Euphytica **84**: 31–42, 1995.

Three-way cluster and component analysis of maize variety trials

P.M. Kroonenberg[1], K.E. Basford[2] & A.G.M. Ebskamp[3]
[1] *Department of Education, Leiden University, The Netherlands;* [2] *Department of Agriculture, University of Queensland, Australia;* [3] *CPRO-DLO, Wageningen, The Netherlands*

Received 28 February 1994; accepted 25 April 1995

Key words: cluster analysis, maize, mixture methods, plant breeding, three-mode component models, variety trials, *Zea mays*

Summary

Data from the Dutch Variety List Trials for maize were analysed with three-way mixture method clustering and three-mode component analysis. The main objective of the paper is to demonstrate the usefulness of such multivariate analysis techniques for plant breeding data. In particular, attention is paid how one may gain insight into the complex patterns that are embodied in this type of data sets.

Introduction

All across the world plant breeders and testing authorities are engaged in trials to assess varieties of commercial crops. In such variety trials they seek to assess which varieties do well both in terms of quantity and quality in different environments. After initial processing, data from such trials often have the form of scores of Varieties or genotypes on several Attributes collected from different Locations during a number of Years. Such data can be referred to as four-way data, because they can be arranged as a four-way data block of Varieties by Locations by Years by Attributes. Given a reasonable number of levels or entities in each way, the number of data points in such data sets can be quite large, and they generally defy simple inspection. What is required to examine the data and the patterns contained therein, are techniques which handle not just ordinary two-way matrices, but which do justice to the multivariate form of the data.

The methods to be presented in this paper are the three-way mixture method of clustering, as fully described in Basford & McLachlan (1985), McLachlan & Basford (1988) and three-way (or three-mode) principal component analysis, see e.g. Kroonenberg (1983, 1988, and their references). Previous analyses of variety trials, especially soybeans and cotton, can be found in Kroonenberg & Basford (1989) and Basford et al. (1990). The presentation of the methods in this paper will primarily be at a conceptual level. However, the technical details of the methodology, can be found in the papers mentioned above.

Data description

The original data set of the Dutch Maize Variety Trials consists of phenotypic mean values per location, and it is a four-way data set of Varieties by Years by Locations by Attributes with a considerable number of missing data due to additions and deletions of varieties over the years, and due to unavailability of some information at specific times for specific locations. To allow a fairly straightforward illustration of the techniques, a special subset was constructed from the available information, so that the data set to be analysed did not contain any missing data. In a later paper, we hope to discuss and illustrate the handling of missing data in variety trials. The three-way data set was obtained by taking averages over the last seven years of the trials (1984–1990). From the original four-way data set, another selection was made in the companion paper by Van Eeuwijk et al. (1995), who analysed variety by environment (year × location) interactions for single attributes with bilinear models and restricted maximum likelihood methods (REML).

Our subset consisted of 14 varieties (Brutus, Splenda, Markant, Vivia, Dorina, Irla, Clipper, Gracia, Sonia, Ascot, LG 20.80, Scana, Presta, and Sogetta). All these 14 varieties were an improvement over existing ones, therefore they all were selected for entering in the Variety List of Field Crops, and they have been in the trials for several years. The varieties were planted in four regions of the Netherlands characterised by the soil and location (Southern Sand, Central Sand, Northern Sand, and River Clay).

As considerable resources have been invested in measuring several attributes on these varieties and the attributes are usually correlated with one another, it would appear that single-attribute analyses would not be sufficient. In most crops, there is one attribute of overriding importance – usually the one which determines the economic return to the farmer – and others of (slightly) less importance. In these data, for instance, the Dry Matter Yield would provide the economic return, but Digestibility would also be an important consideration. The final selection of six attributes included in the analysis consisted of Dry Matter Content (%), Dry Matter Yield (kg/100 m^2), Plant Height (cm), Early Vigour (1–9), Ear Height (cm), and Digestibility (gr/kg).

Although the years might not be intrinsically interesting, they provide the environmental conditions under which the crop is grown. As is explored in the companion paper by Van Eeuwijk et al. (this issue, pp. 9–22), the yearly variability can be related to weather characteristics which are important for the outcomes of the trials. However, if a decision has to be made to reduce the original four-way array to a three-way one, it would seem most reasonable to average over the reduced subset of recent years or else to consider each location-year combination as an environment; for an example of the analogous situation going from three to two ways see Freeman (1975). In the present case, there were insufficient data for the construction of a three-way variety by environment by attribute array, so that it was decided to take the former option, so that, after averaging over years, we ended up with a three-way array of 14 Varieties by 4 Locations by 6 Attributes without missing values.

Three-way methodologies

Cluster analysis

The main aim of using a cluster technique in the analysis of data from plant breeding trials is to group the varieties into several homogeneous groups such that those varieties within a group have a similar response pattern across the locations. It is reasonable to suppose that all the varieties in the trials will not behave completely independently of one another. For instance, those with similar genetic make-up would be expected to behave similarly. If the entire data array containing information on a (usually large) number of varieties can be reduced to the information on a (usually much fewer) number of groups of varieties (within which the varieties have a similar response pattern), then the task of the plant breeder to interpret the information is much simpler.

