UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA TROPICAL

ANÁLISES DISCRIMINANTES NÃO PARAMÉTRICAS APLICADAS AO ESTUDO DA DIVERSIDADE GENÉTICA BASEADO EM DADOS FENOTÍPICOS QUANTITATIVOS

MARCILEIA SANTOS SOUZA

MANAUS-AM
2017

UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA TROPICAL

MARCILEIA SANTOS SOUZA

ANÁLISES DISCRIMINANTES NÃO PARAMÉTRICAS APLICADAS AO ESTUDO DA DIVERSIDADE GENÉTICA BASEADO EM DADOS FENOTÍPICOS QUANTITATIVOS

Dissertação apresentada ao Programa de Pós-Graduação em Agronomia Tropical da Universidade Federal do Amazonas, como requisito para obtenção do título de Mestre em Agronomia Tropical, área de concentração Manejo da Agrobiodiversidade.

ORIENTADOR: Dr. FÁBIO MEDEIROS FERREIRA

MANAUS-AM
2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Souza, Marcileia Santos
S729a Análises discriminantes não paramétricas aplicadas ao estudo da diversidade genética baseado em dados fenotípicos quantitativos / Marcileia Santos Souza. 2017
78 f.: il. color; 31 cm.

Orientador: Fábio Medeiros Ferreira
Tese (Mestrado em Agronomia Tropical) - Universidade Federal do Amazonas.

1. divergência genética. 2. vizinho médio. 3. k-vizinhos mais próximos. 4. melhoramento genético. I. Ferreira, Fábio Medeiros II. Universidade Federal do Amazonas III. Título

MARCILEIA SANTOS SOUZA

ANÁLISES DISCRIMINANTES NÃO PARAMÉTRICAS APLICADAS AO ESTUDO DA DIVERSIDADE GENÉTICA BASEADO EM DADOS FENOTÍPICOS QUANTITATIVOS

Dissertação apresentada ao Programa de Pós-Graduação em Agronomia Tropical da Universidade Federal do Amazonas, como requisito para obtenção do título de Mestre em Agronomia Tropical, área de concentração Manejo da Agrobiodiversidade.

Aprovada em 04 de dezembro de 2017.

BANCA EXAMINADORA

Prof. Dr. Fábio Medeiros Ferreira
Universidade Federal do Amazonas

Profa. Dra. Maria Teresa Gomes Lopes
Universidade Federal do Amazonas

Dr. Inocêncio Júnior de Oliveira
Embrapa Amazônia Ocidental

A meus pais Manoel Ferreira de Souza e Antônia Cícera dos Santos, e irmãs Adriana Santos Souza, Liliane Santos Souza e Mirian Santos Souza pelo incentivo para a realização deste trabalho.

Dedico.

AGRADECIMENTOS

A Deus e seu filho Jesus Cristo, por iluminarem meus caminhos e guiarem meus passos;

Ao professor Dr. Fábio Medeiros Ferreira, pela orientação, ensinamento e apoio constantes;

Ao Dr. Rodrigo Barros Rocha, pesquisador da Embrapa Rondônia, pela cessão dos conjuntos de dados utilizados nesta dissertação;

Aos meus familiares, pais e irmãs pelo apoio e estímulo;

Aos colegas da turma, que auxiliaram no decorrer do mestrado;

A Universidade Federal do Amazonas, pela oportunidade e concessão da bolsa de estudos.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, pela bolsa de estudos concedida.

A Embrapa Rondônia, pela liberação de dados para a elaboração desta dissertação.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.”

(Arthur Schopenhauer)

RESUMO

Os métodos multivariados de análises discriminantes visam identificar as populações nas quais um indivíduo deva pertencer, admitindo previamente, que o indivíduo compõe uma das populações avaliadas. Métodos baseados em funções discriminantes lineares têm sido usados nos estudos preditivos da diversidade no melhoramento genético, quando os dados são fenotípicos quantitativos. Entretanto, este tipo de análise pressupõe a multinormalidade das populações. Objetivou-se avaliar a efetividade das metodologias de análise discriminante não paramétricas do vizinho médio e dos k-vizinhos mais próximos no estudo preditivo da diversidade no melhoramento genético, quando aplicadas à variáveis quantitativas, de modo a classificar satisfatoriamente os genótipos em suas respectivas populações definidas *a priori*. Dois conjuntos de dados foram utilizados: i) 83 matrizes de pupunha, previamente alocadas em três raças primitivas, para sete variáveis do fruto; ii) 122 clones de cafeeiro, previamente alocados entre três variedades botânicas, para dez variáveis agronômicas. Avaliou-se os métodos não paramétricos do vizinho médio e dos k-vizinhos mais próximos sob vários cenários, conforme combinações possíveis entre técnica de análise não paramétrica *x* medida de distância genética *x* valor de k *x* probabilidade *a priori* dos genótipos pertencerem as populações. Comparou-se a alocação dos genótipos nos diferentes cenários e com aquela obtida pelas funções discriminantes de Anderson (considerada padrão) a partir das taxas de erro aparente globais (TEA) de classificação dos indivíduos nas respectivas populações. Utilizou-se o software GENES. Os métodos não paramétricos foram efetivos para classificar os genótipos em suas respectivas populações quando comparados com o método de análise discriminante de Anderson. Não houve diferenças significativas entre as medidas de distâncias Euclidianas. A distância de Gower proporcionou taxas de erro aparente diferente das demais distâncias estudadas. O método de análise discriminante dos k vizinhos mais próximos mostrou ser adequado para populações cuja divergência genética dentro é menor. Já o método do vizinho médio classifica melhor os genótipos em populações em que haja maior diversidade inter ou intrapopulacional.

Palavras-chave: divergência genética; vizinho médio; k-vizinhos mais próximos.

ABSTRACT

The multivariate discriminant analysis methods aim to identify the populations in which an individual should belong, admitting previously, that the individual composes one of the evaluated populations. Methods based on linear discriminant functions have been used in predictive studies of diversity in genetic improvement, when the data are quantitative phenotype. However, this type of analysis presupposes the multinormality of populations. The objective of this study was to evaluate the effectiveness of the non-parametric discriminant methodologies of the middle neighbor and k-Nearest Neighbour in the predictive study of diversity in genetic improvement, when applied to quantitative variables, in order to satisfactorily (re) classify the genotypes in their respective populations defined *a priori*. Two sets of data were used: i) 83 pupunha matrices, previously allocated in three primitive races, for seven variables of the fruit; ii) 122 clones of coffee trees, previously allocated among three botanical varieties, for ten agronomic characteristics. The non-parametric methods of the middle neighbor and the k-Nearest Neighbour were evaluated under various scenarios, according to possible combinations between non-parametric analysis technique *x* genetic distance measure *x* k *x* probability *a priori* of the genotypes belonging to the populations. The genotype allocation was compared in the different scenarios and the one obtained by Anderson's discriminant functions (considered standard) from the global apparent error rates (TEA) of classification of the individuals in the respective populations. The GENES software was used. The nonparametric methods were effective to classify the genotypes in their respective populations when compared with Anderson's discriminant analysis method. There were no significant differences between Euclidean distances measurements. The Gower distance provided apparent error rates different from the other studied distances. The method of discriminant analysis of *the k-Nearest Neighbour* proved to be adequate for populations whose genetic divergence within is smaller. The middle neighbor method, however, classifies the genotypes better in populations where there is greater inter- or intra-population diversity.

Keywords: genetic divergence; middle neighbor; k-Nearest Neighbour.

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

As características (caracteres ou variáveis resposta) são comumente classificadas como qualitativas ou quantitativas (FERREIRA, 2005). No contexto do melhoramento genético vegetal aquelas qualitativas, por exemplo: cor da flor e das folhas, formatos e cor de frutos, resistência a patógenos, etc. apresentam um padrão simples de herança (monogênica ou oligogênica), que se baseia nas proporções das classes fenotípicas, avaliadas nas descendências de cruzamentos. Já os caracteres quantitativos, como produção, teor de óleo, altura da planta, número de ramos produtivos, etc., apresentam herança complexa (poligênica), uma vez que em contraposição aos caracteres qualitativos, são, em sua maioria, condicionados por muitos genes com efeitos individuais pequenos e muito influenciados pelo ambiente (CRUZ, 2005).

Independentemente do tipo de variável, o sucesso do melhoramento genético requer, obrigatoriamente, que o mesmo seja herdável e que haja variação na população em que se pratica a seleção (CRUZ et al., 2014). É comum ao melhorista, dentre as várias estratégias para o desenvolvimento de cultivares, realizar o intercruzamento de materiais genéticos de desempenho superior e divergentes entre si, a fim de se explorar a complementação gênica e a heterose para a formação de híbridos, sejam intraespecíficos ou interespecíficos – quando há compatibilidade gamética. Portanto, é fundamental no estágio inicial do programa de melhoramento conhecer a diversidade genética dos genótipos pertencentes a(s) população(ões) de trabalho.

Na análise da diversidade, técnicas biométricas são empregadas, baseadas na quantificação da heterose – análises dialélicas – ou por métodos preditivos, que se baseiam nas diferenças morfológicas, fisiológicas (CRUZ, 2005), bioquímicas e moleculares entre os genótipos. Métodos biométricos multivariados são comumente aplicados para predizer as divergências entre genitores ou grupos de genitores, entre os quais citam-se as medidas de

distância (dissimilaridade) genética; análise de componentes principais e variáveis canônicas; as técnicas de projeção gráfica; os métodos de agrupamentos de otimização ou aglomerativos; sequências hierárquicas e sem sobreposição e; as análises discriminantes paramétricas e não paramétricas (CRUZ et al., 2011).

A adoção de alguma dessas análises variam de acordo com o padrão de resultado desejado e com a informação disponível (DINIZ FILHO, 2000). Quando se trata de análise discriminante no estudo da diversidade genética, seu propósito é alocar um conjunto de genótipos em suas respectivas populações previamente definidos a qual pertencem.

A análise discriminante procura obter funções que permitam classificar um indivíduo, a partir das informações de um conjunto de características mensuradas, em uma entre várias populações conhecidas, buscando minimizar a probabilidade de má classificação. Assim, devem-se obter funções que permitam alocar um indivíduo na população à qual ele realmente pertence. Constada a eficácia da discriminação, as funções podem ser utilizadas para alocar novos indivíduos, dos quais se desconhece a origem (CRUZ et al., 2011)

Os métodos baseados em funções discriminantes de Fisher, de Anderson, quadráticas e de componentes principais são aqueles que têm sido usados comumente com caracteres quantitativos e requerem pressuposições e probabilidade específica de distribuição, sendo a multinormalidade a mais comum para o estudo das populações (CRUZ et al., 2011), e então, são métodos paramétricos.

Quando se desconhece a distribuição probabilística das variáveis avaliadas ou os dados não seguem normalidade, métodos não paramétricos de análises discriminantes servem como alternativa analítica, a exemplo daquelas características genéticas de marcadores moleculares ou descritores morfológicos, sendo elas de natureza (multi)categórica.

Para Torabi e Ding (1998) quando as amostras são pequenas, o teste não paramétrico deve ser preferido, a não ser que a condição relativa à normalidade seja verificada. Nesse caso, o método paramétrico pode ser utilizado.

Métodos paramétricos de análises discriminantes têm sido aplicados com frequência em espécies vegetais nos estudos da diversidade genética, quando avaliadas variáveis quantitativas, com destaque as funções discriminantes de Anderson em diferentes espécies de braquiária (ASSIS et al., 2003), em acessos de *Capsicum* spp. (SUDRÉ et al., 2006), em soja (NOGUEIRA et al., 2008), entre outras.

Os métodos não paramétricos de análise discriminante têm sido encontrados na literatura com menor frequência. A análise discriminante dos k-vizinhos mais próximos, abordado por Khattree e Naik (2000), foi relatada ter sido utilizada com sucesso em dados qualitativos, a exemplo dos estudos em *Scalesia divisa* e *S. incisa* (Asteraceae) (NIELSEN et al., 2003), batata doce (MCHARO, 2005) e mandioca (OLIVEIRA et al., 2012), todos com marcadores moleculares.

Outra abordagem não paramétrica é a do vizinho médio, presente no software GENES (CRUZ, 2016) e descrita por Cruz et al. (2011). Entretanto, sua aplicabilidade ou relatos em trabalhos científicos não foram encontrados.

Não se conhece a efetividade sobre o uso dos métodos de análise discriminante não paramétrica quando aplicados a dados fenotípicos quantitativos para a discriminação de genótipos nos estudos de diversidade genética e quando comparados aos métodos de análise que geram funções discriminantes. Como estes métodos não paramétricos baseiam-se nas estimativas de medidas de dissimilaridade (distância) genética entre pares de genótipo, cabe responder qual a influência destas medidas sobre o resultado final das classificações dos genótipos em suas populações? Ademais, se desconhece qual o melhor método não paramétrico de classificação com base em informações genéticas (ou fenotípicas): do vizinho

médio ou dos k-vizinhos mais próximos? Também é importante compreender como as alternâncias sobre o número k de vizinhos mais próximos, usados como critério para classificação e, a definição das probabilidades *a priori* das populações, podem interferir na taxa de acerto (ou erro) de classificação dos genótipos. Acredita-se que tais métodos não paramétricos possam classificar genótipos tão bem quanto os métodos paramétricos de análise de funções discriminantes. Deste modo, espera-se com o presente trabalho esclarecer as dúvidas metodológicas sobre a aplicabilidade destas técnicas não paramétricas de discriminação de genótipos e verificar suas potencialidades ou limitações e, então, contribuir com a literatura sobre quais destas técnicas resultam em informações satisfatórias quando aplicadas a dados fenotípicos quantitativos, caso apresentem comprovada eficácia.

2. OBJETIVOS

2.1. Objetivo geral

Avaliar a efetividade das metodologias de análise discriminante não paramétricas do vizinho médio e dos k-vizinhos mais próximos no estudo preditivo da diversidade no melhoramento genético, quando aplicadas à variáveis quantitativas, de modo a classificar satisfatoriamente os genótipos em suas respectivas populações definidas *a priori*.

2.2. Objetivos específicos

Comparar os métodos de análises discriminantes não paramétricos (vizinho médio e dos k-vizinhos mais próximos) com o método de análise discriminante paramétrico (análise discriminante de Anderson).

Avaliar a classificação dos genótipos em suas populações previamente definidas por meio de diferentes medidas de distância genética, usuais para dados quantitativos, quando associadas aos métodos não paramétricos de análise discriminante;

Estimar qual(is) o(s) valor(es) de k de vizinhos mais próximos e quais as probabilidades do genótipo pertencer a uma determinada população são mais adequados no uso da análise discriminante para classificá-los corretamente em suas respectivas populações;

Definir qual método de análise proporciona maiores níveis de acerto quanto à classificação dos genótipos em suas respectivas populações: o método discriminante do vizinho médio ou dos k-vizinhos mais próximos.

3. REVISÃO DE LITERATURA

3.1. Estudo da diversidade genética no melhoramento genético

A diversidade genética corresponde à variação gênica existente dentro de uma espécie ou entre espécies aparentadas. Quando expressa as diferenças de indivíduos de uma mesma espécie é denominada de variabilidade genética, quando de espécies diferentes, resulta na biodiversidade (CRUZ, 2005).

O estudo da diversidade genética no contexto do melhoramento genético se destina à identificação de genitores adequados à obtenção de híbridos com maior efeito heterótico e que proporcionem maior segregação em recombinações, possibilitando o aparecimento de transgressivos (CRUZ et al., 2003), devendo-se na seleção de genitores para cruzamentos, aliar o bom desempenho destes com a divergência genética entre eles.

