UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS AGRÁRIAS
CURSO DE DOUTORADO

NOVAS ESTRATÉGIAS DE SELEÇÃO DE DESCRITORES E COMPARAÇÃO DE MÉTODOS DE AGRUPAMENTO HIERÁRQUICOS E NÃO HIERÁRQUICOS EM ACESSOS DE MAMÃO (*Carica papaya* L.)

ANTONIO LEANDRO DA SILVA CONCEIÇÃO

CRUZ DAS ALMAS - BAHIA
NOVEMBRO - 2019

NOVAS ESTRATÉGIAS DE SELEÇÃO DE DESCRITORES E COMPARAÇÃO DE MÉTODOS DE AGRUPAMENTO HIERÁRQUICOS E NÃO HIERÁRQUICOS EM ACESSOS DE MAMÃO (*Carica papaya* L.)

ANTONIO LEANDRO DA SILVA CONCEIÇÃO
Engenheiro agrônomo
Universidade Federal do Recôncavo da Bahia, 2013

Tese apresentada ao Colegiado do Programa de Pós-Graduação em Ciências Agrárias da Universidade Federal do Recôncavo da Bahia, como requisito parcial para a obtenção do Título de Doutor em Ciências Agrárias (Área de Concentração: Fitotecnia).

Orientador: Prof. Dr. Carlos Alberto da Silva Ledo
Coorientadora: Profa. Dra. Fabiane de Lima Silva
Coorientador: Prof. Dr. Cosme Damião Cruz

CRUZ DAS ALMAS - BAHIA
NOVEMBRO - 2019

FICHA CATALOGRÁFICA

C744n Conceição, Antonio Leandro da Silva.
Novas estratégias de seleção de descritores e comparação de métodos de agrupamento hierárquicos e não hierárquicos em acessos de mamão (*Carica papaya* L.) / Antonio Leandro da Silva Conceição._ Cruz das Almas, BA, 2019.
149f.; il.

Orientador: Carlos Alberto da Silva Ledo.
Coorientadora: Fabiane de Lima Silva.

Tese (Doutorado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias Ambientais e Biológicas - Doutorado em Ciências Agrárias.

1.Mamão – Variabilidade genética. 2.Mamão – Melhoramento genético. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Cruz, Cosme Damião. III.Título.

CDD: 634.651

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas – UFRB.
Responsável pela Elaboração – Antonio Marcos Sarmento das Chagas (Bibliotecário – CRB5 / 1615).
Os dados para catalogação foram enviados pelo usuário via formulário eletrônico.

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS AGRÁRIAS
CURSO DE DOUTORADO

NOVAS ESTRATÉGIAS DE SELEÇÃO DE DESCRITORES E COMPARAÇÃO DE MÉTODOS DE AGRUPAMENTO HIERÁRQUICOS E NÃO HIERÁRQUICOS EM ACESSOS DE MAMÃO (*Carica papaya* L.)

COMISSÃO EXAMINADORA DA DEFESA DE TESE DE ANTONIO LEANDRO DA SILVA CONCEIÇÃO

Realizada em 01 de Novembro de 2019

Prof. Dr. Carlos Alberto da Silva Ledo
Embrapa Mandioca e Fruticultura - EMBRAPA
Examinador Interno (Orientador)

Profa. Dra. Simone Alves Silva
Universidade Federal do Recôncavo da Bahia - UFRB
Examinador Interno

Prof. Dr. Ricardo Franco Cunha Moreira
Universidade Federal do Recôncavo da Bahia - UFRB
Examinador Interno

Dr. Mauricio dos Santos da Silva
Instituto do Meio Ambiente e Recursos Hídricos - INEMA
Examinador Externo

Dra. Daiane Sampaio Almeida
Conselho Regional de Engenharia e Agronomia da Bahia – CREA/BA
Examinador Externo

DEDICATÓRIA

“Dedico esta Tese primeiramente a Deus, pela tua grandeza, pelo seu amor incondicional, pelo carinho, pelo cuidado com minha família, por nunca desistir de mim, por me amparar em todos os momentos”.

Aos meus pais Sinésio e Veralucia.

As minhas avós Maria e Araci

Aos meus sobrinhos, Joice, Tainá, Luiza e João Neto.

Ao meu amigo de 4 patas, Nicky.

“Louvem o nome do SENHOR, pois só o seu nome é exaltado; a sua glória está sobre a terra e o céu”. Salmos 148:13

AGRADECIMENTOS

À Universidade Federal do Recôncavo da Bahia pela oportunidade de realização deste curso.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa durante todo o período de realização deste doutorado.

À Embrapa Mandioca e Fruticultura, que disponibilizou estrutura física e equipamentos adequados para execução do presente trabalho.

Ao meu orientador Prof. Dr. Carlos Alberto da Silva Ledo, pela oportunidade, confiança, apoio e compreensão.

Aos professores do Programa de Pós-Graduação em Ciências Agrárias, pelos ensinamentos transmitidos.

Aos meus coorientadores Prof. Dr. Cosme Damião Cruz e a Profa. Dra. Fabiane de Lima Silva.

À colega de curso e amiga Dra. Gilmara Fachardo, pela disponibilidade e contribuição na realização desse trabalho.

Aos professores Endrigo Sampaio, Ronaldo Fiúza, Sebastião de Oliveira, Ricardo Franco, Simone Alves, Silvia Patrícia, Cláudia Gomes e Kuang Hongyu, pelo apoio e ensinamentos transmitidos.

Aos meus colegas do curso de Pós-graduação em Ciências Agrárias.

Aos meus amigos, Maurício, João, Clailto, Josemario, Jhel, Lili, Bia, Fátima, Mai, Joice e Thaise, pelo incentivo e apoio nessa jornada acadêmica. E a todos os outros que estiveram ao meu lado durante esses anos.

SUMÁRIO

Página

NOVAS ESTRATÉGIAS DE SELEÇÃO DE DESCRITORES E COMPARAÇÃO DE METODOS DE AGRUPAMENTO HIERÁRQUICOS E NÃO HIERÁRQUICOS EM ACESSOS DE MAMÃO (*Carica papaya* L.)

Autor: Antonio Leandro da Silva Conceição
Orientador: Dr. Carlos Alberto da Silva Ledo

RESUMO: O presente estudo teve como objetivo selecionar descritores quantitativos e qualitativos por meio de novas estratégias de seleção e comparar diferentes métodos de agrupamentos hierárquicos e não hierárquicos, com o intuito de obter maior conhecimento da diversidade genética da cultura do mamoeiro. Para tanto, foram utilizados 50 acessos pertencentes ao banco de germoplasma da Embrapa Mandioca e Fruticultura. Esses acessos foram avaliados por meio de 35 descritores quantitativos e 19 descritores qualitativos. Para a seleção de descritores quantitativos, utilizou-se o diagnóstico de multicolinearidade, combinado a análise de componentes principais proposta por Jolliffe e a contribuição relativa de Singh. Para validação da estratégia de seleção utilizou-se o fator de inflação de variância. Resultando na seleção de 24 descritores quantitativos. A seleção dos descritores qualitativos foi realizada por meio do nível de entropia dos descritores de Renyi (H), e por meio da Análise Fatorial Exploratória, utilizando o método da análise paralela proposto por Horn e o critério do autovalor > 1, sugerido por Kaiser para determinação do número de fatores a serem retidos. Os melhores resultados foram obtidos por meio do Nível de entropia e a Análise Fatorial Exploratória (critério de Kaiser) onde foram selecionados 47,37% e 52,63% dos descritores, respectivamente para esses métodos. Os agrupamentos obtidos por meio da análise de componentes principais para os dados quantitativos e da análise de correspondência múltipla para os dados qualitativos apresentaram os melhores resultados na comparação dos métodos hierárquicos e não hierárquicos, apresentando um padrão de agrupamento mais adequado para os conjuntos de dados avaliados.

Palavras-chave: variabilidade genética, análise multivariada, descritores morfoagronômicos

NEW DESCRIPTOR SELECTION STRATEGIES AND COMPARISON OF HIERARCHICAL AND NON-HIERARCHICAL CLUSTERING METHODS IN PAPAYA ACCESSES (*Carica papaya* L.)

Author: Antonio Leandro da Silva Conceição
Advisor: Dr. Carlos Alberto da Silva Ledo

ABSTRACT: The present study aimed to select quantitative and qualitative descriptors through new selection strategies and to compare different methods of hierarchical and nonhierarchical groupings, aiming to gain a better understanding of the genetic diversity of papaya culture. For this, 50 accessions belonging to the germplasm bank of Embrapa Mandioca and Fruticultura were used. These access wered evaluated using 35 quantitative descriptors and 19 qualitative descriptors. For the selection of quantitative descriptors, the multicollinearity diagnosis was used, combined with the principal component analysis proposed by Jolliffe and relative contribution by Singh's. To validate the selection strategy, the variance inflation factor was used. Resulting in the selection of 24 quantitative descriptors. The selection of qualitative descriptors was performed through the entropy level of Renyi (H) descriptors, and through Exploratory Factor Analysis, using the parallel analysis method proposed by Horn and the eigenvalue criterion> 1, suggested by Kaiser for determining the number of factors to be retained. The best results were obtained through Entropy Level and Exploratory Factor Analysis (Kaiser criterion) where 47.37% and 52.63% of the descriptors were selected, respectively for these methods. Those clustering obtained from the principal component analysis for quantitative data and multiple correspondence analysis for qualitative data showed the best results in the comparison of hierarchical and non-hierarchical methods, presenting a more adequate clustering pattern for the evaluated data sets.

Keywords: genetic variability, multivariate analysis, morphoagronomic descriptors.

REFERENCIAL TEÓRICO

Aspectos gerais da cultura do mamoeiro

O mamão (*Carica papaya* L.) é amplamente cultivado em regiões tropicais e subtropicais dos cinco principais continentes do mundo (MANSHARDT, 1992). Sendo seu fruto bastante consumido in natura ou industrializado. O mamão destaca-se por apresentar elevado valor nutricional, sendo rico em açúcares, minerais e compostos bioativos como os carotenoides, vitamina C e polifenóis.

Neste sentido, a Embrapa Mandioca e Fruticultura vem realizando pesquisas que visam à caracterização de variedades desenvolvidas pelo programa de melhoramento genético quanto ao teor de compostos bioativos, a fim de disponibilizar para a sociedade frutos com características iguais ou superiores às variedades comerciais (REIS et al., 2015).

A importância social da cultura do mamoeiro é também de grande relevância, por ser geradora de empregos (diretos e indiretos) e renda, haja vista a absorção de mão de obra durante o ano todo, já que os tratos culturais, a colheita e a comercialização são efetuadas de maneira contínua nas lavouras, além de os plantios serem renovados, em média, a cada 2 ou 3 anos, garantindo a permanência do homem no campo e contribuindo para a redução do êxodo rural (Dantas et al., 2013).

O mamoeiro, *Carica papaya* L., é uma das fruteiras mais cultivadas e consumidas nas regiões tropicais e subtropicais do mundo, com grande expressão econômica dentre as espécies tropicais (CHEN et al., 1991). É de significativa importância para o Brasil, um dos principais produtores da fruta, com produção de 1,58 milhão de toneladas em 2013 (FAO, 2015). Entre os estados brasileiros que produzem mamão destaca-se a Bahia, com 718.726 toneladas em 2013, Espírito Santo com 404.720 toneladas, Minas Gerais com 126.849 toneladas, Ceará com 118.372 toneladas e Rio Grande do Norte com 69.925 toneladas (IBGE, 2017). Em todo o mundo, a Índia lidera o produção de mamão, seguido pelo Brasil, Nigéria, México, Indonésia, República Dominicana, Tailândia, República Democrática do Congo, Peru e Filipinas (FAOSTAT, 2017).

Na cultura do mamoeiro, os elementos climáticos que mais influenciam no seu cultivo são a disponibilidade de água, temperatura e umidade relativa do ar. A cultura apresenta um bom desenvolvimento em locais com precipitação

pluviométrica de 1.500 mm anuais, bem distribuídas, temperatura média anual de 25ºC e umidade relativa entre 60% e 85%. Os solos de textura média, profundos, permeáveis e com bom teor de matéria orgânica são os mais indicados para plantio (MAPA, 2017).

No Brasil, as cultivares de mamão mais comumente exploradas são classificadas conforme o tipo de fruto, o grupo Solo, e o grupo Formosa. Cultivares de mamão do grupo Solo apresentam frutos de cor vermelha e tamanho pequeno (entre 300 e 650g), caracteriza-se também pela precocidade na produção, nesse grupo se encontra a maioria das cultivares utilizadas no mundo. Já as cultivares do grupo Formosa possuem polpa vermelha-alaranjada e tamanho médio (entre 1000 a 1300g), o grupo Formosa é principalmente composto por híbridos comerciais, mas também podem incluem algumas linhagens (DIAS et al., 2011).

Origem do mamoeiro, caracterização botânica e domesticação

Quanto a origem de *Carica papaya* L*.,* é difícil saber com exatidão, dada a sua ampla distribuição pelos espanhóis e sua grande capacidade de adaptação às condições dos ambientes subtropical e tropical, é amplamente distribuída em torno da maioria das regiões subtropicais e tropicais do mundo (MING e MOORE, 2013). No entanto, Vavilov (1987) descreveu três centros de origem da maioria das espécies: o centro mesopotâmico, o Centro Mesoamericano e o centro norte chinês. O Centro Mesoamericano é o centro de origem de importantes culturas tropicais, e foi sugerido como um bom candidato para ser também o centro de origem de *C. papaya* L*.* (HARLAN, 1971).

Carica papaya L., descrito por von Linnaeus (1753), pertence à família Caricaceae que é formada por 6 gêneros e 35 espécies. Os gêneros pertencentes a esta família são Carica (1 espécie), Jarilla (3 espécies), Horovitzia (1 espécie), Jacaratia (7 espécies), Vasconcellea (21 espécies) e Cylicomorpha (2 espécies), de acordo com Badillo (1971,1993, 2000). Exceto para o último gênero, todos os outros cinco gêneros dessa família têm uma origem americana (Scheldeman et al., 2011). Alguns autores sugeriram que *C. papaya* L. se originou no norte da América do Sul (BADILLO, 1971; PRANCE, 1984).

O mamoeiro é normalmente uma planta muito alta, de um só tronco, semi-lenhosa, com crescimento rápido e indeterminado (1 a 3m durante o primeiro

ano). Ocasionalmente, o crescimento vegetativo vigoroso pode induzir a ruptura axilar e se ramificando nas porções mais baixas da planta, que raramente excede alguns centímetros de comprimento (MORTON, 1987).

O mamoeiro se desenvolve muito rapidamente, levando 3 a 8 meses desde a germinação das sementes até floração (fase juvenil) e 9 a 15 meses para colheita (PATERSON et al., 2008). A planta pode viver até 20 anos. No entanto, devido à altura excessiva e as restrições patológicas, a vida comercial de um pomar de mamoeiro é normalmente de 2 a 3 anos (FUENTES; SANTAMARÍA, 2014).

Os frutos bem polinizados podem ter até 600 sementes ou mais, apresentando coloração preta. O embrião é reto e ovoide, com cotilédones achatados (FISHER, 1980). As sementes são revestidas por uma massa mucilaginosa derivada da epiderme pluriestratificada do tegumento externo (ROTH, 1977). O embrião na maturidade fisiológica é fechado em uma sarcotesta gelatinosa. Por baixo pode ser observado uma mesotesta compacta e tegumentos externos e internos. O endosperma é composto de células com paredes finas, com óleo abundante e proteína aleurona nos grãos, sem amido na maturidade. (FISHER, 1980; TEIXEIRA et al., 2007)

A plasticidade fenotípica da raiz do mamoeiro é alta, se ajustando ao longo de todo seu cliclo de vida, em relação ao tamanho da raiz, ao seu número, sua distribuição e orientação ajustam-se facilmente ao longo do perfil do solo e as suas diversas condições (FISHER e MUELLER, 1983; MARLER e DISCEKICI, 1997).

Nas plantas de mamão, o único tronco fornece suporte estrutural, capacidade de armazenamento, substâncias de defesa, altura e habilidade competitiva e traz um fluxo bidirecional de água, nutrientes, vários compostos orgânicos e lançamento de sinais químicos e físicos que regulam as relações com a raiz (REIS et al., 2006).

A planta produz grandes folhas palmadas de aproximadamente 0,6 m^2, com cinco a nove lóbulos pinados de várias larguras (40-60 cm), dispostos em um padrão espiral e agrupados na seção superior de indivíduos adultos (MORTON, 1987; MING et al., 2008).

O mamoeiro são plantas de fotossíntese C_3 (CAMPOSTRINI e GLENN, 2007). A temperatura ideal para o crescimento é de 21 a 33 °C, segundo a qual podem produzir 2 folhas/semana e de 8 a16 frutas/mês. Temperaturas inferiores a 10 ° C não são bem toleradas, Allan (2002, 2005).

O mamoeiro possui três formas florais típicas: flores masculinas, femininas e hermafroditas, que dão origem a plantas masculinas (andróicas), plantas femininas (ginóica) e plantas hermafroditas (andromonóica), (MEDINA; CORDEIRO, 1994).

O sexo das flores determina os diferentes formatos dos frutos do mamoeiro. As plantas masculinas não produzem frutos e quando o fazem, não apresentam valor comercial. As plantas femininas geram frutos de formato arredondado ou ligeiramente ovalados, de baixo valor comercial. Já as plantas hermafroditas produzem frutos valorizados no mercado, sendo estas utilizadas na maioria dos plantios comerciais (COSTA; PACOVA, 2003).

A identificação do sexo das plantas só é feita na época do primeiro florescimento, que ocorre por volta de 4 meses após o plantio. Com isso, os produtores precisam triplicar o número de mudas no campo, para possibilitar a sexagem, com intuito de deixar apenas uma planta hermafrodita por cova. No entanto, esta estratégia aumenta a competição entre as plantas nas fases iniciais de seu desenvolvimento e eleva o custo de produção da cultura (ARANGO et al., 2008).

Os frutos do mamoeiro são bagas e mostram uma grande diversidade em tamanho e forma. Os frutos de plantas hermafroditas tendem a ser alongados e variam de forma cilíndrica a em forma de pera, enquanto os frutos das plantas femininas tendem a ser redondos. O tamanho do fruto podem variar amplamente, variando de menos de 100 g em algumas acessos selvagens a mais de 10 kg em certas variedades. Os frutos do mamoeiro são climatéricos e no amadurecimento a uma alta produção de etileno, que pode começar apenas horas após a colheita. A medida que amadurecem, os frutos do mamão mudam de cor, firmeza, composição de carboidrato e produção de compostos secundários, responsáveis pela cor e fragrância dos frutos. A cor dos frutos maduros pode variar de amarelo para o vermelho salmão (SCHWEIGGERT et al., 2011).

Recursos genéticos e Melhoramento genético do mamoeiro

Os desafios para a segurança alimentar em todo mundo envolvem fatores como crescimento populacional, diminuição da produtividade, redução da base de recursos e as mudanças climáticas. Nesse sentido os recursos genéticos vegetais tem um valor e importância inestimável (SALGOTRA e GUPTA, 2015). Eles

desempenham um grande e crescente papel na segurança alimentar mundial e no desenvolvimento econômico. São cruciais para o crescimento agrícola sustentável e oferecem segurança de subsistência para a sociedade cujo os meios de produção giram em torno da agricultura e/ou do trabalho no campo.

O principal objetivo da conservação de recursos genéticos é explorar, coletar e preservar complexos de genes adaptativos para uso presente ou futuro (HAMMER e TEKLU 2008). A conservação das espécies silvestres que estão geneticamente relacionadas com as espécies cultivadas é muito importante. Não conservar essas espécies pode prejudicar a segurança alimentar do planeta. Um maior uso da diversidade genética das plantas é essencial para enfrentar este e outros desafios futuros (SALGOTRA e GUPTA, 2015). Recursos Geneticos Vegetais são os blocos de construção para a melhoria das culturas agrícolas, para a industria e o setor de agroprocessamento. Recursos genéticos vegetais são os pilares sobre os quais a segurança alimentar mundial depende, especialmente a população global em expansão (OGWU et al., 2014).

Na implantação de um programa de melhoramento genético, uma das principais necessidades do melhorista é o conhecimento do germoplasma disponível e a capacidade de identificar plantas que possuam genes de interesse para o programa (WEEDEN et al., 1994). O grau de relacionamento genético entre genótipos pode ser estimado através de diferentes métodos, podendo ser baseado em dados de genealogia, descritores morfológicos e marcadores moleculares ao nível de DNA (MELCHINGER et al., 1994).

Os programas de melhoramento do mamoeiro visa melhorar características relacionadas à própria planta e ao fruto, como vigor, ausência de ramificação lateral, frutificação precoce, baixo porte, ausência ou ocorrência mínima de carpeloidia, pentandria e esterilidade de verão, resistência a doenças e pragas, alta produção, uniformidade do tamanho do fruto, polpa espessa e cavidade ovariana pequena, alto teor de sólidos solúveis e longevidade dos frutos na pós-colheita (LUNA, 1986; GIACOMETTI; FERREIRA, 1988).

O melhoramento genético do mamoeiro pode contribuir substancialmente para maior produtividade. Este objetivo pode ser alcançado pela aplicação de métodos de melhoramento e seleção de variedades com rendimentos superiores, bem como pela obtenção de linhagens ou híbridos com resistência a doenças e

pragas, o que certamente contribuirá para o melhoramento da cultura, limitada pela grande incidência e distribuição de doenças viróticas (HARKNESS, 1967; ISHII e HOLTZMANN, 1963; GABROVSKA et al., 1967).

Mediante expedições de coletas realizadas pela Embrapa Mandioca e Fruticultura (CNPMF), em parceria com a Embrapa Recursos Genéticos e Biotecnologia (Cenargen), em 1995 teve início a formação do Banco Ativo de Germoplasma de Mamão (BAG- Mamão) da Embrapa. Atualmente, o maior BAG-Mamão do Brasil pertence à Embrapa Mandioca e Fruticultura, fica localizado em Cruz das Almas - BA. Possui quatro espécies (*Carica papaya*, *Vasconcellea cauliflora, Vasconcellea quercifolia* e *Jacaratia spinosa*), compreendidas em 104 acessos de *C. papaya*, 1 acesso de *V. cauliflora*, 1 acesso de *V. quercifolia* e 3 acessos de *J. spinosa.* A conservação dos acessos é feita sob condições de campo e por meio de armazenamento de sementes em câmara fria (10°C) (OLIVEIRA, 2015).

Caracterização e Avaliação do Germoplasma

Na caracterização da diversidade genética das espécies vegetais, animais e de microrganismos, os pesquisadores têm o interesse em agrupar genótipos similares, de maneira que as maiores diferenças ocorram entre os grupos formados. Neste aspecto, técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas e de agrupamento, podem ser aplicadas neste tipo de estudo. A adoção de uma, entre as técnicas citadas, varia de acordo com o padrão de resultado desejado e com a informação disponível, seja ela característica morfológica, fisiológica, ecológica ou genético-molecular (CRUZ et. al., 2011). Dentre estas, pode ser destacada a análise de agrupamento que é muito utilizada pelos pesquisadores tanto da área de melhoramento genético vegetal quanto na caracterização morfológica de novos acessos, ou seja, na caracterização morfológica de coleções de constituições genéticas geralmente mantidas em bancos de germoplasma e ainda pouco conhecidas pelos melhoristas (KOOP et al., 2007).

Nos programas de melhoramento de plantas, a informação quanto à diversidade e à divergência genética, dentro de uma espécie, é essencial para o uso racional dos recursos genéticos. Os estudos sobre a diversidade genética nas

coleções de germoplasma podem ser realizados a partir de descritores morfológicos de natureza qualitativa ou quantitativa. E neste caso várias técnicas estatísticas podem ser utilizadas na predição da diversidade presente (BARBOSA, 2010).

Devido a sua enorme importância para programas de melhoramento e de conservação da diversidade genética, a caracterização dos acessos possibilita a quantificação e a utilização da variabilidade genética de modo eficiente (VALLS, 2007). Tais etapas fazem parte do melhoramento genético de plantas, e este processo, faz com que acessos introduzidos em bancos de germoplasma sejam avaliados de tal forma que auxiliem o melhorista a identificar características desejáveis (novos genes) (CHIORATO, 2004).

A caracterização e a avaliação de um germoplasma visam, basicamente, descrever um acesso pelas suas características morfológicas, fisiológicas, agronômicas, bioquímicas, citogenéticas ou moleculares. Essas características são úteis para a identificação de genes ou genótipos de interesse (RAMANATHA-RAO 2001). De acordo com VALLS (2007), o processo de caracterização e avaliação de germoplasma pode ser classificado em cinco etapas subsequentes e complementares: *a*) identificação botânica; *b*) cadastro dos acessos (preenchimento dos dados de passaporte); *c*) caracterização; *d*) avaliação preliminar; *e*) avaliação.

Seleção de Descritores Morfoagronômicos

Nas coleções de germoplasma, o termo descritor é utilizado para se referir a um atributo ou caráter que se observa ou se mensura nos acessos (QUEROL, 1993), sendo capaz de discriminar um acesso de outro. Nos bancos de germoplasma, frequentemente há um grande número de acessos que necessita ser avaliado, além de ser regra geral as observações e a mensuração de um grande número de descritores (PEREIRA, 1989). Em muitos casos, são obtidos sem nenhum critério sobre sua real contribuição para a variabilidade e esse tipo de procedimento, além de produzir a duplicação da mesma informação, tem contribuído para uma análise multivariada, confusa e de difícil interpretação (DIAS, 1994).

O descarte deve se mostrar efetivo na representação da variação total, além de proporcionar uma redução nos gastos com mão-de-obra e no tempo destinado à tomada de dados. A seleção de descritores tem sido realizada com base em várias análises estatísticas, podendo-se mencionar: a regressão e interdependência de dados, o coeficiente de repetitividade, variáveis canônicas e componentes principais (CRUZ, 1990). Contudo, a análise de componentes principais vem se destacando como a metodologia mais empregada em bancos e/ou coleções de germoplasma, pois além de identificar os descritores mais importantes na contribuição de variação total disponível entre os indivíduos analisados, fornece indicação para eliminar os que pouco contribuem (DIAS, 1994; ALVES, 2002). Porém, pelo fato de alguns trabalhos criticarem o emprego da análise de componentes principais no descarte de descritores, especialmente quando se utiliza o método da seleção direta, considerando um procedimento drástico, faz-se necessária a avaliação de sua eficiência (ALVES, 2002).

Outro método utilizado na seleção de descritores quantitativos é baseado no critério proposto por Singh (1981), levando-se em consideração a contribuição relativa dos descritores para divergência genética, baseado na estatística S.j. onde, considera-se que:

$$D^2_{ii'} = \delta' \psi^{-1} \delta = \sum_{j=1} n \sum_{j'=1} n \ \omega_{jj} d_j d_{j'}$$

em que ω_{jj} é o elemento da j-esima coluna da inversa da matriz de variância e covariância residuais. O total das distâncias envolvendo todos os pares de genótipos é dado por: $SSD^2_{ii'} = S\,D^2_m = SS._j$. Os valores percentuais de S.j constituíram a medida da importância relativa da variável j para o estudo da diversidade genética.

Ao fim das análises que envolvam seleção de descritores é comum a realização do coeficiente de correlação entre os descritores, para verificar a associação entre os descritores descartados e os remanescentes. A significância do coeficiente de correlação é verificada pelo teste de t.

O coeficiente de correlação de Spearman é calculado a partir da seguinte expressão:

$$r_s = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N^3 - N}$$

Em que:

r_s : Correlação de Spearman;

D_i: Diferença entre postos;

N: Número de pares.

O teste t é calculado pela seguinte expressão:

$$t = r_s \sqrt{\frac{N-2}{1-r_s}}$$

Diagnóstico de multicolinearidade

Quando se realiza o processamento dos dados, deve-se ter algumas preocupações básicas, pois em muitos casos encontram-se resultados aparentemente absurdos, principalmente em estudos baseados nas informações de matrizes de correlações. Pode haver um autovalor nulo, o que torna a matriz singular e muitas vezes inadequada para certos estudos, pois não admite inversa comum. Estes casos são frequentes quando existem problemas de multicolinearidade. Quando os descritores estão correlacionados entre si, há uma inter-relação ou multicolinearidade (CRUZ, 2001).

O diagnóstico de dependência linear ou multicolinearidade na matriz de correlação residual, tem por objetivo identificar os coeficientes de correlação elevados, os quais podem causar multicolinearidade, sendo recomendável, portanto, o descarte dos descritores pertinentes. Para esse descarte, pode se consider os fatores de inflação da variância (VIF) na análise descritiva das correlações e o número de condição (NC) (CRUZ, 2001).

O VIF é dado por:

$$VIF_i = \frac{1}{1-R_i^2}$$

onde R_i^2 é o coeficiente de determinação múltipla. É uma medida do grau em que cada variável independente é explicada pelas demais variáveis independentes. Quanto maior for o fator de inflação da variância, mais severa será a multicolinearidade. Valores maiores que 10 correspondem a um coeficiente de determinação R_i^2 > 0,90 e isso é considerado inaceitável e motivo de preocupação. (MARQUARDT,1970; MASON et al., 1989; NETER et al.,1989; KENNEDY, 1992; TAMHANE e DUNLOP, 2000; KUTNER et al., 2004). Outros autores, como Hair (2005), sugerem que os fatores de inflação da variância não devem exceder 4 ou 5, isso dependerá do conhecimento teórico do pesquisador sobre o problema estudado.

Dentre os métodos usados para diagnóstico de multicolinearidade, um dos mais utilizados é o proposto por Montgomery e Peck (1981). Em que o número de condição (NC) consiste na razão entre o maior e o menor autovalor (λ): NC = $\lambda_{máx}$ / $\lambda_{mín}$

Onde o NC<100 é considerado como multicolinearidade fraca e não constitui problema sério, já com o número de condição 100<NC<1000, é considerado uma colinearidade moderada a forte, valores superiores a 1000 é considerada multicolinearidade severa.

A seleção dos descritores qualitativos é geralmente realizada por meio do nível de entropia dos descritores (H), proposto por Renyi (1961), de acordo com o seguinte modelo:

$$\mathrm{H} = -\sum_{i=1}^{s} p_i \ln p_i$$

Onde a Entropia é uma medida da frequência da distribuição de (n) acessos P = (p_1, p_2 ... p_s), sendo: $p_1 = f_i/n$ e ($p_1 + p_2 + ... + p_s = 1$) desde que ($n = f_1 + f_2 + ... + f_s$), onde f_1, f_2, ... f_n, são as contagens de cada uma das classes (s) no descritor considerado.

Análise Fatorial Exploratória

Análise fatorial exploratória (AFE) é uma abordagem estatística que pode ser usada para analisar inter-relações entre um grande número de descritores

(variáveis) e explicar esses descritores em termos de suas dimensões inerentes comuns (fatores). O objetivo é encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas (fatores) com uma perda mínima de informação. Pelo fato de fornecer uma estimativa empírica da estrutura das variáveis consideradas, a análise fatorial se torna uma base objetiva para criar escalas múltiplas. A análise fatorial pode ser utilizada para examinar os padrões ou relações latentes para um grande número de variáveis e determinar se a informação pode ser condensada ou resumida a um conjunto menor de fatores (HAIR, 2009).

Uma série de métodos tem sido sugeridos para determinar o número de fatores a serem retidos na análise fatorial exploratória (AFE). Dois dos métodos mais utilizadas são o da regra de autovalor maior do que um, proposta em 1960 por Kaiser e a Análise Paralela de Horn (1965).

O critério de Kaiser, mais conhecido como *eigenvalue > 1* (PATIL et al., 2008), propõe uma avaliação rápida e objetiva do número de fatores a ser retido. A lógica por trás do critério de Kaiser é simples: cada fator retido apresenta um *eigenvalue* que se refere ao total de variância explicada por este fator. A soma total dos *eigenvalues* é sempre igual ao número de itens utilizados na análise (utilizando uma escala de 10 itens, a soma dos 10 *eigenvalues* retidos é igual a 10). Assim, um componente com *eigenvalue* < 1 apresenta um total de variância explicada menor do que um único item. Como o objetivo das análises fatoriais é reduzir um determinado número de descritores observados em um número menor de fatores, apenas fatores com *eigenvalue* > 1 são retidos (FLOYD e WIDAMAN, 1995). Alguns trabalhos utilizaram o critério de Kaiser-Guttman (CLIFF, 1988; VASSEND e SKRONDAL, 1997; WOOD e KARDASH, 2002; POHLMANN, 2004; HOYLE e DUVALL, 2004; SCHULENBERG e MELTON, 2010; VAN DER EIJK e ROSE, 2015); GOLINO e EPSKAMP, 2017; POPESKA et al., 2018; AUERSWALD e MOSHAGEN, 2019).

A análise paralela (PA) proposta por Horn (1965) leva em consideração a proporção de variância resultante do erro amostral. Análise Paralela surgiu como uma das técnicas mais fortemente recomendadas e é considerado um procedimento adequado para determinar o número de fatores a serem retidos (ZWICK e VELICER, 1986; GLORFELD, 1995; FABRIGAR et al., 1999; VELICER,

EATON e FAVA. ,2000; WANG, 2002, HAYTON, ALLEN, e SCARPELLO, 2004; PERES-NETO, JACKSON e SOMERS, 2005; HENSON e ROBERTS, 2006; PATIL et al., 2007 E 2008; LORENZO-SEVA et al., 2011; RUSCIO e ROCHE, 2012; GARRIDO, ABAD, e PONSODA, 2012; COURTNEY e GORDON, 2013; BAGLIN, 2014; LEDESMA, VALERO-MORA e MACBETH, 2015; ÇOKLUK e KOÇA, 2016; O'BRIEN et al., 2017; HONGYU, 2018; LIM e JAHNG, 2019), entre outros.

Quando não é clara a interpretação dos fatores extraídos utilizando-se as cargas fatoriais, ou seja, existindo mais de um fator e a contribuição dos descritores para cada um deles não é suficientemente clara, um procedimento adequado que pode ser utilizado é a rotação dos eixos coordenados (JOHNSON e WICHERN, 2007). Existem vários métodos de rotação, porém o método Varimax é um método de rotação ortogonal mais comumente utilizado dentre os métodos ortogonais, que procura minimizar o número de descritores que apresentam altas cargas em cada fator (HAIR et al., 2006). O método Varimax procura dar aos fatores maior potencial de interpretabilidade, ou seja, torna a solução fatorial mais simples e pragmaticamente mais significativa (MARDIA et al., 2006; JOHNSON e WICHERN, 2007).

Análise de agrupamento

Antes de aplicar os métodos de agrupamento, o primeiro passo é avaliar se os dados são agrupáveis. Na literatura exitem algumas técnicas para avaliar a tendência de agrupamento, dentre elas a estatística de Hopkins. A estatística Hopkins, que é um teste estatístico hipotético, mede a agrupabilidade (tendência de cluster) de um determinado conjunto de dados (HOPKINS e SKELLAM, 1954; LAWSON e JURS, 1990).

Análise de agrupamentos é um grupo de técnicas multivariadas cuja finalidade principal é agregar objetos com base nas características que eles possuem. Ela tem sido chamada de análise *Q*, construção de tipologia, análise de classificação e taxonomia numérica. Essa variedade de nomes se deve ao uso de métodos de agrupamento nas mais diversas áreas, como psicologia, biologia, sociologia, economia, engenharia e administração. Apesar de os nomes diferirem nas disciplinas, os métodos têm uma dimensão em comum: classificação de acordo com relações entre os objetos sendo agrupados (HAIR et al., 2009). O resultado

obtido a partir da aplicação dessa técnica é um conjunto de grupos com coesão interna e isolamento externo (EVERITT, 1993), ou seja, elementos dentro de um mesmo grupo são tão similares quanto possível e são, ao mesmo tempo, tão dissimilares quanto possível dos elementos presentes nos demais grupos.

A análise de agrupamentos se assemelha à análise fatorial em seu objetivo de avaliar estrutura. Porém, diferem no sentido de que a primeira agrega objetos e a segunda está prioritariamente interessada em agregar variáveis. Além disso, a análise fatorial faz os agrupamentos com base em padrões de variação (correlação) nos dados, enquanto a análise de agrupamentos faz agregados baseados em distância (proximidade) (HAIR et al., 2009).

Muitos dos métodos desenvolvidos focam, especialmente, em dados caracterizados por variáveis contínuas (MINGOTI; LIMA, 2006). Quando há ocorrência de variáveis categóricas, algumas aproximações são usuais: transformá-las em contínuas, atribuindo valores numéricos às suas categorias, ou em binárias, fazendo com que cada uma das suas categorias se torne um descritor que represente presença ou ausência desse determinado atributo, transformar as contínuas em categóricas criando classes de valores ou ainda aplicar aos dados medidas específicas que tratam as observações conjuntamente (JOBSON, 1991-92; MINGOTI, 2005).

A variável estatística em análise de agrupamentos é determinada de maneira muito diferente do que ocorre em outras técnicas multivariadas. A análise de agrupamentos é a única técnica multivariada que não estima a variável estatística empiricamente, mas, em vez disso, usa a variável estatística como especificada pelo pesquisador. O foco da análise de agrupamentos é a comparação de objetos com base na variável estatística, não na estimativa da variável estatística em si. Isso torna a definição da variável estatística feita pelo pesquisador um passo crítico na análise. Com a variável estatística de agrupamento completamente especificada pelo pesquisador, a adição de variáveis ilegítimas ou a eliminação de relevantes podem ter um substancial impacto sobre a solução resultante. Assim, o pesquisador deve tomar muito cuidado com as variáveis usadas na análise, garantindo que elas tenham forte suporte teórico (HAIR et al., 2009).

A aplicação de diversas técnicas estatísticas em uma grande massa de informações objetivando melhorar a análise a partir da redução das dimensões de

observações e variáveis (mineração de dados ou data mining) vem se tornando, portanto, necessária às grandes empresas e a aplicabilidade da análise de agrupamentos é cada vez maior. A diversidade de possibilidades pode levar à dificuldade sobre qual técnica deve ser empregada e, consequentemente, levar a resultados diferentes, devido ao fato de os agrupamentos finais serem fortemente dependentes da metodologia usada. A escolha inadequada da técnica pode comprometer os resultados obtidos. Dessa forma, estudos comparativos contribuiriam na identificação dos métodos mais satisfatórios para uma determinada situação. O desempenho dos algoritmos propostos, combinados com as medidas adequadas, é de fato uma informação relevante que precisa ser cuidadosamente estudada (MATOS, 2007).

Critérios para escolha do número ideal de grupos

Nas análises de agrupamento, uma questão muito importante é de como se deve proceder para escolher o número final de grupos que define a partição do conjunto de dados analisados ou de outra forma, em qual passo *k* o algoritmo de agrupamento deve ser interrompido (MINGOTI, 2007). Infelizmente, não existe qualquer procedimento de seleção padrão e objetivo. Existem inúmeras propostas conforme observado pelo pacote *NbClust* do programa R (R CORE TEAM, 2019).

Como não existe um método totalmente confiável para identificar o número de clusters em um conjunto de dados, a escolha do melhor número de clusters pode muito bem depender do método de agrupamento usado. Portanto, uma análise de agrupamento deve sempre ser realizada para uma faixa (sensível) de diferentes números de grupos. O acesso a essa sequência de soluções é essencial para entender a operação de um algoritmo de agrupamento e para identificar tendências nos dados (HANDL et al., 2005).

Milligan e Cooper (1985) apresentaram um estudo comparativo de 30 critérios de corte para determinação do número de agrupamentos, e utilizando dados artificiais com número conhecido de agrupamentos, mostraram que critérios diferentes podem conduzir a resultados muito discrepantes. Entre as 30 abordagens diferentes, a abordagem proposta por Calinski e Harabasz (1974) supera as demais, esse índice também é conhecido como Pseudo-F. Contudo, com os resultados do estudo, fica evidente que critérios diferentes podem conduzir a

resultados muito discrepantes. Outro critério bastante utilizado, é o índice pseudo t2, proposto por Duda e Hart (1973) inserido no pacote “NbClust” (CHARRAD et al., 2013). Os índices Pseudo-F (CALINSKI e HARABASZ ,1974) e Pseudo T2 (DUDA e HART, 1973) são bons indicadores do número de grupos (MILLIGAN e COOPER, 1985; MINGOTTI, 2005).

As medidas de validação também podem ser utilizadas como critério para definir o número ideal de grupos, essas medidas fornecem quantidades significativas de informações que não podem ser obtidas usando apenas a inspeção visual. Existem ferramentas de validação diferentes e complementares, e o uso de um conjunto dessas ferramentas pode minimizar o risco de interpretar erroneamente os resultados e, assim, maximizar a confiança nos resultados obtidos. Uma boa solução de agrupamento tende a ter um desempenho razoavelmente bom sob várias medidas, se uma solução tiver bom desempenho apenas em uma delas, é provável que seja um artefato dos vieses do algoritmo empregado (HANDL et al., 2005).

Algumas das medidas de validação interna refletem a compactação, a conectividade e a separação das partições dos agrupamentos, como a conexão, que relaciona-se com a extensão em que as observações são colocadas no mesmo grupos que seus vizinhos mais próximos no espaço de dados, e é medida pela conectividade (HANDL et al., 2005), a medida de conectividade varia entre 0 e infinito, e quanto menor melhor. A Largura da silhueta (Rousseeuw, 1987) mede a homogeneidade interna, assume valores entre -1 e 1 e quanto mais próximo de 1 melhor. O Índice de Dunn (Dunn, 1974) quantifica a separação entre os agrupamentos, assume valores entre 0 e 1 e quanto maior melhor.

As medidas de validação de estabilidade comparam os resultados do armazenamento em gupo com base nos dados completos, com o armazenamento em grupo com a remoção de cada coluna, uma de cada vez. Essas medidas funcionam especialmente bem se os dados são altamente correlacionados. As medidas incluídas são APN (average proportion of non-overlap), AD (average distance), ADM (average distance between means) e FOM (figure of merit), (YEUNG et al., 2001; DATTA e DATTA, 2003). A (APN) que é a proporção média de observações não classificadas, assume valor no intervalo [0,1], próximos de 0 indicam agrupamentos consistentes. O (AD) distância média entre observações

classificadas no mesmo cluster nos casos com dados completos e incompletos. Assume valores não negativos, sendo preferíveis valores próximos de zero. O (ADM) esta relacionado com distância média entre os centroides quando as observações estão em um mesmo cluster. Assume valores não negativos, sendo preferíveis valores próximos de zero. Já o (FOM) medida do erro cometido ao usar os centroides como estimativas das observações na coluna removida. Assume valores não negativos, sendo preferíveis valores próximos de zero.

