(47.8)	
Women herbalists (f)	15	42	43	43	36	**179**
(expected)	(15.0)	(37.1)	(39.4)	(41.1)	(46.4)	
Herbalists (m)	24	47	54	59	74	**258**
(expected)	(21.7)	(53.5)	(56.7)	(59.3)	(66.9)	
Non-specialists (m)	4	17	14	23	35	**93**
(expected)	(7.8)	(19.3)	(20.5)	(21.4)	(24.1)	

Box 17. Cross-tabulation statistics for Table 3.

Test statistics	**Value**	**df**	**p (probability)**
Pearson Chi-square	17.34	12	0.14
Likelihood ratio Chi-square	17.59	12	0.13

Coefficient	**Value**	**Asymptotic SE (standard error)**
Phi	0.16	
Cramer V	0.09	
Contingency	0.15	
Goodman-Kruskal Gamma	-0.01	0.04
Spearman Rho	-0.01	0.04
Lambda (column dependent)	0.02	0.02
Uncertainty (column dependent)	0.01	0.00

References

Agresti, A. 1996. An introduction to categorical data analysis. John Wiley & Sons, New York.

Berenson, M.L., Levine, D.M. & Goldstein, M. 1983. Intermediate statistical methods and applications. A computer package approach. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. 1995. Discrete multivariate analysis. Theory and practice.The Massachusetts Institute of Technology, USA.

BMDP Statistical Software Inc. 1999. BMDP Statistical Software. Los Angeles, California, USA.

Bray, J.R. & Curtis, J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monographs. 27: 325-349.

Casgrain, P. 1999. R-package Version 4. University of Montreal, Canada.

Causton, D.R. 1988. Introduction to vegetation analysis. Unwin Hyman, London.

Fischer, H. & Bemmerlein, F. 1986. Numerische Methoden in der Ökologie. Unpublished Manuscript. University of Erlangen, Germany.

Gnanadesikan, R. 1977. Methods for statistical data analysis of multivariate observations. John Wiley & Sons, New York, USA.

Hill, M.O. 1979. TWINSPAN, a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Section of Ecology and Systematics, Cornell University, Ithaca, New York.

Höft, M.G. 1995. Aut-ecological studies on *Tabernaemontana pachysiphon* Stapf and *Rauvolfia mombasiana* Stapf (Apocynaceae) in the Shimba Hills (Kenya) with special reference to their alkaloid contents. Bayreuther Forum Ökologie, vol. 17, Bayreuth, Germany.

Johnson, R.A. & Wichern, D.W. 1988. Applied multivariate statistical analysis. Prentice - Hall, Inc., Englewood Cliffs, New Jersey.

Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R. 1987. Data analysis in community and landscape ecology. Pudoc, Wageningen.

Kachigan, S.K. 1986. Statistical analysis: an interdisciplinary introduction to univariate and multivariate methods . Radius Press.

Kamatenesi, M.M. 1997. Utilization of the medicinal plant 'nyakibazi' (*Rytigynia* spp.) in the multiple use zones of Bwindi Impenetrable National Park, Uganda. Unpublished M.Sc. thesis. Makerere University, Uganda.

Kent, M. & Coker, P. 1994. Vegetation design and analysis. A practical approach. Wiley, Chichester.

Kyoshabire. M. 1998. Medicinal plants and the herbalist preferences around Bwindi Impenetrable Forest, Uganda. Unpublished M.Sc. thesis. Makere University, Uganda.

Lance, G. N. & Williams W. T. 1967. A general theory for classificatory sorting strategies. 1. Hierarchical systems. Computer Journal 9: 373-380.

Legendre, P. & Legendre, L. 1998. Numerical ecology. Elsevier Scientific Publishing Company, Amsterdam.

Legendre, P. & Vaudor, A. (1991). The R Package: Multidimensional analysis, spatial analysis. Département de sciences biologiques, Université de Montréal, Montreal.

Ludwig, J.A. & Reynolds, J.F. 1988. Statistical ecology. A primer on methods and computing. John Wiley & Sons, New York.

MjM Software Design. 1999. PC-ORD for Windows Version 4. Melillo Consulting Inc., Somerset, New Jersey, USA.

Orlóci, L. 1978. Multivariate analysis in vegetation research. Dr. W. Junk b. v., Publishers, The Hague.

PC-ORD. Multivariate analysis of ecological data. MjM Software Design, Glenden Beach.

Peters, C.M. 1996. Beyond nomenclature and use: a review of ecological methods for ethnobotanists. In: M.N. Alexiades (ed.), Selected Guidelines for Ethnobotanical Research. The New York Botanical Garden: 241-276.

Phillips, O.L.B. & Gentry, A.H. 1993a. The useful plants of Tambopata, Peru. I: Statistical hypothesis tests with a new quantitative technique. Economic Botany. 47: 15-32.

Phillips, O.L.B. & Gentry, A.H. 1993b. The useful plants of Tambopata, Peru. II: Additional hypothesis testing in quantitative ethnobotany. Economic Botany. 47: 15-32.

Pielou, E.C. 1984. The interpretation of ecological data. A primer on classification and ordination. Wiley, New York.

Poole, R.W. 1974. An introduction to quantitative ecology. McGraw-Hill, Kosakusha Ltd., Tokyo.

Prance, G.T. 1991. What is ethnobotany today? Journal of Ethnopharmacology. 32: 209-216.

Rohlf, F.J. 1985. NTSYS - Numerical Taxonomy System of Multivariate Statistical Programs. Cornell Theory Center Software Documentation, Ithaca, New York, USA.

