Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v42i1.43195

GENETICS AND PLANT BREEDING

Trait selection using procrustes analysis for the study of genetic diversity in Conilon coffee

Daiana Salles Pontes[1*], **Renato Domiciano Silva Rosado**[1], **Cosme Damião Cruz**[1], **Moysés Nascimento**[1], **Ana Maria Cruz Oliveira**[2] **and Scott Michael Pensky**[3]

[1]Departamento de Estatística Aplicada e Biometria, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, s/n, Campus Universitário, Viçosa, Minas Gerais, Brazil. [2]Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil. [3]Louisiana State University, Baton Rouge, Louisiana, USA. *Author for correspondence. E-mail: salles.yana@gmail.com

ABSTRACT. Trait selection is occasionally necessary to save money and time, as well as accelerate breeding program processes. This study aimed to propose two criteria to select traits based on a Procrustes analysis that are poorly explored in genetic breeding: Criterion 1 (backward algorithm) and Criterion 2 (exhaustive algorithm). Then, these two criteria were further compared with Jolliffe's criterion, which has often been used to select traits in genetic diversity studies. Sixteen agronomic traits were considered, and 40 Conilon coffee (*Coffea canephora*) accessions were evaluated. This study showed that the flexibility in selecting traits by researcher preference, graphical visualization, and Procrustes M^2 statistic through criteria 1 and 2 is a fast and reliable alternative for decision-making. These decisions are based on the removal and addition of traits for phenotyping in studies of Conilon coffee diversity that can be applied to other crops. Other relevant aspects of selection traits criteria were also discussed.

Keywords: principal components; graphical comparison; selection criteria; discard of variables; plant breeding.

Received on June7, 2018.
Accepted on November 21, 2018.

Introduction

Studies on genetic diversity play an important role in breeding programs because they are crucial at the initial phase, called prebreeding, in which it is possible to regenerate, characterize, explore, and promote the conservation of variability available on the base population. Moreover, the information collected at the prebreeding phase is useful in obtaining potential candidates to generate divergent parents. These parents are more likely to promote satisfactory results regarding the genetic potential of derived cultivars or lineages as well as combining their abilities to obtain heterotic hybrids.

Multivariate techniques such as discriminant analysis, principal component analysis, coordinate analysis (Cruz, Ferreira, & Pessoni, 2011), and clustering are used in this kind of study. Several inferences about genetic diversity studies can be made for different purposes of a breeding program using principal component analysis (PCA) (Yousaf et al., 2018; Liu et al., 2017; Muleta, Bulli, Zhang, Chen, & Pumphrey, 2017; Yano, Nonaka, & Ezura, 2018).

The genetic diversity existing among and within populations can be measured by the difference between the phenotypic values of their accessions and is obtained in field experiments using a considerable number of morphological, agronomic, and other traits of the studied cultivars. If a collection of accessions evaluated in a given experiment comes from a population or a germplasm bank, it can be re-evaluated in future studies for a variety of purposes. In some situations with a high cost and degree of difficulty involved to obtain a particular trait(s), it may be valuable to evaluate a smaller number of traits than those recorded in the germplasm bank. However, variability is a factor of extreme importance in the development of new varieties and in the conservation of genetic resources, and it is the breeder's responsibility to investigate the extent of exclusion of one or multiple traits that will affect the present variability in the group of accessions under analysis.

The relative importance of traits in genetic diversity studies can be achieved using the criteria proposed by Singh (1981) and Jolliffe (1972). However, the use of each one is restricted to the initial choice of the researcher regarding the clustering method used in the study of genetic diversity since both methods have different approaches.

The first criterion is used when the diversity is evaluated based on the dissimilarity measuring (distance measured between accession pairs) information to provide the cluster analysis. The second criterion is based on principal components that will generate graphic dispersion information in two or three-dimensional space.

In addition to those traits discarding the abovementioned criteria, there is also another methodology based on Procrustes analysis. Although it is rarely used in genetic diversity studies (especially in traits selection), the Procrustes approach has been used in many different areas including food engineering (Oliveira & Benassi, 2010; Mauricio, Palazzo, Caselato, & Bolini, 2016) and health sciences (Douglas, 2004; Daboul, Ivanovska, Bülow, Biffar, & Cardini, 2018). Thus, its application has shown great promise and has been evaluated in several studies. However, the technique has been poorly explored in genetic breeding (Klingenberg, 2003; Bramardi, Bernet, Asíns, & Carbonell, 2005; García-Peña & Dias, 2009) and, therefore, was the main motivation for this paper.

The Procrustes analysis technique allows a comparison of two configurations or two datasets as long as each line corresponding to the same individual. If two vectors are different from each other but are defined in the same subspace, it is possible to estimate the extent of the differentiation of their respective graphical representations by means of the Procrustes M^2 statistic. Thus, the smaller the value of this statistic is, the more similar the two configurations will be.

Krzanowski (1987) presents a methodology that combines PCA, which is used to obtain the scores of the configurations, and Procrustes analysis to determine how much a subset of traits represents a structure of the original dataset (with all traits). The author discusses a Procrustes analysis from two perspectives: for a selection of traits from the backward elimination algorithm using the Procrustes statistic as a discard criterion and for a comparison of grouping patterns of different trait components resulting from different selection methods using the same statistics.

Based on the strategy proposed by Krzanowski (1987) and using the Procrustes M^2 statistic given by Peres-Neto and Jackson (2001), our objective is to propose two cut-off criteria called the backward algorithm (Criterion 1) and the exhaustive algorithm (Criterion 2) for the selection of traits in the genetic diversity study of Conilon coffee (*Coffea canephora*). To validate the methodology, we will compare both with Jolliffe's criterion (1972), since it is considered a more efficient trait discarding method used for genetic diversity studies, providing more savings in a breeding program.

Material and methods

Database

The databases provided by Ferrão et al. (2008) refer to the means of the characteristics. The experimental design utilized randomized blocks with six replications for PH and DCA and 4 replications for the other characteristics with each plot consisting of two plants. The model effect considered genotypes as fixed, and analyses of the variance of the characteristics were performed based on the average number of plots from the following model:

$$\mathrm{Y_{ij}} = \mu + \mathrm{G_i} + \beta_\mathrm{j} + \mathrm{e_{ij}}$$

where: $\mathrm{Y_{ij}}$ is the phenotypic valor of the ij-th observation referring to the *i*-th genotype in the *j*-th block; μ is the overall mean of the character; G_i is the effect of the i-th genotype (i = 1, 2, ... , 40); β_j is the effect of the j-th block (j = 1, 2, ..., 4 or 6); and $\mathrm{e_{ij}}$ is experimental error, $\mathrm{e_{ij}} \sim \mathrm{N}(0, \mathrm{s}^2)$.

