

SELECTION OF BOTANICAL AND AGRONOMICAL DESCRIPTORS FOR THE CHARACTERIZATION OF CASSAVA (*Manihot esculenta* Crantz) GERMPLASM*

Antonio Vander Pereira¹, Roland Vencovsky² and Cosme Damião Cruz³

ABSTRACT

Several cassava descriptors were studied to evaluate genetic diversity by the use of multivariate analysis. Data on 28 descriptors over 280 accessions from the EMBRAPA's germplasm bank were collected during two crop seasons. The analysis of principal components was used to discard the descriptors which were either redundant or non-discriminating and had a low stability of expression. Through this method it was possible to choose the best 14 descriptors. This reduction in number shall facilitate the characterization of cassava germplasm, without affecting the general information.

INTRODUCTION

Accessions in germplasm banks are usually characterized by botanical and agronomical descriptors, which can be biochemical or morphological in nature. However, the relevance of a particular descriptor is bound to its discriminating capacity and stability of expression.

Several botanical and agronomical descriptors have been proposed for the characterization of cassava germplasm (Conceição, 1979; Silva, 1981, 1982; Silva e Mendes, 1982; Goedert *et al.*, 1982; Gulick *et al.*, 1983). The evaluation of a large number

* Part of a thesis presented by A.V.P. to the Departamento de Genética, ESALQ-USP, in partial fulfillment of the requirements for the Doctoral degree.

¹ EMBRAPA/CNPGL, Rod. MG 133, km 42. 36155 Coronel Pacheco, MG, Brasil. Send correspondence to A.V.P.

² ESALQ/USP, Departamento de Genética, Caixa Postal 82, 13400 Piracicaba, SP, Brasil.

³ Departamento de Biologia, Universidade Federal de Viçosa, 36570 Viçosa, MG, Brasil.

of descriptors has been a general procedure, since there is a lack of precise information about the contribution of most characters to the description of the variability. However, when the number of descriptors is very large, the possibility that many of them be redundant or highly correlated with one another will arise. Consequently, there is an increase in the work done to carry out the evaluation, which does not necessarily improve the precision of the description. As a matter of fact, data analysis and interpretation will prove to be much more complex.

It is known that every character must contribute somehow for the description of germoplasm variability and, in turn, that no single descriptor can account for all existing variation. Hence, it becomes highly desirable to eliminate those descriptors found to be redundant, with low variability, with low stability of expression or difficult to measure, as long as it does not cause a significant loss of the general information.

Sneath and Sokal (1973) presented a wide discussion on the problems associated with selection of characters for taxonomic use, elimination of variables, redundancy of information and character correlation. Jolliffe (1972, 1973) compared eight methods for elimination of variables, which were based on multiple correlation, cluster analysis and principal components. According to the author, all methods were precise in the rejection of characters that were redundant or poor for describing the variation. As well, the elimination of more than half of the number of variables did not modify significantly the analyses. Similar results were obtained through other elimination procedures by Beale *et al.* (1967), Hussaini *et al.* (1977) and Bedigian *et al.* (1986).

The purpose of this study was to select the botanical and agronomical descriptors mostly important for the characterization of cassava germoplasm, taking into account the utilization of such descriptors in multivariate analyses.

MATERIALS AND METHODS

The study involved 280 cassava accessions from the EMBRAPA's Cassava Germplasm Active Bank - BAGM, which were characterized regarding 28 botanical and agronomical descriptors. Two evaluations were carried out, respectively in 1977 and 1978.

Plantings were performed at the EMBRAPA/CNPMPF, Cruz das Almas - BA, the plants being harvested at the age of one year. Data were collected from a single plot in each year. Plots consisted of four utile plants out of a total of 24, which were arranged in three rows measuring eight meters each.

The following botanical and agronomical descriptors were evaluated: easiness of harvest (EH), distance between foliar scars (DBFS), color of stem (CS), habit of ramification (HR), height of first apical branch (HFAB), height of plant (HP), color of unexpanded apical leaves (CUAL), number of foliar lobe (NFL), shape of central lobe

(SCL), length of central lobe (LCL), width of central lobe (WCL), color of petiole (CP), number of days from planting to flowering (NDF), root surface texture (RST), color of root periderm (CRP), easiness of root periderm removal (ERPR), color of outer surface of root cortex (CSRC), color of pulp (CP), length of root (LR), presence of root pedicel (PRP), presence of root constrictions (PRC), yield of fresh root (YFR), percentage of starch (PS), harvest index (HI), resistance to anthracnose (AR), resistance to rust (RR), resistance to oidium (OR) and resistance to mites (MR). Descriptors were evaluated as suggested by Silva (1981), Silva and Mendes (1982) and by Bellotti and Schoonhoven (1978).