Há duas maneiras básicas de se inferir sobre a diversidade genética: com técnicas biométricas de natureza quantitativa ou de natureza preditiva (CRUZ et al., 2011). Entre os métodos de natureza quantitativa de avaliação da diversidade, ou da heterose manifestada nos híbridos, citam-se as análises dialélicas. Nesses métodos é necessário a avaliação de p genitores e de todas (ou amostras) suas combinações híbridas, resultando num total de p (p – 1)/2 híbridos a serem avaliados (CRUZ et al., 2011).

Entre os métodos preditivos da heterose citam-se aqueles que tomam por base as diferenças morfológicas, fisiológicas ou moleculares, quantificando-as em alguma medida de dissimilaridade que expressa o grau de diversidade genética entre os genitores (CRUZ et al., 2011).

De maneira geral, estudos da diversidade genética têm sido realizados a partir de informações das seguintes medidas (CRUZ et al., 2011):

i. Medidas de dissimilaridade obtidas de variáveis quantitativas contínuas ou discretas;

ii. Medidas de dissimilaridade obtidas de variáveis qualitativas binárias;

iii. Medidas de dissimilaridade obtidas de variáveis qualitativas multicategóricas.

Funções de distância são necessárias em muitos algoritmos modernos. As medidas de distância de uma maneira geral podem ser definidas como medidas de similaridade, e dissimilaridade. A primeira é para definir o grau de semelhança entre as instâncias e realizam o agrupamento de acordo com a sua coesão, e a segunda mede as diferenças dos atributos das instâncias (SANTOS, 2015).

Neste ponto, deve ser considerado que uma característica (ou variável) é todo atributo mensurável em uma população, gerando para cada elemento (indivíduo ou família) um determinado valor. Seus valores variam de elemento para elemento e podem assumir grandezas numéricas ou não numéricas. Existem várias maneiras de classificar as variáveis em diferentes tipos, sendo que a mais comum considera dois grandes grupos chamados de variáveis: qualitativas e de quantitativas (CRUZ et al., 2011).

As variáveis classificadas como quantitativas são aquelas que podem ser medidas em escala real, podendo ser contínuas ou discretas. As variáveis contínuas são as que assumem, dentro de um intervalo finito, uma infinidade de valores, incluindo inteiros e fracionários. As variáveis discretas podem assumir apenas um número finito ou infinito contável de valores e, assim, são expressas por valores inteiros. As variáveis qualitativas (ou categóricas) são as que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias ou classes, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais. As variáveis nominais são aquelas em que não existe ordenação dentre as categorias (exemplo, cores variadas) ao contrário do que ocorre com as ordinais (exemplo, tonalidades). Tanto para variáveis do tipo nominal, quanto ordinal pode-se, por razões de conveniência, associar valores numéricos às diferentes categorias. Mas, é importante lembrar que estes valores numéricos não tem significado como tal, nem mesmo no caso de variáveis de tipo

ordinal. Por exemplo, pode-se associar os valores 1 e 2 às categorias tardio e precoce para a variável época de floração. Ou, os valores 1, 2 e 3 às categorias amarela, branca e roxa para cor do bulbo da cebola. Estes números são nada mais que símbolos para representar as categorias e não assumem grandeza e não representam distância métrica (FERREIRA, 2005).

Embora o volume de informações genéticas provenientes de marcadores moleculares, tenha aumentado em grandes proporções para os estudos da diversidade genética, continua-se a dar ênfase ao estudo da diversidade por meio de características fenotípicas, principalmente de natureza quantitativa. Essas características apresentam, geralmente, distribuição contínua, são determinadas por poligenes de pequenos efeitos e influenciadas pelo ambiente. Entretanto, são de grande interesse, tendo em vista sua importância econômica e a necessidade de grandes esforços para maximizar o êxito na escolha adequada de combinações híbridas, de modo a não comprometer o sucesso das estratégias de seleção (CRUZ et al., 2011).

As medidas para caracteres quantitativos mais utilizadas em estudos genéticos são: a distância Euclidiana, a distância Euclidiana média, o Quadrado da distância euclidiana, a distância ponderada e a distância generalizada de Mahalanobis. Os valores de distância são, geralmente, obtidos a partir de informações de g genótipos mensurados em relação a *v* caracteres (CRUZ et al., 2011). As quatro primeiras distâncias genéticas citadas não requerem dados oriundos de delineamentos estatísticos, pois não exigem matrizes de covariâncias e variâncias residuais entre as características mensuradas (CRUZ et al., 2003).

A distância Euclidiana sempre aumenta com o acréscimo do número de características consideradas na análise, e por isso, tem sido usada de forma alternativa, a distância Euclidiana média (CRUZ et al., 2003). Segundo Regazzi (2001), embora a distância Euclidiana seja uma medida de dissimilaridade, às vezes ela é referida como uma medida de

semelhança, pois quanto maior seu valor, menos parecidos são os indivíduos ou unidades amostrais.

O Quadrado da distância euclidiana média é outra forma de expressar a dissimilaridade entre dois genótipos e é preferida quando se deseja manter relação da distância genética entre dois genótipos com a soma de quadrados dos desvios (CRUZ et al., 2003).

A distância de Gower é outra medida interessante para se estimar a divergência genética a partir de dados fenotípicos quantitativos (CRUZ et al., 2011). Uma técnica que permite a análise simultânea de dados quantitativos e qualitativos foi proposta por Gower em 1971, por meio de um algoritmo que estima a similaridade entre dois indivíduos utilizando dados com distribuições contínuas e discretas. Esse tipo de análise tem sido muito utilizado em estudos relacionados à botânica e taxonomia, entretanto, ainda não tem sido bem explorado pelos pesquisadores na área de recursos genéticos vegetais para detecção da variabilidade em coleções de germoplasma.

Malhotra (2001) afirma que o emprego de diferentes medidas de distância pode induzir a resultados diferentes de aglomeração. Assim, é conveniente utilizar medidas diferentes e comparar os resultados. Em todas as distâncias aqui referidas a escala afeta o valor obtido. Adicionalmente, elas são quantificadas em diferentes medidas (peso, comprimento, porcentagem, etc.). Assim, é recomendável o cálculo das distâncias utilizando-se os valores padronizados (CRUZ et al., 2003).

Apesar de as técnicas multivariadas serem conhecidas a longo tempo, sua utilização em maior escala só se tornou possível com a disponibilidade dos recursos computacionais, que possibilitaram a avaliação simultânea de várias características e permitiram que inúmeras inferências pudessem ser feitas a partir do conjunto de dados existentes (CURI, 1983).

A adoção de uma técnica varia de acordo com o padrão de resultado desejado e com a informação disponível, seja ela característica morfológica, fisiológica, ecológica ou genético-molecular (DINIZ FILHO, 2000). Assim sendo, quando o propósito é investigar a diversidade genética por meio da discriminação dos genótipos em populações pré-estabelecidas e então realocá-los, sugere-se realizar as análises discriminantes (CRUZ et al., 2011).

Geralmente, durante a caracterização de uma população, são avaliados vários caracteres, o que gera um grande volume de dados. Para a análise desse banco de dados, as estatísticas multivariadas são de grande utilidade e podem ser aplicadas a um conjunto de caracteres correlacionados, para a caracterização genética de uma população (OLIVEIRA et al., 2007).

3.2. Análise discriminante para o estudo da diversidade genética

A análise discriminante foi inicialmente descrita por Fisher em 1936, em seu estudo sobre o uso de múltiplas medições em problemas taxonômicos, com cultivares de Iris. Abordando o problema da discriminação entre dois ou mais grupos, visando posterior classificação. Por meio de uma combinação linear de características mensuráveis, com um claro poder discriminatório entre populações (MARDIA et al., 1979). Fisher propôs funções matemáticas capazes de classificar um indivíduo *x* (uma observação *x*) em uma de várias populações π_i, $(i = 1, 2, \dots, g)$, com base em medidas de um número *p* de características, buscando minimizar a probabilidade de má classificação, isto é, minimizar a probabilidade de classificar erroneamente um indivíduo em uma população π_i, quando realmente pertence a população π_j, $(i \neq j)\ i, j = 1, 2, \dots, g)$ (REGAZZI, 2000).

Para Marriott (1974), a análise discriminante consiste em investigar como e quando é possível fazer distinções entre os membros de *g* agrupamentos, com base nas observações feitas sobre eles. Segundo o autor supracitado, os objetivos da análise discriminante são testar diferenças estatísticas, significantes a um dado nível de probabilidade, entre *g* agrupamento; determinar o número de funções discriminantes; construir regras de alocações para identificar um indivíduo como membro de um dos *g* agrupamentos e estimar as probabilidades de classificações corretas. Ainda a análise se presta para testar a suficiência de uma série de variáveis discriminantes (LANCHENBRUCH, 1979; MARDIA et al., 1979; RAO; MITRA, 1973).

Cruz et al. (2011) afirmam que a análise discriminante, procura obter funções que permitam classificar um indivíduo, a partir das informações de um conjunto de características mensuradas, em uma entre várias populações conhecidas, buscando minimizar a probabilidade de má classificação. Assim, devem-se obter funções que permitam alocar um indivíduo na população à qual ele realmente pertence. Constada a eficácia da discriminação, as funções podem ser utilizadas para alocar novos indivíduos, dos quais se desconhece a origem.

A análise discriminante tem como objetivo determinar a qual grupo entre dois ou mais definido *a priori*, pertence a um determinado elemento. A determinação é realizada considerando as características das variáveis aleatórias que contribui com informação para referida classificação (FAYYAD, 1996). A eficiência de uma técnica é proporcional à qualidade das informações disponíveis. Isto dá à fase de coleta de dados uma importância fundamental. Independente do método definido, se as variáveis forem selecionadas de forma inadequada acaba por comprometer a eficiência almejada. A Análise Discriminante combina variável em uma ou mais funções determinando os valores para a classificação. Estas funções são construídas de modo que os escores dos elementos de cada grupo se concentrem em torno

do valor médio do grupo, fazendo com que a superposição de escores de elementos de diferentes grupos seja minimizada (HAIR et al., 1995).

A alocação dos objetos ou unidades amostrais em uma das várias populações é um problema multivariado, por isso, é necessário produzir um índice ou um critério bem definido que possa ser utilizado com regra de classificação. Evidentemente, nenhuma regra de decisão será perfeita e, portanto, sempre haverá certa probabilidade de erro de classificação. Tal probabilidade, entretanto, pode ser controlada ou minimizada (KHATTREE; NAIK, 2000).

Nos problemas em análise discriminante os objetos são descritos por um conjunto de variáveis, selecionadas de forma que sejam capazes de diferenciar os objetos com relação às classes consideradas no problema a ser resolvido. Este conjunto de variáveis é denominado vetor de características, que denotamos por $X_T = (X_1, ..., X_p)$ que e suas variáveis são denominadas variáveis preditoras, cujas observações são as mensurações feitas sobre o objeto a ser classificado, quantificando assim características discriminantes, ou seja, aspectos relevantes para distinção de classes. Desta forma, uma das etapas do problema em análise discriminante consiste em observado o vetor de características para objeto cuja classe é desconhecida, associá-lo a uma das classes definidas para o problema que seja a mais apropriada ao valor do vetor de características apresentado pelo objeto. Para abordar esta questão, torna-se necessário desenvolver um procedimento que permita implementar esta alocação, denominado de classificador (COELHO, 2013).

A aplicação da análise discriminante pode ser vista como a construção de um modelo de seis estágios (HAIR et al., 2005):

i. Estabelecimento dos objetivos, avaliando diferenças de grupos em um perfil multivariado e identificando dimensões de discriminação entre grupos;

ii. Planejamento da pesquisa, selecionando as variáveis independentes, determinando o tamanho da amostra total e criando as amostras de análise e de teste;

iii. Certificação das suposições inerentes ao modelo, como normalidade e ausência de multicolinearidade das variáveis independente, linearidade das relações e igualdade das matrizes de dispersão;

iv. Estimação das funções discriminantes e de suas significâncias estatísticas, avaliação da precisão preditiva através da aplicação de uma matriz de classificação;

v. Interpretação das funções discriminantes, determinando quantas funções serão interpretadas e quais variáveis independentes mais contribuem para distinguir os grupos;

vi. Validação dos resultados discriminantes através da utilização de amostra de teste ou da validação cruzada.

A combinação linear para uma análise discriminante, também conhecida como função discriminante, é determinada por uma equação que assume a seguinte forma (HAIR et al., 2005):

$$Z_{jk} = a + W_1 W_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk}$$

Onde:

Z_{jk}: escore *Z* discriminante da função discriminante *J* pra o objeto *K*.

a: constante.

W_{i2}: peso discriminante para a variável independente *i*.

X_{ik}: variável independente *i* para o objeto *k*.

Calculando-se a média dos escores discriminantes para todos os indivíduos dos grupos, obtém-se a média do grupo. Esta média de grupo é chamada centroide e existe uma para cada grupo envolvido na análise. Os centroides indicam o local mais típico de qualquer indivíduo de um grupo particular, e suas posições relativas ao longo da dimensão testada indicam o quão afastados ou divergentes são os grupos avaliados (HAIR et al., 2005).

Ao utilizar procedimentos discriminantes que assumem normalidade multivariada, deve-se primeiramente examinar tal pressuposição. Em situações de não normalidade, a utilização de procedimentos discriminantes que assumem distribuição normal, podem levar a resultados ilusórios e problemas na estimação das funções discriminantes. A opção seria tentar fazer uma transformação para alcançar a normalidade aproximada, ou melhor, optar por métodos de análise discriminantes não paramétricos. (KHATTREE; NAIK, 2000; HAIR et al., 2005).

Khattree e Naik (2000) apresentam a utilização de testes de normalidade multivariada para examinar a validade desta pressuposição, baseado na curtose multivariada de Mardia e na opção gráfica denominada *Q – Q plots*. Há também a possibilidade de se verificar a normalidade através da construção de histogramas para cada variável em cada população e, então, complementar esta análise visual com estatísticas que reflitam a forma da distribuição (assimetria e curtose), bem como um teste estatístico de normalidade.

Hair et al. (2005), relatam que outra questão que pode afetar os resultados de uma análise discriminante (linear) é a multicolinearidade entre as variáveis. Essa consideração se torna parcialmente crítica quando procedimentos de seleção de variáveis (*stepwise*) são empregados. O pesquisador ao interpretar a função discriminante deve estar ciente do nível de multicolinearidade e seu impacto na determinação e seleção de variáveis de maior poder discriminante.

Cabe destacar na literatura três metodologias: análise discriminante linear de Fisher, análise discriminante de Anderson e análise discriminante Canônica. Basicamente, as duas últimas técnicas é que têm sido comumente empregadas nos estudos genéticos em melhoramento de plantas (CRUZ et al., 2011).

A análise discriminante de Fisher, é uma combinação linear das características observadas que apresenta melhor poder de discriminação entre os grupos, constitui a base de

todo o estudo na análise discriminante. Esta função tem a propriedade de minimizar a probabilidade de má classificação, quando as populações apresentam média e variância conhecidas. Contudo, tal situação pode não ocorrer na prática, necessitando-se, portanto de estimativas e métodos de estimação dessas probabilidades ótimas (CRUZ et al., 2012).

Na análise discriminante proposta por Anderson em 1958, consideram-se as informações de indivíduos sabiamente pertencentes a diferentes populações. A partir destas informações são geradas funções, que são combinações lineares das características avaliadas e que têm por finalidade promover a melhor discriminação entre indivíduos, alocando-os em suas devidas populações. Estas funções, uma vez estimadas, passam a ser de grande utilidade, por permitirem classificar novos materiais genéticos, de comportamento conhecido, nas populações já conhecidas. A eficácia das variáveis utilizadas em promover a discriminação também é avaliada, permitindo conhecer a adequação da função estimada (CRUZ et al., 2011).

