Classification of *Coffea canephora* clones in botanical varieties by discriminant analysis of the k-nearest neighbors[1]

Marciléia Santos Souza[2], Fábio Medeiros Ferreira[3], Rodrigo Barros Rocha[4], Maria Teresa Gomes Lopes[2], Leilane Nicolino Lamarão Oliveira[5]*

10.1590/0034-737X202168050007

ABSTRACT

A strategy for genetic improvement of coffee *Coffea canephora* plants is to aggregate through artificial crossings the characteristics of the Conilon botanical variety, such as shorter height and drought resistance, with the higher average grain size and resistance to pests and diseases of the Robusta variety. Efficiently separating the clones into these two groups with the aid of appropriate analytical procedures makes field tasks easier for professionals and, thus, allows the systematic production of intervarietal hybrids. This study verifies if the non-parametric discriminant analyzes of the k-nearest neighbors (k-NN) and k-average neighbors (k-AN) would be able to correctly classify 130 coffee clones in their botanical varieties previously designated as Conilon, Robusta and Intervarietal Hybrids populations from ten quantitative agronomic characteristics, including the processed coffee beans yield, considering the existing population genetic divergence. These characteristics were found to be good discriminatory variables and the discriminant analyzes k-NN and k-AN, based on the principle of similarity by neighborhood, classified the clones with high hit rates. The k-AN discriminant analysis was able to better discriminate intervarietal hybrids from the group clones Conilon. The results correctly reflected the genetic diversity between the botanical varieties and intervarietal hybrids of *Coffea canephora*, allowing us to conclude that these classification methods can assist breeders in the main task of discriminating Conilon from Robusta clones.

Keywords: Conilon; Robusta; genetic diversity; quantitative trait.

INTRODUCTION

In the coffee plant *Coffea canephora* Pierre ex Froehner, two cultivated botanical varieties stand out commercially and exhibit different characteristics (Davis *et al*., 2006). The characteristics of the Robusta botanical variety are greater vigor, erect growth, larger leaves and fruits, late maturation, less tolerance to water deficit, and greater tolerance to pests and diseases. Plants of the Conilon botanical variety have shrubby growth, early flowering, branched stems, elongated leaves, drought resistance, and greater susceptibility to diseases (Ferrão *et al*., 2015).

The crossing of these two varieties occurs naturally, creating hybrid genotypes that can exhibit the best characteristics of each group, associated with the expression of heterosis (Charrier & Berthaud, 1988). Field evaluations seek to add characteristics such as the shorter height and drought resistance of the Conilon variety along with higher average grain size and resistance to pests and diseases of the Robusta variety. The efficient separation of these two botanical varieties allows the systematic production of intervarietal hybrids (Rocha *et al*., 2015).

Thus, plant breeders of *Coffea canephora* need to be able to classify into their respective botanical varieties the most similar genotypes and to identify those that truly diverge, in order to maintain the two populations with a

Submitted on May 14th, 2020 and accepted on February 10th, 2021.

[1]This work is part of the master's dissertation of the first author

[2]Universidade Federal do Amazonas, Faculdade de Ciências Agrárias, Manaus, Amazonas, Brazil. marcileia_souza@hotmail.com; mtglopes@ufam.edu.br;
[3]Universidade Federal do Amazonas, Instituto de Ciências Exatas e Tecnologia, Itacoatiara, Amazonas, Brazil. ferreirafmt@ufam.edu.br;
[4]Empresa Brasileira de Pesquisa Agropecuária, Centro de Pesquisa Agroflorestal, Porto Velho, Rondônia, Brazil. rodrigorocha@embrapa.br;
[5]Prefeitura Municipal de Urucurituba, Secretaria de Meio Ambiente, Urucurituba, Amazonas, Brazil. leilane.nl@gmail.com.
*Corresponding author: ferreirafmt@ufam.edu.br

high heterotic effect, and consequently, to explore the genetic variability of future generations.

Field plant classification is limited due easy to observe characteristics, such as flower and fruit morphology, tend to overlap with other important characteristics used to identify clones in plant breeding populations.

Discriminant analysis statistically distinguishes populations, previously defined by some criterion, from a set of "discriminatory" variables measured in *n* individuals, later classifying them into one of the groups (Hair *et al*., 2005; Khattree & Naik, 2000). Classical discriminant procedures, such as approaches based on linear discriminant functions (Fisher, 1936; Anderson, 1958), commonly employed in plant genetic improvement, are based on the assumptions of multivariate population normality and homoscedasticity of the variance-covariance matrices between evaluated populations (Cruz *et al*., 2020). For example, for the *Coffea canephora* species, Fonseca *et al*. (2004) successfully defined different linear functions, which could classify 32 clones into three varieties of Robusta coffee with different maturation cycles, based on 17 quantitative variables associated with bean production.

In cases of inequality between the variance-covariance matrices, quadratic discriminant functions are recommended (Cruz *et al*., 2020). Graphical analysis, via principal components with establishment of population centroids, also commonly employed in plant breeding studies (Oliveira *et al*., 2018), does not require specific distribution; however, populations must have a common matrix of variances and covariance (Khattree & Naik, 2000).

