Theor Appl Genet (1992) 85:461–469

THEORETICAL AND APPLIED GENETICS

Classifying genotypic data from plant breeding trials: a preliminary investigation using repeated checks

J. K. Bull[1], K. E. Basford[2], I. H. DeLacy[2], M. Cooper[2]

[1] Bureau of Sugar Experiment Stations, PO Box 651, Bundaberg, Queensland 4670, Australia
[2] Department of Agriculture, The University of Queensland, Queensland 4072, Australia

Received January 3, 1992; Accepted June 3, 1992
Communicated by A. R. Hallauer

Summary. Several subjective choices must be made when classifying genotypes based on data from plant breeding trials. One choice involves the method used to weight the contribution each environment makes to the classification. A second involves the use of either genotype-means for each environment or genotype-values for each block, i.e., considering each block to be a different environment. Another involves whether environments (or blocks) in which genotypes are non-significantly different should be included or excluded from such classifications. An alternative to the use of raw or standardized data, is proposed in which each environment is weighted by a discrimination index (*DI*) that is based on the concept of repeatability. In this study the effect of three weighting methods (raw, standardized and *DI*), the choice of using environments or blocks, and the choice of including or excluding environments or blocks in which genotypic effects were not significant, were considered in factorial combination to give 12 options. A data set comprised of five check cultivars each repeated six times in each of three blocks at six environments was used. The effect of these options on the ability of a hierarchical clustering technique to correctly classify the repeats into five groups, each consisting of all the six repeats of a particular check cultivar, was investigated. It was found that the *DI* weighting method generally led to better recovery of the known structure. Using block data rather than environmental data also improved structure recovery for each of the three weighting methods. The exclusive use of environments in which genotypic effects were significant decreased structure recovery while the contrary generally occurred for blocks. The best structure recovery was obtained from the *DI* weighting applied to blocks (whether genotypes were significant or not).

Key words: Cluster Analysis – Genotype × environment interaction – Heritability – Repeatability – Structure recovery

Introduction

In multi-environment plant breeding evaluation trials, the relative performance of genotypes commonly differs across environments. Genotype × environment (G × E) interaction is often of sufficient relative magnitude to complicate the investigation of genotypic performance (DeLacy et al. 1990). Clustering techniques have been found to be useful for reducing the number of genotypic responses that must be examined when investigating performance (Byth et al. 1976; DeLacy 1989; Ghaderi et al. 1982; Hayward et al. 1982; Bull et al. 1992a, b).

When classifying genotypic data from multi-environment trials by a hierarchical procedure, a number of choices regarding the pre-processing and form of data to be used must be made. Firstly, as environments contribute to the entire hierarchy in genotypic classifications in direct relation to their variances, those environments with relatively large variances will dominate such classifications (Burr 1968; Shorter et al. 1977). This has led some workers to standardize (scale) environments so that they contribute equally to the classification of genotypes (Byth et al. 1976; DeLacy et al. 1990; Yau 1991). The desirability of these alternative weighting methods has not been investigated in depth.

A second choice is concerned with the use of either genotype-means for each environment or genotype-

Correspondence to: J. K. Bull

values for each block within each environment to classify genotypes. This is important where within-environment variation exists, because the environmental challenges issued to genotypes may differ among blocks within an environment and cause genotype × block (G × B) interaction. The presence of this interaction may be identified by including repeats (multiple entries) of check cultivars within every block of every environment. This allows an estimate of the variability within blocks to be made and can then be used to test G × B interaction. When present, and of sufficient magnitude, this G × B interaction may influence the reliability and interpretability of genotypic-means over blocks. The effect of such an interaction needs investigation in plant breeding trials.

Thirdly, the consideration of genotypic response across either environments or blocks, irrespective of the statistical significance of genotypic effects within them, may interfere with the determination of the (true) differences among genotypes. The desirability of including or excluding environments (or blocks) in which genotypes are non-significantly different has rarely been considered in analyses of plant breeding trials.

In the present study, the effect of three methods of weighting environmental data on the classification of genotypes is determined. The three methods for weighting or pre-processing data are raw data, standardized data, and discrimination index (*DI*)-weighted data. The *DI* assesses the relative amount of genetic to total variability within an environment and is based on the concept of repeatability, where repeatability is defined as the heritability for fixed genotypes (Fehr 1987).

The effect of G × B interaction on the classification of genotypes is also determined for each of these three weighting methods by classifying genotypes on both genotype-means over blocks within environments and on genotype-plot values within blocks. The effect, on the classification of genotypes, of either including or excluding environments or blocks in which cultivar effects are not significant is also determined for each of the three weighting methods. By the factorial combination of the three weighting methods, the choice of classifying across either environments or blocks, and the inclusion or exclusion of environments or blocks in which genotypes were not significantly different, consideration of 12 options is possible.