A hierarchical clustering methodology was successfully implemented for two-way genotype by environment data for a single attribute by Burt et al. (1971), Mungomery et al. (1974), Byth et al. (1976), Corsten & Denis (1990). The summarization of the information on the individual genotypes by the genotype groups proved very useful to plant breeders. It was deemed essential to retain information on the separate environments (or environment groups, if the clustering was separately applied to the environments) because of the large genotype by environment interaction usually displayed in such data.

However, consider the situation when several attributes have been measured on the varieties in these plant breeding trials. Rather than performing a separate clustering of the varieties for each attribute, it would be more advantageous to perform a clustering of the varieties on the basis of all the information available, i.e. all the attributes measured on the varieties and considered important enough for inclusion.

It appears that the first attempt at developing a clustering technique directly applicable to three-way agricultural data was by Basford & McLachlan (1985) who considered the normal mixture model. They employ a non-hierarchical clustering technique where a specified number of groups of varieties is assumed to adequately summarize the information collected on the individual varieties. More formally, it is assumed that the varieties have been selected from a mixture of several populations (or groups) each with their own particular response pattern across locations. The response pattern for each group is estimated from those varieties belong-

ing to the group and will consist of a mean for each attribute for each environment. Because a group mean is determined for each location, the underlying model allows for groups to do well in some locations, but not in others. This incorporates the possibility of variety by location interaction – an important consideration in variety trials where plant improvers have found considerable evidence of such interaction. The best way to portray the group response pattern is by plotting the group mean for each location for each attribute.

From a practical viewpoint, another important consideration is that the underlying model does not specify the same correlation structure among the attributes within each group. Thus there can be a strong correlation among two attributes within one group, no correlation among them in another group and possibly a strong negative correlation among them for a further group. However, the model explicitly assumes that this correlation structure among the attributes within a group is consistent across the locations.

The clustering procedure is an iterative one in that it starts from a particular grouping of varieties and moves the varieties among the groups until it finds the most satisfactory grouping under the mixture model (using a likelihood criterion). The estimation of a suitable allocation of the varieties into groups takes into account both the mean response pattern across locations and the correlation structure among the attributes within each group. Full details of the underlying statistical model and fitting algorithm can be found in Basford & McLachlan (1985) and Basford et al. (1991).

Ordination

The cluster analysis assumes that the varieties have been selected from an underlying smaller number of groups of varieties, and determines the parameters of these underlying groups. Data reduction and summarization is achieved by representing response patterns for the groups. In this way plant breeders only have to look at the response patterns of a limited number of groups, rather than at those of all individual varieties. Even though the cluster technique uses all the data in the three-way data box, in considering its results one still has to look at the attributes one by one. In other words, the interpretation is essentially univariate, rather than multivariate. Only by comparing the patterns across the different attributes, which is not necessarily easy to do, can one effectively gauge to what extent the varieties exhibit the same patterns across attributes. Another limitation is that the information on the location (or environment) differences has to be interpreted from the response plots, rather than being summarized in a convenient way. Against these disadvantages, its strong advantages must be pitted: In particular its simplicity of comprehension and interpretation.

To provide summarizations of the varieties, the attributes, and the locations, one may use three-way (or three-mode) principal component analysis, as we will illustrate below. This ordination technique aims to provide these three summarizations simultaneously by modelling the analogues of components in the two-way case, but now for each of the three ways. For example, the model provides components for attributes, which indicate to what extent the attributes are alike (across varieties and locations) and to what extent they are different. One way to convey this information is by simply listing the coordinates of the attributes on the components in a table, but one may also use these coordinates to construct a two-, three-, or higher dimensional plot, depending on the number of components. Similar information for the locations and the varieties is produced by the technique, so that one ends up with a table or plot for each way. The choice of the number of components in each of the ways is not always obvious, but is largely guided by the variability accounted for by each of the components. A certain amount of subjective judgement comes into this as well.

The original observations are (mean) scores of varieties on attributes from different locations, and thus the scores represent a relationship between the three ways. It seems natural to expect that in one way or another there should exist some similar connections between the components. In fact, the three-way principal component model also contains this information. In particular, there are values or parameters in the model, which indicate the strength of the relationships between components of the three different ways; these parameters are called core elements, and they are collected in a so-called core array or core matrix. Suppose that there are two components for the locations, three components for the varieties, and three for the attributes, then the model has $2 \times 3 \times 3$ elements in the core matrix, each of which indicates the strength of the relationship between three components. These parameters can be scaled in such a manner that they indicate the amount of variability accounted for by a particular combination of components, so that one can assess their relative importance.

In the present paper, we will not discuss their interpretation directly but use these core elements to con-

Table 1. Clustering of the maize varieties into three groups by the mixture likelihood approach

Group	Solutions		
	1	2	3
A	Brutus, Vivia, Dorina, Irla, Clipper, Presta	- Presta	- Presta & Clipper
B	Splenda, Markant, Gracia, Ascot, Sogetta		
C	Sonia, LG 20.80, Scana	+ Presta	+ Presta & Clipper

struct a special joint plot, which will aid our understanding of the relationships between the components. Details on the construction of these plots can, for instance, be found in Kroonenberg (1983, p. 164, 165).

Application

Cluster analysis

The most usual strategy in searching for a satisfactory clustering is to start with all the varieties in one group and then to increase the number of groups until an adequate description of the data is achieved. Given the relatively small number of varieties being clustered here, it seemed appropriate to select a three-group solution as a summarization of the information in the variety trials. Accordingly, the best three-group solution under the mixture model was determined (Table 1) with the groups being arbitrarily coded as A, B and C.

