Chemometric Classification of Two Peach Palm (*Bactris gasipaes* H.B.K.) Landraces (Juruá and Vaupes)

J.M. Rojas[a,c], **H. Serruya**[b,c,*] and **M.H.S. Bentes**[b,c]

[a]Departamento de Química de la Universidad Nacional de la Amazonía Peruana-UNAP, Iquitos, Perú, [b]Departamento de Química da Universidade Federal do Pará-UFPA, Belém, PA, Brazil and [c]Programa de Pesquisa e Pós-Graduação em Química de Produtos Naturais, UFPA, Belém, Pará, Brazil

A multivariate analysis by Principal Components (PCA), Hierarchical Cluster Analysis, K-Nearest Neighbor (KNN) and Soft Independent Modeling of Class Analogy (SIMCA) methods were used to classify plants to different peach palm races. These statistical operations were applied to a data set of nineteen peach palm plant samples. Each data set contained fifteen variables defined as chemical characteristics of the mesocarp flour and physicochemical characteristics of the oil. The plants belonged to two different races. PCA showed that two principal components separated these races into two classes. KNN and SIMCA confirmed this classification. The final data for the model contained sixteen samples (plants) and eight variables. These results showed the utility of using chemometric methods for the classification of botanical species. These methods should aid the identification of new sources of oleaginous plants.

KEY WORDS: *Bactris gasipaes*, classification, cluster analysis, fatty acids, gas chromatography, multivariate analysis, pattern recognition, peach palm, principal component, vegetable oil.

Peach palm (*Bactris gasipaes* H.B.K.) is becoming an important product for the American economy because of its traditional food application (1) and its potential utilization in the oleochemicals industry. Most analytical data, particularly morphological properties (1–6) of peach palm, have been used to characterize and obtain descriptors for different populations. Such information is important for genetic, agronomic and economic improvement of the crop. Various statistical methods have been used to define these parameters. Univariate analysis with morphological and chemical constituents of fruit, mesocarp and oil was reported by Garcia (7). Discriminant analysis with morphological variables has been used to separate peach palm populations, and step-wise multiple regression has been used by Clement (3) to select descriptors. Multivariate analysis of lipids such as lecithins also has been used to classify oils in terms of their geographical origin (8) and according to process performance (9). In this study, several pattern recognition methods were applied: Principal components analysis (PCA) (10), hierarchical cluster analysis (HCA) (10,11), K-nearest neighbor (KNN) and soft, independent modeling of class analogy (SIMCA) (12,13).

The PCA method is based on variable reduction by linear combinations of the initial variables that define principal components (PC). This treatment of the data allows a researcher to observe the objects of a multidimensional space by projecting them onto the PC space. In other words, it is possible to reduce a p-dimensional space (more than 3) to two or three dimensions, where the objects can be observed without losing essential initial information (11,14). Variable coefficients for a principal component are usually called "loadings", and their values are the correlations of the variables with the principal component. The "scores" are the object projections on the components and represent correlations between objects and components (10).

KNN is based on sample similarity in the p-dimensional space, and it is measured by the distance between points representing samples, the more similar being closer. Differentiation of a sample between two classes is made arbitrarily, according to groups defined by the nearest neighbors.

SIMCA is based on PCA, where separate models are constructed for each class within a set in which the classes are known. Unknown samples may then be classified into the categories whose spatial models are closest.

The problem in this study was to classify peach palms according to the chemical and physicochemical characteristics of the mesocarp and oil. Then multivariate pattern recognition analysis was used to differentiate plants between peach palm races. The objectives in this study were to characterize the peach palm races; determine whether chemical, physicochemical and fatty acid variables could discriminate the two races; define a relevant set of variables and samples (plants) for this separation; and establish models for each race.

MATERIALS AND METHODS

Peach palm plant and fruit samples of two landraces: *Microcarpa Juruá* and *Macrocarpa Vaupes* were obtained from INPA Germplasm Active Bank (Manaus, Brazil) in 1990 and 1991. Each was identified by sample and specimen introduction numbers (Table 1). The mesocarp of the fruits was converted to a flour, and the oil was extracted after the fruits were characterized. The following physicochemical and chemical parameters were determined: Moisture (15), ash (15), oil content by using hexane and the Schwartz and Maxwell (16) and Bentes *et al.* (17) methods, protein (18), carbohydrates (by difference), refraction index (18), and acid, saponification and iodine values (15,18).

