NATÁLIA CAIXETA BARROSO

CATEGORIZAÇÃO DE DADOS QUANTITATIVOS PARA ESTUDOS DE DIVERSIDADE GENÉTICA

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

VIÇOSA
MINAS GERAIS- BRASIL
2010

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T

B277c
2010

Barroso, Natália Caixeta, 1985-
Categorização de dados quantitativos para estudos de diversidade genética / Natália Caixeta Barroso – Viçosa, MG, 2010.
x, 97f. : il. (algumas col.) ; 29cm.

Inclui apêndices
Orientador: Cosme Damião Cruz.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 73-78.

1. Diversidade genética - Métodos estatísticos - Simulação por computador. 2. Análise por agrupamento. I. Universidade Federal de Viçosa. II. Título.

CDD 22.ed. 576.50285

NATÁLIA CAIXETA BARROSO

CATEGORIZAÇÃO DE DADOS QUANTITATIVOS PARA ESTUDOS DE DIVERSIDADE GENÉTICA

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

APROVADA: 15 de dezembro de 2010.

Prof. Luiz Alexandre Peternelli

Prof. Fabyano Fonseca e Silva
(Co-orientador)

Prof. Pedro Crescêncio Souza Carneiro

Prof. Leonardo Lopes Bhering

Prof. Cosme Damião Cruz
(Orientador)

Dedico....

Aos meus pais Jorge e Vanda;

Aos meus irmãos Roberto, Daniel, Lucas;

À minha segunda mãe, Nininha (*in memoriam*);

E a todos que torceram por mim nessa longa caminhada.

AGRADECIMENTOS

À Universidade Federal de Viçosa, pela oportunidade de cursar o mestrado.

À CAPES, pela concessão da bolsa durante parte do curso.

Ao Professor Cosme Damião Cruz, pela orientação, paciência e por todas sugestões valiosas que me deu.

Ao Professor Fabyano Fonseca e Silva, pela co-orientação e pelas sugestões, prestabilidade e disponibilidade para me ajudar.

Ao Professor Paulo Roberto Cecon, pela co-orientação.

Aos professores Leonardo Lopes Bhering, Pedro Crescêncio Souza Carneiro e Luiz Alexandre Peternelli, pela participação na banca de defesa.

Aos meus colegas de mestrado: Adriana, André, Deiciana, Fernanda, Fernando, Maurício e Tiago pelos ótimos anos que convivemos e estudamos juntos.

Às minhas amigas de coração que sempre estiveram comigo em todos os momentos, me acompanhando nessa longa trajetória, muito obrigada pelo carinho.

Ao Francisco Candido Cardoso, por ter me ensinado os primeiros passos na área científica e por ser um grande amigo e companheiro.

Ao Gabriel, por ter me acompanhado nesses quase três anos de mestrado, pelo apoio, carinho, compreensão e incentivo.

Às minhas queridas tias, pelo apoio, amor e atenção que sempre me deram.

Aos meus queridos irmãos Roberto, Daniel e Lucas pelo companheirismo, amizade e por serem modelos de persistência e luta.

Aos meus pais, Jorge e Vanda, por sempre acreditarem em mim, mesmo nos momentos em que nem eu acreditava mais, pelo amor, dedicação, incentivo, apoio, compreensão e paciência.

A Deus, por me dar força sempre para continuar lutando pelos meus sonhos.

SUMÁRIO

RESUMO

BARROSO, Natália Caixeta, M.Sc., Universidade Federal de Viçosa, dezembro de 2010. **Categorização de dados quantitativos para estudos de diversidade genética.** Orientador: Cosme Damião Cruz. Co-Orientadores: Fabyano Fonseca e Silva e Paulo Roberto Cecon.

O estudo da divergência genética é uma ferramenta importante na identificação de indivíduos geneticamente divergentes que, ao serem combinados, possam aumentar o efeito heterótico na progênie. Uma técnica estatística muito aplicada nesse tipo de estudo é a análise de agrupamento. Entretanto, antes dessa técnica ser empregada, deve ser obtida uma matriz de similaridade (ou distância) entre os genótipos. Essas distâncias podem ser calculadas de diversas maneiras, sendo que diferentes propostas são encontradas na literatura para as variáveis quantitativas, binárias e multicategóricas. A transformação de variáveis quantitativas em multicategóricas pode ser utilizada para facilitar sua caracterização com informações preliminares de grande utilidade. Existem vários métodos para se fazer essa transformação, porém estes precisam ser melhor entendidos para que a perda de informações ocorrida na transformação não prejudique significativamente os resultados da análise. Portanto, este trabalho teve como objetivos: verificar quais desses métodos de categorização de variáveis são eficientes; pesquisar a influência da escolha de diferentes coeficientes de dissimilaridades na análise de agrupamentos, feita a partir de dados simulados utilizando variáveis quantitativas e multicategóricas; e averiguar se alguns métodos hierárquicos agrupam com eficiência os dados simulados. Para isto, foram feitas 50 simulações de dez variáveis quantitativas para vinte genótipos de uma espécie de referência como o milho, cada um com quatro repetições. Estes dados foram transformados em multicategóricos através dos métodos: divisão equitativa da amplitude, percentual equitativo, regra do Quadrado, regra de Sturges e distribuição normal. O número de classes tinha que ser estabelecido para os dois primeiros, no caso, foi utilizado quatro e cinco classes para ambos. Foram utilizadas para construir as matrizes de distâncias,

nos dados originais e multicategóricos, as medidas de dissimilaridade: distância euclidiana, euclidiana média, quadrado da distância euclidiana, distância de Mahalanobis e distância ponderada. Posteriormente, o agrupamento foi feito pelo método do vizinho mais próximo e pela ligação média entre grupos (UPGMA). A eficiência destes foi verificada através das estatísticas de eficiência coeficiente de correlação cofenética, estresse e grau de distorção entre as matrizes fenéticas e cofenéticas. Os resultados mostraram que o método de agrupamento UPGMA foi superior ao método do vizinho mais próximo para todas as medidas de distância utilizadas. As distâncias euclidiana e euclidiana média apresentaram a mesma performance em todas as análises de agrupamento feitas. Além disso, essas duas medidas obtiveram os melhores desempenhos em todos os agrupamentos realizados. Todos os métodos de categorização de dados conseguiram um desempenho satisfatório quando agrupados por UPGMA, exceto o método do percentual equitativo com quatro e cinco classes. Contudo, os dados que possuem suas classes estimadas pela regra do Quadrado apresentaram o dendrograma mais semelhante com o obtido por meio dos dados originais, sendo este, então, o método mais recomendado para se fazer a categorização de dados.

ABSTRACT

BARROSO, Natália Caixeta, M.Sc., Universidade Federal de Viçosa, December, 2010**. Categorization quantitative data for studies of genetic diversity.** Adviser: Cosme Damião Cruz. Co-Advisers: Fabyano Fonseca e Silva and Paulo Roberto Cecon.

The genetic diversity study is an important tool in the identification of genetically divergent individuals, wich, can increase the effect of heterosis in the progeny when combinaded. A statistical technique usually applied in this type of study is the cluster analysis. However, before applying this technique, it must be obtained a similarity matrix (or distance) between the genotypes. These distances can be calculated in several ways, which different proposals are found in the literature for quantitative variables, binary and multicategoric. The transformation of quantitative variables in multicategoric can be used to facilitate their characterization with preliminary useful information. There are quite a few methods to make such changes, but they need to be better understood so that the loss of information occurred in such changes does not damage significantly the results of the analysis. Therefore the purposes of this study are: to determine which of these variables categorization methods are efficient; to research the influence of the choice of different coefficients of dissimilarity in cluster analysis, made from simulated data by using quantitative variables and multicategoric; and to investigate whether some hierarchical methods group efficiently the simulated data. For that, there were made 50 simulations of ten quantitative variables to twenty genotypes of a species of reference as corn, each one with four replications. These data were converted in multicategoric using the following methods: equitable division of amplitude, equitable percentage, square rule, Sturges rule and normal distribution. A number of classes had to be established for the first two methods, which were used four and five classes for both. Were used to create distance matrices, in the original data and multicategoric, the dissimilarity measures: Euclidean distance, the average Euclidean, squared Euclidean distance, Mahalanobis distance and weighted distance. Subsequently, the grouping was done by the method of nearest neighbor and the average linkage between groups (UPGMA). The efficiency of these was verified by the statistics of efficiency

cophenetic correlation coefficient, stress and distortion degree between the phenetic and cophenetic matrices. The results showed that the cluster method UPGMA was superior to method of nearest neighbor for all distance measures used. Euclidean distances and average Euclidean showed similar performance in all cluster analysis done. Moreover, these two measures got the best performance in all groups performed. All methods of data categorization achieved a satisfactory performance when grouped by UPGMA, except the method of equal percentage with four and five classes. However, the data which have their classes estimated by the square rule had the most similar dendrogram when compared to the obtained using the original data, and therefore, this is the recommended method to perform the categorization of data.

1-INTRODUÇÃO

O conhecimento sobre a diversidade genética, seja intra ou interpopulacional, é de grande importância ao melhoramento genético de plantas, pois permite ao melhorista direcionar sua estratégia de seleção e realizar mudanças nos sentidos desejados (Ferreira, 2007). As informações sobre a diversidade genética são utilizadas em programas de melhoramento que visam identificar genitores ou combinações promissoras que possibilitem maior efeito heterótico na progênie e maior variabilidade na população segregante (Kamada, 2005).

Métodos estatísticos multivariados, tais como análise de agrupamento, análise discriminante, análise de fatores e análise de componentes principais, podem ser aplicados nesse tipo de estudo (Cruz et al., 2008). Dentre eles, destaca-se a análise de agrupamentos por ser de fácil interpretação e por não exigir pressuposição inicial quanto a distribuição de probabilidade dos dados. Essa técnica é muito utilizada na área de melhoramento genético em estudos de divergência, além de estudos evolutivos (Meyer et. al, 2004). Entretanto, antes da análise de agrupamento ser empregada, deve ser obtida uma matriz de similaridade (ou de distância) entre os genótipos. Essas distâncias podem ser calculadas de diversas maneiras, sendo que diferentes propostas são encontradas atualmente na literatura para as variáveis quantitativas, binárias e multicategóricas (Cruz et al., 2004).

Considerando que a escolha do coeficiente de similaridade pode influenciar nos resultados dos agrupamentos, estes coeficientes precisam ser melhor entendidos, de forma que os mais eficientes, em cada situação específica, possam ser recomendados e empregados. Outro aspecto relevante é que, apesar de existirem inúmeros artigos publicados com o uso de diferentes coeficientes, geralmente não há relato da razão da escolha pelo uso de determinados coeficientes, mostrando a necessidade de estudos adicionais a esse respeito (Meyer et. al, 2004). Uma forma prática e eficiente de se realizar tais estudos é por meio de simulação de dados.

Assim, foi realizado este trabalho com os seguintes objetivos:

i) pesquisar a influência da escolha de diferentes coeficientes de dissimilaridades na análise de agrupamentos, feita a partir de dados simulados utilizando variáveis quantitativas e multicategóricas;
ii) averiguar se alguns métodos hierárquicos agrupam com eficiência os dados simulados na sua forma original, de distribuição quantitativa, ou na transformada em classes multicategóricas;
iii) verificar quais métodos de transformação de variáveis quantitativas em multicategóricas são eficientes;
iv) avaliar o comportamento de diferentes estatísticas de eficiência de análises de agrupamento hierárquicos;

2-REVISÃO DE LITERATURA

2.1 Diversidade Genética

A diversidade genética pode ser definida como a distância genética entre populações, indivíduos ou organismos, baseada em características morfoagronômicas, fisiológicas, bioquímicas e moleculares (Cruz et al., 2004; Barbé, 2008).

A importância da diversidade genética para o melhoramento está no fato de que cruzamentos que envolvam genitores de bom potencial e geneticamente divergentes são mais apropriados para se conseguir maior variabilidade genética das populações segregantes e alto efeito heterótico nos híbridos (Barbé, 2008; Cargnelutti Filho et al., 2008).

A inferência sobre a divergência genética em um grupo de genitores é feita através de técnicas biométricas podendo ser quantitativa e preditiva. Na primeira, citam-se as análises dialélicas, que possibilitam a determinação da capacidade geral e específica de combinação e a heterose manifestada nos híbridos (Barbé, 2008; Cargnelutti Filho et al., 2008). Porém, a necessidade de avaliações de p genitores e de todas as suas combinações híbridas p(p-1)/2, aliada ao fato de que, em algumas culturas, a polinização é difícil de ser executada, onerosa e com pouca probabilidade de êxito na obtenção de semente híbrida, pode inviabilizar o estudo, principalmente quando o valor de p é elevado (Cruz & Carneiro, 2006). Entre os métodos preditivos, a análise de agrupamento destaca-se, pois reúne, por algum critério de classificação, os genitores em grupos, de forma que exista homogeneidade dentro do grupo e heterogeneidade entre os grupos. Essa classificação é importante para identificar os genótipos divergentes e com maior probabilidade de sucesso nos cruzamentos (Cargnelutti Filho et al., 2008).

A escolha do método mais adequado tem sido determinada pela precisão desejada pelo pesquisador, pela forma como os dados foram obtidos e pela facilidade da análise dos mesmos (Cruz et al., 2004).

Estudos de diversidade genética tem sido realizados a partir de vários tipos de variáveis (Dalirsefat et al.,2009; Abreu et al.,2004;Bento et al.,2007).

Em qualquer aplicação, os objetivos da análise de agrupamentos não podem ser separados da seleção de variáveis usadas para caracterizar os objetos a serem agrupados. Os possíveis resultados estão diretamente ligados à seleção das variáveis usadas.

2.2 Tipos de Variáveis

As variáveis podem ser classificadas em dois grandes grupos: o das variáveis quantitativas e o das qualitativas.

As variáveis quantitativas são aquelas que podem ser medidas em escala real, podendo ser contínuas ou discretas. As variáveis contínuas assumem valores dentro de um intervalo contínuo, tipicamente os números reais. As variáveis discretas podem assumir valores dentro de um espaço finito ou enumerável, representadas por números inteiros (Bussab & Morettin, 2002).

As variáveis qualitativas (ou categóricas) são definidas por várias categorias, sendo chamadas de multicategóricas, ou apenas por duas categorias, variáveis binárias. As primeiras podem ser nominais ou ordinais. As nominais são aquelas que não podem ser hierarquizadas ou ordenadas. Ao contrário das nominais, nas ordinais existe uma ordenação entre as categorias (Cruz & Carneiro, 2006).

Já as variáveis binárias, são, por conveniência, codificadas em 0 e 1. Sendo o 0 utilizado para representar a ausência de um determinado padrão da característica, e o 1 para representar a presença deste padrão.

A adequação do uso de certos tipos de variáveis ainda é questionado. A possibilidade de transformação de um tipo de variável em outro é uma alternativa possível de ser adotada pelo pesquisador, mas suas conseqüências precisam ser investigadas. Estudos realizados por meio de simulação poderiam ser de grande valia na elucidação dessas questões.

2.3 Simulação de dados

Simulação, de acordo com Shannon (1975) é o processo de concepção de um modelo representativo de um sistema real e a condução de experimentos com o objetivo de entender o comportamento deste sistema ou

avaliar diferentes estratégias (dentro dos limites impostos por critérios) para sua operação.

Law & Kelton (1991) classificam os modelos de simulação em três dimensões diferentes:

- Dinâmicos ou Estáticos: modelos dinâmicos são formulados para representarem as alterações de estado do sistema ao longo da contagem do tempo de simulação, enquanto que os modelos estáticos são aqueles que visam representar o estado de um sistema em um instante ou que em suas formulações não se leva em conta a variável tempo;
- Determinísticos ou estocásticos: são modelos determinísticos aqueles que em suas formulações não fazem uso de variáveis aleatórias, enquanto os estocásticos podem empregar uma ou mais;
- Discretos ou contínuos: são modelos discretos aqueles em que o avanço da contagem de tempo na simulação se dá na forma de incrementos cujos valores podem ser definidos em função da ocorrência dos eventos ou pela determinação de um valor fixo, nesses casos só é possível determinar os valores das variáveis de estado do sistema nos instantes de atualização da contagem de tempo; enquanto para os modelos contínuos o avanço da contagem de tempo na simulação dá-se de forma contínua, o que possibilita determinar os valores das variáveis de estado a qualquer instante.

O processo de simulação não necessariamente envolve o uso de computadores (Payne,1982). Porém, com os avanços na área de informática, modernos equipamentos e novas linguagens de programação e de simulação têm permitido empregar a técnica de simulação nas diversas áreas do conhecimento humano, fatos que têm propiciado: projetar e analisar sistemas industriais, avaliar performance de hardware e software em sistemas de computação, analisar desempenho de armas e estratégias militares, determinar freqüência de pedidos de compra para recomposição de estoques, projetar e administrar sistemas de transportes como: portos e aeroportos e configurar sistemas de atendimento em hospitais, supermercados e bancos(Law & Kelton,1991).

Na genética, a simulação tem sido de grande utilidade em estudos de populações, do indivíduo ou do próprio genoma. Ela demanda dos geneticistas o desenvolvimento de modelos biológicos adequados, que retratem da melhor forma possível os fenômenos de interesse, e dos programadores as rotinas para o processamento adequado segundo parâmetros e restrições, para que a influência de certos fatores possa ser avaliada (Cruz et. al,2008).

2.4 Uso de variáveis multicategóricas no melhoramento genético

As variáveis multicategóricas são bastante utilizadas na caracterização de cultivares. Essas variáveis estão relacionadas com particularidades morfológicas e estruturais da planta, além de características que conferem qualidade ao produto comercializado. Essas características sofrem pouca influência do ambiente por serem de herança simples, além disso são fáceis e rápidas de se mensurar (Cruz & Carneiro, 2006).

O uso de descritores multicategóricos, com mais de duas classes por variável, tem sido utilizado como ferramenta para estruturação da variabilidade genética. Técnicas de análises multivariadas, como por exemplo a análise de agrupamento, tem sido empregadas para a quantificação da divergência fenotípica em várias espécies de hortaliças (Barelli et al.,2007; Bertan et al., 2006 ; Pereira et al.,2003; Sudré et al. , 2006).

Coimbra et al. (2001), estudando a divergência genética entre 16 genótipos de milho em relação a sete descritores qualitativos, concluíram que a análise de agrupamento utilizando-se matrizes de dissimilaridade obtidas a partir de dados multicategóricos consiste em uma alternativa viável para se avaliar divergência entre genótipos, sendo constatada divergência entre os genótipos avaliados.

Bento et al. (2007) mostraram que a análise de variáveis multicategóricas se mostrou eficiente no agrupamento dos acessos de pimenta estudados, indicando que seu emprego na quantificação da divergência fenotípica e na identificação de grupos heteróticos, pode auxiliar no manejo do banco de germoplasma e na seleção de acessos para programas de melhoramento genético.

A coleta de dados multicategóricos leva vantagem em relação à coleta de dados moleculares e quantitativos por serem facilmente observados e requererem menos tempo e mão de obra na sua realização (Sudré et al., 2006). Entretanto, os limites estabelecidos entre uma e outra classe podem não ser tão facilmente identificados e, assim, as informações coletadas poderão ser carregadas de certa imprecisão tendo em vista a subjetividade de classificação realizada pelo avaliador.

Na abordagem qualitativa deve-se ter a explicitação da subjetividade da percepção do avaliador. Geralmente os critérios adotados têm como base a expectativa de pesquisadores, muitas vezes mutável e flexível, que deve ser sempre conhecida e que depende fortemente de experiências acumuladas. O avaliador tenta muitas vezes disfarçar a presença, necessária, da subjetividade no desenvolvimento da abordagem qualitativa. Isto, inclusive, pode prejudicar a coleta e análise da informação obtida, principalmente no uso da técnica da observação, em que as reflexões do avaliador necessariamente fazem parte da análise. Outro aspecto a ser considerado é que não se pode negar que a avaliação, por se consistir também na emissão de um juízo de valor, está necessariamente permeada pela visão de quem avalia. Portanto, a explicitação de critérios e de suas conseqüências contidos no processo de avaliação, inclusive na interpretação dos resultados, são os caminhos indicados para deixar visível a especificidade buscada na abordagem qualitativa e a sua validade.

Apesar do exposto, não se pode deixar levar pelo princípio de que apenas o que pode ser expresso em números é permeado com a objetividade exigida para dar cientificidade à avaliação. A análise e conclusões obtidas no processo de avaliação adotando-se a abordagem quantitativa também não estão isentas de imprecisões e subjetividades. O mais importante é ser rigoroso, criterioso e atento na execução da avaliação e deixar sempre claro quais os elementos que foram adotados e que permitiram as conclusões obtidas.

3-MATERIAL E MÉTODOS

3.1 Dados

Para analisar os métodos que medem a divergência genética entre genótipos, foram utilizados dados simulados com parâmetros de características da cultura do milho como referência. Os parâmetros necessários para se fazer uma simulação de um ensaio amostral são a média, o coeficiente de variação e a herdabilidade das características

O milho é um cereal de grande importância econômica e social no mundo, sendo utilizado para alimentação humana e animal, e mais recentemente, como fonte de biocombustível (Almeida,2007). Além disso, é uma das culturas geneticamente mais bem estudadas, o que contribuiu para o melhoramento de suas características agronômicas (Freitas, 2001) e abundância de informação sobre suas características de importância. Acredita-se que a espécie *Zea Mays* L. possua aproximadamente 250 raças. Sabendo que dentro de cada raça podem ser identificadas variedades distintas, pode-se dizer que essa espécie possui grande variabilidade, devido principalmente a hibridação e processos de seleção (Cardoso, 2007). Por isso, torna-se necessário o estudo de divergência genética da cultura do milho e informações das mais diversas características, procedimentos biométricos e natureza experimental estão disponíveis.

As variáveis escolhidas para as simulações foram:

1-Altura de planta (AP): medida em cm, tomada na média de plantas ao acaso da parcela;

2-Altura de espiga (AE): medida em cm, tomada na média cinco plantas ao acaso na parcela;

3-Peso de espiga despalhada (PE): medida em $kg.ha^{-1}$;

4-Comprimento de espiga (CE): medida em cm, tomada na média de cinco espigas colhidas ao acaso no meio da parcela;

5-Diâmetro de espiga (DE): medida em cm, tomada na média de espigas colhidas ao acaso na parcela;

6-Produção de grãos (PG): medida em $g.planta^{-1}$, peso de grãos da parcela dividido pelo número total de plantas da parcela;

7-Prolificidade (PRO): quociente entre o número total de espigas produzidas e o número de plantas de cada parcela;

8-Posição relativa da espiga (PRE): quociente entre a altura média da espiga e a altura média da planta;

9-Florescimento Feminino (FF): número de dias entre a semeadura e o florescimento feminino de 50% das plantas da parcela;

10-Florescimento Masculino (FM): número de dias entre a semeadura e o florescimento masculino de 50% das plantas da parcela;

Os valores das médias, herdabilidades e coeficientes de variação necessários para realizar o estudo de simulação das cinco primeiras variáveis foram retirados de Garbuglio et. al (2009). Os parâmetros das outras cinco características foram retirados de Silva (2002), conforme tabela 1.

Tabela1. Parâmetros das variáveis utilizadas nas simulações.

Variáveis	**Média**	**Herdabilidade**	**CV (%)**
AP	197	80	4,4
AE	105	68	6,72
PE	6987	94	8,4
CE	15	88	4,37
DE	4,4	87	3,15
PG	116,78	69,5	14,23
PRO	1,09	64,5	12,97
PRE	0,54	69,5	4,25
FF	66	84	2,28
FM	65,28	84	2

Fonte: Garbuglio et. al, 2009; Silva (2002).

AP: altura da planta; AE: altura da espiga; PE: peso de espiga despalhada; CE: comprimento de espiga; DE: diâmetro de espiga; PG: produção de grãos; PRO: prolificidade; PRE: posição relativa da espiga; FF: florescimento feminino; FM: florescimento masculino.

3.2 Análises estatísticas

Todas as análises genético-estatísticas foram realizadas com o auxílio do programa GENES (Cruz,2001).

3.2.1 Simulação dos dados

Foram feitas 50 simulações das dez variáveis quantitativas, citadas anteriormente, para vinte genótipos da espécie de referência de milho, cada um com quatro repetições. Todas as simulações foram feitas considerando o efeito do genótipo como aleatório, Gi~N(0, σ_G^2).