Sixteen agronomic traits from 40 Conilon coffee accessions were evaluated in the Sooretama municipality located in the Brazilian state of Espírito Santo in the year 2000. An evaluation was conducted for the number of days (D) between flowering and total fruit maturation; the grain yield (GY kg ha^{-1}); the plant height (PH in cm); the diameter crown average (DCA in cm), taken at the "middle third" of the plant; the cherry and coconut dry coffee relationship (ChCo), taken in a 2 kg sample of cherry coffee and its dried weight; the cherry and green coffee relationship (ChBe), taken in a 2 kg sample of cherry coffee and its dried weight after processing; the coconut and green coffee relationship (CoBe), taken in a 2 kg sample of cherry coffee and its dried weight after processing; the coarse grain percentage (CG%); the "flat" grain percentage (FG%); the "mocha" grain percentage (MG%); the grain moisture percentage (GM%); the percentage of grains retained on the sieve mesh size #17 (S17); the percentage of grains retained on the sieve mesh size

#15 (S15); the percentage of grains retained on the sieve mesh size #13 (S13); the percentage of grains retained on the sieve mesh size #11 (S11); and the medium strainer (MS) (medium grain size). According to Ferrão et al. (2008), the coefficients of experimental variation in percentage (CVe) of the characteristics are D (0.05), GY (23.24), PH (5.29), DCA (6.75), ChCo (6.85), ChBe (5.38), CoBe (7.26), CG (65,93), FG (5.20), MG (32.90), GM (11.20) S17 (31.95), S15 (11.17), S13 (15.95), S11 (51.52), and SM (2.16), of which the majority is less than 30% and shows good experimental precision for the coffee crop (Bonomo et al., 2004; Ferrão et al., 2008; Rodrigues, Brinate, Martins, Colodetti, & Tomaz, 2017).

Procrustes analysis

To contextualize the criteria proposed in this work, it is important to first present pertinent information about Procrustes analysis. This technique allows the comparison of two datasets or two configurations, as long as each line corresponds to the same individual. If there are two sets of vectors that differ from one another but that define the same subspace, this technique allows the user to measure the difference between their respective graphical representations by means of the Procrustes M^2 statistic. When the comparison is performed for more than two datasets or configurations, it is defined as a generalized Procrustes analysis.

For the understanding of the technique, consider the triangles $Y: A - B - C$ and $\tilde{Z}: a - b - c$ as the representation of two configurations in a two-dimensional space (matrices of n = 3 individuals and p = 2 traits) with different size, location, and orientation (Figure 1a).

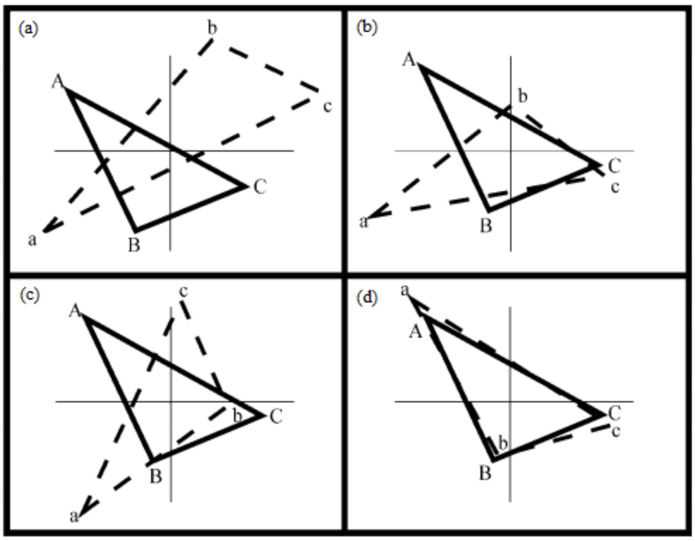

Figure 1. Representation of steps involved in a Procrustes analysis: (a) original configurations where the triangle ABC was used as reference configuration; (b) configurations after standardization (i.e., similar size and common center); (c) configurations after mirror reflection, if necessary; (d) configuration after rotation so that the sum of the squared differences between homologous observations (A/a, B/b, C/c) is a minimum (Peres-Neto & Jackson, 2001).

The difference between these configurations is obtained by means of a Procrustes analysis so that its corresponding points align as well as possible. Procrustes analysis is a procedure that minimizes the trace of the sum of squared differences between two configurations (i.e., two data matrices) in a multivariate Euclidean space (Equation [1]), which is obtained in two steps by adjusting the configuration $\tilde{Z}$ to a reference configuration Y.

$\mathrm{Min}\{\mathrm{trace}[(Y - \tilde{Z})(Y - \tilde{Z})']\}$ Equation [1]

First, the centering (translation) and scaling (dilation) are carried out in Y and $\tilde{Z}$ (Figure 1b), such that $Y = (I - P)Y/\sqrt{\mathrm{tr}[(I - P)\,Y'\,(I - P)]}$ and $\tilde{Z} = (I - P)\tilde{Z}/\sqrt{\mathrm{tr}[(I - P)\tilde{Z}'(I - P)]}$, where I is an identity matrix nxn and P is a matrix *nxn* with all elements equal to 1/n, followed by the reflection (Figure 1c), if necessary, and the rotation of $\tilde{Z}$ (Figure 1d) for its adjustment in Y. That is, $\tilde{Z}$ is rationed to $\tilde{Z}Q$ such that $Q = VU'$ is the rotation matrix, $U\Sigma V'$ is the decomposition of singular values of $\tilde{Z}'Y$ where Σ is a diagonal matrix, and U and V are orthogonal matrices. Finally, we have the statistic M^2 (Equation [2]) as a result of the comparison between Y and $\tilde{\mathrm{Z}}$, referred to as Procrustes statistics or residual sum of squares, ranging from zero to infinity.

$\mathrm{M}^2 = \mathrm{trace}\{\mathrm{YY}' + \tilde{\mathrm{Z}}\tilde{\mathrm{Z}}' - 2\tilde{\mathrm{Z}}\mathrm{Q}'\mathrm{Y}'\} = \mathrm{traço}\{\mathrm{YY}' + \tilde{\mathrm{Z}}\tilde{\mathrm{Z}}' - 2\Sigma\}$ Equation [2]

According to Peres-Neto and Jackson (2001), the variation of the M^2 statistic between 0 and 1 is restricted using the following transformation:

$\mathrm{M}^2 = 1 - (\mathrm{trace}\ \Sigma)^2$ Equation [3]

Procrustes analysis, Procrustes transformation or Procrustes rotation give us the idea that the configurations should be as close as possible (in the same subspace) to compare them. Thus, the configurations under the same referential can be fairly compared, and the "real difference" between them can be quantified.

When Procrustes analysis is performed on the same configuration (Y = Z), we have M^2 = 0, indicating a perfect fit. Thus, the smaller the value of the statistics M^2 are, the more similar the configurations (García-Peña & Dias, 2009).