This partition of the varieties into three groups has some correspondence to date of flowering which was not part of information used to form the clusters. The varieties in Group B have mid-early to mid-late flowering, those in Group A mid-early to early and those in C have early to very early flowering.

It should be noted, however, that this is not the only partition of the varieties into three groups which is consistent with the data. In that sense, the grouping presented should not be considered to be the only acceptable arrangement. Two solutions which gave likelihood values very close to the global maximum have Presta in Group C and both Presta and Clipper in Group C, respectively. For these data, the starting allocation of the varieties into groups strongly influences the final partition obtained.

A further point is that Group C consists only of three members, which limits the generality of the statements that can be made for this group. Moreover, if there are large differences within the group the group results might not be representative for any of the varieties in the group. With larger number per group this is obviously less of a problem.

Once a satisfactory grouping of the varieties has been determined, it is necessary to look at the response pattern for each group to best understand the similarities and differences among these groups. To achieve this, the estimated mean response in each location for each attribute has been plotted (Fig. 1). Each attribute is portrayed on a separate graph, and on each graph the locations are ordered by increasing overall mean for Dry Matter Yield. Equal spacing has been used on the horizontal axis for clarity, although the spacing is often dependent on the actual means.

In order to provide an indication for the variability associated with each point, bars have been placed at the side of each graph to represent the within-group variation. The lengths of these bars correspond to 1.5 times an estimated standard error of the group mean. These have been estimated using the square root of the estimated group variance divided by the number of varieties allocated to the group. This should only be considered as an approximate value which probably underestimates the true value. However, it does prevent one from drawing too much inference from the apparent difference among the group response patterns. If the bars about two group means at a particular location overlap, then these means would not be significantly different. The first five attributes in Fig. 1 show no significant location × group interaction. Clearly, Group A varieties are low on all five attributes, Group B varieties are tall plants with high ears, but low dry matter content with much early vigour. Group C varieties are relatively small plants with high dry matter content and yield. Figure 1 also clearly shows that Digestibility is the only attribute on which there is noticeable location × group interaction. Given that Digestibility is