To verify the feasibility of discarding some of the 28 descriptors, the analysis of principal components was utilized because it allowed for identifying the most and least representative characters in the explanation of total variance. Such procedure is a type of multivariable analysis, whose statistical principles have been explained in detail by Rao (1952), Adams and Wiersma (1978), Mardia *et al.* (1979), Kendall (1980) and Morrison (1981).

Basically, the method consists of transforming a group of p variables x_1, x_2, \dots, x_p , inherent to n individuals (genotypes, cultivars, populations, etc...) into a new group of variables namely, y_1, y_2, \dots, y_p , each y variable being a linear function of the variables x 's but independent from one another. Among all possible y combinations, y_1 will exhibit the greatest variance, y_2 the second one, and so on (Morrison, 1981).

Discarding variables by this method is based on the principle that the principal components' degree of importance or of variance (eigenvalue) decreases gradually from the first to the last component. Thus, the last components concentrate a very small or insignificant fraction of the total variance. In this way, the variable with the highest coefficient or with the greatest importance for the definition of a principal component of reduced eigenvalue (i.e. the last principal components) can be considered as of secondary value in the studied group. This low relevance makes it possible that those variables be discarded without significant losses in the explanation of the total variance.

As recommended by Jolliffe (1972, 1973) and Mardia *et al.* (1979), variables to be discarded were those with the highest coefficient in each principal component with eigenvalue smaller than 0.70. However, it was assumed a new discarding based on the next variable with the greatest magnitude was unnecessary, in the cases where the most important variable associated with a given component (eigenvalue < 0.70) had been previously discarded.

The coefficients of correlation between selected and discarded descriptors were estimated as a measure of the efficiency of the analysis of principal components, to disprize redundant descriptors.

To estimate the correlations, data from the 28 descriptors were subjected to an analysis of variance, plotting as sources of variation the effects of treatments (accessions),

years and treatment x years interaction. Through this procedure, it was possible to obtain estimations of the genetic variances and covariances among descriptors, removing from the mean square and the mean product of treatments the variations attributed to effects of environment and treatment x environment interactions, which were quantified by the mean square and mean product of the interaction.

RESULTS AND DISCUSSION

Table I presents the means and the standard deviations of the 28 botanical and agronomical descriptors studied in the two evaluations. It was noticed that the environmental alteration (years) provoked sensible and differentiated effects on the means of various descriptors, suggesting the existence of differences on the stability of manifestation among the characters.

Table I - Means and standard deviations of 28 botanical and agronomical descriptors, as applied to 280 cassava accessions, evaluated at the EMBRAPA's germplasm bank in 1977 and 1978.

No.	Descriptors Designation	1977		1978		General	
		\bar{x}	s_x	\bar{x}	s_x	\bar{x}	s_x
1	EH	1.343	0.475	1.478	0.500	1.411	0.343
2	DBFS	5.502	3.250	10.081	2.001	7.978	1.779
3	CS	1.614	0.487	1.586	0.493	1.600	0.468
4	HR	2.218	0.579	2.271	0.520	2.244	0.378
5	HFAB	1.013	0.525	0.821	0.316	0.917	0.390
6	HP	2.304	0.386	2.577	0.360	2.441	0.343
7	CUAL	1.893	0.764	1.825	0.651	1.859	0.634
8	NFL	6.307	1.093	6.207	1.250	6.257	1.174
9	SCL	1.053	0.225	1.053	0.225	1.053	0.217
10	LCL	12.814	2.335	15.985	2.096	14.400	1.893
11	WCL	3.862	0.849	4.634	0.859	4.248	0.762
12	CP	2.075	0.997	2.332	0.785	2.203	0.842
13	NDF	1.653	0.476	1.418	0.494	1.536	0.426
14	RST	1.814	0.389	1.764	0.425	1.789	0.369
15	CRP	2.464	0.756	2.536	0.712	2.500	0.692
16	ERPR	1.278	0.449	1.828	0.377	1.553	0.288

Continued

Table I - Continued

No.	Descriptors	1977		1978		General	
		\bar{x}	s_x	\bar{x}	s_x	\bar{x}	s_x
17	CSRC	1.193	0.534	1.250	0.678	1.221	0.548
18	CP	1.039	0.194	1.032	0.176	1.036	0.170
19	LR	26.006	4.750	28.637	4.801	27.321	3.669
20	PRP	1.464	0.499	1.311	0.463	1.387	0.379
21	PRC	1.211	0.408	1.303	0.460	1.257	0.341
22	YFR	19.841	6.526	10.749	3.479	15.294	4.229
23	PS	32.638	2.212	32.319	2.034	32.478	1.871
24	HI	43.129	11.967	50.936	10.667	47.033	9.845
25	RA	1.046	0.242	1.661	0.759	1.353	0.406
26	RR	1.586	0.604	1.353	0.592	1.469	0.454
27	OR	1.675	0.591	1.236	0.473	1.455	0.398
28	MR	2.111	0.747	1.096	0.296	1.603	0.408