Gas chromatography (GC). Fatty acid methyl esters were prepared by mixing oil (50–100 mg) and sodium methoxide/methanol (3 mL). These solutions were heated for 3 min in a water bath. After cooling, BF_3/CH_3OH (3 mL) was added and heated for 3 min to ensure complete transesterification. After cooling, the mixture was diluted with 3 mL hexane, stirred and washed with 50 mL water. The upper phase of the mixture was separated and used for GC analysis (19). GC analysis (Shimadzu CG 14A, Tokyo, Japan) was performed with a 30-m capillary column wide-bore 530 (CG Analytica, São Paulo, Brazil), coated with Carbowax 20 M, and helium was used as the carrier gas. Samples were injected with the following oven temperature program: 100–170°C at 10°C/min, 170–180°C at 1°C/min and 5 min at 180°C. Injector

*To whom correspondence should be addressed at Departamento de Química, Universidade Federal do Pará-UFPA, Campus Universitário-Guamá, Belém, PA, Brazil.

TABLE 1

Sample Collected from INPA Germplasm Active Bank (Manaus, Brazil)

Sample[a]	Introduction	Race[b]	Sample	Introduction	Race
1	91F0189/83 P5	J	10	90F0164/84 P4a	V
2	91F0189/83 P7	J	11	91F0164/84 P4	V
3	91F0199/83 P1	J	12	90F0164/84 P9a	V
4	91F0199/83 P2	J	13	91F0211/84 P2	V
5	91F0199/83 P4	J	14	91F0211/84 P3	V
6	91F0199/83 P7	J	15	90F0215/83 P1a	V
7	91F0201/83 P2	J	16	91F0215/84 P1	V
8	91F0201/83 P8	J	17	90F0215/84 P3a	V
9	91F0201/83 P9	J	18	91F0215/84 P3	V
			19	91F0215/84 P8	V

[a]From 1 to 9 *Microcarpa* (Juruá) and from 10 to 19 *Macrocarpa* (Vaupes) landraces.
[b]J, Jurua and V, Vaupes.

temperature was 200°C, and the detector flame-ionization detector temperature was 220°C. Each sample was analyzed twice.

These data were statistically analyzed by PCA and HCA as exploration methods and KNN and SIMCA (13) as classification methods, from the computational programs ARTHUR of UNICAMP (Campinas, Brazil) and PIROUETTE of Infometrix Inc. (Seattle WA). SIMCA and KNN were used to obtain class separation and to derive classification rules and models for each class.

RESULTS AND DISCUSSION

To discriminate between peach palm races, chemical and physicochemical variables of mesocarp flour and oil (Tables 2 and 3) and fatty acid oil composition (Table 4) were subjected to multivariate analysis. Groupings were observed in these data after projecting many dimensions (in this case, fifteen) down onto a two-dimensional plane. These groupings could be interpreted in terms of known behavior and then used to form the basis for classification.

An exploratory PCA method for nineteen objects (plants) and fifteen variables, plotted in two dimensions, is illustrated in Figures 1 and 2. These plots show the projection of each peach palm tree sample onto a two-dimensional hyperplane in the fifteen-dimensional measurement space that is spanned by the analytical variables (Tables 2–4). Figure 1 shows that sample 10 was grouped with the Juruá class, in spite of being morphologically classified to the Vaupes class, because of its 18:1 content. The first two principal components (PC_1, PC_2) with 69.2% of variance showed (Table 5) the class separation without

TABLE 2

Moisture and Chemical Composition of Two Peach Palm Races, Dry Mesocarp (n = 2)[a]

Sample	Moisture (%)	Ash (%)	Oil (%)	Protein (%)	Carbohydrate (%)
1	62.01	1.64	22.89	7.96	67.50
2	49.24	0.84	21.93	4.90	72.32
3	56.47	1.53	22.58	7.58	68.31
4	58.32	2.32	19.54	9.36	68.78
5	64.86	2.04	19.80	7.49	70.67
6	62.91	2.13	26.06	9.00	62.81
7	49.38	1.22	49.57	6.02	43.20
8	30.48	0.42	52.69	4.60	42.29
9	56.20	1.51	19.39	6.79	72.31
10	70.05	3.30	6.79	7.50	82.40
11	65.40	1.79	4.36	6.27	87.58
12	62.88	2.16	9.29	8.45	80.11
13	67.13	2.47	6.86	10.35	80.33
14	51.19	1.34	7.46	4.45	86.75
15	50.59	1.63	7.45	3.71	87.21
16	57.39	1.17	6.23	4.44	88.16
17	64.88	1.89	10.49	6.07	81.55
18	60.65	1.01	4.00	5.62	89.36
19	58.47	1.27	4.86	7.01	86.86