Agrupamento Hierárquico e Não-Hierárquicos

Informações sobre a estrutura genética de coleções de germoplasma é de suma importância para a conservação e utilização dos recursos genéticos (ODONG et al., 2011). A determinação da estrutura genética das coleções de germoplasma é realizado principalmente por métodos estatísticos multivariados tradicionais, tais como análise de agrupamento hierárquico aglomerativo, componentes principais e escala multidimensional, com base em dados genealógicos, agronômicos, fisiológicos, bioquímicos e moleculares (AMARAL et al., 2010).

As técnicas hierárquicas aglomerativas partem do princípio de que no início do processo de agrupamento tem-se n conglomerados, ou seja, cada conglomerado do conjunto de dados avaliados é considerado como sendo um grupo isolado. Em cada passo do algoritmo, os conglomerados vão sendo agrupados, formando novos grupos até o momento no qual todos os elementos considerados estão num único grupo (MINGOTI, 2007). Nesse sentido, devido a propriedade hierárquica, é possível construir um gráfico chamado de dendrograma que representa a “árvore” ou a história do agrupamento.

Dentre as diferentes técnicas hierárquicas aglomerativas, as mais utilizadas na avaliação dos recursos genéticos têm sido: *i*) UPGMA (*Unweighted Pair Group Method Arithmetic Average*) (SOKAL e MICHENER, 1958), *ii*) WARD (WARD, 1963), e *iii*) Ligação Simples (*Single Linkage*) (SNEATH, 1957) (MOHAMMADI e PRASANNA, 2003). O método UPGMA, agrupa os acessos com base nas médias das distâncias entre estes, a partir do par mais semelhante.

Milligan e Cooper (1985) simularam diferentes níveis de cortes em dendrogramas com base em diversas medidas de distâncias genéticas e concluíram que o método de Ligação Simples revelou resultados menos

consoantes com a estrutura genética dos materiais avaliados, ao passo que WARD e UPGMA permitiram os agrupamentos mais adequados, respectivamente, para tamanhos de grupos idênticos e diferentes.

Ao realizar a análise de agrupamento hierárquico aglomerativo, o interesse está em responder alguns questionamentos: *i*) existe concordância ente a distância original e a distância entre indivíduos representados pelo dendrograma; *ii*) o que pode o dendrograma dizer sobre a estrutura do conjunto de dados; e *iii*) qual é o número ótimo de grupos para um determinado conjunto de dados?

Nesse contexto, uma medida bastante comum de concordância entre a distância original e a distância no dendrograma é o coeficiente de correlação cofenético, no qual correlaciona as duas matrizes (distâncias observadas e as distâncias recuperadas da análise de agrupamento) (SNEATH e SOKAL, 1973).

As medidas de distância mais utilizadas para descritores quantitativos nos estudos genéticos são: a distância euclidiana, a distância euclidiana média, o quadrado da distância euclidiana média, a distância ponderada e a distância generalizada de Mahalanobis (CRUZ e CARNEIRO, 2006).

Para as descritores multicategóricos, características morfológicas atribuídas à estrutura de planta, assim como atributos que conferem qualidade aos produtos comercializados, como forma, cor e sabor – são comumente determinadas utilizando-se a distância de Cole-Rodgers (1997), na qual as características que normalmente não podem ser ordenadas são classificadas em escalas, podendo então ser analisadas como características quantitativas discretas (CRUZ e CARNEIRO, 2003).

Diferentemente dos métodos hierárquicos, nos procedimentos não-hierárquicos (método divisivo) já se sabe, a priori o número k de grupos a serem formados antes mesmo de se iniciar a análise (FERREIRA, 2011).

Uma das principais vantagens dos métodos não-hierárquicos em relação aos métodos hierárquicos é a possibilidade de um padrão poder mudar de agrupamento com a evolução do algoritmo, entretanto, como desvantagem está no fato do número de agrupamentos ter que ser escolhido a priori, o que pode inferir em interpretações errôneas sobre a estrutura dos dados caso o número de agrupamentos não seja o ideal. O problema quando se escolhe erroneamente o

número de agrupamentos é que o método irá impor uma estrutura aos dados, no lugar de buscar a estrutura inerente a estes (FUNG, 2001; KAINULAINEN, 2002).

Entre os diferentes métodos não-hierárquicos, o k-médias ou K-means (HARTIGAN e WONG, 1979) é o mais popular (MINGOTI, 2007; FERREIRA, 2011). O método é composto por quatro etapas: *i*) primeiramente escolhe-se k centroides para se inicializar o processo de partição; *ii*) cada elemento do conjunto de dados é, então, comparado com cada centroide inicial, por meio de uma medida de distância (em geral, a distância Euclidiana). O elemento é alocado ao grupo cuja distância é menor; *iii*) Para cada um dos *n* elementos amostrais recalculam-se os valores dos centroides para cada novo grupo formado, e repete-se o passo *ii*, considerando os centroides destes novos grupos; *iv*) Os passos *ii* e *iii* devem ser repetidos até que todos os elementos amostrais estejam “bem alocados” em seus grupos (MINGOTI, 2007). Geralmente, outro algoritmo de clustering (por exemplo, UPGMA) é executado inicialmente para determinar pontos de partida para os centros de cluster. O k-means é considerado um método de realocação.

Outro conhecido algoritmo de particionamento é o baseado em medoids (KAUFMAN; ROUSSEEUW, 1987). O algoritmo K-médias é sensível a observações aberrantes já que um objeto com valor extremamente grande pode substancialmente distorcer a distribuição dos dados. Ao invés de utilizar o valor médio dos objetos no grupo como um ponto de referência, um medóide pode ser usado, que é o objeto mais centralmente localizado no grupo. Deste modo, o método de particionamento pode ainda ser fornecido baseado no princípio da minimização da soma das dissimilaridades entre cada objeto e seu correspondente ponto de referência. Isto forma a base do método K-medoids (HAN; KAMBER; PEI, 2006; VELMURUGAN; SANTHANAM, 2011). O Particionamento em Torno de Medoids (Partitioning Around Medoids - PAM) foi um dos primeiros algoritmos K-medoids introduzidos.

O algoritmo PAM é baseado na busca de k objetos representativos entre os objetos do conjunto de dados. Neste algoritmo os objetos representativos são os chamados medoids dos grupos. Após encontrar um conjunto de k objetos representativos, os k grupos são construídos atribuindo cada objeto do conjunto de dados para o objeto representativo mais próximo. Alternativamente pode ser usado com a entrada por uma matriz de dissimilaridades entre objetos.

Dentre as medidas de validação externa disponíveis na literatura para comparar o agrupamento dos algoritimos de particionamento, a medida da entropia da distribuição de associações de grupos (MEILA, 2007) é bastante utilizada. Já para comparar os grupos gerados dentro de cada método de particionamento, quanto a semelhança e diferença entre os mesmos, o critério Pearson Gamma (HALKIDI et al., 2001), é uma boa alternativa.

Análise de agrupamento por meio de técnicas de análise de componentes principais (PCA) e análise de correspondência múltipla (MCA)

Na análise de divergência genética, vários métodos multivariados podem ser aplicados, como componentes principais, variáveis canônicas e métodos aglomerativos (CROSSA; FRANCO, 2004). A análise de componentes principais (PCA) consiste em técnica multivariada de transformação de um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais (PC). Todos os PC consistem em uma combinação linear de todas as variáveis originais, ortogonais e estimados com o propósito de reter, em ordem de estimação, o máximo da variação total contida nos dados (HAIR et al., 2009; HONGYU, 2015). O PCA é usado para reduzir a complexidade dos dados multidimensionais para aprimorar manuseio, análise, interpretação e visualização de dados multidimensionais (MACQUEEN, 1967).

Embora o PCA, seja uma ferramenta de análise de dados descritiva amplamente utilizada e adaptativa, ele também possui várias adaptações próprias que o tornam útil para uma ampla variedade de situações e tipos de dados em várias disciplinas. Adaptações de PCA têm sido propostas, entre outras, para dados binários, ordinais, composicionais, discretos, simbólicos ou dados com estrutura especial, como séries temporais ou conjuntos de dados com matrizes de covariância comuns. Abordagens relacionadas à PCA também desempenharam um importante papel direto em outros métodos estatísticos, como a regressão linear e até o agrupamento simultâneo de indivíduos e variáveis (JOLLIFE e CADIMA, 2016).

Na análise de agrupamento com base na PCA, ocorre o agrupamento dos indivíduos de acordo com sua variação (variâncias), ou seja, segundo seu comportamento dentro da população, representado pela variação do conjunto de

características que define o indivíduo. A técnica agrupa os indivíduos de uma população segundo a variação de suas características (HONGYU, 2015).

A análise de correspondência múltipla (MCA) é uma extensão da análise de correspondência simples para resumir e visualizar uma tabela de dados contendo mais de duas descritores categóricos. Também pode ser vista como uma generalização da análise de componentes principais quando os descritores a serem analisados são categóricos em vez de quantitativos (ABDI e WILLIAMS, 2010).

Análise de correspondência múltipla (MCA) faz parte de uma família de métodos descritivos (clustering, análise fatorial e análise de componentes principais) que revelam padrões em conjuntos de dados complexos. No entanto, especificamente, o MCA é usado para representar e modelar conjuntos de dados como "nuvens" de pontos em um espaço euclidiano multidimensional, isso significa que é distinto na descrição geométrica dos padrões, localizando cada variável/unidade de análise como um ponto em um espaço de baixa dimensão. Os resultados são interpretados com base nas posições relativas dos pontos e sua distribuição ao longo das dimensões. A medida que as categorias se tornam mais semelhantes na distribuição, mais próximas (distância entre os pontos) elas são representadas no espaço (JOHNSON e WICHERN, 2007; GREENACRE e HASTIE, 1987).

Embora seja usada principalmente como uma técnica exploratória, pode ser particularmente poderosa, pois “descobre” agrupamentos de categorias variáveis nos espaços dimensionais, fornecendo informações importantes sobre as relações entre as categorias (tratamento multivariado dos dados através da consideração simultânea de múltiplas variáveis categóricas), sem a necessidade de atender a requisitos de premissas, como os exigidos em outras técnicas amplamente usadas para analisar dados categóricos (análise do qui-quadrado e o teste exato de Fischer) (AKTÜRK e KUMUK, 2007). A análise de Correspondência (Múltipla) é uma possibilidade bastante interessante para transformar as variáveis categóricas em poucos componentes principais contínuos, para então serem usados como entrada na análise de agrupamento.

Diante da grande relevância da cultura do mamoeiro, estudos para quantificar de forma mais precisa plantas cultivadas e seus parentes silvestres quanto a diversidade genética, bem como estudos de técnicas de seleção de descritores

morfoagronômicos, são de suma importância, pois proporcionam uma utilização mais efetiva dos acessos de mamão nos programas de melhoramento. Este estudo teve como objetivo avaliar novas estratégias de seleção de descritores e comparar métodos de agrupamento hierárquicos e não hierárquico em acessos de mamão, visando obter maior conhecimento acerca do grau de variabilidade presente nesse conjunto de indivíduos.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley interdisciplinary reviews**: **computational statistics**, v. 2, n. 4, p. 433-459, 2010.

ALLAN, P. Carica papaya responses under cool subtropical growth conditions. In: International Symposium on Tropical and Subtropical Fruits. **Acta Horticulturae**, v. 575, p. 757-763, 2000.

ALLAN, P. Phenology and production of Carica papaya'Honey “Honey Gold” cool subtropical conditions. **Acta Horticulturae**, v. 740, p. 217-223, 2005.

ALVES, R. M. **Caracterização genética de populações de cupuaçuzeiro, Theobroma grandiflorum (Will ex Spreng) Schum., por marcadores microssatélites e descritores botânico-agronômicos**.Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.146, 2002.

AMARAL, A. T.; VIANA, A. P.; GONÇALVES, L. S. A.; BARBOSA, C. D. Procedimentos multivariados em recursos genéticos. In: Telma Nair Santana Pereira. (Org.). **GERMOPLASMA: Conservação, Manejo e Uso no Melhoramento de Plantas**. 1ed.VIÇOSA: ARKA, p. 205- 250, 2010.

ARANGO, L.V.; REZENDE, C.R.; CARVALHO, S.P. Identificação antecipada do sexo do mamoeiro pelos caracteres físicos das sementes e padrões isoenzimáticos

das mudas. **Revista Corpoica - Ciencia y Tecnología Agropecuaria**, v.1, p.22-29, 2008.

AUERSWALD, M.; MOSHAGEN, M. How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. **Psychological methods**, 2019.

BADILLO, V. M. Monografia de la família Caricaeae. Editorial Nuestra América C. A. Maracay, Venezuela, p. 221, 1971.

BADILLO, V.M. (1993) *Caricaceae* e. Segundo Esquema. **Rev Fac Agron Univ Centr Venezuela**. v. 43, p. 1–111, 1993.

BADILLO, V.M. Carica L. VS. Vasconcella St.Hil. (Caricaceae) con la rehabilitacion de este último. Ernestia. Maracay, Venezuela. p. 74-79, 2000.

BAGLIN, J. Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR. **Practical Assessment, Research & Evaluation**, v. 19, n. 5, p. 2, 2014.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, n. 1, p. 1-27, 1974.

CAMPOSTRINI, E.; GLENN, D. M. Ecophysiology of papaya: a review. **Brazilian Journal of Plant Physiology**, v. 19, n. 4, p. 413-424, 2007.

COLE-RODGERS, P.; SMITH, D. W.; BOSLAND, P. W. A novel statistical approach to analyze genetic resource evaluations using Capsicum as an example. Crop Science, v. 37, n. 3, p. 1000-1002, 1997.

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

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos Biométricos Aplicados ao Melhoramento Genético**, vol. 2. Viçosa: UFV, p. 585, 2006.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. NbClust: An examination of indices for determining the number of clusters. **R package version**, v. 1, 2013.

CHIORATO, A. F. **Divergência genética em acessos de feijoeiro (Phaseolus vulgaris L.) do Banco de Germoplasma do Instituto Agronômico-IAC**. Tese de Doutorado. Dissertação (mestrado em agronomia), Campinas, Instituto Agronômico, p. 85, 2004.

CLIFF, N. The eigenvalues-greater-than-one rule and the reliability of components. **Psychological bulletin**, v. 103, n. 2, p. 276, 1988.

ÇOKLUK, Ö.; KOÇAK, D. Using Horn's Parallel Analysis Method in Exploratory Factor Analysis for Determining the Number of Factors. **Educational Sciences: Theory and Practice**, v. 16, n. 2, p. 537-551, 2016.

COSTA, A.F.S.; PACOVA, B.E.V. **A cultura do mamoeiro: tecnologias de produção**. Vitória, ES: Incaper, 2003.

COURTNEY, M. G. R.; GORDON, M. Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2. 0 to make more judicious estimations. **Practical assessment, research & evaluation**, v. 18, n. 8, p. 1-14, 2013.

CROSSA, J.; FRANCO, J. Statistical methods for classifying genotypes. **Euphytica**, v. 137, pp. 19-37, 2004.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas**. Tese (Doutorado em genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.188, 1990.

CRUZ, C. D. **Programa Genes: versão Windows; aplicativo computacional em genética e estatística**. UFV, 2001.

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

DANTAS, J.L.L.; JUNGHANS, D.T.; LIMA, J.F. **Mamão: o produtor pergunta, a Embrapa responde**. 2.ed. Brasília, p.176, 2013.

DATTA, S.; DATTA, S. Somnath. Comparisons and validation of statistical clustering techniques for microarray gene expression data. **Bioinformatics**, v. 19, n. 4, p. 459-466, 2003.

DIAS, L. A. dos S. **Divergência genética e análise multivariada na predição de híbridos e preservação de germoplasma de cacau** (Theobroma cacao L.). Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba. p.94, 1994.

DIAS, N. L. P.; DE OLIVEIRA, E. J.; DANTAS, J. L. L. Avaliação de genótipos de mamoeiro com uso de descritores agronômicos e estimação de parâmetros genéticos. **Pesquisa Agropecuária Brasileira**, v. 46, n. 11, p. 1471-1479, 2011.

DUDA, R. O.; HART, P. E. Pattern classification and scene analysis. John Wiley & Sons: New York, p.189–225, 1973.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. **Journal of cybernetics**, v. 4, n. 1, p. 95-104, 1974.

EVERITT, B. S. Cluster Analysis. New York: John Wiley & Sons, Inc., 1993.

FABRIGAR, L. R.; WEGENER, D. T.; MACCALLUM, R. C.; STRAHAN, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. **Psychological methods**, v. 4, n. 3, p. 272, 1999.

FAO. Food and Agriculture Organization. **The State of Food and Agriculture**. Disponível em :<ftp://ftp.fao.org/docrep/fao/009/a0800e/a0800e.pdf>. Acesso em abril de 2017.

FERREIRA, D.F. **Estatística multivariada**. Lavras, Universidade Federal de Lavras. 676p, 2011.

FISHER, J. B.; MUELLER, R. J. Reaction anatomy and reorientation in leaning stems of balsa (Ochroma) and papaya (Carica). **Canadian journal of botany**, v. 61, n. 3, p. 880-887, 1983.

FISHER, J. B. The vegetative and reproductive structure of papaya (*Carica papaya*). **Lyonia**. v.1, p.191–208. 1980.

FLOYD, F. J.; WIDAMAN, K. F. Factor analysis in the development and refinement of clinical assessment instruments. **Psychological assessment**, v. 7, n. 3, p. 286, 1995.

FUENTES, G.; SANTAMARÍA, J. M. Papaya (*Carica papaya* L.): Origin, Domestication, and Production. In: **Genetics and Genomics of Papaya**. [s.l.] Springer, p. 3–15, 2014.

FUNG, G. A Comprehensive Overview of Basic Clustering Algorithms. June 22, 2001.

GABROVSKA I.; VALDIVIESO, A.S.; BECQUER, A.; SAENZ, B. Las enfermidades virosas de la fruta bomba (*Carica papaya* L.) en Cuba. **Revista de Agricultura**, Piracicaba, v.1, p.1-21, 1967.

GARRIDO, L. E.; ABAD, F. J.; PONSODA, V. Performance of Velicer's minimum average partial factor retention method with categorical variables. **Educational and Psychological Measurement**, v. 71, n. 3, p. 551-570, 2011.

GIACOMETTI, D. C.; FERREIRA, F. R. Melhoramento genético do mamão no Brasil e perspectivas. In: Simpósio brasileiro sobre a cultura do mamoeiro, 2, 1988, Jaboticabal, SP. **Anais**... Jaboticabal, SP: FCA/UNESP, p.377-38, 1988.

GLORFELD, L. W. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. **Educational and psychological measurement**, v. 55, n. 3, p. 377-393, 1995.

GOLINO, H. F.; EPSKAMP, S. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. **PloS one**, v. 12, n. 6, p. e0174035, 2017.

GONÇALVES, L. S. A.; RODRIGUES, R.; AMARAL JÚNIOR, A. T.; KARASAWA, M.; SUDRÉ, C. P. Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. **Genetics and Molecular Research**, v. 7, p.1289-1297, 2008.

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

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Multivariate data analysis** (Vol. 6). 2006.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. Bookman Editora, 2009.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of intelligent information systems**, v. 17, n. 2-3, p. 107-145, 2001.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3rd. ed. San Francisco: Morgan Kaufmann Publisher, 2006.

HARLAN, J. R. Agricultural origins: centers and noncenters. **Science**, v. 174, n. 4008, p. 468-474, 1971.

HARKNESS, R.W. Papaya growing in Florida. Florida: Fla. Agr. Ext. Serv., 1967.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 28, n. 1, p. 100-108, 1979.

HAYTON, JAMES C.; ALLEN, DAVID G.; SCARPELLO, VIDA. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. **Organizational research methods**, v. 7, n. 2, p. 191-205, 2004.

HENSON, R. K.; ROBERTS, J. K. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. **Educational and Psychological measurement**, v. 66, n. 3, p. 393-416, 2006.

HONGYU K.; SANDANIELO V.L.M.; OLIVEIRA JUNIOR G.J. Análise de componentes principais: resumos teórico, aplicação e interpretação. **Enginneering and Science**, 5:1. 2015.

HONGYU, Kuang. Análise Fatorial Exploratória: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, v. 7, n. 4, p. 88-103, 2018.

HOPKINS, B.; SKELLAM, J. G. A new method for determining the type of distribution of plant individuals. **Annals of Botany**, v. 18, n. 2, p. 213-227, 1954.

HORN, J. L. A rationale and test for the number of factors in factor analysis. **Psychometrika**, v. 30, n. 2, p. 179-185, 1965.

HOYLE, R. H.; DUVALL, J. L. Determining the number of factors in exploratory and confirmatory factor analysis. **Handbook of quantitative methodology for the social sciences**, p. 301-315, 2004.

IBGE. Produção Agrícola Municipal, 2013. Disponível em: <ftp://ftp.ibge.gov.br/Producao_Agricola/Producao_Agricola_Municipal_[anual]/2013/pam2013.pdf>. Acesso em: 10 jun. 2017.

ISHII, Y.; HOLTZMANN, O.W. Papaya mosaic disease in Hawaii. **Plant Disease Reporter**, Beltsville, v. 47, p. 947-951, 1963.

JOBSON, J. D. **Applied Multivariate Data Analysis**. New York: Springer, 1991–92.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 6th. **New Jersey, US: Pearson Prentice Hall**, 2007.

JOLLIFE, I.T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions Royal Society** A, 374: 20150202. 2016.

KAINULAINEN, P.; NISSINEN, A.; PIIRAINEN, A.; JARMO K. HOLOPAINEN, T. K.; Essential oil composition in leaves of carrot varieties and preference of specialist and generalist sucking insect herbivores**. Agricultural and Forest Entomology**, v. 4, p. 211–216, 2002.

KAISER, H. F. The application of electronic computers to factor analysis. **Educational and psychological measurement**, v. 20, n. 1, p. 141-151, 1960.

KAUFMAN, L.; ROUSSEEUW, P. J. Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm. **Y. Dodge, Ed**, p. 405-416, 1987.

KAUFMAN, L.; ROUSSEEUW, P.J. Partitioning around medoids (program pam). **Finding groups in data: an introduction to cluster analysis**, p. 68-125, 1990.

KENNEDY, P.A. **Guide to Econometrics**, Blackwell, Oxford, 1992.

KOOP, M. M.; SOUZA, V.Q.; COIMBRA, J.L.M.; LUZ, V.K.; MARINI, N.; OLIVEIRA, A.C. Melhoria da correlação cofenética pela exclusão de unidades experimentais na construção de dendrogramas. **Revista da Faculdade de Zootecnia, Veterinária e Agronomia (Uruguaiana)**, v. 14, p. 46-53, 2007.

KUTNER, M. H.; NACHTSHEIM, C; NETER, J. **Applied linear models**. 5th ed. New York: McGraw-Hill Irwin, 2004.

LAWRENCE, C. J.; KRZANOWSKI, W. J. Mixture separation for mixed-mode data. **Statistics and Computing**. v. 6, p. 85–92, 1996.

LAWSON, R. G.; JURS, P. C. New index for clustering tendency and its application to chemical problems. **Journal of chemical information and computer sciences**, v. 30, n. 1, p. 36-41, 1990.

LEDESMA, R. D.; VALERO-MORA, P.; MACBETH, G. The scree test and the number of factors: a dynamic graphics approach. **The Spanish journal of psychology**, v. 18, 2015.

LIM, S.; JAHNG, S. Determining the number of factors using parallel analysis and its recent variants. **Psychological methods**, v. 24, n. 4, p. 452, 2019.

LINNAEUS, C. V. Species plantarum, 2 vols. **Laurentii Salvii, Holmiae**, 1753.

LORENZO-SEVA, U.; TIMMERMAN, M. E.; KIERS, H. A. The Hull method for selecting the number of common factors. **Multivariate behavioral research**, v. 46, n. 2, p. 340-364, 2011.

LUNA, J.V.U. Variedades de mamoeiros. Epamig, Belo Horizonte, MG. **Informe Agropecuário**, v.12, n.134, p.14-18, 1986.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate Analysis**. Califórnia: Academic Press, 2006.

MANSHARDT, R. “Papaya,” in *Biotechnology of perennial Fruit Crops*, eds F. A. Hammerschlag and R. E. Litz (Oxford: Cambridge University Press), 489–511, 1992.

MAPA. **Ministério da Agricultura, Pecuária e Abastecimento**. Portaria 251/2011. Disponível em: <http://sistemasweb.agricultura.gov.br/sislegis/action/.do?method=visualizarAtoPortalMapa&chave=825295627>. Acesso em: 26 fev. 2017.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, p. 281-297, 1967.

MARQUARIDT, D. W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. **Technometrics**, v. 12, n. 3, p. 591-612, 1970.

MASON, R. L.; GUNST, R. F.; HESS, J. L. **Statistical design and analysis of experiments: Applications to engineering and science**. New York: Wiley,1989.

MARLER, T. E.; DISCEKICI, H. M. Root development of ‘Red Lady’papaya plants grown on a hillside. **Plant and Soil**, v. 195, n. 1, p. 37-42, 1997.

MATOS, R. A. **Comparação de metodologias de análise de agrupamentos na presença de variáveis categóricas e contínuas**. Dissertação (mestrado em Estatística), Universidade Federal Minas Gerais, p. 16, 2007.

MEDINA, V.M.; CORDEIRO, Z.J.M. **Mamão para exportação**: aspectos técnicos da produção. Brasília, DF: EMBRAPA-SPI. 52p. 1994.

MEILĂ, M. Comparing clusterings—an information based distance. **Journal of multivariate analysis**, v. 98, n. 5, p. 873-895, 2007.

MELCHINGER, A.E.; GRANER, A.; SINGH M.; MESSMER, M. M. Relationships among European barley germplasm: I. Genetic diversity among winter and spring cultivars revealed by RFLPs. **Crop Science**, v. 34, p. 1191-1199, 1994.

MILLIGAN, G. W.; COOPER, M. An examination of procedures for determining the number of cluster in a data set. **Psychometrika**, v. 50, p. 159-179, 1985.

MING, R.; YU, Q.; BLAS, A.; CHEN, C.; NA, J. K.; MOORE, P. H. Genomics of Papaya a Common Source of Vitamins in the Tropics. **Genomics of tropical crop plants**, p. 405-420, 2008.

MING, R.; MOORE, P. H. **Genetics and genomics of papaya**. Springer Science & Business Media, 2013.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada**. Belo Horizonte: Editora UFMG, 2005.

MINGOTI, S. A.; LIMA, J. O. Comparing SOM neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. **European Journal of Operational Research**, v. 174, n. 3, p. 1742–1759, 2006.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada***.* Uma Abordagem Aplicada. Ed. UFMG, 2007.

MOHAMMADI, S. A.; PRASANNA, B. M. Analysis of genetic diversity in crop plants-salient statistical tools and considerations. **Crop Science**, v. 43, p. 1235-1248, 2003.

MONTGOMERY, D.C.; PECK, E.A. **Introduction to linear regression analysis**. New York: John Wiley & Sons, p. 504, 1981.

MORTON, J. **Papaya**. In: Fruits of warm climates. Julia F. Morton, Miami, p. 336–346, 1987.

NETER, J.; WASSERMAN, W.; KUTNER, M. G. Applied linear regression analysis. **Homewood, IL: Irwin**, 1989.

O'BRIEN, R.; PAN, X.; COURVILLE, T.; BRAY, M. A.; BREAUX, K.; AVITIA, M.; CHOI, D. Exploratory factor analysis of reading, spelling, and math errors. **Journal of Psychoeducational Assessment**, v. 35, n. 1-2, p. 7-23, 2017.

ODONG, T. L.; VAN HEERWAARDEN, J.; JANSEN, J.; VAN HINTUM, T. J.; VAN EEUWIJK, F. A. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data?. **Theoretical and applied genetics**, v. 123, n. 2, p. 195-205, 2011.

OGWU, M. C.; OSAWARU, M. E.; AHANA, C. M. Challenges in conserving and utilizing plant genetic resources (PGR). **International Journal of Genetics and Molecular Biology**, v. 6, n. 2, p. 16-23, 2014.

OLIVEIRA, G.A.F. **Identificação, caracterização e validação de marcadores minissatélites para o mamoeiro**. Dissertação de mestrado em Recursos Genéticos Vegetais, Universidade Federal do Recôncavo da Bahia e Embrapa Mandioca e Fruticultura, Cruz das Almas-BA. p. 15, 2015.

PATERSON, A. H.; FELKER, P.; HUBBELL, S. P.; MING, R. The fruits of tropical plant genomics. **Tropical Plant Biology**, v. 1, n. 1, p. 3-19, 2008.

PATIL, V. H.; SINGH, S. N.; MISHRA, S.; DONAVAN, D. T. Parallel analysis engine to aid determining number of factors to retain [Computer software]. **Instruction and Research Server**, University of Kansas, 2007.

PATIL, V. H.; SINGH, S. N.; MISHRA, S.; DONAVAN, D. T. Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one'criterion. **Journal of Business Research**, v. 61, n. 2, p. 162-170, 2008.

PEREIRA, V. A. **Utilização de análise multivariada na caracterização de germoplasma de mandioca (Manihot esculenta Crantz.)**. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, p. 180, 1989.

PERES-NETO, P. R.; JACKSON, D. A.; SOMERS, K. M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. **Computational Statistics & Data Analysis**, v. 49, n. 4, p. 974-997, 2005.

POHLMANN, J. T.Use and interpretation of factor analysis in The Journal of Educational Research: 1992-2002. **the Journal of Educational research**, v. 98, n. 1, p. 14-23, 2004.

POPESKA, B.; JOVANOVA-MITKOVSKA, S.; CHIN, M. K.; EDGINTON, C.; MO CHING MOK, M.; GONTAREV, S. Implementation of Brain Breaks® in the classroom and effects on attitudes toward physical activity in a Macedonian school setting. **International journal of environmental research and public health**, v. 15, n. 6, p. 1127, 2018.

PRANCE, G. T. The pejibaye, Guilielma gasipaes (HBK) Bailey, and the papaya, Carica papaya L. In: Stone D. (ed.), Pre-Columbian Plant Migration. **Papers of the Peabody Museum of Archaeology and Ethnology**, vol. 76, Harvard University Press, Cambridge, p. 85–104, 1984.

QUEROL, D. **Recursos genéticos, nosso tesouro esquecido**. Tradução Joselita Wasniewski. Rio de Janeiro: ASPTA, p. 206, 1993.

R Development Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2017.

REIS, O. F.; CAMPOSTRINI, E.; SOUSA, E. F.; SILVA, M. G. Sap flow in papaya plants: laboratory calibrations and relationships with gas exchanges under field conditions. **Scientia horticulturae**, v. 110, n. 3, p. 254-259, 2006.

REIS, R. C.; VIANA, E. S.; JESUS, J. L.; LIMA, L. F.; NEVES, T. T.; CONCEIÇÃO, E. A. Compostos bioativos e atividade antioxidante de variedades melhoradas de mamão. **Ciência Rural**, Santa Maria, v. 45, n. 11, p. 2076-2081, 2015.

ROTH, I. Fruits of angiosperms. Handbuch der Pflanzenanatomie X, 1. **Berlin: Borntraeger**, (Handbuch der Pfl anzenanatomie spezieller Teil, Band x, Teil 1). Anatomy and Morphology (KR, 197706808). p. 675, 1977.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, p. 53-65, 1987.

SALGOTRA, R. K.; GUPTA, B. B. **Plant Genetic Resources and Traditional Knowledge for Food Security**. Springer Science+Business Media, Singapore, 2015.

SCHELDEMAN, V.; KYNDT, T.; COPPENS D'EECKENBRUGGE G.; MING. R.; DREW. R.; VAN DROOGENBROECK. B.; VAN DAMME P.; MOORE P.H. Vasconcellea. In: Kole C (ed) **Wild crop relatives: genomic and breeding resources**. Springer Science+Business Media, New York, p. 213–249, 2011.

SCHULENBERG, S. E.; MELTON, A. M. A confirmatory factor-analytic evaluation of the purpose in life test: Preliminary psychometric support for a replicable two-factor model. **Journal of Happiness Studies**, v. 11, n. 1, p. 95-111, 2010.

SCHWEIGGERT, R. M.; STEINGASS, C. B.; HELLER, A.; ESQUIVEL, P.; CARLE, R. Characterization of chromoplasts and carotenoids of red-and yellow-fleshed papaya (Carica papaya L.). **Planta**, v. 234, n. 5, p. 1031, 2011.

SINGH, D. The relative importance of characters affecting genetic divergence. **The Indian Journal of Genetic and Plant Breeding**, New Delhi, v. 41, p. 237-245, 1981.

TEIXEIRA, S.J.A.; RASHID Z, NHUT D.T.; SIVAKUMAR, D.; GERA, A.; SOUZA, M.T JR.; TENNANT, P.F. Papaya (*Carica papaya* L.) biology and biotechnology. **Tree Forest Sci Biotechnol**. v. 1, n. 1, p. 47–73, 2007.

SNEATH, P. H. A.; SOKAL, R. R. The comparison of dendrograms by objective methods. **Taxon**, vol. 11, p. 33-40, 1973.

SOKAL, R.R.; MICHENER, C.D. A statistical method for evaluating systematic relationships, *Univ. Kansas. Sci. Bull. V.*38, p**.** 1409-1438, 1958.

SOKAL, R.R.; ROHLF, F.J. The comparison of dendrograms by objective methods. **Taxon**, Berlin, v.11, p.30-40, 1962.

TAMHANE, A. C. & DUNLOP D. D*.* **Statistics and Data Analysis** – from elementary to, intermediate. Upper Saddle River: Prentice-Hall, 2000.

VALLS, J. F. M. Caracterização de recursos genéticos vegetais. In: NASS, L.L. (ed.). **Recursos genéticos vegetais**. Brasília, Embrapa Recursos Genéticos e Biotecnologia, p. 281-305, 2007.

VAN DER EIJK, C.; ROSE, J. Risky business: factor analysis of survey data–assessing the probability of incorrect dimensionalisation. **PloS one**, v. 10, n. 3, p. e0118900, 2015.

VASSEND, O.; SKRONDAL, A. Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled?. **European Journal of Personality**, v. 11, n. 2, p. 147-166, 1997.

VAVILOV, N. I. **Origin and geography of cultivated plants**. Cambridge University Press, 1987.

VELICER, W. F.; FAVA, J. L. Affects of variable and subject sampling on factor pattern recovery. **Psychological methods**, v. 3, n. 2, p. 231, 1998.

VELMURUGAN, T.; SANTHANAM, T. A survey of partition based clustering algorithms in data mining: An experimental approach. **Information Technology Journal**, v. 10, n. 3, p. 478-484, 2011.

WANG, C. N.; WENG, L. J. Evaluating the use of exploratory factor analysis in Taiwan: 1993-1999. **Chinese Journal of Psychology**, v. 44, n. 2, p. 239-252, 2002.

WEEDEN, N.F.; TIMMERMAN, G.M.; LU, J. Identifying mapping genes of economic significance. **Euphytica**, Wageningen, v. 73, p. 191-198, 1994.

WOOD, P.; KARDASH, C. Critical elements in the design and analysis of studies of epistemology. **Personal epistemology: The psychology of beliefs about knowledge and knowing**, p. 231-260, 2002.

YEUNG, K. Y.; HAYNOR, D. R.; RUZZO, W. L. Validating clustering for gene expression data. **Bioinformatics**, v. 17, n. 4, p. 309-318, 2001.

ZWICK, W. R.; VELICER, W. F. Comparison of five rules for determining the number of components to retain. **Psychological bulletin**, v. 99, n. 3, p. 432, 1986.

ARTIGO 1

NOVA ESTRATÉGIA DE SELEÇÃO DE DESCRITORES QUANTITATIVOS PARA CARACTERIZAÇÃO DE ACESSOS DE MAMÃO (*Carica papaya* L.)[1]

[1]Artigo a ser ajustado para posterior submissão ao Comitê Editorial do periódico científico Pesquisa Agropecuária Brasileira, em versão na língua inglesa.

NOVA ESTRATÉGIA DE SELEÇÃO DE DESCRITORES QUANTITATIVOS PARA CARACTERIZAÇÃO DE ACESSOS DE MAMÃO (*Carica papaya* L.)

Autor: Antonio Leandro da Silva Conceição
Orientador: Dr. Carlos Alberto da Silva Ledo

RESUMO: O objetivo deste estudo foi selecionar descritores quantitativos importantes para caracterizar e avaliar acessos representativos de mamão do banco de germoplasma da Embrapa Mandioca e Fruticultura. Foram utilizados 35 descritores na caracterização de 50 acessos. Foi aplicada uma nova estratégia de seleção visando a escolha dos descritores mais importantes, que não ocasionem problemas de multicolinearidade e representem a variação total existente no conjunto de individuos em estudo. Para isso, utilizou-se o diagnostico de multicolinearidade baseado no número de condição proposto por Montgomery e Peck, combinado a análise de componentes principais proposta por Jolliffe e a contribuição relativa de cada descritor para a divergência genética proposta por Singh. As técnicas propostas por Jolliffe e Singh foram usadas na decisão final de descarte para cada par de descritores mais auto correlacionados, indicadas durante o diagnóstico com base no número de condição. Para validação dessa estratégia foi utilizado o fator de inflação das variâncias proposto por Berk. Com base nos resultados, conclui-se que 31,43% dos descritores analisados foram relativamente redundantes. Já os descritores mais informativos foram: comprimento dos internódios 8 meses; comprimento dos internódios: 12 meses; Largura da folha: 8 meses; comprimento do pecíolo da folha: 8 meses; comprimento do pecíolo da folha: 12 meses; nº de frutos: 8 meses; nº de frutos carpelóides: 8 meses; nº de frutos: 12meses; nº de frutos carpelóides: 12 meses; nº de frutos por axila; altura dos primeiros frutos; comprimento do pedúnculo do fruto; nº de flores por pedúnculo; comprimento do pedúnculo da inflorescência; comprimento da corola da flor hermafrodita; comprimento do fruto; firmeza dos frutos; diâmetro da cavidade central; peso fresco de sementes do fruto; peso fresco de 100 sementes; acidez; vitamina C; pH e sólidos solúveis totais.

Palavras-chave: Análise multivariada, recursos genéticos, melhoramento

NEW STRATEGY FOR SELECTION OF QUANTITATIVE DESCRIPTORS FOR CHARACTERIZATION OF PAPAYA (*Carica papaya* L.)

Author: Antonio Leandro da Silva Conceição
Advisor: Dr. Carlos Alberto da Silva Ledo

ABSTRACT: The objective of this study was to select important quantitative descriptors to characterize and evaluate representative accessions of papaya from the germplasm bank of Embrapa Mandioca and Fruticultura. For this, 35 descriptors were used to characterize 50 accessions. A new selection strategy was applied aiming at choosing the most important descriptors that do not cause multicollinearity problems and represent the total variation existing in the set of individuals under study. For this, the multicollinearity diagnosis based on the condition number proposed by Montgomery and Peck was used, combined the principal component analysis proposed by Jolliffe and with a relative contribution of each descriptor to the genetic divergence proposed by and the relative contribution of each descriptor to the genetic divergence proposed by Singh. The techniques proposed by Jolliffe and Singh were used in the final disposal decision for each pair of correlating descriptors, indicated during diagnosis based on the condition number. To validate this strategy, the variances inflation factor of proposed by Berk was used. Based on the results, it is concluded that 31.43% of the analyzed descriptors were relatively redundant. The most informative descriptors were: length of internodes 8 months; length of internodes: 12 months; Leaf Width: 8 months; leaf petiole length: 8 months; leaf petiole length: 12 months; number of fruits: 8 months; number of carpeloid fruits: 8 months; number of fruits: 12 months; number of carpeloid fruits: 12 months; number of fruits per axilla; height of the first fruits; peduncle length of the fruit; number of flowers per peduncle; inflorescence peduncle length; corolla length of the hermaphrodite flower; fruit length; firmness of the fruits; diameter of central cavity; fresh weight of fruit seeds; fresh weight of 100 seeds; acidity; Vitamin C; pH and total soluble solids.

Key words: Multivariate analysis, genetic resources, improvement

INTRODUÇÃO

O Brasil é o segundo maior produtor de mamão (Carica *papaya* L.) do mundo, superado apenas pela Índia (FAO, 2017). O país também é um dos maiores exportadores da cultura, sua produção está voltada para o mercado interno de frutas frescas, bem como os mercados de exportação de frutas frescas e processamento industrial (CARDOSO, et al., 2017).

As variedades de mamoeiro mais cultivadas pertencem aos grupos Solo e Formosa. Contudo, os sistemas de produção são baseados no cultivo de um número reduzido de variedades, o que resulta em restrita variabilidade genética. Esta prática pode levar à maior vulnerabilidade a doenças, pragas e variações edafoclimáticas, e comprometer a sustentabilidade desse cultivo. Assim, a busca pelo aumento da variabilidade genética, por meio do desenvolvimento de novos genótipos, pode garantir maior competitividade e sustentabilidade à cultura do mamoeiro (DIAS et al., 2012).

O conhecimento da variabilidade genética, disponível no germoplasma da espécie, é um pré-requisito para a indicação de potenciais genitores, para se combinarem alelos relacionados a características de importância econômica, no direcionamento dos cruzamentos em programas de melhoramento (DIAS, et al., 2012).

A tentativa de verificar o grau de variabilidade presente em um conjunto de indivíduos usando todas as informações disponíveis, ou seja, todos os descritores, parece ser a melhor alternativa. Pois se investe muito tempo e dinheiro na coleta de dados e informações. Porém, nem todos os descritores possuem a capacidade de contribuir para discriminar o conjunto de indivíduos em estudo, ou algumas podem apresentar problemas de colinearidade resultando em distorções nos padrões de agrupamento.

Quando o número de descritores é elevado, muitos deles podem contribuir pouco para a discriminação dos indivíduos avaliados. Essa situação aumenta o trabalho de caracterização, mas não melhora a precisão, além de tornar mais complexa a análise e interpretação dos dados (PAIVA et al., 2010).

Análises multivariadas são instrumentos úteis na identificação de descritores com maior conteúdo informativo na caracterização de germoplasma e

melhoramento genético, além de fornecer indicação para eliminar os descritores que pouco contribuem na variação total disponível (CRUZ et al., 2004).

