Original article

Selecting genotypes by clustering, for qualitative genotype by environment interaction, using a non-symmetric inferiority score

J Moro [1]*, JB Denis [2]

[1]*Department of Forestry, Cifor-Inia, Apdo 8111, 28080 Madrid, Spain;*
[2]*Laboratoire de biométrie, Inra, F-78026 Versailles cedex, France*

(Received 16 January 1997; accepted 10 June 1997)

Summary — A clustering method to reduce the difficulties in selection found in plant trials by the frequently existing qualitative or crossover interaction is described. It is based on an order-constrained inferiority score and avoids the need to postulate a model for the interaction. The resulting dendrogram is analyzed to arrive at a proper stopping point. Each final cluster includes a 'leader' genotype, which has the lowest inferiority score and would show no significant crossover (rank change) with the others. The method can be seen as a preselection of some genotypes, each one uniformly not inferior to others in the same group; the latter can be disregarded for subsequent analysis. When the procedure is compared with some previous proposals it seems to provide a better and simpler solution.

genotype-by-environment interaction / crossover interaction / qualitative interaction / clustering / inferiority score

Résumé — Sélection des génotypes par classification, pour l'interaction qualitative entre génotypes et milieux, selon un score d'infériorité non-symétrique. Les interactions qualitatives génotype–milieu (changement de rang des génotypes d'un environnement à un autre) sont une difficulté majeure pour la sélection des meilleurs génotypes. Pour traiter cette difficulté une méthode nouvelle de classification des génotypes est proposée. Contrairement à d'autres approches récentes, basées sur une modélisation paramétrique, la méthode ne suppose aucune forme particulière de l'interaction et s'appuie sur un indice non symétrique calculé pour tout couple de génotypes sur l'ensemble des milieux (score d'infériorité). La classification obtenue est de type hiérarchique et à chaque étape les groupes de génotypes formés sont représentés par un génotype dit supérieur (leader) supposé uniformément meilleur que tous les génotypes de son groupe. L'idée finale est que le sélectionneur puisse restreindre son choix à un sous-ensemble de génotypes supérieurs; les autres génotypes ne présentent plus d'intérêt puisque inférieurs à au moins un génotype supérieur dans chacun des milieux. La règle d'arrêt utilisée pour l'interprétation du dendrogramme est fondée sur un test statistique de non changement de rang.

interaction génotype–milieu / interaction qualitative / classification / score d'infériorité

Article communicated by Max Rives (Avignon)

* Correspondence and reprints
Tel: (34) 1 347 68 32; fax: (34) 1 357 22 93; e-mail: jmoro@inia.es

INTRODUCTION

The basic objective of plant genotype trials is to select those new genotypes that are the best (with high expected performance in a univariate perspective) across locations and years. This selection is seriously complicated by the existence of genotype by location interaction when it implies rank changes of genotypes from one location to another.

Numerous statistical approaches and models have been developed for interpreting the interaction term added to the additive scheme. For a review of recent developments, see for instance Kang and Gauch (1996). Most of the models are based on a simplified form of interaction. Among the main families of such models are the biadditive (or multiplicative) models, the regression models and the clustering approaches. Biadditive models (Gauch, 1992; Denis and Gower, 1996) constrain the interaction to bilinearity expressed by multiplicative terms; they are very flexible because the number of multiplicative terms can be adapted to the complexity of the phenomenon. Regression models, such as joint regression (Finlay and Wilkinson, 1963) or factorial regression (Denis, 1988), incorporate internal or external additional information into the analysis. Their efficiency has a direct relationship with the appropriateness of the covariates. The usual purpose of clustering methods (Lin and Thompson, 1975; Lefkovitch, 1985; Corsten and Denis, 1990; Denis and Moro, 1995; Moro and Denis, 1995) is to find groups of genotypes and/or locations such that additivity be acceptable within groups.

Two main criticisms can be raised with respect to these well established procedures: a) qualitative interaction (Azzalini and Cox, 1984) or synonimously, crossover interaction or COI (Gail and Simon, 1985; Baker, 1988; Virk and Mangat, 1991; Cornelius et al, 1992), which is in fact the main concern for plant breeders, is not distinguished from simpler forms of departures from additivity; b) good and bad genotypes are equally considered for choosing and fitting a compromise model, whereas plant breeders are only interested in good genotypes.

