Crop Breeding and Applied Biotechnology 4:285-289, 2004
Brazilian Society of Plant Breeding. Printed in Brazil

Discriminant analysis for the classification and clustering of robusta coffee genotypes

Aymbiré Francisco Almeida da Fonseca[1*], Tocio Sediyama[2], Cosme Damião Cruz[3], Ney Sussumu Sakiyama[2], Romário Gava Ferrão[4], Maria Amélia Gava Ferrão[1], and Scheilla Marina Bragança[5]

Received 4 March 2004

Accepted 7 July 2004

***ABSTRACT** - This study evaluated the adequacy of the composition of three clonal* Coffea canephora *varieties recommended for the State of Espírito Santo by a multivariate method designated discriminant analysis. This method consists in the establishment of functions that enable the classification of a given individual into one, among various distinct populations, reducing the probability of a misclassification. It simultaneously considers measures of several traits, in order to give the new variety homogeneity. The original classification of genotypes in the three studied varieties, based on agronomical criteria, maintained expressive concordance with the results of the discriminant analysis, with an apparent deviation rate of only 6.25%. Corrected discriminant functions were also proposed, capable of classifying a new genotype into one of the three clonal varieties to be used in improvement programs, eliminating the subjectivity of the clustering process.*

Key words: *Coffea canephora*, clonal varieties, improvement, robusta coffee

INTRODUCTION

Natural allogamy restricts the trait fixation of a particular material in sexual propagation, so the natural *Coffea canephora* populations are highly heterozygous, with broad genetic variability in practically all traits of interest (Vossen 1985, Carvalho et al. 1991, Fonseca 1996).

The vegetative propagation of elite plants maintains the selected traits (Charrier and Berthaud 1988). Obtaining clonal varieties has therefore come to be one of the most commonly applied strategies in improvement programs with the species (Berthaud 1986, Charrier and Berthaud 1988). In many countries, clonal varieties are currently the basic material for the production of this coffee species (Dublin 1967, Ferwerda 1969, Vossen 1985, Bouharmont et al. 1986).

A clonal *C. canephora* variety is gradually obtained through a series of stages. Once the clones of interest, bearers of the desirable trait that is to be improved, are selected, they have to be clustered. Besides, the genetic compatibility among them and other common traits have to be taken into account to give the new variety homogeneity in plant height, architecture, bean weight, yield index of processing, and

[1] Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (INCAPER), CRDR, Centro Serrano, Rodovia BR 262, Km 94, Fazenda do Estado, 29375-000, Venda Nova do Imigrante, ES, Brasil. *E-mail: aymbire@incaper.es.gov.br
[2] Departamento de Fitotecnia, Universidade Federal de Viçosa (UFV), 36570-000, Viçosa, MG, Brasil
[3] Departamento de Biologia Geral, BIOAGRO, UFV
[4] INCAPER, Rua Afonso Sarlo, 29, 29052-010, Vitória, ES, Brasil
[5] INCAPER, CRDR, 29900-190, Linhares, ES, Brasil

especially, in fruit maturation, among others (Fonseca 1995, 1996, Ferrão et al. 1999). In such cases, researchers are not only interested in isolated studies of a given trait, but in the simultaneous performance of many of them.

In numerous studies into plant breeding, the use of the theory of the multivariate analyses appears to be promising, as it allows the combination of all information held in the experimental unit, so that the inferences are based on a complex of variables. Therefore, one of the aims of discriminant analysis consists in the establishment of functions that allow the classification of a given individual, based on the measures of several traits, in one among several distinct populations. This is an attempt to minimize the probability of a misclassification, i.e., to classify the referred individual into a population, when it would actually belong to another one (Fisher 1936).

This kind of situation was initially described by Fisher (1936) by a linear combination of the observed traits with a clearer discrimination power among the groups. This combination is called the Linear Discriminant Function of Fisher, based on a thorough study of the discriminant analysis. This function minimizes the probability of misjudging a classification when the populations are distributed normally with known mean and variance.