[a]Moisture (%) = moisture percent of wet mesocarp; Ash (%) = ash percent of dry mesocarp; Oil (%) = oil percent of dry mesocarp; Protein (%) = protein percent of dry mesocarp; Carbohydrate = carbohydrate percent of dry mesocarp.

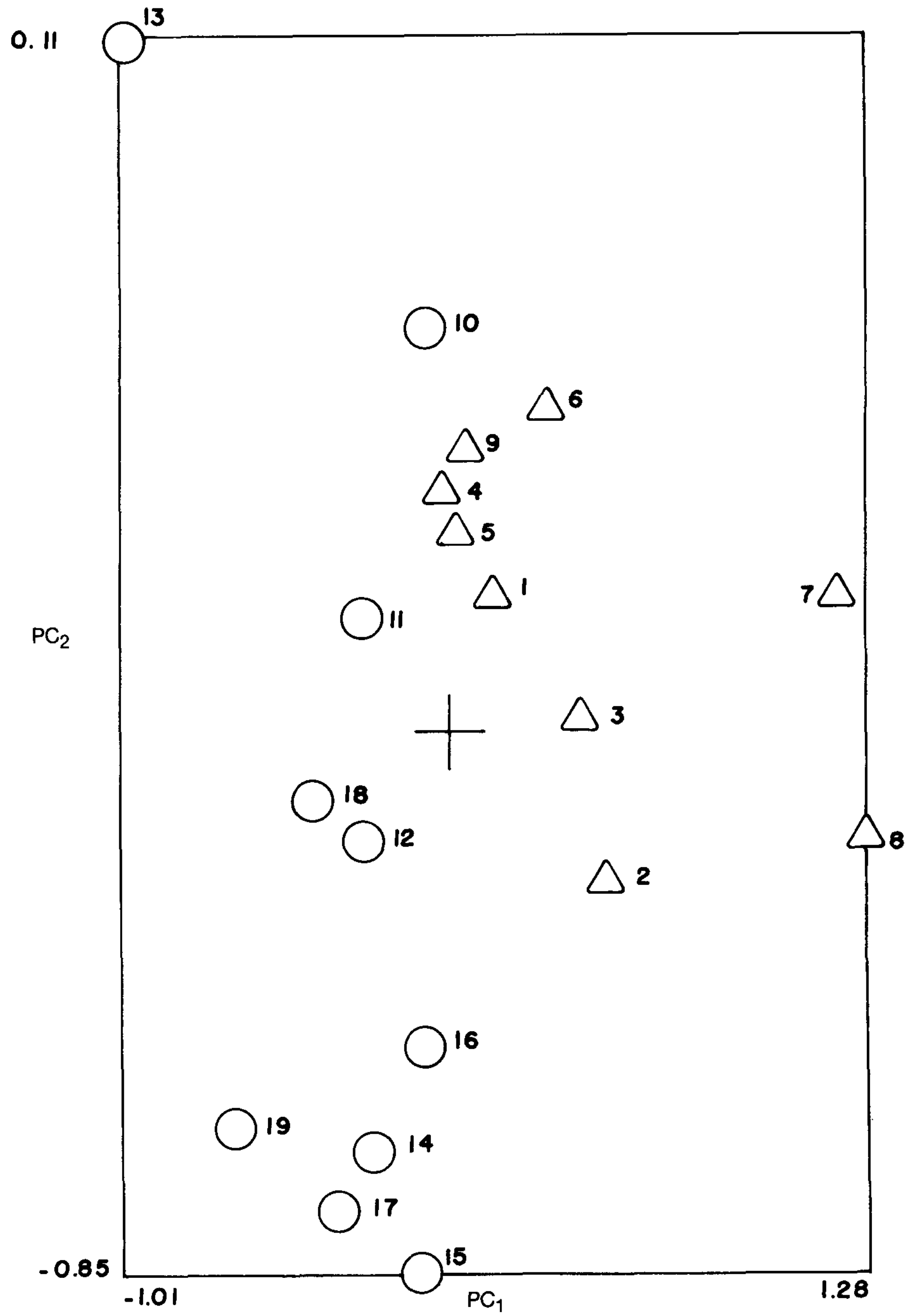

FIG. 1. Objects originally in a fifteen-dimensional space are projected on the principal component plane PC_1–PC_2, 19 × 15 matrix. The cross is (0,0); △ = Juruá; ○ = Vaupes.

considering object 10. In Figure 2, PC_1 and PC_3 (48.3% of variance) gave a better class separation, which revealed a structure similar to that based on morphological fruit properties (3).

The HCA (Fig. 3) application with 100% of variance showed that the original groupings were formed at a similarity level of 0.57, according to the hopped classes. However, object 10 was located in the wrong class for the same reason explained above; objects 7, 8 (with the highest oil content, 49.57 and 52.69% respectively, Table 2) and 13 (the highest content of 16:0, 49.93%, Table 4) were potential outliers.

Variables and samples considered irrelevant according to PCA application were removed. Only variables with high absolute value of loadings of the first PC (Fig. 5) and good modeling power criterion (10,20,21) (Table 6) were considered. The reduced matrix (16 × 8) without samples 7, 8 and 13, with eight relevant variables (Table 5,