De acordo com Cruz (2006), o programa GENES considera o seguinte modelo estatístico para simulação dos dados:

$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$

em que:

Y_{ij}: observação simulada de uma dada característica;

μ: média geral da característica, cujo valor é especificado pelo usuário;

G_i: efeito associado ao i-ésimo genótipo;

B_j: efeito associado ao j-ésimo bloco; e

ε_{ij}: erro aleatório, sendo que $\varepsilon_{ij} \sim N(0,\sigma^2)$

A simulação de dados, considerando a distribuição normal com média e variância conhecida, é feita de acordo com o método de Box-Muller (Box & Muller, 1958). Para isso, são utilizadas as variáveis:

$$x = \sqrt{-2\log_e(RND)V}\,Cos(2\pi RND)$$

e

$$y = \sqrt{-2\log_e(RND)V}\,Sen(2\pi RND)$$

sendo RND um número aleatório uniforme entre 0 e 1. As variáveis x e y, obtidas dessa forma, têm distribuição normal com média zero e variância V. Se o processo de simulação demanda a obtenção de n dados com média (μ) e variância (σ^2), pode-se utilizar a estratégia de gerar cada um destes dados (z) por meio da expressão:

$z = \mu + \frac{1}{2\theta}\sum_{i=1}^{\theta}(x_i + y_i)$

sendo:

$V = 2\theta\,\sigma^2$

em que θ representa a repetibilidade de cada ponto simulado. Quanto maior o valor de θ utilizado, maior a precisão da simulação, porém mais lenta.

A explicação dos demais itens simulados também foram retiradas do manual do GENES (Cruz, 2006).

Simulação dos efeitos de blocos

O quadrado médio de blocos no aplicativo é considerado 1,5 vezes maior que o quadrado médio do resíduo. Como a variância residual (σ^2) é indiretamente estabelecida pelo usuário (a partir do fornecimento da média e do coeficiente de variação experimental), o valor do quadrado médio de bloco (QMB) pode ser estimado. Obtidos estes quadrados médios (QMB e QMR), o componente de variabilidade associado ao efeito fixo de bloco (Φ_b) pode ser estimado através de:

$$\Phi_b = \frac{QMB - QMR}{g}$$

sendo g o número de genótipos simulados.

Sabe-se que em um conjunto de dados contendo n valores, em progressão aritmética, de razão r e média $\overline{X}$, em que o primeiro termo é denotado por X_1 e o último por X_n, a variância é dada por:

$$S^2 = \frac{n(n+1)}{3(n-1)^2}(Xn - \overline{X})^2$$

Assim, para estimar os efeitos de blocos admite-se a existência de b efeitos fixos, cujos valores configuram uma progressão aritmética de razão r, com a particularidade de que $B_1 = -B_b$ e $\overline{B} = 0$. Logo, o valor B_b é estimado por meio de:

$$B_b = \frac{(n-1)\sqrt{3\phi_b}}{\sqrt{n(n+1)}}$$

e os demais efeitos são estabelecidos considerando a razão da progressão aritmética dada por:

$$r = \frac{B_b - B_1}{b - 1}$$

Simulação dos efeitos dos genótipos

Para estimar os efeitos dos genótipos é necessário saber o valor da variância genética, que é obtida a partir das informações sobre o coeficiente de variação experimental (CVe) e a herdabilidade (h^2). Primeiramente, é obtido o valor da variância ambiental:

$$\sigma^2 = \left(\frac{\mu CV_e}{100} \right)^2$$

Sabe-se que:

$$h^2 = \frac{100\sigma_G^2}{\sigma_G^2 + \frac{1}{b}\sigma^2}$$

logo:

$$\sigma_G^2 = \frac{\sigma^2 h^2}{b(100 - h^2)}$$

Dessa forma, são estabelecidos:

Efeito aleatório de genótipos

Neste caso, considera-se que $G_i \sim NID(0, \sigma_G^2)$. Como o valor de σ_G^2 é conhecido, os efeitos podem ser estimados através da função randômica de Box-Miller descrita anteriormente.

Simulação dos erros aleatórios

Neste caso, considera-se que $\varepsilon_i \sim NID(0, \sigma^2)$. Como é fornecido ao programa o CVe e a média da característica, o valor de σ^2 torna-se conhecido, portanto, os erros aleatórios e independentes podem ser também estimados pela função randômica citada no item anterior.

Estabelecimento dos valores fenotípicos

Conhecido o valor da média da característica e dos efeitos envolvidos, os valores fenotípicos, de cada variável, são estabelecidos de acordo com o modelo:

$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$

3.2.2 Transformação de variáveis quantitativas em multicategóricas

As dez variáveis quantitativas criadas foram transformadas em variáveis qualitativas multicategóricas ordinais por meio dos seguintes critérios:

i. Estabelecimento do número de classe: método arbitrário, uso da Regras de Sturges e da regra do Quadrado
ii. Alocação dos intervalos de classes: uso da característica da Distribuição Normal, divisão Equitativa da Amplitude e Percentual Equitativo.

Assim, foram adotados os seguintes procedimentos biométricos:

Estabelecimento do Número de Classes

Arbitrário:

Para a divisão equitativa da amplitude e percentual equitativo é preciso definir previamente o número de classes, foram feitas transformações utilizando dois valores de classes: quatro e cinco.

Regra de Sturges:

O número de classes (K) é definido da seguinte forma:

$K=1 + 3{,}3\log_{10}N$

sendo N o número de observações

Regra do Quadrado:

Neste método o número de classes(K) é definido pela raiz quadrada do número de observações.

K= $\sqrt{N}$

Estabelecimento dos Intervalos de Classes

Depois de definido o número de classes, é preciso delimitar o intervalo de cada classe. Para isso, é necessário saber a amplitude da amostra, que é dada pela diferença entre o valor máximo e o mínimo dos valores amostrados. A amplitude de cada classe é dada pela divisão da amplitude da amostra pelo número de classes.

Para categorizar os dados, também foram utilizados os métodos de Divisão Equitativa da Amplitude, do Percentual Equitativo e da Distribuição Normal. Estes são descritos a seguir.

Divisão Equitativa da Amplitude

Neste método calcula–se a amplitude da amostra e a divide igualmente pelo número de classes pré-definido. Assim, se o valor mínimo é 1 e máximo é 4, a amplitude será igual a 3. Se o número de classes for 3, os intervalos serão: [1,2[, [2,3[e [3,4]. A quantidade de valores dentro da classe é variável e imprevisível.

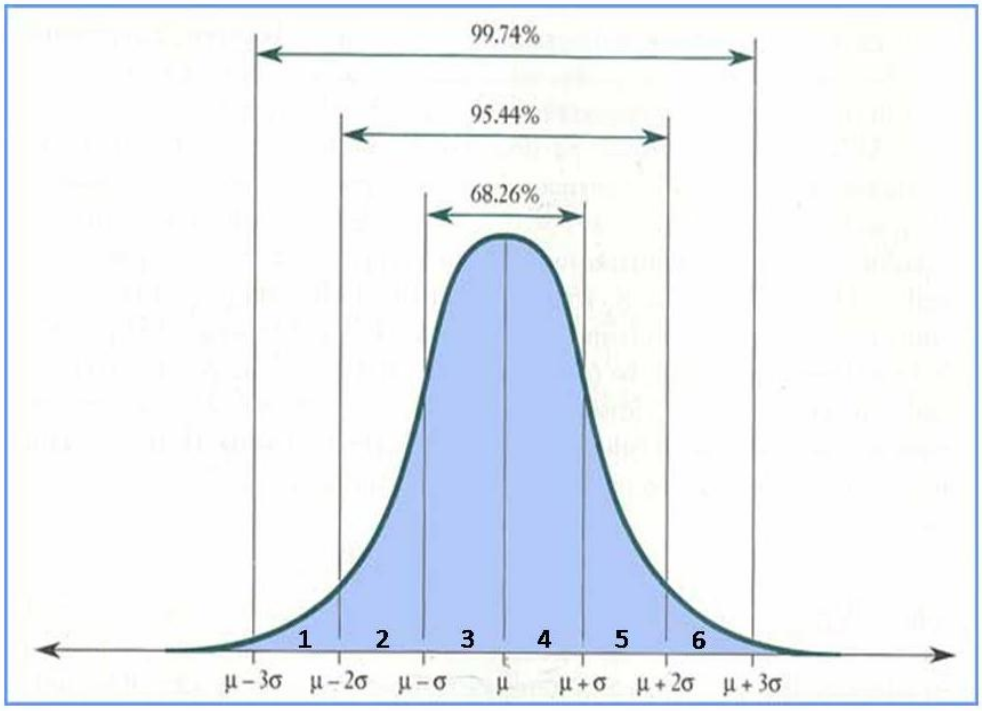
Percentual Equitativo

Neste método adota-se o critério de que todas as classes tenham o mesmo percentual de dados. Esse percentual é definido de acordo com o número de classes escolhido. Assim, se for estabelecido 4 classes, deve-se alocar 25% das observações dentro de cada uma delas e, portanto, a amplitude de variação das classes será variável e dependerá de como os dados estão distribuídos.

Distribuição Normal

O número de classes da transformação por este método foi fixado em seis classes. Esta transformação se baseia na média e no desvio padrão da amostra para construir as classes. A primeira classe é formada pelos valores compreendidos entre a média menos três desvios padrão e a média menos dois desvios padrão, os valores da segunda classe estão entre a média menos dois desvios padrão e a média menos um desvio padrão. A formação das outras quatro classes é mostrada na figura (Figura 1) a seguir. O percentual de valores dentro de cada classe é variável, mas previsível dentro das propriedades da distribuição normal.

Figura 1. Divisão de classes baseada numa distribuição normal.



3.2.3 Medidas de dissimilaridade

As medidas de dissimilaridade são importantes em estudos de diversidade genética, pois identificam genitores possíveis de serem utilizados em programas de melhoramento.

As medidas de dissimilaridade são diferentes para cada grupo de variáveis: quantitativas, binárias e multicategóricas. Porém, serão abordadas

apenas as medidas de dissimilaridade obtidas por variáveis quantitativas e multicategóricas.

Variáveis quantitativas

Para que uma medida de distância seja considerada métrica é necessário satisfazer algumas propriedades (Johnson & Wichern, 1992). Considere os indivíduos (genótipos) i e j e d_{ij} a medida de distância entre eles. Desse modo, têm-se as seguintes propriedades:

i) $d_{ij} \geq 0$

ii) $d_{ij} = d_{ji}$

iii) $d_{ij}=0$ se i=j

iv) $d_{ij} \leq d_{ik} + d_{jk}$ (inequação triangular)

As medidas mais utilizadas para caracteres 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 & Carneiro, 2006).

Distância Euclidiana

Dado que Y_{ij} é a observação no i-ésimo genótipo para a j-ésima característica, define-se que a distância euclidiana entre o par de genótipos i e i' seja representada por meio da expressão:

$$d_{ii'} = \sqrt{\sum_j (Y_{ij} - Y_{i'j})^2}$$

Distância Euclidiana Média

A distância euclidiana média tem sido utilizada de forma alternativa à distância euclidiana, pois o valor desta sempre aumenta com o acréscimo do número de características consideradas na análise. A distância euclidiana média é dada por:

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

sendo v o número de características estudadas.

Quadrado da Distância Euclidiana Média

A dissimilaridade entre dois genótipos pode ser expressa também através do quadrado da distância euclidiana média, dado por:

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

Em todas as distâncias euclidianas, a escala afeta o valor obtido. Além disso, elas são quantificadas em diferentes medidas (porcentagens, comprimento, peso, etc...) sendo, portanto, recomendável padronizar os dados antes de efetuar o cálculo das distâncias.

A padronização é feita do seguinte modo:

$$y_j = \frac{Y_j}{\hat{\sigma}_j}$$

em que $\hat{\sigma}_j$ é o desvio padrão associado à j-ésima característica.

Distância Ponderada

Considera-se no cálculo da distância euclidiana, a diferença de precisão de cada variável mensurada, de forma que aquelas que foram medidas com maior precisão contribuam mais para o valor da diversidade entre pares de acessos. Para o cálculo dessa distância, é necessário conhecer a variação residual, que pode ser obtida em análise prévia de variância em modelos estatísticos apropriados, de forma que a distância possa ser calculada da seguinte forma:

$$d^2_{ii'} = \sum_{j=1}^{v}\frac{d_j^2}{\hat{\sigma}_j^2}$$

em que $\hat{\sigma}_j^2$ é o quadrado médio do resíduo associado à j-ésima variável.

Distância Generalizada de Mahalanobis

A distância generalizada de Mahalanobis leva em consideração as variâncias e as covariâncias residuais que existem entre as características mensuradas, possíveis de serem quantificadas quando as avaliações são realizadas em genótipos avaliados em delineamentos experimentais. Essa é uma vantagem em relação às distâncias euclidianas.

Quando se dispõe de informações de ensaios experimentais é possível se obter a matriz de dispersão residual (ψ) e as médias das características. Com base nessas informações, obtêm-se as estimativas das distâncias de Mahalanobis por meio da expressão:

$$D^2_{ii'} = \delta' \psi^{-1} \delta$$

em que:

$D^2_{ii'}$: é a distância de Mahalanobis entre os genótipos i e i';

ψ:matriz de variâncias e covariâncias residuais;

$\delta' = [d_1 \ d_2 \ ... \ d_v]$, sendo $d_j = Y_{ij} - Y_{i'j'}$;

Y_{ij}: é a média do i-ésimo genótipo em relação à j-ésima variável.

Variáveis multicategóricas

No melhoramento vegetal os caracteres multicategóricos são comumente avaliados, principalmente aqueles relacionados com particularidades morfológicas e estruturais da planta, além de se ter grande interesse em certos atributos que conferem qualidade ao produto comercializado, como a coloração, a forma e o sabor do fruto (Cruz & Carneiro, 2006).

Para avaliar as características multicategóricas, pode ser utilizado o índice de coincidência simples ou a distância euclidiana média ponderada, por exemplo. Contudo, esses índices não serão utilizados nas análises, pois como as classes multicategóricas criadas podem ser ordenadas, estabelecendo-se uma escala de valores para elas, é possível tratá-las como variáveis

quantitativas discretas (Sneath & Sokal, 1973). A teoria aprofundada desses índices pode ser encontrada em Cruz et al. (2008).

Após ser feita a transformação das variáveis quantitativas em multicategóricas, todas as medidas de dissimilaridade, citadas acima, foram utilizadas para medir a dissimilaridade entre cada par de genótipos.

Comparação entre as matrizes de distâncias genéticas

As matrizes das variáveis transformadas, construídas para todas medidas de dissimilaridade, são comparadas com as obtidas pelos dados originais pelo teste de Mantel (Manly,1997).

O valor Z de Mantel é dado por:

$$Z = \sum_{i,j=1}^{n} X_{ij} Y_{ij} \ ,$$

onde X_{ij} e Y_{ij} são elementos das matrizes X e Y a serem comparadas. A significância desse valor de Z pode ser obtida comparando-se esse valor observado com valores de uma distribuição sob hipótese nula, recalculando-se os valores de Z diversas vezes, aleatorizando, em cada uma delas, a ordem dos elementos de uma das matrizes. Este Z calculado após permutações aleatórias é chamado de Z randômico (Z_{rnd}). A estatística Z possui uma relação monotônica com o r de Pearson entre as matrizes (correlação matricial), de modo que ela é de fato utilizada para testar a significância do r (Manly, 1997).

A correlação calculada pelo teste de Mantel varia de -1 a +1, e mede a correlação entre duas matrizes com relação ao Z randômico (Z_{rnd}). Para valores negativos de r, quanto menor a freqüência do $Z_{rnd} \leq Z_{obs,}$ maior a correlação entre as duas matrizes. Para valores positivos de r, quanto menor a freqüência de $Z_{rnd} \geq Z_{obs,}$ maior a correlação.

Neste trabalho, 100 permutações aleatórias foram utilizadas para se testar a significância das correlações matriciais.

3.2.4 Análise de agrupamento

Segundo Cruz et al. (2004), as técnicas de agrupamento permitem a divisão de um grupo em vários menores, de tal forma que exista homogeneidade dentro dos grupos e heterogeneidade entre eles. Esta técnica depende de medidas de dissimilaridade estimadas previamente, sendo a mais utilizada pela comunidade científica a distância generalizada de Mahalanobis (Barbé, 2008).

Existe um grande número de métodos de agrupamento, que se distinguem pelo tipo de resultado a ser fornecido e pelas diferentes formas de definir a proximidade entre um indivíduo e um grupo já formado ou entre dois grupos quaisquer (Faria, 2009). Entretanto, os mais utilizados no melhoramento de plantas são os métodos de otimização e os hierárquicos (Cruz & Carneiro, 2006).

As técnicas hierárquicas são as mais amplamente difundidas (Siegmund et al., 2004) e envolvem basicamente duas etapas. A primeira se refere à estimação de uma medida de similaridade ou dissimilaridade entre os indivíduos e a segunda, à adoção de uma técnica de formação de grupos (Santana & Malinovski,2002).

Segundo Cruz et. al (2004), existem várias formas de representar esta estrutura de agrupamento, dentre elas o método do vizinho mais próximo e o método UPGMA (ligação média entre grupos).

No método do vizinho mais próximo, as conexões entre genótipos e grupos ou entre grupos são feitas por ligações simples entre pares de genótipos, ou seja, a distância entre os grupos é definida como sendo aquela entre os genótipos mais parecidos entre esses grupos. Os dendrogramas resultantes deste procedimento são geralmente pouco informativos, devido à informação dos indivíduos intermediários que não são evidentes (Meyer et. al, 2004).

O método UPGMA é o mais utilizado em diversidade quando se trabalha com populações silvestres e tendo vantagem sobre os demais métodos por considerar médias aritméticas das medidas de dissimilaridade, o que evita caracterizar a dissimilaridade por valores extremos entre os indivíduos considerados (Cruz & Carneiro, 2006; Meyer et. al,2004).

Nos métodos de otimização realiza-se a partição do conjunto de indivíduos em subgrupos não vazios e mutuamente exclusivos por meio da minimização ou maximização de alguma medida predefinida. Um dos métodos mais comumente utilizados na área de melhoramento genético é o proposto por Tocher (Cruz & Carneiro, 2006).

Porém, no presente trabalho, não foi utilizado o método de Tocher, pois os métodos escolhidos para avaliar as técnicas de agrupamento não são adequados para os métodos de otimização.

Métodos hierárquicos

Nos métodos hierárquicos, os genótipos são agrupados por um processo que se repete em vários níveis, até que seja estabelecido o diagrama de árvore ou o dendrograma. Neste caso, o maior interesse está na "árvore" e nas suas ramificações e não no número ótimo de grupos. As delimitações podem ser estabelecidas por um exame visual do dendrograma, em que se avaliam pontos de alta mudança de nível, tomando-os em geral como delimitadores do número de genótipos para determinado grupo (Cruz & Carneiro, 2006).

Serão abordados o método do vizinho mais próximo e o de ligação média entre grupos - UPGMA (Unweighted Pair-Group Method Using Arithmetic Average).

Método do vizinho mais próximo

Neste método, identifica-se, na matriz de dissimilaridade, os indivíduos mais similares, os quais formarão o grupo inicial. Em seguida, são calculadas as distâncias desse grupo aos demais indivíduos e, posteriormente, em relação a outros grupos formados.

A dimensão da matriz de dissimilaridade diminui cada vez que são identificados os grupos ou indivíduos mais similares. Esse processo só se finaliza quando todos os indivíduos estiverem reunidos em um único grupo.

A distância entre um indivíduo k e um grupo formado pelos indivíduos i e j é dada por:

$d_{(ij)k} = \min \{d_{ik}; d_{jk}\}$,

ou seja, $d_{(ij)k}$ é dada pelo menor elemento do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k).

A distância entre dois grupos é dada por:

$d_{(ij)(kl)} = \min \{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$,

ou seja, a distância entre dois grupos formados, respectivamente, pelos progenitores (i e j) e (k e l) é dada pelo menor elemento do conjunto, cujos elementos são as distâncias entre os pares de indivíduos (i e k), (i e l), (j e k) e (j e l).

Método de ligação média entre grupos – UPGMA (Unweighted Pair-Group Method Using Arithmetic Averages).

Neste método, o dendrograma é estabelecido pelos indivíduos com maior similaridade, sendo que a distância entre um indivíduo k e um grupo formado pelos indivíduos i e j é dada por:

$$d_{(ij)k} = \text{média } \{d_{ik}; d_{jk}\} = \frac{d_{ik} + d_{jk}}{2},$$

onde $d_{(ij)k}$ é a média do conjunto das distâncias dos pares (i e k) e (j e k). A distância entre os dois agrupamentos é definida por:

$$d_{(ij)(kl)} = \text{média } \{d_{ik}; d_{il}; d_{jk}; d_{jl}\} = \frac{d_{ik} + d_{il} + d_{jk} + d_{jl}}{4},$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é a média do conjunto, cujos elementos são as distâncias entre os pares de indivíduos (i e k), (i e l), (j e k) e (j e l).

3.2.4.1 Dendrogramas

Antes de construir os dendrogramas, foi feita uma análise dos métodos de agrupamento para cada medida de dissimilaridade de acordo com três estatísticas de eficiência: coeficiente de correlação cofenética, distorção entre a matriz de dissimilaridade e a matriz cofenética, e também o estresse entre essas duas matrizes.

Apenas para os agrupamentos que obtiveram bons resultados é que foram construídos os dendrogramas, ou seja, somente para aqueles que mostraram coeficiente de correlação cofenética alto, nível de estresse e grau de distorção baixos.

Coeficiente de correlação cofenética

O coeficiente de correlação cofenética (CCC) mede o grau de ajuste da matriz de dissimilaridade (matriz fenética) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz cofenética).

O CCC foi obtido por (Bussab et al., 1990):

$$r_{Cof} = r_{FC} = \frac{C\hat{o}v(F,C)}{\sqrt{\hat{V}(F).\hat{V}(C)}}$$

em que:

Côv(F,C): covariância entre os elementos da matriz fenética e cofenética;

$\hat{V}$(F): variância dos elementos da matriz fenética;

$\hat{V}$(C): variância dos elementos da matriz cofenética.

Distorção entre a matriz de dissimilaridade e a matriz cofenética

Esse parâmetro mede a distorção entre a matriz de dissimilaridade e a matriz cofenética.

O grau da distorção (1-α) é obtido, segundo Kruskal, 1964, por:

$$\alpha = \frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n} c_{ij}}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} d_{ij}}$$

em que:

c_{ij}: valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

d_{ij}: valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz de dissimilaridade.

Estresse entre a matriz de dissimilaridade e a matriz cofenética

A estatística do estresse determina a precisão do ajuste obtido com a representação da matriz de dissimilaridade no dendrograma. O estresse foi classificado de acordo com os critérios apresentados na Tabela 2.

Tabela 2. Classificação do estresse.

Nível de estresse(%)	**Ajuste**
40 ├ 100	Insatisfatório
20 ├ 40	Regular
10 ├ 20	Bom
5 ├ 10	Muito Bom
0 ├ 5	Excelente

Fonte: Tabela modificada de Kruskal, 1964.

O valor do estresse foi calculado por:

$$S = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(d_{ij} - c_{ij})^2}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} d_{ij}}}$$

em que:

c_{ij}: valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

d_{ij}: valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz de dissimilaridade.

3.2.5 Análise das variáveis transformadas

As variáveis originais e categorizadas foram submetidas à Análise de Variância (ANOVA) univariada e posteriormente seus resultados foram comparados para avaliar o efeito da transformação sobre as medidas estatísticas. O modelo estatístico utilizado foi

$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$

em que:

Y_{ij}: observação simulada de uma dada característica;

μ: média geral da característica, cujo valor é especificado pelo usuário;

G_i: efeito associado ao i-ésimo genótipo;

B_j: efeito associado ao j-ésimo bloco; e

ε_{ij}: erro aleatório, sendo que $\varepsilon_{ij} \sim N(0,\sigma^2)$.

O esquema da ANOVA está apresentado no Quadro 1.

Quadro 1. Esquema da ANOVA.

Fator de Variação	**GL**	**QM**	**F**
Blocos	r-1	QMB	
Tratamentos	g-1	QMT	QMT/QMR
Resíduo	(r-1)(g-1)	QMR	

Além disso, foram construídos histogramas para mostrar a distribuição dos dados dentro das classes para cada variável de todos os métodos de transformação. Isto foi feito para observar se havia alguma relação entre a distribuição dos dados e a eficiência do método de agrupamento.

Para avaliar a correlação entre as variáveis originais e categorizadas foi utilizado o coeficiente de correlação de Pearson. Este pode ser encontrado através da seguinte fórmula

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde $x_1,x_2,....,x_n$ e $y_1,y_2,...,y_n$ são os valores medidos de ambas as variáveis. Para além disso, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ são as médias aritméticas de ambas as variáveis.

3.2.6 Análise das estatísticas de eficiência dos métodos de agrupamento

Para observar o comportamento das médias e dos coeficientes de variação das estatísticas de eficiência dos métodos de agrupamento, coeficiente de correlação cofenética, estresse e distorção entre as matrizes fenética e cofenética, foram construídos histogramas.