Trait selection criteria

To reduce the number of agronomic traits, based on the Conilon coffee dataset, we initially selected a subset of traits using Procrustes analysis according to the methodology presented by Krzanowski (1987). The methodology presented by Krazanowski (1987) combines principal component analysis (PCA) and a Procrustes analysis (Figure 2) to determine how much a subset of traits represents the structure of the set of p original traits. Thus, after performing PCA on the matrices of the set of original traits $\mathrm{X}_{\mathrm{nxp}}$ and the subset of q traits $\tilde{\mathrm{X}}_{\mathrm{nxq}}$, the novel matrices obtain the configurations $\mathrm{Y}_{\mathrm{nxk}}$ and $\tilde{\mathrm{Z}}_{\mathrm{nxk}}$ represented by the scores of the data matrices to be compared. Thus, if the true dimensionality of the data is k, then $\mathrm{Y}_{\mathrm{nxk}}$ will be the true configuration and $\tilde{\mathrm{Z}}_{\mathrm{nxk}}$ is the corresponding approximation of the configuration based on only q traits. The difference between the two configurations $\mathrm{Y}_{\mathrm{nxk}}$ and $\tilde{\mathrm{Z}}_{\mathrm{nxk}}$ was calculated by the statistic M^2 from the differences between the corresponding scores of these settings. The loss of information due to the exclusion of (p-q) represents the residue produced when only the traits of q were used instead of all p traits.

$$\begin{tikzcd}
X(n \times p) \arrow[r, "Select"] \arrow[d, "PCA"] & \tilde{X}(n \times q) \arrow[d, "PCA"'] \\
Y(n \times k) \arrow[r, leftrightarrow, "Procrustes"'] & \tilde{Z}(n \times k)
\end{tikzcd}$$

Figure 2. Diagram illustrating Procrustes analysis data by Krzanowski (1987).

The choice of k in the most different areas has generally been based on the first k principal components to explain the total variance as much as possible, while also maintaining as much information contained in the original dataset as possible. A fixed k = 2 value was used in order to compare two-dimensional graphical dispersions, which is very useful in genetic breeding. Therefore, the M^2 statistic will characterize the disagreement between those two graphical representations, based on the distance between accessions presented on a single 2D chart.

The strategy proposed by Krzanowski (1987) uses the scores of the PCA to obtain the configurations. However, since his M^2 statistic goes from zero to infinity, an immeasurable space, it is better to use the Procrustes M^2 statistic provided by Peres-Neto and Jackson (2001), which gives a limited space. Accordingly, we established two criteria: backward (Criterion 1) and exhaustive (Criterion 2) algorithms for the selection of traits in the study of genetic diversity. We compare them with Jolliffe's criterion (1972) to validate the methodology. It is assumed that there is a trait's subset that satisfactorily represents the original dataset structure with a minimal loss of information (represented by M2) regarding the original dataset. That is, the residue produced by the loss of information due to the discard of some of the traits is minimal, therefore, the cluster pattern of the evaluated accessions is not significantly affected.

Criterion 1: The Backward Algorithm

Based on the backward algorithm proposed by Krzanowski (1987) and considering M^2 given by Equation [3], there is no stopping rule. The result is a sequence of (p-k) traits and their respective estimated M^2 values. It is important to remember that k=2 is the number of principal components chosen to graphically evaluate the genetic diversity. Moreover, the decision about which traits to retain in the study is arbitrary.

To select the subset of traits by means of this algorithm, the purpose of Criterion 1 is to establish a cut-off value for the M^2 statistic, called the M^2_{critical}. The resulting subset of traits, named optimal selection, is the subset that has the M^2 estimated value closer (less than or equal) to the M^2_{critical}.

Criterion 2: The Exhaustive Algorithm

Considering all combinations with k, k+1 until (p-1) traits totalizing $C_p^k, C_p^{k+1}, \dots, C_p^{p-1}$ subsets, respectively, it is possible find the optimal solution from the same M^2_{critical}. Thus, a total of $\sum_{i=k}^{p-1} C_p^i$ analyses were performed on all subsets and characterized a new procedure referred to as exhaustive algorithm, which certainly demands greater computational effort than Criterion 1.

From $M^2_{\text{critical}} = 0.1$, this procedure provided a series of subsets of traits with M^2 values lower than the M^2_{critical}. However, the optimal selection was the one that resulted in the M^2 estimated to be less than or equal to the M^2_{critical}.

Jolliffe's criterion: Traits discarded by the principal component's technique

The trait subsets obtained by Criteria 1 and 2 were compared with the subset obtained according to Jolliffe's criterion (1972), which considered the removal of traits with greater weights for less important components (minor variance). Considering standardized traits for genetic diversity studies, Cruz et al. (2011) recommend that the number of traits to discard should be equal to the number of components with eigenvalues less than 0.7.

To avoid the traits with a greater variance affecting the grouping result, the data standardization is commonly used before the PCA since the PCA is obtained from the covariance matrix (Σ). Here, statistical standardization of the data was performed where each value was subtracted by the mean and divided by the standard deviation of its respective variable. After standardization, the principal components are obtained from the covariance matrix of the standardized data, according Mingoti (2005), which is the same as obtaining the principal components from the correlation matrix (R) of the original dataset. Thus, we have $\Sigma = R$.

All statistics were performed in GENES software version 2016 (Cruz, 2013; 2016). GENES software is available on http://www.ufv.br/dbg/genes/genes.htm.

Results and discussion

A $M^2_{\text{critical}} = 0.1$ value obtained a subset with six, eight, and seven Conilon coffee traits according to criterion 1 (backward algorithm), criterion 2 (exhaustive algorithm), and Jolliffe's criterion, respectively (Table 1). Additionally, common traits existed between the subsets given by optimal selection from criteria 1 and 2, such as MG%, S15, and MS. The importance of these traits to accessions variability of the Conilon coffee is shown since the subsets selected by both criteria provided an increase in total variance explained by the first two principal components.

Table 1. Optimal selection using Procrustes and Jolliffe criteria.

Criteria	Optimal selection	M^2estimated	VTa%†	Discarded traits	Solution
Procrustes (backward)	GY, ChCo, MG%, S17, S15, and MS	0.0895	62.57	CoBe, S11, D, PH, S13, FG%, ChBe, DCA, GM%, and CG%	Only
Procrustes (exhaustive)	DCA, ChBe, CoBe, MG%, GM%, S15, S11, and MS	0.1	56.82	D, GY, PH, ChCo, CG%, FG%, S17, and S13	1 in 9,841
Jolliffe	D, DCA, ChCo, CoBe, CG%, FG%, and S17	0.3359	51.50	S13, MS, MG%, ChBe, S11, GY, PH, S15, and GM%	Only

†VTa%: Cumulative percent of the total variation explained by the two primary components.

The 2D graphical dispersion of accessions of Conilon coffee, considering all 16 traits (Figure 3a), represents the original data configuration and explains 49.35% of total variance. Although the 40 accessions could not be grouped in clusters, the graphical dispersion was considered useful in making inferences about Conilon coffee accessions diversity in this study. Figure 3a shows that accessions 13 and 8 are divergent, and according to their *per se* potential, they can be used in a cross to explore vigor and increase variability.

