Theor Appl Genet (1990) 79:225–234

THEORETICAL AND APPLIED GENETICS

Multiattribute evaluation of regional cotton variety trials

K. E. Basford [1], P. M. Kroonenberg [2], I. H. DeLacy [1] and P. K. Lawrence [3]

[1] Department of Agriculture, University of Queensland, St. Lucia, Queensland 4067, Australia
[2] Department of Education, University of Leiden, Leiden, The Netherlands
[3] Department of Primary Industries, Biloela, Queensland 4715, Australia

Received June 6, 1989; Accepted October 20, 1989
Communicated by A.R. Hallauer

Summary. The Australian Cotton Cultivar Trials (ACCT) are designed to investigate various cotton [*Gossypium hirsutum* (L.)] lines in several locations in New South Wales and Queensland each year. If these lines are to be assessed by the simultaneous use of yield and lint quality data, then a multivariate technique applicable to three-way data is desirable. Two such techniques, the mixture maximum likelihood method of clustering and three-mode principal component analysis, are described and used to analyze these data. Applied together, the methods enhance each other's usefulness in interpreting the information on the line response patterns across the locations. The methods provide a good integration of the responses across environments of the entries for the different attributes in the trials. For instance, using yield as the sole criterion, the excellence of the namcala and coker group for quality is overlooked. The analyses point to a decision in favor of either high yields of moderate to good quality lint or moderate yield but superior lint quality. The decisions indicated by the methods confirmed the selections made by the plant breeders. The procedures provide a less subjective, relatively easy to apply and interpret analytical method of describing the patterns of performance and associations in complex multiattribute and multilocation trials. This should lead to more efficient selection among lines in such trials.

Key words: Three way data – Clustering via mixtures – Principal component analysis

Introduction

Two methods for the analysis of three-way data from regional variety trials are described using a cotton [*Gossypium hirsutum* (L.)] breeding program as an example. The aim is to enhance the researcher's ability to make informed decisions about the results of these trials.

At the time of these trials, four cotton breeding programs were operating in Australia, three in New South Wales (NSW) and one in Queensland (Qld). Beginning in 1974/75, the cotton breeders at the Commonwealth Scientific, Industrial, and Research Organization (CSIRO) and the Queensland Department of Primary Industries (QDPI) have jointly been conducting the Australian Cotton Cultivar Trials (ACCT) at 6–11 locations per year throughout the major cotton growing districts in NSW and Qld (Fig. 1). In any given year, from 16 to 30 cotton lines are evaluated by measuring lint yield (tons/ha) and other lint quality characteristics, the most important of these being lint strength (g/tex), lint micronaire (combined measure of fiber diameter and maturity), and lint length (inches). The units used are the industry standards for these characteristics.

Each year, a three-way data array classified as lines by locations by attributes must be evaluated to assess the performance of the cotton lines. Interpreting the underlying complex interactions in such a three-way array is difficult. If the evaluation of the lines is made using only one attribute such as yield then, even though this may be considered to be the most important attribute, much of the available data is being ignored. Separate analysis for each attribute is not satisfactory because of the difficulty of successfully combining the results and also because this procedure explicitly ignores the correlations among the data. Line assessment then depends very much on the "ability and experience" of the particular plant breeder. Therefore, it would seem advantageous to use statistical techniques that will simultaneously analyze more than one attribute at a time.

Generally for single attributes, cluster techniques and ordination techniques are used jointly and in a supple-

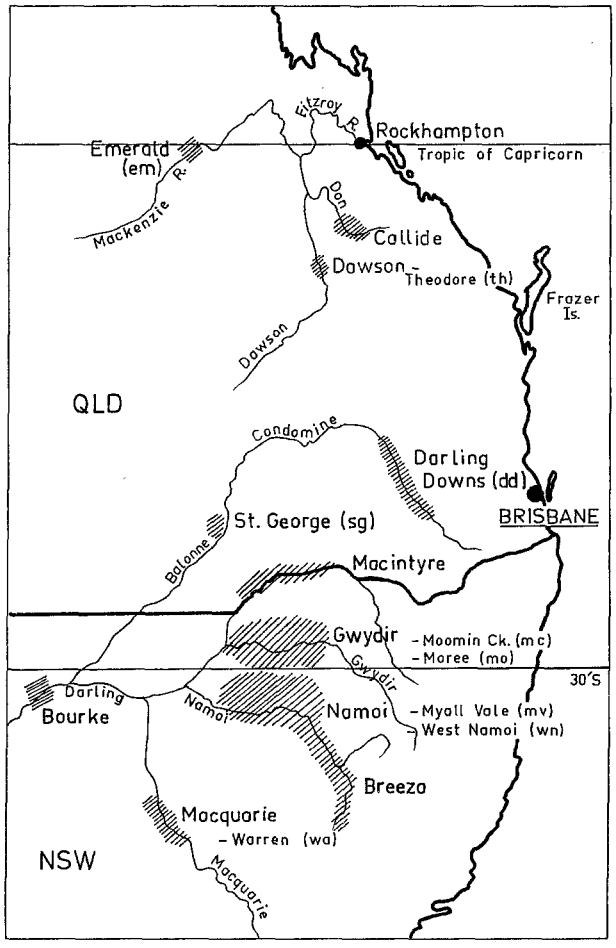

Fig. 1. The eleven locations which represent the major cotton growing districts in eastern Australia used for the Australian Cotton Cultivar Trials (ACCT)

mentary fashion to evaluate relative performance of genotypes over environments (Williams 1976; DeLacy 1981). Similarly, cluster analysis and ordination can be used for evaluation of three-way data. Here, one representative of each class of multivariate techniques, the mixture maximum likelihood method of clustering and three-mode principal component analysis, are discussed. They have each been used separately to analyze soybean [*Glycine max* (Merr.)] data of this form (Basford and McLachlan 1985a; Kroonenberg and Basford 1989), but they are not techniques regularly employed by plant breeders. Both of these approaches will be discussed briefly and then illustrated by the analysis of the multiattribute data collected on 25 cotton lines grown in each of nine locations in eastern Australia in the summer of 1980/81 as part of the ACCT.