A redução no número de descritores morfoagronômicos é relatada em vários estudos, os quais utilizaram diversas técnicas de descarte, com o objetivo de otimizar o trabalho de coleta dos dados e identificar os descritores com maior contribuição na divergência genética. Para DIAS (2011), estudando 27 genótipos de mamão, por meio de 51 descritores morfoagronômicos, foi definida que a lista de descritores mínimos do mamoeiro para fins de proteção de variedades e classificação dos genótipos foi constituída por 18 descritores quantitativos e 13 multicategóricos. Oliveira et al., (2014), analisando acessos de mandioca (*Manihot esculenta*) utilizando-se de 51 descritores morfoagronômicos, constataram que 32 deles foram suficientes em razão de sua alta capacidade para discriminar o germoplasma de mandioca e de sua habilidade de manter alguns descritores agronômicos preliminares, úteis para a caracterização inicial do germoplasma. Silva et al. (2017), para caracterizar 262 acessos de mandioca representativos da Amazônia oriental, utilizaram 21 descritores, sendo verificado que apenas 13 desses foram capazes de discriminar os acessos e representar a variabilidade morfológica.

É importante despertar o interesse por pesquisas que expõem a combinação de métodos, uma vez que, podem resultar num aperfeiçoamento dos resultados de uma análise. Ainda que cada técnica possua suas particularidades e objetivos específicos de pesquisa, o ajuste de duas ou mais técnicas podem proceder a invenção de uma nova técnica (ALVES, 2007).

O objetivo deste trabalho foi reduzir a dimensão do conjunto original de descritores com menor perda de informação possível, eliminando as informações redundantes em decorrência da correlação entre descritores, e eliminar as que contribuem pouco para explicar a variação total e que não ocasione distorções nos padrões de agrupamentos em análises posteriores.

MATERIAIS E MÉTODOS

Foram utilizados 50 acessos pertencentes ao banco de germoplasma de mamão (BAG-Mamão) da Embrapa Mandioca e Fruticultura (Tabela 1). O plantio

dos acessos foi realizado do dia 26 a 29 de agosto de 2014. As avaliações foram realizadas de outubro de 2014 a dezembro de 2015. Foi utilizado espaçamento de 3,0 m entre linhas e 2,0 m entre plantas, adotando-se as práticas culturais e os tratos fitossanitários preconizados para a cultura (Martins & Costa, 2003).

Tabela 1. Classificação por tipo e origem dos acessos de mamão avaliados, que compõem o Banco de Germoplasma (BAG-Mamão) da Embrapa Mandioca e Fruticultura. Cruz das Almas-BA, 2019.

Acesso	Tipo de fruto	Origem	Sigla
BGM 01	Formosa	Costa Rica	BGM 01 FC
BGM 02	Formosa	Taiwan	BGM 02 FT
BGM 03	Formosa	Havaí	BGM 03 FH
BGM 04	Solo	Havaí	BGM 04 SH
BGM 05	Solo	Havaí / Taiwan	BGM 05 SHT
BGM 06	Formosa	Malásia	BGM 06 FM
BGM 07	Formosa	Costa Rica	BGM 07 FC
BGM 08	Solo	Malásia	BGM 08 SM
BGM 09	Formosa	Malásia	BGM 09 FM
BGM 10	Formosa	Taiwan	BGM 10 FT
BGM 11	Formosa	Taiwan	BGM 11 FT
BGM 12	Formosa	Brasil	BGM 12 FB
BGM 13	Solo	Taiwan	BGM 13 ST
BGM 14	Formosa	Malásia	BGM 14 FM
BGM 15	Formosa	Taiwan	BGM 15 FT
BGM 16	Formosa	*	BGM 16 F
BGM 17	Formosa	Costa Rica	BGM 17 FC
BGM 18	Formosa	*	BGM 18 F
BGM 19	Formosa	Costa Rica	BGM 19 FC
BGM 20	Formosa	*	BGM 20 F
BGM 21	Solo	*	BGM 21 S
BGM 22	Formosa	Brasil	BGM 22 FB
BGM 23	Formosa	Brasil	BGM 23 FB
BGM 24	Formosa	Brasil	BGM 24 FB
BGM 25	Formosa	Brasil	BGM 25 FB
BGM 26	Formosa	Brasil	BGM 26 FB
BGM 27	Solo	Brasil	BGM 27 SB
BGM 28	Solo	Brasil	BGM 28 SB
BGM 29	Solo	Brasil	BGM 29 SB
BGM 30	Formosa	Havaí	BGM 30 FH
BGM 31	Formosa	Brasil	BGM 31 FB
BGM 32	Solo	Brasil	BGM 32 SB
BGM 33	Solo	Havaí	BGM 33 SH
BGM 34	Formosa	Havaí	BGM 34 FH
BGM 35	Solo	Brasil	BGM 35 SB
BGM 36	Formosa	Brasil	BGM 36 FB
BGM 37	Formosa	Brasil	BGM 37 FB
BGM 38	Solo	*	BGM 38 S
BGM 39	Solo	Brasil	BGM 39 SB
BGM 40	Formosa	*	BGM 40 F

Tabela 1. (Continuação)

Acesso	Tipo de fruto	Origem	Sigla
BGM 41	Formosa	*	BGM 41 F
BGM 42	Solo	Havaí	BGM 42 SH
BGM 43	Solo	Havaí	BGM 43 SH
BGM 44	Solo	Havaí	BGM 44 SH
BGM 45	Formosa	África do Sul	BGM 45 FA
BGM 46	Formosa	África do Sul	BGM 46 FA
BGM 47	Solo	África do Sul	BGM 47 SA
BGM 48	Formosa	Brasil	BGM 48 FB
BGM 49	Formosa	*	BGM 49 F
BGM 50	Formosa	*	BGM 50 F

As avaliações foram realizadas em Cruz das Almas, Bahia, Brasil (12º48'38"S e 39º6'26"W), na área experimental da Embrapa Mandioca e Fruticultura. Os dados climatológicos do município, obtidos pelo Instituto Nacional de Meteorologia (INMET), (INMET,2017), durante o período de condução do experimento estão representados na Figura 1.

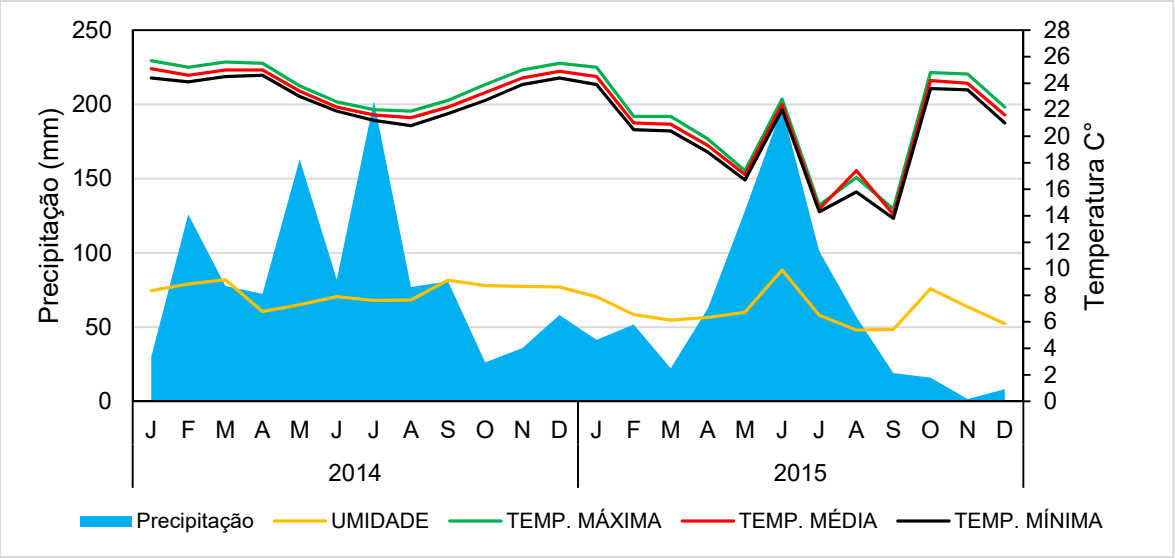


Figura 1. Temperatura máxima, média e mínima; umidade e precipitação acumulada, para os meses do ano de 2014 e 2015 no município de Cruz das Almas-Ba. INMET/2017.

A avaliação foi realizada baseada no Manual de Descritores para Mamão [Catálogo de Germoplasma de Mamão (*Carica papaya* L.)], adaptado pela Embrapa Mandioca e Fruticultura a partir dos descritores inicialmente estabelecidos pelo

International Board for Plant Genetic Resources (IBPGR, 1988), atualmente Bioversity International, com algumas alterações sugeridas por Dantas et al. (2000).

Para o presente estudo foram utilizados 35 descritores quantitativos: Comprimento (cm) médio dos internódios: 8 meses (CI8); comprimento (cm) médio dos internódios: 12 meses (CI12); Comprimento (cm) da folha: 8 meses (CF8); comprimento (cm) da folha: 12 meses (CF12); Largura (cm) da folha: 8 meses (LF8); Largura (cm) da folha: 12 meses (LF12); comprimento (cm) do pecíolo da folha: 8 meses (CPF8); comprimento do pecíolo da folha: 12 meses (CPF12); altura (m) da planta (8 meses) (ALT8); altura (m) da planta: 12 meses (ALT12); diâmetro (cm) do caule: 8 meses (DC8); diâmetro (cm) do caule: 12 meses (DC12); número de frutos: 8 meses (NF8); número de frutos comerciais: 8 meses (NFCo8); número de frutos carpelóides: 8 meses (NFCa8); número de frutos: 12 meses (NF12); número de frutos comerciais: 12 meses (NFCo12); número de frutos carpelóides: 12 meses (NFCa12); número de frutos por axila (NFaxi); altura (cm) dos primeiros frutos (ALT1F); Comprimento (cm) do pedúnculo do fruto (CPF); peso (kg) do fruto (PF); comprimento (cm) do fruto (CF); diâmetro (cm) do fruto (DF) ;firmeza dos frutos (MFF) medida em libras, com uso de penetrômetro manual modelo FT 327 (McCormick Fruit Tech, Yakima, WA, EUA); diâmetro (cm) da cavidade central (DCC); número de flores por pedúnculo (NFP); comprimento (cm) do pedúnculo da inflorescência (CPI); comprimento (cm) da corola da flor hermafrodita (CCFher); peso (g) fresco de sementes do fruto (PFS); peso (g) fresco de 100 sementes (PS); Acidez (AC), acidez total titulável, expressa em gramas de ácido cítrico por 100 g de suco; vitamina C (VITC); pH (PH) e Sólidos solúveis totais, medidos em ºBrix (BRIX).

As medições de descritores relacionados a flores e frutos foram realizadas aos 12 meses após o plantio. Foram avaliadas três folhas por planta. Na análise, foram colhidos, aleatoriamente, oito frutos oriundos de plantas hermafroditas, no estágio 2 de amadurecimento, quando os frutos apresentavam até 25% da superfície amarela. Já as análises físico-químicas foram realizadas quando os frutos atingiram o estágio 5, quando os frutos apresentavam 100% da superfície amarela.

A nova estratégia proposta e abordada nesse estudo tem como base a análise multivariada de dados, aplicada a seleção de descritores, onde foi utilizado o

diagnóstico de multicolinearidade combinado com analises de componentes principais e o critério de Singh (1981), com o fator de inflação da variância das variáveis (VIF), utilizado para validação dos resultados.

O método utilizado para diagnóstico foi o exame do número de condição (NC), proposto por Montgomery e Peck (1981), (Tabela 2). A integração dessa técnica com a análise de componentes principais e a contribuição relativa para divergência proposta por Singh (1981), tem como objetivo a escolha de quais descritores serão excluídos a cada passo do diagnóstico de multicolinearidade pelo (NC), até que está se enquadre no grau de colinearidade fraca, a qual não apresentará problemas de multicolinearidade.

Tabela 2. Tabela de Classificação de Montgomery e Peck (1981).

Número de Condição (NC)	Multicolinearidade
NC < 100	Fraca (Não constitui problema sério)
100 < NC < 1000	Moderada a forte
> 1000	Severa

Número de condição (NC)

A escolha dos descritores a serem descartados durante diagnóstico de multicolinearidade pelo (NC) foi realizada por dois procedimentos, detalhados abaixo:

i) Método proposto por Jolliffe (1972, 1973), sendo indicado para descarte todo descritor que apresentou maior coeficiente de ponderação em valor absoluto (autovetor), no componente principal de autovalor menor, partindo-se do último descritor até aquele que seu autovalor não excedeu 0,70. O processo de descarte consiste em considerar o autovetor (coeficientes do CP) correspondente ao menor autovalor e rejeitar o descritor associado ao maior coeficiente (valor absoluto). Então, o próximo menor autovetor é avaliado. Esse processo continua até que o autovetor associado ao autovalor inferior a 0,7 seja considerado. A razão para isso é que descritores altamente correlacionados aos componentes principais de menor variância representam variação praticamente insignificante (MARDIA et al., 1997).

ii) Método baseado no critério proposto por Singh (1981), levando-se em consideração a contribuição relativa dos descritores para divergência genética, baseado na estatística S.j. Considerou-se que:

$$D^2_{ii'} = \delta' \psi^{-1} \delta = \sum_{j=1}^{n} \sum_{j'=1}^{n} \omega_{jj} d_j d_{j'}$$

em que ω_{jj} é o elemento da j-esima coluna da inversa da matriz de variância e covariância residuais. O total das distâncias envolvendo todos os pares de genótipos é dado por: $SSD^2_{ii'} = S\,D^2_m = SS_{\cdot j}$. Os valores percentuais de S.j constituíram a medida da importância relativa da variável j para o estudo da diversidade genética.

Para validação dessa nova estratégia de seleção de descritores, foi utilizado o fator de inflação da variância das variáveis (VIF) (BERK, 1977), onde esta validação foi realizada junto ao processo de descarte, para detectar quais apresentavam maior VIF e se essas estavam dentro do grupo de descritores descartados, o VIF também foi realizado após o descarte para avaliar a precisão e consistência do método de seleção proposto, sendo avaliada nessa segunda etapa apenas os descritores selecionados, com o intuito de verificar se os descritores remanescentes apresentavam o VIF abaixo de 10, o que é considerado ideal. O VIF é dado por:

$$VIF_i = \frac{1}{1-R^2_i}$$

onde R^2_i é o coeficiente de determinação múltipla. É uma medida do grau em que cada variável independente é explicada pelas demais variáveis independentes. Quanto maior for o fator de inflação da variância, mais severa será a multicolinearidade. Valores maiores que 10 correspondem a um coeficiente de determinação R^2_i > 0,90 e isso é considerado inaceitável e motivo de preocupação. (MARQUARDT,1970; MASON et al., 1989; NETER et al.,1989; KENNEDY, 1992; TAMHANE e DUNLOP, 2000; KUTNER et al., 2004). Outros autores, como HAIR (2005), sugerem que os fatores de inflação da variância não devem exceder 4 ou 5, isso dependerá do conhecimento teórico do pesquisador sobre o problema estudado.

Inicialmente foi testada a normalidade dos dados pelo teste de Shapiro–Wilk, e em seguida, para auxiliar na decisão de descarte, foram estimados os coeficientes de correlação de Spearman entre todos os descritores, para verificar a associação entre os descritores descartados e os remanescentes. A significância do coeficiente de correlação foi verificada pelo teste de t. Essas análises foram realizadas com o auxílio do programa estatístico R (R CORE TEAM, 2017).

As análises de componentes principais, multicolinearidade e Fator de Inflação das Variâncias (VIF), também foram realizadas pelo programa estatístico R (R CORE TEAM, 2017), com auxílio dos pacotes FactoMineR, Faraway e Agricolae. A análise de contribuição relativa de cada descritor para divergência genética pelo método proposto por Singh (1981), foi obtida pelo programa Genes (CRUZ, 2013).

RESULTADOS E DISCUSSÃO

Na Tabela 3 estão apresentadas as estatísticas descritivas dos descritores quantitativos. Nesta, pode-se observar a amplitude dos valores apresentados para os descritores estudados. Observou-se que os coeficientes de variação (CV) oscilaram de 2,68 a 270,36, para o pH (PH) e número de frutos carpelóides: 12 meses (NFCa12) respectivamente. O número de frutos carpelóides aos 12 meses variou de 0 a 4, valores próximos aos encontrados por Dantas et al. (2015), avaliando linhagens e híbridos de mamoeiro, onde foi observada uma variação de 0,92 a 4,69 para número de frutos carpeloides (deformados).

Foram observados altos valores para o CV (%) em alguns descritores avaliados (Tabela 3). Altos valores de CV, também foram observados nos estudos de Silva et al. 2007; Oliveira et al. (2010) e Dantas et al. (2015). Segundo Dantas et al. (2015), altos valores obtidos devem-se ao fato de que os descritores avaliados são de natureza poligênica e bastante influenciadas pelo ambiente.

Dentre os valores mínimos e máximos observados, o diâmetro do caule: aos 8 e 12 meses (DC8 e DC12), respectivamente e peso (kg) do fruto (PF) apresentaram grande variação, sendo esses dois descritores muito importantes, pois segundo Fraife Filho et al. (2001) e Silva et al. (2007) sugerem que a seleção de plantas de mamoeiro com maior diâmetro do caule pode resultar em plantas mais produtivas, em virtude da alta correlação genética entre esses descritores.

Os sólidos solúveis totais apresentaram uma variação de 7,90 a 16,12. Na literatura foram relatadas diferentes variações nos teores de sólidos solúveis, em frutos de mamoeiro, como por Viegas, (1992) (12-13 ºBrix); Fioravanço et al. (1992) (8,68-11,66 ºBrix); Fagundes e Yamanishi (2001) (9,9-12,5 ºBrix); Santana et al. (2004) (9-14 ºBrix); Ocampo et al. (2006) (4,6-13,3 ºBrix); Silva et al. (2008) (10,24-12,27 ºBrix); Brito Neto et al. 2011 (7-14 ºBrix); Dias et al. (2011) (7,25 a 11,53 ºBrix); Schweiggert et al. (2012) (7,9-13,6 ºBrix); Reis et al. (2015) (9,58-14,76 °Brix); Viana et al. (2015) (11,15-15,10 ºBrix); Dantas et al. (2015) (12,6 a 13,3 ºBrix); e Barros et al. (2017) (12,10-15,05 ºBrix). Oliveira et al. (2010) encontraram variação de 5,0 a 16,2 no °Brix, esses resultados são semelhantes aos encontrados no presente estudo (Tabela 3).

O teor de Vitamina C (VITC) variou de 47,90 a 135,48 mg.100g^{-1}. Silva et al. (2018) estudaram a caracterização de frutos de mamão do grupo Formosa e Solo, encontrando valores de VITC entre 102 e 137 mg.100g^{-1}. Reis et al. (2015), estudando variedade dos mesmos grupos, encontraram valores entre 91,47 e 115,43 mg.100g^{-1}. Já Zaman et al. (2006), obtiveram um valor médio para VITC em quatro variedades de mamão vermelho e amarelo de 41,8 mg.100g^{-1}. Nesses estudos pode ser observado valores bastante variáveis. Isso demonstra grande variabilidade apresentada para esse descritor e seu potencial para trabalhos de melhoramento da cultura.

Com base no teste de normalidade de Shapiro-Wilks (W) (Tabela 3), observou-se que a maioria dos descritores não seguiu distribuição normal. Exceto os descritores comprimento (cm) médio dos internódios: 8 meses (CI8); comprimento (cm) da folha: 8 meses (CF8); comprimento (cm) da folha: 12 meses (CF12); comprimento (cm) do pecíolo da folha: 8 meses (CPF8); comprimento do pecíolo da folha: 12 meses (CPF12); diâmetro (cm) do caule: 8 meses (DC8); peso (kg) do fruto (PF); comprimento (cm) do fruto (CF); pH (PH) e Sólidos solúveis totais (BRIX) quem apresentaram distribuição normal. Portanto calculou-se a correlação de Spearman para medir a relação entre os descritores e para auxiliar na decisão de descarte, verificando a associação entre descritores descartados e os remanescentes.

Tabela 3. Estatística descritiva e teste de normalidade dos descritores quantitativos de acessos de mamão estudadas. Cruz das Almas, BA. 2019.

Descritores	Mínimo	Máximo	Média	Desvio Padrão	CV (%)	W
AC	0,05	0,44	0,09	0,05	61,37	0,40**
ALT1F	0,40	2,10	0,95	0,31	32,29	0,93**
ALT8	0,80	2,40	1,48	0,37	24,77	0,97^{ns}
ALT12	1,15	3,10	1,88	0,42	22,46	0,94**
BRIX	7,90	16,12	12,91	1,60	12,38	0,97^{ns}
CCFher	2,50	5,00	3,66	0,57	15,57	0,99^{ns}
CF	9,15	26,25	17,80	3,61	20,3	0,98^{ns}
CF8	27,68	55,90	40,11	5,92	14,75	0,97^{ns}
CF12	20,60	43,80	31,45	4,70	14,96	0,98^{ns}
CI8	8,70	20,94	13,45	2,91	21,63	0,97^{ns}
CI12	9,00	21,90	13,58	2,96	21,77	0,92**
CPF	2,32	10,35	3,87	1,38	35,65	0,76**
CPF8	38,40	89,60	62,02	9,63	15,53	0,99^{ns}
CPF12	24,40	71,10	49,17	9,59	19,51	0,99^{ns}
CPI	0,64	7,42	2,21	1,26	57,22	0,80**
DC8	5,40	13,20	8,27	1,70	20,51	0,97^{ns}
DC12	6,60	17,50	10,21	2,08	20,39	0,89**
DCC	23,05	81,46	42,31	10,71	25,3	0,95**
DF	5,27	14,70	8,96	1,77	19,71	0,98^{ns}
LF8	42,96	86,60	62,02	8,57	13,82	0,98^{ns}
LF12	29,00	69,80	48,25	7,51	15,56	0,99^{ns}
MFF	0,67	4,40	1,79	0,84	46,96	0,92**
NF8	1,00	49,00	12,72	11,77	92,56	0,83**
NF12	1,00	43,00	9,86	9,40	95,29	0,81**
NFaxi	1,00	3,00	1,40	0,67	47,86	0,62**
NFCa8	0,00	11,00	0,74	1,80	243,9	0,46**
NFCa12	0,00	4,00	0,28	0,76	270,36	0,43**
NFCo8	0,00	48,00	9,82	9,61	97,88	0,82**
NFCo12	0,00	41,00	7,64	8,71	114,07	0,73**
NFP	2,60	17,60	6,59	3,04	46,16	0,89**
PS	5,70	14,65	10,94	1,98	18,08	0,98^{ns}
PF	0,16	1,71	0,73	0,33	44,76	0,96^{ns}
PFS	5,64	168,88	62,56	31,61	50,54	0,89**
PH	4,78	5,62	5,25	0,14	2,68	0,96^{ns}
VITC	47,90	135,48	81,36	20,82	25,59	0,94**

Comprimento médio dos internódios: 8 meses (CI8); comprimento médio dos internódios: 12 meses (CI12); Comprimento da folha: 8 meses (CF8); comprimento da folha: 12 meses (CF12); Largura da folha: 8 meses (LF8); Largura da folha: 12 meses (LF12); comprimento do pecíolo da folha: 8 meses (CPF8); comprimento do pecíolo da folha: 12 meses (CPF12); altura (m) da planta (8 meses) (ALT8); altura (m) da planta: 12 meses (ALT12); diâmetro do caule: 8 meses (DC8); diâmetro do caule: 12 meses (DC12); número de frutos: 8 meses (NF8); número de frutos comerciais: 8 meses (NFCo8); número de frutos carpelóides: 8 meses (NFCa8); número de frutos: 12 meses (NF12); número de frutos comerciais: 12 meses (NFCo12); número de frutos carpelóides: 12 meses (NFCa12); número de frutos por axila (NFaxi); altura dos primeiros frutos (ALT1F); Comprimento do pedúnculo do fruto (CPF); número de flores por pedúnculo (NFP); comprimento do pedúnculo da inflorescência (CPI); comprimento da corola da flor hermafrodita (CCFher); peso (kg) do fruto (PF); comprimento do fruto (CF); diâmetro do fruto (DF) ;firmeza dos frutos (MFF); diâmetro da cavidade central (DCC); peso (kg) fresco de sementes do fruto (PFS); peso (g) fresco de 100 sementes (PS); Acidez (AC); vitamina C (VITC); pH (PH) e Sólidos solúveis totais, medidos em ºBrix (BRIX).

Na Figura 2 e na Tabela 4 são apresentadas as estimativas dos autovalores associados aos componentes principais e suas respectivas variâncias relativas e acumuladas obtidas para os 35 descritores morfológicos quantitativos, percebe-se que os dois primeiros componentes conseguiram explicar 41,98%. O primeiro componente explicou 26,92%, o segundo 15,06% (Figura 2). Onde a variância acumulada foi concentrada até o 13º componente principal, que respondeu por 89,98% de toda a variação relativa observada (Figura 2) e (Tabela 4).

Resultados semelhantes foram obtidos por Oliveira et al. (2012), no estudo da seleção de descritores em cultivares de mamão, onde a variação total nos dois primeiros componentes foi de 52,09%. Percentuais próximos que variaram aproximadamente de 24 a 45% na soma dos dois primeiros componentes, foram encontrados nos estudos com mamão e de outras culturas (Sugimura et al., 1997; Strapasson et al., 2000; Oliveira et al., 2006; Asudi et al., 2010; Aikpokpodion, (2012); Afonso et al., 2014). A distribuição da variância é mais concentrada nos dois primeiros componentes principais quando poucos descritores são avaliados ou quando pertencem a partes específicas da planta, ou seja, quando são avaliados por exemplo, apenas descritores relacionados aos frutos (Pereira et al., 1992).

O primeiro descritor que foi sugerido a ser descartado na seleção direta foi o DC12, que apresentou o coeficiente ponderado mais alto em módulo com o último componente principal (0,43), seguido pelo DC8, NFCo8, CPF8 e ALT12, cujos respectivos autovetores em módulo ocorreram nos componentes principais 34, 33, 32 e 31, respectivamente (Tabela 4).

Utilizando este procedimento, 21 descritores (60%) dos 35 avaliados foram considerados redundantes e foram indicados para serem descartados na seguinte ordem: DC12 (**1D**); DC8 (**2D**); NFCo8 (**3D**); CPF8 (**4D**); ALT12 (**5D**); CF8 (**6D**); ALT8 (**7D**); PF (**8D**); NFCo12 (**9D**); PFS (**10D**); NFaxi (**11D**); ALT1F (**12D**); NFP (**13D**); CPI (**14D**); CF12 (**15D**); CI12 (**16D**); NF8 (**17D**); PS (**18D**); DF (**19D**); LF12 (**20D**) e CF (**21D**) (Tabela 4). Resultados semelhantes foram obtidos por Oliveira et al. (2012), em que foi relatada, também, uma redução de 60% dos descritores.

Foi observado que a eliminação de descritores com base na seleção direta apresentou um rigor bastante elevado, considerando que descritores relacionados aos frutos foram quase todos indicados para o descarte, como o peso do fruto (PF), a altura dos primeiros frutos (ALT1F) e diâmetro do fruto (DF). De acordo com

Oliveira et al. (2012), a maioria desses descritores são extremamente importantes na caracterização de uma variedade de mamão e também permite a classificação e o uso final dos frutos para diferentes tipos de mercados. Pelo fato de alguns trabalhos criticarem o emprego da análise de componentes principais no descarte de descritores, especialmente quando se utiliza o método da seleção direta, considerando um procedimento drástico, faz-se necessária a avaliação de sua eficiência (ALVES, 2002), e a busca de novas alternativas.

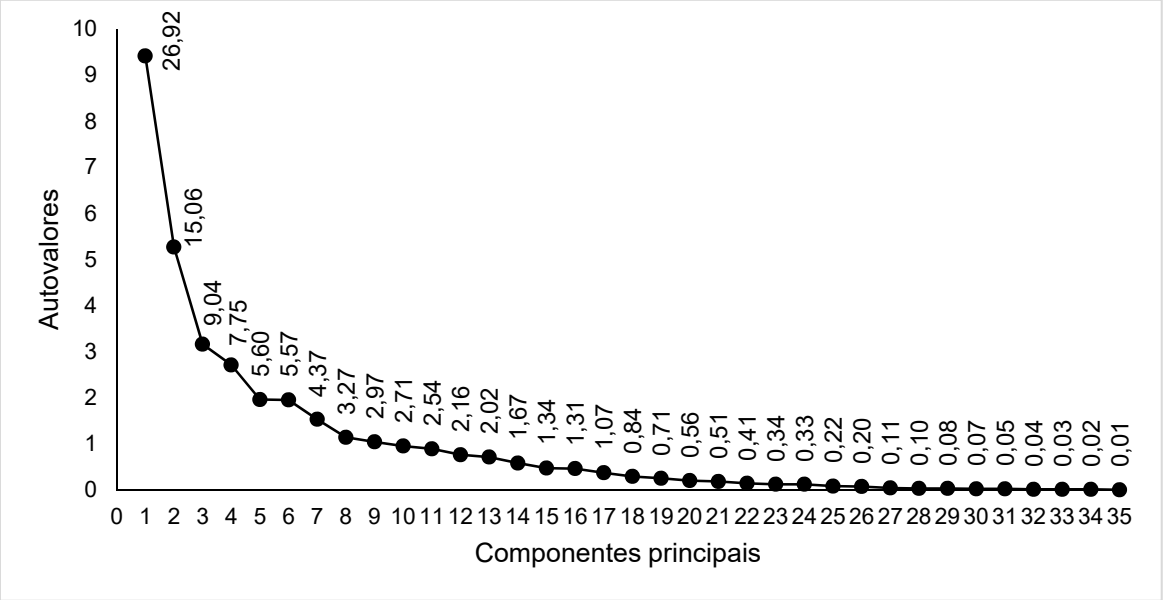


Figura 2. Autovalores (eixo y) e porcentagem da variação original armazenada em cada um dos 35 componentes principais (eixo dos x).

Tabela 4. Estimativas dos coeficientes de ponderação associados aos componentes principais de autovetores inferiores 0,70 e identificação dos descritores com indicação para exclusão, em cada componente, pela seleção direta dos 50 acessos de mamão. Cruz das Almas-BA, 2019.

Descritores	Componentes Principais																					
	CP14	CP15	CP16	CP17	CP18	CP19	CP20	CP21	CP22	CP23	CP24	CP25	CP26	CP27	CP28	CP29	CP30	CP31	CP32	CP33	CP34	CP35
CI8	-0,06	-0,07	0,03	0,01	0,02	0,02	0,22	0,20	0,20	0,23	0,09	0,16	0,16	0,12	0,02	0,04	0,11	0,09	-0,03	-0,07	0,02	-0,02
CI12						**0,32**	0,00	0,06	0,06	0,07	-0,03	-0,26	-0,13	-0,26	-0,21	-0,21	-0,18	-0,07	0,04	-0,06	-0,19	0,06
CF8																**0,34**	0,24	0,06	-0,07	-0,34	0,02	-0,35
CF12							**-0,25**	0,09	0,14	0,17	0,01	0,04	0,14	0,01	0,09	-0,06	-0,21	-0,15	0,03	-0,13	-0,27	0,06
LF8	-0,06	-0,24	0,27	0,29	-0,19	0,27	-0,11	0,02	0,10	-0,04	0,39	0,02	0,00	0,16	-0,27	0,14	0,10	0,15	-0,13	0,05	0,24	0,06
LF12		**0,36**	0,04	0,04	0,37	0,10	0,03	-0,19	-0,09	-0,19	0,29	0,07	-0,01	0,04	0,03	0,22	0,29	0,27	0,12	0,33	-0,27	0,28
CPF8																			**0,56**	0,19	0,15	-0,01
CPF12	-0,20	0,06	0,14	0,10	0,20	0,26	0,18	0,00	0,32	0,29	0,15	0,04	-0,31	0,13	0,41	0,14	0,04	-0,19	0,03	-0,13	-0,10	-0,23
ALT8															**0,42**	0,22	-0,11	-0,10	0,44	0,07	-0,07	0,25
ALT12																		**0,52**	-0,13	0,00	-0,21	0,24
DC8																					**0,39**	0,12
DC12																						**0,43**
NF8					**0,58**	-0,25	0,07	-0,12	0,04	-0,03	-0,09	0,04	0,18	0,00	-0,05	0,02	-0,15	0,04	-0,07	0,03	-0,04	0,19
NFCo8																				**0,41**	0,18	-0,10
NFCa8	-0,02	-0,18	-0,19	-0,25	0,23	0,20	-0,20	0,12	-0,05	0,00	-0,06	-0,06	-0,20	0,00	-0,30	0,27	0,11	-0,17	0,17	-0,08	-0,24	-0,01
NF12																	**0,53**	-0,29	-0,05	0,06	-0,19	0,15
NFCo12													**-0,32**	0,01	-0,25	0,31	-0,02	-0,15	-0,10	-0,12	0,13	0,11
NFCa12	0,18	0,04	-0,22	-0,15	0,06	-0,07	0,14	0,10	-0,04	-0,02	0,31	0,12	0,02	0,12	-0,12	-0,05	-0,20	0,02	-0,02	-0,32	0,01	0,28
NFaxi											**-0,39**	0,07	-0,09	-0,06	-0,08	0,07	-0,11	0,08	0,35	-0,22	0,38	0,09
ALT1F										**0,49**	0,24	0,09	0,08	-0,04	**-0,21**	-0,01	0,05	0,00	-0,04	0,06	-0,04	-0,12
CPF	-0,15	0,04	0,01	-0,06	-0,02	0,29	-0,18	-0,10	-0,35	0,18	0,16	0,25	-0,05	0,31	-0,05	-0,33	-0,08	-0,01	0,14	0,10	0,22	0,04
NFP									**-0,46**	0,41	0,05	-0,12	-0,05	-0,09	0,08	0,28	-0,18	0,13	-0,06	0,15	-0,14	-0,05
CPI								**-0,36**	0,21	0,02	-0,24	0,02	-0,14	0,32	-0,21	-0,19	-0,05	-0,16	0,07	0,23	-0,05	-0,09
CCFher	0,19	-0,16	0,12	0,07	0,00	0,08	0,17	0,01	-0,17	0,00	-0,13	-0,05	0,01	0,11	0,22	-0,29	0,19	0,15	0,02	0,22	-0,03	-0,20
PF														**0,35**	-0,18	-0,09	-0,26	0,07	0,06	0,08	-0,22	-0,07
CF	**0,31**	-0,20	-0,18	-0,11	0,08	0,17	0,22	-0,30	0,09	0,31	-0,02	0,02	-0,05	-0,16	0,04	-0,15	0,41	-0,13	0,04	-0,17	0,24	0,28
DF			**-0,29**	0,06	-0,04	0,27	0,12	-0,23	0,07	-0,19	0,05	-0,17	0,16	0,03	-0,13	0,18	-0,09	-0,14	0,05	0,05	0,16	-0,03
MFF	-0,29	0,03	0,04	-0,10	0,09	0,00	-0,05	0,07	-0,03	0,03	-0,01	0,01	-0,11	0,04	-0,03	0,00	0,02	0,01	0,04	-0,16	0,06	0,16
DCC	0,07	0,00	-0,12	0,02	0,02	0,04	-0,14	0,13	0,02	-0,06	0,05	-0,15	-0,01	0,26	0,05	-0,09	0,02	0,03	0,01	-0,14	-0,01	0,21
PFS												**0,52**	-0,19	-0,32	-0,02	-0,05	-0,09	0,05	-0,14	0,07	-0,06	-0,06
PS				**0,45**	0,02	0,03	0,03	-0,03	-0,02	0,08	-0,05	-0,16	0,07	-0,02	0,00	0,07	0,09	-0,04	0,00	0,08	0,01	-0,07
AC	0,02	0,07	-0,03	0,03	0,06	0,06	-0,16	-0,06	-0,13	0,24	0,03	-0,31	0,18	0,01	-0,02	0,14	-0,02	0,01	0,03	0,05	0,09	-0,11
VITC	-0,11	0,00	0,01	-0,02	0,02	0,05	-0,05	-0,01	-0,01	0,01	0,03	0,08	-0,01	-0,09	0,01	-0,01	-0,03	0,03	-0,01	0,00	0,04	0,00
PH	0,10	0,09	-0,01	0,02	0,03	0,04	0,15	-0,04	-0,07	-0,03	0,03	-0,23	0,22	-0,05	-0,01	0,11	-0,02	0,00	0,00	0,07	0,14	-0,10
BRIX	-0,24	-0,04	0,22	-0,34	0,01	0,23	0,11	-0,11	0,03	-0,13	-0,03	0,33	0,01	-0,35	-0,05	0,13	-0,11	-0,04	-0,18	0,18	0,08	-0,05

Comprimento dos internódios: 8 meses (CI8); Comprimento dos internódios: 12 meses (CI12); Comprimento da folha: 8 meses (CF8); Comprimento da folha: 12 meses (CF12); Largura da folha: 8 meses (LF8); Largura da folha: 12 meses (LF12); Comprimento do pecíolo da folha: 8 meses (CPF8); Comprimento do pecíolo da folha: 12 meses (CPF12); Altura da planta: 8 meses (ALT8); Altura da planta: 12 meses (ALT12); Diâmetro do caule :8 meses (DC8); Diâmetro do caule: 12 meses (DC12); Nº frutos: 8 meses (NF8); Nº frutos comerciais: 8 meses (NFCo8); Nº frutos carpelóides: 8 meses (NFCa8); Nº frutos: 12meses (NF12); Nº frutos comerciais: 12 meses (NFCo12); Nº frutos carpelóides: 12 meses (NFCa12); Nº frutos por axila (NFaxi); Altura dos primeiros frutos (ALT1F); Comprimento do pedúnculo do fruto (CPF); Número de flores por pedúnculo (NFP); Comprimento do pedúnculo da inflorescência (CPI); Comprimento da corola da flor hermafrodita (CCFher); Peso do fruto (PF); Comprimento do fruto (CF); Diâmetro do fruto (DF); Média firmeza dos frutos (MFF); Diâmetro da cavidade central (DCC); Peso fresco de sementes do fruto (PFS); Peso fresco de 100 sementes (PS); Acidez (AC); Vit C (VITC); pH (PH); Sólidos solúveis totais: Brixº (BRIX).

Segundo o método de Singh (Figura 3), os descritores que proporcionaram maiores contribuições relativas foram o peso fresco de sementes do fruto (PFS), com 41,83% de contribuição quanto à divergência genética dos acessos, sendo este responsável pela maior percentagem de toda variabilidade dos dados. Como a propagação comercial do mamoeiro é realizada por sementes, a avaliação de características relacionadas as mesmas, são de grande importância, pois o uso de sementes de alta qualidade é essencial para o estabelecimento de mudas vigorosas e sadias (MARCOS FILHO, 2005). O segundo descritor com maior contribuição foi a vitamina C (VITC), com 18,14%. O resultado apresentado para VITC é importante para os programas de melhoramento de mamão, a fim de aumentar os teores desse nutriente nos frutos.

Já o terceiro descritor com maior contribuição, foi o número de frutos: 8 meses (NF8) com 5,80%, esses três descritores contribuíram com 65,77% da distribuição total (Figura 3). Os descritores que apresentaram contribuição intermediária e juntos representaram 28,70% de contribuição, foram LF8; LF12; CPF8; CPF12; NFCo8; NF12; NFCo12 e DCC.

As menores contribuições foram para os descritores: CI8; CI12; CF8; CF12; ALT8; ALT12; DC8; DC12; NFCa8; NFCa12; NFaxi; ALT1F; CPF; NFP; CPI; CCFher; PF; CF; DF; MFF; PS; AC; PH e BRIX. Estes últimos 24 descritores totalizaram apenas 5,53% da importância relativa e, portanto, foram considerados de pouca importância na caracterização dos genótipos de mamão e podem ser descartadas (Figura 3).

Dentre os descritores indicados com menor contribuição, a altura dos primeiros frutos (ALT1F), os sólidos solúveis totais (BRIX), peso do fruto (PF), número de frutos carpelóides (NFCa12) e firmeza dos frutos (MFF) são extremamente importantes na caracterização de uma variedade de mamão e também permite a classificação e o uso final dos frutos para diferentes tipos de mercados (OLIVEIRA et al., 2012). De acordo com os resultados obtidos pelo critério de Singh, salienta-se que o pesquisador deve ter bastante atenção na seleção de descritores baseados exclusivamente nesse critério, pois este, nem sempre proporciona uma distribuição real da variabilidade inerente aos descritores avaliados do conjunto de dados em estudo.

O descritor número de frutos carpelóides (NFCa12) também chamados de frutos deformados, é de fundamental importância para os programas de

melhoramento, pois está diretamente relacionada à produtividade. Quanto maiores os valores de frutos deformados, consequentemente, menor vai ser a produtividade, indicando que a seleção de genótipos de mamoeiro deve ser realizada visando selecionar aqueles que apresentam menores valores para essa característica (DANTAS et al., 2015).

Os resultados obtidos por meio da análise de componentes principais e pelo critério de Singh irão auxiliar na escolha dos descritores que serão eliminados na estratégia combinada com diagnóstico de multicolinearidade pelo (NC), pois os descritores com menor contribuição segundo o critério de Singh terão maior peso na decisão de descarte, bem como os primeiros descritores indicados para descarte utilizando componentes principais, que também terão maior peso nesse processo.

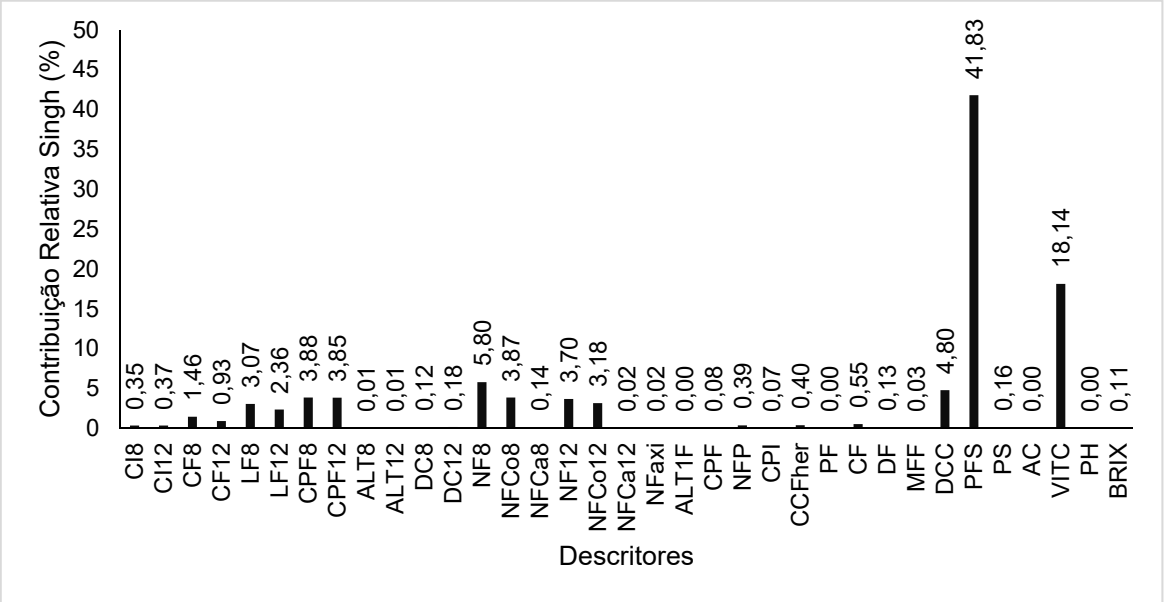


Figura 3. Contribuição relativa para divergência genética segundo o critério de Singh 1981, dos 35 descritores quantitativos avaliados, em 50 acessos de mamoeiro (*Carica papaya* L.) do Banco de Germoplasma da Embrapa Mandioca e Fruticultura.

