DAIANA SALLES PONTES

SELEÇÃO DE VARIÁVEIS NO ESTUDO DA DIVERSIDADE GENÉTICA VIA ANÁLISE DE PROCRUSTES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
P814s
2016

Pontes, Daiana Salles, 1989-
Seleção de variáveis no estudo da diversidade genética via análise de procrustes / Daiana Salles Pontes. – Viçosa, MG, 2016.
x, 37f. : il. (algumas color.) ; 29 cm.

Orientador: Cosme Damião Cruz.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Análise multivariada. 2. Diversidade genética. 3. Café - Melhoramento genético - Métodos estatísticos. 4. Biometria. 5. Genética quantitaiva. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.535

DAIANA SALLES PONTES

SELEÇÃO DE VARIÁVEIS NO ESTUDO DA DIVERSIDADE GENÉTICA VIA ANÁLISE DE PROCRUSTES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 24 de fevereiro de 2016.

_____
Paulo Roberto Cecon
(Coorientador)

_____
Pedro Crescêncio Souza Carneiro

_____
Cosme Damião Cruz
(Orientador)

“Para quem tem fé, o céu não é o limite.”

Autor desconhecido

"Quem caminha sozinho pode até chegar mais rápido, mas aquele que vai acompanhado, com certeza vai mais longe."

Clarice Lispector

“Não haverá borboletas se a vida não passar por longas e silenciosas metamorfoses.”

Rubem Alves

“Sabemos que todas as coisas cooperam para o bem daqueles que amam a Deus...”

Romanos 8, 28

AGRADECIMENTOS

A Deus pelo dom da vida, pela fé e força, essenciais na superação de todos os desafios.

Aos meus pais, especialmente, à minha mãe Maria pelo carinho e zelo dedicados a minha irmã e a mim, pelo apoio e encorajamento. Você é o meu maior exemplo de coragem e força.

À minha família, em especial, à tia Linda, por seu exemplo de amor e fé, pelos ensinamentos e conselhos. Você é mais do que madrinha, é minha segunda mãe.

Ao meu querido esposo e principal incentivador, Jadimilson, pelo amor e carinho, companheirismo, compreensão e paciência. Essa conquista não é só minha, é nossa. Obrigada por estar sempre ao meu lado.

À Universidade Federal de Viçosa, particularmente ao Departamento de Estatística que por meio ótimos professores ofereceu um ensino de excelência contribuindo não apenas para minha formação profissional como também para o meu crescimento pessoal.

Ao meu orientador, Cosme Damião Cruz, pelo saber transmitido, pelas conversas e conselhos, pela amizade.

Ao professor e coorientador Moysés pelas valiosas correções.

Ao professor Paulo Roberto Cecon, por ter aceitado o convite para coorientação.

Ao professor Pedro Crescêncio Souza Carneiro, por ter aceitado o convite para participar da banca.

Aos colegas de turma do mestrado, principalmente ao Vitor companheiro de estudos de longa data, obrigada por sua amizade.

Aos colegas do Laboratório de Bioinformática, pela amizade e companheirismo, em particular, Haroldo, Renato e Vinícius pelos conhecimentos compartilhados e que muito contribuíram para a realização deste trabalho.

À secretária da Pós-Graduação Carla Zinato Campos por todo apoio, incentivo, atenção, amizade e empenho em ajudar.

À CAPES pelo apoio financeiro para o desenvolvimento deste trabalho.

BIOGRAFIA

DAIANA SALLES PONTES, filha de Josemar Reginaldo Salles e Maria das Graças Viana Salles, nasceu dia 3 de abril de 1989, em Vitória, Espírito Santo.

Em março de 2007, iniciou o curso de Estatística na Universidade Federal do Espírito Santo, concluindo-o em fevereiro de 2014.

Em março de 2014, ingressou no mestrado *Stricto Sensu* no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação no dia 24 de fevereiro de 2016.

SUMÁRIO

RESUMO vii

ABSTRACT ix

1. INTRODUÇÃO 1

2. REVISÃO DE LITERATURA 4

2.1. Diversidade Genética no Melhoramento 4

2.2. Café: sua importância e avaliação de caracteres 5

2.3. Análise de Componentes Principais 6

2.4. Estudo da diversidade por meio da projeção de escores 10

2.5. Critérios de descarte de variáveis 11

2.6. Análise de Procrustes 12

3. MATERIAL E MÉTODOS 17

3.1. Dados 17

3.2. Proposta para critério de seleção de variáveis via Análise de Procrustes 17

3.3. Comparação de resultados 19

4. RESULTADOS E DISCUSSÃO 20

4.1. Análise de Componentes Principais 20

4.2. Redução de variáveis para fins de estudo da diversidade genética 22

5. CONCLUSÕES 33

REFERÊNCIAS BIBLIOGRÁFICAS 34

RESUMO

PONTES, Daiana Salles, M. Sc., Universidade Federal de Viçosa, fevereiro de 2016, **Seleção de Variáveis no estudo da Diversidade Genética via Análise de Procrustes.** Orientador: Cosme Damião Cruz; Coorientadores: Moysés Nascimento e Paulo Roberto Cecon.

Para o sucesso de um programa de melhoramento é indispensável que população de trabalho disponha de variabilidade genética de forma que a prática de seleção seja viável. Nesse sentido, a avaliação da divergência genética têm sido de grande importância por fornecerem parâmetros para a identificação de combinações híbridas cujo cruzamento proporcione maior efeito heterótico e maior probabilidade de recuperar genótipos superiores nas gerações segregantes. O estudo sobre diversidade genética elucida relações genéticas, quantifica ou prediz o nível de variabilidade total existente e sua distribuição entre indivíduos, acessos de bancos de germoplasma, linhagens e cultivares ou dentro de populações e espécies. Conhecimento que tem proporcionado, dentre outras coisas, importantes contribuições ao melhoramento genético, ao gerenciamento de bancos de germoplasma e à conservação de recursos genéticos. Assim, o interesse maior, em estudos de caracterização da diversidade genética das espécies vegetais, animais e de microrganismos consiste na identificação de grupos de genótipos similares de forma que a maior diferença entre os grupos formados seja realçada. Para isso, algumas técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas e de agrupamento podem ser utilizadas nesse tipo de estudo. Contudo, de modo geral, tais técnicas ainda exigem a utilização de todas as variáveis para a avaliação dos indivíduos/acessos, o que nem sempre é possível devido ao alto custo ou mesmo o grau de dificuldade envolvido na obtenção de determinadas variáveis. É necessária, portanto, a aplicação de algum método de seleção de variáveis ou de um critério de seleção baseado em alguma técnica analítica, como é o caso do critério apresentado por Jolliffe (1972). Baseado na técnica de componentes principais, esse critério é usualmente utilizado na determinação da importância relativa de caracteres no estudo da diversidade de modo que caracteres de menor importância serão desconsiderados do estudo. Há também outra metodologia baseada em Análise de Procrustes ainda pouco utilizada em estudos de diversidade genética, sobretudo para

este fim, por meio da qual é possível selecionar variáveis com base no padrão de dissimilaridade ou similaridade entre acessos. Desta forma, este trabalho tem por objetivo propor um critério baseado em Análise de Procrustes como nova possibilidade para a seleção de variáveis no estudo da diversidade genética. Em seguida, comparar o critério apresentado com o critério proposto por Jolliffe (1972) - ambos os critérios estabelecidos por meio do uso de componentes principais. Para elucidar a teoria apresentada, foram consideradas informações de 40 acessos de café Conilon avaliados em Sooretama/ES no ano 2000 segundo 16 caracteres agronômicos. As técnicas apresentadas neste trabalho demonstram ser vantajosas na seleção (ou descarte) de variáveis proporcionando relevante contribuição para os estudos sobre diversidade genética. A técnica apresentada, baseada em análise de Procrustes, torna-se uma alternativa mais eficaz do que o uso do critério de Jolliffe (1972) para fins de estudo da diversidade genética.

ABSTRACT

PONTES, Daiana Salles, M. Sc., Universidade Federal de Viçosa, February, 2016, **Variables Selection in the study of Genetic Diversity by Procrustes Analysis.** Adviser: Cosme Damião Cruz; Co-Advisers: Moysés Nascimento and Paulo Roberto Cecon.

For the success of a breeding program, it is essential the existence of genetic variability in the base population so that selection practice can be feasible. In this context, the evaluation of genetic diversity have been very important because gives parameters to identify hybrid combinations whose cross provides the greatest heterotic effect and the highest probability of recovering superior genotypes in segregating generations. The study about genetic diversity aims to elucidate genetic relationships, quantify or predict the level of total variability existing and its distribution among genotypes, accessions of germplasm banks, landraces and cultivars or within populations and species. This knowledge have provided, among other things, important contributions to genetic breeding, the management of germplasm banks and the conservation of genetic resources. So, the main interest in characterization studies of genetic diversity of vegetable, animals and microorganisms species consists of identifying groups of similar genotypes, so that the largest difference between groups formed is highlighted. Some multivariate techniques such as discriminant analysis, principal component analysis, coordinate analysis and cluster analysis can be used in these studies. However, in general, these techniques still require the use of all variables for evaluation of genotypes - accessions, which is not always possible due to high cost or even the degree of difficulty involved in obtaining some variables. Thus, it is required the application of a variables selection method or a selection criterion based in an analytical technique, such as the criterion presented by Jolliffe (1972). Based on principal component technique, this criterion is commonly used to determine the relative importance of traits in diversity study, so that less important traits will be eliminated of the study. There is also another methodology based on Procrustes analysis still little used in genetic diversity studies, particularly for this purpose. By this methodology is possible to select variables based on the pattern of dissimilarity or similarity among accessions. Thus, this work aims to propose a criterion based on Procrustes analysis as a new possibility to select

variables in genetic diversity study. Then, compare the criterion proposed with the criterion presented by Jolliffe (1972) - both criteria established through the use of principal components. To elucidate the theory presented, it was used the information from 40 accessions of Conilon coffee evaluated in Sooretama/ ES in the year 2000 regarding to 16 agronomic traits. The techniques presented in this work demonstrated to be advantageous in the selection (or discarding) of variables and consequently providing relevant contributions for genetic diversity study. The technique presented, based on Procrustes analysis, it becomes an alternative more effective than the criterion of Jolliffe (1972) for genetic diversity study.

1. INTRODUÇÃO

Para o sucesso de um programa de melhoramento é indispensável que exista variabilidade genética na população de trabalho viabilizando, assim, a prática da seleção. E, para a formação de população-base, os melhoristas têm recomendado o intercruzamento entre cultivares superiores e divergentes (FERRÃO, 2004).

Os estudos sobre divergência genética desempenham importante papel em um programa de melhoramento, pois fazem parte da sua fase inicial, denominada pré-melhoramento, por meio do qual é possível viabilizar a regeneração, caracterização, conservação e exploração da variabilidade disponível na população de trabalho. Ainda nessa fase é que se obtêm informações acerca dos candidatos a progenitores mais divergentes que terão maiores probabilidades de promover resultados satisfatórios em um programa de melhoramento.

O interesse em estudos de caracterização da diversidade genética das espécies vegetais, animais e de microrganismos reside na identificação de grupos de genótipos similares de forma que a maior diferença entre os grupos formados seja realçada. Para isso, algumas técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas e de agrupamento podem ser utilizadas nesse tipo de estudo.

É possível retratar tais técnicas sobre duas abordagens: baseadas em análise de agrupamento ou cluster e baseadas em dispersão gráfica. Na análise de agrupamento, como é o caso do método de Tocher e dos métodos SAHN (do inglês, sequencial, agglomerative, hierarchical, nonoverlapping), dentre os quais podemos ressaltar o método do vizinho mais próximo e o método do vizinho mais distante, é necessário definir, *a priori*, a medida de dissimilaridade (distância) a ser utilizada na formação da matriz de distâncias entre pares de indivíduos. Nas técnicas baseadas em dispersão gráfica, consideram-se as posições relativas de acessos em gráficos bi ou tridimensionais como, por exemplo, a técnica de componentes principais e a da projeção de distâncias 2D ou 3D. Para Cruz et. al. (1994), a escolha pela abordagem mais adequada varia, em geral, segundo a precisão desejada, a facilidade de análise e a natureza do material, que é determinada pela forma com que os dados foram obtidos.