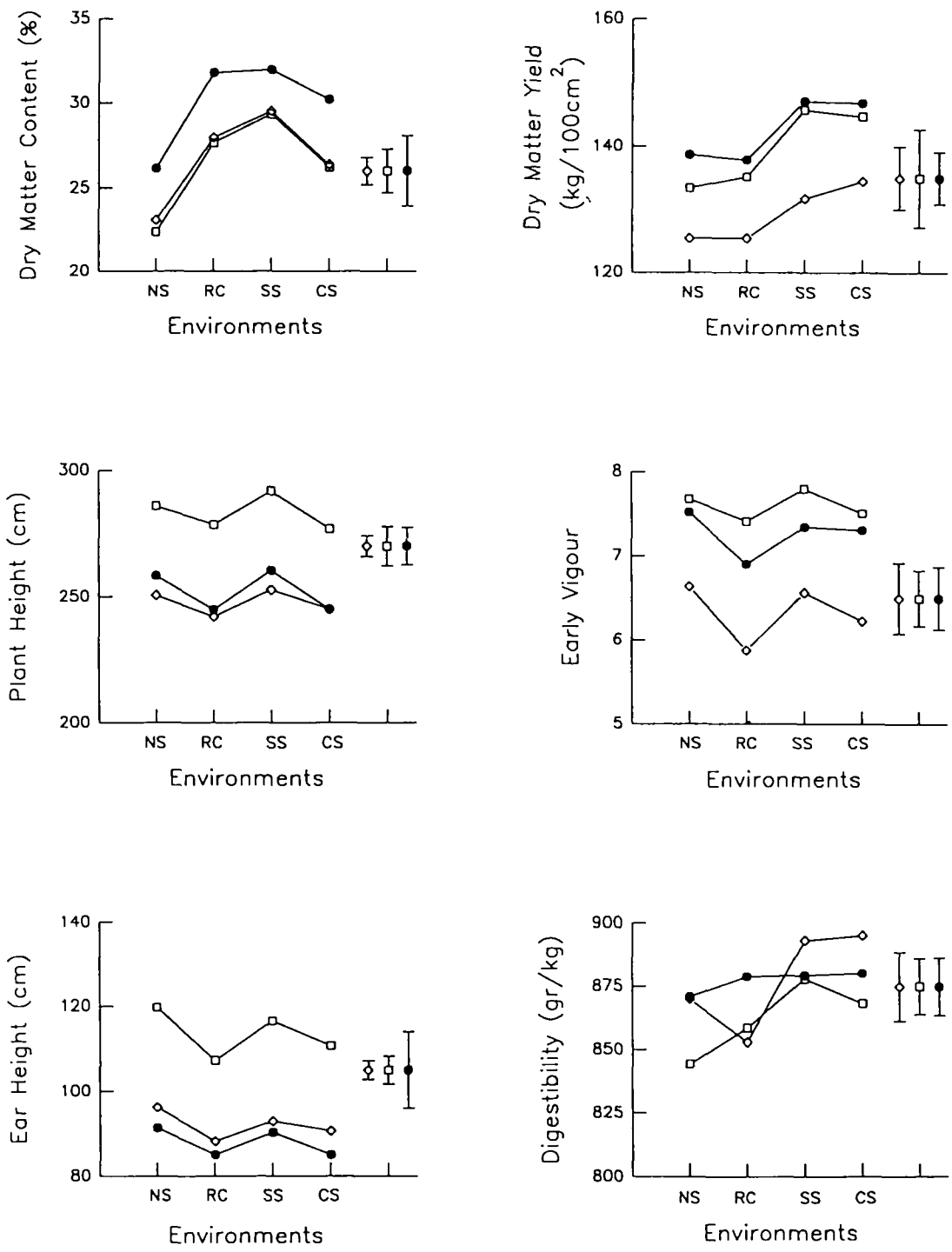

Fig. 1. Estimated mean for each group (◇ Group A, □ Group B; • Group C) for each attribute for each environment (NS = Northern Sand, RC = River Clay; SS = Southern Sand; CS = Central Sand).

an attribute of increasing importance to maize breeders, this interaction suggests that further investigation might be warranted.

The other information provided on the groups by the mixture method of clustering is an estimated correlation matrix among the attributes for each group. In the present case, one needs to interpret such matrices

with considerable care, particularly as only relatively few varieties are contributing to the calculation, be it that each variety is measured in four locations. That these matrices are determined by pooling the average squared deviation about the group means within each environment definitely improves the situation. Given the small number of varieties in each group and the ensuing unreliability of the estimates, these correlation matrices will not be discussed here (see Kroonenberg et al., in press, for a situation where they could be interpreted).

Table 2. Location components[a]

Location	1	2
Southern Sands	**0.94**	- 0.08
Central Sands	**0.86**	**- 0.23**
Northern Sands	**0.92**	- 0.04
River Clay	**0.90**	**0.33**
Explained Variability	82.3%	4.2%

[a] Components have been scaled as 'loadings' in two-way PCA [i.e. length = $\sqrt{}$ (no. of levels * proportional fit) = $\sqrt{}$ 4* (0.823 or 0.042)].

Ordination

Because we intended to use the ordination for a parsimonious description of the data, a solution with a limited number of components was sought for the three-way principal component analysis. The explained variability and the sizes of the elements of the core matrix can provide guidance, but the detail of description desired should be considered as well. In addition, substantive, interpretational criteria should play an important role as well. Here, we selected a solution with an overall fitted variability of 86%, which is partitioned by the three variety components as 50%, 29%, and 7%, by the three attribute components as 50%, 26%, and 10%, and by the two location components as 82% and 4%.

Locations. From the above percentages it is clear that the first location component is by far the largest, and that the contribution of the second component is much smaller. From Table 2 we see that the first location component is nearly constant. In other words, the first component indicates that the varieties react essentially similar on the attributes in the four Dutch locations. The size of the second component is anything but impressive, on the other hand it does describe whatever the most important, be it small, differences in location are. The primary difference is between the varieties grown on River Clay and those on the Central Sands. Which attribute-varieties combination are responsible for this difference will become clear when we examine the joint plots. If these (small) differences are not of interest, one could have selected a single component for the locations, or possibly have averaged over locations. The disadvantage is, of course, that differential location information can be of prime importance to the farmer when choosing from variety lists.

Varieties. In Fig. 2 the three-dimensional representation is given of the three components for the varieties. The lengths of the components have been scaled as is common in two-way principal component analysis, i.e. they have unit lengths and they are centred. (The latter is due to the original centring of the data.) In the figure the optimal clustering of the varieties is drawn, but one can see that from the point of view of the continuous representation the other two clusterings mentioned would have been acceptable. Note that whichever way defined, the clusters are well separated and it is possible to connect the outer-most points of each cluster without overlap. The first two dimensions are clearly related to the cluster separation, while the third is not, be it that it provides the reason for the non-uniqueness of the cluster solution, in the sense that on the third dimension Brutus and Presta are similar, but both are dissimilar with Clipper, while they are close together in the first two dimensions. These contrasts make it difficult for the clustering method to unequivocally allocate Presta and Clipper to cluster A or cluster C.

In the figure an arrow has been drawn from roughly South-West to North-East to indicate the direction of later flowering. This arrow indicates that the (spatial) arrangement of the varieties can be partially explained by the data of flowering. Methods have been developed to include such external variables as flowering date into three-way analysis (see Franc, 1992), in a similar way as was done with factorial regression in the companion paper by Van Eeuwijk et al. (this issue, pp. 9–22). To get a more complete picture of the similarities between the varieties as depicted in the figure we have to relate the spatial arrangement of varieties with that of the attributes.

Attributes. The attributes are depicted in Fig. 3. The attributes have been drawn as arrows to emphasize their difference from varieties. It also helps to provide a proper impression of the relationships among them.