Na figura 4, são apresentados os valores do fator de inflação da variância (VIF), para todos os descritores estudados, onde na primeira etapa de validação da seleção dos descritores utilizando o VIF, foi detectado que 71,43% dos descritores apresentaram o valor acima de 10, sendo essas: ALT8; ALT12; LF8; CF8; PF; DC8; NFCo12; DF; LF12; NF8; CF12; NF12; NFCo8; CPF8; NFaxi; ALT1F; CPF12; CI8;

CF; CPI; DC12; CI12; VITC; DCC e AC. Observa-se que os maiores valores do VIF são para as vaiáveis ALT8 e ALT12 com 69,93 e 56,54 respectivamente (Figura 4).

Segundo Kutner et al. (2004), quanto maior for o VIF, mais severa será a multicolinearidade. Estes altos valores de VIF são explicados devido à forte relação entre ALT8 e ALT12, o que foi verificado aplicando a correlação de Spearman, que foi de 0,92** (Tabela 5). O mesmo cenário pode ser observado nos dois próximos descritores com os maiores valores de VIF, o descritor LF8 e o CF8 com 54,22 e 54,11 respectivamente, que também apresentam uma alta correlação de 0,95** (Tabela 5).

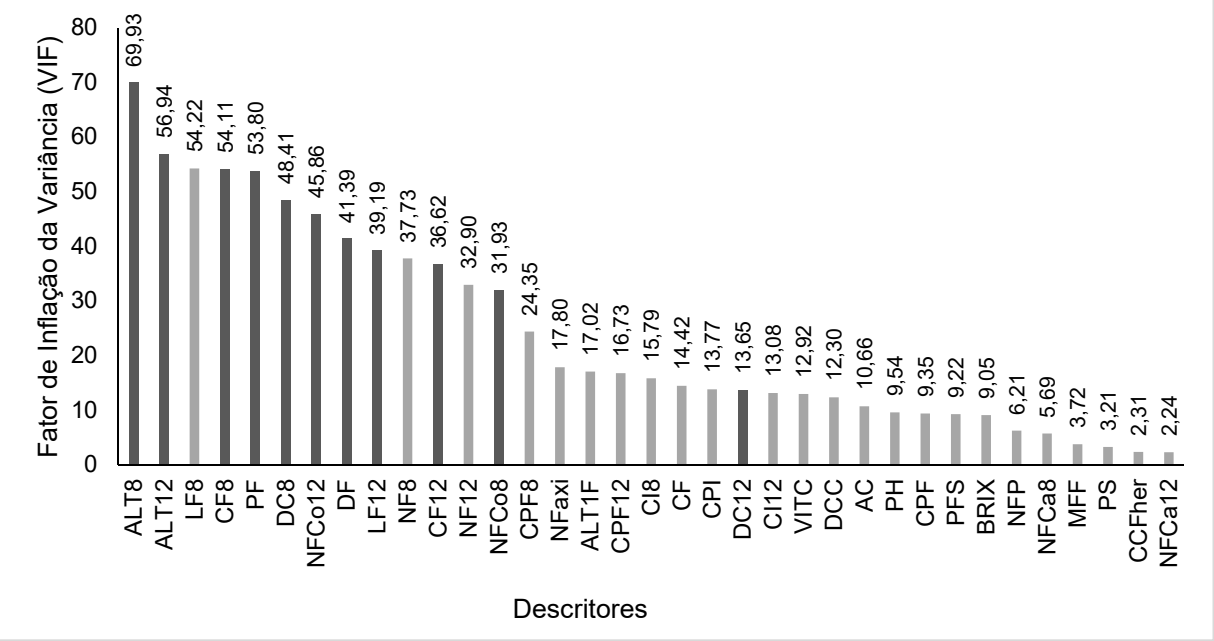


Figura 4. Fator de Inflação da Variância (VIF) entre os 35 descritores quantitativos avaliados, em 50 acessos de mamoeiro (*Carica papaya* L.) do Banco de Germoplasma da Embrapa Mandioca e Fruticultura.

É apresentado na tabela 5, os resultados da estratégia combinada de seleção, onde o primeiro diagnóstico de multicolinearidade realizado, levando em consideração a correlação entre todos os trinta e cinco descritores em estudo, que o número de condição (NC), situou-se na faixa considerada Severa (NC>1000) com o valor de 28972,26, sendo indicado para o descarte um dos descritores que apresentaram a maior correlação, onde, o número de frutos: 8 meses (NF8) e número de frutos comerciais: 8 meses (NFCo8) foram os descritores com maior valor de associação (ρ =0,96), com o auxílio da análise de componentes principais

e contribuição de Singh. Foi escolhido para exclusão o descritor NFCo8, pois este apresentou a menor contribuição para divergência genética segundo Singh e na ordem de descarte por componentes principais foi o terceira a ser descartado, já a NF8 foi o decimo sétimo descritor indicado para um possível descarte (Tabela 5). A colinearidade ou a correlação excessiva entre os descritores explicativos podem complicar ou interferir na identificação de um conjunto ótimo de descritores explicativos para um modelo estatístico.

Foram realizados onze diagnósticos de multicolinearidade (Tabela 5), até que no último foi observado o número de condição com o valor de 98,63, considerada uma colinearidade fraca (NC<100), onde envolveu a correlação entre Comprimento do pecíolo da folha: 12 meses (CPF12) e Largura da folha: 12 meses (LF12), com valor de associação (ρ =0,69), sendo excluída o descritor LF12, levando em consideração a ordem de descarte por componentes principais e a maior contribuição segundo Singh.

A combinação de informações dos métodos na estratégia de seleção, permitiu que 31,43% dos descritores fossem eliminados, ou seja, onze dos trinta e cinco descritores avaliados ao longo de todo o processo. Oliveira et al. (2012), estudando seleção de descritores em acessos e mamão, obteve uma proporção de 40% de descritores eliminados.

Os Descritores descartados nessa estratégia foram: número de frutos comerciais: 8 meses (NFCo8); comprimento da folha: 8 meses (CF8); o número de frutos comerciais: 12 meses (NFCo12); Altura da planta: 12 meses (ALT12); Comprimento da folha: 12 meses (CF12); Diâmetro do caule: 12 meses (DC12); Peso do fruto (PF); Diâmetro do caule:8 meses (DC8); Diâmetro do fruto (DF); Altura da planta: 8 meses (ALT8) e Largura da folha: 12 meses (LF12), (Tabela 5).

Nessa estratégia de seleção, alguns descritores importantes citados por Morais et al., (2007); Oliveira et al., (2012) e Dantas et al., (2015), foram preservados, como a altura dos primeiros frutos (ALT1F), sólidos solúveis totais (BRIX); número de frutos carpelóides (deformados) (NFCa12) e firmeza dos frutos (MFF).

A altura de inserção dos primeiros frutos (ALTF1) é um descritor interessante e de fundamental importância nos programas de melhoramento de mamoeiro, pois quanto menor o valor obtido para este caráter, mais precocemente a planta começa a produzir frutos, indicando precocidade e maior facilidade para a colheita de frutos

em ciclos de produção mais avançados (DIAS et al., 2011; Dantas et al., 2015). Muitos autores também recomendam cultivares de mamão que exibem menor ALTF1 associada a menor PH (ALONSO et al., 2008; OLIVEIRA et al., 2010).

O diâmetro da cavidade central (DCC), foi outro descritor não descartado no presente estudo. Segundo Dias et al. (2011), este descritor está relacionado à qualidade dos frutos, pois aqueles com menor diâmetro da cavidade interna, geralmente, apresentam maior quantidade de polpa.

A firmeza do fruto, é um atributo de qualidade muito importante que estabelece a vida útil pós-colheita, uma vez que frutos com baixa firmeza apresentam menor resistência ao transporte, armazenamento e ao manuseio, influenciando diretamente na comercialização (FAGUNDES; YAMANISHI, 2001; MORAIS et al., 2007; FONTES et al., 2012).

Os atributos, PH, AC e BRIX, estão diretamente relacionados à qualidade dos frutos e, em alguns casos, são considerados cruciais na comercialização do mamão (SILVA et al., 2018). Segundo Dantas et al. (2015) os Sólidos solúveis totais (BRIX) são um dos principais parâmetros de qualidade de frutos de mamão. A exigência para comercialização de mamão para exportação, é que apresente °Brix superior a 12 (FAGUNDES e YAMANISHI (2001); SCHWEIGGERT et al., 2012; DANTAS et al., 2015).

De acordo com Oliveira et al. (2012), as análises simultâneas com vários métodos parecem ser uma estratégia eficaz para minimizar erros na eliminação de descritores. O emprego de mais de um procedimento no descarte dos descritores redundantes é um procedimento que tem sido adotado para dar maior segurança na seleção de descritores (CASTRO et al., 2012; SILVA et al., 2013; AFONSO et al., 2014; OLIVEIRA et al., 2014; SILVA et al., 2017; SILVA et al., 2018).

Tabela 5. Estratégia de seleção combinada com Diagnóstico de Multicolinearidade (NC), com auxílio dos métodos (Singh e CPA). Descritores Descartados (em negrito) e Análise Descritiva das Correlações. Cruz das Almas-Ba, 2019.

Singh	CPA	Descritores	Correlação ρ	NC	Colinearidade	Ord. DM
5,80	17D	NF8	0,96	28972,26	Severa	1
3,87	3D	**NFCo8**				
3,07	nd	LF8	0,95	19227,70	Severa	2
1,46	6D	**CF8**				
3,70	nd	NF12	0,93	1754,33	Severa	3
3,18	9D	**NFCo12**				
0,01	7D	ALT8	0,92	1670,96	Severa	4
0,01	5D	**ALT12**				
2,36	20D	LF12	0,90	627,77	Mod/forte	5
0,93	15D	**CF12**				
0,12	2D	DC8	0,88	469,55	Mod/forte	6
0,18	1D	**DC12**				
0,13	19D	DF	0,86	403,40	Mod/forte	7
0,00	8D	**PF**				
3,88	4D	CPF8	0,85	372,42	Mod/forte	8
0,12	2D	**DC8**				
0,13	nd	DCC	0,73	245,91	Mod/forte	9
4,80	19D	**DF**				
0,00	12D	ALT1F	0,71	192,82	Mod/forte	10
0,01	7D	**ALT8**				
3,85	nd	CPF12	0,69	98,63	Fraca	11
2,36	20D	**LF12**				

Contribuição de Singh (Singh); Descarte por Análise de Componentes Principais (CPA); Ordem de descarte (1D, 2D...20D); Não descartada utilizando componentes principais (nd); Correlação de Spearman (ρ); Número de condição (NC); Ordem em sequência dos diagnósticos de multicolinearidade (Ord. DM); Moderada a forte (Mod/forte); Nº frutos comerciais: 8 meses (NFCo8); Comprimento da folha: 8 meses (CF8); Nº frutos comerciais: 12 meses (NFCo12); Altura da planta: 12 meses (ALT12); Comprimento da folha: 12 meses (CF12); Diâmetro do caule: 12 meses (DC12); Peso do fruto (PF); Diâmetro do caule :8 meses (DC8); Diâmetro do fruto (DF); Altura da planta (8 meses) (ALT8); Largura da folha: 12 meses (LF12); Nº frutos: 8 meses (NF8); Largura da folha: 8 meses (LF8); Nº frutos: 12meses (NF12); Comprimento do pecíolo da folha: 8 meses (CPF8); Diâmetro da cavidade central (DCC); Altura dos primeiros frutos (ALT1F); Comprimento do pecíolo da folha: 12 meses (CPF12).

Na Figura 5, são apresentados os valores do fator de inflação da variância (VIF) para validação dos descritores selecionados, onde é mostrado que todos

expressaram valores do VIF abaixo de 10. Segundo Tamhane & Dunlop (2000); Kutner et al. (2004), valores do VIF para serem considerados aceitáveis devem ser inferiores a 10.

Observa-se ainda na Figura 5, que o menor valor de VIF foi de 1,50 e o maior de 8,84 para os descritores NFCa12 e CPF8, respectivamente. Esses resultados evidenciam não haver indícios de multicolinearidade, com isso, a estratégia de seleção aplicada mostra-se eficaz na seleção de descritores

.

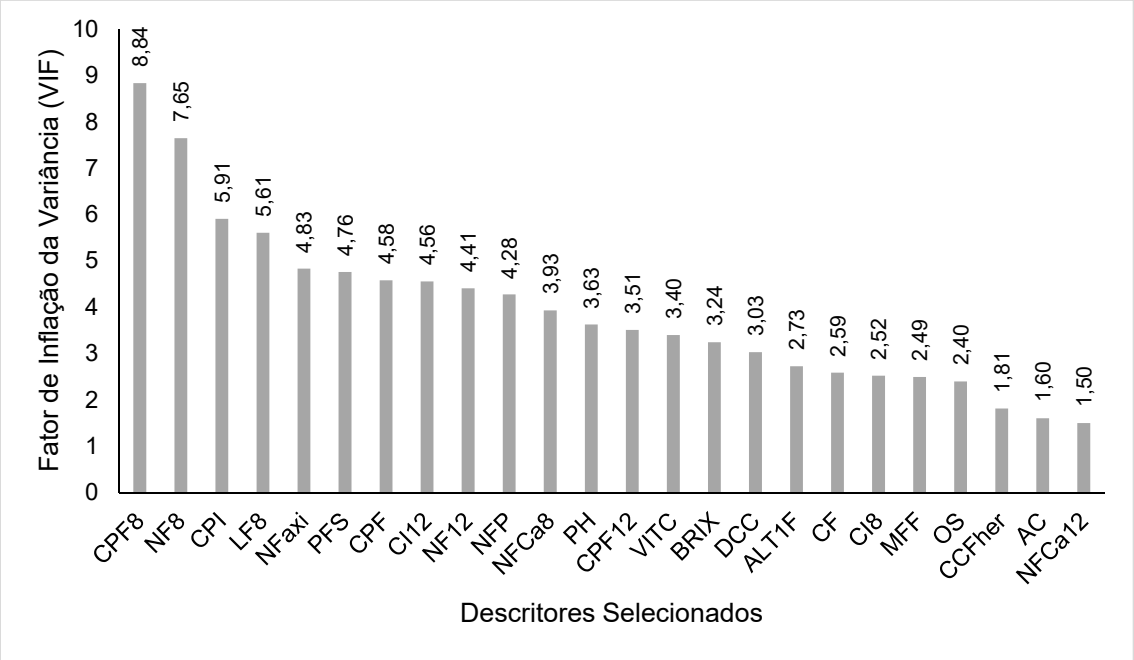


Figura 5. Fator de Inflação da Variância (VIF) entre os 24 descritores quantitativos selecionados, avaliadas em 50 acessos de mamoeiro (*Carica papaya* L.) do Banco de Germoplasma da Embrapa Mandioca e Fruticultura.

Nesse estudo as estimativas dos coeficientes de correlação de Spearman entre os descritores remanescentes com os descartados são apresentados na Tabela 6, em que foi observado correlação positiva e altamente significativa, com alta magnitude entre os descritores descartados e os selecionados. Como foi apresentado entre os descritores que expressaram as maiores correlações: número de frutos comerciais aos 8 meses (NFCo8) com número de frutos aos 8 meses NF8 com (r=0,96**), onde o (NFCo8) é proporcional ao (NF8). Para a correlação entre CF8 e LF8 com (r=0,95**), observa-se que quando a um aumento do comprimento da folha (CF8), também aumenta-se a largura da folha.

A menor magnitude das correlações entre os descritores descartados e os selecionados foi observado entre a largura da folha (LF12) e o comprimento do pecíolo da folha (CPF12), porém essa correlação foi positiva e significativa com valor de (r=0,69**), onde possuem uma relação proporcional (Tabela 6). O pecíolo tem uma forte relação com a folha por esta ligado a sustentação da mesma.

Foi observada também correlação positiva (r=0,76) entre altura da planta: 12 meses (ALT12) e altura dos primeiros frutos (ALT1F). No estudo realizado por Ocampo et al., 2006; Ide (2008); Lucena (2013) e Nascimento (2018) foram encontrados valores semelhantes para relação entre esses dois descritores. Segundo Lucena (2013), a redução no porte da planta implicará em menor altura dos primeiros frutos. A altura de inserção dos primeiros frutos no presente estudo apresentou média variando de 40,00 a 210,0 cm. Marin et al. (1989) recomendam selecionar mamoeiro com altura de inserção inferior a 80 cm. De acordo com Nascimento (2018) essa alta correlação entre altura da planta e altura dos primeiros frutos, indica que ao selecionar genótipos com classes fenotípicas definidas para tais características, estes terão um maior número de alelos efetivos e com baixa interferência do efeito de dominância facilitando a seleção de novas cultivares.

O diâmetro do fruto (DF) e o diâmetro da cavidade central (DCC) apresentaram correlação elevada e positiva (r=0,73), esse resultado indica que frutos com maior diâmetro tendem a apresentar maior diâmetro da cavidade central. Correlações elevadas e positivas entre esses descritores foram observadas nos trabalhos de Oliveira et al., (2010) e Reis et al., (2015) estudando plantas de mamoeiro. Segundo Oliveira et al., (2010), o aumento na cavidade central pode levar à redução em sólidos solúveis totais (BRIX), o que é indesejável para o consumidor da fruta. Além disso, frutos com maior diâmetro da cavidade interna são mais suscetíveis às perdas ocorridas durante o transporte e armazenamento.

As estimativas da correlação de Spearman, entre o conjunto de descritores redundantes e o dos selecionados, demonstram que o descarte de parte dos descritores no presente estudo se mostrou eficiente, pois os mesmos apresentam altas correlações significativas com os descritores remanescentes. Com isso, o conjunto reduzido de descritores após o descarte se mostrou adequado na representação da variação total e sem indícios de multicolinearidade (Tabela 5). Assim, torna-se possível a eliminação de descritores sem perda de informação, pois

os mesmos podem estar correlacionados a outros que permaneceram na análise, conforme demonstrado neste trabalho.

Tabela 6. Estimativas dos coeficientes de correlação de Spearman entre os descritores morfoagronômicos remanescentes e os descartados, avaliados em 50 acessos do banco de germoplasma de mamão da Embrapa Mandioca e Fruticultura. Cruz das Almas-BA, 2019.

R	Descartados (D)										
	1D	**2D**	**3D**	**4D**	**5D**	**6D**	**7D**	**8D**	**9D**	**10D**	**11D**
AC	-0,20^{ns}	0,05^{ns}	0,06^{ns}	-0,06^{ns}	-0,04^{ns}	0,09^{ns}	-0,05^{ns}	0,08^{ns}	-0,15^{ns}	-0,10^{ns}	-0,01^{ns}
ALT1F	-0,40**	0,34*	-0,05^{ns}	0,76**	0,32*	0,59**	0,12^{ns}	0,49**	0,04^{ns}	0,70**	0,29^{ns}
BRIX	-0,31*	-0,01^{ns}	-0,06^{ns}	-0,03^{ns}	0,07^{ns}	0,07^{ns}	-0,16^{ns}	-0,05^{ns}	-0,18^{ns}	-0,07^{ns}	-0,01^{ns}
CF	-0,03^{ns}	0,46**	-0,22^{ns}	0,18^{ns}	0,50**	0,35*	0,69**	0,23^{ns}	0,43**	0,19^{ns}	0,26^{ns}
CI8	-0,05^{ns}	0,18^{ns}	-0,12^{ns}	0,46**	0,05^{ns}	0,41**	-0,06^{ns}	0,41**	-0,11^{ns}	0,48**	-0,07^{ns}
CI12	0,12^{ns}	0,38*	0,10^{ns}	0,60**	0,21^{ns}	0,52**	0,04^{ns}	0,66**	-0,05^{ns}	0,67**	0,06^{ns}
CPF	0,16^{ns}	0,42**	0,12^{ns}	0,14^{ns}	0,14^{ns}	0,33*	0,22^{ns}	0,29^{ns}	0,05^{ns}	0,23^{ns}	-0,03^{ns}
CPF8	0,05^{ns}	0,56**	0,26^{ns}	0,62**	0,55**	0,79**	0,33*	0,86**	0,29^{ns}	0,61**	0,56**
CPF12	-0,34*	0,26^{ns}	0,13^{ns}	0,33*	0,64**	0,52**	0,23^{ns}	0,39*	0,27^{ns}	0,17^{ns}	0,69**
CPI	-0,17^{ns}	0,58**	0,16^{ns}	0,28^{ns}	0,40**	0,42**	0,19^{ns}	0,33*	0,11^{ns}	0,20^{ns}	0,27^{ns}
DCC	-0,05^{ns}	0,09^{ns}	-0,21^{ns}	0,01^{ns}	0,19^{ns}	0,15^{ns}	0,64**	0,11^{ns}	0,73**	0,00^{ns}	0,08^{ns}
LF8	-0,14^{ns}	0,95**	-0,04^{ns}	0,52^{ns}	0,76**	0,66**	0,44**	0,59**	0,31*	0,50**	0,33*
MFF	-0,26^{ns}	0,15^{ns}	-0,38*	-0,06^{ns}	0,14^{ns}	0,02^{ns}	0,31*	0,01^{ns}	0,30*	-0,05^{ns}	0,01^{ns}
NF8	0,96**	-0,03^{ns}	0,40**	-0,04^{ns}	-0,25^{ns}	-0,23^{ns}	-0,22^{ns}	0,07^{ns}	-0,26^{ns}	0,13^{ns}	-0,33*
NF12	0,33*	0,04^{ns}	0,93**	0,30*	-0,08^{ns}	0,11^{ns}	-0,38*	0,19^{ns}	-0,36*	0,28^{ns}	0,15^{ns}
NFaxi	0,21^{ns}	-0,09^{ns}	0,56**	0,17^{ns}	-0,12^{ns}	0,09^{ns}	-0,39**	0,16^{ns}	-0,26^{ns}	0,15^{ns}	-0,01^{ns}
NFCa8	0,14^{ns}	-0,06^{ns}	-0,17^{ns}	-0,27^{ns}	-0,14^{ns}	-0,22^{ns}	-0,10^{ns}	-0,11^{ns}	-0,06^{ns}	-0,15^{ns}	-0,34*
NFCa12	0,10^{ns}	0,15^{ns}	0,08^{ns}	0,07^{ns}	0,05^{ns}	0,00^{ns}	-0,12^{ns}	0,01^{ns}	-0,06^{ns}	0,11^{ns}	-0,12^{ns}
NFP	-0,26^{ns}	0,42**	0,09^{ns}	0,18^{ns}	0,34*	0,37**	0,22^{ns}	0,26^{ns}	0,10^{ns}	0,20^{ns}	0,14^{ns}
PS	-0,25^{ns}	0,24^{ns}	-0,11^{ns}	0,16^{ns}	0,18^{ns}	0,24^{ns}	0,30*	0,22^{ns}	0,24^{ns}	0,19^{ns}	-0,03^{ns}
PFS	-0,20^{ns}	0,21^{ns}	-0,18^{ns}	0,27^{ns}	0,23^{ns}	0,36*	0,57**	0,34*	0,54**	0,29^{ns}	0,08^{ns}
PH	-0,19^{ns}	-0,12^{ns}	-0,19^{ns}	0,00^{ns}	-0,09^{ns}	0,04^{ns}	0,14^{ns}	-0,01^{ns}	0,25^{ns}	-0,03^{ns}	0,14^{ns}
VITC	-0,28^{ns}	-0,19^{ns}	-0,15^{ns}	-0,09^{ns}	-0,16^{ns}	-0,16^{ns}	-0,17^{ns}	-0,30*	-0,22^{ns}	-0,12^{ns}	-0,13^{ns}

Comprimento dos internódios: 8 meses (CI8); Comprimento dos internódios: 12 meses (CI12); Largura da folha: 8 meses (LF8); Comprimento do pecíolo da folha: 8 meses (CPF8); Comprimento do pecíolo da folha: 12 meses (CPF12); Nº frutos: 8 meses (NF8); Nº frutos carpelóides: 8 meses (NFCa8); Nº frutos: 12meses (NF12); Nº frutos carpelóides: 12 meses (NFCa12); Nº frutos por axila (NFaxi); Altura dos primeiros frutos (ALT1F); Comprimento do pedúnculo do fruto (CPF); Número de flores por pedúnculo (NFP); Comprimento do pedúnculo da inflorescência (CPI); Comprimento do fruto (CF); Firmeza dos frutos (MFF); Diâmetro da cavidade central (DCC); Peso fresco de sementes do fruto (PFS); Peso fresco de 100 sementes (PS); Acidez (AC); Vit C (VITC); pH (PH); Sólidos solúveis totais: Brixº (BRIX). Nº frutos comerciais: 8 meses (**1D**); Comprimento da folha: 8 meses **(2D**); Nº frutos comerciais: 12 meses (**3D**); Altura da planta: 12 meses (**4D**); Comprimento da folha: 12 meses (**5D**); Diâmetro do caule: 12 meses (**6D**); Peso do fruto (**7D**); Diâmetro do caule: 8 meses (**8D**); Diâmetro do fruto (**9D**); Altura da planta: 8 meses (**10D**); Largura da folha: 12 meses (**11D**). Remanescentes (R).

A aplicação e criação de técnicas que visem a seleção de descritores de maneira mais precisas são fundamentais, pois é através dessa seleção que serão

caracterizados o conjunto de indivíduos de determinada espécie a serem estudados, sendo esse processo crucial, pois irá contribuir de maneira significativa para o sucesso dos programas de melhoramento e conservação da espécie. Além de proporcionar uma redução nos gastos com mão-de-obra e no tempo destinado à tomada de dados em futuros experimentos envolvendo esses descritores e a cultura avaliada.

CONCLUSÕES

A nova estratégia aplicada para seleção de descritores, mostrou-se eficaz no estudo dos acessos de *Carica papaya* L., sem causar perda considerável de informação. Preservando também, descritores importantes como a altura dos primeiros frutos, os sólidos solúveis totais, número de frutos carpelóides, diâmetro da cavidade central e firmeza dos frutos.

Os descritores analisados, relativamente redundantes podem ser descartados em experimentos futuros na proporção de 31,43%.

Os descritores selecionados são: comprimento dos internódios 8 meses; comprimento dos internódios: 12 meses; Largura da folha: 8 meses; comprimento do pecíolo da folha: 8 meses; comprimento do pecíolo da folha: 12 meses; nº frutos: 8 meses; nº frutos carpelóides: 8 meses; nº frutos: 12meses; nº frutos carpelóides: 12 meses; nº frutos por axila; altura dos primeiros frutos; comprimento do pedúnculo do fruto; nº de flores por pedúnculo; comprimento do pedúnculo da inflorescência; comprimento da corola da flor hermafrodita; comprimento do fruto; firmeza dos frutos; diâmetro da cavidade central; peso fresco de sementes do fruto; peso fresco de 100 sementes; acidez; vitamina C; pH e sólidos solúveis totais..

Com o descarte de descritores redundantes, espera-se uma economia relativa de tempo e de custos em experimentos futuros.

AGRADECIMENTOS

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa durante todo o período de realização deste doutorado e a UFRB e Embrapa Mandioca e Fruticultura, que disponibilizaram estrutura física e equipamentos adequados para execução do presente trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

AFONSO, S.D.J.; LEDO, C.A. da S.; MOREIRA, R.F.C.; SILVA, S. de O. e; LEAL, V.D. de J.; CONCEIÇÃO, A.L. da S. Selection of descriptors in a morphological characteristics considered in cassava accessions by means of multivariate techniques. **Journal of Agriculture and Veterinary Science**, v.7, p.13-20, 2014.

AIKPOKPODION, P. O. Assessment of genetic diversity in horticultural and morphological traits among papaya (Carica papaya) accessions in Nigeria. **Fruits**, v. 67, n. 3, p. 173-187, 2012.

ALONSO, M.; TORNET, Y.; RAMOS, R.; FARRÉS, E.; CASTRO, J.; RODRÍGUEZ, M. C. Evaluación de três cultivares de papaya del Grupo Solo basada em caracteres de crecimiento y productividad. **Cultivos Tropicales**, La Habana, v. 29, n. 2, p. 59-64, 2008.

ALVES, R. M. **Caracterização genética de populações de cupuaçuzeiro, Theobroma grandiflorum (Will ex Spreng) Schum., por marcadores microssatélites e descritores botânico-agronô**micos.Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.146, 2002.

ASUDI, G. O.; OMBWARA, F. K.; RIMBERIA, F. K.; NYENDE, A. B.; ATEKA, E. M.; WAMOCHO, L. S.; ONYANGO, A. Morphological diversity of Kenyan papaya germplasm. **African Journal of Biotechnology**, v. 9, n. 51, p. 8754-8762, 2010.

BARROS, F. L. D. S.; KUHLCAMP, K. T.; ARANTES, S. D.; MOREIRA, S. O. Productivity and quality of Formosa and Solo papaya over two harvest seasons. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 599-606, 2017.

BERK, K. N. Tolerance and condition in regression computations. **Journal of the American Statistical Association**, v. 72, n. 360a, p. 863-866, 1977.

BRITO NETO, J. F.; PEREIRA, W. E.; CAVALCANTI, L. F.; da COSTA ARAÚJO, R.; DE LACERDA, J. S. Produtividade e qualidade de frutos de mamoeiro Sunrise Solo em função de doses de nitrogênio e boro. **Semina: Ciências Agrárias**, Londrina, v. 32, n. 1, p. 69-80, 2011.

CARDOSO, D.L.; VIVAS, M.; PINTO, F.O.; VIANA, A. P.; AMARAL JÚNIOR, A.T.; PEREIRA, M.G. Diallel mixed-model analysis of papaya fruit deformities. **Ciência Rural**, Santa Maria, v. 47, n. 5, 2017.

CASTRO, J. A.; NEVES, C. G.; de JESUS, O. N.; de OLIVEIRA, E. J. Definition of morpho-agronomic descriptors for the characterization of yellow passion fruit. **Scientia Horticulturae**, v. 145, p. 17-22, 2012

CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. 1.ed. Viçosa, MG: UFV, Imprensa Universitária, p. 390, 1994.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 3.ed. Viçosa: UFV, v. 1, Cap. 8, p. 377-413, 2004.

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

DANTAS, J. L. L.; PINTO, R. M. S.; LIMA, J.L.; FERREIRA, F. R. **Catálogo de germoplasma de mamão** (*Carica papaya* L.). Cruz das Almas-BA: Embrapa Mandioca e Fruticultura, (Embrapa Mandioca e Fruticultura, Documentos, 94), p. 40, 2000.

DANTAS, J. L. L.; LUCENA, R. S.; VILAS BOAS, S. A. AVALIAÇÃO AGRONÔMICA DE LINHAGENS E HÍBRIDOS DE MAMOEIRO. **Revista Brasileira de Fruticultura**, Jaboticabal, v. 37, n. 1, p. 138-148, Mar. 2015.

DIAS, N. L. P.; DE OLIVEIRA, E. J.; DANTAS, J. L. L. Avaliação de genótipos de mamoeiro com uso de descritores agronômicos e estimação de parâmetros genéticos. **Pesquisa Agropecuária Brasileira**, v. 46, n. 11, p. 1471-1479, 2011.

FAGUNDES, G. R.; YAMANISHI, O. K. Características físicas e químicas de frutos do mamoeiro do grupo “Solo” comercializados em quatro estabelecimentos de Brasília-DF. **Revista Brasileira de Fruticultura**, Jaboticabal, v. 23, n. 3, p. 541-545, 2001.

FAO. Faostat - Statistics Database (2014). FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. Disponível em: < http://www.fao.org/faostat/en/#data/QC/visualize>. Acesso em: 22 set. 2017.

FIORAVANÇO, J. C.; PAIVA, M. C.; CARVALHO, R. I. N.; MANICA, I. Qualidade de mamão solo que são comercializados em Porto Alegre de outubro/91 a junho/92. **Rev. Ciên. Agron**, v. 23, n. 3, p. 1-5, 1992.

FONTES, R. V.; VIANA, A. P.; PEREIRA, M. G.; OLIVEIRA, J. G.; VIEIRA, H. D. Manejo da cultura do híbrido de mamoeiro (*Carica papaya* L.) do grupo 'formosa' UENF/CALIMAN-01 para melhoria na qualidade do fruto com menor aplicação de adubação NPK. **Revista Brasileira de Fruticultura, Jaboticabal**, v. 34, n. 1, p. 143-151, 2012.

FRAIFE FILHO, G. de A.; DANTAS, J.L.L.; LEITE, J.B.V.; OLIVEIRA, J.R.P. Avaliação de variedades de mamoeiro no extremo sul da Bahia. **Magistra**, v.13, p.37-41, 2001.

HAIR, Jr.; J.H.; ANDERSON, R. E.; TATHAM, R.L.; BLACK, W.C. trad. Adonai Schlup Sant'Ana e Anselmo Chaves Neto. **Análise Multivariada de Dados**. 5 ed. Porto Alegre: Bookman. 2005.

IDE, C.D. **Melhoramento genético do mamoeiro (*Carica papaya* L.): Parâmetros genéticos e capacidade combinatória em ensaios de competição de cultivares**. Tese de Doutorado em Genética e Melhoramento de Plantas - Universidade Estadual do Norte Fluminense, Campos dos Goytacazes, RJ. 139p., 2008.

INMET - INSTITUTO NACIONAL DE METEOROLOGIA. Disponível em: <http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>. Acesso em: 20 set. 2017.

IBPGR - INTERNATIONAL BOARD FOR PLANT GENETIC RESOURCES. **Descriptor list for papaya**. Rome: IPGRI, 1988. Disponivel em: <http://pdf.usaid.gov/pdf_docs/PNABC145.pdf>.

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

JOLLIFFE, I.T. Discarding variables in a principal component analysis. II: real data. **Journal of the Royal Statistical Society**. Series C (Applied Statistics), v.22, p. 21-31, 1973.

KENNEDY, P.A. **Guide to Econometrics**, Blackwell, Oxford, 1992.

KUTNER, M. H.; NACHTSHEIM, C; NETER, J. **Applied linear models**. 5th ed. New York: McGraw-Hill Irwin, 2004.

LUCENA, R. S. **Caracterização agronômica de novas linhagens e híbridos de mamoeiro (Carica papaya L.)**. Dissertação (mestrado em Recursos Genéticos Vegetais), Universidade Federal do Recôncavo da Bahia, p. 67, 2013.

MARCOS FILHO, J. M. F. **Fisiologia de sementes de plantas cultivadas**. Fealq, 2005.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis**. 6.ed. London: Academic Press, p. 518p, 1997.

MARQUARIDT, D. W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. **Technometrics**, v. 12, n. 3, p. 591-612, 1970.

MARTINS, D. DOS S.; COSTA, A. DE F. S. DA. **A cultura do mamoeiro:** tecnologias de produção.Vitória: Incaper, p. 497, 2003.

MASON, R. L.; GUNST, R. F.; HESS, J. L. **Statistical design and analysis of experiments: Applications to engineering and science**. New York: Wiley,1989.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to linear regression analysis**. New York: John Wiley & Sons, p. 504, 1981.

MORAIS, P. L. D.; DA SILVA, G. G.; MENEZES, J. B.; MAIA, F. N.; DANTAS, D. J.; JÚNIOR, R. Pós-colheita de mamão híbrido UENF/Caliman 01 cultivado no Rio Grande do Norte. **Revista Brasileira de Fruticultura**, Jaboticabal, v.29, n.3, p. 666-670, 2007.

NETER, J.; WASSERMAN, W.; KUTNER, M. G. Applied linear regression analysis. **Homewood, IL: Irwin**, 1989.

OCAMPO, J.; D'EECKENBRUGGE, G. C.; BRUYÈRE, S.; DE BELLAIRE, L. D. L.; OLLITRAULT, P. Organization of morphological and genetic diversity of Caribbean and Venezuelan papaya germplasm. **Fruits**, Montpellier, v. 61, n. 1, p. 25-37, 2006.

OLIVEIRA, M. do S.P. de; FERREIRA, D.F.; SANTOS, J.B. Seleção de descritores para caracterização de germoplasma de açaizeiro para produção de frutos. **Pesquisa Agropecuária Brasileira**, Brasília, v. 41, n. 7, p. 1133-1140, 2006.

OLIVEIRA, E. J.; LIMA, D. S.; LUCENA, R. S.; MOTTA, T. B. N.; DANTAS, J. L. L. Correlações genéticas e análise de trilha para número de frutos comerciais por planta em mamoeiro. **Pesquisa Agropecuária Brasileira**, v. 45, p. 855-862, 2010.

OLIVEIRA, E. J.; DIAS, N. L. P.; DANTAS, J. L. L. Selection of morpho-agronomic descriptors for characterization of papaya cultivars. **Euphytica**, v. 185, n. 2, p. 253-265, 2012.

OLIVEIRA, E.J.; OLIVEIRA FILHO, O.S. de.; SANTOS, V. da. S. Selection of the most informative morphoagronomic descriptors for cassava germplasm. **Pesquisa Agropecuária Brasileira**, Brasilia, DF, v.49, n.11, p.891-900, nov. 2014.

PAIVA, A.L.C.; TEIXEIRA, R. B.; YAMAKI, M.; MENEZES, G.R.O.; LEITE, C. D. S.; TORRES, R.A. Análise de componentes principais em características de produção de aves de postura. **Revista Brasileira de Zootecnia**, Viçosa, MG, v. 39, n. 2, p. 285-288, fev. 2010.

PEREIRA, A.V.; VENCOVSKY, R.; CRUZ, C.D. Selection of botanical and agronomical descriptors for the characterization of cassava (*Manihot esculenta* Crantz.) germplasm. **Revista Brasileira de Genética**, v.15, p.115-124, 1992.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2017. Disponível em: <http://www.R-project.org/>.

REIS, R. C.; DE SOUZA VIANA, E.; DE JESUS, J. L.; DANTAS, J. L. L.; LUCENA, R. S. Caracterização físico-química de frutos de novos híbridos e linhagens de mamoeiro. **Pesquisa Agropecuária Brasileira**, v. 50, n. 3, p. 210-217, 2015.

SANTANA, L. R. R.; MATSUURA, F. C. A. U.; CARDOSO, R. L. Genótipos melhorados de mamão (Carica papaya L.): avaliação sensorial e físico-química dos frutos. **Ciência e Tecnologia de Alimentos**, Campinas, v. 24, n. 2, p. 217-222, 2004.

SCHWEIGGERT, R. M.; STEINGASS, C. B.; ESQUIVEL, P.; CARLE, R. Chemical and morphological characterization of Costa Rican papaya (Carica papaya L.)

hybrids and lines with particular focus on their genuine carotenoid profiles. **Journal of agricultural and food chemistry**, v. 60, n. 10, p. 2577-2585, 2012.

SILVA, F.F. DA; PEREIRA, M.G.; RAMOS, H.C.C.; DAMASCENO JUNIOR, P.C.; PEREIRA, T.N.S.; IDE, C.D. Genotypic correlations of morpho-agronomic traits in papaya and implications for genetic breeding. **Crop Breeding and Applied Biotechnology**, v.7, p.345-352, 2007.

SILVA, F.F. da; PEREIRA, M.G.; RAMOS, H.C.C.; DAMASCENO JUNIOR, P.C.; PEREIRA, T.N.S.; GABRIEL, A.P.C.; VIANA, A.P.; DAHER, R.F.; FERREGUETTI, G.A. Estimation of genetic parameters related to morpho-agronomic and fruit quality traits of papaya. **Crop Breeding and Applied Biotechnology**, v.8, p.65-73, 2008.

SILVA, W. C.; de CARVALHO, S. I. C.; DUARTE, J. B. Identification of minimum descriptors for characterization of Capsicum spp. germplasm. **Horticultura Brasileira**, v.31, n. 2, p. 190- 202, 2013.

SILVA, R.S.; MOURA, E.F.; FARIAS-NETO, J.T.; LEDO, C.A.S.; SAMPAIO, J.E. Selection of morphoagronomic descriptors for the characterization of accessions of cassava of the Eastern Brazilian Amazon. **Genetics and Molecular Research**. v. 16, n. 2, 2017.

SILVA, M. DE S.; LEONEL, S.; SOUZA, J. M. A.; MODESTO, J. H.; FERREIRA, R. B.; BOLFARINE, A. C. B. Evaluation of papaya genotypes using agronomic descriptors and estimation of genetic parameters. **Bioscience Journal**, v. 34, n. 4, 8 Aug. 2018.

SINGH, D. **The relative importance of characters affecting genetic divergence**. The Indian Journal of Genetic and Plant Breeding, New Delhi, v. 41, p. 237-245, 1981.

STRAPASSON, E.; VENCOVSKY, R.; BATISTA, L. A. R. Seleção de descritores na caracterização de germoplasma de Paspalum sp. por meio de componentes principais. **Revista Brasileira de Zootecnia**, v. 29, n. 2, p. 373-381, 2000.

SUGIMURA, Y.; ITANO, M.; SALUD, C. D.; OTSUJI, K.; YAMAGUCHI, H. Biometric analysis on diversity of coconut palm: cultivar classification by botanical and agronomical traits. **Euphytica**, v. 98, n. 1-2, p. 29-35, 1997.

TAMHANE, A. C. & DUNLOP D. D. **Statistics and Data Analysis** – from elementary to, intermediate. Upper Saddle River: Prentice-Hall, 2000.

VIANA, E. D. S.; REIS, R. C.; DA SILVA, S. C. S.; DAS NEVES, T. T.; DE JESUS, J. L. Avaliação físico-química e sensorial de frutos de genótipos melhorados de mamoeiro1. **Pesquisa Agropecuária Tropical**, v. 45, n. 3, 2015.

VIEGAS, P. R. A. **Características químicas e físicas do mamão (Carica papaya L.) cultivares 'Sunrise Solo' e 'Formosa' relacionados ao ponto de colheita**. Dissertação de mestrado, Viçosa: UFV, 1992.

ZAMAN, W.; BISWAS, S. K.; HELALI, M. O. H.; IBRAHIM, M.; HASSAN, P. Physico-chemical composition of four papaya varieties grown at Rajshahi. **Journal of Biosciences**, v. 14, p. 83-86, 2006.