A análise discriminante Canônica é uma técnica de redução de dimensão semelhante à técnica de componentes principais, pois permite a simplificação no conjunto de dados, resumindo as informações, originalmente contidas em um grupo de *n* variáveis, em poucas variáveis, que apresentam as propriedades de reterem o máximo da variação originalmente disponível e serem independentes entre si (CRUZ et al., 2003)

Uma vez obtida às funções discriminantes, é fundamental avaliar a sua eficácia, que é dependente do grau de dissimilaridade entre as populações analisadas e, principalmente da quantidade e qualidade das variáveis consideradas na discriminação. Como as funções discriminantes são obtidas a partir de análises prévias de observações que se supõe serem, de fato, pertencente às populações consideradas, pode-se calcular a probabilidade de má classificação, reclassificando toda observação até então disponível. A classificação de uma observação pertencente a uma população π_j em outra é indicativo de menor eficiência da

função discriminante estimada, contribuindo para o acréscimo na taxa de erro aparente (CRUZ et al., 2011).

Segundo Huberty (1994) a análise discriminante implica na estimativa das densidades de probabilidades específicas nas diferentes populações. Para estimar essas densidades específicas, utilizam-se duas abordagens, a paramétrica e a não paramétrica. Em relação à escolha da função discriminante, segundo Webb (2002), ela pode depender do conhecimento prévio dos padrões que serão utilizados no processo de classificação ou pode-se optar por utilizar uma forma funcional específica com parâmetros estimados utilizando o conjunto de treinamento. Na abordagem paramétrica assume-se que os dados seguem uma distribuição normal.

Os métodos de análises discriminantes normalmente empregados na classificação de genótipos, a partir de dados moleculares binários, são ditos não paramétricos, uma vez que não é assumida nenhuma forma paramétrica de distribuição para a função de densidade das variáveis analisadas (KHATTREE; NAIK, 2000).

A técnica de análise discriminante não paramétrica consiste em, incialmente, estimar a medida de dissimilaridade entre cada indivíduo estudado, considerando um índice apropriado. Para isso, é necessário também estabelecer a probabilidade *a priori* inerente às várias populações avaliadas em um determinado estudo. E caso não haja informações prévias para classificação dos indivíduos, pode-se pressupor que as probabilidades sejam iguais para todas as populações analisadas. Outra opção é admitir que as probabilidades *a priori* sejam proporcionais ao tamanho de cada população (CRUZ et al., 2011).

Uma das técnicas de análises discriminantes que vêm sendo utilizadas com sucesso no melhoramento genético são o método dos k-vizinhos mais próximos e o método do vizinho médio (CRUZ et al., 2011).

O método dos k-vizinhos mais próximos foi proposto inicialmente por Fix e Hodges em 1951, em um relatório técnico não publicado da School of Aviation USAF Medicine, sendo responsáveis pela utilização de regras da alocação de vizinhos mais próximos para análises discriminantes não paramétricas (SILVERMAN et al., 1989). O método dos k-vizinhos mais próximos é baseado em certo critério envolvendo distâncias entre indivíduos imediatos, ou seja, dado um grupo de observações a serem classificadas, um algoritmo procura, para uma particular observação, aquelas que lhe sejam mais próximas. Assim, cada observação é alocada na classe que contenha a maior proporção de k-vizinhos mais próximos (KHATTREE; NAIK, 2000; BEHARAV; NEVO, 2003).

A regra de classificação, a função que calcula a distância entre dois pontos e a escolha do valor de k, são três parâmetros importantes no método dos k-vizinhos mais próximos . A regra de classificação diz respeito à relevância de cada um dos k elementos selecionados. A função de distância mede a distância no espaço multidimensional. A escolha do valor de k permite escolher qual a fronteira na vizinhança do padrão a ser classificado a ser utilizada na classificação (SANTOS, 2015).

O método do vizinho médio, proposto por Cruz et al. (2011) é baseado na alocação de um indivíduo em uma população de acordo com a classe de seus vizinhos mais próximos, seguindo critérios específicos (algoritmo, medida de distância genética). Com base na média de todas as distâncias genéticas possíveis de serem estimadas entre os indivíduos, com exceção do indivíduo com ele mesmo, independentemente da população a qual ele pertence, define-se a população de que ele esteja mais próximo, alocando-o nesta (CRUZ et al., 2011; CRUZ, 2016).

As aplicações de análise discriminante ocorrem em diversas áreas de estudos em ciência e tecnologia. Engenharia, biologia, psicologia, medicina, marketing, visão computacional, sensoriamento remoto, inteligência artificial, são alguns exemplos de áreas em

que há necessidade de classificar objetos em classes definidas para o problema considerado. Exemplos de algumas situações (COELHO, 2013): Detectar tipos de poluição industrial, num espaço geográfico; Detectar células anormais em imagens digitais de amostras de sangue; Identificar peças defeituosas em um processo de produção por meio de imagens digitais; Identificar alvos por meio de sinais de radar; Classificar plantas em diferentes espécies; Identificar suspeitos de crimes por meio da impressão digital; Classificar assinaturas em imagens de documentos como falsas ou verdadeiras; Classificação de diferentes tipos de solo em imagens de satélites.

Para as diversas situações, que são muitas vezes solucionadas com esforço humano, muitas pesquisas em ciência e tecnologia almejam resolvê-las de maneira mais prática, automatizando tanto quanto possível os procedimentos necessários. Com a disponibilidade cada vez maior de recursos computacionais, muitos destes processos de automatização já são perfeitamente viáveis (JAIN et al., 2000; HASTIE et al., 2009).

3.3. Aplicações das análises discriminantes no melhoramento genético de plantas

Ferreira et al. (1995), discriminaram 20 genitores de arroz em tolerantes e sensíveis à toxidez de alumínio, com quatro características fenotípicas (comprimento da raiz, peso da matéria seca da raiz e da parte aérea e altura de plantas) a partir das funções discriminantes de Anderson. Foi necessário excluir a variável peso da matéria seca total, pois esta é estabelecida pela soma de pesos de matéria seca da raiz e parte aérea e proporcionava uma matriz de (co)variâncias singular, em função da multicolinearidade que se manifesta. Com os resultados, concluiu-se, em geral, que as cultivares de arroz de sequeiro são tolerantes a toxidez de alumínio, e as irrigadas, são sensíveis.

Moltalván et al. (1998), utilizaram as concentrações relativas de 16 frações de proteínas de grãos como variáveis para a estimativa das funções discriminantes de Fisher, entre 58 genótipos de arroz brasileiro e nove japoneses. Com o teste T^2 de Hotelling, diferenças significativas foram encontradas entre as frações proteicas dos grupos brasileiro e japonês, assim como entre os genótipos melhorados e não melhorados do Brasil. Os autores puderam determinar através dos coeficientes de ponderação as frações proteicas que tiveram contribuições importantes no processo de discriminação. Contudo, alertaram que o uso individualizado ou conjunto somente dessas variáveis foi insuficiente para diferenciar de maneira satisfatória os grupos. O estudo permitiu indicar a origem geográfica e o nível de melhoramento dos cultivares.

Em outro trabalho Ebdon et al. (1998), avaliaram a eficácia da análise discriminante de Anderson e quadrática em distinguir 61 cultivares de capim-do-prado (*Poa pratensis* L., KGB) em relação ao padrão de uso de água (baixo e alto), com a premissa de que provavelmente esses dois grupos iriam diferir também quanto as suas propriedades morfológicas (14 características), e gerar funções de classificações capazes de reconhecer as diferenças destes padrões entre os grupos. As probabilidades *a priori* ($\pi_1 = \pi_2 = 50\%$) foram consideradas para interpretar os resultados da classificação, ou seja, essas são as probabilidades esperadas em relação à má classificação. Os autores selecionaram as variáveis com maior poder discriminatório pelo processo *stepwise*, verificando a taxa de classificação correta através da taxa de erro aparente e validação cruzada. Ao utilizarem números diferentes de variáveis (1 a 7) puderam verificar em cada caso quais as melhores proporções de classificações corretas, entre as funções lineares e quadráticas. Os resultados mostraram que com apenas duas das variáveis estudadas, a função linear teve a melhor proporção de casos classificados corretamente (75,4%) a partir da validação cruzada. Eles ressaltaram que os coeficientes de ponderação refletiram importantes interpretações biológicas na classificação

de uma observação baseada nos escores discriminantes e concluíram que a análise discriminante foi uma ferramenta eficiente e útil na predição do padrão de uso de água de novas cultivares com base em poucas variáveis rotineiramente avaliadas pelos melhoristas.

Vaylay e Van Saten (2002) investigaram a diversidade genética de cultivares da espécie forrageira *Festuca arundinacea* Schreb. em resposta as forças da seleção natural usando a análise discriminante canônica. Do ponto de vista do melhoramento de plantas, os autores concluíram que tal análise é útil em identificar a variação genética e as características que mais afetam a variação genética de populações de plantas. As cargas canônicas das características morfológicas e agronômicas de uma cultivar indicam a magnitude da variação genética. Segundo eles, as características importantes são aquelas que respondem as forças da seleção natural.

Bantte e Prasanna (2003) utilizaram a análise discriminante canônica com 23 linhagens QPM (alto teor de proteína) de milho, das quais 13 eram linhagens endogâmicas da Índia (DMRQPM) e 10 eram linhagens endogâmicas do México (CIMMYT). Com a análise de agrupamento o grupo de linhagens do CYMMIT foi subdivido em três grupos. Os grupos foram discriminados com base no comprimento das bandas de quatro locos microssatélites que se encontravam dentro de uma faixa de leitura no gel de agarose de 100 bp (peso molecular). O padrão de classificação foi concordante com as informações de pedigree.

Fonseca et al. (2004) averiguaram a adequação da composição de três variedades clonais, a partir de 32 clones de *Coffea canephora* recomendadas para o Espírito Santo com base nas funções discriminantes de Anderson e características agronômicas. Os autores relataram que algumas das características avaliadas foram descartadas, em virtude da existência de multicolinearidade. A classificação original dos genótipos nas três variedades estudadas manteve expressiva concordância com os resultados obtidos pela análise discriminante, com uma taxa de erro aparente (TEA) de apenas 6,25%. Os dois únicos clones

maus classificados foram realocados. Assim, as funções discriminantes corrigidas foram propostas a novas classificações em uma das três populações em questão, a serem utilizadas em programas de melhoramento, eliminando, segundo os autores, a subjetividade do processo de agrupamento.

Nielsen et al. (2003) alocaram 112 indivíduos em seis subpopulações, sendo duas para cada espécie de *Scalesia divisa* e *S. incisa* e as outras duas de uma população diferenciada. Os pesquisadores utilizaram características morfológicas e marcadores AFLP (polimorfismo de comprimento de fragmentos de amplificação). Optou-se pela análise discriminante não paramétrica do vizinho mais próximo, sendo as marcas codificadas em presença (1) e ausência (0). Os altos níveis de classificação incorreta foram encontrados em duas populações, o que auxiliou os autores, junto a outros tipos de análises, a concluir sobre o padrão de variação das espécies.

Mcharo (2005) a partir de clones de batata doce, validou a análise discriminante não paramétrica com informações moleculares para a diferenciação de grupos quanto a: resistência e susceptibilidade a nematoide e alto e baixo teor de açúcar, em populações com pequeno tamanho amostral. O autor, ao selecionar marcas com maior poder discriminatório, sugeriu uma ligação entre elas e as características avaliadas (resistência a nematoide e teor de açúcar), e consequentemente, estas marcas poderiam ser usadas na seleção assistida por marcadores.

Oliveira et al. (2012) realizaram a caracterização molecular e avaliaram a diversidade de 17 novos clones de mandioca em relação a aptidão comercial da cultura, ou seja, para o processo industrial e consumo humano "*in natura*". Os autores obtiveram 13 *primers* microssatélites, cujas marcas serviram como variáveis para atingir os objetivos do estudo. Utilizando o método dos k-vizinhos mais próximos, com número $k = 3$, foi possível discriminar os clones corretamente nos grupos mandioca "mesa"e mandioca indústria, tendo

apenas uma classificação diferente (errada) conforme a alocação *a priori* dos clones em seus grupos pré-definidos.

Anderson et al. (2003) descreveram o método canônico de análise discriminante generalizada com base em uma matriz de dissimilaridade para testar diferenças em grupos a priori de observações multivariadas. Em seus estudos apresentaram estatísticas de teste e suas distribuições de permutação assintótica para uma análise canônica com base em matrizes de dissimilaridade geral. Propondo um novo método para colocar uma nova observação no espaço canônico, com base apenas em distâncias entre pontos, e damos critérios para escolher o número apropriado de eixos de coordenadas principais a serem usados para a análise canônica. Demostrando o uso desta abordagem para a ordenação, testes canônicos e classificação, com exemplos ecológicos. Ao final de seus estudos afirmaram que para qualquer situação em que a informação ou os dados sejam talvez melhor representados por uma matriz de distância ou de dissimilaridade ou onde também existam Muitas variáveis para uma análise tradicional, a análise discriminante generalizada baseada em distâncias que propomos aqui pode ser usada para resultados frutíferos e interpretação.

Devillard et al. (2010) em seus estudos, apresentaram a Análise Discriminante de Componentes Principais (ADCP), um método multivariante concebido para identificar e descrever grupos de indivíduos geneticamente relacionados. Com conjuntos de dados simulados avaliamos o desempenho do método, que também foram analisados usando software STRUCTURE como referência. Além disso, ilustramos o método analisando o polimorfismo de microsatélites em populações humanas em todo o mundo e a variação da sequência gênica da hemaglutinina na gripe sazonal. A análise de dados simulados revelou que a ADCP mostrou-se tão precisa como o software STRUCTURE na detecção de grupos ocultos de população dentro de modelos simples da população insular. Além disso, o ADCP foi mais adequado para desvendar a estruturação subjacente em modelos de genética

populacional mais complexos. Outra vantagem importante do ADCP sobre as abordagens de agrupamento bayesiano é a possibilidade de gerar uma representação gráfica da relação entre os clusters inferidos. Aplicado a dois conjuntos de dados empíricos altamente contrastados, nosso método foi capaz de identificar padrões biológicos não significativos e significativos. Um dos principais ativos da ADCP é a sua grande versatilidade. De fato, a ADCP não depende de um modelo particular de genética populacional e, portanto, está livre de pressupostos sobre equilíbrio de Hardy-Weinberg ou desequilíbrio de ligação. Como tal, deve ser útil para uma variedade de organismos, independentemente da sua ploidia e taxa de recombinação genética. Além disso, contrariamente aos métodos de agrupamento bayesiano, o ADCP pode ser aplicado a conjuntos de dados muito grandes dentro de um tempo computacional insignificante (todas as análises apresentadas neste documento demoraram menos de um minuto para serem executadas em um computador padrão). Além disso, o método não é restringido aos dados genéticos e pode ser aplicado a qualquer dado quantitativo, como dados morfométricos. Este recurso é particularmente interessante porque permite dividir os efeitos de covariáveis indesejáveis, como diferentes protocolos de seqüência, ou estruturas genéticas triviais que poderiam obscurecer padrões menores e mais interessantes.