A diversidade genética existente entre e/ou dentro de populações pode ser mensurada pela diferença entre os valores fenotípicos de seus indivíduos que, por sua vez, são obtidas no experimento de campo em que um número considerável de características morfológicas, agronômicas, dentre outras, da espécie/cultivar estudada são medidas. Ao passo que, se uma coleção de indivíduos avaliada em dado experimento estiver contida em uma população ou em um banco de germoplasma, ela poderá ser reavaliada em estudos futuros para fins diversos. Porém, em alguns casos, pode ser de interesse do melhorista avaliar uma quantidade menor de características do que as registradas no banco de germoplasma, seja pelo alto custo, precisão da informação ou mesmo o grau de dificuldade envolvido na obtenção de determinada(s) variável(is).

Visto que a variabilidade é um fator de extrema importância no desenvolvimento de novas variedades de uma espécie assim como na conservação dos recursos genéticos da mesma, é responsabilidade do melhorista investigar até que ponto a exclusão de um ou outro caráter do estudo irá afetar a variabilidade presente no grupo de acessos avaliados. Desta forma, é recomendável a aplicação de algum método de seleção de variáveis ou de um critério de seleção baseado em alguma técnica analítica para seleção e, ou, descarte de variáveis que é geralmente estabelecido por meio da determinação da importância relativa de cada uma delas.

Quanto à determinação da importância relativa de caracteres no estudo da diversidade destacam-se o critério proposto por Singh (1981), baseado na distância D^2 de Mahalanobis, e o critério baseado na técnica de componentes principais tal como preconizado por Jolliffe (1972). Ambos os critérios têm sido amplamente utilizados no descarte de variáveis, porém, o seu uso está restrito à escolha inicial do pesquisador quanto ao método de agrupamento utilizado no estudo da diversidade genética, já que ambos possuem abordagens distintas. Enquanto o primeiro critério avalia o padrão de agrupamento por meio da análise de agrupamento, o segundo avalia por meio da adequação do conjunto de variáveis em prover informações em dispersão gráfica no espaço bi ou tridimensional.

Na literatura encontram-se vários trabalhos sobre o estudo da diversidade para melhoramento de culturas de grande impacto socioeconômico das quais podemos citar a soja, o feijão, o milho e o café. Destas, a cafeicultura está entre as de maior importância

econômica e social. O Brasil é, atualmente, o maior produtor mundial de café do mercado internacional e o segundo mercado consumidor. Devido à relevância socioeconômica dessa cultura, os programas de melhoramento de café têm investido no desenvolvimento de variedades com qualidade cada vez mais elevada, que associe alta produtividade com boa qualidade do produto e características que atendam o produtor, a indústria e o consumidor. O alcance desses objetivos exige uma série de informações acerca das características biológicas intrínsecas da espécie a ser melhorada que, às vezes, envolve algumas variáveis de difícil obtenção seja por limitações de recursos humanos, financeiros ou técnicos. Nesses termos, justifica-se a disponibilidade de métodos e/ou critérios para a seleção de caracteres no estudo da diversidade genética dentro de um programa de melhoramento do cafeeiro bem como para outras culturas de importância socioeconômica.

Além dos critérios de descarte de variáveis já mencionados, há também outra metodologia baseada em *análise de procrustes* ainda pouco utilizada em estudos de diversidade genética, sobretudo para este fim, por meio da qual é possível selecionar variáveis com base no padrão de dissimilaridade ou similaridade entre acessos. Esta metodologia, conforme apresentada por Krzanowski (1987), combina componentes principais e análise de procrustes para determinar o quanto um subconjunto de variáveis representa a estrutura do conjunto de variáveis originais. Mediante esta abordagem, busca-se um subconjunto de variáveis que represente a mesma estrutura do conjunto original de variáveis sem perda de informação, ou ainda que a perda de informação devida à exclusão de algumas variáveis seja mínima de tal forma que o padrão de agrupamento não seja afetado.

Diante do exposto, este estudo tem por objetivo propor um critério baseado em análise de procrustes, conforme retratado por Krzanowski (1987), como nova possibilidade para a seleção de variáveis no estudo da diversidade genética e em seguida, compará-lo com o critério proposto por Jolliffe (1972) - ambos os critérios estabelecidos por meio do uso de componentes principais.

2. REVISÃO DE LITERATURA

2.1. Diversidade Genética no Melhoramento

Para o sucesso de um programa de melhoramento é indispensável que exista variabilidade genética na população de trabalho. E para a formação de população-base os melhoristas têm recomendado o intercruzamento entre cultivares superiores e divergentes (FERRÃO, 2004).

A avaliação da divergência genética em programas de melhoramento têm sido de grande importância por fornecerem parâmetros para a identificação de combinações híbridas cujo cruzamento proporcione maior efeito heterótico e maior probabilidade de recuperar genótipos superiores nas gerações segregantes (CRUZ et. al. 2004; CRUZ, 2005).

PESSONI (2007) esclarece que o estudo da diversidade visa elucidar relações genéticas entre indivíduos ou populações, quantificar ou predizer o nível de variabilidade total existente e sua distribuição entre e/ou dentro do material genético, quer eles sejam indivíduos, acessos de bancos de germoplasma, linhagens, cultivares, populações e espécies. Conhecimento que tem proporcionado, dentre outros aspectos, importantes contribuições ao melhoramento genético para fins de orientação de cruzamentos, ao gerenciamento de bancos de germoplasma e à conservação de recursos genéticos.

O interesse em estudos de caracterização da diversidade genética das espécies vegetais, animais e de microrganismos reside na identificação de grupos de genótipos similares de forma que a maior diferença entre os grupos formados seja realçada. Para isso, algumas técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas, análise de projeção e técnicas de agrupamento podem ser utilizadas nesse tipo de estudo (CRUZ et. al., 2011).

É possível relatar tais técnicas sobre duas abordagens: baseadas em análise de agrupamento (ou cluster) e baseadas em dispersão gráfica. Na análise de agrupamento, como é o caso do método de Tocher e dos métodos SAHN (do inglês, sequencial, agglomerative, hierarchical, nonoverlapping), dentre os quais podemos ressaltar o método do vizinho mais próximo e o método do vizinho mais distante, é necessário definir a priori a medida de dissimilaridade (distância) a ser utilizada na formação da

matriz de distâncias entre pares de indivíduos. Nas técnicas baseadas em dispersão gráfica, consideram-se as posições relativas de acessos em gráficos bi ou tridimensionais como, por exemplo, a técnica de componentes principais e a projeção de distâncias 2D ou 3D. Para Cruz et. al. (1994), a escolha pela abordagem mais adequada varia, em geral, segundo a precisão desejada, a facilidade de análise e a natureza do material, que é determinada pela forma com que os dados foram obtidos.

Na literatura encontram-se vários trabalhos sobre o estudo da diversidade para melhoramento de culturas de grande impacto socioeconômico das quais podemos citar a soja (PELUZIO et. al., 2009; ALMEIDA et. al., 2011; RIGON et. al., 2012), o feijão (COELHO et. al., 2007; BERTINI, 2009), o milho (ROTILI et. al., 2012) e o café (FONSECA et. al., 2006; GUEDES et. al., 2013; ROCHA et. al., 2013).

2.2. Café: sua importância e avaliação de caracteres

Das diferentes atividades ligadas ao negócio agrícola em nível mundial, a cafeicultura está entre as de maior importância econômica e social. O Brasil é, atualmente, o maior produtor mundial de café (responsável por 30% do mercado internacional, o equivalente à soma da produção dos outros seis maiores países produtores) e o segundo mercado consumidor, precedido pelos Estados Unidos (ABIC, 2016). Além disso, a importância da cultura café no Brasil deve-se fundamentalmente pela geração de empregos e receitas pela manutenção de famílias no campo, sobretudo nos estados responsáveis pelo maior volume da produção nacional, a saber: Minas Gerais, São Paulo, Espírito Santo, Paraná, Rondônia e Bahia.

A produção de café no Brasil e no mundo concentrava-se unicamente na espécie *Coffea arabica*. Contudo, devido a um grande surto de ferrugem que afetou os cafezais do sul e leste da Ásia a partir do fim do século XIX, a espécie *Coffea canephora*, que se mostrava resistente à doença, passou a ser alvo de estudos científicos objetivando sua exploração econômica (VAN DER VOSSEN, 1985; CHARRIER; BERTHAUD, 1988, citado por FERRÃO et. al., 2011 pag.37).

Das principais espécies do gênero Coffea descritas por Carvalho (1946), citado por Ferrão (2004), a *Coffea arabica* e a *Coffea canephora* constituem quase todo o café

produzido e comercializado no mundo, contribuindo com cerca de 70% e 30%, respectivamente (FERRÃO, 2004).

De acordo com Chavalier (1929, 1944), citado por Ferrrão (2004), a espécie *Coffea canephora*, conhecida como Robusta, inclui diversas variedades tais como: 'Kouilou', 'Robusta', 'Sankutu', 'Bakaba', 'Niaculi', entre outras. Destas, a 'Kouilou', denominada no Brasil, Conilon é a mais importante em nosso país pelo seu volume de produção e valor industrial. Em função de sua menor acidez e maior quantidade de sólidos solúveis, o café Conilon tem sido amplamente utilizado pela indústria na fabricação dos cafés solúveis, em misturas com o café arábica, chegando a participar com até 50% dos *blends* e também como fonte de alelos favoráveis para resistência à pragas, doenças e nematoides (FERRÃO, 2004; BELING, 2005; FERRÃO et. al., 2011).

E em função da importância dessa cultura para país, os programas de melhoramento de café têm investido no desenvolvimento de variedades com qualidade cada vez mais elevada, que associe alta produtividade com boa qualidade do produto e características que atendam o produtor, a indústria e o consumidor.

O alcance desses objetivos exige a adoção de estratégias de melhoramento mais condizentes com a realidade do programa de forma que seja possível a identificação de genótipos superiores que confiram maiores ganhos genéticos com economia de tempo, esforço e dinheiro (FERRÃO, 2004).

Além disso, devido a características biológicas intrínsecas da espécie ser melhorada, que no caso do café poderíamos citar seu porte elevado, algumas variáveis tornam-se de difícil obtenção seja por limitações de recursos humanos, financeiros ou técnicos. Nesses termos, justifica-se a disponibilidade de métodos e/ou critérios para a seleção de caracteres no estudo da diversidade genética dentro de um programa de melhoramento.

2.3. Análise de Componentes Principais

Proposta originalmente por PEARSON (1901) e aplicada mais tarde em diversas áreas da ciência por HOTELLING (1933, 1936), a técnica dos componentes principais objetiva a simplificação estrutural de variação dos dados por meio da

transformação de um conjunto original de variáveis em outro conjunto, os componentes principais, de dimensões equivalentes.

Segundo Mingoti (2005), seu objetivo principal é o de explicar a estrutura de variância e covariância de um vetor aleatório, composto de *p*-variáveis aleatórias, por meio da construção de combinações lineares das variáveis originais. Estas combinações são chamadas de componentes principais e são independentes entre si. No entanto, deseja-se em geral obter "redução do número de variáveis a serem avaliadas e interpretação das combinações lineares construídas", ou seja, a informação contida nas *p*-variáveis originais é substituída pela informação contida em k ($k<p$) componentes principais não correlacionados. Desta forma, o sistema de variabilidade do vetor aleatório que contém as k componentes principais representará a variabilidade contida nos dados inicias.

