Combined analysis of categorical and numerical descriptors of australian groundnut accessions using nonlinear principal component analysis

Kroonenberg, P.M.; Harch, B.D.; Basford, K.E.; Cruickshank, A.

Combined Analysis of Categorical and Numerical Descriptors of Australian Groundnut Accessions Using Nonlinear Principal Component Analysis

P.M. KROONENBERG, B.D. HARCH, K.E. BASFORD, and A. CRUICKSHANK

For users of germplasm collections, the purpose of measuring characterization and evaluation descriptors, and subsequently using statistical methodology to summarize the data, is not only to interpret the relationships between the descriptors, but also to characterize the differences and similarities between accessions in relation to their phenotypic variability for each of the measured descriptors.

The set of descriptors for the accessions of most germplasm collections consists of both numerical and categorical descriptors. This poses problems for a combined analysis of all descriptors because few statistical techniques deal with mixtures of measurement types. In this article, nonlinear principal component analysis was used to analyze the descriptors of the accessions in the Australian groundnut collection. It was demonstrated that the nonlinear variant of ordinary principal component analysis is an appropriate analytical tool because subspecies and botanical varieties could be identified on the basis of the analysis and characterized in terms of all descriptors. Moreover, outlying accessions could be easily spotted and their characteristics established.

The statistical results and their interpretations provide users with a more efficient way to identify accessions of potential relevance for their plant improvement programs and encourage and improve the usefulness and utilization of germplasm collections.

Key Words: Genetic diversity; Mixture of data types; Ordinal data; Oleic-linoleic ratio; Ordination; *Arachis hypogaea* L.

1. INTRODUCTION

Germplasm collections contain large numbers of accessions (samples of germplasm material of a crop) on which several characteristics are measured. For users of germplasm collections, the purpose of collecting these measurements, and subsequently using multivariate statistical techniques, is not only to acquire an insight into the relationships between the descriptors, but also to characterize the differences and similarities between accessions in relation to their phenotypic variability.

P.M. Kroonenberg is Associate Professor, Department of Education, Leiden University, Wassenaarseweg 52, Leiden, The Netherlands. B.D. Harch is Statistician, CSIRO, Mathematical & Information Sciences, Private Bag 2, Glen Osmond, SA 5064, Australia. K.E. Basford is Associate Professor, Department of Agriculture, The University of Queensland, Brisbane, Qld 4072, Australia. A. Cruickshank is Peanut Breeder, J. Bjelke-Petersen Research Station, P.O. Box 23, Kingaroy Qld 4610, Australia.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 2, Number 3, Pages 294–312

This article focuses on obtaining information about the phenotypic variability in the Australian groundnut (*Arachis hypogaea* L.) germplasm collection. Eight hundred and thirty-five (835) groundnut accessions were sown during the 1990/1991 growing season, and several descriptors varying in measurement type were recorded. For instance, stem color was binary (green or purple), pod constriction was ordinal or ordered multicategory (absent, slight, moderate, deep, and very deep), and weight per hundred seeds (or 100-seed weight) was numerical. Full details, analyses, and references with respect to the Australian groundnut germplasm collection can be found in Harch (1996; see also Harch et al. 1995; Harch et al. 1996a).

Although Wynne and Coffelt (1982) and Stalker (1989) reported extensive phenotypic variability in the characteristics of *Arachis hypogaea* L., Gregory et al. (1951) and, more recently, Krapovickas and Gregory (1994) have devised a taxonomy for distinguishing the subspecies and botanical varieties of *Arachis hypogaea* L. In the Australian groundnut collection, two subspecies and three botanical varieties of *Arachis hypogaea* L. can be identified:

1. *subspecies hypogaea*
   var. hypogaea (Virginia type: Bunch and Runner)
2. *subspecies fastigiata*
   2.1 *var. fastigiata* (Valencia type)
   2.2 *var. vulgaris* (Spanish type)

Summarizing the phenotypic variability in germplasm data, such as that contained in the Australian groundnut collection, can be undertaken using multivariate statistical techniques. The results from these techniques allow users (e.g., plant breeders) to interpret patterns or the lack of patterns found in the data. These interpretations often involve using either the descriptors that are distinguishing most amongst the accessions or taxonomic information, or both. Together, the summary information and interpretations provide users with a more time-efficient way to identify accessions of potential relevance for their particular plant improvement programs and ultimately encourage and improve the usefulness and utilization of germplasm collections.

As mentioned previously, germplasm collection descriptors have different types or levels of measurement; that is, some of the descriptors are numerical and others are categorical. Although this may pose serious problems for standard multivariate statistical procedures, a relatively new technique, *nonlinear principal component analysis,* is especially geared toward handling datasets in which descriptors have different types of measurement. The statistical theory, methods, algorithms, and programs, as well as the history of the subject, have been fully described in a book by Gifi (1990, chap. 4).

In this article, the Australian groundnut data are explored using nonlinear principal component analysis. The aim is to provide an overall summary of the collection using all descriptors irrespective of their measurement type. Such a unified analysis should give plant breeders a comprehensive overview of the available phenotypic diversity for groundnut accessions (Bretting et al. 1990; Perry and McIntosh 1991; Singh et al. 1991). Given the relative unfamiliarity of the technique, a conceptual introduction into nonlinear principal component analysis is presented to provide sufficient background for understanding the analysis of the groundnut germplasm data.

2. EXPERIMENTAL DETAILS

The Australian groundnut germplasm collection comprises 835 accessions, of which 693 are cultivars and advanced breeding lines and 142 are land races. These accessions were grown in 1990/91 at the J. Bjelke-Petersen Research Station, Kingaroy, Queensland (26° 35′ S and 150°0′ E), in a single replicate, completely random design with grid plot checks. Grid plot check data were not provided for analysis. Details of the plots and growing conditions are outlined in Harch et al. (1995).

Accessions were evaluated for 16 descriptors, including plant characteristics, seed characteristics, and fatty acid composition, following the IBPGR and ICRISAT (1992) groundnut descriptor guidelines. This information is made available to plant breeders and other researchers for use in their breeding programs through the Australian Tropical Field Crops Genetic Resource Center. Details of the Australian groundnut germplasm collection, its objectives, format and use of databases, and the status, location, regeneration, and evaluation of accessions are outlined in Lawrence (1989).

Of the 835 accessions, 831 have been used in this study. The 16 descriptors have been partitioned into three data types: five binary, five ordinal (or ordered multicategory), and six numerical descriptors (Table 1). Details of the descriptor measurements taken are provided in IBPGR and ICRISAT (1992) and the methods used to obtain the fatty acid samples are given in Harch et al. (1995).