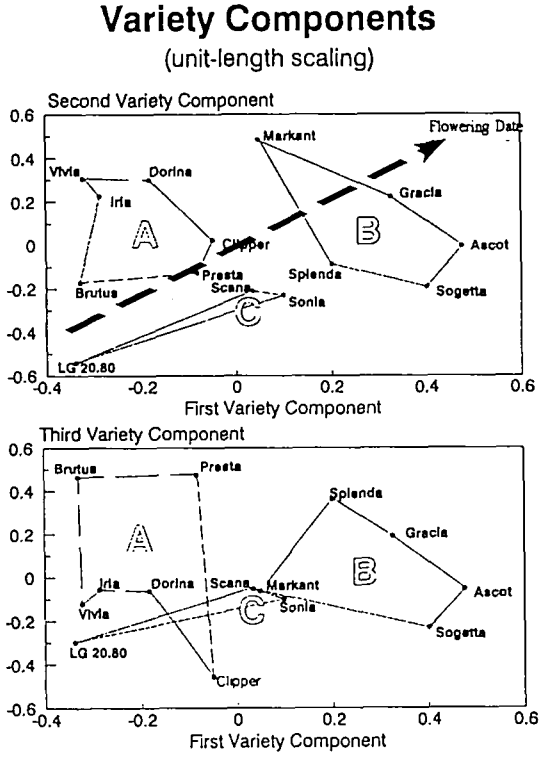

Fig. 2. Components for the varieties from the Three-mode principal component analysis (Top: First against second component; Bottom: First against third component). The varieties residing in the same cluster have been joined. The arrow in the top part indicates the direction of increasing flowering date.

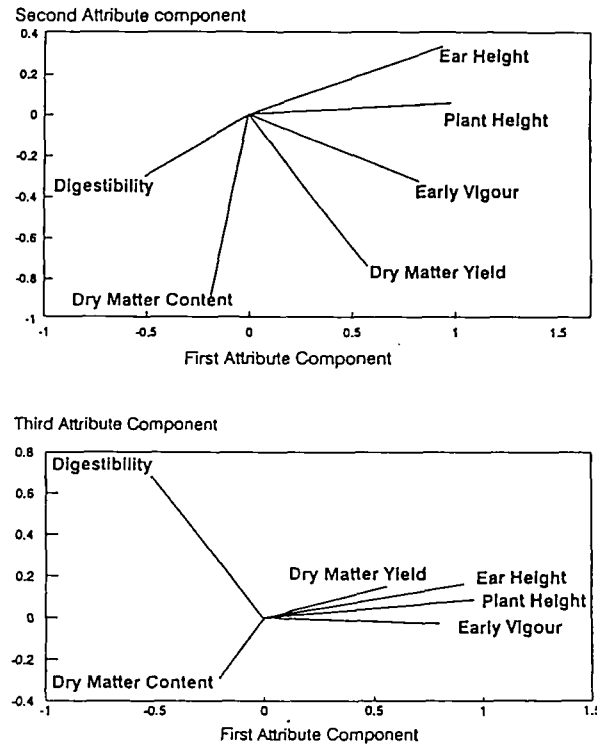

Fig. 3. Components for the attributes from the Three-mode principal component analysis (Top: First against second component; Bottom: First against third component). The lines connect the origin with the location of each attribute in the plots.

In particular, the angles between arrows have a direct relation with the correlation between attributes in as far as they are represented in this figure. Small, acute angles indicate highly positive correlations, perpendicular arrows indicate lack of correlation, and large, obtuse angles indicate highly negative correlations. Given that the representation is three-dimensional, a bit of care should be exercised in judging the correlation from the figure.

Looking only at the spatial arragement of the attributes, it can be seen that Ear Height and Plant Height are highly related, as their arrows have a very acute angle in three-dimensional space. On the other hand, the correlation of Dry Matter Yield and Dry Matter Content is rather low, because of their near 90° angle.

Core matrix. The core matrix is presented in Table 3. Due to the severe rounding it is obvious that there are only four important elements in the core matrix. The core matrix represents a further partitioning of the fitted variability. In particular, the combination of the first components of each of the ways explains 49%, that of the second components of the varieties and attributes with the first of the locations 26% and the combination of the third components of the varieties and attributes with the first of the locations 6%. The only important combination connected with the second location component, which reflects the differences between the locations, is that of the third component of the varieties with the second of the attributes and the second location component. As we have not explicitly given names (or an interpretation) to the components of the varieties or attributes, it is very difficult to indicate the meaning of the particular combinations of components. In some data sets, interpretations of components are available and then a direct interpretation of the core matrix is feasible. However, of all the components in the present analysis only the two location components have a clearly defined meaning (see above), and thus a direct interpretation of the core elements can not be given. This means that we will have to find another way of interpreting the relationships in the core matrix.

Table 3. Core matrix (percentages[a])

Location component 1 (82%)		Attribute Components		
		1	2	3
Variety	1	**49**	0	0
Components	2	0	**26**	0
	3	0	0	**6**
Location component 2 (4%)				
Variety	1	0	0	0
Components	2	0	0	**2**
	3	0	0	**0**

[a] All percentages smaller than 1 have been indicated as 0.

Joint plots. Varieties and Attributes. Graphics are ideally suited to acquire insight in the interplay between the components of the various modes, but making a single graph containing three different kinds of components is not an easy matter. Here we have chosen to present a joint plot for the varieties and the attributes, one corresponding to each location component. The first plot shows how the varieties and attributes are related for the first location component and it indicates that the locations are not much different from each other (see Table 3). The plot for the second, much less important, location component indicates that there are some differences between the River Clay location and the Sands, especially the Central Sands. This implies that the joint plot for this component displays those attributes which have a different reaction pattern for varieties in the River Clay location compared to the Sands. However, this effect is small (4%) compared to the overall similarity expressed via the first component (82%).