The recommendation to use the data transformation method does not always satisfy this analytical condition (Khattree & Naik, 2000). On the other hand, the literature proposes simple techniques of discriminant analysis, known as non-parametric, which are free from the assumption of normality, like the k-nearest neighbors method (k-NN), which is based on allocating the genotype based on the greater probability of classifying it with a group of genotypes - the closest neighbors - belonging to one of the populations evaluated, whose proximity is defined from a distance measure (Silverman *et al*., 1989). Following the idea of discrimination by neighborhood, Cruz *et al*. (2020) proposed that the allocation would be due to the shortest average distance of the genotype in relation to the other genotypes belonging to each of the predefined populations, referring to a concept of average neighborhood and, thus, called the k-average neighbors (k-AN).

The literature refers to the application of k-NN successfully in genetic studies both for data on molecular markers (Mcharo & LaBonte, 2010; Oliveira *et al*., 2012; Zhang *et al*., 2005) and for quantitative agronomic characteristics (Bannayan & Hoogenboom, 2009; Nielsen *et al*., 2003).

The k-NN technique proved to be effective for correctly allocating genotypes of rice (Dheer & Singh, 2019) and wheat (Dheer *et al*., 2019) in their populations, compared to Fisher's linear discriminant analysis, logistic regression, and the Näive Bayes classifier, based on agronomic variables of continuous distribution. K-NN is considered one of the simplest machine learning algorithms, in terms of classification, implementation, and understanding, in addition to being robust and producing good results as a popular classifier in several areas (Salari *et al*., 2014). The literature does not report applications of the k-average neighbors method.

Considering that breeding programs of the species *Coffea canephora* have focused on the exploration of promising artificial crosses between the Conilon and Robusta types, our objective was to verify if the discriminant analyzes k-NN and k-AN could classify and correctly allocate *Coffea canephora* clones into botanical varieties or in intervarietal hybrids, based on agronomic characteristics commonly measured in the field and considering the existing population genetic divergence.

MATERIAL AND METHODS

Field experiment

In December 2011, the coffee was sowed in the experimental field of Embrapa Rondônia in the municipality of Ouro Preto do Oeste, RO, Brazil (10° 37' 03'' S and 62° 51' 50'' W). The competition assay and evaluation of genetic variability between 130 *C. canephora* coffee clones was delineated in four complete randomized blocks, with four plants per plot, spaced at 3 × 2 m. The management and cultural treatment of planting followed the recommendations, according to Marcolan *et al*. (2009).

The evaluated clones represented the botanical varieties Conilon (73 clones) and Robusta (38 clones), in addition to intervarietal hybrids (19 clones) from these two varieties. The categorization of the genotypes in these three populations was through field observations in relation to agronomic behavior, characteristic of each group (such as disease resistance, drought tolerance, vigor, size and architecture of the plant, size of leaves and sieve-size, and quality of the drink) (Musoli *et al*., 2009).

Ten agronomic characteristics were measured: i) plant height (m), measured from the soil level to the final growth point of the plant; ii) number of productive plagiotropic branches; iii) number of rosettes per plagiotropic branch, obtained from the average of three evaluations; iv) length (m) of the plagiotropic branch, measured from the initial insertion of the orthotropic branch to its final growth point; v) distance (cm) between rosettes of the intermediate part of the plagiotropic branch, obtained from the average of

three evaluations; vi) number of fruits per rosette, obtained from the average of three evaluations; vii) leaf length (cm), measured from the leaf insertion in the petiole until its end; viii) leaf width (cm), measured at the widest part of the leaf; ix) number of days for maturation, registering day between flowering and harvest; x) production of processed coffee beans (bags of 60 kg.ha^{-1}).

Discriminant analysis and classification of clones

The data used in the analyzes were represented by the arithmetic averages of the evaluated plants of each of the clones, obtained three years after planting in the 2015/2016 agricultural harvest. Preliminarily, multivariate normality was verified in each population, based on the asymmetry and kurtosis tests proposed by Mardia (1970). In addition, the Box's M test (Box, 1949) was applied to verify the multivariate homoscedasticity of the population variances and covariance matrices. For all tests, a significance level of 0.05 was adopted. The intra-population and inter-population pairwise Euclidean distances were also estimated, to comparatively assess genetic diversity and divergence.

Two arrangements in the dataset were established for the analyzes. The first analysis included all clones of the two botanical varieties, plus the intervarietal hybrids (CHR set). In the second dataset, the hybrid clones were removed, so only the clones of the botanical varieties Conilon and Robusta remained (CR set). All other procedures described below were performed for both datasets.

Discriminant analyzes of the k-nearest neighbors (k-NN) (Fix & Hodges, 1951; Fix & Hodges, 1952) and k-average neighbors (k-AN) (Cruz *et al*., 2020) were used. The data were standardized beforehand to prevent the units of measurement used for the agronomic characteristics from arbitrarily affecting the similarity between the genotypes for the characteristics to contribute equally to the evaluation.

To implement the k-NN method, first, all Euclidean distances ($d_{ii'}$) between pairs of clones were estimated, for every $i \neq i'$. After that, the value of k was defined, which corresponded to the number of nearest clones (neighbors), with less genetic distance, in relation to the plant that was desired to be classified in one of the studied populations. Thus, the k value established the maximum number of nearest neighbors that could be obtained in each simulation.

In this way, all possible k values for the two datasets were evaluated. This served to verify the disagreements for the classification of clones and the variations in the number of nearest neighbors. The k values ranged from 1 to 18 (= n_l^* - 1) for the CHR set, and from 1 to 37 for the CR set, where n_l^* is the size of the smallest population, which is the Hybrids Intervarietals for CHR set and Robusta clones for CR data.