3. NONLINEAR PRINCIPAL COMPONENT ANALYSIS

3.1 General Description

Nonlinear principal component analysis is an extension of ordinary principal component analysis to handle descriptors of any measurement type. Thus, the descriptors need not be numerical, but may be categorical (binary, unordered multicategory, or ordered multicategory). The additional generality introduces some complexities in interpretation, but the major principles behind ordinary principal component analysis are maintained. In particular, the first principal component is a new descriptor resulting from a linear combination of the original descriptors, which on its own explains as much of the variation in the descriptors as possible. One way to express this is that the new descriptor should have an average squared correlation with the original descriptors as high as possible. How to achieve this with only numerical descriptors is part of the standard literature on multivariate analysis (e.g., see Joliffe 1986). When some of the descriptors are categorical, the technical complexity to achieve the same goal is considerably increased, but not the basic idea of maximizing the average squared correlation between the descriptors and the component.

The new aspect in nonlinear principal component analysis is that the correlations of the categorical descriptors with the component have to be determined. In nonlinear principal component analysis, this is achieved by assigning numerical values to the categories in a specific way. This assignment of numerical values to the categories of a categorical descriptor is called quantification. For instance, stem color green might be assigned a value of, say, −1.09 and stem color purple might be assigned a value .32.

Table 1. Descriptors Observed From the Australian Groundnut Germplasm Collection (Containing 831 Accessions)

Abbreviation	*Description*	*Category definitions*
Binary descriptors:		
Branch	branching pattern	1=alternate; 2=sequential
Stem	stem pigmentation	1=green; 2=purple
Peg	peg pigmentation	1=absent; 2=present
Petal	petal colour	1=yellow; 2=orange
Sdcol	seed colour	1=non-variegated; 2=variegated
Ordinal descriptors:		
Habit	growth habit	1=procumbent & decumbent1; 2=decumbent2; 3=decumbent3 & erect
Beak	pod beak	1=absent; 2=slight; 3=moderate; 4=prominent; 5=very prominent
Constr	pod constriction	1=absent; 2=slight; 3=moderate; 4=deep & very deep
Retic	pod reticulation	1=absent; 2=slight; 3=moderate; 4=prominent; 5=very prominent
Seeds	most frequent number of seeds per pod	1=1 seed; 2=2 seeds; 3=3 or 4 seeds
Numeric descriptors:		
Shell	shelling percentage (%)	1≤58; 2=58,59; 3=60,61; 4=62,63; 5=64,65; 6; 66,67; 7=68,69; 8=70,71; 9=72,73; 10≥73
Height[†]	estimated plant height (cm; nearest multiple of 5)	1≤25; 2=25; 3=30; 4=35; 5=40; 6=45; 7=50; 8≥50
Width	estimated plant width (cm; nearest multiple of 5)	1≤65; 2=65,70; 3=75,80; 4=85,90; 5=95,100; 6=105,110; 7=115,120; 8≥120
Weight	100-seed weight	1≤30; 2=30 to 40; 3=40 to 50; 4=50 to 60; 5=60 to 70; 6=70 to 80; 7=80 to 90; 8=90 to 100; 9≥100
Oil	oil content (%)	1≤47; 2=47 to 48; 3=48 to 49; 4=49 to 50; 5=50 to 51; 6=51 to 52; 7=52 to 53; 8=53 to 54; 9≥54
Ol/Lin	logarithm of oleic-linoleic ratio	1≤ −.3; 2=−.3 to −.2; 3=−.2 to −.1; 4=-.1 to .0; 5=.0 to .1; 6=.1 to .2; 7=.2 to .3; 8=.3 to .4; 9=.4 to .5; 10≥.5

[†] Descriptor was treated as an unordered multicategory descriptor for the analyses reported in this article.

The assignment will be such that the newly quantified descriptor stem color will have as high a correlation with the first component as possible, given the other descriptors. The assignment of values is thus related to both the other descriptors in the dataset (we have to maximize the average squared correlations) and to the component. When more than one component is desired, two possibilities with respect to quantification are available—the same quantification for the categories of a descriptor for all components (called single quantification) or a separate quantification for each component (called multiple quantification). The rationale behind the latter option is that one type of contrast between the categories might be related to the descriptors determining the first component, and another type of contrast between the categories might be related to the descriptors determining another component. With two categories, only one contrast (single quantification) is possible.

3.2 Interpretation

One of the major interpretative tools of standard principal components analysis is the matrix of correlations between the descriptors and the components. In psychology these correlations are mostly referred to as loadings, but the use of the term is not always unambiguous. In nonlinear principal component analysis, similar correlations may be computed using the quantified (or optimally scaled) descriptors (also referred to as component-quantified descriptor correlations). For descriptors with multiple quantified categories, such as plant height (see Section 4), the correlations refer to a different quantification for each component. Squared multiple correlations for the regression of the descriptors on the components (often called communalities) indicate how well the components succeed in accounting for the variability of the quantified descriptors. The proportional variance accounted for by the component is the average of the squared multiple correlations with the component. Small numbers of categories often limit the variability of a descriptor and thus it has an averse effect on percentages variance accounted for by a component. However, relatively low percentages of variance accounted for should not necessarily be taken as an indication of a lack of structure.

One caveat must be expressed with respect to the interpretation in nonlinear principal component analysis when there are missing data. In that case, the correlations are no longer exact correlations but only approximations to them. When there are a limited number of missing data, as is the case here (.7%), the deviations are not serious (see Gifi 1980, pp. 136–140).

3.3 Technical Background: Nature of the Data

In order to gain a deeper understanding of the way nonlinear principal component analysis works, it is necessary to briefly discuss the philosophy about data and measurement types underlying nonlinear multivariate analysis as contained in Gifi (1990). This philosophy can be summarized as "All data are categorical (measured with finite precision) and the measurement type is determined by the transformations that may be applied to the categories."

With ordinal data, we may assign the values 1, 2, 3, and so on to the categories provided category 3 has more of the property measured by the descriptor than category 2 has, and 2 has in turn more of the property than category 1 has. However, only the order of the values 1, 2, and 3 is important, not the numerical values themselves. The values 5, 9, and 20 would have done as well. In fact, any order-preserving or monotonic transformation of the values 1, 2, and 3 may be used without changing the meaning of the categories. In nonlinear principal component analysis, we are using this transformational freedom to find the monotonic transformation that leads to maximum correlation between the descriptor and the component, given the other descriptors.

For binary data, there are no restrictions on the values assigned to the two categories. Thus, any transformation that will produce a high descriptor–component correlation may be used. However, only a single quantification for each category is possible, because two categories can only have one contrast. For unordered multicategory descriptors, the transformations are unrestricted, but a choice between single and multiple quantifications exists. As mentioned previously, ordinal descriptors, or ordered multicategory descriptors, are defined by monotonic transformations. In practice, only single quantifications are considered even though multiple quantifications could theoretically be envisaged.

Finally, given that the measured values are in the correct scale, the only transformations allowed for numeric data are linear in the category values. Thus, equidistant observed values have to remain equidistant after transformation. When all descriptors are numeric, the results from nonlinear and ordinary principal component analysis will be the same. Also, if the measured scale is not the "natural" one, log-transformations and other power transformations may be used. In nonlinear principal component analysis, a problem may arise with numeric descriptors in that most observed values are only observed a limited number of times, mostly once. This might cause practical problems during analyses when all distinct values are treated as separate categories. Practice has shown that it is often advantageous to reduce numerical descriptors to a more limited number of categories, say, 7 to 10, preferably covering equal intervals except for the end points. Gifi (1990) indicated that for balanced analyses most categories should preferably not have too low a frequency, say, smaller than 5.