To gain an objective measure of the effect these 12 options have on genotypic classifications, a data set with known genotypic structure was used (Bull et al. 1992a). In this data set all the genotypes (check cultivars) were repeated (planted a number of times) in each block of each environment. Conventionally, the mean over repeats for each genotype would be found and these means then classified. But, in this study each repeat of each cultivar is considered to be a distinct genotype and these repeats are then classified. By doing so, it is reasonable to assume that a known number of groups and a known group composition should be found (Bull et al. 1992a); i.e., each group should include all the repeats of a particular cultivar (termed cultivar-group).

However, in such a data set there is no unique way to associate the performance of a repeat in one block with the performance of a repeat in another block, whether the blocks are from the same environment or not (Bull et al. 1992b). Thus when forming a G × E data matrix, in which each repeat is treated as a different genotype, the repeats within each cultivar may be randomly associated across blocks within each environment. These associations may be used to form a mean over blocks for each repeat and these means may then be randomly associated across environments to give a constructed response pattern for each repeat (termed construct). A similar process may be used when forming a G × B data set, except that the constructs would involve a response pattern across all blocks. Irrespective of the type of the response pattern considered the constructs may then be classified. This procedure may be repeated many times with each classification being based on a different random association of the repeats across blocks and environments (Bull et al. 1992a, b).

The level of recovery of known genotypic structure was assessed for the 12 options using the Hubert and Arabie (1985) simple matching coefficient adjusted for chance (Milligan and Cooper 1986; Milligan 1989) of the partitions (groups at a particular truncation level) identified with the known solution. From this assessment, preliminary recommendations are made regarding the general utility of these options in the analysis of plant breeding trials.

Background

Weighting data

In the general field of classification (numerical taxonomy), many methods have been used to weight variables (here environments or blocks). DeSoete et al. (1985) and DeSarbo et al. (1984) provide reviews of this topic, while Davies and Boratyński (1979), Milligan and Cooper (1988) and Milligan (1989) outline the practical implications of many approaches to variable weighting.

Pragmatically, Lumelsky (1982) states that if variables are to be used in classification they must allow good differentiation between individuals (here constructs) and must be interpretable. Lumelsky (1982) considered that the use of raw data or standardized data would fulfil these criteria. The deliberate weighting of variables by the proportion of informative (here genetic) variability to total variability (here total phenotypic) would satisfy Lumelsky's criteria.

In a study similar to that presented here, Johnson (1982) weighted variables by the relative variability between duplicate pairs (two repeats) compared with the total variability for each variable. This gave rise to three weighting methods: Variance ratio (Flake et al. 1969), *F*-statistic, and Log (*F*-statistic). These weighting methods attempt to measure the discriminatory worth of a variable. However, they ignore the appropriate divisors for the variance components implicit in their calculation. Also, the ratios of mean squares, as in the last two weighting methods, have no upper bound and their utility is therefore questioned. Johnson (1982) gave no clear recommendations about which weighting method was superior, but all three generally improved classificatory results over those obtained using standardized data.

The *DI* used here is conceptually similar to these measures but takes into account the appropriate divisors for the variance components used in its calculation and has an upper and lower bound. The effect of the following weighting methods on the recovery of the known cultivar-group structure was therefore considered:

(1) Raw data – environments intrinsically weighted by their phenotypic variance.
(2) Standardized data – genotypes within environments were scaled to unit variance (but not centred) and therefore each environment was equally weighted.
(3) *DI* weighted data – environments were weighted by the ratio of genetic variability on total phenotypic variability for each environment.

Blocks or environments

Gauch and Zobel (1988) state that blocking is a dubious means of error control when G × B interaction exists. By using repeats of check cultivars within a block, such an interaction may be identified and its contribution to the total variation determined. The presence of G × B interaction suggests that the blocks within an environment pose different environmental challenges to the genotypes. An average across such blocks may confound real differences in the performance patterns of the genotypes and thus decrease the utility of resultant classifications. It is contended that blocks should be considered to be different environments when G × B interaction is present and substantial. The effect of this interaction on the recovery of known structure was investigated here by comparing the recovery of the cultivar-group structure for classifications of constructs across environments with that across blocks.