ARTIGO 2

COMPARAÇÃO DE ESTRATÉGIAS DE SELEÇÃO DE DESCRITORES QUALITATIVOS EM ACESSOS DE MAMÃO[2]

[2]Artigo a ser ajustado para posterior submissão ao Comitê Editorial do periódico científico Acta Scientiarum. Agronomy, em versão na língua inglesa.

COMPARAÇÃO DE ESTRATÉGIAS DE SELEÇÃO DE DESCRITORES QUALITATIVOS EM ACESSOS DE MAMÃO

Autor: Antonio Leandro da Silva Conceição
Orientador: Carlos Alberto da Silva Ledo

RESUMO: No processo de caracterização do germoplasma, a utilização de descritores morfológicos permitem facilmente a distinção dos fenótipos. Geralmente são descritores de alta herdabilidade que podem ser detectados visualmente sendo pouco influenciados pelo ambiente. Com isso, o objetivo desse trabalho foi a seleção de descritores qualitativos que permitam a melhor caracterização de 50 acessos de mamão pertencentes ao banco de germoplasma da Embrapa Mandioca e Fruticultura, sendo utilizados para isso, 19 descritores qualitativos. A seleção dos descritores qualitativos foi realizada por meio do nível de entropia dos descritores (H), proposto por Renyi e por meio da Análise Fatorial Exploratória, utilizando o método da análise paralela proposto por Horn e o critério autovalor > 1, proposto por Kaiser para determinação do número de fatores a serem retidos. Considerando o Nível de entropia, foram selecionados 47,37% dos descritores. Já na análise fatorial utilizando o critério da análise paralela, cinco fatores foram considerados como ideais na representação, selecionando 57,89%, no critério de Kaiser, sete fatores foram apresentados como ideais, com isso foram selecionados 52,63%. O resultado obtido pelo Nível de Entropia e por meio da Análise Fatorial Exploratória combinada ao critério de Kaiser apresentou maior consistência e representatividade. O descarte realizado por mais de um método possibilita uma visão mais ampla acerca da redução dos descritores, possibilitando avaliar a técnica mais adequada para reduzir a dimensionalidade do conjunto de dados e selecionar os mais representativos de acordo com a variabilidade presente, com a menor perda de informação possível.

Palavras-chave: Análise multivariada, nível de entropia, análise fatorial exploratória, *Carica papaya* L., rotação Varimax.

COMPARISON OF DESCRIPTOR SELECTION STRATEGIES QUALITATIVE IN PAPAYA ACCESSES

Author: Antonio Leandro da Silva Conceição
Advisor: Carlos Alberto da Silva Ledo

ABSTRACT: In the process of characterization of germplasm, the use of morphological descriptors easily allows the distinction of phenotypes. They are usually descriptors of high heritability that can be detected visually and are little influenced by the environment. Thus, the objective of this work was the selection of qualitative descriptors that allow the best characterization of 50 papaya accessions belonging to the Embrapa Mandioca and Fruticultura germplasm bank, using 19 qualitative descriptors. The selection of qualitative descriptors was performed through the entropy level of the descriptors (H) proposed by Renyi and through Exploratory Factor Analysis, using the parallel analysis method proposed by Horn and the eigenvalue criterion > 1 proposed by Kaiser for. determining the number of factors to be retained. Considering the Entropy Level, were selected of 47.37% the descriptors. In the factor analysis using the parallel analysis criterion, five factors were considered as ideal in the representation, selecting 57.89%, in the Kaiser criterion, seven factors were presented as ideal, with 52.63% were selected. The result obtained by Entropy Level and by Exploratory Factor Analysis combined with the Kaiser criterion showed greater consistency and representativeness. Discarding by more than one method allows a broader view of the reduction of descriptors, allowing the evaluation of the most appropriate technique to reduce the dimensionality of the data set and select the most representative according to the present variability, with the lowest loss of data of possible information.

Keywords: Multivariate analysis, entropy level, exploratory factor analysis, *Carica papaya* L., Varimax rotation.

INTRODUÇÃO

O mamão é um fruto amplamente cultivado e consumido, rico em vitaminas A, B1, B2 e C, niacina, riboflavina, ferro, cálcio, e fibras, sendo considerado um fruto altamente nutricional. Além da papaína compreender uma importante enzima proteolítica, que auxilia na digestão de alimentos ricos em proteínas, com diversas aplicações farmacêuticas (HUERTA-OCAMPO et al., 2012; ARAVIND et al., 2013).

Os principais estados produtores brasileiros de mamão são Espírito Santo, Rio Grande do Norte, Bahia, Ceará e Minas Gerais, respectivamente (IBGE, 2017), que fazem do Brasil um dos principais produtores mundiais de mamão (*Carica papaya* L.). Em 2017, a produção nacional foi de 1,5 milhão de toneladas, correspondendo a 11,6% da produção mundial de mamão (FAOSTAT, 2017).

Quando se utiliza um grande número de descritores, as técnicas multivariadas têm se tornado muito difundidas em várias áreas do conhecimento, devido a facilidade na realização das análises e interpretação. Dentre as áreas de grande aplicação atualmente estão: agronomia, zootecnia, ecologia, florestal, psicologia, etc. (HONGYU, 2018).

Os estudos de seleção de descritores morfoagronômicos são cruciais em programas de melhoramento e conservação, pois estes tem o objetivo de reduzir a dimensão dos dados originais em um conjunto de dados menor, sem perder informação considerável, ou seja, resultando em um conjunto mínimo de descritores que proporcione uma boa capacidade de discriminar indivíduos de determinada espécie, proporcionando também uma economia de tempo e dinheiro em avaliações. A grande maioria dos estudos de seleção de descritores qualitativos são realizados utilizando o nível de entropia dos descritores (H), proposto por Renyi (1961). O nível de entropia pode ser utilizado para quantificar a variabilidade presente em descritores qualitativos por meio da observação das frequências relativas das classes para cada descritor avaliado (LEDO et al. 2011).

Na literatura encontram-se vários estudos de avaliação e/ou seleção de descritores morfológicos em acessos de mamão (Ocampo et al. 2006; Coppens-d'Eeckenbrugge et al. 2007; Rieger (2009); Dias (2011); Aikpokpodion (2012); Brown et al. (2012); Oliveira et al. (2012); Dantas et al. (2013); Moore (2013); Nobre (2016); Octaviani, Hafsah e Hayati, 2018; Nishimwe (2019). Trabalhos recentes com seleção de descritores qualitativos envolvendo outras culturas, também são

encontrados, como os estudos de acessos de *Manihot esculenta Crantz* (Oliveira et al. (2014); Afonso et al. (2014) e Silva et al. (2017) *Capsicum baccatum* e *Capscicum chinense* (Padilha, Sosinski junior, e Barbieri, 2016); *Physalis angulata* L., (Silva et al. 2018) e *Mangifera indica* L. (Souza, 2018).

A análise fatorial exploratória (AFE) ou “exploratory factoranalysis” é uma técnica dentro da análise fatorial cujo objetivo abrangente é identificar as relações subjacentes entre os descritores avaliados. A AFE é uma técnica estatística que estuda correlações entre um grande número de descritores agrupando-os em fatores. Essa técnica permite a redução de dados, identificando os descritores mais representativos ou criando um novo conjunto de variáveis, bem menor que o original, ou seja, essa técnica tem como objetivo encontrar um meio de condensar a informação contida em diversos descritores (variáveis originais) em um conjunto menor de variáveis estatísticas (fatores) com uma perda mínima de informação (HAIR et al., 2009; KIRCH et al., 2017; HONGYU, 2018). Mesmo com o evidente potencial da análise fatorial exploratória em vários tipos de estudos, não foram encontrados trabalhos utilizando essa técnica visando a seleção de descritores qualitativos em acessos de mamão. Portanto, os resultados aqui obtidos são relevantes, pois constituem informações importantes que podem ser utilizadas em trabalhos futuros.

Este trabalho teve como objetivo reduzir a dimensão do conjunto original de descritores com menor perda de informação possível e descartar os que contribuem pouco para distinguir os acessos de mamoeiro avaliados, bem como comparar os métodos utilizados para seleção de descritores.

MATERIAIS E MÉTODOS

Foram utilizados 50 acessos pertencentes ao banco de germoplasma de mamão (BAG-Mamão) da Embrapa Mandioca e Fruticultura (Tabela 1). O plantio dos acessos foi realizado do dia 26 a 29 de agosto de 2014. As avaliações foram realizadas de outubro de 2014 a dezembro de 2015. Foi utilizado espaçamento de 3,0 m entre linhas e 2,0 m entre plantas, adotando-se as práticas culturais e os tratos fitossanitários preconizados para a cultura (Martins & Costa, 2003). As avaliações foram realizadas em Cruz das Almas, Bahia, Brasil (12º48'38"S e 39º6'26"W), na área experimental da Embrapa Mandioca e Fruticultura.

Tabela 1. Classificação por tipo e origem dos acessos de mamão avaliados, que compõem o Banco de Germoplasma (BAG-Mamão) da Embrapa Mandioca e Fruticultura. Cruz das Almas-BA, 2019.

Acesso	Tipo de fruto	Origem	Sigla
BGM 01	Formosa	Costa Rica	BGM 01 FC
BGM 02	Formosa	Taiwan	BGM 02 FT
BGM 03	Formosa	Havaí	BGM 03 FH
BGM 04	Solo	Havaí	BGM 04 SH
BGM 05	Solo	Havaí / Taiwan	BGM 05 SHT
BGM 06	Formosa	Malásia	BGM 06 FM
BGM 07	Formosa	Costa Rica	BGM 07 FC
BGM 08	Solo	Malásia	BGM 08 SM
BGM 09	Formosa	Malásia	BGM 09 FM
BGM 10	Formosa	Taiwan	BGM 10 FT
BGM 11	Formosa	Taiwan	BGM 11 FT
BGM 12	Formosa	Brasil	BGM 12 FB
BGM 13	Solo	Taiwan	BGM 13 ST
BGM 14	Formosa	Malásia	BGM 14 FM
BGM 15	Formosa	Taiwan	BGM 15 FT
BGM 16	Formosa	*	BGM 16 F
BGM 17	Formosa	Costa Rica	BGM 17 FC
BGM 18	Formosa	*	BGM 18 F
BGM 19	Formosa	Costa Rica	BGM 19 FC
BGM 20	Formosa	*	BGM 20 F
BGM 21	Solo	*	BGM 21 S
BGM 22	Formosa	Brasil	BGM 22 FB
BGM 23	Formosa	Brasil	BGM 23 FB
BGM 24	Formosa	Brasil	BGM 24 FB
BGM 25	Formosa	Brasil	BGM 25 FB
BGM 26	Formosa	Brasil	BGM 26 FB
BGM 27	Solo	Brasil	BGM 27 SB
BGM 28	Solo	Brasil	BGM 28 SB
BGM 29	Solo	Brasil	BGM 29 SB
BGM 30	Formosa	Havaí	BGM 30 FH
BGM 31	Formosa	Brasil	BGM 31 FB
BGM 32	Solo	Brasil	BGM 32 SB
BGM 33	Solo	Havaí	BGM 33 SH
BGM 34	Formosa	Havaí	BGM 34 FH
BGM 35	Solo	Brasil	BGM 35 SB
BGM 36	Formosa	Brasil	BGM 36 FB
BGM 37	Formosa	Brasil	BGM 37 FB
BGM 38	Solo	*	BGM 38 S
BGM 39	Solo	Brasil	BGM 39 SB
BGM 40	Formosa	*	BGM 40 F
BGM 41	Formosa	*	BGM 41 F
BGM 42	Solo	Havaí	BGM 42 SH
BGM 43	Solo	Havaí	BGM 43 SH
BGM 44	Solo	Havaí	BGM 44 SH
BGM 45	Formosa	África do Sul	BGM 45 FA
BGM 46	Formosa	África do Sul	BGM 46 FA
BGM 47	Solo	África do Sul	BGM 47 SA
BGM 48	Formosa	Brasil	BGM 48 FB
BGM 49	Formosa	*	BGM 49 F
BGM 50	Formosa	*	BGM 50 F

*Origem desconhecida.

A avaliação foi realizada baseada no Manual de Descritores para Mamão [Catálogo de Germoplasma de Mamão (*Carica papaya* L.)], adaptado pela Embrapa Mandioca e Fruticultura a partir dos descritores inicialmente estabelecidos pelo International Board for Plant Genetic Resources (IBPGR, 1988), atualmente Bioversity International, com algumas alterações sugeridas por Dantas et al. (2000).

Para o presente estudo foram utilizados 19 descritores qualitativos: cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), formato dos bordos foliares (FBF), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI), densidade de flores nas inflorescência (DFI), forma das folhas (FF), forma da cavidade do limbo (FCL), presença de pelos (Ppel), presença de cera (Pcer), coloração do pedúnculo da inflorescência (CPI), coloração dos lóbulos da corola (CLC), uniformidade de distribuição de frutos (UDF), coloração da casca do fruto imaturo (8 meses) (CCFI), tipo de florescimento (TF) e mudança de sexo da flor (MSF). Na tabela 2, são apresentados esses descritores e suas respectivas classes avaliadas.

Tabela 2. Relação dos descritores qualitativos de 50 acessos de mamão estudados. Cruz das Almas, BA. 2019.

Descritores qualitativos	Classes
Cor do caule	Esverdeada
	Cinza claro
	Cinza com manchas arroxeadas
	Arroxeadas
	Outras
Pigmentação do caule	Parte basal
	Parte mediana
	Parte superior
	Indiscriminada
Cor do pecíolo	Verde pálido
	Verde normal
	Verde escuro
	Verde com manchas arroxeadas
	Arroxeadas
	Outros
Forma das folhas	Forma 1
	Forma 2
	Forma 3
	Forma 4
	Forma 5
	Forma 6
	Forma 7
	Forma 8

Tabela 2. (Continuação)

Descritores qualitativos	Classes
Forma das folhas	Forma 9
	Forma 10
	Forma 11
	Forma 12
	Forma 13
	Forma 14
Forma dos bordos foliares	Reta
	Convexa
	Côncava
	Outras
Forma da cavidade do limbo	Aberta
	Levemente aberta
	Levemente fechada
	Fechada
	Outras
Presença de Pêlos	Presença de pelos
	Ausência de pelos
Presença de Cera	Presença de cera
	Ausência de cera
Coloração do pedúnculo da inflorescência	Esverdeado
	Púrpura
	Roxo
	Outros
Coloração dos lóbulos da corola	Branco
	Creme
	Amarelo
	Alaranjado
	Esverdeado
	Verde escuro
	Amarelo/Verde com manchas arroxeadas
	Vermelho arroxeado
	Outras
Coloração das flores hermafroditas	Branco
	Creme
	Amarelo
	Alaranjado
	Esverdeado
	Verde escuro
	Amarelo/Verde com manchas arroxeadas
	Vermelho arroxeado
	Outras
Uniformidade de distribuição de frutos	Uniforme
	Desuniforme
Coloração da casca do fruto imaturo (8 meses)	Amarelo a
	Amarelo b
	Amarelo c
	Amarelo d
	Laranja a
	Laranja b
	Laranja c
	Laranja d
	Verde a
	Verde b

Tabela 2. (Continuação)

Descritores qualitativos	Classes
Coloração da casca do fruto imaturo (8 meses)	Verde c
	Verde d
	Verde e
	Vermelho claro
	Vermelho
	Vermelho escuro
Formato dos frutos	Globular
	Arredondado (afilada)
	Altamente arredondado
	Elíptico
	Oval
	Oblongo
	Oblongo - elipsóide
	Oblongo - maciço
	Elongata
	Alongado - cilíndrico
	Forma de pera
	Forma de clava
	Forma de flor com extremidade cônica
	Oblongo com extremidade cônica
	Reniforme
	Forma de pião
	Forma de ameixa
	Alongado - afilado
	Vela
	Alongado – forma de pera
	Oblongo – forma de pera
	Oval - forma de pera
Tipo de hermafroditismo	Tipo 1
	Tipo 2
	Tipo 3
	Tipo 4
	Tipo 5
	Tipo 6
Tipo de florescimento	Flores isoladas
	Inflorescência
	Ambas
Densidade da inflorescência	Densa
	Média
	Esparsa
Densidade de flores na inflorescência	Densa
	Média
	Esparsa
Mudança de sexo da flor	MSF 1
	MSF 2
	MSF 3
	SMSF

Cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), formato dos bordos foliares (FBF), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI), densidade de flores nas inflorescência (DFI), forma das folhas (FF), forma da cavidade do limbo (FCL), presença de pelos (Ppel), presença de cera (Pcer), coloração do pedúnculo da inflorescência (CPI), coloração dos lóbulos da corola (CLC), uniformidade de distribuição de frutos (UDF), coloração da casca do fruto imaturo (8 meses) (CCFI), tipo de florescimento (TF) e mudança de sexo da flor (MSF).

Seleção dos descritores qualitativos foi realizada por duas estratégias distintas:

1 - Nível de entropia dos descritores (H), proposto por Renyi (1961)

2 - Análise Fatorial

1.1 **Nível de entropia dos descritores (H), proposto por Renyi (1961)**

A seleção dos descritores qualitativos realizada por meio do nível de entropia dos descritores (H), proposto por Renyi (1961), de acordo com o seguinte modelo:

$$H = -\sum_{i=1}^{s} p_i \ln p_i$$

Onde a Entropia é uma medida da frequência da distribuição de (n) acessos P = (p_1, p_2 ... p_s), sendo: $p_1 = f_i/n$ e ($p_1 + p_2 + ... + p_s = 1$) desde que ($n = f_1 + f_2 + ... + f_s$), onde f_1, f_2, ... f_n, são as contagens de cada uma das classes (s) no descritor considerado.

A entropia de um descritor qualquer será tão maior quanto maior for o número de classes fenotípicas desse e quanto mais homogêneo for o balanço entre a frequência dos acessos nas diferentes classes fenotípicas (VIERA et al., 2007). Neste trabalho valores baixos de H (<0,70) foram utilizados como critério de descarte.

2.1 Modelo teórico da análise fatorial

O modelo da análise de fatores pode ser escrito conforme Equação 1 (JOHNSON & WICHERN, 2007):

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + e_i \qquad (1)$$

em que X_i é o i-ésimo escore depois dele ter sido padronizado (média 0 e desvio-padrão 1); i = 1, ⋯ , p; p é o número de variáveis; a_{i1}, a_{i2}, ⋯ , a_{im} são as cargas dos fatores para o i-ésimo teste; F_1,F_2, ⋯ , F_m são m fatores comuns não correlacionados, cada um com média 0 e variância 1 e e_i é um fator específico

somente para o i-ésimo teste que é não correlacionado com qualquer dos fatores comuns e tem média zero.

Os *p* valores observados X_p são expressos em termos de *p* + *m* variáveis aleatórias não observáveis (F_1, F_2, ⋯ , F_m; ε_1, ε_2,⋯ , ε_p). Isso distingue o modelo fatorial do modelo de regressão múltipla, no qual as variáveis independentes podem ser observadas, e cujas posições são ocupadas por F no modelo fatorial. Matricialmente, o modelo é expresso pela Equação 2:

$$\boldsymbol{X}_{(p\times 1)} = \boldsymbol{\Lambda}_{(p\times m)}\,\mathbf{F}_{(m\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)} \qquad (2)$$

Já o modelo AF é expresso na Equação 3:

$$\begin{aligned} Var(X_j) &= a_j{}^2{}_1 Var(F_1) + a_j{}^2{}_2 Var(F_2) + \cdots + a_{jm}{}^2\, Var(F_m) + Var(e_j) \\ &= a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + Var(e_j) \end{aligned} \qquad (3)$$

em que $a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2$ é chamada a comunalidade de X_j (a parte da sua variância que é explicada pelos fatores comuns). A comunalidade não pode exceder 1, sendo necessário que fique compreendida entre −1 ≤ a_{ij} ≤ +1. Pode também ser estabelecido que a correlação entre X_j e $X_{j'}$ seja dada pela Equação 4:

$$r_{jj'} = a_{j1}a_{j'1} + a_{j2}a_{j'2} + \cdots + a_{jm}a_{j'm} \qquad (4)$$

Consequentemente, duas variáveis (descritores) somente serão altamente correlacionadas se elas tiverem altas cargas no mesmo fator (JOHNSON & WICHERN, 2007; NEISSE & HONGYU, 2016).

Para verificar se a aplicação da análise fatorial tem validade para as variáveis escolhidas, foi utilizado o critério de KaiserMeyer-Olkin (KMO). O KMO é calculado por meio do quadrado das correlações totais dividido pelo quadrado das correlações parciais, das variáveis analisadas, cuja expressão é dada na Equação 5 (DZIUBAN & SHIRKEY, 1974):

$$KMO = \frac{\sum_{j=1}^{p}\sum_{m=1,m\neq j}^{p} r_{jm}^2}{\sum_{j=1}^{p}\sum_{m=1,m\neq j}^{p} r_{jm}^2 + \sum_{j=1}^{p}\sum_{m=1,m\neq j}^{p} r_{pjm}^2}$$

em que $r_{jm}{}^2$ é o coeficiente de correlação linear entre *Xj* e *Xm*; $r_{pjm}{}^2$ é o coeficiente de correlação parcial amostral entre *Xj* e *Xm*, definido como sendo o coeficiente de correlação linear entre os resíduos.

2.2 Retenção de Fatores

Uma série de métodos tem sido sugeridos para determinar o número de fatores a serem retidos na análise fatorial exploratória (EFA). Porém, no presente estudo serão utilizadas dois dos mais utilizados. O primeiro critério é o de Kaiser-Guttman, sugerido por Guttman (1954) e adaptado e popularizado por Kaiser. (1960, 1961) e é comumente apresentado como "critério Guttman-Kaiser", também conhecido como critério do *eigenvalue* (autovalor) maior do que 1,0 (λ_i>1). O segundo critério é conhecido como o método das análises paralelas (AP) (HORN, 1965), que leva em consideração a proporção de variância resultante do erro amostral. Esse critério vem sendo cada vez mais consolidado na literatura internacional, com implementação no *software* R no pacote "*psych*" (HONGYU, 2018).

Após verificar se a base de dados está adequada e determinar a técnica de extração e o número dos fatores, o próximo passo foi a realização da rotação dos fatores. A rotação *varimax* é a mais bem avaliada e a mais utilizada nas pesquisas entre os métodos ortogonais (DAMÁSIO, 2012).

De acordo com Pallant (2007), o tipo de rotação ortogonal *Varimax* é o mais comumente utilizado, pois esse método procura minimizar o número de descritores que apresentam altas cargas em cada fator. Por esse motivo, esse trabalho utilizou esse tipo de rotação de fatores.

Todas as análises foram realizadas por meio de rotinas computacionais implementadas no software R 3.5.1 (R Development Core Team, 2017) com os pacotes "*corrplot*", "*vegan*", "*factoextra*", "*psych*", "*dplyr", "vegan" e "entropy".*

RESULTADOS E DISCUSSÃO

Na Tabela 3, estão apresentados os descritores qualitativos, suas classes fenotípicas, frequência percentual dos acessos em cada uma das classes e o nível de entropia de Renyi. Foram considerados como descritores descartados todos aqueles que apresentaram nível de entropia inferior a 0,70.

Os descritores descartados inicialmente por não serem capazes de diferir os acessos são apresentados na análise de entropia com uma frequência de 100%, ou seja, esses acessos ficaram concentrados na mesma classe dos descritores em questão, apresentando nível de entropia igual a zero: Presença de Pêlos (Ppel),

Presença de Cera (Pcer), Coloração da casca do fruto imaturo avaliado aos 8 meses (CCFI) e Tipo de florescimento (TF). Resultados semelhantes foram obtidos por Oliveira et al. (2012), no estudo da seleção de descritores em cultivares de mamão, onde os descritores Ppel e Pcer também apresentaram 100% da sua frequência em apenas uma classe. No estudo realizado por Aikpokpodion (2012), onde foram avaliados descritores morfológicos em acessos de mamão na Nigéria, a Coloração da casca do fruto imaturo (CCFI) também não apresentou variação considerável, concentrando sua frequência em apenas duas classes para esse descritor, sendo essas: verde (70%) e verde claro (30%) (Tabela 3).

Os descritores que apresentaram maiores valores estão relacionadas ao Formato dos frutos (Ffrut): (H=2,14), Cor do pecíolo (CP): (H=1,24), cor do caule (CC) (H=1,12) e Tipo de hermafroditismo (Ther): (H=1,06), em função de apresentarem elevado número de classes e um maior equilíbrio na proporção entre a frequência dos acessos nas diferentes classes fenotípicas. Isso revela variabilidade genética entre os acessos estudados (Tabela 3). De acordo com Dantas et al. (2013), o formato do fruto (Ffrut) é um descritor muito importante, pois está relacionado com a preferência do mercado interno e externo, onde esses preferem frutos de formato piriforme, característicos de plantas com flores hermafroditas, portanto, são frutos de maior valor comercial. Nos trabalhos de Asudi et al. (2010); Aikpokpodion (2012); Brown et al. (2012) e Nishimwe et al. (2019), também foram observadas uma grande variação para o descritor formato do fruto (Ffrut). Ainda no trabalho de Aikpokpodion (2012), foi observada alta distribuição da frequência percentual para os descritores, CP e CC. Esses resultados evidenciam a importância desses descritores para caracterização da variabilidade presente na cultura.

No estudo de Nobre (2016), foram encontrados valores semelhantes ao do presente trabalho para CC, DI, DFI e para tipo de hermafroditismo (Ther), essa última apresentando valor de entropia de (H=0,99) (Tabela 3). Estudos relacionadas à caracterização do sexo do mamoeiro são muito importantes, para que os produtores possam identificar precocemente as plantas hermafroditas, uma vez que, estas são as que apresentam maior relevância econômica dentre os sexos que o mamoeiro pode apresentar (TRINDADE e OLIVEIRA, (2000); DANTAS et al. 2013).

Para o descritor referente a coloração das flores, toda frequência ficou concentradas em três classes (Tabela 3). Resultados semelhantes foram observados por Aikpokpodion (2012), onde toda a frequência ficou concentrada em duas classes, amarelo (45%) e branco (55%).

Os descritores que apresentaram os menores valores em relação ao nível de entropia foram: Forma das folhas (FF): H=0,10, Mudança de sexo da flor (MSF): (H=0,10), Forma da cavidade do limbo (FCL): H=0,20, Uniformidade de distribuição de frutos (UDF): (H=0,28), Coloração dos lóbulos da corola (CLC): (H=0,33) e Coloração do pedúnculo da inflorescência (CPI): (H=0,33). Com isso, serão descartados por não apresentarem um nível aceitável de entropia que possa ser determinante para discriminação dos acessos em estudo (Tabela 3). No estudo realizado por Oliveira et al. (2012), os descritores FF e CPI também se apresentaram pouco informativos, com H=0,00 para ambos.

De acordo com Ledo et al. (2011), o nível de entropia pode ser utilizado para quantificar a variabilidade presente em descritores qualitativos por meio da observação das frequências relativas das classes para cada descritor avaliado. Desta forma, baixos valores para entropia estão associados a uma menor quantidade de classes fenotípicas para o descritor utilizado e a um maior desequilíbrio na proporção entre a frequência dos acessos nas diferentes classes fenotípicas.

Utilizando o nível de entropia foram selecionados 47,37% dos descritores qualitativos avaliados, no total de nove, sendo eles: cor do caule, pigmentação do caule, cor do pecíolo, formato dos bordos foliares, coloração das flores hermafroditas, formato dos frutos, tipo de hermafroditismo, densidade da inflorescência e densidade de flores nas inflorescência.

Os descritores descartados foram: forma das folhas, forma da cavidade do limbo, presença de pêlos, presença de cera, coloração do pedúnculo da inflorescência, coloração dos lóbulos da corola, uniformidade de distribuição de frutos, coloração da casca do fruto imaturo (8 meses), tipo de florescimento e mudança de sexo da flor. O descarte realizado possibilitará a redução no tempo, na mão-de-obra e nos custos das atividades de avaliação e melhor caracterização da cultura.

Tabela 03. Descritores qualitativos avaliados, classes fenotípicas, frequência percentual e nível de entropia de acessos de mamão do banco de germoplasma da Embrapa Mandioca e Fruticultura. Cruz das Almas-BA, 2019.

Descritores qualitativos	Siglas	Classes	Frequência percentual	Nível de Entropia
Cor do caule	CC	Esverdeada	8	1,12
		Cinza claro	28	
		Cin + manchas arroxe.	0	
		Arroxeadas	10	
		Outras	54	
Pigmentação do caule	PigC	Parte basal	32	0,78
		Parte mediana	4	
		Parte superior	0	
		Indiscriminada	64	
Cor do pecíolo	CP	Verde pálido	2	1,24
		Verde normal	4	
		Verde escuro	2	
		Verde + manchas arroxe	44	
		Arroxeadas	10	
		Outros	38	
Forma das folhas	FF	Forma 1	98	0,10
		Forma 2	0	
		Forma 3	0	
		Forma 4	0	
		Forma 5	0	
		Forma 6	0	
		Forma 7	0	
		Forma 8	0	
		Forma 9	2	
		Forma 10	0	
		Forma 11	0	
		Forma 12	0	
		Forma 13	0	
		Forma 14	0	
Forma dos bordos foliares	FBF	Reta	50	0,91
		Convexa	8	
		Côncava	42	
		Outras	0	
Forma da cavidade do limbo	FCL	Aberta	96	0,20
		Levemente aberta	2	
		Levemente fechada	0	
		Fechada	2	
		Outras	0	
Presença de Pêlos	Ppel	Presença de pelos	0	0,00
		Ausência de pelos	100	

Tabela 3. (Continuação)

Descritores qualitativos	Siglas	Classes	Frequência percentual	Nível de Entropia
Presença de Cera	Pcer	Presença de cera	0	0,00
		Ausência de cera	100	
Coloração do pedúnculo da inflorescência	CPI	Esverdeado	90	0,33
		Púrpura	10	
		Roxo	0	
		Outros	0	
Coloração dos lóbulos da corola	CLC	Branco	10	0,33
		Creme	0	
		Amarelo	90	
		Alaranjado	0	
		Esverdeado	0	
		Verde escuro	0	
		A/V + manchas arroxea.	0	
		Vermelho arroxeado	0	
		Outras	0	
Coloração das flores hermafroditas	Cfher	Branco	10	0,71
		Creme	0	
		Amarelo	76	
		Alaranjado	0	
		Esverdeado	0	
		Verde escuro	0	
		A/V + manchas arroxea.	14	
		Vermelho arroxeado	0	
		Outras	0	
Uniformidade de distribuição de frutos	UDF	Uniforme	8	0,28
		Desuniforme	92	
Coloração da casca do fruto imaturo (8 meses)	CCFI	Amarelo a	0	0,00
		Amarelo b	0	
		Amarelo c	0	
		Amarelo d	0	
		Laranja a	0	
		Laranja b	0	
		Laranja c	0	
		Laranja d	0	
		Verde a	0	
		Verde b	100	
		Verde c	0	
		Verde d	0	
		Verde e	0	
		Vermelho claro	0	
		Vermelho	0	
		Vermelho escuro	0	

Tabela 3. (Continuação)

Descritores qualitativos	Siglas	Classes	Frequência percentual	Nível de Entropia
Formato dos frutos	Ffrut	Globular	2	2,14
		Arredondado (afilada)	0	
		Altamente arredondado	0	
		Elíptico	12	
		Oval	0	
		Oblongo	0	
		Oblongo - elipsóide	4	
		Oblongo - maciço	0	
		Elongata	28	
		Alongado - cilíndrico	0	
		Forma de pera	14	
		Forma de clava	8	
		For de flor- extre. cônica	4	
		Oblongo- extre. cônica	0	
		Reniforme	0	
		Forma de pião	0	
		Forma de ameixa	12	
		Alongado - afilado	0	
		Vela	4	
		Alongado – for. de pera	0	
		Oblongo – for. de pera	8	
		Oval - forma de pera	4	
Tipo de hermafroditismo	Ther	Tipo 1	66	1,06
		Tipo 2	16	
		Tipo 3	4	
		Tipo 4	0	
		Tipo 5	4	
		Tipo 6	10	
Tipo de florescimento	TF	Flores isoladas	0	0,00
		Inflorescência	100	
		Ambas	0	
Densidade da inflorescência	DI	Densa	10	0,93
		Média	54	
		Esparsa	36	
Densidade de flores na inflorescência	DFI	Densa	20	0,97
		Média	58	
		Esparsa	22	
Mudança de sexo da flor	MSF	MSF 1	0	0,10
		MSF 2	0	
		MSF 3	2	
		SMSF	98	

Análise Fatorial Exploratória (AFE) aplicada na seleção de descritores qualitativos

Na aplicação da AFE não foram considerados para análise os descritores que apresentaram nível de entropia igual a zero: Presença de Pêlos (Ppel), Presença de Cera (Pcer), Coloração da casca do fruto imaturo avaliado aos 8 meses (CCFI) e Tipo de florescimento (TF), pois estes não são capazes de diferir os acessos estudados.

Número de fatores retidos (Critério da análise paralela e Critério de Kaiser)

Na Figura 1, é apresentado o Screeplot que demonstra os autovalores dos componentes principais e da análise fatorial. Observa-se que, pelo método das análises paralelas foram indicados cinco fatores como ideais para serem extraídos para continuação da análise fatorial (AF).

Já de acordo com o critério de Kaiser (Figura 1), o ScreePlot, permitiu a observação da ordem dos autovalores e a dispersão dos componentes. Nele foi verificado que a maior porcentagem da variação total foi explicada pelos sete primeiros fatores, com isso, o critério de Kaiser sugere que deve-se extrair sete fatores: o primeiro apresenta um autovalor de 2,35, carregando 15,69% da variância, o segundo fator apresenta eigenvalue de 2,08, carregando de 13,86% da variância, o eigenvalue com suas respectivas variâncias do terceiro ao sétimo são apresentadas a seguir, com 1,79 (11,95%), 1,53 (10,22%) 1,41 (9,41%), 1,11 (7,38) e 1,03 (8,88%). Em conjunto, esses sete fatores explicam 75,41% da variância das variáveis originais.

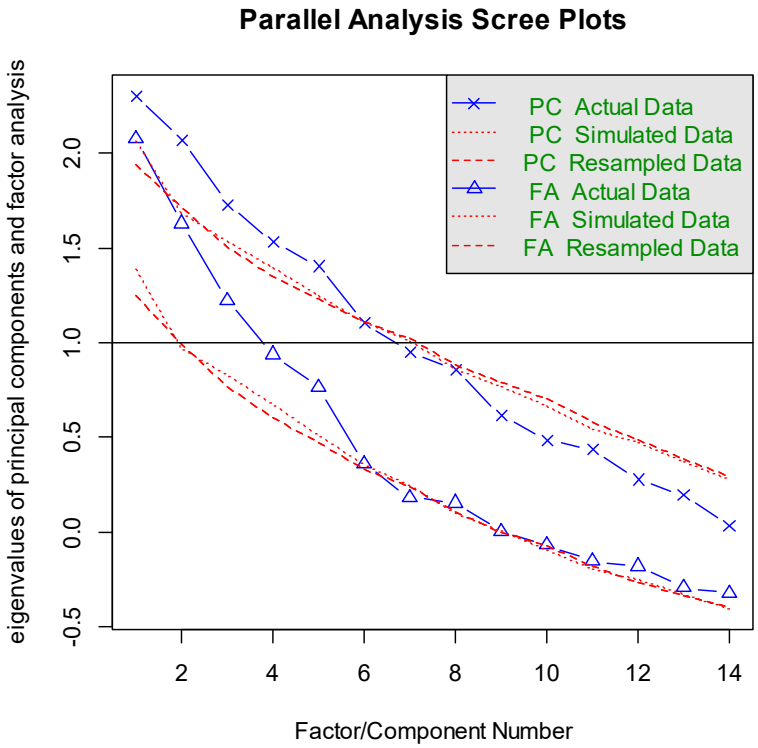


Figura 1. Analise paralela e Critério de Kaiser aplicados para determinar o número de fatores a serem utilizados na Análise Fatorial Exploratória. Componentes principais (PC); Análise fatorial (FA).

Na Tabela 4 e 5 também são apresentados os valores das cargas fatoriais rotacionadas pelo método Varimax para os 5 fatores retidos pela AP e para os sete fatores retidos pelo critério de Kaiser em relação a cada descritor avaliado e as suas comunalidades. De acordo com Pallant (2007), o tipo de rotação Varimax é o mais comumente utilizado, pois esse método procura minimizar o número de descritores que apresentam altas cargas em cada fator. E por ser um tipo de rotação ortogonal, essas são mais fáceis de reportar e de interpretar.

Os resultados das variâncias totais encontradas no presente estudo foram de 49% e 64% para o critério da análise paralela (AP) e o critério de Kaiser, respectivamente. Constata-se uma boa representatividade da variabilidade dos dados para ambos os critérios. Contudo, utilizando o Critério de Kaiser, a proporção da variância total explicada é 15% superior à obtida no critério da analise paralela (Tabela 4 e 5). De acordo com Hair et al. (2006) um patamar de 60% é sugerido como boa representatividade.

As cargas fatoriais dos descritores relacionadas a seus respectivos fatores retidos pelo método da análise paralela são apresentadas a seguir: Fator 1 (AP1),

foram apresentadas pelos descritores forma das folhas (FF), com valor de 0,99≅1 e forma da cavidade do limbo (FCL) com carga de 0,94, esses dois descritores são características relacionadas a folha. Já para o fator 2 (AP2), os descritores Cfher (0,88) e CPI (0,68) foram os que apresentaram as maiores cargas, esses dois descritores são ligados a coloração da flor. Com relação ao fator 3 (AP3), foram os descritores DI (0,99≅1) e DFI (0,71) que apresentaram as maiores cargas, características estas, ligadas a inflorescência (Tabela 4 e 5).

No fator 4 (AP4) os descritores com maior carga foram PigC (-0,59), CLC (0,56) e Ffrut (-0,50), esses descritores possuem baixas cargas fatoriais, a baixa comunalidade entre esse grupo de descritores é outro indício de que eles podem estar fracamente relacionados e pouco explicados pelo seu respectivo fator. Sendo assim, esse fator pode ser considerado pouco consistente e representativo. E por fim, o fator 5 (AP5), tendo os descritores Ther (-0,71) e MSF (0,50) com as maiores cargas, esses descritores estão ligados ao sexo da planta (Tabela 4 e Figura 2). De acordo com o método da análise paralela, os descritores CP, UDF e CC por não apresentarem relação com nenhum dos fatores retidos e o descritor FBF por apresentar carga inferior a 0,5 podem ser descartados.

A comunalidade, que é a percentagem de variabilidade (h^2) explicada de cada um desses descritores quando agrupado nesses fatores, os maiores valores de h^2 observados quando utilizado a analise paralela (AP) foram de 0,99 para FF, Ther e Cfher e 0,90 para FCL (Tabela 4). As comunalidades representam a proporção da variância para cada descritor incluído na análise que é explicada pelos componentes extraídos (SCHAWB, 2007). Usualmente o valor mínimo aceitável para comunalidade é de 0,50 (FIGUEIREDO FILHO e SILVA JÚNIOR, 2010).

Os fatores retidos pelo critério de Kaiser foram divididos e nomeados da seguinte forma: descritores relacionados a folha fator 1 (K1), descritores relacionados a cor da flores e inflorescência fator 2 (K2), o fator 3 (K3), contempla descritores ligados a densidade de flores e inflorescência, o fator 4 (K4) apresenta descritores ligados ao sexo da flor, no fator 5 (K5) esse fator caracteriza a coloração do caule, no fator 6 (F6) temos o descritor ligado ao pecíolo. No fator 7 (K7) apresenta o descritor relacionado ao fruto. As maiores cargas apresentadas para AF dentro dos seus respectivos fatores quando utilizou esse critério foram para os descritores FF=0,99 ≅1 (K1); Cfher=0,99 ≅1 (k2); DI=0,99≅1 (K3); CP=0,99(K6); Ther=0,95 (K4); PigC=0,84 (K5) e Ffrut=0,97 ≅ 1(K7), por possuírem as maiores

cargas, esses descritores tem maior efeito nos seus componentes. Os maiores valores de h² observados foram apresentados pelos descritores CP, FF, Cfher, Ffrut, DI, Ther e FCL com valores acima de 0,90 (Tabela 5). Por meio desse critério os descritores CLC, UDF, MSF, CC e FBF podem ser descartados, pois não apresentam relação com nenhum dos fatores retidos, ou apresentaram carga inferior a 0,5. Esses descritores também apresentaram comunalidade muito baixa (Tabela 5). Valores muito baixos de comunalidade entre um grupo de descritores é um indício de que eles não estão linearmente correlacionados e, por isso, não devem ser incluídos na análise fatorial (FIGUEIREDO FILHO; SILVA JUNIOR, 2010).

Apesar do método das análises paralelas (AP) ser considerado um procedimento adequado para determinar o número de fatores a serem retidos (GLORFELD, 1995; PATIL et al., 2008; LORENZO-SEVA, TIMMERMAN, e KIERS, 2011). Observa-se que os descritores qualitativos foram melhor distribuídos ao longo dos sete fatores retidos seguindo o critério de Kaiser, uma vez que a nomeação de cada um dos mesmos foi realizada de maneira mais clara e concisa, possuindo uma representação mais satisfatória com uma maior relação dos descritores presentes e seus respectivos fatores, pois foram observadas cargas fatoriais e valores de comunalidade(h²) mais altos para grande maioria dos descritores incluídos na análise, e com isso sendo melhor explicados pelos fatores extraídos. O critério de Kaiser, também apresentou uma maior proporção da variância total explicada (64%), sendo cerca de 15% superior à obtida no critério da análise paralela (Tabela 4 e 5).

O número de 50 acessos e 15 descritores qualitativos avaliados no estudo se mostrou adequado para utilização da AFE direcionada a seleção de descritores. Reis (1997) e Hair et al. (1998) sugerem que o número de observações deva ser de no mínimo 5 vezes o número de descritores. Hair et al. (1998) enfatiza que ela não deve ser utilizada em amostras inferiores a 50 observações. Apesar do valor do Kaiser-Meyer-Olkin KMO ser usualmente empregado para averiguar a adequação dos dados a aplicação da AFE, onde alguns autores relatam que o valor acima de 0,5 indicam que a análise fatorial é adequada ao estudo. No presente estudo o valor de KMO foi de 0,48, estando aproximadamente dentro da medida de adequacidade de amostra de Kaiser-Meyer-Olkin (KMO).