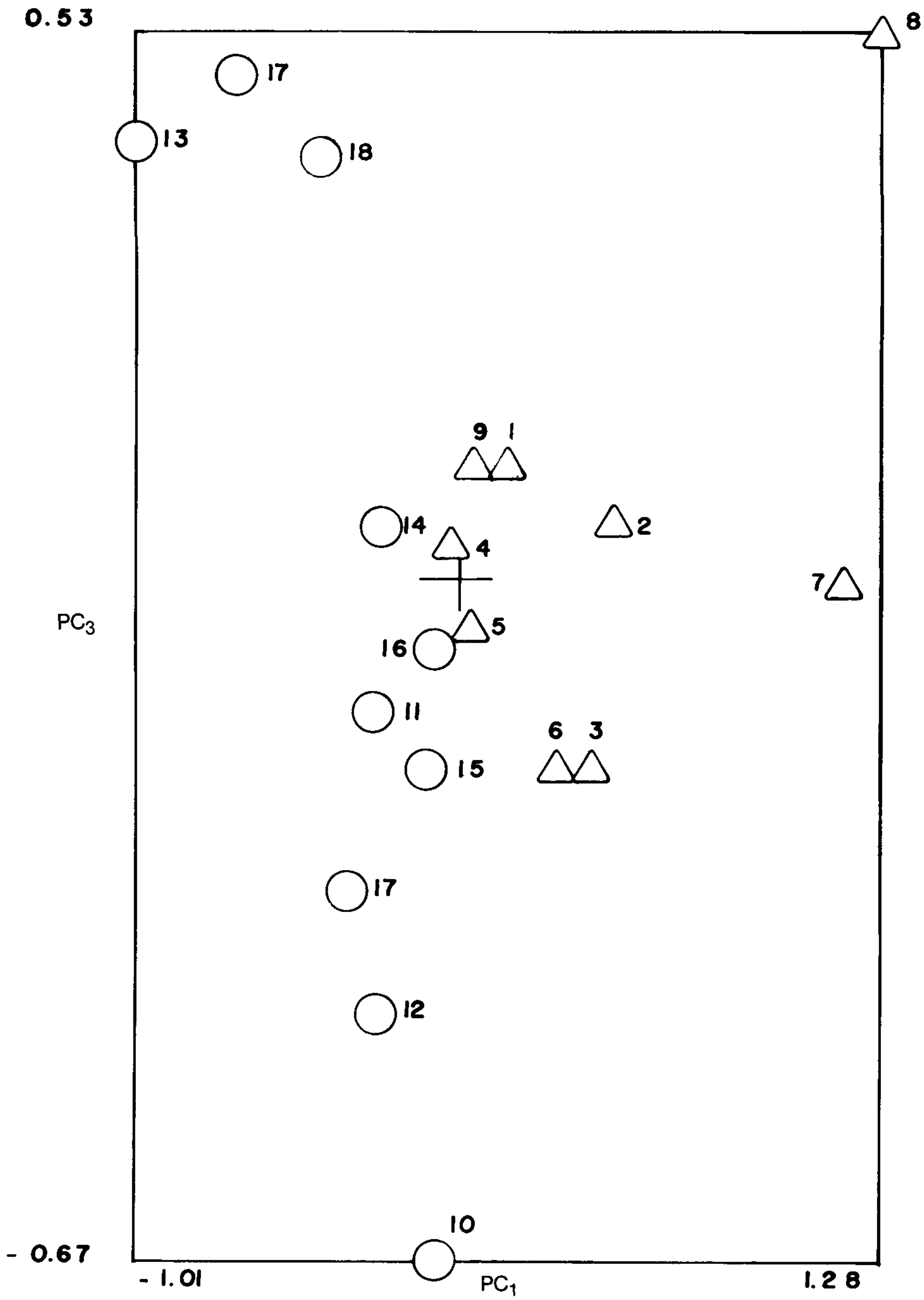

FIG. 2. Objects originally in a fifteen-dimensional space are projected on the principal component plane PC_1–PC_3, 19 × 15 matrix. The cross is (0,0); △ = Juruá; ○ = Vaupes.

underlined), was subjected to recalculation with PCA giving and gave the same original structure. A SIMCA (Table 7) application to the 16 × 8 matrix led to a separated class model for *Microcarpa Juruá* with zero PCs and *Macrocarpa Vaupes* with two PCs, calculated according to the cross-validation criterion (22). By means of the modeling power for each variable, its influence on the classification can be seen (Table 6).

Each sample was fitted to the models according to the distance between sample and each model. This gave more probability of correct classification to those that were closer to the calculated category, and it resulted in a perfect fit of all objects into appropriate classes (Table 6), including sample 10, due to the restrictions imposed by the SIMCA model in relation to the distance. A KNN application also showed a good separation of the classes with an accuracy of 100% for 1 to 8 nearest neighbors (no missed samples) and a minimum

TABLE 3

Chemical and Physicochemical Analysis of Mesocarp Oils from Two Peach Palm Races[a]

Sample	RI (40°C)[a]	SV	AV	IV
1	1.4481	196.70	88.14	63.24
2	1.4467	195.18	104.46	71.88