Romesburg, H.C. 1984. Cluster analysis for researchers. Wadsworth Inc., Lifetime Learning, Belmont, California.

SAS Institute Inc. 1999. SAS for Windows, Version 6.12. SAS Institute Inc., Cary, North Carolina, USA.

Sneath, P.H.A. & Sokal, R.R. 1973. Numerical taxonomy. Freeman, San Francisco, CA.

Sokal, R.R. & Rohlf, F.J. 1995. Biometry. Freeman, New York.

SPSS Inc. 1999. SPSS Version 10. SPSS Inc. Chicago, Illinois, USA.

SPSS Inc. 1998. Systat Version 8. SPSS Inc. Chicago, Illinois, USA.

Tabachnick, B.G. & Fidell, L.S. 1996. Using multivariate statistics. Harper Collins, New York.

Ter Braak, C.J.F. 1988a. CANOCO - a Fortran program for canonical community ordination by (partial) (detrended) (canonical) correspondence analysis, principal component analysis and redundancy analysis. Agricultural Mathematics Group, Wageningen.

Ter Braak, C.J.F. 1988b. CANOCO - an extension of DEDORANA to analyze species-environment relationships. Vegetation 75: 159-160.

Wanja Waiganjo, F. 1999. Forest plants used in Ragati, Mt. Kenya: their taxonomy, exploitation, economic values and conservation status. Unpublished M.Sc. thesis, Kenyatta University, Kenya.

Zar, J. H. 1996. Biostatistical Analysis. Third edition, Prentice- Hall International Inc., New Jersey.

Further reading:

Bailey, N.T. 1981. Statistical methods in biology. Hodder and Stoughton. London, Sydney, Auckland, Toronto.

Digby, P.G.N. & Kempton, R.A. 1987. Multivariate analysis of ecological communities. Chapman & Hall, London.

Feoli, E. & Orlóci, L. 1991. Computer assisted vegetation analysis. In: H. Lieth: Handbook of vegetation science. Kluwer, Dordrecht.

Gauch, H.G. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge.

Greig-Smith, P. 1983. Quantitative plant ecology. Blackwell Scientific Publishers, Oxford.

Krebs, C.J. 1989. Ecological methodology. Harper & Row, New York.

Krzanowski, W.J. & Mariott, H.H.C. 1994. Multivariate analysis. Arnold, London.

Acknowledgements

Having a manuscript on statistical methods for ethnobotanists has been a wish expressed by many students supported over the years by the People and Plants Initiative. The current working paper evolved from a training course on quantitative methods in ethnobiology which took place from 20 August to 1 September in Nairobi and Kilifi, Kenya. This Course was held essentially by Javier Caballero Nieto and Gary Martin and used the Kenyan woodcarving industry as an example. From the material developed during the Course S.K. Barik wrote a manual on cluster analysis. This was complemented by a discussion of the application of several multivariate and statistical methods using various kinds of data sets, mostly from work supported by People and Plants.

The authors wish to thank Remigius Bukenya-Ziraba, Tony Cunningham, Jeremy Midgley and John Tabuti for their invaluable suggestions and comments on various versions of this Working Paper. Malcolm Hadley, Timothy Johns and Ebi Kimanani carefully read the manuscript and made useful suggestions for improvement. Nevertheless, we realize that this document still has shortcomings and possibly some errors which may have been overlooked and the authors assume the full responsiblity for those.

The multidisciplinary Sahel-Sudan Environmental Research Initiative (SEREIN) of the Danish International Development Agency DANIDA through financial support enabled Anne Mette Lykke to contibute to this publication.

Appendix

Table 4. Basic data matrix for the 'Wood identification task'.

Species	Respondents															
	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}
S_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S_2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
S_3	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	0
S_4	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
S_5	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0
S_6	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	0
S_7	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
S_8	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0

Table 5. Basic data matrix for the 'Paired comparison of wood species'.

Species	Respondents															
	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}
S_1	1	1	1	2	2	1	2	1	1	1	1	1	1	4	1	2
S_2	5	5	5	5	4	5	4	5	5	2	5	5	5	5	5	5
S_3	3	3	1	2	3	4	5	4	4	2	4	4	4	1	3	3
S_4	2	2	1	1	1	2	1	3	2	5	2	2	2	1	2	1
S_5	4	4	1	4	4	3	3	2	3	2	3	3	3	1	3	4

Tables 4 through 14 and figures 11 and 12 detail the results of matrix transformations applied in cluster analysis. The necessary transformation steps are summarized in Box 8 (page 18).

Table 4 presents the responses (yes/no) of 16 woodcarvers who were asked to identify pieces of wood from eight different species. To demonstrate the following transformations, the last three columns and rows of the basic wood identification data matrix (Table 4) have been taken separately to constitute a smaller 3 x 3 data matrix (Table 7). The matrix represents the ability of the respondent (R_{14}...R_{16}) to identify the wood species (S_6...S_8). In a normal mode analysis the degree of similarity of the responses must be determined for each pair of respondents (i.e. columns of the data matrix). In this example, R_{14} and R_{16} could identify only one species i.e. S_7, while R_{15} could identify all the three species. Hence, R_{14} and R_{16} are more similar to each other than R_{15}.