Our main objective in presenting these analyses is to show that it is possible to treat several attributes in one analysis, to make both global and detailed statements about the relative performance of the cotton lines, and to emphasize the usefulness of treating three-way data in this way.

Materials and methods

Experimental details

In the 1980/81 growing season, the nine locations used in the ACCT were, from north to south, Emerald, Theodore, Darling Downs, St. George, Moomin Creek, Mooree, Myall Vale, West Namoi, and Warren (Fig. 1). The 25 cotton lines planted are listed in Table 1, and the industry standard at the time was dp61. The individual experiments were randomized complete block designs with three replications in each location in Queensland and square lattice designs with three replications in each location in New South Wales. Among other attributes, lint yield (tons/ha) and three lint quality characters – lint strength (g/tex), lint micronaire (combined measure of fiber diameter and maturity), and lint length (inches) – were measured on all lines in all locations. This gives a three-way array of 25 lines by nine locations by four attributes that plant breeders need to interpret.

Details of the trials, entries, and locations are contained in a paper (Reid et al. 1989) on regional evaluation of cotton cultivars in eastern Australia 1974–1985. Before lines are entered in the ACCT, they have been tested in trials at two to three locations for 2 years. This data together with the ACCT data is used to select entries for the next year's trials. Selection was based on yield, three fiber characteristics (the three listed above), lint percent (percent of whole seed harvested which is lint), and field notes based on agronomic type, etc. However, yield and fiber quality were the most important. On these criteria, lines c310, c315, m220, dp55, dp61, sic1, sic2, 39h, mo63, and 286f were selected for the subsequent year's experimentation, while lines nam (namcala) and dp16 were also retained for genetic reasons (checks).

To avoid possible confusion, the "lines" or "entries" and "locations" in the experiment will henceforth be referred to as "genotypes" and "environments", respectively.

Mixture maximum likelihood method of clustering

Data collected from regional variety trials are often in the form of a large three-way array, designated as genotypes by environments by attributes in Basford (1982) and Basford and McLachlan (1985a). If the genotypes can be clustered or grouped such that the genotypes within a group have similar response patterns for each of the attributes across environments, then the plant breeder can examine a much smaller data set and, hence, more easily integrate the information inherent in the trials. The mixture maximum likelihood method of clustering is a model-based technique, which can be applied in such cases to produce a grouping of genotypes based on the simultaneous use of attributes and environments.

As detailed in McLachlan and Basford (1988), the technique of clustering uses the measurements on a set of elements (genotypes in the present context) to identify clusters or groups in which the elements are relatively homogeneous, while they are heterogeneous between the clusters. In using the mixture method of clustering, it is assumed in the first instance that there is a specified number, e.g., g, of underlying groups. A likelihood is formed under the assumption that the elements are a sample from a mixture in various proportions of these groups; this is why it is called the mixture method. The most common assumption, and the one used here, is that the underlying distribution of the attributes in each group is multivariate normal. In the

Table 1. Group composition and estimated means (with standard errors in parentheses) for the four attributes formed by the clustering technique

Attribute	Group A	Group B	Group C	Group D	Mean
Lint yield (t/ha)	1.21 (L) (0.03)	1.32 (M) (0.09)	1.45 (H) (0.03)	1.33 (M) (0.04)	1.37 (0.03)
Strength (g/tex)	22.0 (L) (0.4)	23.6 (M) (0.4)	23.5 (M) (0.2)	25.4 (H) (0.5)	24.0 (0.3)
Micronaire (diameter and maturity)	4.21 (L) (0.10)	4.81 (H) (0.11)	4.57 (M) (0.07)	4.39 (L) (0.08)	4.52 (0.06)
Lint length (inches)	1.09 (L) (0.01)	1.09 (L) (0.01)	1.13 (M) (0.01)	1.16 (H) (0.01)	1.13 (0.01)
Membership	m8 rex	g106 st7AN 286f 28/1	dp16 dp55 dp61 dp80 st7A sic1 sic2 39h 286h 28/3 m220	nam c310 c312 c315 c511 c600 mo63 572n	

High (H), medium (M), and low (L) mean values for the groups, with high micronaire indicating low lint quality

model, the groups have different mean vectors and different correlation matrices.

One of the objectives of the analysis is to estimate these unknown parameters in the model. This is achieved by consideration of the likelihood (Dempster et al. 1977) described above. The probability that each element belongs to each of the underlying groups is calculated by replacing the unknown parameters in the appropriate probability expression with their likelihood estimates; this is why it is called the mixture maximum likelihood method. Each element is then allocated to the group for which it has the largest estimated (posterior) probability. This results in an allocation of the elements into groups or clusters.

Basford and McLachlan (1985a) showed how this approach can be extended to the type of three-way data described in the previous section. The model assumes that each underlying population has its own mean vector, which can be different from one environment to another; that is, a group may yield well in one environment but poorly in another. However, the correlation structure between the attributes in that group is the same across environments; that is, within the group the same correlation structure between attributes holds across environments. The model does allow the correlation matrices between the attributes to be different for the different groups. This allows for the general situation where there may be interaction between genotypes and environments; for example, in one group there may be a positive correlation between yield and lint length, while in another group this may not be so. Indeed, in the current example there is a highly significant genotype by environment interaction.