Variable selection

The use of a sub-set of variables in which the effect of interest is significant may be anticipated to facilitate interpretation because variables in which the effects of interest are not significant may simply represent random 'noise' (Johnson 1982; DeSarbo et al. 1984; Thorpe 1985). Sokal and Sneath (1963) advocate using a large number of variables to obtain a stable classification. Such a recommendation has been criticised since the addition of variables that are predominantly or completely composed of 'noise' may unduly influence classificatory results (Johnson 1982; Milligan 1980, 1989). This may be of particular importance when raw or standardized data are being used as too much weight may be placed on doubtful variables (Milligan 1989).

There is an analogy between the variables measured on an individual in numerical taxonomy and the environments or blocks used to measure the performance of a genotype in plant breeding experiments. That is, in plant breeding trials each of the environments or blocks considered may be thought of as a different variable. Given the above considerations we here evaluate the effect of including and excluding those environments or blocks in which cultivars were not significantly different on the recovery of the known cultivar-group structure.

Assessment criteria

The Hubert and Arabie (1985) matching coefficient adjusted for chance (MCAC) has been recommended by Milligan and Cooper (1986) as a method of measuring the agreement between two partitions. It was used here to gain a general assessment of the impact each of the 12 options had on the ability of the classificatory technique to recover the known cultivar-group structure. The average, over the observed partitions, of the similarity between the known solution and the observed partitions was determined.

Experimental details

The experiment was described by Bull et al. (1992a) and involved the evaluation of five sugarcane cultivars (CP44-101, H56-752, Q110, Q137, Q141) for tonnes of cane per hectare (TCH). The five cultivars were each repeated six times in each of three blocks at three sites in both plant and first ratoon crops (2 crop-years). Thus by the combination of the three sites by 2 crop-years six environments may be considered. More detailed site descriptions were given by Hogarth and Bull (1990). Each plot of a repeat consisted of a single row 10 m long. For each cultivar, the repeats were planted into three adjacent rows and two contiguous plots per row. These cultivar plots were randomly positioned in each block at each site in the plant crop.

The effect of non-randomly positioning the six repeats of each cultivar in each block in every environment may influence error estimates for this experiment. Although a random allocation of repeats to plots would be preferable, it was considered that the analyses conducted were appropriate and, therefore, that some limited inference could be made to wider applications. Further, the present arrangement was regarded as adequate for demonstrating the application of the 12 options considered.

Practical implementation

Construction of response patterns

The typical treatment of repeats, which involves the formation of a mean over repeats for each cultivar, was not used in this study. Instead each of the six repeats of a cultivar was treated as separate entry of that cultivar. By treating each repeat as a separate genotypic entry, however, there was no unique way to associate the performance of a repeat from a particular cultivar in one block with the performance of a repeat from that cultivar in another block whether the blocks were from the same or different environments (Bull et al. 1992b). Consequently, Bull et al. (1992a, b) randomly associated the performance of each of the six repeats from a particular cultivar with the performance of repeats from the same cultivar across the blocks of an environment. Such associations across blocks within environments were then used to form six mean values over blocks for each cultivar. These six mean values for each of the five cultivars were then associated across the six environments to form 30 constructed response patterns (one response pattern for each of the six repeats of the five cultivars). In forming a G × B data set, the repeats were similarly associated as for the G × E data set but the mean over the three blocks from each of the six environments was not found. The number of possible ways of associating repeats across the environments (see Bull et al. 1992b for further details) or blocks used for each of the 12 options will be considered later.

A hierarchical agglomerative clustering procedure (Williams 1971), with Ward's (1963) method as the fusion strategy, and squared Euclidean distance as a dissimilarity measure (Wishart 1969; Burr 1970), was used for all classifications of the constructed responses. This technique was chosen as it was found to be of high performance, relative to a number of fusion strategies, in cluster validation studies (Blashfield 1976; Milligan and Cooper 1988), and has previously been found to be useful in the investigation of genotypic adaptation in plant breeding trials (Byth et al. 1976; Bull et al. 1992a, b).

The 30 constructs were derived by a random association (without replacement) of repeats across environments or blocks. Thus at each particular truncation level the classificatory procedure could return different partitionings of the constructs depending on the randomizations used. For each of the 12 options, 1,000 classifications were considered. Each of these 12,000 classifications was based on a different random association of repeats. If there were differences in levels and/or patterns of response among the check cultivars of a magnitude greater than experimental error, then five cultivar-groups could be formed with each group consisting of all the repeats of a cultivar. Since the data were known to consist of five cultivars all classifications were examined at the five-group level.

Methods of weighting data

Using a similar nomenclature to that employed by Milligan and Cooper (1988), the three methods of weighting variables before classification were raw data (denoted by Z_0), standardized data (denoted by Z_1), and DI weighted data (denoted by Z_2).