3.4 Technical Background: Algorithm

A compact, simplified description of one-dimensional nonlinear principal component analysis is that, simultaneously, (non)linear transformations of the descriptors and a linear combination of the transformed descriptors are sought such that the average squared correlation of the transformed descriptors and the linear combination is as large as possible. Thus, the technique consists of a combination of two distinct processes. The first consists of transforming the descriptors, and these transformations should be optimal with respect to the aim of achieving as high an average squared correlation between the quantified descriptors and a component as possible. Therefore, this process is called optimal scaling. The other process is the formation of linear combinations of transformed descriptors. The latter process is identical to ordinary principal component analysis, and it aims to achieve as high a variance as possible for the component, given the quantified descriptors. However, neither the optimal transformations nor the best linear combinations are known beforehand, so they have to be determined simultaneously. In practice, the way to do this is to start with some particular transformation for each of the descriptors, perform a principal component analysis on the transformed descriptors, readjust the transformations to suit the derived components, search again for the linear combinations, and so forth until the procedure converges and both the optimal transformations and the best linear combinations are found. This procedure is the basis of the program PRINCALS, which is part of the Category package contained in SPSS (SPSS, Inc. 1990), and was used for all analyses presented in this article.

4. DATA PREPARATION OF PEANUT ACCESSIONS

Before the analysis proper, all values of the numerical descriptors were grouped into 7–10 categories in such a way that, except for the end points, the new categories spanned

Table 2. Correlations Between Optimally Quantified Variables and Components (Loadings) for All 821 Accessions

	*Component**		*Variance*
Descriptor	*1*	*2*	*accounted for*
Branching pattern	−**.773**	.375	.738
Log Oleic/Linoleic ratio	**.685**	−.332	.579
Shelling percentage	**.654**	.061	.431
100-seed weight	**.607**	−.293	.454
Growth habit	−**.516**	.428	.450
Seeds per pod	−**.503**	−.489	.492
Pod constriction	.480	.094	.239
Plant height (1st quant.)[†]	−.455		
Stem pigmentation	−.450	−.142	.223
Pod reticulation	−**.519**	−**.616**	.649
Plant height (2nd quant.)[†]		−**.570**	.531
Plant width	.271	−**.541**	.367
Peg pigmentation	.327	.462	.320
Petal colour	−.414	−.458	.381
Seed colour	−.214	−.371	.183
Oil content	−.002	.136	.018
Pod beak	.252	−.088	.071
Variance accounted for	.235	.148	.383

[†] Because Plant height was treated as an unordered multicategory descriptor, it received separate independent quantifications for each dimension and thus the correlations between the two components and Plant height pertain to these two independent quantifications.

* Values larger than .50 are set in **bold**.

equal intervals and no category had fewer than 5 accessions. For ordinal descriptors, categories were combined with their neighboring categories if they contained fewer than 5 accessions. This was only necessary for end categories (see Table 1). Categories were combined to prevent rare categories unduly influencing the analysis. Oleic-linoleic ratio was first logtransformed with natural logarithms to make the descriptor symmetric with respect to oleic and linoleic content.

For the final analysis reported here, plant height was treated as an unordered multicategory descriptor because preliminary analyses revealed that the descriptor had a nonlinear relationship with other numerical descriptors (see Fig. 1), and multiple quantifications within nonlinear principal component analysis can be used to handle this. The effectiveness of treating plant height as an unordered multicategory descriptor is highlighted in the following section.

5. RESULTS FOR THE OVERALL ANALYSIS

5.1 Descriptor–Component Correlations

Table 2 contains the component-quantified descriptor correlations for all accessions, as well as the squared multiple correlations for the regression of the descriptors on the

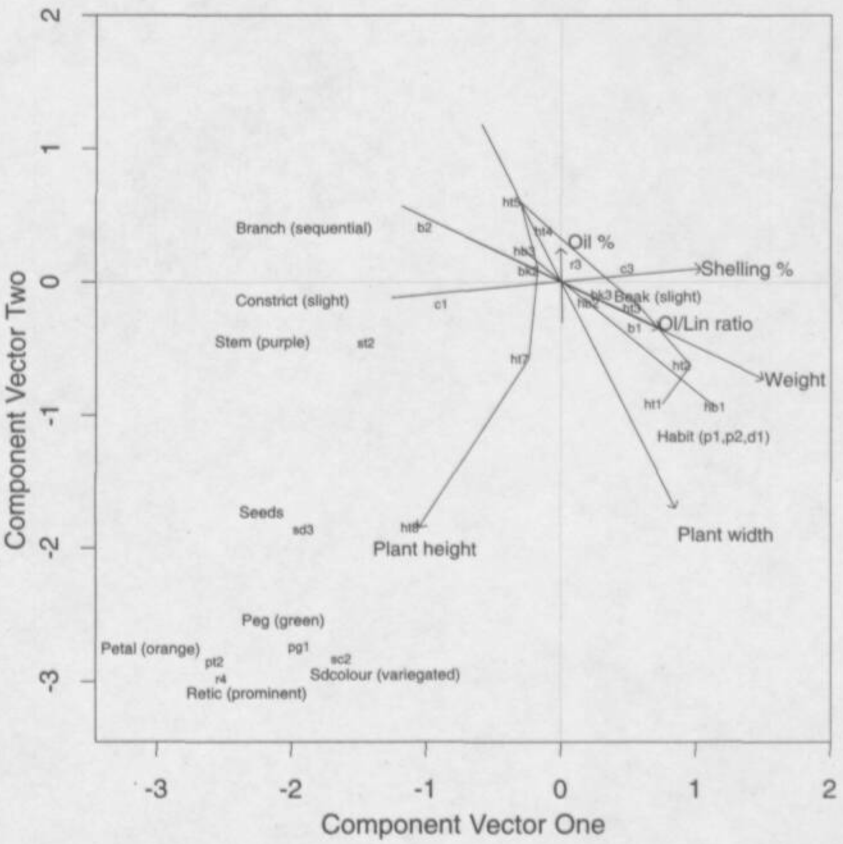

Figure 1. Plot of the Optimal Scaled Values for 16 Descriptors Along the 1st and 2nd Principal Component Vectors, Based on the Entire Australian Groundnut Germplasm Collection Containing 831 Accessions.

components (or communalities). The overall proportion variance accounted for by the components, .38, is the average of the squared multiple correlations (variance accounted for of the descriptors by the components) in the last column. As mentioned in the previous section, the relatively low percentage of variance accounted for can be partly attributed to the presence of descriptors with a limited number of categories and should not be taken as an indication of a lack of structure, as will become evident in the sequel.

From Table 2, descriptors like branching pattern, the log oleic/linoleic ratio, shelling percentage, 100-seed weight, growth habit, seeds per pod, pod reticulation, plant height, and plant width are important in distinguishing between the accessions, while, for instance, oil content and pod beak are not.