Contudo, a qualidade dessa aproximação dependerá do número de componentes mantidos no sistema e é medida através da proporção da variação explicada pelo número de componentes considerados na análise. Espera-se que os primeiros componentes retenham o máximo da informação, em termos de variação total, contida nas variáveis originais. A seguir será descrito, com alguns detalhes, os princípios básicos da técnica conforme descrito por Mingoti (2005).

Seja $X = (X_1, X_2, \dots, X_p)'$ um vetor aleatório com vetor de médias $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$ e matriz de covariâncias Σ_{pxp}. Sejam $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ os autovalores da matriz Σ_{pxp}, com seus respectivos autovetores normalizados $e_1, e_2, \dots, e_p$, isto é, os autovetores e_i satisfazem as seguintes condições:

i) $e_i' e_j = 0$, para todo $i \neq j$;

ii) $e_i' e_i = 1$, para todo $i = 1,2, \dots, p$;

iii) $\Sigma_{pxp} e_i = \lambda_i e_i$, para todo $i = 1,2, \dots, p$;

com $e_i = (e_{i1}, e_{i2}, \dots, e_{ip})'$. Considere o vetor aleatório $Y = O'X$, dado por p combinações lineares das variáveis aleatórias do vetor X, cujo vetor de médias igual a $O'\mu$ e matriz de covariâncias igual a Λ_{pxp}:

$$O_{pxp} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{p1} \\ e_{12} & e_{22} & \dots & e_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1p} & e_{2p} & \cdots & e_{pp} \end{bmatrix} = [e_1\ e_2 \dots e_p] \text{ e } \Lambda_{pxp} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

Definição 1: O *j*-ésimo componente principal da matriz $\Sigma_{pxp}, j = 1,2, \dots, p$, é definido como:

$$Y_j = e_j'X = e_{j1}X_1 + e_{j2}X_2 + \cdots + e_{jp}X_p.$$

Com esperança e variância dadas, respectivamente, por:

$$E[Y_j] = e_j'\mu = e_{j1}\mu_1 + e_{j2}\mu_2 + \cdots + e_{jp}\mu_p$$

$$Var[Y_j] = e_j'\Sigma_{pxp}e_j = \lambda_j$$

sendo $Cov[Y_j, Y_k] = 0,\ j \neq k$. Cada autovalor λ_j representa a variância de um componente principal Y_j. Como os autovalores estão ordenados em ordem decrescente, a primeiro componente é o de maior variabilidade e o *p*-ésimo é o de menor.

Na prática, Σ_{pxp} é desconhecida e precisa ser estimada pela matriz de covariâncias amostral, S_{pxp}, calculada a partir dos dados amostrais como segue:

$$S_{pxp} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}.$$

sendo $s_{ij} = s_{ji},\ j \neq i$, e s_{ii} e s_{ij}a variância e covariância amostral, definidas respectivamente por:

$$s_{ii} = \frac{\sum_{l=1}^{n}(X_{il} - \bar{X}_i)^2}{n-1} \qquad \text{e} \qquad s_{ij} = \frac{\sum_{l=1}^{n}(X_{il} - \bar{X}_i)(X_{jl} - \bar{X}_j)}{n-1}.$$

Assim têm-se $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ os autovalores da matriz S_{pxp}, com seus respectivos autovetores normalizados $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$. Então, o *j*-ésimo componente principal estimado é definido por:

$$\hat{Y}_j = \hat{e}_j'X = \hat{e}_{j1}X_1 + \hat{e}_{j2}X_2 + \cdots + \hat{e}_{jp}X_p.$$

Observação: É comum o uso do termo "a componente principal" por alguns autores, principalmente nos livros de estatística em que se faz referência ao componente principal como uma variável aleatória.

Definição 2: Pelo teorema da decomposição espectral, a matriz de covariâncias S_{pxp} pode ser expressa como:

$$S_{pxp} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{e}_j \hat{e}_j{}'.$$

Definição 3: A importância relativa do *j*-ésimo componente estimado, dada pela proporção da variância total de *X* que ele explica, é definida como:

$$\frac{Var[\hat{Y}_j]}{Variância\ Total\ Estimada\ de\ X} = \frac{\hat{\lambda}_j}{traço(S_{pxp})} = \frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}$$

Definição 4: A correlação estimada entre o *j*-ésimo componente principal amostral e a variável aleatória $X_i, i = 1,2, \ldots, p$ é dada por:

$$r_{\hat{Y}_j, X_i} = \frac{\hat{e}_{ji}\sqrt{\hat{\lambda}_j}}{s_{ii}}.$$

Além da correlação que mede o quão forte é a associação linear entre o componente principal e cada variável, há também uma outra medida, conforme relata Cruz *et al.* (2011), denominada grau de distorção, $1 - \alpha$, que capta a distorção provocada pela técnica de componentes principais ao se fazer uma simplificação do espaço p-dimensional para o *k*-dimensional ($k < p$) .

Definição 5: O valor de α é dado pela relação matemática entre as estimativas das distâncias gráficas (ou distâncias euclidianas baseadas nos escores dos *k* primeiros componentes) e a distância euclidiana baseada nos dados originais *p*-dimensional:

$$\alpha = \frac{\Sigma\Sigma_{i<i'} d^2 cp_{ii'}}{p\Sigma\Sigma_{i<i'} d^2_{ii'}},$$

em que $d^2 cp_{ii'}$ é o quadrado da distância euclidiana estimado a partir dos escores de *k* componentes principais; e $d^2_{ii'}$ é o quadrado da distância euclidiana média estimado a partir das *p* variáveis originais.

De acordo Cruz *et al.* (2011), admite-se satisfatória a representação gráfica quando os valores de distorção forem inferiores a 20%. Ainda de acordo com os autores, deduz-se que α é a porcentagem da variância explicada pelas k componentes principais.

Os componentes principais são obtidos a partir da matriz de covariâncias e, por conta disso, sofrem influencia das variáveis de maior variância, podendo ocorrer uma discrepância muito acentuada entre essas variâncias causada muitas vezes pela unidade de medida. Uma solução para evitar que este problema influencie nos resultados de agrupamento é a padronização dos dados. Em estatística, a padronização de uma variável refere-se a subtrair de cada valor a sua média e dividi-lo pelo seu desvio-padrão, e assim, os dados padronizados terão médias iguais a 0 e variâncias iguais a 1. Neste caso, os componentes principais são obtidos a partir da matriz de covariâncias das variáveis originais padronizadas, o que equivale a extraírem-se os componentes principais utilizando-se a matriz de correlação R_{pxp} das variáveis originais dadas pela matriz X.

2.4. Estudo da diversidade por meio da projeção de escores

A análise de componentes principais consiste em transformar um conjunto original de variáveis, por exemplo, altura de planta, produtividade, etc., em outro conjunto de dimensão equivalente, mas com propriedades importantes, que são de grande interesse em certos estudos de melhoramento. A viabilidade de sua utilização em estudos de divergência genética dependerá da possibilidade de resumir o conjunto de variáveis originais em poucos componentes, o que significará ter boa aproximação do comportamento dos indivíduos (genótipos), oriundo de um espaço p-dimensional (p: número de caracteres estudados) em um espaço bi ou tridimensional (CRUZ et al., 2011).

Quando a diversidade entre os grupos de acessos é avaliada por meio da dispersão gráfica, inspeciona-se o agrupamento dado pelos escores dos dois primeiros componentes sendo, portanto, desejável que a proporção da variância total acumulada nestes exceda 80%. Nos casos em que esse limite não é atingido nos dois primeiros componentes, a análise é complementada com a dispersão gráfica em relação ao terceiro e ao quarto componentes.

Além de possibilitar o estudo da diversidade genética de um grupo de acessos, a técnica dos componentes principais tem a vantagem de possibilitar a avaliação da importância de cada caráter estudado sobre a variação total disponível entre os genótipos avaliados. O interesse nessa avaliação reside na possibilidade de se descartarem caracteres que são muitas vezes redundantes e que pouco contribuem para a discriminação do genótipo avaliado, reduzindo, dessa forma, mão-de-obra, tempo e custo despendidos na experimentação agrícola (CRUZ *et al.,* 2011). Da mesma forma, se o objetivo do estudo for a caracterização dos recursos genéticos contidos em um banco de germoplasma, pode ser de interesse do melhorista avaliar uma quantidade menor de caracteres do que os registrados, seja pelo alto custo, precisão da informação ou mesmo o grau de dificuldade envolvido na obtenção de determinada(s) variável(is).

É importante ressaltar que a técnica de componentes principais baseia-se apenas nas informações individuais de cada acesso. Caso os dados tenham vindo de um experimento com mais de uma informação por acesso, esta técnica não é adequada sendo recomendado, portanto, o uso da análise por variáveis canônicas que é indicada quando se dispõe de dados com repetição, sendo possível estimar uma matriz de variâncias e covariâncias residuais requerida na análise.

2.5. Critérios de descarte de variáveis

Como já discutido, a variabilidade é um fator de extrema importância no desenvolvimento de novas variedades de uma espécie bem como na conservação dos recursos genéticos da mesma sendo, portanto, responsabilidade do melhorista investigar até que ponto a exclusão de um ou outro caráter do estudo irá afetar a variabilidade presente no grupo de acessos avaliados.

Em geral, o descarte de variáveis é estabelecido por meio da determinação da importância relativa de cada uma delas para o conjunto de dados. Quanto à determinação da importância relativa de caracteres no estudo da diversidade destacam-se o critério proposto por Singh (1981), baseado na distância D^2 de Mahalanobis, e o critério baseado na técnica de componentes principais tal como preconizado por Jolliffe (1972, 1973). Ambos os critérios têm sido amplamente utilizados, porém, o seu uso está restrito à escolha inicial do pesquisador quanto ao método de agrupamento utilizado no estudo da diversidade genética, já que o primeiro critério avalia o padrão de agrupamento por meio da análise de agrupamento e o segundo por meio da adequação

do conjunto de variáveis em prover informações em dispersão gráfica no espaço bi ou tridimensional.

Jolliffe (1972) propôs um critério simples para o descarte de variáveis, baseado na técnica de componentes principais, em que seriam descartadas as variáveis com maiores coeficientes (em valores absolutos) em relação ao componente (autovetor) que possuir a menor variância. Assim, se q variáveis devem ser descartadas de um conjunto contendo p variáveis, descarta-se as que têm os maiores coeficientes (em valores absolutos) nos q componentes correspondentes aos q últimos autovalores.

Há duas possibilidades para se descartar variáveis por meio desse critério. A primeira delas é excluindo as q variáveis de uma só vez a partir de uma única análise de componentes principais. A segunda, retratada por Jolliffe como uma versão backward da primeira, é descartar uma a uma até que as q variáveis sejam excluídas, sendo que a cada exclusão é realizada uma nova análise de componentes principais.

Vale destacar que, quando o descarte simultâneo é considerado (como no primeiro caso citado), pode acontecer de, em um componente de menor variância, o maior coeficiente de ponderação estar associado a uma variável já previamente descartada. De acordo com Cruz *et al.* (2011), nesse caso, tem-se optado por não fazer nenhum outro descarte com base nos coeficientes daquele componente, mas prosseguir a identificação da importância relativa das variáveis no outro componente de variância imediatamente superior.

Nos estudos de diversidade genética não há uma regra que estabeleça um prévio de número de variáveis a serem descartadas e normalmente consideram-se as variáveis de maior peso nos autovetores cujos autovalores sejam nulo(s) ou, de forma mais flexível, aqueles cuja magnitude seja inferior a 0,7 (CRUZ *et al.,* 2011).

2.6. Análise de Procrustes

Além dos critérios de descarte de variáveis já discutidos, há também outra metodologia baseada em Análise de Procrustes ainda pouco utilizada em estudos de diversidade genética, sobretudo para fins de seleção de variáveis com base no padrão de diversidade ou similaridade entre acessos.