According to Anderson (1958), when several populations are available and one wants to know to which one of them a new individual should belong, an important procedure, besides assuming some distribution to obtain the discriminant functions, is the establishment of *a priori* probabilities for the various populations. The reason is that there are cases where the probability that a certain individual belongs to a given population can be quite distinct from the one that it would belong to another, so that the researcher's experience becomes an extremely important factor.

This study had the objective of assessing the adequacy of the composition of the first three clonal varieties of *C. canephora* recommended for the State of Espírito Santo, EMCAPA 8111, EMCAPA 8121, and EMCAPA 8131, formed by 9, 14, and 9 clones respectively, based on the multivariate method designated discriminant analysis. Furthermore, an adjustment of the discriminant functions, enabling the non-subjective classification of a new selected genotype into one of the cited three populations was proposed.

MATERIAL AND METHODS

Initially, we tried to verify the adequacy of the composition of three robusta coffee (*Coffea canephora*) varieties: EMCAPA 8111 (early maturation), EMCAPA 8121 (intermediate maturation) and EMCAPA 8131 (late maturation) in relation to the classification of its clones, considering 17 traits. The multivariate technique proposed by Anderson (1958), designated discriminant function, was used. It is assumed to optimize the genotype classification when a set of traits from each genotype is considered simultaneously.

To establish discriminant functions, information on genetic material that is known to belong to a certain group, bearer of proper and well-defined traits was considered. Groups 1, 2, and 3 were taken into consideration which contained, respectively, clones of the varieties EMCAPA 8111 (9 clones), EMCAPA 8121 (14 clones), and EMCAPA 8131(9 clones), as described by Ferrão et al. (1999).

The following traits were used: mean weight of 1.000 flat beans (P1000); percentage of "mocha" beans retained in sieve 13 (M13); percentage of "peaberry" beans retained in sieve 12 (M12); percentage of flat beans (GCH); percentage of shell-shaped beans (GCO); yield index of processing (IR), determined by the ratio freshly harvested coffee biomass: coffee biomass after processing; sieve mean of flat beans (PGCH); harvest time (EC); percentage of flat beans with a higher sieve mean than 13 (PS13); number of orthotropic shoots per plant (NHP); mean canopy diameter (MCD); mean plant height (MPH); mean yield per plant in 1989 (P89), 1990 (P90), 1991 (P91), and 1992 (P92); and mean of the four years (PMG).

Other available traits with a strong correlation to the aforementioned were not taken into consideration since this would lead to a strong colinearity, thus affecting the analysis results (Cruz 2001).

Data were obtained in an experiment set up in March 1987 in Marilândia, State of Espírito Santo, in a randomized complete block design with four replications. The plots contained six plants, spaced 3.0 x 1.5 m. To establish the discriminant functions, we considered the mean of the four evaluation replications carried out during the fourth harvest, with exception of the traits P89, P90, P91, P92, and PMG.

Let:

π_1 = population bearing trait of group 1;

u_1 = vector of means of the p traits evaluated in population π_1;

$\sum_1$ = matrix of the covariance among the evaluated traits in population π_1;

π_2 = population bearing trait of group 2;

u_2 = vector of means of the p traits in population π_2;

$\sum_2$ = matrix of covariance among the traits evaluated in population π_2;

π_3 = population bearing trait of group 3;
u_3 = vector of means of the p traits in population π_3;
$\sum_3$ = matrix of covariance among the evaluated traits in population π_3; and
χ = vector of representative variables of the traits involved in the analysis.

Considering that there is homogeneity among the matrixes of covariances $\sum_1$, $\sum_2$, and, $\sum_3$ the matrix $\sum$, brought forth by the combination of $\sum_1$, $\sum_2$, and $\sum_3$ is obtained , observing the respective freedom degrees.