Ivoglo (2007) estudando a divergência genética entre 21 progênies de meios-irmãos de *Coffea canephora*, em relação a 14 características morfo-agronômicas, com o intuito de indicar as progênies mais divergentes para a definição de populações-base para os programas de seleção e produção de híbridos. Utilizando dados amostrais obtidos de um experimento instalado no Polo Regional do Nordeste Paulista (APTA Regional/Mococa, SP), com delineamento de blocos ao acaso, com 21 tratamentos (progênies) e 24 repetições, no espaçamento de 4,0 x 3,0 m, com uma planta útil por parcela. A divergência genética foi estudada por procedimentos multivariados empregando-se a distância generalizada de

Mahalanobis e os métodos de agrupamento de Tocher, UPGMA e dispersão gráfica no plano tridimensional. Em todos os caracteres estudados verificaram-se diferenças significativas ($p < 0,01$ ou $p < 0,05$), indicando a existência de variabilidade genética entre as progênies para todas as características avaliadas. As médias de todas as características foram comparadas pelo teste de Scott-Knott a 5% de probabilidade. No estudo da divergência genética observou-se, pela distância generalizada de Mahalanobis, dissimilaridade entre os genótipos variando de 1,17 a 19,65. O método de Tocher reuniu as 21 progênies em quatro grupos distintos. Os grupos 1 e 2 foram subdivididos em quatro e três subgrupos, respectivamente, para melhor representar a divergência entre as progênies. Nos métodos UPGMA e projeção das distâncias no plano 3D, verificou-se concordância parcial com o método de Tocher.

4. MATERIAIS E MÉTODOS

4.1. Conjuntos de dados

A execução dos métodos de análise discriminante, vizinho médio e k-vizinhos mais próximos, foi realizada a partir de dados reais. O primeiro conjunto de dados foi referente a 83 genótipos (matrizes) de pupunha (*Bactris gasipaes*), oriundas da Embrapa em Porto Velho - RO, e sete variáveis (físicas e químicas) do fruto (pupunha) mensuradas – massa do fruto (em g), polpa (%), matéria seca (%), teores de óleo (%), fibras (%), cinzas (%) e proteínas (%). Inicialmente, as matrizes foram classificadas conforme a definição de raças primitivas (MORA URPÍ; CLEMENT, 1988), baseada no peso dos frutos, nas categorias microcarpa (43 matrizes), mesocarpa (32 matrizes) e macrocarpa (8 matrizes), (SANTOS et al., 2017). Denominou-se este conjunto de dados como PUPUNHA.

O outro conjunto de dados referiu-se a 122 genótipos (clones) de cafeeiro (*Coffea canephora*), oriundos da Embrapa em Porto Velho - RO, em que foram avaliadas dez variáveis agronômicas, a saber: altura de plantas a partir do nível do solo (em m), número de ramos plagiotrópicos produtivos; número de rosetas por ramo plagiotrópico; comprimento do ramo plagiotrópico (em m); distância entre rosetas da parte intermediária do ramo plagiotrópico (em cm); número de grãos por roseta da parte intermediária do ramo plagiotrópico; comprimento e a largura das folhas (em cm); época de maturação com o registro da data de colheita (em dias) e; o valor genotípico da produção de grãos de café beneficiado (em sacas de 60 kg por hectare). Estes clones foram previamente categorizados em variedades botânicas Conilon (72 clones) e Robusta (28 clones), e em Híbridos intervarientais (22 clones) obtidos em cruzamentos naturais entre estas variedades botânicas. Denominou-se este conjunto de dados como CAFÉ.

4.2. Análises estatísticas

4.2.1. Descrições populacionais

Admitiu-se a multinormalidade para os conjuntos PUPUNHA e CAFÉ, a partir dos testes de Lillieffors (LILLIEFORS, 1967), assimetria e curtose (P > 0,01).

Testou-se também a dissimilaridade entre pares de populações (raças primitivas para pupunha e variedades botânicas para café), admitindo-se a multinormalidade sobre as variáveis avaliadas, com matriz de variâncias e covariâncias comum, cujo teste foi efetuado por meio de (CRUZ et al., 2014):

$$F_{calc} = \frac{n_l + n_{l'} - v - 1}{v(n_l + n_{l'} - v - 2)} \frac{n_l n_{l'}}{n_l + n_{l'}} D^2_{ll'}$$

Em que:

$D^2_{ll'}$: refere-se a distância de Mahalanobis entre as populações P_l e $P_{l'}$, para todo $l \neq l$`; n_l e $n_{l'}$ representam os tamanhos amostrais das respectivas populações, ou seja, o número de genótipos pertencentes às populações P_l e $P_{l'}$, respectivamente; v é o número de variáveis consideradas.

O valor da estatística F calculada (F_{calc}) tem distribuição F com g_1 e g_2 graus de liberdade, em que $g_1 = v$ e $g_2 = n_l + n_{l'} - v - 1$. Testou-se em nível de significância de 5%.

Neste caso, quando o teste é significativo indica para o par de populações que elas são bastante distintas e, portanto, a classificação de um novo genótipo em uma das populações será com maior probabilidade de acerto (CRUZ et al., 2014).

4.2.2. Análises discriminantes

Foram executados três métodos de análises discriminantes para cada conjunto de dados: um método paramétrico (análise discriminante linear de Anderson, 1958) e dois não paramétricos (k-vizinhos mais próximos de Fix e Hodges, 1951; e vizinho médio, proposto por Cruz et al. 2011), descritos a seguir.

4.2.2.1. Análise discriminante paramétrica

a) Funções discriminantes lineares

Realizou-se a análise discriminante de Anderson (1958), para relacionar as variáveis com a classificação original estabelecida *a priori* em PUPUNHA e CAFÉ, e poder comparar tais resultados de classificação com as técnicas de discriminação não paramétricas. A função discriminante para uma população é definida por:

$$D_l(\tilde{y}) = \ln(\pi_l) + (\tilde{y} - \frac{1}{2}\mu_l)\sum^{-1}\mu_l$$

Em que:

$D_l(\tilde{y})$: escore de classificação da l-ésima população; $\sum^{-1}$: inversa da matriz de covariâncias; $\tilde{y}$ = vetor de genótipos das populações em análise; μ_l: vetor de médias; π_l: probabilidade *a priori* de um genótipo pertencer a l-ésima população.

O critério de decisão para alocar o genótipo em uma das populações é definido pelo maior escore de classificação estimado, sendo: $D_l(\tilde{y}) = \text{máx}\left[D_{1\tilde{y}}, D_{2\tilde{y}}, D_{3\tilde{y}}\right]$.

4.2.2.2. Análises discriminantes não paramétricas

a) Pelo método do vizinho médio

Primeiramente foram estimadas todas as distâncias genéticas entre os genótipos, para todo $i \neq i`$ e, então, calculou-se as dissimilaridades médias do i-ésimo genótipo em relação a l-ésima população, expressas por:

$$\overline{D}_{il} = \frac{\sum_{i=1}^{n_l} d_{ii'}}{n_l}$$

Em que:

n_l: é o tamanho da l-ésima população; $d_{ii'}$ é a medida de dissimilaridade do genótipo *i*, a ser classificado, em relação ao genótipo *i'* da l-ésima população.

Com base na dissimilaridade média do genótipo *i* com cada população, o mesmo foi alocado naquela cujo valor de $\overline{D}_{il}$ foi o menor, ou seja, decidiu-se alocar o genótipo na população quando: $\overline{D}_{il} = \text{mín}\,[\overline{D}_{i1}, \overline{D}_{i2}, \overline{D}_{i3}]$.

As dissimilaridades do genótipo com ele mesmo (d_{ii}) foram descartadas.

b) Pelo método dos k-vizinhos mais próximos

Inicialmente foram estimadas todas as dissimilaridades ($d_{ii'}$) entre os genótipos, para todo $i \neq i'$. Após isto, definiu-se o valor de *k*, que representa o número de genótipos (vizinhos) mais próximos de um genótipo qualquer do conjunto de dados que se deseja alocar em uma das populações estudadas. Os vizinhos mais próximos são aqueles genótipos com menor distância genética em relação ao genótipo a ser classificado. Assim, o valor *k* estabeleceu o número máximo de genótipos mais próximos que se poderia obter e, consequentemente, ajudou a definir qual população o genótipo foi (re)alocado.

Dentre esses *k* genótipo mais próximos, k_l podem ser provenientes de uma das L populações, cuja probabilidade *a priori* de um genótipo a ela pertencer foi π_l. Então, a probabilidade de um genótipo pertencer a l-ésima população foi estimada por:

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

Em que:

n_l: é o número de genótipo da l-ésima população; k_l é o número de vizinhos mais próximos do genótipo *i* (Y_i) da l-ésima população, dentre os k-vizinhos mais próximos encontrados. Seja *l* = 1, 2,..., L, sendo L = 3 populações (tanto para PUPUNHA quanto CAFÉ).

O genótipo Y_i foi (re)alocado na l-ésima população quando $\hat{P}(Y_i, P_l)$ foi a maior probabilidade entre as L populações avaliadas.

4.2.3. Medidas de dissimilaridade utilizadas

Para a execução dos métodos não paramétricos de análise discriminante propostos, considerando as variáveis quantitativas avaliadas, foram utilizadas quatro diferentes medidas de dissimilaridade (distâncias), conforme descrição a seguir:

a) Distância Euclidiana

A distância Euclidiana entre o par de genótipos *i* e *i'* se dá por meio da expressão:

$$d'_{ii} = \sqrt{\sum_{j} \left(Y_{ij} - Y_{i'j}\right)^2}$$

Em que:

Y_{ij}: refere-se a observação no i-ésimo genótipo para a j-ésima característica.

b) Distância Euclidiana Média

A distância Euclidiana média entre o par de genótipos *i* e *i'* se dá por meio da expressão:

$$d_{ii'} = \sqrt{\frac{1}{v}\sum_{j}\left(Y_{ij} - Y_{i'j}\right)^2}$$

Em que:

v: é o número de variáveis avaliadas.

c) Quadrado da Distância Euclidiana Média

O Quadrado da distância euclidiana média entre o par de genótipos *i* e *i'* se dá por meio da expressão:

$$d^2_{ii'} = \frac{1}{v}\sum_{j}\left(Y_{ij} - Y_{i'j}\right)^2$$

d) Distância de Gower

A distância entre o par de genótipos *i* e *i'* se dá por meio da expressão:

$$d_{ii'} = \frac{1}{v}\sum_{j=1}^{V}\frac{\left|Y_{ij} - Y_{i'j}\right|}{R_j}$$

Em que:

R_j: é a amplitude total de variação verificada para a j-ésima característica, ou seja, o valor máximo observado subtraído do valor mínimo da característica.

Para todas as medidas (distâncias) foram utilizados valores observados padronizados (Y^*_{ij}), a fim de evitar que a escala de medida para cada característica afetasse o valor da

dissimilaridade entre dois genótipos quaisquer. Ademais, com a padronização, as variáveis contribuíram igualmente na avaliação da dissimilaridade entre indivíduos. Padronizou-se por meio de:

$$Y_{ij}^{*} = \frac{Y_{ij}}{\hat{\sigma}_j}$$

Em que:

Y_{ij}: é o valor observado do i-ésimo genótipo para a j-ésima característica e; $\hat{\sigma}_j$ é o desvio-padrão associado a j-ésima característica.

4.2.4. Cenários para execução das análises discriminante

Para as três técnicas de análises discriminantes definiu-se as probabilidades *a priori* de um genótipo pertencer a uma determinada população (π_l) da seguinte forma: *i)* iguais (1/L) e *ii)* proporcionais ao tamanho das populações (n_l/N), em que N (= $\sum_{l=1}^{L} n_l$), corresponde ao total de genótipos avaliados em cada conjunto de dados.

Deste modo, tanto para o conjunto PUPUNHA quanto CAFÉ as probabilidades *a priori* iguais foram de 1/3 (= 0,3333). As probabilidades *a priori* proporcionais foram definidas da seguinte maneira: no conjunto PUPUNHA, as raças assumiram valores de π_1= 0,5180, para microcarpa; π_2= 0,3855, para mesocarpa e; π_3= 0,0963, para macrocarpa. No conjunto CAFÉ, as variedades assumiram os valores π_1= 0,5901, para Conilon; π_2 = 0,2295, para Robusta e; π_3 = 0,1803, para Híbridos intervarientais.

Para o método k-vizinhos mais próximos, os valores de k variaram de 1 até $n_l - 1$, sendo n_l o tamanho da menor população avaliada.

Sendo assim os cenários testados para discriminar as populações de PUPUNHA e de CAFÉ e para avaliar a efetividade das técnicas de análise discriminantes não paramétricas em alocar corretamente os genótipos em suas respectivas populações, variaram conforme as possibilidades de combinações: (*técnica de análise discriminante – três*) x (*medida de distância genética - quatro*) x (*valores de* π_i *- dois tipos*) x (*valor de k – sete para PUPUNHA e vinte e um para CAFÉ*), o que permitiu gerar 66 cenários para os dados PUPUNHA (dois cenários para análise discriminante linear, oito para o método do vizinho médio e 56 para o método k-vizinhos mais próximos) e 178 cenários para os dados CAFÉ (dois cenários para análise discriminante linear, oito para o método do vizinho médio e 168 para o método k-vizinhos mais próximos).

4.2.5. Eficiência das análises discriminantes

Estimou-se a taxa de erro aparente para medir a eficiência dos métodos avaliados. A soma de todos os casos desfavoráveis (má classificações) encontrados em cada população forneceu a taxa de erro aparente (TEA), dada por:

$$\frac{\sum_{l=1}^{L} m_l}{N}$$

Em que:

m_l: é o número de genótipos classificados erroneamente na l-ésima população, uma vez que estavam alocados previamente em outra.

Para todas estas análises foi utilizado o software GENES, versão 1990.2017.26 (CRUZ, 2016).

5. RESULTADOS E DISCUSSÃO

5.1. Raças primitivas de pupunha

5.1.1. Caracterização das populações

As raças primitivas de pupunha (microcarpa, mesocarpa e macrocarpa) apresentaram dissimilaridade multivariada significativa ($p < 0,01$), o que evidencia a diferença entre as populações dos conjuntos de dados PUPUNHA (Tabela 1).

É procedimento primordial verificar se as populações são diferentes o suficiente para que a análise discriminante tenha algum significado, quando admitida a distribuição multinormal (KHATTREE; NAIK, 2000).

O teste significativo para dissimilaridade de populações por meio da distância generalizada de Mahalanobis é indicativo da real distinção entre elas e, portanto, a classificação de um novo genótipo em uma das populações se fará com maior chance de sucesso (CRUZ et al., 2014).

Embora exista diferença estatística entre as populações microcarpa e mesocarpa, a estimativa da distância de Mahalanobis entre este par revelou ser a menor ($D^2 = 2,24$) dentre as demais comparações (Tabela 1). Para as raças microcarpa x mesocarpa, estimou-se a maior distância.

TABELA 1. Dissimilaridades entre os pares de pupunha (*Bactris gasipaes*) das diferentes raças primitivas microcarpa, mesocarpa e macrocarpa expressas pela distância generalizada de Mahalanobis (D^2) e testado pelo teste F a 1% de probabilidade.

Par de Populações	**D^2**	**G.l***	**$F_{calculado}$**	**P > F (%)**
Microcarpa x Mesocarpa	2,24	7, 67	5,40	0,00
Microcarpa x Macrocarpa	9,44	7, 43	7,98	0,00
Mesocarpa x Macrocarpa	6,32	7, 32	4,86	0,08

*Grau de liberdade.

A análise univariada sobre a existência de diferença entre estas populações de pupunha demonstrou haver variabilidade significativa ($p < 0,05$) entre as raças primitivas para seis das sete variáveis avaliadas, sendo a exceção o teor de cinzas (SANTOS et al., 2017). Neste mesmo estudo, a análise da diversidade genética via componentes principais (com capitação de 79% da variação total) permitiu agrupar bem as matrizes designadas como microcarpa, mesocarpa e macrocarpa, a partir de doze variáveis físicas e químicas da pupunha, embora algumas delas tenham tido alocação dúbia para as raças primitivas, conforme definição dos grupos feitos pelos autores.