4- RESULTADOS E DISCUSSÕES

4.1 Análise de agrupamento por meio de medidas de dissimilaridade obtidas de variáveis quantitativas contínuas

Método do vizinho mais próximo

As medidas de dissimilaridade que obtiveram os melhores desempenhos dentro do método de agrupamento do vizinho mais próximo, para variáveis quantitativas, foram as distâncias euclidiana e euclidiana média. Nota-se que as duas obtiveram os mesmos resultados (Tabela 3). Isso ocorre porque a distância euclidiana média é igual ao valor da distância euclidiana dividido pela raiz do número de variáveis utilizadas na análise, no caso, dez. Apesar da distância euclidiana média fornecer distâncias menores entre os acessos, ela mantém a mesma relação entre os acessos que foi construída pela distância euclidiana. Ou seja, os acessos que eram mais similares na matriz da distância euclidiana serão também os mais similares na matriz da distância euclidiana média. Portanto, não altera a formação do dendrograma nem a qualidade deste, ou seja, o valor dos parâmetros de eficiência não é alterado.

Analisando apenas o coeficiente de correlação cofenética não se percebe uma diferença muito grande deste entre as medidas de dissimilaridade, tanto que o coeficiente de variação desta estatística não chegou a cinco por cento. Porém, quando se observa o nível de distorção e o de estresse, se torna perceptível a superioridade de desempenho da distância euclidiana e da euclidiana média em relação às demais medidas.

Esses desempenhos não foram satisfatórios, pois o CCC está muito abaixo de 0,7, que é o valor referencial recomendado pelos estudiosos da área. Para o grau de distorção não se tem um parâmetro exato do que é considerado bom, mas uma distorção de mais de 44% não pode ser considerado um resultado interessante. O nível de estresse ficou entre 20 e 40%, o que é considerado um resultado apenas regular.

Tabela 3. Desempenho médio das medidas de dissimilaridade calculadas com os dados originais sendo posteriormente agrupados pelo método do vizinho mais próximo.

Dissimilaridade	*CCC*	*Distorção*	Estresse
Euclidiana	0,6157	44,8155	29,3557
Euclid. Média	0,6157	44,8155	29,3557
Quad. Euclid.	0,5837	72,3503	52,7046
Mahalanobis	0,5611	79,9300	60,9554
Dis. Ponderada	0,5704	78,1841	58,9193
Média	0,5893	64,0191	46,2581
CV (%)	4,3076	27,7318	33,9967

Em alguns casos relatados na literatura, dois ou três coeficientes de similaridade são utilizados com o mesmo conjunto de dados com a expectativa de que se os resultados são robustos, os diferentes coeficientes devem revelar essencialmente o mesmo padrão de diversidade. Se dois coeficientes de similaridade revelam padrões um pouco diferente das relações entre indivíduos, raramente há qualquer justificativa apresentada para sugerir qual modelo é mais válido, e muitas vezes apenas um dos padrões é apresentado na publicação Kosman & Leonard (2005).

Johns et al.(1997) verificaram que diferentes coeficientes de similaridade basicamente, não modificaram o agrupamento de variedades de feijão comum do Chile, em grupos, correspondentes ao Mesoamericano e Andino. Duarte et al. (1999) também provaram que a utilização de diferentes coeficientes de similaridade provocou poucas alterações na classificação dos 27 cultivares de feijão. Esses resultados foram obtidos com base em dados moleculares.

O agrupamento feito a partir da matriz de distâncias de Mahalanabis obteve um desempenho abaixo do obtido com a distância euclidiana. Esse resultado não condiz com o realizado por Machado et al. (2002), que constatou maior eficiência da distância de Mahalanobis em classificar populações segregantes de feijão com maior potencial de variabilidade mesmo quando provenientes de genitores formados por cultivares/linhagens aparentadas. Santos (2005) realizou a seleção de pré-cultivares de soja baseada em índices, utilizando as técnicas de distância euclidiana e Mahalanobis, indicando também

maior eficiência em discriminar as linhagens quando foi utilizada a distância de Mahalanobis.

O resultado apresentado na Tabela 3 também não está de acordo com a consideração apresentada por Krzanowski & Marriot (1995) de que para dados normalmente distribuídos a distância de Mahalanobis seria uma medida adequada. Vale relembrar que, neste trabalho, as variáveis quantitativas foram construídas com base numa distribuição normal. Uma possível alternativa seria utilizar a raiz quadrada da distância de Mahalanobis nos agrupamentos. Isto foi feito apenas para uma das cinqüenta simulações realizadas. Essa medida parece ser eficaz para melhorar o desempenho da distância de Mahalanobis quando agrupada pelo método do vizinho mais próximo. O novo coeficiente de correlação cofenética obtido foi 0,6845, melhorando, portanto, em 0,13 o resultado anterior. A distorção provocada pelo agrupamento diminuiu, passou de 79,93% para 51,76%. O mesmo ocorreu para o fator estresse, que foi reduzido em 26%.

Entretanto, este é apenas um resultado bom obtido com a transformação da distância de Mahalanobis. Mas para se afirmar que essa transformação sempre aumenta a eficácia do agrupamento, teriam que ser feitos mais estudos sobre isso.

Ligação média entre grupos (UPGMA)

Para o método de agrupamento UPGMA, as medidas de dissimilaridade que obtiveram os melhores desempenhos também foram a distância euclidiana e euclidiana média. Novamente, essas duas medidas tiveram o mesmo desempenho. Apesar do CCC não ter sido maior que 0,7, ele se aproxima bastante desse valor. O estresse é considerado um resultado bom e o grau de distorção pequeno.

Nota-se que, para todas as medidas de distância, o desempenho obtido com o agrupamento feito pelo método UPGMA (Tabela 4) foi superior ao obtido pelo método do vizinho mais próximo (Tabela 3).

Esse resultado está de acordo com Gonçalves et al. (2008), que obtiveram melhores resultados com o método de agrupamento UPGMA do que

com o método do vizinho mais próximo, com base no coeficiente de correlação cofenética. Rocha et al. (2009) também encontraram um coeficiente de correlação cofenética mais alto para o método UPGMA que para o método do vizinho mais próximo.

De acordo com Hair Jr. et al. (1995), o método de agrupamento UPGMA leva vantagem em relação ao método do vizinho mais próximo, por não utilizar valores extremos e a formação dos grupos ser baseada em todos os membros deste, em vez de ser baseada em um único par de membros extremos, que é o que ocorre no método do vizinho mais próximo.

O resultado obtido com a distância de Mahalanobis não foi melhor que o da euclidiana, repetindo o resultado encontrado pelo método do vizinho mais próximo. O agrupamento foi refeito com a raiz da distância de Mahalanobis e novamente houve uma melhora nos resultados das estatísticas de eficiência. O coeficiente de correlação cofenética obtido foi 0,7533, tendo uma melhora de 0,9; a distorção que era de 11,5% diminuiu para 2,96% e o estresse também diminuiu, mudou de 33,82% para 17,21%.

Tabela 4. Desempenho médio das medidas de dissimilaridade calculadas com os dados originais sendo posteriormente agrupados pelo método UPGMA.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6925	2,2082	14,8329
Euclid. Média	0,6925	2,2081	14,8329
Quad. Euclid.	0,6580	8,2391	28,6561
Mahalanobis	0,6604	11,4951	33,8237
Dis. Ponderada	0,6527	10,7763	32,7264
Média	0,6712	6,9854	24,9744
CV (%)	2,9237	64,7870	37,8625

4.2 Transformação de dados quantitativos em multicategóricos

Existem vários métodos para transformar uma variável contínua em multicategórica, porém não se tem informações sobre qual método é o mais

eficiente quando se trata de variáveis contínuas, em especial aquelas que seguem distribuição normal. Além disso, na literatura fala que o número de classes criadas com a transformação teria que estar entre cinco e vinte classes (Triola, 2008), não tendo, portanto, um número ótimo de classes definido.

Encontrar um método de transformação que não acarrete em perda grande de informações capaz de alterar os resultados obtidos com as variáveis originais, seria de grande valia para o pesquisador. Por exemplo, se o pesquisador pudesse dividir a variável contínua altura de planta em cinco classes: planta anã, pequena, média, grande e gigante, sem modificar a análise dos dados, economizaria tempo e mão de obra na coleta dessa variável.

Antes de expor o resultado das análises de agrupamento feitas com as variáveis transformadas, será feita uma comparação entre as variáveis quantitativas e as transformadas para ver qual transformação distorce menos os dados originais, considerando diferentes particularidades.

4.2.1 Efeito da transformação sobre medidas estatísticas univariadas

Para esta comparação utilizou-se apenas uma das variáveis simuladas (altura de planta), sem perda de generalidade, procurando verificar o efeito da transformação, no contexto univariado, sobre os dados originais. Assim, comparando-se o resultado obtido pelas ANOVAs das variáveis transformadas com o obtido pelas variáveis originais, descritos na Tabela 5, constata-se que a significância não foi alterada com a transformação dos dados. O processo de transformação certamente altera o valor da média, pois a escala quantitativa de qualquer que seja a amplitude de valores é transformada em escala discreta variando de 1 a c, sendo c o número de classes estabelecido arbitrariamente ou calculado a partir de certos estimadores.

A alteração da razão da variabilidade entre tratamentos e total pode ser verificada pelo coeficiente de herdabilidade. A característica herdabilidade mede a relação entre a variação genotípica e a variação total, que é a variação genotípica acrescida da variação residual. Se a taxa de herdabilidade for alta, significa dizer que a variação residual foi pequena. Portanto, quanto maior a herdabilidade, menor a variação residual. O valor dessa estatística está intimamente ligada à estatística F calculada na ANOVA. Essa estatística mede a relação entre a variância dos tratamentos e a variância residual.

O resultado da Anova feita para a variável altura de planta original e para as variáveis transformadas mostra que os dados que tiveram suas classes estimadas pela regra do Quadrado e pelo percentual equitativo com cinco classes conseguiram maiores valores para a estatística F que os dados originais (Tabela 5). Mas este resultado não se repete para as outras variáveis. Para 50% delas, os dados originais conseguiram maiores valores de F que os dados transformados. Os dados transformados pelo percentual equitativo com cinco classes conseguiram apenas mais um resultado melhor que o obtido pelos dados originais. Os dados com classes estimadas pela regra do Quadrado conseguiram mais dois resultados melhores que os dados originais. Os métodos divisão equitativa da amplitude com quatro e cinco classes, distribuição normal e regra de Sturges conseguiram apenas um resultado melhor que o encontrado para os dados originais. O único que não conseguiu, para nenhuma variável, obter resultado melhor que o dos dados originais foi o percentual equitativo com quatro classes. Porém, para todas as variáveis, exceto para a variável produção de grãos, essa alteração ocorrida no valor de F não alterou a significância deste. Para esta variável a alteração da significância ocorreu nos métodos de transformação divisão equitativa com quatro classes, percentual equitativo com quatro e cinco classes, classes estimadas por Sturges e distribuição normal.

Para a variável em estudo constata-se que a alteração ocorre podendo ampliar ou reduzir esta relação. Assim, isto é um indicativo de que a transformação das variáveis contínuas e discretas poderá afetar o estudo da diversidade genética que é feito no contexto multivariado, mas que a magnitude deste efeito pode não ser tão pronunciado em certos tipos de transformações. Encontra-se, em anexo, os resultados comparativos da análise de variância para as demais variáveis.

Tabela 5. Resultado da análise de variância da variável altura de planta com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de		DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
Variação	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	47,83		0,083333		0,8125		0,233333		0,433333		1,11		0,65		0,033	
Tratamentos	19	405,2	5,25**	1,68	4,04**	2,70	4,7**	3,03	4,13**	5,34	5,32**	9,47	6,18**	5,24	4,88**	2,55	3,89**
Residuo	57	77,13		0,416667		0,575658		0,733333		1		1,53		1,08		0,67	
Média			197		2,53		3,06		2,5		3		5,06		4,04		3,5
CV (%)			4,46		25,56		24,77		34,25		33,39		24,46		25,69		23,14
Herdabilidade			80,96		75,22		78,73		75,77		81,22		83,81		79,49		74,3

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA_ 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE_5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

4.2.2. **Efeito sobre o padrão de distribuição**

Observa-se na Figura 1 que o gráfico dos dados originais construído para a variável altura de planta se assemelha muito ao gráfico da distribuição normal. Este era o padrão esperado, já que todas as dez variáveis consideradas neste estudo foram simuladas baseadas nessa distribuição. Os dados transformados que mantiveram uma distribuição simétrica, semelhante à normal, dos valores dentro das classes foram: divisão equitativa com quatro e cinco classes, classes estimadas pela regra do Quadrado, classes estimadas por Sturges e distribuição normal.

O único método que não manteve uma distribuição simétrica dos dados foi o percentual equitativo com quatro e cinco classes. Isto ocorre porque a freqüência de cada classe tem que ser a mesma neste método, não sendo possível então ter classes com frequências diferentes.

Como os dados quantitativos simulados foram gerados segundo uma distribuição normal, esperava-se que os métodos de transformação de dados que mantivessem um padrão semelhante ao original tivessem bons desempenhos. E foi exatamente o que ocorreu. Somente os métodos de percentual equitativo com quatro e cinco classes não tiveram desempenhos satisfatórios, quando agrupados pelo método do UPGMA.

O mesmo padrão de distribuição de dados pode ser observado para as outras variáveis. Encontra-se, em anexo, os gráficos das distribuições de freqüências das classes de cada método de transformação de dados da segunda até a décima variável.

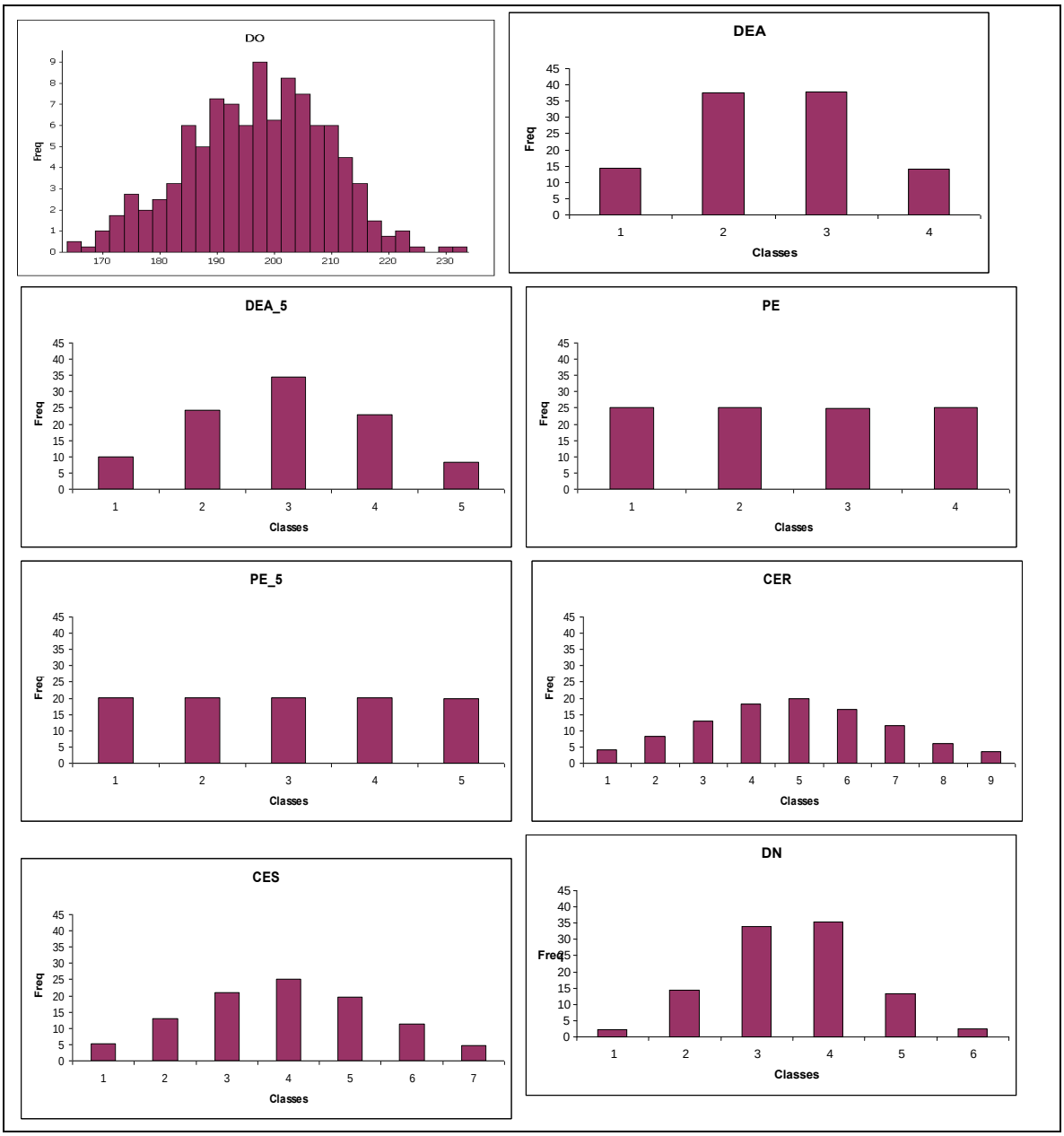


Figura 1. Dispersão dos valores da variável 1 (altura da planta) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

4.2.3. Correlação entre variáveis multicategóricas transformadas e as originais contínuas

Outra maneira de medir a discrepância entre dados originais e aqueles obtidos pela transformação em um conjunto discreto e multicategórico de valores é por meio da associação linear, medida pela correlação de Pearson. Para todas as variáveis, a correlação entre os dados originais e os dados transformados foi alta, independentemente do método de transformação utilizado (Tabela 6). Os dados transformados que possuem suas classes estimadas pela regra do Quadrado e por Sturges obtiveram as maiores correlações com os dados originais, para todas as variáveis (Tabela 6).

Apenas para a variável comprimento de espiga, ocorreram correlações abaixo de 0,9, sendo 0,897 a correlação entre os dados originais e os gerados pelo método da divisão equitativa com quatro classes, e 0,881 a correlação entre os dados originais e os criados pelo método do percentual equitativo com quatro classes. Nota-se que, para todas as variáveis, a correlação encontrada entre os dados originais e os transformados pela divisão equitativa da amplitude com cinco classes é sempre maior que a correlação deste método com os dados originais quando são criadas apenas quatro classes. Este mesmo fato ocorre também com o método percentual equitativo. Portanto, o aumento do número de classes, nesse caso, melhora a relação entre as variáveis transformadas e as originais.

Tabela 6.. Coeficiente de correlação entre os dados originais e os transformados para as dez variáveis.

Variável	*DEA*	*DEA_5*	*PE*	*PE_5*	*CER*	*CES*	*DN*
AP	0,9439*	0,956*	0,9172*	0,9376*	0,9882*	0,9791*	0,9631*
AE	0,9413*	0,9628*	0,9341*	0,9484*	0,9886*	0,9801*	0,9608*
PE	0,9548*	0,9693*	0,9475*	0,9644*	0,9911*	0,9844*	0,9651*
CE	0,897*	0,9388*	0,881*	0,9026*	0,9793*	0,9609*	0,9495*
DE	0,9232*	0,9549*	0,912*	0,9365*	0,9893*	0,9752*	0,9623*
PG	0,9373*	0,9566*	0,9144*	0,931*	0,9877*	0,9755*	0,9677*
PRO	0,9322*	0,9578*	0,9252*	0,9418*	0,9865*	0,9775*	0,9639*
PRE	0,9411*	0,9622*	0,9358*	0,9574*	0,9892*	0,9848*	0,9624*
FF	0,9185*	0,951*	0,9058*	0,9283*	0,9869*	0,9729*	0,9579*
FM	0,9325*	0,9629*	0,9161*	0,9325*	0,9896*	0,9808*	0,962*

AP: altura da planta; AE: altura da espiga; PE: peso de espiga despalhada; CE: comprimento de espiga; DE: diâmetro de espiga; PG: produção de grãos; PRO: prolificidade; PRE: posição relativa da espiga; FF: florescimento feminino; FM: florescimento masculino.
DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
*: significativo a 1% pelo teste t

4.2.4. Efeito sobre os genótipos mais similares e dissimilares encontrados pelos dados originais e transformados

Para o estudo da diversidade genética é fundamental que se avalie as alterações no valor de dissimilaridade (por diferentes medidas de dissimilaridade) entre cada par de tratamentos e, num contexto global, no padrão de agrupamento obtido por diferentes métodos de agrupamento.

Neste estudo consideraram-se, preliminarmente, os valores obtidos na matriz de dissimilaridade estabelecida a partir dos dados originais e sete outras matrizes de dissimilaridade obtidas a partir dos dados transformados, segundo diferentes critérios de transformação. Isto foi feito apenas para a distância euclidiana.

Os genótipos 15 e 16 foram os mais similares entre os dados originais, baseado na distância euclidiana, e 1 e 2 foram os mais divergentes. Apenas o método de transformação do percentual equitativo com quatro classes encontrou resultados diferentes. Neste método, os acessos mais próximos

foram 15 e 16, porém os mais dissimilares foram 4 e 10. Todos os outros métodos de transformação de dados tiveram seus genótipos mais e menos similares iguais aos dos dados originais (Tabela 7).

A distância euclidiana entre 15 e 16 nos dados originais foi de 1,59. As distâncias mínimas dos métodos de transformação variaram de 1,46 a 2,05, sendo o menor valor encontrado nos dados que tiveram suas classes estimadas pela regra do Quadrado, e o maior valor foi obtido pelos dados transformados com base na distribuição normal.

A maior distância euclidiana encontrada nos dados originais foi de 7,87. Nos dados transformados, as maiores distâncias variaram de 6,49 a 7,88, sendo que o menor valor pertence ao percentual equitativo com cinco classes e o maior valor, pertence ao método das classes estimadas pela regra do Quadrado.

Tabela 7. Acessos mais e menos similares das variáveis originais e transformadas, baseado na distância euclidiana.

	Mais similares		Menos Similares	
	Acessos	Distância	Acessos	Distância
DO	15;16	1,59	1;2	7,87
DEA	15;16	1,47	1;2	7,97
DEA_5	15;16	1,56	1;2	7,76
PE	15;16	1,82	4;10	6,49
PE_5	15;16	1,84	1;2	6,49
CER	15;16	1,46	1;2	7,88
CES	15;16	1,8	1;2	7,59
DN	15;16	2,05	1;2	7,68

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA_5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE_5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.

4.3. Padrão de agrupamentos obtidos pelos diversos métodos aplicados ao conjunto de dados multicategóricos e ao conjunto de dados originais contínuos

4.3.1. Divisão Equitativa da Amplitude com quatro classes

Método do vizinho mais próximo

Assim como verificado para a análise com os dados originais de distribuição contínua, a distância euclidiana e euclidiana média obtiveram desempenhos idênticos e se destacaram com relação as demais medidas de dissimilaridade (Tabela 8). Os valores encontrados para as estatísticas de eficiência, distorção e estresse entre a matriz fenética e a cofenética, apresentaram grande diferença entre as medidas de dissimilaridade citadas e as demais. Em contrapartida, o coeficiente de correlação cofenética não apresentou grande diferença de valores entre as medidas de distâncias. O desempenho de nenhuma delas foi satisfatório, pois além do CCC está muito abaixo de 0,7, o grau de distorção foi alto, acima de 44%, e o nível de estresse foi de quase 30%, o que é considerado um resultado regular.

Comparando este resultado com o obtido pelas variáveis originais (Tabela 3) percebe-se que a média do CCC foi a única estatística de eficiência que piorou após a transformação dos dados, passou de 0,59 para 0,57. As médias do estresse e da distorção diminuíram um pouco, mas não o suficiente para acarretar uma mudança relevante no desempenho do agrupamento.

Foi feita também, para os dados transformados, a mudança na distância de Mahalanobis. Os resultados mudaram significativamente. Após tirar a raiz quadrada de Mahalanobis e refazer o agrupamento, o CCC aumentou para 0,69, sendo maior que o obtido pelos dados originais após a modificação, 0,68. A distorção e o estresse diminuíram para 54% e 36%, respectivamente. Mas essas duas estatísticas de eficiência não conseguiram alcançar resultados melhores do que aqueles obtidos pelos dados originais. Além disso, a nova distância de Mahalanobis conseguiu melhor CCC do que a distância euclidiana.