Based on the theory of Anderson (1958), the discriminant functions are obtained for groups 1, 2, and 3, respectively, by the expressions:

$$D_1(\mathrm{x}) = \ln(p_1) + \left(\mathrm{x} - \tfrac{1}{2}u_1\right)\sum^{-1}u_1$$

$$D_2(\mathrm{x}) = \ln(p_2) + \left(\mathrm{x} - \tfrac{1}{2}u_2\right)\sum^{-1}u_2$$

$$D_3(\mathrm{x}) = \ln(p_3) + \left(\mathrm{x} - \tfrac{1}{2}u_3\right)\sum^{-1}u_3$$

Hence, $D_1(x)$, $D_2(x)$, and $D_3(x)$ are the discriminant functions that make it possible to obtain scores for the genotype classification in the populations π_1 , π_2, and π_3, considered to be bearers of the desired traits in groups 1, 2, and 3, respectively, and p_1, $p_{2,}$ and p_3 are the *a priori* probabilities to belong to the populations π_1, π_2, and π_3, respectively. In this study, these values were considered 0.333, as there was no previous information on the performance of the material to be classified.

Anderson's criterion (1958) classifies the ith genotype with the mean X_i vector in population π_j (j = 1, 2, or 3) if, and only if $D_j(x_i)$ is the largest element of the set $\{D_1(x_i), D_2(x_i), D_3(x_i)\}$. A particular genotype would therefore be classified as fitting better into group 1 in cases where the function D_1 (x) presented the highest value among the three obtained functions, and so on.

Using the discriminant functions and the data of the proper populations π_1, π_2 , and π_3, the apparent deviation rate that measures the efficiency of the discriminant function to classify genotypes correctly was estimated, whether it has the traits that include it in one or another group or not. Finally, the discriminant functions were estimated, considering an apparent deviation rate of zero, in other words, taking already classified clones as basis in each one of the populations.

RESULTS AND DISCUSSION

The discriminant functions of the groups 1 $D_1(\chi)$, 2 $D_2(\chi)$, and 3 $D_3(\chi)$ were obtained, using data of the proper populations π_1(EMCAPA 8111), π_2 (EMCAPA 8121), and π_3(EMCAPA 8131). Consequently, each discriminant function is a linear combination of the 17 traits observed in this analysis, according to the following expressions:

$$D_1(\chi) = -2312.426 - 3.855x\mathrm{P}1000 + 0.919x\mathrm{M}13 - 7.249x\mathrm{M}12 + 2.594xG C\mathrm{H} - 89.900xGC\mathrm{O} +$$
$$206.524x\mathrm{I}R + 277.802x\mathrm{P}GC\mathrm{H} - 0.289x\mathrm{E}C - 8.220x\mathrm{P}S13 - 29.777x\mathrm{NHP} + 121.329x$$
$$\mathrm{M}CD + 259.574x\mathrm{MPH} - 110.445x\mathrm{P}89 - 110.385x\mathrm{P}90 - 110.451x\mathrm{P}91 - 110.398x$$

$$D_2(\chi) = -2380.046 - 3.867x\mathrm{P}1000 + 0.952x\mathrm{M}13 - 7.378x\mathrm{M}12 + 2.649xGC\mathrm{H} - 91.897xGC\mathrm{O} +$$
$$+ 210.571x\mathrm{I}R + 282.248x\mathrm{P}GC\mathrm{H} - 0.010x\mathrm{E}C - 8.405x\mathrm{P}S13 - 30.076x\mathrm{NHP} + 124.425$$
$$x\mathrm{M}CD + 259.822x\mathrm{MPH} - 110.254x\mathrm{P}89 - 110.187x\mathrm{P}90 - 110.258x\mathrm{P}91 - 110.207x$$
$$\mathrm{P}92 + 441.008x\mathrm{PM}G$$