Among these closest k genotypes, k_l may come from one of the L populations studied, whose *a priori* probability of a genotype belonging to it was π_l. Then, the probability of classifying a genotype to belong to the *lth* population was estimated by

$$\hat{P}(Y_i,P_l) = \frac{\pi_l\left(\frac{k_l}{n_l}\right)}{\Sigma_{l=1}^{L}\pi_l\left(\frac{k_l}{n_l}\right)}$$

where: n_l is the genotype number of the *lth* population; k_l is the number of neighbors nearest the genotype i (Y_i), belonging to *lth* population P_l, among the k-nearest neighbors found. The clone Y_i was allocated to the *lth* population when $\hat{P}(Y_i,P_l)$ was the highest probability among the L populations evaluated.

The *a priori* probabilities of a genotype belonging to a given population (π_l) were defined proportional to the size of the populations (n_l/N), where $N = \Sigma_{l=1}^{L} n_l$, which corresponds to the total of clones evaluated. Thus, in the CHR sample, the values $\pi_1 = 0.5615$ were considered, for Conilon; $\pi_2 = 0.1462$, for Intervarietal Hybrids; and $\pi_3 = 0.2923$, for Robusta. In the CR sample, $\pi_1 = 0.6577$ and $\pi_3 = 0.3423$.

Based on the allocation of genotypes in one or more populations – in case of tie $\hat{P}(Y_i,P_l)$ - a classification matrix could be established for each of the different k values adopted, which allowed allocation of the clones to be determined by the k-NN method. The classification matrices were represented by the percentages of total correct classification and by population (P_c), as well as the number of clones correctly classified (n_c). P_c values were estimated as the arithmetic complement of apparent error rates (AERs) (Cruz *et al*., 2014), calculated by $P_c = 1 - \Sigma_{l=1}^{L}\pi_l\frac{m_l}{n_l} = 1 - \text{AER}$, where: m_l is the number of genotypes wrongly classified in the *lth* population, since they previously belonged to another. The k-NN analyzes performed for different k values made, it possible to calculate the minimum, average, and maximum values; standard deviations; and the variation coefficients for P_c and n_c.

The method of k-average neighbors (k-AN), follows the similarity reasoning by neighborhood; however, without variations in k values. The allocation of a clone i *(Y_i)* in a population was defined based on the average of the Euclidean distances $\bar{D}_l$ between Y_i and the clones belonging to one of the populations, discarding the estimates of distances in which $i = i'$. Thus, this average value was calculated by: $\bar{D}_l = \frac{\Sigma_{i'=1}^{n_l} d_{ii'}}{n_l}$, where $\bar{D}_l$ is the

average dissimilarity of clone Y_i to be classified for each population; $d_{ii'}$ refers to the Euclidean distance between Y_i and $Y_{i'}$ of the *lth* population.

For the lowest value of $\bar{D}_l$, the investigated clone is allocated to this *lth* population, and a tie is possible. From these classifications, a confusion matrix was also established, with P_c and n_c values.

Although the P_c values provide precision in the allocation of clones in the respective populations, it is not known whether these values represented acceptable levels. For this issue, the criterion of maximum chance was adopted to compare the precision of joint and individualized classification by population, and the value of P_c should be at least 1.25 greater than the chance of being allocated to all clones, randomly, in the population most likely (Hair *et al*., 2005)

To check whether the classification pattern was in line with the expected genetic divergence, principal component analysis was conducted and preliminary tests for multivariate normality and homoscedasticity were performed using the Past 3.20 software (Hammer *et al*., 2001). Discriminant analyzes were performed using the Genes program version 1990.2017.26 (Cruz, 2016).

RESULTS AND DISCUSSION

In the multivariate tests presented in Table 1, only the group Intervarietal Hybrids met the assumption of normality. The pairs of populations Conilon × Hybrids and Robusta × Hybrids can be represented by a common variance and covariance matrix, and not the three populations together.

Given this scenario, the discriminant analysis of the k-nearest neighbors (k-NN) and the k-average neighbors (k-AN) was performed as an analytical alternative in relation to attempts at data transformation to meet the above mentioned assumptions.

As shown in Table 2, the mean percentages of correct classification (Pc) considering the evaluation of both the clones jointly and by population were generally higher for the CR dataset – which includes the botanical varieties Conilon and Robusta and excludes Intervarietal Hybrids – than for the CHR set, for both methods of discriminant analysis employed.

For the dataset with the three populations included, the mean (79.62%), minimum (77.69%) and maximum (81.54%) values of P_c by the method of the k-nearest neighbors exceeded the value of P_c (70.00% and n_c = 91 clones) of the k-AN method, and reflected differences between methods in terms of the number of classification hits from ten to fifteen clones, depending on the k value adopted by the k-NN. However, the correct classification for the hybrid clones was much higher in the k-AN method (P_c = 73.68% and corresponded to n_c = 14 hits) than the mean value obtained by the k-NN technique (P_c = 16.08% and n_c = 3.06, only). For the Conilon clones, the classification errors were smaller by the k-NN method (P_c = 96.88%) (Table 2).