Tabela 8. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do vizinho mais próximo, para os dados transformados pela divisão equitativa da amplitude com quatro classes.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,5926	44,8653	29,3887
Euclid. Média	0,5926	44,8653	29,3887
Quad. Euclid.	0,5565	72,3707	52,7458
Mahalanobis	0,5545	78,8639	59,6853
Dis. Ponderada	0,5319	77,3191	58,0073
Média	0,5656	63,6569	45,8432
CV (%)	4,6774	27,2102	33,2382

Ligação média entre grupos

Novamente distância euclidiana e euclidiana média conseguiram os melhores resultados, além de terem apresentado performances idênticas (Tabela 9). Todas as medidas obtiveram um coeficiente de correlação cofenética razoável, com média de 0,66. Houve uma grande variação de distorção e estresse entre as medidas de dissimilaridade. Sendo o coeficiente de variação do primeiro maior que 60%. Apesar disso, o grau de distorção apresentado por todas as medidas de distâncias foi baixo, menor que 11%. Os resultados obtidos com a distância euclidiana e euclidiana média podem ser considerados satisfatórios, pois o coeficiente de correlação cofenética foi próximo de 0,7, o grau de distorção foi baixo e estresse de quase 15%, o que é considerado bom.

Estes resultados comprovam, mais uma vez, que todas as medidas de dissimilaridade conseguem melhores resultados quando são agrupadas pelo método do UPGMA.

Fazendo uma comparação das estatísticas de eficiência mostradas na Tabela 9 com aquelas obtidas pelos dados originais (Tabela 4), percebe-se que houve uma leve diminuição nas médias do CCC, do grau de distorção e do estresse. Como essa mudança nos valores das estatísticas foi ínfima, pode-se

considerar o resultado obtido pelos dados transformados pela divisão equitativa da amplitude com quatro classes igual ao obtido pelos dados originais.

Agrupou-se também, pelo UPGMA, a raiz quadrada da distância de Mahalanobis. Houve um aumento de 0,10 no CCC, ou seja, o novo valor obtido foi 0,76. O grau de distorção baixou para 3% e o nível de estresse para 17%. Analisando apenas o CCC, a nova distância de Mahalanobis fica com um resultado melhor que o obtido pela distância euclidiana e melhor também do que o CCC da distância de Mahalanobis modificada dos dados originais (0.75).

Tabela 9. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados pela divisão equitativa da amplitude com quatro classes.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6819	2,1825	14,7446
Euclid. Média	0,6819	2,1825	14,7446
Quad. Euclid.	0,6442	8,1843	28,5491
Mahalanobis	0,6559	10,7770	32,6971
Dis. Ponderada	0,6504	9,9879	31,5277
Média	0,6629	6,6628	24,4526
CV (%)	2,6955	62,9844	36,7661

4.3.2 Divisão Equitativa da Amplitude com cinco classes

Método do vizinho mais próximo

As distâncias euclidiana e euclidiana média conseguiram novamente obter o mesmo desempenho para todas as estatísticas de eficiência (Tabela 10). O grau de distorção e o nível de estresse variaram muito entre as distâncias, ficando o coeficiente de variação dessas duas estatísticas em torno de 30%. O coeficiente de correlação cofenética, ao contrário, não variou muito, menos que quatro por cento. Nenhuma medida de dissimilaridade teve um desempenho satisfatório, pois o coeficiente de correlação cofenética de todas

foi muito inferior a 0,7, o grau de distorção foi acima 40% e o nível de estresse variou de 29 a 60%, ou seja, de regular a insatisfatório.

Apesar do desempenho desse agrupamento não ter sido bom, ele foi bem parecido com o obtido com os dados originais. Outra vez, com a transformação, ocorreu uma diminuição na média das três estatísticas de eficiência, mas nenhuma alteração que seja significante.

Outra comparação que pode ser feita é a dos resultados dos dados transformados com quatro (Tabela 8) e cinco classes (Tabela 10) pela divisão equitativa da amplitude. Percebe-se que quando se utiliza cinco classes, ao invés de quatro, os resultados das estatísticas de eficiência melhoram. Esse fato pode ter ocorrido porque a correlação deste método de transformação com os dados originais também aumenta com o aumento do número de classes.

Para este método também foi feita a modificação na distância de Mahalanobis para verificar se ocorreria uma melhora no seu desempenho. O CCC aumentou 0,08, ou seja, passou de 0,56 para 0,63. Houve uma diminuição de 30% e 27% no grau de distorção e no nível de estresse, respectivamente. Mesmo essas duas estatísticas tendo diminuído significativamente, não foi suficiente para serem menores do que as encontradas para a distância euclidiana. Já o novo CCC de Mahalanobis conseguiu ultrapassar o valor obtido pela distância euclidiana.

No caso da nova distância de Mahalanobis, não se pode falar que o aumento do número de classes melhora o desempenho de todas as estatísticas de eficiência. Os dados transformados com quatro classes tiveram o CCC maior 0,05 do que o CCC dos dados transformados com cinco classes. Entretanto, para as outras estatísticas de eficiência, estes conseguiram melhores resultados que os primeiros. O grau de distorção passou de 54% para 48% com o aumento do número de classes, e o nível de estresse passou de 37% para 33%, não alterando a classificação do resultado, que é tido como regular.

Tabela 10. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do Vizinho mais próximo, para os dados transformados pela divisão equitativa da amplitude com cinco classes.

Dissimilaridade	*CCC*	*Distorção*	Estresse
Euclidiana	0,5972	44,5091	29,1267
Euclid. Média	0,5972	44,5092	29,1267
Quad. Euclid.	0,5617	71,9731	52,3526
Mahalanobis	0,5595	79,1706	60,0117
Dis. Ponderada	0,5516	77,4608	58,0963
Média	0,5734	63,5246	45,7428
CV (%)	3,8387	27,6446	33,7277

Ligação média entre grupos

Como era esperado, as distâncias euclidiana e euclidiana média apresentaram resultados idênticos (Tabela 11). O coeficiente de correlação cofenética não variou muito entre as medidas, mantendo uma média de 0,66. Apesar do grau de distorção apresentar um alto coeficiente de variação, maior que 60%, ele se manteve baixo em todas as medidas. O nível de estresse também variou muito, de 14 a 33%, ou seja, de bom a regular.

Os resultados obtidos com as distâncias euclidiana e euclidiana média podem ser considerados satisfatórios, visto que o grau de distorção foi baixo, o nível de estresse foi considerado bom e o coeficiente de correlação cofenética próximo de 0,7.

Quando se compara os resultados obtidos pelos dados transformados com os dos dados originais, percebe-se que nos primeiros houve uma pequena diminuição nos valores das três estatísticas de eficiência em todas as medidas de distância. Mesmo com essas mudanças, os resultados mostrados pelos dois tipos de dados, continuaram sendo muito parecidos entre eles.

Analisando as diferenças de resultados obtidas entre os dados divididos em quatro (Tabela 9) e cinco classes (Tabela 11), nota-se que nos primeiros o desempenho das distâncias de Mahalanobis e distância ponderada foram melhores. Além disso, os dados divididos em quatro classes obtiveram as melhores médias das três estatísticas de eficiência. Por outro lado, os dados

divididos em cinco classes tiveram desempenhos mais altos para as medidas de dissimilaridade mais eficientes. Então, o aumento de número de classes, nesse caso, só pode ser considerado melhor se tiverem sendo comparadas apenas as distâncias euclidiana e euclidiana média.

Para o agrupamento UPGMA, a modificação na distância de Mahalanobis também melhorou os resultados obtidos por essa medida de distância. O CCC passou de 0,65 para 0,7. A distorção diminuiu de 11% para 3%. O nível de estresse também diminuiu, passou de 33% para 17%. Esta nova distância de Mahalanobis conseguiu resultados satisfatórios, porém só teve desempenho melhor que a distância euclidiana no CCC.

A distância de Mahalanobis modificada obteve um resultado pior para o CCC quando os dados foram divididos em cinco classes ao invés de quatro. O CCC passou de 0,76 para 0,7, com o aumento no número de classes. O nível de estresse quase não se alterou, diminuiu 0.20%, e o grau de distorção se manteve o mesmo. Portanto, para a distância de Mahalanobis modificada, o aumento do número de classes não melhorou os resultados das estatísticas de eficiência.

Tabela 11. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados pela divisão equitativa da amplitude com cinco classes.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6851	2,1468	14,6180
Euclid. Média	0,6851	2,1468	14,6180
Quad. Euclid.	0,6469	8,0561	28,3209
Mahalanobis	0,6486	11,0931	33,2200
Dis. Ponderada	0,6481	10,1973	31,8544
Média	0,6628	6,7280	24,5263
CV (%)	3,0785	64,2861	37,5921

4.3.3 Percentual Equitativo com quatro classes

Método do vizinho mais próximo

As distâncias euclidiana e euclidiana média obtiveram os melhores resultados, porém nem estes foram satisfatórios. O nível de estresse e o grau de distorção variaram muito entre os coeficientes de dissimilaridade e, em geral, foram bastante altos. O coeficiente de correlação cofenética não apresentou grande variação entre as diversas medidas, porém manteve uma média baixa, de 0,493 (Tabela 12).

Entre todos os métodos de transformação já mostrados, este foi o que obteve os piores CCC quando comparados com os obtidos pelos dados originais (Tabela 3). A média do CCC passou de 0,59 para 0,49. Por outro lado, o grau de distorção e o nível de estresse foram mais baixos nas variáveis transformadas, mas a diferença com os dados originais foi bem pequena. Uma possível explicação para este fato é que o método do percentual equitativo com quatro classes foi o que obteve os menores valores de correlação, para todas as variáveis, com os dados originais. Em outras palavras, parece que as distorções ocorridas nas variáveis pioraram os resultados do agrupamento.

A modificação da distância de Mahalanobis também foi feita com os dados transformados pelo percentual equitativo da amplitude com quatro classes. O CCC passou de 0,47 para 0,48. O grau de distorção diminuiu 20%. O nível de estresse baixou 25%, passando de insatisfatório para regular. Nesse caso, a distância de Mahalanobis modificada não conseguiu ultrapassar o resultado do CCC obtido pela distância euclidiana. Além disso, apenas neste método a nova distância de Mahalanobis não conseguiu obter bom desempenho no CCC.

Tabela 12. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do Vizinho mais próximo, para os dados transformados pelo percentual equitativo com quatro classes.

Dissimilaridade	*CCC*	*Distorção*	Estresse
Euclidiana	0,5214	44,0324	28,7777
Euclid. Média	0,5214	44,0324	28,7777
Quad. Euclid.	0,4747	71,4559	51,8420
Mahalanobis	0,4692	78,5955	59,2072
Dis. Ponderada	0,4511	76,6595	57,1052
Média	0,4930	62,9551	45,1420
CV (%)	6,1812	27,7503	33,6216

Ligação média entre grupos

As medidas de dissimilaridade que obtiveram os melhores desempenhos foram as distâncias euclidiana e euclidiana média. Mais uma vez, essas medidas apresentaram resultados iguais (Tabela 13). O grau de distorção variou bastante entre os coeficientes de dissimilaridade, entretanto sua média permaneceu baixa. Em contrapartida, a média do nível de estresse foi alta, pois a maioria dos coeficientes de dissimilaridade teve um elevado nível de estresse. Não houve muita variação entre os coeficientes de correlação cofenética, porém sua média foi baixa. Devido a esses fatores, nenhuma medida de distância teve uma boa performance.

A média do CCC se alterou bastante após os dados originais terem sido transformados pelo método do percentual equitativo com quatro classes, passou de 0,67 para 0,62. Houve uma diminuição também no grau de distorção e no nível de estresse apresentados pelas variáveis transformadas, porém diminuição no valor dessas variáveis é uma mudança positiva e não negativa, como é o caso do CCC. A alteração ocorrida nessas duas estatísticas de eficiência não foi relevante.

Fazendo uma comparação com os resultados obtidos pelos métodos de transformação, este foi o que obteve o pior CCC. Este fato reforça a idéia de que os dados transformados que sofreram maiores distorções são os que

apresentaram os piores desempenhos em relação à estatística de eficiência CCC.

A distância de Mahalanobis modificada também foi agrupada pelo método UPGMA. O CCC melhorou, passou de 0,62 para 0,66, mas não o suficiente para ficar próximo de 0.7. O grau de distorção diminuiu 7% e o nível de estresse passou de regular para bom. Com esse resultado, a nova distância de Mahalanobis conseguiu obter um melhor CCC do que a distância euclidiana, porém teve desempenho pior nas outras estatísticas.

Tabela 13. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados pelo percentual equitativo com quatro classes.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6483	2,0374	14,2421
Euclid. Média	0,6483	2,0374	14,2421
Quad. Euclid.	0,5991	7,7799	27,8219
Mahalanobis	0,6157	10,0557	31,6190
Dis. Ponderada	0,6121	9,1124	30,1032
Média	0,6247	6,2046	23,6057
CV (%)	3,5875	62,68	36,6603

4.3.4 Percentual Equitativo com cinco classes

Método do vizinho mais próximo

Quando o número de classes foi aumentado para cinco, o desempenho do CCC aumentou 0,07. Entretanto, o CCC continuou baixo. O grau de distorção e o nível de estresse quase não se alteraram com o aumento do número de classes. Portanto, para o agrupamento pelo vizinho mais próximo, o aumento do número de classes proporcionou um melhor desempenho das medidas de distância.

Como já era esperado, as distâncias euclidiana e euclidiana média apresentaram desempenhos idênticos e melhores que as demais medidas (Tabela 14). Em geral, o coeficiente de correlação cofenética não variou muito, porém se manteve baixo em todas as medidas de distâncias. O grau de

distorção foi alto em todos os coeficientes de dissimilaridade. E o nível de estresse variou de regular a insatisfatório.

O CCC médio das variáveis transformadas foi menor 0,02 que o CCC das variáveis originais. A média das outras estatísticas quase não variou. Logo, quando esse método dividiu os dados em cinco classes, o desempenho deste passou a ser mais semelhante ao dos dados originais.

Para este também foi calculada a raiz quadrada da distância de Mahalanobis. Neste caso, a modificação na distância de Mahalanobis não melhorou o desempenho do CCC quando comparada com a média da medida original. A média do CCC era 0,56, com a mudança passou a ser 0,52. Porém, diminuiu em 26% o grau de distorção e o nível de estresse passou de insatisfatório a regular. Por esta razão, a nova distância de Mahalanobis não obteve desempenho melhor que a distância euclidiana em nenhuma das estatísticas.

Entretanto, em relação ao resultado da nova distância de Mahalanobis obtido para este método com quatro classes, o que tem cinco classes levou vantagem apenas no desempenho do CCC. Este aumentou 0,04, ou seja, mudou de 0,48 para 0,52. Mas para as outras estatísticas de eficiência, o aumento do número de classe diminuiu o desempenho do método em menos de 2%. Como a mudança no CCC foi mais relevante, pode-se dizer que o aumento do número de classes também melhorou o desempenho da distância de Mahalanobis modificada.

Tabela 14. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do Vizinho mais próximo, para os dados transformados pelo percentual equitativo com cinco classes.

Dissimilaridade	CCC	Distorção	Estresse
Euclidiana	0,5972	44,5091	29,1267
Euclid. Média	0,5972	44,5092	29,1267
Quad. Euclid.	0,5615	71,9731	52,3526
Mahalanobis	0,5595	79,1706	60,0117
Dis. Ponderada	0,5516	77,4608	58,0963
Média	0,5734	63,5246	45,7428
CV (%)	3,8436	27,6446	33,7277

Ligação média entre grupos

Para o método de agrupamento UPGMA, as medidas de dissimilaridade que obtiveram os melhores desempenhos também foram a distância euclidiana e euclidiana média. Outra vez, essas medidas conseguiram obter performances idênticas (Tabela15). O grau de distorção foi baixo para todos os coeficientes de dissimilaridade, não ultrapassando 12%. Já o nível de estresse foi bom para as duas primeiras medidas, e regular para as demais. O coeficiente de correlação cofenética não variou muito e foi baixo para todos.

Os resultados obtidos com as duas primeiras medidas de dissimilaridade não podem ser considerados satisfatórios, muito menos os resultados das demais medidas, pois apesar de terem conseguido um grau de distorção baixo e um nível de estresse bom, o coeficiente de correlação cofenética foi muito abaixo do considerado ideal.

Com a transformação dos dados, houve uma queda de 0,05 na média do CCC, este passou de 0,67 para 0,62. Nas outras duas estatísticas de eficiência, houve uma pequena melhora nos resultados obtidos quando os dados foram transformados, porém essa alteração foi muito pequena.

O aumento do número de classes não melhorou o desempenho dos dados transformados, por este método, quando agrupados por UPGMA. As médias de todas estatísticas quase não se alteraram, podendo, então, considerar que os dados transformados com quatro e cinco classes tiveram o mesmo resultado.

Para este agrupamento também foi calculada a raiz quadrada da distância de Mahalanobis. A mudança melhorou o desempenho do CCC, este passou de 0,61 para 0,64, porém esse valor ainda ficou distante do ideal, 0,7. O grau de distorção diminuiu 9% e o nível de estresse mudou de regular para bom. Os resultados das estatísticas da nova Mahalanobis ficaram bem semelhantes aos resultados apresentados pela distância euclidiana, mas não superiores.

Não houve melhora no desempenho da nova distância de Mahalanobis quando o número de classes foi aumentado de quatro para cinco. O CCC diminui 0,02 no seu rendimento com o aumento do número de classes. O grau

de distorção se manteve em 3% e o nível de estresse também continuou bom. Contudo, nota-se que o aumento do número de classes manteve o desempenho do estresse e da distorção, porém diminuiu o desempenho do CCC.

Tabela 15. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados pelo percentual equitativo com cinco classes.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6444	2,0568	14,3152
Euclid. Média	0,6444	2,0568	14,3152
Quad. Euclid.	0,6007	7,7535	27,7911
Mahalanobis	0,6111	11,8813	32,8803
Dis. Ponderada	0,6095	9,3565	30,4660
Média	0,6220	6,621	23,9536
CV (%)	0,0335	66,7381	37,4927

4.3.5 Classes estimadas pela Regra do Quadrado

Método do vizinho mais próximo

As distâncias euclidiana e euclidiana média conseguiram o mesmo desempenho nas três estatísticas de eficiência (Tabela 16). Mais uma vez, essas duas medidas alcançaram os melhores resultados, porém estes não foram satisfatórios. O coeficiente de correlação cofenética foi abaixo de 0,62 para todas medidas. O nível de estresse variou de regular, das duas primeiras, a insatisfatório para as demais. O grau de distorção variou muito, porém se manteve acima de 40% em todos os coeficientes de dissimilaridade.

Este método de transformação de dados foi o que obteve resultados mais semelhantes aos obtidos pelos dados originais. O CCC dos dados originais e transformados foi 0,59. Eles também tiveram o mesmo desempenho no grau de distorção e estresse, quando os resultados foram arredondados. Portanto, pode ser considerado que os dados originais e os transformados pela

regra do Quadrado tiveram resultados idênticos. Este fato pode ter ocorrido porque este método de transformação de dados obteve os maiores coeficientes de correlação com os dados originais para todas as dez variáveis

Foi feita também para este método a alteração na distância de Mahalanobis. O CCC aumentou 0,17 no seu desempenho, ou seja, passou de 0,56 para 0,73. O grau de distorção diminuiu 27% e o nível de estresse mudou de insatisfatório para regular. Mesmo o CCC estando acima de 0,73, esta nova medida não obteve desempenho satisfatório, pois o grau de distorção continuou alto, acima de 50%, e o nível de estresse foi apenas regular. A nova distância de Mahalanobis só conseguiu resultado melhor que a distância euclidiana no CCC.

Tabela 16. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do Vizinho mais próximo, para os dados transformados por classes estimadas pela regra do Quadrado.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6153	44,9771	29,4864
Euclid. Média	0,6153	44,9771	29,4864
Quad. Euclid.	0,5827	72,3928	52,7713
Mahalanobis	0,5584	80,1290	61,1798
Dis. Ponderada	0,5651	78,1518	58,9482
Média	0,5874	64,1256	46,3740
CV (%)	4,5978	27,6171	33,9001

Ligação média entre grupos

Como já era esperado, as distâncias euclidiana e euclidiana média tiveram resultados idênticos (Tabela 17). Além disso, as duas medidas conseguiram obter os melhores desempenhos dentre todas as medidas de distância. O nível de estresse foi considerado bom para estas e regular para as demais medidas. O grau de distorção foi baixo para todos os coeficientes de dissimilaridade, não ultrapassando 12%. O coeficiente de correlação cofenética foi alto para todas as medidas, mas só os dois primeiros ficaram bem próximos de 0,7.

Pode-se dizer, então, que as distâncias euclidiana e euclidiana média tiveram resultados satisfatórios, já que seus coeficientes de correlação cofenética chegaram bem próximo do ideal, o grau de distorção foi bem baixo e o nível de estresse foi bom.

Os dados agrupados por este método também conseguiram melhor desempenho, entre todos os métodos de transformação, quando agrupados por UPGMA. Os resultados das três estatísticas de eficiência foram praticamente idênticos aos resultados obtidos pelos dados originais. Isso reforça o fato de que quanto maior a correlação dos dados transformados com os dados originais, mais similares serão seus desempenhos.

O CCC da raiz quadrada da distância de Mahalanobis foi 0,12 maior que o CCC da distância original, ou seja, era 0,66 e passou a ser 0,78, melhorando bastante o desempenho dessa estatística. O grau de distorção passou a ser 3% e o nível de estresse passou de regular para bom. Mesmo com essa melhora significativa, a nova distância de Mahalanobis só conseguiu resultado superior ao obtido pela distancia euclidiana no CCC.

Tabela 17. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados por classes estimadas pela regra do Quadrado.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6989	2,1988	14,7893
Euclid. Média	0,6989	2,1988	14,7893
Quad. Euclid.	0,6642	8,1478	28,4775
Mahalanobis	0,6608	11,5647	33,3962
Dis. Ponderada	0,6457	10,9507	32,9806
Média	0,6737	7,0122	24,8866
CV (%)	3,5676	65,2990	37,8403

4.3.6 Classes estimadas por Sturges

Método do vizinho mais próximo

Mais uma vez, os melhores resultados foram obtidos pelas distâncias euclidiana e euclidiana média, que tiveram performances idênticas (Tabela 18). Essas duas apresentaram um coeficiente de correlação um pouco maior que as outras, porém não foi suficiente para alcançar o valor ideal de 0,7. O grau de distorção foi alto para todos os coeficientes de dissimilaridade e o nível de estresse variou muito, sendo regular para as duas primeiras e insatisfatório para as demais.

Logo, pode-se dizer que nem as distâncias euclidiana e euclidiana média, que foram as que tiveram os melhores desempenhos, tiveram resultados satisfatórios.

A média do CCC das variáveis transformadas foi menor que a média obtida pelos dados originais. A média destes foi 0,59, enquanto a média dos primeiros foi 0,57. A média do grau de distorção dos dados transformados foi menor 1% comparada com a média dos dados originais. E o nível de estresse se manteve o mesmo. Apenas no CCC que ocorreu uma maior alteração no resultado após os dados terem sido transformados.

Para estes dados transformados também foi calculada a raiz quadrada da distância de Mahalanobis. Com a mudança, o CCC aumentou para 0.66, o grau de distorção diminuiu para 52% e o nível de estresse passou de insatisfatório para regular. Apesar da grande melhora ocorrida nos resultados, este não pode ser considerado satisfatório, pois nenhuma das estatísticas de eficiência obteve um resultado bom. Mas, comparando com os resultados da distância euclidiana, a distância modificada de Mahalanobis obteve um CCC superior ao dela.

Tabela 18. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do Vizinho mais próximo, para os dados transformados por classes estimadas por Sturges.

Dissimilaridade	*CCC*	*Distorção*	Estresse
Euclidiana	0,6062	44,9801	29,4267
Euclid. Média	0,6062	44,9801	29,4267
Quad. Euclid.	0,5739	72,5160	52,7881
Mahalanobis	0,5367	79,7348	60,7213
Dis. Ponderada	0,5466	78,0327	58,6804
Média	0,5739	62,6050	46,2086
CV (%)	5,6566	25,9502	33,7474

Ligação média entre grupos

Para o método de agrupamento UPGMA, as medidas de dissimilaridade que obtiveram os melhores desempenhos também foram as distâncias euclidiana e euclidiana média (Tabela 19). O grau de distorção de todas as medidas foi baixo, não ultrapassando 12%. Somente para as duas primeiras medidas o nível de estresse foi bom, para as outras este foi apenas regular. O coeficiente de correlação cofenética não variou muito, mas se aproximou do ideal apenas nas duas primeiras medidas de dissimilaridade.

Logo, os resultados obtidos com as distâncias euclidiana e euclidiana média podem ser considerados satisfatórios.

Os dados transformados por Sturges conseguiram desempenhos idênticos aos obtidos pelos dados originais nas três estatísticas de eficiência, quando os dados foram arredondados. Isso pode ter ocorrido porque as variáveis transformadas por Sturges possuem a segunda correlação mais alta com os dados originais e quanto mais alta a correlação entre os dados transformados e os originais, mais similares são seus desempenhos.