Common patterns for all locations. In Fig. 4 the joint plot is presented corresponding to the first location component. In the same plot all positions of the varieties and the attributes are portrayed, so that it is possible to describe in which sense the varieties have different or similar scores on the attributes. A problem is the three- dimensionality of the figure, but by presenting the axes in pairs, we hope to have given sufficient insight about the relative positions. The interpretation of such plots have been set forth, for instance, by Gabriel (1971, 1981), and is based on the inner (scalar) product relationship between varieties and attributes. In Table 4 the relevant inner products have been given, arranged in such a way that the table can be fairly easily read. The information in the table and the figure are essentially the same. For instance, the table shows that Sogetta has comparatively the highest Dry Matter Yield, followed by Ascot, while Dorina, Irla, and Vivia have the lowest values. In the figure this is displayed by the projections of the varieties on the Dry Matter Yield vector (here only indicated for Sogetta). The point in the middle of the plot represents a (non-existing) variety which has average scores on all attributes. Whereas the table has the advantage of numerical precision, the figure has the advantage of a general overview. Careful scrutiny of the information will show that the patterns bear a close relationship with the estimated cluster means from the previous section. What is added is the behaviour within the clusters. For instance, the rather different pattern of LG 20.80 in cluster C, and that of Markant in cluster B. Figure 4 clearly shows their extreme position in their group. Given the limited number of varieties, there is no way the cluster analysis could have created a special single member group for either of them. The table also shows the correlation patterns mentioned in the previous section, for instance the high positive correlation between Plant Height and Ear Height. The overall strong similarity between the cluster analysis and the joint-plot and inner-product information related to the first location components is primarily caused by the very strong similarity of the locations or the limited variety-location (genotype by environment) interaction.

Location differences. A second joint plot could be constructed for the second location component, i.e. one that portrays the contrast between the River Clay location and the Sands, especially the Central Sands. However, as it turns out this plot is very much one dimensional, and only really involves one attribute, Digestibility, and 4 varieties, i.e. Vivia (- 1.5), Markant (- 1.3), Dorina (- 1.2), and Irla (- 1.2). In particular, these varieties have comparatively lower Digestibility on the River Clay location than on the Sands. The numbers given serve as 'corrections' of the values in Table 5 for the combination of Digestibility and these varieties. Thus Vivia would have a lower value than the - 0.3 on the River Clay location and a higher value on the Sands locations. It should be borne in mind that these locations differences do not amount to more than 4% of the total variability.

Note that also in Fig. 1, Digestibility was responsible for the variety × location interaction, and that primarily Group A, which contains Vivia, Dorina, and Irla (but not Markant), showed a deviant pattern. On the

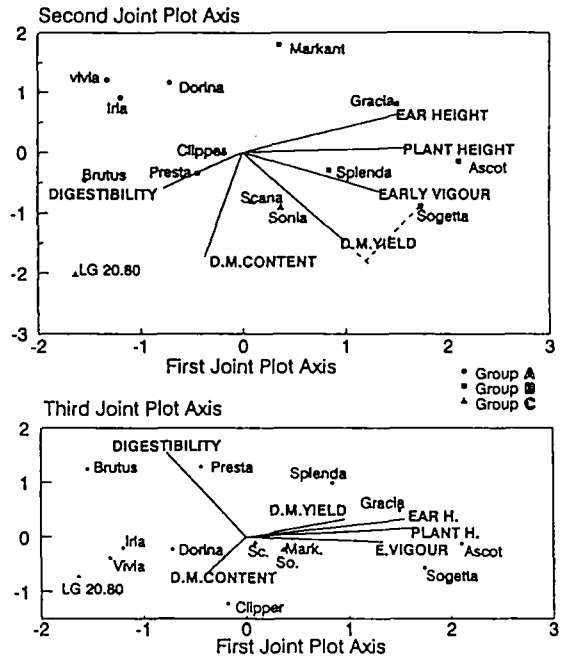

Fig. 4. Joint plot of attributes and varieties for the first location component (Top: First against second joint plot axis; Bottom: First against third joint plot axis). In the top part Sogetta is projected onto the attribute Dry Matter Yield (see text).

Table 4. Inner-Products between Varieties and Attributes (First Location Component)

Variety	DM. Yield	DM. Content	Digest.	Pl. Height	Ear Height	E. Vigour
Cluster B						
Sogetta	**2.7**	1.2	- 1.7	**2.6**	1.9	**3.0**
Ascot	2.2	- 0.5	- 1.7	**3.4**	**3.1**	**2.9**
Splenda	1.5	0.4	1.0	1.5	1.4	1.2
Gracia	0.5	- 2.3	- 0.9	**2.6**	**3.0**	1.5
Markant	- 2.3	**- 3.1**	- 1.7	0.7	1.6	- 0.7
Cluster C						
Sonia	1.6	1.6	- 0.1	0.5	- 0.1	1.1
Scana	1.2	1.4	0.2	0.1	- 0.4	0.7
LG 20.80	1.1	**4.6**	1.3	**- 2.9**	**- 4.1**	- 0.8
Cluster A						
Presta	0.5	- 0.0	**2.5**	- 0.5	- 0.5	- 0.5
Brutus	- 0.4	0.6	**3.4**	- 2.3	- 2.2	- 1.9
Clipper	- 0.6	0.9	- 1.7	- 0.5	- 0.7	- 0.1
Dorina	- 2.4	- 1.6	- 0.5	- 1.1	- 0.4	- 1.7
Irla	**- 2.5**	- 1.0	0.1	- 1.9	- 1.3	- 2.2
Vivia	**- 3.1**	- 1.3	- 0.3	- 2.1	- 1.4	**- 2.5**

The varieties have been ordered within clusters with respect to Dry Matter Yield.

other hand, these four varieties are the ones which have the largest numbers of missing data. Without looking into this matter more closely, we cannot give further insights into the reasons for the interaction, but at least we know now where to look.