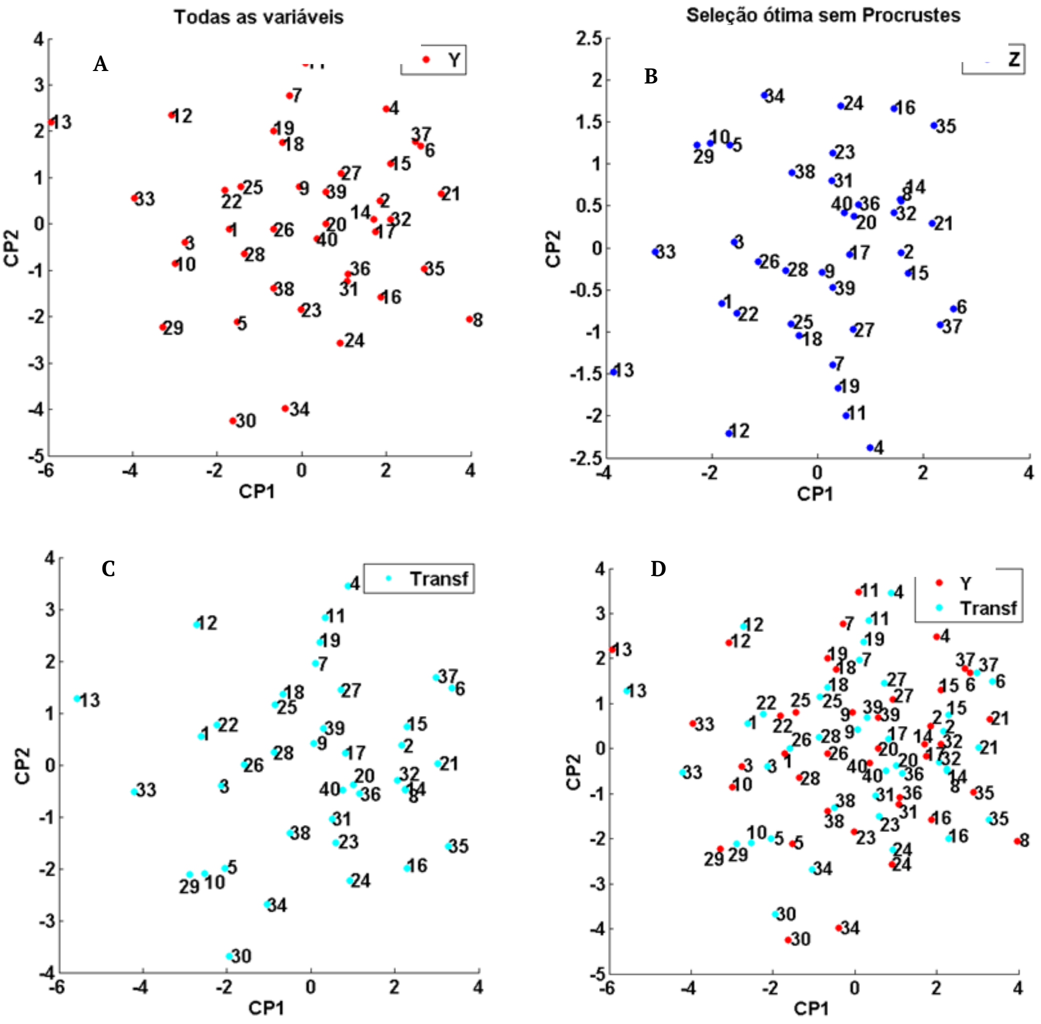

Figure 3. Dispersion ranking of Conilon coffee accessions using two principal components (CP1 and CP2), according to: A – evaluation of 16 agronomic traits, B – Criteria 1 without Procrustes transformation, C – Criteria 1 with Procrustes transformation, and D – A and C superimposed ($M^2 = 0.089445$).

Figure 3b shows the scores graphical dispersion of the accessions in relation to the first two components for the optimal selection resulted by Criterion 1. Note the change on the position of the accessions since they were reflected around the origin of the component 2 in relation to the original configuration given by Figure 3a (accessions now positive but were previously negative). To make the matching between these configurations feasible, following the steps described in material and methods and illustrated in Figure 2, the Procrustes analysis adjusted the configuration of Figure 3b in 3a such that the distance between them is minimal. After the Procrustes transformation on the optimal selection, it was then possible to calculate its true difference in relation to the original dataset estimated by means of $M^2 = 0.0895$.

It was verified that accession 8, which was previously divergent such as accession 13, was now in the same group of genotypes that included accession 14 (Figure 3c). Thus, we have the optimal selection with six traits that did not satisfactorily represent the diversity pattern from the dispersion given by the original dataset (Figure 3a). We can better visualize the change in the clustering pattern of accessions by superimposing the graphs a and c (Figure 3d). It is worth pointing out that even if the optimal selection included the characteristics of interest, it was not adequate for evaluating the diversity of Conilon coffee accessions.

Criterion 1 provided a sequence of traits whose exclusion at each step of the backward algorithm provided the lowest estimation of the M^2 value (Table 2). Note that the estimated M^2 value is increased by discarding the traits in each step of the algorithm. This was expected, as the discarding of variables increased the residue produced by the loss of information compared to the original dataset. As a six-variable subset resultant of criterion 1 did not satisfactorily represent the structure of accessions diversity, the researcher can choose a different M^2_{critical} value (higher or lower than the last one used). Therefore, more or fewer traits are considered for a re-evaluation of the clustering pattern among accessions according to how much loss of information the researcher can tolerate. If the subsequent M^2_{critical} values are inappropriate, the researcher can follow the study considering the information of all sixteen traits or evaluate different subsets given by Criterion 2.

Table 2. Variables excluded by the backward algorithm for the Conilon coffee.

	Discarded											Selected				
Excluded variables	CoBe	S11	D	PH	S13	FG%	ChBe	DCA	GM%	CG%	MS	GY	ChC	S15	MG%	S17
M^2	0.0074	0.0111	0.0239	0.0236	0.0331	0.0399	0.0528	0.0569	0.0678	0.0894	0.1535	0.1903	0.2538	0.4325	-	-

CoBe: Coconut and green coffee relationship, S11: Percentage of grains retained on the sieve mesh size #11, D: Number of days, plant height (PH, in cm), S13: Percentage of grains retained on the sieve mesh size #13, FG%:"Flat" grain percentage, ChBe: Cherry and green coffee relationship, DCA: Diameter crown average (cm), GM%: Grain moisture percentage, CG%: Coarse grain percentage, MS: Medium strainer (medium grain size), GY: Grain yield (kg ha^{-1}), ChCo: Cherry and coconut dry coffee relationship, S15: Percentage of grains retained on the sieve mesh size #15, MG%: "Mocha" grain percentage and S17: Percentage of grains retained on the sieve mesh size #17.