Analysis of genotype by environment interactions by ranks (Hühn et al, 1993; Hühn, 1996) indeed deals only with qualitative interactions but it seems to us that the loss of information implied by transforming yields into ranks is too drastic to be the only way of analyzing such costly experiments.

Other statistical approaches use order constraints to estimate and/or test the expected response. Nevertheless, these proposals are either based on a priori orders, typically associated with a quantitative factor (Hirotsu, 1978; Salama and Quade, 1981; Marcus and Talpaz, 1983; Mansouri, 1989), or are not very advanced (Denis, 1979, 1982; Guénoche et al, 1994).

In fact, Cornelius and coworkers (Cornelius et al, 1992, 1993; Crossa et al, 1993, 1995, 1996) seem to be the first authors to have tackled both points a) and partially b). They associate the clustering of genotypes and SHMM with one multiplicative term, independently within each cluster. SHMM (shift multiplicative model) was proposed by Seyedsadr and Cornelius (1992, 1993); it is a biadditive model that with one multiplicative term is equivalent to the implicit model suggested by Tukey (1949) and studied later by Mandel (1971) and many others. This relaxes the standard requirement for additivity and avoids COI if the point of concurrence of regression lines is outside the range of location main effects. They state: "... Clustering would allow breeders and growers to restrict selection of cultivars to the better ones in each cluster".

Here we propose something beyond this first step, ending with a set of 'leader' genotypes, as small as possible, such that every non-leader genotype is surpassed in performance by at least one 'leader' in each location. The selection of 'leader' genotypes is not model-based as in the previous approach but it is attained by hierarchical clustering based on an inferiority score. An analysis of the resulting dendrogram is made to justify the stopping or cutting point on a statistical basis. The flexibility and power of the proposed method is demonstrated with the data set of Cornelius et al (1993).

METHOD

Clustering

In a multisite trial of I genotypes tested at J locations the general response or yield y_{ij} of the ith genotype at the jth location is usually assumed to be distributed as an independent normal variate with mean μ_{ij} and constant variance σ^2. An estimate $\hat{\sigma}^2$ of this variance is supposed to be available based on g degrees of freedom. Let the inferiority of genotype i with respect to the genotype i' at the jth location be:

$$l_{ii'.j} = \min(0, d_{ii'.j})$$

$$d_{ii'.j} = y_{ij} - y_{i'j}$$

with min*(a,b)* denoting the minimum of *a* and *b*.

The inferiority score of the *i*th genotype with respect to the *i'*th is defined, as

$$L_{ii'} = \sum_{j=1}^{J} l^2_{ii'.j} / \hat{\sigma}^2$$

This score is non-negative, being zero when the first genotype is uniformly superior to the second. It increases when the first genotype is being outperformed by the second. Obviously, $L_{ii} = 0$ and, in general, $L_{ii'} \neq L_{i'i}$. Moreover, $L_{ii'}$ can be greater than $L_{ii''} + L_{i''i'}$. Thus, no important metric properties can be assigned to this score. The introduction of σ in the definition of $L_{ii'}$ is unnecessary in this case, but we prefer to make it appear for easier interpretation of computation results and for the possible generalizations to the heteroscedastic or to the multivariate case.

Other expressions could have been used to define an inferiority score. Any order inverting mapping of the $l_{ii'.j}$ into R^+, for instance the absolute value function, implying in this case only a reflection around the origin, would provide similar classification results. We made the above choice to maintain some parallelism with the sums of squares of random normal variables, which are useful quantities to test statistical hypotheses.

The clustering algorithm is very simple. First, the non-symmetric matrix $\mathbf{L} = \{L_{ii'}\}$ of order I is computed once and for all. Then a pair of genotypes, say (*a, b*) such that L_{ab} is minimal is identified. Both genotypes form the first cluster. This assignment is stored and the corresponding score is kept for later use in scaling the dendrogram. The dimension of the **L** matrix is reduced by 'voiding' the row and column of the inferior genotype. The cluster is represented by the scores of *a*, the superior or 'leader' genotype. Then, the algorithm iterates. The following minimal value, say (*a', b'*), is found. If it happens that *a' = a* the genotype *b'* is added to the first cluster. If not, a second cluster is started with *a'* as its 'leader'. In this case, if *b' = a*, then genotypes *a* and *b* are included in this cluster. The procedure follows in this way until the **L** matrix is reduced to only one element and all genotypes are in the same cluster led by one of them.