When only Conilon and Robusta populations were evaluated (CR data), the levels of correct answers (P_c) were more than 10% higher than the data with intervarietal hybrids included (CHR). The classification rates of the clones were similar for both methods, whose P_c value = 94.60% for the k-AN discriminant analysis was close to the maximum value obtained by the k-NN method (P_c = 95.50%, with k = 3). The Conilon population had its clones very well classified, regardless of the method.

The vast majority of Robusta clones were correctly classified; however, the classification rates were lower than that of the botanical Conilon variety (Table 2). This is consistent with the genetic diversity of the populations, which was higher within the population of Robusta clones and lower within the Conilon and Hybrid groups (Table 1). Of the accessions that make up the germplasm bank of Embrapa Rondônia, the Robusta group has wider molecular diversity (Souza *et al*., 2013). In addition, the Conilon and Robusta groups were the most divergent, and between the Conilon and Intervarietal Hybrids groups had greater genetic similarity (Table 1), as observed by Oliveira *et al*. (2018), and this provided higher rates of classification error in the CHR set compared to the CR data.

Table 1: The Asymmetry and Kurtosis tests of Mardia analyzed population multivariate normality and the Homogeneity test analyzed the variance and covariance matrices (Box M) among ten agronomic characteristics evaluated in the three populations of *Coffea canephora*; and the prediction of intra and interpopulation genetic diversity (Euclidean distance)

Populations	Mardia Multivariate Test		Homogeneity between of (co)variant matrices	Genetic Diversity
	Asymmetry	Kurtosis		
Conilon	332.20*	2.14*	294.95**	0.58
Hybrid	75.85ns	-1.57ns		0.56
Robusta	219.50*	3.74*		0.72
Conilon × Hybrid	-	-	86.60ns	0.63
Conilon × Robusta	-	-	296.36**	0.88
Robusta × Hybrid	-	-	110.63ns	0.83

* ($P < 0.05$) and ns ($P > 0.05$), for Mardia asymmetry and kurtosis tests and for Box M test.

When the k-NN analysis was performed, 25 of the 37 k values adopted for the CR dataset achieved 100% correct classification for the Conilon clones, and 29 hits were obtained for Robusta clones (P_c > 76.00% hits).

The maximum value of 106 hits for the CHR set was reached when k was equal to 3, 5, and 6. When k = 1, the number of classification hits was 103 clones, and using k = 18, 102 hits out of 130 evaluated clones were obtained. For the CR set, the value of 106 hits occurred when k = 3, with 100% and 86.84% of correct classifications for Conilon and Robusta clones, respectively (data not shown).

From the low estimates of standard deviation and coefficient of variation related to P_c and n_c, presented in Table 2, the classifications changed very little with the k variant. The literature has no clear definition of this. Khattree & Naik (2000) report that in studies, especially with large samples, the choice of the k value is irrelevant.

In rice (Zhang *et al*., 2005) and sweet potatoes (Machro & Labonte, 2010), using data from molecular markers, k = 1 was adopted, which made the method more practical, as the decision to allocate the genotype in one of the populations is based only on the definition of a single nearest neighbor – with no chance of a tie in the classification of the genotype and weights do not need to be defined for the probabilities *a priori*.

In cassava, Oliveira *et al*. (2012) arbitrarily established k = 3. Studies that classified seven wheat varieties (Dheer *et al*., 2019) and eight rice varieties (Dheer & Singh, 2019) from quantitative agronomic variables found that k = 20, was the neighborhood that promoted the most accurate classifications, among the variants of k = 3 to 100.

To verify if the accuracy in the classifications was greater than a percentage obtained by chance, both for clones evaluated together and by population, the criterion of maximum chance was applied, comparing them to the P_c values. The accuracy of the classifications, or proportions of correct answers, must be greater than at least 70.20% for the CHR dataset and greater than 82.21% for the CR data, considering that this probability corresponds to the percentage of clones correctly classified if all were allocated to the group with the highest probability of occurrence, plus 25% over this percentage (Hair *et al*., 2005).

For CHR data, the mean, minimum, and maximum P_c value (79.62, 77.69, and 81.54%, respectively) obtained by k-NN exceeded the stipulated "maximum chance", which did not occur with the method of the k-average neighbors. In addition, the P_c value for the Hybrid type in the k-NN method and for the Conilon type in the k-AN technique, were also less than 70.20% (Table 2). In the CR data, for all k variants, the maximum chance criterion was exceeded,

Table 2: Percentage of correct classification (P_c) and number of clones correctly classified (n_c) of *Coffea canephora*, in their respective botanical varieties Conilon and Robusta and, in Intervarietal Hybrids, from the non-parametric discriminant analyzes of the k-nearest neighbors (k-NN) and k-average neighbors (k-AN)

Method	Measurement[&]	Dataset[#]			
		CHR		CR	
		P_c[$]	n_c[$]	P_c	n_c
k-NN	Mean	79.62	103.50	91.58	101.69
	Minimal	77.69	101.00	88.29	98.00
	Maximum	81.54	106.00	95.50	106.00
	Standard deviation	1.19	1.54	1.44	1.60
	Coefficient of variation (%)	1.49	1.49	1.58	1.57
k-AN	-	70.00	91.00	94.60	105.00
	Population	P_c	n_c	P_c	n_c
k-NN	Conilon	96.88	70.72	99.51	72.64
	Hybrid	16.08	3.06	-	-
	Robusta	78.22	29.72	76.46	29.06
k-AN	Conilon	64.38	47.00	98.63	72.00
	Hybrid	73.68	14.00	-	-
	Robusta	78.95	30.00	86.84	33.00

[#]The CHR dataset includes the botanical varieties Conilon and Robusta and the Intervarietal Hybrids, with 73, 19, and 38 clones, respectively, previously allocated, totaling 130 genotypes. The CR dataset includes only the clones of the Conilon and Robusta groups, totaling 111 clones.