Figure 4b shows that the accessions given by the optimal selection of Criterion 2 were reflected, as in Criterion 1, but now in relation to the origin of components 1 and 2, simultaneously. According to the dispersion of the accessions presented by the optimal selection resulting from Criterion 2, no change in the clustering pattern (Figure 4d) was observed. Therefore, the optimal selection (Figure 4c) provided a global dispersion satisfactorily close to the given dispersion of the original dataset (Figure 4a).

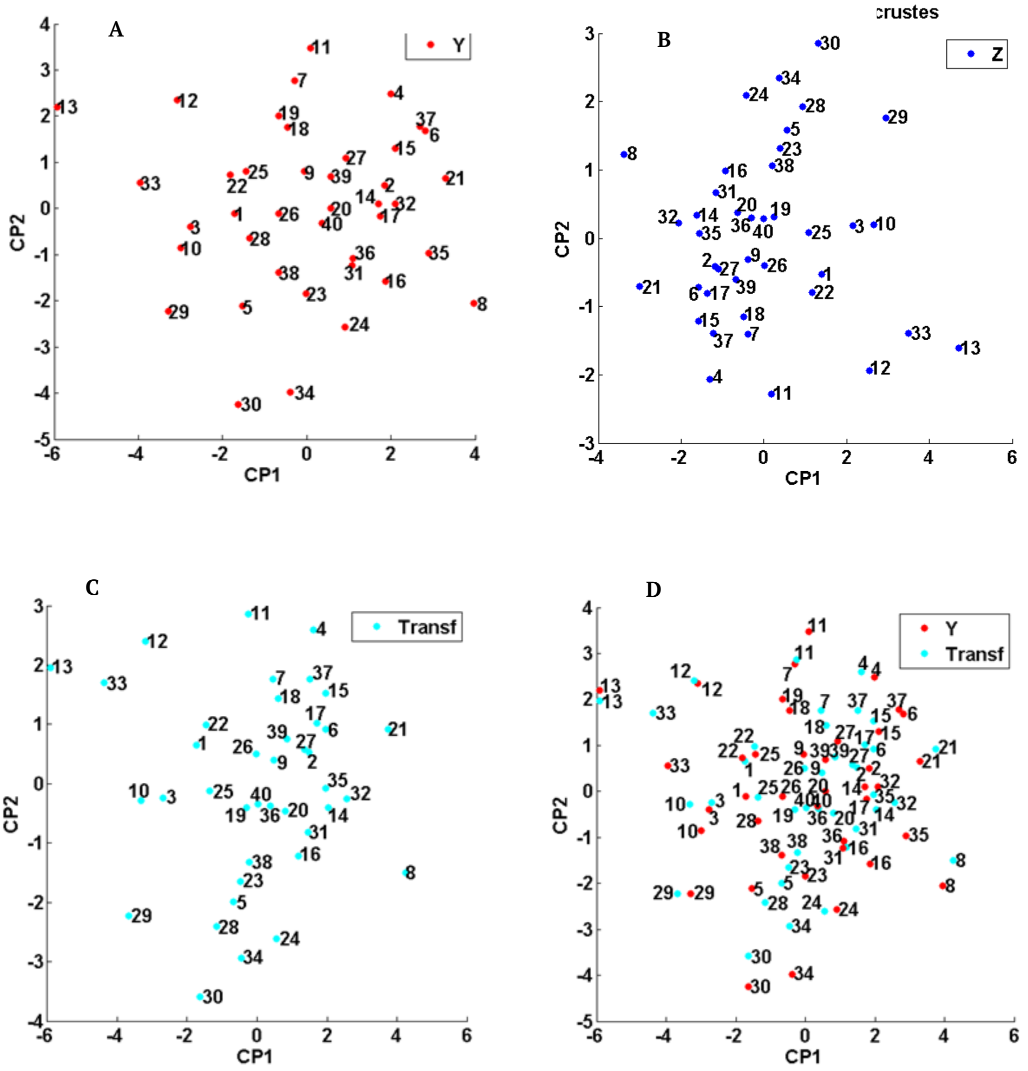

Figure 4. Dispersion ranking of Conilon coffee accessions using two principal components (CP1 and CP2), according to A – evaluation of 16 agronomic traits, B – Criteria 2 without Procrustes transformation, C – Criteria 2 with Procrustes transformation, and D – A and C superimposed ($M^2 = 0.1$)

Criterion 2 provided a total of 9,841 combinations (subsets) that resulted in M^2 values lower than M^2_{critical} (Table 3), which include the optimal selection resulting from Criterion 1. If the optimal selection of

Criterion 2 does not satisfy the breeder's purposes, it is possible to evaluate the diversity of other subsets with more or fewer traits.

Table 3. Total subsets determined by the exhaustive algorithm with $M^2_{\text{critical}} < 0.1$.

Traits	Subsets
5	2
6	64
7	322
8	735
9	1,538
10	2,490
11	2,504
12	1,502
13	548
14	120
15	16
Total	9,841

From the data presented in Table 4, it is possible to identify the relative importance of the traits on the genetic diversity of the Conilon coffee accessions through which the deletion must be performed. According to a criterion presented by Jolliffe (1972) and suggested by Cruz et. al. (2011), from the last to the ninth principal component, the traits of greatest weights were S13, MS, MG%, ChBe, S11, GY, PH, S15, and GM%. Accordingly, the optimal selection was given by the subset of traits: D, DCA, ChCo, CoBe, CG%, FG%, and S17.

Table 4. Eigenvalue estimates from the correlation matrix, containing 16 traits and associated eigenvectors (components).