A abordagem de Procrustes tem sido utilizada nas mais diversas áreas como engenharia de alimentos (ZENG et. al., 2007; KOBAYASHI e BENASSI, 2010) e medicina (DOUGLAS, 2004), dentre outras. Com isso, sua aplicação tem se mostrado promissora tornando-a de grande interesse em certos estudos, sobretudo no melhoramento (KLINGENBERG, 2003; BRAMARDI et. al., 2005; GARCÍA-PEÑA e DIAS, 2009).

Esta técnica permite a comparação de duas configurações ou dois conjuntos de dados desde que cada linha corresponda ao mesmo indivíduo. Pensando na possibilidade de existir dois conjuntos de vetores que sejam diferentes um do outro, mas estejam definindo o mesmo subespaço, é necessária uma técnica analítica que disponha de uma medida numérica de quanto duas ou mais representações gráficas diferem.

Krzanowski (1987) apresentou uma metodologia que combina componentes principais e análise de procrustes para determinar o quanto um subconjunto de variáveis representa a estrutura do conjunto de variáveis originais (Figura 1). O autor aborda a análise de procrustes sob duas perspectivas: para a seleção de variáveis a partir do algoritmo de eliminação *backward* utilizando como critério de descarte a estatística de procrustes M^2 e para a comparação de padrões de agrupamento provenientes de diferentes subconjuntos de variáveis também por meio da mesma estatística.

Para o entendimento desta técnica considere que para *n* indivíduos tenham sido observadas *p* variáveis (características) $X_1, X_2, \dots, X_p$ caracterizando assim o conjunto de variáveis originais X_{nxp} e que deseja-se selecionar q das p variáveis em que $q < p$ (Figura 1). A partir da análise de componentes principais - ACP ou PCA do termo em inglês, Principal Component Analysis - realizada sobre as matrizes X_{nxp} (denominada matriz de referência) e $\tilde{X}_{nxq}$ (dada pelo subconjunto de q variáveis) têm-se suas respectivas matrizes de escores Y_{nxp} e Z_{nxq}. Contudo, admitindo que a matriz de escores Y_{nxk} (denominada configuração de referência) seja uma boa aproximação da matriz X_{nxp}, padronizada ou não, visto que os k ($k \leq q$) primeiros componentes conservem o máximo da sua informação tem-se sua correspondente $\tilde{Z}_{nxk}$ (dada pelos escores dos k primeiros componentes de Z_{nxq}) como a melhor aproximação de dimensão k para a configuração de dimensão q definida por $\tilde{X}_{nxq}$.

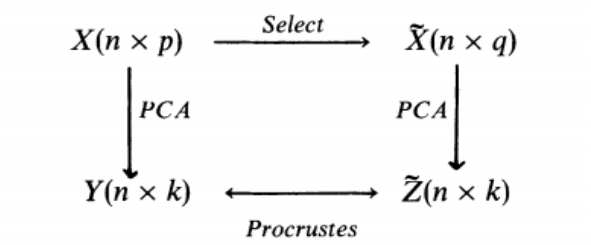

Figura 1: Diagrama ilustrativo da análise de procrustes dada por Krzanowski (1987).

Como devidamente esclarecido por Guedes e Ivanqui (1998), se a verdadeira dimensão dos dados é de fato k, então Y_{nxk} pode ser vista como a verdadeira configuração, enquanto $\tilde{Z}_{nxk}$ é a aproximação correspondente da configuração baseada em somente q variáveis. A discrepância existente entre a verdadeira configuração (Y_{nxk}) e a configuração obtida pelo subconjunto de q variáveis ($\tilde{Z}_{nxk}$) é na verdade o resíduo produzido pela perda de informação devido à exclusão de $p - q$ variáveis ($k \leq q$) e é obtida por meio da Análise de Procrustes.

Logo, a análise de procrustes busca minimizar o traço da matriz de quadrados da diferença entre as duas configurações num espaço Euclidiano multivariado:

$$Min\{traço[(Y - \tilde{Z})(Y - \tilde{Z})']\}. \qquad (1)$$

A solução da equação (1) é obtida em duas etapas cujos passos são ilustrados na Figura 2. Considere os triângulos Y: $A - B - C$ e $\tilde{Z}$: $a - b - c$ representando duas configurações num espaço bidimensional com tamanho, localização e orientação diferentes (Figura 2a). Primeiro efetua-se a translação/centralização e escala/dilatação em Y e $\tilde{Z}$ (Figura 2b), tal que $Y = (I - P)Y/\sqrt{tr[(I - P)Y'(I - P)]}$ e $\tilde{Z} = (I - P)\tilde{Z}/\sqrt{tr[(I - P)\tilde{Z}'(I - P)]}$ onde I é uma matriz identidade nxn e P uma matriz nxn com todos os elementos iguais a $1/n$. Depois, se considera uma das matrizes como configuração de referência, nesse caso a matriz Y, e efetua-se a reflexão quando necessário (Figura 2c) e rotação de $\tilde{Z}$ (Figura 2d) para seu ajuste em Y. Ou seja, $\tilde{Z}$ é rotacionando para $\tilde{Z}Q$ tal que $Q = VU'$ é a matriz de rotação, $U\Sigma V'$ é a decomposição de valores singulares de $\tilde{Z}'Y$ em que Σ é uma matriz diagonal e U e V são matrizes ortogonais. Desta forma, tem-se a estatística M^2 - dada por (2) - como resultado da comparação entre Y e $\tilde{Z}$ denominada soma de quadrados residual de Procrustes ou estatística de Procrustes. Esta soma de quadrados é calculada a partir das

diferenças entre os escores correspondentes às configurações e mede a perda de informação sobre a estrutura dos dados quando apenas as q variáveis são utilizadas no lugar das p variáveis originais.

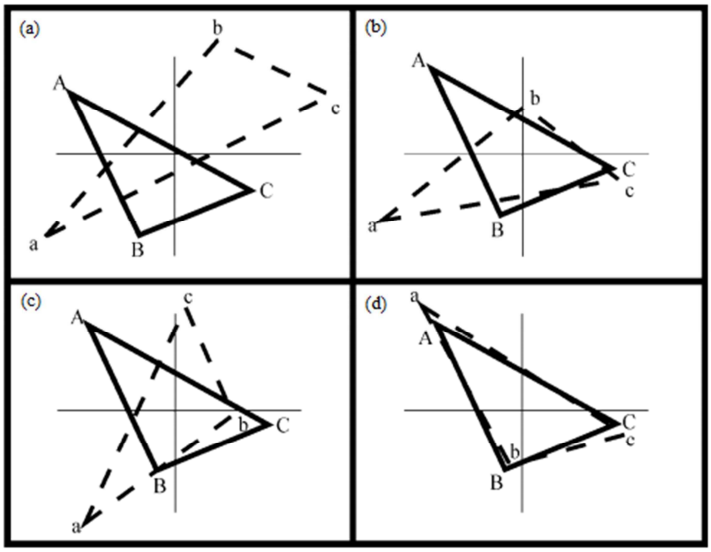

Figura 2: Passos da análise de procrustes (PERES-NETO, 2000).

$$M^2 = traço\{YY' + \tilde{Z}\tilde{Z}' - 2\tilde{Z}Q'Y'\} = traço\{YY' + \tilde{Z}\tilde{Z}' - 2\Sigma\} \qquad (2)$$

Segundo Peres-Neto e Jackson (2001), para restringir a variação da estatística M^2 entre 0 e 1 basta utilizar a seguinte transformação:

$$M^2 = 1 - (traço\ \Sigma)^2 \qquad (3)$$

A ideia existente por traz da Análise de Procrustes, encontrada na literatura também como transformação de Procrustes ou rotação de Procrustes, é que essas configurações estejam tão próximas quanto possível para que a partir de um mesmo referencial (mesmo subespaço) elas possam ser comparadas, digamos, de maneira justa e de tal forma que a "real" diferença entre as configurações possa então ser quantificada. Deste modo, tem-se inicialmente a translação e a escala realizadas simultaneamente (Figura 2b). Conforme detalhado por García-Peña e Dias (2009), a translação consiste no deslocamento de todos os pontos através de uma distância constante no mesmo sentido enquanto a escala é dada pelo estiramento ou encolhimento uniforme de toda a configuração. Depois a reflexão (Figura 2c) que é dada pelo "espelho" da configuração em torno de qualquer um dos eixos ou de todos os k eixos

considerados. E por fim a rotação (Figura 2d) que é dada pelo deslocamento fixo de todos os pontos através de um ângulo constante, mantendo a distância de cada ponto ao centróide.

O valor dado pelas equações (2) e (3) pode ser obtido para qualquer subconjunto selecionado com q variáveis e reflete a proximidade com que a configuração desse subconjunto representa a configuração original. Em outras palavras, quaisquer duas configurações serão mais similares quanto menores forem os valores de M^2. Por consequência, tem-se que o "melhor" subconjunto de q variáveis será aquele que apresentar o menor valor de M^2 dentre todos os subconjuntos possíveis de q variáveis, podendo ser obtido por qualquer método de seleção de variáveis.

Se os subconjuntos comparados forem obtidos por diferentes métodos de seleção de variáveis e por essa razão contiverem quantidades distintas, o que melhor representa o conjunto original de dados é aquele com o menor valor de M^2 estimado e que apresente uma configuração semelhante a do conjunto original (GUEDES E IVANQUI, 1998).

Krzanowski (1987) apresentou como sugestão para a seleção desse subconjunto de variáveis o algoritmo de eliminação *backward*, conforme descrito abaixo:

(I) Inicialmente tome $q = p$, e para k fixo calcule a matriz Y de escores das componentes principais. Tome $Z = Y$.

(II) Obtenha as matrizes $\tilde{Z}$ de escores dos componentes principais resultantes da eliminação de uma variável por vez.

(III) Calcule M^2, por meio da equação (2), entre Y e cada uma das matrizes $\tilde{Z}$ de escores e identifique a variável X_i cuja exclusão forneça o menor valor de M^2.

(IV) Exclua a variável X_i. Tome $\tilde{X} = (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_p)$ e retorne ao passo II com $(p-1)$ variáveis. Repita o procedimento até que somente q variáveis permaneçam.

3. MATERIAL E MÉTODOS

3.1. Dados

Para apresentar a proposta de análise de dados e obter resultados comparativos por diferentes técnicas de análise com ênfase em diversidade genética, foram utilizados dados referentes a 40 acessos de café Conilon avaliados em Sooretama/ES no ano 2000 em que foram mensurados 16 caracteres agronômicos. Para cada acesso foi considerada informação da média de 3 repetições (Fonte: FERRÃO, 2004).

Os caracteres avaliados foram: **C** - Período, em número de dias, da florada principal à completa maturação dos frutos colheita; **PMG** - Produtividade média de grãos (Produtividade média de grãos beneficiados da parcela convertida para kg/ha, após corrigida para 14% de umidade); **AP** - Altura média da planta (Distância da superfície do solo à extremidade do ramo ortotrópico, expressa em centímetros); **DC** - Diâmetro médio da copa (Tomada no "terço médio" da planta e expressa em centímetros); **CeCo** - Relação café cereja e café em coco (Relação baseada em uma amostra de 2 kg de café cereja e seu peso após a secagem); **CeBe** - Relação café cereja e café beneficiado (Relação baseada em uma amostra de 2 kg de café cereja e seu peso após a secagem e beneficiamento); **CoBe** - Relação café coco e beneficiado (Relação baseada no peso da amostra de café em coco seco e seu peso após o beneficiamento); **GCHO -** Porcentual de grãos chochos; **GCHA** - Porcentual de grãos "chatos"; **GMO** - Porcentual de grãos "moca"; **UMI** - Porcentagem de umidade dos grãos; **P17** - Porcentual de grãos retidos na peneira 17; **P15** - Porcentual de grãos retidos na peneira 15; **P13** - Porcentual de grãos retidos na peneira 13; **P11** - Porcentual de grãos na peneira 11; e **PM** - Peneira média (Tamanho médio dos grãos).