In order to compute the similarity index it is necessary to be familiar with the concept and terms relating to an association or similarity contingency table. Table 8 shows a typical contingency table using binary (yes/no coded as 1/0) data for two respondents (R_{14} and R_{15}). The case that both respondents recognize the species occurs one time, the case that none of them recognizes the species occurs zero times, etc. A resemblance matrix for all respondents is subsequently built from the 2 x 2 similarity contingency tables (Table 9).

Table 6. The standardized data matrix

	R_1	R_2	R_3	R_4	R_5
S_1	0.538	0.538	0.538	0.691	0.691
S_2	0.394	0.394	0.394	0.394	0.867
S_3	0.109	0.109	1.852	0.980	0.109
S_4	0.122	0.122	0.854	0.854	0.854
S_5	1.065	1.065	1.942	1.065	1.065

Table 7. A 3 x 3 data matrix constructed from the wood identification data matrix (Table 4) by taking its last 3 rows and columns for demonstration of the computation of similarity functions.

Species	Respondents R_{14}	R_{15}	R_{16}
S_6	0	1	0
S_7	1	1	1
S_8	0	1	0

Table 8. A 2 x 2 contingency similarity table based on the data from Table 7 for the Respondents R_{14} and R_{15}.

		Respondent R_{15} Yes	No	
Respondent R_{14}	Yes	1	0	1 + 0 = 1
	No	2	0	2 + 0 = 2
		1 + 2 = 3	0 + 0 = 0	1 + 0 + 2 + 0 = 3

Table 9. Resemblance matrix based on Jaccard's index (measure of similarity) for 'Wood identification task' data matrix. The matrix is derived from the basic data matrix (Table 4).

R_1	1.000														
R_2	1.000	1.000													
R_3	1.000	1.000	1.000												
R_4	1.000	1.000	1.000	1.000											
R_5	1.000	1.000	1.000	1.000	1.000										
R_6	0.875	0.875	0.875	0.875	0.875	1.000									
R_7	0.875	0.875	0.875	0.875	0.875	0.750	1.000								
R_8	0.750	0.750	0.750	0.750	0.750	0.625	0.625	1.000							
R_9	0.750	0.750	0.750	0.750	0.750	0.625	0.625	1.000	1.000						
R_{10}	0.625	0.625	0.625	0.625	0.625	0.714	0.500	0.571	0.571	1.000					
R_{11}	1.000	1.000	1.000	1.000	1.000	0.875	0.875	0.750	0.750	0.625	1.000				
R_{12}	0.875	0.875	0.875	0.875	0.875	0.750	1.000	0.625	0.625	0.500	0.875	1.000			
R_{13}	0.625	0.625	0.625	0.625	0.625	0.500	0.500	0.833	0.833	0.429	0.625	0.500	1.000		
R_{14}	0.625	0.625	0.625	0.625	0.625	0.500	0.714	0.571	0.571	0.429	0.625	0.714	0.667	1.000	
R_{15}	1.000	1.000	1.000	1.000	1.000	0.875	0.875	0.750	0.750	0.625	1.000	0.875	0.625	0.625	1.000
R_{16}	0.375	0.375	0.375	0.375	0.375	0..250	0.429	0.500	0.500	0.143	0.375	0.429	0.600	0.600	0.375	1.000

for the 'Paired comparison of wood species' data matrix.

R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}
0.538	0.691	0.538	0.538	0.538	0.538	0.538	0.538	3.148	0.538	0.691
0.394	0.867	0.394	0.394	3.388	0.394	0.394	0.394	0.394	0.394	0.394
0.763	1.634	0.763	0.763	0.980	0.763	0.763	0.763	1.852	0.109	0.109
0.122	0.854	1.098	0.122	3.050	0.122	0.122	0.122	0.854	0.122	0.854
0.063	0.063	0.939	0.063	0.939	0.063	0.063	0.063	1.942	0.063	1.065

Table 10. Resemblance matrix based on Bray-Curtis Index (measure of dissimilarity) for the 'Paired comparison of wood species'. The resemblance matrix was derived from the standardized data matrix (Table 6).

R_1	0.000															
R_2	0.000	0.000														
R_3	0.000	0.000	0.000													
R_4	2.46	2.461	0.000	0.000												
R_5	4.027	4.027	0.000	8.824	0.000											
R_6	1.078	1.078	0.000	4.423	7.315	0.000										
R_7	3.878	3.878	0.000	4.964	4.627	2.949	0.000									
R_8	2.250	2.250	0.000	6.339	10.398	1.251	4.371	0.000								
R_9	1.078	1.078	0.000	4.423	7.315	0.000	2.949	1.251	0.000							
R_{10}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000						
R_{11}	1.078	1.078	0.000	4.423	7.315	0.000	2.949	1.251	0.000	0.000	0.000					
R_{12}	1.078	1.078	0.000	4.423	7.315	0.000	2.949	1.251	0.000	0.000	0.000	0.000				
R_{13}	1.078	1.078	0.000	4.423	7.315	0.000	2.949	1.251	0.000	0.000	0.000	0.000	0.000			
R_{14}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		
R_{15}	1.157	1.157	0.000	16.471	0.000	1.184	8.691	4.016	1.184	0.000	1.184	1.184	1.184	0.000	0.000	
R_{16}	1.039	1.039	0.000	0.580	1.133	2.049	2.161	3.083	2.049	0.000	2.049	2.049	2.049	88.112	2.865	0.000

The tree matrix in Table 11 shows similarities linearly transformed into distances. Since the tree does not exactly represent the data matrix, a correlation coefficient is calculated to get some measure of matching degree between tree and resemblance matrix. This relationship is graphically depicted in Figures 11 and 12 (Page 39). Tables 13 and 14 show the results of cophenetic matching between each pair of respondents.