Three-mode principal component analysis

In cluster analysis the environments and attributes are jointly used to find an optimal separation of the genotypes into groups or clusters. After the clusters have been found, mean values for all environments are graphed for each attribute separately, to evaluate the relative performance of the clusters with respect to the environments and attributes. In cluster analysis no direct attempt is made to describe the commonalities and differences between environments and/or attributes. Furthermore, the differences between genotypes are described only insofar as they align with the one cluster structure discovered. Other important sources of variability between genotypes might exist that gives rise to an ordering of genotypes that is not commensurate with the primary cluster structure. It is, therefore, useful to supplement the cluster analysis with ordination techniques, thereby achieving a different investigation of possible structure in the data for genotypes, environments, and attributes.

Common ordination techniques for two-way data are principal component analysis, principal coordinate analysis, multidimensional scaling, and correspondence analysis. For the three-way genotype by environment by attribute ($G \times E \times A$) data, an extension of the first method will be described, i.e., three-mode principal component analysis, which was devised by Tucker (1966) and for which an (alternating) least-squares algorithm was developed by Kroonenberg and De Leeuw (1980) (see also Kroonenberg 1983).

The basic aim of the model underlying the method is to represent each of the ways or modes (genotypes, environments, and attributes) as well as possible (i.e., accounting for as much variation as possible) in a low-dimensional space by forming linear combinations (components) of the levels of the modes. Furthermore, the model describes how the components of the different modes interact. There are various ways to present the condensed information in terms of (functions of) the parameters of the model. Since the model is a simultaneous description of all three modes, it is possible to emphasize the description of one of the modes in any presentation.

As discussed more fully in Kroonenberg and Basford (1989), an attractive way to present the results from data when a description of the genotypes is emphasized is to make scatter diagrams (plots) displaying simultaneously the position of the genotypes in an attribute space and the attributes in a genotype space. Since both spaces are directly comparable (they have been

scaled), the scatter diagrams can be superimposed. In such "joint plots" each genotype and attribute is represented by a vector emanating from the origin, and the relationships between genotypes and attributes follow from the lengths and angles of the particular vectors. For similar plots of two-way data, called bi-plots, see Gabriel (1971) and Kempton (1984). The strength of the relationship is measured by the inner (or scalar) product of the two vectors (i.e., the product of their lengths times the cosine of the angle between them), and these can be presented in table form (Kroonenberg and Basford 1989).

Since it is usual to emphasize the description of one of the modes (here genotypes) in terms of the other two, the vectors (lines from the origin to the points) of only one of the modes (here attributes) is drawn. The strength of the relationship (inner product) between the genotypes and the attributes can then be ascertained from the projections of the genotypes on the attribute vectors. In the case of one dimension effectively explaining all the variability, the joint plots collapse into a single line and the inner products become simply products of lengths of colinear vectors. For such single-dimension line plots, it is possible to include the vectors of the third mode as well, creating what could be called "tri-plots". In such a case, the strength of the trivariate (or tri-componental) relationship can be determined as the product of the lengths of vectors from each of the modes. The clusters found with the mixture method can be readily drawn on the joint plots, so that the information from the two techniques can be evaluated jointly.

Using the residuals from a three-mode principal component analysis, information is also provided about how well the genotypes, attributes, and/or environments fit the model. The overall fit of the model can be assessed and the relative importance of the components of the modes and their combinations can be evaluated with the squared multiple correlation between observed and estimated data.

Cluster analysis versus principal component analysis

One of the striking differences between the techniques is that cluster analysis can very efficiently describe the characteristics of groups of genotypes, but it can do so only in one way. On the other hand, the component analysis provides no clear grouping, but gives a spatial representation of each mode as well as of combinations of modes.

In cluster analysis, a genotype can have an estimated (posterior) probability of belonging to several groups, with the natural proviso that these probabilities add to one over all the groups. To obtain a partitioning into non-overlapping groups, each genotype is allocated to that group for which it has the largest such probability. This non-allocation of genotypes to a group or cluster until the final stage is one of the advantages of the mixture method of clustering (McLachlan and Basford 1988). McLachlan and Basford recommend the examination of these probability estimates of group membership both as an aid in the choice of the number of underlying groups and also to provide information on the strength of the association of an element with a particular group. Several examples are quoted where the estimates of (posterior) probabilities are useful in the latter context. However, these probabilities do not appear to be as informative for three-way data, because the maximum values seem artificially high. For example, in the present case they are all equal to one.

The component analysis provides no clear grouping, but gives a spatial representation of each mode as well as of combinations of modes. In the interpretation, there is no reduction in the number of elements to inspect; for instance, all genotypes make up a spatial representation, but it is of low dimensionality. This makes for a more complex but also a more detailed interpretation. There is no restriction on the position of single genotypes, nor on the formations of different groups of genotypes in different dimensions.

It is primarily the combination of the global organization with fairly straightforward interpretation and the detailed organization with a rather sophisticated interpretation that provides the usefulness of employing the techniques in conjunction.

Results

The results of the cluster and ordination techniques will be discussed below in a relatively independent way. In this manner, the different and supplementary character of the two techniques can be demonstrated more clearly.