A distância de Mahalanobis modificada também foi agrupada por UPGMA. O CCC aumentou para 0,73, quando comparada com a distância de Mahalanobis original. O grau de distorção diminuiu para 3% e o estresse passou de regular para bom com a modificação da distância. O resultado

obtido por essa nova distância pode ser considerado satisfatório. Além disso, conseguiu obter um maior CCC que a distância euclidiana.

Tabela 19. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados por classes estimadas por Sturges.

Dissimilaridade	*CCC*	*Distorção*	Estresse
Euclidiana	0,6931	2,1523	14,6425
Euclid. Média	0,6931	2,1523	14,6425
Quad. Euclid.	0,6569	8,0706	28,3567
Mahalanobis	0,6480	11,2944	33,5067
Dis. Ponderada	0,6418	10,4952	32,3237
Média	0,6666	6,8330	24,6944
CV (%)	3,7201	64,9012	37,9529

4.3.7 Transformação considerando a distribuição Normal

Método do vizinho mais próximo

Os melhores resultados obtidos utilizando o método do vizinho mais próximo foram, novamente, os que usaram como medidas de dissimilaridade a distância euclidiana e a euclidiana média (Tabela 20). O coeficiente de correlação cofenética não variou muito, porém nenhuma das medidas teve um coeficiente próximo de 0,7. O nível de estresse foi regular para a distância euclidiana e euclidiana média e insatisfatório para as demais medidas. O grau de distorção variou bastante, mas não obteve valores baixos para nenhum coeficiente de dissimilaridade.

Portanto, nem os desempenhos obtidos com as distâncias euclidiana e euclidiana média, que foram as medidas que tiveram os melhores resultados, foram satisfatórios.

Os resultados obtidos pelos dados transformados foram bastante parecidos com os resultados dos dados originais. A média do CCC e do nível de estresse dos dados transformados foram as mesmas dos dados originais. Apenas no grau de distorção houve uma variação de um por cento. Pelo exposto, os resultados dos dados transformados pela distribuição dormal podem ser considerados idênticos aos obtidos pelos dados originais. Esse

resultado já era esperado, pois as variáveis originais foram geradas com base em uma distribuição normal, e as variáveis transformadas também seguiram o mesmo padrão dos dados originais.

Foi calculada para os dados transformados pela distribuição normal, a raiz quadrada da distância de Mahalanobis. O CCC aumentou para 0,78, este foi o maior CCC obtido quando os dados foram agrupados pelo método do vizinho mais próximo. O grau de distorção diminuiu,mas se manteve alto, em torno de 50%. O nível de estresse passou de insatisfatório para regular. Comparando apenas CCC, o desempenho deste ficou bem superior ao obtido pela distância euclidiana.

Tabela 20. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o do Vizinho mais próximo, para os dados transformados por distribuição normal.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6140	44,7194	29,2830
Euclid. Média	0,6140	44,7194	29,2830
Quad. Euclid.	0,5808	72,3098	52,6457
Mahalanobis	0,5616	78,6459	59,5900
Dis. Ponderada	0,5568	77,0574	57,7627
Média	0,5854	63,4904	45,7129
CV (%)	4,7101	27,2377	33,2790

Ligação média entre grupos

Mais uma vez, as medidas de dissimilaridade que obtiveram os melhores desempenhos foram as distâncias euclidiana e euclidiana média (Tabela 21). O grau de distorção se manteve baixo em todas as medidas de dissimilaridade. O nível de distorção foi considerado bom para os dois primeiros coeficientes e regular para os demais. As distâncias euclidiana e euclidiana média obtiveram um coeficiente de correlação cofenética bem próximo do ideal, mas as outras não.

Portanto, apenas as duas primeiras medidas tiveram resultados satisfatórios nas três estatísticas observadas.

Para o método UPGMA, também foram obtidos resultados idênticos entre os dados originais e as variáveis transformadas pelo método da distribuição Normal. Esse resultado já era esperado, pelo mesmo motivo apresentado anteriormente.

Foi feito também o agrupamento pelo método UPGMA da distância de Mahalanobis modificada. O CCC aumentou para 0,79 o seu desempenho. Foi o mais alto CCC encontrada para esse método de agrupamento. O grau de distorção passou a ser 3% e o nível de estresse mudou de regular para bom. A nova medida de Mahalanobis obteve um desempenho satisfatório para todas estatísticas.

Tabela 21. Desempenho médio das medidas de dissimilaridade tendo utilizado como método de agrupamento o UPGMA, para os dados transformados por distribuição normal.

Dissimilaridade	***CCC***	***Distorção***	**Estresse**
Euclidiana	0,6911	2,1997	14,8075
Euclid. Média	0,6911	2,1997	14,8075
Quad. Euclid.	0,6548	8,2493	28,6733
Mahalanobis	0,6601	11,0237	33,1018
Dis. Ponderada	0,6497	10,2577	31,9355
Média	0,6694	6,7860	24,6651
CV (%)	3,0154	63,4770	37,0724

4.4 Comparação entre estimativas de dissimilaridade obtidas a partir de dados originais e multicategóricos.

De modo geral, o coeficiente de correlação de Mantel foi alto entre as matrizes de dados originais e transformados, independente da medida de distância utilizada. O método de transformação de dados que obteve os menores coeficientes correlação foi o do percentual equitativo com quatro e cinco classes. Esse resultado está de acordo com o obtido pelas estatísticas de eficiência, que mostraram pior desempenho também para este método. Os maiores coeficientes foram obtidos pelos métodos que tem suas classes estimadas pela regra do Quadrado e por Sturges. Esse resultado corrobora com o que foi encontrado também pelas estatísticas de eficiência.

Tabela 22. Coeficientes de correlação de Mantel entre valores de dissimilaridade das matrizes dos dados originais e transformados para todas as medidas de dissimilaridade.

	DEA	DEA_5	PE	PE_5	CER	CES	DN
Euclidiana	0,95	0,96	0,91	0,93	0,99	0,98	0,97
Euclid. Média	0,95	0,96	0,91	0,93	0,99	0,98	0,97
Quad. Euclid.	0,95	0,97	0,92	0,93	0,99	0,98	0,97
Mahalanobis	0,95	0,95	0,88	0,83	0,98	0,97	0,89
Dis. Ponderada	0,96	0,97	0,91	0,89	0,98	0,99	0,96

DEA: divisão equitativa da amplitude com quatro classes; DEA_5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE_5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal. * todos os coeficientes foram significativos a 1% de probabilidade.

4.5 Análise do desempenho geral dos métodos de transformação

Método do vizinho mais próximo

Analisando cada estatística de eficiência (CCC, estresse e distorção) separadamente, observa-se que os dados transformados pelo método da distribuição normal e pela regra do Quadrado foram os que obtiveram coeficientes de correlação cofenética mais parecidos com o CCC dos dados originais (Tabela 21). Por outro lado, quando se observa o grau de distorção e o nível se estresse, percebe-se que todos os métodos tiveram desempenhos semelhantes. Porém, como já foi dito anteriormente, nenhum método de transformação de dados conseguiu, para nenhuma das medidas de dissimilaridade, um resultado satisfatório quando utilizou-se como método de agrupamento o vizinho mais próximo.

É interessante ressaltar que, quando se analisa as diversas medidas de dissimilaridade dentro de cada método de transformação de dados, a estatística que varia menos entre as medidas é o coeficiente de correlação cofenética. Entretanto, esse foi o que obteve maior variação quando se observa apenas o melhor desempenho de cada método de transformação de dados. Com o grau de distorção e nível de estresse acontece justamente o contrário. Quando se observa o melhor desempenho de cada método de transformação de dados ao mesmo tempo, essas duas estatísticas quase não variam, menos

de um por cento. Mas quando se analisa as diversas medidas de dissimilaridade dentro de cada método de transformação de dados, a variação delas fica em torno de 30%.

Tabela 21. Desempenho médio dos métodos de transformação de dados utilizando como medida de dissimilaridade a distância euclidiana e o Vizinho mais próximo como método de agrupamento.

Métodos	***CCC***	***Distorção***	**Estresse**
DO	0,6157	44,8155	29,3557
DEA	0,5926	44,8653	29,3887
DEA_5	0,5972	44,5091	29,1267
PE	0,5214	44,0324	28,7777
PE _5	0,5239	44,1160	28,8208
CER	0,6153	44,9771	29,4864
CES	0,6062	44,9801	29,4267
DN	0,6140	44,7194	29,2830
Média	0,5815	44,5999	29,1871
CV (%)	7,0583	0,8856	0,9913

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.

Ligação Média entre grupos

Os dados transformados pela regra do Quadrado, regra de Sturges e pela distribuição normal foram os que tiveram desempenhos mais parecidos com o dos dados originais (Tabela 22). Vale ressaltar que os dois primeiros obtiveram resultados um pouco melhor que o obtido pelos dados originais. Somente o método percentual equitativo com quatro e cinco classes não obtiveram resultados satisfatórios, pois seus coeficientes de correlação cofenética ficaram muito abaixo de 0,7.

Comparando os resultados obtidos com o método do vizinho mais próximo, verifica-se que somente para o CCC houve uma diminuição no coeficiente de variação, para as demais estatísticas, este foi maior no método do UPGMA. A média do CCC aumentou 0,10, passou de 0,58 para 0,68. O grau de distorção médio passou de 45% para 2% e o nível de estresse mudou

de regular para bom. Portanto, percebe-se que quando se agrupa os dados pelo método do UPGMA, os resultados de todas as estatísticas de teste melhoram significativamente.

Outro fato importante observado é que os melhores resultados pertencem aos métodos de transformação com os maiores números de classes. Os métodos que tem suas classes estimadas pela regra do Quadrado e por Sturges possuem nove e sete classes, respectivamente.

Tabela 22. Desempenho médio dos métodos de transformação de dados utilizando como medida de dissimilaridade a distância euclidiana e o UPGMA como método de agrupamento.

Métodos	***CCC***	***Distorção***	**Estresse**
DO	0,6925	2,2082	14,8329
DEA	0,6819	2,1825	14,7446
DEA 5	0,6851	2,1468	14,6180
PE	0,6483	2,0374	14,2421
PE 5	0,6444	2,0568	14,3152
CER	0,6989	2,1988	14,7893
CES	0,6931	2,1523	14,6425
DN	0,6911	2,1997	14,8075
Média	0,6775	2,1392	14,5942
CV (%)	3,2513	3,1049	1,5597

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.

4.6. Comparação de dendrogramas obtidos utilizando como medida de dissimilaridade a distância euclidiana

Método de agrupamento UPGMA

Um problema comum aos métodos hierárquicos é a dificuldade em se estabelecer um número ótimo de grupos formados. Geralmente, esse procedimento é feito de forma subjetiva, sendo recomendado que se faça uma

análise visual de pontos onde ocorrem mudanças acentuadas de níveis, que possibilitem a demarcação dos grupos (Barbé, 2008; Cruz et al., 2004).

Analisando os 20 genótipos simulados pelo método UPGMA tendo utilizado como medida de dissimilaridade a distância euclidiana, com base nas dez variáveis quantitativas ou multicategóricas, verificou-se a formação de seis dendrogramas (Figura 2), onde as porcentagens das distâncias entre os genótipos estão representadas no eixo X, e no eixo Y estão representados os genótipos. Não foram construídos dendrogramas para o método do percentual equitativo com quatro e cinco classes porque estes não tiveram desempenhos satisfatórios.

Observou-se no dendrograma dos dados originais que ao se fazer um corte vertical em 60%, tem-se a formação de oito grupos. Ao se fazer um corte a 70%, nota-se a ocorrência de apenas quatro grupos distintos. Considerando que a maior discriminação ocorreu no corte a 60%, considerou-se esse percentual como ponto de corte de todos dendrogramas deste trabalho, para facilitar a comparação entre eles.

Apenas no dendrograma formado pelo método de divisão equitativa da amplitude com quatro classes o corte foi feito a 70%. Neste dendrograma, houve a formação de quinze grupos quando o corte foi realizado a 60%, ou seja, não havia se formado quase nenhum grupo, pois estão sendo analisados apenas vinte genótipos. Portanto, o corte foi feito a 70%, definindo-se dez grupos diferentes.

Os genótipos 15 e 16 estão presentes na composição do grupo 1 em todos os dendrogramas formados. Além disso, também é comum a todos dendrogramas o fato dos genótipos 2 e 4 serem os menos similares aos outros genótipos. Será comparado separadamente cada dendrograma formado por algum método de transformação de dados com o dendrograma construído com os dados originais,

O número de grupos formados pelo método de divisão equitativa da amplitude com quatro classes foi dez, enquanto os dados originais criaram oito grupos. O grupo 1 formado pelo dendrograma do método de divisão equitativa da amplitude com quatro classes engloba praticamente todos os genótipos que

formaram o grupo 1 e 2 no dendrograma dos dados originais, apenas o genótipo 6 não entrou nesse grupo. O grupo 5 foi formado por dois elementos pertencentes também ao grupo 5 dos dados originais. E neste dendrograma, o genótipo menos similar foi o 4 e o segundo menos similar foi o 2, no dendrograma original ocorre justamente o inverso.

No agrupamento feito com o método de divisão equitativa da amplitude com cinco classes houve a formação de nove grupos. Todos elementos do grupo 1 desse método estão dentro do grupo 1 formado pelos dados originais. O grupo 2 dos dois agrupamentos são compostos pelos mesmos genótipos. O grupo 3 possui apenas o genótipo 5 em comum com o original. E os genótipos menos similares nos dois agrupamentos são os mesmos.

O agrupamento formado pelo método de classes estimadas pela regra do Quadrado produziu oito grupos distintos. Os grupos 1,2,7 e 8 são idênticos aos respectivos grupos formados pelos dados originais. Nos demais grupos não existem elementos em comum com os grupos originais.

O método de classes estimadas por Sturges também formou oito grupos. Sendo os grupos 2,4,5,6,7,8 iguais aos formados pelo agrupamento feito com os dados originais. O grupo 1 formado por Sturges possui três elementos em comum ao dos dados originais e o grupo 3 apenas um genótipo em comum. A formação dos oito grupos gerados por Sturges só não é idêntica a original devido a dois elementos do grupo 1 dos dados originais que estão no grupo 3 da formação feita por Sturges.

O agrupamento formado pelo método de distribuição normal criou oito grupos. Dentre estes, apenas três são idênticos aos grupos formados pelos dados originais, que são os grupos 1,7 e 8. O grupo 2 tem um elemento a mais que o grupo 2 do agrupamento original. E os grupos restantes não possuem elementos em comum aos respectivos grupos formados pelos dados originais.

Logo, dos métodos de transformação de dados estudados, o método de classes estimadas por Sturges foi o que obteve um dendrograma mais parecido com o formado pelos dados quantitativos.

A formação dos grupos pode ser conferida na Tabela 23.

Tabela23. Formação dos grupos gerados pelos dendrogramas dos dados originais e multicategóricos.

G	*DO*	*DEA*	*DEA_5*	*CER*	*CES*	*DN*
1	8;9;15;16;20	8;9;13;14;15;16; 17;18;19;20	9;15;16	8;9;15;16;20	15;16;9	8;9;15;16;20
2	6;13;14;17;18;19	6	6;13;14;17; 18;19	6;13;14;17; 18;19	6;13; 14;17; 18;19	6;11;13;14;17; 18; 19
3	5	11	5;8;20	1;10;12	5;8;20	1;10;12
4	7	3	7	3;11	7	3
5	1;10;12	10;12	3;11	5	1;10;12	7
6	3;11	1	10;12	7	3;11	5
7	4	5	1	4	4	4
8	2	7	4	2	2	2
9	*	2	2	*	*	*
10	*	4	*	*	*	*

G: grupo; DO:dados originais;DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.

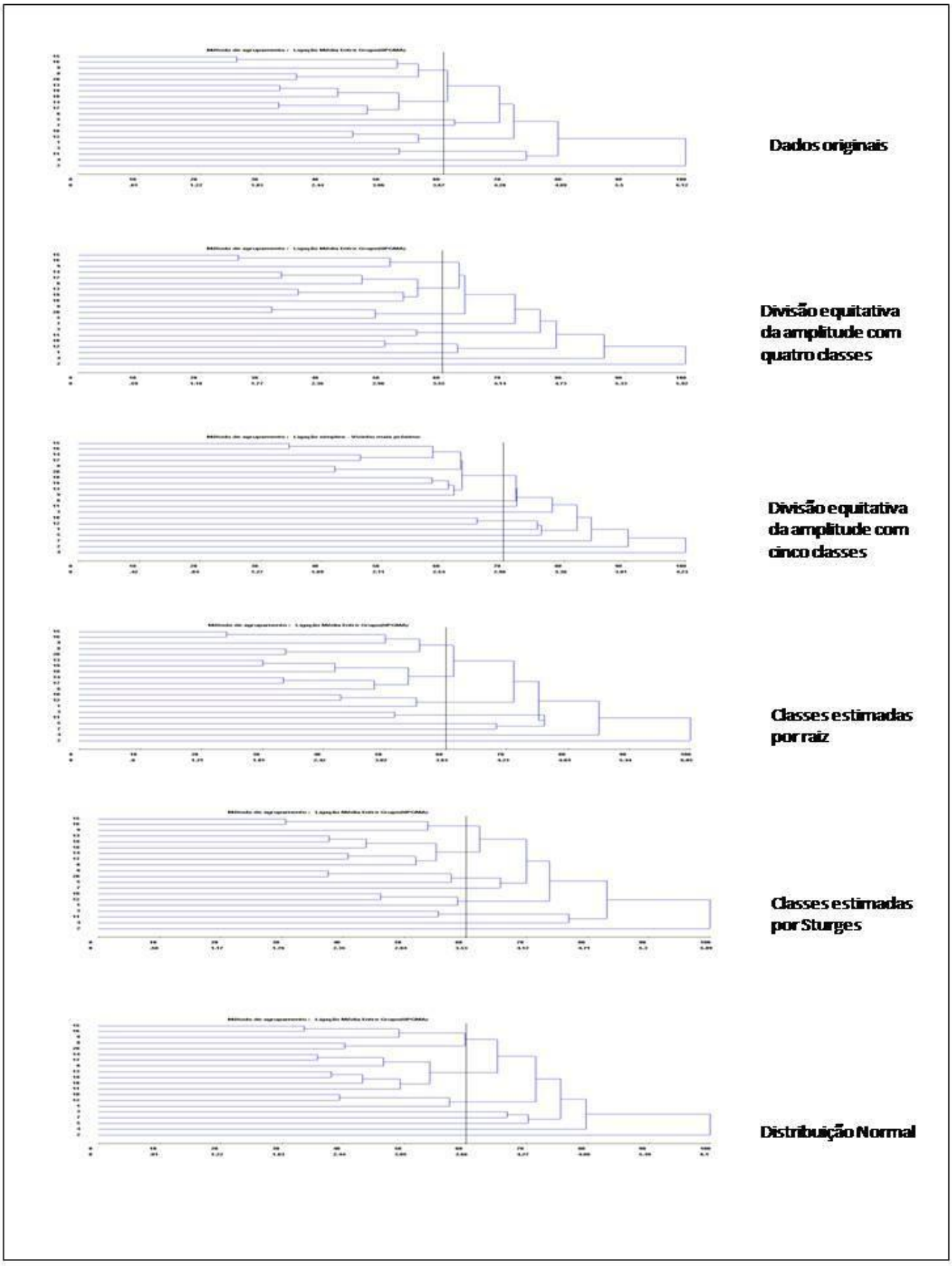


Figura 2 Dendrogramas do método UPGMA utilizando como medida de dissimilaridade a distância euclidiana.

Método do Vizinho Mais Próximo

Apesar de não terem obtido rendimentos satisfatórios, foram construídos dendrogramas para o método do vizinho mais próximo utilizando como medida de dissimilaridade a distância euclidiana (Figura 3). Isto foi feito com o objetivo de observar se há diferenças entre os grupos formados pelo VMP e UPGMA. Apenas para o método percentual equitativo com quatro e cinco classes não foram feitos os dendrogramas, porque estes não foram construídos nem para o método UPGMA devido ao seu desempenho ruim.

O corte foi realizado a 70% ao invés de 60%, como foi feito nos dendrogramas construídos pelo UPGMA, porque a 60% a maioria dos genótipos ainda não havia sido agrupada.

Houve a formação de nove grupos pelos dados originais pelo VMP (Tabela 24), enquanto pelo UPGMA foram formados oito grupos (Tabela 23). O grupo 1 do primeiro englobou o grupo 1 e 2 formados pelo UPGMA. O grupo 5 do VMP teve um componente a menos que o do UPGMA, apenas o genótipo 1 ficou fora desse grupo. O único grupo que foi idêntico ao do UPGMA foi o 7, cujo componente era apenas o genótipo 4. Portanto, percebe-se que o tipo de agrupamento utilizado altera bastante a constituição dos grupos formados.

No método da divisão equitativa da amplitude com quatro classes foram formados 10 grupos. Os grupos 5,6,8 e 9 foram idênticos aos formados pelos dados originais. O grupo 1 dos dados transformados só não foi idêntico ao grupo 1 dos dados originais porque o genótipo 6 não faz parte da formação do primeiro. Quando compara-se os grupos formados pelos dados transformados com aqueles formados pelos dados originais agrupados por UPGMA, percebe-se que não há nenhum grupo idêntico entre eles. O grupo 1 do primeiro engloba todos os elementos do grupo 1 e 2 dos dados originais, exceto o genótipo 6. Este pertence ao grupo 2 também dos dados transformados. O grupo 5 é quase igual ao dos dados originais, exceto por não incluir o genótipo 1 na sua formação.

Nove grupos foram formados pela divisão equitativa da amplitude com cinco classes. Apenas os grupos 3,6 e 9 foram idênticos aos formados pelos dados originais. O grupo 1 difere do grupo correspondente nos dados originais devido a inclusão de dois genótipos neste e exclusão de outro que está no grupo 1 dos dados originais. O grupo 5 só tem o genótipo 12 em comum com a

formação original. Quando estes grupos são comparados com os formados pelos dados originais agrupados pelo UPGMA, não se encontra nenhum grupo idêntico entre eles. O grupo 1 dos dados transformados possui cinco genótipos em comum com a formação dos dados originais e os grupos 2 e 5 possuem apenas um.

Para os dados transformados pela regra do Quadrado, ocorreu a formação de nove grupos. Dentres eles, apenas o primeiro e o último foram idênticos aos formados pelos dados originais. Quando estes dados transformados foram comparados com os originais agrupados por UPGMA, notou-se, mais uma vez, que não há nenhum grupo igual entre eles. Outro fato perceptível é que o grupo 1 dos dados transformados é a junção do grupo 1 e 2 dos dados originais. Este evento também ocorreu quando foram comparados os grupos formados pelos dados originais agrupados por VMP e UPGMA.

O agrupamento formado pelo método de Sturges criou onze grupos. O grupo 1 dos dados originais é a junção do grupo 1,2 e 3 de Sturges. Apenas o último grupo foi idêntico ao formado pelos dados originais. Estes dados transformados também foram comparados ao agrupamento dos dados originas pelo UPGMA. Novamente não há ocorrência de nenhum grupo idêntico entre eles. O grupo 1 possui apenas dois genótipos em comum com o grupo 1 formado pelos dados originais e o grupo 2, três genótipos.

Dez grupos foram formados pelo agrupamento dos dados transformados pelo método de distribuição normal. Os grupos 5,6,7 e 9 dos dados transformados são idênticos aos formados pelos dados originais. Os grupos 1 e 2 dos dados transformados correspondem ao grupo 1 dos dados originais. Quando se faz um paralelo destes dados com os dados originais agrupados por UPGMA, observa-se que apenas o grupo sete teve a mesma formação para os dois. O grupo 1 dos dados transformados possui quatro genótipos em comum com o correspondente nos dados originais, o grupo 2 possui três e o grupo 5, apenas dois.

A formação dos grupos ocorrida nos dados transformados pela divisão equitativa com quatro classes foi a mais semelhante à obtida pelos dados originais agrupados por VMP. Por outro lado, a distribuição normal foi a mais similar com a formação obtida pelos dados originais agrupados por UPGMA. Este é tomado como parâmetro, já que obteve melhores resultados nas

estatísticas de eficiência do que os dados originais agrupados por VMP. Portanto, fica evidente que se o desempenho nessas estatísticas é ruim, a formação dos grupos fica bem diferente da considerada ideal.

Tabela24. Formação dos grupos gerados pelos dendrogramas dos dados originais e multicategóricos.