[&]The mean, minimum, and maximum values, standard deviation and variation coefficient for P_c and n_c were obtained in the k-NN method from the variations in the k values, ranging from 1 to 18 in the CHR set, and for k varying from 1 to 37 in the CR set. By the k-AN method, mean values of P_c and n_c did not exist, as there is no variant k.

as well as in the k-AN method. The exception was the lower value of P_c = 76.46% for the Robusta population, in the k-nearest neighbors. The adjustment to the maximum chance criterion by an additional 0.25 times in relation to the greater probability of occurrence among the populations aimed to correct the upward bias, that is, to overestimate the predictive accuracy of classification when using the analysis sample (or training) in discriminating procedures (Cruz *et al*., 2014).

After assessing the general adjustment of the discriminant analysis, the allocation of the observations should be individually examined for predictive accuracy, to identify, especially, the poorly classified cases (Hair *et al*., 2005). Thus, the diagnosis was made for the clones that had the highest frequencies of poor classification when performing the analysis via k-NN, especially for the Conilon and Robusta groups (Table 3, Figures 1 and 2), since the programs to improve *C. canephora* coffee has focused its strategy on the hybridization of these botanical varieties.

When considering the frequencies of poor classification of all discriminant k-NN analyzes performed, the Conilon group had the least number of clones classified incorrectly (Figure 1A and D). Only clone C890 had an allocation error greater than half of 18 simulations (assumed k values) in CHR data. Among the Hybrid clones, only H910 had the number of bad classifications – two – less than half of the k simulations (Figure 1B). Among the Robusta, eight clones (R11170, R11191, R13161, R13171, R13281, R81102, R101H, and R160H) stood out with poor classifications of more than half the number of k simulations, for both datasets.

Table 3: Average Euclidean distance of eleven *Coffea canephora* clones in relation to the botanical varieties Conilon (C) and Robusta (R), and Intervarietal Hybrids (H) - and their respectively poor classifications based on the method of the k-nearest neighbors

Clone[#]	Dataset[$]	Average distance from population *			*Clone was misclassified when k[@] was equal to a…				
		C	H	R	1	3	6	18	37
C39	CHR	0.68	**0.61**	0.61	Yes	No	No	No	-
	CR		-		Yes	No	Yes	No	Yes
C890	CHR	0.84	**0.64**	0.87	No	No	No	Yes	-
	CR		-		No	No	Yes	No	No
R10342	CHR	**0.69**	0.71	**0.69**	No	No	No	No	-
	CR				No	No	No	Yes	Yes
R11170	CHR	0.72	**0.65**	0.67	Yes	Yes	Yes	Yes	-
	CR		-		Yes	Yes	Yes	Yes	Yes
R11191	CHR	0.66	**0.54**	0.60	Yes	Yes	Yes	Yes	-
	CR		-		Yes	No	Yes	Yes	Yes
R13161	CHR	0.66	**0.53**	0.82	Yes	Yes	Yes	Yes	-
	CR		-		Yes	Yes	Yes	Yes	Yes
R13171	CHR	**0.67**	0.76	0.68	No	Yes	Yes	Yes	-
	CR		-		No	Yes	Yes	Yes	Yes
R13281	CHR	0.67	**0.55**	0.63	No	No	Yes	Yes	-
	CR		-		No	No	Yes	Yes	Yes
R81102	CHR	**0.64**	0.70	0.74	Yes	No	Yes	Yes	-
	CR		-		Yes	No	Yes	Yes	Yes
R101H	CHR	**0.59**	**0.59**	0.87	Yes	Yes	Yes	Yes	-
	CR		-		Yes	Yes	Yes	Yes	Yes
R160H	CHR	0.79	**0.69**	0.93	Yes	Yes	Yes	Yes	-
	CR		-		Yes	Yes	Yes	Yes	Yes

[#]Clones that had the greatest number of poor classifications, when the k values were varied in the discriminant analysis k-NN. The initials C and R that precede the clone code, represent the botanical varieties Conilon and Robusta, respectively.

[$]The CHR dataset included clones of the two botanical varieties and Intervarietal Hybrids. The CR set only included data on botanical varieties.

*The mean distance values in bold indicate the greatest mean similarity between the clone and the population considered (C, H, or R) for the CHR dataset. The underlined mean distance values indicate the greatest mean similarity between the clone and the population for the CR dataset.

[@]k value = 1 represents the minimum value used in the k-MVP method; k = 3 and 6, represented the values of nearest neighbors in which the percentage of correct classification (Pc) was the highest for the CR and CHR sets, respectively; k = 18 and 37, represented the maximum k values used in the k-NN method for the CHR and CR sets, respectively.

Principal components analysis elucidated that the most poorly classified Conilon and Robusta clones (Figure 1) were located closer to the center of the graphic dispersion, at the intersection with the genotypes of the other botanical variety (Figure 2). The genetic distance of clone C39, although more similar to Robusta or Hybrid populations (0.61), varied its classification for the different k values (Table 3).