As in Burr's implementation formulae (Burr 1968), these weighting methods were defined by:

$$Z_0 = Y$$

$$Z_1 = \frac{Y}{s}$$

$$Z_2 = \frac{Y}{s}\sqrt{DI}$$

where:

Y = TCH value (cane yield) of a construct within any particular environment or block

s = standard deviation of the TCH values for that particular environment or block over all repeats.

DI = discrimination index of the TCH values for that particular environment or block. This term is explicitly defined shortly.

Before doing so, it is necessary to estimate variance component values derived from the expected mean squares using a fully random effects model given by:

$$y_{ijk} = m + c_i + b_j + (cb)_{ij} + (rcb)_{kij},$$

$$i = 1, \ldots, n_c; \quad j = 1, \ldots, n_b; \quad k = 1, \ldots, n_{r/c}$$

within each environment, where

y_{ijk} = TCH value of repeat k within cultivar i and block j

m = general mean

c_i = effect of cultivar i assumed to be distributed as $N(0, \sigma_c^2)$

b_j = effect of block j assumed to be distributed as $N(0, \sigma_b^2)$

$(cb)_{ij}$ = interaction effect of cultivar i and block j assumed to be distributed as $N(0, \sigma_{cb}^2)$

$(rcb)_{kij}$ = effect of repeat k within cultivar i and block j assumed to be distributed as $N(0, \sigma_{r/cb}^2)$

n_c = number of cultivars

n_b = number of blocks
$n_{r/c}$ = number of repeats within cultivars

and:

$$y_{ik} = m + c'_i + (rc')_{ki},$$

within each block, where

y_{ik} = TCH value of repeat k within cultivar i
m = general mean
c'_i = effect of cultivar i assumed to be distributed as $N(0, \sigma^2_{c'})$
$(rc')_{ki}$ = effect of repeat k within cultivar i assumed to be distributed as $N(0, \sigma^2_{r/c'})$.

Then, the *DI* for each environment and block was calculated as

$$DI_{\text{environment}} = \frac{\sigma^2_c}{\sigma^2_c + \dfrac{\sigma^2_{cb}}{n_b} + \dfrac{\sigma^2_{r/cb}}{n_b n_{r/c}}}$$

and

$$DI_{\text{block}} = \frac{\sigma^2_{c'}}{\sigma^2_{c'} + \dfrac{\sigma^2_{r/c'}}{n_{r/c}}}$$

respectively.

Formulae for the *DI*s were based on repeatability, as defined by Fehr (1987), and in each case summarize the relative proportion of genetic variability to toal variability for the defined reference unit. However, the term '*DI*' is considered to be more appropriate than the term 'repeatability' as it is the relative power of the environment or block to discriminate among genotypes that is of interest here, not their repeatability per se. Theoretically, the *DI* has a lower bound of zero and an upper bound of one. If negative variance components are encountered and not assumed to be zero, *DI* values may lie outside this range. In this study, negative variance components were assumed to have no meaning (Thomson and Moore 1963) and were, therefore, set equal to zero. Environments or blocks with zero *DI* did not contribute to the classification of constructs.

Although the performance of the repeats (constructed responses) from these cultivars is subsequently assessed by classification, thus implying cultivars were a fixed effect, a random effect was assumed in the above model. This was because, in practice, plant breeders are interested in both selection and the relative magnitude of variance components. This assumption was further justified because inference from these analyses is to the wider application of these 12 options to data from plant breeding trials.

The three weighting methods were applied to the following types of cultivar data sets: all environments (denoted C × E), all blocks (denoted C × B), only those environments in which cultivar effects were significant (denoted C × E*), and only those blocks in which cultivar effects were significant (denoted C × B*).

Assessment criteria

For these 12 options the matching coefficient adjusted for chance (MCAC) of the partitions identified with the cultivar-group solution was used to define the recovery of known structure. The MCAC provides an overall summary of the similarity between two partitions by assessing the similarity in allocation of constructs to groups in them and ranges from below zero to one. A negative or zero MCAC indicates that the similarity in the allocation of constructs to groups between the two partitions is only due to chance, while a MCAC of one means that the allocation of constructs to groups in the two partitions is identical. The determination of significant differences among options was conservatively estimated by assesing whether the confidence intervals (mean MCAC ± twice the standard error) for the options overlapped.

Results and discussion

For brevity, the individual analyses conducted within environments and within blocks will not be presented. However, the estimated variance components and the statistical significance of their associated mean square is presented for these analyses (Tables 1a, b). Cultivar effects were significant in one environment and highly significant in one other (Table 1a). Cultivar × block (C × B) interaction effects were significant in two environments and highly significant in two others. The general lack of significant cultivar effects may be due to only five cultivars being evaluated. Further, by considering cultivars that are somewhat adapted to each of these test environments, the chance of finding significant effects is less than that for unselected clones. Typically at this stage of selection, there would be an additional 120 clones under evaluation. The opportunity for significant clone effects and clone × block interaction effects in such trials may, therefore, be anticipated to be greater than that found in this experiment.