5.2 Plotting Descriptors and Accessions

To get a proper view of the relationships among the descriptors and the accessions, one needs to look at their joint representation, especially with some descriptors being

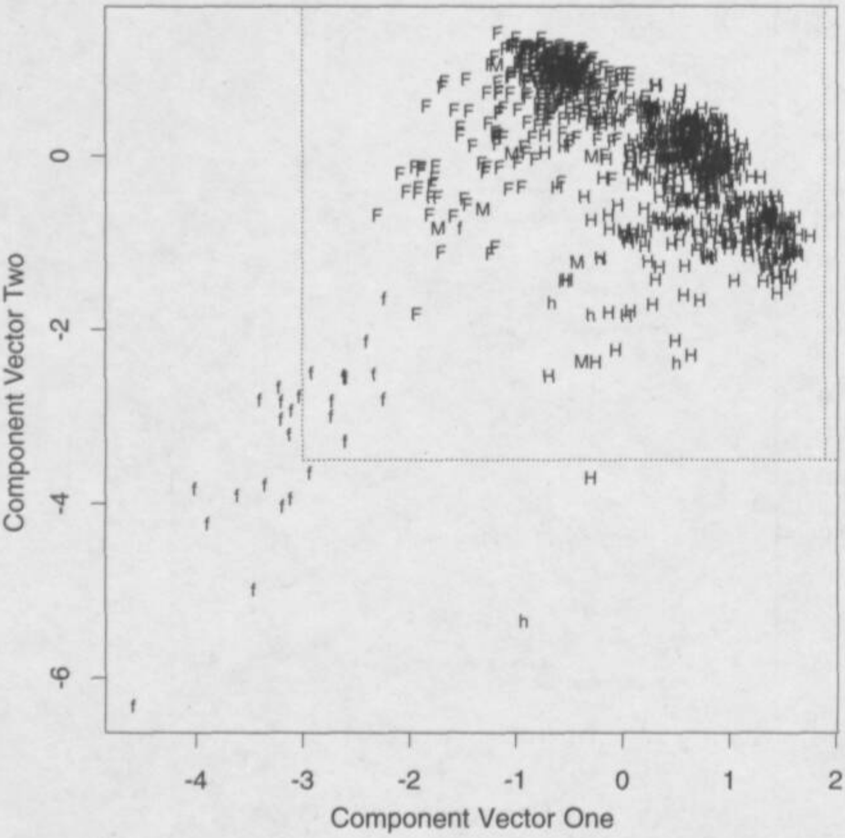

Figure 2. Plot of Accession Scores Along the 1st and 2nd Principal Component Vectors for the Entire Australian Groundnut Germplasm Collection Containing 831 Accessions. Accession points are labeled with their Branching pattern as either "F and f" (sequential), "H and h" (alternate), or "M" (unavailable information). Lower case letters refer to accessions removed in subsequent analyses. Dashed lines indicate where Figure 1 should be superimposed.

discrete rather than continuous. This can be done by constructing displays with the optimal scale values for the categories of the descriptors (Fig. 1) and the accession scores (Fig. 2). Ideally, they should be presented in a single plot, but the large number of descriptor categories and the large number of accessions makes separate plots easier to interpret. The two figures may be superimposed after equalizing the physical scales of the plots. The dashed lines on Figure 2 indicate where Figure 1 should be superimposed.

Interpretation of the plots is based on the fact that nonlinear principal component analysis attempts to place categories in the center of gravity of the accessions scoring in this category. For numerical and ordinal descriptors the categories lie on a straight line through the origin, and for the numerical descriptors we may employ the standard biplot interpretation by projecting accessions on the descriptor vectors shown in Figure 1 (Gabriel 1971).

5.3 INTERPRETATION OF THE DESCRIPTOR DISPLAY

Whereas Table 2 provided the summary measures for the relationships between the descriptors, Figure 1 allows a more detailed inspection of the descriptors and their categories, particularly with there being so many categorical descriptors.

Figure 1 clearly shows the high correlations between the quantified descriptors of the log oleic/linoleic ratio, plant width, and 100-seed weight, as their arrows all point in the same direction. At the same time, procumbent and slightly decumbent (hb1, hb2) accessions with an alternate branching pattern (b1) generally produce wide plants with large 100-seed weight and high oleic versus linoleic content in their seeds, while decumbent and erect (hb3) accessions with a sequential branching pattern tend to produce narrow plants with small 100-seed weight and high linoleic versus oleic content in their seeds. Furthermore, the lengths of the arrows of the continuous and ordered descriptors generally reflect the importance of the descriptors for distinction between the accessions. As remarked previously, oil percentage with its small arrow is not important, while plant width and 100-seed weight are. Similarly, the spread of the categories of an unordered descriptor also reflects this importance; that is, the descriptor pod reticulation is important for the distinction between accessions but pod beak is not, because all pod beak category points are close to the origin of the plot.

In the lower left hand corner, there is a clustering of categories from several descriptors. In particular, there seem to be a group of accessions that tend to have tall plants in excess of .5 m (ht8), orange petals (pt2), prominent pod reticulation (r4), variegated seed coloring (sc2), three to four seeds per pod (sc3), and green pegs (pg1).

5.4 INTERPRETATION OF THE ACCESSIONS DISPLAY

In Figure 2 the majority of the accessions (about 750 of them) roughly form an ellipse with its major axis running from northwest to southeast, with increased saturation indicating large numbers of overlapping accessions. There is also a group of 30–40 "stragglers" located in the southwestern direction of the plot. As mentioned previously, Figure 1 can be superimposed on Figure 2 so that we can establish which accessions have particular characteristics. When describing the patterns in Figure 1, we have implicitly described the accessions as well. To evaluate which characteristics a particular (or group of) accession(s) has, we may drop perpendiculars on the continuous descriptors and evaluate the relevance of the descriptor for that accession, analogous to the way this is done on biplots. To get an overview of the extent to which categorical descriptors succeed in distinguishing between accessions, one may label each accession with its category value for a particular descriptor. This enables insight into the extent of overlap existing between the categories, and it gives the opportunity to identify outlying values if they exist. It also allows searching for accessions with specific or unusual characteristics. Multivariate information about the descriptors is already given in Figure 1, but labeling individual accessions with single (categorical) descriptors illustrates the importance of separate descriptors for discriminating amongst the accessions.

The groundnuts in the Australian collection can be distinguished according to two main subspecies, *Arachis hypogaea* L. *spp. hypogaea* (Virginia: Bunch and Runner) and

Arachis hypogae L. *spp. fastigiata* (Spanish and Valencia). In particular, they are distinguished on the basis of their branching patterns (i.e., alternate and sequential, respectively). Table 2 shows that branching pattern is one of the most discriminating descriptors. In other words, many differences between accessions are strongly related to subspecies. To illustrate this distinction, each accession has been labeled according to its subspecies in Figure 2, that is, by "H" or "h" (alternate branching - *spp. hypogaea*), "F" or "f" (sequential branching - *spp. fastigiata*), or "M", with M referring to accessions for which no information about subspecies is available; lower case letters refer to accessions that will be removed from subsequent analyses (see next section). The discriminatory power of the subspecies designation is evident because the subspecies clearly occupy different parts of the plot. Note that from the location of the accessions in the plot, one could make an intelligent guess about the branching patterns (and thus subspecies) of accessions labeled with an "M".