3	1.4511	195.12	105.22	71.39
4	1.4495	196.76	103.44	62.13
5	1.4509	196.63	112.14	61.45
6	1.4517	195.63	92.34	61.78
7	1.4560	194.42	29.33	68.93
8	1.4510	195.34	41.94	70.59
9	1.4505	196.06	105.79	57.00
10	1.4590	195.23	100.64	60.53
11	1.4600	197.78	91.19	64.96
12	1.4584	195.69	107.02	78.70
13	1.4655	200.20	120.96	39.14
14	1.4560	196.90	114.83	79.23
15	1.4520	195.99	121.02	84.66
16	1.4542	106.24	106.72	70.00
17	1.4561	196.35	107.44	88.20
18	1.4562	198.36	98.29	63.79
19	1.4423	198.96	123.82	74.83

[a]RI (40°C) = refraction index; IV = iodine value; AV = acid value; SV = saponification value.

TABLE 4

Oil Fatty Acid Composition of Two Peach Palm Races

Sample	16:0	16:1	18:0	18:1	18:2	18:3
1	33.05	4.18	2.59	52.29	7.17	0.7
2	24.93	3.69	2.74	58.49	9.60	0.5
3	23.74	4.57	1.31	62.90	7.47	—
4	33.62	4.05	3.04	50.93	8.36	—
5	31.60	5.45	2.51	55.51	4.93	—
6	29.49	2.30	3.95	59.28	4.98	—
7	21.46	3.40	3.13	67.68	4.33	—
8	23.66	5.76	1.71	62.10	6.76	—
9	30.94	3.48	10.97	46.88	7.72	—
10	26.11	3.69	6.16	61.81	2.23	—
11	28.23	5.33	4.80	52.37	7.98	0.3
12	25.43	5.32	4.02	47.70	14.96	2.5
13	49.93	7.12	7.90	32.49	2.56	—
14	27.39	9.60	1.07	44.64	15.32	1.9
15	21.17	10.67	0.69	51.89	12.30	3.2
16	24.37	9.47	1.22	52.93	10.61	1.4
17	22.81	10.61	1.57	43.39	17.82	3.8
18	34.24	11.54	2.41	42.33	9.49	—
19	34.86	13.03	0.80	32.72	16.21	2.3

of 93.8% for 9 to 10 nearest neighbors (one missed sample).