Sometimes, particularly at the early steps when there are inferior genotypes dominated by many others, it may happen that the minimum value is attained by different pairs of genotypes. Though the handling of these equalities is secondary as the interest would normally center on the last clustering steps, one has to distinguish two cases, as follows.

(i) The case of the existence of several potential 'leaders' a_1, a_2,... Then the algorithm chooses the first genotype with the minimum total marginal

$$\min_i \sum_k L_{ik},\ i \in \{a_1, a_2, \ldots\}$$

In case of ties, instead of the first genotype, a random choice among the candidates could be performed.

(ii) The case of several genotypes b_1, b_2,... that could be joined to the same 'leader' *a*. Then, the algorithm chooses that with the greatest inferiority score with respect to *a*. More formally: let $S_a = \{b_1, b_2, \ldots\}$ be the set of genotypes with index value

$$L_{ab_1} = L_{ab_2} = \ldots = \min L_{ii'}$$

The genotype chosen is the first *b* with L_{ba} = max L_{ha}, $h \in S_a$. Alternatively, in case of ties, instead of the first genotype, one may be chosen at random within the candidates.

One difference with other clustering methods based on distances or similarities is that they define distances so that clusters adequately represent all their members. But the present procedure disregards the inferior genotypes and uses only the scores of the 'leader' to represent the cluster. In this it is similar to the Leader Algorithms in chapters 3 and 9 of Hartigan (1975), but these are fast clustering algorithms using a priori threshold values and whose results depend on the initial ordering of the objects.

Apart from cases of ties, the clustering results are invariant to the order of levels of the factors as well as to changes of scale and location. To the usual dendrogram showing the clustering steps one should add the indication of the 'leader' genotypes. To achieve more visual clarity in drawing the dendrogram we have used the successive accumulation of the inferiority scores as horizontal scale.

Analysis of the dendrogram

To determine the stopping point we propose to examine the groups by proceeding backwards following the branches of the dendrogram from the root to the leaves, starting from a unique best genotype and looking for more 'leaders'. The question is to decide if within a group there is no significant crossover interaction of the 'leader' genotype with the rest. If the hypothesis of no COI is rejected the branch is followed and the group is split. If it is not rejected the tree is pruned at the current point and no further progress is made on this branch. The process stops when the tree is completely pruned and there is no COI declared significant.

The testing of the no COI hypothesis is made for each tetrad involving the 'leader' genotype with any of the others. A tetrad $\theta_{ii'.jj'} = d_{ii'.j} - d_{ii'.j'}$ is a COI tetrad if sign($d_{ii'.j}$) ≠ sign($d_{ii'.j'}$). It is declared significant if both standardized differences are greater in absolute value than some critical deviate on the Student's *t*-distribution for a probability value α_1; this value is determined to maintain the global type I error equal to the chosen significance level α.

Therefore, a COI tetrad is declared significant at the level α if

$$\left\{|d_{ii'.j}| > t_{g;\alpha_1} * \sqrt{2} * \hat{\sigma}\right\} \text{ and } \left\{|d_{ii'.j'}| > t_{g;\alpha_1} * \sqrt{2} * \hat{\sigma}\right\}$$

with $t_{g:\alpha_1}$ the Student's deviate with g degrees of freedom at level α_1. As determined by Azzalini and Cox (1984), α_1 must be equal to $\sqrt{\alpha / 2}$ to provide a joint α level.

This test is used by Cornelius et al (1993), who also proposed the use of the global test for COI derived also by Azzalini and Cox (1984). This last test is not adequate for the present procedure. If one wants to achieve a more global level of protection, eg, at the cluster level, the test derived by Gail and Simon (1985), can be used after adjusting the significance level for the total number of comparisons within cluster (Baker, 1988).

One should realize that with the application of the procedure, a 'leader' genotype is not declared to be confidently superior to other genotypes at all locations. Thus, some care should be exercised before the elimination of inferior genotypes.