Table 5. Ranking of the Varieties based on the Inner Products between Varieties and Attributes, in relation to their Admission to and Deletion from the Variety List

Variety	Quality Rating	Year of Admission	Year of Deletion
LG 20.80	10.7	1988	——
Brutus	10.0	1980	——
Presta	7.5	1988	1993
Splenda	6.4	1983	1993
Sonia	4.5	1986	1993
Scana	4.4	1988	——
Sogetta	1.5	1988	1993
Ascot	- 1.2	1987	1992
Gracia	- 4.0	1985	1991
Clipper	- 5.4	1985	1991
Irla	- 5.7	1981	1991
Dorina	- 7.9	1980	1990
Vivia	- 8.4	1983	1990
Markant	- 12.8	1983	1988

The Quality Rating (QR) is calculated from the inner product values of Table 5 as: QR = 3 × Digestibility + 2 × Dry Matter Yield + 1 × Dry Matter Content.

Discussion

From a plant breeders' viewpoint one may ask to what extent such analyses as these may assist in breeding decisions. Partly this depends on the attributes that are deemed important. If all one is interested in is Dry Matter Yield, then an analysis such as this one is of no use at all for the decision which varieties look most promising. If, on the other hand, one would like to take more than one attribute into account, then analyses such as these, can be used to make recommendations. However, when more than one variable is analysed, it has to be realised that their relative importance in the analysis is determined by their statistical characteristics, which not necessarily coincide with their agronomic importance. For the present data Dry Matter Contents and Digestibility are economically more important than, for instance Plant Height.

Some further agricultural background

In the beginning of the eighties there was a growing demand for tall leafy maize varieties with a high yield of dry matter. In general these varieties (to be found in Group B) are middle-early till middle-late, have a mediocre digestibility, and have a similar genetic background. Genetically, they differ radically from the varieties in the other two groups, which are much more alike.

Towards the end of the decade there arose a trend towards earlier varieties with a very good digestibility. LG 20.80 was the first one of this kind to be admitted to the Variety List, and by now many more very early varieties have been added to the list. From an agricultural point of view LG 20.80 is in a class of its own compared to the other varieties in the present sample.

Cluster analysis

With the mixture method of clustering the best three-group solution was determined, but two other groupings produced roughly the same value of the likelihood criterion, notably by moving Presta, or Presta and Clipper. The basic assumption of the cluster analysis is that the resulting groups are homogeneous, so that decisions about varieties can be taken on the basis of the groups rather than the individual varieties. However, within the groups there are some varieties which do not fit well into their groups.

In Group B, Markant has a different behaviour with respect to yield, plant height, and early vigour, which is in accordance with its somewhat different genetic make-up. In addition, Sogetta has a much higher dry matter content than the other varieties in this group. LG 20.80 does not fit into Group C given its chracteristics. In fact, it should be in a group of its own, which unfortunately was not possible in the present analysis. If more of these very early varieties had been included in the data set, this would not have been a problem. In Group A Brutus and Presta stand out because of their good digestibility, and Clipper is somewhat different because of its earliness. Irla, Vivia, and Dorina form a very homogeneous subgroup, which can be explained by their common pollinator. Presta and Clipper also have the same pollinator but their female single is quite different.

Using Figs 1 and 3, and Table 4, we can come to a general characterisation of the groups. In particular, Group B varieties are tall plants with high ears with a high yield of dry matter, moderate dry matter content, mediocre digestibility, and good early vigour. If LG 20.80 is ignored, Group C varieties are characterised by rather small plants with rather low ears, good yield of dry matter, rather high dry matter content and moderate digestibility and early vigour. Finally, Group A

is characterised by small plants with low ears, low yield of dry matter, moderate dry matter content, good digestibility, and rather poor early vigour. Furthermore, overall there is a large difference between Group B and the other two. Groups A and C are much more similar.

Ordination

Group memberships provides only limited information about the individual variety. To gain further insight into the varieties, ordination can be especially helpful. Individual differences and similarities between members of a group become more visible. This visibility depends, of course, on the appropriate presentation of the information. On the kind of figures presented here with two dimensions at a time, it is not always easy to grasp the true extent of a relationship. For instance, the digestibility of LG 20.80 looks 'over-estimated' in Fig. 4A and 'under-estimated in Fig. 4B. To assist with the interpretation the inner-products in Table 4 are rather helpful because they supply the values of the relationships between attributes and varieties. It provides the more detailed information necessary for selection. For instance, it shows quite clearly that even though LG 20.80 fits best in Group C it is quite different from the other two. The situation for Markant is similar. The inclusion of Sogetta was clearly based on its good yield, dry matter content, and early vigour, notwithstanding its poor digestibility. Presta has a good digestibility with reasonable yield in its favour, while LG 20.80 is outstanding in dry matter content. Finally, Scana was apparently included due to its overall reasonable performance, even though it is not outstanding on any one attribute.