Results of this study showed that chemical characteristics of peach palm can be used to form a data set relevant to the classification of landraces. The multivariate analysis was able to separate peach palm samples into two races. One race (Vaupes) revealed a more complicated structure than Juruá. A final set of eight variables was selected for these statistical analyses to discriminate races. They included: moisture %, oil content %, carbohydrates %, saponification value, acid value, palmitic acid %, palmitoleic acid % and oleic acid %.

TABLE 5

Eigenvectors, Variance (%) and "Loadings" in Principal Component Analysis 19 × 15 Matrix

Component/loading	PC_1[a]	PC_2	PC_3	PC_4	PC_5
Moisture (%)	−.30045	.17837	−.33816	.10875	−.02846
Ash (%)	−.19249	.24450	−.46354	.09535	−.20917
Oil (%)	.39010	.08028	.16880	.06866	−.30172
Protein (%)	−.13348	.34085	−.05704	.39235	−.44672
Carbohydrate (%)	−.36846	−.13789	−.14159	−.12588	.37452
RI (40°C)	−.12645	.15931	−.20213	−.78509	−.29979
SV	−.33968	.06152	.40959	−.08388	−.07982
AV	−.35869	−.06150	−.13687	.29414	.31882
IV	.06461	−.42083	−.22689	.03710	−.15370
16:0	−.25979	.26065	.38943	.09599	−.09056
16:1	−.22972	−.31099	.24048	−.18587	−.05915
18:0	−.07389	.33319	−.04175	−.17974	.12193
18:1	.36588	.06079	−.31053	−.01172	.22019
18:2	−.13852	−.38744	−.01309	.12893	−.25700
18:3	−.15470	−.35673	−.17977	.02145	−.40246
Eigenvector	5.492E + 00	4.884E + 00	1.756E + 00	1.001E + 00	7.078E − 01
% Variance (%)	36.6	32.6	11.7	6.7	4.7
Cumulated V	36.6	69.2	80.9	87.6	92.3

[a]PC_1–PC_5 are linear combinations of all the variables. V = variance. See Table 3 for other abbreviations.