$$D_3(\chi) = -2321.157 - 3.856x\mathrm{P}1000 + 0.942x\mathrm{M}13 - 7.270x\mathrm{M}12 + 2.690xGC\mathrm{H} - 93.932xGC\mathrm{O}$$
$$+ 207.415x\mathrm{I}R + 277.992x\mathrm{P}GC\mathrm{H} + 2.350x\mathrm{E}C - 8.222x\mathrm{P}S13 - 30.138x\mathrm{NHP} +$$
$$122.886x\mathrm{M}CD + 255.327x\mathrm{MPH} - 113.366x\mathrm{P}89 - 113.303x\mathrm{P}90 - 113.374x\mathrm{P}91 -$$
$$113.320x\mathrm{P}92 + 453.462x\mathrm{PM}G$$

Table 1 presents the classification of the genotypes in the three groups 1, 2, and 3, as the respective estimates of the discriminant functions, according to the methodology of Anderson (1958). Presence of an adequate cluster for the great majority of the genotypes was verified. Genotypes ES 15 and ES 23 were exceptions, as they were classified in groups 1 and 3, respectively, although they had originally been classified as belonging to group 2, resulting in an apparent deviation rate of 6.25%.

$$\mathrm{P}92 + 441.777x\mathrm{PM}G$$

These results sufficiently back up the statement that the formation of the genotype groups that compose the clonal varieties EMCAPA 8111 and EMCAPA 8131 is quite adequate, but can be further improved in variety EMCAPA 8121 by the exclusion of genotypes ES 15 and ES 23.

However, as indicated by the methodology used, the inclusion of these genotypes into the clone group that represents the other varieties is possible, if care is taken so as not to affect their performance, once the important aspects such as genetic compatibility among the clones of a certain variety must be considered.

Although the apparent deviation rate was low, the estimate of the discriminant functions was repeated for the use of this methodology to classify other genotypes into these three groups. It is advisable that the genotypes ES 15 and ES 23 be placed in the indicated groups or then eliminated from the reference population to obtain an apparent deviation rate of zero. The new functions for the discrimination of genotypes with an unknown performance are thus provided with a greater statistic consistence, similarly to what Ferreira (1995) observed when he classified unknown rice genotypes according to their aluminum toxicity-tolerance. In this case, however, the performance of the genotypes used to obtain the estimates of the discriminant functions for aluminum toxicity-tolerance or intolerance was already well-known.

Table 1. Estimates of the discriminant functions for the *Coffea canephora* genotypes which are components of the varieties EMCAPA 8111, EMCAPA 8121, and EMCAPA 8131, and their respective classification according to the methodology of Anderson (1958)

Original group	Genotype	$D_1(x)$	$D_2(x)$	$D_3(x)$	Obtained classification[1]
Group 1	ES 01	2359.546	2358.133	2356.007	1
	ES 02	2272.189	2271.738	2268.526	1
	ES 05	2360.238	2358.807	2356.885	1
	ES 37	2263.887	2261.327	2259.349	1
	ES 07	2253.489	2250.332	2250.666	1
	ES 08	2320.021	2318.232	2317.413	1
	ES 09	2331.150	2330.367	2329.837	1
	ES 10	2291.992	2289.566	2289.029	1
	ES 22	2339.572	2339.200	2338.939	1
Group 2	ES 11	2467.127	2468.490	2466.540	2
	ES 12	2536.076	2540.416	2536.236	2
	ES 13	2314.282	2315.984	2313.055	2
	ES 14	2414.339	2415.302	2413.916	2
	ES 15	2313.445	2312.622	2311.363	1
	ES 16	2259.783	2260.202	2258.525	2
	ES 18	2433.636	2435.936	2433.493	2
	ES 19	2337.133	2338.025	2336.144	2
	ES 20	2388.374	2390.766	2388.102	2
	ES 23	2347.729	2348.314	2348.479	3
	ES 24	2462.522	2464.426	2461.438	2
	ES 25	2297.549	2299.850	2297.393	2
	ES 30	2311.407	2312.013	2309.869	2
	ES 28	2384.141	2387.568	2384.452	2
Group 3	ES 26	2316.126	2317.207	2319.474	3
	ES 27	2337.493	2337.407	2340.104	3
	ES 31	2265.599	2267.371	2269.697	3
	ES 34	2240.195	2238.839	2241.862	3
	ES 36	2344.685	2345.856	2346.999	3
	ES 92	2390.710	2389.547	2391.900	3
	ES 38	2277.619	2278.848	2281.086	3
	ES 39	2353.117	2353.854	2355.566	3
	ES 21	2319.686	2321.860	2323.970	3