5.1.2. Análises discriminantes paramétrica e não paramétrica

5.1.2.1. Análise discriminante linear (de Anderson, paramétrica)

Foram definidas três funções discriminantes segundo metodologia de Anderson (1958), para as situações da probabilidade *a priori* de um genótipo pertencer a uma determinada população (π_l) serem iguais e proporcionais. Com base no cenário de probabilidades *a priori* iguais, estimou-se as seguintes funções:

$$D_1(\tilde{y}) = - 596,79 - 0,20Y_1 + 8,15Y_2 + 5,63Y_3 - 0,27Y_4 + 17,96Y_5 + 36,93Y_6 + 12,72Y_7$$

$$D_2(\tilde{y}) = - 607,85 - 0,21Y_1 + 8,34\ Y_2 + 5,57Y_3 - 0,24Y_4 + 17,56Y_5 + 36,34Y_6 + 12,78Y_7$$

$$D_3(\tilde{y}) = - 584,76 - 0,07Y_1 + 8,03Y_2 + 5,52Y_3 - 0,25Y_4 + 17,93Y_5 + 36,17Y_6 + 12,56Y_7$$

Em que:

$D_1(\tilde{y})$, $D_2(\tilde{y})$ e $D_3(\tilde{y})$ referem-se às funções discriminantes para as populações microcarpa, mesocarpa e macrocarpa, respectivamente e; Y_1, Y_2, Y_3, Y_4, Y_5, Y_6 e Y_7 referem-

se às variáveis massa do fruto, porcentagem de polpa, teor de matéria seca, teor de óleo, teor de fibras, teor de cinzas e teor de proteínas, respectivamente.

Baseado no cenário de probabilidades *a priori* proporcionais, estimou-se as seguintes funções:

$$D_1(\tilde{y}) = - 596{,}35 - 0{,}20Y_1 + 8{,}16Y_2 + 5{,}63Y_3 - 0{,}27Y_4 + 17{,}96Y_5 + 36{,}93Y_6 + 12{,}72Y_7$$

$$D_2(\tilde{y}) = - 607{,}71 - 0{,}21Y_1 + 8{,}34\ Y_2 + 5{,}57Y_3 - 0{,}24Y_4 + 17{,}56Y_5 + 36{,}34Y_6 + 12{,}78Y_7$$

$$D_3(\tilde{y}) = - 586{,}00 - 0{,}07Y_1 + 8{,}03Y_2 + 5{,}52Y_3 - 0{,}25Y_4 + 17{,}93Y_5 + 36{,}17Y_6 + 12{,}56Y_7$$

Percebeu-se que há apenas uma mudança nas funções discriminantes em razão das probabilidades *a priori* definidas: na constante da equação. Com esta técnica, uma vez geradas as funções especificas das populações é possível classificar qualquer novo genótipo, de comportamento desconhecido, nas populações já conhecidas (CRUZ et al., 2014).

O maior número de matrizes classificadas incorretamente ocorreu no tipo microcarpa, com probabilidade de má classificação de 11,63%, fornecida por cinco alocações erradas, independentemente das probabilidades *a priori* – de um genótipo pertencer a uma determinada população (Tabela 2 e 3).

Na classificação das matrizes da raça mesocarpa, houve má classificação de 9,37% (três matizes) e 6,25% (duas matrizes), para os cenários de probabilidades *a priori* iguais e proporcionais, respectivamente. Na população macrocarpa, a porcentagem de má classificação foi nula e de 25% (representada por duas matrizes), para os cenários de probabilidades *a priori* iguais e proporcionais, respectivamente.

A taxa de erro aparente (TEA) foi relativamente baixa com 9,36% (Tabelas 2) e 10,84% (Tabela 3).

TABELA 2. Resumo da classificação de 83 acessos de pupunha (*Bactris gasipaes*) nas diferentes raças primitivas microcarpa (1), mesocarpa (2), macrocarpa (3), com base em sete variáveis físicas e químicas da pupunha, conforme análise discriminante de Anderson, admitindo probabilidades *a priori* iguais* dos genótipos pertencerem às populações (raças) estudadas.

População	% de classificação			Total de Ordenações	Acertos	Erros	Taxa de erro %
	1	2	3				
1	88,37	11,63	0,00	43,00	38,00	5,00	
2	3,12	90,63	6,25	32,00	29,00	3,00	
3	0,00	0,00	100,00	8,00	8,00	0,00	
Total				83,00	75,00	8,00	9,63

*Probabilidades *a priori* iguais: 0,3333.

TABELA 3. Resumo da classificação de 83 acessos de pupunha (*Bactris gasipaes*) nas diferentes raças primitivas microcarpa (1), mesocarpa (2), macrocarpa (3), com base em sete variáveis físicas e químicas da pupunha, conforme análise discriminante de Anderson, admitindo probabilidades *a priori* proporcionais* (π_l) dos genótipos pertencerem às populações (raças) estudadas.

População	% de classificação			Total de Ordenações	Acertos	Erros	Taxa de erro %
	1	2	3				
1	88,37	11,63	0,00	43,00	38,00	5,00	
2	6,25	93,75	0,00	32,00	30,00	2,00	
3	0,00	25,00	75,00	8,00	6,00	2,00	
Total				83,00	74,00	9,00	10,84

*Probabilidade *a priori* proporcionais: $\pi_1 = 0,5180$; $\pi_2 = 0,3855$; $\pi_3 = 0,0963$.

A análise discriminante de Anderson mostrou boa eficiência na classificação das matrizes, pois as classificações erradas concordaram com a expectativa biológica (genética) e geográfica de diferenciação. Uma revisão do gênero *Bactris* caracterizou populações silvestres com frutos muito pequenos como sendo da variedade botânica *chichagui*, e populações cultivadas representantes da variedade *gasipaes* (HENDERSON, 2000).

As populações de *Bactris gasipaes* var. gasipaes, estão distribuídas em três raças ao longo dos rios Solimões e Amazonas, identificadas como as raças Pará, Solimões e Putumayo. A raça Pará (microcarpa) está distribuída desde o litoral norte de Pará e o Estado do Amapá, no leste, até o baixo rio Solimões e rio Negro, no oeste, incluindo todos os tributários entre estes pontos, com ênfase no rio Madeira. Os frutos desta raça são pequenos, com alto teor de

óleo e bastante fibrosos. A raça Solimões (mesocarpa) está distribuída desde a cidade de Coari até próximo a cidade de Fonte Boa, ao longo do rio Solimões. Seus frutos são de tamanho mediano e possuem quantidades medianas de óleo e amido. A raça Putumayo (macrocarpa) está distribuída desde os arredores da cidade de Fonte Boa até Iquitos, Peru e Colômbia (MORA URPÍ et al., 1997). A variedade botânica *B. gasipaes* var. *gasipaes* apresenta maior variabilidade no tamanho, na massa e na composição dos frutos, Os frutos microcarpa variam entre 10 a 20 g, mesocarpo entre 20 e 70 g e macrocarpa acima de 70 g (CLEMENT et al., 2009). Portanto é mais provável que alocações errôneas ocorram entre genótipos do tipo microcarpa x mesocarpa e macrocarpa x mesocarpa, aspecto evidenciado nestas análises discriminantes, em que nenhuma classificação errada foi detectada entre microcarpa e macrocarpa.

Em Rondônia, na década de 1980, ocorreram introduções de sementes de pupunha de diferentes procedências, com destaque para as populações de Benjamin Constant (macrocarpa) e Yurimaguas (mesocarpa) (LOCATELLI; RAMALHO, 2005).

A efetividade das análises discriminantes linear foram avaliadas no presente estudo pela TEA. Esta taxa é de fato subestimada, pois a taxa de erro calculada foi obtida com o mesmo conjunto de dados que gerou as funções discriminantes (KHATTREE; NAIK, 2000; CRUZ et al., 2014), ou seja, o conjunto de treinamento – que gera as funções – é o mesmo conjunto de teste do desempenho da análise de discriminação. No entanto, este critério foi adotado para todas as análises discriminantes (paramétricas e não paramétricas). Outros métodos têm sido propostos pela literatura, como a validação cruzada, teste de permutação e uso de dados de treinamento e de teste diferentes (KHATTREE; NAIK, 2000; BEHARAV; NEVO, 2003; CRUZ et al., 2011) e o programa GENES, dentre estes métodos citados, executa apenas a validação cruzada para a análise discriminante linear.

Contudo, em uma pesquisa publicada, comparou-se três diferentes maneiras de avaliar o desempenho da classificação pelas análises discriminantes, teste de permutação (1); validação cruzada (2) e taxa de erro aparente (TEA) na própria amostra de análise (amostra de treinamento) (3), entre duas ou mais populações com grande número de marcadores RAPD (polimorfismo de DNA amplificado ao acaso), mas com pequeno número de indivíduos por população (BEHARAV e NEVO, 2003). Além disso, os autores selecionaram as marcas moleculares através do procedimento *stepwise*. Eles obtiveram substancial diferença entre os resultados do método 1 e o método 3, nas várias situações designadas à análise discriminante. Por outro lado, verificaram que uma alta taxa de classificações corretas também foi obtida com o teste de permutação, principalmente, quando o número de populações a serem discriminadas era pequeno. Por outro lado, a taxa de classificações corretas obtidas com a TEA a partir dos dados originais foi, em geral, mais significativa do que a porcentagem obtida pelo teste de permutação. Nenhuma ou pequenas diferenças na taxa de classificação correta foram observadas entre o procedimento de validação cruzada e o método TEA, especialmente, quando se considerou um menor número de locos selecionados pelo método *setpwise*. Os autores concluíram que a seleção de um número menor de variáveis "discriminatórias" é mais adequada à execução da análise discriminante.

5.1.2.2. Análises discriminantes não paramétricas

O método de análise discriminante pelo vizinho médio foi proposto como procedimento alternativo à avaliação de dados que não seguem distribuição específica, ou não (multi)normais, como ocorre com variáveis categóricos, a exemplo do descritores morfológicos e, com marcadores moleculares aplicados aos estudos de diversidade genética

vegetal (NIELSEN et al., 2003); MCHARO, 2005; OLIVEIRA et al., 2012). Quando criado, esse algoritmo pretendia agregar a ideia principal de vizinhança mais próxima, na verdade, em relação a um grupo mais próximo (similar) do genótipo a qual desejava-se alocar.

As distâncias Euclidiana e Euclidiana média apresentaram o menor valor de TEA, quando comparadas as medidas de dissimilaridade Quadrado da distância euclidiana e de Gower (Tabela 4). No entanto, estes valores de TEA, superaram em média 2,36 vezes a TEA da análise discriminante linear. Entretanto, para todas as quatro medidas de dissimilaridade utilizadas, em nenhuma delas houve alocação invertida das matrizes das raças primitivas do tipo microcarpa e macrocarpa. Assim como na análise discriminante linear, os erros de classificação ocorreram nas populações microcarpa *x* mesocarpa e mesocarpa *x* macrocarpa (dados não apresentados).

Independentemente das distâncias genéticas utilizadas, houve uma classificação 100% correta dos genótipos da raça macrocarpa (dados não apresentados), ou seja, quando a população está bem caracterizada por suas variáveis e é verdadeiramente divergente de outra(s), o método do vizinho médio mostrou eficiência na classificação dos genótipos.

TABELA 4. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica do vizinho médio dos conjuntos de dados das raças primitivas de pupunha (*Bactris gasipaes*) com diferentes medidas de distâncias genéticas.

Distância genética	**TEA (%)**
Euclidiana	22,89
Euclidiana média	22,89
Quadrado da distância euclidiana	25,30
Gower	25,30
Média	24,10

O agrupamento de cultivares de feijão avaliado com as distâncias Euclidiana, Euclidiana média e Quadrado da distância euclidiana, mostrou relação direta ou quadrática entre elas. As distâncias Euclidiana e Euclidiana média proporcionaram maior consistência nos padrões de agrupamento (CARGNELUTTI FILHO et al., 2010), a exemplo do que foi observado no presente estudo no quesito alocação de genótipos.

No método dos k-vizinhos mais próximos para PUPUNHA, as medidas de distância apresentaram TEA's semelhantes para um mesmo valor de k (vizinhos) nos dois diferentes cenários de probabilidades *a priori* – iguais e proporcionais – do genótipo pertencer a cada uma das raças primitivas estudadas (Tabela 5 e 6). As TEA's referentes às medidas com propriedades euclidianas apresentaram valores iguais em cada cenário de probabilidade *a priori*, para um mesmo valor de k.

A distância Euclidiana foi originalmente proposta para variáveis quantitativas (DIAS, 1998). Portanto, trata-se de uma medida sensível à correlação entre variáveis e, assim, de utilidade restrita a variáveis independentes. Entretanto, em estudos de melhoramento genético é praticamente impossível avaliar um conjunto de características não relacionadas e o uso da distância Euclidiana tem sido de grande utilidade mesmo nas situações em que a independência entre as características mensuradas não é constatada (CRUZ et al., 2011).

Exemplos de sucesso com o uso da distância Euclidiana nos estudos genéticos e de melhoramento em vegetais têm sido reportado na literatura, como em avaliação das estratégias de condução e seleção de população segregantes de soja portadores do gene RR (SILVA, 2015) e para se predizer a divergência genética entre acessos de mandioca-de-mesa por meio de características morfoagronômicas (ZUIN et al., 2009).

TABELA 5. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as diferentes raças primitivas de pupunha (*Bactris gasipaes*), admitindo probabilidades *a priori* iguais* dos indivíduos pertencerem às respectivas populações (raças) com diferentes medidas de distâncias genéticas para variados valores de k.

Distância genética	**Valor de k**			**Média**
	1	**4**	**7**	
Euclidiana	10,66	11,41	12,96	11,68
Euclidiana média	10,66	11,41	12,96	11,68
Quadrado da distância euclidiana	10,66	11,41	12,96	11,68
Gower	12,48	10,37	11,92	11,59
Média	11,12	11,15	12,70	11,66

*Probabilidades *a priori* iguais: 0,3333.

TABELA 6. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as diferentes raças primitivas de pupunha (*Bactris gasipaes*), admitindo probabilidades *a priori* proporcionais* dos indivíduos pertencerem às respectivas populações (raças) com diferentes medidas de distâncias genéticas para variados valores de k.

Distância genética	**Valor de k**			**Média**
	1	**4**	**7**	
Euclidiana	9,63	16,86	15,66	14,05
Euclidiana média	9,63	16,86	15,66	14,05
Quadrado da distância euclidiana	9,63	16,86	15,66	14,05
Gower	12,04	14,45	14,45	13,65
Média	10,23	16,26	15,36	13,95

*Probabilidade *a priori* proporcionais: Microcarpa 0,5180; Mesocarpa 0,3855; Mesocarpa 0,0963.

A medida de Gower tendeu a apresentar menores taxas de erros quando o valor de k cresceu (Tabela 5 e 6), na comparação com as outras medidas de dissimilaridade utilizadas. Esta medida é uma opção para análise simultânea de variáveis categóricas e quantitativas e trata-se de um procedimento de pouca complexidade e que tem produzido resultados confiáveis, embora ainda pouco explorados pelos pesquisadores que atuam na área de recursos genéticos vegetais para detecção da variabilidade em bancos de germoplasma (QUINTAL et al., 2012). O emprego da distância de Gower para estudos de predição da diversidade genética encontram-se na literatura com uso de descritores morfológicos e agronômicos, a exemplo de *Brassica napus* L. (RODRÍGUEZ et al., 2005), pimentas do

gênero *Capsicum* spp. (MOURA et al., 2010) e tomateiro do grupo cereja (ROCHA et al., 2010).