†λ	VT%	VTa%	Traits															
			D(Days)	GY(kg/ha)	PH(cm)	DCA(cm)	ChCo	ChBe	CoBe	CG%	FG%	MG%	GM%	S17(%)	S15(%)	S13(%)	S11(%)	MS
4.7455	29.6592	29.6592	0.1232	0.2484	0.0986	0.0411	-0.0074	-0.2062	-0.2427	-0.2021	-0.0141	0.004	0.1996	0.342	0.3336	-0.4028	-0.4001	0.4227
3.1508	19.6928	49.352	-0.2224	-0.2399	-0.2168	-0.3252	-0.3129	-0.2551	0.0065	0.1987	0.4322	-0.4125	-0.298	0.1998	-0.0455	-0.1696	0.0173	0.1449
1.8687	11.6794	61.0314	-0.4702	0.3076	0.4762	0.4349	-0.2447	-0.1216	0.1346	0.2525	0.0954	-0.0763	-0.1413	-0.166	0.1811	0.0727	-0.1023	-0.0282
1.4027	8.7669	69.7983	-0.1501	-0.02	-0.2756	-0.0342	-0.3937	-0.3792	-0.1923	0.1677	-0.4756	0.5244	-0.1734	0.0329	-0.0332	-0.024	0.0379	-0.0253
1.2503	7.8145	77.6128	-0.0331	-0.1849	-0.1226	-0.1889	-0.2062	0.3865	0.6281	0.018	-0.1457	0.1555	-0.0683	-0.0699	0.3876	-0.0783	-0.3189	0.1195
1.0303	6.4394	84.0522	-0.288	0.1108	-0.011	0.0506	0.4514	0.2802	0.1072	0.3921	-0.1596	0.1161	-0.1815	0.4425	-0.2793	-0.2514	0.0874	0.195
0.7983	4.9891	89.0414	0.1637	-0.159	0.2618	-0.23	-0.0063	-0.2259	0.1373	0.592	-0.0278	0.0232	0.6126	-0.0197	0.056	-0.1	0.1183	-0.0434
0.526	3.2875	92.3288	0.4582	0.4745	-0.4142	0.3515	-0.265	0.0038	0.2984	0.2031	0.1034	-0.1152	0.0385	0.086	-0.1336	-0.045	0.1274	0.0155
0.404	2.5249	94.8538	-0.0673	0.4471	-0.2347	-0.3525	0.1958	0.1661	-0.298	0.2442	0.044	-0.04	-0.113	-0.3364	0.4588	-0.1351	0.1391	-0.1559
0.3091	1.9316	96.7853	0.4837	0.1882	0.5185	-0.3681	-0.1591	0.0768	-0.0435	0.1166	-0.0752	0.0513	-0.4681	0.0842	-0.1564	0.1078	-0.0927	0.0172
0.2653	1.658	98.4434	-0.3596	0.476	0.0239	-0.4555	-0.223	0.0711	0.157	-0.2914	0.0094	0.0394	0.3586	0.0962	-0.3195	0.1564	0.0545	0.0509
0.1043	0.652	99.0954	0.0063	-0.109	0.2333	0.0856	-0.3495	0.3045	-0.071	-0.2121	-0.0187	0.0504	0.0186	0.0451	0.107	-0.4886	0.6371	0.0065
0.0971	0.6071	99.7025	-0.0315	-0.1175	-0.0767	0.1131	-0.3579	0.568	-0.499	0.2734	0.1014	0.0066	0.2121	0.0733	-0.1139	0.1845	-0.288	0.0371
0.0306	0.1915	99.8939	0.0435	-0.0098	0.0044	0.009	0.0832	-0.0273	0.019	0.0275	0.6156	0.6287	-0.0318	-0.1773	-0.0176	0.0978	0.1208	0.3976
0.0166	0.1037	99.9976	0.0087	-0.033	-0.0218	0.0062	-0.0013	0.0124	-0.0486	0.0578	-0.3497	-0.3066	-0.0044	-0.3427	-0.0304	0.228	0.2194	0.7467
0.0004	0.0024	100	0.0037	-0.0035	-0.0031	0.003	0.0051	-0.0025	-0.0023	0.0041	0.0068	0.0053	0.0021	0.5659	0.4906	0.5772	0.3249	0.0126

†λ: Eigenvalue. VT%: Percentage of total variation explained by the i-th principal component. VTa%: Cumulative percentage by components.

Figure 5b shows the dispersion of the accessions scores in relation to the first two principal components for the subset of seven traits established by Jolliffe’s criterion. As in previous cases, the change of accessions position occurred due to the exclusion of some traits, which were reflected around the origin of component 1 and component 2. After the Procrustes transformation on the optimal selection, its real difference in relation to the original set was estimated by $M^2 = 0.3359$. The estimated magnitude of M^2 translated the nonproximity between the coffee accessions corresponding to the configurations (Figure 5d), which revealed a significant change in the pattern of clustering of the accessions. This difference can be seen in accession 19, which was fitted to accession group 13 after transformation, as well as accessions 16, 17, and 35, all belonging to accession group 8 after the transformation (Figure 5c).

From the moment the researcher knows which traits are of greater biological importance on the characteristic to be improved, their use can reflect their importance and lead to saving time and financial resources, making breeding programs more sustainable. Thus, if the breeder has an interest in a specific subset, its diversity can be evaluated graphically and its estimated value of M^2 compared to that obtained by optimal selection of the exhaustive or backward algorithm as a way of guiding discovery of how the magnitude of M^2 is affecting the dispersion of its accessions group.