[1] Groups 1, 2 and 3. Apparent deviation rate equal to 6.25% (ES 15 - original classification = group 2, and ES 23 original classification = group 2)

The estimates of the discriminant functions were therefore repeated, considering genotypes ES 15 in group1 and ES 23 in group 3, as indicated by the analysis. The new discriminant functions presented in the following were obtained this way, which may be useful to fit a new genotype into one of the three groups.

$$\begin{gathered} D_1(\chi) = -2312.792 - 3.861x\mathrm{P1000} + 0.945x\mathrm{M13} - 7.270x\mathrm{M12} + 2.583xGC\mathrm{H} - 89.885xGC\mathrm{O} \\ + 205.946x\mathrm{I}R + 277.960x\mathrm{P}GC\mathrm{H} - 0.256x\mathrm{E}C - 8.214x\mathrm{P}S13 - 29.708x\mathrm{NHP} + 121.158 \\ x\mathrm{M}CD + 260.113x\mathrm{MPH} - 111.952x\mathrm{P89} - 111.894x\mathrm{P90} - 111.959x\mathrm{P91} - 111.906x\mathrm{P92} \\ + 447.809x\mathrm{PM}G \end{gathered}$$

$$\begin{gathered} D_2(\chi) = -2389.027 - 3.887x\mathrm{P1000} + 0.957x\mathrm{M13} - 7.383x\mathrm{M12} + 2.672xGC\mathrm{H} - 91.779xGC\mathrm{O} \\ + 210.571x\mathrm{I}R + 282.248x\mathrm{P}GC\mathrm{H} - 0.010x\mathrm{E}C - 8.405x\mathrm{P}S13 - 30.076x\mathrm{NHP} + 124.425 \\ x\mathrm{M}CD + 258.295x\mathrm{MPH} - 109.233x\mathrm{P89} - 109.165x\mathrm{P90} - 109.237x\mathrm{P91} - 109.185x\mathrm{P92} \\ + 436.921x\mathrm{PM}G \end{gathered}$$

$$\begin{gathered} D_3(\chi) = -2324.180 - 3.830x\mathrm{P1000} + 0.917x\mathrm{M13} - 7.267x\mathrm{M12} + 2.675xGC\mathrm{H} - 94.084xGC\mathrm{O} \\ + 208.556x\mathrm{I}R + 277.569x\mathrm{P}GC\mathrm{H} + 2.363x\mathrm{E}C - 8.215x\mathrm{P}S13 - 30.212x\mathrm{NHP} + 122.180 \\ x\mathrm{M}CD + 257.094x\mathrm{MPH} - 112.793x\mathrm{P89} - 112.728x\mathrm{P90} - 112.800x\mathrm{P91} - 112.747x\mathrm{P92} \\ + 451.166x\mathrm{PM}G \end{gathered}$$

CONCLUSIONS

The original classification of the genotypes into the three clonal varieties under study, based on agronomical criteria, presented expressive concordance with the results obtained by the discriminant analysis, with an apparent deviation rate of only 6.25%. The corrected discriminant functions can be used to include a new genotype into one of the three groups.