From the analyses conducted for each block within each environment (18 blocks in all), cultivar effects were found to be significant in three blocks and highly significant in nine others (Table 1b). Since cultivar effects were not significant in all environments or blocks, the effect on the recovery of known structure of including or excluding those environments or blocks in which cultivars were not significantly different could be investigated. The number of environments or blocks used in each classification and the corresponding number of possible allocations of repeats over them were determined (Table 2). For each option, the number

Table 1a. The environmental designation, estimated variance components and associated standard errors (SE) for cultivar (σ_c^2), cultivar × block interaction (σ_{cb}^2) and sampling error ($\sigma_{r/cb}^2$), and *DI* values for each of the six environments

Environment			Variance components			*DI*(%)
No.	Site	Crop-class[a]	$\sigma_c^2 \pm (SE)$	$\sigma_{cb}^2 \pm (SE)$	$\sigma_{r/cb}^2 \pm (SE)$	
1	Bingera	P	198.57*(143.26)[b]	89.24*(63.46)	308.67(49.75)	80.89
2	Bingera	1R	355.32**(216.78)	28.84(27.08)	184.17(29.68)	94.71
3	Nambour	P	136.73(122.75)	169.23**(92.84)	227.54(36.67)	66.44
4	Nambour	1R	−36.69(39.17)	121.75*(101.82)	617.39(99.50)	0[c]
5	Station	P	114.76(123.09)	187.89**(117.86)	445.71(71.83)	56.77
6	Station	1R	26.72(39.79)	22.82(50.53)	515.06(83.02)	42.45

[a] P, plant crop; 1R, first ratoon crop

[b] Standard error

[c] *DI* assumed to be zero

* Mean square significant at $P \leq 0.05$

** Mean square significant at $P \leq 0.01$

Table 1b. The environmental designation, estimated variance components and associated standard errors (SE) for cultivar ($\sigma_{c'}^2$) and sampling error ($\sigma_{r/c'}^2$), and *DI* value for each of the three blocks within each of the six environments

Environment				Variance components		*DI*(%)
Block no.	Site	Crop-class[a]	Block	$\sigma_{c'}^2 \pm (SE)$	$\sigma_{r/c'}^2 \pm (SE)$	
1	Bingera	P	A	167.21*(140.95)	446.35(121.49)	69.21
2	Bingera	P	B	599.88**(364.67)	189.38(51.54)	95.00
3	Bingera	P	C	96.34*(84.59)	290.27(79.00)	66.57
4	Bingera	1R	A	479.41**(290.14)	137.99(37.56)	95.42
5	Bingera	1R	B	382.50**(241.44)	212.18(57.75)	91.54
6	Bingera	1R	C	290.56**(187.45)	202.33(55.07)	89.60
7	Nambour	P	A	446.28**(281.45)	244.94(66.66)	91.62
8	Nambour	P	B	170.40**(125.34)	273.82(74.52)	78.88
9	Nambour	P	C	301.19**(189.80)	163.85(44.59)	91.69
10	Nambour	1R	A	113.48(107.44)	418.16(113.81)	61.95
11	Nambour	1R	B	4.78(38.35)	336.90(91.69)	7.84
12	Nambour	1R	C	127.92(186.20)	1097.12(298.60)	41.16
13	Station	P	A	555.95**(368.34)	485.30(132.08)	87.30
14	Station	P	B	229.90**(171.83)	396.52(107.92)	77.67
15	Station	P	C	122.12(116.17)	455.33(123.92)	61.67
16	Station	1R	A	127.80*(106.68)	330.79(90.03)	69.86
17	Station	1R	B	−40.83(36.11)	526.51(143.30)	0[b]
18	Station	1R	C	61.67(106.47)	687.86(187.21)	34.98

[a] P, plant crop; 1R, first ratoon crop

[b] *DI* assumed to be zero

* Mean square significant at $P \leq 0.05$

** Mean square significant at $P \leq 0.01$

of different allocations possible was much larger than the number of classifications (1,000) considered.

The (raw) Z_0 weighting for the six environments (based on the average variance of ten different randomizations) and for the 18 blocks was determined. The *DI* for each environment and block was also determined (Tables 1a, b). From the scatterplots of the Z_0 weighting versus *DI*, where each was expressed as a percentage of their maximum value found, it appeared that different emphasis was placed on environments or blocks by these two weighting methods (Fig. 1a, b), as evidenced by deviations from linear correspondence.