3.2. Proposta para critério de seleção de variáveis via Análise de Procrustes

Visando descartar variáveis, ou caracteres agronômicos mensurados, incialmente foi selecionado um subconjunto de variáveis por meio da análise de Procrustes conforme metodologia apresentada por Krzanowski (1987), que combina análise de componentes principais e análise de procrustes (Figura 3) para determinar o quanto um subconjunto de variáveis representa a estrutura do conjunto de variáveis originais. Após a realização da análise de componentes principais sobre as matrizes do conjunto de p variáveis originais X_{nxp} e do subconjunto de q variáveis $\tilde{X}_{nxq}$,

respectivamente, finalmente a análise de Procrustes foi efetuada entre as configurações Y_{nxk} e $\tilde{Z}_{nxk}$ representadas pelos escores das matrizes de dados a serem comparadas. Guedes e Ivanqui (1998) esclarecem que se a verdadeira dimensão dos dados é de fato k, Y_{nxk} pode ser vista como a verdadeira configuração, enquanto $\tilde{Z}_{nxk}$ é a aproximação correspondente da configuração baseada em somente q variáveis. Desta forma, a discrepância existente entre a configuração Y_{nxk} e a configuração $\tilde{Z}_{nxk}$ representa o resíduo produzido pela perda de informação devido à exclusão de (*p-q)* variáveis.

A escolha de *k* nas mais diversas áreas, geralmente, tem sido feita com base nos *k* primeiros componentes principais tal que estes conservem o máximo da informação possível contida no conjunto de dados original. Contudo, este trabalho impõe um valor fixo de *k*=2 com a finalidade de comparar dispersões gráficas bidimensionais, de grande utilidade na área de melhoramento genético, de tal modo que a estatística M^2 traduza exatamente a distância entre os acessos vistos no gráfico 2D.

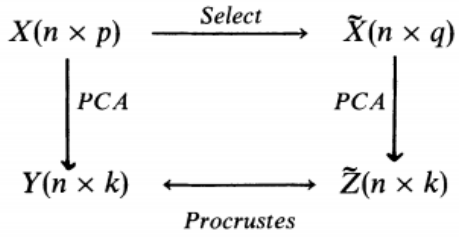

Figura 3: Diagrama ilustrativo da análise de procrustes dada por Krzanowski (1987).

Baseado na estatística M^2 de Procrustes dada por (3), conforme definida por Peres-Neto e Jackson (2001), cuja variação restringe-se entre 0 e 1, e por meio do uso de componentes principais pretende-se estabelecer um critério para a seleção de variáveis no estudo da diversidade genética. Ou seja, busca-se por um subconjunto que represente a mesma estrutura do conjunto original de variáveis sem perda de informação, ou ainda que o resíduo produzido pela perda de informação devida à exclusão de algumas variáveis seja mínimo de tal forma que o padrão de agrupamento não seja afetado.

Critério 1: o algoritmo *backward*. Para selecionar o subconjunto de variáveis ótimo Krzanowski (1987) propôs o algoritmo *backward*. Em cada estágio desse algoritmo é descartada aquela variável que proporcione o menor M^2 e o processo se repete até que, no final, têm-se uma sequência de (*p-k)* variáveis (no nosso caso *k*=2) e

seus respectivos valores de M^2 estimados. Considerando que por meio de M^2 dado por (3), *a priori* não há uma regra de parada e que, com base nesse algoritmo, a tomada de decisão sobre quais variáveis manter no estudo é arbitrária, propõe-se neste trabalho o seguinte critério: estabelecer/fixar um valor para a estatística M^2 de Procrustes ($M^2_{crítico}$) e buscar pelo subconjunto de variáveis que possua M^2 mais próximo do $M^2_{crítico}$, denominado seleção ótima.

Critério 2: o algoritmo *exaustivo.* Para adoção desta estratégia, considera-se que, a partir do mesmo $M^2_{crítico}$, é possível buscar também uma segunda seleção ótima dentre todos os subconjuntos de variáveis possíveis. Assim, são feitas todas as análises considerando números de variáveis iguais a $C_p^1, C_p^2, \ldots, C_p^{p-k}$ tendo-se, portanto, todas as possibilidades de combinações caracterizando um novo procedimento que, aqui, será denominado de algoritmo exaustivo e que certamente demanda maior esforço computacional que o critério anteriormente citado. Este procedimento fornece além da seleção ótima baseada no valor de $M^2_{crítico}$, uma série de outros subconjuntos de variáveis cujos valores de M^2 são inferiores ao $M^2_{crítico}$.

Critério 3: exclusão de variáveis por técnica de componentes principais. A terceira seleção de variáveis ótima foi obtida por meio do critério de Jolliffe (1972) que consiste em descartar aquelas variáveis com maior peso nos últimos componentes de menor importância. Em estudos de diversidade genética com caracteres padronizados, Cruz *et.al.* (2011) recomendam a exclusão de tantas variáveis quantos forem os autovalores dos componentes inferiores a 0,7.

3.3. Comparação de resultados

Finalmente, após a obtenção dos subconjuntos de variáveis dados pela seleção ótima por meio dos critérios previamente estabelecidos, tais subconjuntos foram comparados pela Análise de Procrustes. Isto é, o subconjunto que melhor representa o conjunto de variáveis originais será aquele com o menor valor de M^2 estimado e que apresente uma configuração semelhante à configuração de referência.

Importante destacar a padronização dos dados que foi realizada previamente à análise de componentes principais para evitar que variáveis de maior variância influenciem nos resultados de agrupamento já que os componentes principais são obtidos a partir da matriz de covariâncias. De modo análogo, após a padronização

(subtraindo de cada valor a média de sua respectiva variável e dividindo pelo seu desvio-padrão) os componentes principais são obtidos a partir da matriz de covariâncias das variáveis padronizadas, o que equivale a extraírem os componentes principais utilizando-se a matriz de correlação das variáveis originais.

Todas as análises foram realizadas no programa GENES (CRUZ, 2013) versão 2015.

4. RESULTADOS E DISCUSSÃO

4.1. Análise de Componentes Principais

De acordo com os resultados da análise de componentes principais (ACP) sobre o conjunto original de 16 caracteres agronômicos mensurados em 40 acessos de café Conilon, obteve-se uma proporção da variância total explicada até segundo componente de 49.35% (Tabela 1). Desta forma, avaliando a dispersão gráfica dos escores do primeiro e segundo componentes (Figura 4) assume-se então que há uma distorção de 50,65% no gráfico 2D.

Esta distorção tem sido considerada elevada para fins de estudo do grau de dissimilaridade entre pares de acessos tanto em espécies vegetais quanto animais recomendando-se, neste caso, incluir mais um eixo e optando-se pela representação gráfica 3D. A dificuldade de se reduzir o espaço *p* dimensional para apenas 2 ou 3 dimensões pode ser um inconveniente estatístico mas retrata um aspecto biológico importante, pois significa que o pesquisador tem um conjunto de informações não-redundantes e pode observar a diversidade genética a partir da ação de diferentes complexos gênicos.

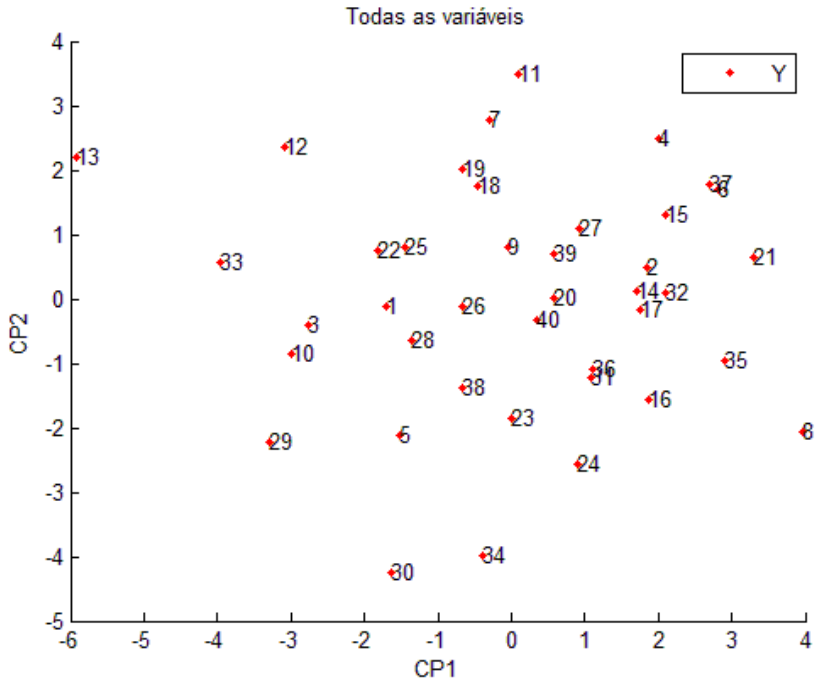

Figura 4: Dispersão dos acessos de café em relação a eixos representativos dos dois primeiros componentes principais obtidos pela análise de 16 caracteres agronômicos.

Para fins deste estudo vamos admitir que a representação gráfica, mesmo com a distorção elevada, foi considerada útil para inferir sobre o padrão de similaridade ou dissimilaridade entre os acessos de café. Assim, analisando a Figura 3, algumas evidências podem ser estabelecidas destacando, por exemplo, o fato dos acessos de café 13 e 8 serem divergentes o que pode ser indicativo, dependendo do potencial *per se* destes acessos, de uma recomendação de cruzamento para explorar vigor em combinação híbrida e maior variabilidade em população segregante. Várias outras inferências, para diferentes propósitos de um programa de melhoramento, poderão ser feitas a partir de informações tal como demonstrada na Figura 4. Após estas inferências, de grande utilidade no contexto do melhoramento genético, ainda permanecerá uma questão sobre a possibilidade de realizar este mesmo estudo, que conduziria a resultados similares, mas com economia de recurso por demandar a mensuração de um conjunto menor de variáveis. Ou seja, questionamos se é possível ter um padrão gráfico de diversidade tão similar quanto ao apresentado na Figura 4, mas com a utilização de um número menor de variáveis e, em caso afirmativo, quais e quantas seriam as variáveis descartadas. Tais questões poderão ser elucidadas considerando as abordagens apresentadas a seguir.

4.2. Redução de variáveis para fins de estudo da diversidade genética

Como resultado do algoritmo *backward*, obteve-se uma sequência de variáveis cuja exclusão em cada passo do algoritmo proporcionou o menor valor de M^2 estimado. Veja na Tabela 2 que o valor de M^2 cresce à medida que se excluem variáveis, o que era de se esperar dado que à medida que cada variável é excluída maior é o resíduo produzido pela perda de informação em relação ao conjunto original. Para um $M^2_{\text{crítico}} = 0.1$, obteve-se a seleção ótima pelo seguinte subconjunto de variáveis: PM, PMG, CeCo, P15, GMO e P17.

Tabela 1: Estimativas dos autovalores obtidos da matriz de correlação das 16 variáveis e autovetores (componentes) associados.

λ	VT%	VTa%	Variáveis														
			C(Dias)	**PMG(kg/ha)**	**AP(cm)**	**DC (cm)**	**CeCo**	**CeBe**	**CoBe**	**GCHO(%)**	**GCHA(%)**	**GMO(%)**	**UMI(%)**	**P17(%)**	**P15(%)**	**P13(%)**	**P11(%)**	**PM**
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: Autovalor. **VT%**: Percentual da variância total explicada pelo *i*-ésimo componente principal. **VTa%**: Percentual acumulado pelos componentes.

Tabela 2: Sequência de variáveis excluídas pelo algoritmo *backward*.