Análise discriminante para classificação e agrupamento de genótipos de café conilon

RESUMO - *Este trabalho averiguou a adequação da composição de três variedades clonais de* Coffea canephora *recomendadas para o Estado do Espírito Santo com base no método multivariado denominado análise discriminante. Este método consiste na obtenção de funções que permitem classificar um determinado indivíduo em uma, dentre várias populações distintas, minimizando a probabilidade de classificação equivocada. Baseia-se em medidas de várias características, consideradas simultaneamente, de forma a proporcionar homogeneidade à nova variedade. A classificação original dos genótipos nas três variedades estudadas, baseada em critérios agronômicos, manteve expressiva concordância com os resultados obtidos através da análise discriminante, com uma taxa de erro aparente de apenas 6,25%. Foram propostas funções discriminantes corrigidas, capazes de permitir a classificação de um novo genótipo em uma das três populações em questão, a serem utilizadas em programas de melhoramento, eliminando a subjetividade do processo de agrupamento.*

Palavras-chave: *Coffea canephora*, variedades clonais, melhoramento, café robusta

REFERENCES

Anderson TW (1958) **An introduction to multivariate statistical analysis**. John Wiley, New York, 374p.

Berthaud J (1986) **Les ressources génétiques pour l'amériolation des caféiers africains diploides. Evaluation de la richesse génétique des populations sylvestres et de ses mécanismes organisateurs. Conséquences pour l'application**. ORSTOM, Paris, 379p. (Document ORSTOM 188).

Bouharmont P, Lotodé R, Awemo A and Castaing X (1986) La sélection générative du caféier robusta au Cameroun: analyse des résultats d'hybrides diallèle partiel implanté en 1973. **Café Cacao, Thé 30**: 93-112.

Charrier A and Berthaud J (1988) Principles and methods in coffee plant breeding: *Coffea canephora* Pierre. In: Clark RJ and Macrae R (eds.) **Coffee agronomy**. Elsevier, London, p. 167-195.

Carvalho A, Medina Filho, HP, Fazuoli LC, Guerreiro Filho and Lima MNA (1991) Aspectos genéticos do cafeeiro. **Revista Brasileira de Genética 14**: 135-183.

Cruz CD (2001) **Programa Genes – Versão Windows: aplicativo computacional em genética e estatística**. Editora UFV, Viçosa, 642 p.

Dublin P (1967) L'amélioration du caféier robusta en République Centrafricaine: dix années de sélection clonale. **Café Cacao Thé 11**: 101-138.

Ferreira RP (1995) Identificação de cultivares de arroz tolerantes à toxidez de alumínio por técnicas multivariadas. **Pesquisa Agropecuária Brasileira 30**: 789-795.

Ferrão RG, Fonseca AFA and Ferrão MAG (1999) Programa de melhoramento genético de café robusta no Brasil. In: Paiva R (ed.) **Anais do III Simpósio de atualização em genética e melhoramento de plantas**. Universidade Federal de Lavras, Lavras, p. 50-65.

Ferwerda FP (1969) Breeding of *Coffea canephora*. In: Ferwerda FP and Wit F (eds.) **Coffee: *Coffea arabica* L. and *Coffea canephora* Pierre ex Froehner**. Agricultural University, Wageningen, p. 216 -241. (Miscellaneous Papers 4).

Fisher RA (1936) The use of multiple measurements in taxonomic problems. **Annals of Eugenics 7**: 179-188.

Fonseca AFA (1995) Variedades clonais de café conilon. In: **I Simpósio Estadual de Café**. CETCAF, Vitória, p. 29-33.

Fonseca AFA (1996) Propagação assexuada de *Coffea canephora* no Estado do Espírito Santo. In: Paiva R (ed) **Workshop sobre avanços na propagação assexuada de plantas lenhosas**. Editora UFLA, Lavras, p. 31-34.

Vossen HAM (1985) Coffee selection and breeding. In: Clinffort MN and Willson KC (eds.) **Coffee - botany, biochemistry and production of beans and beverage**. Croom Helm, London and Sidney, p. 48-96.