Table 11. Tree matrix for the 'Wood identification task' data. The clustering was performed on the resemblance matrix presented in Table 9.

R_1	1.000
R_2	1.000
R_3	1.000
R_4	1.000
R_5	1.000
R_{15}	1.000
R_{11}	0.875
R_6	0.859
R_7	1.000
R_{12}	0.671
R_8	1.000
R_9	0.833
R_{13}	0.624
R_{14}	0.578
R_{10}	0.405
R_{16}	-----

Table 12. Tree matrix for the 'Paired comparison of wood species' data. The clustering was performed on the resemblance matrix presented in Table 10.

R_1	0.0000000
R_2	0.0000000
R_{10}	0.0000000
R_{14}	0.5390000
R_6	0.0000000
R_{13}	0.0000000
R_9	0.0000000
R_{11}	0.0000000
R_{12}	1.1950000
R_8	3.3638500
R_5	0.0000000
R_{15}	4.2995208
R_3	0.0000000
R_4	0.2900000
R_{16}	2.3750000
R_7	---

Figure 11. Matrix comparison plot showing the relationship between cophenetic matrix and resemblance matrix in the 'Wood identification' task.

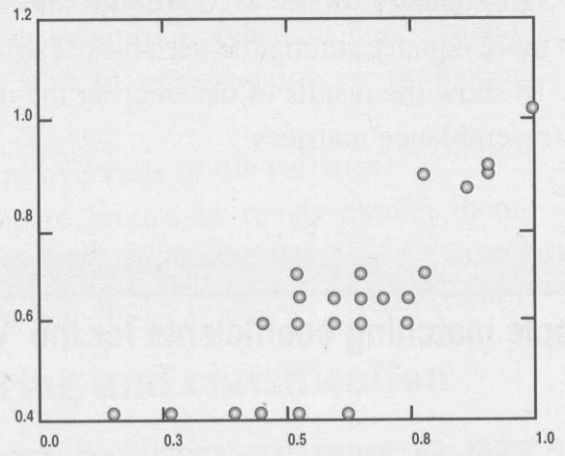


Figure 12. Matrix comparison plot showing the relationship between cophenetic matrix and resemblance matrix in 'Paired comparison of wood species'.

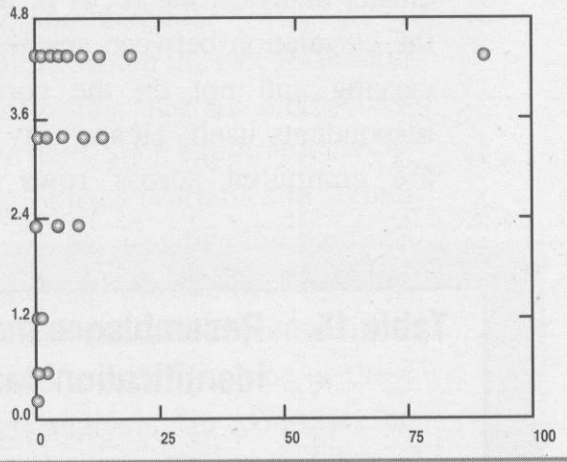


Table 13. Cophenetic matrix for the 'Wood identification task' data. The cophenetic matrix was derived from the tree matrix presented in Table 11.

R_1	1														
R_2	1.000	1													
R_3	1.000	1.000	1												
R_4	1.000	1.000	1.000	1											
R_5	1.000	1.000	1.000	1.000	1										
R_6	0.875	0.875	0.875	0.875	0.875	1									
R_7	0.859	0.859	0.859	0.859	0.859	0.859	1								
R_8	0.671	0.671	0.671	0.671	0.671	0.671	0.671	1							
R_9	0.671	0.671	0.671	0.671	0.671	0.671	0.671	1.000	1						
R_{10}	0.578	0.578	0.578	0.578	0.578	0.578	0.578	0.578	0.578	1					
R_{11}	1.000	1.000	1.000	1.000	1.000	0.875	0.859	0.671	0.671	0.578	1				
R_{12}	0.859	0.859	0.859	0.859	0.859	0.859	1.000	0.671	0.671	0.578	0.859	1			
R_{13}	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.833	0.833	0.578	0.671	0.671	1		
R_{14}	0.624	0.624	0.624	0.624	0.624	0.624	0.624	0.624	0.624	0.578	0.624	0.624	0.624	1	
R_{15}	1.000	1.000	1.000	1.000	1.000	0.875	0.859	0.671	0.671	0.578	1.000	0.859	0.671	0.624	1
R_{16}	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405	0.405

Table 14. Cophenetic matrix for the 'Paired comparison of wood species' data. The cophenetic matrix was derived from the tree matrix presented in Table 12.