An interesting use of the inner products would be to apply a differential weighting to the values in Table 4. In particular, when we weight digestibility by a factor 3, dry matter yield by 2, and dry matter content by 1, we get the values ('quality rating') of Table 5, also included in the table are the years of admission and deletion from the Variety List. There is a remarkable correlation between the quality ratings and the year of deletion from the list. All varieties with negative ratings have been deleted before 1993 in the order of their ratings. The two top varieties are still on the list, while all other except for Scana have now been removed. Table 5 suggests that combinations of various attributes probably have played an important role in the selection of the varieties.

Location differences

That we have not found any serious variety × location interaction does not mean that there is no such interaction in the Netherlands. We have to keep in mind that all varieties included in this data set were already in the Variety List. It is most likely that varieties which showed considerable interaction were probably withdrawn from the trials at a much earlier stage, because they did not perform as well as varieties less susceptible to interaction.

We have remarked before about the surprising fact that the varieties which showed interaction with respect to digestibility, were exactly the ones with incomplete data before aggregation over time. On the other hand, Groups B and C differed significantly on the Northern Sands, but not elsewhere. Only further investigation can shed light on these interactions. Whether they are genuine, a type I error, or an artifact from missing data.

Conclusion

Plant breeders have to deal with many new varieties with a widely varying genetic background and considerable variation in the expression of the characteristics. In a breeding programme the cluster analysis can be a valuable aid to divide the total range into a manageable number of groups of comparable varieties. Also in situations with many interactions it is useful to have groups of varieties which react in the same way.

However, at the end of the selection procedures, when it is a matter of selecting a few outstanding varieties from a relatively small number, the ordination is more suitable, because it provides information about the specific characteristics of individual varieties with respect to the other ones. In this stage of the process the ordination seems pre-eminently suited for the work of testing authorities.

At the same time, it seems that the combination of the two techniques on the same data, provides an even better perspective on the varieties, their similarities and differences. The main reason is that using more than one method provides an opportunity for mutual validation of results, especially because the methods use the available information in different ways.

Acknowledgements

The first author received grants from the Netherlands Organisation of Scientific Research (NWO), the Royal Netherlands Academy of Sciences (KNAW), and the Australian Academy for the Social Sciences. The second author was financially supported by the Bilateral Science, and Technology Collaboration Program of the Australian Department of Industry, Technology, and Commerce.

References

Basford, K.E. & G.J. McLachlan, 1985. The mixture method of clustering applied to three-way data. J. of Classification 2: 109–125.

Basford, K.E., P.M. Kroonenberg & I.H. DeLacy, 1991. Three-way methods for multiattribute genotype × environment data: an illustrated partial survey. Field Crops Res. 27: 131–157.

Basford, K.E., P.M. Kroonenberg, I.H. DeLacy & P.K. Lawrence, 1990. Multiattribute evaluation of regional cotton variety trials. Theor. Appl. Genet. 79: 225–234.

Burt, R.L., L.A. Edye, W.T. Williams, B. Grof & C.H.L. Nicholson, 1971. Numerical analysis of variation patterns in the genus Stylosanthes as an aid to plant introduction and assessment. Aust. J. Agric. Res. 22: 737–757.

Byth, D.E., R.L. Eisemann & I.H. DeLacy, 1976. Two-way pattern analysis of a large data set to evaluate genotypic adaptation. Heredity 37: 215–230.

Corsten, L.C.A. & J.B. Denis, 1990. Structuring interaction in two-way tables by clustering. Biometrics 46: 207–215.

Franc, A., 1992. Étude algebrique des multitableaux: Apports de l' algèbre tensorielle (Algebraic study of multi-way arrays: Contributions of tensor algebra). Unpublished doctoral thesis, Université de Montpellier II, Sciences et Techniques du Languedoc, Montpellier, France.

Freeman, G.H., 1975. Analysis of interactions in incomplete two-way tables. Applied Statistics 24: 46–55.

Gabriel, K.R., 1971. The biplot graphical display of matrices with applications to principal components. Biometrika 58: 453–467.

Gabriel, K.R., 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. p. 147–173. In: V. Barnett (Ed). Interpreting Multivariate Data. Wiley, Chicester, UK.

Kroonenberg, P.M., 1983. Three-mode principal component analysis. Theory and application. DSWO Press, Leiden.

Kroonenberg, P.M., 1988. Three-mode analysis. p. 231–236. In: S. Kotz & N.L. Johnson (Eds). Encyclopedia of Statistical Sciences, Vol. 9. Wiley, New York.

Kroonenberg, P.M. & K.E. Basford, 1989. An investigation of multiattribute genotype response across environments using three-mode principal component analysis. Euphytica 44: 109–123.

Kroonenberg, P.M., K.E. Basford & M. Van Dam, (in press). Classifying infants in the strange situation with three-way mixture method clustering. British Journal of Psychology.

McLachlan, G.J. & K.E. Basford, 1988. Mixture models: Inference and applications to clustering. Marcel Dekker, New York.

Mungomery, V.E., R. Shorter & D.E. Byth, 1974. Genotype × environment interactions and environmental adaptation. I. Pattern analysis – application to soya bean populations. Aust. J. Agric. Res. 25: 59–72.

Van Eeuwijk, F.A., L.C.P. Keizer & J.J. Bakker, 1995. Linear and bilinear models for the analysis of multi-environment trials: II. An application to data from the Dutch Maize Variety Trials. Euphytica 84: 9–22.