G	DO	*DEA*	DEA_5	CER	CES	DN
1	6;8;9;13;14; 15;16;17;18; 19; 20	8;9;13;14;15; 16;17;18;19; 20	5;6;8;9;11; 13;14;15;16; 17;19; 20	6;8;9;13;14; 15;16;17; 18;19;20	6;14; 15;16; 17	6;8;14;15 16;17;20
2	11	6	18	3	9;13; 18;19	9;13;18; 19
3	3	11	3	11	8;20	11
4	5	3	10	10;12	10;12	3
5	10;12	10;12	12	1	11	10;12
6	1	1	1	5	5	1
7	4	5	7	7	3	4
8	7	7	4	4	1	5
9	2	2	2	2	7	2
10	*	4	*	*	4	7
11	*	*	*	*	2	*

G: grupo; DO:dados originais;DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.

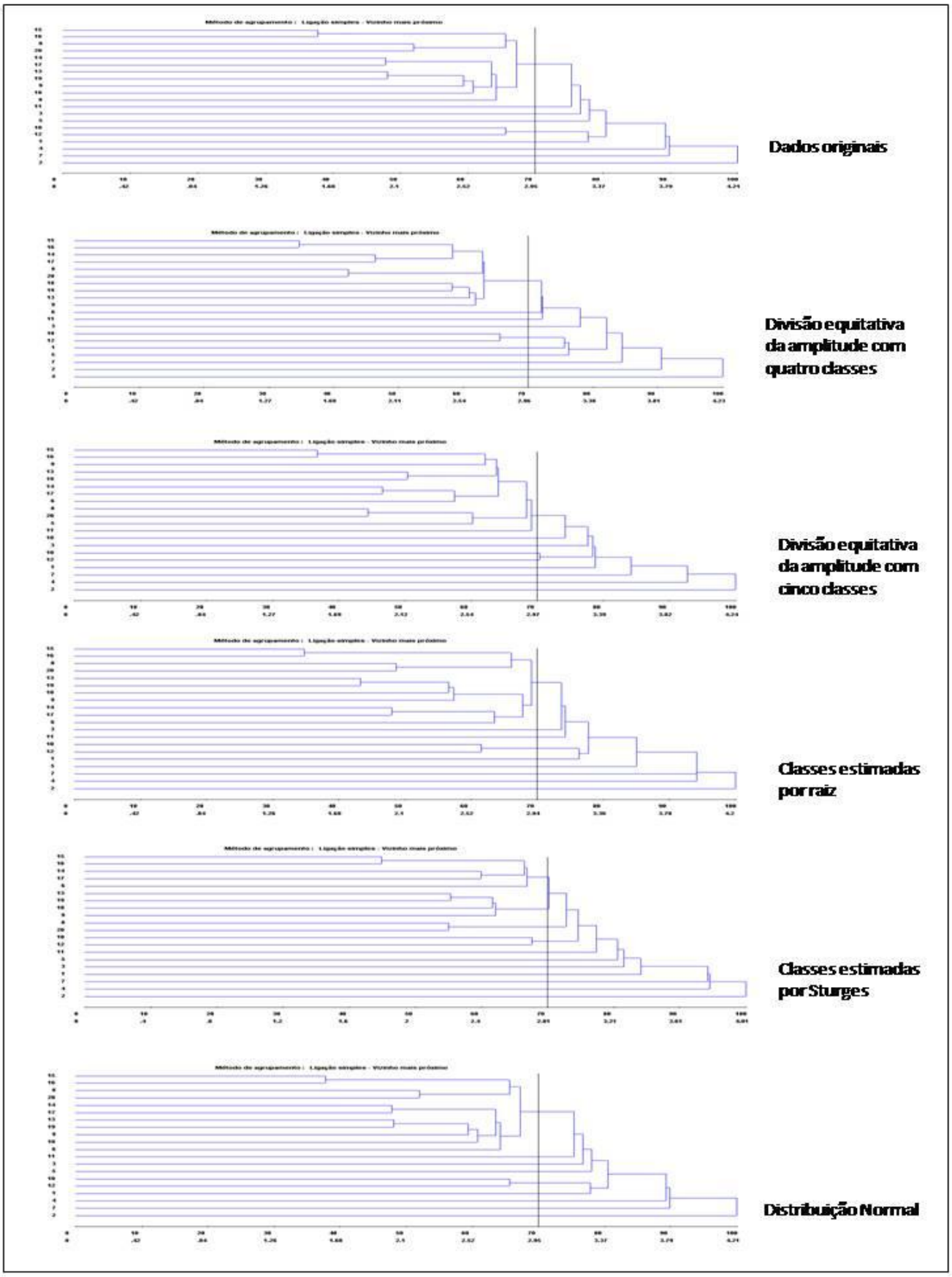


Figura 3. Dendrogramas do método VMP utilizando como medida de dissimilaridade a distância euclidiana.

4.7 Análise do comportamento das estatísticas de eficiência dos métodos de agrupamento

O método de transformação de dados que obteve a menor média do coeficiente de correlação cofenética foi o percentual equitativo com quatro classes quando agrupado pelo método do vizinho mais próximo. De todos que foram agrupados por esse, os dados originais foram os que obtiveram a maior média. Percebe-se, pelo gráfico (figura 4), que quando se muda de método de agrupamento, o valor da média aumenta. Porém, a média não aumenta muito quando são utilizados os métodos de percentual equitativo com quatro e cinco classes. As médias dos outros métodos de transformação foram bem parecidas, além de serem relativamente maiores que as citadas anteriormente.

O menor coeficiente de variação, 0,03%, do CCC foi do método de percentual equitativo com cinco classes quando agrupado por UPGMA. O segundo menor coeficiente, que é dos dados transformados pela divisão equitativa da amplitude com quatro classes e agrupados por UPGMA, alcança quase três por cento de variação. A maioria dos métodos de transformação teve seus coeficientes de variação entre três e cinco por cento. Os únicos que obtiveram coeficientes acima desta porcentagem foram os dados transformados pelo percentual equitativo com quatro classes, agrupados por VMP, e os que tiveram suas classes estimadas por Sturges, também agrupados por VMP.

As médias da estatística de eficiência distorção entre a matriz fenética e a cofenética obtiveram um valor próximo de seis por cento para todos os métodos de transformação agrupados por UPGMA. Por outro lado, os métodos de transformação agrupados por VMP tiveram suas médias variando entre 62 e 64%.

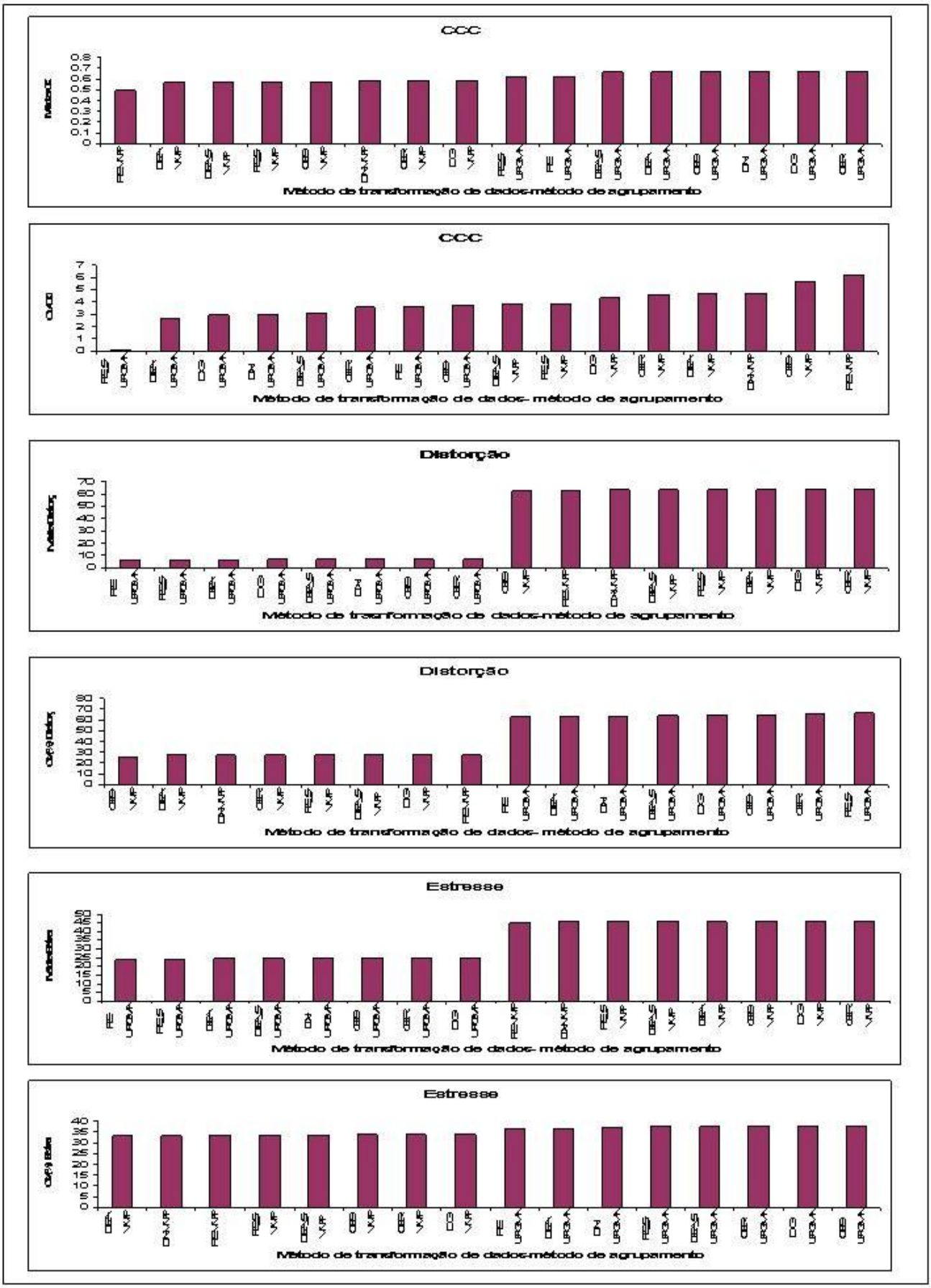
Os coeficientes de variação da estatística distorção encontrados para os métodos de transformação agrupados por VMP foram menores que os coeficientes calculados quando os métodos foram agrupados por UPGMA. Os primeiros ficaram em torno de trinta por cento, e os outros na faixa de 62 a 67%.

O padrão mantido pelas médias de estresse entre a matriz fenética e a cofenética é parecido com o que foi encontrado para as médias de distorção.

Porém os valores não são os mesmos. As médias dos métodos de transformação que foram agrupados por UPGMA ficaram em torno de 25%, e para os que foram agrupados por VMP as médias ficaram próximas de 45%.

O coeficiente de variação da estatística estresse não variou muito entre os métodos de transformação. Os métodos que foram agrupados por VMP tiveram seu valores de coeficiente de variação em torno de 33% e os agrupados por UPGMA, entre 36 e 37%.

Pode-se dizer que o CCC apresentou resultados coerentes com as demais estatísticas, porém ele não fornece coeficientes com valores discrepantes entre as medidas de dissimilaridade, diferentemente do que ocorre com o nível de estresse e o grau de distorção. Portanto, esses dois últimos avaliadores podem ser considerados mais sensíveis do que o coeficiente de correlação cofenética para detectar diferenças nos desempenhos das medidas de dissimilaridade.



DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas por raiz; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal; DO: dados originais.

Figura 4. Histogramas das médias e coeficientes de variação do coeficiente de correlação cofenética, estresse e distorção entre as matrizes fenéticas e cofenéticas.

5-CONCLUSÕES

Em função dos resultados obtidos, conclui-se que:

- Em relação às medidas de dissimilaridades

- As distâncias euclidiana e euclidiana média tiveram a mesma performance em todas as análises de agrupamento feitas;
- As medidas de dissimilaridade que obtiveram os melhores desempenhos em todos os agrupamentos foram as distâncias euclidiana e euclidiana média;
- Tirar a raiz quadrada da distância de Mahalanobis parece uma boa solução para melhorar o desempenho desta medida, quando se tem dados que seguem distribuição normal;

- Em relação aos métodos de agrupamentos

- Para as variáveis quantitativas e multicategóricas simuladas utilizando parâmetros genéticos da cultura do milho, o método do vizinho mais próximo não produziu nenhum agrupamento que tivesse um desempenho satisfatório, de acordo com as estatísticas de eficiência utilizadas neste trabalho;
- O método de agrupamento UPGMA foi superior ao método do vizinho mais próximo para todas as medidas de distância utilizadas;

- Em relação aos métodos de transformação de dados quantitativos em multicategóricos

- O único método de transformação que não conseguiu obter um desempenho satisfatório quando agrupado por UPGMA foi o percentual equitativo com quatro e cinco classes;
- O dendrograma construído com os dados transformados pela regra de Sturges foi o mais parecido com o dendrograma formado pelos dados originais, quando agrupados por UPGMA;

- Em relação às estatísticas utilizada como medida de eficiência do padrão de agrupamento

- As estatísticas de eficiência que conseguiram detectar diferenças maiores entre as medidas de dissimilaridade, facilitando a distinção dos desempenhos dessas medidas, foram distorção e estresse entre as matrizes fenéticas e cofenéticas.

6-BIBLIOGRAFIA

ABREU, F.B.; LEAL, N.R.; RODRIGUES, R.; AMARAL JÚNIOR, A.T.; SILVA, D.J.H. Divergência genética entre acessos de feijão-de-vagem de hábito de crescimento indeterminado. **Horticultura Brasileira**, Brasília, v.22, n.3, p.547-552, jul-set 2004.

ALMEIDA, R. V. **Parâmetros genéticos e alterações nas frequências alélicas em três ciclos de seleção divergente para tolerância ao alumínio em milho**. Viçosa,2007. 65p. Dissertação (Mestrado em Genética e Melhoramento)- Universidade Federal de Viçosa, 2007.

BARBÉ, T.C. **Estimativa de divergência genética entre linhas de feijão-de-vagem (Phaseolus vulgaris L.) por meio de análise multivariada e associação com a genealogia**. Campos dos Goytacazes, 2008. 95p. Dissertação(Mestrado em Produção Vegetal)- Universidade Estadual do Norte Fluminense , 2008.

BARELLI,M.A.A.; SCHAWINSKI, E.C.; NEVES, L.G. Avaliação de acessos de *Manihot esculenta,* Crantz. Com uso de variáveis multicategóricas no município de Cáceres-MT. **Raízes e Amidos Tropicais.** Botucatu: CERAT/UNESP. V.3,2007.

BENTO, C.S.; SUDRÉ, C.P.; RODRIGUES, R.; RIVA, E.M.; PEREIRA, M.G. Descritores qualitativos e multicategóricos na estimativa da variabilidade fenotípica entre acessos de pimentas. **Scientia Agraria**, v.8, n.2, p.149-156. 2007.

BERTAN,I.;CARVALHO,F.I.F.C.;OLIVEIRA,A.C.;VIEIRA,E.A. Comparação de métodos de agrupamento na representação da distância morfológica entre genótipos de trigo. **R. Bras. Agrociência**, Pelotas, v. 12, n. 3, p. 279-286, jul-set, 2006

BOX, G. E.P.; MULLER, M.E. A Note on the Generation of Random Normal Deviates. **The Annals of Mathematical Statistics** , Princenton, v.29, p. 610-611, jan.1958.

BUSSAB, W. DE O.; MIAZAKI, E.S.; ANDRADE D.F. **Introdução à análise de agrupamentos.** São Paulo: Associação Brasileira de Estatística, 1990. 105p.

BUSSAB, W. DE O., MORETTIN, P. **Estatistica Básica***.* 5ed. São Paulo: Saraiva,2002. 526p.

CARDOSO, W. **Variabilidade de genótipos de milho quanto à composiçao de carotenóides nos grãos visando a biofortificação**. Viçosa,2007. 67p. Dissertação (Mestrado em Fitotecnia)- Universidade Federal de Minas Gerais, 2007.

CARGNELUTTI FILHO, A.; RIBEIRO, N.D.; REIS, R.C.P; SOUZA, J.R.; JOST,E. Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. **Ciência Rural***,* Santa Maria, v.38, n.8 , p. 2138-2145, nov.2008.

CHEN, X.; REYNOLDS, C.H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching:Comparison of Structural Descriptors and Similarity Coefficients. **J. Chem. Inf. Comput. Sci.** 42: 1407-1414, 2002.

COIMBRA, R.R.; MIRANDA, G.V.; MOREIRA, G.R.; SILVA, D.J.H.; CRUZ, C.D.; CARNEIRO, P.C.S.; SOUZA, L.V.; GUIMARÃES, L.J.M.; MARCASSO, R.C.; CANIATO, F.F. Divergência genética de cultivares de milho baseada em descritores qualitativos. In: **Simpósio de recursos genéticos para a América Latina e Caribe**, III SIRGEALC.,2001,Londrina. *Anais...* Londrina, PR, 2001. p. 266-268.

CRUZ, C. D. **Programa Genes***:* Aplicativo computacional em genética e estatística*.* Viçosa: UFV,2001.

CRUZ, C. **Programa Genes- Análise Multivariada e Simulação**. Viçosa: UFV,2006. 175p.

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

CRUZ, C.D.; FERREIRA, F.M.; PESSONI, L.A. **Biometria Aplicada ao estudo da diversidade genética.** Viçosa, 2008. 539p.

CRUZ, C.D.; REGAZZI, A.J; CARNEIRO, P.C.S. **Modelos Biométricos Aplicados ao Melhoramento Genético**, vol.1. Viçosa: UFV,2004. 480p.

DALIRSEFAT, S.; MEYER, A.; MIRHOSEINI, S. Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, *Bombyx mori*. **Journal of Insect Science** 9:71, 2009.

DUARTE, J.M.; SANTOS, J.B. ; MELO, L.C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics and Molecular Biology**, 22,3, 427-432 (1999).

FARIA, P. **Avaliação de métodos para determinação do número ótimo de clusters em estudo de divergência genética entre acessos de pimenta.** Viçosa,2009.67p. Dissertação (Mestrado em Estatística Aplicada e Biometria)- Universidade Federal de Viçosa, 2009.

FERREIRA, F.M. **Diversidade em populações simuladas com base em locos multialélicos.** Viçosa,2007. 177p. Tese (Doutorado em Genética e Melhoramento)- Universidade Federal de Viçosa, 2007.

FREITAS, F. **Estudo genético-evolutivo de amostras modernas e arqueológicas de milho (*Zea mays mays* L.) e feijão (*Phaseolus vulgaris,* L.)**.Piracicaba,2001. 144p.Tese (Doutorado em genética e melhoramento de plantas)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo,2001.

GARBUGLIO, D. D. ; MIRANDA FILHO, J. B. ; CELLA, M. . Variabilidade genética em famílias S1 de diferentes populações de milho. **Acta Scientiarum** (UEM), v. 31, p. 209-213, 2009.

GONÇALVES, L.S.A.; RODRIGUES, R.; AMARAL JÚNIOR, A.T.; KARASAWA, M.; SUDRÉ, C.P. Comparison of multivariate statitical algorititthms to cluster

tomato heirloom accessions. **Genetics and Molecular Research** 7(4): 1289-1297 (2008).

HAIR JR., J.F.; ANDERSON, R.E.; RONALD, L.T. **Multivariate Data Analysis:** with Readings. 4ed. New York : Macmillan Publishing Company, 1995.

JONHS, M.A.; SKROCH, P.W.; NIENHUIS, J.; KINRICHSEN, P.; BASCUR,G.;MUÑOZ-SCHICK,C. Gene pool classification of common bean landraces from Chile based on RAPD and morphological data. **Crop Sci.** 37:605-613. 1997.

JOHNSON,R.A.; WICHERN,D.W. **Applied multivariate statistical analysis**. 3 ed. New Jersey: Prentice Hall, 1992. 642p.

KAMADA, T. **Avaliação da diversidade genética de populações de Fáfia (Pfaffia glomerata (Spreng.) Pedersen) por RAPD, caracteres morfológicos e teor de beta-ecdisona.** Viçosa,2005. 106p. Tese(Doutorado em Genética e Melhoramento)- Universidade Federal de Viçosa,2005.

KOSMAN, E.; LEONARD, K.J. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid and polyploidy species. **Molecular Ecology** 14: 415-424, 2005.

KRUSKAL, J.B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. **Psychometrika** 29:1-27, 1964.

KRZANOWSKI, W.J.; MARRIOTT, F.H.C. **Multivariate Analysis**, Part 2: Classification, covariance strutures and repeated measurements. London: Edward Arnold, 1995. 280p.

LAW, A.; KELTON, W.D. **Simulation Modeling and Analysis,** McGrawHill, 1991.

MACHADO, C.F; NUNES, G.H.S.; FERREIRA, D.F.; SANTOS, J.B. Divergência genética entre genótipos de feijoeiro a partir de técnicas multivariadas. **Ciência Rural**, v.32, p.251-258, 2002.

MANLY, B.F.J. **Randomization, bootstrap and Monte Carlo methods in biology.** London: Chapman & Hall, 1997.281p.

MEYER, A. S. ; Garcia, A.A.F. ; SOUZA, A. P. ; SOUZA JR., C. L. . Comparison of similarity coefficients used for cluster analysis with dominant

markers in maize (Zea mays L). **Genetics and Molecular Biology**, v. 27, p. 83-91, 2004.

PAYNE, J.A. **In**

troduction to simulation: programming techniques and methods of analysis, McGrawHill, 1982.

PEREIRA, F.H.F.; PUIATTI, M.; MIRANDA, G.V.; SILVA, D.J.H.; FINGER, F.L. Divergência genética entre acessos de taro utilizando caracteres morfoqualitativos de inflorescência. **Horticultura Brasileira**, Brasília, v. 21, n. 3, p. 520-524, julho-setembro 2003.

ROCHA, M.C.; GONÇALVES, L. S. A.; CORRÊA, F.M.; RODRIGUES,R.; SILVA, S.L.; ABBOUD, A.C.S.; CARMO, M.G.F. Descritores quantitativos na determinação da divergência genética entre acessos de tomateiro do grupo cereja. **Ciência Rural**, Santa Maria, v.39, n.3, p.664-670, mai-jun, 2009.

SANTOS, V.S. **Seleção de pré-cultivares de soja baseada em índices.** 2005. 104p. Tese (Doutorado)- Universidade de São Paulo, Piracicaba.

SHANNON,R.E. **System Simulation: The Art and Science**, Prentice –Hall, Englewood Cliffs, N.J., 1975.

SNEATH, P.H.; SOKAL, R.R. **Numerical taxonomy**: The principles and practice of numerical classification. San Francisco: W.H. Freeman, 1973. 573p.

SANTANA, C.M.;MALINOVSKI, J.R. Uso da análise multivariada no estudo de fatores humanos em operadores de motosserra. **Cerne**, v.8, n.2, p-101-107, 2002.

SIEGMUND, K.D; LAIRD,P.W.; LAIRDOFFRINGA, I.A. A comparision of clusters analysis methods using DNA methylation data. **Bioinformatics**, v.20, n.12, p.1896-1904,2004.

SILVA, A. **Análise genética de caracteres quantitativos em milho com o delineamento III e marcadores moleculares.** Piracicaba, 2002. 154p. Tese (Doutorado em Genética e Melhoramento de plantas)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, 2002.

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

TRIOLA, M.F. **Introdução à Estatística.** LTC. 10a edição **2008**. 722p

APÊNDICES

APÊNDICE 1

Tabelas dos resultados das análises de variância da segunda variável ate a décima.