The average over all partitions of the MCAC between each of the 1,000 partitions and the assumed cultivar-group solution, was determined for each of the 12 options (Fig. 2). For the classifications based on the 30 × 6 C × E data sets (Fig. 2), transformation Z_2 produced the best recovery of known structure followed

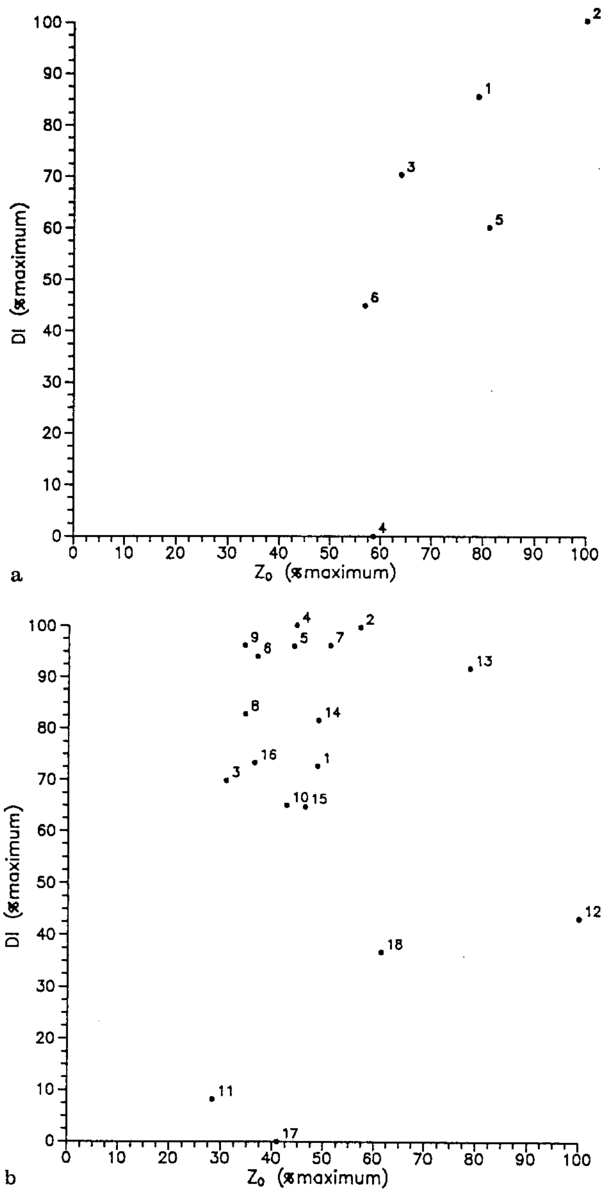

Fig. 1 a, b. The correspondence between environmental variability and *DI*. **a** For the six environments (environment numbers defined in Table 1a). Environmental variability was determined from the average variance of the 30 construct means over ten randomizations. **b** For the 18 blocks (block numbers defined in Table 1b)

Table 2. A summary for each of the 12 options considered of the number of environments or blocks actually used and the number of combinations of repeats that was possible

Option	Number of environments or blocks used	Number of different combinations possible
Z_0		
C × E	6	$(6!)^{(6*3-1)*5}$
C × B	18	$(6!)^{(18-1)*5}$
C × E*	2	$(6!)^{(2*3-1)*5}$
C × B*	12	$(6!)^{(12-1)*5}$
Z_1		
C × E	6	$(6!)^{(6*3-1)*5}$
C × B	18	$(6!)^{(18-1)*5}$
C × E*	2	$(6!)^{(2*3-1)*5}$
C × B*	12	$(6!)^{(12-1)*5}$
Z_2		
C × E	5[a]	$(6!)^{(5*3-1)*5}$
C × B	17[b]	$(6!)^{(17-1)*5}$
C × E*	2	$(6!)^{(2*3-1)*5}$
C × B*	12	$(6!)^{(12-1)*5}$