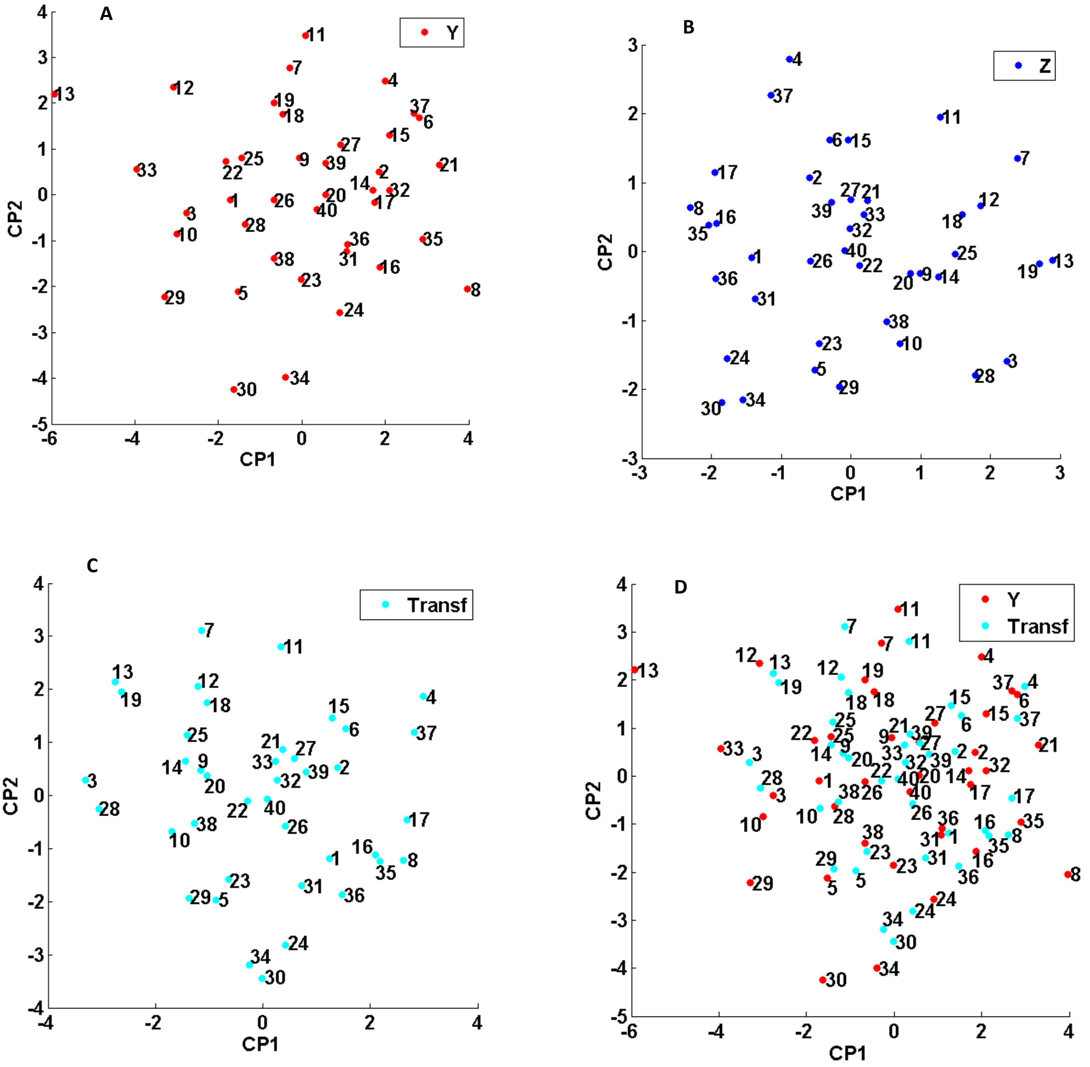

Figure 5. Dispersion ranking of Conilon coffee accessions using two principal components (CP1 and CP2), according to A – evaluation of 16 agronomic traits, B – Jolliffe criteria without Procrustes transformation, C – Jolliffe criteria with Procrustes transformation, and D – A and C superimposed (M^2 = 0.33585).

According to the obtained results it can be observed that the optimal selection given by Criterion 1 provided the lowest value of M^2 estimated and the smallest number of traits. However, this did not adequately represent the Conilon coffee diversity considering the PCA from the set with all 16 traits (Figure 5a). Furthermore, the subset selected by Criterion 2, despite having a greater number of traits, satisfactorily represented the diversity among the accessions. Note that the subset selected by Jolliffe's criterion provided a high value of estimated M^2, which was 3 times more than the critical M^2_{critical}, revealing a change in the cluster pattern of the accessions and making this criterion relatively less efficient than the others.

Based on the Procrustes analysis, the number of solutions of each criterion should be taken into account. In the case of Criterion 1 and Jolliffe's criterion, only one optimal selection was provided, while Criterion 2 provided all subsets of traits with an estimated value of M^2 below M^2_{critical} (Table 3). This opens a range of possibilities for the researcher's decision-making since the M^2_{critical} and the backward algorithm may not include some variables that present biological importance into the process of genetic improvement of the culture. Additionally, the subset selected by Criterion 1 may not reveal graphical scatter equivalent to that obtained by the analysis of the original set.

We also must pay attention to the process of obtaining solutions. Unlike Criterion 1, which excludes one variable at a time in each step of the backward algorithm, Criterion 2 uses the exhaustive algorithm that evaluates all possibilities of discarding traits - one by one, two by two, etc. The stepwise algorithm, which is useful in selecting traits in linear regression models, is different from the method for Criterion 2 because it establishes the importance of traits by a different decision rule and the exclusion or inclusion of traits is made interactively.

It is possible to verify the total analyses performed by the exhaustive algorithm according to the number of traits studied (Table 5). Note that as the number of traits increases, the number of analyses performed by Criterion 2 increases considerably. Thus, Criterion 2 becomes uninteresting in cases of high-order data matrices whose handling involves high computational cost, and processing the results may take months. However, the researcher must consider its use by computational resources as well as the time it has since there currently are no studies that establish the computational cost of this algorithm in relation to the number of traits.

Table 5. Total number of analyzes performed by exhaustive algorithm on the Conilon coffee.

# of variables	Analyses
10	1,012
16	65,518
20	1,048,554
30	1,073,741,792
50	1,125,899,906,842,570

Notice that the M^2 statistic used in this work ranged from 0 to 1, and the $\mathrm{M}^2_{\text{crítical}}$ value can be interpreted as the percentage loss of information acceptable resulting from the selected subset of traits. Thus, the researcher must consider that even if a relatively small loss is established, the resulting subset of traits may or may not satisfactorily represent the genetic diversity of the original dataset. This occurs because the breeder's considerations of the biological importance of a variable may be different from the statistical significance. Therefore, the optimal selection must include all traits that are important to the breeder and represent the level of diversity in the original dataset.

Another interesting aspect about the criteria based on the Procrustes analysis concerns the value of $\mathrm{M}^2_{\text{critical}} = 0.1$ suggested in this study. It is worth mentioning that in Criteria 1 and 2, the critical value M^2 can be slightly relaxed according to the number of traits that the breeder wishes to discard. In this sense, we suggest a variation interval from a minimum value of 0.05 to a maximum of 0.15, as long as the clustering pattern of accessions of the culture is maintained. These limits do not constitute a rule since there are no other studies that discard traits using these specific limits for genetic diversity, and therefore, the researcher must decide them. However, it is worth remembering that the Procrustes statistic adopted in this work ranges from 0 to 1, $\mathrm{M}^2_{\text{critical}}$ and was selected assuming that the residual produced by the loss of information with the resulting subset of traits would be 10% at most.

Based on the strategy proposed by Krzanowski (1987) and the Krzanowski (1996) backward algorithm, Munita, Barroso, and Oliveira (2013) obtained their results with a subset of only eight traits sufficient to interpret the data in two axes (k = 2 principal components) that explained 76.6% of the total variation without substantial loss of information. The dataset represented the concentration of 13 chemical elements (traits) obtained by activation with neutrons in a set of 40 samples of ceramic fragments, whose first two components explained 79.9% of the total variation. Guedes and Ivanqui (1998) obtained similar results in a medical study with simulated data regarding 14 traits related to liver cancer, whose first two main components explained 93.61% of the total variation. Based on the backward algorithm without a stop rule, a subset with 8 traits was established by Procrustes analysis with configuration similar to the original set with representation in two axes that explained 93.66% of the data variation.

The results obtained in this study also showed that even with minimal explanation of the total variation of the data by the first two principal components, it was possible to obtain a satisfactory representation of the accessions diversity in two axes according to the optimal selection obtained by Procrustes analysis. This confirmed the importance of the contribution of the proposed criteria and the technique presented for the selection of traits in the study of genetic diversity. Finally, the exhaustive procedure, which suggests enormous potential for genetic studies, is highlighted by the number of resulting optimal solutions.

The Procrustes analysis presents wide applicability and has interesting approaches. For the plant breeding, there is currently no literature using Procrustes analysis to select phenotypic traits, which further highlights the relevance of this study for genetic improvement. Although Procrustes analysis has been minimally explored in the area of plant breeding, García-Peña and Dias (2009) used the analysis to compare different techniques of uni- and multivariate analysis by the AMMI methodology in the genotypic versus environmental interaction study. The joint use of the Procrustes and PCA techniques presents enormous potential and its application in genetic improvement extends beyond the selection of variables, including the possibility of evaluating the genetic diversity that is important for a breeding program through graphic dispersion.