Apart from their branching pattern, Virginias and the Spanish and Valencias differ in many other aspects. To illustrate this, we need to look along the long axis of the ellipse that can be drawn around the main body of accessions in Figure 2, which more or less coincides with the line connecting the two categories of branching pattern. This axis is highly correlated with 100-seed weight, plant width, and the log oleic/linoleic ratio. In particular, the Virginias located in the southeastern corner of Figure 1 have predominantly larger 100-seed weight, procumbent or slight decumbent growth habit (hb1 in Fig. 1) coupled with higher log oleic/linoleic ratios, large plant widths, and somewhat higher shelling percentages. The Spanish and Valencias on the opposite northwestern side have smaller 100-seed weight, decumbent or erect growth habits (hb3 in Fig. 1) coupled with lower log oleic/linoleic ratios, smaller plant widths, and lower shelling percentages. Note that the vector oil percentage is more or less independent of the distinction between the two subspecies.

As mentioned previously, groundnuts in the Australian collection can also be distinguished by a botanical classification into the varieties Valencia (*Arachis hypogaea* L. *spp. fastigiata var. fastigiata*), Spanish (*Arachis hypogaea* L. *spp. fastigiata var. vulgaris*), and Virginia (*Arachis hypogaea* L. *spp. hypogaea var. hypogaea*), with Virginia having a bunched habit type (Virginia Bunch) and a runner habit type (Virginia Runner).

The additional subdivision of subspecies *spp. fastigiata* into Spanish and Valencia is generally based on more than one characteristic. The Valencias are primarily located in the southwestern part of the plot (i.e., the "stragglers" in Fig. 2) as they generally have prominent pod reticulation (r4 in Fig. 1), three seeds per pod (sd3 in Fig. 1), and sequential branching patterns (b2 in Fig. 1).

6. RESULTS FOR THE BULK OF THE ACCESSIONS

Due to the deviating patterns of the small group of Valencias, a clear view on the similarities and differences between the major groups of accessions is somewhat obscured. To gain a clearer insight, the analysis was repeated without the outlying Valencias; in particular, 34 accessions with prominent pod reticulation (r4) were not included. The results are shown in Figure 3 (descriptors) and Figure 4 (accessions) for the first two component vectors, which account for 33% of the total variance.

Table 3. Correlations Between Optimally Quantified Variables and Components (Loadings) for the Main 797 Accessions

	*Component**		*Variance*
Descriptor	*1*	*2*	*accounted for*
Branching pattern	**−.857**	−.035	.735
Log Oleic/Linoleic ratio	**.754**	−.001	.568
100-seed weight	**.701**	.343	.609
Growth habit	**−.602**	**.574**	.691
Shelling percentage	**.584**	.063	.345
Plant height (1st quant.)[†]	−.488		
Plant height (2nd quant.)[†]		**.557**	.548
Pod beak	.286	**.541**	.375
Plant width	.415	**−.524**	.447
Pod reticulation	−.048	.445	.200
Stem pigmentation	−.398	−.435	.347
Pod constriction	.388	.359	.280
Seeds per pod	−.193	−.251	.101
Oil content	−.066	−.159	.029
Petal colour	−.058	−.147	.025
Peg pigmentation	−.034	.118	.015
Seed colour	−.016	−.004	.000
Variance Accounted For	.209	.123	.332

[†] Because Plant height was treated as an unordered multicategory descriptor, it received separate independent quantifications for each dimension and thus the correlations between the two components and Plant height pertain to these two independent quantifications.

[*] Values larger than .50 are set in **bold**.

6.1 Descriptor–Component Correlations

In Table 3, we have presented the component–descriptor correlations for the analysis based on 797 accessions. The first component more or less coincides with the alternate-sequential distinction, as is evident from the very high correlation (.856), and it also coincides with the long axis of the main body of accessions in Figure 2 (see Fig. 4). From Table 3, it is clear that the descriptors log oleic/linoleic ratio and shelling percentage are almost exclusively related to subspecies distinction, but that 100-seed weight, growth habit, and plant width also differentiate between accessions independent from the subspecies distinction. Several descriptors fail to contribute to differences between the majority of accessions, such as seeds per pod, oil content, pod beak, petal color, peg pigmentation, and seed color.

6.2 Interpretation of the Descriptor and Accession Displays

Figure 4 shows essentially the accessions in the ellipse of Figure 2, even though the ellipse is no longer recognizable as such. The basic patterns for the descriptors in Figure 3 are unchanged, except that plant height is much more linear, and some of the categorical and ordinal descriptors can be more easily evaluated than before. The accession plot shows a better separation into the two groups with sequential and alternate

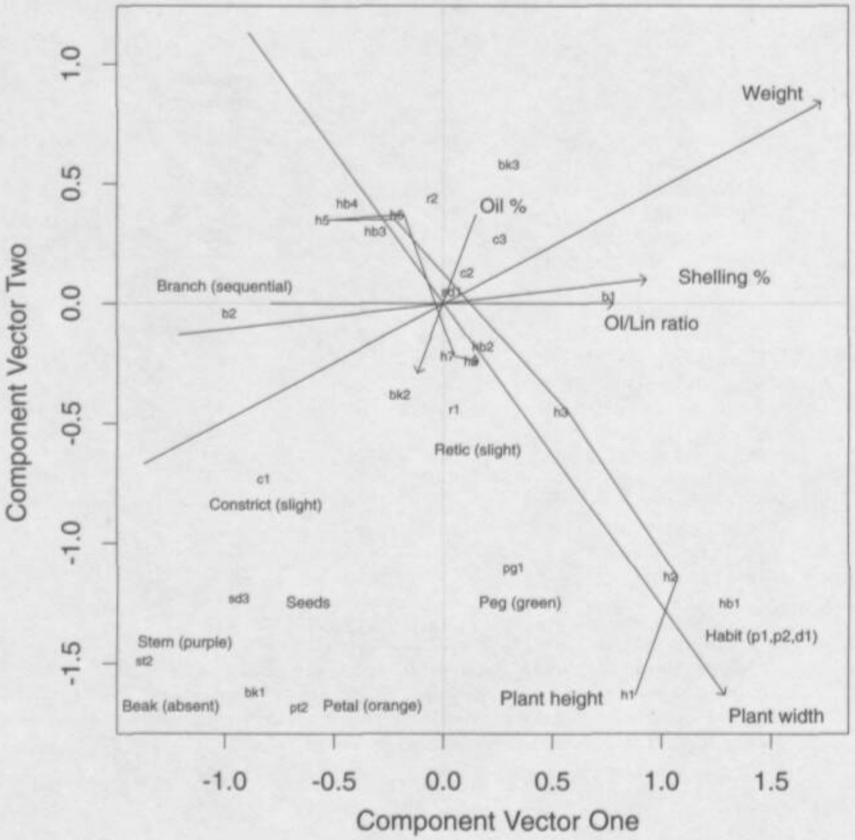

Figure 3. Plot of the Optimal Scaled Values for 16 Descriptors Along the 1st and 2nd Principal Component Vectors, Based on a Restricted Subset From the Australian Groundnut Germplasm Collection Containing 797 Accessions.

branching patterns (labeled "F" and "H" in Fig. 4a), the Valencias and Spanish and the Virginias, respectively. Moreover, there is a suggestion of further grouping within the main subspecies.