Tabela A1. Resultado da análise de variância da variável altura de espiga com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de		DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
Variação	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	25,95		0,3125		0,545833		0,533333		0,566667		1,95		1,15		5	
Tratamentos	19	131,27	2,75**	1,34	2,6**	2,04	2,27**	2,53	2,86**	3,84	2,57**	6,99	2,83**	4,33	2,52**	2,16	2,96**
Residuo	57	47,75		0,514254		0,896711		0,884211		1,5		2,47		1,72		0,73	
Média			105		2,56		3,14		2,5		3		5,26		4,14		3,5
CV (%)			6,58		27,98		30,18		37,61		40,78		29,88		31,66		24,38
Herdabilidade			63,62		61,59		56,02		65		61,05		64,61		60,35		66,26

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A2. Resultado da análise de variância da variável peso de espiga despalhada com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de Variação	GL	DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
		QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	115224,8		0,079		0,083		0,066		0,433		0,78		0,017		0,2167	
Tratamentos	19	6537568,5	20,48**	3,22	15,29**	4,79	14,97**	4,29	13,36**	7,18	18,45**	17,6	19,71**	9,56	15,15**	3,87	15,91**
Residuo	57	319292,4		0,210		0,320		0,321		0,389		0,893		0,63		0,24	
Média			6987		2,74		3,33		2,5		3		5,56		4,43		3,48
CV (%)			8,09		16,77		17,02		22,66		20,8		16,99		17,95		14,19
Herdabilidade			95,12		93,46		93,32		92,52		94,58		94,93		93,4		93,71

DO:dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A3. Resultado da análise de variância da variável comprimento de espiga com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de Variação		DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	0,1855		0,2458		0,1458		0,0667		0,0667		0,5667		0,3		0,35	
Tratamentos	19	3,166	7,79**	1,218	5,78**	1,8	5,18**	3,29	5,03**	5,12	4,84**	6,89	8,68**	3,64	7,39**	2,89	7,51**
Residuo	57	0,406		0,210746		0,347588		0,654386		1,06		0,794737		0,49		0,39	
Média			15		2,34		2,76		2,5		2,95		4,5		3,65		3,48
CV (%)			4,25		19,64		21,34		32,36		34,87		19,81		19,24		17,86
Herdabilidade			87,16		82,7		80,71		80,11		79,34		85,21		86,46		86,68

DO:dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A4. Resultado da análise de variância da variável diâmetro de espiga com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de Variação	GL	DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
		QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	0,023		0,1792		0,55		0,3125		1,05		1,3		1,02		0,48	
Tratamentos	19	0,1223	7,03**	1,407	4,4**	2,4	5,19**	3,17	4,65**	4,87	4,86**	8,67	6,76**	5,71	6,45**	2,73	6,35**
Residuo	57	0,017		0,3195		0,4623		0,6809		1		1,28		0,89		0,43	
Média			4,4		2,74		3,33		2,46		2,94		5,5		4,53		3,53
CV (%)			3		20,65		20,45		33,51		34,08		20,4		20,79		18,62
Herdabilidade			85,78		77,29		80,71		78,49		79,41		85,21		84,49		84,25

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A5. Resultado da análise de variância da variável produção de grãos com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de		DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
Variação	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	724,11		1,479		1,55		2,63		3,9		6,1		4,1		2,85	
Tratamentos	19	705,96	2,5**	1,17	2,13*	1,85	2,64**	1,66	1,56	3	1,87*	6,2	2,43**	3,25	2,12*	1,87	2,13*
Residuo	57	282,15		0,549342		0,699123		1,06		1,6		2,55		1,53		0,88	
Média			116,78		2,49		2,93		2,5		3		4,9		3,9		3,48
CV (%)			14,38		29,8		28,59		41,24		42,19		32,57		31,71		26,94
Herdabilidade			60,03		53,06		62,1		35,87		46,61		58,88		52,89		53,03

DO:dados originais;DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
*Significativo a 5% de probabilidade, pelo teste F.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A6. Resultado da análise de variância da variável prolificidade com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de		DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
Variação	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	0,0332		0,9125		0,85		0,9833		2,41		5,18		3,55		1,35	
Tratamentos	19	0,0723	5,18**	1,723	4,36**	2,45	4,81**	2,81	3,68**	4,54	3,95**	8,38	5,18**	4,76	4,46**	2,78	4,76**
Residuo	57	0,1396		0,3950		0,5079		0,76404		1,15		1,62		1,07		0,58	
Média			1,09		2,49		2,98		2,48		2,99		5,04		4,03		3,51
CV (%)			10,84		25,28		23,96		35,32		35,89		25,25		25,67		21,73
Herdabilidade			80,68		77,08		79,23		72,84		74,68		80,68		77,58		79

DO:dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A7. Resultado da análise de variância da variável posição relativa da espiga com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de		DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
Variação	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	0,0010		0,7792		0,8833		1,08		1,25		4,08		2,62		1,08	
Tratamentos	19	0,0021	4,07**	1,533	2,92**	2,18	3,9**	2,56	3,07**	3,87	3,54**	9,03	3,7**	5	3,69**	2,43	3,72**
Residuo	57	0,0005		0,5248		0,5588		0,8336		1,09		2,44		1,35		0,65	
Média			0,54		2,59		3,03		2,41		2,83		5,04		4,28		3,38
CV (%)			3,69		28		24,71		37,84		36,99		31		27,21		23,95
Herdabilidade			75,41		67,78		74,39		67,44		71,79		73		72,92		73,15

DO:dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A8. Resultado da análise de variância da variável florescimento feminino com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de Variação	GL	DO		DEA		DEA_5		PE		PE_5		CER		CES		DN	
		QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	4,284		0,6		0,645833		1,6		2,87		2,73		1,63		1,21	
Tratamentos	19	147,187	7,66**	1,64	7,09**	2,12	5,07**	3,34	6,01**	5,26	5,84**	7,67	6,86**	4,37	6,06**	2,88	6,4**
Residuo	57	1,92		0,231579		0,417763		0,55614		0,901754		1,12		0,72		0,45	
Média			66		2,15		2,51		2,5		3		4,05		3,25		3,46
CV (%)			2,1		22,38		25,73		29,83		31,65		26,12		26,13		19,36
Herdabilidade			86,94		85,9		80,27		83,36		82,87		85,41		83,49		84,37

DO: dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

Tabela A9. Resultado da análise de variância da variável florescimento masculino com os dados originais e com os dados transformados para padrão discreto multicategórico.

Fonte de		**DO**		**DEA**		**DEA_5**		**PE**		**PE_5**		**CER**		**CES**		**DN**	
Variação	GL	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F	QM	F
Blocos	3	6,15		1,283		1,3		1,35		2,33		5,08		3,9		1,75	
Tratamentos	19	9,77	5,43**	1,576	5,56**	2,43	4,46**	3,05	5,12**	5,29	5,74**	8,22	5,37**	5,33	5,81**	2,16	3,64**
Residuo	57	1,8		0,2833		0,5456		0,5956		0,9211		1,53		0,92		0,59	
Média		65,28		2,73		3,4		2,48		3		5,63		4,6		3,48	
CV (%)		2,06		19,53		21,73		31,18		31,99		21,99		20,82		22,14	
Herdabilidade		81,57		82,03		77,56		80,47		82,59		81,39		82,77		72,53	

DO:dados originais; DEA: divisão equitativa da amplitude com quatro classes; DEA 5: divisão equitativa da amplitude com cinco classes; PE: percentual equitativo com quatro classes; PE 5: percentual equitativo com cinco classes; CER: classes estimadas pela regra do Quadrado; CES: classes estimadas por Sturges; DN: classes estimadas por distribuição normal.
** Significativo a 1% de probabilidade, pelo teste F.

APÊNDICE 2

Distribuição dos dados dentro das classes em cada método de transformação de dados, da variável 2 até a 10.

Variável 2 (Altura de espiga)

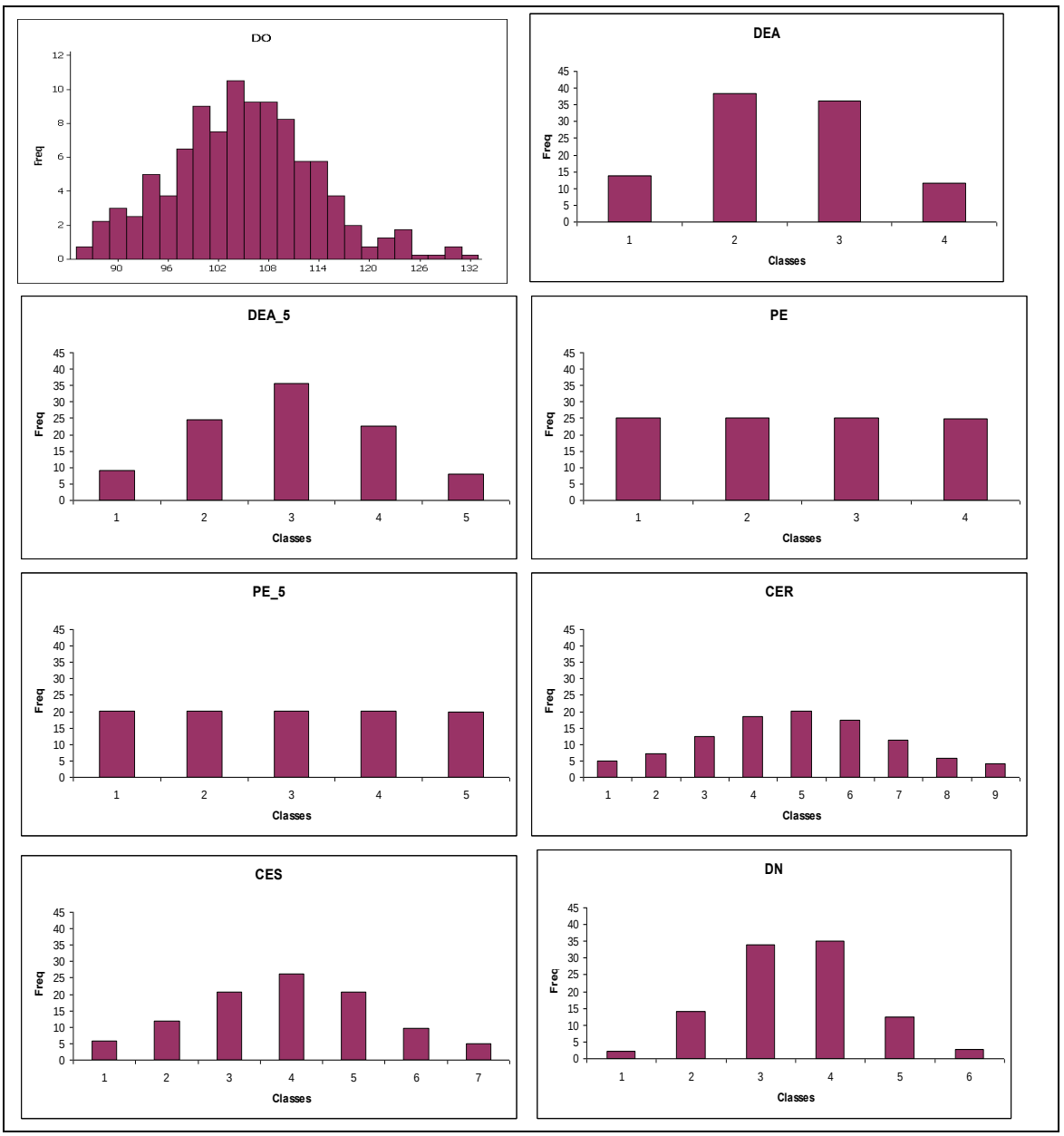


Figura A1. Dispersão dos valores da variável 2 (altura de espiga) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 3 (Peso de espiga despalhada)

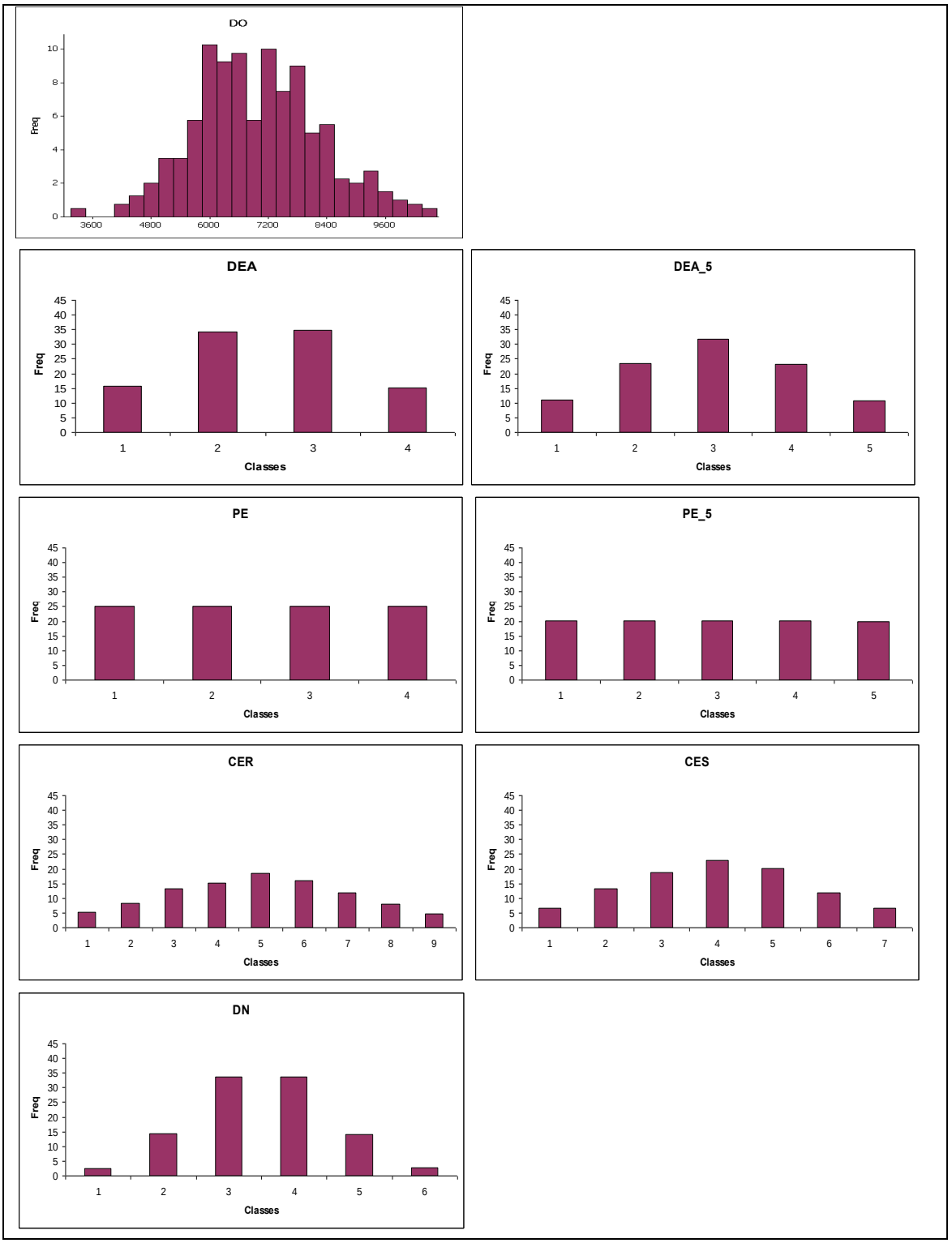
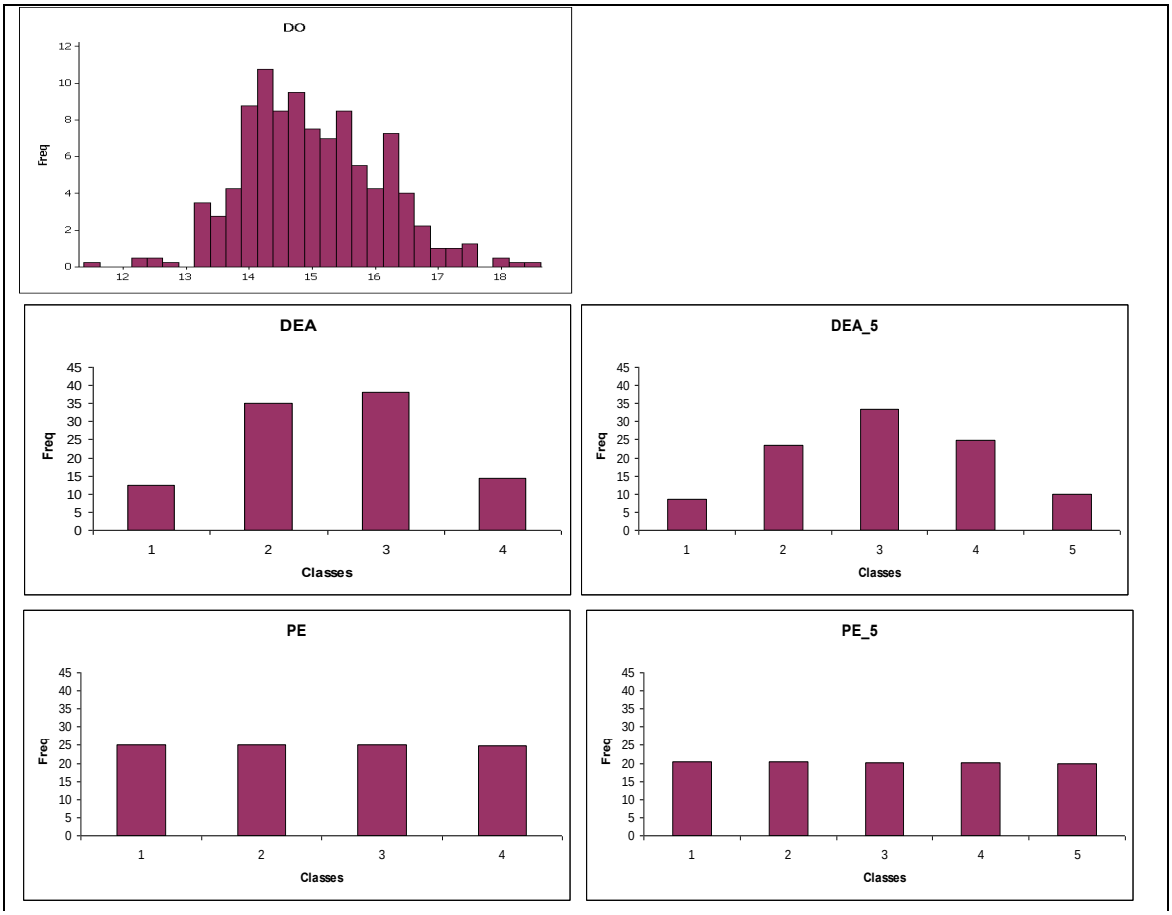


Figura A2. Dispersão dos valores da variável 3 (peso de espiga despalhada) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes

estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 4 (Comprimento de espiga)



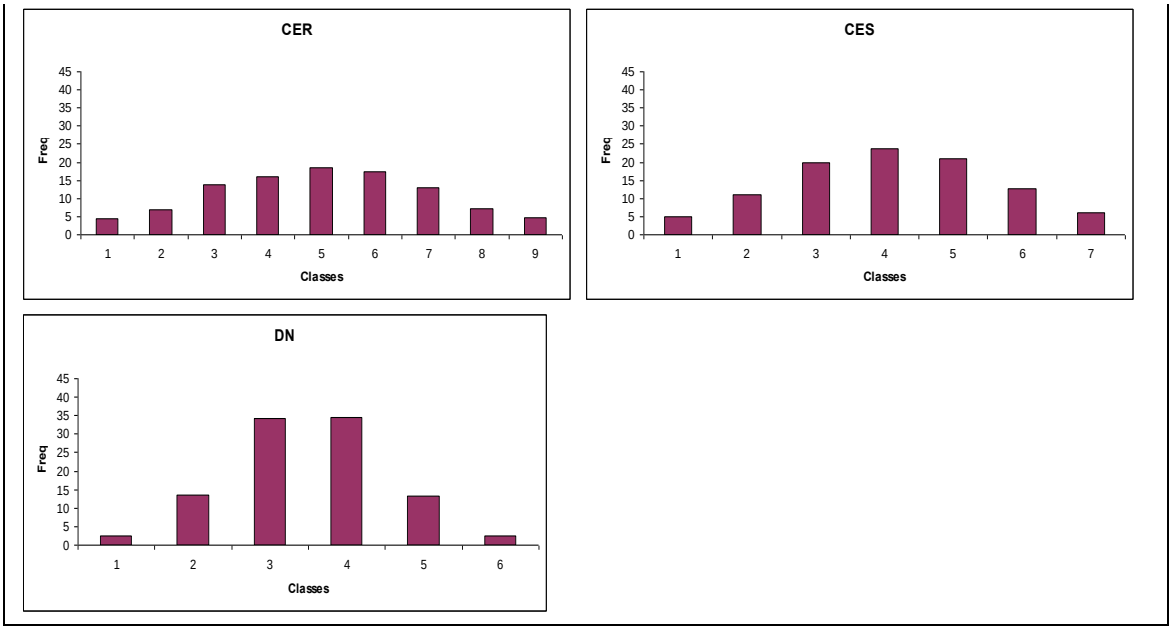
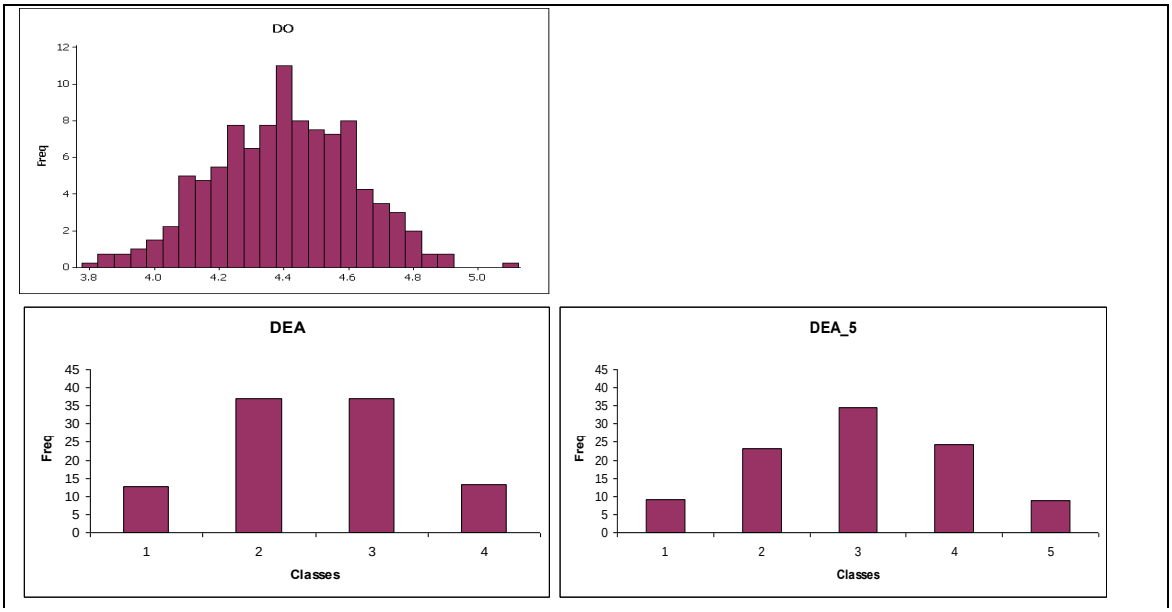


Figura A3. Dispersão dos valores da variável 4 (comprimento de espiga) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 5 (Diâmetro de espiga)



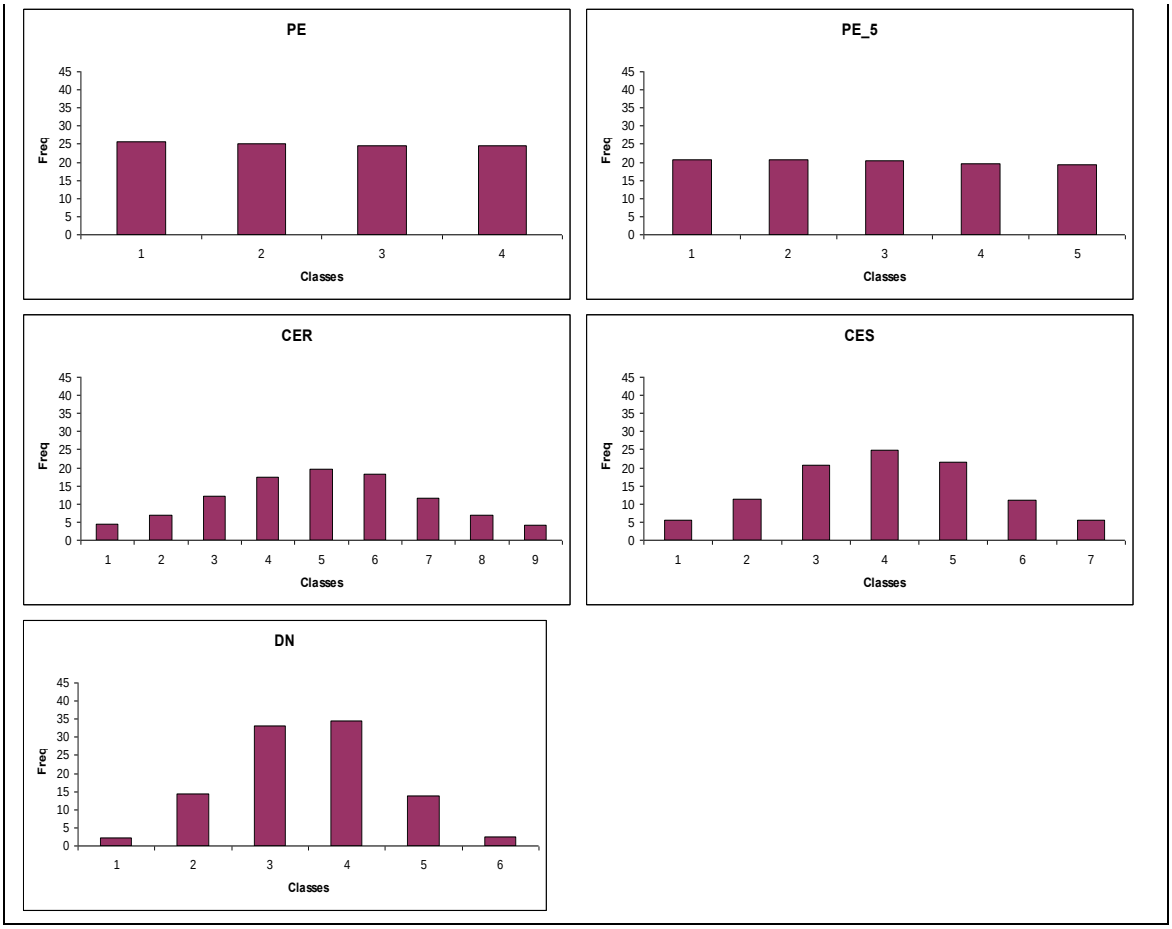


Figura A4. Dispersão dos valores da variável 5 (diâmetro de espiga) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 6 (Produção de grãos)