R_1	0.000														
R_2	0.000	0													
R_3	4.300	4.300	0												
R_4	4.300	4.300	0.000	0											
R_5	3.364	3.364	4.300	4.300	0										
R_6	0.539	0.539	4.300	4.300	3.364	0									
R_7	4.300	4.300	2.375	2.375	4.300	4.300	0								
R_8	1.195	1.195	4.300	4.300	3.364	1.195	4.300	0							
R_9	0.539	0.539	4.300	4.300	3.364	0.000	4.300	1.195	0						
R_{10}	0.000	0.000	4.300	4.300	3.364	0.539	4.300	1.195	0.539	0					
R_{11}	0.539	0.539	4.300	4.300	3.364	0.000	4.300	1.195	0.000	0.539	0				
R_{12}	0.539	0.539	4.300	4.300	3.364	0.000	4.300	1.195	0.000	0.539	0.000	0			
R_{13}	0.539	0.539	4.300	4.300	3.364	0.000	4.300	1.195	0.000	0.539	0.000	0.000	0		
R_{14}	0.000	0.000	4.300	4.300	3.364	0.539	4.300	1.195	0.539	0.000	0.539	0.539	0.539	0	
R_{15}	3.364	3.364	4.300	4.300	0.000	3.364	4.300	3.364	3.364	3.364	3.364	3.364	3.364	3.364	0
R_{16}	4.300	4.300	0.290	0.290	4.300	4.300	2.375	4.300	4.300	4.300	4.300	4.300	4.300	4.300	4.300

Tables 15 through 22 detail the matrix transformations necessary for principal component analysis. The steps involved are summarised in Box 9 (page 20). In contrast to cluster analysis, the focus in this exercise is on the correlation between species used for woodcarving and not on the correlation between respondents itself. Hence, any derived matrices are computed across rows and not across columns. Computation of the resemblance matrices for paired comparison of wood species (Tables 15 and 16) is carried out on the standardised data matrix (Table 6). Decentering is subsequently done to distribute the distances more equally among the variables. Tables 17 and 18 show the results of decentering the respective resemblance matrices.

Table 15. Resemblance matrix based on simple matching coefficients for the 'Wood identification' task data.*

S_1	1.0000000							
S_2	0.9375000	1.0000000						
S_3	0.7500000	0.8125000	1.0000000					
S_4	0.9375000	0.8750000	0.6875000	1.0000000				
S_5	0.8125000	0.8750000	0.8125000	0.7500000	1.0000000			
S_6	0.7500000	0.8125000	0.6250000	0.6875000	0.8125000	1.0000000		
S_7	0.8750000	0.8125000	0.6250000	0.9375000	0.6875000	0.6250000	1.0000000	
S_8	0.6250000	0.6875000	0.8750000	0.6875000	0.8125000	0.6250000	0.6250000	1.0000000

*The analysis is across the rows and was performed on the basic data matrix at Table 3.

Table 16. Resemblance matrix based on variance-covariance across the rows for the 'Paired comparison of wood species' data. The resemblance matrix was derived from the standardized data matrix (Table 6).

S_1	1.0000000				
S_2	0.0193613	1.0000000			
S_3	-0.4193588	0.1190259	1.0000000		
S_4	-0.4895840	-0.6254190	0.0708739	1.0000000	
S_5	-0.2103228	0.1421269	0.4148225	-0.2037374	1.0000000

Table 17. The transformed matrix after decentering of the resemblance matrix (Table 15) for the 'Wood identification task'.

S_1	0.123							
S_2	0.045	0.092						
S_3	-0.064	-0.018	0.248					
S_4	0.076	-0.002	-0.111	0.154				
S_5	-0.049	-0.00	0.014	-0.096	0.154			
S_6	-0.033	0.014	-0.096	-0.080	0.045	0.311		
S_7	0.061	-0.018	-0.127	0.139	-0.111	-0.096	0.248	
S_8	-0.158	-0.111	0.154	-0.080	0.045	-0.064	-0.096	0.311

Table 18. The transformed matrix after decentering of the resemblance matrix (Table 16) for the 'Paired comparison of wood species'.

S_1	1.145				
S_2	0.014	0.843			
S_3	-0.531	-0.144	0.631		
S_4	-0.315	-0.601	-0.011	1.205	
S_5	-0.313	-0.112	0.055	-0.277	0.648

Eigen-values are then computed for the principal components (sources of variation for each species) to arrive at some measure of variance for a particular principal component (Tables 19 and 20). Eigen-vectors pertain to the variables (species) itself and define the principal component (Tables 21 and 22). From Tables 19 and 20 it can be seen that the first three principal components account for 85.13% and 89.39% (cumulative percentages) of the variances in the 'Wood identification' and the 'Paired comparison of wood species' tasks, respectively. Principal components accounting for the remaining percentages can be left aside in further analysis. Tables 21 and 22 show the 'loading' (Eigen-vectors) that each species is assigned by each of the three most important principal components. A higher (more positive) loading indicates a greater importance of the principal component for the respective species i. In both examples the loading by the principal component 1 is high for species 1 and 2.

Table 19. Eigen-values (λ) for the principal components for the data on the 'Wood identification task'. The values are derived from the decentred matrix presented in Table 17.

i	Eigen value	Percent	Cumulative
1	0.70775	43.14	43.14
2	0.45658	27.83	70.97
3	0.23233	14.16	85.13
4	0.12500	7.62	92.75
5	0.06686	4.08	96.82
6	0.04334	2.64	99.47
7	0.00877	0.53	100.00
8	0.00000	0.00	100.00

Table 20. Eigen-values (λ) for the principal components for the data on 'Paired comparison of wood species'. The values are derived from the decentred matrix presented in Table 18.

i	Eigen-value	Percent	Cumulative
1	1.84713	41.30	41.30
2	1.42966	31.96	73.26
3	0.72162	16.13	89.39
4	0.47447	10.61	100.00
5	0.00000	0.00	100.00

Table 21. Eigen-vector matrix (U) with the loading of each character in each principal component for the data on 'Wood identification task'.