To highlight these groupings (more specifically, the distinctions among the botanical varieties), Figure 4a is redrawn as Figs. 4b, 4c and 4d, but with the accessions marked with the categories of the more discriminating descriptors. Figure 4b uses stem pigmentation to show the distinction in the subspecies *spp. fastigiata*, which can be mainly attributed to differences in Valencias ("P" - purple; southwest region of Fig. 4b) and Spanish ("G" - green; northwest) botanical varieties, while Figs. 4c and 4d use growth habit and plant height to show distinctions in the subspecies *spp. hypogaea*, which can be attributed to differences in the types Virginia Runner ["P" - procumbent, decumbent-1 (Fig. 4c) and "L" - $\leq$ 30 mm (Fig. 4d); southeast region of plot] and Virginia Bunch ["2" - decumbent-2, "3" - decumbent-3, "E" - erect (Fig. 4c) and "M" - 35 to 40 mm, "H" - $\geq$ 40 mm (Fig. 4d); northeast].

Again, the categorical descriptors have been singled out to label the accessions in the plots based on the multivariate analysis, because it is easier to assess the coherence of groups of accessions with particular characteristics. We could have labeled the same plot with two or more discrete descriptors, but this would have complicated the interpretation of the plot.

7. DISCUSSION AND CONCLUSION

In this article, nonlinear principal component analysis was used to analyze both categorical and numerical descriptors of the Australian groundnut germplasm collection. The resulting plots provided a global picture of the diversity available for use in plant improvement programs and showed the major relationships between all descriptors, together with the extent to which they contributed to distinguishing the accessions. For the analysis that included all of the accessions, the two subspecies of *Arachis hypogaea* L. *spp. hypogaea* and *Arachis hypogaea* L. *spp. fastigiata* could be clearly distinguished. The results from the analysis with outliers removed enabled a more detailed characterization of the accessions, providing not only an identification of the two subspecies, but also allowing a clearer distinction between the three botanical varieties (Spanish, Valencia, Virginia) as well as the separation of the Virginia types by their growth habit (Virginia Runner and Virginia Bunch). The plots also clearly showed accessions that had different characteristics from the main body of accessions.

The use of both the accession and descriptor plots is seen as valuable because it allows data interpretation when there is a need for plant breeders to look for different sources of variability to accommodate various breeding needs. For example, the domestic market may demand larger sized groundnut seeds, whereas export markets may require smaller sized groundnut seeds (known as *cultural requirements*; see Henning et al. 1982). Consequently, the accessions with high 100-seed weight, which are suitable for the domestic market, can be easily identified on accession plots (mainly Virginia types) in relation to the direction of the 100-seed weight vector in the descriptor plot and similarly for the accessions with low 100-seed weight (mainly Spanish and Valencias types). Thus, perceiving the various breeding requirements as descriptor profiles enables easy identification of relevant accessions from the accession and descriptor plots.

The graphics can also assist by providing information when data are incomplete (i.e., "M" on Fig. 2). The position of these accessions in the plots can indicate the most likely subspecies, botanical variety, and so on, to which they may belong.

Compared to biplots constructed on the basis of numerical descriptors, the present descriptor plots require more interpretational efforts, primarily because there is an emphasis on categories along with the descriptors. The introduction of transformations for the values of descriptors requires an intimate knowledge of the data to decide on the proper measurement level of the descriptors and to judge the acceptability of the transformations. The nonlinear behavior of plant height, which only came to the foreground during the analysis, emphasizes this point.

The information contained in these plots has the potential to simplify the identification of valuable accessions, reduce the amount of time that it has previously taken for evaluating relevant accession material for use in plant improvement programs, and ultimately improve the usefulness and utilization of germplasm collections by plant breeders (Knauft and Gorbet 1989; Smartt 1994, chap. 17).

The advantage of using nonlinear principal component analysis is that descriptors of different measurement levels can be combined into a single analysis. For efficiency purposes this meant that the numerical descriptors had to be categorized into 7 to 10 categories, but the loss in precision this entails is relatively minor.

Previously, mixed measurement level data were often converted to separate matrices of similarities between accessions for each descriptor using a similarity measure appropriate for the measurement level in question (see Gower 1971; Romesburg 1984). An example of this procedure, using the same data taken from the Australian germplasm collection, is contained in Harch et al. (1996a). They averaged the range-standardized similarity matrices for the binary, ordered multicategory and quantitative descriptors (using equal and unequal weighting for the data types) and performed standard principal component analysis and hierarchical clustering [Ward's (1963) method] on the averaged similarity matrix. Although the computational approach taken by Harch et al. (1996a) acknowledges the different data types within its algorithm and enables one complete analysis to be performed, in contrast to the analysis presented here, the similarities amongst the descriptors could not be included in the analysis along with the similarities amongst the accessions. One possible avenue that could be explored to address this is to apply individual differences scaling to the set of similarity matrices, but this will not be explored in this article.

Both sets of analyses (equal and unequal weighting) found that the descriptors distinguishing among the accessions along the first principal component vector were branching pattern, 100-seed weight, shelling percentage, and the log oleic/linoleic ratio. These results, like the results found here, were reflecting the main differences between the subspecies of *Arachis hypogaea* L. *spp. hypogaea* (Virginia) and *Arachis hypogaea* L. *spp. fastigiata* (Spanish and Valencia). Equal weighting of the data types provided additional information about distinguishing accessions with respect to their pod beak and pod reticulation characteristics. It was uncertain whether this would apply to other datasets.

As found using nonlinear principal component analysis, further distinction between the botanical varieties of Spanish (*Arachis hypogaea* L. *spp. fastigiata var vulgaris*) and Valencia types (*Arachis hypogaea* L. *spp. fastigiata var fastigiata*) was also illustrated in the ordination plots of the first and third component vectors. However, the distinction between the subspecies of the Virginia types (Runner and Bunch) was not shown in Harch et al. (1996a). Harch et al. (1996a) and Harch et al. (1996b) outlined the approach taken by Esquivel et al. (1993a, 1993b) for the analysis of Cuban groundnut germplasm data and the analysis taken by Holbrook et al. (1993) for the United States groundnut germplasm data, respectively. Both of these datasets contained mixed data types. Harch et al. (1996b) also proposed methodology for analyzing mixed data types from the world groundnut database (12,160 accessions). Esquivel et al. (1993a, 1993b), Holbrook et al. (1993), and Harch et al. (1996b) all found that their results reflected groundnut taxonomy.