APPLICATION

As an example we apply the procedure to the data of table I in Cornelius et al (1993). They correspond to the yield of 41 cultivars grown in seven locations. The letters r and c have been added, respectively, to the identification of cultivars (rows of the data table) and locations (columns of the data table). The within variance estimate is taken from their table II to be equal to 426 with 846 degrees of freedom. The dendrogram is presented in figure 1. Twenty-four clustering steps (not shown in the dendrogram) were processed before the inferiority score, used for merging, took its first positive value. At this point there were four clusters and 12 isolated cultivars. The analysis of the dendrogram requires moving from the root, numbered step 1 for convenience, to the 10th step to achieve no significant COI in any of the groups at the 5% level. If one pursues a clusterwise protection against the type I error with the more conservative Gail-Simon test, the 7th step is enough. It would define four groups and three isolated cultivars but five cultivars would still be involved in significant COI according to the test of individual tetrads at the same level.

A summary of the results at step 10 is presented in table I. Six groups and four isolated cultivars are obtained. There are 52 COI tetrads but none is significant at the 5% level. Group 1, the largest, is presented in table II. The first line corresponds to r26, the 'leader' cultivar. Figure 2a, b is presented as an example and corresponds to groups 2 and 7, respectively. For each cultivar in these groups the points given by the location main effect and its standardized yield ($y_{.j}$, $y_{ij} / \hat{\sigma}$) are linked by line segments. The dominance of the 'leader' cultivar is clearly appreciated in each figure.

Table I. Summary of the analysis of the clusters at step 10 of the dendrogram of figure 1.

Group	*'Leader'*	*Number*	*N Tetr*	*N Sig*	*N no*	*N*	*N Sig*	*Mean COI*	*Max.*
			ads	*Tetrads*	*COI*	*COI*	*COI*		*COI**
1	r26	19	378	53	354	24	0	28.208	51.000
2	r13	6	105	16	105	0	0		
3	r24	2	21	2	16	5	0	30.900	55.000
4	r36	5	84	9	67	17	0	28.265	58.500
5	r12	1	0	0	0	0	0		
6	r17	1	0	0	0	0	0		
7	r38	3	42	8	42	0	0		
8	r23	1	0	0	0	0	0		
9	r28	1	0	0	0	0	0		
10	r37	2	21	2	15	6	0	29.500	52.000

'Leader' = code of the majorizing cultivar. Number = number of cultivars composing the group. N tetrads = total number of tetrads involving the 'leader' cultivar. N Sig. Tetrads = of these tetrads those being significant after a standard t-test at the 5% level. N no COI = number of tetrads which are not COI. N COI = number of tetrads being COI. N Sig COI = number of COI tetrads being significant according to the joint t-test at $\alpha = 0.05$. * The last two columns refer to all COIs (mean and maximum semisum of absolute values of $d_{ii'.j}$) No significant COI after Gail-Simon test (0.05) at this step.

Table II. Performances of cultivars composing group 1 at step 10 of the dendrogram.

	c1	*c2*	*c3*	*c4*	*c5*	*c6*	*c7*	*Mean*	*Order*
r26	361	379	341	566	492	426	238	400.43	1
r1	185	287	191	490	408	299	133	284.71	16
r2	198	323	109	497	379	340	124	281.43	17
r3	183	302	143	442	390	337	85	268.86	18
r5	307	408	353	526	500	400	165	379.86	3
r7	306	369	341	566	503	450	233	395.43	2
r9	225	259	222	479	472	303	179	305.57	12
r10	258	264	339	505	463	314	193	333.71	8
r11	253	267	282	398	448	368	111	303.86	13
r14	232	314	311	469	442	319	172	322.71	10
r16	181	345	208	525	347	380	170	308.00	11
r19	237	262	242	461	403	348	98	293.00	15
r20	282	226	222	424	434	327	180	299.29	14
r25	280	342	315	527	432	397	201	356.29	5
r29	314	338	301	503	505	402	197	365.71	4
r32	245	280	319	472	426	349	185	325.14	9
r33	278	309	298	529	445	374	182	345.00	6
r39	248	231	248	279	359	305	117	255.29	19
r40	274	340	266	475	463	369	176	337.57	7
Mean	255.11	307.63	265.84	480.68	437.42	358.26	165.21	324.31	

The first line corresponds to the 'leader' cultivar. The last column is the order according to the mean.