Adaptações de técnicas para estudos dos quais não foram originalmente criadas são relevantes, pois podem resultar em bons resultados, otimizando assim o processo de análise. O próprio nível de entropia (H) de Rényi que foi adaptado com muito sucesso em estudos de seleção de descritores qualitativos, é uma extensão da entropia de Shannon. O índice de Shannon (Shannon, 1948) foi criado no contexto de Teoria da Informação e se expandiu posteriormente para outros domínios da engenharia, informática, estatística, economia e adaptado para diversidade biológica. Outro exemplo de adaptação de técnicas, é o método da análise paralela (AP) que Inicialmente foi desenvolvido para ser utilizado como critério de retenção de componentes. Entretanto, tem sido adaptado para o uso no contexto das AFEs (VELICER, EATON, e FAVA, 2000; CRAWFORD et al., 2010).

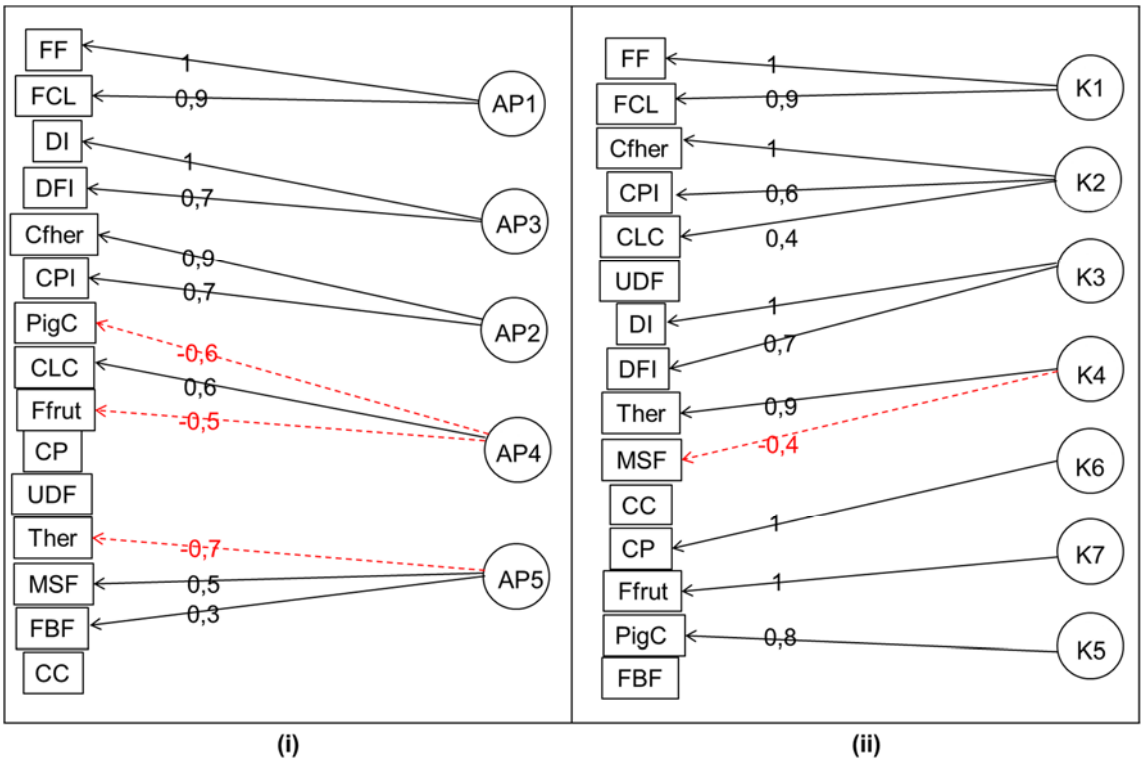


Figura 2. Diagramas com os descritores qualitativos que compõem os fatores definidos pela análise paralela (i) e pelo critério de Kaiser (ii) e suas respectivas cargas fatoriais.

Tabela 4. Cargas fatoriais rotacionadas pelo método Varimax e comunalidade para cada descritor qualitativo com número de cargas definidas pela análise paralela.

Descritores	AP1	AP3	AP2	AP4	AP5	h²
CC	-0,14	-0,01	0,1	0,18	-0,22	0,111
PigC	-0,19	0,24	0,38	-0,59	-0,22	0,641
CP	-0,09	0,01	0,03	0,21	0,01	0,053
FF	0,99	-0,11	-0,04	-0,06	0,05	0,999
FBF	0,04	0,16	0,04	0,14	0,33	0,156
FCL	0,94	-0,04	0,09	-0,03	0,06	0,904
CPI	0,01	0,05	0,68	-0,05	0,04	0,476
CLC	0,1	0,09	0,22	0,56	-0,09	0,388
Cfher	0,02	0,04	0,88	0,49	0,09	0,992
UDF	0,05	-0,11	0,12	0,18	0,04	0,063
Ffrut	0,09	-0,03	0,05	-0,46	0,17	0,247
Ther	-0,01	0,18	-0,08	0,07	-0,71	0,55
DI	0,05	0,99	0,01	0,04	0,08	0,995
DFI	-0,15	0,71	0,08	-0,03	-0,04	0,53
MSF	0	0,02	0,01	-0,07	0,5	0,252
Soma do quadrado das cargas	1,98	1,64	1,49	1,26	1,02	
Proporção da Variância	0,13	0,11	0,1	0,08	0,07	
Proporção da Variância acumulada	0,13	0,24	0,34	0,42	0,49	

Cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), formato dos bordos foliares (FBF), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI), densidade de flores nas inflorescência (DFI), forma das folhas (FF), forma da cavidade do limbo (FCL), coloração do pedúnculo da inflorescência (CPI), coloração dos lóbulos da corola (CLC), uniformidade de distribuição de frutos (UDF) e mudança de sexo da flor (MSF).

Tabela 5. Cargas fatoriais rotacionadas pelo método *Varimax* e comunalidade (h²) para cada descritor qualitativo com número de cargas definidas pelo critério de Kaiser.

Descritores	K1	K2	K3	K4	K6	K7	K5	h²
CC	-0,17	0,15	-0,02	0,23	0,03	-0,04	-0,08	0,112
PigC	-0,11	0,13	0,23	0,08	-0,04	0,17	0,84	0,82
CP	-0,05	0,03	0,03	0,04	0,99	-0,04	-0,06	0,997
FF	0,99	-0,03	-0,1	-0,04	-0,03	0,04	-0,09	0,998
FBF	0,01	0,12	0,19	-0,23	0	0,11	-0,26	0,179
FCL	0,95	0,09	-0,03	-0,05	-0,03	0,02	-0,06	0,914
CPI	0,01	0,63	0,06	-0,05	-0,08	0,09	0,19	0,451
CLC	0,04	0,4	0,09	0,17	0,04	-0,16	-0,36	0,354
Cfher	-0,01	0,99	0,05	-0,03	0,16	-0,11	-0,09	0,992
UDF	0,04	0,18	-0,11	-0,05	-0,08	-0,16	-0,07	0,088
Ffrut	0,09	-0,05	-0,05	-0,16	-0,09	0,97	0,11	0,999
Ther	-0,03	-0,06	0,15	0,95	-0,17	0,04	-0,03	0,957
DI	0,04	0,03	0,99	-0,03	-0,06	0	-0,04	0,988
DFI	-0,14	0,04	0,71	0,04	0,08	-0,01	0,13	0,544
MSF	-0,01	0,02	0,05	-0,42	-0,13	0,07	-0,1	0,213
Soma do quadrado das cargas	1,95	1,63	1,63	1,25	1,1	1,07	1,01	
Proporção da Variância	0,13	0,11	0,11	0,08	0,07	0,07	0,07	
Proporção da Variância acumulada	0,13	0,24	0,35	0,43	0,5	0,58	0,64	

Cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), formato dos bordos foliares (FBF), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI), densidade de flores nas inflorescência (DFI), forma das folhas (FF), forma da cavidade do limbo (FCL), coloração do pedúnculo da inflorescência (CPI), coloração dos lóbulos da corola (CLC), uniformidade de distribuição de frutos (UDF) e mudança de sexo da flor (MSF).

Na Figura 3, pode-se observar de maneira mais clara os resultados das técnicas de seleção aplicadas no presente trabalho. Onde foram descartados por meio do nível de entropia (H), os descritores FF, FCL, CPI, CLC, UDF e MSF que apresentaram valor de H<0,70. Os descritores selecionados que apresentaram valores de H≥70, foram CC, PigC, CP, FBF, Cfher, Ffrut, Ther, DI e DFI. Os descritores descartados pelo método da análise paralela foram CP, UDF, CC e FBF. Já pelo critério de Kaiser, os descritores CLC, UDF, MSF, CC e FBF foram indicados para descarte.

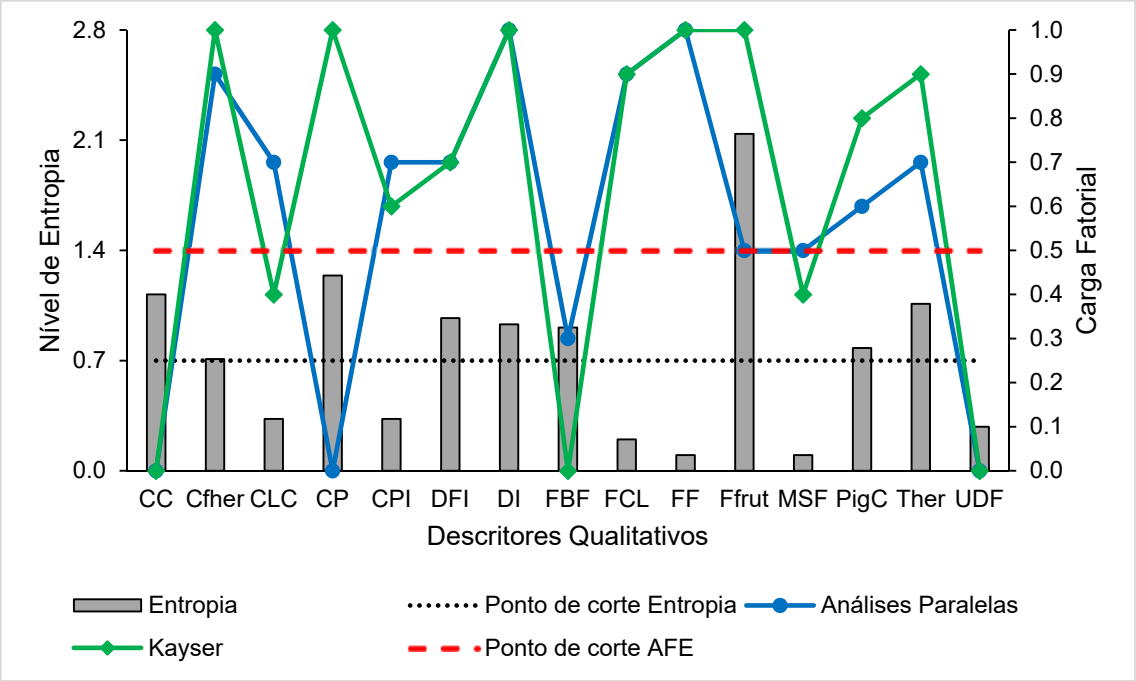


Figura 3: Apresentação gráfica resumida dos resultados do descarte de descritores qualitativos utilizando o Nível de Entropia e a análise fatorial exploratória (Método da análise paralela e critério de Kaiser).

Dos quatro descritores (Ffrut, CP, CC e Ther) com maior distribuição da frequência (%) de suas respectivas classes entre os 50 acessos de mamão avaliados, ou seja, os que apresentaram os mais altos valores para o nível de entropia (H), três destes (Ffrut, CP e Ther) foram concordantes quando utilizada a AFE com o critério de Kaiser, apresentando valores altos para comunalidade (h^2), além de estarem fortemente correlacionados com seus respectivos fatores (Figura 4). Estes métodos além de apresentarem maior concordância, foram mais consistentes e representativos.

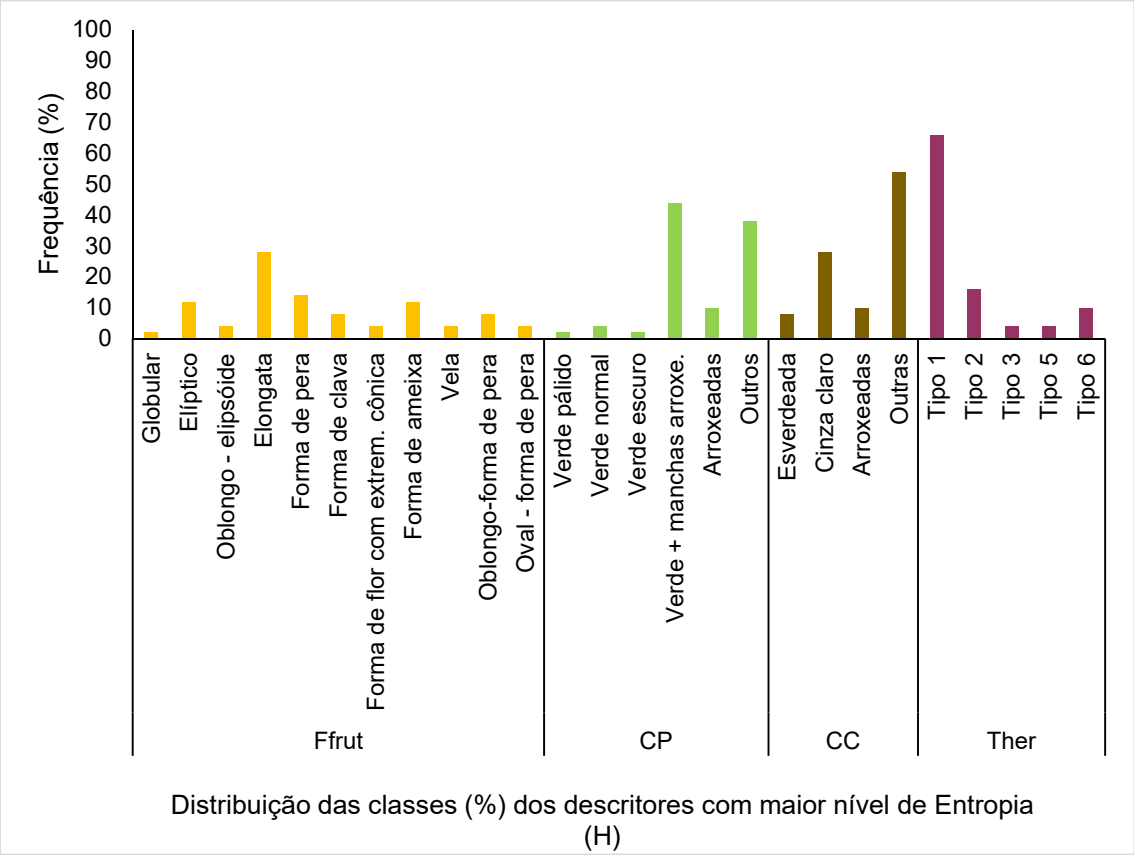


Figura 4: Gráfico dos descritores e suas respectivas classes que apresentaram maior distribuição da frequência percentual.

Para o descritor Ffrut, que apresentou a maior variabilidade no presente trabalho, foi observada a seguinte distribuição nas classes avaliadas: Elongata (28%), Forma de pera (14%), Globular (2%), Elíptico, (12%), Oblongo – elipsoide (4%), Forma de clava (8%), For de flor- extremamente cônica (4%), Forma de ameixa (12%), Vela (4%), Oblongo – forma de pera (8%) e Oval - forma de pera (4%). Valores próximos para as classes com maior frequência avaliadas do Ffrut no presente estudo, também foram observadas no trabalho de Aikpokpodion (2012), com as seguintes frequências percentuais: Elongata (20%), elíptico (16,7%), forma de pera (10%) e forma de ameixa (8%) onde foram avaliados descritores morfológicos em acessos de mamão na Nigéria. Nishimwe et al. (2019) no estudo da morfologia de linhas híbridas de mamão recém-desenvolvidas no Quênia, encontrou também grande variação para o Ffrut dentro das linhas e entre as linhas hibridas: Linha 1 (56,7% de frutas com formato oval, 26,7% arredondados, 6,7 elípticas, 3,3% de globulares e 3,3% altamente arredondadas), Linha 2

(Oblongo – maciço 57,6%, 20% elongata 16,7% elípticos e 3,3% globulares), Linha 3 (70% oblongo maciço e 13,3% alongado). As Linhas 4, 5, 6, 7 e 8 variaram muito, mas foram divididas em apenas em duas formas: Linha 4 (70% forma de pera e 30% elípticos), Linha 5 (63.3%em forma de pera e 36,7 alongados), Linha 6 (50% Oblongo e 50% Oblongo elipsóide) e Linha 7 (73,3% alongados e 26,7% em forma de pera) e Linha 8 (46,7% alongados e 53,3% elípticos).

Esses descritores são de grande importância, principalmente o formato do fruto (Ffrut), que além de ser bastante relevante para discriminação dos acessos estudados, está relacionado comercialmente com a preferência do mercado interno e externo. Segundo Barrett et al. (2010); Zhou et al. (2014) e Dantas et al. (2013), o formato do fruto é uma dos principais características que determinam o preço de mercado e as notas para exportações de frutas.

CONCLUSÕES

A análise fatorial exploratória foi eficiente para classificar os descritores qualitativos em um grupo de fatores, como também em reduzir os dados em um conjunto menor de descritores sem perder muita informação.

A análise fatorial exploratória aplicada a seleção de descritores qualitativos em acessos de mamão, mostrou que a quantidade de grupos de descritores foi diferente para cada tipo de critério utilizado para retenção do número de fatores, na análise paralela foram formados 5 grupos de variáveis estatísticas e no critério de Kaiser, 7 grupos.

Os descritores selecionados pelo nível de entropia e pela análise fatorial exploratória utilizando o critério de Kaiser, apresentam resultados mais consistentes e representativos.

Considerando os descritores inicialmente descartados (Ppel, Pcer, CCFI e TF), e os descartados em cada procedimento, a porcentagem de descritores remanescentes foi de: 47,37% dos descritores selecionados pelo nível de entropia (CC, Cfher, CP, DFI, DI, FBF, Ther, PigC e Ffrut). Já na análise fatorial utilizando o método da análise paralela, foram selecionandos 57,89% (Cfher, CLC, CPI, DFI, DI, FLC, FF, Ther, MSF, PigC e Ffrut), no critério de Kaiser selecionados 52,63% (Cfher, CP, CPI, DFI, DI, FCL, FF, Ther, PigC e Ffrut) dos descritores.

AGRADECIMENTOS

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa durante todo o período de realização deste doutorado e a UFRB e Embrapa Mandioca e Fruticultura, que disponibilizaram estrutura física e equipamentos adequados para execução do presente trabalho.

REFERÊNCIAS

AFONSO, S.D.J.; LEDO, C.A. da S.; MOREIRA, R.F.C.; SILVA, S. de O. e; LEAL, V.D. de J.; CONCEIÇÃO, A.L. da S. Selection of descriptors in a morphological characteristics considered in cassava accessions by means of multivariate techniques. **Journal of Agriculture and Veterinary Science**, v.7, p.13-20, 2014.

AIKPOKPODION, P. O. Assessment of genetic diversity in horticultural and morphological traits among papaya (Carica papaya) accessions in Nigeria. **Fruits**, v. 67, n. 3, p. 173-187, 2012.

ARAVIND, G.; BHOWMIK, D.; DURAIVEL, S.; HARISH, G. Traditional and medicinal uses of Carica papaya. **Journal of Medicinal Plants Studies**, v. 1, n. 1, p. 7-15, 2013.

ASUDI, G. O.; OMBWARA, F. K.; RIMBERIA, F. K.; NYENDE, A. B.; ATEKA, E. M.; WAMOCHO, L. S.; ONYANGO, A. Morphological diversity of Kenyan papaya germplasm. **African Journal of Biotechnology**, v. 9, n. 51, p. 8754-8762, 2010.

BARRERA-PACHECO, A.; MENDOZA-HERNÁNDEZ, G.; DE LEÓN-RODRÍGUEZ, A.; DE LA ROSA, A. P. B. Proteomic analysis of differentially accumulated proteins during ripening and in response to 1-MCP in papaya fruit. **Journal of proteomics**, v. 75, n. 7, p. 2160-2169, 2012.

BARRETT, D. M.; BEAULIEU, J. C.; SHEWFELT, R. Color, flavor, texture, and nutritional quality of fresh-cut fruits and vegetables: desirable levels, instrumental

and sensory measurement, and the effects of processing. **Critical reviews in food science and nutrition**, v. 50, n. 5, p. 369-389, 2010.

BROWN, J. E.; BAUMAN, J. M.; LAWRIE, J. F.; ROCHA, O. J.; MOORE, R. C. The structure of morphological and genetic diversity in natural populations of Carica papaya (Caricaceae) in Costa Rica. **Biotropica**, v. 44, n. 2, p. 179-188, 2012.

COPPENS-D'EECKENBRUGGE, G.; RESTREPO, M. T.; JIMÉNEZ, D.; MORA-NEWCOMER, E. Morphological and isozyme characterization of common papaya in Costa Rica. Caracterización morfológica y mediante isoenzimas de la papaya común en Costa Rica. **Acta Horticulturae.**, n. 740, p. 109-120, 2007..

CRAWFORD, A. V.; GREEN, S. B.; LEVY, R.; LO, W. J.; SCOTT, L.; SVETINA, D.; THOMPSON, M. S. Evaluation of parallel analysis methods for determining the number of factors. **Educational and Psychological Measurement**, v. 70, n. 6, p. 885-901, 2010.

DAMÁSIO, B. F. Uso da análise fatorial exploratória em psicologia. Avaliação Psicológica, v.11, n.2, p. 213-228, 2012.

DANTAS, J. L. L.; PINTO, R. M. S.; LIMA, J.L.; FERREIRA, F. R. **Catálogo de germoplasma de mamão** (*Carica papaya* L.). Cruz das Almas-BA: Embrapa Mandioca e Fruticultura, (Embrapa Mandioca e Fruticultura, Documentos, 94), p. 40, 2000.

DANTAS, J.L.L.; JUNGHANS, D.T.; LIMA, J.F. **Mamão: o produtor pergunta, a Embrapa responde**. 2.ed. Brasília, p.176, 2013.

DIAS, N. L. P. Caracterização morfoagronômica de genótipos de mamoeiro (Carica papaya L.) e seleção de descritores visando a proteção de cultivares. Dissertação (mestrado em Recursos Genéticos Vegetais), **Universidade Federal do Recôncavo da Bahia**, p.39, 2011.

DZIUBAN, C.D. SHIRKEY, E, S. **When is a correlation matrix appropriate for factor analysis? Some decision rules**. Psychological Bulletin, 81(6), 358-361, 1974.

FAOSTAT - Organização de Alimentos e Agricultura Corporate Statistical Database. Disponível em: <Disponível em: http://faostat3.fao.org/browse/Q/QC/E >. Acessado em 19 de dezembro de 2017.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. D. Visão além do alcance: uma introdução à análise fatorial. **Opinião pública**, v. 16, n. 1, p. 160-185, 2010.

GLORFELD, L. W. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. **Educational and psychological measurement**, v. 55, n. 3, p. 377-393, 1995.

GUTTMAN, L. Some necessary conditions for common-factor analysis. **Psychometrika**, v. 19, n. 2, p. 149-161, 1954.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; WILLIAM, C. **Multivariate Data Analysis**, 5. ed., New Jersey: Prentice Hall, 1998.

HAIR, Jr; BLACK, W. C; BABIN, B. J; ANDERSON, R. E e TATHAM, R. L. **Multivariate Data Analysis**. 6ª edição. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. Bookman Editora, 2009.

HORN, J. L. A rationale and test for the number of factors in factor analysis. **Psychometrika**, v. 30, n. 2, p. 179-185, 1965.

HONGYU, K.; SANDANIELO, V. L. M.; DE OLIVEIRA JUNIOR, G. J. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, v. 5, n. 1, p. 83-90, 2016.

HONGYU, Kuang. Análise Fatorial Exploratória: resumo teórico, aplicação e interpretação. **E&S Engineering and Science**, v. 7, n. 4, p. 88-103, 2018.

HUERTA-OCAMPO, J. Á.; OSUNA-CASTRO, J. A.; LINO-LÓPEZ, G. J.; LORENZO-SEVA, U.; TIMMERMAN, M. E.; KIERS, H. A. The Hull method for selecting the number of common factors. **Multivariate behavioral research**, v. 46, n. 2, p. 340-364, 2011.

IBGE. **Censo agro 2017,** 2017. Maiores produções de Mamão // Brasil. Disponível em : < https://censos.ibge.gov.br/agro/2017/templates/censo_agro/resultadosagro/agricultura.html?localidade=0&tema=76343>. Acesso em 26/01/2019.

IBPGR - INTERNATIONAL BOARD FOR PLANT GENETIC RESOURCES. **Descriptor list for papaya**. Rome: IPGRI, 1988. Disponivel em: <http://pdf.usaid.gov/pdf_docs/PNABC145.pdf>.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 6ª Edição. Madison: Prentice Hall International, p. 816 2007.

KAISER, H. F. The application of electronic computers to factor analysis. **Educational and psychological measurement**, v. 20, n. 1, p. 141-151, 1960.

KAISER, H. F. A note on guttman's lower bound for the number of common factors 1. **British Journal of Statistical Psychology**, v. 14, n. 1, p. 1-2, 1961.

KIRCH, J. L.; HONGYU, K.; SILVA, F. L.; DIAS, C. T. S. Análise Fatorial para Avaliação dos Questionários de Satisfação do Curso de Estatística de uma Instituição Federal. **E&S Engineering and Science**, v.6, n.1, 2017.

LEDO, C. A. S.; ALVES, A.; da SILVEIRA, T. C.; de OLIVEIRA, M. M.; SANTOS, A.; TAVARES FILHO, L. D. Q. **Caracterização morfológica da coleção de espécies silvestres de Manihot (Euphorbiaceae – Magnoliophyta) da Embrapa**

Mandioca e Fruticultura. Cruz das Almas-BA**:** Embrapa Mandioca e Fruticultura (Boletim de Pesquisa e Desenvolvimento), 2011.

LORENZO-SEVA, U.; TIMMERMAN, M. E.; KIERS, H. A. The Hull method for selecting the number of common factors. **Multivariate behavioral research**, v. 46, n. 2, p. 340-364, 2011.

MARTINS, D. S.; COSTA, A. F. S. **A cultura do mamoeiro:** tecnologias de produção.Vitória: Incaper, p. 497, 2003

MOORE, P. H. Phenotypic and genetic diversity of papaya. In: **Genetics and Genomics of Papaya**. Springer, New York, NY, 2014. p. 35-45.

NEISSE, A. C.; HONGYU, K. Aplicação de componentes principais e análise fatorial a dados criminais de 26 estados dos EUA. **E&S Engineering and Science**, v. 5, n. 2, p. 105-115, 2016.

NISHIMWE, G. Characterization of Morphological and Quality Characteristics of New Papaya (Carica papaya L) Hybrids Developed at JKUAT. Tese de Doutorado. **JKUAT-AGRICULTURE**, 2019.

NISHIMWE, G.; KOSGEI, J. C. E.; OKOTH, E. M.; OCHIENG'ASUDI, G.; RIMBERIA, F. K. Evaluation of the morphological and quality characteristics of new papaya hybrid lines in Kenya. **African Journal of Biotechnology**. Vol. 18(2), p.58-67, 9 January, 2019.

NOBRE, V. F. Caracterização Morfo-Agronômica de Acessos de Mamoeiro (*Carica Papaya* L.) da Embrapa Mandioca e Fruticultura e Estudo do Coeficiente De Variaçao em Experimentos da Espécie. Dissertação de Mestrado. **Universidade Federal do Recôncavo da Bahia**. Centro de Ciências Agrárias, Ambientais e Biológicas, Cruz das Almas, 38 f., 2016.

OCAMPO, J.; D'EECKENBRUGGE, G. C.; BRUYÈRE, S.; DE BELLAIRE, L. D. L.; OLLITRAULT, P. Organization of morphological and genetic diversity of Caribbean and Venezuelan papaya germplasm. **Fruits**, v. 61, n. 1, p. 25-37, 2006.

OLIVEIRA, E. J.; DIAS, N. L. P.; DANTAS, J. L. L. Selection of morpho-agronomic descriptors for characterization of papaya cultivars. **Euphytica**, v. 185, n. 2, p. 253-265, 2012.

OLIVEIRA, Eder Jorge de; OLIVEIRA FILHO, Osvaldo Sebastião de; SANTOS, Vanderlei da Silva. Selection of the most informative morphoagronomic descriptors for cassava germplasm. **Pesquisa Agropecuária Brasileira**, v. 49, n. 11, p. 891-900, 2014.

OCTAVIANI, F.; HAFSAH, S.; HAYATI, R. Chemical properties and morphological characteristics some genotype papaya (Carica papaya L.) in Aceh province. **International of Agronomy and Agricultural Research** (IJAAR), vol-13, p. 64-72, 2018.

PADILHA, H. K. M.; SOSINSKI JUNIOR, E. E.; BARBIERI, R. L. Morphological diversity and entropy of peppers (capsicum baccatum and capsicum chinense, solanaceae). **Internacional Journal of Current Research**, v8, p.42758-42766, 2016.

PADILHA, H. K. M.; SOSINSKI JUNIOR, E. E.; BARBIERI, R. L. Morphological diversity and entropy of peppers (capsicum baccatum and capsicum chinense, solanaceae). **International Journal of Current Research,** Vol. 8, Issue, 12, pp.42758-42766, 2016.

PALLANT, J. A step by step guide to data analysis using SPSS for windows. In: **SPSS Survival manual**. Open University Press, 2007.

PATIL, V. H.; SINGH, S. N.; MISHRA, S.; DONAVAN, D. T. Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one'criterion. **Journal of Business Research**, v. 61, n. 2, p. 162-170, 2008.

R DEVELOPMENT CORE TEAM. R: **A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, Vienna, 2017.

REIS, E. **Estatística Multivariada Aplicada**. Edições Sílabo, Lisboa, 2ª ed, p.342, 1997.

RENYI, A. **On Measures of Entropy and Information**. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics, University of California Press, v. 01, p.547-561, 1961.

RIEGER, J. E. Genetic and morphological diversity of natural populations of Carica papaya. Tese de Doutorado. **Miami University**, 2009.

SCHAWB, A.J. **Eletronic Classroom**. [Online], 2007. Disponível em: <http://www.utexas.edu/ssw/eclassroom/schwab.html> Acesso em: [22 jan. 2017].

SILVA, R. S.; MOURA, E. F.; FARIAS NETO, J. T.; LEDO, C. A.; SAMPAIO, J. E. Selection of morphoagronomic descriptors for the characterization of accessions of cassava of the Eastern Brazilian Amazon. **Genetics and Molecular Research**, Ribeirão Preto, v. 16, n. 2, p. 1-11, 2017.

SILVA, H. K.; PASSOS, A. R.; SCHNADELBACH, A. S.; MOREIRA, R. F. C.; CONCEIÇÃO, A. L. S.; LIMA, A. P. Selection of Morphoagronomic Descriptors in Physalis angulata L. Using Multivariate Techniques. **Journal of Agricultural Science**, v. 11, p. 289-302, 2018.

SHANNON, C. E. A mathematical theory of communication. **Bell system technical journal**, v. 27, n. 3, p. 379-423, 1948.

SOUZA, E. C. Caracterização morfológica, seleção de descritores e diversidade genética entre acessos de mangueira do banco ativo de germoplasma da Embrapa Semiárido. Dissertação (mestrado em Recursos Genéticos Vegetais), **Universidade Federal do Recôncavo da Bahia**, 2018.

TRINDADE, A.V.; OLIVEIRA, J.R.P.; **Mamão. Produção: aspectos técnicos, Embrapa Mandioca e Fruticultura** (Cruz das Almas, BA), Comunicação para Transferência de Tecnologia, 2000.

VELICER, W. F.; EATON, C. A.; FAVA, J. L. Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In: **Problems and solutions in human assessment**. Springer, Boston, MA, 2000. p. 41-71.

VIEIRA, E. A.; DE FREITAS FIALHO, J.; SILVA, M. S.; FUKUDA, W. M. G.; FALEIRO, F. G. Variabilidade genética do banco de germoplasma de mandioca da Embrapa cerrados acessada por meio de descritores morfológicos. **Científica**, v. 36, n. 1, p. 56-67, 2007.

ZOU, L.; PAUL, R. E.; CHEN, N. J. Papaya: Post Harvest Quality-Maintenance Guidelines. **College of Tropical Agriculture and Human Resources. University of Hawaii at Manoa**. Disponivel em:< https://www.ctahr.hawaii.edu/oc/freepubs/pdf/F_N-34.pdf> 2014.

ZWICK, W. R.; VELICER, W. F. Comparison of five rules for determining the number of components to retain. **Psychological bulletin**, v. 99, n. 3, p. 432, 1986.

ARTIGO 3

ANÁLISE DE AGRUPAMENTO POR MEIO DE MÉTODOS HIERÁRQUICOS E NÃO HIERÁRQUICOS NA CARACTERIZAÇÃO DE ACESSOS DE MAMÃO[3]

[3]Artigo a ser ajustado para posterior submissão ao Comitê Editorial do periódico científico Genetics and Molecular Research, em versão na língua inglesa.

ANÁLISE DE AGRUPAMENTO POR MEIO DE MÉTODOS HIERÁRQUICOS E NÃO HIERÁRQUICOS NA CARACTERIZAÇÃO DE ACESSOS DE MAMÃO

Autor: Antonio Leandro da Silva Conceição
Orientador: Carlos Alberto da Silva Ledo

RESUMO: Neste trabalho, objetivou-se comparar métodos de agrupamento hierárquicos e não hierárquicos, com o propósito de identificar o método mais adequado para quantificar a variabilidade genética existente entre 50 acessos do BAG-Mamão da Embrapa Mandioca e Fruticultura. Para o estudo foram utilizados 24 descritores quantitativos e 9 descritores qualitativos. O método de agrupamento hierárquico utilizado para os dados quantitativos foi o UPGMA (método de ligação média entre grupos) combinado com a distância euclidiana. Os métodos não hierárquicos de particionamento utilizados foram o algoritmo K-médias e o Particionamento em Torno de Medoids (PAM), os dados também foram agrupados por meio da análise de componentes principais combinado ao algoritmo HCPC (Clustering Hierárquico em Componentes Principais). Para obtenção dos agrupamentos dos dados qualitativos, foi utilizado o método UPGMA combinado com a distância de Cole-Rogers e o agrupamento com base na Análise de Correspondência Múltipla combinada ao algoritmo HCPC. Para determinação do melhor número de grupos foram utilizados os resultados das medidas de validação interna, estabilidade, além dos índices Pseudo-T2 e pseudo-F. Para as análises combinadas ao algoritimo HCPC também foi considerado o ganho de inércia interna para determinação do número ótimo de grupos. Os agrupamentos obtidos por meio da análise de componentes principais e da análise de correspondência múltipla apresentaram grupos bem distribuídos e mais consistentes, com maior semelhança dos acessos dentro dos grupos e maior a diferença entre os grupos, ou seja, um padrão de agrupamento mais adequado para os conjuntos de dados avaliados.

Palavra-chave: Análise multivariada, divergência genética, algoritmos de agrupamento

CLUSTERING ANALYSIS THROUGH HIERARCHICAL AND NON-HIERARCHICAL METHODS IN CHARACTERIZATION OF PAPAYA ACCESSES

Author: Antonio Leandro da Silva Conceição
Advisor: Carlos Alberto da Silva Ledo

ABSTRACT: This study aimed to compare hierarchical and non-hierarchical clustering methods, with the purpose of identifying the most appropriate method to quantify the genetic variability existing between 50 accessions of BAG-Papaya from Embrapa Mandioca and Fruticultura. For the study 24 quantitative descriptors and nine qualitative descriptors were used. The hierarchical clustering method used for quantitative data was the UPGMA (mean group bonding method) combined with Euclidean distance. The non-hierarchical partitioning methods used were the K-averaging algorithm and Medoids Partitioning (PAM), the data were also grouped through principal component analysis combined with the HCPC (Principal Component Hierarchical Clustering) algorithm. To obtain the qualitative data groupings, the UPGMA method combined with the Cole-Rogers distance and the clustering based on the Multiple Correspondence Analysis combined with the HCPC algorithm was used. To determine the best number of groups, the results of the internal validation measures, stability, and the Pseudo-T2 and pseudo-F indices were used. For the analyzes combined with the HCPC algorithm, the internal inertia gain was also considered to determine the optimal number of groups. The groupings obtained through principal component analysis and multiple correspondence analysis showed well distributed and more consistent groups, with greater similarity of accesses within the groups and greater the difference between the groups, that is, a clustering pattern best suited for the evaluated data sets.

Key words: Multivariate analysis, genetic divergence, clustering algorithms

INTRODUÇÃO

O mamoeiro, *Carica papaya* L., é uma das fruteiras mais cultivadas no mundo, seu fruto apresenta elevada importância econômica e fonte de renda para o Brasil, que é o segundo maior produtor mundial desta fruta, atrás apenas da Índia (FAO, 2017).

Informações sobre a estrutura genética de coleções de germoplasma é de suma importância para a conservação e utilização dos recursos genéticos (ODONG et al., 2011). Assim como identificar métodos mais eficientes de exploração dos dados, para quantificar de forma mais precisa a divergência genética existente.

A análise de agrupamento (cluster) visa definir grupos significativos a partir de observações, fazendo com que as observações atribuídas no mesmo grupo sejam semelhantes entre si (AHMAD e KHAN, 2019). Muitos procedimentos de cluster utilizam-se de distâncias, onde os clusters são obtidos definindo primeiro uma medida de distância apropriada e depois aplicando um algoritmo que atribui observações próximas um ao outro no mesmo cluster. No agrupamento baseado em distância, uma quantificação de similaridade entre objetos é baseada nessa distância. Depois que a distância é definida, um algoritmo hierárquico ou de particionamento é aplicado (BUDIAJI e LEISCH, 2019).

Diferentemente dos métodos hierárquicos, nos procedimentos não-hierárquicos (método divisivo) já se sabe, a priori o número k de grupos a serem formados antes mesmo de se iniciar a análise (FERREIRA, 2011). Dentre os principais métodos de particionamento se destacam o K-means (HARTIGAN; WONG, 1979) e o K-medóides (PAM) (KAUFMAN e ROUSSEEUW, 1987).

A análise de componentes principais permite a interpretação de múltiplos descritores e é eficaz na predição da divergência entre genótipos, permitindo que sejam agrupados para que haja homogeneidade nos grupos e heterogeneidade entre os grupos (CRUZ et al., 2011). A análise de correspondência múltipla (MCA) é uma extensão da análise de correspondência simples e também pode ser vista como uma generalização da análise de componentes principais quando as descritores a serem analisados são categóricos em vez de quantitativos (ABDI e WILLIAMS, 2010).

A análise de agrupamento, de preferência, deve ser realizada usando vários algoritmos de agrupamento conceitualmente distintos, ou seja, algoritmos que não

são direcionados para o mesmo tipo de agrupamentos (Handl et al. (2005). Com isso, aumentará as chances de se obter o agrupamento mais adequado para os conjuntos de dados em análise, resultando em respostas mais precisas e consistentes acerca dos grupos formados.

Alguns estudos utilizando métodos hierárquicos e/ou não hierárquicos para avaliar a divergência genética em mamoeiro podem ser encontrados na literatura (ASUDI et al., 2010; AIKPOKPODION, 2012; SUDHA et al., 2013; SARAN et al., 2015; ARA et al., 2016; SILVA et al., 2017; PRIHATINI, BUDIYANTI e NOFLINDAWATI, 2019; NASCIMENTO et al., 2019).

Um dos principais problemas que um pesquisador enfrenta quando se quer agrupar um determinado conjunto de dados, é escolher a melhor técnica para agrupá-los, como determinar um número ideal de grupos e como validar os resultados dos agrupamentos gerados. A realização de estudos com base em dados quantitativos ou qualitativos são importantes, pois nem sempre se dispõe de dados de diferentes naturezas para aplicação de uma análise simultânea, ou a depender do propósito da pesquisa não seja necessário utilizar dados de origens distintas.

Este trabalho teve por objetivo comparar os agrupamentos hierárquicos e não hierárquicos em acessos de mamão (*Carica papaya* L.) por meio da análise de dados quantitativos e qualitativos, com o propósito de identificar o método mais adequado para quantificar a variabilidade genética existente.

MATERIAL E MÉTODOS

Foram avaliados 50 acessos de mamão (*Carica papaya* L.) pertencentes ao banco de germoplasma de mamão (BAG-Mamão) da Embrapa Mandioca e Fruticultura (Tabela 1). O plantio dos acessos foi realizado do dia 26 a 29 de agosto de 2014. As avaliações foram realizadas de outubro de 2014 a dezembro de 2015. Foi utilizado espaçamento de 3,0 m entre linhas e 2,0 m entre plantas, adotando-se as práticas culturais e os tratos fitossanitários preconizados para a cultura (MARTINS e COSTA, 2003). As avaliações foram realizadas em Cruz das Almas, Bahia, Brasil (12º48'38"S e 39º6'26"W), na área experimental da Embrapa Mandioca e Fruticultura.

Para as análises foram utilizados 24 descritores quantitativos e 9 descritores qualitativos. Os descritores quantitativos foram: comprimento (cm) dos internódios 8 meses (CI8); comprimento (cm) dos internódios: 12 meses (CI12); Largura (cm) da folha: 8 meses (LF8); comprimento (cm) do pecíolo da folha: 8 meses (CPF8); comprimento (cm) do pecíolo da folha: 12 meses (CPF12); nº de frutos: 8 meses (NF8); nº frutos carpelóides: 8 meses (NFCa8); nº de frutos: 12meses (NF12); nº de frutos carpelóides: 12 meses (NFCa12); nº de frutos por axila (NFaxi); altura (m) dos primeiros frutos (ALT1F); comprimento (cm) do pedúnculo do fruto (CPF); nº de flores por pedúnculo (NFP); comprimento (cm) do pedúnculo da inflorescência (CPI); comprimento (cm) da corola da flor hermafrodita (CCFher); comprimento (cm) do fruto (CF); firmeza dos frutos (MFF); diâmetro da cavidade central (DCC); peso (g) fresco de sementes do fruto (PFS); peso (g) fresco de 100 sementes (PS); acidez (AC); vitamina C (VITC); pH (PH) e sólidos solúveis totais (BRIX).

Os descritores qualitativos utilizados foram: cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), formato dos bordos foliares (FBF), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI) e densidade de flores nas inflorescência (DFI).

Tabela 1. Classificação por tipo e origem dos acessos de mamão avaliados, que compõem o Banco de Germoplasma (BAG-Mamão) da Embrapa Mandioca e Fruticultura. Cruz das Almas-BA, 2019.