This study provides the breeder with a technique based on Procrustes analysis to assist him in the decision-making regarding the exclusion of redundant characters. In practical terms, character exclusion can reduce possible measurement errors and reduce experiment time and costs since the excluded character may require a high cost of measurement or be difficult to measure. Technically, Procrustes analysis in diversity studies allows for visualization of the pattern of grouping of accessions after discarding variables. This allows the breeder to graphically evaluate the selected subset of traits, either by an automated selection method or determined by the breeders themselves. In addition, it allows quantification, through the statistics M^2, of the loss of information of a reduced subset of selected traits in relation to the set of all traits.

Conclusion

The flexibility in selecting traits by the researcher, graphical visualization, and Procrustes M^2 statistics through Criteria 1 and 2 becomes a fast and reliable alternative for decision-making of traits for phenotyping in studies of Conilon coffee diversity as well as other crops. Procrustes analysis is advantageous in selecting traits and provides a relevant contribution to genetic diversity studies as an efficient alternative to Jolliffe’s criterion.

Acknowledgements

We would like to thank FAPEMIG, CAPES, and CNPq for the financial support and for conceding scholarships to the project development.

References

Bonomo, P., Cruz, C. D., Viana, J. M. S., Pereira, A. A., Oliveira, V. R., & Carneiro, P. C. S. (2004). Seleção antecipada de progênies de café descendentes de “híbrido de timor” X “catuaí amarelo” e “catuaí vermelho. *Acta Scientiarum. Agronomy*, *26*(1), 91-96. DOI: 10.4025/actasciagron.v26i1.1969

Bramardi, S. J., Bernet, G. P., Asíns, M. J., & Carbonell, E. A. (2005). Simultaneous agronomic and molecular characterization of genotypes via the Generalised Procrustes Analysis. *Crop Science*, *45*(4), 1603-1609. DOI: 10.2135/cropsci2004.0633

Cruz, C. D. (2013). Genes: a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum. Agronomy*, *35*(3), 271-276. DOI: 10.4025/actasciagron.v35i3.21251

Cruz, C. D. (2016). Genes Software-extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum. Agronomy*, *38*(4), 547-552. DOI: 10.4025/actasciagron.v38i4.32629

Cruz, C. D., Ferreira, F. M., & Pessoni, L. A. (2011). *Biometria aplicada ao estudo da diversidade genética.* Visconde do Rio Branco, MG: Suprema.

Daboul, A., Ivanovska, T., Bülow, R., Biffar, R., & Cardini, A. (2018). Procrustes-based geometric morphometrics on MRI images: An example of inter-operator bias in 3D landmarks and its impact on big datasets. *PLoS ONE*, *13*(5), e0197675. DOI: 10.1371/journal.pone.0197675

Douglas, T. S. (2004). Image processing for craniofacial landmark identification and measurement: a review of photogrammetry and cephalometry. *Computerized Medical Imaging and Graphics*, *28*(7), 401-409. DOI: 10.1016/j.compmedimag.2004.06.002

Ferrão, R. G. Cruz, C. D., Ferreira, A., Cecon, P. R., Ferrão, M. A. G., Fonseca, A. F. A. D., ... Silva, M. F. D. (2008). Parâmetros genéticos em café Conilon. *Pesquisa Agropecuária Brasileira*, *43*(1), 61-69. DOI: 10.1590/S0100-204X2008000100009

García-Peña, M., & Dias, C. T. S. (2009). Análise dos modelos aditivos com interação multiplicativa (AMMI) bivariados. *Revista Brasileira de Biometria*, *27*(4), 586-602.

Guedes, T. A., & Ivanqui, I. L. (1998). Análise procrustes aplicada à seleção de variáveis. *Acta Scientiarum. Technology*, *20*, 505-509. DOI: 10.4025/actascitechnol.v20i0.3073

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *21*(2), 160-173. DOI: 10.2307/2346488

Klingenberg, C. P. (2003). Quantitative genetics of geometric shape: heritability and the pitfalls of the univariate approach. *Evolution*, *57*(1), 191-195. DOI: 10.1111/j.0014-3820.2003.tb00230.x

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *36*(1), 22-33. DOI: 10.2307/2347842

Krzanowski, W. J. (1996). A stopping rule for structure-preserving variable selection. *Statistics and Computing*, *6*(1), 51-56. DOI: 10.1007/BF00161573

Liu, S., Zheng, X., Yu, L., Feng, L., Wang, J., Gong, T., ... Xu, R. (2017). Comparison of the genetic structure between *in situ* and *ex situ* populations of Dongxiang wild rice (*Oryza rufipogon* Griff.). *Crop Science*, *57*(6), 3075-3084. DOI: 10.2135/cropsci2017.01.0015

Mauricio, A. A., Palazzo, A. B., Caselato, V. M., & Bolini, H. M. A. (2016). Generalized procrustes analysis and external preference map used to consumer drivers of diet gluten free product. *Food and Nutrition Sciences*, *7*(9), 711-723. DOI: 10.4236/fns.2016.79072

Mingoti, S. A. (2005). *Análise de dados através de métodos de estatística multivariada*: uma abordagem aplicada. Belo Horizonte, MG: Editora UFMG.

Muleta, K. T., Bulli, P., Zhang, Z., Chen, X., & Pumphrey, M. (2017). Unlocking diversity in germplasm collections via genomic selection: A case study based on quantitative adult plant resistance to stripe rust in spring wheat. *The Plant Genome*, *10*(3), 1-15. DOI: 10.3835/plantgenome2016.12.0124

Munita, C. S., Barroso, L. P., & Oliveira, P. M. (2013). Variable selection study using Procrustes analysis. *Open Journal of Archaeometry*, *1*(e7), 31-35. DOI: 10.4081/arc.2013.e7

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, *129*(2), 169-178. DOI: 10.1007/s004420100720

Oliveira, A. P. V., & Toledo Benassi, M. de (2010). Avaliação sensorial de pudins de chocolate com açúcar e dietéticos por perfil livre. *Ciência e Agrotecnologia*, *34*(1), 146-154. DOI: 10.1590/S1413-70542010000100019

Rodrigues, W. N., Brinate, S. V., Martins, L. D., Colodetti, T. V., & Tomaz, M. A. (2017). Genetic variability and expression of agro-morphological traits among genotypes of *Coffea arabica* being promoted by supplementary irrigation. *Genetics and Molecular Research*, *16*(2). DOI: 10.4238/gmr16029563

Singh, D. (1981). The relative importance of characters affecting genetic divergence. *Indian Journal of Genetics and Plant Breeding*, *41*(2), 237-245.

Yano, R., Nonaka, S., & Ezura, H. (2018). Melonet-DB, a grand RNA-Seq gene expression atlas in melon (*Cucumis melo* L.). *Plant and Cell Physiology*, *59*(1), e4-e4. DOI: 10.1093/pcp/pcx193

Yousaf, M. I., Hussain, K., Hussain, S., Ghani, A., Arshad, M., Mumtaz, A., & Hameed, R. A. (2018). Characterization of indigenous and exotic maize hybrids for grain yield and quality traits under heat stress. *International journal of Agriculture and Biology*, *20*(2), 333-337. DOI: 10.17957/IJAB/15.0493