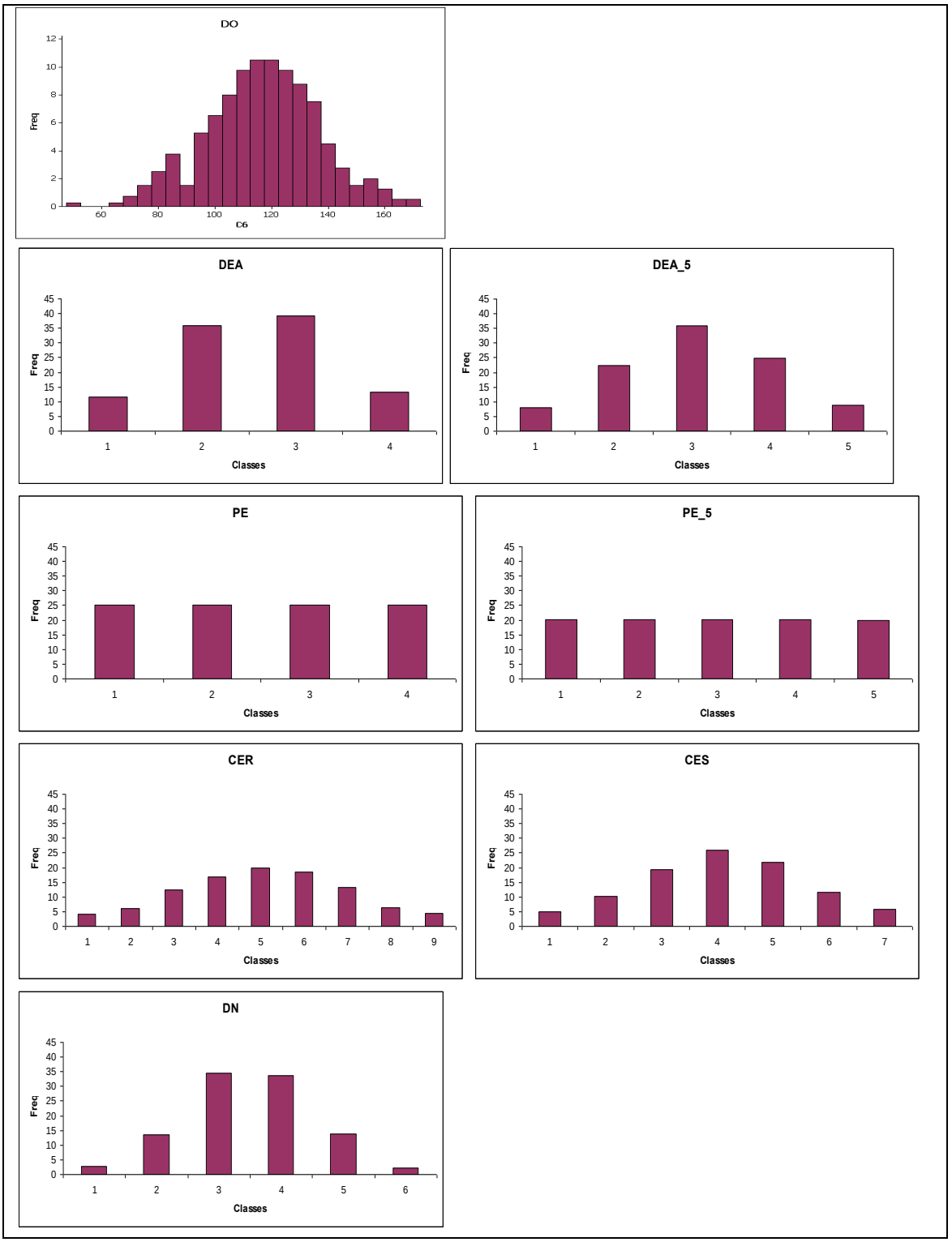


Figura A5. Dispersão dos valores da variável 6 (produção de grãos) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 7 (Prolificidade)

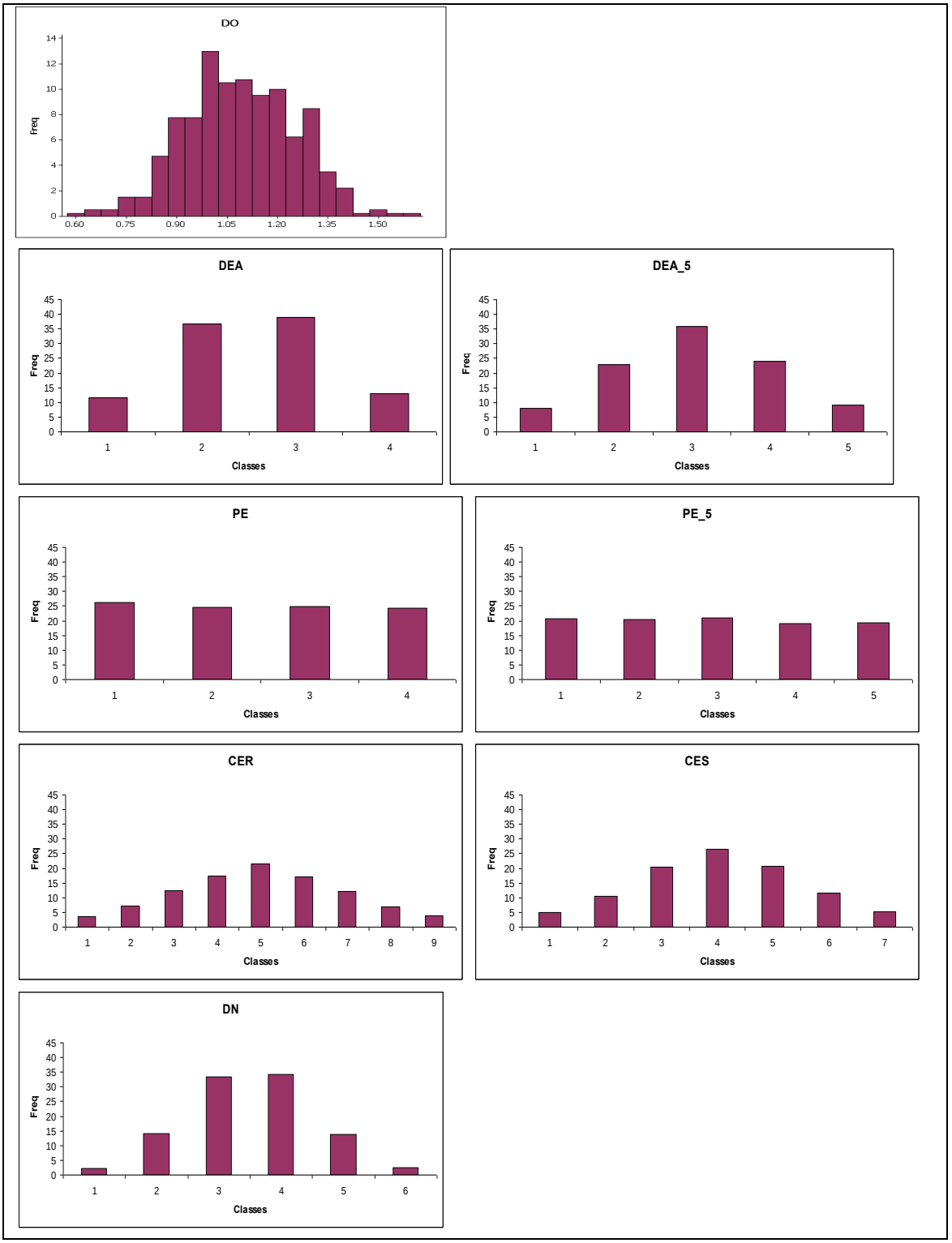
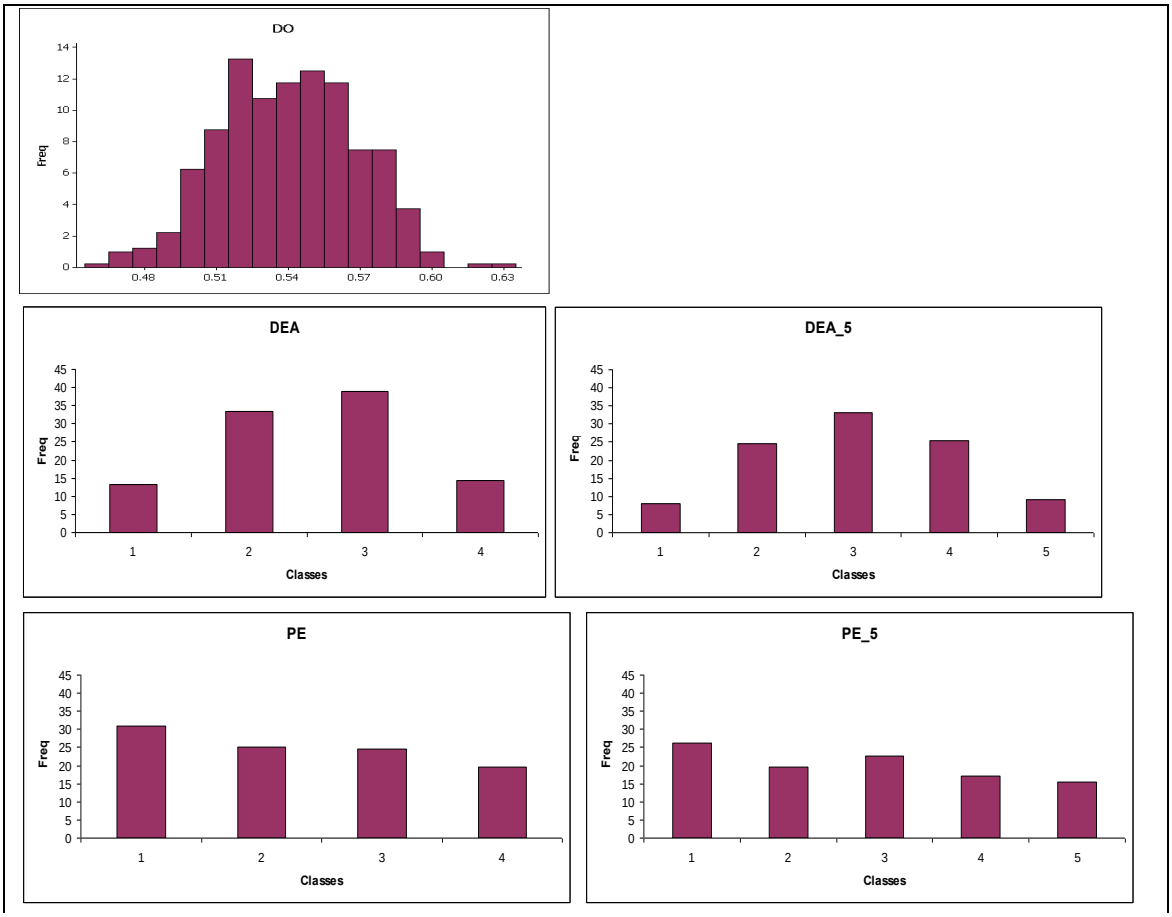


Figura A6. Dispersão dos valores da variável 7 (prolificidade) dentro das classes em cada método de transformação de dados: divisão equitativa da

amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 8 (Posição relativa da espiga)



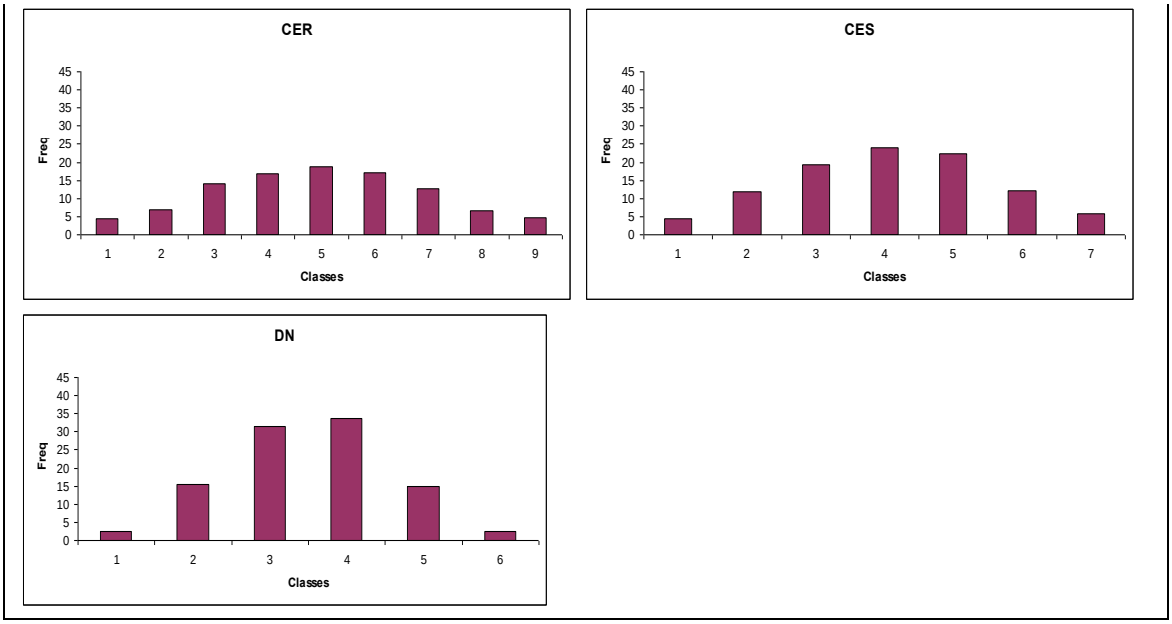
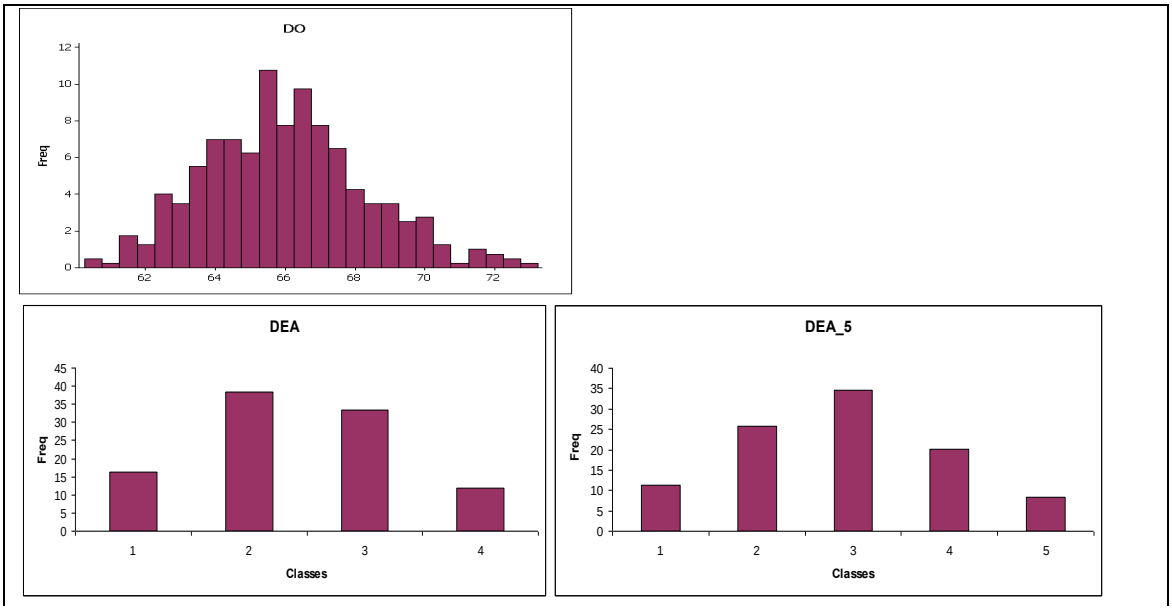


Figura A7. Dispersão dos valores da variável 8 (posição relativa da espiga) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 9 (Florescimento Feminino)



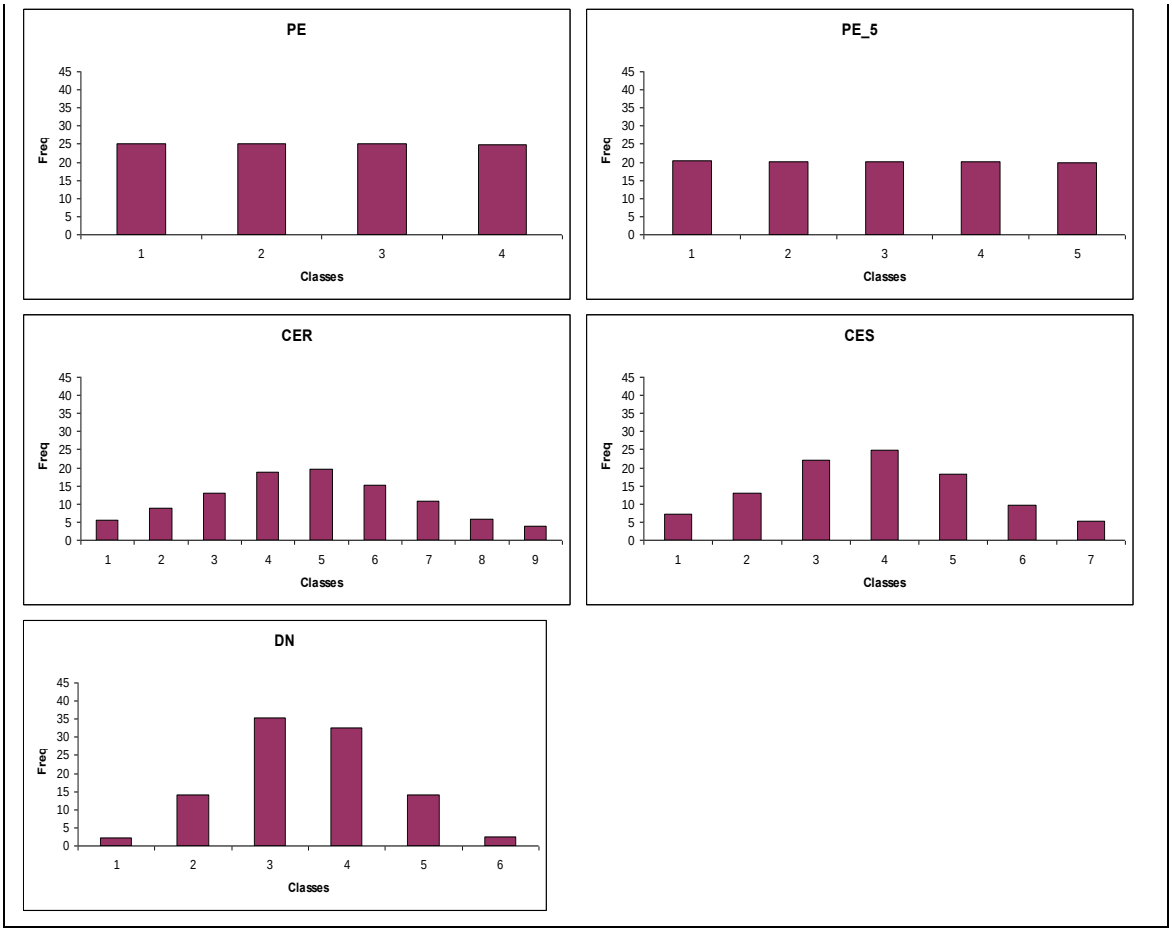


Figura A8. Dispersão dos valores da variável 9 (Florescimento feminino) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).

Variável 10 (Florescimento Masculino)

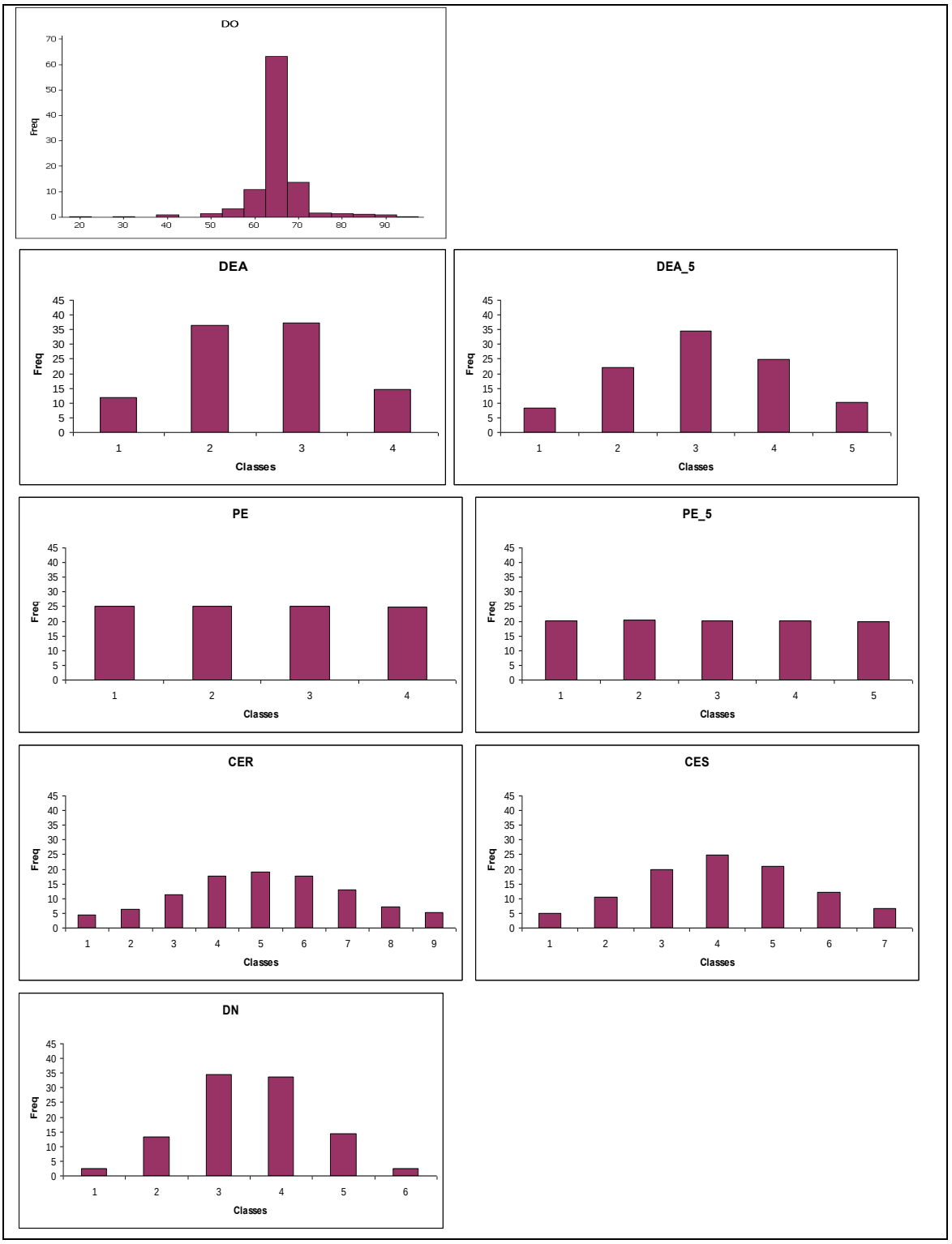


Figura A9. Dispersão dos valores da variável 10 (florescimento masculino) dentro das classes em cada método de transformação de dados: divisão equitativa da amplitude com quatro classes (DEA), divisão equitativa da amplitude com cinco classes (DEA_5), percentual equitativo com quatro classes (PE), percentual equitativo com cinco classes (PE_5), classes estimadas pela regra do Quadrado (CER), classes estimadas por Sturges (CES), classes estimadas por distribuição normal (DN).