Table II shows the selected and discarded descriptors, among the 28 ones originally considered, in the two evaluations. Through the analysis of principal components, it was possible to discard those botanical and agronomical descriptors that were non-discriminant and/or redundant, which were in number of 10 and 12, for the evaluations of 1977 and 1978, respectively. Among them, two groups were identified, the first one consisting of eight descriptors (CS, HFAB, NFL, LCL, NDF, CRP, HI and MR), which were discarded in both evaluations, thus with no use for the estimation of genetic diversity in the BAGM. And, still, the second group, formed by six descriptors (CUAL, SCL, CSRC, LR, RR and OR) which were either selected (discriminant) or discarded (non-discriminant or redundant). Such results indicate these groups are unnecessary for conducting characterization of cassava germplasm, especially when studies are targeted at examining bank diversity through multivariate methods. This is understandable, since their utilization would not enrich the information provided by the selected descriptors.

In reality, their utilization would only increase the work of evaluation and data processing, leading to more complex analysis and interpretations, adding nothing to the precision of estimations.

Fourteen descriptors were found to be common to the two groups of selected variables (which included, respectively, 18 descriptors in 1977 and 16 descriptors in

1978). They were therefore regarded as discriminant and stable to the purpose of studying diversity in the BAGM.

Table II - Subgroups of discarded or selected descriptors, arranged from the total 28 by analysis of principal components.

Subgroups	Numbering of descriptors*	
	1977	1978
Discarded descriptors (redundant or non-discriminant)	3, 5, 8, 9, 10, 13, 15, 24, 27, 28	3, 5, 7, 8, 10, 13, 15, 17, 19, 24, 26, 28
Discarded descriptors (common to both evaluations)	3, 5, 8, 10, 13, 15, 24, 28	3, 5, 8, 10, 13, 15, 24, 28
Selected descriptors in each year	1, 2, 4, 6, 7, 11, 12, 14, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26	1, 2, 4, 6, 9, 11, 12, 14, 16, 18, 20, 21, 22, 23, 25, 27
Selected descriptors (not common to both evaluations)	7, 17, 19, 26	9, 27
Selected descriptors common to both evaluations (discriminant and stable)	1, 2, 4, 6, 11, 12, 14, 16, 18, 20, 21, 22, 23, 25	1, 2, 4, 6, 11, 12, 14 16, 18, 20, 21, 22, 23, 25

* - According to Table I.

By applying this method, it was possible to reduce the number of descriptors by 50%. Such a reduction, shall simplify cassava germplasm characterization, given the much smaller number of variables involved without significant loss of general information.

The estimations of phenotypic coefficients of correlation among selected and discarded descriptors, and those from each group per se, are exhibited on Tables III and IV, respectively. Correlations among selected and discarded descriptors (Table III) generated 58 pair combinations with statistically significant values. However, the combinations HP x HI, WCL x LCL, CS x RST, RST x CRP and YFR x HI were the only ones to show coefficients greater than $r = 0.5$, a level that is normally considered of high magnitude.

Table III - Estimations of the phenotypic coefficients of correlation* among selected (down) and discarded (across) descriptors for the 28 botanical and agronomical descriptors, in two evaluations.