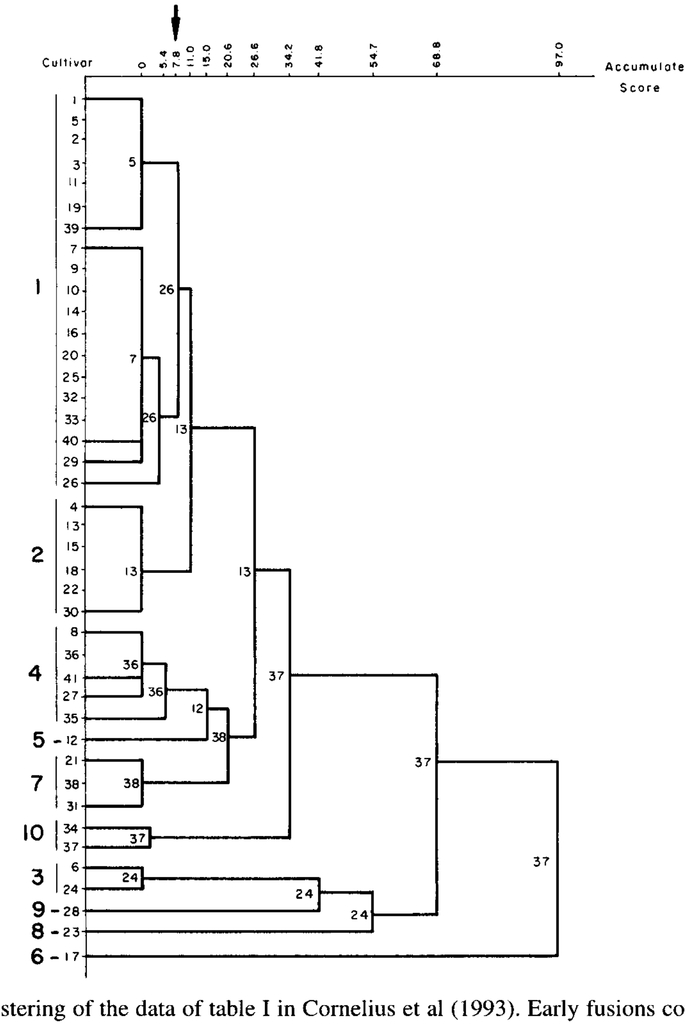

Fig 1. Dendrogram from clustering of the data of table I in Cornelius et al (1993). Early fusions corresponding to inferiority scores equal to zero are not depicted. The fusions shown start at step 25 with L = 0.15, joining cultivars 6(r6) and 24(r24). The arrow points to the proposed cutting or stopping point. The big numbers on the left identify the resulting clusters.