i	PC1	PC2	PC3
1	0.2695094	-0.0062328	-0.1723919
2	0.1000234	-0.1082403	-0.2080325
3	-0.3723312	0.1829196	-0.2211452
4	0.3116364	0.1593744	0.0754053
5	-0.2026852	-0.1634770	-0.0267113
6	-0.0330392	-0.5098736	0.1677226
7	0.3784826	0.2194557	0.1632280
8	-0.4515962	0.2260740	0.2219250

Table 22. Eigen-vector matrix (U) with the loading of each character in each principal component for the data on the 'Paired comparison of wood species'.

i	PC1	PC2	PC3
1	0.7290292	-0.7280290	0.2209689
2	0.6063664	0.3678820	-0.5407599
3	-0.3844371	0.4585559	-0.0635704
4	-0.8926225	-0.5732191	-0.1974509
5	-0.0583360	0.4748102	0.5808122

Table 23 details steps in the constructing of the log-linear model. While the example below is two-dimensional, log-linear models may be used to identify structures of multi-dimensional nature in ethnobotanical data. Contingency tables with three or more dimensions (three or more variables) are often analysed by employing log-linear models. The term 'model' is an expression for how the observed frequencies (or counts) are affected by variables and combinations of variables. 'Log-linear' refers to a procedure whereby a multiplicative relationship is transformed to a linear relationship by the use of logarithms. In the terminology of log-linear models, interactions of variables are tested. A null hypothesis for no interactions between variables implies that all the variables are independent. If the null hypothesis is rejected, the variables are said to be associated. In the example below professional category and habitat preference are highly associated.

Table 23. Output for log-linear model estimates of Table 3.

(data after Kyoshabire, 1999; see text page 33).

Deviations = Observed-Expected

==========

CATEGORY	HABITAT				
	Garden	Early fallow	Old fallow	Bushland	Forest
Traditional birth attendants	1.546	3.86	5.54	-3.26	-7.78
Women herbalists	-0.04	4.90	3.64	1.89	-10.38
Male specialists	-0.26	-6.48	-2.73	-0.26	7.15
Male non-specialists	-3.82	-2.28	-6.45	1.64	10.90

Standardized Deviates = (Obs-Exp)/sqrt(Exp)

==========

CATEGORY	HABITAT				
	Garden	Early fallow	Old fallow	Bushland	Forest
Traditional birth attendants	0.39	0.63	0.87	-0.50	-1.11
Women herbalists	-0.01	0.80	0.58	0.29	-1.52
Male specialists	0.50	-0.89	-0.36	-0.03	0.87
Male non-specialists	-1.36	-0.52	-1.43	0.35	2.22

Pearson Chi-square = (Obs-Exp)^2/Exp

==========

CATEGORY	HABITAT				
	Garden	Early fallow	Old fallow	Bushland	Forest
Traditional birth attendants	0.15	0.39	0.76	0.25	1.24
Women herbalists	0.0	0.65	0.34	0.09	2.32
Male specialists	0.25	0.78	0.13	0.00	0.77
Male non-specialists	1.86	0.27	2.03	0.13	4.93

Likelihood Ratio Deviance = 2*(Exp-Obs+Obs*log(Obs/Exp))

==========

CATEGORY	HABITAT				
	Garden	Early fallow	Old fallow	Bushland	Forest
Traditional birth attendants	0.15	0.38	0.73	0.26	1.31
Women herbalists	0.0	0.62	0.33	0.09	2.52
Male specialists	0.24	0.82	0.13	0.00	0.74
Male non-specialists	2.27	0.28	2.29	0.12	4.32

Table 23. continued.

Freeman-Tukey Deviates = sqrt(Obs)+sqrt(Obs+1)-sqrt(4*Exp+1)

CATEGORY	HABITAT				
	Garden	Early fallow	Old fallow	Bushland	Forest
Traditional birth attendants	0.44	0.65	0.88	-0.47	-1.12
Women herbalists	0.05	0.81	0.60	0.33	-1.57
Male specialists	0.53	-0.88	-0.33	-0.00	0.88
Male non-specialists	-1.44	-0.47	-1.48	0.40	2.05

Contribution to log(likelihood) = -Exp+Obs*log(Exp)-log(gamma(Obs+1))

CATEGORY$	HABITAT$				
	Garden	Early fallow	Old fallow	Bushland	Forest
Traditional birth attendants	-2.41	-2.98	-3.20	-2.88	-3.42
Women herbalists	-2.28	-3.10	-2.96	-2.84	-3.97
Male specialists	-2.63	-3.26	-2.98	-2.96	-3.44
Male non-specialists	-2.77	-2.48	-3.39	-2.55	-4.86

Log-Linear Effects (Lambda)

THETA 3.44

CATEGORY$

Traditional birth attendants	Women herbalists	Male specialists	Male non-specialists
0.09	0.07	0.43	-0.59

HABITAT$

Garden	Early fallow	Old fallow	Bushland	Forest
-0.80	0.10	0.16	0.21	0.33

Standard Error of Lambda

THETA 3.44

CATEGORY$

Traditional birth attendants	Women herbalists	Male specialists	Male non-specialists
0.07	0.07	0.06	0.08

HABITAT$

Garden	Early fallow	Old fallow	Bushland	Forest
0.11	0.08	0.07	0.07	0.07

Table 23. continued.