Cluster analysis

Although it might seem more realistic to allow the correlation matrices to be different for different groups, the results of applying the mixture maximum likelihood method of clustering with a common correlation matrix for all groups (and, hence, estimating less parameters in the model) may be quite informative. Therefore, the mixture method was applied under both the conditions of equal and unrestricted correlation matrices for the underlying populations. Both methods gave the same allocation of the lines to groups for $g=3$ and $g=4$, but there was a difference at the five-group level. Tests on the log-likelihood values indicated that a significant extra amount of the variation was being accounted for by increasing the number of groups to five, but because of the inconsistency of the membership at the five-group level and because of subjective assessment of the posterior probabilities, the four-group level was chosen as an appropriate representation of the data.

The four groups (Table 1 and Fig. 2) had, for each attribute, distinct properties and distinct patterns of response across the locations. The properties and response patterns for the groups reflected different selectional and genetic backgrounds of the entries within them. Group C is related to the deltapine germ plasm and has a yield advantage at all locations with moderate to good lint quality. Group D, which consists of namcala- and coker-derived entries, has moderate yield but excellent quality. Groups A and B did not possess good yield or quality characteristics. It is clear from this grouping of genotypes that all four attributes played a role in arriving at the group composition. The low micronaire at Emerald (Fig. 2) resulted from harvesting the trial when the cotton was immature due to a late season, but there is no explanation for the drop in lint strength at Theodore.

When clustering at the five-group level there is strong evidence that the genotype nam should form a separate, single-member group. Presently, this technique does not

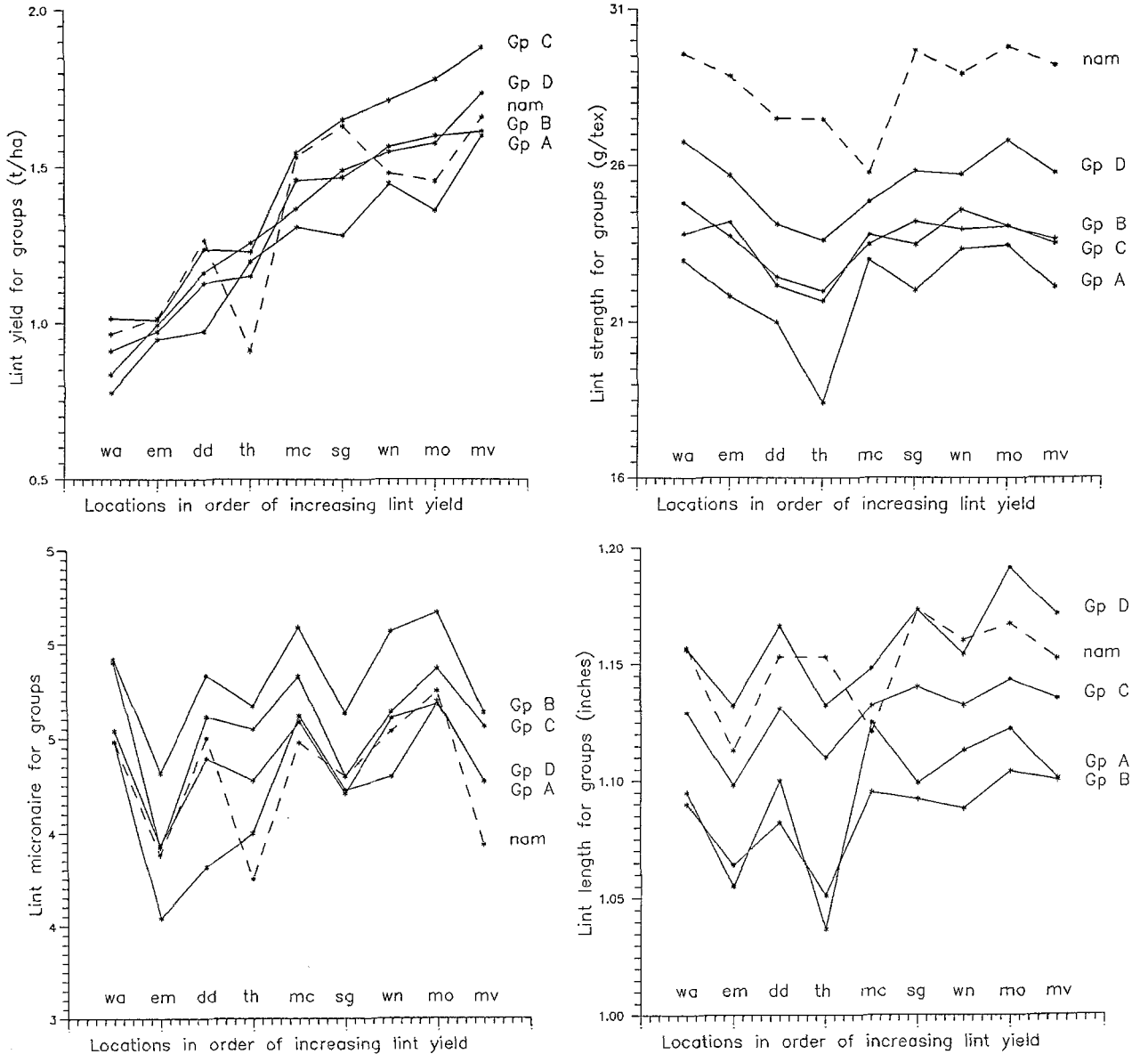

Fig. 2. The expected means for four groups (formed by mixture maximum likelihood) for lint yield and three lint quality attributes plotted against locations. The response for nam (namcala) has been added separately. (For environment abbreviations, see Fig. 1)

allow convergence to a single-member group, as no correlation matrix is estimable for a sample of size one. Using arbitrary correlation matrices, nam and m220 separated from the others, while with an assumed common matrix, nam and mo63 formed another group. The next best local maximum for these two conditions had mo63 with nam and m220 with nam, respectively. The closeness of the log-likelihood values for these local maximum solutions (606.7 compared with 601.0 for arbitrary correlation matrices and 500.2 compared with 488.6 for a common matrix) indicates that the $g=4$ solution was a very good summary of the data, while either of the $g=5$ solutions could be acceptable.