By the k-AN method, clone C890 was more similar to the Hybrid population (0.64) when evaluated in the CHR dataset. But with the removal of hybrid clones (CR data), it was allocated to its previously designated population. The Robusta clones R101H, R160H, and R13161 were poorly classified for all datasets and methods of analysis, sometimes resembling the Hybrids group and sometimes resembling the Conilon clones, as shown in Table 3. Clones R11170, R11191, R13171, R13281, and R81102 also fluctuated considerably in terms of their classifications under the different k values (Figure 1 and Table 3).

The Robusta genotypes, R13161 and R13281, poorly classified by discriminant analyzes, were selected in the work by Oliveira et al. (2018) to compose a diallel scheme with ten clones of each botanical variety, based on their genetic divergences and their phenotypic values for bean production. The authors also stated that these same 38 Robusta genotypes, which are distant from their centroid indicated polymorphisms not characteristic of the botanical variety or, even a mixture between varieties; therefore, these are not recommended for hybridization with the Conilon group.

Among the selected clones, Oliveira *et al.* (2018) also pointed out that the Conilon 890 and Robusta 13161 clones were among those chosen with the greatest potential to gain from recombination of selected matrices in partial diallel scheme, but they did not propose a cross between them.

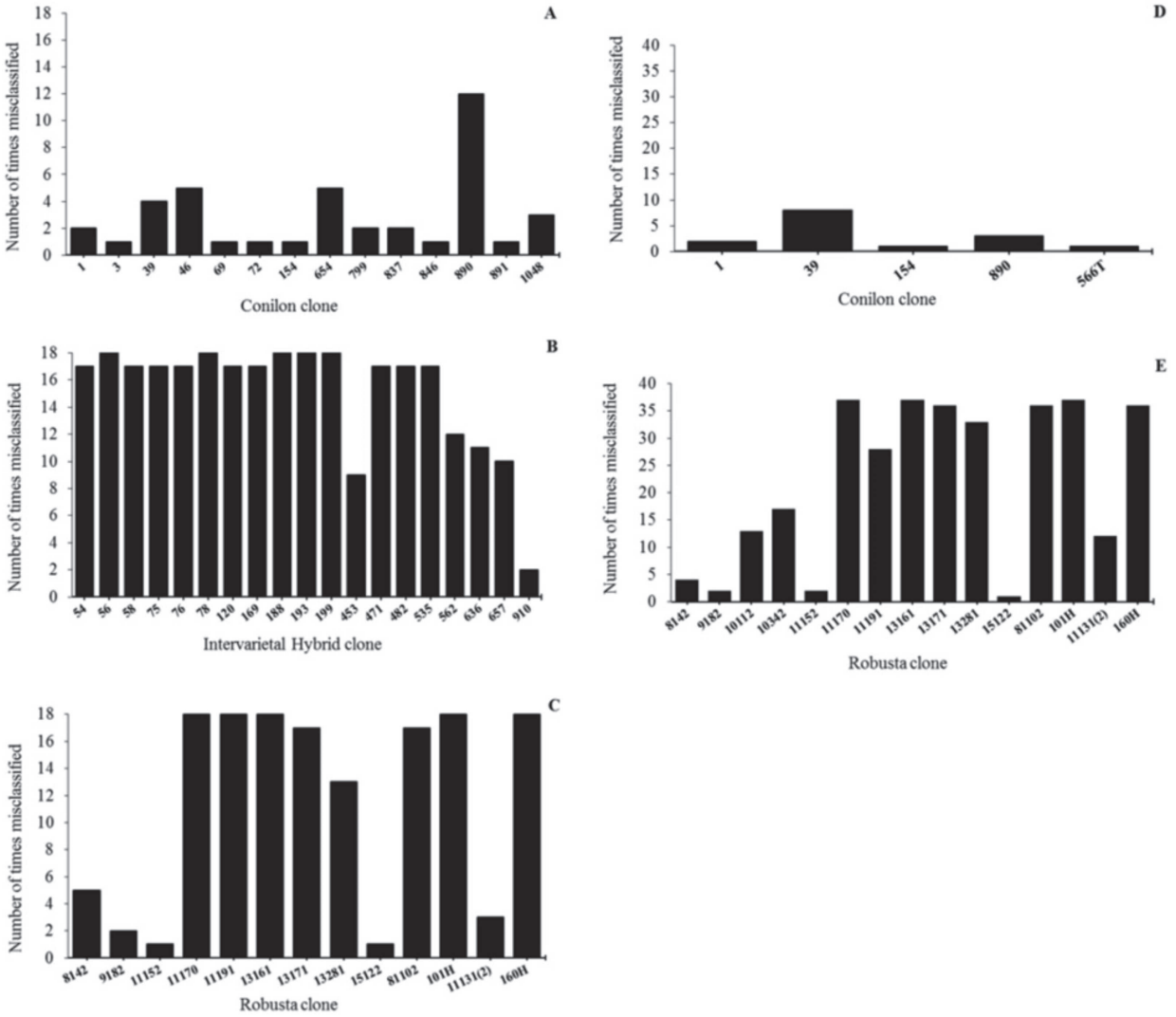

Figure 1: Poorly classified clones and total number of bad ratings obtained in allocations made with the discriminant analysis k-NN for different values of k, from the dataset in which the Conilon (A), Intervarietal Hybrids (B), and Robusta (C) populations were considered as well as the dataset whose populations were only the botanical varieties Conilon (D) and Robusta (E).