Lambda / SE(Lambda)

==========

THETA 3.44

CATEGORY$			
Traditional birth attendants	Women herbalists	Male specialists	Male non-specialists
1.41	0.99	7.24	-7.05

HABITAT$				
Garden	Early fallow	Old fallow	Bushland	Forest
0.11	1.37	2.20	2.83	4.67

Multiplicative Effects = exp(Lambda)

==========

THETA 31.33

CATEGORY$			
Traditional birth attendants	Women herbalists	Male specialists	Male non-specialists
1.10	1.07	1.07	1.54

HABITAT$				
Garden	Early fallow	Old fallow	Bushland	Forest
0.45	1.11	1.18	1.23	1.39

Model ln(MLE): -61.373

==========

Term tested	The model without the term				Removal of term from model		
	ln(MLE)	Chi-Sq	df	p-value	Chi-Sq	df	p-value
CATEGORY$	-101.862	98.57	15	0.0000	80.98	3	0.0000
HABITAT$	-100.124	95.09	16	0.0000	77.50	4	0.0000

The following explanations start with a two-dimensional design (after Bishop *et al.*, 1995). The logarithm of the relative odds (cross-product ratio α of cells of the contigency table) can also be expressed as the linear contrast of the log-probabilities of the elementary cells:

$$\log \alpha = \log p_{11} - \log p_{12} - \log p_{21} + \log p_{22}$$

A linear model in the natural logarithms of the cell probabilities can be constructed by analogy with analysis of variance models:

$$\log f_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12\,(ij)}$$

$$i = 1,2$$
$$j = 1,2$$

where u is the grand mean of the logarithms of the probabilities:

$$u = 1/4\,(\log p_{11} + \log p_{12} + \log p_{21} + \log p_{22})$$

and $u + u_{1(i)}$ is the mean of the logarithms of the probabilities at level i of the first variable:

$$u + u_{1(i)} = 1/2\,(\log p_{i1} + \log p_{i2}) \qquad i = 1,2$$

Similarly for the jth level of the second variable:

$$u + u_{2(j)} = 1/2\,(\log p_{j1} + \log p_{j2}) \qquad j = 1,2$$

Since $u_{1(i)}$ and $u_{2(j)}$ represent deviations from the grand mean u:

$$u_{1(1)} + u_{1(2)} = u_{2(1)} + u_{2(2)} = 0$$

Similarly, $u_{12(ij)}$ represents a deviation from $u + u_{1(i)} + u_{2(j)}$, so that:

$$u_{12(11)} = - u_{12(12)} = - u_{12(21)} = u_{12(22)}$$

The additive properties imply that each *u*-term has one absolute value for dichotomous variables. By analogy with ANOVA models the grand mean can be written:

$$u = \frac{\log p+ +}{4}$$

In contrast to ANOVA, the interest in log-linear models is to test for interactions of factors, while in ANOVA, the focus is on main effects. Instead of using probabilities, the actual counts or frequencies are refered to as m in the following. A three dimensional model would look like this:

$$m_{ijk} = \exp[(u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}]$$

The direct estimate would read:

$$\hat{m} = \frac{\chi_{i+k} * \chi_{+jk}}{\chi_{++k}}$$

Other notations in publications for log-linear model parameters include the following:

$$\log m_{ij} = \mu + \lambda_i^F + \lambda_j^S + \lambda_{ij}^{FS}$$

$$\xi_{ij} = \Theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

$$G_{ij} = \Theta + \lambda_i^F + \lambda_j^S + \lambda_{ij}^{FS}$$

or in multiplicative form:

$$F_{ij} = \eta r_i^A r_j^B r_{ij}^{AB}$$

where: $\xi = \log(F_{ij})$,
$\Theta = \log\eta$,
$\lambda_i^A = \log(r_i^A)$ *etc.*

Like χ^2 and G^2 tests, log-linear models are based on contingency tables, however, higher order interactions are incorporated. A model involves a linear combination of parameters that are calculated on the basis of the contingency table. The natural logarithm of expected cell frequencies (ln *f*) is described by the following linear function:

$$\ln(f_{ijk}) = \Theta + \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^{AB}_{ij} + \lambda^{AC}_{jk} + \lambda^{BC}_{ik} + \lambda^{ABC}_{ijk}$$

where:

f_{ijk} = expected frequency in row *i*, column *j* and depth *k* of a three-way contingency table;

Θ = is an overall mean effect, calculated of the logarithms of the expected frequencies;

λ = parameters, summing to zero over the levels of the row factors and the column factors.

λ^A_i, λ^B_j and λ^C_k = main effects of the categories i, j and k of variables A, B and C;

λ^{AB}_{ij}, λ^{AC}_{ik} and λ^{BC}_{jk} = second order interaction terms expressing the dependence of a category of one factor on a category of another factor;.

λ^{ABC}_{ijk} = third order interaction term expressing the mutual dependency of all three variables on each other.

For each cell, the logarithm of the expected frequency is the sum of Θ and λ's. Each cell therefore has its own combination of parameters that are used to predict cell frequency. The observed cell frequencies are calculated exactly by a saturated model that contains all main effects and interactions. Three steps are suggested for building the model:

1. screening through frequency analysis: variables that are not found to be significant in simple χ^2 and G^2 tests are excluded from the model;
2. choice of an appropriate model;
3. evaluation and interpretation of the selected model.