At the four-group level the only entry of any magnitude in the pooled estimate of the common correlation matrix (assuming equal correlation matrices for the underlying populations) is the positive correlation between lint yield and micronaire, which indicates that high yield tends to concur with coarse micronaire (Table 2). This was evident to varying degrees in the estimated correlation matrices for each group. The only other correlation estimates of any magnitude were 0.41 and 0.33 between length and strength and length and yield, respectively, in group B. The four attributes appear to contain relatively independent information about the size and quality of the harvest.

Table 2. Pooled estimate of the common correlation matrix from the cluster analysis

	Yield	Strength	Micronaire	Length
Yield	1.00			
Strength	−0.12	1.00		
Micronaire	0.33	−0.14	1.00	
Length	0.07	−0.06	0.04	1.00

Table 3. Components from three-mode PCA

A Environments (unit length)			
	E1	E2	R^2
Theodore	0.34	0.66	0.64
Emerald	0.26	0.44	0.49
St. George	0.37	0.15	0.71
Warren	0.34	−0.08	0.64
Myall Vale	0.36	−0.10	0.75
Darling Downs	0.36	−0.16	0.73
Mooree	0.39	−0.24	0.77
West Namoi	0.30	−0.31	0.66
Moomin Creek	0.24	−0.39	0.60
R^2	0.65	0.02	

B Lint attributes (unit length)				
	A1	A2	A3	R^2
Length	−0.65	0.37	0.45	0.80
Micronaire	0.33	0.44	−0.56	0.46
Strength	−0.67	0.01	−0.69	0.83
Yield	0.10	0.82	0.09	0.60
R^2	0.34	0.22	0.11	

Three-mode PCA

Three-mode principal component analysis is used, not only to give extra information on the relationships among attributes and environments in the way they describe the variability among genotypes, but also to enable a more detailed description of the relationships between the attributes and the clusters obtained with the mixture maximum likelihood method.

Following Kroonenberg and Basford (1989), the data were first corrected (centered) for the mean of each attribute environment (location) combination and then standardized (scaled) by the standard deviation for each attribute over all environments. In any analysis of multi-attribute or multienvironment data, careful consideration must be given to what, if any, and in what order centering and scaling are applied to the data (Harshman and Lundy 1984). Here the data were corrected because the relative performance of genotypes is of interest and not the overall differences between environments. The variability of the centered scores for each attribute was equalized so that each contributed equally to the analysis. Components were then computed for the genotypes, attributes, and environments.

A model that had three components for genotypes and attributes and two components for environments was considered adequate, as it accounted for two-thirds of the total variability in the data (overall R^2 between data and predictions estimated with the model was 0.67). The three components for the genotypes partitioned this variability (R^2) into 0.33, 0.23, and 0.12, respectively; those for the attributes into 0.34, 0.22, and 0.11; and the two components for the environments into 0.65 and 0.02 (Table 3). The results showed that there is considerable variability among the scores on their respective components for both the genotypes (not tabled) and attributes (Table 3 B). This is especially noticeable for the latter as there are only four of them. The relative independence of the attributes had already been expressed in the pooled estimate of the common correlation matrix (Table 2) in the cluster analysis at the four-group level. It is expressed here in that three components are needed to explain the differences among four attributes. Each of the three components expresses a different contrast (comparison) among the four attributes. In comparison, the scores for the environments are rather homogeneous (Table 3 A, first component), i.e., the patterns of the genotypes over attributes are rather similar across all environments, with somewhat lower values for Moomin Creek and Emerald. The major difference among the environments is between the central Queensland locations and the southern Queensland and New South Wales locations (Fig. 1 and Table 3 A, second component). This difference may be associated with temperature (day degrees) differences between the cotton growing regions.

An assessment of how well the variability of each genotype, environment, and attribute is accounted for by the model can be made using R^2 values. For instance, predicted values for genotypes dp80, m220, and 28/3 account for 20% or less of their variability while, on average, 67% of the variability was accounted for. All attributes and environments fit more or less equally, and thus contributed in a comparable manner to the solution.

The interrelationships between the different modes of the data given by the three-mode principal component analysis are portrayed as scatter diagrams by the use of joint plots (Fig. 3 and 4) or in tabular form by the "inner products" of the vectors in the reduced space (Tables 4 and 5). As described above, both attributes and genotypes are vectors from the origin, but as the relationship among the genotypes in terms of the attributes is being emphasized, only the attributes are shown as such. The strength of these relationships can be measured by the inner products between the vectors (Tables 4 and 5). The values of the inner products of the first and second joint plots may be directly compared, as they are presented on the same scale. For a single attribute, the sizes of these