Variável excluída	Descartadas										Selecionadas					
	CoBe	**P11**	**C**	**AP**	**P13**	**GCHA**	**CeBe**	**DC**	**UMI**	**GCHO**	**PM**	**PMG**	**CeCo**	**P15**	**GMO**	**P17**
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	-	-

Na Figura 5 é apresentada a dispersão dos escores dos acessos em relação aos dois primeiros componentes para o subconjunto de 6 variáveis estabelecido a partir da seleção ótima realizada por meio do algoritmo *backward*. Nota-se que a configuração ficou refletida em torno da origem do componente 2.

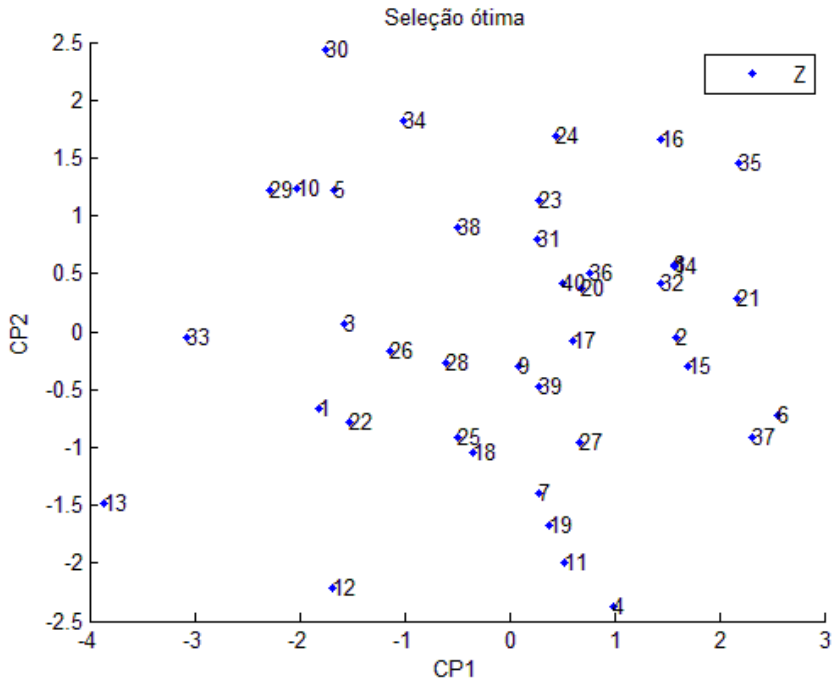

Figura 5: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o subconjunto estabelecido pela seleção ótima segundo o algoritmo *backward.*

Após a transformação de Procrustes sobre a seleção ótima pôde-se então observar sua real diferença em relação ao conjunto de variáveis original, diferença estimada por meio de um $M^2 = 0.0895$ (Figura 6). Verificou-se que houve mudança no padrão de agrupamento dos acessos avaliados uma vez que, por exemplo, o acesso 8, antes divergente assim como o acesso 13, passou a pertencer a um grupo de genótipos juntamente com o acesso 14 (Figura 7). Sendo assim, a seleção ótima mesmo que inclua as variáveis de interesse não é adequada dado que não proporcionou uma dispersão dos acessos satisfatoriamente próxima da dispersão dada pelo conjunto original conforme mostra a Figura 4. E, neste caso, com base unicamente no critério estabelecido ($M^2_{\text{crítico}} = 0.1$) por meio do algoritmo *backward*, a recomendação seria alterar o valor de $M^2_{\text{crítico}}$ e avaliar novamente o padrão de agrupamento entre os acessos ou, se as seleções ótimas subsequentes não forem pertinentes quanto as variáveis selecionadas, prosseguir o estudo utilizando a informação das 16 variáveis.

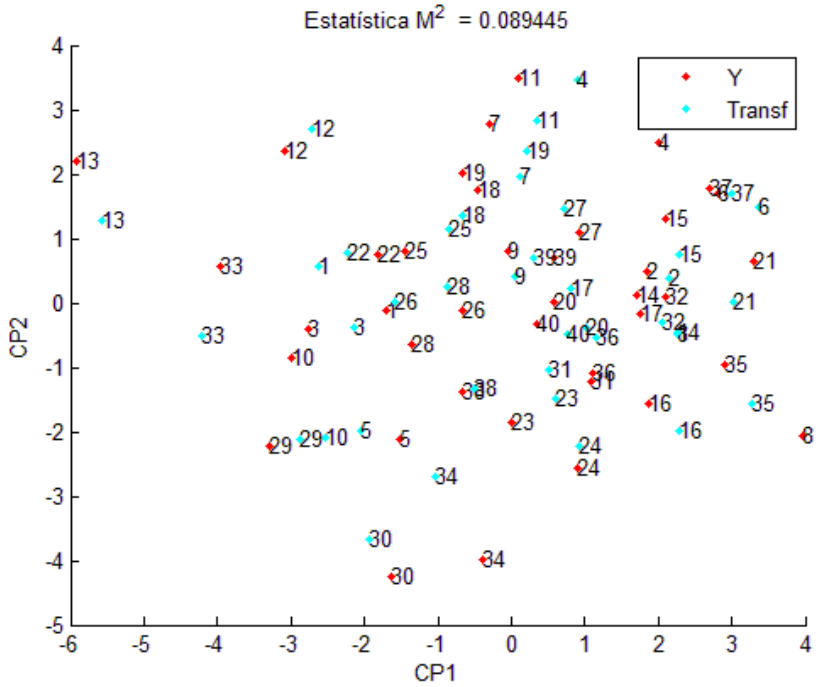

Figura 6: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o conjunto de 16 variáveis e para o subconjunto estabelecido pela seleção ótima segundo o algoritmo *backward* após a transformação de Procrustes.

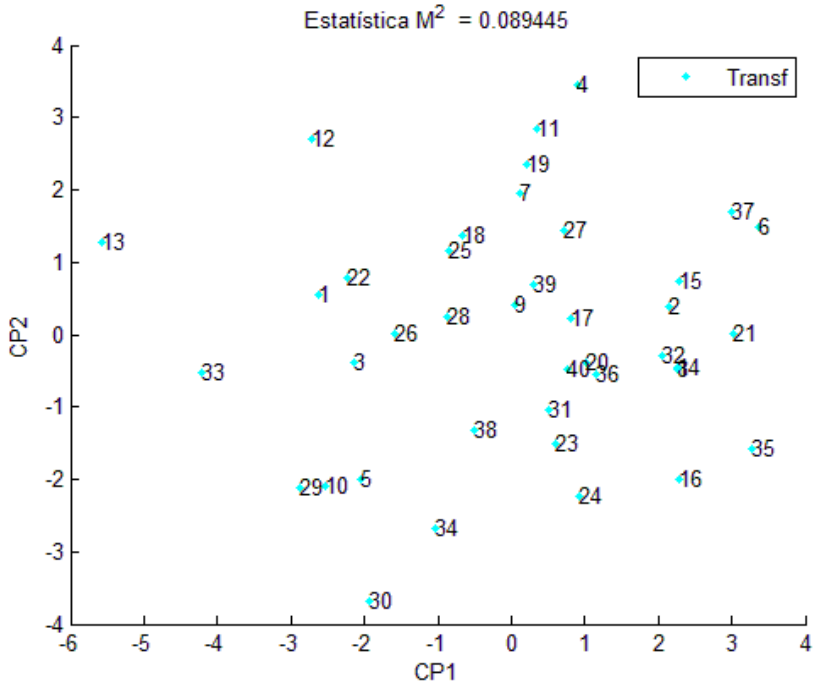

Figura 7: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o subconjunto estabelecido pela seleção ótima segundo o algoritmo *backward* após a transformação de Procrustes.

A seleção ótima resultante do algoritmo exaustivo, para um $M^2_{\text{crítico}} = 0.1$, é dada pelo seguinte subconjunto de variáveis: DC, CeBe, CoBe, GMO, UMI, P15, P11 e PM. Há um total de 9841 combinações (subconjuntos) que apresentaram valor de M^2 inferior ao $M^2_{\text{crítico}}$ (Tabela 3), dentre elas a seleção ótima resultante do algoritmo *backward*.

Tabela 3: Total de subconjuntos com valor abaixo do $\boldsymbol{M^2_{\text{crítico}} = 0.1}$.

Variáveis	Subconjuntos
5	2
6	64
7	322
8	735
9	1538
10	2490
11	2504
12	1502
13	548
14	120
15	16
Total	**9841**

Na Figura 8 é apresentada a dispersão dos escores dos acessos em relação aos dois primeiros componentes para o subconjunto de 8 variáveis estabelecido a partir da seleção ótima realizada por meio do algoritmo exaustivo. Mais uma vez, a configuração dos acessos da pela seleção ótima ficou refletida, porém agora em relação à origem dos componentes 1 e 2 simultaneamente.

Após a transformação de Procrustes sobre a seleção ótima sua real diferença em relação ao conjunto original foi estimada por meio de $M^2 = 0.1$ (Figura 9). Observou-se que não houve alteração no padrão de agrupamento (Figura 10) de modo que a seleção ótima proporcionou uma dispersão global satisfatoriamente próxima da dispersão dada pelo conjunto original conforme retrata a Figura 4.

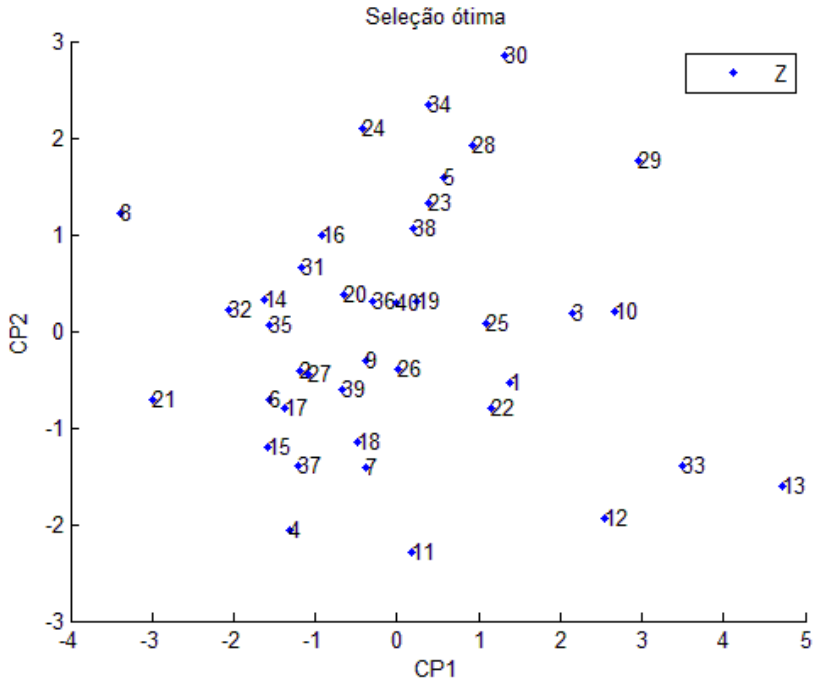

Figura 8: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o subconjunto estabelecido pela seleção ótima segundo o algoritmo exaustivo.

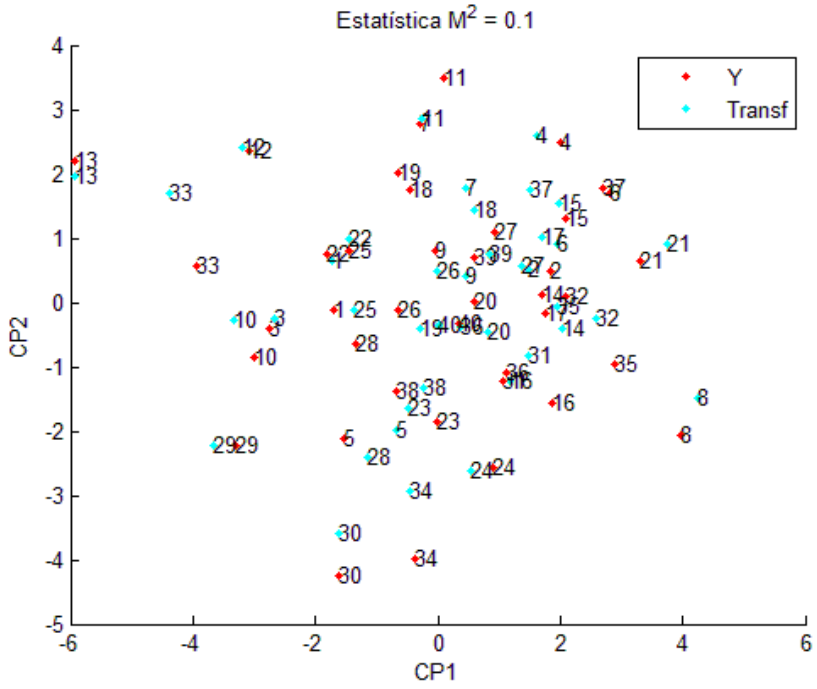

Figura 9: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o conjunto de 16 variáveis e para o subconjunto estabelecido pela seleção ótima segundo o algoritmo exaustivo após a transformação de Procrustes.