Acesso	Tipo de fruto	Origem	Sigla
BGM 01	Formosa	Costa Rica	BGM 01 FC
BGM 02	Formosa	Taiwan	BGM 02 FT
BGM 03	Formosa	Havaí	BGM 03 FH
BGM 04	Solo	Havaí	BGM 04 SH
BGM 05	Solo	Havaí / Taiwan	BGM 05 SHT
BGM 06	Formosa	Malásia	BGM 06 FM
BGM 07	Formosa	Costa Rica	BGM 07 FC
BGM 08	Solo	Malásia	BGM 08 SM
BGM 09	Formosa	Malásia	BGM 09 FM
BGM 10	Formosa	Taiwan	BGM 10 FT
BGM 11	Formosa	Taiwan	BGM 11 FT
BGM 12	Formosa	Brasil	BGM 12 FB
BGM 13	Solo	Taiwan	BGM 13 ST
BGM 14	Formosa	Malásia	BGM 14 FM
BGM 15	Formosa	Taiwan	BGM 15 FT
BGM 16	Formosa	*	BGM 16 F
BGM 17	Formosa	Costa Rica	BGM 17 FC
BGM 18	Formosa	*	BGM 18 F

Tabela 1. (Continuação)

Acesso	Tipo de fruto	Origem	Sigla
BGM 19	Formosa	Costa Rica	BGM 19 FC
BGM 20	Formosa	*	BGM 20 F
BGM 21	Solo	*	BGM 21 S
BGM 22	Formosa	Brasil	BGM 22 FB
BGM 23	Formosa	Brasil	BGM 23 FB
BGM 24	Formosa	Brasil	BGM 24 FB
BGM 25	Formosa	Brasil	BGM 25 FB
BGM 26	Formosa	Brasil	BGM 26 FB
BGM 27	Solo	Brasil	BGM 27 SB
BGM 28	Solo	Brasil	BGM 28 SB
BGM 29	Solo	Brasil	BGM 29 SB
BGM 30	Formosa	Havaí	BGM 30 FH
BGM 31	Formosa	Brasil	BGM 31 FB
BGM 32	Solo	Brasil	BGM 32 SB
BGM 33	Solo	Havaí	BGM 33 SH
BGM 34	Formosa	Havaí	BGM 34 FH
BGM 35	Solo	Brasil	BGM 35 SB
BGM 36	Formosa	Brasil	BGM 36 FB
BGM 37	Formosa	Brasil	BGM 37 FB
BGM 38	Solo	*	BGM 38 S
BGM 39	Solo	Brasil	BGM 39 SB
BGM 40	Formosa	*	BGM 40 F
BGM 41	Formosa	*	BGM 41 F
BGM 42	Solo	Havaí	BGM 42 SH
BGM 43	Solo	Havaí	BGM 43 SH
BGM 44	Solo	Havaí	BGM 44 SH
BGM 45	Formosa	África do Sul	BGM 45 FA
BGM 46	Formosa	África do Sul	BGM 46 FA
BGM 47	Solo	África do Sul	BGM 47 SA
BGM 48	Formosa	Brasil	BGM 48 FB
BGM 49	Formosa	*	BGM 49 F
BGM 50	Formosa	*	BGM 50 F

Antes de aplicar os métodos de agrupamento, o primeiro passo foi avaliar se os dados são agrupáveis, um processo definido como a avaliação da tendência de agrupamento, para essa avaliação foi utilizada a estatística de Hopkins.

A estatística Hopkins (Lawson e Jurs, 1990) é usada para avaliar a tendência de agrupamento de um conjunto de dados por medir a probabilidade de que um determinado conjunto de dados tenha origem de uma distribuição uniforme. Em outras palavras, ele testa a aleatoriedade espacial dos dados.

Por exemplo, seja D um conjunto de dados real. A estatística Hopkins pode ser calculada como:

A fórmula é definida da seguinte forma:

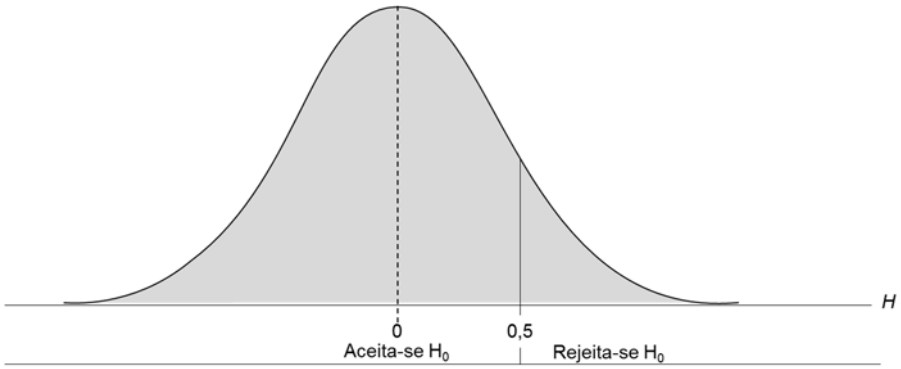
$$H = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi + \sum_{i=1}^{n} yi}$$

Um valor de H com cerca de 0,5 significa que $\sum_{i=1}^{n} Yi$ e $\sum_{i=1}^{n} xi$ estão próximos um do outro e, portanto, os dados D são distribuídos uniformemente.

As hipóteses nulas e alternativas são definidas da seguinte forma:

Hipótese nula (**H_0**): o conjunto de dados D é distribuído uniformemente (ou seja, nenhum clusters)

Hipótese alternativa (**H_1**): o conjunto de dados D não é distribuído uniformemente (ou seja, contém clusters significativos)



Se o valor da estatística Hopkins for próximo de zero (abaixo de 0,5), pode-se rejeitar a hipótese nula e concluí-se que o conjunto de dados D é significativamente um dado agrupável (HAN; KAMBER; PEI, 2012; PRASAD, 2016; KRISHNA; BABU; KUMAR, 2018).

Após a avaliação da tendência dos agrupamentos, foram aplicados os métodos de agrupamento hierárquicos e não hierárquicos nos conjuntos de dados estudados. Esses métodos foram empregados para avaliar a previsibilidade dos agrupamentos ideais combinados a métodos estatísticos de validade interna, estabilidade e os índices Pseudo-F e PseudoT² .

Dados quantitativos

O método de agrupamento hierárquico utilizado para os dados quantitativos foi o de ligação média entre os grupos, o UPGMA (Unweighted Pair-Group Method Using Arithmetic Averages) combinado com a distância euclidiana. Os métodos não hierárquicos de particionamento utilizados foram o K-médias (K-means) proposto por (HARTIGAN; WONG, 1979) e o de Particionamento em Torno de Medoids (PAM) baseado em medoids (KAUFMAN; ROUSSEEUW, 1987). Os

dados quantitativos também foram agrupados utilizando a análise de componentes principais (PCA) combinado ao algoritimo ((HCPC) Hierarchical Clustering on Principal Components) (HUSSON; JOSSE; PAGES, 2010; HUSSON et al, 2019). A análise de componentes principais foi inicialmente descrita por Pearson (1901) e uma descrição de métodos computacionais práticos veio muito mais tarde com Hotelling (1933, 1936). A HCPC foi realizada em duas etapas, sendo a primeira uma análise de componentes principais sobre os descritores numéricos selecionados. Por combinação desses descritores, foram obtidos os componentes principais, ou seja, dimensões subjacentes refletindo as correlações dos descritores originais. Em seguida, os escores fatoriais dos sujeitos sobre os componentes principais foram submetidos a uma análise hierárquica de agrupamentos.

Dados qualitativos

Para obtenção dos agrupamentos dos dados qualitativos, foi utilizado o método UPGMA combinado com a distância de Cole-Rogers et al. (1997). Para esse conjunto de dados, também foi utilizada a Análise de Correspondência Múltipla (MCA). A MCA foi usada para transformar as descritores categóricos em poucos componentes principais contínuos, para então serem usados como entrada da análise de agrupamento. É um método de componente principal aplicável aos métodos qualitativos (dados categóricos). Onde as dimensões representam os principais componentes, que são ordenados para que a primeira e a segunda dimensão (Dim1 e Dim2) expliquem a maior parte da variação dos dados.

Na MCA, a matriz indicadora é usada e associações entre descritores são descobertas calculando as distâncias entre as categorias de descritores e entre os acessos. Então, as associações são visualizadas e interpretadas. Após a realização da MCA foi aplicado o algoritmo HCPC para melhor visualização dos dados aplicados nas primeiras dimensões, sendo extraido o essencial das informações dos dados originais.

Criterios para determinação do melhor número de grupos

Como as medidas de validação também podem ser utilizadas como critério para definir o número ideal de grupos, fornecendo quantidades significativas de informações que não podem ser obtidas usando apenas a inspeção visual (HANDL et al., 2005). Foram usadas três medidas de validade interna: conectividade (HANDL et al., 2005), a Largura da silhueta (ROUSSEEUW, 1987) e o Índice de Dunn (DUNN, 1974). Já as medidas de estabilidade incluídas foram APN (average proportion of non-overlap), AD (average distance), ADM (average distance between means) e FOM (figure of merit), (DATTA e DATTA, 2003; YEUNG et al., 2001). As medidas de validade interna e estabilidade aqui empregadas, foram comparadas umas com as outras, a fim de obter o resultado mais consistente e representativo para os conjuntos de dados quantitativos utilizados. Essas medidas foram usadas para comparar os métodos não hierárquicos de particionamento K-means e PAM e o método hierárquico UPGMA combinado a distância euclidiana.

Para avaliar o desempenho dos algoritimos de particionamento também foi calculada a distribuição de associações de cluster por meio da entropia (MEILA, 2007) e para comparar a semelhança dos grupos formados em cada método foi utilizado o critério Pearson Gamma (HALKIDI et al. (2001). Para avaliar a consistência dos agrupamentos hierárquicos foi utilizado além das medidas de validação, o coeficiente de correlação cofenético (SNEATH e SOKAL, 1973).

Os índices Pseudo-F (CALINSKI e HARABASZ ,1974) e Pseudo T2 (DUDA e HART, 1973) são bons indicadores do número de Grupos (MILLIGAN e COOPER, 1985; MINGOTTI, 2005). Com isso, esses critérios também foram utilizados para auxiliar na determinação do número de grupos ideal.

Nas análises dos dados quantitativos com base na PCA e dos qualitativos com base na MCA, o número de grupos também foi considerado pelo ganho de inércia, ou seja, a hierarquia foi representada por um dendograma que foi indexado pelo ganho de inércia interna. Sendo que, a inércia dentro do grupo caracteriza a homogeneidade de um cluster. A função HCPC (Clustering Hierárquico nos Componentes Principais) implementa esse cálculo depois de ter construído a hierarquia e sugere um “ótimo” nível de divisão. É sugerido uma divisão em Q clusters quando o aumento da inércia entre $Q - 1$ e Q é muito maior que entre Q e $Q + 1$ clusters. Um critério empírico pode formalizar essa ideia. Seja $\Delta(Q)$ a inércia aumenta quando se desloca de Q - 1 a Q clusters, o critério proposto é:

$$\frac{\Delta(Q)}{\Delta(Q\ +\ 1)}.$$

As características dos grupos (clusters) obtidos por meio da análise de PCA e MCA também foram auxiliadas pelo valor do teste v. O valor do teste v indica a importância do descritor no cluster. Um valor positivo indica que o valor médio do descritor no cluster é maior que a média geral, já um valor negativo indica que o valor médio do descritor no cluster é menor que a média geral.

Todas as análises foram realizadas no programa R (R CORE TEAM, 2018), com exceção da análise de agrupamento dos dados qualitativos por meio da distância de Cole-Rogers et al. (1997), a qual foi obtida no programa Genes (CRUZ, 2013).

RESULTADOS E DISCUSSÃO

Inicialmente os dados quantitativos e qualitativos foram testados para avaliar a tendência dos agrupamentos, sendo verificado por meio da estatística de Hopkins (H). Onde pode-se observar que os dados avaliados apresentaram valores inferiores a 0,5, sendo observado que os dados quantitativos e qualitativo apresentaram valor de 0,34 e 0,41, respectivamente para esse teste. Em virtude disso, constata-se que os dados são agrupáveis. Atualmente trabalhos em diversas áreas têm utilizado a estatística de Hopkins (H) para verificação da tendência de agrupamento, como o de Krishna; Babu; Kumar, (2018), estudando determinação ideal de grupos em métodos de agrupamento não hierárquicos, encontrando valor de 0,2357, indicando que os dados são altamente agrupáveis.

Poucos pesquisadores que trabalham principalmente com análises de divergência genética vegetal tem se interessado em realizar verificações pré-agrupamento como a de tendência de agrupamento, porém esses profissionais devem estar atentos quanto a esses tipos de verificações iniciais, pois qualquer método de agrupamento agrupará seus dados, mesmo que não haja grupos, daí a busca a posteriori por “validade de agupamentos”.

Nas Figuras 1 e 2 são apresentadas as informações intrínsecas dos dados quantitativos para avaliar a qualidade do agrupamento. O agrupamento hierárquico com dois grupos foi o que apresentou os resultados mais consistentes de acordo

com as medidas de validação interna utilizadas. De acordo com Handl (2005), uma boa solução de agrupamento tende a ter um desempenho razoavelmente bom sob várias medidas. Os escores ótimos para conectividade foi de (6,48), Dunn (0,44) e Silhueta (0,51). Segundo Brock et al. (2008), os valores de conectividade variam entre 0 e infinito e devem ser minimizadas, ou seja quanto mais próximo de zero melhor, já os valores das medidas de Dunn e Silhueta devem ser maximizadas, seus valores variam de 0 e -1 respectivamente, a 1.

Em relação a validação de estabilidade dos agrupamentos, as medidas APN e ADM foram concordantes com o resultado das medidas de validação interna, apontando para o método hierárquico com 2 grupos como o mais estável. As medidas de AD e FOM apesar da concordância quanto ao método de agrupamento hierárquico, divergiram das demais quanto ao número de grupos, essas medidas apresentaram uma tendência a diminuir à medida que o número de clusters aumenta (Figura 2).

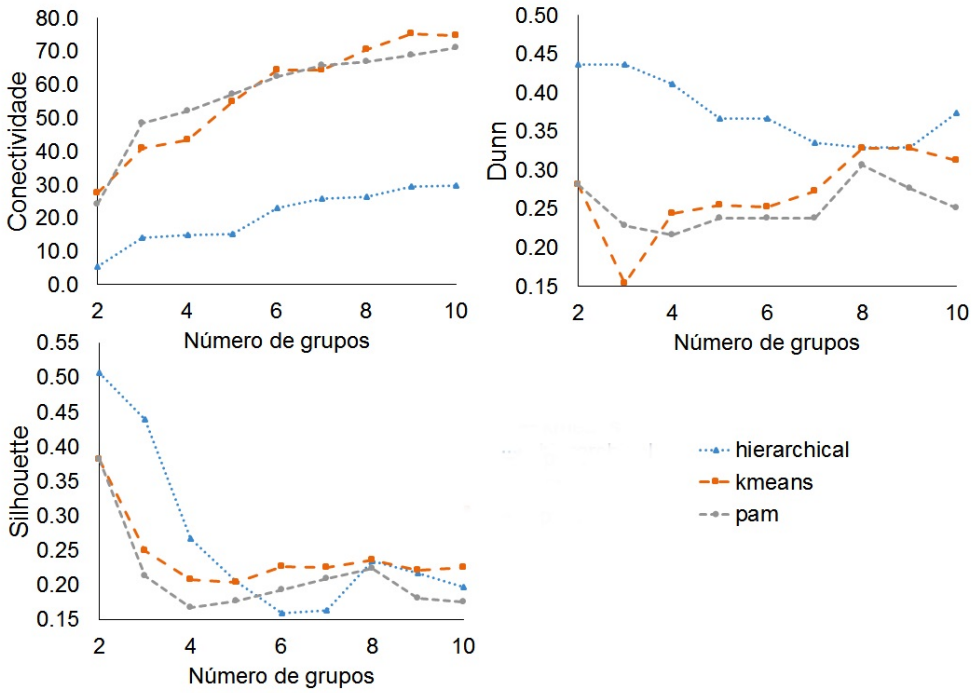


Figura 1: Validação Interna dos dados quantitativos. Medidas de Conectividade, Largura da silhueta (silhouette) e o Índice de Dunn (Dunn).

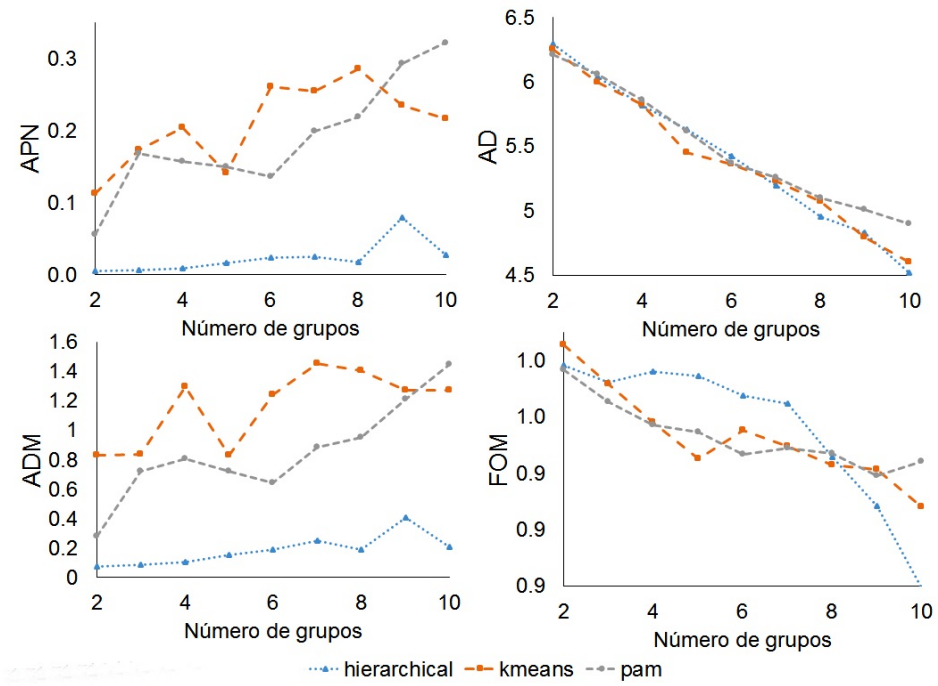


Figura 2: Validação de Estabilidade dos dados quantitativos. AD (Distância Média); ADM (Distância Média entre as Médias); APN (Proporção Média de Não-sobreposição) e FOM (Figura de Mérito).

As técnicas de agrupamento empregadas convergiram para um número de grupos comum, sendo determinado a formação de dois grupos como a ideal para o agrupamento pelo método UPGMA e pelos métodos de particionamento K-means e PAM. Sendo que desses métodos citados, o método hierárquico UPGMA foi indicado como o mais estável e consistente. Já o agrupamento resultante da análise de componentes principais (PCA), o número de grupos representado no dendograma foi indexado pelo ganho de inércia interna, onde foi observado a sugestão de 3 grupos distintos.

No agrupamento dos dados quantitativos pelo método hierárquico UPGMA (figura W), obtido por meio da distância euclidiana, com valor do coeficiente de correlação cofenético (CCC) de 0,83**, observa-se que os dois grupos formados, apresentaram uma distribuição dos acessos bastante influenciada pelo acesso BGM3, o qual se apresenta como o mais divergente de acordo com a matriz de dissimilaridade (Figura 3). O grupo 1 (G1), é formado justamente pelo acesso BGM3, e o grupo 2 (G2) formado pelos demais acessos (Figura 3 (d)). Apesar da consistência e estabilidade obtido método UPGMA, esse método não apresentou um padrão de agrupamento muito plausível, em relação aos grupos e a distribuição

dos acessos nos mesmos. Os agrupamentos obtidos pelos métodos não hierárquicos K-means e PAM, que apesar de apresentar uma maior distribuição dos acessos em dois grupos, não resultou em uma maior homogeneidade interna dos grupos, como pode ser observado na Figura 3 (a, b).

A validação entre os algoritmos de particionamento (Figura 3c), foi calculada por meio da entropia da distribuição de associações de cluster (MEILA, 2007), onde o resultado mostra que o agrupamento K-means apresentou 0,65 e o PAM 0,55, sendo assim o algoritimo PAM apresentou um agrupamento com qualidade um pouco melhor que o k-means, sendo a entropia uma medida negativa, quanto menor a entropia, melhor o agrupamento. No trabalho realizado por Sripada e Rao (2011), comparando agrupamentos pelos métodos k-means e fuzzy c-means, a medida de entropia foi eficiente na indicação do melhor agrupamento. Contudo, no presente estudo, ambos os métodos de particionamento apresentaram valores baixos para o critério Pearson Gamma (HALKIDI et al. (2001), com o K-means apresentando o valor de 0,27 e o PAM 0,31 (Figura 3c). De acordo com esse critério, zero significa o mesmo agrupamento, e um significa agrupamentos diferentes, ou seja, os grupos 1 e 2 gerados em ambos os métodos foram bastante semelhantes entre si, isso implica dizer que houve heterogeneidade bastante restrita entre os grupos formados, sendo que, o que se busca é uma maior homogeneidade possível dentro do grupo e maior heterogeneidade entre os grupos.

Esses resultados evidenciam que os agrupamentos obtidos por meio dos algoritmos de particionamento não apresentaram uma separação significativa e satisfatória dos acessos avaliados. Levando em consideração os resultados estatísticos, bem como a inspeção visual dos agrupamentos gerados, juntamente com conhecimento e verificação da estrutura dos dados originais, o padrão de agrupamento apresentado pela PCA com a separação em 3 grupos se mostrou mais adequado, exibindo uma distribuição mais condizente dos acessos.

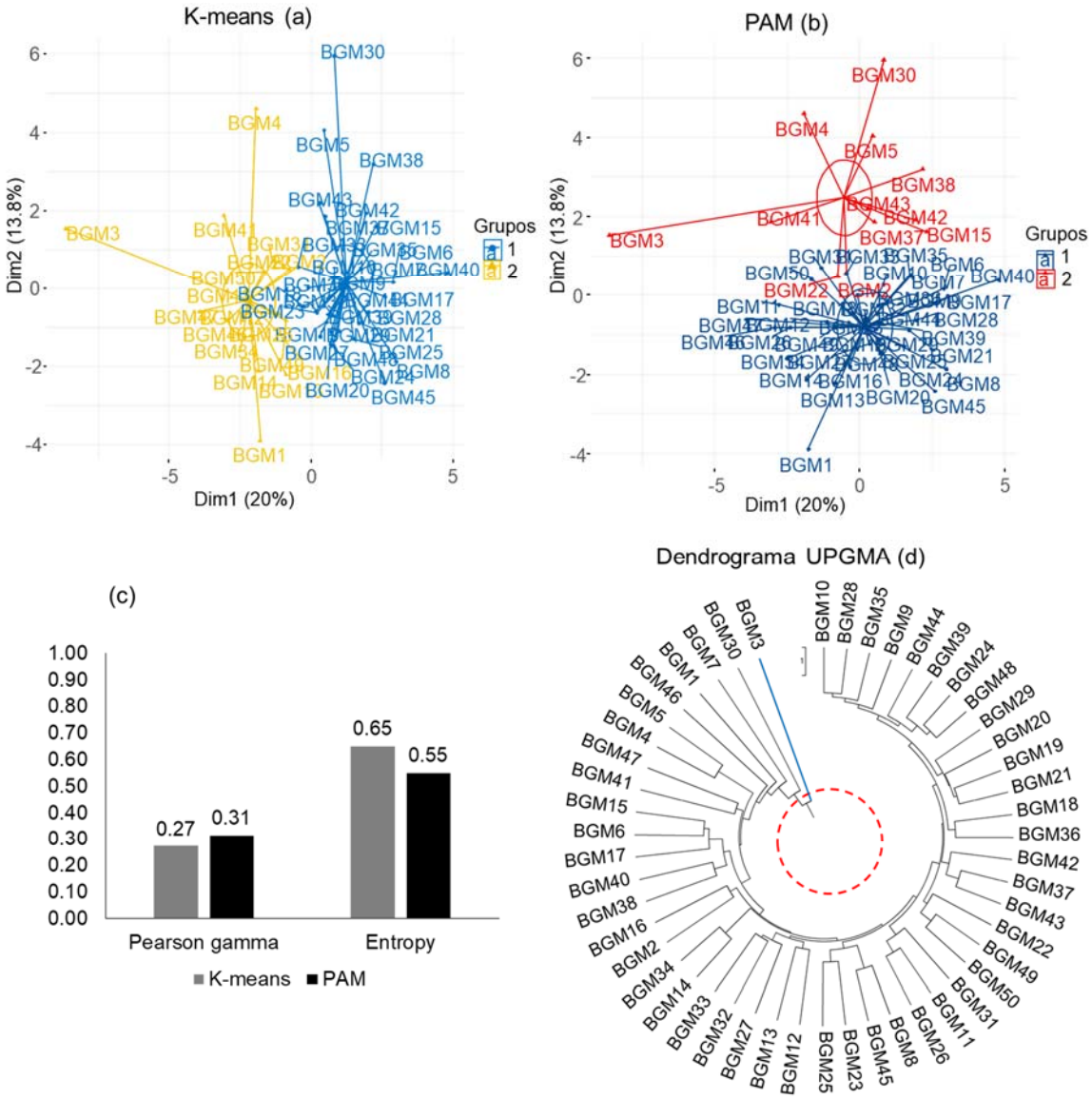


Figura 3. Agrupamentos dos dados quantitativos obtidos por meio dos métodos de particionamento K-means (a), PAM (b), gráfico com os critérios de qualidade e semelhança dos agrupamentos pelos métodos de particionamento (c) e o dendrograma do método hierárquico UPGMA (d).

Na análise de componentes principais (PCA), ao analisar as estimativas dos autovalores associados aos componentes principais e suas respectivas variâncias relativas e acumuladas obtidas para os 24 descritores quantitativos, percebe-se que os sete primeiros componentes conseguiram explicar 71,11% da variação total (Figura 4). Resultados semelhantes quanto a explicação da variação total foram obtidos nos trabalhos de divergência morfológica em mamão, de Asudi et al. (2010),

com 72,9% e Aikpokpodion (2012), com 73,47% da variação total explicada pelos sete primeiros componentes.

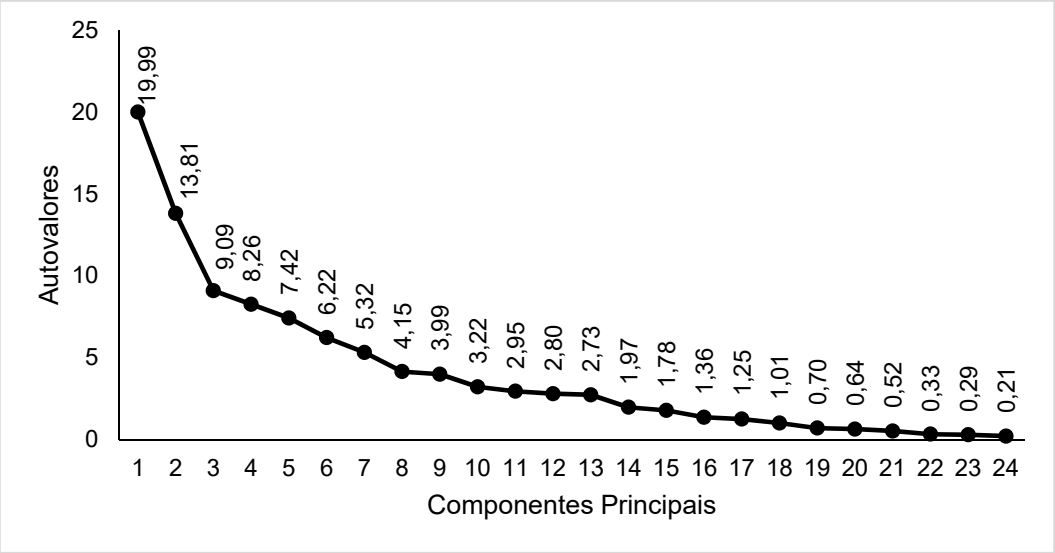


Figura 4. Autovalores (eixo y) e porcentagem da variação original armazenada em cada um dos 24 componentes principais (eixo dos x).

O dendrograma obtido pelo algoritmo HCPC apresentou resultados satisfatórios ao distinguir os grupos e diferenciar os acessos mais divergentes, gerando 3 grupos distintos (Figura 5). Estudos utilizando a técnica de PCA para avaliar a divergência genética em mamoeiro, encontraram números de grupos próximos ao encontrado no presente estudo. Sudha et al. (2013) obtiveram 4 grupos para 73 genótipos de mamão, Saran et al. (2015) também obtiveram 4 grupos, avaliando 24 genótipos e Ara et al. (2016) onde obtiveram 4 grupos, avaliando 14 genótipos de mamão.

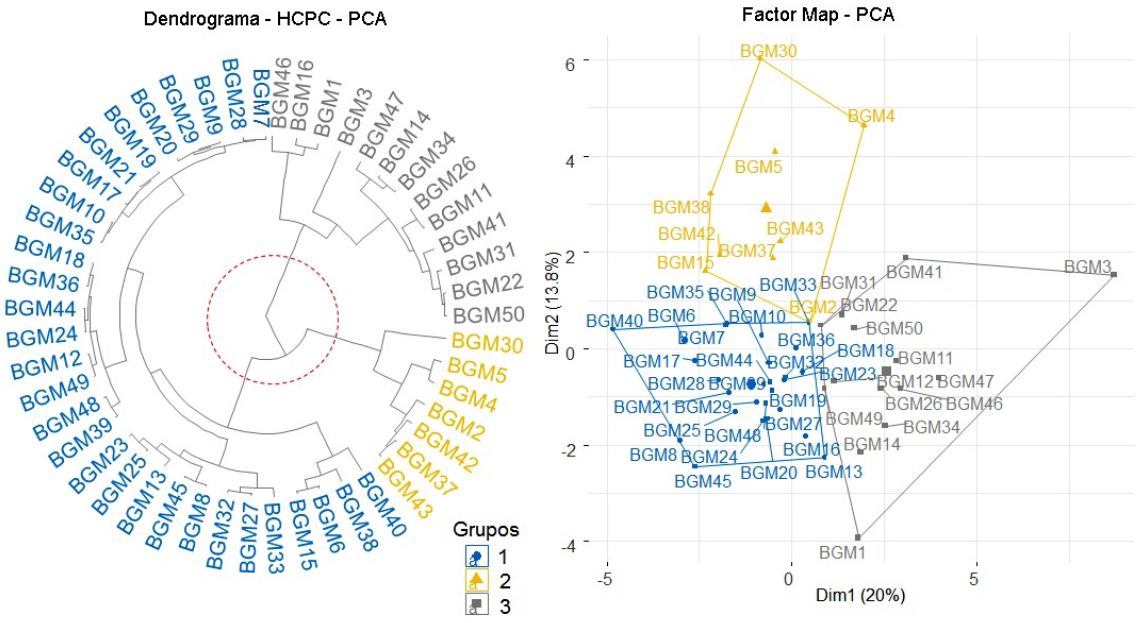


Figura 5: Dendrograma e Fator Map da distribuição dos grupos de 50 acessos de mamão do Banco de germoplasma da Embrapa Mandioca e Fruticultura obtido por meio da PCA, combinado ao algoritmo HCPC.

A partir da Tabela 2, pode-se observar que os descritores VITC, NF8, CI12 e NFaxi estão mais significativamente associadas ao grupo 1 (G1), com destaque para VITC que apresentou uma média de 90,48, sendo superior à média geral de 81,36. As dimensões 1 e 2 são as mais significativamente influentes na formação desse grupo (Tabela 3). As características mais importantes nesse grupo são acessos com os maiores valores para vitamina C, menor quantidade de frutos aos 8 e 12 meses e menor quantidade de frutos por axila. Esse grupo é composto por 30 acessos (BGM6, BGM7, BGM8, BGM9, BGM10, BGM12, BGM13, BGM15, BGM17, BGM18, BGM19, BGM20, BGM21, BGM23, BGM24, BGM25, BGM27, BGM28, BGM29, BGM32, BGM33, BGM35, BGM36, BGM38, BGM39, BGM40, BGM44, BGM45, BGM48 e o BGM49), (Figura 5). O grupo 1 também possui os acessos com os maiores valores para o BRIX, cerca de 86% desses acessos possuem faixas entre 12 e 15 ºBrix. Características como altos valores de Brix e Vitamica C são bastante relevantes em trabalhos de melhoramento da cultura. Os sólidos solúveis totais (BRIX) são um dos principais parâmetros de qualidade de frutos de mamão. A exigência para comercialização de mamão para exportação, é que apresente °Brix superior a 12 (FAGUNDES e YAMANISHI (2001); SCHWEIGGERT et al., 2012; DANTAS et al., 2015).

Já o grupo 2 (G2) é formado por 7 acessos (BGM2, BGM4, BGM5, BGM30, BGM37, BGM42 e o BGM43) (Figura 5). As dimensões 2 e 3 são as mais significativamente influentes na formação desse grupo (Tabela 3). Sendo caracterizado por apresentar acessos com maior quantidade de frutos aos 8 (NF8) e aos 12 meses (NF12), como também por apresentar maior quantidade de frutos por axila (NFaxi) e por apresentar os menores valores para ALT1F. A altura de inserção dos primeiros frutos (ALTF1) é um descritor de fundamental importância nos programas de melhoramento do mamoeiro, pois quanto menor o valor obtido para este caráter, mais precocemente a planta começa a produzir frutos, indicando precocidade e maior facilidade para a colheita de frutos em ciclos de produção mais avançados (DIAS et al., 2011; DANTAS et al., 2015).

A maioria dos acessos desse grupo não apresentou a formação de frutos carpelóides, porém apenas um desses acessos se destaca negativamente em relação a esse descritor, o BGM30, pois apresentou 11 frutos carpelóides aos 8 meses (NFCa8). Segundo Dantas et al. (2015), quanto maiores os valores desse descritor, consequentemente, menor vai ser a produtividade, indicando que a seleção de genótipos de mamoeiro deve ser realizada visando a seleção daqueles que apresentem menores valores para essa característica.

E por fim o grupo 3 (G3), composto por 13 acessos (BGM1, BGM3, BGM11, BGM14, BGM16, BGM22, BGM26, BGM31, BGM34, BGM41, BGM46, BGM47 e o BGM50 (Figura 5). A dimensão 1 é a mais significativamente influente na formação desse grupo (Tabela 3), que é caracterizado principalmente por apresentar os maiores valores para comprimento do pedúnculo da inflorescência (CPI), Largura da folha: 8 meses (LF8), altura dos primeiros frutos (ALT1F), comprimento do fruto (CF).

Para maior detalhamento dos três grupos obtidos, na Figura 12 (ANEXO), são apresentados os box-plots contendo a variação de cada um dos 24 descritores nesses grupos. Sendo mostrado que a separação entre os 3 grupos é bastante distinguível por meio da magnitudes dos valores medianos para grande maioria desses descritores.

Tabela 2. Descritores quantitativos que mais descrevem cada grupo resultante da análise de componentes principais (PCA).

Grupos	Descritores	v.test	Média no Grupo	Média geral	p.valor
G1	VITC	3,354	90,477	81,363	7,98E-04
	NF8	-2,324	9,148	12,72	2,01E-02
	CI12	-2,693	12,544	13,584	7,08E-03
	NFaxi	-2,88	1,148	1,4	3,98E-03
	CPF	-2,924	3,347	3,875	3,46E-03
	LF8	-3,063	58,593	62,019	2,19E-03
	NFP	-3,105	5,356	6,588	1,90E-03
	NF12	-3,117	6,037	9,86	1,83E-03
	CPF8	-3,684	57,387	62,019	2,30E-04
	CPI	-3,937	1,561	2,211	8,24E-05
G2	NF8	5,175	31,111	12,72	2,28E-07
	NFaxi	4,065	2,222	1,4	4,80E-05
	NF12	3,81	20,667	9,86	1,39E-04
	NFCa8	3,129	2,444	0,74	1,76E-03
	PFS	-2,124	42,286	62,556	3,36E-02
	MFF	-2,128	1,252	1,793	3,34E-02
	PH	-2,908	5,128	5,251	3,63E-03
	ALT1F	-3,075	0,667	0,952	2,10E-03
G3	CPI	4,375	3,466	2,211	1,22E-05
	LF8	4,269	70,317	62,019	1,96E-05
	NFP	4,098	9,414	6,588	4,16E-05
	CPF8	4,01	70,777	62,019	6,08E-05
	CPF	3,695	5,032	3,875	2,20E-04
	ALT1F	2,943	1,157	0,952	3,25E-03
	CPF12	2,835	55,336	49,168	4,58E-03
	CF	2,596	19,929	17,801	9,44E-03
	CCFher	2,146	3,934	3,657	3,19E-02
	PS	2,132	11,893	10,937	3,30E-02
	CI12	1,962	14,9	13,584	4,97E-02
	VITC	-2,282	70,591	81,363	2,25E-02

Comprimento dos internódios: 12 meses (CI12); Largura da folha: 8 meses (LF8); Comprimento do pecíolo da folha: 8 meses (CPF8); Comprimento do pecíolo da folha: 12 meses (CPF12); Nº frutos: 8 meses (NF8); Nº frutos carpelóides: 8 meses (NFCa8); Nº frutos: 12meses (NF12); Nº frutos carpelóides: 12 meses (NFCa12); Nº frutos por axila (NFaxi); Altura dos primeiros frutos (ALT1F); Comprimento do pedúnculo do fruto (CPF); Número de flores por pedúnculo (NFP); Comprimento do pedúnculo da inflorescência (CPI); Comprimento do fruto (CF); Firmeza dos frutos (MFF); Peso fresco de sementes do fruto (PFS); Peso fresco de 100 sementes (PS); Vit C (VITC); pH (PH). Teste de desvio da média da população (v.test). Grupo 1 (G1), Grupo 2 (G2), Grupo 3 (G3)

A estatística Hopkins, utilizando os descritores quantitativos originas apresentou um valor de 0,33. Já o agrupamento considerando os 3 primeiros PCs apresentou um valor de 0,28. Sendo assim, a agrupabilidade das informações transmitidas pelas pontuações do PCA com 3 componentes é melhor do que a

agrupabilidade das informações transmitidas pelas variáveis (descritores) explicativas originais. Esse resultado reforça a capacidade de agrupamento das informações transmitidas pelas pontuações dos 3 primeiros PCs (Tabela 3), mostando ser ideal para representar o melhor padrão de agrupamento. Resultados semelhantes podem ser observados no trabalho de Mylevaganam (2017), estudando o índice de Análise do Desenvolvimento Humano (IDH) para categorizar os Estados-Membros das Nações Unidas (ONU), onde a agrupabilidade das informações transmitidas pelas pontuações do PCA foi melhor do que a agrupabilidade das informações transmitidas pelas variáveis explicativas.

Tabela 3. Dimensões principais mais associadas aos grupos obtidos na PCA.

Grupos	Dimensões	v.test	Média na categoria	Média geral	p.valor
G1	Dim.2	-3,021	-0,725	-1,62E-16	0,00252
	Dim.1	-3,833	-1,107	-1,80E-15	0,00013
G2	Dim.2	5,251	2,915	-1,62E-16	1,51E-07
	Dim.3	-2,117	-0,953	-5,68E-17	3,43E-02
G3	Dim.1	5,127	2,573	-1,80E-15	2,94E-07

Grupo 1 (G1), Grupo 2 (G2), Grupo 3 (G3). Teste de desvio da média da população (v.test).

Para os dados qualitativos o agrupamento obtido pelo método UPGMA combinado com a distância de Cole-Rogers et al., (1997), (Figura 6) apresentou um CCC de 0, 67, esse valor é considerado como baixa consistência. Conforme sugerem Bussab et al. (1990), análises de agrupamento são aceitáveis se produzirem um coeficiente de correlação cofenético a partir de 0,80. Segundo Mohammad e Prasanna (2003), quanto maior CCC, menor será a distorção provocada ao agrupar os acessos. Essa baixa consistência estatística foi constada com a inspeção visual dos grupos formados, com base na verificação minuciosa dos dados qualitativos originais e suas respectivas categorias. Sendo que, independente do número de grupos sugeridos pelos índices pseudoT^2 (9 grupos) e pelo índice Pseudo-F (4 grupos), em todas as situações os grupos exibiram baixíssima homogeneidade interna. Fazendo com que o agrupamento hierárquico UPGMA combinado a distância de Cole-Rogers, se mostrasse frágil e com baixa precisão. Já os grupos formados por meio da análise de correspondência múltipla (MCA) apresentou uma ótima distribuição e separação dos acessos, apresentando grande homogeneidade dentro dos grupos e heterogeneidade bastante acentuada

entre esses grupos, nesse método foram formados 4 grupos indexados pelo ganho de inércia interna, esse número de grupos foi concordante com o sugerido pelo índice Pseudo-F.

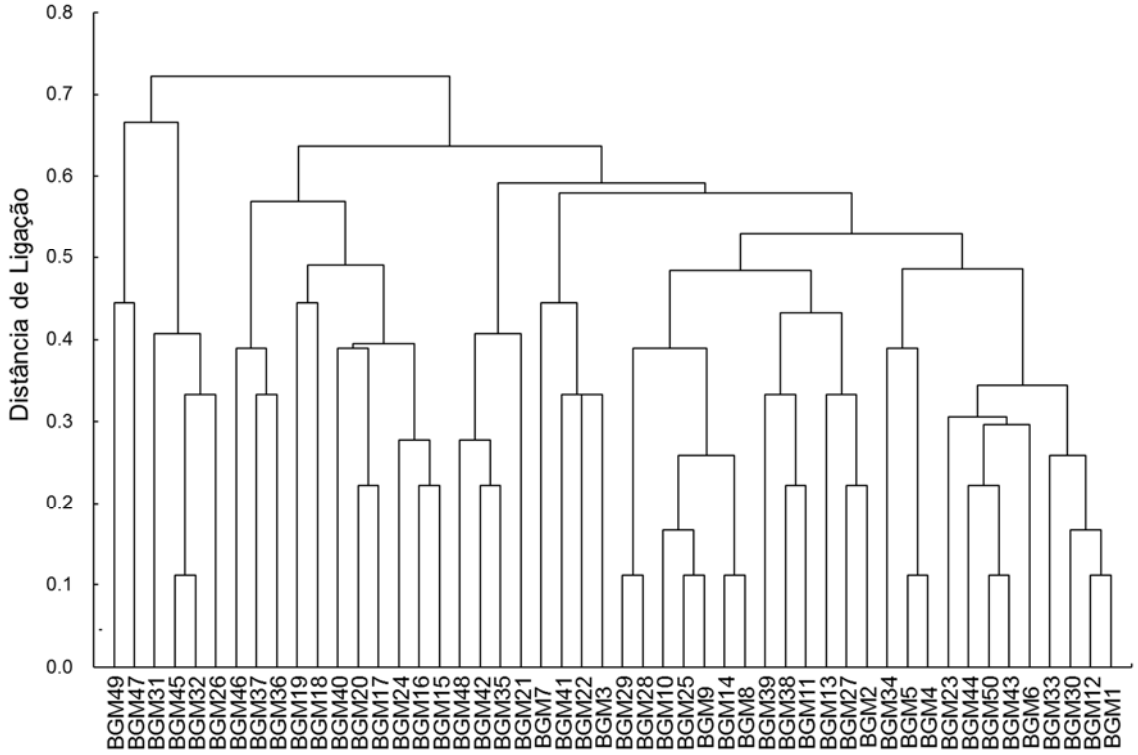


Figura 6. Dendrograma obtido do agrupamento de 9 descritores qualitativos por meio do método UPGMA combinado a distância distância de Cole-Rogers et al., (1997).