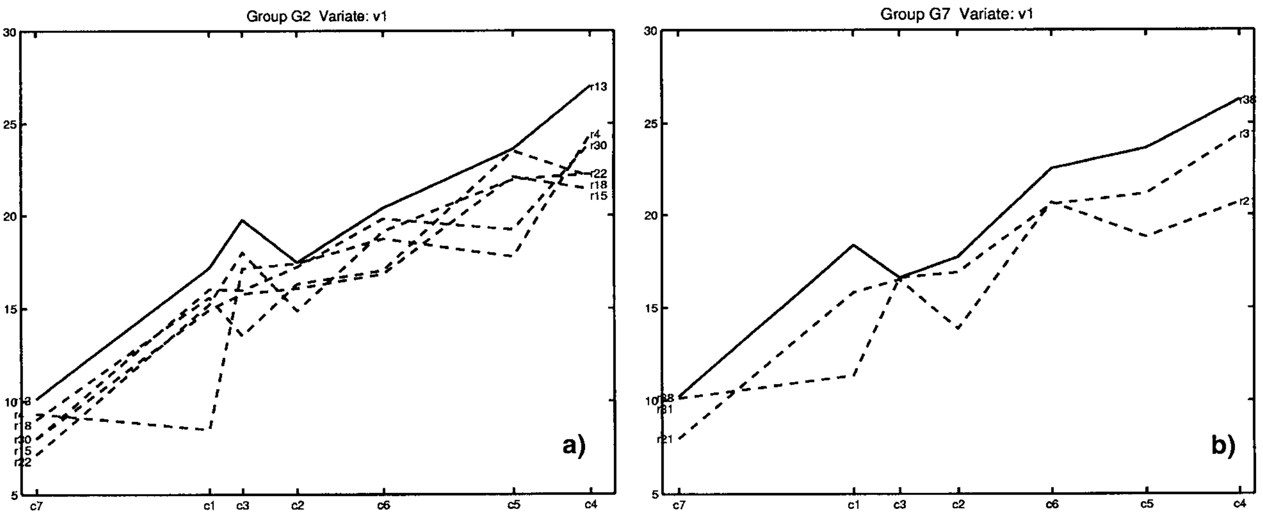

Fig 2. Groups 2 and 3. Ordinates are the standardized yields at each location. On the abscissas the location main effects. a. Cultivars forming group 2. The 'leader' is r13. b. Cultivars forming group 7. The 'leader' is r38.

The analysis of Cornelius et al produced eight clusters plus six isolated cultivars. It separates cultivars r4, 6, 7, 16 and 21. However, cultivar r26 is uniformly greater than r16 and almost the same as r7 (see table II). The same happens with cultivar r4, which is majorized by r13 in group 2 (fig 2a), r6 by r24 in group 3 and r21 inferior to r38 in group 7 (fig 2b). On the other hand Cornelius et al do not separate r17, which is included in the same group as r13, both showing important COI.

The response of the selected cultivars is shown in figure 3. Similarly to previous figures it shows the broken lines joining the points of coordinates $(j, y_{ij} / \hat{\sigma})$ for $i \in \{i_1, i_2, \ldots, i_K\}$, i_k indicating the 'leader' cultivar and K the number of retained groups. Locations are ordered after their main effects. The picture indicates the existence of a complicated pattern of interaction but the size of the problem has been considerably reduced. The application of some methods from the full panoply available to explain and model the interaction can now be explored, including those proposed by Cornelius and coworkers.

The duality of cultivar and locations for this problem should make possible the use of the same tools under a somewhat modified strategy to partition the locations for further simplifying of the problem.

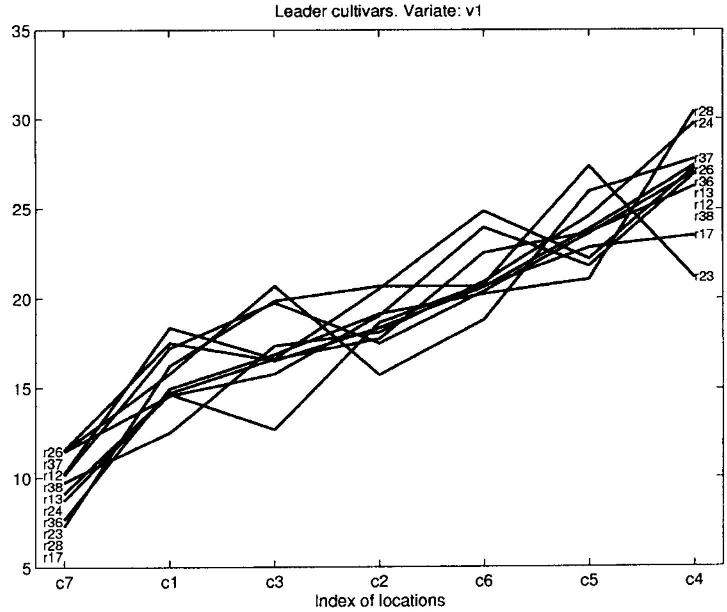

Fig 3. Joint response of the ten 'leader' cultivars. Ordinates are the standardized yields at each location. On the abscissas is the rank of location main effects.

ACKNOWLEDGMENT

We would like to thank the editor and an anonymous referee for his/her useful observations. This work has been carried out under the auspices of AIR3-CT94-1330 (J Moro); Inra/Inia-Cifor Bilateral Agreement (JB Denis).