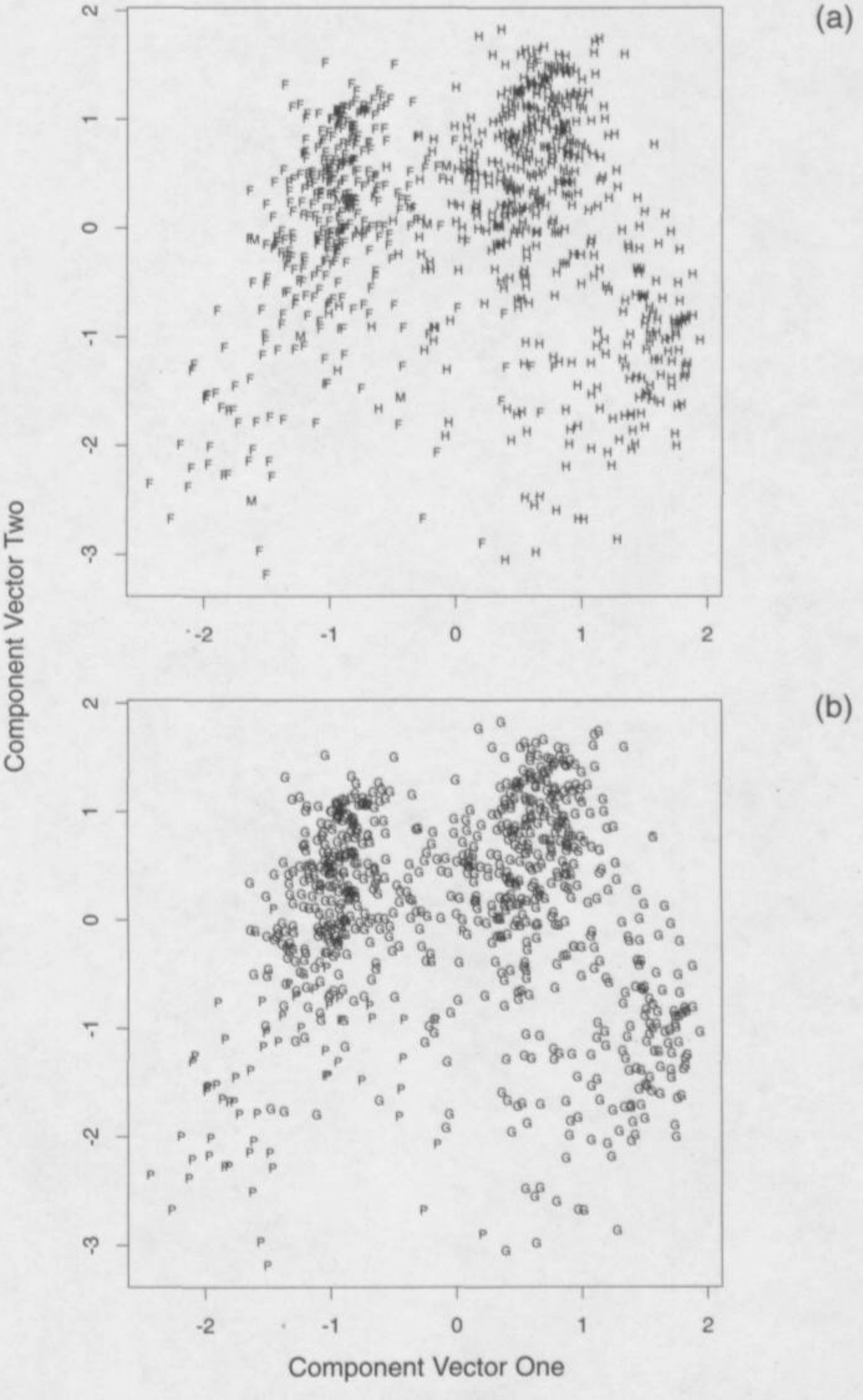

Figure 4. Plot of Accession Scores Along 1st and 2nd Principal Component Vectors Based on a Restricted Subset of Australian Groundnut Germplasm Collection (797 Accessions). (a) Branching pattern of points are "F" (sequential), "H" (alternate), or "M" (unavailable information). (b) Stem Pigmentation is "G" (green) or "P" (purple).