Os valores de k variaram de 1 até 7 porque a menor população (macrocarpa) possui apenas oito genótipos. Então, para k = 8, certamente, ter-se-ia pelo menos um genótipo não pertencente a população original macrocarpa, uma vez que não se classifica a vizinhança do genótipo com ele mesmo. Em média, as TEA's variaram e tenderam a aumentar conforme incrementou-se o valor k. Quando k = 1 obteve-se as menores TEA's (Tabela 5 e 6).

Tabela 7. Medidas de disposição das Taxas de erro aparente (TEA) pelo método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as diferentes raças primitivas de pupunha (*Bactris gasipaes*), admitindo probabilidades *a priori* iguais e proporcionais* dos indivíduos pertencerem às respectivas populações (variedades) para as medidas de distâncias genéticas de Gower e Euclidiana em diferentes valores de k.

	Valor de k						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Média geral	11,20	11,02	10,87	13,27	14,45	15,37	13,75
Variância	1,27	2,47	9,36	6,54	4,86	11,55	2,03
Desvio - padrão	1,13	1,57	3,06	2,56	2,20	3,40	1,42
C.V. (%)	10,08	14,26	28,16	19,27	15,25	22,11	10,36
Mínimo	9,63	8,82	7,27	10,37	10,64	11,15	11,92
Máximo	12,48	13,25	15,66	16,86	15,85	19,28	15,66
Amplitude total	2,85	4,43	8,39	6,49	5,21	8,13	3,74

*Probabilidades *a priori* iguais: 0,3333. Probabilidade *a priori* proporcionais: Microcarpa 0,5180; Mesocarpa 0,3855; Mesocarpa 0,0963.

Em relação às TEA's, admitindo probabilidades *a priori* iguais e proporcionais pelas distâncias Euclidiana e de Gower, a menor variância encontrada foi de 1,27 e a maior de 11,55 para os valores de k =1 e k = 6 respectivamente. A média geral variou de 10,87 a 15,37, indicando uma pequena diferença das TEA's entre os valores de k. E a amplitude das TEA's variou de 7,27 a 19,28 para os valores de k = 3 e k = 6 respectivamente (Tabela 7).

O desvio-padrão, em relação à variância, variou entre 1,13 a 3,40 para os valores de k = 1 e k = 6 respectivamente, corroborando os valores da variância. O coeficiente de variação

em relação ao desvio-padrão e a média geral observado foi de 10,08% a 28,16%, comprovando o grau de diferença entre as TEA's para diferentes valores de k (Tabela 7).

Negreiros (2013) estudando a Divergência genética entre progênies de pupunheira quanto a caracteres de palmito, obteve um coeficiente de variação das variáveis agronômicas – massa do palmito de primeira e de segunda e plantas por parcela – foi maior do que 30% e, quanto à massa da base do palmito, foi maior do que 27%, tendo sido, portanto, satisfatórios, uma vez que foram detectadas diferenças significativas para a maioria das variáveis avaliadas entre as progênies.

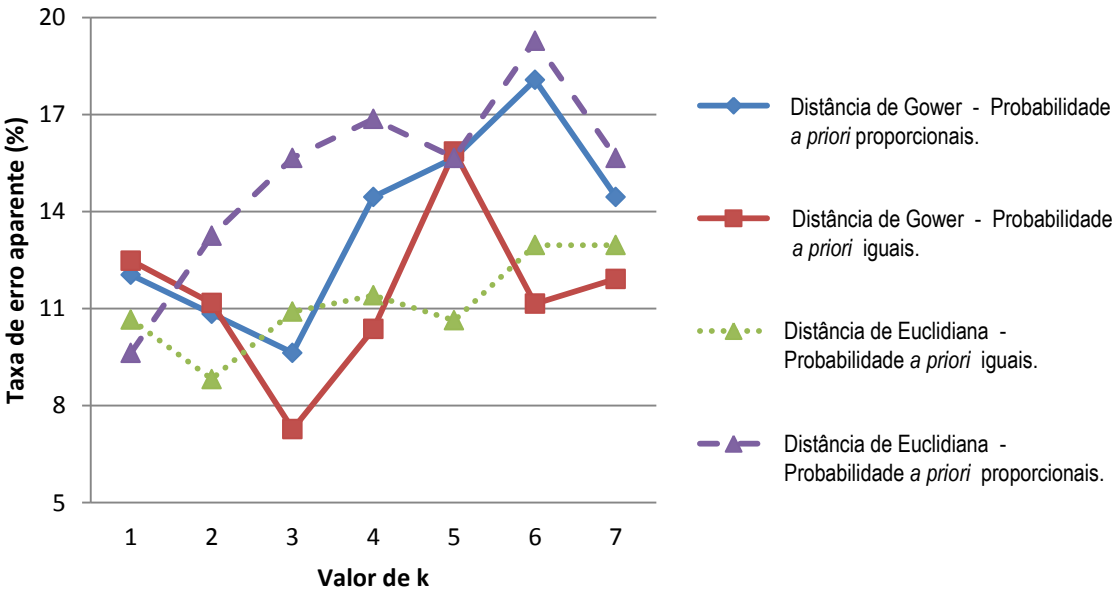

Gráfico 1. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as diferentes raças primitivas de pupunha (*Bactris gasipaes*), admitindo probabilidades *a priori* iguais e proporcionais* dos indivíduos pertencerem às respectivas populações (variedades) para as medidas de distâncias genéticas de Gower e Euclidiana em diferentes valores de k.

*Probabilidades *a priori* iguais: 0,3333. Probabilidade *a priori* proporcionais: Microcarpa 0,5180; Mesocarpa 0,3855; Mesocarpa 0,0963.

As distâncias Euclidiana e de Gower possuem uma tendência a valores de k mais elevados (Gráfico 1), quando comparados pela probabilidade *a priori iguais*. Nota-se que a

distancia genética Euclidiana, admitindo a probabilidade *a priori* proporcionais possui os valores mais elevados de k.

Pela literatura recente, não há uma definição clara quanto à escolha do melhor valor para k. Contudo, em alguns estudos, especialmente com grandes amostras, a escolha do valor de k é irrelevante (KHATTREE; NAIK, 2000), inclusive há trabalhos cujo valor adotado foi de k = 1 (BEHARAV; NEVO, 2003; BRESSAN; VITRI, 2003).

Assim como o método do vizinho médio, para todas as quatro medidas de dissimilaridade utilizadas, em nenhuma delas houve alocação invertida das matrizes das raças primitivas do tipo microcarpa e macrocarpa. Os erros de classificação ocorreram nas populações microcarpa x mesocarpa e mesocarpa x macrocarpa (dados não apresentados). Há um aspecto interessante, que para valores de k > 1 é possível que um genótipo fique com classificação indefinida, uma vez que a decisão sobre a alocação recai sobre estimativas de probabilidade, e não apenas na quantidade de vizinhos mais próximos, podendo, então, existir genótipos classificados em mais de uma população.

5.2. Variedades botânicas de café "*canephora*"

5.2.1. Caracterização das populações

As variedades botânicas do cafeeiro (Conilon, Robusta e Híbrido intervarientais) apresentaram dissimilaridade multivariada significativa ($p < 0,05$), o que evidencia a diferença entre as populações dos conjuntos de dados CAFÉ (Tabela 8). Embora exista diferença estatística entre as populações Conilon e híbridos, a estimativa da distância de

Mahalanobis entre estas variedades revelou ser a menor (D^2 = 4,14) dentre as demais comparações (Tabela 8). Para as populações Conilon e Robusta, estimou-se a maior distância.

Diferenças significativas ($p < 0,05$) foram detectadas entre os clones de cafeeiro "canephora" para cada uma das dez variáveis avaliadas. Mais ainda, por meio da técnica de componentes principais, em dois anos agrícolas, verificou-se que os clones do tipo Híbridos intervarientais apresentaram maior similaridade com a variedade botânica Conilon. Alguns genótipos Robusta se agruparam distantes do ponto centroide de seu grupo, indicando ou um evento raro de segregação ou mais provavelmente, uma mistura entre os clones (OLIVEIRA, 2017).

TABELA 8. Dissimilaridades entre os pares de café (*Coffea canephora*) nas diferentes variedades botânicas Conilon, Robusta e Híbridos intervarientais, expressas pela distância generalizada de Mahalanobis (D^2) e testado pelo teste F a 1% de probabilidade.

Par de Populações	**D^2**	**G.l***	**$F_{calculado}$**	**P > F (%)**
Conilon x Robusta	4,26	(10, 89)	7,80	0,00
Conilon x Híbridos intervarientais	1,44	(10, 83)	2,19	2,58
Robusta x Híbridos intervarientais	3,53	(10, 39)	3,54	0,21

*Grau de liberdade.

5.2.2. Análises discriminantes paramétrica e não paramétrica

5.2.2.1. Análise discriminante linear (de Anderson, paramétrica)

Foram definidas três funções discriminantes, segundo metodologia de Anderson (1958), para as situações cujas probabilidades *a priori* de um genótipo pertencer a uma determinada população (π_l) seriam iguais e proporcionais. Com base no cenário de probabilidades *a priori* iguais, estimou-se as seguintes funções:

$$D_1(\tilde{y}) = -233,78 + 20,32Y_1 + 0,24Y_2 + 2,85Y_3 + 1,66Y_4 + 9,16Y_5 + 0,92Y_6 + 1,11Y_7 - 0,05Y_8 + 2,03Y_9 - 5,59Y_{10}$$

$D_2(\tilde{y}) = -\ 248{,}15 + 20{,}58Y_1 + 0{,}23Y_2 + 2{,}91Y_3 - 1{,}69Y_4 + 9{,}85Y_5 + 0{,}90Y_6 + 1{,}13Y_7 - 0{,}10Y_8 + 2{,}51Y_9 - 5{,}02Y_{10}$

$D_3(\tilde{y}) = -\ 226{,}88 + 21{,}64Y_1 + 0{,}24Y_2 + 2{,}76Y_3 + 1{,}07Y_4 + 9{,}68Y_5 + 0{,}98Y_6 + 1{,}07Y_7 - 0{,}06Y_8 + 1{,}84Y_9 - 4{,}68Y_{10}$

Em que:

$D_1(\tilde{y})$, $D_2(\tilde{y})$ e $D_3(\tilde{y})$ referem-se às funções discriminantes para as populações Conilon, Robusta e Híbridos intervarientais, respectivamente e; Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8, Y_9 e Y_{10} referem-se as variáveis altura da planta a partir do nível do solo, número de ramos plagiotrópicos produtivos; número de rosetas por ramo plagiotrópico; comprimento do ramo plagiotrópico; distância entre rosetas da parte intermediária do ramo plagiotrópico; número de grãos por roseta da parte intermediária do ramo plagiotrópico; comprimento e a largura das folhas; época de maturação com o registro da data de colheita e; o valor genotípico da produção de grãos beneficiados, respectivamente.

Baseado no cenário de probabilidades *a priori* proporcionais, estimou-se as seguintes funções:

$D_1(\tilde{y}) = -\ 233{,}21 + 20{,}32Y_1 + 0{,}24Y_2 + 2{,}85Y_3 + 1{,}66Y_4 + 9{,}16Y_5 + 0{,}92Y_6 + 1{,}11Y_7 - 0{,}05Y_8 + 2{,}03Y_9 - 5{,}59Y_{10}$

$D_2(\tilde{y}) = -\ 248{,}53 + 20{,}58Y_1 + 0{,}23Y_2 + 2{,}91Y_3 - 1{,}69Y_4 + 9{,}85Y_5 + 0{,}90Y_6 + 1{,}13Y_7 - 0{,}10Y_8 + 2{,}51Y_9 - 5{,}02Y_{10}$

$D_3(\tilde{y}) = -\ 227{,}50 + 21{,}64Y_1 + 0{,}24Y_2 + 2{,}76Y_3 + 1{,}07Y_4 + 9{,}68Y_5 + 0{,}98Y_6 + 1{,}07Y_7 - 0{,}06Y_8 + 1{,}84Y_9 - 4{,}68Y_{10}$

As maiores taxas de erro aparente (TEA, %) na classificação dos clones ocorreram entre as variedades Conilon e Híbridos intervarientais (Tabelas 9 e 10) e houveram más

classificações em relação a variedade botânica Robusta, corroborando com os resultados da Tabela 8 e da análise preditiva da diversidade genética (OLIVEIRA, 2017) entre estes clones de cafeeiro. A TEA média, considerando os cenários de probabilidades *a priori* iguais e proporcionais, foi de aproximadamente 20,00%.

A espécie *Coffea canephora* se caracteriza por apresentar plantas de duas variedades botânicas distintas, denominadas Conilon e Robusta (SUNARHARUM et al., 2014). A variedade botânica Robusta se caracteriza por apresentar maior vigor, crescimento ereto, folhas e frutos de maior tamanho, maturação tardia, menor tolerância ao déficit hídrico e maior tolerância a pragas e doenças. Por sua vez, a variedade botânica Conilon se caracteriza por apresentar plantas de crescimento arbustivo, maturação precoce, caules ramificados, folhas alongadas, resistência a seca e maior suscetibilidade a doenças (FONSECA et. al, 2015; MONTAGNON et al., 2012). Ambas as variedades botânicas apresentam plantas de ciclo de maturação precoce (240 dias), média ou intermediárias (270 dias), tardias (300 dias) e extremamente tardias (330 dias) (BRAGANÇA et al., 2001; FERRÃO et al., 2008). Híbridos naturais (intervarientais) que apresentam a arquitetura de copa, precocidade e resistência à seca do 'Conilon', com o vigor, tamanho de frutos e resistência a pragas e doenças do 'Robusta', têm naturalmente se destacado nas avaliações de campo (ROCHA et al., 2015).

TABELA 9. Resumo da classificação dos 122 acessos de café (*Coffea canephora*) nas diferentes populações Conilon (1), Robusta (2) e Híbridos intervarientais (3), com base em dez variáveis agronômicas do cafeeiro, conforme análise discriminante de Anderson, admitindo probabilidade *a priori* igual* de os indivíduos pertencerem às respectivas populações (variedades).

População	**% de classificação**			**Total de Ordenações**	**Acertos**	**Erros**	**Taxa de erro %**
	1	**2**	**3**				
1	77,78	1,39	20,83	72,00	56,00	16,00	
2	0,00	92,86	7,14	28,00	26,00	2,00	
3	27,27	13,64	59,09	22,00	13,00	9,00	
Total				122,00	95,00	27,00	22,13

*Probabilidades *a priori* iguais: 0,3333.

TABELA 10. Resumo da classificação dos 122 acessos de café (*Coffea canephora*) nas diferentes variedades botânicas Conilon (1), Robusta (2) e Híbridos intervarientais (3), com base em dez variáveis agronômicas do cafeeiro, conforme análise discriminante de Anderson, admitindo probabilidade *a priori* proporcionais* de os indivíduos pertencerem às respectivas populações (variedades).

População	**% de classificação**			**Total de Ordenações**	**Acertos**	**Erros**	**Taxa de erro %**
	1	**2**	**3**				
1	95,83	0,00	4,17	72,00	69,00	3,00	
2	14,29	85,71	0,00	28,00	24,00	4,00	
3	63,64	9,09	27,27	22,00	6,00	16,00	
Total				122,00	99,00	23,00	18,85

*Probabilidade *a priori* proporcionais: Conilon 0,5901; Robusta 0,2295; Híbridos intervarientais 0,1803.