Descriptor**	3	5	7	8	9	10	13	15	17	19	24	26	27	28
1	0.005	0.073	0.003	-0.074	-0.007	-0.011	0.034	0.041	-0.071	0.115	-0.071	0.126	0.049	-0.004
2	-0.057	0.056	-0.064	0.092	-0.023	0.052	-0.098	-0.139	0.011	0.017	0.017	0.019	0.081	-0.062
4	-0.047	-0.257	-0.027	-0.095	0.036	-0.113	0.056	-0.061	0.027	-0.128	-0.125	-0.081	0.055	-0.037
6	-0.043	0.367	-0.037	0.069	-0.116	0.139	-0.088	0.071	0.089	-0.190	-0.534	0.190	0.134	-0.062
11	-0.054	0.193	-0.282	0.357	-0.489	0.562	-0.444	-0.019	-0.188	0.062	-0.045	0.193	-0.011	-0.316
12	-0.031	-0.071	-0.123	0.103	-0.040	0.017	0.012	-0.009	-0.177	-0.089	0.002	0.091	0.016	-0.033
14	0.588	0.014	0.041	0.021	0.052	-0.017	-0.014	0.873	0.076	0.010	0.128	0.004	0.106	0.080
16	0.112	0.018	0.056	-0.066	0.025	-0.042	-0.008	0.197	0.027	-0.072	0.059	-0.022	-0.057	0.051
18	0.021	-0.137	0.101	-0.295	-0.050	-0.297	0.198	-0.029	0.419	-0.038	-0.150	-0.110	0.048	0.123
20	0.013	0.026	-0.081	-0.107	-0.003	-0.104	0.055	0.034	0.081	0.136	-0.075	-0.035	-0.004	-0.069
21	0.144	0.079	-0.142	0.188	-0.065	0.125	-0.002	-0.083	-0.042	0.130	0.056	-0.042	0.091	-0.102
22	0.060	0.080	-0.146	0.207	0.129	0.150	-0.113	0.104	-0.113	0.379	0.780	-0.049	0.030	-0.135
23	0.016	-0.007	-0.175	0.216	-0.063	0.181	-0.253	0.024	-0.308	0.193	-0.053	0.015	-0.094	-0.073
25	0.039	-0.207	0.149	-0.157	0.028	-0.190	-0.082	0.137	0.138	-0.036	-0.004	-0.009	-0.018	0.059

* Coefficients of correlation with absolute value greater than 0.12 are significant at the level of 1% (t Test).

** Numbering according to Table I.

Table IV - Estimations of the phenotypic coefficients of correlation* among descriptors inside each group, of either selected (lower diagonal arrangement) or discarded (upper diagonal arrangement) descriptors, as averaged from two evaluations.

Descriptor**	5	7	8	9	10	13	15	17	19	24	26	27	28
	-0.050	0.077	-0.077	0.044	-0.100	0.031	0.661	0.112	-0.015	0.056	-0.118	-0.029	0.053
2	-0.046	-0.134	0.272	0.004	0.391	-0.269	0.042	-0.044	-0.043	0.062	0.142	0.040	-0.158
4	-0.031	0.001	-0.232	0.120	-0.216	0.217	0.065	0.211	-0.056	-0.093	-0.174	-0.099	0.375
6	0.030	-0.002	-0.038	-0.096	0.653	-0.536	0.010	-0.244	0.181	0.092	0.194	0.024	-0.166
11	0.086	0.080	-0.034	0.042	-0.096	0.134	0.047	-0.025	-0.028	0.163	-0.138	-0.013	0.119
12	-0.095	0.119	-0.050	-0.038	0.053	-0.562	-0.028	-0.328	0.162	-0.008	0.237	0.003	-0.262
14	0.049	-0.107	-0.066	0.093	-0.045	0.023	-0.021	0.211	-0.123	0.041	-0.230	-0.017	0.138
16	0.030	-0.154	-0.071	-0.063	0.012	0.029	0.182	0.125	-0.036	0.103	-0.025	0.062	0.105
18	-0.051	0.033	-0.064	0.122	-0.220	-0.013	-0.021	-0.039	-0.189	-0.106	-0.070	0.029	0.157
20	0.136	0.031	-0.020	0.025	-0.096	-0.113	0.022	-0.035	0.060	0.280	0.025	-0.007	-0.134
21	0.074	0.061	0.045	-0.059	0.093	0.023	-0.158	0.078	-0.004	-0.018	-0.139	-0.021	-0.034
22	-0.076	0.039	-0.097	-0.292	0.040	0.029	0.126	-0.095	-0.137	0.047	0.225	-0.104	0.034
23	-0.004	0.066	-0.033	0.006	0.175	0.056	0.037	-0.005	-0.074	0.047	0.016	0.034	0.034
25	-0.017	0.059	0.088	-0.144	-0.065	0.040	-0.129	-0.009	-0.039	-0.044	-0.064	-0.108	-0.108
1													
2													
4													
6													
11													
12													
14													
16													
18													
20													
21													
22													
23													
25													

* Coefficients of correlation with absolute value greater than 0.12 are significant at the level of 1% (t Test).

** Numbering according to Table I.

From inside the selected and discarded group of descriptors (Table IV), there were found, respectively, 13 and 39 characters combinations whose coefficients of correlation were significant. In the group of discarded descriptors, four estimations of coefficients of correlation were greater than $r = 0.5$, whereas all estimations from the group of selected descriptors were smaller than that value.