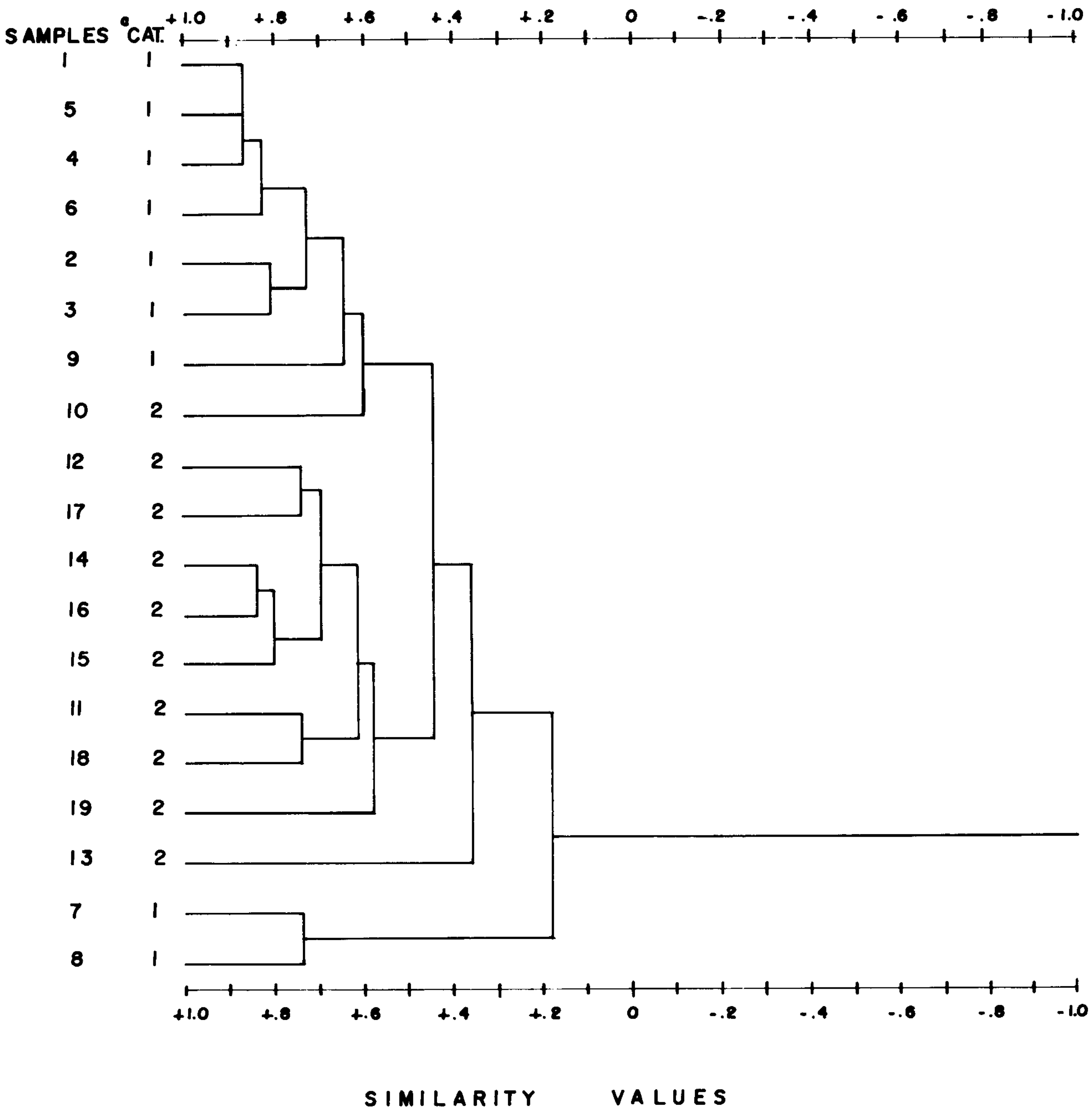

FIG. 3. Cluster analysis dendogram for 19 × 15 matrix. Cat. = Category.

TABLE 6

Modeling Power of the Variables Taken Over the Principal Components of the Model 16 × 8 Matrix[a]

	Modeling power	
Variable	Juruá Class	Vaupes Class
Moisture (%)	0.62544	0.65808
Oil (%)	0.78962	0.72246
Carbohydrate (%)	0.66109	0.82723
SV	0.76984	0.74592
AV	0.40849	0.48147
16:0	0.84733	0.62279
16:1	0.54710	0.53363
18:1	0.77683	0.78590

[a]See Table 3 for abbreviations.

TABLE 7

SIMCA Application Results for the 16 × 8 Matrix[a]

	Category				
Sample	True	Calc[b]	Distance	Category	Distance
1	1	1	*** 1.676E-01	2	* 3.066E-01
2	1	1	*** 1.964E-01	2	* 2.817E-01
3	1	1	*** 1.751E-01	2	** 2.531E-01
4	1	1	**** 1.181E-01	2	* 2.889E-01
5	1	1	*** 1.598E-01	2	** 2.636E-01
6	1	1	*** 1.526E-01	2	* 3.345E-01
9	1	1	**** 1.212E-01	2	** 2.690E-01
10	2	2	**** 1.070E-01	1	* 3.079E-01
11	2	2	**** 1.122E-01	1	* 3.454E-01
12	2	2	**** 1.159E-01	1	** 2.381E-01
14	2	2	**** 9.639E-02	1	* 3.613E-01
15	2	2	**** 8.508E-02	1	4.132E-01
16	2	2	**** 9.319E-02	1	* 3.241E-01
17	2	2	*** 1.584E-01	1	* 3.385E-01
18	2	2	**** 9.958E-02	1	4.391E-01
19	2	2	**** 1.140E-01	1	5.518E-01

[a]More asterisks designate increased probability of correct classification.
[b]Category calculated by the program.

ACKNOWLEDGMENTS

The authors thank Eng. S. Ferreira for supplying peach palm samples, Professor Dr. Roy E. Bruns for critical reading and granting the computational programs ARTHUR, Dr. Lamar Scott Ramos for assisting us in using PIROUETTE and for valuable advice; CAPES, FINEP and SUDAM for supporting this research.