5.2.2.2. Análises discriminantes não paramétricas

As TEA's pelo método do vizinho médio (Tabela 11) equipararam-se as probabilidades de má classificação das análises discriminantes linear para CAFÉ, em contrapartida as TEA's obtidas para o conjunto PUPUNHA, que superaram em média 2,36 vezes as taxas de erro da análise discriminante linear.

A utilização da distância Euclidiana apresenta o inconveniente de ser alterada, quando medida a partir de variáveis originais, com a mudança da escala de medições, com o número de características e pela correlação entre elas. Para solucionar os dois primeiros problemas a utilização da distância Euclidiana média e a padronização dos dados são indicadas (CRUZ et al., 1994), embora não resolva o problema de correlações entre as características analisadas. A distância Euclidiana média para variáveis quantitativas serve quando o germoplasma encontra-se instalado sem casualização e ou sem controle local (CRUZ et al., 2004). O uso da distância Euclidiana média em clones de cafeeiro *C. canephora* para estudos de divergência genética encontra-se na literatura (FONSECA, 1999).

TABELA 11. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica do vizinho médio dos conjuntos de dados das variedades botânicas de café (*Coffea canephora*) com diferentes medidas de distâncias genéticas.

Distância genética	TEA (%)
Euclidiana	21,31
Euclidiana média	21,31
Quadrado da distância euclidiana	18,85
Gower	20,49
Média	20,49

O método dos k-vizinhos mais próximos, para probabilidades *a priori* iguais, apresentou TEA's superiores cerca de 1,8 a 2,0 vezes ao método de análise discriminante linear e o método do vizinho médio para o conjunto CAFÉ (Tabela 12), assim como k = 1 para probabilidades *a priori* proporcionais (Tabela 13), pois nesta condição as probabilidades pré definidas em relação ao genótipo pertencer a uma das populações não são computadas para discriminar e, portanto, não influenciam a classificação.

Para valores de k = 5, 9, 13, 17 e 21 (Tabela 13), as porcentagens de classificação errônea (TEA's) se assemelharam aquelas geradas pelas análises discriminantes linear e vizinho médio.

Para ambos os métodos de análise discriminante não paramétrica, ocorreram classificações incorretas entre Conilon *x* Robusta, Conilon *x* Híbridos e Robusta *x* Híbridos, coadunando com a informação de grande variabilidade entre e dentro das populações naturais de *C. canephora* (SOUZA et al., 2013).

Novamente, as TEA's foram iguais para aquelas distâncias que resguardam propriedades euclidianas (Tabela 12 e 13) para cada k simulado. A medida de Gower, admitindo probabilidades *a priori* proporcionais foi a que apresentou as menores taxas de erro, na avaliação deste conjunto de dados.

TABELA 12. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as diferentes variedades botânicas de café (*Coffea canephora*), admitindo probabilidades *a priori* iguais* dos indivíduos pertencerem às respectivas populações (variedades) com diferentes medidas de distâncias genéticas para variados valores de k.

Distância genética	Valor de k						Média
	1	5	9	13	17	21	
Euclidiana	50,39	36,54	33,38	34,57	35,91	33,47	37,38
Euclidiana média	50,39	36,54	33,38	34,57	35,91	33,47	37,38
Quadrado da distância média	50,39	36,54	33,38	34,57	35,91	33,47	37,38
Gower	51,25	36,68	33,76	34,72	30,82	34,31	36,92
Média	50,61	36,58	33,48	34,61	34,64	33,68	37,27

*Probabilidades *a priori* iguais: 0,3333.

TABELA 13. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as variedades botânicas de café (*Coffea canephora*), admitindo probabilidades *a priori* proporcionais* dos indivíduos pertencerem às respectivas populações (variedades) com diferentes medidas de distâncias genéticas para variados valores de k.

Distância genética	Valor de k						Média
	1	5	9	13	17	21	
Euclidiana	38,52	24,59	22,13	20,49	21,31	22,95	25,00
Euclidiana média	38,52	24,59	22,13	20,49	21,31	22,95	25,00
Quadrado da distância média	38,52	24,59	22,13	20,49	21,31	22,95	25,00
Gower	36,88	23,77	21,31	20,49	20,49	20,49	23,91
Média	38,11	24,39	21,93	20,49	21,11	22,34	24,73

*Probabilidade *a priori* proporcionais: Conilon 0,5901; Robusta 0,2295; Híbridos intervarientais 0,1803.

No estudo da diferenciação sexual de *Araucaria angustifolia*, utilizando o método dos k-vizinhos mais próximos , foi possível predizer o sexo através da comparação de seu perfil metabólico com o de plantas com sexo conhecido. Também confirmou-se a eficiência do método com a utilização de sensores óticos na detecção de agentes patogênicos em plantas, mostrando ser possível separar com sucesso as folhas controladas e infetadas, usando os espectros medidos com uma taxa de sucesso superior a 90% (CARVALHO, 2012).

Em outro estudo que se analisou categorias de reconhecimento visual, obteve-se uma taxa de classificação correta de 59,05% (± 0,56%) em 15 imagens (dados) de treinamento por

aula e 66,23% (± 0,48%) em 30 imagens (dados) de treino, ao que os autores consideraram ser eficaz o método dos k-vizinhos mais próximos (ZHANG et al., 1995).

Com relação às TEA's, admitindo probabilidades a priori iguais e proporcionais pelas distâncias Euclidiana e de Gower, a menor variância encontrada foi de 24,48 e a maior de 175,43 para os valores de k =11 e k = 2 respectivamente. A média geral variou de 26,03 a 44,26, indicando uma relativa diferença das TEA's entre os valores de k. E a amplitude das TEA's variou de 19,67 a 54,77 para os valores de k = 8 e k = 2 respectivamente (Tabela 14).

O desvio-padrão, em relação a variância, variou de 4,95 a 13,25 para os valores de k = 11 e k = 2 respectivamente, corroborando os valores de k para a variância. O coeficiente de variação em relação ao desvio-padrão e a média geral observado foi de 14,90% a 34,90 % para k = 1 e k = 2 respectivamente (Tabela 14).

Para Oliveira (2017) o coeficiente de variação experimental (CV %) está associado à magnitude do componente de variância genotípica e a precisão em que cada uma das características foi avaliada. Em seu estudo, a característica produção de café da roça com 29,42% apresentou a maior CV%. De acordo com Ferrão (2008) o CV (%) está associado a uma boa condução experimental.

Tabela 14. Medidas de disposição das Taxas de erro aparente (TEA) pelo método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as variedades botânicas de café (*Coffea canephora*), admitindo probabilidades *a priori* iguais e proporcionais* dos indivíduos pertencerem às respectivas populações (variedades) para as medidas de distâncias genéticas de Gower e Euclidiana em diferentes valores de k.

	Valor de k										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Média Geral	44,26	37,95	36,74	31,14	30,39	26,43	27,73	28,17	27,65	26,03	26,88
Variância	43,49	175,43	66,19	97,15	38,70	30,70	45,12	60,32	35,22	27,75	24,48
Desvio - padrão	6,60	13,25	8,14	9,86	6,22	5,54	6,72	7,77	5,93	5,27	4,95
C.V. (%)	14,90	34,90	22,14	31,65	20,47	20,96	24,22	27,57	21,47	20,23	18,40
Mínimo	36,88	24,59	27,05	20,49	23,77	20,49	20,49	19,67	21,31	20,49	21,31
Máximo	51,25	54,77	45,09	41,53	36,68	32,19	36,82	37,30	33,76	32,79	33,40

Amplitude total	14,37	30,18	18,04	21,04	12,91	11,70	16,33	17,62	12,45	12,30	12,09

Valor de k	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Média Geral	30,02	27,57	28,04	26,99	26,44	27,13	27,14	27,81	28,33	27,80
Variância	94,99	50,08	57,08	42,39	32,92	42,16	42,90	46,45	44,66	37,87
Desvio - padrão	9,75	7,08	7,56	6,51	5,74	6,49	6,55	6,82	6,68	6,15
C.V. (%)	32,47	25,67	26,94	24,12	21,70	23,93	24,13	24,51	23,59	22,13
Mínimo	20,49	20,49	20,49	20,49	20,49	20,49	20,49	20,49	20,49	20,49
Máximo	42,45	34,72	35,95	34,06	34,06	35,91	36,16	37,09	35,57	34,31
Amplitude total	21,96	14,23	15,46	13,57	13,57	15,42	15,67	16,60	15,08	13,82

* Probabilidades *a priori* iguais: 0,3333. Probabilidade *a priori* proporcionais: Conilon 0,5901; Robusta 0,2295; Híbridos intervarientais 0,1803.

Em quase todas as circunstâncias (Euclidiana – probabilidades iguais e proporcionais, Gower – probabilidades iguais e proporcionais) quando os valores de k = 1 (Gráfico 2), obtém-se as maiores Taxas de erro aparente, em comparação a outros valores de k. As TEA's para a probabilidade *a priori* iguais, no geral, superaram aquelas obtidas para probabilidades a priori proporcionais, independente da medida de distância utilizada (no caso, Euclidiana e Gower), quando considerado um mesmo valor de k.

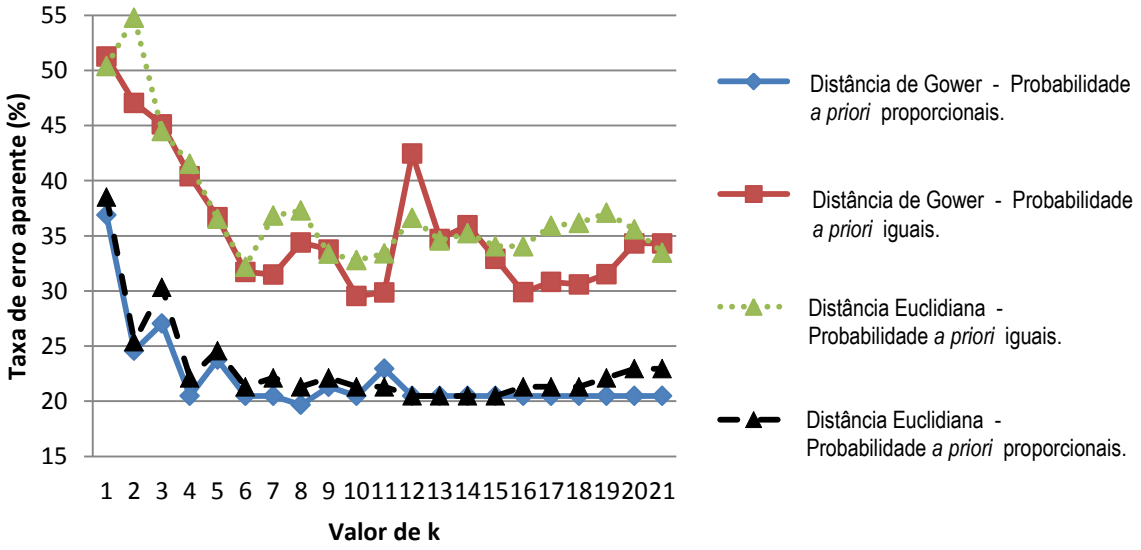

Gráfico 2. Taxas de erro aparente (TEA) para o método de análise discriminante não paramétrica dos k-vizinhos mais próximos para as variedades botânicas de café (*Coffea canephora*), admitindo probabilidades *a priori* iguais e proporcionais* dos indivíduos pertencerem às respectivas populações (variedades) para as medidas de distâncias genéticas de Gower e Euclidiana em diferentes valores de k.

* Probabilidades *a priori* iguais: 0,3333. Probabilidade *a priori* proporcionais: Conilon 0,5901; Robusta 0,2295; Híbridos intervarientais 0,1803.

Ao comparar o método dos k-vizinhos mais próximo, com a análise discriminante quadrática e análise discriminante linear para reconhecimento de padrões de sinais eletromiográficos a partir de diferentes direções de movimento do pulso, constatou uma taxa média de reconhecimento de 84,9%, sendo superior aos demais métodos empregados. O método dos k-vizinhos mais próximos foi estatisticamente diferente da análise discriminante ($p < 0,05$) e, consequentemente, o mais confiável para classificar as direções do movimento do pulso (KIM et al., 2011). Os autores ainda consideraram o valor de $k = 5$ (utilizaram de 1 a 10) como aquele que obteve a melhor taxa de reconhecimento e com menor desvio padrão.

As populações dos conjuntos de dados PUPUNHA e CAFÉ apresentaram tamanhos amostrais satisfatórios para o estudo, pois tratam-se de genótipos que compõem bancos de germoplasma da Embrapa em Porto Velho - RO. Os dois conjuntos apresentam característica diferente quanto a divergência de suas populações: PUPUNHA tem suas populações (raças primitivas) mais dissimilares, portanto mais fácil de discriminá-las, ao contrário das populações (variedades botânicas) CAFÉ (ver projeção gráfica em SANTOS et al. 2017 e OLIVEIRA, 2017). Isto permitiu, de certa forma, responder as indagações acerca dos métodos não paramétricos de análise discriminante: método do vizinho médio e dos k-vizinhos mais próximos.

No geral, o método dos k-vizinhos mais próximos apresentou maior número de classificações corretas quando os genótipos das populações mostraram-se mais agrupados e menos dispersos (dados PUPUNHA). Em contrapartida o método do vizinho médio mostrou maior eficiência na classificação de populações cujos genótipos são mais divergentes dentro e entre grupos, podendo causar mais misturas na classificação. Uma explicação para isso é a de que, em decorrência da filosofia analítica do método, a média das distâncias dos genótipos para cada grupo deva captar melhor as diferenças multivariadas entre as populações.

A distância Euclidiana e suas extensões apresentaram valores de taxa de erro idênticos. A distância de Gower estimou taxas de erro semelhantes às distâncias Euclidianas e mostra ser eficiente para classificar genótipos aliado aos métodos de análise discriminante.

As alternâncias nos valores de k apresentaram taxas de erro semelhantes quando utilizadas diferentes medidas de dissimilaridade. Neste estudo, os valores de k = 1 tenderam a apresentar menores taxas de erro, quando os genótipos são melhor agrupados em suas respectivas populações (PUPUNHA). Quando os indivíduos de uma população estão mais dispersos ou longe de seus centroides, o valor de k = 1 deve ser evitado. Realmente, o valor de k nãos é o mais limitante para compor boas classificações. A eficácia das funções discriminantes está associada principalmente à quantidade e à qualidade das variáveis consideradas na discriminação (CRUZ et al., 2004).

Estudos futuros devem ser realizados para se investigar o quanto que as variáveis são capazes aumentar a capacidade discriminatória das análises discriminantes não paramétricas, quando adicionadas ou retiradas da análise.

6. CONCLUSÕES

Os métodos não paramétricos foram efetivos para classificar os genótipos em suas respectivas populações quando comparados com o método de análise discriminante de Anderson.

Não há diferenças expressivas de classificação entre as medidas de distâncias Euclidiana. A distância de Gower proporciona taxas de erro aparente distinta das demais distâncias estudadas, mas não tão diferentes.

O método de análise discriminante dos k-vizinhos mais próximos mostrou ser adequado para populações cuja divergência genética dentro é menor. Já o método do vizinho médio classifica melhor os genótipos em populações em que haja maior diversidade inter ou intrapopulacional.

7. REFERÊNCIAS

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. New York: John Wiley & Sons, Inc, 242 p. 1958.

ANDERSON, M. J.; ROBINSON, J. **Generalized discriminant analysis based on distances**. University of Auckland and University of Sydney. Aust. N. Z. J. Stat. 45(3), 301-318, 2003.

ASSIS, G. M. L. da.; EUCLYDES, R. F.; CRUZ, C. D.; VALLE, C. B. do. Discriminação de Espécies de *Brachiaria* Baseada em Diferentes Grupos de Caracteres Morfológicos. **R. Bras. Zootec.**, v.32, n.3, p. 576-584, 2003.