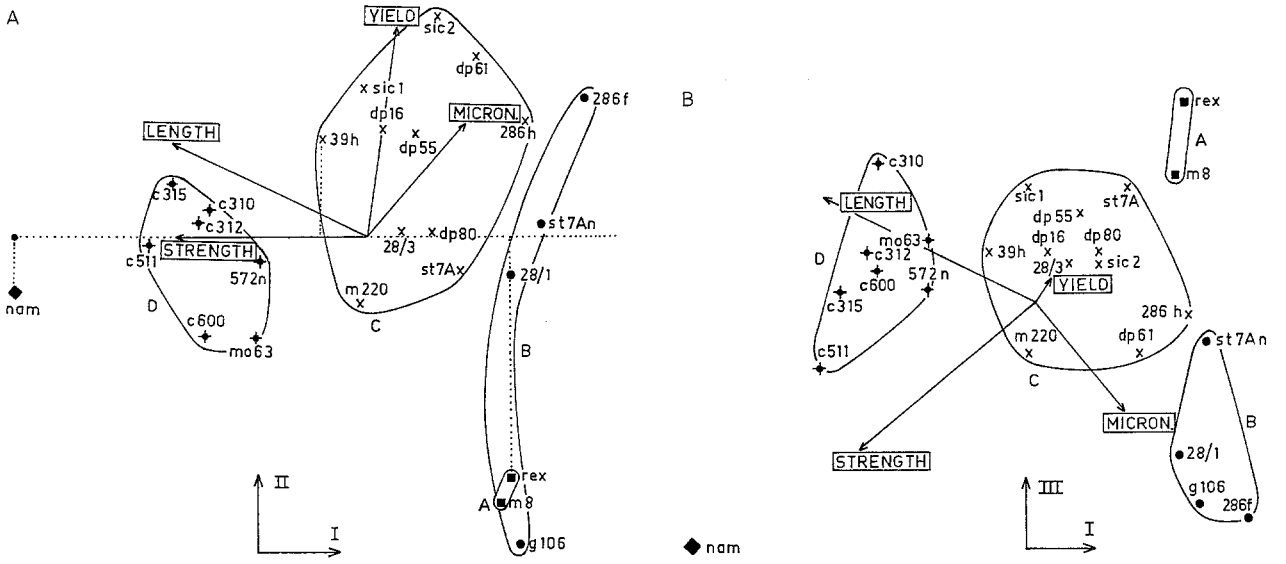

Fig. 3. **A** Joint plot associated with first environment component. Axis I versus Axis II. **B** Joint plot associated with first environment component. Axis I versus Axis III

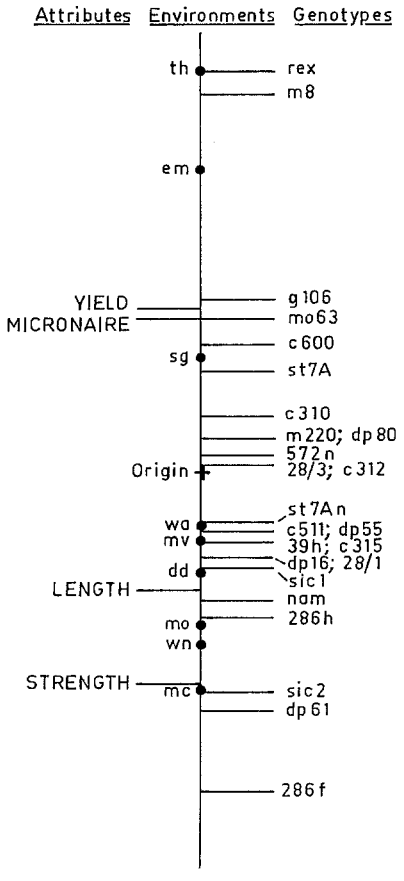

Fig. 4. Joint plot associated with second environment component – Axis I. (For environment abbreviations, see Fig. 1)

Table 4. Inner products between genotypes and attributes [a] (first joint plot)

Cluster	Genotype	Length	Strength	Micronaire	Yield
A	rex	−2.5	−4.6	−3.1	−3.5
	m8	−3.2	−3.5	−2.6	−4.1
B	g106	−6.3	0.5	0.4	−5.4
	286f	−4.1	−0.5	5.2	2.4
	28/1	−3.6	−0.2	2.3	−0.5
	st7An	−3.0	−2.0	1.9	0.4
C	sic2	0.9	−1.5	2.4	4.1
	dp61	−0.7	−1.0	3.1	3.2
	sic1	2.0	−1.3	0.4	2.8
	dp55	0.8	−1.9	0.5	2.2
	286h	−1.7	−2.2	2.5	2.2
	39h	1.8	0.0	0.1	1.6
	dp16	0.9	−1.0	0.7	1.2
	dp80	−0.6	−1.6	0.1	0.3
	28/3	−0.3	−0.9	0.0	0.2
	st7A	−0.9	−2.8	−0.6	−0.2
	m220	−0.8	0.8	−0.2	−1.3
D	c310	3.8	0.8	−2.3	0.4
	c315	3.6	3.1	−1.1	0.5
	c312	3.1	2.1	−1.6	0.1
	c511	2.8	4.2	−1.1	−0.7
	c600	2.0	2.1	−2.6	−2.1
	572n	1.6	1.7	−1.2	−0.6
	mo63	1.4	0.9	−2.4	−1.9
E	nam	3.1	8.6	−0.8	−2.1

[a] A value of zero means average on an attribute

Remarks: A – weak short lint, low yield, but fine micronaire; B – very short weak lint, coarse micronaire, mixed yield; C – good yield (especially first half of lines), reasonable micronaire, weak lint of mixed length; D – long strong lint, fine micronaire, average yield; E – low yield, long very strong lint, fine micronaire

Table 5. Inner products between genotypes and attributes [a] (second joint plot – Central Queensland versus Southern Queensland and New South Wales)

Cluster	Genotype	Length	Strength	Micronaire	Yield
A	rex	−0.9	−1.5	0.8	0.8
	m8	−0.9	−1.4	0.8	0.7
B	286f	0.1	0.9	−0.8	−1.2
E	nam	0.7	1.0	0.2	0.3

[a] Only those genotypes listed with at least one value $\geq|0.8|$
Remarks: rex and m8 – stronger/longer with higher yield and coarser micronaire in south than in north; 286f – stronger lint but finer micronaire and lower yield in north than in south; nam – stronger in north than in south

inner products with the genotypes and, therefore, the strength of the relationship are directly proportional to their projections on the attribute vector. Therefore, these projections can be used to compare the importance of an attribute for a genotype or cluster of genotypes. As an example, the projections for nam (namcala), 39h, and rex on lint strength are shown in Fig. 3A. Clearly, nam has considerable lint strength, rex has little lint strength, and 39h has nearly average lint strength, as it projects nearly into the origin.