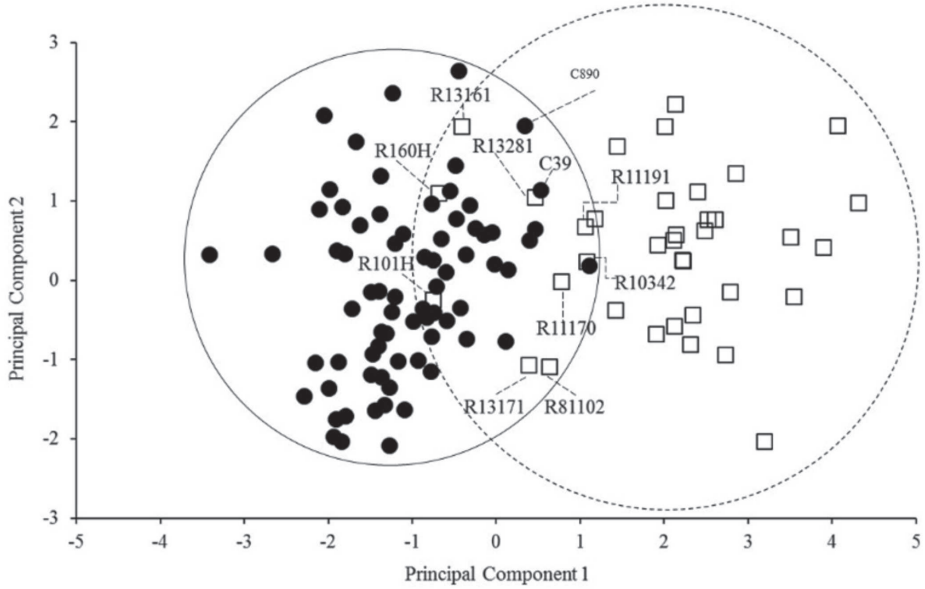

Figure 2: Graphical dispersion of the first two principal components for 111 coffee clones (*Coffea canephora* Pierre ex Froenher), established by the correlation matrix and the linear combination of ten agronomic characteristics. These two principal components accumulated 52.51% of the total variation. There were 73 clones grouped in the botanical variety conilon – C – (Ï%) and 38 clones in the variety Robusta – R – (¡%). The identified clones refer to those that were most poorly classified in the discriminant k-NN analyzes.

Discriminant analyzes based on neighborhood (k-NN and k-AN) proved to be a useful tool for decision-making related to intervarietal crosses, whose results added to a breeder's experience in the field can help to differentiate and better classify the genetic materials. The classifications occurred as expected and the results obtained correctly reflected the genetic diversity and divergence between clones of the botanical varieties and intervarietal hybrids of *C. canephora* coffee.

Finally, the efficiency of the discriminant analyzes is mainly associated with the degree of differentiation of the populations and the quantity and quality of the variables used (Cruz *et al*., 2014). The production of processed beans is one of the characteristics that most contributes to the divergence between coffee genotypes (Guedes *et al*., 2013; Giles *et al*., 2018) and in the present study, together with the other agronomic characteristics, proved to be good variables to discriminate between the Conilon and Robusta groups.

CONCLUSIONS

The k-NN and k-AN discriminant analyzes efficiently classified *C. canephora* coffee genotypes, in the botanical varieties Conilon and Robusta, taking into account the expression of a set of quantitative, agronomic characteristics and the processed coffee yield.

The k-AN discriminant analysis was able to better discriminate intervarietal hybrids from the group clones Conilon, since this botanic variety is more similar to hybrids. These classification methods can assist coffee breeders in directing hybridizations between different parents, discarding atypical polymorphisms, characteristic of hybrid plants, which are not part of the botanical varieties.

ACKNOWLEDGEMENTS, FINANCIAL SUPPORT AND FULL DISCLOSURE

We thank the Brazilian Coffee Research and Development Consortium for supporting the project "Genetic Improvement of Coffee Conilon and Arabica for Productivity and Drink Quality in the Western Amazon". The National Council for Scientific and Technological Development (CNPq) provided financial support and granting the scholarship to the first author. The Amazonas State Research Support Foundation (FAPEAM) for granting the scholarship to the fifth author. The Higher Education Personnel Improvement Coordination (CAPES) for supporting the Post-graduate Programs. We declare that there is no conflict of interest.

REFERENCES

Anderson TW (1958) An introduction to multivariate statistical analysis. New York, John Wiley & Sons. 242p.

Bannayan M & Hoogenboom G (2009) Using pattern recognition for estimating cultivar coefficients of a crop simulation model. Field Crops Research, 111:290–302.

Box GEP (1949) A general distribution theory for a class of likelihood criteria. Biometrika, 36:317–346.

Charrier A & Berthaud J (1988) Principles and methods of coffee plant breeding: *Coffea canephora*. In: Clarcke RJ & Macrae R (Ed.) Coffee, Volume 4: Agronomy. London, Elsevier Applied Science. p.167-198.

Cruz CD, Carneiro PCS & Regazzi AJ (2014) Modelos biométricos aplicados ao melhoramento genético. 3rd ed. Viçosa, UFV. 668p.