In the case of selected descriptors, it is important to note that, besides only few estimations of correlation being significant, their magnitude can be considered very low (Table IV). These results indicate that each selected descriptor must be responsible for one and only type of biological information and that, in the interaction of those descriptors, they would complement one another to provide a general description of the germplasm.

On the other hand, the large number of combination pairs among selected and discarded descriptors showing significant correlation demonstrates that the analysis of principal components is an efficient statistical tool to discard redundant variables, i.e. those which explain the same phenomenon similarly.

It is worth mention that this method for discarding variables takes into account only statistical criteria, as a means to avoid redundancy of information. Hence, descriptors of interest to the breeder, such as HI and NDF, were eliminated because they ended up correlated to other selected descriptors. Therefore, when selecting the group of descriptors for germplasm characterization, one should take this aspect into consideration.

ACKNOWLEDGMENTS

The authors are very grateful to EMBRAPA/CNPMPF for providing the germplasm data, and to Dr. Leônidas P. Passos (EMBRAPA/CNPGL) for his English translation of the manuscript.

Research supported by EMCAPA and CNPq.

Publication supported by FAPESP.

RESUMO

O estudo visou a seleção de descritores botânico-agronômicos de mandioca para avaliação da diversidade genética presente no germoplasma, através de métodos multivariados. Foram considerados dados referentes a 28 descritores e 280 acessos do banco de germoplasma de mandioca da EMBRAPA, em duas avaliações. A análise de componentes principais foi empregada para descartar os descritores considerados redundantes, não-discriminantes e com baixa estabilidade de expressão. O método de descarte de variáveis permitiu desprezar 50% dos descritores considerados, resultando na seleção de 14 variáveis discriminantes e estáveis. Esta redução do número de descritores deverá facilitar o trabalho de caracterização do germoplasma, sem perda significativa da informação geral.

REFERENCES

- Adams, M.W. and Wiersma, J.V. (1978). An adaptation of principal component analysis to an assessment of genetic distance. *Research Report* (Michigan State University) 347: 2-7.
- Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). The discarding of the variables in multivariate analysis. *Biometrika* 54: 357-365.
- Bedigian, D., Smith, C.A. and Harlan, J.R. (1986). Patterns of morphological variation in *Sesamum indicum*. *Econ. Bot.* 40: 353-365.
- Bellotti, A. and Schoonhoven, A. (1978). *Plagas de la yuca y su control*. CIAT, Cali. pp. 73.
- Conceição, A.J. (1979). *A mandioca*. UFBA/EMBRAPA/BRASCAN NORDESTE, Cruz das Almas. pp. 382.
- Goedert, C.O., Silva, S.O. and Mendes, R.A. (1982). Manual de caracterização e avaliação de mandioca. EMBRAPA/CENARGEN, Brasília. pp. 46.
- Gulick, P., Hershey, C.H. and Alcazar, J.E. (1983). *Genetic resources of cassava and wild relatives*. IBPGR, Rome. pp. 56.
- Hussaini, S.H., Goodman, M.M. and Timothy, D.H. (1977). Multivariate analysis and the geographical distribution of the world collection of finger millet. *Crop Science* 17: 257-263.
- Jolliffe, I.T. (1972). Discarding variables in a principal component analysis. I. Artificial data. *Applied Statistics* 21: 160-173.
- Jolliffe, I.T. (1973). Discarding variables in a principal component analysis. II. Real data. *Applied Statistics* 22: 21-31.
- Kendall, M. (1980). *Multivariate analysis*. Charles Griffin, High Wycombe. pp. 209.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate analysis*. Academic Press, London. pp. 521.
- Morrison, D.F. (1981). *Multivariate statistical methods*. 2 ed. McGraw Hill, Tokyo. pp. 415.
- Rao, A.V. (1952). *Advanced statistical methods in biometrics research*. John Wiley and Sons, New York. pp. 390.
- Silva, S.O. (1981). *Instalação e caracterização botânico-agronômica de coleções de mandioca*. EMBRAPA/CNPMPF, Cruz das Almas. pp. 51.
- Silva, S.O. (1982). Germoplasma de mandioca no Brasil. Congresso Brasileiro de Mandioca, 2, Vitória, 1981. *Anais*. EMBRAPA/CNPMPF, Cruz das Almas. 1: 85-91.
- Silva, S.O. and Mendes, R.A. (1982). *Banco de germoplasma e caracterização de cultivares*. EMBRAPA/CNPMPF, Cruz das Almas. pp. 30.
- Sneath, P.H. and Sokal, R.R. (1973). *Numerical taxonomy; the principles and practice of numerical taxonomy*. W.H. Freeman, San Francisco. pp. 573.

(Received December 7, 1990)