The clusters derived by the mixture method are also indicated in the joint plots for the first environment component (Fig. 3A and B), except that nam (namcala) is isolated from cluster D (and referred to as a single-member cluster E), as it seems to be rather far away from the other genotypes in that cluster (see also Fig. 2). As mentioned before, a joint plot can be made for each component of the environments. As all environments have approximately equal loadings on their first component (0.33 ± 0.05), the inner products (Fig. 3, Table 4) are of equal value to these environments, which means that they indicate what the environments have in common. On the other had, the line plot (i.e., one-dimensional joint plot) of Fig. 4 associated with the second environment component shows how certain relationships between attributes and genotypes are different for the environments, in particular with respect to the central Queensland locations, Emerald and Theodore, and the Namoi, Gwydir locations, West Namoi, Moomin Creek, and Mooree (Fig. 1 and Table 3). The joint plot for the second environment component is one-dimensional, as the second and third axes for attributes and genotypes are effectively zero in length.

The major conclusions from the first environment component (Fig. 3A and B, Table 4) are as follows.

(1) The major differences between clusters are associated with varying lint strengths, i.e., namcala has stronger lint than cluster D (namcala- and coker-derived varieties), which are stronger than cluster C (primarily related to deltapine germ plasm), which are stronger than clusters A and B.

(2) There is a difference within the clusters associated with lint yield with, on the average, slightly higher yield for cluster C compared to D, and particularly low yields for rex, m8, and g106.

(3) Cluster D is distinguished by its long lint and fine micronaire, while cluster B has coarse micronaire and short lint.

(4) Namcala is different from the other cluster D genotypes because it is so strong. These results obviously confirm the cluster analysis conclusion about the properties of the cluster D genotypes, but they also provide additional information, e.g., that within this cluster, c310 probably has the best combination of attributes.

The genotypes, m8 and rex, dominate on one side of the line plot of the second environment component (Fig. 4), and 286f, dp61, and sicot2 dominate on the other side. To facilitate the interpretation of this plot, the positions of the environments have been indicated, i.e., all three vector plots can be superimposed. The interpretation proceeds as in a two-dimensional plot, but the inner products (Table 5), which represent the strength of the relationships, are now simply the product of the vector lengths. Furthermore, each inner product of a genotype and attribute should in turn be multiplied by the vector length of an environment. High positive values of these triple products indicate that the particular combination has a high, above average score. Thus, at Theodore (and Emerald), rex (and m8) had, relatively speaking, higher yields and coarser micronaire (the product of these vector lengths is positive; they are all on the same side of the axis), but at the other locations, rex (and m8) had comparatively lower yield [the product is negative; loc (−), rex (+), yield (+)]. This is confirmed from the values of the inner products (Table 5). Theodore and Emerald, m8 and rex, and lint micronaire and yield are all on the same side of the axis, while the other environments are on the other side (Fig. 4). This indicates that m8 and rex had relatively coarser micronaire and higher yield in central Queensland compared to the other locations, and that they had rather weaker and shorter lint in the central Queensland locations. The reverse patterns are present for 286f, dp61, and sicot2, as they have relatively stronger and longer lint with finer micronaire and higher yields in the southern locations compared to the more northern ones.

Discussion

The information obtained from the joint analysis of the data from the Australian Cotton Cultivar Trials can be summarized as follows.

(1) Both the obtained clusters and the three-way principal component analysis gave a sensible and useful integration of the data from this regional variety trial. However, considerably more detail and interpretation were available through the complementary use of the two methods, especially in examining the relationship among, and the variation within clusters. This addresses the practical problem for plant breeders that, although such clusters are easier to look at than many individual lines, selection has to be made for individual lines.

(2) The methods have successfully integrated the yield and quality data. Using yield as the sole criterion, the excellence of the namcala and coker group for quality is overlooked. The analyses point to a decision in favor of either high yields of moderate to good quality lint or moderate yield but superior lint quality.

(3) Namcala deserves special consideration. It has especially strong lint and is among the best lines for long lint and fine micronaire. Namcala is included in the trials as a benchmark for high quality lint. However, it just does not yield enough to be a viable proposition. The dp61 and sic2 quality is "good enough" for most "good" quality cotton.

Before genotypes are entered in the ACCT, they have been previously tested in trials at two to three locations for approximately 2 years. These data together with the ACCT data are used to select entries for the next year's trials. From the above analyses the "best" members from cluster C would be selected on high yield and adequate quality, and the best from cluster D on the basis of good quality and reasonable yield, and namcala would be retained for its outstanding quality. In fact, all of the higher yielding members (Fig. 3 A) of C (sic1, sic2, dp16, dp55, dp61, and 39h) except 286/h were selected. This entry was rejected because it has a hairy leaf character that produces poor quality cotton. M220 was selected as it was the best of the early maturing lines in the trial. C315 and c310 were selected as the best of the coker lines and this corresponds to the analyses described here (Fig. 3 A). Mo63 was selected as the best quality line from the coker group and for its high yield at Emerald. This was not confirmed in subsequent trials and this entry was dropped from the ACCT after one more year's trials. Namcala was retained as a benchmark for quality and 286f was retained as it was the best of the lines with a genetic character, frego bract which, it was hoped, confers some resistance to insect attack. In consequence, these analyses represented the data in the way that they were seen by the breeders who conducted the trials. Differences occurred where extra information not available to the methods influenced the decision of the plant breeders.