Cruz CD, Ferreira FM & Pessoni LA (2020) Diversidade genética baseada em informações moleculares. In: Cruz CD, Ferreira FM & Pessoni LA (Ed.). Biometria aplicada ao estudo da diversidade genética. Visconde do Rio Branco, Suprema. p.321-428.

Cruz CD (2016) Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum Agronomy, 38:547-552.

Davis A, Govarerts R, Bridson M & Stoffelen P (2006) An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). Botanical Journal of the Linnean Society, 152:465-512.

Dheer P, Purshottam & Singh V (2019) Classifying wheat varieties using machine learning model. Journal of Pharmacognosy and Phytochemistry, 8:47-49.

Dheer P & Singh RK (2019) Identification of indian rice varieties using machine learning classifiers. Plant Archives, 19:155-158.

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

Fix E & Hodges JL (1951) Nonparametric discrimination: consistency properties. Randolf Field, School of Aviation Medicine. 21p.

Fix E & Hodges JL (1952) Discriminatory analysis - nonparametric discrimination: small sample performance. Randolf Field, School of Aviation Medicine. 43p.

Fonseca AFA, Sediyama T, Cruz CD, Sakiyama NS, Ferrão RG, Ferrão MAG & Bragança SM (2004) Discriminant analysis for the classification and clustering of robusta coffee genotypes. Crop Breeding and Applied Biotechnology, 4:285-289.

Ferrão RG, Ferrão MAG, Fonseca AFA, Mistro JC, Volpi PS, Verdin Filho AC, Mauri AL & Lani JA (2015) Cultivares. In: Fonseca A, Sakiyma NS e Borém A (Ed.) Café Conilon do plantio à colheita. Viçosa, UFV. p.29-49.

Giles JAD, Partelli FL, Ferreira A, Rodrigues JP, Oliosi G & Silva FHL (2018) Genetic diversity of promising 'conilon' coûee clones based on morpho-agronomic variables. Anais da Academia Brasileira de Ciências, 90:2437-2446.

Guedes JM, Vilela DJM, Rezende JC, Silva FL, Botelho CE & Carvalho, SP (2013) Divergência genética entre cafeeiros do germoplasma Maragogipe. Bragantia, 72:127-132.

Hair JF, Anderson RE, Tatham RL & Black WC (2005) Análise multivariada de dados. 5th ed. Porto Alegre, Bookman. 593p.

Hammer Ø, Harper DAT & Ryan PD (2001) PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica, 4:1-9.

Khattree R & Naik DN (2000) Applied multivariate statistical with SAS Software. Cary, SAS Institute Inc. 338p.

Marcolan AL, Ramalho AR, Mendes AM, Teixeira CAD, Fernandes CF, Costa JNM, Vieira Júnior JR, Oliveira SJM, Fernandes SR & Veneziano W (2009) Cultivo dos cafeeiros conilon e robusta para Rondônia. 3rd ed. Porto Velho, Embrapa Rondônia/Emater-RO. 61p.

Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika, 57:519-530.

Mcharo M & LaBonte DR (2010) Multivariate selection of AFLP markers associated with â-carotene in sweetpotatoes. Euphytica, 175:123-132.

Musoli P, Cubry P, Aluka P, Billot C, Dufour M, De Bellis F, Pot D, Bieysse D, Charrier A & Leroy T (2009) Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda. Genome, 52:634-646.

Nielsen LR, Cowan RS, Siegismund HR, Adsersen H, Philipp M & Fay MF (2003) Morphometric, AFLP and plastid microsatellite variation in populations of *Scalesia divisa* and *S. incise* (Asteraceae) from the Galápagos Islands. Botanical Journal of the Linnean Society, 143:243-254.

Oliveira MVC, Baliza DP, Souza GA, Carvalho SP & Assis LHB (2012) Caracterização de clones de mandioca utilizando marcadores microssatélites. Revista Ciência Agronômica, 43:170-176.

Oliveira LNL, Rocha RB, Ferreira FM, Spinelli VM, Ramalho AR & Teixeira AL (2018) Selection of *Coffea canephora* parents from the botanical varieties Conilon and Robusta for the production of intervarietal hybrids. Revista Ciência Rural, 48:1-7.

Rocha RB, Teixeira AL, Ramalho AR & Souza FF (2015) Melhoramento de *Coffea canephora* – Considerações e Metodologias. In: Marcolan AL & Espindula MC (Ed.). Café na Amazônia. Brasília, Embrapa. p.101-122.

Salari N, Shohaimi S, Najafi F, Nallappan M & Karishnarajah I (2014) A Novel Hybrid Classification Model of Genetic Algorithms, Modified k-Nearest Neighbor and Developed Backpropagation Neural Network. PLoS ONE, 9:1-50.

Silverman BW & Jones MC (1989) E. Fix & J. L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation. International Statistical Review, 57: 233-247.

Souza FF, Caixeta ET, Ferrão LFV, Pena GF, Sakiyama NS, Zambolim EM, Zambolim L & Cruz CD (2013) Molecular diversity in *Coffea canephora* germplasm conserved and cultivated in Brazil. Crop Breeding and Applied Biotechnology, 13:221-227.

Zhang N, Xu Y, Akash M, McCouch S & Oard JH (2005) Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. Theoretic Applied Genetics, 110:721–729.