[a] The *DI* for one environment was zero

[b] The *DI* for one block was zero

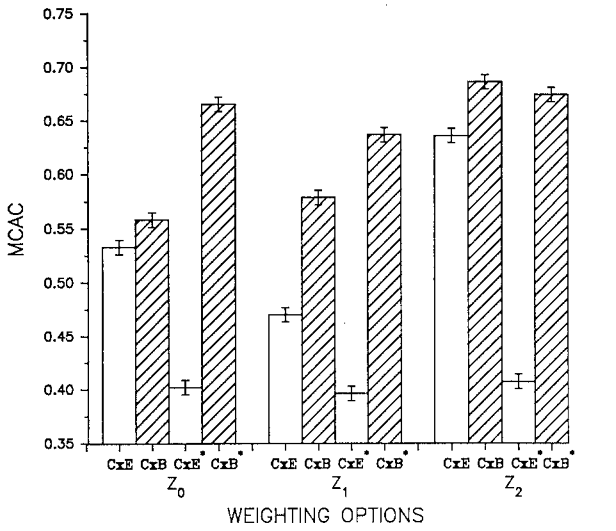

Fig. 2. The average, for each of the 12 options, of the MCAC of the 1,000 observed partitions with the cultivar-group solution. The confidence interval (mean ± twice the standard error) is indicated for each option

by Z_0 and Z_1, respectively. The poor structure recovery for transformation Z_1 may be due to the more variable environments containing relatively more genetic information and less 'noise' (high *DI*) than lower variance environments (Table 1a, Fig. 1a). Thus by placing equal weight on each environment, an increased weighting was placed on more 'noisy' environments and a reduced weighting was placed on the less 'noisy' environments compared with the Z_0 and Z_2 weightings (Fig. 1a).

Considering the classifications based on the 30 × 18 C × B data sets (Fig. 2), transformation Z_2 produced the best structure recovery, followed by transformation Z_1 and Z_0, respectively. The reversal in effectiveness of transformations Z_0 and Z_1, relative to the classifications based on the C × E data sets, may be due to low-variance blocks containing relatively less 'noise' (high *DI*) than high-variance blocks (Table 1b). This is reflected in the lack of positive association between block *DI* and variance (Fig. 1b) in contrast to the positive association for environments (Fig. 1a).

Classifications based on the C × B data sets for each of the three weighting methods gave superior structure recovery to those based on the C × E data sets. This may be due to C × B interaction being significant within most of the environments (four out of six). Forming an average over blocks when this interaction is significant may serve to obscure the pattern of adaptation of cultivars to the conditions present within each environment. Thus the pattern of interaction across blocks for each cultivar contributed in a meaningful way to the discrimination between constructs.

For classifications based on the 30 × 2C × E* data sets (Fig. 2), all three weighting methods produced non-significantly different, and somewhat poor, recoveries of known structure. This poor structure recovery may be because only two environments were considered and that such a small number of environments limited the resolving power of the classification.

Thus, for each weighting method, classifications based on the C × E data sets gave better structure recovery than classifications based on the C × E* data sets. But, if there was truly no useful information concerning cultivar performance in the environments in which cultivar effects were not significant then their inclusion, regardless of weighting, should have decreased structure recovery. Thus there was useful information, although sometimes statistically non-significant, regarding the performance patterns of these cultivars in each environment.

For classifications based on the 30 × 12C × B* data sets (Fig. 2), transformations Z_2 and Z_0 gave the best and non-significantly different structure recoveries. All weighting methods gave somewhat high recoveries of known structure. For the transformations Z_0 and Z_1, the average MCAC was greater than those for classifications based on the C × B data sets, respectively. But, for the Z_2 transformation, classifications based on the C × B and C × B* data sets gave the same degree of structure recovery. Thus the inclusion of blocks in which cultivar effects were not significant did not impact negatively on structure recovery using this weighting method. By contrast, transformations Z_0 and Z_1 were unduly affected by 'noisy' blocks as they produced poorer recovery of structure when classifying over the C × B data sets than when classifying over the C × B* data sets. Thus Z_2 applied to the C × B (or C × B*) data sets produced the best recovery of known structure of the 12 options considered.

Conclusions

By using a data set that contained several cultivars that were each repeated, it was possible to examine objectively the effect of different methods of data pre-processing and different choices of data form on the results from classification. As only one data set was considered one needs to be somewhat circumspect in making inferences from these results to wider applications in plant breeding trials or other types of experiments. Nonetheless the following conclusions and tentative generalizations may be made:

(1) Weighting the contribution environments or blocks made to the classification of constructs by the *DI* led to better recoveries of known structure than were obtained by using raw or standardized data. More generally, as raw and standardized data weight environments or blocks arbitrarily their utility may be questionable.
(2) Classifications of constructs based on block values rather than environment means lead to better structure recovery. It is considered that if G × B interaction is present within an environment, the blocks within that environment may be issuing different challenges to genotypes. Then the investigation of genotypic performance based on a mean over blocks may be of limited value.
(3) Classifying constructs based on *DI* weighted data that contained environments or blocks in which genotype effects were not significant did not decrease structure recovery over the exclusive use of environments or blocks in which cultivar effects were significant. In general it is considered that the use of a formal test of significance (the *F*-test) to exclude environments or blocks in which cultivar effects are non-significant is unnecessary if the *DI* is to be applied. That is, the *DI* will give a low weighting to those environments or blocks that would be excluded using an *F*-test. Thus these approaches only differ by relative conservatism.