One aim of modelling is to determine the minimum of parameters necessary to adequately describe the data set. An incomplete model with fewest effects may be preferred over the complete model, since a smaller number of parameters eases the interpretation of data. Ideally, one uses its knowledge on the subject matter of the study to determine which effects to include in the model. In practice, stepwise exclusion of parameters from log-linear models is done while maintaining an adequate fit of expected to observed cell frequencies. However, if many models are tested in search for the ideal model, one must bear in mind that the *p* value associated with a specific test is valid for one test only. *P* values may be used as relative measures when testing several models. In the end the model includes only those parameters necessary to reproduce the observed frequency.

Testing for significance of a parameter in a model is done by comparing two models, one wich includes the parameter in question and one which excludes it. The G^2 statistic is computed for each model, and the difference between the G^2 values calculated. The calculated difference (a G^2 statistic itself) tests the (goodness of) fit between observed and expected frequencies, and

a good model thus is one where G^2 is not significant and H_0 retained. In order not to have too many good models, Tabachnick and Fidell (1996) propose to use a higher than 0.05 α–significance level, *e.g.* 0.1 or 0.25.

However, testing of hypotheses in log-linear models is not a goal in itself. The purpose of modelling in ethnobotany is to identify interpretable structures in the data and the significance values of hypotheses are used to guide the modelling process.

Types of models

Generally, two types of models exist, hierachical and non-hierachical models. Where significant interactions are present, lower order terms for all possible combinations of variables involved in the interaction must be included in hierachical models. In contrast, non-hierachical models can be build without this restriction.

Hierachical modelling starts with the most complex models. The three-way interaction in a three-way contingency table is tested for first. A non significant G^2 value indicates little evidence for three factor interaction, and the l ABC_{ijk} term is omitted from the model. A significant three-factor interaction term implies that the degree of association between any pair of variables depends upon the different categories of the third variable. In such case reduced models should not be used (Sokal and Rohlf 1995). Analoguous to classical analysis of varicane, lower order effects cannot be interpreted unambiguously if there are higher order effects.

Non-hierachical models are mainly of interest in testing pre-specified models. As there are no clear statistical criteria for choosing among non-hierachical models, they are not recommended for model building in general.

The parameters in the model (*l*'s) represent increase or decrease of m for a particular combinations of row, column and depth variables. Investigation of these parameters yields information about the effect of different categories and interactions on the cell frequencies. A positive *l* for a main effect indicates that the frequency in this category is above average. The interaction parameter indicates how much difference there is between the individually and collectively taken sums of effects of variables. They represent the "boost" or interference associated with a particular combination of variables (SAS).

Effects with larger standardized parameter estimates are more important in predicting a cell's frequency than effects with smaller standardized parameter estimates.

Already published in this series:

1. Cunningham, A. B. 1993. *African medicinal plants: Setting priorities at the interface between conservation and primary healthcare.* (This publication is also available in Spanish.)

2. Cunningham, A. B. and Mbenkum, F.T. 1993. *Sustainability of harvesting* Prunus africana *bark in Cameroon: A medicinal plant in international trade.*

3. Aumeeruddy, Y. 1994. *Local representations and management of agroforests on the periphery of Kerinci Seblat National Park, Sumatra, Indonesia.* (This publication is also available in French and Spanish.)

4. Cunningham, A. B. 1996. *People, park and plant use: Recommendations for multiple-use zones and development alternatives around Bwindi Impenetable National Park, Uganda.* (This publication is also available in French.)

5. Wild, R. and Mutebi, J. 1996. *Conservation through community use of plant resources. Establishing collaborative management at Bwindi Impenetrable and Mgahinga Gorilla National Parks, Uganda.* (This publication is also available in French.)

The People and Plants Initiative

was started in July 1992 by WWF, UNESCO and the Royal Botanic Gardens, Kew to promote the sustainable and equitable use of plant resources through providing support to ethnobotanists from developing countries.

The initiative stems from the recognition that people in rural communities often have detailed and profound knowledge of the properties and ecology of locally occurring plants, and rely on them for many of their foods, medicines, fuel, building materials and other products. However, much of this knowledge is being lost with the transformation of local ecosystems and local cultures. Over-harvesting of non cultivated plants is increasingly common, caused by loss of habitat, increase in local use and the growing demands of trade. Long-term conservation of plant resources and the knowledge associated with them is needed for the benefit of the local people and for their potential use to local communities in other places.

The diversity of traditional plant-resource management practices runs through a spectrum from "cultivation" through to gathering "wild" plants, all of which are included in the People and Plants approach.

Ethnobotanists can work together with local people to study and record the uses of plant resources, identify cases of over-harvesting of non-cultivated plants, find sustainable harvesting methods and investigate alternatives such as cultivation.

The People and Plants initiative is building support for ethnobotanists from developing countries who work with local people on issues related to the conservation of both plant resources and traditional ecological knowledge. Key participants organize participatory workshops, undertake discussion and advisory visits to field projects and provide literature on ethnobotany, traditional ecological knowledge and sustainable plant resource use. It is hoped that a network of ethnobotanists working on these issues in different countries and regions can be developed to exchange information, share experience and collaborate on field projects.

Contact addresses:

WWF International
Plant Conservation Officer
Panda House, Weyside Park
Godalming, Surrey GU7 1XR
UNITED KINGDOM
Fax: 44 1483 426409

WWF

Division of Ecological Sciences
Man and the Biosphere Programme
UNESCO, 7 Place de Fontenoy
75352 Paris Cedex 07 SP
FRANCE
Fax: 33 1 45685804

UNESCO

The Director
Royal Botanic Gardens, Kew
Richmond
Surrey TW9 3AB
UNITED KINGDOM
Fax: 44 181 3325278

ROYAL BOTANIC GARDENS KEW