BANTTE, K.; PRASANNA, B. M. Simple sequence repeat polymorphism in Quality Protein Maize (QPM) lines. **Euphytica**, v. 129, p. 337-344, 2003.

BEHARAV, A.; NEVO, E. Predictive validity of discriminant analysis for genetic data. **Genetica**, v. 119, p. 259-267, 2003.

BRAGANÇA, S. M.; CARVALHO, C. H. S. de; FONSECA, A. F. A da; FERRÃO, R. G. Variedades clonais de café Conilon para o Estado do Espírito Santo. **Pesquisa Agropecuária Brasileira**, Brasília, vol. 36, n. 5, p. 765-770, 2001.

BRESSAN, M.; VITRIÀ, J. Nonparametric discriminant analysis and nearest neighbor classification. **Pattern Recognition Letters**, v. 24, p. 2743–2749, 2003.

CARGNELUTTI FILHO, A.; RIBEIRO, N. D.; BURIN, C. Consistência do padrão de agrupamento de cultivares de feijão conforme medidas de dissimilaridade e métodos de agrupamento. **Pesquisa agropecuária brasileira**, Brasília, v. 45, n. 3, p. 236-243, 2010.

CARVALHO, B. G. **Diferenciação sexual de *Araucaria angustifolia* por RMN HR-MAS e análise multivariada**. Dissertação (Mestrado). Universidade Federal de Goiás, Instituto de Química (IQ), Programa de Pós-Graduação em Química, Goiânia, 59 f. 2012.

CLEMENT, C. R. Domestication of the pejibaye palm (*Bactris gasipaes*): past and present. **Advances Economic Botany**, v.6, p.155-174, 1988.

CLEMENT C. R.; KALIL FILHO, N. A.; MODOLO, V. A.; YUYAMA, K.; RODRIGUES, D. P.; VAN LEEUWEN, J.; FARIAS NETO, J. T.; CRISTO-ARAUJO, M.; FLORES, W. B. C. Domesticação e melhoramento de pupunha. *In*: BORÉM, A.; LOPES, M. T. G; CLEMENT, C. R. **Domesticação e melhoramento: Espécies amazônicas**. (Ed.) Viçosa: UFV, p. 363-394. 2009.

COELHO, C. F. **Misturas finitas de normais assimétricas e de t assimétricas aplicadas em análise discriminante**. Dissertação de Mestrado. Manaus: UFAM/ICE, 2013.

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

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. v. 2. Editora UFV, Viçosa, 585 p. 2003.

CRUZ, C. D.; REGAZZI, J. A.; CARNEIRO, P. C. S. Divergência genética. In: CRUZ, C.D.; REGAZZI, J.A.; CARNEIRO, P.C.S. (Ed.). **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, v. 1, p. 377-413. 2004.

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

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

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. 5. ed. Viçosa, Ed. UFV, 480 p. 2012.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético** 2. 4. ed. vol. 2. Viçosa, MG. Ed. UFV, 668 p. 2014.

CRUZ, C. D. Genes Software – extended and integrated with the R, Matlab and Selegen. **Revista Acta Scientiarum Agronomy**, Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec. 2016.

CURI, P. R. Análise de agrupamento: métodos sequenciais, aglomerativos e hierárquicos. **Ciência e Cultura**, São Paulo, v. 35, n. 10, p. 1416-1429, 1983.

DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. Jombart et al. **BMC Genetics** 11:94. 2010.

DIAS, L. A. S. Análises multidimensionais. In: ALFENAS, A. C. (Ed.) **Eletroforese de isoenzimas e proteínas afins**. Viçosa-MG: Ed. UFV, 1998. P. 405-475.

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

EBDON, J. S.; PETROVIC, A. M.; SCHWAGER, S. J. Evalution of discriminant analysis of low- and high-water use Kentucky bluegrass cultivars. **Crop Science**, v. 38, p. 152-157, 1998.

FAYYAD, U. M. **Advances in Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press/MIT Press, 611p. 1996.

FERRÃO, R. G.; CRUZ, C. D.; FERREIRA, A.; CECON, P. R.; FERRÃO, M. A. G.; FONSECA, A. F. A.; CARNEIRO, P. C. D.; SILVA, M. F. Genetic parameters in Conilon coffee. **Pesquisa Agropecuária Brasileira**, v. 43, n. 1, p. 61-69, 2008.

FERREIRA, R. P; CRUZ, C. D.; SEDIYAMA, C. S.; FAGERIA, N. K. Identificação de cultivares de arroz tolerantes à toxidez de alumínio por técnica multivariada. **Pesquisa Agropecuária Brasileira**, v. 30, n. 6, p. 789-795, jun. 1995.

FERREIRA, D. F. **Estatística básica**. Lavras: UFLA, 664 p. 2005.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 10, p. 422-429, 1936.

FIX, E.; HODGES, J. L. *In*: SILVERMAN, B. W.; JONESSOURCE, M. C. E. (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). **International Statistical Review**, v. 57, n. 3, p. 233-238, 1989.

FONSECA, A. F. A. **Análise biométrica em café Conilon (*Coffea canephora* Pierre)**. Tese (Doutorado) - Universidade Federal de Viçosa, Viçosa, MG. 1999.

FONSECA, A. F.A.; SEDIYAMA, T.; CRUZ, C. D.; SAKIYAMA, N. S.; FERRÃO, R. G.; FERRÃO, M. A. G.; BRAGANÇA, S. M. Discriminant analysis for the classification and clustering of Robusta coffe genotypes. **Crop Improvement and Applied Biotechnology**, v. 4, n. 3, p. 285-289, 2004.

FONSECA, A. F. A. da (Org.); SAKIYAMA, N. S. (Org.); BORÉM, A (Org.). **Café Conilon: do plantio à colheita**. Ed. Viçosa, MG: UFV, 2015.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, Arlington, v. 27, n. 3, p. 857-871, 1971.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Multivariate Data Analysis: With Readings**. 4 ed. Upper Sadle River: Prentice Hall, 1995.

HAIR, J. F.; ANDERSON, R. E.; TATHAM R. L.; BLACK, W. C. Trad: SANT'ANNA, A. S.; NETO, A. C. **Análise multivariada de dados**. 5. ed. Porto Alegre, Bookman, 593 p. 2005.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. **The Elements of Statistical Learning**. Data Mining, Inference, and Prediction. Second Edition. Springer, USA. (2009).

HUBERTY, C. J. **Applied Discriminant Analysis. Wiley Series in Probability and Mathematical Statistics**. Ed. John Wiley & Sons, Inc., 1994.

IVOGLO, M. G. **Divergência genética entre progênies de café Robusta**. Dissertação de Mestrado. Instituto Agronômico. Campinas, 75fls, 2007.

JAIN, A. K., DUIN, R. P. W. & MAO, J. Statistical Pattern Recognition: A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 22(1), 4–37. 2000.

KHATTREE, R.; NAIK, D. N. **Applied multivariate statistical with SAS Software**. 2. ed., a co publication of Cary, NC: SAS Institute Inc. and New York, John Wiley & Sons, 338p. 2000.

KIM, K. S.; CHOI, H. H.; MOON, C. S.; MUN, C. W. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. **Current Applied Physics**. v. 11, p. 740-745, 2011.

LANCHENBRUCH, P. A. **Discriminant analysis**. New York: Hatner Press, p. 128, 1979.

LILLIEFORS, H. W. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. **Journal of the American Statistical Association**, vol. 62, n. 318, p. 399-402. 1967.

MALHOTRA, N. K. **Pesquisa de marketing: uma orientação aplicada**. Porto Alegre: Bookman, 2001.

MARDIA, K. V.; KENT, J. T.; BIBBY, M. J. **Multivariate analysis**. New York, Academic Press, 1979.

MARRIOTT, F. H. C. The interpretation multivariate observation. London, **Academic Press**, 117 p. 1974.

MCHARO, T. M. **Association molecular markers with phenotypes in sweetpotatoes and liriopogons using multivariate statistical modeling**. 111f. Thesis (Ph.D), Louisiana State University, Lousiana, 2005.

MOLTALVÁN, R; ANDO, A.; ECHEVERRIGARAY, S. Use of seed protein polymorphism for discrimination of improvement level and geographic origin f upland rice cultivars. **Genetic and Molecular Biology**, v. 21, n. 4, p. 531-534, 1998.

MONTAGNON C., CUBRY P.; LEROY T. Amélioration génétique du caféier *Coffea canephora* Pierre: connaissnaces acquises, stratégieset perspectives. **Cahiers Agricultures**. n. 21: p.2-3. 2012.

MORA URPÍ, J.; CLEMENT, C. R. Races and population of peach palm found in the Amazon basin. FINAL report: **Peah Palm Germplasm Bank**. INPA, Manaus, p.78-94. 1988.

MORA URPÍ, J.; WEBER J. C.; CLEMENT, C. R. **Peach palm, Bactris gasipaes Kunth. Promoting the conservation and use of underutilized and neglected crops**. 20. Institute of Plant Genetics and Crop Plant Research, Gatersleben/ International Plant Genetic Resources Institute, Rome, Italy, 83p. 1997.

MOURA, M. C. C. L.; GONÇALVES, L. S. A.; SUDRÉ, C. P.; RODRIGUES, R.; AMARAL JÚNIOR, A. T.; PEREIRA, T. N. S. Algoritmo de Gower na estimativa da divergência genética em germoplasma de pimenta. **Horticultura Brasileira**. v. 28, n. 2, p. 155-161, 2010.

NEGREIROS, J. R. S.; BERGO, C. L.; MIQUELONI, D. P.; Lunz, A. M. P. Divergência genética entre progênies de pupunheira quanto a caracteres de palmito. Pesq. Agropec. Bras., Brasília, v.48, n.5, p.496-503, 2013.....

NIELSEN, L. R.; COWAN, R. S.; SIEGISMUND, H. R.; et al. 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, 2003.

NOGUEIRA, A. P. O.; SEDIYAMA, T.; CRUZ, C. D.; REIS, M. S.; PEREIRA, D. G.; JANGARELLI, M. Novas características para diferenciação de cultivares de soja pela análise discriminante. **Ciência Rural**, Santa Maria, v. 38, n. 9, p. 2427-2433, 2008.

OLIVEIRA, M. de S. P. de; FERREIRA, D. F.; SANTOS, J. B. dos. Divergência genética entre acessos de açaizeiro fundamentada em descritores morfoagronômicos. **Pesquisa Agropecuária Brasileira**, v. 42, p. 501-506, 2007.

OLIVEIRA, M. V. C.; BALIZA, D. P.; SOUZA, G. A.; CARVALHO, S. P.; ASSIS, L. H. B. Caracterização de clones de mandioca utilizando marcadores microssatélites. **Revista Ciência Agronômica**, v. 43, n. 1, p. 170-176, 2012.

OLIVEIRA, L. N. L. **Divergência genética em clones superiores de *coffea canephora* pierre ex froenher em Rondônia**. Dissertação de Mestrado. Universidade Federal do Amazonas. Manaus. 40 f, 2017.

QUINTAL, S. S. R.; VIANA, A. P.; GONÇALVES, L. S. A.; PEREIRA, M. G.; JÚNIOR, A. T. A. Genetic divergence among papaya accessions by morphoagronomic traits. Semina: **Ciências Agrárias**, Londrina, v. 33, n. 1, p. 131-142, 2012.

RAO, C. R.; MITRA, S. K. **Linear statistical inference and its applications**. New York, John Wiley, 1973.

REGAZZI, A. J. **Análise multivariada, notas de aula INF 766**. Departamento de Informática da Universidade Federal de Viçosa, v. 2, 2000.

REGAZZI, A. J. INF 766 - **Análise multivariada**. Viçosa: Universidade Federal de Viçosa, Centro de Ciências Exatas e Tecnológicas. Departamento de Informática, 166p. Apostila de disciplina. 2001.

ROCHA, M. C.; GONÇALVES, L. S. A.; RODRIGUES, R.; SILVA, P. R. A. da.; CARMO, M. G. F. do.; ABBOUD, A. C. S. Uso do algoritmo de Gower na determinação da divergência

genética entre acessos de tomateiro do grupo cereja. **Acta Scientiarum Agronomy**, Maringá, v. 32, n. 3, p. 423-431, 2010.

ROCHA, R. B.; TEIXEIRA, A. L.; RAMALHO, A. R.; SOUZA, F. F. Melhoramento de *Coffea canephora* – Considerações e Metodologias. In: MARCOLAN, A. L. & ESPINDULA, M. (Eds.). Café na Amazônia. Brasília, DF: **Embrapa**, v. 1, p. 217-236. 2015.

RODRÍGUEZ, V. M.; CARTEA, M. E.; PADILLA, G.; VELASCO, P.; ORDÁS, A. The nabicol: a horticultural crop in northwestern Spain. **Euphytica**, Wageningen, v. 142, n. 3, p. 237-246, 2005.

SANTOS, B. W. C. dos; FERREIRA, F. M.; SOUZA, V. F. de; CLEMENT, C. R.; ROCHA, R. B. Análise discriminante das características físicas e químicas de frutos de pupunha (*Bactris gasipaes* Kunth) do alto Rio Madeira, Rondônia, Brasil. **Revista Científica**, Jaboticabal, v. 45, n. 2, p. 154-161, 2017.

SILVA, F. M da. **Estratégias de condução de populações segregantes de soja portadoras do gene RR e seleção por meio de análises uni e multivariada**. Tese (Doutorado). Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, 76 p. 2015.

SILVERMAN, B. W.; JONESSOURCE, M. C. E. Fix and J. L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). **International Statistical Review**, v. 57, n. 3, p. 233-238, 1989.

SOUZA, F. de F.; CAIXETA, E. T.; FERRÃO, L. F. V.; PENA, G. F.; SAKIYAMA, N. S.; ZAMBOLIM, E. M.; ZAMBOLIM, L.; CRUZ, C. D. Molecular diversity in Coffea canephora germplasm conserved and cultivated in Brazil. **Crop Breeding and Applied Biotechnology**, Londrina, v. 13, p. 221-227, 2013.

SUDRÉ, C. P; CRUZ, C. D; RODRIGUES, R.; RIVA, E. M; AMARAL JÚNIOR, A. T; SILVA, D. J. H; PEREIRA, T. N. S. Variáveis multicategóricas na determinação da divergência genética entre acessos de pimenta e pimentão. **Horticultura Brasileira**, v. 24, n. 1, p. 88-93, 2006.

SUNARHARUM, W.B.; WILLIAMS, D.J.; SMYTH, H.E. **Complexity of coffee flavor: A compositional and sensory perspective**. Food Research International, v. 62, p. 315-325, 2014.

TORABI, M. R.; DING, K. Selected measurement and statistical issues in health education evaluation and research. **The International Electronic Journal of Health Education**, v. 1, p. 26-38, 1998.

WEBB, A. **Statistical Pattern Recognition**. Ed. John Wiley & Sons,Inc, 2 ed, 2002.
VAYLAY, R.; VAN SANTEN, E. Application of canonical discriminant analysis for the assessment of genetic variation in tall fescue. **Crop Science**, v. 42, p. 534-539, 2002.

ZHANG, H.; BERG, A. C.; MAIRE, M.; MALIK, J. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 25, 1995.

ZUIN, G. C.; FILHO, P. S. V.; KVITSCHAL, M. V.; GONÇALVES-VIDIGAL, M. C.; COIMBRA, G. K. Divergência genética entre acessos de mandioca-de-mesa coletados no município de Cianorte, região Noroeste do Estado do Paraná. **Ciências Agrárias**, Londrina, v. 30, n. 1, p. 21-30, jan./mar. 2009.