Na análise de correspondência multipla (MCA), ao analisar as estimativas dos autovalores associados às dimensões principais e suas respectivas variâncias relativas e acumuladas obtidas de 9 descritores qualitativos contendo um total de 41 categorias avaliadas, essas categorias foram totalmente explicadas por 32 dimensões, sendo que 12 dessas dimensões explicam mais de 70% da variação dos dados (Figura 7).

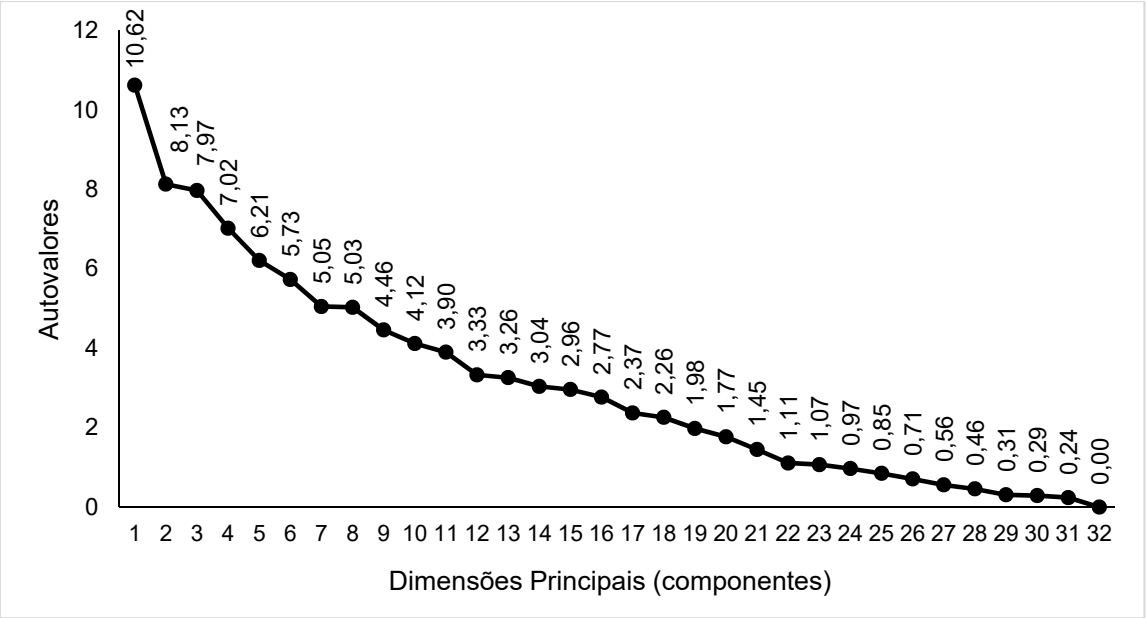


Figura 7. Autovalores (eixo y) e porcentagem da variação original armazenada em cada uma das 32 dimensões principais (eixo dos x).

O gráfico exibido na Figura 8, identifica os descritores mais correlacionados com cada dimensão. Onde as correlações ao quadrado entre os descritores e as dimensões foram usadas como coordenadas. O eixo horizontal mede a contribuição dos descritores para o primeiro componente (Dim1) do MCA, o que explica 10,6% do total da variabilidade dos descritores (que têm x categorias cada), enquanto o eixo vertical mede a contribuição para o segundo componente mais importante (Dim2), explicando 8,1% do total variabilidade. Pode-se observar que, os descritores CP, CC e Cfher são as mais correlacionadas com a dimensão 1 (Dim1). Da mesma forma, os descritores DFI, DI e Ffrut são os mais correlacionados com a dimensão 2 (Dim2).

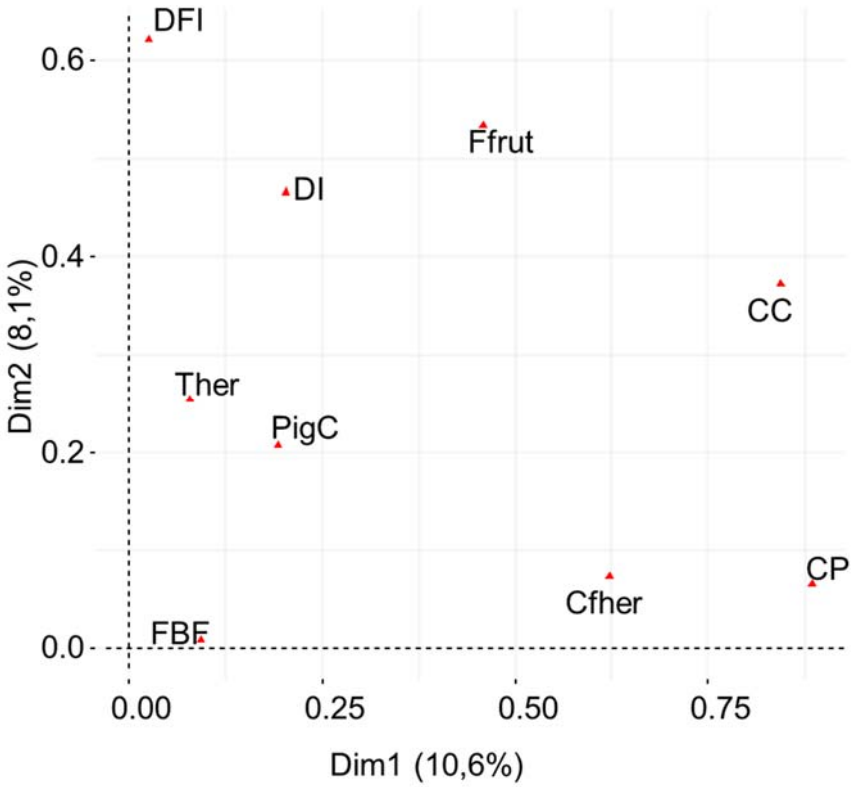


Figura 8. Descritores mais correlacionadas nas dimensões 1 e 2. Cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI), densidade de flores nas inflorescência (DFI), coloração dos lóbulos da corola, formato dos bordos foliares (FBF).

Os relacionamentos entre as categorias dos descritores são apresentados abaixo, onde as categorias variáveis com um perfil semelhante são agrupadas, a distância entre os pontos da categoria e a origem mede a qualidade da categoria variável no mapa de fatores, com isso os pontos de categoria que estão longe da origem estão bem representados no mapa de fatores. No gráfico também foi inserido curvas de densidade para visualização prévia das zonas altamente concentradas pelos acessos (Figura 9). Os 50 acessos distribuidos na Figura 10, estão relacionados às categorias variáveis próximas a eles. As categorias variáveis podem ser vistas como características que descrevem os acessos agrupados próximos uns aos outros.

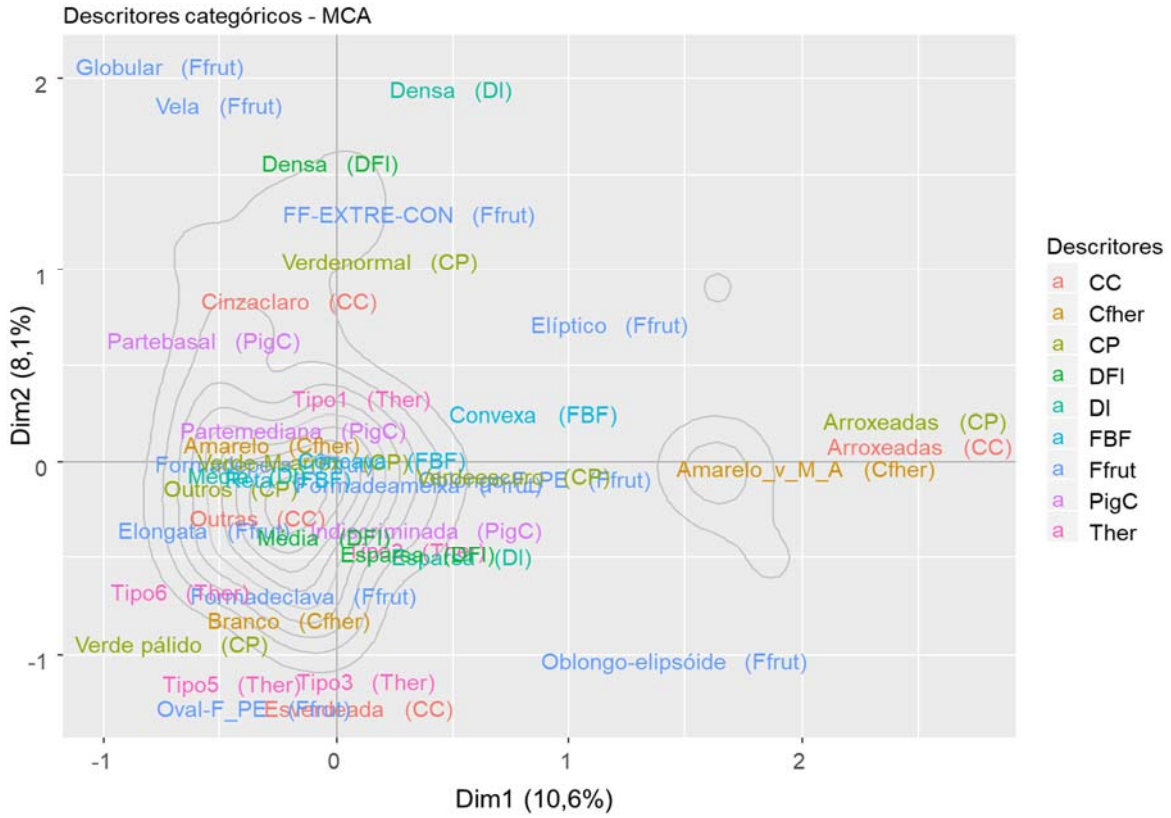


Figura 9. Relacionamentos entre as categorias dos descritores qualitativos e as dimensões e curvas de densidade das zonas mais concentradas pelos acessos de mamão avaliados.

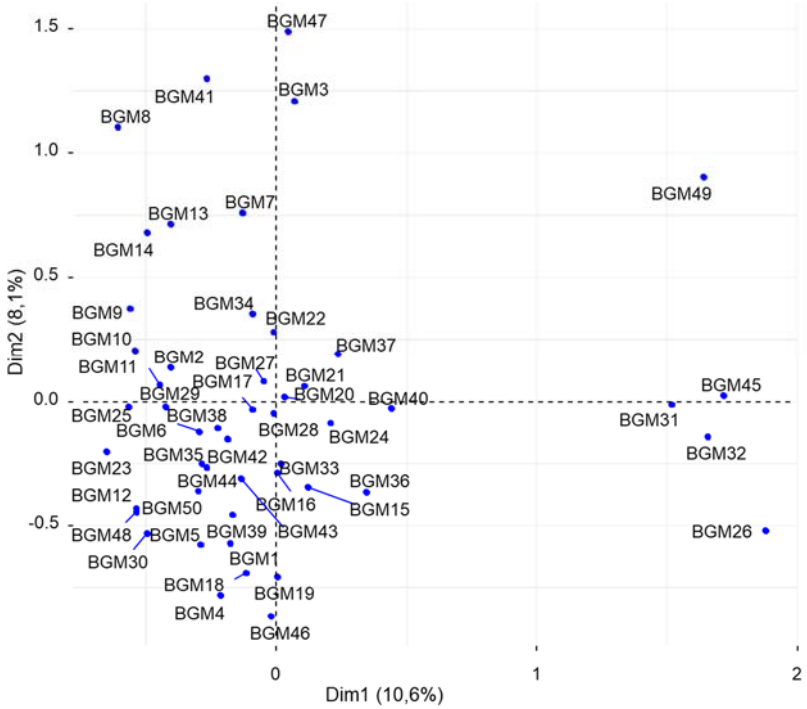


Figura 10. Distribuição dos acessos nas duas primeiras dimensões obtidas por meio da MCA.

No agrupamento por meio da MCA combinado com o algoritmo HCPC, observa-se que houve a formação de 4 grupos bem definidos pelas dimensões (variáveis estatísticas) que mais se correlacionaram com os descritores qualitativos e suas respectivas categorias (Figura 11). O resultado do número ótimo de grupos indexado pelo ganho de inércia interna foi concordante com o resultado obtido pelo índice de Calinski-Harabasz (CH), propiciando o melhor agrupamento.

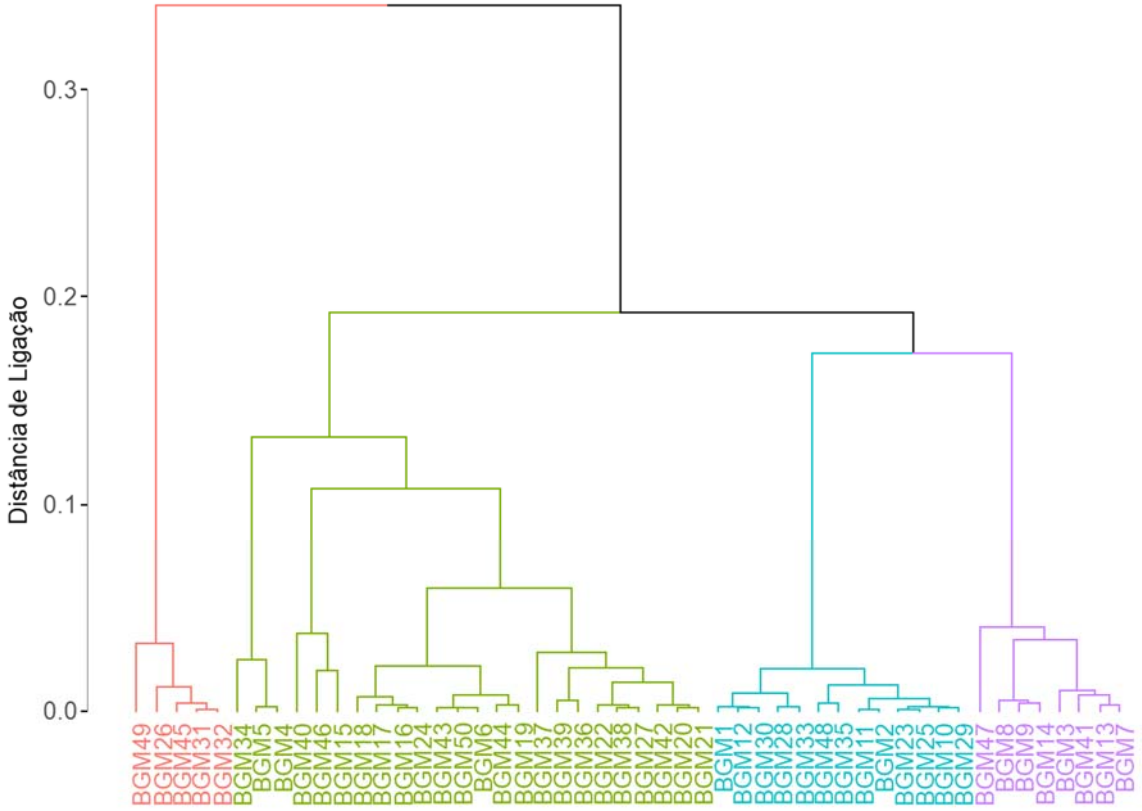


Figura 11. Dendrograma obtido por meio da análise de correspondência múltipla (MCA) combinado com o algoritmo HCPC.

O grupo 1 (G1) é formado por quatorze acessos, BGM1, BGM2, BGM9, BGM10, BGM11, BGM12, BGM23, BGM25, BGM28, BGM29, BGM30, BGM33, BGM35 e BGM48 (Figura 11). Esse grupo tem maior influência significativa das dimensões 1 e 3 (Tabela 5), que por sua vez estão mais bem correlacionadas com os seguintes descritores e suas respectivas categorias (Tabela 4). Esse grupo é caracterizado principalmente por conter 85,71% de todos os acessos avaliados que possuem o formato do fruto elongata, 63,16% dos acessos avaliados que possuem

a Cor do Pecíolo categorizado como “Outros”, cerca de 56,25% dos indivíduos avaliados que tem a Pigmentação do caule categorizada como Parte basal, 44,44% de todos os acesso que possuem a Densidade da inflorescência do tipo Média e a mesma porcentagem para a Cor do caule categorizada como “Outras”, e 80% dos acessos estudados que possuem o Tipo de hermafroditismo (Ther), na categoria (Tipo 6), essa categoria apresenta poucas flores hermafroditas perfeitas e muitas carpelóides e pentândricas.

Já o grupo 2 (G2) tem maior influência significativa da dimensão 2 (Tabela 5), esse grupo é constituído por sete acessos: BGM3, BGM7, BGM8, BGM13, BGM14, BGM41 e BGM47 (Figura 11). Onde os descritores e as respectivas categorias que mais o caracteriza são (Tabela 4): DFI (Densa=70,00%), DI (Densa=80,00%), CC (Cinza claro=42,86%), PigC (Parte basal=37,50%) e Ffrut (Vela=100%), ou seja, esse grupo é caracterizado principalmente por possuir todos os acessos avaliados no estudo com formato do fruto em forma de vela (100%), 80% do total de acessos avaliados que possuem densidade da inflorescência categorizada como densa e 70% dos acessos estudados que possuem densidade das flores na inflorecencias categorizada também como densa.

O grupo 3 (G3) é formado por vinte e quanto acessos: BGM4, BGM5, BGM6, BGM15, BGM16, BGM17, BGM18, BGM19, BGM20, BGM21, BGM22, BGM24, BGM27, BGM34, BGM36, BGM37, BGM38, BGM39, BGM40, BGM42, BGM43, BGM44, BGM46 e o BGM50 (Figura 11). Sendo caracterizado principalmente pelos seguintes descritores e respectivas categorias (Figura 4), PigC (Indiscriminada=65,63%), CP (Verde com manchas arroxeadas=72,73%), DI (Esparsa=72,22%) e CC (Esverdeada=100,00%). Este grupo tem maior influência das dimensões 2, 3 e 5 (Tabela 5). Um descritor pouco influente nas dimensões em relação a esse grupo, porém importante para caracterização, é o formato do fruto com forma de pera, onde esse grupo agrega 71,43% de todos os acessos avaliados que possuem esse formato.

E Por fim, o grupo 4 (G4), que reúne cinco acessos: BGM26, BGM31, BGM32, BGM45 e o BGM49 (Figura 11). Esse grupo possui influência mais significativa da dimensão 1 (Tabela 4), e é cacarterizado por conter todos os acessos avaliados que possuem a Cor do Pecíolo (CP) arroxeados (100%) e todos os acessos estudados com Cor do Caule (CC) arroxeadas (100%), como também

71,43% de todos os acessos que possuem a cor da flor hermafrodita (Cfher) amarelo verde com manchas arroxeadas.

Tabela 4. Descrição dos grupos por meio dos descritores/categorias mais significantes e representativas em cada grupo.

Grupos	Descritor	Categoria	Cla/Mod (%)	p.valor	v.test
G1	Ffrut	Elongata	85,71	6,22E-08	5,412
	CP	Outros	63,16	2,68E-05	4,199
	PigC	Partebasal	56,25	4,24E-03	2,859
	DI	Média	44,44	5,72E-03	2,764
	CC	Outras	44,44	5,72E-03	2,764
	Ther	Tipo6	80,00	1,89E-02	2,348
	PigC	Indiscriminada	12,50	2,03E-03	-3,085
	CP	Verde-M-arrox	4,55	9,64E-04	-3,301
G2	DFI	Densa	70,00	9,11E-06	4,437
	DI	Densa	80,00	1,44E-03	3,186
	CC	Cinzaclaro	42,86	4,00E-03	2,879
	PigC	Partebasal	37,50	9,86E-03	2,581
	Ffrut	Vela	100,00	2,29E-02	2,276
	Ther	Tipo1	24,24	2,59E-02	2,228
	PigC	Indiscriminada	6,25	2,11E-02	-2,306
	DI	Esparsa	0,00	1,96E-02	-2,334
	DFI	Média	3,45	7,04E-03	-2,695
G3	PigC	Indiscriminada	65,63	0,00020	3,713
	CP	Verde-M-arrox	72,73	0,00102	3,286
	DI	Esparsa	72,22	0,00687	2,703
	CC	Esverdeada	100,00	0,03845	2,070
	DI	Densa	0,00	0,03810	-2,074
	CP	Arroxeadas	0,00	0,03810	-2,074
	CC	Arroxeadas	0,00	0,03810	-2,074
	CP	Outros	26,32	0,03399	-2,120
	DFI	Densa	10,00	0,01214	-2,508
	Cfher	Amarelo_v_M_A	0,00	0,00889	-2,616
	Ffrut	Elongata	14,29	0,00572	-2,764
	PigC	Partebasal	6,25	8,65E-05	-3,926
G4	CP	Arroxeadas	100,00	4,72E-07	5,037
	CC	Arroxeadas	100,00	4,72E-07	5,037
	Cfher	Amarelo_v_M_A	71,43	9,91E-06	4,419
	CP	Verde-M-arrox	0,00	4,64E-02	-1,992
	CC	Outras	0,00	1,59E-02	-2,412
	Cfher	Amarelo	0,00	3,74E-04	-3,558

Cor do caule (CC), pigmentação do caule (PigC), cor do pecíolo (CP), coloração das flores hermafroditas (Cfher), formato dos frutos (Ffrut), tipo de hermafroditismo (Ther), densidade da inflorescência (DI), densidade de flores nas inflorescência (DFI), coloração dos lóbulos da corola, formato dos bordos foliares (FBF). Grupo 1 (G1), Grupo 2 (G2), Grupo 3 (G3) e Grupo 4 (G4). Porcentagem total dos acessos avaliados que possuem determinada categoria e que compõe o grupo. Porcentagem de acessos com essa categoria que estão neste grupo (Cla/Mod). Teste de desvio da média da população (v.test).

Tabela 5. Dimensões principais mais associadas aos grupos.

Grupos	Dimensões	v.test	Média na categoria	Média geral	p.value
G1	Dim.1	-2,861	-0,403	-8,84E-17	4,22E-03
	Dim.3	-4,609	-0,562	1,09E-17	4,04E-06
G2	Dim.2	5,410	0,952	-4,64E-17	6,29E-08
G3	Dim.3	4,022	0,331	1,09E-17	5,8E-05
	Dim.5	-2,377	-0,173	-3,94E-17	0,01747
	Dim.2	-3,079	-0,256	-4,64E-17	0,00208
G4	Dim.1	6,378	1,680	-8,84E-17	1,79E-10

Grupo 1 (G1), Grupo 2 (G2), Grupo 3 (G3) e Grupo 4 (G4). Dimensão (Dim.). Teste de desvio da média da população (v.test).

Não há uma estratégia universal e unânime para avaliar agrupamentos de acessos, por isso devem ser utilizadas técnicas que estiverem disponíveis e no domínio do pesquisador, buscando assim métodos de agrupamento distintos, pois como foi observado no presente estudo, os resultados apresentam maior robustez estatística e com isso auxiliam na melhor tomada de decisão, propiciando resultados mais precisos e adequados aos conjuntos de dados em análise.

O pesquisador deve ficar atento e reunir informações que resultem numa avaliação combinada levando em consideração o número ótimo de grupos determinado pelos critérios utilizados, sugestão visual do agrupamento apresentado pelos métodos escolhidos, conhecimento e suposições biológicas da estrutura dos dados e principalmente atrelar os resultados com uma verificação minuciosa do conjunto de dados originais, os quais caracterizam os grupos formados, para que esses grupos reúnam acessos com características semelhantes, sendo o mais homogêneo possivel, como também apresentar maior heterogeneidade entre os grupos formados.

CONCLUSÃO

A estatística de Hopkins foi eficiente em identificar grupos significativos nos conjuntos de dados avaliados.

Os agrupamentos obtidos por meio da análise de componentes principais (PCA) e da análise de correspondência múltipla (MCA) apresentaram grupos bem distribuídos e mais consistentes, com maior semelhança dos acessos dentro dos grupos (homogeneidade) e maior a diferença entre os grupos (heterogeneidade),

apresentando um padrão de agrupamento mais adequado para os conjuntos de dados avaliados.

A MCA é uma ferramenta poderosa para explorar os dados qualitativos com suas respectivas categorias, fornecendo resultados consistentes para formação dos agrupamentos.

Não há um índice universal para obter o número ótimo de grupos, bem como um método de agrupamento que seja ideal para todos os conjuntos de dados. O conhecimento biológico da cultura e dos dados em estudo, atrelado a inspeção visual dos agrupamentos resultantes dos métodos estatísticos, sejam de suma importância, a fim de se obter uma solução final com obtenção de agrupamentos compactos, porém equilibrados, bem separados, consistentes e parcimoniosos.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley interdisciplinary reviews: computational statistics**, v. 2, n. 4, p. 433-459, 2010.

AHMAD, A.; KHAN, S. S. Pesquisa de algoritmos avançados de agrupamento de dados mistos. **Acesso IEEE**, v. 7, p. 31883-31902, 2019.

AIKPOKPODION, P. O. Assessment of genetic diversity in horticultural and morphological traits among papaya (Carica papaya) accessions in Nigeria. **Fruits**, v. 67, n. 3, p. 173-187, 2012.

ARA, N.; MONIRUZZAMAN, M.; BEGUM, F.; KHATOON, R. Genetic divergence analysis in papaya (Carica papaya L.) Genotypes. **Bangladesh Journal of Agricultural Research**, v. 41, n. 4, p. 647-656, 2016.

ASUDI, G. O.; OMBWARA, F. K.; RIMBERIA, F. K.; NYENDE, A. B.; ATEKA, E. M.; WAMOCHO, L. S.; ONYANGO, A. Morphological diversity of Kenyan papaya germplasm. **African Journal of Biotechnology**, v. 9, n. 51, p. 8754-8762, 2010.

BROCK, G.; PIHUR, V.; DATTA, S.; DATTA, S. clValid: An R package for cluster validation. **Journal of Statistical Software**, 25(4), March, 2008.

BUDIAJI, W.; LEISCH, F. Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. **Algorithms**, v. 12, n. 9, p. 177, 2019.

BUSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. Introdução à análise de agrupamento. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, São Paulo. Anais ... São Paulo: ABE, p. 105, 1990.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, n. 1, p. 1-27, 1974.

COLE-RODGERS, P.; SMITH, D. W.; BOSLAND, P. W. A novel statistical approach to analyze genetic resource evaluations using *Capsicum* as an example. **Crop Science**, v. 37, p. 1000 - 1002, 1997.

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

DANTAS, J. L. L.; LUCENA, R. S.; VILAS BOAS, S. A. AVALIAÇÃO AGRONÔMICA DE LINHAGENS E HÍBRIDOS DE MAMOEIRO. **Revista Brasileira de Fruticultura**, Jaboticabal, v. 37, n. 1, p. 138-148, Mar. 2015.

DATTA, S.; DATTA, S. Somnath. Comparisons and validation of statistical clustering techniques for microarray gene expression data. **Bioinformatics**, v. 19, n. 4, p. 459-466, 2003.

DIAS, N. L. P.; DE OLIVEIRA, E. J.; DANTAS, J. L. L. Avaliação de genótipos de mamoeiro com uso de descritores agronômicos e estimação de parâmetros genéticos. **Pesquisa Agropecuária Brasileira**, v. 46, n. 11, p. 1471-1479, 2011.

DUDA, R. O.; HART, P. E. Pattern classification and scene analysis. John Wiley & Sons: New York, p.189–225, 1973.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. **Journal of cybernetics**, v. 4, n. 1, p. 95-104, 1974.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of intelligent information systems**, v. 17, n. 2-3, p. 107-145, 2001.

HAN, J.; KAMBER, M. Data Mining: C d h Concepts and Techniques. 2012.

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. **Bioinformatics**, v. 21, n. 15, p. 3201-3212, 2005.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 28, n. 1, p. 100-108, 1979.

HOTELLING, H. Analysis of a complex of statistical variables into principal components.**The Journal Educational Psychology**, Cambridge, v.24, p.498-520, 1933.

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

HUSSON, F.; JOSSE, J.; PAGES, J. Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data. **Applied Mathematics Department**, p. 1-17, 2010.

HUSSON, F.; JOSSE, J.; LE, S.; MAZET, J.; HUSSON, M. F. Package 'FactoMineR'. **Package FactorMineR**, 2019.

FAGUNDES, G. R.; YAMANISHI, O. K. Características físicas e químicas de frutos do mamoeiro do grupo “Solo” comercializados em quatro estabelecimentos de Brasília-DF. **Revista Brasileira de Fruticultura**, Jaboticabal, v. 23, n. 3, p. 541-545, 2001.

FAO. Food and Agriculture Organization. **The State of Food and Agriculture**. Disponível em :<ftp://ftp.fao.org/docrep/fao/009/a0800e/a0800e.pdf>. Acesso em abril de 2017.

FERREIRA, D.F. **Estatística** Multivariada – 2.ed.rev e ampl. – Lavras: Ed. UFLA, 2011.

KAUFMAN, L.; ROUSSEEUW, P. J. Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm. **Y. Dodge, Ed**, p. 405-416, 1987.

KRISHNA, T. S.; BABU, A. Y.; KUMAR, R. K. Kiran. Determination of optimal clusters for a Non-hierarchical clustering paradigm K-Means algorithm. In: **Proceedings of International Conference on Computational Intelligence and Data Engineering**. Springer, Singapore, p. 301-316, 2018.

MACQUEEN, J. Some methods for classifcation and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, v. 1, p. 281–297, 1967.

MARTINS, D. DOS S.; COSTA, A. DE F. S. DA. **A cultura do mamoeiro:** tecnologias de produção.Vitória: Incaper, p. 497, 2003.

MEILĂ, M. Comparing clusterings—an information based distance. **Journal of multivariate analysis**, v. 98, n. 5, p. 873-895, 2007.

MILLIGAN, G. W.; COOPER, M. An examination of procedures for determining the number of cluster in a data set. **Psychometrika**, v. 50, p. 159-179, 1985.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada**. Belo Horizonte: Editora UFMG, 2005.

MOHAMMADI, S. A.; PRASANNA, B. M. Analysis of genetic diversity in crop plants-salient statistical tools and considerations. **Crop Science**, v. 43, p. 1235-1248, 2003.

MYLEVAGANAM, S. The Analysis of Human Development Index (HDI) for Categorizing the Member States of the United Nations (UN). **Open Journal of Applied Sciences**, v. 7, n. 12, p. 661, 2017.

NASCIMENTO A.L.; SCHMILDT, O.; FERREGUETTI, G.A.; KRAUSE, W.; SCHMILDT, E.R.; CAVATTE, P.C. & AMARAL, J.A.T. Genetic diversity of segregating *Carica papaya* genotypes using the Ward-MLM strategy. **Genetics and Molecular Research**, v. 18, n. 2, 2019.

ODONG, T. L.; VAN HEERWAARDEN, J.; JANSEN, J.; VAN HINTUM, T. J.; VAN EEUWIJK, F. A. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data?. **Theoretical and applied genetics**, v. 123, n. 2, p. 195-205, 2011.

PEARSON, Karl. Principal components analysis. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 6, n. 2, p. 559, 1901.

PRIHATINI, R.; BUDIYANTI, T.; NOFLINDAWATI, N. Genetic Variability Of Indonesian Papaya Accessions As Revealed By Random Amplified Polymorphic Dna And Morphological Characterization. **Indonesian Journal of Agricultural Science**, v. 20, n. 1, p. 1-8, 2019.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2018. Disponível em: <http://www.R-project.org/>.

SARAN, P. L.; CHOUDHARY, R.; SOLANKI, I. S.; PATIL, P.; KUMAR, S. Genetic variability and relationship studies in new Indian papaya (Carica papaya L.) germplasm using morphological and molecular markers. **Turkish Journal of Agriculture and Forestry**, v. 39, n. 2, p. 310-321, 2015.

SCHWEIGGERT, R. M.; STEINGASS, C. B.; HELLER, A.; ESQUIVEL, P.; CARLE, R. Characterization of chromoplasts and carotenoids of red-and yellow-fleshed papaya (Carica papaya L.). **Planta**, v. 234, n. 5, p. 1031, 2011.

SILVA, C. A.; NASCIMENTO, A. L.; FERREIRA, J. P.; SCHMILDT, O.; MALIKOUSKI, R. G.; ALEXANDRE, R. S.; ... & SCHMILDT, E. R. Genetic diversity among papaya accessions. **African Journal of Agricultural Research**, v. 12, n. 23, p. 2041-2048, 2017.

SNEATH, P. H. A.; SOKAL, R. R. The comparison of dendrograms by objective methods. ***Taxon***, vol. 11, p. 33-40, 1973.

SRIPADA, S. C.; RAO, M. S. Sreenivasa. Comparison of purity and entropy of k-means clustering and fuzzy c means clustering. **Indian journal of computer science and engineering**, v. 2, n. 3, p. 343-346, 2011.

SUDHA, R.; SINGH, D. R.; SANKARAN, M.; SINGH, S.; DAMODARAN, V.; SIMACHALAM, P. Genetic diversity analysis of papaya (Carica papaya L.) genotypes in Andaman Islands using morphological and molecular markers. **African Journal of Agricultural Research**, v. 8, n. 41, p. 5187-5192, 2013.

YEUNG, K. Y.; HAYNOR, D. R.; RUZZO, W. L. Validating clustering for gene expression data. **Bioinformatics**, v. 17, n. 4, p. 309-318, 2001.

CONSIDERAÇÕES FINAIS

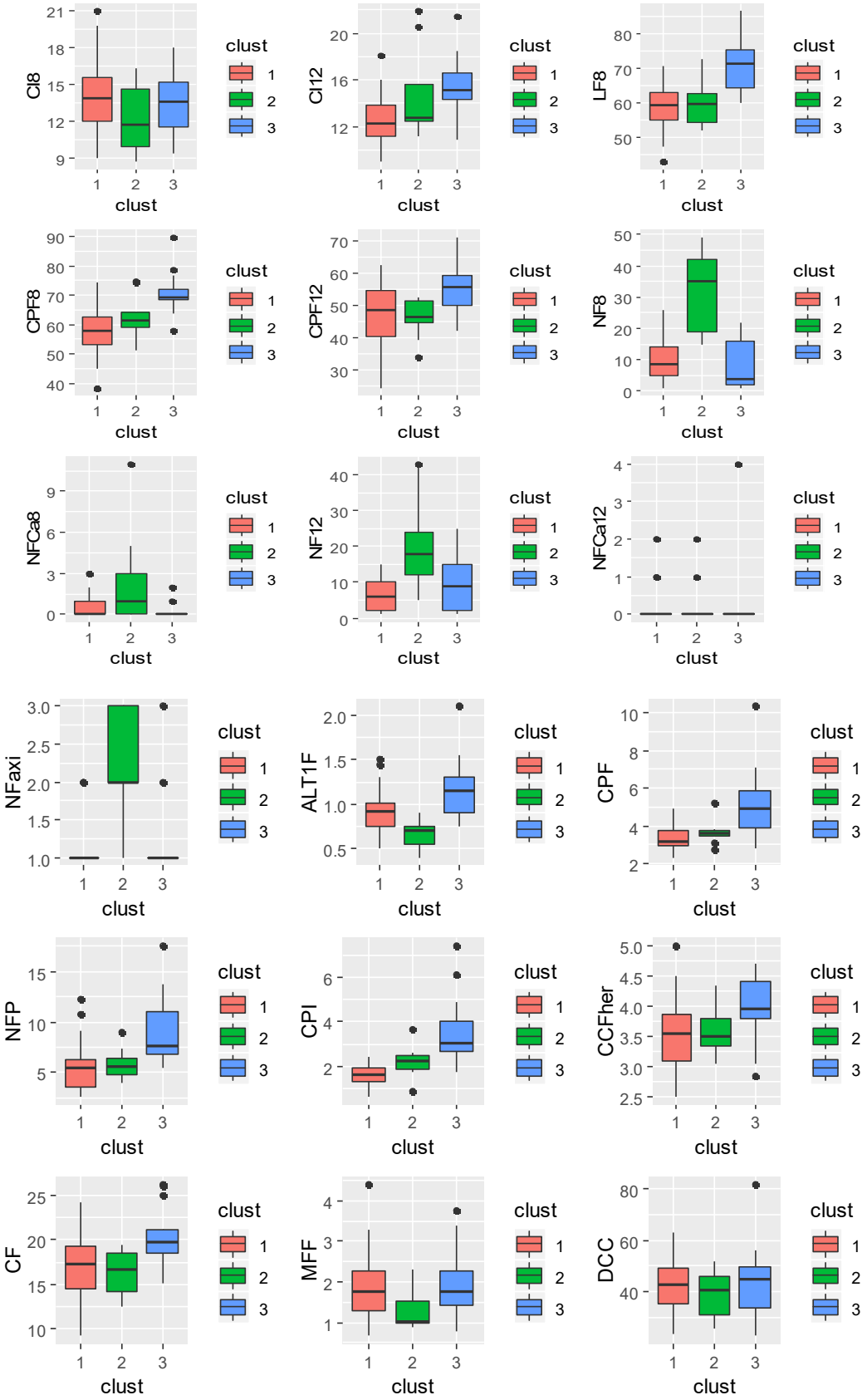
As aplicações desenvolvidas neste estudo salienta a importância de pesquisas relativas ao estudo de novas técnicas direcionadas a seleção de descritores, uma vez que o conhecimento do grau de variabilidade dos descritores e suas relações, possibilita identificar de maneira mais precisa os acessos em estudo.

Atualmente ainda não são encontradas muitas técnicas aplicadas a avaliação da capacidade discriminatória dos descritores qualitativos em acessos de mamão, com o intuito de gerar informações sobre o número mínimo que permitiria a caracterização adequada dos indivíduos.

Espera-se que as informações aqui presentes possam auxiliar os pesquisadores na condução da técnica de análise fatorial exploratória (AFE) aplicada a seleção de descritores qualitativos. Outros estudos com a AFE, com números distintos de acessos e descritores, testes com outros critérios de retenção do número de fatores, com outros métodos de rotação, se fazem necessário para maior implementação da técnica voltada para seleção de descritores.

O uso de novas metodologias ou adaptações e combinações de metodologias, devem ser analisadas com critério, para que se promova uma real otimização dos resultados.

ANEXO



Continua...

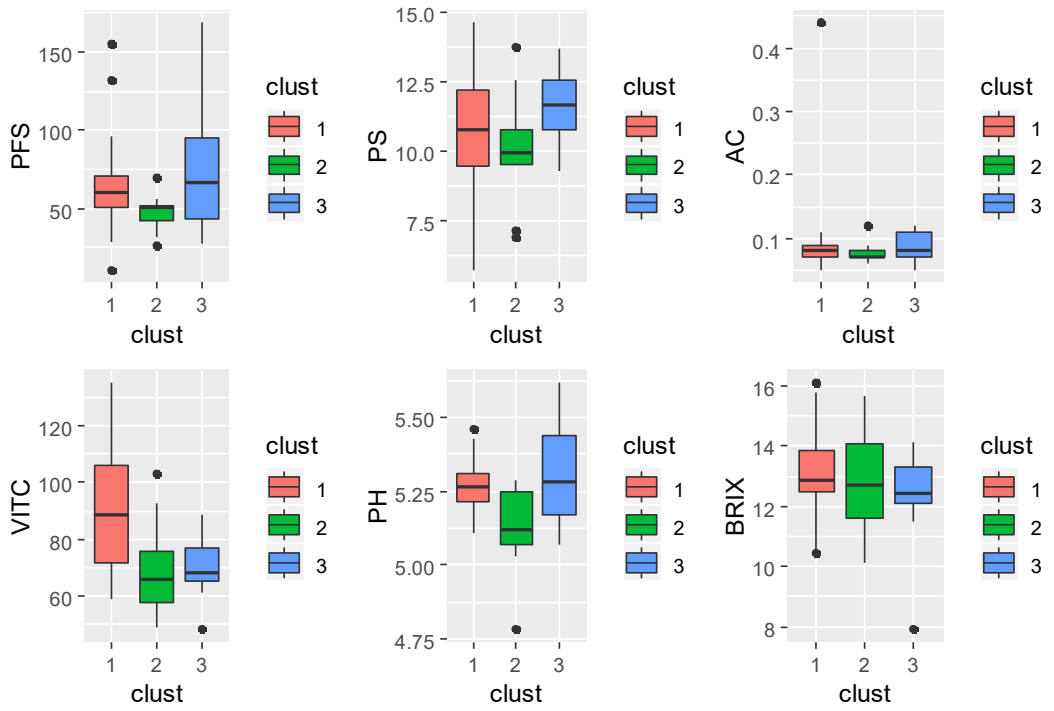


Figura 12: Boxplots apresentando a variação dos dados observados dos 24 descritores quantitativos entre os três grupos obtidos por meio da análise de PCA combinada ao algoritmo HCPC. Comprimento dos internódios 8 meses (CI8); comprimento dos internódios: 12 meses (CI12); Largura da folha: 8 meses (LF8); comprimento do pecíolo da folha: 8 meses (CPF8); comprimento do pecíolo da folha: 12 meses (CPF12); nº frutos: 8 meses (NF8); nº frutos carpelóides: 8 meses (NFCa8); nº frutos: 12meses (NF12); nº frutos carpelóides: 12 meses (NFCa12); nº frutos por axila (NFaxi); altura dos primeiros frutos (ALT1F); comprimento do pedúnculo do fruto (CPF); nº de flores por pedúnculo (NFP); comprimento do pedúnculo da inflorescência (CPI); comprimento da corola da flor hermafrodita (CCFher); comprimento do fruto (CF); firmeza dos frutos (MFF); diâmetro da cavidade central (DCC); peso fresco de sementes do fruto (PFS); Peso fresco de 100 sementes (PS); acidez (AC); vitamina C (VITC); pH (PH) e sólidos solúveis totais (BRIX).