REFERENCES

Azzalini A, Cox DR (1984) Two new tests associated with the analysis of variance. *J R Stat Soc* 46, 335-343

Baker RJ (1988) Tests for crossover genotype-environment interactions. *Can J Plant Sci* 68, 405-410

Cornelius PL, Seyedsadr MS, Crossa J (1992) Using the shifted multiplicative model to search for separability in crop cultivars trials. *Theor Appl Genet* 84, 161-172

Cornelius PL, Van Sanford DA, Seyedsadr MS (1993) Clustering cultivars into groups without rank-change interactions. *Crop Sci* 33, 1193-1200

Corsten LCA, Denis JB (1990) Structuring interaction in two way Anova tables by clustering. *Biometrics* 46, 207-215

Crossa J, Cornelius PL, Seyedsadr M, Byrne P (1993) A shifted multiplicative model cluster analysis for grouping environments without genotypic rank change. *Theor Appl Genet* 85, 577-586

Crossa J, Cornelius PL, Sayre K, Ortiz-Monasterio JI (1995) A shifted multiplicative model fusion method for grouping environments without cultivar rank change. *Crop Sci* 35, 54-62

Crossa J, Cornelius PL, Seyedsadr MS (1996) Using the shifted multiplicative model cluster methods for crossover genotype-by-environment interaction. In: *Genotype by Environment Interaction* (Kang MS, Gauch HG Jr eds), CRC Press, Boca Raton, 175-198

Denis JB (1979) Sous modèles interactifs respectant des contraintes d'ordre. *Biom Praxim* 19, 49-58

Denis JB (1982) Test de l'interaction sous contraintes d'ordre. *Biom Praxim* 22, 29-45

Denis JB (1988) Two way analysis using covariates. *Statistics* 19, 123-132

Denis JB, Gower JC (1996) Asymptotic confidence regions for biadditive models, interpreting genotype-environment interactions. *Appl Stat* 45(4), 479-493

Denis JB, Moro J (1995) Multivariate generalizations for modeling two-way interaction. Defining and estimating models. *Biuletyn Oceny Odmian* 26-27, 43-56

Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant breeding programme. *Aust J Agric Res* 14,742-754

Gail M, Simon R (1985) Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41, 361-372

Gauch HG Jr (1992) *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs.* Elsevier, Amsterdam, 278pp

Guénoche A, Vandeputte-Riboud B, Denis JB (1994) Selecting varieties using a series of trials and a combinatorial ordering method. *agronomie* 14, 363-375

Hartigan J (1975) *Clustering Algorithms.* John Wiley, New York, 351pp

Hirotsu C (1978) Ordered alternatives for interaction effects. *Biometrika* 65, 561-570

Hühn, M (1996) Non parametric analysis of genotype x environment interactions by ranks. In: *Genotype by Environment Interaction* (Kang MS, Gauch HG Jr, eds), CRC Press, Boca Raton, 235-271

Hühn M, Lolito S, Piepho HP (1993) Relationships between genotype x environment interactions and rank orders for a set of genotypes tested in different environments. *Theor Appl Genet* 86, 943-950

Kang MS, Gauch HG Jr (eds) (1996) *Genotype by Environment Interaction.* CRC Press, Boca Raton, 416 pp

Lefkovitch LP (1985) Multicriteria clustering in genotype-environment interaction problems. *Theor Appl Genet* 70, 585-589

Lin CS, Thompson BK (1975) An empirical method of grouping genotypes based on a linear function of the genotype-environment interaction. *Heredity* 34, 255-263

Mandel J (1971) A new analysis of variance model for non-additive data. *Technometrics* 13, 1-18

Mansouri H (1989) Linear rank tests for homogeneity against ordered alternatives in anova and anocova. *Comm Stat A Theory Methods* 18, 4321-4334

Marcus R, Talpaz H (1983) On testing homogeneity of t normal means against ordered alternatives in r groups. *Comm Stat A Theory Methods* 12(24), 2897-2902

Moro J, Denis JB (1995) Multivariate generalizations for modeling two-way interaction. II: Interpreting models and examples. *Biuletyn Oceny Odmian* 26-27, 57-72

Salama JA, Quade D (1981) Using weighted ranking test against ordered alternatives in complete blocks. *Comm Stat A Theory Methods* 10, 385-399

Seyedsadr MS, Cornelius PL (1992) Shifted multiplicative models for nonadditive two way tables. *Comm Stat B Simul Comp* 21, 807-832

Seyedsadr MS, Cornelius PL (1993) Hypothesis testing for components of the shifted multiplicative model for a non-additive two way table. *Comm Stat B Simul Comp* 22, 1065-1078

Tukey J (1949) One degree of freedom for non additivity. *Biometrics* 5, 232-242.

Virk DS, Mangat BK (1991) Detection of crossover genotype x environment interactions in pearl millet. *Euphytica* 52, 193-199