Acknowledgments. This trial and the development of the analytical procedure was conducted with partial financial support from the Australian Sugar Research and Development Corporation.

References

Blashfield RK (1976) Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods. Psychol Bull 83:377–388

Bull JK, Basford KE, DeLacy IH, Cooper M (1992a) Determining appropriate group number and composition for data sets containing repeated check cultivars. Field Crops Res (in press)

Bull JK, Cooper M, DeLacy IH, Basford KE, Woodruff DR (1992b) Utility of repeated checks for hierarchical classification of data from plant breeding trials. Field Crops Res 30:79–95

Burr EJ (1968) Cluster sorting with mixed character types. I. Standardization of character values. Aust Comput J 1: 97–99

Burr EJ (1970) Cluster sorting with mixed character types. II. Fusion strategies. Aust Comput J 2:98–103

Byth DE, Eisemann RL, DeLacy IH (1976) Two-way pattern analysis of a large data set to evaluate genotypic adaptation. Heredity 37:215–230

Davies RG, Boratyński KL (1979) Character selection in relation to the numerical taxonomy of some male Diaspididae (Homoptera: Coccoidea). Biol J Linn Soc 12:95–165

DeLacy IH (1989) Analysis and interpretation of pattern of response of agricultural adaptation experiments. In: DeLacy IH (ed) Analysis of data from agricultural adaptation experiments. ACNARP, Bangkok, pp 50–70

DeLacy IH, Eisemann RL, Cooper M (1990) The importance of genotype-by-environment interaction in regional variety trials. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding. Louisiana State University, Baton Rouge, Louisiana, pp 287–300

DeSarbo WS, Carroll JD, Clark LA, Green PE (1984) Synthesized clustering: a method for amalgamating alternative clustering bases with differential weighting of variables. Psychometrika 49:57–78

DeSoete G, DeSarbo WS, Caroll JD (1985) Optimal variable weighting for hierarchical clustering: an alternating least-squares algorithm. J Classif 2:173–192

Fehr WR (1987) Principles of cultivar development, vol 1: theory and technique. Macmillan Publishing Company, New York, pp 95–105

Flake RH, von Rudloff E, Turner BL (1969) Quantitative study of clinal variation in *Juniperus virginiana* using terpenoid data. Proc Natl Acad Sci USA 64:487–494

Gauch HG, Zobel RW (1988) Predictive and postdictive success of statistical analyses of yield trials. Theor Appl Genet 76: 1–10

Ghaderi A, Adams MW, Saettler AW (1982) Environmental response patterns in commercial classes of common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 63:17–22

Hayward MD, DeLacy IH, Tyler BF, Drake DW (1982) The application of pattern analysis for the recognition of adaptation in a collection of *Lolium multiflorum* populations. Euphytica 31:383–396

Hogarth DM, Bull JK (1990) The implications of genotype × environment interactions for evaluation of sugarcane families. I. Effect on selection. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding. Louisiana State University, Baton Rouge, Louisiana, pp 335–344

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2: 193–218

Johnson RW (1982) Effect of weighting and the size of the attribute set in numerical classification. Aust J Bot 30:161–174

Lumelsky VJ (1982) A combined algorithm for weighting the variables and clustering in the clustering problem. Pattern Rec 15:53–60

Milligan GW (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45:325–342

Milligan GW (1989) A validation study of a variable weighting algorithm for cluster analysis. J Classif 6:53–71

Milligan GW, Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Mult Behav Res 21:441–458

Milligan GW, Cooper MC (1988) A study of standardization of variables in cluster analysis. J Classif 5:181–204

Shorter R, Byth DE, Mungomery VE (1977) Genotype × environment interactions and environmental adaptation. II. Assessment of environmental contributions. Aust J Agric Res 28:223–235

Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. Freeman, San Francisco

Thompson WA, Moore JR (1963) Non-negative estimates of variance components. Technometrics 5:441–449

Thorpe RS (1985) The effect of insignificant characters on the multivariate analysis of simple patterns of geographic variation. Biol J Linn Soc 26:215–223

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Williams WT (1971) Principles of clustering. Annu Rev Ecol Systems 2:303–326

Wishart D (1969) Mode analysis: a generalisation of nearest neighbour which reduces chaining effects. In: Cole AJ (ed) Numerical taxonomy. Academic Press, London, pp 282–311

Yau SK (1991) Need of scale transformation in cluster analysis of genotypes based on multi-location yield data. J Genet Breed 45:71–76