The present description of the application of a cluster analysis technique and three-mode principal component analysis looks reasonably straightforward. However, this is not completely the case, as we have not mentioned several technical details. For example, the mixture method of clustering is applied via the EM algorithm introduced by Dempster et al. (1977). It is an iterative technique which is repeated for various starting values in an attempt to locate all local maxima of the likelihood, but the global maximum is not necessarily obtained. In this case, a satisfactory solution was obtained by using the results of hierarchical clustering techniques on individual attributes at the appropriate group level as initial allocations for the mixture approach. Basford and McLachlan (1985 b) detail some of the problems with the non-uniqueness of the solution in the two-way situation.

Similarly, testing for the actual number of clusters from which the sample is drawn is an important but difficult problem. McLachlan and Basford (1988) discussed this at some length and recommended the adoption of the likelihood ratio criterion for testing the hypothesis of g_1 versus g_2 groups ($g_1 < g_2$) as suggested by Wolfe (1971). This is only an approximation and should not be rigidly interpreted, but rather used as a guide to the possible number of underlying groups. Examination of the estimated posterior probabilities of group membership for the genotypes for values of g near to the value accepted according to the likelihood ratio test can be useful in leading to the final decision on the number of groups, but this seems more reliable for two-way rather than three-way data.

With respect ot the three-mode PCA, few technical issues arise, apart from an adequate choice of the number of components in all three modes. Interpretation of the results is not always easy, especially in the initial stages when acquiring experience with the technique. However, several guidelines are contained in Kroonenberg (1983) along with worked examples.

The clustering of three-way data is described in detail in Basford and McLachlan (1985 a) and McLachlan and Basford (1988). The latter reference contains the listing of a FORTRAN program to perform the required calculations on a mainframe IBM machine. On request, K. E. Basford will supply a copy of the program, along with sample input and output files, on floppy disk suitable for a mainframe machine or a personal computer running MS-DOS. Kroonenberg and Basford's study (1989) contains an in-depth example of the application of three-mode PCA of a plant breeding experiment on soybeans. The program used is documented in a manual by Kroonenberg and Brouwer (1985) and is available from P. M. Kroonenberg in a form suitable for running on mainframe machines.

The major advantage of these methods is that they allow the data set to be treated in the form of a three-way array. An overall picture of response is obtained and, in the case of the mixture approach, used to allocate the cotton lines to groups. The important genotype by environment interaction present in such trials is incorporated

directly into the underlying models. Similarly, the representation of the cotton lines in a reduced space allows a quicker appreciation of the major differences inherent in the data. The three-way PCA allows possible structure in the environments and attributes to be extracted. The techniques provide complementary information that can be readily displayed in common figures. They are useful, reasonably easy-to-apply techniques which should be commonly employed in the statistical analysis of such three-way data.

References

Basford KE (1982) The use of multidimensional scaling in analysing multi-attribute genotype response across environments. Aust J Agric Res 33:473–480

Basford KE, McLachlan GJ (1985a) The mixture method of clustering applied to three-way data. J Classification 2:109–125

Basford KE, McLachlan GJ (1985b) Likelihood estimation with normal mixture models. Appl Stat 34:282–289

DeLacy IH (1981) Cluster analysis for the interpretation of genotype by environment interaction. In: Byth DE, Mungomery VE (eds) Interpretation of plant response and adaptation to agricultural environments. Queensland Branch, Australian Institute of Agricultural Science, Brisbane, pp 277–292

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38

Gabriel KR (1971) The biplot graphical display of matrices with applications to principal components. Biometrika 58:452–462

Harshman RA, Lundy ME (1984) Data preprocessing and the extended PARAFAC model. In: Law HG, Snyder CW Jr, Hattie JA, McDonald RP (eds) Research methods for multimode data analysis. Praeger, New York, pp 216–284

Kempton RA (1984) The use of bi-plots in interpreting variety by environment interactions. J Agric Sci 103:123–135

Kroonenberg PM (1983) Three-mode principal component analysis: Theory and applications. DSWO Press, Leiden

Kroonenberg PM, Basford KE (1989) An investigation of multiattribute genotype response across environments using three-mode principal component analysis. Euphytica 44:109–123

Kroonenberg PM, Brouwer P (1985) User's guide to TUCKALS3 (version 4.0). Technical report, University of Leiden, Department of Education

Kroonenberg PM, De Leeuw J (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 45:69–97

McLachlan GJ, Basford KE (1988) Mixture models: Inference and applications to clustering. Marcel Dekker, New York

Reid PE, Thomson NJ, Lawrence PK, Luckett DJ, McIntyre GT, Williams ER (1989) Regional evaluation of cotton cultivars in eastern Australia 1974–1985. Aust J Exp Agric 29:679–689

Tucker LR (1966) Some mathematical notes on three-mode factor analysis. Psychometrika 31:279–311

Williams WT (ed) (1976) Pattern analysis in agricultural science. Elsevier, Amsterdam

Wolfe JH (1971) A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2. U.S. Naval Personnel and Training Research Laboratory, San Diego