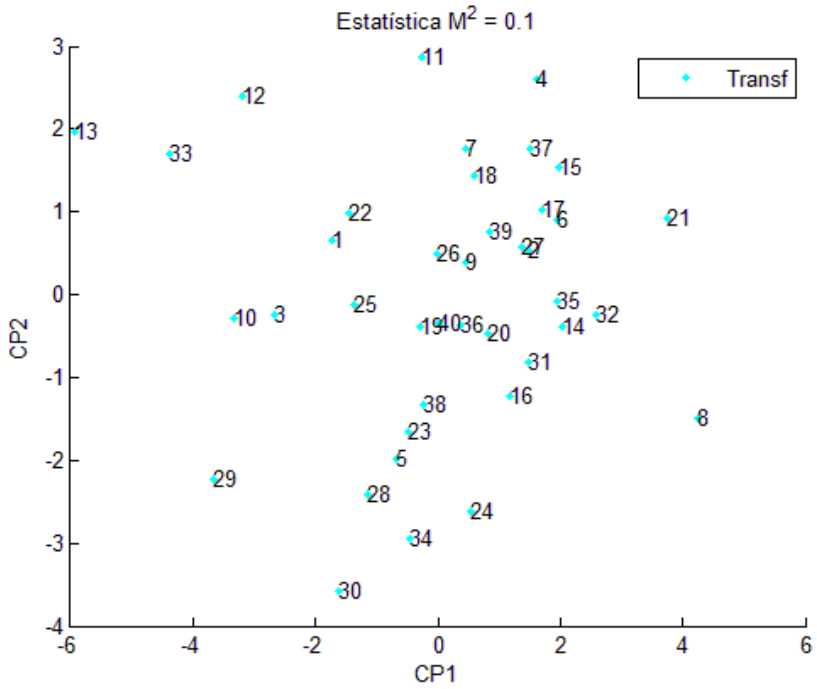

Figura 10: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o subconjunto estabelecido pela seleção ótima segundo o algoritmo exaustivo após a transformação de Procrustes.

A partir dos dados apresentados na Tabela 1 é possível identificar a importância relativa das variáveis sobre a diversidade genética dos acessos, por meio da qual o descarte deve ser executado. Conforme critério apresentado por Jolliffe (1972) e sugestão de Cruz et. al. (2011), partindo-se do último até o nono componente as variáveis de maiores pesos foram, respectivamente: P13, PM, GMO, CeBe, P11, PMG, AP, P15 e UMI. Desta forma, a seleção ótima é dada pelo subconjunto de variáveis: C, DC, CeCo, CoBe, GCHO, GCHA e P17.

Na Figura 11 é apresentada a dispersão dos escores dos acessos em relação aos dois primeiros componentes para o subconjunto de 7 variáveis estabelecido a partir da seleção ótima realizada por meio do critério de Jollife. Assim como casos anteriores, ocorreu a mudança de posição dos acessos em função da exclusão de algumas variáveis, que nesse caso ficaram refletidos em torno da origem dos componente 1 e componente 2.

Após a transformação de Procrustes sobre a seleção ótima sua real diferença em relação ao conjunto original foi estimada por meio de $M^2 = 0.3359$ (Figura 12). A magnitude de M^2 traduziu a não proximidade entre os acessos de café correspondentes às configurações que, nesse caso, revelou a alteração no padrão de agrupamento dos

acessos avaliados tal qual podemos destacar o acesso 19 que agora pertence ao grupo do acesso 13 bem como os acessos 16, 17 e 35 que pertencem ao mesmo grupo do acesso 8 (Figura 13).

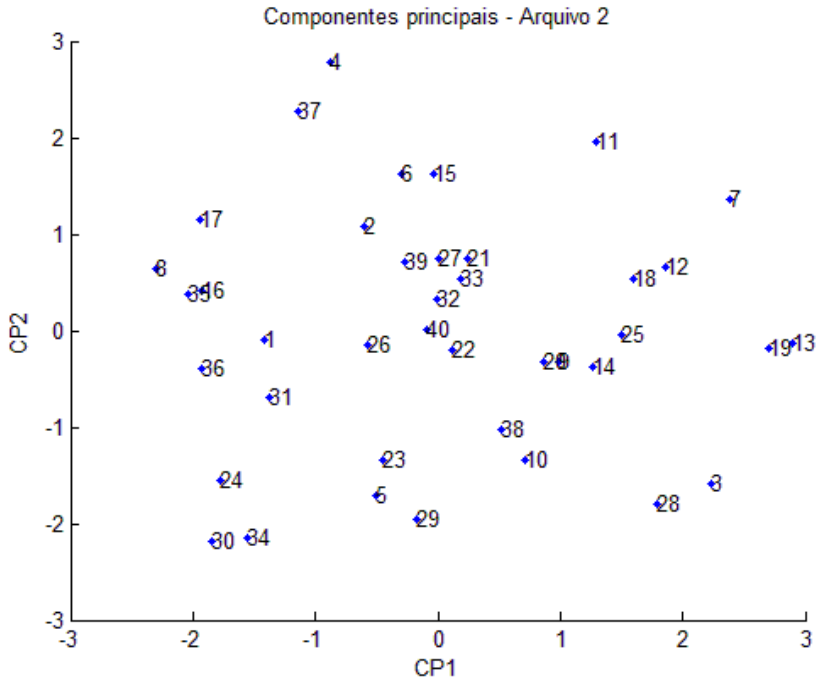

Figura 11: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o subconjunto estabelecido pela seleção ótima segundo o critério de Jolliffe.

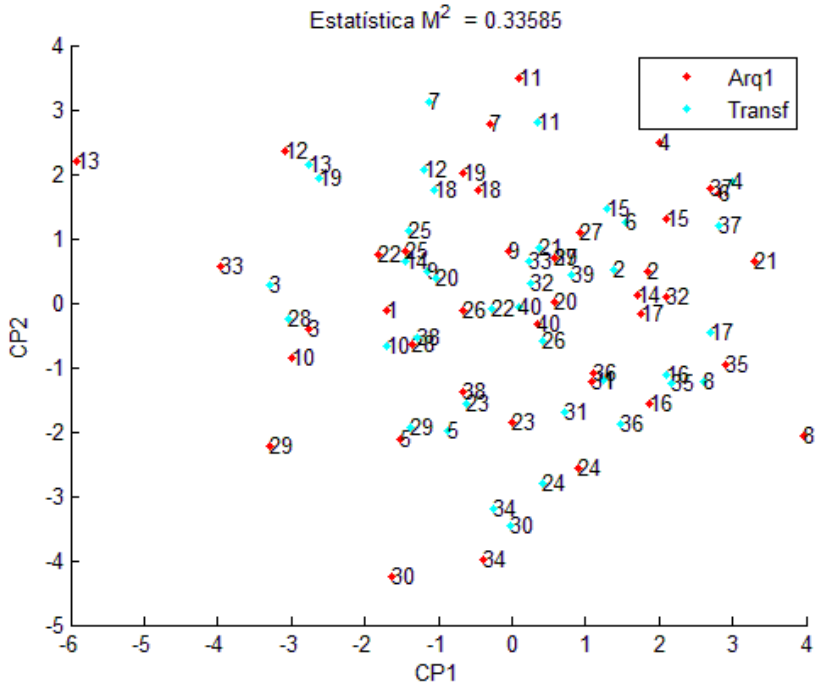

Figura 12: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o conjunto de 16 variáveis e para o subconjunto estabelecido pela seleção ótima segundo o critério de Jolliffe após a transformação de Procrustes.

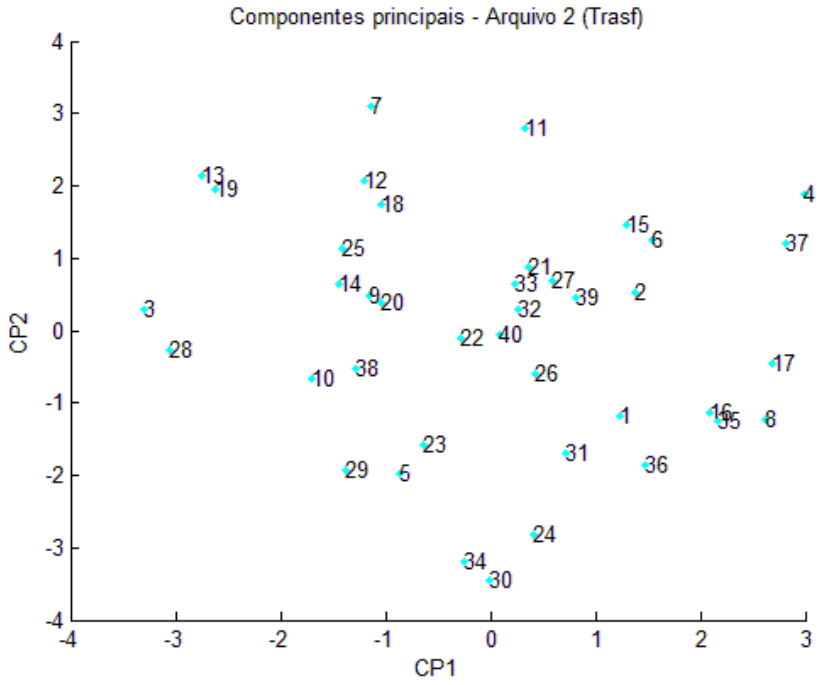

Figura 13: Dispersão dos escores dos acessos em relação aos dois primeiros componentes principais para o subconjunto estabelecido pela seleção ótima segundo o critério de Jolliffe após a transformação de Procrustes.

Na Tabela 4 é dado um quadro com o resumo dos resultados obtidos para todos os critérios avaliados. Algumas variáveis foram comuns aos subconjuntos estabelecidos pela seleção ótima por meio dos algoritmos *backward* e exaustivo (ambos os critérios baseados em análise de Procrustes) como, por exemplo, as variáveis: GMO, P15 e PM. Isso pressupõe a importância dessas variáveis para a variabilidade do grupo de acessos avaliados uma vez que os subconjuntos selecionados por ambos os critérios proporcionaram ganho na variância total explicada pelos dois primeiros componentes.

Tabela 4: Seleção ótima pelos critérios de Procrustes e Jolliffe.

Critério	Seleção ótima	M^2 estimado	VTa%	Variáveis eliminadas	Solução
Procrustes (backward)	PMG, CeCo, GMO, P17, P15, e PM	0.0895	62.57	CoBe, P11, C, AP, P13, GCHA, CeBe, DC, UMI e GCHO	Única
Procrustes (exaustivo)	DC, CeBe, CoBe, GMO, UMI, P15, P11 e PM	0.1	56.82	C, PMG, AP, CeCo, GCHO, GCHA, P17 e P13	1 em 9841
Jolliffe	C, DC, CeCo, CoBe, GCHO, GCHA e P17	0.3359	51.50	P13, PM, GMO, CeBe, P11, PMG, AP, P15 e UMI	Única

VTa%: Percentual acumulado da variância total explicada pelos dois primeiros componentes.