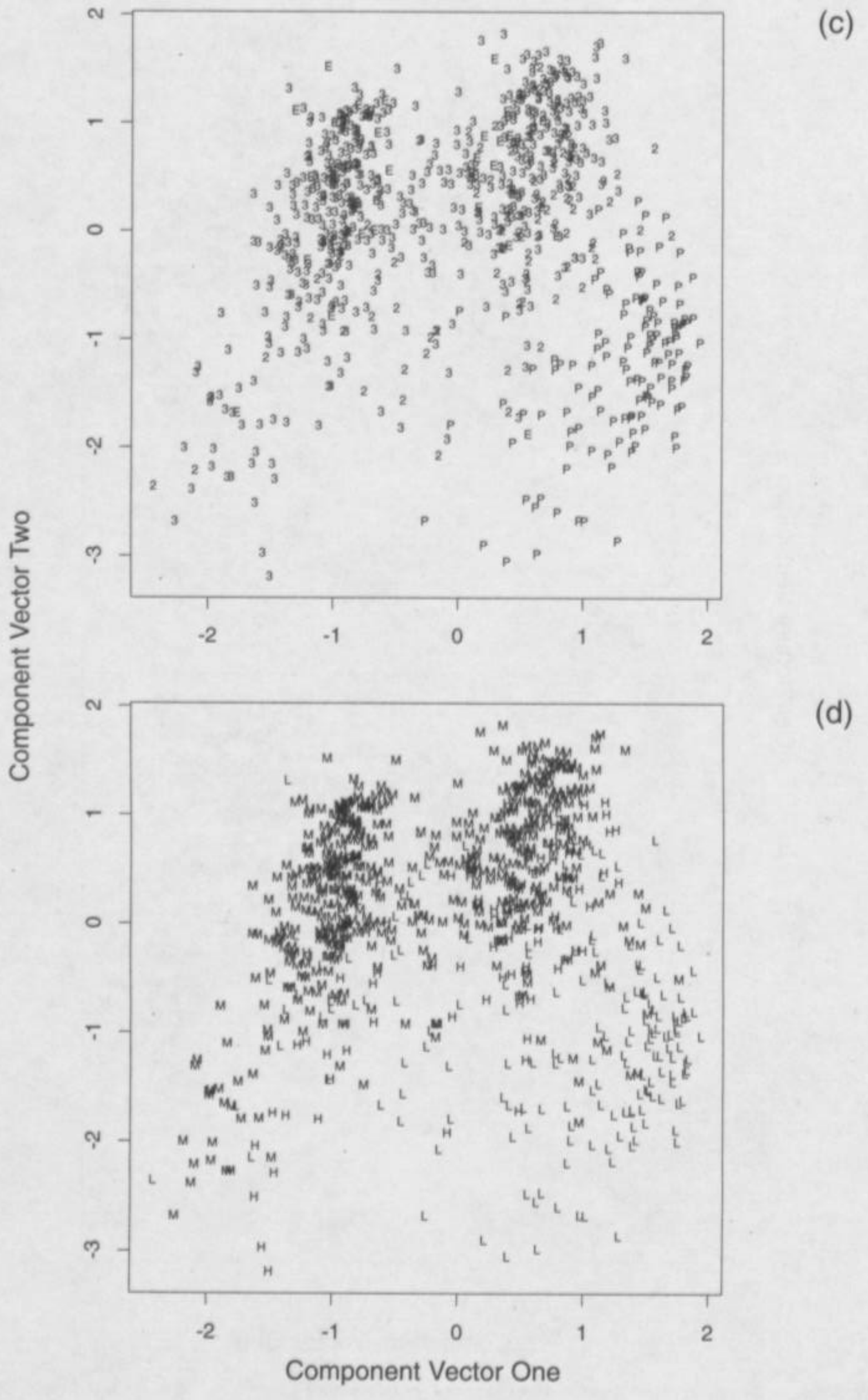

Figure 4. (*continued*). (*c*) *Growth habit is "P" (procumbent, decumbent-1), "2" (decumbent-2), "3" (decumbent-3), or "E" (erect). (d) Plant height is "L" (≤ 30 mm), "M" (35–40 mm), or "H" (≥ 40 mm).*

ACKNOWLEDGMENTS

This article was partially written while the first author was on leave at the Department of Agriculture of the University of Queensland. He was financially supported by the Netherlands Organization for Scientific Research (NWO) and by a grant to Dr. K.E. Basford from the Rural Industries and Development Corporation, Canberra, ACT. The second author contributed to this work while being supported by the Grains Research and Development Corporation, Canberra, Australia (Junior Research Fellowship JRF3B) at the Department of Agriculture, The University of Queensland, Australia.

We acknowledge Peter Lawrence (Department of Primary Industries, Biloela, Queensland) for providing the appropriate databases, John Tonks for growing and characterizing the groundnut accessions at the J. Bjelke-Petersen Research Station (Department of Primary Industries), Kingaroy, Queensland, and the gas chromatography work done by Cathy McLeod at the Department of Primary Industries Chemistry Laboratories, Indooroopilly, Brisbane, Queensland.

[Received September 1996. Revised March 1997.]

REFERENCES

Bretting, P. K., Goodman, M. M., and Stuber, C. W. (1990), "Isozymatic Variation in Guatemalan Races of Maize," *American Journal of Botany*, 77, 211–225.

Esquivel, M., Barrios, M., Waln, L., and Hammer, K. (1993a), "Peanut (*Arachis hypogaea* L.) Genetic Resources in Cuba. I. Collecting and Characterisation," *FAO/IBPGR Plant Genetic Resources Newsletter 91/92*, 9–15.

——— (1993b), "Peanut (*Arachis hypogaea* L.) Genetic Resources in Cuba. II. Preliminary Germplasm Evaluation," *FAO/IBPGR Plant Genetic Resources Newsletter 91/92*, 17–20.

Gabriel, K. R. (1971), "The Biplot-Graphical Display of Matrices With Application to Principal Components Analysis," *Biometrika*, 58, 453–467.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Chichester, UK: Wiley.

Gower, J. C. (1971), "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, 27, 857–872.

Gregory, W. C., Smith, B. W., and Yarbrough, J. A. (1951), "A Radiation Breeding Experiment With Peanuts. II. Characterisation of the Irradiated Population (NC4-18.5 kR)," *Radiation Botany*, 8, 85–93.

Harch, B. D. (1996), "Statistical Evaluation of Germplasm Collections," unpublished Ph.D. thesis, Department of Agriculture, The University of Queensland, Brisbane, Australia.

Harch, B. D., Basford, K. E., DeLacy, I. H., Lawrence, P. K., and Cruickshank, A. (1995), "Patterns of Diversity in Fatty Acid Composition in the Australian Groundnut Germplasm Collection," *Genetic Resources & Crop Evolution*, 42, 243–256.

——— (1996a), "Mixed Data Types and the Usage of Pattern Analysis on the Australian Groundnut Germplasm Data," *Genetic Resources and Crop Evolution*, 43, 363–376.

Harch, B. D., Basford, K. E., DeLacy, I. H., and Lawrence, P. K. (1996b), "The Analysis of Large Scale Incomplete Data Taken From the World Groundnut Germplasm Collection. II. Two-Way Data with Mixed Data Types," Centre for Statistics Research Report 54, Department of Mathematics, The University of Queensland, Brisbane, Australia.

Henning, R. J., Allison, A. H., and Tripp, L. D. (1982), "Cultural Practices," in *Peanut Science and Technology*, eds. H. E. Pattee and C. T. Young, Yoakum, TX: American Peanut Research & Education Society, Inc., pp. 123–138.

Holbrook, C. C., Anderson, W. F., and Pitman, R. N. (1993), "Selection of a Core Collection from the U.S. Germplasm Collection of Peanut," *Crop Science*, 33, 859–861.

International Board for Plant Genetic Resources (IBPGR), and International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) (1992), *Descriptors for Groundnut*, Rome, Italy and Patancheru, India: Authors.

Joliffe, I. T. (1986), *Principal Components Analysis*, New York: Springer-Verlag.

Knauft, D. A., and Gorbet, D. W. (1989), "Genetic Diversity Among Peanut Cultivars," *Crop Science*, 29, 1417–1422.

Krapovickas, A., and Gregory, W. C. (1994), "Taxonomía del Género *Arachis* (*Leguminosae*)," *Bonplandia,* 8, 1–186.

Lawrence, P. (1989), "The Australian Tropical Field Crops Genetic Resource Centre," *Australian Plant Introduction Review,* 20(2), 1–5.

Perry, M. C., and McIntosh, M. S. (1991), "Geographical Patterns of Variation in the USDA Soybean Germplasm Collection: I. Morphological Traits," *Crop Science,* 31, 1350–1355.

Romesburg, H. C. (1984), *Cluster Analysis for Researchers,* Belmont, CA: Lifetime Learning Publications.

Singh, S. P., Nodari, R., and Gepts, P. (1991), "Genetic Diversity in Cultivated Common Bean: I. Allozymes," *Crop Science,* 31, 19–23.

Smartt, J. P. (1994), "The Future of the Groundnut Crop," in *The Groundnut Crop,* ed. J. P. Smartt, London: Chapman and Hall, pp. 700–720.

SPSS Inc. (1990), *Categories,* Chicago: Author.

Stalker, H. T. (1989) "Utilising Wild Species for Crop Improvement," in *IBPGR Training Courses, Lecture Series 2. Scientific Management of Germplasm: Characterisation, Evaluation and Enhancement,* eds. H. T. Stalker and C. Chapman, Rome: IBPGR, and Raleigh, NC: Department of Crop Science, North Carolina State University, pp. 139–154.

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association,* 58, 236–244.

Wynne, J. C., and T. A. Coffelt (1982) "Genetics of *Arachis hypogaea* L.", in *Peanut Science and Technology,* ed. H. E. Patee and C. T. Young, Yoakum, TX: American Peanut Research and Education Society Inc., pp. 95–122.