REFERENCES

1. Clement C.R., Descriptores mínimos para el Pejibaye (*Bactris gasipaes* H.B.K.) y sus implicancias filogenéticas, M.S. Thesis, Costa Rica, 1986.
2. Morera, M.J.A., Descripción Sistemática de la Colección Panamá de Pejibaye (*Bactris gasipaes* H.B.K.) del CATIE, M.S. Thesis, Turrialba, Costa Rica, 1981.
3. Clement, C.R., *Final Report, Peach Palm (*Bactris gasipaes *H.B.K.)* germplasm bank, edited by C.R. Clement and L. Coradin, United States Agency for International Development project report, pp. 34–81, 1985.
4. Mora, A, Descriptores de la semilla de tres poblaciones de pejibaye (*Bactris gasipaes* H.B.K.) y sus implicaciones filogenéticas. Thesis, University of Costa Rica, San Jose, 1986.
5. Valle, L., Descriptores dela inflorescencia del pejibaye (*Bactris gasipaes* H.B.K.) de cuatro poblaciones y sus implicaciones filogenéticas, Thesis, University of Costa Rica, San Jose, 1986.
6. Valverde, M.E., Descriptores de la flor de pejibaye (*Bactris gasipaes* H.B.K.) en cuatro populaciones y sus posibles implicaciones filogenéticas, Thesis, University of Costa Rica, San Jose, 1986.
7. Garcia, T.D.E., Caracterização da fração lipidica do mesocarpo de tres raças de pupunha (*Bactris gasipaes* H.B.K.), mantidas no Banco Ativo de Germoplasma do INPA, M.S. Thesis, Instituto de Amazona and Fundacão Universidade da Amazona, Brazil, 1989.
8. Forina, M., C. Armanino, S. Lanteri and E. Tisciornia, in *Food Research and Data Analysis*, edited by H. Martens, and H. Russwurm, Jr., Applied Science Publishers, London, 1983.
9. Kaufmann, P., U. Olsson and B.G. Herslof, *J. Am. Oil Chem. Soc. 67*:537 (1990).
10. Massart, D.L., *Chemometrics*, Elsevier Publishers B.V., Netherlands, 1988.
11. Kowalski, B.R., *Chemometrics: Mathematics and Statistics in Chemistry Series C. Mathematical and Physical Sciences*, Vol. 138, Reidel Publishers, Holland, 1984.
12. Kowalski, B.R., *Chemometrics: Theory and Application*, American Chemical Society, Washington D.C., 1977.
13. Bruns, R.E., E. Faigle and J.F.G. Quimiometria, *Quimica Nova 4*:84 (1985).
14. Delaney, M.F., A.N. Papas and M.J. Walters, *J. Chromatogr. 410*:31, (1987).
15. *Official Methods and Recommended Practices of the American Oil Chemists' Society*, American Oil Chemists Society, 1989, Methods Ca 2B-38, Ca 11-55, Cd 3a-63, Cd 3-25, Cd 1-25.
16. Schwartz, N.O.A., and R.J. Maxwell, *J. Am. Oil Chem. Soc. 56*:634 (1979).
17. Bentes, M.H.S., H. Serruya, M.D.B. Silva and D.J. Oliveira, Lipidios em Vegetais-Novo esquema de extração, 6°Encontro de Profissionais de Química da Amazônia Manaus-Brasil, 1988.
18. Instituto Adolfo Lutz, *Normas Anal i ticas do Instituto Adolfo Lutz. Métodos quimicos e f isicos para análise de alimentos*, 2nd ed. Sao Paulo, Vol. 1, 1985, pp. 159, 189, 190.
19. Khan, G.R., and F. Scheinmann, *Prog. Chem. Fats and Other, Lipids 15*:343 (1978).
20. Wold, S., and M. Sjostrom, *SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy in: Kowalski*, 1977.
21. Bruns, R.E., and S.I. Scarminio, *Manual do Usuario do sistema Computacional Arthur para Microcomputadoras Campinas SP., s.d.*, 1980.
22. Wold, S., *Technometrics 20(4)*:397 (1978).

[Received March 2, 1993; accepted October 7, 1993]