Embora o subconjunto selecionado pelo algoritmo *backward* tenha proporcionado o menor valor de M^2 estimado bem como o que possui o menor número de variáveis, este não representou adequadamente a diversidade dos acessos entre os acessos considerando a ACP a partir do conjunto com todas as 16 variáveis conforme Figura 4. Por outro lado, o subconjunto selecionado pelo algoritmo exaustivo, apesar de possuir um M^2 estimado maior e maior número de variáveis que o primeiro critério, representou satisfatoriamente a diversidade entre os acessos.

Entretanto, é importante destacar alguns pontos característicos de ambos os procedimentos baseados em Análise de Procrustes. Quanto ao número de soluções, temos que o algoritmo *backward* assim como o critério de Jolliffe fornece apenas uma seleção ótima enquanto o algoritmo exaustivo fornece todos os subconjuntos de variáveis que possuem um valor de M^2 estimado abaixo do $M^2_{\text{crítico}}$ tal como mostrado na Tabela 3. Isso abre um leque de possibilidades ao pesquisador uma vez que, a partir do $M^2_{\text{crítico}}$ estabelecido, o subconjunto de variáveis apontado como a seleção ótima pelo procedimento *backward* pode não incluir alguma variável com importância biológica e de interesse para seu estudo. Ou ainda, o subconjunto selecionado por este critério pode não revelar dispersão gráfica equivalente à obtida pela análise considerando o conjunto original, como neste caso.

O segundo ponto refere-se à obtenção das soluções. Diferentemente do procedimento backward que exclui uma variável por vez em cada passo do algoritmo, o procedimento exaustivo avalia todas as possibilidades de descarte de variáveis (uma a uma, duas a duas, etc). Isso faz com que este último procedimento talvez não seja interessante em casos de matrizes de dados de alta ordem cujo manuseio envolve elevado custo computacional e o processamento dos resultados pode levar dias ou até meses. A Tabela 5 mostra o total de análises realizadas pelo algoritmo exaustivo de acordo com o número de variáveis estudadas. Observe que à medida aumenta-se o número de variáveis, a quantidade de análises realizadas por esse algoritmo cresce consideravelmente.

Tabela 5: Total de analises realizada pelo algoritmo exaustivo.

N° de variáveis	Análises
10	1 012
16	65 518
20	1 048 554
30	1 073 741 792
...	...
50	1 125 899 906 842 570

Outro aspecto interessante sobre os critérios baseados em Análise de Procrustes diz respeito ao valor de $M^2_{\text{crítico}} = 0.1$ estabelecido neste estudo, que representou uma opinião pessoal do pesquisador. Vale salientar que no algoritmo *backward* bem como no algoritmo exaustivo, o valor de $M^2_{\text{crítico}}$ pode ser levemente relaxado de acordo com o número de variáveis que o pesquisador/melhorista deseja descartar sugerindo-se, portanto uma variação entre os limites mínimo e máximo de 0.5 e 0.15 desde que seja mantido o padrão de agrupamento de acessos da cultura estudada.

Quanto ao subconjunto selecionado pelo critério de Jolliffe (1972), observou-se um elevado valor de M^2 estimado que revelou alteração do padrão de agrupamento dos dados fazendo com esse critério seja relativamente menos eficiente que os demais.

Por fim, destaca-se o procedimento exaustivo, como de grande potencialidade para fins de estudos genéticos, pelo número de soluções ótimas resultantes.

Em estudos futuros, pretende-se por meio de um estudo de simulação estabelecer um limiar para a estatística M^2 de procrustes de modo que esse limite seja

capaz de captar o momento em que o descarte de variáveis irá alterar o padrão de agrupamento, dando ao melhorista e pesquisador o poder de decisão sobre quais caracteres excluir do seu estudo sem prejudicar a variabilidade genética do grupo de acessos estudado.

5. CONCLUSÕES

As técnicas apresentadas neste trabalho demonstram ser vantajosas na seleção (ou descarte) de variáveis proporcionando relevante contribuição para os estudos sobre diversidade genética.

A técnica apresentada, baseada em análise de Procrustes, torna-se uma alternativa mais eficaz do que o uso do critério de Jolliffe (1972) para fins de estudo da diversidade genética.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, R. D. DE; PELUZIO, J. M.; AFFÉRRI, F. S. Divergência genética entre cultivares de soja, sob condições de várzea irrigada, no sul do Estado Tocantins. **Revista Ciência Agronômica**, v. 42, n. 1, p. 108-115, 2011.

ASSOCIAÇÃO BRASILEIRA da INDÚSTRIA do CAFÉ. *História: O café brasileiro na atualidade. Disponível em:* <http://www.abic.com.br/publique/cgi/cgilua.exe/sys/start.htm?sid=38#58>. Acesso em: 22 jan. 2016.

BELING, R. R. Anuário brasileiro do café 2005. **Santa Cruz do Sul: Gazeta Santa Cruz**, 2005. 136 p.

BERTINI, C. H. C. M.; TEÓFILO, E. M.; DIAS, F. T. C. Divergência genética entre acessos de feijão-caupi do banco de germoplasma da UFC. **Revista Ciência Agronômica**, v. 40, n. 1, p. 99-105, 2009.

BRAMARDI, S.J.; BERNET, G. P.; ASÍNS, M. J.; CARBONELL, E. A. Simultaneous agronomic and molecular characterization of genotypes via the generalized procrustes analysis: an application to cucumber. **Crop Science,** v. 45, p. 1603-1609, 2005.

CHARRIER, A.; BERTHAUD, J. Principles and methods in coffee plant breeding: Coffea canephora Pierre. In: CLARKE, R. J.; MACRAE, R. (Eds.). ***Coffee*: Agronomy. London: Elsilvier Applied Science**, 1988, v. 3, p. 167-195.

COELHO, C. M. M.; COIMBRA, J. L. M.; SOUZA, C. A. de; BOGO, A.; GUIDOLIN, A. F. Diversidade genética em acessos de feijão (Phaseolus vulgaris L.). **Ciência Rural**, v. 37, n. 5, p. 1241-1247, 2007.

CRUZ, C. D.; VENCOVSKY, R.; CARVALHO, S. P. Estudos sobre divergência genética. III. Comparação de técnicas multivariadas. **Ceres**, v. 41, n. 234, p. 191-201, 1994.

CRUZ, C. D. Genes: a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum. Agronomy**, v. 35, n. 3, p. 271-276, 2013.

CRUZ, C. D. **Princípios de genética quantitativa.** Viçosa: UFV, 2005. 394 p.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético.** 4 ed, v. 1,. Viçosa: UFV, p.514. 2012.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética.** Visconde do Rio Branco: Suprema, 2011. 620p.

DINIZ FILHO, J. A. **Métodos filogenéticos comparativos.** Ribeirão Preto: Holos, 2000. 120 p.

DOUGLAS, T. S. Image processing for craniofacial landmark identification and measurement: a review of photogrammetry and cephalometry. **Computerized Medical Imaging and Graphics**, v. 28, n. 7, p. 401-409, 2004.

FERRÃO, R. G. **Biometria aplicada ao melhoramento genético do café Conilon.** 2004. 256 f. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, MG, 2004.

FONSECA, A. F. A. DA; SEDIYAMA, T.; CRUZ, C. D.; SAKAIYAMA, N. S.; FERRÃO, M. A. G.; FERRÃO, R. G.; BRAGANÇA, S. M. Divergência genética em café conilon. **Pesquisa agropecuária brasileira, Brasília**, v. 41, n. 4, p. 599-605, 2006.

GARCÍA-PEÑA, M.; DIAS, C. T. S. Análise dos modelos aditivos com interação multiplicativa (AMMI) bivariados. **Revista Brasileira de Biometria, São Paulo**, v. 27, n. 4, p. 586-602, 2009.

GUEDES, T.A.; IVANQUI, I. L. Análise procrustes aplicada à seleção de variáveis. **Acta Scientiarum. Technology**, v. 20, p. 505-509, 1998.

GUEDES, J. M.; VILELA, D. J. M.; REZENDE, J. C.; SILVA, F. L.; BOTELHO, C. E.; CARVALHO, S. P. Divergência genética entre cafeeiros do germoplasma Maragogipe. **Bragantia**, v. 72, n. 2, p. 127-132, 2013.

HOTELLING, H. Analisys of a complex of statistical variables into principal components. **Journal of Education Psychology,** v. 24, p.417-441, 1993.

HOTELLING, H. Simplified calculation of principal components. **Psychometrika,** v. 1, n. 1, p. 27-35, 1936.

JOLLIFFE. I.T. Discarding variables in a principal components analysis. I: Artificial data. **Applied statistics**. v. 21, p.160-173, 1972.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. II: Real data. **Applied statistics**, p.21-31, 1973.

KLINGENBERG, C. P. Quantitative genetics of geometric shape: heritability and the pitfalls of the univariate approach. **Evolution,** v. 57, n. 1, p. 191-195, 2003.

KOBAYASHI, M. L.; BENASSI, M. T. Caracterização sensorial de cafés solúveis comerciais por Perfil Flash. **Semina: Ciências Agrárias**, v. 33, n. 6Supl2, p. 3081-3092, 2013.

KRZANOWSKI, W. J. Selection of variables to preserve multivariate data structure, using principal components. **Applied Statistics**, p. 22-33, 1987.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. 6 ed. Londres: Academic press, 1979.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** 2ª reimpressão. Belo Horizonte: Editora UFMG, 2005.

PEARSON, K. On lines and planes of closet fit to systems of points in space. **Philosophical Magazine,** v. 2, p. 559-572, 1901.

PELUZIO, J. M., VAZ-DE-MELO, A., AFFÉRRI, F. S., SILVA, R. R., BARROS, H. B., NASCIMENTO, I. R.; FIDELIS, R. R. Variabilidade genética entre cultivares de soja, sob diferentes condições edafoclimáticas. **Revista Brasileira de Tecnologia Aplicada nas Ciências Agrárias**, v. 2, n. 3, p. 21-40, 2009.

PERES-NETO, P. R.; JACKSON, D. A. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. **Oecologia**, v. 129, n. 2, p. 169-178, 2001.

PESSONI, L. A. **Estratégias de análise da diversidade em germoplasma de cajueiro (*Anacardium* spp. L.)**. 2007. 159 f. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, MG, 2007.

RIGON, J. P. G.; CAPUANI, S.; DE BRITO NETO, J. F.; ROSA, G. M. DA, WASTOWSKI, A. D.; RIGON, C. A. G. Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos.**Revista Ceres**, v. 59, n. 2, p. 233-240, 2012.

ROCHA, R. B., SANTOS, D. V., RAMALHO, A. R., & TEIXEIRA, A. L. Caracterização e uso da variabilidade genética de banco ativo de germoplasma de Coffea canephora Pierre ex Froehner. **Coffee Science**, v. 8, n. 4, p. 478-485, 2013.

ROTILI, E. A.; CANCELLIER, L. L.; DOTTO, M. A.; PELUZIO, J. M.; CARVALHO, E. V. Divergência genética em genótipos de milho, no Estado do Tocantins. **Revista Ciência Agronômica,** v. 43, n. 3, p. 516-521, 2012.

SINGH, Daljit. The relative importance of characters affecting genetic divergence. **Indian Journal of Genetics and Plant Breeding (The)**, v. 41, n. 2, p. 237-245, 1981.

VAN DER VOSSERN, H. A. M. Coffea selection and breeding. In: CLIFFORD, M. N.; WILLSON. K. C. (Eds.). *Coffe*: botany, biochemistry and production of beans and beverage. London: Croom Helm, Westport Conn, 1985. P.48-96.

ZENG, X.; Intelligent sensory evaluation: Concepts, implementations and applications. **Mathematics and Computers in Simulation**, v. 77, n. 5, p. 443-452, 2008.